AN ABSTRACT OF THE DISSERTATION OF

Cheyenne N. Jarman for the degree of Doctor of Philosophy in Integrative Biology presented on October 9, 2026.

Title: Scientific Machine Learning for Identifying Mechanisms of Population Dynamics and Forecasting Community State in Ecological Time Series

Abstract approved: _____

Mark Novak

Ecosystems are shaped by complex, nonlinear interactions that unfold across multiple time scales, making their governing dynamics difficult to characterize from field observations alone. Mathematical models and data-driven techniques offer complementary tools for addressing this challenge, and the emerging field of scientific machine learning (SciML), which imports a wide variety of tools primarily from engineering and the physical sciences to embed machine learning within mechanistic modeling frameworks, has shown particular promise for recovering underlying mechanisms and/or producing accurate forecasts directly from time series data. However, because these methods were developed and validated on data from their field of origin, they have typically been tested under conditions of high sampling frequency and minimal noise – conditions that ecological monitoring data rarely meet. This dissertation assesses whether, and under what conditions, two SciML

methods (symbolic regression and Universal Differential Equations) can be reliably applied to the sparse, noisy, and irregularly sampled data that characterize real ecological systems, using a combination of synthetic data experiments, an empirical case study, and dynamical modeling.

In Chapter 2, I evaluated the utility of symbolic regression using synthetically generated time series that varied in sampling density, cycle asymmetry, process noise, and the inclusion of spurious variables. I also compared two data preprocessing options for estimating population growth rates and four evaluation workflows for selecting a final equation from among the candidates. Equation and variable recovery were governed primarily by a trade-off between sampling density and process noise: symbolic regression failed below 25 sampling points per cycle, while higher process noise improved recovery rates above that threshold. Importantly, the correct equation was often present among the candidates but not consistently chosen by existing selection workflows, revealing a need for more robust criteria for identifying the correct equation among competing candidates.

In Chapter 3, I applied symbolic regression combined with time-delay embedding to long-term monitoring data from a persistent giant kelp (*Macrocystis pyrifera*) forest, testing the method's ability to recover credible mechanisms of per capita population growth rate. Across two candidate variable sets (one deliberately including a spurious variable), adult kelp abundance at a six-month lag emerged as the dominant, consistently negative predictor of growth rate, while the spurious variable was reliably excluded. However, the equations that best reconstructed the empirical growth rate-abundance relationship were not the same equations

that produced the most accurate forward-simulation forecasts, demonstrating that attractor reconstruction and forecasting skill are distinct, non-interchangeable criteria.

In Chapter 4, I assessed the reliability of universal dynamic equations (UDEs), which embed trainable neural networks within mechanistic differential equation models, for both recovering unknown species interactions and forecasting future system states, using synthetic data generated from a coral reef model under varying sampling density, initial conditions, and noise levels. I looked at two different UDE formulations, one where all model parameters except for the unknown term were fixed to their true value, and one where all model parameters were estimated. Internal validation (forecasting an unseen portion of the training trajectory) was consistently strong, but interaction-term recovery succeeded for only the fixed-parameter formulation of the UDE, and even then only under a narrow subset of conditions. Both term recovery and forecasting on novel trajectories were highly sensitive to a trajectory's initial distance to the boundary separating alternative ecosystem states, and while models often captured the correct qualitative outcome for unseen trajectories, none achieved fine-grained quantitative accuracy – underscoring the need to stress-test UDEs before their outputs inform ecological forecasting or management decisions.

Together, these chapters combine synthetic data experiments, long-term empirical monitoring data, and dynamical modeling to systematically assess where SciML succeeds and fails under realistic ecological data conditions. This dissertation demonstrates that SciML can recover biologically meaningful structure from

ecological time series, but its performance is highly sensitive to sampling density, noise, and the specific evaluation criterion used. The results throughout this dissertation point to the same conclusion from different angles: SciML methods are only as good as the data underlying them. Rather than reducing the need for sustained field observation, the rise of these tools raises the stakes for it – long-term monitoring data remain the foundation without which no amount of methodological sophistication can substitute.

©Copyright by Cheyenne N. Jarman
October 9, 2026
All Rights Reserved

Scientific Machine Learning for Identifying Mechanisms of Population Dynamics and Forecasting Community State in Ecological Time Series

by

Cheyenne N. Jarman

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented October 9, 2026
Commencement June 2027

Doctor of Philosophy dissertation of Cheyenne N. Jarman presented on October 9, 2026.

APPROVED:

Major Professor, representing Integrative Biology

Head of the Department of Integrative Biology

Dean of the Graduate Education

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Cheyenne N. Jarman, Author

ACKNOWLEDGEMENTS

Nothing in science is done alone. We all have a community behind us, and we should not forget that. Many people supported me through this dissertation. Starting with my dissertation committee: Rebecca Hutchinson, Taal Levi, Dave Lytle, and Bruce Menge. Thank you all for your investment in me as a scientist and for sticking with me as the dissertation took on different forms. Your probing questions, thoughtful feedback, and genuine excitement for science have helped me build the confidence and skills to see myself as a scientist. In particular, I am grateful to Bruce for allowing me to join in on Lubchenco-Menge Lab meetings and the opportunities to join them in the field. Bruce, your observations and nimbleness in the field are something I aspire to. And to Taal for being a collaborator on my second chapter.

My advisor, Mark Novak, gets his own paragraph because I simply cannot summarize my gratitude in just a couple of sentences, even if they do run on for far too long (it seems to be an academic inherited trait). Mark, thank you for supporting me on this journey. I knew moving to a new state and starting a Ph.D. in 2020 was not going to be an easy task, but my gut said this was the right choice, and year after year that feeling has only grown stronger. You have taught me how to think carefully, to develop mathematical intuition, and to write with clarity, especially about complex ideas. Thank you for letting me freely explore my own interests (even when they ventured far from your area of expertise), while still always being there as a safety net in case my leaps were too big. Being a good

scientist and a good mentor is no small feat, and I truly appreciate the care you bring to each.

I would also like to thank James Watson for allowing me to join in on SEA Lab meetings, introducing me to wonderful, career-expanding opportunities, and inviting me to collaborate with the Model Enabled Machine Learning group: Emerson Arehart, Jorge Arroyo-Esquivel, Jack Buckner, Nathan Fitzpatrick, Ariel Greiner, Lisa McManus, Zechariah Meunier, and Kunal Rathore. It has been a pleasure to be collaborators and friends, and I look forward to future work!

Thank you to the Novak SuperLab members, past and present, for your friendship and support, especially: Rod Alves, Augie Kalytiak-Davis, Ricardo Noe Reyes Grimaldo, Shannon Hennesey, Angelika Kurthen, Bryan K. Lynn, Arina Martin, Timeyin Pajiah, Isaac Julio-Pereira, Zach Randell, Andrés Pinos-Sánchez, and Leah Mupas Segui. It has been amazing to learn from and alongside such amazing scientists.

There is so much paperwork involved in science, from completing forms to booking travel. A huge thank you to the IB office staff, Melissa, Rowan, Tara, Traci, and Tresa, who helped me keep track of the many logistics involved in completing a dissertation. You all really keep the department afloat.

My research was made possible by an NSF GRFP, the Animal Behavioral Society, the International Women's Fishing Association, the Integrative Biology Research Award, the College of Science Student Travel Award, the Graduate School Scholarly Presentation Award, the Holt Marine Education grant, the Oregon Lottery Scholarship, the Gaulke Center for Marine Innovation and Technology, an

IB Summer Graduate Research Assistantship, and several IB Graduate Teaching Assistantships.

I am immensely grateful to my family and friends for their endless support and encouragement. Thank you to my mom, Kristina, for showing me the wonders of the natural world even from our urban apartment. From taking my summer camp to feed Rhonda the rhino at the Los Angeles Zoo, to breeding snakes and hand-rearing baby cockatiels, to flushing out rabbits in a vineyard with a Harris hawk on my head... how could I not fall in love with nature? Your drive and stubbornness to accomplish whatever you set your mind to were always an inspiration. To my stepdad, Rick, your support has truly meant the world, and it is no exaggeration to say that I would not be in this position without it. Thank you to Kate Sancer for your endless encouragement and for helping me edit so many bad graduate school application drafts. To the Fjordbeck family, Lisa, Keith, Katie, and Linda, thank you for welcoming me into your family and your home. A special thank you to Lisa for always having snacks to feed the bird (me) and Keith for Gossip Hour.

Thank you to the fellow friends who have been on the graduate school journey with me: Kris Bauer, Summer Blanco, Gina Cerimele, Megan Davis, and Kaitlyn Tonra. Having a community to share this unique experience with has been invaluable and extremely grounding... especially during qualifying exams. And thank you to the friends who remind me that life is about balance: Thank you to Mady Maszk for the hikes, book clubs, and art chats; to Annika Stensland for sharing early morning miles on the trails; to Kayna Robison and Lulu McMahon for the most majestic backpacking trips; and to my Puffin Posse, Aidan Bassett, Alex

Denton, Rachel Finegold, and Audrey Sherman, for always being up for adventure.

And to the many pets that kept me company throughout graduate school and the dissertation writing: Sherman and Lily, Charley, Emma and Jackie, Juicy and Wowie, and of course my own two menaces, Gonzo and Fran.

Last, but certainly not least, to my partner Erik Fjordbeck. Erik, it feels fitting to be writing this as we celebrate our sixth anniversary. You have been by my side for this whole process. From me watching Zoom lectures in your studio to being my assistant in the tank farm to meeting me in Paris. Thank you for always being up for my shenanigans, for being calm, loving, and level-headed, and for your support through the highs and lows. An honorary degree is definitely in order.

CONTRIBUTION OF AUTHORS

Chapter 2:

C.N. Jarman - Conceptualization, Investigation, Formal Analyses, Visualization, Writing - Original Draft

T. Levi - Conceptualization, Writing - Review & Editing

M. Novak - Conceptualization, Writing - Review & Editing

Chapter 3:

C.N. Jarman - Conceptualization, Investigation, Formal Analyses, Visualization, Writing - Original Draft

M. Novak - Conceptualization, Writing - Review & Editing

Chapter 4:

C.N. Jarman - Conceptualization, Investigation, Formal Analyses, Visualization, Writing - Original Draft

J.H. Buckner - Conceptualization, Writing - Review & Editing

A. Greiner - Conceptualization, Writing - Review & Editing

Z.D. Meunier - Writing - Review & Editing

L.C. McManus - Writing - Review & Editing

J.R. Watson - Writing - Review & Editing

E. Arehart - Conceptualization, Investigation, Writing - Review & Editing

Declaration of AI and AI-assisted technologies: Claude Sonnet (Anthropic) and Google Gemini were used to debug code for Chapters 3 and 4.

TABLE OF CONTENTS

Page

1 Introduction 1
1.1 The role of mathematical models in ecology 1
1.2 The value of long-term monitoring data and some inherent limitations 4
1.3 Data-driven approaches and the emergent field of scientific machine learning 6
1.4 Assessing SciML reliability under realistic ecological data conditions . 8
1.5 Dissertation overview 9

2 Symbolic regression for empirically realistic population dynamic time series 17
2.1 Introduction 18
2.2 Methods 22
2.2.1 The generative model 22
2.2.2 Case studies 24
2.2.3 Symbolic regression 28
2.3 Results 33
2.4 Discussion 38
2.4.1 Equation generation versus equation selection 42
2.4.2 Data informativeness 44
2.4.3 Cycle asymmetry and data preprocessing 45
2.4.4 Variable identification 46
2.4.5 Conclusions 47
2.5 Code and Data Availability 48
2.6 Acknowledgments 48
A Symbolic regression and PySR 50
B Parameters used to simulate time series 54
C Time series period lengths and aliasing 55
D Asymptotic equivalence of discrete and continuous per capita growth rates. 56
E Equation recovery when using only $A(t)$ and $A(t-2)$ 57
F Attractor reconstruction 58

TABLE OF CONTENTS (Continued)

Page

3 Identifying biotic drivers and forecasting the population dynamics of persistent giant kelp (*Macrocystis pyrifera*) . . . 64
3.1 Introduction . . . 65
3.2 Methods . . . 69
3.2.1 Long-term monitoring at San Nicolas Island, California, USA 69
3.2.2 Implementing symbolic regression . . . 72
3.2.3 Evaluation Workflow . . . 73
3.3 Results . . . 76
3.3.1 Symbolic regression output . . . 76
3.3.2 Evaluation workflow . . . 79
3.4 Discussion . . . 92
A Pareto frontier for with stars run . . . 98

4 Recovering and forecasting simulated coral reef dynamics with universal dynamic equations . . . 106
4.1 Introduction . . . 107
4.2 Methods . . . 113
4.2.1 Data generation . . . 114
4.2.2 UDE models . . . 121
4.2.3 Model fitting . . . 123
4.2.4 Evaluation . . . 124
4.3 Results . . . 126
4.3.1 H1: Process noise, observation error, and sampling frequency 127
4.3.2 H2: Transient length . . . 131
4.3.3 H3: UDE model formulation . . . 134
4.3.4 Qualitative external validation . . . 136
4.4 Discussion . . . 136
4.4.1 The role of noise as an informative signal depends on its interaction with transient dynamics . . . 138
4.4.2 Transient length shapes term recovery and forecasting . . . 140
4.4.3 UDE formulation involves a trade-off between term recovery and forecasting . . . 142

TABLE OF CONTENTS (Continued)

Page

4.4.4 The difference between qualitative and quantitative forecast accuracy . . . 143
4.4.5 Limitations and future directions . . . 145
4.5 Conclusion . . . 147
A Parameter Values Used for Analyses . . . 150
B Sensitivity Analysis for the rRMSE Recovery Threshold . . . 150
B.1 Constructing grazing functions of known error . . . 150
B.2 Forward simulation and equilibrium classification . . . 151
B.3 Sensitivity varies with distance from the basin boundary . . . 151
B.4 Selecting the recovery threshold . . . 152

5 Conclusion and Future Directions . . . 165

Appendices . . . 171
A Symbolic regression on a state-variable-reliant equation . . . 172

LIST OF FIGURES

Figure Page

2.1 Time series illustrating our six case studies and the consequences of varying sampling density within each. A total of 15 cycles were used in our analyses, but only two are plotted for visualization. The light gray lines show the dynamics of the instantaneous deterministic Bence-Nisbet model for reference. The overlaid colored points show the (downsampled and interpolated) data from which the per capita growth rates and (lagged) putative predictor variables were obtained and provided to symbolic regression. Per capita growth rates were calculated using a discrete- or continuous-time preprocessing approach (see main text). 25

2.2 (a) Time series data were analyzed using either a discrete-time approach (orange line, difference in population sizes) or a continuous-time approach (green line, tangent to interpolated population sizes) to calculate the response variable, the per capita growth rate (pcgr), that we provided to symbolic regression. (b) The additional inputs provided to symbolic regression were the suites of predictor variables ((lagged) population sizes) and mathematical operators (including the custom function $[x]_+$). (c) In PySR, equation evolution occurs in two stages, (*i*)"within population evolution" wherein multiple equations exist within a single generation and parental equations are selected based on their fitness to create new equations in subsequent generations, and (*ii*)"between population evolution" wherein equation components can migrate between concurrent populations. (d) The output at the end of a run entails a list of the best (lowest MSE) equation for each level of complexity, along with its MSE values and complexity level. (e) The Pareto frontier of equations may be examined to select the equation affecting the largest drop in MSE (represented by the orange lines) as done in the first of the four workflows that we evaluated. 29

LIST OF FIGURES (Continued)

Figure | Page

2.3 The success rate of symbolic regression and the four model selection workflows in selecting equations that contained the correct variables or that contained the correct variables and matched the formulation of the Bence-Nisbet model with which we generated the data. Confidence intervals reflect binomial uncertainties given a sample size of 100 independent runs. The black bars indicate the proportion of the 100 independent runs in which the Bence-Nisbet model occurred, regardless of whether a workflow selected it or whether it occurred in the top 10 equations. 36

2.4 The proportional frequency of variable combinations present among the up to top 10 (lowest MSE) equations of each complexity level in the range of 10–18 (combining all 100 runs of each case-study sampling-density combination). The Bence-Nisbet model with which the data were generated has a complexity of 14 and contains only $A(t)$ and $A(t-2)$, making the inclusion of $A(t-1)$ and $A(t-3)$ spurious. All possible variable combinations are depicted in the legend; however, the combination of $A(t)$ and $A(t-3)$ never occurred. . . 39

2.5 The proportion of the top 10 (lowest MSE) equations that correctly matched the Bence-Nisbet model in each complexity level in the range 10–18 (combining all 100 runs of each case-study sampling-density combination). Lightgray backgrounds indicate instances in which at least one equation occurred that contained only the two variables of the Bence-Nisbet model, $A(t)$ and $A(t-2)$, as shown in Fig. 2.4. The complexity level of the Bence-Nisbet model is 14. Matches at complexities greater than 14 occurred when the equations could be algebraically simplified to the Bence-Nisbet model. Note that matches at odd-valued complexity levels are not possible given the even-valued complexity of the Bence-Nisbet model and the necessity of including an operator for every additional parameter or variable. 40

LIST OF FIGURES (Continued)

Figure | Page

2.6 The maximum MSE-rank of the Bence-Nisbet model when it was recovered, reflecting the number of equations that exhibited a lower MSE than it in the set of up to 100 equations returned by the 100 runs of each case-study sampling-density combination. The complexity level of the Bence-Nisbet model is 14 (indicated by asterisks), hence lower levels are not possible. Ranks at complexities greater than 14 (indicated by points) occurred when the equations could be algebraically simplified to the Bence-Nisbet model. The Bence-Nisbet model was never recovered for time series of 10 and 5 points per cycle or for the deterministic discrete-time processed case study at 25 points per cycle (see Fig. 2.5). 41

S.1 The maximum MSE-rank of the Bence-Nisbet model when it was recovered, returned by the 100 runs from the asymmetric, 100 points per cycle data, with a continuous-time and discrete-time approach, respectively. We assessed the complexity range from 10 to 20. The lowest complexity that the Bence-Nisbet model occurred was complexity 14, which is the complexity level of the Bence-Nisbet model is 14 (indicated by asterisks). Ranks at complexities greater than 14 (indicated by points) occurred when the equations could be algebraically simplified to the Bence-Nisbet model. 57

S.2 Phase portraits of population size and per capita growth rate (log ratio or interpolated for discrete-time or continuous-time, respectively) provide insight into the influence of sampling density on symbolic regression's success. As sampling density decreases, the apparent attractor tends to shrink relative to the true attractor of the Bence-Nisbet model (overlaid in light gray), except for the cases with process noise. 58

3.1 Giant kelp abundance over time from A) West Dutch Harbor and B) East Dutch Harbor, San Nicolas Island, CA. The lines connecting the points are to assist in readability and are not intended to be read as interpolation. 70

LIST OF FIGURES (Continued)

Figure Page

3.2 Retention percentage of the variables included in the top-10 lowest-loss equations across complexities 1–30, for the no stars run (blue) and with stars run (red). Variables included adult kelp A, juvenile kelp J, sea urchins U (and, for the with stars run, bat stars, S), each at the current value and four temporal lags (6, 12, 18, and 24 months). Both runs had 280 equations each (10 per complexity level across complexities 1–30, excluding complexity levels with no returned equations). 78

3.3 Structural diversity of the top-10 lowest-loss equations at each complexity level for the A) no stars and B) with stars runs. At each complexity level, the 10 equations were grouped by shared functional form, and stacked bar segments show the number of equations belonging to each group, ranked by group size from largest (Rank 1, dark blue) to smallest (Rank 10, red). Ranks are not conserved across complexity levels. A single solid-colored bar indicates that all 10 equations share one functional form at that complexity level. Bars divided into multiple colors indicate multiple distinct forms. Complexity levels 2 and 5 are omitted because PySR did not return equations at those complexity levels. 80

3.4 Pareto frontier for the no stars PySR run, showing the best-performing equation's loss (MSE) at each complexity level across 100 independent searches. The red markers indicate the eight complexity levels selected for further evaluation, at which increasing complexity produced a relatively large drop in loss relative to the preceding step. The equations at these complexity levels were carried forward to Evaluation Steps 1-3. The with stars run produced a near-identical frontier and selected the same eight complexity levels (see Supplementary Materials). 82

3.5 Predicted versus observed per capita growth rate (pcgr) for the selected equations, using the no stars run's coefficients except where noted. The complexity 26 panels show both the no stars and with stars equations, since this was the one selected complexity where the equation structure differed between the two runs. The dashed red line indicates the 1:1 line of perfect prediction. 87

LIST OF FIGURES (Continued)

Figure | Page

3.6 Residuals versus fitted values for the selected equations, using the no stars run's coefficients except where noted. The dashed red line marks zero residual. 88

3.7 Population size forecasting for six equations × 4 randomly selected transects (24 combinations total) to represent the range of forward-simulation performance observed: complexity 13 and 17 at West Dutch transect 10R, and East Dutch transect 22R, complexity 20 at West Dutch transect 10R, and complexity 26's no stars and with stars equations at West Dutch transect 10R. Red points show the observed time series; the dashed line shows the equation's forward-simulated trajectory beginning at the point the solver takes over from the empirical data. R^2, RMSE, and NRMSE are reported for each panel. 91

S.1 Pareto frontier for the with stars PySR run, showing the best-performing equation's loss (MSE) at each complexity levels across 100 independent runs. The red stars indicate equations where increasing complexity level produced a relatively large drop in loss relative to the preceding step. These were the complexities we looked at for our further analyses. The no stars run produced a near-identical frontier and selected the same eight complexities (see Pareto frontier in the main text). 99

4.1 Example deterministic time series across all four coral initial condition values and all four sampling frequencies. Rows show initial coral cover (Coral IC), and columns show sampling frequency (every 6 months, every 1 year, every 2 years, and every 4 years, corresponding to $\delta t = 0.5, 1, 2$, and 4, respectively). Fractional coverage represents the proportion of benthic reef substrate occupied by each cover type. Points show coral (orange points) and macroalgae (green triangles) cover over time. 117

Figure Page

4.2 Example time series illustrating the stochastic data-generating processes evaluated in this study, all initialized at a coral initial condition value of 0.44 and sampled every 1 year ($\delta t = 1$). Top row: process noise on the grazing term at low, medium, and high magnitude. Bottom row: observation error at low, medium, and high magnitude. Fractional coverage represents the proportion of benthic reef substrate occupied by each cover type. Points show coral (orange points) and macroalgae (green triangles) cover over time. . 120

4.3 Term recovery, internal validation, and external validation performance across process noise for both UDE formulations. Columns show term recovery rRMSE (%; left), internal validation MAE (%; center), and external validation MAE (%; right), all on a log scale. Point color indicates sampling frequency (every 6 months, every 1 year, every 2 years, every 4 years); point shape indicates initial coral cover. Rows indicate model formulation (CE-UDE, FP-UDE). Dashed lines mark the 10% term recovery threshold and the 5% forecast MAE threshold. Process noise did not improve term recovery or forecasting for either UDE formulation, and external validation MAE never fell below the threshold under any condition. 128

4.4 Term recovery, internal validation, and external validation performance across observation error for both UDE formulations. Columns show term recovery rRMSE (%; left), internal validation MAE (%; center), and external validation MAE (%; right), all on a log scale. Point color indicates sampling frequency (every 6 months, every 1 year, every 2 years, every 4 years); point shape indicates initial coral cover. Rows indicate model formulation (CE-UDE, FP-UDE). Dashed lines mark the 10% term recovery threshold and the 5% forecast MAE threshold. Observation error degraded forecasting for both formulations and term recovery for FP-UDE, as its magnitude increased, while external validation MAE never fell below the threshold under any condition. 129

LIST OF FIGURES (Continued)

Figure Page

4.5 Term recovery, internal validation, and external validation performance for short-transient ($C(0) = 0.1, 0.8$) versus long-transient ($C(0) = 0.42, 0.44$) initial conditions under deterministic training data, for both UDE formulations. Columns show term recovery rRMSE (%; left), internal validation MAE (%; center), and external validation MAE (%; right), all on a log scale. Point color indicates sampling frequency (every 6 months, every 1 year, every 2 years, every 4 years); point shape indicates initial coral cover. Rows indicate model formulation (CE-UDE, FP-UDE). Dashed lines mark the 10% term recovery threshold and the 5% forecast MAE threshold. Longer transients improved term recovery and external validation, but modestly worsened internal validation forecast error, though nearly all internal validation forecasts remained within threshold regardless of transient length. 132

4.6 Comparison of the two UDE formulations across all 112 conditions tested per formulation, shown as the percentage of runs meeting versus not meeting each success threshold (solid vs. faded fill). The first column is term recovery (rRMSE $< 10\%$). The second is internal validation forecasting error (MAE $< 5\%$), and the third is external validation forecasting error (MAE $< 5\%$). FP-UDE achieved better term recovery than CE-UDE, while CE-UDE achieved modestly better internal validation forecast success; neither formulation met the external validation threshold in any run. 135

4.7 Distribution of external validation state-match rates (the fraction of the four external benchmark trajectories correctly classified) across all 112 conditions tested per formulation. 137

S.1 Phase-plane trajectories under the true grazing term (gray) and under perturbed grazing functions of increasing target rRMSE (dashed, colored from purple at 1% to green at 200%), for each of the four initial conditions. Stable equilibria (coral-dominated and macroalgae-dominated) are shown as black circles; the interior saddle point is shown as a dark red X. For IC = 0.1 and IC = 0.8, low-error trajectories track the true trajectory closely before diverging at higher rRMSE; for IC = 0.42 and IC = 0.44, even modestly perturbed trajectories diverge from the true trajectory near the saddle point, consistent with their proximity to the basin boundary. 154

LIST OF FIGURES (Continued)

Figure | Page

S.2 Comparison of coral and macroalgae nullclines under the true grazing term (solid, faint) and under a grazing function with 50% rRMSE (dashed), for each of the four initial conditions. The shift in the algae nullcline under 50% error moves the effective saddle point, redirecting the 50% rRMSE trajectory (purple, dotted) into the coral-dominated basin even where the true trajectory (gray) converges to the macroalgae-dominated equilibrium. 154

LIST OF TABLES

Table — Page

S.1 Parameter values used to generate time series. 54

S.2 Parameter values used to introduce process noise. 54

3.1 Comparison of the selected equations from no stars and with stars PySR runs. All equations use the same variables between the no star and with star runs, and all but complexity 26 are the same structural form. Coefficient values are rounded to the nearest 0.01; symbols with identical rounded values are reported once. The letters e, i, and t were excluded from coefficient renaming to avoid confusion with the exponential constant, the imaginary unit, and the time variable, respectively. 84

3.2 R^2, RMSE, and NRMSE for all six selected equations across all four transects assessed for population size forecasting. Metrics were calculated only over the forward-simulated period following the solver handoff; observations used to construct the history function seeding each equation's initial condition were excluded from these calculations. 90

4.1 Summary of hypotheses, predictions, and outcomes across term recovery, internal validation, and external validation. CE-UDE: coestimated-parameter formulation; FP-UDE: fixed-parameter formulation. 148

S.1 The parameter values used in the Mumby et al. model. Values come from Fung et al. (2011) and Baskett et al. (2014). Values were calibrated as annual rates. 150

S.2 First target rRMSE at which the simulated trajectory under a perturbed grazing function converged to the wrong equilibrium (i.e., a different basin of attraction than the true trajectory), by initial condition, alongside each initial condition's approximate distance from the separatrix (evaluated along its fixed coral cover value). 153

S.3 Summary of grazing-term recovery (term recovery rRMSE, %) for the two UDE formulations split by noise type and level. Values are median [25th – 75th percentile], followed by the raw range (min–max). N gives the number of runs (four initial conditions × four sampling densities) underlying each row. 155

LIST OF TABLES (Continued)

Table | Page

S.4 Summary of grazing-term recovery under extrapolation (extrapolation rRMSE, %, evaluated over the full biologically feasible grid rather than only the training trajectory, for the two UDE formulations split by noise type and level. Values are median [25th–75th percentile], followed by the raw range (min–max). N gives the number of runs underlying each row. 156

S.5 Summary of internal validation forecast error (coral-macroalgae-averaged MAE, % absolute cover) for the two UDE formulations split by noise type and level. N gives the number of runs underlying each row. 157

S.6 Summary of external validation forecast error (coral-macroalgae-averaged MAE, % absolute cover, averaged across all four external-validation benchmark trajectories) for the two UDE formulations split by noise type and level. N gives the number of runs underlying each row. 158

S.7 External validation state-match rates (%; the proportion of the four external-validation benchmark trajectories for which the forecasted final state matched the true state) by initial condition and sampling frequency, under deterministic training data ($N = 1$ per cell). Initial conditions $C(0) = 0.1$ and 0.8 are short-transient trajectories; $C(0) = 0.42$ and 0.44 are long-transient trajectories 159

LIST OF APPENDIX FIGURES

Figure Page

A.1 Pareto frontier for the four datasets examined when PySR included all variables (A, J, G, F) and the respective lags. The top-performing equation from each complexity is plotted, and the star indicates the equation selected based on the trade-off between producing the largest drop in MSE for the lowest complexity. 176

A.2 Observed vs. predicted per capita growth rate (pcgr) from the PySR searches that used all explanatory variables and the associated lags. The predicted per capita growth rate value comes from the starred equations from the Pareto frontier (Fig. A.1). The red line indicates the one-to-one line. 177

LIST OF APPENDIX TABLES

Table Page

A.1 Comparison of starred equations from the all-variable PySR searches. The true equation was $\frac{1}{A}\frac{dA}{dt} = \frac{0.7J}{A} - 1$. Coefficients were rounded to two decimal places. 178

For my mother, *Kristina*.

Chapter 1: Introduction

1.1 The role of mathematical models in ecology

Theoretical ecology is the branch of the discipline concerned with building formal, typically mathematical, representations of ecological processes – population growth, species interactions, community assembly – in order to derive general principles that extend beyond any single observed system. Its central premise is that ecological patterns, however idiosyncratic they may appear in any one place or time, often arise from a small number of underlying mechanisms that can be captured, studied, and tested in the abstract (MacArthur, 1984; May, 1981; Pickett et al., 2013). Theory in this sense does not merely summarize what has been observed; it is theory-dependent in a stronger sense, shaping which questions ecologists ask and the data they collect, which methods they judge appropriate, and how they interpret the patterns they find (Kuhn, 1970; Layman and Rypel, 2023). Because no single model can simultaneously maximize generality, realism, and precision, ecological models have historically been built by deliberately sacrificing one of those properties to gain the other two, in service of goals (explaining, predicting, or modifying a system) that only partly overlap (Levins, 1966). This tension between abstraction and applicability has structured theoretical ecology from its outset.

Mathematical models have shaped ecological thinking since the field's earliest formal efforts to describe population change. The logistic growth equation (Verhulst, 1838), Lotka-Volterra predator-prey and competition models (Lotka, 1925; Volterra, 1926), and their many descendants, such as stage-structured population models (Leslie, 1945), gave ecologists a shared symbolic language. A language that has allowed the field to express hypotheses about density dependence (Hassell, 1975), inter- and intraspecific interaction (Gause, 1934; Connell, 1961; Post et al., 2008), and coexistence (Chesson, 2000). These frameworks did more than describe particular systems. They generated general, transferable predictions – that intraspecific competition should stabilize populations (Hassell, 1975), that certain interaction structures should permit or exclude coexistence (Saavedra et al., 2017; Song et al., 2020), that complexity and stability trade off in predictable ways (May, 1972). These predictions could be tested, debated, and extended across taxa and systems far removed from where they originated (May, 1975; O'Gorman and Emmerson, 2009; Preston et al., 2018; Adler et al., 2026). This tradition of building simplified, analytically tractable representations of ecological processes to derive general principles has given rise to modern coexistence theory, structured population models, and metapopulation theory (Levins, 1969; Caswell, 2001; Chesson, 2000).

Yet this same strength of abstraction from any single system's particulars is also the tradition's central limitation. Mathematical models in ecology are built on simplifying assumptions that rarely hold exactly in any real population or community, and the field has, over decades, generated an increasingly large and only loosely

reconciled collection of formal frameworks, each embedding different assumptions about what matters and why. Song (2026) argues that this proliferation has outpaced the field's effort to synthesize it. As such, much of what looks like scientific disagreement in community ecology is in fact semantic ambiguity: different theories, built with different formalizations and applied toward different functions, are talked past one another rather than genuinely tested against each other and appropriately retired once proven non-applicable. Song's proposed remedy is to organize theoretical approaches along two independent axes, formalization (how theory is built) and function (why it is used). This is a call to be more explicit about what any given mathematical model can and cannot tell us, and about which of its assumptions are load-bearing versus incidental.

This dissertation inherits that concern in a specific and practical form. If mathematical models are powerful because they encode explicit, testable assumptions about mechanism, then a central open question is how effectively such mechanisms can be recovered directly from the data (rather than assumed in advance by the modeler) and under what conditions that recovery is trustworthy. Before turning to that question, however, it is worth examining the other major tradition in ecology from which any data-driven approach must draw on: empirical, long-term observation.

1.2 The value of long-term monitoring data and some inherent limitations

Where mathematical theory offers generality, sustained empirical observations offer grounding. The proliferation of long-term ecological monitoring programs over the past half-century, from the U.S. Long Term Ecological Research (LTER) network to system-specific time series maintained by individual groups, has transformed what ecologists can know about population and community dynamics (Hobbie et al., 2003; Magnuson, 1990; Pauly, 1995). These datasets have revealed multi-year and multi-decadal population cycles invisible in shorter studies, documented community responses to disturbance and recovery over time scales that no single field season could capture, and provided the empirical foundation against which theoretical predictions about stability, coexistence, and regime shifts can be tested (Mduma et al., 1999; Menge et al., 2019; Scheffer et al., 1993; Kéfi et al., 2007; Kenner and Tinker, 2018).

However, working with monitoring data requires reckoning with several ways in which it departs from the idealized time series that most quantitative methods, especially those imported from other disciplines, are designed around. One major aspect is sampling density. Synthetic and laboratory-based studies can generate observations as often as a system's dynamics require. Field-based monitoring cannot: logistical, financial, and seasonal constraints impose hard limits on how often a population or community can be sampled, and those limits are rarely set by the dynamics of the system itself. When sampling density falls below what a

given signal requires, the underlying dynamics become more difficult or impossible to recover, regardless of how sophisticated the downstream analysis method is. There is also, of course, the prior problem of knowing what the system's underlying dynamics are, and therefore what sampling regime they would necessitate.

Another aspect is noise, which can take on different forms. *Process noise* refers to genuine stochasticity in the ecological process itself where the true state of the system is buffeted by unpredictable environmental or demographic fluctuations, independent of how well it is observed. *Observation noise*, by contrast, refers to error introduced in measurement – imprecision in counting, detecting, or estimating the true state. These two noise sources are frequently conflated in ecological data but have different implications for what can be recovered from a time series and how. Moreover, noise is not uniformly an obstacle to be filtered away: Boettiger (2018) highlights that noise, specifically process noise, can be informative, acting as a series of mini perturbation experiments that help reveal the underlying system. Any method applied to ecological time series must therefore contend not only with noise's magnitude, but with its type, and with the possibility that noise is doing dynamical work of its own.

These properties define the gap this dissertation is centrally concerned with: the difference between the data conditions under which many quantitative methods are developed and validated, and the data conditions that long-term ecological monitoring actually provides.

1.3 Data-driven approaches and the emergent field of scientific machine learning

Partly in response to the limits of purely theory-driven modeling described above, ecology has increasingly turned to data-driven approaches; methods that infer structure directly from observed data rather than beginning from a pre-specified mechanistic form (Song, 2026). The appeal of these methods is that they allow the data to inform the model structure with little to no *a priori* specification compared to the traditional approaches. Machine learning methods – from random forests and neural networks to more specialized time series models – have found growing use in ecology for exactly this reason (Christin et al., 2019; Pichler and Hartig, 2022), often achieving strong performance without requiring the modeler to first commit to a particular mechanistic hypothesis.

This flexibility, however, comes at a cost that mirrors the strength of the mathematical modeling tradition described in Section 1.1: standard machine learning methods are typically poor at yielding the kind of interpretable, mechanistic insight that has historically been theoretical ecology's central currency. A highly accurate but uninterpretable model can forecast a population's trajectory without revealing why it moves that way, and indeed such forecasts can serve an important role, for instance in warning a manager that a population or community may be approaching collapse. But a key part of theoretical ecology is to understand *why*, not just the *what*.

Scientific machine learning (SciML) has emerged as a hybrid response to this

trade-off (Rackauckas et al., 2021). Rather than treating mechanism and data-driven flexibility as competing approaches, SciML methods incorporate prior mechanistic knowledge into an otherwise data-driven learning process. This distinguishes SciML from standard machine learning in ecology: where conventional ML methods typically treat the entire input-output relationship as a black-box, SciML methods build in whatever aspects of the system's dynamics are at least partly understood and worth preserving, asking the data-driven component to resolve only what remains unknown.

Two such methods are the focus of this dissertation. Symbolic regression searches over a space of candidate mathematical expressions to identify equations that both fit observed data and remain human-readable, aiming to recover a governing equation directly from the focal data. Universal dynamic equations (UDEs) take a more targeted approach, embedding a trainable neural network within an otherwise mechanistic differential equation model, so that only a specific, unknown component (e.g., an unmodeled species interaction) is learned from the data while the remaining structure is fixed in advance. Both methods have shown early promise on synthetic and laboratory time series, but both were developed and benchmarked primarily using data far better behaved, in terms of sampling density and noise, than what ecological monitoring time series typically provide.

1.4 Assessing SciML reliability under realistic ecological data conditions

The gap identified in Sections 1.1 through 1.3, between the data conditions under which SciML methods were built and validated, and the data conditions ecological systems actually generate, is an untested constraint on whether these tools can reliably be adopted in ecology. Assuming that strong performance in one data regime will transfer to another without rigorous testing is a real risk. As SciML tools become easier to apply "out of the box", it becomes correspondingly easier to treat their outputs as validated scientific findings rather than as model outputs whose reliability depends entirely on how well the application context matches the conditions for which the method was built. In ecology, where independent ground truth is rarely available, and results can inform conservation or management decisions, the cost of misplaced confidence in a spurious equation, an artificial interaction term, or an incorrect forecast is huge.

Addressing this gap requires deliberately stress-testing these methods under conditions that resemble what ecological data actually offers, rather than assuming that performance demonstrated elsewhere will transfer. This dissertation begins to take up that task using two complementary strategies. First, synthetic data experiments allow direct, controlled manipulation of the specific data properties identified above against a known ground truth, making it possible to isolate exactly what conditions cause a method to succeed or fail. Second, an empirical case study applies these methods to real long-term monitoring data, leveraging

the intuition gained from synthetic data experiments together with independent domain knowledge as a partial substitute for the ground truth, even as the full complexity and confounded structure of a real ecological system remain unavoidably present. Together, these two approaches provide both the internal validity of controlled experimentation and an external, proof-of-concept demonstration of a genuine ecological application.

1.5 Dissertation overview

This dissertation addresses this question across three chapters using two SciML methods, symbolic regression and UDEs, evaluated through a combination of synthetic experiments and long-term empirical monitoring data.

In Chapter 2, I evaluate the utility of symbolic regression using synthetic time series that vary in sampling density, cycle asymmetry, process noise, and the inclusion of spurious variables. I compare two approaches to estimating population growth rates and four workflows for selecting a final equation among candidates. I find that equation and variable recovery are governed primarily by a trade-off between sampling density and process noise, with symbolic regression failing outright below 25 sampling points per cycle. Importantly, even when the correct equation was recovered among the generated candidates, existing post-hoc selection workflows did not consistently identify it, pointing to a need for more robust criteria for selecting the correct equation among many potential candidates.

In Chapter 3, I apply symbolic regression combined with time-delay embedding

to long-term monitoring data from a persistent giant kelp (*Macrocystis pyrifera*) forest, testing the method's ability to recover credible mechanisms of kelp per capita growth rate, and whether those recovered equations could in turn forecast population dynamics. Across two candidate variable sets, one deliberately including a spurious variable, adult kelp abundance at a six-month lag emerged as the dominant, consistently negative predictor of growth rate, while the spurious variable was reliably excluded, supporting the method's ability to distinguish real from spurious drivers even in noisy field data. However, the equations that best reconstructed the empirical growth rate-abundance relationship were not the same equations that produced the most accurate population size forecasts, demonstrating that attractor reconstruction and forecasting skill are distinct, non-interchangeable criteria for evaluating a candidate equation.

In Chapter 4, I assess the reliability of UDEs for both recovering an unknown species interaction and forecasting future ecosystem states, using synthetic data generated from a coral reef model under varying sampling density, initial conditions, and noise levels. Internal validation, which assesses forecasting success on an unseen portion of the training trajectory, was consistently strong, but interaction-term recovery succeeded for only one of two UDE formulations tested, and even then only under a narrow subset of conditions. Both term recovery and forecasting on novel trajectories were highly sensitive to a trajectory's initial distance to the boundary separating alternative ecosystem states. While models often captured the correct qualitative outcome for unseen trajectories, none achieved fine-grained quantitative accuracy.

Together, these three chapters use synthetic experiments, long-term monitoring data, and dynamical modeling to systematically assess where SciML succeeds and where it fails under realistic ecological conditions. These results converge on a consistent conclusion from three different angles: these methods are only as trustworthy as the data and the evaluation criteria used to assess them. This dissertation also underscores the continued, and even growing, need for sustained empirical observation, which is the basis for data-driven methods, and for empiricists whose natural history intuition helps validate, interpret, and inform the models these methods produce.

Bibliography

Adler, P. B., Detto, M., Ellner, S. P., Gibbs, T. L., Gold, Z. J., Leonard, S. J., Levine, J. I., Schiffer, A. E., Song, C., Stemkovski, M., Vahsen, M. L., and Levine, J. M. (2026). What Have We Learned From Empirical Applications of Modern Coexistence Theory? *Ecology Letters*, 29(6):e70404.

Boettiger, C. (2018). From noise to knowledge: how randomness generates novel phenomena and reveals information. *Ecology Letters*, 21(8):1255–1267.

Caswell, H. (2001). Matrix population models: construction, analysis, and interpretation. In *Matrix population models: construction, analysis, and interpretation*, pages 722–722.

Chesson, P. (2000). Mechanisms of Maintenance of Species Diversity. *Annual*

Review of Ecology and Systematics, 31(1):343–366.

Christin, S., Hervet, É., and Lecomte, N. (2019). Applications for deep learning in ecology. *Methods in Ecology and Evolution*, 10(10):1632–1644.

Connell, J. H. (1961). The Influence of Interspecific Competition and Other Factors on the Distribution of the Barnacle Chthamalus Stellatus. *Ecology*, 42(4):710–723.

Gause, G. F. (1934). *The Struggle for Existence*.

Hassell, M. P. (1975). Density-Dependence in Single-Species Populations. *Journal of Animal Ecology*, 44(1):283–295.

Hobbie, J. E., Carpenter, S. R., Grimm, N. B., Gosz, J. R., and Seastedt, T. R. (2003). The US Long Term Ecological Research Program. *BioScience*, 53(1):21.

Kenner, M. C. and Tinker, M. T. (2018). Stability and Change in Kelp Forest Habitats at San Nicolas Island. *Western North American Naturalist*, 78(4):633.

Kuhn, T. S. (1970). *The structure of scientific revolutions*. International encyclopedia of unified science. Foundations of the unity of science, v. 2, no. 2. University of Chicago Press, Chicago, [2d ed., enl edition.

Kéfi, S., Rietkerk, M., Alados, C. L., Pueyo, Y., Papanastasis, V. P., ElAich, A., and De Ruiter, P. C. (2007). Spatial vegetation patterns and imminent desertification in Mediterranean arid ecosystems. *Nature*, 449(7159):213–217.

Layman, C. A. and Rypel, A. L. (2023). Beyond Kuhnian paradigms: Normal science and theory dependence in ecology. *Ecology and Evolution*, 13(7):e10255.

Leslie, P. H. (1945). On the Use of Matrices in Certain Population Mathematics. *Biometrika*, 33(3):183–212.

Levins, R. (1966). The Strategy of Model Building in Population Biology. *American Scientist*, 54(4):421–431.

Levins, R. (1969). Some Demographic and Genetic Consequences of Environmental Heterogeneity for Biological Control. *Bulletin of the Entomological Society of America*, 15(3):237–240.

Lotka, A. J. (1925). *Elements of Physical Biology*. Williams and Wilkins, Baltimore.

MacArthur, R. H. (1984). *Geographical Ecology: Patterns in the Distribution of Species*. Princeton University Press.

Magnuson, J. J. (1990). Long-Term Ecological Research and the Invisible Present. *BioScience*, 40(7):495–501.

May, R. M. (1972). Will a Large Complex System be Stable? *Nature*, 238(5364):413–414. Number: 5364.

May, R. M. (1975). Biological populations obeying difference equations: Stable points, stable cycles, and chaos. *Journal of Theoretical Biology*, 51(2):511–524.

May, R. M. (1981). The Role of Theory in Ecology. *American Zoologist*, 21(4):903–910.

Mduma, S. A. R., Sinclair, A. R. E., and Hilborn, R. (1999). Food regulates the Serengeti wildebeest: a 40-year record. *Journal of Animal Ecology*, 68(6):1101–1122.

Menge, B. A., Milligan, K., Caselle, J. E., Barth, J. A., Blanchette, C. A., Carr, M. H., Chan, F., Cowen, R. K., Denny, M., Gaines, S. D., Hofmann, G. E., Kroeker, K. J., Lubchenco, J., McManus, M. A., Novak, M., Palumbi, S. R., Raimondi, P. T., Somero, G. N., Warner, R. R., Washburn, L., and White, J. W. (2019). Pisco: Advances Made Through the Formation of a Large-Scale, Long-Term Consortium for Integrated Understanding of Coastal Ecosystem Dynamics. *Oceanography*, 32(3):16–25.

O'Gorman, E. J. and Emmerson, M. C. (2009). Perturbations to trophic interactions and the stability of complex food webs. *Proceedings of the National Academy of Sciences*, 106(32):13393–13398.

Pauly, D. (1995). Anecdotes and the shifting baseline syndrome of fisheries. *Trends in Ecology & Evolution*, 10(10):430.

Pichler, M. and Hartig, F. (2022). Machine learning and deep learning—A review for ecologists. *Methods in Ecology and Evolution*. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.14061.

Pickett, S. T. A., Kolasa, J., and Jones, C. G. (2013). *Ecological Understanding: The Nature of Theory and the Theory of Nature*. Elsevier. Google-Books-ID: x_AkBQAAQBAJ.

Post, D. M., Palkovacs, E. P., Schielke, E. G., and Dodson, S. I. (2008). Intraspecific Variation in a Predator Affects Community Structure and Cascading Trophic Interactions. *Ecology*, 89(7):2019–2032.

Preston, D. L., Henderson, J. S., Falke, L. P., Segui, L. M., Layden, T. J., and Novak, M. (2018). What drives interaction strengths in complex food webs? A test with feeding rates of a generalist stream predator. *Ecology*, 99(7):1591–1601.

Rackauckas, C., Ma, Y., Martensen, J., Warner, C., Zubov, K., Supekar, R., Skinner, D., Ramadhan, A., and Edelman, A. (2021). Universal Differential Equations for Scientific Machine Learning. arXiv:2001.04385 [cs].

Saavedra, S., Rohr, R. P., Bascompte, J., Godoy, O., Kraft, N. J. B., and Levine, J. M. (2017). A structural approach for understanding multispecies coexistence. *Ecological Monographs*, 87(3):470–486.

Scheffer, M., Hosper, S., Meijer, M.-L., Moss, B., and Jeppesen, E. (1993). Alternative equilibria in shallow lakes. *Trends in Ecology & Evolution*, 8(8):275–279.

Song, C. (2026). Toward a Conceptual Commons for Theory in Community Ecology.

Song, C., Von Ahn, S., Rohr, R. P., and Saavedra, S. (2020). Towards a Proba-

bilistic Understanding About the Context-Dependency of Species Interactions. *Trends in Ecology & Evolution*, 35(5):384–396.

Verhulst, P. (1838). Notice on the law that a population follows in its growth. *Correspondence Mathematique et Physique*, 10:113–121.

Volterra, V. (1926). Fluctuations in the Abundance of a Species considered Mathematically1. *Nature*, 118(2972):558–560.

Chapter 2: Symbolic regression for empirically realistic population dynamic time series

Published article

Jarman, C. N., Levi, T. & Novak, M. Symbolic regression for empirically realistic population dynamic time series. Ecological Informatics 99, 103988 (2026).

Funding

CNJ was supported by an NSF Graduate Research Fellowship (Grant No. 2234662) and the Gaulke Center for Marine Innovation and Technology.

Abstract

Applications of machine learning in ecology are rapidly expanding. Symbolic regression is gaining particular attention for its success in reverse-engineering human-readable explanatory population models, including the logistic growth and Lotka-Volterra equations, from simulated and laboratory-based population time series. However, field-based populations often lack the characteristics of the idealized time series used in prior assessments. We evaluated the utility of symbolic regression for such time series by quantifying its success for synthetic data varying in sampling

density, population cycle asymmetry, process noise, and the erroneous consideration of spurious variables. We further compared two data preprocessing options for estimating population growth rates (i.e., a discrete-time vs. continuous-time approach), and four evaluation workflows for selecting equations (two subjective, where gains in performance as a function of model complexity are assessed visually, and two objective, where quantitative metrics of performance determine model selection). Results indicate that a trade-off between sampling density and process noise primarily drives equation and variable recovery. Symbolic regression failed to recover the underlying equation at sampling densities below 25 points per cycle; however, at higher sampling densities, process noise did increase equation recovery rates. Importantly, although the true model was frequently recovered at sampling densities of 25 or more points per cycle, it was not consistently selected by the evaluation workflows. This discrepancy highlights a need for more robust post-algorithm selection criteria to identify the focal equation among competing candidates.

2.1 Introduction

Ecologists are motivated by the goal of describing patterns and process through observation, experimentation, and mathematical modeling. Mathematical models are particularly important for abstracting high-dimensional systems into equations that bring focus to key variables and mechanisms. For over a century, such models have been constructed based on *a priori* conceptualizations. For example, the

variables, parameters, and formulations of the logistic growth (Verhulst, 1838) and Lotka-Volterra (Volterra, 1926; Lotka, 1925) models, as well as the myriad other population dynamics models descended from these, were specified by logic, intuition, and first principles (Hilborn and Mangel, 1997). Unfortunately, although this approach may work for well-studied or relatively simple systems, it is susceptible to biases stemming from preconceived notions, paradigms, and scientific inertia (Bonnaffé et al., 2021; Ellner et al., 2002; Kuhn, 1970; Novak et al., 2025). In recent years, several approaches have therefore begun to take a more data-driven approach to dynamical model building (Solomatine and Ostfeld, 2008). Tools like neural networks (Bonnaffé et al., 2021), empirical dynamical modeling (Sugihara and May, 1990; Munch et al., 2020), and symbolic regression (Bongard and Lipson, 2007) in particular are increasingly being applied to time series to infer the data-generating processes from the data, rather than specifying a particular model or set of models *a priori*.

As a form of scientific machine learning (Baker et al., 2019), symbolic regression stands out among these data-driven approaches for its goal of providing human-readable, mechanistically interpretable equations in the spirit of the classical equations of theoretical ecology (Gerwin, 1974; Langley, 1977). Given data on a response variable and a set of putative predictor variables, symbolic regression algorithms use a process akin to biological evolution to evolve generations of equations through mutation, the recombination of parental equation components, and the selection and transmission of equations based on their fit to the data. (See the *Supplementary Materials* for an overview.) The first ecological application of

symbolic regression to population time series appears to have been Bongard and Lipson (Bongard and Lipson, 2007) who found it to successfully recover the two-species Lotka-Volterra competition equations even when the simulated time series were subject to observation noise. They also found it to produce similar equations and mechanistic interpretations to previous *a priori* models of lynx-hare interactions when applied to the classic Hudson Bay population time series. Gaucel et al. (Gaucel et al., 2014) extended these findings to demonstrate symbolic regression's ability to recover the Lotka-Volterra predator-prey equations from simulated time series subject to observation noise under sampling strategies in which the time elapsed between observations was either regular or irregular. This was followed by Martin et al. (Martin et al., 2018) who demonstrated symbolic regression's success in reverse-engineering the logistic growth equation, the Lotka-Volterra predator-prey equations, and a stage-structured population model from empirical time series of single and paired predator-prey laboratory populations. Finally, Chen et al. (Chen et al., 2019) assessed the method's ability to recover and reverse-engineer the population growth and nonlinear functional response equations of predator-prey populations from simulated and laboratory time series, focusing specifically on the data's dynamical informativeness through a contrast of time series exhibiting a stable limit cycle versus a transient approach to a fixed-point equilibrium. Although symbolic regression was only successful at recovering the generative equation in the cycling case, they concluded that its failure in the transient case was a limitation of the data rather than of symbolic regression itself (Chen et al., 2019). Altogether, the current literature therefore provides a unanimous endorsement of

symbolic regression for learning population dynamic models from time series data.

Despite prior demonstrations of its success, symbolic regression's ability to reverse-engineer the equations of more complex ecological time series with field-relevant sampling characteristics remains unknown. While symbolic regression offers a powerful framework for recovering mechanistic models, it has yet to be fully operationalized for generating insights from empirical systems. This lag in broad-scale application persists because we lack a clear understanding of how it handles the reality of field data compared to idealized simulations. In this regard it is particularly unclear whether the sampling densities of typical population time series are sufficient for symbolic regression to succeed. Prior assessments have assumed sampling intervals many times shorter than typical population cycle period lengths or generation times (e.g., Inchausti and Halley, 2001). Further knowledge gaps remain regarding time-series preprocessing practices (i.e., the use of time series interpolation methods) and how process noise rather than observation noise, the symmetry of population cycles, and the erroneous consideration of spurious predictor variables (previous benchmarks were restricted to *a priori* known drivers) affects the method's reliability. Finally, there has been little consideration of the alternative methods by which the suite of equations returned by symbolic regression can be assessed to identify and select the correct or "best" equation.

By generating empirically relevant time series from a mechanistic model of giant kelp populations, we evaluated symbolic regression's success in regards to each of these factors in six case studies, with each subjected to five different sampling densities. We included a contrast of four alternative evaluation workflows for

selecting candidate equations from along the Pareto frontier, the set of equations representing the optimal trade-off between complexity and explanatory power. We also examined the diversity of returned equations along and behind the frontier to gain insight into the effectiveness of each workflow. We structure our manuscript by introducing our generative population model, further motivating and describing the case studies, and detailing our implementation of symbolic regression and means of quantifying its rates of success for each workflow. Overall, we identified significant limitations under certain field-relevant conditions, particularly in regards to post-algorithm selection criteria for identifying the true model among the equations returned by symbolic regression.

2.2 Methods

2.2.1 The generative model

We used the Bence-Nisbet delay-differential model for giant kelp (*Macrocystis pyrifera*) to generate population time series (Bence and Nisbet, 1989). This model qualitatively reproduces the cycles observed in empirical giant kelp populations by encapsulating an effect of adult kelp on the recruitment of juvenile kelp in competing for space. The model describes the population growth rate of adult kelp A at time t as

$$\frac{dA(t)}{dt} = se^{-\alpha\tau}[1 - a_A A(t-\tau)]_+ - mA(t), \tag{2.1}$$

where s represents the juvenile settlement rate, α and m represent the juvenile and adult per capita mortality rates, a_A represents the proportion of space occupied by an adult individual, and τ represents the temporal lag of juveniles maturing to adults (i.e., juveniles are represented implicitly by a time-delayed effect of adult population size). The operator $[x]_+ = x$ if $x \geq 0$ else 0 provides an abrupt stop in growth once no space is available. We modified this model to include the possibility of multiplicative process noise in juvenile settlement and adult mortality (i.e., $s \to s\exp[\epsilon_1]$ and $m \to m\exp[\epsilon_2]$ with $\epsilon_i \sim \mathcal{N}(0,\ \sigma_i)$). Previous studies have only considered unbiased observation noise (i.e., measurement error), the effect of which diminishes with increased sampling density and time-series replication, while neglecting process noise (i.e., the inherent stochasticity of population dynamic processes). Notably, although process noise may increase the challenge of reverse engineering equations just like observation noise, it may also make time series more informative of their underlying processes by pushing systems to explore a larger space of variable states and conditions (Boettiger, 2018).

We chose the Bence-Nisbet model for several additional reasons as well. First, because it can generate both symmetrical and asymmetrical population cycles in a biologically justified manner. Many real-world populations exhibit asymmetric cycles (e.g., a rapid increase followed by a slow decay (Turchin, 2003)), yet with the exception of (Chen et al., 2019), all prior studies have considered only symmetrically cycling populations. This may be important because the combination of asymmetry and low sampling density could bias estimated rates of population change due to the higher chance of capturing the longer, slower phase, particularly

under sampling regimes that are regular (i.e., periodic), as most field-based time series are. Second, because it is more complex than previously used generative models by virtue of a time-delayed variable and use of exponential and $[\ ]_+$ operators. Its inclusion of a time-delayed variable was of particular interest because (*i*) many population time series entail time-delayed processes arising from higher-dimensional multi-species and age- or stage-structured processes (Turchin, 2003; De Roos and Persson, 2003), (*ii*) several time-series methods (e.g., empirical dynamical modeling) rely on the representation of multi-dimensional systems using time-delay embedding (i.e., Takens' theorem (Takens, 1981)), and (*iii*) it provided a means to introduce non-trivial spurious variables in our assessments; all prior studies provided their symbolic regression algorithms only with variables known to influence dynamics, which is an unlikely scenario in field contexts.

2.2.2 Case studies

2.2.2.1 Time series generation

We used the model to generate time series for six case studies, evaluating both a discrete-time and a continuous-time approach to calculating the per capita growth rate from symmetrically and asymmetrically cycling deterministic time series (cases 1–4; Fig. 2.1 top four rows) and a continuous-time approach from asymmetrically cycling stochastic time series under low and high levels of process noise (cases 5–6; Fig. 2.1 bottom two rows). These cases were designed to isolate specific empirical

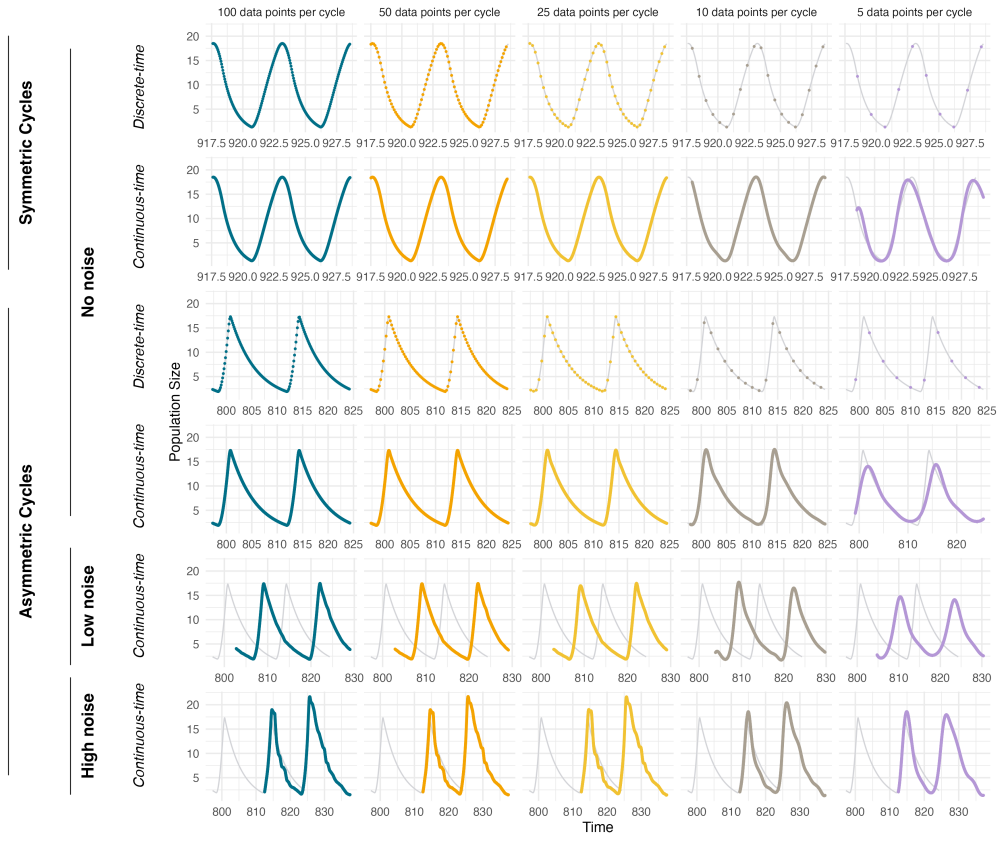


Figure 2.1: Time series illustrating our six case studies and the consequences of varying sampling density within each. A total of 15 cycles were used in our analyses, but only two are plotted for visualization. The light gray lines show the dynamics of the instantaneous deterministic Bence-Nisbet model for reference. The overlaid colored points show the (downsampled and interpolated) data from which the per capita growth rates and (lagged) putative predictor variables were obtained and provided to symbolic regression. Per capita growth rates were calculated using a discrete- or continuous-time preprocessing approach (see main text).

data-relevant challenges (e.g., how often data points can be collected, a modeler's choice between discrete- and continuous-time growth rate estimations, whether a population exhibits symmetric or asymmetric cycles, and the inclusion of process noise). The juvenile maturation rate was set to $\tau = 2$ following (Bence and Nisbet, 1989) for all case studies. Most other parameter values were specific to each case study (Tables S.1-S.2). Each time series was simulated for 1000 time steps, from which the last 15 complete population cycles were retained. The cycle period lengths, as determined by Fourier analysis (see Supplemental Materials), were 5.5 and 13.5 time steps in the symmetric and asymmetric deterministic cases, and averaged 13.2 and 12.5 time steps for the symmetric stochastic cases under low and high process noise, respectively. The time series thereby excluded transient dynamics and were of a length representative of long-term empirical datasets. To generate the actual time series that we provided to symbolic regression, we downsampled each case study's time series at sampling densities of 100, 50, 25, 10, or 5 time points per cycle. Although not reported, prior studies used synthetic time series with sampling densities of hundreds of time points per cycle and laboratory time series with sampling densities no lower than approximately 12 time points per cycle.

2.2.2.2 Response and predictor variables

Each downsampled time series was used to generate the response and predictor variables that we provided to symbolic regression. As the response variable, we

used the per capita growth rate, estimated differently depending on the use of a discrete- or continuous-time approach (Fig. 2.2a). For the discrete-time approach we calculated the per capita growth rate using the log-ratio of successive abundances, $\ln\left(A_{t+\Delta t}/A_t\right)/\Delta t$, as this converges on the instantaneous per capita growth rate $\frac{1}{A}\frac{dA}{dt}$ as Δt decreases (see *Supplementary Materials*). For the continuous-time approach we calculated the first derivative of the cubic spline interpolated log-transformed abundances (using SciPy, Virtanen et al., 2020) since $\frac{d\ln(A)}{dt} = \frac{1}{A}\frac{dA}{dt}$. Our comparison of these two approaches was motivated by prior simulation- and laboratory-based studies that have used continuous-time preprocessing, relying on sufficiently high sampling densities to avoid issues associated with alternative interpolation techniques. Our use of log-transformed population size differs from prior studies as it avoided the interpolation of negative population sizes to which the use of natural-scale population sizes is susceptible when sampling density is low.

All prior studies of ecological applications of symbolic regression to population time series provided their algorithms with state variables that were known to pertain directly to the focal dynamic. Similarly, we conducted a preliminary study to ensure that the symbolic regression algorithm we used could recover the generative equation when only $A(t)$ and $A(t-2)$ were considered as putative predictors (see the *Supplementary Materials*). However, in empirical contexts the true causal variables are typically not known. Our analyses therefore also considered the extraneous predictor variables $A(t-1)$ and $A(t-3)$. Note that only $A(t)$ and $A(t-2)$ appear in the generating equation (eqn. 2.1), making $A(t-1)$ and $A(t-3)$

autocorrelated but spurious determinants of the dynamics. For both the discrete-time and continuous-time approaches, the lagged variables were determined by back-transforming interpolated log-transformed population sizes.

2.2.3 Symbolic regression

Symbolic regression aims to find simple yet accurate equations to explain a given dataset. Using a process akin to biological evolution, generations of equations are evolved, where equations with a better fit to the data have a higher chance of producing offspring equations in the following generation. Symbolic regression begins with the creation of an initial generation of equations based on the input variables and operators (Fig. 2.2b). Parental equations of high fitness produce offspring equations for the next generation of equations with additions, deletions, recombinations, or mutations of their components. This equation evolution process occurs for a set amount of time or generations (Fig. 2.2c). The output is a set of top-performing equations that span a range of complexity values. For a more detailed overview, see the *Supplementary Materials*.

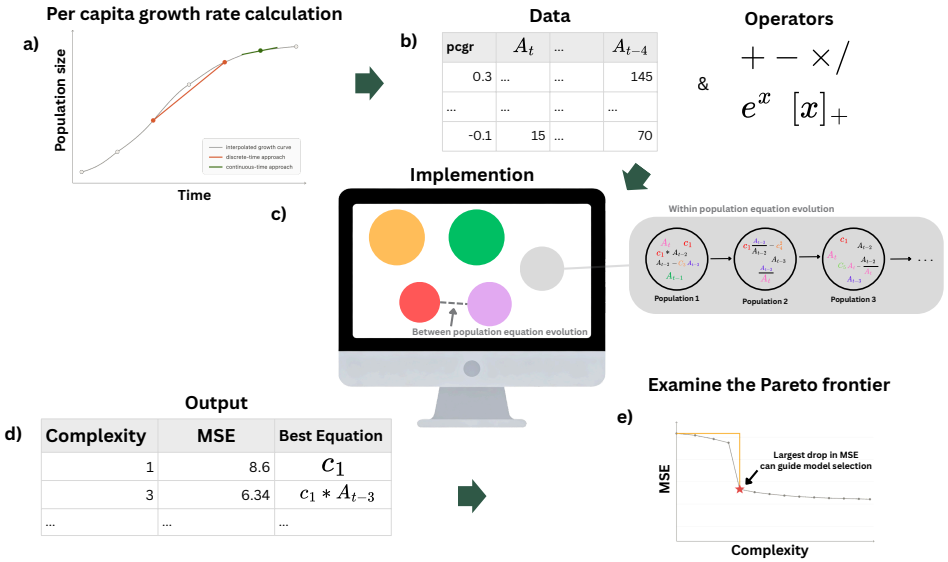


Figure 2.2: (a) Time series data were analyzed using either a discrete-time approach (orange line, difference in population sizes) or a continuous-time approach (green line, tangent to interpolated population sizes) to calculate the response variable, the per capita growth rate (pcgr), that we provided to symbolic regression. (b) The additional inputs provided to symbolic regression were the suites of predictor variables ((lagged) population sizes) and mathematical operators (including the custom function $[x]_+$). (c) In PySR, equation evolution occurs in two stages, (i)"within population evolution" wherein multiple equations exist within a single generation and parental equations are selected based on their fitness to create new equations in subsequent generations, and (ii)"between population evolution" wherein equation components can migrate between concurrent populations. (d) The output at the end of a run entails a list of the best (lowest MSE) equation for each level of complexity, along with its MSE values and complexity level. (e) The Pareto frontier of equations may be examined to select the equation affecting the largest drop in MSE (represented by the orange lines) as done in the first of the four workflows that we evaluated.

2.2.3.1 Implementation

Prior ecological studies used custom-built symbolic regression algorithms. We used PySR (v1.3.1; Cranmer, 2023), a Python library with a Julia backend that has seen frequent use in Physics. Unlike prior algorithms, PySR evolves equations using multiple semi-independent populations of equations between which migration can occur in order to explore a greater volume of equation space (see the *Supplementary Materials* for further details). With mean squared error (MSE, a.k.a. L2-loss) as the metric of equation fitness, we used 96 populations and ran each downsampled time series through PySR 100 independent times. Each of the $6 \times 5 \times 100 = 3000$ searches was run on a single compute node of a high performance computing cluster equipped with 64 CPU cores and a total of 514 GB of RAM for 22 hours or until an equation reached an MSE of $2 \cdot 10^{-9}$, though in almost all cases the search time was exhausted first. (Preliminary experiments revealed that these stopping criteria approached the limit where performance gains were exhausted.) Running 100 independent searches allowed us to quantify rates of equation recovery in a probabilistic manner rather than the binary success/failure manner of prior studies.

2.2.3.2 Equation and variable recovery success

We compared four workflows — two subjective and two objective — to select candidate equations from among the learned equation and assess symbolic regression's ability to recover the Bence-Nisbet model. Success rates were quantified in two ways for each workflow: as the proportion of 100 runs in which the selected equa-

tion (i) included only the two correct predictor variables of the Bence-Nisbet model (i.e., $A(t)$ and $A(t-2)$), and (ii) included only the two correct predictor variables and matched the Bence-Nisbet model in its functional form and parameter values (to three significant digits).

Prior studies assessed success using several approaches, including by subjectively inspecting the Pareto frontier of the learned equations to search for the true or a recognizable equation. The Pareto frontier represents the equations that maximize fitness (e.g., minimize MSE) for their given level of complexity; it reflects the most efficient trade-off between equation fitness and complexity. PySR quantifies an equation's complexity as the total count of all parameters, variables, and operators it contains (which for the Bence-Nisbet model is 14). Typically, the simplest model that provides the greatest increase in fitness along the Pareto frontier is selected as the best performing model, although a user's goals may cause the selection of other models. For instance, a goal of high predictive performance may influence a user to select a more complex model than if the goal is mechanistic interpretation (Shmueli, 2010). For Workflow 1, we followed the common practice of visually inspecting plots of MSE versus complexity and selecting the simplest equation that affected the largest additive decrease in MSE relative to all other equations along the frontier (e.g., Fig. 2.2e). Workflow 2 repeated this visual inspection using ln(MSE) to select the simplest model affecting the largest multiplicative decrease in MSE relative to all other equations along the frontier. PySR itself provides an objective, quantitative selection criterion equivalent to Workflow 2 called *score* that is calculated as the negative discrete log loss change

per unit change in complexity (Cranmer, 2023), which is what we used for Workflow 3. In Workflow 4, we used the Bayesian Information Criteria (BIC) for model selection (Schwarz, 1978). BIC evaluates equation performance by penalizing the equation's fit (its MSE) by the count of its parameters and the sample size of the data (Höge et al., 2018; Hilborn and Mangel, 1997). Next to AIC, BIC remains the most commonly used information criterion but has the property of being asymptotically consistent as the sample size increases (i.e., will select the true model — rather than the most efficient model — when it is present). For all workflows, we only considered equations with a complexity of up to 20 as equations of greater complexity became challenging for human readability, algebraic manipulation, and interpretation.

2.2.3.3 Explanatory variable diversity and selection success

In addition to quantifying rates of equation and variable recovery success for each workflow as above, we also (*i*) looked for the Bence-Nisbet model and (*ii*) characterized the diversity of variable combinations present within the top (lowest MSE) equations of each complexity level returned by PySR, combining all 100 runs of each case-study sampling-density combination. We did the latter by counting the number of unique variable combinations that occurred in the up to the top 10 equations, restricting ourselves to equations in the complexity range of 10–18. (While we considered up to 10 equations, fewer were recovered for certain complexities across the 100 searches. Because these instances were rare, we maintain the 'top

10' terminology throughout for brevity.) Without necessitating an equation selection workflow, this analysis was motivated by the context in which a user does not know the true driving variables or is not expecting classical population models to apply — hence would not look for or be able to identify these along the Pareto frontier or in any suite of returned equations — but could gain confidence and insight by virtue of symbolic regression returning a low diversity of equations that combined just a few dominant variables at or near the Pareto frontier.

Finally, in our most judicious assessment aligned with those of several prior studies, we determined whether the Bence-Nisbet model appeared in any of the returned equations of a given complexity in the range of 10–20, combining all 100 runs of each case study sampling-density combination. This permitted us to distinguish the ability of symbolic regression to evolve the correct generating equation from the ability of the four workflows to identify it along the Pareto frontier. Thus, when the Bence-Nisbet model did occur, we determined its maximum MSE-rank by counting the number of equations of the same complexity that exhibited a lower MSE than it.

2.3 Results

2.3.0.1 Equation and variable recovery success

We observed widely varying success rates (0–75% of 100 runs) in recovering the Bence-Nisbet model using the four equation selection workflows (Fig. 2.3). Over-

all, success rates were most influenced by sampling density and the presence of process noise, with no substantive or consistent effects of cycle symmetry versus asymmetry or preprocessing method. Although none of the workflows proved reliable, the two objective workflows (3 and 4) generally performed more poorly than the two subjective workflows (1 and 2) when success was observed. The highest rates of success were observed when applying Workflow 1 to the continuous-time preprocessed asymmetric time series having high process noise subject to sampling densities of 50 or more time points per cycle. Modest rates of success (10–20%) were also observed when applying Workflows 1 and 2 to otherwise equivalent time series having low process noise subject to sampling densities of 50 or more time points per cycle. However, rates of success were very low ($< 10\%$) or zero for the vast majority of other case studies, workflows, and sampling densities. In fact, the Bence-Nisbet was never recovered by any workflow when sampling densities were less than 50 time points per cycle, except for a few instances in the presence of process noise (19 of 7,200 instances; 18 cases $\times$ 100 runs $\times$ 4 workflows).

We observed higher — albeit still low and widely varying (0–100% of 100 runs) — rates of success in using the workflows to select equations which contained only the correct variables of the Bence-Nisbet model along the Pareto frontier (Fig. 2.3). A rate of 100% correct variable recovery occurred anomalously for Workflows 1 and 3 in the case of discrete-time preprocessed deterministic asymmetric cycles subject to a sampling density of only 5 time points per cycle. (None of these equations matched the Bence-Nisbet model's structure, see below.) Similarly, an anomalously high rate of correct variable recovery occurred when applying Workflow 3 in

the case of discrete-time preprocessed deterministic symmetric cycles at a sampling density of 50 time points per cycle. Beyond these exceptions, rates of correct variable recovery generally paralleled the patterns observed in the equation recovery success, tending to be low overall ($< 20\%$) and higher in the presence of process noise when applying Workflow 1. Unlike for equation recovery, correct variable recovery success rates remained above zero at sampling densities less than 50 time points per cycle in several deterministic cases.

2.3.0.2 Explanatory variable diversity and selection success

The diversity of variable combinations among the top 10 (lowest MSE) equations at each complexity level was low for a given case study (Fig. 2.4). No set of the top 10 equations contained more than 7 of the 11 possible pairwise-or-higher combinations of variables that could occur at complexities above 9, with agreement among all top equations for a single combination of variables or just two combinations of variables being by far the most frequent (48.7% and 27.7% of 267, respectively). Out of the 2581 equations examined in this analysis, $A(t)$ and $A(t-2)$ occurred most frequently (31.8%), followed by $A(t)$, $A(t-1)$ and $A(t-2)$ (23.2%). For most case studies, the frequency of equations that contained only the two correct variables, $A(t)$ and $A(t-2)$, increased substantially at all complexity levels as sampling density increased, with the majority of the top equations per complexity containing only the two correct variables in several high sampling density case studies. (The notable exceptions were the case study considering deterministic

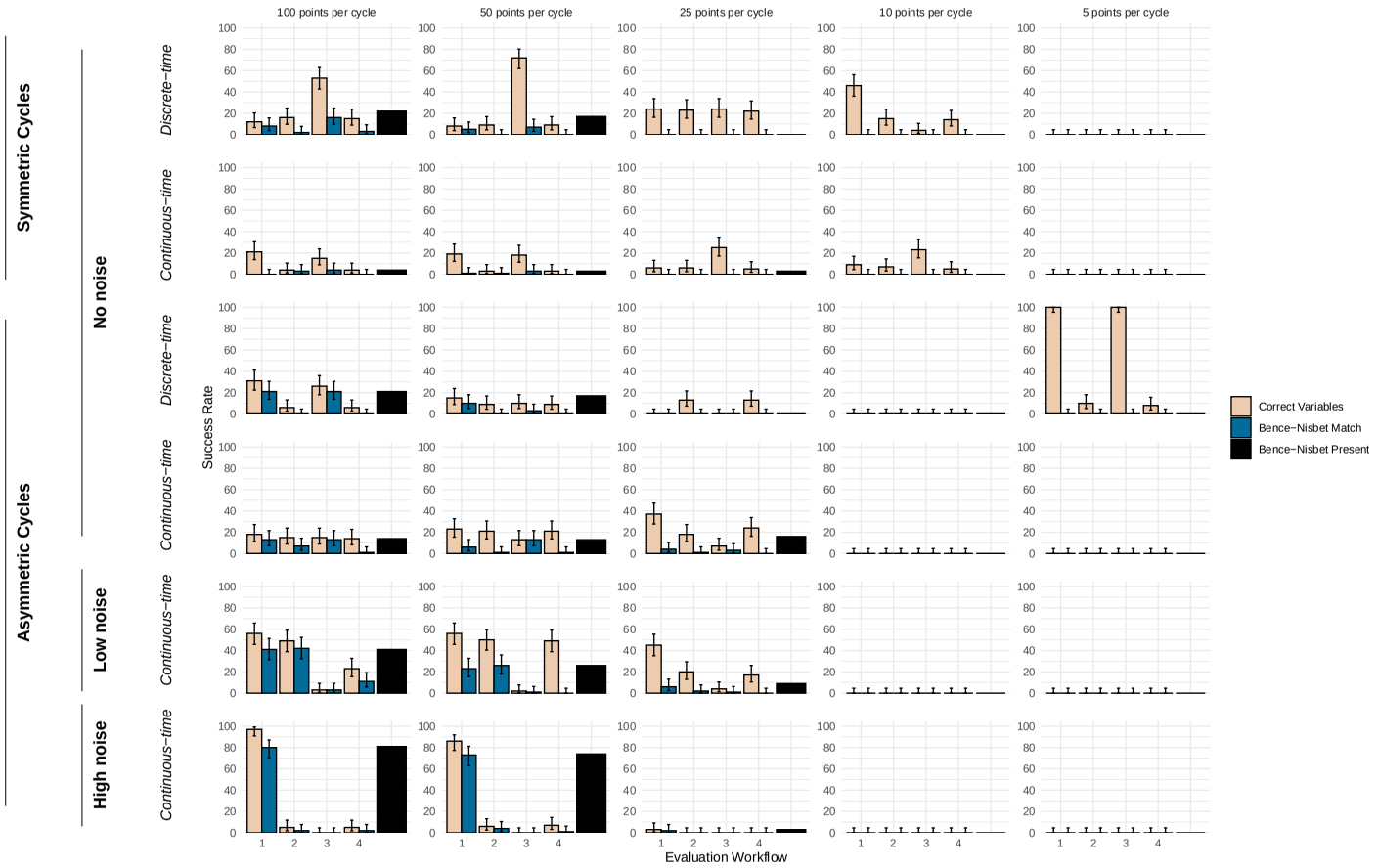


Figure 2.3: The success rate of symbolic regression and the four model selection workflows in selecting equations that contained the correct variables or that contained the correct variables and matched the formulation of the Bence-Nisbet model with which we generated the data. Confidence intervals reflect binomial uncertainties given a sample size of 100 independent runs. The black bars indicate the proportion of the 100 independent runs in which the Bence-Nisbet model occurred, regardless of whether a workflow selected it or whether it occurred in the top 10 equations.

symmetric cycles preprocessed with the continuous-time approach, for which the frequency of the correct variable equation remained low for all sampling densities, and the previously mentioned anomalous case of discrete-time preprocessed deterministic asymmetric cycles at a sampling density of 5 time points per cycle.) For all case studies, sampling densities less than 50 points per cycle tended to result in equations containing spurious variables, with the frequency of equations containing the combinations of $A(t)$, $A(t-1)$ and either $A(t-2)$ or $A(t-3)$ increasing particularly in the presence of process noise. Equations containing $A(t)$, $A(t-1)$, and $A(t-2)$ specifically also occurred more frequently for the two symmetric cycle case studies at high sampling densities. Equations that included only the two spurious variables, $A(t-1)$ and $A(t-3)$, never occurred, except in four cases of asymmetric cycles (deterministic, discrete- and continuous-time preprocessed at a sampling density of 10 points per cycle, deterministic, discrete-time preprocessed at a sampling density of 5 points per cycle, and low process noise, continuous-time preprocessed at a sampling density of 10 points per cycle; 178 of 36,000 equations (4 cases $\times$ 9 complexity levels $\times$ 100 runs $\times$ 10 top equations).

Despite not being consistently selected from along the Pareto frontier of any given run using the four workflows, the Bence-Nisbet model was evolved by symbolic regression for nearly all case studies (black bars in Fig. 2.3), and occurred in the top 10 (lowest MSE) equations of all case studies, given a sampling density of 25 or more points per cycle (Fig. 2.5). In the case studies in which the Bence-Nisbet model occurred within the top 10 as complexity 14, it had the lowest MSE, ranking first among all returned equations, 60% of the time. However, in four instances it

held a lower rank (higher MSE), and in two cases it failed to appear entirely, not placing along the Pareto frontier (Fig. 2.6). In most of the cases where the Bence-Nisbet model occurred, equations of higher complexity which were equivalent to the Bence-Nisbet model (i.e., equations that had extraneous constants that could be removed by algebraic simplification) also occurred (Fig. 2.5). These equations often also ranked first at their complexity level, but were more likely to drop in rank and out of the top 10 as their complexity increased (Fig. 2.6).

2.4 Discussion

We evaluated the utility of symbolic regression for recovering the generative population model of synthetic time series that varied in sampling density, process noise, cycle asymmetry, and preprocessing approach, and assessed four workflows for selecting candidate equations from among those returned. In contrast to prior studies based on densely sampled time series, our results indicate that both equation recovery and equation selection are substantially less reliable under characteristics typical of field-based time series. In particular, recovery success depended strongly on sampling density and, to a lesser extent, on the presence of process noise, while population cycle asymmetry and preprocessing approach had comparatively little influence. Even when the true equation was evolved, none of the four workflows proved consistently reliable at identifying it.

Nevertheless, our findings also suggest reasons for cautious optimism. Indeed, a central insight of this study is the need to distinguish among three related but

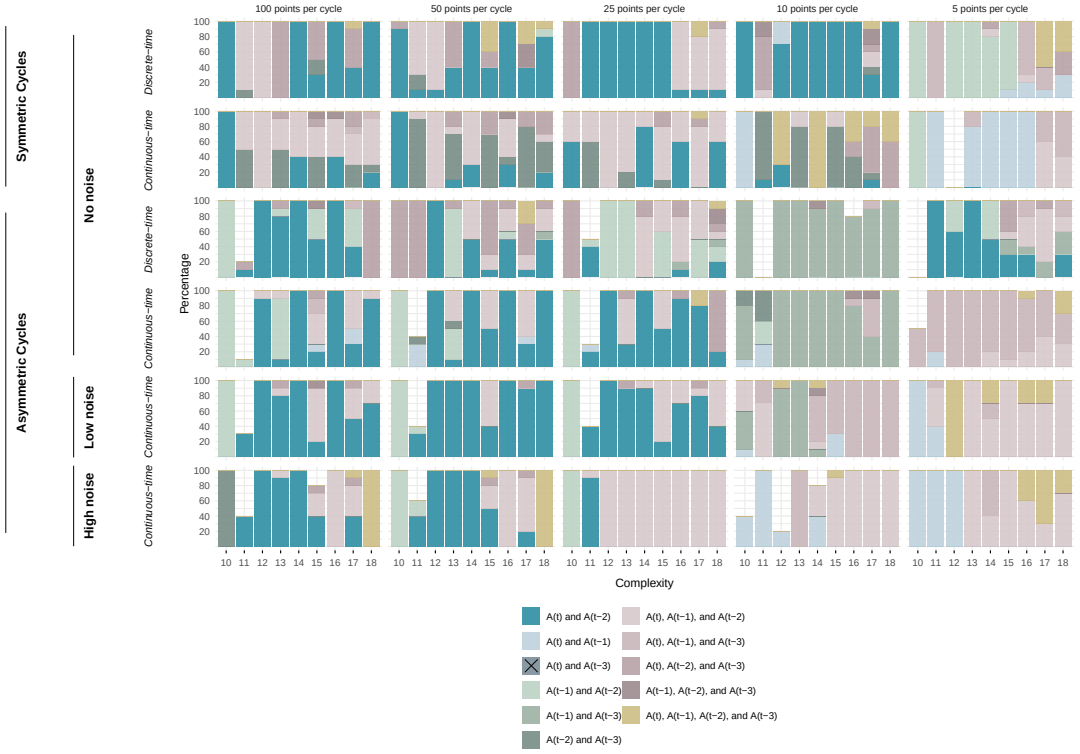


Figure 2.4: The proportional frequency of variable combinations present among the up to top 10 (lowest MSE) equations of each complexity level in the range of 10–18 (combining all 100 runs of each case-study sampling-density combination). The Bence-Nisbet model with which the data were generated has a complexity of 14 and contains only $A(t)$ and $A(t-2)$, making the inclusion of $A(t-1)$ and $A(t-3)$ spurious. All possible variable combinations are depicted in the legend; however, the combination of $A(t)$ and $A(t-3)$ never occurred.

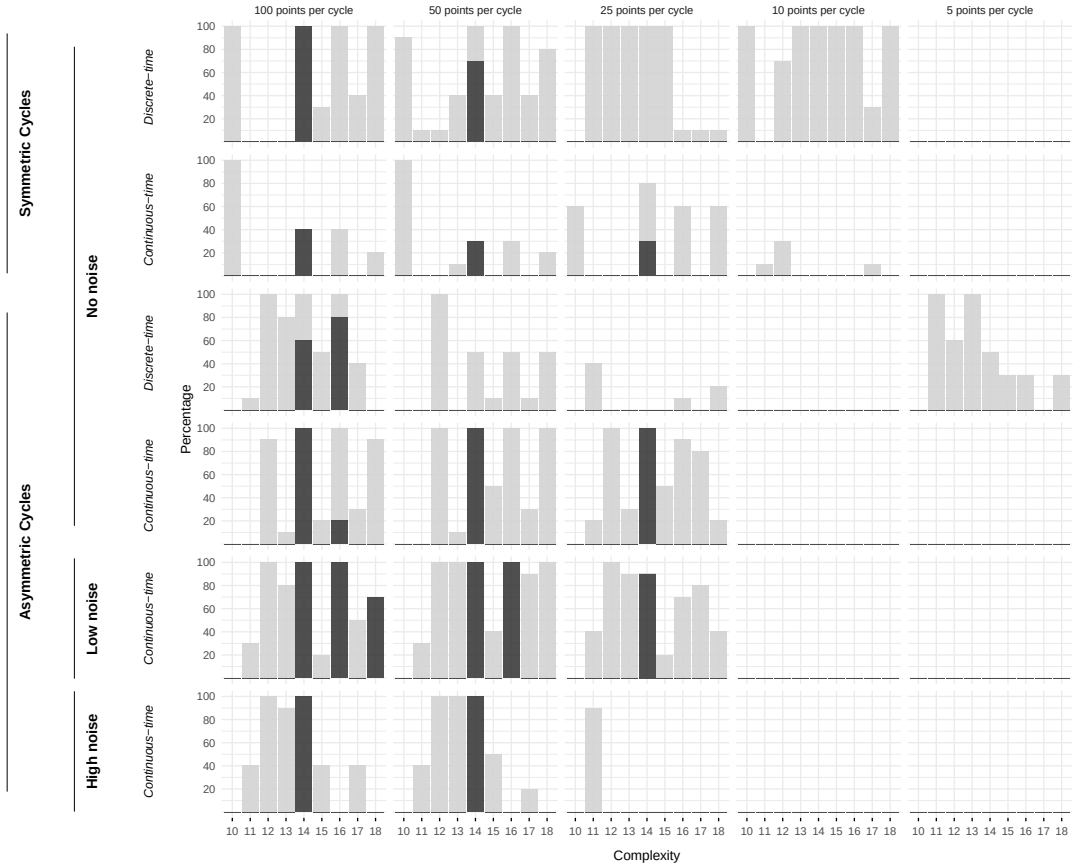


Figure 2.5: The proportion of the top 10 (lowest MSE) equations that correctly matched the Bence-Nisbet model in each complexity level in the range 10–18 (combining all 100 runs of each case-study sampling-density combination). Lightgray backgrounds indicate instances in which at least one equation occurred that contained only the two variables of the Bence-Nisbet model, $A(t)$ and $A(t-2)$, as shown in Fig. 2.4. The complexity level of the Bence-Nisbet model is 14. Matches at complexities greater than 14 occurred when the equations could be algebraically simplified to the Bence-Nisbet model. Note that matches at odd-valued complexity levels are not possible given the even-valued complexity of the Bence-Nisbet model and the necessity of including an operator for every additional parameter or variable.

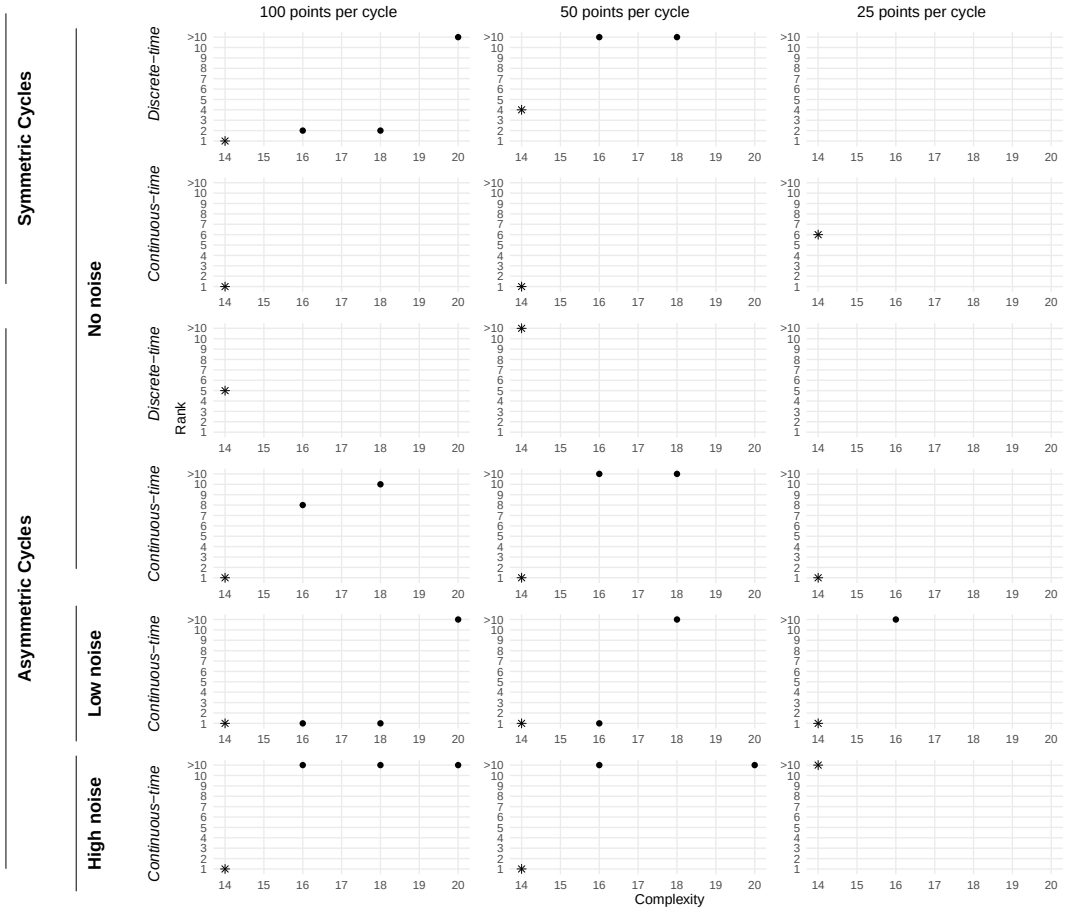


Figure 2.6: The maximum MSE-rank of the Bence-Nisbet model when it was recovered, reflecting the number of equations that exhibited a lower MSE than it in the set of up to 100 equations returned by the 100 runs of each case-study sampling-density combination. The complexity level of the Bence-Nisbet model is 14 (indicated by asterisks), hence lower levels are not possible. Ranks at complexities greater than 14 (indicated by points) occurred when the equations could be algebraically simplified to the Bence-Nisbet model. The Bence-Nisbet model was never recovered for time series of 10 and 5 points per cycle or for the deterministic discrete-time processed case study at 25 points per cycle (see Fig. 2.5).

distinct components of symbolic regression analyses: the algorithm's ability to evolve the correct equation, the position of that equation relative to and along the Pareto frontier, and the performance of workflows in selecting an equation. Although the generative model we used was more complex than those of prior studies, symbolic regression frequently did evolve the correct equation at least once across repeated runs when sampling density was sufficiently high. In these cases, the correct equation often occurred among the top-performing equations of its complexity level and typically ranked first or near-first in terms of MSE across all runs. Moreover, the diversity of equations near the Pareto frontier was generally low and dominated by models containing the correct variables. Together, these results underscore the value of long-term ecological time series of high density (Hughes et al., 2017) and indicate that symbolic regression can recover meaningful mechanistic structure from them, but also reveal that identifying this structure requires improved criteria for robust equation selection.

2.4.1 Equation generation versus equation selection

The low rate of successful recovery on a per-run basis raises the question of whether failures arose from limitations of our symbolic regression implementation or from limitations of the data and evaluation workflows. Our results suggest that the latter two, rather than computational constraints, were the primary factors. All searches were run for substantially more runs, longer durations, and across more populations of equations than in prior ecological studies, making it unlikely that

recovery failures reflect insufficient exploration of equation space.

Instead, symbolic regression frequently evolved incorrect equations that fit the data as well as or better than the true model. This outcome is consistent with extensive prior work demonstrating that structurally distinct models can generate indistinguishable dynamics when data do not sufficiently constrain the mechanism (e.g., Perretti et al., 2013; Novak and Stouffer, 2021; Song and Levine, 2025). In such cases, equation selection criteria that rely on fit and various measures of complexity alone cannot be depended upon to identify the true equation, even when it is present. In addition, in our study, the correct equation not only occurred behind the Pareto frontier but more commonly was present on the frontier, yet was not selected by an equation selection workflow, particularly for the workflows that were metric-based and hence entirely objective. Together, these findings highlight the need for criteria that better reflect structural identifiability rather than goodness-of-fit and traditional measures of complexity alone (e.g., Novak and Stouffer, 2021). Approaches that explicitly consider equations both on and behind the Pareto frontier, or that incorporate additional dynamical diagnostics like the ability to reproduce key statistical measures of the time series (e.g., Song and Levine, 2025), may also help to improve inference.

2.4.2 Data informativeness

2.4.2.1 Effects of sampling density

Sampling density exerted the strongest influence on symbolic regression's performance. At sampling densities below 25–50 points per cycle, the correct equation was rarely evolved or reliably selected. Because all our time series were sampled with densities determined on a per-cycle basis and the lowest assessed sampling frequency exceeded the Nyquist-Shannon frequency of the simulated dynamics, differences in recovery across case studies are unlikely to reflect classical aliasing effects (see the *Supplementary Materials*). Instead, reduced sampling density appears to limit recovery by diminishing data informativeness, affecting the accuracy of interpolated abundances and the estimation of growth rates (particularly under a continuous-time approach), and consequently the ability of the data to discriminate among alternative equation variables and structures.

Our analyses therefore emphasize that sampling density must be considered relative to a system's dynamical frequencies or characteristic timescales when applying equation-discovery methods. For strongly seasonal systems, sampling once or even twice a year, as is often the case for field-based population time series, will not be enough. Because achieving high sampling density is often challenging in field-based contexts, alternative strategies — such as irregular sampling (Gaucel et al., 2014), compressive sampling (Candes and Wakin, 2008), adaptive sampling (Henrys et al., 2024), or the use of state-space rather than rate-versus-state equation formulations that avoid the error-sensitive estimation of growth rates (Munch

et al., 2020) — may be useful in mitigating these constraints and warrant further investigation.

2.4.2.2 Effects of process noise

A more favorable result for field-based contexts was that process noise consistently increased recovery success. Indeed, as increasingly recognized in other contexts (Boettiger, 2018; Coulson et al., 2004; Dennis et al., 2006), stochastic perturbations appear to enhance data informativeness by expanding coverage of the system's attractor (see the *Supplementary Materials*) and revealing responses to variation that are absent with deterministic cycles. These effects likely reduce the set of dynamically equivalent models and help distinguish among competing equations. That said, contrary to our expectations, the highest noise level that we assessed often performed as well as or better than the low-noise case. Identifying the threshold at which process noise transitions from informative to detrimental thus remains an important direction for future work.

2.4.3 Cycle asymmetry and data preprocessing

Population cycle asymmetry had little effect on symbolic regression's performance, with asymmetric cycles performing comparably to or slightly better than symmetric ones. This result contrasts with our expectation that asymmetric cycles would be more vulnerable to under-sampling, but suggests that asymmetry alone does

not strongly interact with sampling density to constrain recovery when sampling density is sufficiently high (i.e., > 25 points) on a per-cycle basis.

Similarly, we observed no consistent advantage of either discrete-time or continuous-time preprocessing for estimating per capita growth rates. Although both approaches can introduce distinct sources of error, these effects were generally secondary to those of sampling density and process noise. Stronger differences may emerge under alternative dynamics or more extreme asymmetries.

2.4.4 Variable identification

Introducing spurious predictor variables revealed that symbolic regression can correctly identify the true driving variables when sampling density is sufficiently high (i.e., $\gtrsim 25$ points per cycle), even when the true equation structure is not evolved or identified. At lower sampling densities, however, autocorrelation among lagged predictors likely favored the inclusion of spurious but dynamically redundant variables. This pattern suggests that variable selection failures at low sampling densities again arise from data informativeness rather than from an inherent limitation of symbolic regression. The implementation of methods like convergent cross-mapping to remove spurious variables during preprocessing may be useful for improving inference (Munch et al., 2020; Sugihara and May, 1990). That said, the observation that the combination of $A(t)$, $A(t-1)$, and $A(t-2)$ specifically was the second most common combination of variables is particularly interesting given that it reflects the variables by which the true dynamics of the our presumed model may

be faithfully represented by time-delay embedding using Takens' theorem (Takens, 1981). Future work should examine the degree to which selection biases arise when spurious state variables differ in their degree of autocorrelation with the system's true driving variables, as well as the extent to which time-delay embedding may interfere with the identification of a system's true causal structure. Additionally, investigating the frequency of specific equation structures, rather than just variable occurrences, would be valuable. Such an effort will necessitate the automation of algebraic and structural analyses given the large number of equations involved.

2.4.5 Conclusions

Our results indicate that symbolic regression can recover mechanistically meaningful population models, but only under conditions of sufficiently informative data and with careful attention to equation selection. Given sufficient sampling, the primary limitation identified here lies not in equation generation but in identifying the correct model from among dynamically equivalent alternatives. Improving post hoc evaluation criteria — particularly those that move beyond exclusive reliance on the Pareto frontier — thus represents a key opportunity for methodological advancement. Such improvements are needed for robust generic guidelines to be developed regarding the specific time series characteristics that are necessary to ensure correct ecological inferences using symbolic regression.

2.5 Code and Data Availability

All code and data are available at `https://zenodo.org/records/19339107` and `https://github.com/cheyennejarman/SymbReg_KelpForest_Simulations.git`.

2.6 Acknowledgments

We thank Ricardo Noe Gerardo Reyes Grimaldo, Rebecca Hutchinson, Bryan K. Lynn, Dave Lytle, Bruce Menge, Steve Munch, and the Model Enable Machine Learning group for their insight throughout the project development. We also thank members of the Novak Lab for their feedback on early drafts of the manuscript.

Supplementary Materials for

Symbolic regression for empirically realistic population dynamic time series

Cheyenne N. Jarman, Taal Levi & Mark Novak

A Symbolic regression and PySR

A.0.1 Symbolic regression

Symbolic regression aims to find simple yet accurate equations to explain a given dataset. Using a process akin to biological evolution, generations of equations are evolved, where individual equations with a better fit to the data have a higher chance of producing offspring equations in the following generation. Symbolic regression begins with the creation of an initial generation of equations. Two equations with high fitness are selected as "parental equations". These parental equations produce an offspring equation for the next generation of equations through additions, deletions, recombinations, or mutations of their components. Multiple pairs of parental equations are selected within any given generation. Additionally, equations from previous generations may move forward into subsequent generations. This equation evolution process occurs for a set amount of time or generations. This algorithm can be implemented in various ways – the four previous ecological studies utilized custom-built algorithms. We implemented symbolic regression with PySR (Cranmer, 2023), an open-source Python library with a Julia backend. We will describe symbolic regression through the PySR lens, noting that while there are overarching similarities in the workflow, specific implementation details differ from other algorithms.

PySR begins with the user providing a list of predictor variables, a single response variable, and a suite of mathematical operators (both built-in and custom, user-defined operators) to combine terms (Fig. 2.2b). PySR differs from other symbolic regression algorithms in that evolution can occur within and between populations of equations (Fig. 2.2c). The number of populations is pre-specified as a user-input argument. PySR initiates the specified number of populations, each containing an initial generation of equations created through random assemblages of predictor variables, parameters, and mathematical operators. Within each population, equation evolution occurs through tournament selection, where the population is randomly subsampled and the individuals with the highest fit-

ness values are selected as parental candidates (Cranmer, 2023). Mean squared error (MSE) is commonly used as the fitness metric, thus the equation with the highest fitness will have the lowest MSE. The two parental candidates exchange subexpressions of their equation to produce a new equation via crossover. Mutation can also occur and only needs a single parental candidate. In this case, a subexpression of the parental equation is modified (e.g., a change in a parameter value or the operator used). Equation evolution is also done between populations (i.e., gene flow) where equations from one population randomly migrate to another population and vice versa, allowing for potentially new equation genes to enter a given population (Fig. 2.2c) (Cranmer, 2023). PySR's output is a file containing the best-performing equations and the respective MSE across a range of complexity levels, where complexity is defined as the sum of the numbers of parameters, variables, and operators present in a given equation (Fig. 2.2d). To help with equation selection, the user can plot the equation's MSE against the complexity level (Fig. 2.2e).

A.0.2 PySR

PySR allows for the specification and customization of many different arguments. We specified that the following operators could be used: addition, subtraction, multiplication, division, and the exponential function. We also created a custom function that matched the non-linear function, $[\ ]_+$. This was the only custom function we used. PySR's *populations* argument defines how many populations of equations are evolving. More populations will increase the diversity of equations and run time (Cranmer, 2023). We used 96 populations in all analyses. The user can also specify how many iterations, or the total number of generations, PySR will run for. The PySR documentation recommends setting this argument to a high value, so we specified a maximum of $1 \cdot 10^{10}$ iterations. Finally, the user can control how much time PySR spends searching by specifying the timeout argument. We specified our search to run for 22 hours. Equation overfitting and

bloating are known concerns for symbolic regression algorithms. PySR manages these concerns by defining a maximum equation size (user-specified or default value 30, we used the latter). Additionally, PySR ensures relatively even equation complexity representation by tracking the frequency of particular complexity levels and the complexity levels of the most recent equations (Cranmer, 2023). While there is no clear recommendation for the optimal number of runs needed to ensure the global optimum is reached, we ran each simulated data set through PySR independently 100 times. Previous studies also conducted (but combined) multiple search runs, with the number ranging from 3–30.

```
X = #array of predictor variables
y = #single response variable (e.g., log.ratio)

model = PySRRegressor(
early_stop_conditions = 2e-09,
timeout_in_seconds = 792000,
niterations = 1e10,
populations = 96,
ncyclesperiteration = 5000,
loss = "L2DistLoss()",
binary_operators=["+", "-", "*", "/"],
unary_operators=["exp","mygreater(x) = x >= 0 ? 1.0f0 : 0.0f0"],
extra_sympy_mappings={
  "mygreater": lambda x:
      Piecewise((1, Eq(x, 0) | (x > 0)), (0, True)), }, )
model.fit(X, y)
```

B Parameters used to simulate time series

Our modifiication of the (Bence and Nisbet, 1989) model incorporated process noise as

$$\frac{dA}{dt} = (se^{\epsilon_1})e^{-\alpha\tau}\left[1 - a_A A(t-\tau)\right]_+ - (me^{\epsilon_2})A(t) \tag{S.1}$$

where ϵ_1 and ϵ_2 are random variables with $\epsilon_i \sim \mathcal{N}(0, \sigma_i)$ and all parameter values are given in Tables S.1 & S.2.

Table S.1: Parameter values used to generate time series.

Parameter	Symbol	Symmetric	Asymmetric
Juvenile settlement rate	s	55	55
Juvenile mortality rate	α	0.35	0.15
Space occupied by individual adult	a_A	0.1	0.35
Adult mortality rate	m	1	0.2
Juvenile maturation delay	τ	2	2

Table S.2: Parameter values used to introduce process noise.

Parameter	σ_1	σ_2
No noise	0	0
Low	0.05	0.2
High	0.1	0.8

C Time series period lengths and aliasing

The period length of cycles produced by the model for each of the four cases was determined by Fourier analysis using *SciPy*. In the deterministic cases, the period lengths were 5.5 and 13.5 for the symmetric and asymmetric cases, respectively. For the stochastic cases, we simulated the time series for 1000 time steps and took the last 500 time steps to average the period length, which was 13.2 and 12.5 time steps for the symmetric stochastic cases under low and high process noise, respectively.

Since in every case we examined the cycles had periods of at least five sampling intervals — i.e., we observed a minimum of five time points per cycle — their fundamental frequencies were well below the Nyquist-Shannon cutoff (Shannon, 1949). Consequently, the down-sampled series are unlikely to suffer aliasing; only if very high-frequency harmonics exceeded half the sampling rate would they have risked being reflected at lower apparent frequencies.

D Asymptotic equivalence of discrete and continuous per capita growth rates.

The log ratio $\frac{N_{t+\Delta t}}{\Delta t}$ is asymptotically equivalent to the instantaneous per capita growth rate $\frac{1}{N}\frac{dN}{dt}$ as Δt approaches 0. That is,

$$\lim_{\Delta t \to 0} \frac{\ln\left(\frac{N_{t+\Delta t}}{N_t}\right)}{\Delta t} = \frac{1}{N}\frac{dN}{dt}\,. \tag{S.2}$$

This may be shown as follows.

Using the quotient rule for logs, we can rewrite the log ratio as the difference between $N_{t+\Delta t}$ and N_t,

$$\lim_{\Delta t \to 0} \frac{\ln\left(\frac{N_{t+\Delta t}}{N_t}\right)}{\Delta t} = \lim_{\Delta t \to 0} \frac{\ln(N_{t+\Delta t}) - \ln(N_t)}{\Delta t}\,. \tag{S.3}$$

The term on the right-hand side is equivalent to the definition of a derivative, thus we can rewrite the expression as

$$\lim_{\Delta t \to 0} \frac{\ln(N_{t+\Delta t}) - \ln(N_t)}{\Delta t} = \frac{d}{dt}\ln(N_t)\,. \tag{S.4}$$

Lastly, applying the chain rule shows how the time derivative of the logarithm of population size relates to the instantaneous per capita growth rate. Specifically, since $\ln(N_t)$ is a function of N_t, which itself varies with time, we differentiate $\ln(N_t)$ with respect to N_t and then multiply by the rate at which N_t is changing over time. Because the derivative of $\ln(N_t)$ is $\frac{1}{N_t}$, this yields

$$\frac{d}{dt}\ln(N_t) = \frac{1}{N}\frac{dN}{dt}\,. \tag{S.5}$$

E Equation recovery when using only $A(t)$ and $A(t-2)$

We simulated the Bence-Nisbet model using parameter values that generated asymmetric cycles (as in Table S.1) and provided PySR with data downsampled to a sampling density of 100 time points per cycle with the per capita growth rate using either the continuous-time or the discrete-time preprocessing approach. We performed 100 independent runs of PySR for each time series. For each run, we evaluated all resulting equations and ranked them according to their MSE. We then examined the rankings to determine whether the Bence-Nisbet model was present (Fig. S.1). In both the continuous- and discrete-time approaches, the Bence-Nisbet model was recovered among the top 10 ranked models.

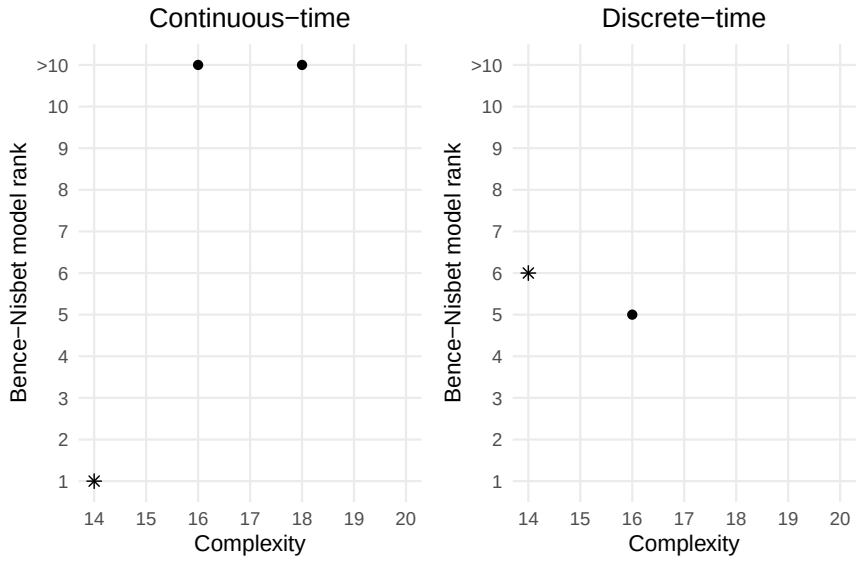


Figure S.1: The maximum MSE-rank of the Bence-Nisbet model when it was recovered, returned by the 100 runs from the asymmetric, 100 points per cycle data, with a continuous-time and discrete-time approach, respectively. We assessed the complexity range from 10 to 20. The lowest complexity that the Bence-Nisbet model occurred was complexity 14, which is the complexity level of the Bence-Nisbet model is 14 (indicated by asterisks). Ranks at complexities greater than 14 (indicated by points) occurred when the equations could be algebraically simplified to the Bence-Nisbet model.

F Attractor reconstruction

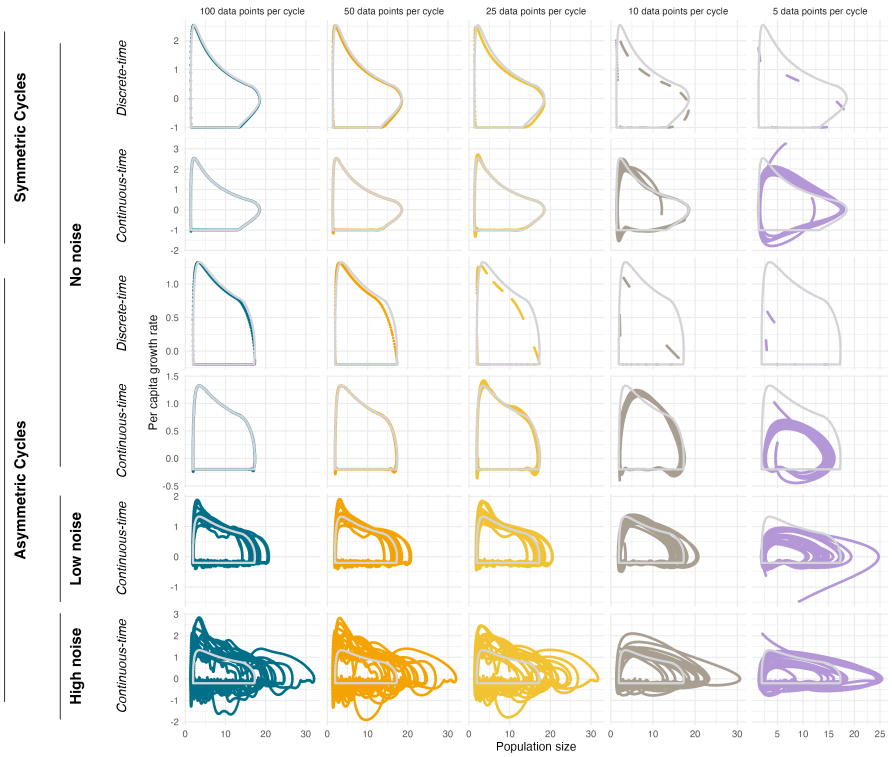


Figure S.2: Phase portraits of population size and per capita growth rate (log ratio or interpolated for discrete-time or continuous-time, respectively) provide insight into the influence of sampling density on symbolic regression's success. As sampling density decreases, the apparent attractor tends to shrink relative to the true attractor of the Bence-Nisbet model (overlaid in light gray), except for the cases with process noise.

Bibliography

Baker, N., Alexander, F., Bremer, T., Hagberg, A., Kevrekidis, Y., Najm, H., Parashar, M., Patra, A., Sethian, J., Wild, S., Willcox, K., and Lee, S. (2019). Workshop report on basic research needs for scientific machine learning: Core technologies for artificial intelligence. Technical Report 1478744.

Bence, J. R. and Nisbet, R. M. (1989). Space-Limited Recruitment in Open Systems: The Importance of Time Delays. *Ecology*, 70(5):1434–1441. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.2307/1938202.

Boettiger, C. (2018). From noise to knowledge: how randomness generates novel phenomena and reveals information. *Ecology Letters*, 21(8):1255–1267.

Bongard, J. and Lipson, H. (2007). Automated reverse engineering of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 104(24):9943–9948.

Bonnaffé, W., Sheldon, B. C., and Coulson, T. (2021). Neural ordinary differential equations for ecological and evolutionary time-series analysis. *Methods in Ecology and Evolution*, 12(7):1301–1315. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13606.

Candes, E. J. and Wakin, M. B. (2008). An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30.

Chen, Y., Angulo, M. T., and Liu, Y.-Y. (2019). Revealing Complex Ecological Dynamics via Symbolic Regression. *BioEssays*, 41(12):1900069. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bies.201900069.

Coulson, T., Rohani, P., and Pascual, M. (2004). Skeletons, noise and population growth: the end of an old debate? *Trends in Ecology & Evolution*, 19(7):359–364.

Cranmer, M. (2023). Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl.

De Roos, A. M. and Persson, L. (2003). Competition in size-structured populations: mechanisms inducing cohort formation and population cycles. *Theoretical Population Biology*, 63(1):1–16.

Dennis, B., Ponciano, J. M., Lele, S. R., Taper, M. L., and Staples, D. F. (2006). Estimating Density Dependence, Process Noise, and Observation Error. *Ecological Monographs*, 76(3):323–341. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1890/0012-9615%282006%2976%5B323%3AEDDPNA%5D2.0.CO%3B2.

Ellner, S. P., Seifu, Y., and Smith, R. H. (2002). Fitting population dynamic models to time-series data by gradient matching. *Ecology*, 83(8):2256–2270.

Gaucel, S., Keijzer, M., Lutton, E., and Tonda, A. (2014). Learning Dynamical Systems Using Standard Symbolic Regression. In Nicolau, M., Krawiec, K., Heywood, M. I., Castelli, M., García-Sánchez, P., Merelo, J. J., Rivas Santos, V. M., and Sim, K., editors, *Genetic Programming*, volume 8599, pages 25–36. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.

Gerwin, D. (1974). Information processing, data inferences, and scientific generalization. *Behavioral Science*, 19(5):314–325.

Henrys, P. A., Mondain-Monval, T. O., and Jarvis, S. G. (2024). Adaptive sampling in ecology: Key challenges and future opportunities. *Methods in Ecology and Evolution*, 15(9):1483–1496.

Hilborn, R. and Mangel, M. (1997). *The Ecological Detective: Confronting Models with Data*. Princeton University Press, Princeton, USA.

Höge, M., Wöhling, T., and Nowak, W. (2018). A primer for model selection: The decisive role of model complexity. *Water Resources Research*, 54(3):1688–1715.

Hughes, B. B., Beas-Luna, R., Barner, A. K., Brewitt, K., Brumbaugh, D. R., Cerny-Chipman, E. B., Close, S. L., Coblentz, K. E., de Nesnera, K. L., Drobnitch, S. T., Figurski, J. D., Focht, B., Friedman, M., Freiwald, J., Heady, K. K., Heady, W. N., Hettinger, A., Johnson, A., Karr, K. A., Mahoney, B., Moritsch, M. M., Osterback, A.-M. K., Reimer, J., Robinson, J., Rohrer, T., Rose, J. M., Sabal, M., Segui, L. M., Shen, C., Sullivan, J., Zuercher, R., Raimondi, P. T., Menge, B. A., Grorud-Colvert, K., Novak, M., and Carr, M. H. (2017). Long-term studies contribute disproportionately to ecology and policy. *BioScience*, 67(3):271–281.

Inchausti, P. and Halley, J. (2001). Investigating long-term ecological variability using the global population dynamics database. *Science*, 293(5530):655–657.

Kuhn, T. S. (1970). *The structure of scientific revolutions*. International encyclopedia of unified science. Foundations of the unity of science, v. 2, no. 2. University of Chicago Press, Chicago.

Langley, P. (1977). BACON: A production system that discovers empirical laws. *IJCAI*.

Lotka, A. J. (1925). *Elements of Physical Biology*. Williams & Wilkins.

Martin, B. T., Munch, S. B., and Hein, A. M. (2018). Reverse-engineering ecological theory from data. *Proceedings of the Royal Society B: Biological Sciences*, 285(1878):20180422.

Munch, S. B., Brias, A., Sugihara, G., and Rogers, T. L. (2020). Frequently asked questions about nonlinear dynamics and empirical dynamic modelling. *ICES Journal of Marine Science*, 77(4):1463–1479.

Novak, M., Coblentz, K. E., and DeLong, J. P. (2025). In defense of type i functional responses: The frequency and population dynamic effects of feeding on multiple prey at a time. *The American Naturalist*, 206(4):347–361.

Novak, M. and Stouffer, D. B. (2021). Geometric complexity and the information-theoretic comparison of functional-response models. *Frontiers in Ecology and Evolution*, 9:776.

Perretti, C. T., Munch, S. B., and Sugihara, G. (2013). Model-free forecasting outperforms the correct mechanistic model for simulated and experimental data. *Proceedings of the National Academy of Sciences*, 110(13):5253–5257.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

Shannon, C. E. (1949). Communication in the presence of noise. *Proceedings of the I.R.E.*

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, 25(3).

Solomatine, D. P. and Ostfeld, A. (2008). Data-driven modelling: some past experiences and new approaches. *Journal of Hydroinformatics*, 10(1):3–22.

Song, C. and Levine, J. M. (2025). Rigorous validation of ecological models against empirical time series. *Nature Ecology & Evolution*.

Sugihara, G. and May, R. M. (1990). Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature*, 344(6268):734–741.

Takens, F. (1981). *Detecting strange attractors in turbulence*, volume 898. Springer Berlin Heidelberg.

Turchin, P. (2003). *Complex Population Dynamics: A Theoretical/Empirical Synthesis*. Princeton University Press, Princeton, USA.

Verhulst, P. (1838). Notice on the law that a population follows in its growth. *Correspondence Mathematique et Physique*, 10:113–121.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt,

S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Volterra, V. (1926). Fluctuations in the abundance of a species considered mathematically. *Nature*, 118(2972):558–560.

Chapter 3: Identifying biotic drivers and forecasting the population dynamics of persistent giant kelp (*Macrocystis pyrifera*)

Funding

CNJ was supported by an NSF Graduate Research Fellowship (Grant No. 2234662) and the Gaulke Center for Marine Innovation and Technology.

Abstract

Identifying the biotic drivers of population dynamics, and the time scales over which they operate, remains a central challenge in ecology, particularly for foundation species whose fluctuations reshape entire communities. Giant kelp (*Macrocystis pyrifera*) forests exemplify this challenge. Decades of monitoring show substantial population variability, but the relative roles of density dependence, grazing, and competition remain unresolved. We combined symbolic regression with time-delay embedding to recover candidate governing equations for kelp per capita growth rate directly from long-term monitoring data of a persistent kelp population. We compared two explanatory variable sets, one of which included a spurious variable as a means of testing the method's ability to discern biologically important drivers. We evaluated equations along the Pareto frontier in three steps: biological

feasibility of selected variables, reconstruction of the empirical per capita growth rate-abundance relationship, and forward-simulation forecasting skill. Adult kelp abundance at a 6-month lag emerged as the dominant, consistently negative predictor of growth rate across complexity levels and both variable sets, while the spurious variable was consistently excluded, supporting the method's ability to distinguish biologically meaningful drivers from spurious ones. The negative effect of adult kelp abundance, together with a recurring negative effect of juvenile kelp abundance two years prior, is consistent with density-dependent self-thinning. Equations that best reconstructed the empirical growth rate relationship, however, were not those that forecast population trajectories most accurately, and forecasting skill also varied by transect independent of equation choice. These results demonstrate that symbolic regression can generate biologically feasible, interpretable hypotheses about kelp population drivers from empirical data, but also show that attractor reconstruction and dynamical forecasting are distinct, non-interchangeable criteria for evaluating candidate equations.

3.1 Introduction

Population dynamics in ecological systems emerge from the interplay of biotic interactions, abiotic forcing, and stochastic disturbances operating across multiple time scales (Turchin, 1999; Kendall et al., 1999; Levin, 1992; Bjørnstad and Grenfell, 2001). Identifying which of these drivers dominate, and over what temporal lags, is central to explaining how populations, and the communities that depend on them,

respond to environmental change. This challenge is especially acute for foundation species, species that have a disproportionate role in a community by shaping or defining it (Dayton, 1972). When natural or anthropogenic disturbances impact a foundation species, the resulting population fluctuations propagate outward to reshape the physical structure and biodiversity of entire ecosystems (Ellison et al., 2005; Angelini et al., 2011). Yet disentangling the drivers of population change from observational data is difficult, in part because standard modeling approaches require the relevant drivers and the form of their effects to be specified before a model is fit (Kendall et al., 1999).

Giant kelp forests are a tractable and important system in which to address this challenge: they are highly dynamic and ecologically important. Decades of long-term monitoring make them an ideal test case for disentangling the drivers of population dynamics (Watson and Estes, 2011; Kenner and Tinker, 2018; Krumhansl et al., 2016). Giant kelp (*Macrocystis pyrifera*) forms dense stands in shallow waters (5–20 m) across the coasts of California, Alaska, Washington, Australia, and Chile, creating complex three-dimensional habitats that hundreds of species rely on for food, habitat, and refugia (Schiel and Foster, 2015; Miller et al., 2018). Giant kelp is a perennial species; individuals can persist for multiple years, and new individuals recruit throughout the year, with seasonal pulses, maintaining the structural integrity of the kelp forest year-round (Parnell et al., 2026). Some kelp forests, including the sites studied here, have maintained continuous kelp cover for decades rather than cycling between kelp-dominated and urchin-dominated states (Kenner et al., 2013; Filbee-Dexter and Scheibling, 2014).

Despite this persistence, giant kelp densities exhibit fluctuations that cannot be attributed to seasonality alone (Bell et al., 2015; Randell, 2021). The mechanisms underlying these fluctuations remain debated, with proposed drivers spanning physical disturbance, bottom-up forcing, and biotic interactions (Castorani et al., 2022; Karatayev et al., 2021; Giraldo-Ospina et al., 2025; Detmer et al., 2021). Abiotic and oceanographic conditions tend to dominate population variability at broad spatial scales. Biotic interactions and stochastic processes are thought to be more influential at smaller, within-site spatial scales (Lamy et al., 2018; Edwards and Konar, 2020). Here, we focus on this small spatial scale, examining biotic mechanisms such as density dependence, grazer pressure, and intra- and interspecific competition (Santelices and Ojeda, 1984; Bence and Nisbet, 1989; Harrold and Reed, 1985; Graham, 1997). Because these drivers can act at different time scales – from near-instantaneous physical removal to delayed recruitment failure – identifying them requires methods that can resolve lags (i.e., the time interval between a driver and its effect on population growth) rather than assuming a static functional form.

Data-driven methods combined with long-term *in situ* monitoring let the data identify influential drivers directly, rather than presupposing them. Symbolic regression, in particular, stands out for its ability to produce easily interpretable equations and thereby shed light on potential drivers. There is a growing body of literature showcasing the successful application of symbolic regression to ecological data (Chen et al., 2019; Martin et al., 2018; Hovhannisyan, 2023). These applications often use ordinary differential equations, continuous-time equations

that require data for each state variable they include – a demand that is rarely met in ecology, where we often only have a single-variable time series. Delay differential equations offer a solution: temporal delays of that single state variable can reconstruct the focal dynamics via time-delay embedding (Takens, 1981; Sugihara et al., 2012). These delays can reflect a true temporal dependency (e.g., current recruitment depending on past adult density) or serve as a proxy for an unmeasured variable. Previous work combining symbolic regression with delay differential equations found that equation recovery depends on sampling density, succeeding above a threshold and improving further with process noise (Jarman et al., 2026), and that simpler underlying equations can be recovered even under sparse sampling (Appendix A). Working with empirical data, we have no known equation to check recovered variables against. As such, we treat the recovered equations and the variables they select as candidate mechanistic hypotheses about the underlying drivers, which can then be evaluated for biological feasibility, assessed via forecasting, and eventually tested experimentally. Because good fit to observed growth rate does not guarantee realistic dynamics when an equation is simulated forward, we evaluate these criteria separately.

Here, we apply symbolic regression with time-delay embedding to long-term monitoring data from giant kelp forests to identify which biotic drivers, and over what time lags, most plausibly shape kelp population dynamics. We evaluate candidate equations in three steps: assessing the biological feasibility of the variables and the functional forms recovered, testing their ability to reconstruct the empirical relationship between per capita growth rate and population size, and

evaluating their performance when forward-simulated as forecasts of population trajectories. Across these evaluation steps, adult kelp abundance at a 6-month lag consistently emerged as the dominant explanatory variable of kelp growth rate, potentially indicating lagged density-dependent intraspecific competition for resources or a seasonality effect. However, the equations that best reconstructed observed growth rates did not correspond to the equations that forecast population trajectories most accurately, indicating that reconstructing observed growth rates and forecasting population trajectories are not interchangeable measures of model quality. Beyond kelp, this approach offers a way to move from long-term monitoring data to testable hypotheses, and our results caution that how candidate equations are evaluated shapes which ones are chosen.

3.2 Methods

3.2.1 Long-term monitoring at San Nicolas Island, California, USA

Since the 1980s, the giant kelp forests surrounding San Nicolas Island, California, USA, have been sampled near-continuously via SCUBA surveys twice a year by the United States Geological Survey. The data used in our analysis come from two adjacent sites at the southern end of the island, West Dutch Harbor and East Dutch Harbor, which are the most wave-protected sites on the island and have maintained persistent kelp populations since monitoring began (Kenner et al., 2013)(Fig. 3.1). Surveys are conducted along five permanent 10x2 meter tran-

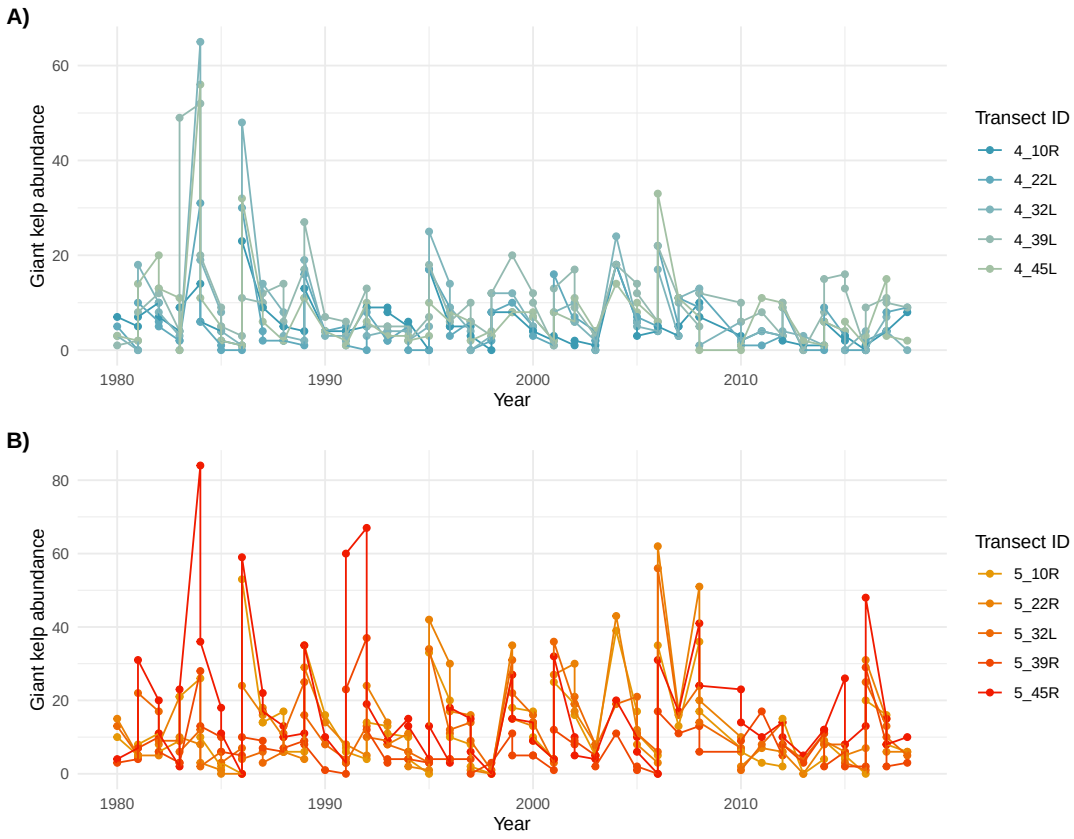


Figure 3.1: Giant kelp abundance over time from A) West Dutch Harbor and B) East Dutch Harbor, San Nicolas Island, CA. The lines connecting the points are to assist in readability and are not intended to be read as interpolation.

sects at each site, recording the densities of adult and juvenile giant kelp, purple sea urchins (*Strongylocentrotus purpuratus*), red sea urchins (*Mesocentrotus franciscanus*), and bat stars (*Patiria miniata*). Additional algal, invertebrate, and fish species are surveyed using complementary methods at the same sites (see Kenner et al. 2013 for full monitoring program protocols).

We selected four variables to be used as explanatory variables in our symbolic

regression analyses. Three of these were biologically expected predictors: adult kelp (A), juvenile kelp (J), and sea urchins (U, combining red and purple urchins). The fourth was a spurious variable, bat stars (S), which are detritivores with no documented direct interaction with live juvenile and adult giant kelp. West and East Dutch Harbor were combined for our analysis, as they are in proximity to each other and exhibit similar giant kelp dynamics (Fig. 3.1). By doing so, we were also able to increase the number of data points provided to the symbolic regression algorithm (transect identity was preserved throughout the analysis). We used two different interpolation procedures for two different purposes. First, to calculate our response variable, the per capita growth rate of adult giant kelp, we added 0.1 to any instance where adult kelp density equaled zero, reasoning that a measured zero reflects the absence of adult kelp within the limited area of a 10x2 m transect rather than the true local extinction of adult kelp at the site, then interpolated the logarithrm-transformed values of adult kelp using a cubic spline (Virtanen et al., 2020) and took the first derivative since $\frac{d\ln(A)}{dt} = \frac{1}{A}\frac{dA}{dt}$. Second, to construct our temporal lags, we interpolated all explanatory variables to fill in 9 of the 76 surveys missing across both sites (7 isolated single-period gaps, and 1 two-period gap, making one year the largest missing interval). Given that there were relatively few points missing, we opted to interpolate the missing periods rather than remove these periods. An initial attempt using cubic splines produced anomalously large interpolated values for juvenile kelp – a known failure mode of cubic splines in areas of data sparsity. We therefore interpolated all explanatory variables using a piecewise cubic Hermite interpolating polynomial (PCHIP), which

preserves monotonicity and stays bounded within the local data range, avoiding the overshoot behavior that cubic splines exhibited (Fritsch and Carlson, 1980). All variables were then lagged for four time steps, with each step representing a 6-month increment (e.g., $A(t), A(t-6), A(t-12), A(t-18)$, and $A(t-24)$). All transects were concatenated into a single dataset; transect identity was preserved throughout interpolation, per capita growth rate calculation, and lag construction.

3.2.2 Implementing symbolic regression

Symbolic regression evolves generations of equations from an input of a single response variable and a set of explanatory variables. Over successive generations, equations are evaluated by how well they minimize the mean squared error between their predicted and observed values of the response variable, with higher-performing equations more likely to serve as parents for the next generation. Several symbolic regression algorithms exist; we used PySR (v 1.3.1) (Cranmer, 2023) to search for candidate governing equations from our focal dataset (see Jarman et al. (2026) for a detailed description). We compared two explanatory variable sets, both using per capita growth rate of adult giant kelp as the response variable, each evaluated in its own run of PySR searches. In one run (hereafter the "no stars run"), we included biologically relevant variables: adult kelp (A), juvenile kelp (J), sea urchins (U), and their lags. In our second run (hereafter the "with stars run"), we added a spurious variable, bat stars (S), alongside A, J, U, and their lags. For the operator library, we included addition, subtraction, multiplication, divi-

sion, and the exponential function, with exponential self-nesting disallowed (e.g., no $exp(exp())$). Each run included 100 independent searches through PySR with L2 loss (mean squared error) as the equation fitness metric. We ran analyses on a single compute node of a high-performance computing cluster equipped with 64 CPU cores and 514 GB of RAM, with a maximum runtime of 22 hours (justified in Jarman et al. 2026).

3.2.3 Evaluation Workflow

3.2.3.1 Equation selection from the Pareto frontier

The Pareto frontier represents the set of equations that minimize MSE at each level of complexity, reflecting the most efficient trade-off between equation fitness and complexity. PySR quantifies an equation's complexity as the total number of parameters, variables, and operators it contains. We plotted the Pareto frontier by identifying the best-performing equation (i.e., lowest MSE) at each complexity level across the 100 independent searches within a run, with equation complexity on the x-axis and MSE on the y-axis. Following common practice, we visually inspected the Pareto frontier and selected the simplest equation that produced the largest decrease in MSE relative to less complex equations along the frontier; we did not apply an explicit quantitative threshold for this decrease, as such selection is inherently subjective. Equations selected by this workflow were carried forward to in-depth evaluation across three steps: variable selection and equation formulation,

per capita growth rate reconstruction, and population size forecasting.

3.2.3.2 Evaluation Step 1: Variable selection and equation formulation

For each equation selected from the Pareto frontier, we examined which variables were selected and how they were structured within the equation, assessing biological feasibility where interpretable. Specifically, for each retained variable, we computed the sign of its partial derivative with respect to predicted per capita growth rate at every observed data point ($n = 720$), and assessed whether that sign was consistent with the expected direction of a positive or negative biological influence and, when necessary, its consistency across data points. We additionally compared variable selection and equation form between the no stars run and the with stars run to evaluate whether the spurious variable was retained in high-performing equations and whether it appeared alongside or in place of biologically relevant terms.

3.2.3.3 Evaluation Step 2: Per capita growth rate reconstruction

We evaluated how each equation selected from the Pareto frontier reconstructed the empirical relationship between per capita growth rate and population size directly from the observed data. To do so, we compared predicted against observed per capita growth rate and examined residuals against fitted (predicted) values

to identify systematic biases (e.g., consistent over- or under-prediction at particular population densities and across specific time periods). Equations that reconstructed the per capita growth rate well (i.e., closer fits on the 1:1 line) and showed no strong residual structure were carried forward to population size forecasting; equations with poor reconstruction or clear residual patterns were excluded at this step.

3.2.3.4 Evaluation Step 3: Population size forecasting

Equations passing per capita growth rate reconstruction were forward-integrated as delay differential equations to evaluate whether they reproduced the qualitative dynamics of the observed system over time. Delay differential equations require a history function (i.e., the state of the system over a past interval, not just a single initial point), since the equation depends on lagged values. Because these equations were recovered from empirical data rather than derived from a known mechanistic model, no true solution was available to generate a history function analytically. Instead, we constructed history and forcing functions by interpolating the observed time series for a given site and transect, allowing us to generate values of $A(t)$ and any other required variables at arbitrary time points chosen by the numerical solver during integration, rather than being restricted to the discretely sampled survey time points. For any explanatory variable other than A appearing in a selected equation (i.e., J, U, or S), we treated its observed time series as external forcing rather than jointly simulating it, since no corresponding governing equation for

that variable was available. Equations were then integrated forward from a single seed time point across the full length of the time series (i.e., whole-trajectory forecasting, without resetting to observed values at intermediate steps), and the resulting trajectory was compared against the subsequent observed trajectory via visual inspection and goodness-of-fit metrics (R^2, root mean square error (RMSE), and normalized root mean square error (NRMSE)). For this analysis, we randomly selected two transects from each site (West Dutch 10R, West Dutch 32L, East Dutch 22R, and East Dutch 39R), rather than reporting all 10 transect $\times$ equation combinations.

3.3 Results

3.3.1 Symbolic regression output

Across 100 independent searches for the no stars run, PySR returned equations spanning complexity 1–30, with best-fit loss (MSE) ranging from 8.04 at the lowest complexity to 3.38 at the highest. Because the 100 independent searches did not all converge on equations with comparably low loss at a given complexity level, we examined the 10 lowest-loss equations at each complexity level (280 total equations). We found that of the 15 candidate variables (adult kelp, juvenile kelp, and sea urchins, each at four temporal lags plus a current value), 13 appeared in at least one top-performing equation, but selection frequency was not uniform (Fig. 3.2). $A(t-6)$ was present in 96.4% of the 280 equations, and current-value

terms for adult and juvenile kelp each appeared 67.9% and 71.4%, respectively. Sea urchin terms were rare, occurring in under 2% of equations and with no lag terms selected at any complexity. Two lag terms, $A(t-18)$ and $J(t-6)$, never appeared in a top-10 equation at any complexity level.

For the with stars run, PySR returned equations spanning complexity 1–30, with MSE ranging from 8.04 to 3.38 – nearly identical to the range obtained in the no stars run. Examining the 10 lowest-loss equations at each complexity level (280 equations), the same pattern of kelp-variable dominance held: $A(t-6)$ was retained in 96.4% of equations, and current-value terms for adult and juvenile kelp appeared in 67.5% and 71.4% of equations, respectively (Fig. 3.2). Neither sea urchins nor bat stars appeared frequently: sea urchin terms did not appear in any top-10 equation at any complexity level, and bat stars appeared in only 9.28% of equations, with 7.1% represented by a single lag term $S(t-24)$. The remaining bat star features (current value and three lags) were retained in fewer than 2.1% of equations. As in the no stars run, $A(t-18)$ and $J(t-6)$ never appeared in a top-10 equation at any complexity level. The same was true for $S(t-12)$ and all five sea urchin terms.

For the no stars run, structural diversity among the top-10 equations varied non-monotonically with complexity: the top 10 converged on a single shared functional form at each complexity level from 1–13, increased in diversity at complexity 14–15, returned to a single shared functional form at each complexity level from 16–18, then increased from complexity 19 onward, with near-complete diversity (i.e., 9–10 distinct forms out of 10) typical from complexity ≥ 23 onward (Fig. 3.3A).

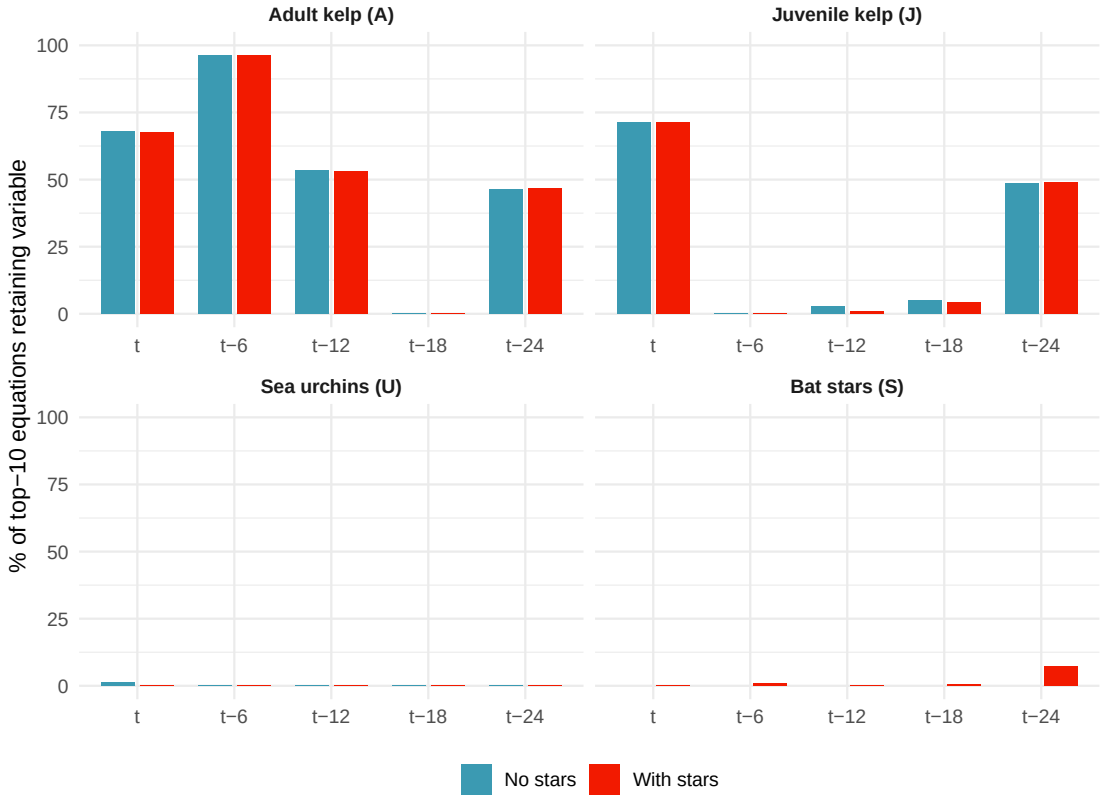


Figure 3.2: Retention percentage of the variables included in the top-10 lowest-loss equations across complexities 1–30, for the no stars run (blue) and with stars run (red). Variables included adult kelp A, juvenile kelp J, sea urchins U (and, for the with stars run, bat stars, S), each at the current value and four temporal lags (6, 12, 18, and 24 months). Both runs had 280 equations each (10 per complexity level across complexities 1–30, excluding complexity levels with no returned equations).

The with stars run showed a broadly similar pattern: the top 10 converged on a single shared functional form at each complexity level from 1–13, increased at complexity 14–16, then settled into 4–9 distinct forms through complexity 20–30 (Fig. 3.3B).

3.3.2 Evaluation workflow

3.3.2.1 Equation selection from the Pareto frontier

The aggregated Pareto frontier (Fig. 3.4) showed a steep initial drop in loss between complexity 1 and 4 (MSE: 8.04 to 5.42), followed by a series of smaller, repeated step-downs rather than a single dominant inflection point. Because the frontier took this step-wise shape, consistent with several successive, comparably sized gains in fit, rather than one clear broken-stick-like form, we did not restrict equation selection to a single selected complexity level. Instead, we visually identified each point along the frontier at which increasing complexity produced a relatively large drop in loss relative to the preceding step, identifying eight complexity levels of interest: 4, 7, 11, 13, 15, 17, 20, and 26 (MSE = 5.42, 4.67, 4.19, 4.07, 3.94, 3.83, 3.63, and 3.47, respectively), and carrying forward the corresponding equation at each. The no stars and with stars runs produced near-identical frontiers, and the same eight complexity levels were identified in both (see Supplementary Materials). These complexity levels produced the same or closely equivalent equations across the no stars and with stars runs. At every selected complexity, the same

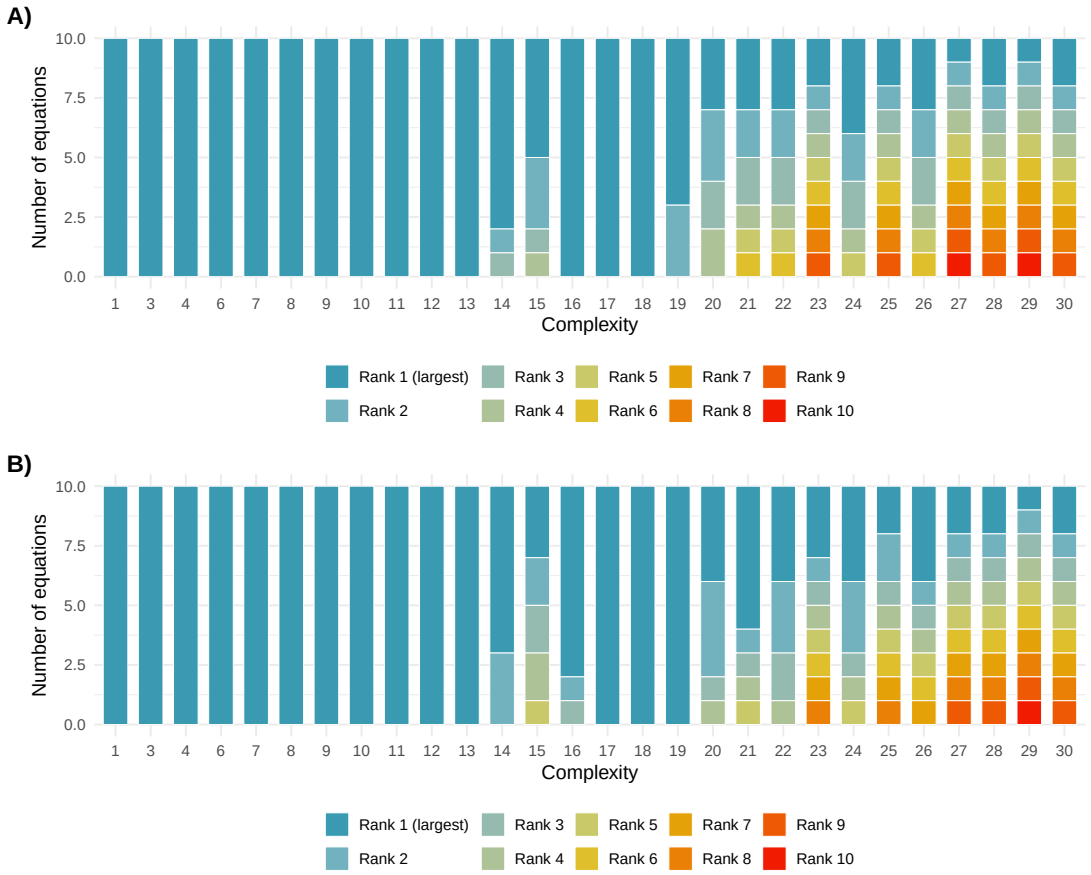


Figure 3.3: Structural diversity of the top-10 lowest-loss equations at each complexity level for the A) no stars and B) with stars runs. At each complexity level, the 10 equations were grouped by shared functional form, and stacked bar segments show the number of equations belonging to each group, ranked by group size from largest (Rank 1, dark blue) to smallest (Rank 10, red). Ranks are not conserved across complexity levels. A single solid-colored bar indicates that all 10 equations share one functional form at that complexity level. Bars divided into multiple colors indicate multiple distinct forms. Complexity levels 2 and 5 are omitted because PySR did not return equations at those complexity levels.

variables appeared, and the underlying functional relationships were equivalent: differences were limited to independently refit numeric coefficients (e.g., $h = 16.16$ vs. $k = 16.29$ in complexity 11, Table A.1, as expected since each search is a separate optimization), and some algebraically equivalent sign-flipped rearrangements. The sole structural exception was complexity 26, which converged on a different but comparably performing combination of the same variables. Bat stars did not appear in any of the eight equations selected, meaning the spurious variable's occasional appearance elsewhere on the frontier never corresponded to a large enough drop in loss to be selected for further evaluation.

3.3.2.2 Evaluation Step 1: Variable selection and equation formulation

Across both the no stars and with stars runs, the equations selected along the Pareto frontier retained an identical set of variables at every selected complexity: adult kelp at current time t and lags of 6, 12, and 24 months ($A(t)$, $A(t-6)$, $A(t-12)$, $A(t-24)$) and juvenile kelp at current time and a 24-month lag ($J(t)$, $J(t-24)$). No sea urchin or bat star terms were retained at any selected complexity in either run.

For complexities 4, 7, 11, 13, 15, 17, and 20, the equations were structurally identical between the no stars and with stars runs, with coefficients differing only in lower-order decimal places (Table A.1). This indicates that convergence on the same underlying functional form held not only within the 100 independent searches

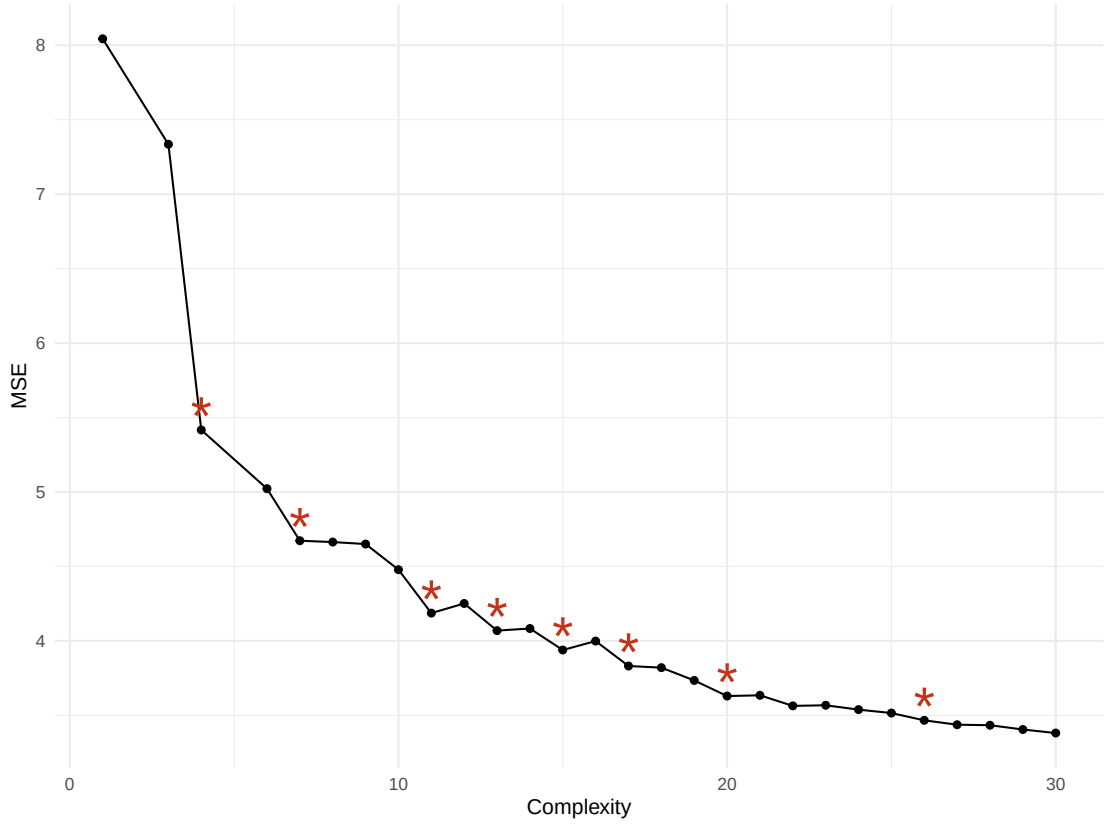


Figure 3.4: Pareto frontier for the no stars PySR run, showing the best-performing equation's loss (MSE) at each complexity level across 100 independent searches. The red markers indicate the eight complexity levels selected for further evaluation, at which increasing complexity produced a relatively large drop in loss relative to the preceding step. The equations at these complexity levels were carried forward to Evaluation Steps 1-3. The with stars run produced a near-identical frontier and selected the same eight complexity levels (see Supplementary Materials).

of a given variable set, but also across the two variable sets themselves, 200 independent searches in total. The complexity 26 equation was the exception: while both runs retained the same variables, the two runs' equations combined them differently, indicating that at this complexity PySR converged on two comparably-performing but structurally distinct formulations of the same variable set.

Table 3.1: Comparison of the selected equations from no stars and with stars PySR runs. All equations use the same variables between the no star and with star runs, and all but complexity 26 are the same structural form. Coefficient values are rounded to the nearest 0.01; symbols with identical rounded values are reported once. The letters e, i, and t were excluded from coefficient renaming to avoid confusion with the exponential constant, the imaginary unit, and the time variable, respectively.

Complexity	No stars	With stars
4	$\frac{a}{\exp(A(t-6))}$	$\frac{a}{\exp(A(t-6))}$
7	$\frac{b}{A(t-6)+c} - d$	$\frac{b}{A(t-6)+c} - d$
11	$\frac{-f}{\frac{-g-A(t-6)}{J(t)+h}} - j$	$\frac{f}{\frac{g+A(t-6)}{J(t)+k}} - j$
13	$\frac{l}{A(t-6)+m} + \frac{-n}{J(t)+o+A(t)}$	$\frac{r}{A(t-6)+m} + \frac{-p}{J(t)+q+A(t)}$
15	$\frac{s}{\frac{A(t-6)}{J(t)+u}+v} - \frac{A(t-6)}{A(t)+w}$	$\frac{s}{\frac{A(t-6)}{J(t)+x}+v} - \frac{A(t-6)}{A(t)+w}$
17	$\frac{y}{A(t-6)+z} - \frac{J(t-24)+\hat{a}}{A(t-12)+J(t)+A(t)+\hat{b}}$	$\frac{\hat{c}}{A(t-6)+z} - \frac{J(t-24)+\hat{a}}{A(t-12)+J(t)+A(t)+\hat{b}}$
20	$\frac{\exp\left(-\hat{e}\frac{J(t-24)}{J(t)+\hat{d}}\right)}{\frac{A(t-6)}{A(t-24)+A(t-12)+A(t)}+\hat{f}} - \hat{g}$	$\frac{\exp\left(-\hat{e}\frac{J(t-24)}{J(t)+\hat{h}}\right)}{\frac{A(t-6)}{A(t-24)+A(t-12)+A(t)}+\hat{f}} - \hat{i}$
26	$\frac{\hat{j}}{A(t-6)+\hat{k}} + \frac{-\hat{l}-A(t-6)}{\frac{A(t-24)+A(t-12)-J(t)}{\exp(\hat{m}\,J(t-24))}+J(t)+\hat{n}+A(t)}$	$\frac{\hat{o}}{A(t-6)+\hat{p}} + \frac{-\hat{q}-A(t-6)}{A(t)+\big(A(t-24)+A(t-12)\big)\Big(\exp(-g\,J(t-24))+g\,J(t)\Big)+\hat{r}}$

$a = 6.59$, $b = 9.36$, $c = 1.12$, $d = 1.77$, $f = 0.29$, $g = 0.04$, $h = 16.16$, $j = 1.63$, $k = 16.29$, $l = 11.76$, $m = 1.44$, $n = 63.18$, $o = 18.66$, $p = 63.06$, $q = 18.62$, $r = 11.75$, $s = 0.2$, $u = 3.25$, $v = 0.03$, $w = 3.75$, $x = 3.24$, $y = 17.39$, $z = 2$, $\hat{a} = 136.46$, $\hat{b} = 33.94$, $\hat{c} = 17.38$, $\hat{d} = 0.71$, $\hat{e} = 0.02$, $\hat{f} = 0.11$, $\hat{g} = 2.86$, $\hat{h} = 0.73$, $\hat{i} = 2.85$, $\hat{j} = 12.64$, $\hat{k} = 1.48$, $\hat{l} = 44.28$, $\hat{m} = 0.05$, $\hat{n} = 6.31$, $\hat{o} = 12.08$, $\hat{p} = 1.43$, $\hat{q} = 50.27$, $\hat{r} = 7.8$

Across the eight selected equations in the no stars and with stars runs, the direction of each retained variable's effect on predicted per capita growth rate was highly consistent for most terms, with some notable exceptions at higher complexities. $A(t-6)$, the most consistently retained term, had a negative effect on predicted growth in every equation and at every observed data point. Current-value adult and juvenile kelp abundance ($A(t)$, $J(t)$) and longer adult lags ($A(t-12)$, $A(t-24)$) were positive in 94-100% of observed data points across nearly all equations in which they appeared. $J(t-24)$ was negative in every equation where it appeared except the no stars equation at complexity 26, where its effect was mixed (25.4% positive). Sign consistency was lower at the two highest selected complexities (20 and 26), where several terms fell to 66–94% directional consistency rather than the near-uniform signs seen at lower complexity. In particular, the no stars and with stars equations at complexity 26 differed in the direction of $J(t)$'s effect (66.1% positive in the no stars equation versus 99.6% in the with stars equation), a further divergence beyond their structural difference.

3.3.2.3 Evaluation Step 2: Per capita growth rate reconstruction

Predicted versus observed per capita growth rate (Fig. 3.5) showed a clear improvement with increasing complexity. At complexity 4 and 7, predicted values were compressed into a narrow range and failed to capture much of the observed variation, particularly at higher observed growth rates where the equations consistently under-predicted. Complexity 11 showed a wider range but still largely

deviated from the 1:1 line across much of the data. From complexity 13 onward, predicted values tracked the 1:1 line more closely and captured a broader portion of the observed range, with complexity 13, 15, 17, 20, and both the no stars and with stars equations at complexity 26 showing the tightest correspondence between predicted and observed values among the selected equations.

Residuals versus fitted values (Fig. 3.6) were consistent with this pattern. Complexities 4 and 7 showed a banded residual structure with substantial spread at low predicted values, reflecting both equations' dependence on a single lagged adult kelp term that took a limited set of repeated values across the dataset. Complexity 11's residuals were more broadly scattered but still showed larger spread at low predicted values than the higher-complexity equations. Complexity 13, 15, 17, 20, and 26 (both equations) showed residuals more evenly distributed around zero across a range of predicted values.

Based on this reconstruction and residual patterns, we carried the equations from complexity 13, 15, 17, 20, and both the no stars and with stars equations at complexity 26 forward to population size forecasting, excluding equations from complexities 4, 7, and 11 due to their comparatively poor reconstruction and larger, more structured residuals.

3.3.2.4 Evaluation Step 3: Population size forecasting

Population size forecasting performance did not track per capita growth rate reconstruction quality (Evaluation Step 2). Full results for all six selected equations

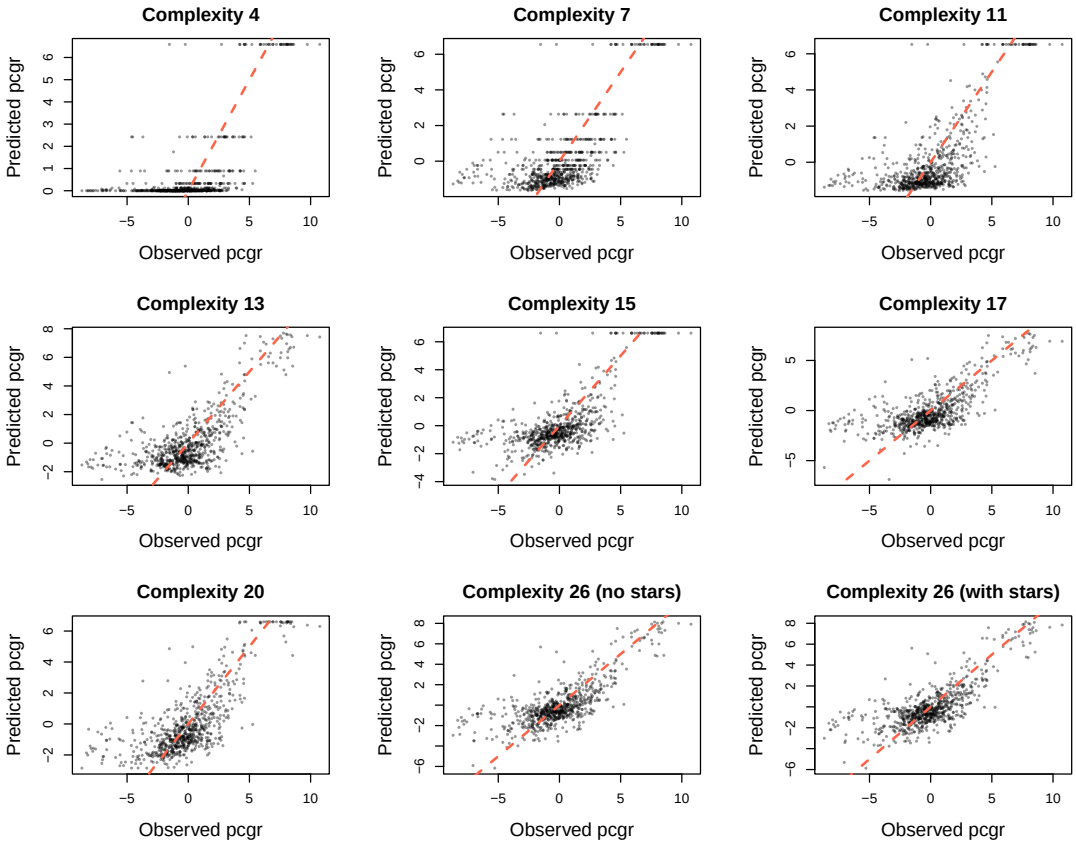


Figure 3.5: Predicted versus observed per capita growth rate (pcgr) for the selected equations, using the no stars run's coefficients except where noted. The complexity 26 panels show both the no stars and with stars equations, since this was the one selected complexity where the equation structure differed between the two runs. The dashed red line indicates the 1:1 line of perfect prediction.

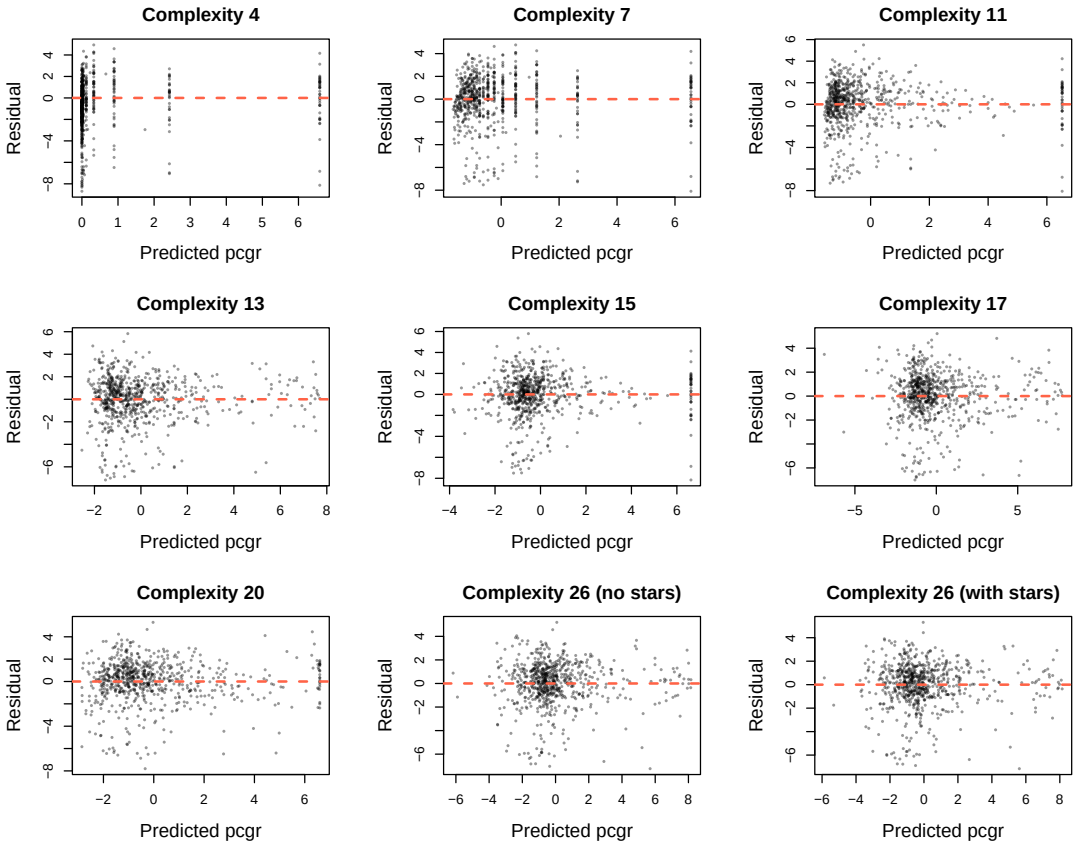


Figure 3.6: Residuals versus fitted values for the selected equations, using the no stars run's coefficients except where noted. The dashed red line marks zero residual.

across all four randomly selected transects (two from each site) are reported in Table 3.2. The equations that most tightly reconstructed the per capita growth rate (complexity 20 and 26) performed worst under forward integration, while lower-complexity equations (13 and 17) that showed looser reconstruction fits proved more stable when simulated forward in time. Complexity 13 achieved the best overall performance among the six equations in forecasting the West Dutch 10R transect trajectory with $R^2 = 0.249$ (Fig. 3.7). However, this same equation, along with complexity 17, failed to forecast the East Dutch 22R transect trajectory ($R^2 = -0.103$ and -0.105, respectively), indicating that forward-simulation performance depended on transect-specific dynamics in addition to equation choice. This transect-dependence may reflect differences in the underlying dynamics between transects rather than chaotic sensitivity to initial conditions. Complexity 20 performed worst among all six equations, with the abundance moving to zero early on in the simulation for all four transects tested. Complexity 26's no stars and with stars equations produced comparably poor performance to each other across transects, with R^2 negative for almost all transect/equation combinations tested (e.g., at West Dutch transect 10R, $R^2 = -0.266$ and -0.212, respectively), consistent with the similar per capita growth rate reconstruction and sign-consistency behavior these two formulations showed despite their algebraic structure noted in *Evaluation Step 1: Variable selection and equation formulation.*

Table 3.2: R^2, RMSE, and NRMSE for all six selected equations across all four transects assessed for population size forecasting. Metrics were calculated only over the forward-simulated period following the solver handoff; observations used to construct the history function seeding each equation's initial condition were excluded from these calculations.

Equation	Transect	R^2	RMSE	NRMSE
Complexity 13	West Dutch 10R	0.249	3.717	0.162
Complexity 13	West Dutch 32L	0.056	7.524	0.116
Complexity 13	East Dutch 22R	-0.103	13.740	0.222
Complexity 13	East Dutch 39R	0.095	6.306	0.170
Complexity 15	West Dutch 10R	0.062	4.155	0.181
Complexity 15	West Dutch 32L	0.015	7.684	0.118
Complexity 15	East Dutch 22R	-0.405	15.510	0.250
Complexity 15	East Dutch 39R	0.027	6.539	0.177
Complexity 17	West Dutch 10R	0.202	3.832	0.167
Complexity 17	West Dutch 32L	0.031	7.620	0.117
Complexity 17	East Dutch 22R	-0.105	13.760	0.222
Complexity 17	East Dutch 39R	0.196	5.945	0.161
Complexity 20	West Dutch 10R	-0.917	5.941	0.258
Complexity 20	West Dutch 32L	-0.172	8.381	0.129
Complexity 20	East Dutch 22R	-1.041	18.690	0.301
Complexity 20	East Dutch 39R	-0.772	8.824	0.238
Complexity 26 (no stars)	West Dutch 10R	-0.266	4.827	0.210
Complexity 26 (no stars)	West Dutch 32L	-0.136	8.253	0.127
Complexity 26 (no stars)	East Dutch 22R	-0.379	15.360	0.248
Complexity 26 (no stars)	East Dutch 39R	0.067	6.404	0.173
Complexity 26 (with stars)	West Dutch 10R	-0.212	4.723	0.205
Complexity 26 (with stars)	West Dutch 32L	-0.119	8.191	0.126
Complexity 26 (with stars)	East Dutch 22R	-0.298	14.910	0.240
Complexity 26 (with stars)	East Dutch 39R	-0.168	7.163	0.194

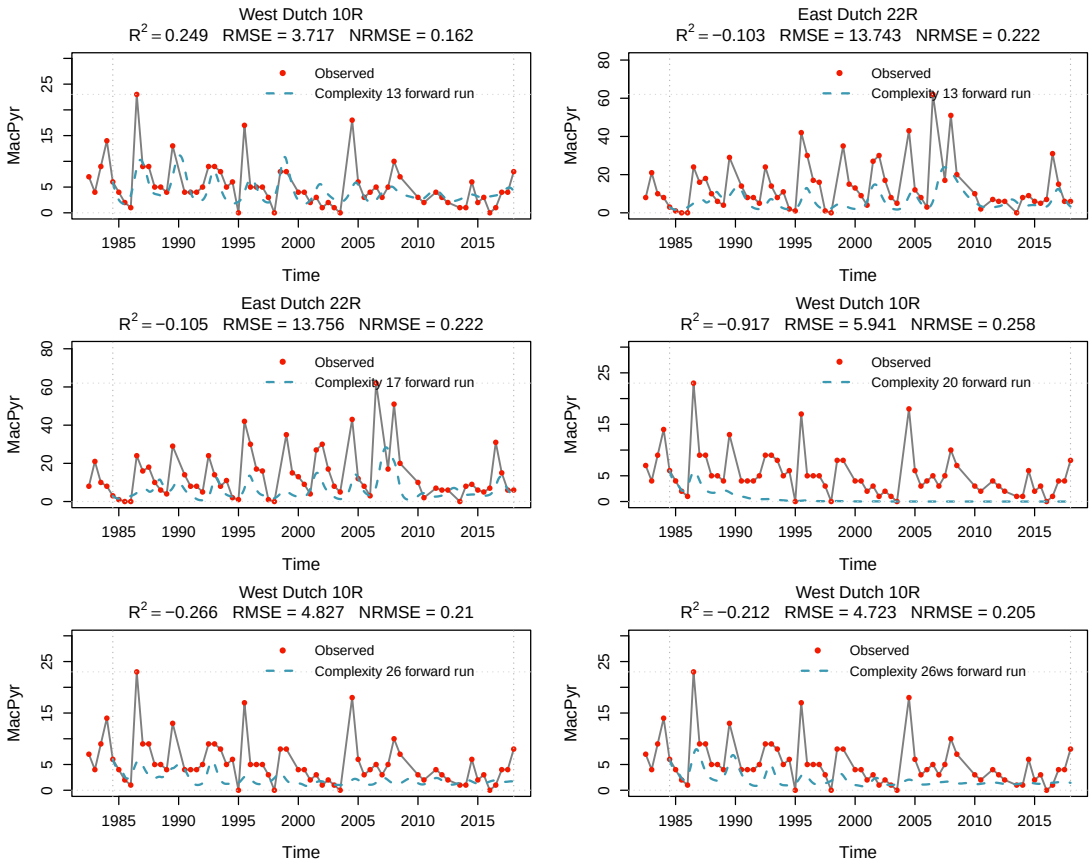


Figure 3.7: Population size forecasting for six equations × 4 randomly selected transects (24 combinations total) to represent the range of forward-simulation performance observed: complexity 13 and 17 at West Dutch transect 10R, and East Dutch transect 22R, complexity 20 at West Dutch transect 10R, and complexity 26's no stars and with stars equations at West Dutch transect 10R. Red points show the observed time series; the dashed line shows the equation's forward-simulated trajectory beginning at the point the solver takes over from the empirical data. R^2, RMSE, and NRMSE are reported for each panel.

3.4 Discussion

Symbolic regression consistently identified adult kelp abundance, particularly at a 6-month lag, as the dominant predictor of per capita growth rate across independent searches, complexity levels, and both the no stars and with stars runs. Equations retaining this term showed a negative relationship with predicted growth in every equation and at every observed data point, while current-value adult and juvenile kelp density and longer adult lags were predominantly positive. Notably, variable selection remained stable across the full range of complexity levels explored. The ten lowest-loss equations converged on a single shared functional form through complexity 19 in both variable sets, with only isolated exceptions (complexity 14-16) before diversity increased consistently from complexity 20 onward. The same set of flagged equations was recovered from the Pareto frontier in both variable sets at every complexity level except the highest (complexity 26), where the two runs converged on different functional forms using the same variables. Per capita growth rate reconstruction improved steadily with complexity, but this did not translate into better population size forecasts. This separation highlights that reconstruction of the per capita growth rate and forecasting of population size probe distinct functions of the candidate equations, and that strong performance on one does not guarantee suitability for the other (Song, 2026). The equations that most tightly reconstructed the observed relationship between per capita growth rate and population size (complexity 20 and 26) performed worst under forward integration, while comparatively simpler equations (complexity 13 and 17) avoided

this divergent behavior when simulated forward in time, despite showing poorer fit in the reconstruction. Forward-simulation performance also varied by transect independent of equation choice, with East Dutch 22R transect proving difficult to forecast regardless of which equation was used.

Beyond their statistical performance, the selected equations suggest a consistent biological narrative. Across complexities 13, 17, 20, and 26, per capita growth rate declines with adult kelp abundance six months prior. It increases with current adult and juvenile kelp abundance and with longer adult lags, a pattern consistent with self-thinning at high density followed by recovery as the cohort ages past the density-dependent bottleneck (but see limitations paragraph below). The recurring appearance of $J(t-24)$, generally with a negative effect, further suggests that juvenile density two years prior constrains present-day growth. Because juveniles from 24 months ago are a cohort that would later contribute to the canopy formation, this term may act as a proxy for the resulting canopy density and its persistent structural effects on the substrate (Bence and Nisbet, 1989) or some other limitation. These patterns are broadly consistent with density-dependent self-regulation previously documented in giant kelp populations (Graham, 1997; Santelices and Ojeda, 1984), lending biological feasibility to the recovered equations beyond their MSE values. However, because our lags also coincide with seasonal cycles, we cannot fully rule out that some portion of this signal reflects seasonal forcing rather than density dependence per se (see limitations paragraph below).

Across both aggregated Pareto frontiers and the top-10 performing equations at every complexity level, bat stars (which we included as a variable with no known di-

rect mechanistic link to giant kelp population dynamics) were retained only rarely, and never at any of the eight complexity levels selected for detailed evaluation. Where bat star terms did appear elsewhere along the with stars run frontier, their inclusion did not correspond to any meaningful improvement in fit. This result supports the reliability of the method's variable selection: a spurious biotic variable with no established interaction with kelp was not promoted to explanatory status.

In contrast to bat stars, sea urchins are a biotic variable with an established, direct grazing relationship to giant kelp, yet urchin terms were not retained in any top-performing equation, at any complexity level, in either variable set. This absence is plausible given the ecological context of this dataset. In a persistent kelp forest such as the one sampled here, sea urchins are known to shift toward passive grazing on drift algae rather than active grazing on standing kelp when kelp abundance is high (Harrold and Reed, 1985; Filbee-Dexter and Scheibling, 2014). This site is additionally characterized by high substrate rugosity, which retains higher concentrations of drift algae and thus further sustains consumption of drift algae rather than live kelp, likely weakening any signal linking urchin abundance to kelp per capita growth rate (Randell et al., 2022). The absence of sea urchins from the recovered equations therefore likely reflects a genuine, site-specific decoupling between sea urchin abundance and kelp dynamics under persistent, high rugosity conditions, rather than a failure of symbolic regression to detect a real relationship.

Several aspects of this approach constrain what can be concluded from the recovered equations. First, the temporal resolution of the survey data (approxi-

mately every 6 months) sets the resolution of every lag term considered, including the dominant $A(t-6)$ term retained across nearly all top-performing equations. Because 6-month lags coincide with the seasonal cycle between early fall (when kelp population size is typically highest) and early spring (when kelp population size is typically lowest), we cannot say with full certainty whether this lag represents a genuine delayed density-dependent process, a seasonal effect correlated with survey timing, or some combination of the two. Distinguishing between these possibilities would require either a finer or seasonally stratified sampling regime.

Second, the choice of using lags was motivated by Takens' theorem, which establishes that the dynamics of a system can be reconstructed from time-delayed observations of a single state variable, given a sufficient number of lags spaced at an appropriate, consistent delay interval. This justifies using lagged kelp abundance as a proxy for unobserved state variables when reconstructing the system's attractor, but it does not follow that the specific variables and lags retained in a given equation correspond to causal drivers of the underlying dynamics. A lagged variable's inclusion in a recovered equation may reflect a genuine mechanistic relationship, a proxy for another correlated but unmeasured driver, or an artifact of the particular embedding used (Sugihara et al., 2012; Runge et al., 2019), and these possibilities cannot necessarily be distinguished within the symbolic regression framework. Establishing whether the relationships identified in this study are causal, rather than simply descriptive of a correctly reconstructed attractor, would require complementary techniques explicitly designed to separate correlation from causation, such as convergent cross mapping or other causal-inference methods

suited to time series data (Munch et al., 2022; Runge et al., 2019; Suzuki et al., 2022).

This study used symbolic regression to recover candidate equations governing giant kelp population dynamics directly from long-term, biannual field survey data, without assuming a mechanistic model *a priori*. The recovered equations consistently identified biologically established drivers of kelp population growth while excluding a variable with no known mechanistic relevance, providing indirect support for the reliability of this method's variable selection even in the absence of a known governing equation. This reliability held even as the functional form of the top-performing equations diversified with increasing complexity, suggesting that consistency in variable selection may be a more robust signal of a meaningful predictor than convergence on any single equation structure. At the same time, the finding that the per capita growth rate reconstruction and population size forecasting did not agree in which equations performed best indicates that fit to an empirical attractor is, on its own, an insufficient criterion for selecting equations intended for forward simulation, a discrepancy that warrants further investigation in other systems and datasets. Symbolic regression applied to empirical time series therefore represents a promising complement to traditional mechanistic modeling, particularly in theory-poor systems, bridging model-driven and data-driven approaches to ecological inference by allowing structure to be discovered from data while still yielding an interpretable, testable equation (Song, 2026).

Supplementary Materials for

Identifying biotic drivers of persistent giant kelp *Macrocystis pyrifera* population dynamics

Cheyenne N. Jarman & Mark Novak

A Pareto frontier for with stars run

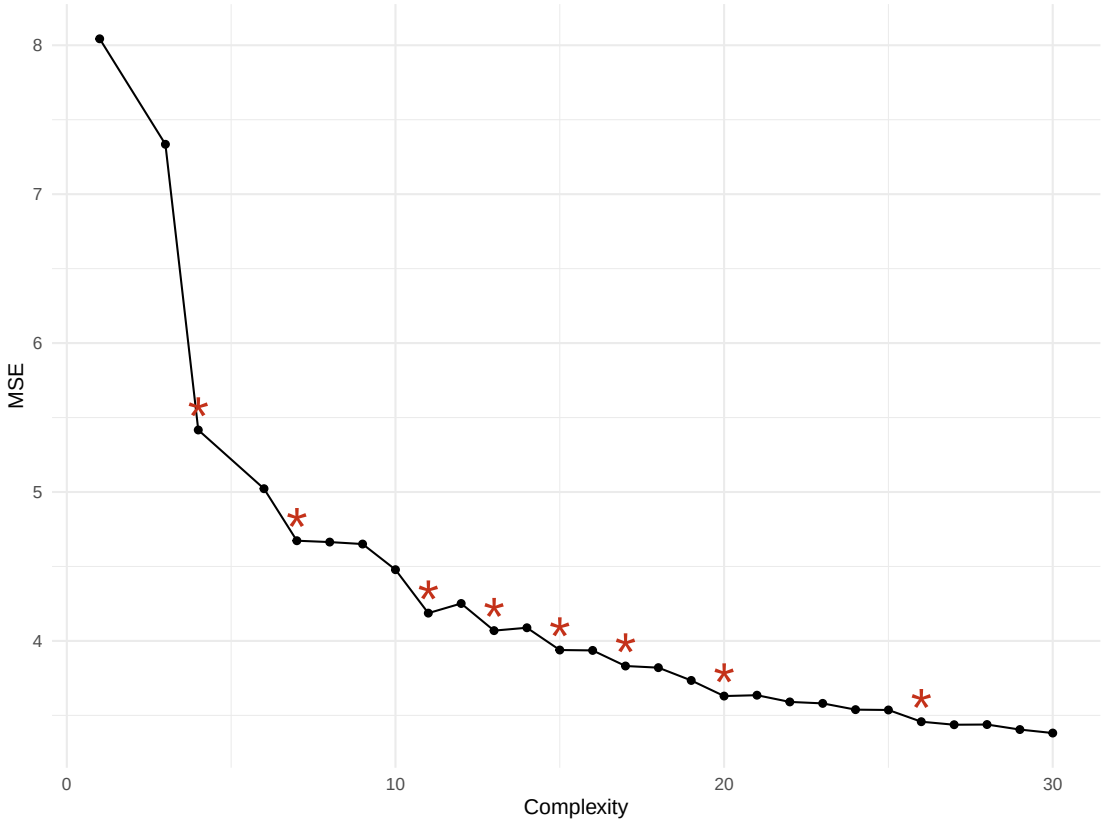


Figure S.1: Pareto frontier for the with stars PySR run, showing the best-performing equation's loss (MSE) at each complexity levels across 100 independent runs. The red stars indicate equations where increasing complexity level produced a relatively large drop in loss relative to the preceding step. These were the complexities we looked at for our further analyses. The no stars run produced a near-identical frontier and selected the same eight complexities (see Pareto frontier in the main text).

Bibliography

Angelini, C., Altieri, A. H., Silliman, B. R., and Bertness, M. D. (2011). Interactions among Foundation Species and Their Consequences for Community Organization, Biodiversity, and Conservation. *BioScience*, 61(10):782–789.

Bell, T. W., Cavanaugh, K. C., Reed, D. C., and Siegel, D. A. (2015). Geographical variability in the controls of giant kelp biomass dynamics. *Journal of Biogeography*, 42(10):2010–2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jbi.12550.

Bence, J. R. and Nisbet, R. M. (1989). Space-Limited Recruitment in Open Systems: The Importance of Time Delays. *Ecology*, 70(5):1434–1441. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.2307/1938202.

Bjørnstad, O. N. and Grenfell, B. T. (2001). Noisy Clockwork: Time Series Analysis of Population Fluctuations in Animals. *Science*, 293(5530):638–643.

Castorani, M. C. N., Bell, T. W., Walter, J. A., Reuman, D. C., Cavanaugh, K. C., and Sheppard, L. W. (2022). Disturbance and nutrients synchronise kelp forests across scales through interacting Moran effects. *Ecology Letters*, 25(8):1854–1868. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ele.14066.

Chen, Y., Angulo, M. T., and Liu, Y.-Y. (2019). Revealing Complex Ecological Dynamics via Symbolic Regression. *BioEssays*, 41(12):1900069. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bies.201900069.

Cranmer, M. (2023). Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl. arXiv:2305.01582 [astro-ph, physics:physics].

Dayton, P. K. (1972). Toward an understanding of community resilience and the potential effects of enrichments to the benthos at McMurdo Sound, Antarctica. *Proceedings of the Colloquium on Conservation Problems Allen Press*, pages 81–96.

Detmer, A. R., Miller, R. J., Reed, D. C., Bell, T. W., Stier, A. C., and Moeller, H. V. (2021). Variation in disturbance to a foundation species structures the dynamics of a benthic reef community. *Ecology*, 102(5):e03304. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ecy.3304.

Edwards, M. S. and Konar, B. (2020). Trophic downgrading reduces spatial variability on rocky reefs. *Scientific Reports*, 10(1):18079.

Ellison, A. M., Bank, M. S., Clinton, B. D., Colburn, E. A., Elliott, K., Ford, C. R., Foster, D. R., Kloeppel, B. D., Knoepp, J. D., Lovett, G. M., Mohan, J., Orwig, D. A., Rodenhouse, N. L., Sobczak, W. V., Stinson, K. A., Stone, J. K., Swan, C. M., Thompson, J., Von Holle, B., and Webster, J. R. (2005). Loss of foundation species: consequences for the structure and dynamics of forested ecosystems. *Frontiers in Ecology and the Environment*, 3(9):479–486.

Filbee-Dexter, K. and Scheibling, R. (2014). Sea urchin barrens as alternative stable states of collapsed kelp ecosystems. *Marine Ecology Progress Series*, 495:1–25.

Fritsch, F. N. and Carlson, R. E. (1980). Monotone Piecewise Cubic Interpolation. *SIAM Journal on Numerical Analysis*, 17(2):238–246.

Giraldo-Ospina, A., Bell, T., Carr, M. H., and Caselle, J. E. (2025). Drivers of spatiotemporal variability in a marine foundation species. *Ecological Applications*, 35(1):e3092. _eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1002/eap.3092.

Graham, M. H. (1997). Factors determining the upper limit of giant kelp, Macrocystis pyrifera Agardh, along the Monterey Peninsula, central California, USA. *Journal of Experimental Marine Biology and Ecology*, 218(1):127–149.

Harrold, C. and Reed, D. C. (1985). Food Availability, Sea Urchin Grazing, and Kelp Forest Community Structure. *Ecology*, 66(4):1160–1169.

Hovhannisyan, K. (2023). Inference of Bacterial Growth Dynamics via Symbolic Regression. Master's thesis, University of Helsinki.

Jarman, C. N., Levi, T., and Novak, M. (2026). Symbolic regression for empirically realistic population dynamic time series. *Ecological Informatics*, 99:103988.

Karatayev, V. A., Baskett, M. L., Kushner, D. J., Shears, N. T., Caselle, J. E., and Boettiger, C. (2021). Grazer behaviour can regulate large-scale patterning of community states. *Ecology Letters*, 24(9):1917–1929. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ele.13828.

Kendall, B. E., Briggs, C. J., Murdoch, W. W., Turchin, P., Ellner, S. P., McCauley, E., Nisbet, R. M., and Wood, S. N. (1999). Why Do Populations Cycle? A Synthesis of Statistical and Mechanistic Modeling Approaches. *Ecology*, 80(6):1789–1805. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1890/0012-9658%281999%29080%5B1789%3AWDPCAS%5D2.0.CO%3B2.

Kenner, M. C., Estes, J. A., Tinker, M. T., Bodkin, J. L., Cowen, R. K., Harrold, C., Hatfield, B. B., Novak, M., Rassweiler, A., and Reed, D. C. (2013). A multi-decade time series of kelp forest community structure at san nicolas island, california. *Ecological Archives E*, 94:244.

Kenner, M. C. and Tinker, M. T. (2018). Stability and Change in Kelp Forest Habitats at San Nicolas Island. *Western North American Naturalist*, 78(4):633.

Krumhansl, K. A., Okamoto, D. K., Rassweiler, A., Novak, M., Bolton, J. J., Cavanaugh, K. C., Connell, S. D., Johnson, C. R., Konar, B., Ling, S. D., Micheli, F., Norderhaug, K. M., Pérez-Matus, A., Sousa-Pinto, I., Reed, D. C., Salomon, A. K., Shears, N. T., Wernberg, T., Anderson, R. J., Barrett, N. S., Buschmann, A. H., Carr, M. H., Caselle, J. E., Derrien-Courtel, S., Edgar, G. J., Edwards, M., Estes, J. A., Goodwin, C., Kenner, M. C., Kushner, D. J., Moy, F. E., Nunn, J., Steneck, R. S., Vásquez, J., Watson, J., Witman, J. D., and

Byrnes, J. E. K. (2016). Global patterns of kelp forest change over the past half-century. *Proceedings of the National Academy of Sciences*, 113(48):13785–13790.

Lamy, T., Reed, D. C., Rassweiler, A., Siegel, D. A., Kui, L., Bell, T. W., Simons, R. D., and Miller, R. J. (2018). Scale-specific drivers of kelp forest communities. *Oecologia*, 186(1):217–233.

Levin, S. A. (1992). The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture. *Ecology*, 73(6):1943–1967. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.2307/1941447.

Martin, B. T., Munch, S. B., and Hein, A. M. (2018). Reverse-engineering ecological theory from data. *Proceedings of the Royal Society B: Biological Sciences*, 285(1878):20180422.

Miller, R. J., Lafferty, K. D., Lamy, T., Kui, L., Rassweiler, A., and Reed, D. C. (2018). Giant kelp, *Macrocystis pyrifera* , increases faunal diversity through physical engineering. *Proceedings of the Royal Society B: Biological Sciences*, 285(1874):20172571.

Munch, S. B., Rogers, T. L., and Sugihara, G. (2022). Recent developments in empirical dynamic modelling. *Methods in Ecology and Evolution*, n/a(n/a). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13983.

Parnell, P. E., Lennert-Cody, C. E., Ladah, L. B., Riser, K. L., Bulach, B., Leichter, J. J., Latker, A. K., Schroeter, S. C., and Dayton, P. K. (2026). Demography and dynamics of giant kelp cohorts across four decades: Lessons for conservation and resilience planning. *Ecological Applications*, 36(1):e70181.

Randell, Z. (2021). *Grazer-resource Behavior, Interactions, and Feedbacks Control Kelp-forest Dynamics and Stability*. Ph.D., Oregon State University.

Randell, Z., Kenner, M., Tomoleoni, J., Yee, J., and Novak, M. (2022). Kelp-forest dynamics controlled by substrate complexity. *Proceedings of the National Academy of Sciences*, 119(8):e2103483119.

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., van Nes, E. H., Peters, J., Quax, R., Reichstein, M., Scheffer, M., Schölkopf, B., Spirtes, P., Sugihara, G., Sun, J., Zhang, K., and Zscheischler, J. (2019). Inferring causation from time series in Earth system sciences. *Nature Communications*, 10(1):2553.

Santelices, B. and Ojeda, F. (1984). Population dynamics of coastal forests of Macrocystis pyrifera in Puerto Toro, Isla Navarino, Southern Chile. *Marine Ecology Progress Series*, 14:175–183.

Schiel, D. R. and Foster, M. S. (2015). *The Biology and Ecology of Giant Kelp Forests*.

Song, C. (2026). Toward a conceptual commons for theory in community ecology. *Annual Review of Ecology, Evolution, and Systematics*, 57.

Sugihara, G., May, R., Ye, H., Hsieh, C.-h., Deyle, E., Fogarty, M., and Munch, S. (2012). Detecting Causality in Complex Ecosystems. *Science*, 338(6106):496–500.

Suzuki, K., Matsuzaki, S.-i. S., and Masuya, H. (2022). Decomposing predictability to identify dominant causal drivers in complex ecosystems. *Proceedings of the National Academy of Sciences*, 119(42):e2204405119.

Takens, F. (1981). Detecting strange attractors in turbulence. In Rand, D. and Young, L.-S., editors, *Dynamical Systems and Turbulence, Warwick 1980*, volume 898, pages 366–381. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Mathematics.

Turchin, P. (1999). Population Regulation: A Synthetic View. *Oikos*, 84(1):153.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones,

E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Watson, J. and Estes, J. A. (2011). Stability, resilience, and phase shifts in rocky subtidal communities along the west coast of Vancouver Island, Canada. *Ecological Monographs*, 81(2):215–239.

Chapter 4: Recovering and forecasting simulated coral reef dynamics with universal dynamic equations

Abstract

Understanding the dynamics of ecological communities requires recovering the mechanisms that drive species interactions, while forecasting requires only that a model's predictions be accurate. These two goals demand different kinds of model performance, and strong performance on one does not guarantee strong performance on the other. Universal dynamic equations (UDEs), which embed trainable neural networks within a mechanistic differential equation framework, offer a promising way to combine theoretical structure with data-driven flexibility, but their reliability under the sparse, noisy conditions typical of ecological monitoring remains poorly understood. We evaluated UDEs on both fronts – recovering the functional form of an unknown species interaction and forecasting future system states – using synthetic data generated from a coral reef model, a system well served by existing mechanistic theory yet characteristically data-limited in practice. We compared two UDE formulations that differ in how much mechanistic structure is fixed versus co-estimated alongside the unknown term, across a range of sampling frequencies, initial conditions, and process noise and observation error levels. We assessed forecasting under both internal validation (a withheld portion of the

training trajectory) and external validation (independently simulated trajectories never seen during training). Recovery of the unknown term met our threshold in 28.6% of runs for the fixed parameter formulation and 4.5% for the co-estimated formulation, whereas the co-estimated formulation showed a modest internal validation advantage. Internal validation forecast success declined markedly under high noise, particularly high observation error. Under deterministic training data, longer transient periods improved term recovery and external validation. No UDE met our quantitative threshold under external validation, although the true model did, and mean UDE state-match rates (57–58%) fell between those of a constant-state forecast (50%) and a forecast based on initial dominance (75%). Together, these results show that forecasting a trajectory does not establish that a UDE has learned its mechanism, underscoring the need to stress-test these tools against known dynamics and simple baselines before their outputs are used with confidence in ecological monitoring and management.

4.1 Introduction

Ecological communities are dynamic systems whose behavior emerges from many interacting species, feedbacks, and external drivers, making them difficult to understand mechanistically and forecast reliably, two central goals of ecology (Doak et al., 2008; Beckage et al., 2011). These goals — recovering the mechanisms that govern species interactions and predicting how systems will change over time — have historically demanded different modeling philosophies. The mechanis-

tic philosophy treats a model as a hypothesis about how low-level processes can lead to emergent behaviors at higher levels of biological organization. It starts from first principles, deriving the model from assumptions about how a subset of these processes, like birth, death, and predation, drive population change. It then judges success by how well the model captures emergent behaviors (e.g., the Lotka-Volterra predator-prey model capturing population cycles (Lotka, 1925; Volterra, 1926; Elton and Nicholson, 1942)). The predictive philosophy treats a model as a tool for anticipating outcomes. It starts from data, imposes comparatively little structure, and judges success by accuracy on new observations regardless of whether the model's internal structure corresponds to anything mechanistic (Breiman, 2001; Shmueli, 2010). Because the two philosophies define success differently, a model that succeeds under one need not succeed under the other.

In ecology, these philosophies are instantiated in two broad modeling traditions with complementary strengths and limitations. Mechanistic models (e.g., Lotka-Volterra type or other multi-species differential equation models of species interactions) offer interpretable, parsimonious representations of system processes grounded in ecological theory, but are frequently criticized for oversimplifying the dynamics they aim to capture (Duarte et al., 2003). Because their structure is meant to reflect the mechanisms actually generating the dynamics, rather than patterns specific to a given dataset, mechanistic models are in principle able to generalize across a wide range of conditions, including conditions absent from the data used to build them (Petchey et al., 2015; Urban et al., 2016). In contrast, data-driven approaches, for example, statistical or machine learning models fit directly

to time series, prioritize predictive accuracy over mechanistic interpretability, but the fitted structure can be difficult to interpret and often sensitive to data quality, sampling resolution, and overfitting. These models are typically optimized to reproduce patterns within the range of conditions represented in the training data, and their accuracy is not guaranteed (and often degrades) when applied outside that range (Petchey et al., 2015; Urban et al., 2016). Each tradition therefore addresses one of the two goals more readily than the other, leaving a gap that either alone struggles to close.

Scientific machine learning (SciML) has emerged as a promising synthesis of these two traditions, producing models that leverage theoretical structure while using observed data to inform model form and parameterization (Rackauckas et al., 2021; Karniadakis et al., 2021). Within SciML, universal dynamic equations (UDEs) offer a particularly flexible approach, embedding trainable model components directly within an otherwise mechanistic dynamical system so that unknown or poorly understood terms can be learned from data while known structure is preserved. Their continuous-time form, universal differential equations, uses neural networks as the trainable components within a differential equation (Rackauckas et al., 2021; Buckner et al., 2026). This hybrid architecture gives the modeler meaningful control. Ecological intuition and prior theory can be encoded into the known terms, while the data inform recovery of terms that are difficult to observe or characterize directly. UDEs are therefore potentially well-suited to systems where theory is rich, but data are limited or noisy, conditions common in empirical ecology. Because this method is new to ecology, however, it remains un-

clear whether this potential is realized in practice, and the answer may depend on which of ecology's two central goals a UDE is asked to serve. Learning an unknown term from data (i.e., term recovery) is a mechanistic goal, whereas forecasting future system states is a predictive one, and success at one does not guarantee success at the other, so performance on each must be assessed independently before UDEs can be confidently deployed.

Whether either goal can be achieved also depends on how UDEs respond to the limitations of real ecological data. The first is uncertainty, which comes in two forms with different consequences for model fitting. Process noise refers to variation in the rates of natural processes (e.g., birth, death, predation) and can perturb a system's actual state, whereas observation error obscures the true state without altering the dynamics. Not all noise is a nuisance to be minimized (Boettiger, 2018; de Valpine and Hastings, 2002; Jarman et al., 2026). Process noise can continually push a system away from equilibrium, exposing dynamics that a smooth, noise-free trajectory would never reveal and potentially making the underlying process easier to recover, whereas observation error adds uncertainty without adding information. UDEs implemented within a state-space framework are designed to accommodate both, separately estimating the true, unobserved system state and the error in its observation (Buckner et al., 2026), but how well this separation holds as each source of uncertainty grows is unknown. The second limitation is sampling frequency. Monitoring data are often collected infrequently because of logistical constraints, and how often a system is sampled can strongly affect whether its underlying mechanisms can be recovered (Jarman et al., 2026),

because sparse observations may miss the changes that reveal how a process operates. The third limitation concerns which dynamics a monitoring record captures. Data-driven approaches, including SciML methods, have shown that chaotic and oscillating systems, often assumed to be hard to model, can be more identifiable or forecastable than expected, because their sustained, recurring dynamics repeatedly expose a model to the processes generating them (Buckner et al., 2026; Rogers et al., 2022; Bonnaffé et al., 2021). Systems with alternative stable states offer a different and comparatively unexplored source of informative dynamics through transients, whose length depends on how close a trajectory begins to the boundary between basins of attraction (Hastings et al., 2018). A trajectory initialized near that boundary passes through an extended transient before settling into one state, exposing it to a broader region of state space, whereas a trajectory initialized well within a basin approaches equilibrium quickly. Because monitoring records are often short relative to these timescales, how much of a transient a record captures may shape how much it reveals about the underlying dynamics.

Coral reefs offer a useful system in which to examine these limitations. They are among the most ecologically and economically significant yet threatened marine ecosystems, facing accelerating degradation from climate change, overfishing, and local stressors (Hughes et al., 2017). Reefs are served by a rich tradition of theoretical modeling that provides strong mechanistic priors (Blackwood et al., 2018), yet their monitoring is constrained by logistical and financial limits on the frequency and consistency of surveys and by uncertainty in how cover is estimated. They also exhibit well-documented, often abrupt transitions between coral-dominated

and macroalgae-dominated states (Fung et al., 2011; Tebbett et al., 2023; Mumby et al., 2007), so their dynamics include the boundary-dependent transients described above. These transitions are governed in part by grazing on macroalgae, a process central to reef state yet notoriously difficult to characterize mechanistically or quantify in the field (Bruggemann et al., 1994), making it the kind of poorly constrained but consequential term a UDE could learn from data. Whether a UDE can recover this grazing term from realistic reef data, and whether it can forecast reef state transitions, remains unresolved.

This study addresses three related questions. First, how do data quality, including the magnitude of process noise and observation error and the frequency of sampling, affect a UDE's ability to (a) recover an unknown term and (b) forecast community dynamics? We hypothesized that process noise would serve as an informative signal, improving both term recovery and forecasting success, whereas observation error and reduced sampling frequency would hinder both goals (H1). Second, does the length of the transient a training trajectory passes through before settling into a state, as determined by its initial condition, alter term recovery or forecasting success? We hypothesized that both goals would be most successful for models trained on trajectories with a longer transient, which expose a model to a broader region of state space (H2). Third, does constraining more of a UDE's mechanistic structure, by fixing known parameters rather than co-estimating them alongside the unknown term, create a trade-off between term recovery and forecasting success? Fixing known parameters constrains what the unknown term must account for, which could aid term recovery. The added structure could also help

forecasting by anchoring the model to values known to be correct, but documented trade-offs between models built to explain and models built to predict suggest that it may instead come at a cost (Shmueli, 2010). We therefore hypothesized that fixing known parameters would improve term recovery relative to co-estimating them, at the cost of forecasting success (H3).

To address these questions, we generated synthetic datasets from the Mumby et al. 2007 coral model (hereafter the Mumby et al. model), which let us compare what a UDE learned against a ground truth. We varied process noise, observation error, and sampling frequency to reflect common challenges in reef monitoring data, and varied initial conditions to place training trajectories at different distances from the basin boundary, thereby generating variation in transient length. We compared two UDE formulations that differ in whether the model's mechanistic parameters are fixed at their true values or co-estimated alongside an unknown grazing term. Term recovery was evaluated by comparing the grazing term each UDE learned to the true grazing function. We evaluated forecasting under two complementary validation schemes: internal validation, which forecasts a withheld later portion of the training trajectory, and external validation, which forecasts independently simulated trajectories never seen during training.

4.2 Methods

To evaluate UDE performance for term recovery and forecasting under ecologically realistic conditions, we constructed synthetic datasets from the Mumby et

al. model, a canonical representation of coral-macroalgae dynamics. We chose this model for its balance of simplicity and behavioral richness: with only two independent state variables and no spatial, age, or additional functional-group structure, it is among the simplest models capable of reproducing the coral-macroalgae bistability that motivates this study. Starting from a deterministic version of the model, we generated training datasets that varied in initial coral cover (four levels), sampling frequency (four levels), and either process noise in the grazing term or observation error in the state variables (three magnitudes each, plus a deterministic baseline with neither). Because process noise and observation error were applied separately rather than in combination, this yielded $4 \times 4 \times (1 + 3 + 3) = 112$ training conditions. We fit each training dataset with two UDE formulations, one with the Mumby et al. model parameters (excluding grazing) fixed at their true values and one with them co-estimated, for a total of 224 model fits, each trained on a single time series. We then evaluated every fit on term recovery and on forecasting under internal and external validation. We describe each of these components in turn below.

4.2.1 Data generation

In the Mumby et al. model, reef benthic cover can exist in one of two alternative states: a coral-dominated state, in which reef-building corals occupy substantial cover, and a macroalgae-dominated state, in which macroalgae outcompete corals for space and light. Which state a reef occupies, and whether it will transition be-

tween states, is governed by the balance between coral growth, macroalgae growth, and grazing. Herbivorous fish and invertebrates, which are not modeled explicitly but are represented by the grazing term, graze down macroalgae that would otherwise overgrow coral or turf, freeing space for coral larvae to settle and for existing coral to expand. When grazing pressure is high, corals can outcompete algae for space, but when grazing is suppressed (e.g., by overfishing or disease-driven sea urchin die-offs), macroalgae can overgrow both turf and coral, potentially locking the reef into the macroalgae-dominated state. Algal turf occupies whatever substrate is not already held by coral or macroalgae, and modulates the grazing pressure macroalgae experience.

Mathematically, the Mumby et al. model describes the dynamics of coral ($C(t)$), macroalgae ($M(t)$), and algal turf ($T(t)$) and is defined by the following coupled ordinary differential equations:

$$\frac{dC}{dt} = rTC - dC - aMC \tag{4.1a}$$

$$\frac{dM}{dt} = aMC - \gamma(M,T) + qMT \tag{4.1b}$$

where $T = 1 - C - M$ and $C, M, T \in [0,1]$. Because turf cover is fully determined by coral and macroalgae cover, the system's dynamics are governed by two independent state variables. The grazing term is given by

$$\gamma(M,T) = \frac{gM}{M+T} \tag{4.2}$$

The parameters r, d, a, and q represent the rate of coral overgrowth on turf, coral mortality, macroalgal overgrowth on coral, and macroalgal overgrowth on turf, respectively, while g represents the grazing pressure governing $\gamma(M, T)$, the rate at which algae is grazed (with $g = g_0 = 0.3$ in the deterministic core; Table S.1).

We refer to Equation 4.1 as the deterministic core. It represents an idealized case in which reef dynamics evolve exactly as prescribed by the model equations, with no process noise and no observation error. We simulated it from four initial coral cover values $C(0) = 0.1, 0.42, 0.44$, and 0.8, with initial turf fixed at 0.1, and initial macroalgae cover defined by $M(0) = 1 - 0.1 - C(0)$. Given the parameter values used, these initial conditions span the macroalgae-dominated basin ($C(0) = 0.1$), the coral-dominated basin ($C(0) = 0.8$), and the region near the separatrix dividing the two basins of attraction ($C(0) = 0.42$ or 0.44), where trajectories pass through extended transients before settling into a state (Fig. 4.1).

Real reef dynamics are never perfectly deterministic. Ecological rates fluctuate over time due to processes the model does not explicitly represent; for instance, short-term variability in herbivore abundance or foraging behavior, driven by factors like weather or predator activity, causes grazing pressure to deviate from a fixed constant. We represented this process noise by replacing the constant grazing rate with a stochastic one, so that grazing became

$$\gamma_p(M, T) = \frac{g_p(t)M}{M + T} \tag{4.3}$$

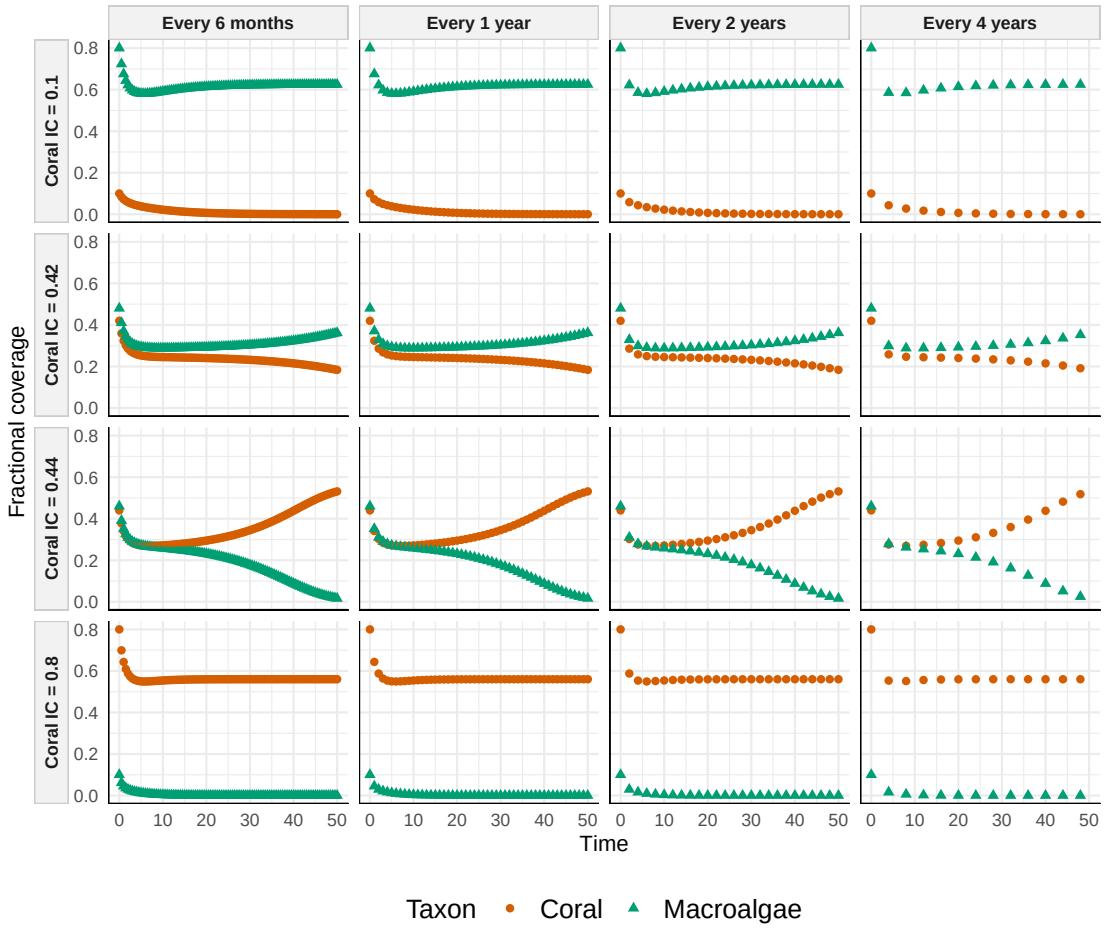


Figure 4.1: Example deterministic time series across all four coral initial condition values and all four sampling frequencies. Rows show initial coral cover (Coral IC), and columns show sampling frequency (every 6 months, every 1 year, every 2 years, and every 4 years, corresponding to $\delta t = 0.5, 1, 2$, and 4, respectively). Fractional coverage represents the proportion of benthic reef substrate occupied by each cover type. Points show coral (orange points) and macroalgae (green triangles) cover over time.

where $g_p(t)$ is a piecewise constant stochastic process:

$$g_p(t) = g_0 \exp(\epsilon_k), \quad \epsilon_k \sim \mathcal{N}(\frac{-\sigma_g^2}{2}, \sigma_g^2) \quad \text{for } t \in [t_k, t_{k+1}) \tag{4.4}$$

Here, $g_0 = 0.3$ is the baseline grazing rate, $t_{k+1} - t_k = 0.5$ years is the interval between independent draws, and σ_g sets the magnitude of the noise (0.2 = low, 0.6 = medium, and 0.8 = high; Fig. 4.2, top row). We centered the noise distribution at $\frac{-\sigma_g^2}{2}$ so that $\mathbb{E}[g_p(t)] = g_0$, preserving the expected grazing rate across noise levels. Process noise was simulated at each magnitude from all four initial conditions.

Even a perfectly deterministic reef would not be observed perfectly. Field surveys of coral cover typically rely on divers or photo-quadrat analysis to estimate the proportion of substrate occupied by coral, macroalgae, and turf along a transect, introducing both random imprecision (e.g., a diver's estimate varying with viewing angle) and systematic bias (e.g., one observer consistently scoring cover differently than another). We modeled only the imprecision component; systematic bias is not represented in our simulated observations.

To incorporate this imprecision, we added variability to the deterministic trajectories using the Dirichlet distribution (Frigyik et al., 2010), which generates draws centered on the true proportions $\mathbf{P}_t$ with dispersion controlled by a concentration parameter. We chose this distribution because it is multivariate and constrains values to lie between 0 and 1. At any point in time, all reef substrate in the Mumby et al. model must be classified as either coral, macroalgae, or turf, so the three cover values must sum to one ($C + M + T = 1$). The Dirichlet dis-

tribution naturally enforces that constraint while adding realistic error around the true proportions, making it consistent with the fractional coverage definitions of the Mumby et al. model. The noisy observations, $\mathbf{O}_t = \{C_{obs}, M_{obs}, T_{obs}\}$, were generated as

$$\mathbf{O}_t \sim \mathrm{Dir}(\alpha_0 \cdot \mathbf{P}_t) \tag{4.5}$$

where $\mathbf{P}_t$ is the vector of true proportions $\{C_t, M_t, T_t\}$ from the deterministic simulation, and α_0 is a total concentration parameter controlling the precision of the observations. Rather than specifying α_0 directly, we parameterized observation error in terms of a target variance σ^2, computing the corresponding concentration as

$$\alpha_0 = \frac{K-1}{K^2\sigma^2} - 1, \tag{4.6}$$

where $K = 3$ is the number of state variable categories (coral, macroalgae, and turf). Smaller σ^2 therefore corresponds to larger α_0 and lower observation error (i.e., draws more tightly concentrated around $\mathbf{P}_t$), while larger σ^2 corresponds to smaller α_0 and noisier, more dispersed observations. We simulated three observation error levels: low ($\sigma^2 = 0.0001$), medium ($\sigma^2 = 0.001$), and high ($\sigma^2 = 0.01$) (Fig. 4.2 bottom row).

Finally, to evaluate the effect of sparse sampling, we sampled each simulated trajectory, whether deterministic, with process noise, or with observation error, at four intervals (i.e., $\delta t = 0.5, 1, 2$, and 4). Because the Mumby et al. model's parameters are calibrated as annual rates, these intervals correspond to observations collected every 6 months, and every 1, 2, and 4 years, respectively. All simula-

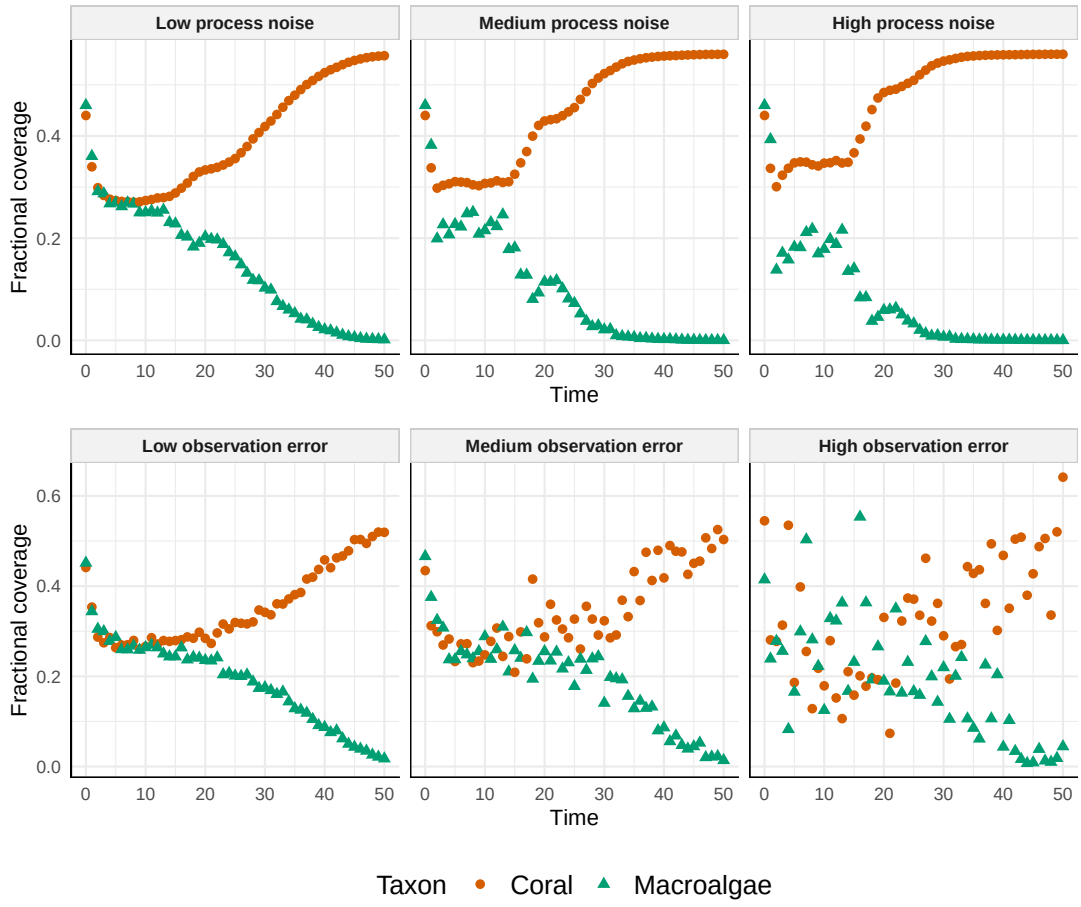


Figure 4.2: Example time series illustrating the stochastic data-generating processes evaluated in this study, all initialized at a coral initial condition value of 0.44 and sampled every 1 year ($\delta t = 1$). Top row: process noise on the grazing term at low, medium, and high magnitude. Bottom row: observation error at low, medium, and high magnitude. Fractional coverage represents the proportion of benthic reef substrate occupied by each cover type. Points show coral (orange points) and macroalgae (green triangles) cover over time.

tions ran for 50 model years, a duration chosen to let trajectories complete their transient dynamics and approach equilibrium. Although this exceeds the length of most reef monitoring records, which typically span a decade or two, it allowed us to isolate the role of sampling frequency independent of record length. Sampling therefore produced datasets ranging from 101 observations at 6-month intervals to 13 observations at 4-year intervals (51 and 26 observations at 1- and 2-year intervals, respectively; Fig. 4.1). For external validation, we generated four benchmark trajectories, one from each initial condition, using low-magnitude process noise ($\sigma_g = 0.2$) from a realization never used for training. Each benchmark was sampled annually ($\delta t = 1$) over the same 50-year period as the training data.

4.2.2 UDE models

UDEs combine a mechanistic differential equation model with a neural network that learns an unknown or weakly specified process directly from data (Rackauckas et al., 2021). We implemented both UDE formulations with the Julia package `UniversalDiffEq.jl` (Buckner et al., 2026), which fits UDEs as state-space models. A neural network is a flexible function approximator built from layers of simple computational units ("neurons"), each of which computes a weighted sum of its inputs, applies a nonlinear transformation (an "activation function"), and passes the result to the next layer. A network with a single hidden layer of sufficient width can, in principle, approximate any continuous function (Hornik et al., 1989). Rather than specifying the functional form of grazing *a priori*, we replaced

the grazing term in Equation 4.1b with a neural network while retaining the rest of the Mumby et al. model's mechanistic structure:

$$\frac{dC}{dt} = rTC - dC - aMC \tag{4.7a}$$

$$\frac{dM}{dt} = aMC - \Gamma(C, M) + qMT \tag{4.7b}$$

where

$$\Gamma(C, M) = \log\left(1 + \exp\left(\mathrm{NN}_\theta(C, M)\right)\right) \tag{4.8}$$

Here, NN_θ is a neural network with parameters θ, and the softplus transformation $\log(1 + e^x)$ keeps the learned grazing term positive with a smooth gradient near zero. The network takes the current state (C, M) as input and consists of a single hidden layer of 10 units with hyperbolic tangent (tanh) activation (the *SimpleNerualNetwork()* function in the package).

A state-space model distinguishes the true, unobserved state of the system $\hat{u}_t = (\hat{C}_t, \hat{M}_t)$ from the observations made of it. Its process model describes how the true state changes from one observation time to the next according to the UDE, allowing for process noise, while its observation model describes how each observation relates to the true state at that time, allowing for observation error. During fitting, the true state at each observation time is estimated alongside the model parameters, so that variation in the data can be attributed either to the dynamics or to how they were observed.

We compared two UDE formulations, distinguished by their treatment of the remaining mechanistic parameters (r, d, a, q). In the co-estimated-parameter for-

mulation (CE-UDE), these parameters were estimated jointly with the network, initialized at $r = 0.8$, $d = 0.3$, $a = 0.4$, and $q = 0.5$ (true values: $r = 1$, $d = 0.44$, $a = 0.1$, $q = 0.8$) to reflect a plausible but imperfect prior understanding of the system, and constrained to remain positive by estimating them on a log scale. In the fixed-parameter formulation (FP-UDE), the parameters were held at their true values, so the network's only task was to recover the unknown grazing term.

4.2.3 Model fitting

We trained both UDE formulations by minimizing the conditional likelihood loss implemented in the package, which uses point estimates of the true system states and requires the process and observation error covariances, Σ_{obs} and Σ_{proc}, to be specified rather than estimated. We used the package's default for both covariances (a variance of 0.025 for each state variable), held constant across all fits rather than tuned to the noise in each dataset, to reflect applications in which error magnitudes are not known or estimated beforehand. The marginal likelihood, which integrates over uncertainty in the states and estimates process error, is more accurate in theory. However, Buckner et al. (2026) found that conditional and marginal likelihood training produced similar forecasting accuracy in their case studies, while marginal-likelihood training scaled poorly with model complexity and was more computationally expensive. Given the 224 model fits required, we used conditional likelihood.

For all analyses, we used a 70:30 temporal split, in which the first 70% of

each time series was used for training. In addition to the likelihood terms, the training objective included normally distributed priors over the neural network weights (regularization weight $= 1 \times 10^{-4}$), penalizing large weights to control the smoothness of the learned grazing function. We optimized the model parameters using the Adam algorithm (Kingma and Ba, 2017) in two sequential stages: an initial coarse optimization (500 iterations, learning rate $= 0.25$), followed by fine-tuning (750 iterations, learning rate $= 0.01$), for a total of 1250 iterations per model. No convergence tolerance or early-stopping criterion was applied, and the procedure was identical across both formulations.

4.2.4 Evaluation

We evaluated each fit on term recovery and on forecasting under internal and external validation. We quantified recovery of the grazing term using the relative root mean squared error (rRMSE) between the learned term $\Gamma(C, M)$ and the true grazing term $\frac{gM}{M+T}$. We computed two different versions of this metric that differed in where the error was evaluated, but shared a common normalization. Term recovery rRMSE was evaluated only at the (C, M) pairs actually visited by the training portion of a given trajectory, and is the primary metric we report in the main text. Extrapolated term recovery rRMSE was evaluated over a fixed 50×50 grid spanning $C, M \in [0, 0.9]$, restricted to the region where $C + M < 1$, regardless of which states a trajectory visited. This version penalized models that fit their own training trajectory well but learned an incorrect functional form elsewhere in

state space; because it performed poorly across all conditions, we report it only in the *Supplementary Materials*. In both cases,

$$\text{rRMSE} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\Gamma(C_i, M_i) - \frac{gM_i}{M_i+T_i}\right)^2}}{\overline{\left(\frac{gM}{M+T}\right)}} \times 100\%, \tag{4.9}$$

where the sum is taken over the N evaluation points for each version (training points or the full grid), and the denominator is the mean true grazing term evaluated over the full feasible grid, held fixed across both versions so they share a common scale. We selected a 10% rRMSE threshold for acceptable term recovery based on a sensitivity analysis of how error in the learned grazing function propagates to trajectory and equilibrium predictions (see *Supplementary Materials*).

We assessed forecast success under two validation schemes. Internal validation tested whether a model trained on the first 70% of a trajectory could forecast that trajectory's own withheld final 30%, which tests extrapolation forward in time from a single observed trajectory, the most direct application of fitting a UDE to an existing monitoring record. External validation tested whether a model's learned dynamics generalize beyond its training trajectory, analogous to training a UDE on one reef and forecasting a nearby reef. Every fitted model was used to forecast all four benchmark trajectories, regardless of the initial condition, sampling frequency, or noise level it was trained under, with each forecast initialized from the benchmark's true initial state and integrated forward without further parameter updates.

For both schemes, we report the mean absolute error (MAE) of forecasted

coral and macroalgae cover, averaged across the two state variables and across time points; for external validation, we then averaged the four benchmark values to give a single score per model. We considered an MAE below 5% cover to indicate high forecasting accuracy. For external validation, we also assessed the qualitative forecast, defining a trajectory's state as whichever of coral or macroalgae had higher cover at the final time point and scoring whether the forecast state matched the benchmark's.

To place external validation performance in context, we compared UDE forecasts with three reference forecasts. The true deterministic model was integrated from each benchmark's true initial state using the true parameter values and baseline grazing rate g_0, and was scored with the same MAE and state-match metrics as the UDEs. A constant-state forecast predicted the same final state for all four benchmarks, and an initial-dominance forecast predicted that whichever of coral or macroalgae had higher initial cover would remain dominant.

4.3 Results

We evaluated our three hypotheses concerning the effects of noise and sampling resolution (H1), transient length (H2), and model formulation (H3) on grazing term recovery and forecast success, comparing the co-estimated parameter (CE-UDE) and fixed parameter (FP-UDE) UDE formulations. For term recovery, we report term recovery rRMSE; extrapolated term recovery rRMSE, which performed poorly across all conditions (0% of runs met the 10% threshold in every tested

condition), is reported in the *Supplementary Materials* (Table S.4). For forecast success, we report the coral-macroalgae-averaged MAE under both internal and external validation. Under external validation, we report the quantitative MAE (i.e., how close the forecasted cover values were to the true values) and the qualitative state-match rate (i.e., did a forecast correctly identify which basin of attraction a trajectory settled into). Summary tables for each model run and condition are provided in the *Supplementary Materials*.

4.3.1 H1: Process noise, observation error, and sampling frequency

We hypothesized that process noise would act as an informative signal, improving term recovery and forecasting success, while observation error and reduced sampling resolution would hinder both goals.

Contrary to this hypothesis, process noise did not improve performance (Fig. 4.3). The CE-UDE formulation showed no benefit at any process noise level, with only 1 of 16 runs (6.2%) meeting the 10% term recovery threshold at each level of process noise. The FP-UDE formulation showed a slightly higher fraction of runs meeting the threshold under low process noise relative to the deterministic baseline (10 of 16, 62.5% vs. 9 of 16, 56.2%), but this benefit disappeared entirely at medium and high process noise (0 out of 16 runs meeting the threshold at either level). Internal validation success declined with increasing process noise for FP-UDE (15 of 16, 9 of 16, 7 of 16, and 6 of 16 runs meeting the 5% MAE threshold at none, low, medium, and high process noise, respectively); CE-UDE showed a similar decline

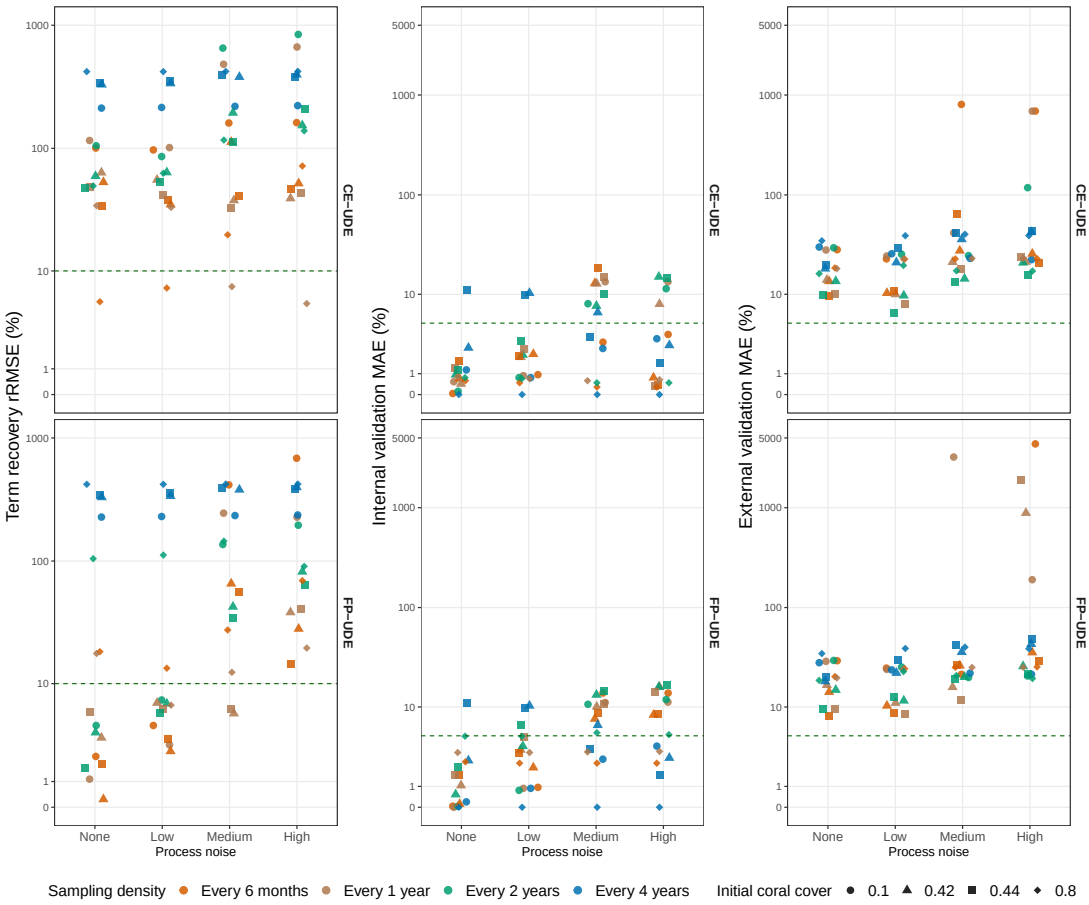


Figure 4.3: Term recovery, internal validation, and external validation performance across process noise for both UDE formulations. Columns show term recovery rRMSE (%; left), internal validation MAE (%; center), and external validation MAE (%; right), all on a log scale. Point color indicates sampling frequency (every 6 months, every 1 year, every 2 years, every 4 years); point shape indicates initial coral cover. Rows indicate model formulation (CE-UDE, FP-UDE). Dashed lines mark the 10% term recovery threshold and the 5% forecast MAE threshold. Process noise did not improve term recovery or forecasting for either UDE formulation, and external validation MAE never fell below the threshold under any condition.

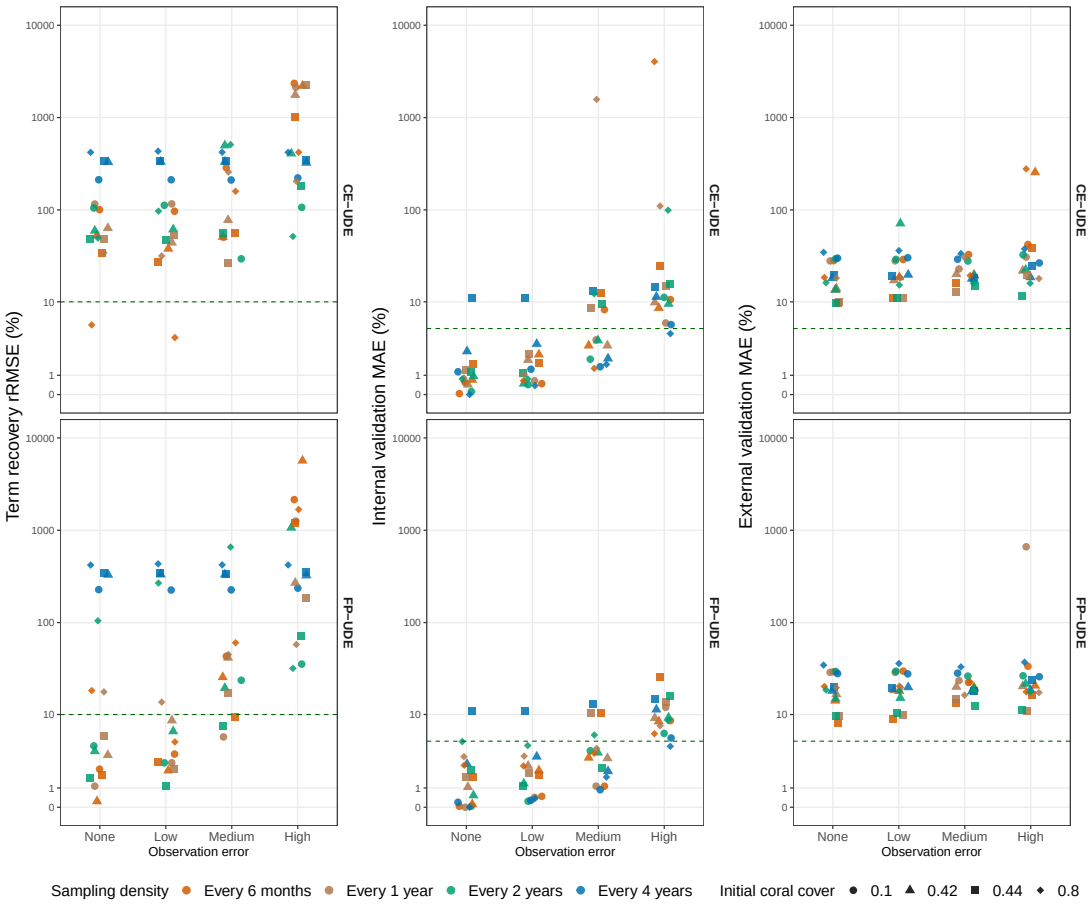


Figure 4.4: Term recovery, internal validation, and external validation performance across observation error for both UDE formulations. Columns show term recovery rRMSE (%; left), internal validation MAE (%; center), and external validation MAE (%; right), all on a log scale. Point color indicates sampling frequency (every 6 months, every 1 year, every 2 years, every 4 years); point shape indicates initial coral cover. Rows indicate model formulation (CE-UDE, FP-UDE). Dashed lines mark the 10% term recovery threshold and the 5% forecast MAE threshold. Observation error degraded forecasting for both formulations and term recovery for FP-UDE, as its magnitude increased, while external validation MAE never fell below the threshold under any condition.

through medium noise but a slight uptick at high noise (15 of 16, 13 of 16, 10 of 16, and 11 of 16).

Observation error degraded forecasting for both formulations, and term recovery for FP-UDE specifically, as its magnitude increased (Fig. 4.4). The CE-UDE formulation's fraction of runs meeting the term recovery threshold was uniformly low across all observation error levels (1 of 16, 6.2%, at both deterministic baseline and low observation error, falling to 0 of 16 at medium and high observation error). The FP-UDE formulation, by contrast, showed a clear decline, from 9 of 16 (56.2%) and 10 of 16 (62.5%) at the deterministic baseline and low observation error to 3 of 16 (18.8%) at medium observation error and 0 of 16 at high levels. Internal validation success declined sharply with increasing observation error for both formulations, most severely between medium and high noise (CE-UDE: 15 of 16, 15 of 16, 9 of 16, and 1 of 16 runs meeting the 5% MAE threshold at none, low, medium, and high observation error, respectively; FP-UDE: 15 of 16, 15 of 16, 12 of 16, and 1 of 16).

Reduced sampling frequency also hindered term recovery as hypothesized, most severely at the lowest frequency (once every 4 years; Figs. 4.3 and 4.4). For FP-UDE, the fraction of runs meeting the 10% term recovery threshold was comparable across the three higher sampling frequencies (11 of 28, 11 of 28, and 10 of 28 at sampling every 6 months, every 1 year, and every 2 years, respectively) but fell to 0 of 28 at the lowest frequency (every 4 years). CE-UDE showed a similar pattern, though term recovery was already poor even at higher frequencies (3 of 28 and 2 of 28 runs meeting the threshold at sampling every 6 months and every

1 year, respectively), with no runs meeting the threshold at sampling every 2 or 4 years. Internal validation success, by contrast, did not decline monotonically with reduced sampling frequency for either formulation: CE-UDE met the 5% MAE threshold in 21 of 28, 19 of 28, 15 of 28, and 19 of 28 runs at sampling every 6 months, every 1 year, every 2 years, and every 4 years, respectively, and FP-UDE showed a similar pattern (17 of 28, 15 of 28, 14 of 28, and 19 of 28).

Under external validation, no run of either formulation met the 5% MAE threshold at any level of process noise, observation error, or sampling frequency.

4.3.2 H2: Transient length

We hypothesized that models trained on trajectories with a longer transient period, i.e., those initialized near the basin boundary ($C(0) = 0.42$ or 0.44), would show more successful term recovery and forecasting than those trained on trajectories with a short transient ($C(0) = 0.1$ or 0.8).

This hypothesis was partially supported for term recovery; the UDE recovered the grazing term more accurately when trained on trajectories with a longer transient. Median term recovery rRMSE was lower for long-transient trajectories than short-transient trajectories under both CE-UDE (56.3% vs. 102.8%) and FP-UDE (3.6% vs. 18.0%), and a greater proportion of long-transient runs met the 10% threshold under FP-UDE (6 out of 8, 75.0% vs. 3 out of 8, 37.5%) (Fig. 4.5 first column). Under CE-UDE, however, no long-transient runs met the 10% threshold despite the lower median rRMSE (0 of 8 vs. 1 of 8, 12.5%, for short-transient runs).

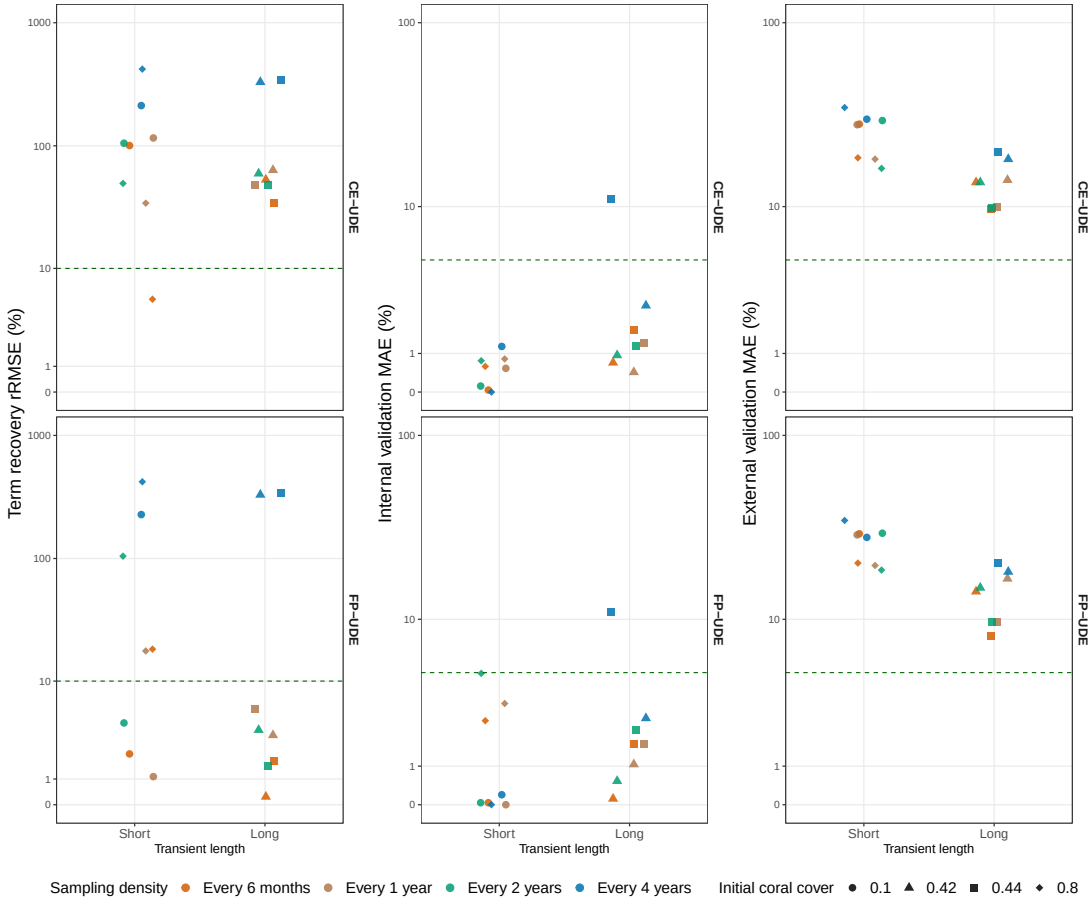


Figure 4.5: Term recovery, internal validation, and external validation performance for short-transient ($C(0) = 0.1, 0.8$) versus long-transient ($C(0) = 0.42, 0.44$) initial conditions under deterministic training data, for both UDE formulations. Columns show term recovery rRMSE (%; left), internal validation MAE (%; center), and external validation MAE (%; right), all on a log scale. Point color indicates sampling frequency (every 6 months, every 1 year, every 2 years, every 4 years); point shape indicates initial coral cover. Rows indicate model formulation (CE-UDE, FP-UDE). Dashed lines mark the 10% term recovery threshold and the 5% forecast MAE threshold. Longer transients improved term recovery and external validation, but modestly worsened internal validation forecast error, though nearly all internal validation forecasts remained within threshold regardless of transient length.

Even so, CE-UDE's median term recovery rRMSE remained well above threshold for both transient lengths, and FP-UDE's median rRMSE for short-transient trajectories also remained well above threshold, indicating that a longer transient improved term recovery in a relative sense without reliably bringing rRMSE below the 10% threshold.

Internal validation MAE, by contrast, was higher for long-transient than short-transient trajectories under both formulations (CE-UDE 1.25% vs. 0.63%; FP-UDE 1.65% vs. 0.15%), even though nearly all forecasts for both transient lengths fell below the 5% MAE threshold (Fig. 4.5 second column). The single exception occurred under the long-transient condition: $C(0) = 0.44$ at the lowest sampling frequency (i.e., sampling every 4 years) exceeded the threshold under both CE-UDE and FP-UDE (internal MAE of 10.95% in each case), which is likely attributable more to the reduced sampling frequency rather than to transient length.

External validation followed the pattern predicted by the hypothesis for both formulations. Long-transient trajectories produced substantially higher external validation state-match rates (i.e., did the UDE forecast the correct state qualitatively) than short-transient trajectories for CE-UDE (84.4% vs. 50.0%) and for FP-UDE (75.0% vs. 50.0%), with correspondingly lower external validation MAE (CE-UDE median: 13.6% vs. 28.1%; FP-UDE median: 14.6% vs. 28.4%; Fig. 4.5 third column; full state-match values by initial condition and sampling frequency are provided in Table S.7, Supplementary Materials).

Overall, H2 was partially supported for term recovery, where longer transients lowered median rRMSE under both formulations but met the threshold more often

only under FP-UDE. It was supported for external validation, where longer transients improved performance, but not for internal validation, where longer-transient trajectories showed modestly higher forecast error despite nearly all forecasts remaining well within the accuracy threshold.

4.3.3 H3: UDE model formulation

We hypothesized that fixing known mechanistic parameters (FP-UDE) would improve term recovery relative to co-estimating them (CE-UDE), at some cost to forecasting flexibility. This hypothesis was supported on both counts (Fig. 4.6). The CE-UDE formulation recovered the grazing term successfully in only 5 of 112 runs (4.5%), meeting the 10% rRMSE term recovery threshold far less often than the FP-UDE formulation, which succeeded in 32 of 112 runs (28.6%) (Fig. 4.6 first column). Conversely, the CE-UDE formulation achieved modestly better internal validation forecast success than FP-UDE: 74 of 112 runs (66.1%) met the 5% internal validation MAE threshold under CE-UDE, compared to 65 of 112 (58.0%) under FP-UDE (Fig. 4.6 second column). The two formulations performed similarly on external validation (Fig. 4.6 third column), with comparable median external validation MAE (19.8% for CE-UDE and 20.2% for FP-UDE).

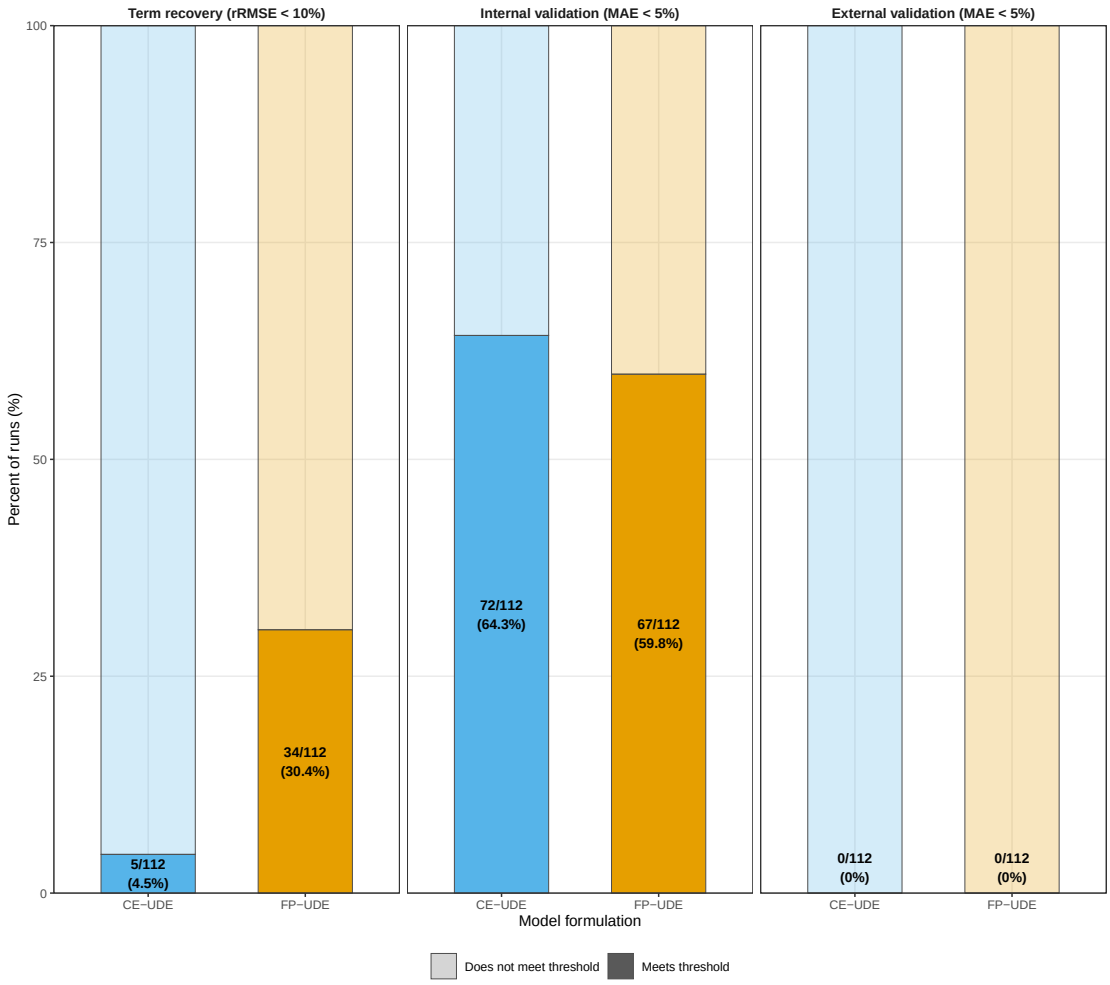


Figure 4.6: Comparison of the two UDE formulations across all 112 conditions tested per formulation, shown as the percentage of runs meeting versus not meeting each success threshold (solid vs. faded fill). The first column is term recovery (rRMSE $< 10\%$). The second is internal validation forecasting error (MAE $< 5\%$), and the third is external validation forecasting error (MAE $< 5\%$). FP-UDE achieved better term recovery than CE-UDE, while CE-UDE achieved modestly better internal validation forecast success; neither formulation met the external validation threshold in any run.

4.3.4 Qualitative external validation

Because no run met the 5% MAE threshold under external validation, we also examined a coarser, qualitative measure of external validation success: whether a model correctly identified which basin of attraction each benchmark trajectory settled into, regardless of the numerical accuracy of its forecast. Both formulations performed better on this measure, and similarly to one another (Fig. 4.7). Mean state-match rates were 57% for CE-UDE and 58% for FP-UDE, and 51 CE-UDE runs (45.5%) and 50 FP-UDE runs (44.6%) correctly classified at least 3 of the 4 external benchmark trajectories. The most common outcomes were correct classification of 3 of 4 trajectories (CE-UDE: 44 runs, 39.3%; FP-UDE: 42 runs, 37.5%) and 2 of 4 (CE-UDE: 36 runs, 32.1%; FP-UDE: 39 runs, 34.8%). Only a single run, a CE-UDE run, failed to classify any benchmark trajectory correctly. For comparison, the true deterministic model classified all four benchmark trajectories correctly (100%) with a mean MAE of 0.67%. A constant-state forecast classified two of the four correctly (50%) regardless of the state chosen, and an initial-dominance forecast classified three (75%), misclassifying only the $C(0) = 0.44$ benchmark.

4.4 Discussion

We evaluated three hypotheses concerning the reliability of UDEs for recovering an unknown mechanistic term and forecasting community dynamics under ecologically realistic conditions. Support for these hypotheses was mixed and, in places,

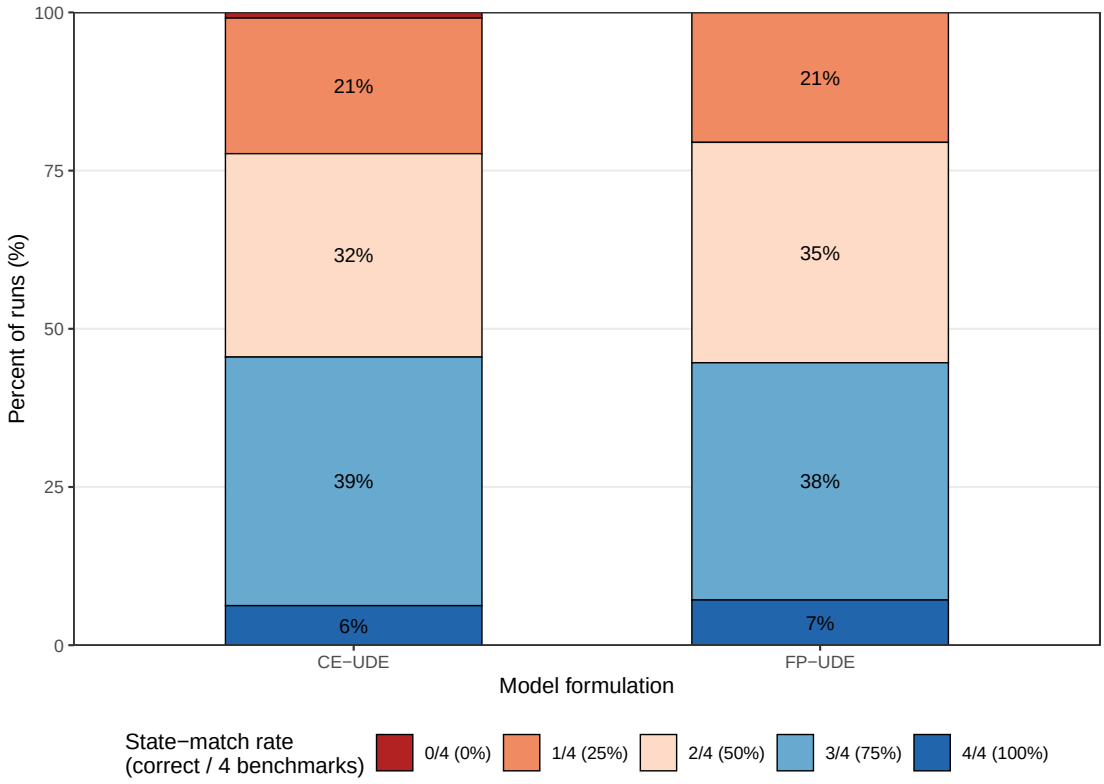


Figure 4.7: Distribution of external validation state-match rates (the fraction of the four external benchmark trajectories correctly classified) across all 112 conditions tested per formulation.

informatively contradicted our expectations (Table 4.1). Process noise did not act as the informative signal we hypothesized (H1); however, observation error and reduced sampling did degrade both goals largely as expected. Longer transient periods improved both term recovery and external validation as hypothesized (H2), but showed the opposite pattern for internal validation, where forecast error was higher, though nearly always within threshold, for long-transient trajectories. Fixing known mechanistic parameters improved term recovery relative to co-estimating them, while co-estimating parameters modestly improved internal validation forecasting as we hypothesized (H3), but not quite to the degree we anticipated. A further pattern emerged underlying all three hypotheses: regardless of formulation, noise condition, sampling frequency, or transient length, no model in this study met our 5% quantitative MAE threshold on external validation, even though nearly half of the models correctly forecast the qualitative outcome for at least 3 of the 4 external benchmark trajectories.

4.4.1 The role of noise as an informative signal depends on its interaction with transient dynamics

We hypothesized that process noise would act as an informative signal, improving both term recovery and forecast success, based on the idea that noise can expand the range of state space a system explores, thereby revealing information about underlying dynamics that a purely deterministic trajectory would not (Boettiger, 2018). Our results did not support this expectation. Term recovery under pro-

cess noise was, at best, indistinguishable from the deterministic case, and internal validation forecast success generally declined as process noise increased.

However, this may reflect a feature of our particular system rather than a general failure of process noise as an informative signal. The coral-macroalgae system used here moves toward one of two stable equilibria relatively quickly under the deterministic dynamics, and much of the information relevant to identifying the grazing term is concentrated in the transient period before a trajectory settles near its eventual equilibrium (Hastings, 2004; Hastings et al., 2018). Rather than expanding the range of state space explored, as Boettiger's mechanism describes, process noise applied to the grazing parameter in our simulations visibly shortened this transient: example time series show trajectories under process noise settling toward equilibrium more quickly than their deterministic counterparts, particularly at medium and high magnitudes (Fig. 4.2, top row). Process noise in our system therefore may have shortened, rather than lengthened, the effective transient available for learning – the opposite condition under which Boettiger's (2018) mechanism predicts a benefit. Additionally, we only tested one form and a small set of magnitudes of process noise, and other noise regimes may also be informative: noise that causes flickering between alternative states, for instance, has been shown to reveal information about an otherwise unobserved driver (Rathore et al., 2025).

4.4.2 Transient length shapes term recovery and forecasting

We motivated our second hypothesis by noting that existing work applying data-driven approaches to chaotic and oscillating systems has found that these dynamics, which are often assumed to be inherently hard to model because of their sustained complexity, can in fact be comparatively identifiable and forecastable (Bonnaffé et al., 2021; Buckner et al., 2026). Our results extend this idea to a different, but common, source of complexity. Transient dynamics are a common feature of ecological systems (Rogers et al., 2022; Scheffer et al., 2001; Beisner et al., 2003). That a UDE can extract useful information from this kind of transient suggests the informational benefit of complex dynamics is not specific to chaos or cycling, but reflects something more general about how much a trajectory's dynamics actually reveal about the process generating them (Hastings, 2004; Hastings et al., 2018).

When forecasting the withheld 30% test data (i.e., internal validation), error was modestly higher for long-transient trajectories, although nearly all forecasts met the threshold. We interpret this through Pennekamp et al.'s (2019) distinction between intrinsic and realized predictability: a trajectory already resting near equilibrium poses an easy forecasting problem almost regardless of model quality, because the correct answer is largely "stay where you are", while a trajectory still moving through its transient is intrinsically harder to forecast even when its underlying dynamics have been correctly learned. In the Mumby et al. model specifically, most training trajectories (even those with a longer transient) have

largely completed their approach toward equilibrium before the end of the 70% training window, so the withheld final 30% falls within the equilibrium region for nearly all trajectories. While the exact convergence rate depends on the specific parameter values used here, rapid convergence relative to more structurally complex reef models is a general property of the Mumby et al. model's simple, three-variable form, and so represents a limitation of our focal model rather than of internal validation as a framework. That is, the Mumby et al. model's relatively fast convergence to equilibrium leaves little opportunity for a withheld forecast window to fall within an active transient. A system with slower relaxation dynamics, or a shorter training window relative to the timescale of the transient, would be needed to test whether the same near-equilibrium explanation still holds when a trajectory has not yet settled by the time a forecast is made. This limitation is compounded by our choice to train each model on a single time series: a trajectory initialized at $C(0) = 0.8$, for instance, is almost exclusively defined by near-zero macroalgae cover, leaving the network with essentially no informative examples of the grazing term's shape regardless of how long that single trajectory runs. Training on multiple trajectories spanning a range of initial conditions, discussed further below, would directly address this.

The same intrinsic-predictability logic complicates how a recovery threshold like ours should be applied near a system's basin boundary. Our sensitivity analysis for term recovery (*Supplementary Materials*) shows that trajectories initialized close enough to the coral-macroalgae separatrix can have their predicted long-term outcome reversed by only a small fraction of the grazing term error tolerated in the

system. This is not a flaw specific to UDEs or to our particular threshold choice; it is a structural feature of any bistable system, and it means a fixed error threshold cannot be uniformly conservative across trajectories that differ substantially in their distance from a separatrix. A given rRMSE value therefore means something different depending on how close its trajectory sits to a known regime boundary.

4.4.3 UDE formulation involves a trade-off between term recovery and forecasting

Our third hypothesis was supported in both directions: fixing known mechanistic parameters (FP-UDE) improved term recovery relative to co-estimating them (CE-UDE), while co-estimating parameters modestly improved internal validation forecast success. This pattern mirrors a long-standing distinction in the statistical and machine learning literature between models built to explain a process and models built to predict new outcomes from it, consistent with the broader observation that added flexibility often trades interpretability for predictive performance (Breiman, 2001; Pichler and Hartig, 2022).

The source of each formulation's advantage differs. With four parameters fixed at their true values, the neural network's only remaining task is learning the single unknown grazing function, which sharpens recovery of that specific term. CE-UDE's flexibility, by contrast, operates by absorbing parameter uncertainty rather than by adding capacity to the neural network itself. That flexibility apparently helps the fitted trajectory adapt to a training window's idiosyncrasies in ways that

carry over, modestly, to the withheld final 30% of the same trajectory (internal validation), even though it does nothing to help the mechanism itself, and by diluting the grazing-term signal, likely hinders it.

Despite this trade-off, neither formulation forecasted the external validation benchmarks well: median external validation MAE was similar for both (19.8% for CE-UDE, 20.2% for FP-UDE), and neither formulation produced a single run across all 112 conditions tested that met our 5% accuracy threshold. Whatever internal-validation advantage CE-UDE's flexibility provided did not extend to genuinely new conditions, consistent with meta-analytic evidence from short-term ecological population forecasting that added model complexity, even when it improves in-sample fit, does not reliably translate into better out-of-sample (external validation) forecasts and can sometimes make them worse (Ward et al., 2014).

4.4.4 The difference between qualitative and quantitative forecast accuracy

UDEs performed better on the qualitative measure of external validation than against the 5% MAE threshold, which no run met regardless of formulation, noise condition, sampling frequency, or transient length. This distinction could matter in applied settings where uses of a reef forecast (e.g., determining whether a reef is degrading or prioritizing sites for intervention) depend on a coarser qualitative outcome rather than cover estimates accurate to within a few percentage points. This qualitative performance should nevertheless be interpreted cautiously, be-

cause the four benchmark trajectories provide a coarse test that simple rules can also pass. Two benchmarks ended coral-dominated and two ended macroalgae-dominated. A constant-state forecast that always predicted the same state would classify 50% of the benchmarks correctly without using any information about a trajectory. A forecast based only on each trajectory's initially dominant cover type would classify 75% correctly, failing only for our $C(0) = 0.44$ benchmark, which began with slightly more macroalgae than coral but ended coral-dominated. The true deterministic model, by contrast, classified all four correctly (100%) with a mean MAE of 0.67%, showing that accurate quantitative and qualitative forecasts of this benchmark set were achievable given the correct dynamics. UDEs correctly classified 57% (CE-UDE) and 58% (FP-UDE) of trajectories on average, with approximately 45% of runs classifying at least three of the four correctly. On average, then, the UDEs outperformed a constant-state forecast but not a forecast based on initial dominance, and fell well short of the true model. Because each model was evaluated on only four trajectories, these rates should be interpreted as an indicator at this time. Read this way, the qualitative measure does not rescue external validation performance so much as clarify its limits. If the goal is fine-grained quantitative accuracy, such as precise estimates of future cover to inform detailed management thresholds, our results suggest that UDEs (at least in the form tested here) are not yet reliable beyond their training trajectory. If the goal is a qualitative forecast of which state the reef is heading toward, UDEs still did not outperform a simple rule based on initial dominance on average. Because the true model achieved high accuracy on these benchmarks, this shortfall likely

reflects the limited information a UDE can extract from a single trajectory. As such, training UDEs on multiple trajectories is a natural next step to determine whether external validation can be improved.

Together, these findings suggest a measured conclusion. UDEs trained on a single reef's monitoring record showed some ability to anticipate whether a different reef would end up coral- or macroalgae-dominated, but on average did not outperform a simple rule based on initial dominance or forecast that reef's trajectory with the precision our quantitative threshold demands. For applied uses, a UDE's qualitative forecast (as used here) should therefore at this time be compared against baselines before it is used to inform management decisions.

4.4.5 Limitations and future directions

Several methodological choices in this study limit the generality of our conclusions. One of these is addressed above: the Mumby et al. model's rapid convergence to equilibrium curtailed the range of transient lengths we could test. Testing UDE performance on reef models with richer transient or oscillatory dynamics would help distinguish which of our findings reflect a general property of UDEs under realistic data constraints and which instead reflect the specific structure of the Mumby et al. model (see Blackwood et al. 2018 for a review of coral reef models spanning a range of structural complexity).

A second limitation is more consequential for how far specific numerical thresholds and failure modes should be extrapolated. Our results speak only to models

trained on a single time series. This intentional design choice let us isolate the effects of a trajectory's own initial condition, sampling frequency, and noise level, but it also produced the identifiability failure at $C(0) = 0.8$ discussed above. This is a limitation of the single trajectory design itself, not just of the Mumby et al. model or any one noise or sampling condition, and our results may understate what a UDE trained on multiple trajectories could achieve. At the same time, our models had advantages unlikely to hold for real reefs, including correct mechanistic structure outside the grazing term, relatively longer time records, a single source of noise in each dataset, and, for FP-UDE, the true parameter values, so our results should not be read as a general lower or upper bound on UDE performance. Training a UDE on multiple trajectories spanning a range of initial conditions could directly address this limitation, and may also help close the extrapolation and external validation gaps described above. This design would also more closely reflect the kind of data many reef monitoring programs collect, where several nearby reefs are surveyed over a common, often short time window (Australian Institute of Marine Science, 2026). Such monitoring programs also typically survey reefs at intervals closer to our two lowest tested sampling frequencies (once every 2–4 years) than to the higher frequencies under which term recovery was most successful in our study, underscoring that closing the gap between multi-trajectory training data and realistic sampling frequency, together, may be necessary before UDEs can be reliably applied to real reef monitoring records.

4.5 Conclusion

UDEs' combination of mechanistic and data-driven components makes them an appealing candidate for ecological applications, but our results show that this promise is only partly realized when incorporating aspects of realistic data constraints. Term recovery and quantitative external validation proved difficult, and qualitative external validation did not exceed a simple baseline on average. Internal validation succeeded more often, with the acknowledgement that this was a comparatively easier test. Our aim in this study was to evaluate the conditions under which UDEs can reliably recover mechanistic understanding and forecast ecological dynamics, given the messy, limited data that field monitoring typically provides. Stress-testing these tools under conditions that mimic real monitoring constraints, as we have done here, is a necessary step before their outputs can be used with confidence to inform ecological understanding or management decisions. Our results suggest that, under such constraints, UDEs are not yet a reliable qualitative forecasting tool: on average, they outperformed a constant-state forecast but not a simple forecast based on initial dominance. Reliable term recovery and precise quantitative forecasting will likely require richer training data, through multiple monitoring trajectories, rather than what a single, sparsely-sampled time series can provide.

Table 4.1: Summary of hypotheses, predictions, and outcomes across term recovery, internal validation, and external validation. CE-UDE: co-estimated-parameter formulation; FP-UDE: fixed-parameter formulation.

Hypothesis	**Prediction**	**Term recovery**	**Internal validation**	**External validation**
H1: Noise & sampling frequency	Process noise improves both goals; observation error and reduced sampling frequency hinder both	Not supported for process noise; supported for observation error and sampling frequency	Not supported for process noise (success generally declined); supported for observation error; not supported for sampling frequency (no monotonic decline)	No run met threshold under any condition
H2: Transient length	Longer transients improve term recovery and forecasting	Partially supported (lower median rRMSE under both formulations; threshold reliably met only under FP-UDE)	Not supported (higher MAE for long transients, though still mostly within threshold)	Supported (higher state-match rate, lower MAE)
H3: Model formulation	Fixing parameters (FP-UDE) improves term recovery at a cost to forecasting	Supported (32 of 112 for FP-UDE vs. 5 of 112 for CE-UDE)	Supported (FP-UDE underperformed CE-UDE)	Not distinguishable between formulations
Qualitative state match	—	—	—	At least 3 of 4 benchmark trajectories correctly classified in 45.5% of CE-UDE and 44.6% of FP-UDE; similar between formulations (57% CE-UDE, 58% FP-UDE mean match rate)

Supplementary Materials for

Recovering and forecasting simulated coral reef dynamics with universal dynamic equations

Cheyenne N. Jarman, Jack H. Buckner, Ariel Greiner, Zechariah D. Meunier, Lisa C. McManus, James R. Watson, and Emerson Arehart

A Parameter Values Used for Analyses

Table S.1: The parameter values used in the Mumby et al. model. Values come from Fung et al. (2011) and Baskett et al. (2014). Values were calibrated as annual rates.

Biological meaning	Parameter	Value
Rate of coral overgrowth on turf	r	1
Coral mortality	d	0.44
Rate of macroalgae overgrowth on coral	a	0.1
Grazing rate	g	0.3
Rate of macroalgae overgrowth on turf	q	0.8

B Sensitivity Analysis for the rRMSE Recovery Threshold

To select an rRMSE threshold for acceptable grazing-term recovery, we conducted a sensitivity analysis examining how error in the learned grazing function propagates to downstream trajectory and equilibrium predictions, independent of any UDE model fit. Rather than relying on model-fitted grazing functions, we constructed a family of synthetic "learned" grazing functions with known, controllable rRMSE relative to the true grazing term, and assessed how simulated trajectories under these functions diverged from the true trajectory as a function of rRMSE.

B.1 Constructing grazing functions of known error

We defined a perturbation field

$$\phi(C, M) = (M - C)\sin(\pi M), \tag{S.1}$$

and constructed a family of perturbed grazing functions

$$\Gamma_\epsilon(C, M) = \frac{gM}{M + T} + \epsilon\,\phi(C, M), \tag{S.2}$$

where ϵ scales the magnitude of the perturbation. For each of the 19 target rRMSE values (1, 2, 3, 5, 7, 10, 12, 15, 17, 20, 22, 25, 30, 40, 50, 75, 100, 150, and 200%), we solved numerically for the value of ϵ that produced that exact rRMSE (Equation 4.9, evaluated over the same 50×50 feasible grid used in the main text) between Γ_ϵ and the true grazing term, generating a set of grazing functions with precisely known recovery error. This perturbation targets a single systematic error pattern rather than the full space of possible error shapes a fitted neural network could produce; we discuss this scope below.

B.2 Forward simulation and equilibrium classification

The deterministic system has three equilibria: a coral-dominated stable sink at $(C, M) = (0.560, 0.000)$, a macroalgae-dominated stable sink at $(C, M) = (0.000, 0.625)$, and an interior saddle point at $(C, M) = (0.249, 0.283)$ separating their basins of attraction. For each of the four initial conditions used in the main analysis ($C(0) = 0.1, 0.42, 0.44, 0.8$), we forward-simulated the true system and each perturbed system (Γ_ϵ for every target rRMSE level) for 500 time units, well beyond the timescale needed for trajectories to converge to a stable equilibrium (Fig. S.1). We then classified each simulation's final state by the proximity to the three equilibria and compared this classification to that of the corresponding true trajectory.

B.3 Sensitivity varies with distance from the basin boundary

The rRMSE at which a perturbed trajectory first converged to the wrong equilibrium varied substantially across initial conditions: 15% (IC = 0.1), 2% (IC = 0.42), 22% (IC = 0.44), and 25% (IC = 0.8) (Table S.2). This variability broadly tracked

each initial condition's proximity to the separatrix: coral IC = 0.42, whose trajectory begins just 0.007 (in M, along its fixed $C(0)$) from the basin boundary, was by far the most sensitive to grazing-term error, with even a 2% rRMSE sufficient to reverse its predicted long-term outcome. Coral IC = 0.44, despite also being classified as a near-boundary condition in the main text, is somewhat farther away from the boundary (0.036) and tolerated substantially more error (22%) before its outcome flipped, indicating that sensitivity depends on more than distance from the boundary alone – the geometry of a trajectory's approach to the saddle likely also matters. Coral IC = 0.8 showed no basin transition along its fixed coral cover value within the range of M we tested, consistent with it lying deep within the coral-dominated basin (Table S.2).

B.4 Selecting the recovery threshold

We selected 10% rRMSE as our threshold for acceptable term recovery in the main text. This threshold is conservative for three of the four initial conditions tested (IC = 0.1, 0.44, and 0.8), remaining below the rRMSE at which their predicted long-term outcome would be reversed. It is not conservative for coral IC = 0.42, the trajectory positioned closest to the basin boundary, which is already sensitive enough to grazing-term error that our lowest tested error level (1%) approached its failure point and 2% exceeded it. This is an expected consequence of the boundary's sensitivity rather than an inconsistency in the threshold – trajectories near a basin boundary are, by construction, more sensitive to any source of model error, and no fixed rRMSE threshold can be simultaneously conservative for both boundary-adjacent and interior trajectories – in principle, a trajectory positioned arbitrarily close to the separatrix can be made arbitrarily sensitive, so no single threshold, however small, could remain conservative for every possible near-boundary initial condition. We therefore treat 10% as a threshold appropriate for interpreting recovery success away from the immediate neighborhood of the basin boundary, and note that recovery claims for near-boundary trajectories such

as coral IC = 0.42 warrant additional caution regardless of their reported rRMSE. We also note that this analysis characterizes sensitivity to one specific, systematic pattern of grazing-term error; a fitted neural network could produce error with a different spatial structure (e.g., concentrated near the saddle point rather than distributed smoothly across state space), which could plausibly alter the rRMSE at which a given trajectory's predicted outcome flips. The reported thresholds should therefore be interpreted as informative benchmarks for this error pattern rather than a guarantee that any grazing function achieving a given rRMSE will preserve the correct qualitative outcome.

Table S.2: First target rRMSE at which the simulated trajectory under a perturbed grazing function converged to the wrong equilibrium (i.e., a different basin of attraction than the true trajectory), by initial condition, alongside each initial condition's approximate distance from the separatrix (evaluated along its fixed coral cover value).

Initial condition ($C(0)$)	First failure rRMSE (%)	Distance from separatrix (M)
0.10	15.0	0.679
0.42	2.0	0.007
0.44	22.0	0.036
0.80	25.0	N/A (no basin transition detected)

Summary Performance Tables

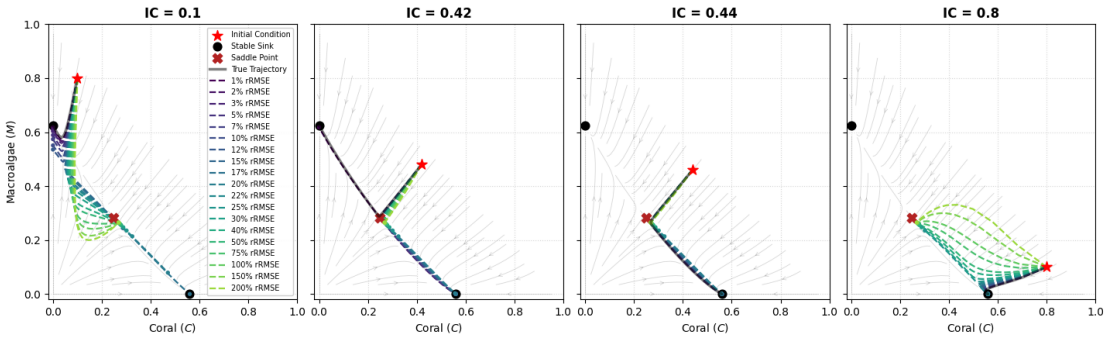


Figure S.1: Phase-plane trajectories under the true grazing term (gray) and under perturbed grazing functions of increasing target rRMSE (dashed, colored from purple at 1% to green at 200%), for each of the four initial conditions. Stable equilibria (coral-dominated and macroalgae-dominated) are shown as black circles; the interior saddle point is shown as a dark red X. For IC = 0.1 and IC = 0.8, low-error trajectories track the true trajectory closely before diverging at higher rRMSE; for IC = 0.42 and IC = 0.44, even modestly perturbed trajectories diverge from the true trajectory near the saddle point, consistent with their proximity to the basin boundary.

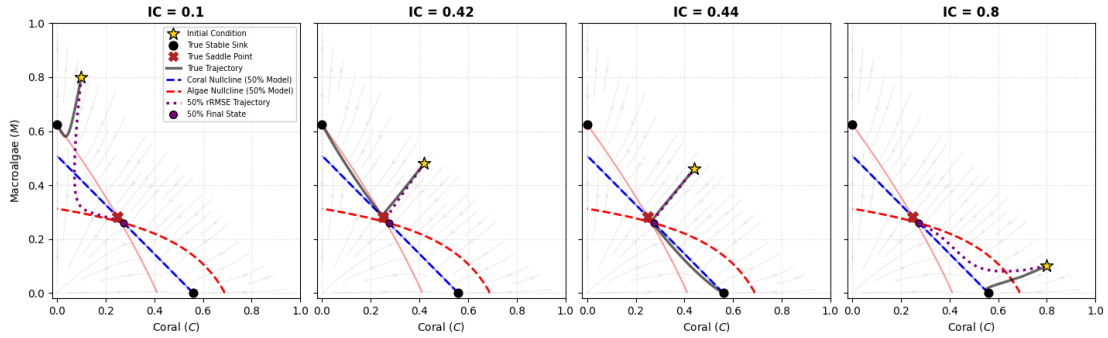


Figure S.2: Comparison of coral and macroalgae nullclines under the true grazing term (solid, faint) and under a grazing function with 50% rRMSE (dashed), for each of the four initial conditions. The shift in the algae nullcline under 50% error moves the effective saddle point, redirecting the 50% rRMSE trajectory (purple, dotted) into the coral-dominated basin even where the true trajectory (gray) converges to the macroalgae-dominated equilibrium.

Table S.3: Summary of grazing-term recovery (term recovery rRMSE, %) for the two UDE formulations split by noise type and level. Values are median [25th – 75th percentile], followed by the raw range (min–max). N gives the number of runs (four initial conditions × four sampling densities) underlying each row.

Formulation / noise	Level	N	Term recovery rRMSE	Range	%<10%
CE-UDE					
Deterministic	–	16	61.5 [48.3–139.9]	5.5–420	6.2
Process noise	Low	16	63.3 [40.9–129.8]	7.2–420	6.2
Process noise	Med	16	138.7 [40.2–383.7]	7.4–652.1	6.2
Process noise	High	16	157.8 [50.7–386.2]	5.3–842.1	6.2
Observation error	Low	16	79.1 [42.7–140.2]	3.9–432.4	6.2
Observation error	Med	16	185.0 [54.8–333.8]	26.5–509.6	0.0
Observation error	High	16	414.4 [217.7–1851.4]	51.6–2346.7	0.0
FP-UDE					
Deterministic	–	16	5.1 [2.1–135.3]	0.3–420	56.3
Process noise	Low	16	6.9 [5.4–141.3]	2.5–420	62.5
Process noise	Med	16	100.9 [32.7–278.6]	5.6–421.4	12.5
Process noise	High	16	86.2 [40.3–273.1]	14.5–683.8	0.0
Observation error	Low	16	5.7 [2.7–235.9]	1.1–432.4	62.5
Observation error	Med	16	42.4 [18.9–252.6]	5.6–658.1	18.8
Observation error	High	16	337.0 [155.9–1223.7]	31.9–5677.8	0.0

Table S.4: Summary of grazing-term recovery under extrapolation (extrapolation rRMSE, %, evaluated over the full biologically feasible grid rather than only the training trajectory, for the two UDE formulations split by noise type and level. Values are median [25th–75th percentile], followed by the raw range (min–max). N gives the number of runs underlying each row.

Formulation / noise	Level	N	Extrapolation rRMSE	Range	%<10%
CE-UDE					
Deterministic	–	16	326.6 [72.1–972.3]	65.9–6000.0	0.0
Process noise	Low	16	327.5 [78.3–574.2]	61.7–6639.1	0.0
Process noise	Med	16	602.1 [348.2–5067.3]	201.2–10563.6	0.0
Process noise	High	16	1964.2 [351.1–6664.7]	102.2–11367.0	0.0
Observation error	Low	16	327.0 [70.6–681.6]	56.8–6619.3	0.0
Observation error	Med	16	851.9 [327.4–2071.9]	79.7–9711.4	0.0
Observation error	High	16	430.2 [334.4–2463.0]	103.7–5268.7	0.0
FP-UDE					
Deterministic	–	16	76.5 [56.0–331.8]	19.0–9861.5	0.0
Process noise	Low	16	90.3 [65.5–331.8]	29.2–9925.2	0.0
Process noise	Med	16	886.6 [271.0–3112.0]	72.8–11239.7	0.0
Process noise	High	16	2292.5 [380.4–3322.2]	327.2–12354.4	0.0
Observation error	Low	16	145.9 [65.1–341.0]	20.3–10955.4	0.0
Observation error	Med	16	326.9 [165.1–526.7]	81.4–11772.8	0.0
Observation error	High	16	1219.9 [337.2–8192.3]	83.3–17455.1	0.0

Table S.5: Summary of internal validation forecast error (coral-macroalgae-averaged MAE, % absolute cover) for the two UDE formulations split by noise type and level. N gives the number of runs underlying each row.

Formulation / noise	Level	N	Internal validation MAE	Range	%<5%
CE-UDE					
Deterministic	–	16	0.83	0.00–10.95	93.8
Process noise	Low	16	1.45	0.00–10.30	87.5
Process noise	Med	16	7.05	0.00–18.60	43.8
Process noise	High	16	2.20	0.00–15.05	68.8
Observation error	Low	16	1.10	0.45–11.00	93.8
Observation error	Med	16	3.63	1.40–1571.20	56.2
Observation error	High	16	11.28	4.35–4041.35	6.2
FP-UDE					
Deterministic	–	16	1.35	0.00–10.95	93.8
Process noise	Low	16	3.23	0.00–10.30	81.3
Process noise	Med	16	8.10	0.00–14.70	31.3
Process noise	High	16	8.35	0.00–16.90	37.5
Observation error	Low	16	1.88	0.30–11.00	93.8
Observation error	Med	16	3.38	0.90–13.20	75.0
Observation error	High	16	8.85	4.35–25.85	6.2

Note: Extreme upper values in some ranges (e.g., observation error, medium and high) reflect a small number of individual runs with poor optimization convergence, not a typical outcome for that condition.

Table S.6: Summary of external validation forecast error (coral-macroalgae-averaged MAE, % absolute cover, averaged across all four external-validation benchmark trajectories) for the two UDE formulations split by noise type and level. N gives the number of runs underlying each row.

Formulation / noise	Level	N	External validation MAE	Range	%<5%
CE-UDE					
Deterministic	–	16	18.2	9.7–34.7	0.0
Process noise	Low	16	21.9	6.4–39.0	0.0
Process noise	Med	16	23.9	13.2–805.3	0.0
Process noise	High	16	23.5	15.6–691.0	0.0
Observation error	Low	16	19.0	10.9–71.3	0.0
Observation error	Med	16	19.6	12.8–33.6	0.0
Observation error	High	16	25.5	11.7–277.8	0.0
FP-UDE					
Deterministic	–	16	19.2	8.1–34.6	0.0
Process noise	Low	16	23.3	8.5–38.9	0.0
Process noise	Med	16	23.6	11.7–3218.5	0.0
Process noise	High	16	32.3	19.4–4356.7	0.0
Observation error	Low	16	19.6	9.0–36.0	0.0
Observation error	Med	16	19.5	12.5–33.1	0.0
Observation error	High	16	20.6	11.0–662.2	0.0

Note: Extreme upper values in some ranges (e.g., process noise, medium and high) reflect a small number of individual runs with poor optimization convergence, not a typical outcome for that condition.

Table S.7: External validation state-match rates (%; the proportion of the four external-validation benchmark trajectories for which the forecasted final state matched the true state) by initial condition and sampling frequency, under deterministic training data ($N = 1$ per cell). Initial conditions $C(0) = 0.1$ and 0.8 are short-transient trajectories; $C(0) = 0.42$ and 0.44 are long-transient trajectories

	Sampling frequency (δt)				
Initial condition ($C(0)$)	0.5	1.0	2.0	4.0	Mean
CE-UDE					
0.10 (short)	25	25	25	25	25.0
0.42 (long)	75	75	75	75	75.0
0.44 (long)	100	100	100	75	93.8
0.80 (short)	75	75	75	75	75.0
FP-UDE					
0.10 (short)	25	25	25	25	25.0
0.42 (long)	75	50	75	75	68.8
0.44 (long)	75	100	75	75	81.2
0.80 (short)	75	75	75	75	75.0

Bibliography

Australian Institute of Marine Science (2026). Great Barrier Reef Annual Summary Report Coral Reef Condition. Technical report, Australian Institute of Marine Science.

Baskett, M. L., Fabina, N. S., and Gross, K. (2014). Response Diversity Can Increase Ecological Resilience to Disturbance in Coral Reefs. *The American Naturalist*, 184(2):E16–E31.

Beckage, B., Gross, L. J., and Kauffman, S. (2011). The limits to prediction in ecological systems. *Ecosphere*, 2(11):art125.

Beisner, B., Haydon, D., and Cuddington, K. (2003). Alternative stable states in ecology. *Frontiers in Ecology and the Environment*, 1(7):376–382.

Blackwood, J. C., Okasaki, C., Archer, A., Matt, E. W., Sherman, E., and Montovan, K. (2018). Modeling alternative stable states in Caribbean coral reefs. *Natural Resource Modeling*, 31(1):e12157. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/nrm.12157.

Boettiger, C. (2018). From noise to knowledge: how randomness generates novel phenomena and reveals information. *Ecology Letters*, 21(8):1255–1267.

Bonnaffé, W., Sheldon, B. C., and Coulson, T. (2021). Neural ordinary differential equations for ecological and evolutionary time-series analysis. *Methods in Ecology and Evolution*, 12(7):1301–1315. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13606.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3):199–215.

Bruggemann, J. H., van Oppen, M. J. H., and Breeman, A. M. (1994). Foraging by the stoplight parrotfish Sparisoma viride. I. Food selection in different, socially determined habitats. *Marine Ecology Progress Series*, 106(1/2):41–55.

Buckner, J. H., Meunier, Z. D., Arroyo-Esquivel, J., Fitzpatrick, N., Greiner, A., McManus, L. C., and Watson, J. R. (2026). Recovering complex ecological dynamics from time series using state-space universal dynamic equations. *Communications Earth & Environment*, 7(1):112.

de Valpine, P. and Hastings, A. (2002). Fitting Population Models Incorporating Process Noise and Observation Error. *Ecological Monographs*, 72(1):57–76.

Doak, D. F., Estes, J. A., Halpern, B. S., Jacob, U., Lindberg, D. R., Lovvorn, J. R., Monson, D. H., Tinker, M. T., Williams, T. M., Wootton, J. T., Carroll, I., Emmerson, M., Micheli, F., and Novak, M. (2008). Understanding and predicting ecological dynamics: Are major surprises inevitable? *Ecology*, 89(4):952–961.

Duarte, C., Amthor, J., DeAngelis, D., Joyce, L., Maranger, R., Pace, M., Pastor, J., and Running, S. W. (2003). The Limits to Models in Ecology. In *Models in ecosystem science*, pages 437–451.

Elton, C. and Nicholson, M. (1942). The ten-year cycle in numbers of the lynx in canada. *Journal of Animal Ecology*, 11(2):215–244.

Frigyik, B. A., Kapila, A., and Gupta, M. R. (2010). Introduction to the Dirichlet distribution and related processes. Technical Report UWEETR-2010-0006, University of Washington.

Fung, T., Seymour, R. M., and Johnson, C. R. (2011). Alternative stable states and phase shifts in coral reefs under anthropogenic stress. *Ecology*, 92(4):967–982.

Hastings, A. (2004). Transients: the key to long-term ecological understanding? *Trends in Ecology & Evolution*, 19(1):39–45.

Hastings, A., Abbott, K. C., Cuddington, K., Francis, T., Gellner, G., Lai, Y.-C., Morozov, A., Petrovskii, S., Scranton, K., and Zeeman, M. L. (2018). Transient phenomena in ecology. *Science*, 361(6406):eaat6412.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366.

Hughes, T. P., Barnes, M. L., Bellwood, D. R., Cinner, J. E., Cumming, G. S., Jackson, J. B. C., Kleypas, J., Van De Leemput, I. A., Lough, J. M., Morrison, T. H., Palumbi, S. R., Van Nes, E. H., and Scheffer, M. (2017). Coral reefs in the Anthropocene. *Nature*, 546(7656):82–90.

Jarman, C. N., Levi, T., and Novak, M. (2026). Symbolic regression for empirically realistic population dynamic time series. *Ecological Informatics*, 99:103988.

Karniadakis, G., Kevrekidis, Y., Lu, L., Perdikaris, P., Wang, S., and Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, 3:422–440.

Kingma, D. P. and Ba, J. (2017). Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG].

Lotka, A. J. (1925). *Elements of Physical Biology*. Williams and Wilkins, Baltimore.

Mumby, P. J., Hastings, A., and Edwards, H. J. (2007). Thresholds and the resilience of Caribbean coral reefs. *Nature*, 450(7166):98–101.

Pennekamp, F., Iles, A. C., Garland, J., Brennan, G., Brose, U., Gaedke, U., Jacob, U., Kratina, P., Matthews, B., Munch, S., Novak, M., Palamara, G. M., Rall, B. C., Rosenbaum, B., Tabi, A., Ward, C., Williams, R., Ye, H., and Petchey, O. L. (2019). The intrinsic predictability of ecological time series and its potential to guide forecasting. *Ecological Monographs*, 89(2):e01359.

Petchey, O. L., Pontarp, M., Massie, T. M., Kéfi, S., Ozgul, A., Weilenmann, M., Palamara, G. M., Altermatt, F., Matthews, B., Levine, J. M., Childs, D. Z., McGill, B. J., Schaepman, M. E., Schmid, B., Spaak, P., Beckerman, A. P., Pennekamp, F., and Pearse, I. S. (2015). The ecological forecast horizon, and examples of its uses and determinants. *Ecology Letters*, 18(7):597–611. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ele.12443.

Pichler, M. and Hartig, F. (2022). Machine learning and deep learning—A review for ecologists. *Methods in Ecology and Evolution*. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.14061.

Rackauckas, C., Ma, Y., Martensen, J., Warner, C., Zubov, K., Supekar, R., Skinner, D., Ramadhan, A., and Edelman, A. (2021). Universal Differential Equations for Scientific Machine Learning. arXiv:2001.04385 [cs].

Rathore, K. J., Buckner, J. H., Meunier, Z. D., Arroyo-Esquivel, J., and Watson, J. R. (2025). Predicting an Unobserved Driver of Refime Shifts in Social-Ecological Systems with Universal Dynamic Equations.

Rogers, T. L., Johnson, B. J., and Munch, S. B. (2022). Chaos is not rare in natural ecosystems. *Nature Ecology & Evolution*, 6(8):1105–1111.

Scheffer, M., Carpenter, S., Foley, J. A., Folke, C., and Walker, B. (2001). Catastrophic shifts in ecosystems. *Nature*, 413(6856):591–596.

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, 25(3).

Tebbett, S. B., Connolly, S. R., and Bellwood, D. R. (2023). Benthic composition changes on coral reefs at global scales. *Nature Ecology & Evolution*, 7(1):71–81.

Urban, M. C., Bocedi, G., Hendry, A. P., Mihoub, J.-B., Pe'er, G., Singer, A., Bridle, J. R., Crozier, L. G., De Meester, L., Godsoe, W., Gonzalez, A., Hellmann, J. J., Holt, R. D., Huth, A., Johst, K., Krug, C. B., Leadley, P. W., Palmer, S. C. F., Pantel, J. H., Schmitz, A., Zollner, P. A., and Travis, J. M. J. (2016). Improving the forecast for biodiversity under climate change. *Science*, 353(6304):aad8466.

Volterra, V. (1926). Fluctuations in the abundance of a species considered mathematically. *Nature*, 118:558–560.

Ward, E. J., Holmes, E. E., Thorson, J. T., and Collen, B. (2014). Complexity is costly: a meta-analysis of parametric and non-parametric methods for short-term population forecasting. *Oikos*, 123(6):652–661.

Chapter 5: Conclusion and Future Directions

I initially started this dissertation thinking I could dive straight into applying SciML methods, specifically symbolic regression, to the long-term monitoring data I knew so well, having worked in kelp forest ecosystems throughout my undergraduate research career. However, after months (or years) of puzzling over equations — recovered from applying symbolic regression out of the box — that were not aligning with my natural history understanding of giant kelp, I realized that the method was not ready to take that leap, which led to the creation of the second chapter presented here. While immersed in the symbolic regression work that would become Chapters 2 and 3, a collaboration emerged around universal dynamic equations (UDEs), also a new scientific machine learning method still very much active in development. Its hybrid structure offered a different strategy for incorporating prior knowledge than symbolic regression's approach, and its capacity for forecasting, a complement to the explanatory power I had been prioritizing in Chapters 2 and 3, made it appealing. Stress testing UDEs offered an opportunity to ask the same underlying question that had motivated my earlier work — can this method be trusted under realistic ecological conditions? — while also probing a dimension of SciML's promise, forecasting, that my other two chapters had not directly addressed. This line of questioning reflects a broader pattern: as machine learning techniques proliferate across ecology, they arrive faster than the field's

capacity to evaluate whether they are suited to the data ecologists actually have. This dissertation is my attempt to slow that process down for two such methods. Taken together, this dissertation assesses the applicability of SciML for ecological time series, but it does not provide a single yes-or-no answer. Each chapter tested a different piece of the larger question, and together they capture a snapshot of the current offerings of these tools and highlight future developments that can be made to make them readily available for ecologists.

Chapter 2 systematically varied sampling density, process noise, cycle asymmetry, and the presence of spurious variables to evaluate whether symbolic regression could recover a known equation from synthetic data. I demonstrate that equation recovery is most strongly influenced by sampling density, with recovery failing below 25 points per cycle. This threshold raises concerns for empirical applications, since most long-term monitoring programs in giant kelp forests sample far less frequently. Above that threshold, however, the presence of process noise improved recovery rates. This is encouraging for empirical applications, because process noise is an unavoidable feature of field data, and my results suggest that it can help rather than hinder recovery, likely by pushing trajectories through a broader region of state space and revealing more of the underlying dynamics. This chapter also highlighted the need for improved post-hoc equation selection. Even when symbolic regression recovered the correct equation, that equation was not always selected from among the candidates. This suggests that the limitation lies not only in symbolic regression's ability to find the correct equation, but also in how candidate equations are evaluated afterward. Whether these findings would hold

for real ecological data, where the true equation is unknown and sampling is often sparse, was the question I took up in Chapter 3.

Chapter 3 applied symbolic regression to long-term monitoring data to test whether it could recover credible biological drivers of giant kelp population growth rate. Because this chapter was a substantial step beyond the controlled conditions of Chapter 2, I first ran additional analyses testing symbolic regression's reliability on an underlying equation simpler than that of Chapter 2, across a range of sampling densities (Appendix A). Symbolic regression recovered the true equation even at the lowest sampling density tested, which was sparser than that of Chapter 3's focal data, suggesting that the threshold of 25 points per cycle from Chapter 2 reflects the complexity of that system's governing equation rather than a universal limit. These results gave me reasonable confidence to proceed with the empirical analyses. I found that symbolic regression could distinguish real from spurious drivers in noisy field data. However, the equations that best reconstructed the population growth rate were not the same as those that best forecast population size. Moreover, because there is no ground truth against which to compare the recovered equations, I argue that they are best treated as hypotheses requiring further testing through field observation or experimentation, rather than as causal mechanistic models in their own right.

Chapter 4 turned to a different SciML method, UDEs. Using synthetic coral reef data, I tested whether UDEs could recover an unknown species interaction term and forecast future ecosystem states across two model formulations and a range of noise and sampling conditions. I found that UDEs forecast reliably more

often than they recover correct mechanistic structure, echoing the separation between reconstruction and forecasting observed in Chapter 3. I also found that a trajectory's initial distance from a system's alternative state boundary governed both outcomes.

Collectively, these results converge on a broader claim: a SciML method's reliability is not a fixed property of the method, but a joint property of the method, the specific data conditions it is given, and the specific criterion used to judge its output. These results also return to a concern raised in this dissertation's introduction: the limited availability and quality of long-term monitoring data are not problems these methods can simply route around. Rather, the sensitivity, especially to sampling density, documented across Chapters 2 through 4 raises the value of sustained, long-term ecological monitoring.

Future Directions

If the reliability of SciML methods depends as much on the data as on the methods themselves, then the most important investments ecologists can make may lie not in the next algorithm, but in how we observe the systems we study. Empirical observations are the backbone of ecology, and they will only become more valuable as data-hungry methods continue to enter the field. Fortunately, the tools available for observing ecosystems are continually advancing. Drones, satellite imagery, and remotely operated vehicles now make it possible to observe systems like kelp forests and coral reefs at frequencies and spatial extents that would be impractical with

diver-based surveys alone. Pairing these technologies with established monitoring programs could raise sampling densities toward the thresholds that methods like symbolic regression require, while preserving the continuity and ground-truthing that make long-term datasets so valuable.

Increasing sampling frequency is not the only path forward, however, and it will not always be feasible. Other fields have long grappled with how to capture the essential information in a system whose underlying dynamics are unknown, and ecology stands to learn from them. Information theory and signal processing offer principled frameworks for deciding how often, and when, a system must be observed to capture its dynamics, from classical results on minimum sampling rates to approaches such as compressed sensing and adaptive sampling that recover signals from sparse or strategically timed observations. Investing in research at this intersection could make long-term monitoring more efficient, directing limited field effort toward the observations that carry the most information about a system's dynamics rather than simply collecting more data.

Finally, and perhaps most ambitiously, ecology needs better tools for inferring causality. I began this work hoping that SciML methods could recover the mechanisms underlying the focal dynamics, and the results were mixed. Symbolic regression could separate real from spurious drivers, and UDEs could forecast more reliably, but neither consistently recovered the correct mechanistic structure. A further challenge also emerged: multiple distinct models could reproduce the same dynamics, a problem of non-identifiability that complicates any attempt to infer mechanism from model fit alone, otherwise known as equifinality. There is real

potential in these methods, but further development is needed before their outputs can be interpreted causally. Other approaches designed explicitly for causal inference in nonlinear dynamical systems, such as empirical dynamic modeling and convergent cross mapping, are worth pursuing in these systems, independently or in concert with SciML methods. The payoff for successfully identifying causality would be substantial. Conservation and restoration efforts depend on knowing which levers to pull, and methods that could reliably identify the causal drivers of ecosystem change from monitoring data would allow managers to intervene where it matters most. Reaching that point will require ecologists and SciML developers to work in close partnership. Ecologists can contribute natural history, data, and a practical sense of what can feasibly be collected in the field, while developers can build methods tested against the messy, sparse, and noisy conditions that real ecosystems present. There is also a need for people who sit between the two, and that is where I have spent the last several years: a jack of several trades, master of none, but, as of this dissertation, a doctor of philosophy nonetheless.

APPENDICES

Appendix A: Symbolic regression on a state-variable-reliant equation

A.1 Introduction

In order to make the jump between Chapter 2 and 3, further synthetic data experiments were needed.

A.2 Synthetic data

To generate our simulated time series, we used the following system of ordinary differential equations inspired by the Bence and Nisbet model (used in Chapter 2),

$$\begin{cases} \frac{dG}{dt} = se^{-\alpha}[1 - a_A A]_+ - \gamma G \\ \frac{dJ}{dt} = \gamma G - gJ - \beta J \\ \frac{dA}{dt} = gJ - mA \end{cases} \tag{A.1}$$

where s is the settlement rate of gametophytes G, α is the mortality rate of gametophytes, a_A is the amount of space occupied by adult kelp A at time t, γ is the conversion rate of gametophytes into juveniles J, g is the conversion rate of juveniles into adults, β is the mortality rate of juveniles, and m is the mortality rate of adults. The notation $[\ ]_+$ is defined as $[x]_+ = x$ if $x \geq 0$ else $x = 0$, and

ensures that the available free space does not drop below 0. In our simulation, we included an unrelated state variable F where,

$$\frac{dF}{dt} = bF(1 - dF) + csin(ft) \tag{A.2}$$

to assess whether symbolic regression could distinguish between intrinsic and irrelevant state variables.

We simulated the system for 200 time steps and discarded the first 165 to remove transient dynamics. From the remaining data, we created four time series that were sampled at different densities relative to the period length of the kelp population cycles: 100, 25, 10, and 5 data points per cycle collected uniformly. Due to differences in the cycle period between the kelp population and F, our time series included approximately 15 post-transient kelp cycles and four cycles of F. Consequently, we did not maintain the intended gradient of sampling densities relative to cycle length for F; in fact, we sampled F at a higher effective density across all time series. However, we did not view this as a limitation of our study, as species in nature often have inherently different cycle lengths that may not be explicitly accounted for in experimental designs or long-term monitoring programs.

To prepare each dataset for symbolic regression, we first calculated the instantaneous per capita growth rate of adult giant kelp by taking the first derivative of the Cubic Spline of $\ln(A)$ since $\frac{d\ln(A)}{dt} = \frac{1}{A}\frac{dA}{dt}$. We then generated four temporal lags for each variable (A, J, G, and F) by interpolating each variable using a piecewise cubic Hermite interpolating polynomial (PCHIP) interpolator and then

lagging the variables 1-4 full time steps back, such that if $t = 4.25$, a lag of 1 would be the value at $t = 3.25$, a lag of 2 would be the value at $t = 2.25$, etc. The interpolation was needed since not all datasets had time points needed to correctly produce the lags.

A.3 Symbolic regression implementation

We used the PySR algorithm to search for candidate governing equations from our datasets (see Jarman et al. 2026 for a detailed description). The calculated instantaneous per capita growth rate was used as the response variable. For each dataset, we used all four variables (A, J, G, F), along with the respective temporal lags as our set of explanatory variables. The operator library included addition, subtraction, multiplication, division, and the exponential function. We disallowed exponential self-nesting (e.g., no $exp(exp())$). For each dataset, we conducted 100 independent searches in PySR, with mean squared error (MSE) as the metric of equation fitness, to check for equation convergence, each running for 22 hours or until MSE reached 2×10^{-9} (justified in Jarman et al. 2026). All searches ($2 \times 4 \times 100 = 800$ total) were conducted on a single compute node of a high-performance computing cluster equipped with 64 CPU cores and 514 GB of RAM.

A.4 Examining the candidate equations

The complexity versus the MSE of the recovered equations for a given dataset and explanatory variable set were plotted, creating the Pareto frontier. The equation that produced the largest drop in MSE with the lowest complexity was selected as the best-performing equation. This equation was then evaluated for its reconstruction of the per capita growth rate-population size relationship and match to the true equation.

A.5 Results

We observed the successful recovery of the true underlying equations for all four datasets examined. The Pareto frontier showed a clear equation selection of complexity 7 for each dataset (Fig. A.1). Each selected equation produced a one-to-one mapping of the true underlying per capita growth rate-population size relationship (Fig. A.2). We found that the recovered equations matched the true equation, including the coefficient values (rounded to the 0.01 place), when evaluating equation formulation match (Table A.1, all variables column).

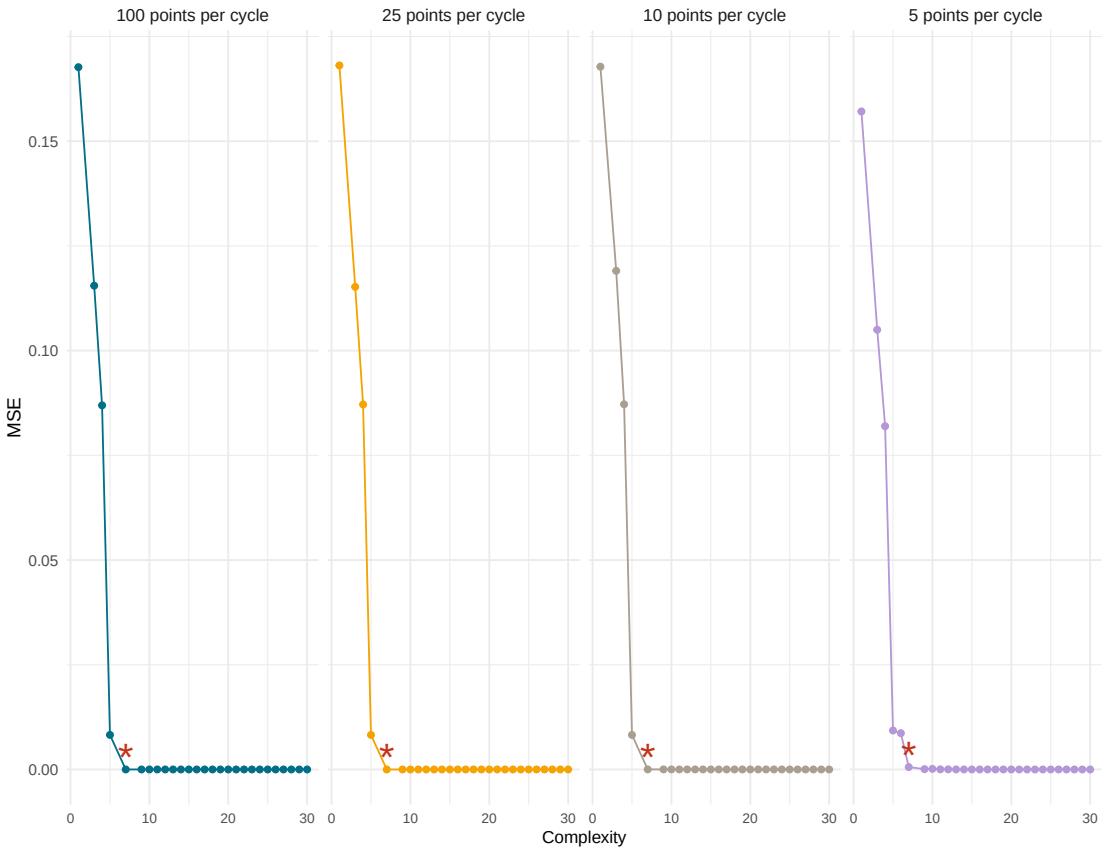


Figure A.1: Pareto frontier for the four datasets examined when PySR included all variables (A, J, G, F) and the respective lags. The top-performing equation from each complexity is plotted, and the star indicates the equation selected based on the trade-off between producing the largest drop in MSE for the lowest complexity.

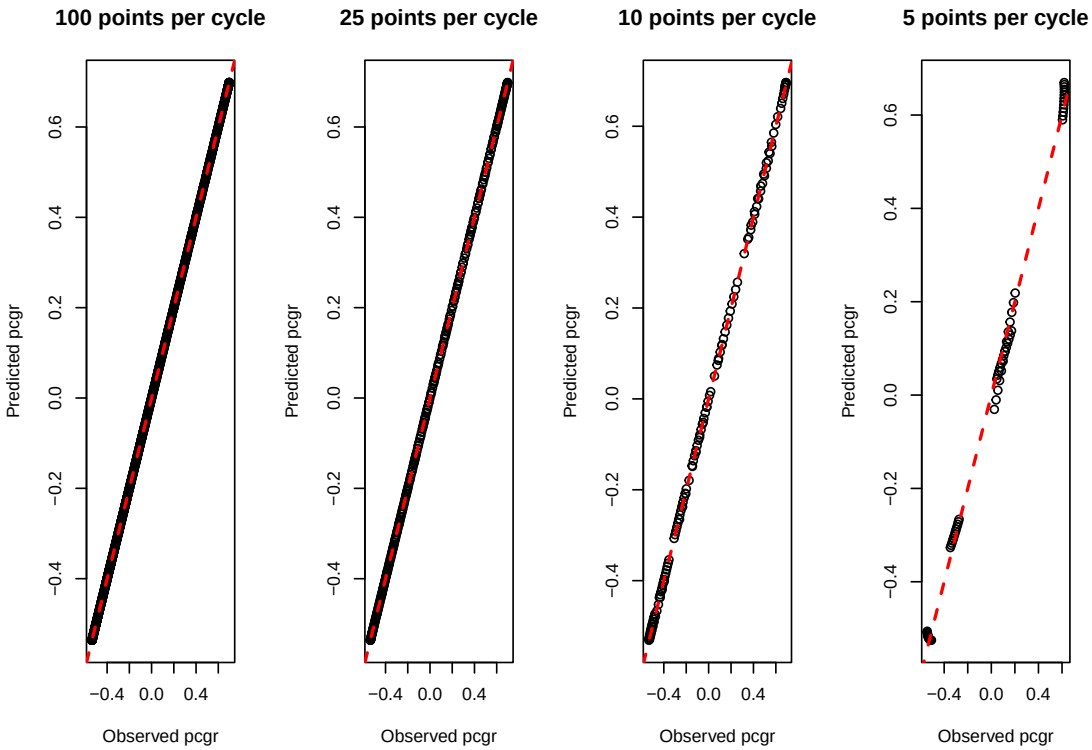


Figure A.2: Observed vs. predicted per capita growth rate (pcgr) from the PySR searches that used all explanatory variables and the associated lags. The predicted per capita growth rate value comes from the starred equations from the Pareto frontier (Fig. A.1). The red line indicates the one-to-one line.

Table A.1: Comparison of starred equations from the all-variable PySR searches. The true equation was $\frac{1}{A}\frac{dA}{dt} = \frac{0.7J}{A} - 1$. Coefficients were rounded to two decimal places.

Points per cycle	**All variables**
100	$\frac{0.7J}{A} - 1$
25	$\frac{0.7J}{A} - 1$
10	$\frac{0.7J}{A} - 1$
5	$\frac{0.7J}{A} - 1$