J. Jake Nichol

Candidate

Computer Science

Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Melanie E. Moses, Chair

G. Matthew Fricke, Member

Abdullah Mueen, Member

Tobias P. Fischer, Member

Laura P. Swiler, Member

Seeking Structure in Complex Systems: From Feature Analysis to Space-Time Causal Discovery with Earth Science Applications

BY

J. Jake Nichol

Bachelor of Science, Computer Science, University of New Mexico, 2016

Master of Business Administration, University of New Mexico, 2017

DISSERTATION

Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

Computer Science

The University of New Mexico

Albuquerque, New Mexico

July, 2025

DEDICATION

To my wife, Whitney, for your unwavering love, support, and stubborn encouragement. You bring peace to my restless mind.

To my parents, Carolyn and Jeff, for setting me on the path to become who I am. I learned more from you than you know.

To my brother, Frank, for being my first partner and confidant. I learned more from you than you should know.

"Whatsoever is contrary to nature is contrary to reason, and whatsoever is contrary to reason is absurd."

—Baruch Spinoza

ACKNOWLEDGMENTS

This dissertation could not have been possible without the significant help of several mentors and colleagues. Their hard work, guidance, and support were an inspiration and an example I will strive to follow throughout my life.

First, thank you to my co-advisors, Melanie Moses and Matthew Fricke. They have given me many opportunities and guidance since I met them during my undergraduate 11 years ago. From swarming robots and the NASA Swarmathon to my dissertation work, they have continuously encouraged all of my ideas while pushing me to be more rigorous, clear, and persuasive. Thank you both for believing in me and giving me the tools to confidently set forth in my career.

Thank you to Laura Swiler. Laura led the Attribution team on the CLDERA grand challenge LDRD project at Sandia National Laboratories, which funded much of my research. She provided expert guidance on my research with excellent feedback and encouragement. She continuously prioritized my education and degree progress and found ways to align it with CLDERA's priorities.

Thank you to Michael Weylandt, who in our very first meeting pressed me to describe how my work would get me closer to graduating. His insistence on excellence and determined direction has made me a more careful scientist and clearer writer. I am certain this dissertation would be of a lesser quality and completed at a later date without Michael's mentorship.

Thank you to Matt Peterson for bringing me into Sandia National Labs as his intern and introducing me to machine learning and my first climate science project. That serendipity was the foundation of this dissertation's success. Your mentorship and guidance built my confidence and gave me the tools to navigate both Sandia and graduate school.

Thank you to Diana Bull and Kara Peterson for your tireless work leading the projects that funded most of this work. I'm so fortunate that when I asked about more Earth science work, you had ideas ready and were willing to bring me aboard. Thank you to everyone who supported my work and worked alongside me at Sandia, including Irina Tezaur, Lyndsay Shand, Kenny Chowdhary, Joey Kart, Warren Davis, Meredith Brown, and Mark Smith. Thank you to everyone on

Sandia's CLDERA team, particularly Tom Ehrmann, Benj Wagman, Hunter Brown, and Joe Hollowed, for help me understand the atmospheric dynamics we sought to capture, developing the E3SMv2-SPA and HSW-V Earth system models, preparing the data, and providing their expertise.

Thank you to my early advisors, Lydia Tapia and Marina Kogan. Lydia took a chance on me to bring me into the computer science Ph.D. program, and while I did not remain in her lab, I am sincerely grateful for the opportunities and lessons I learned. Marina generously took me into her lab while her own path was unclear. She taught me an invaluable lesson: a successful Ph.D. requires a tolerance and acceptance of uncertainty. I took that to heart, and I believe it will continue to serve me throughout my career.

Thank you to Josh Hecker for showing me that a Ph.D. can be fun, attainable, and valuable. I'll always value our friendship and work on the iAnts, Swarmies, and Swarmathon; it was an inspiration.

Thank you to the many postdocs, graduate students, and undergraduate students who had lasting impacts on me throughout my education including Satomi Sugaya, John Ericksen, Jarett Jones, Julie Hayes, Will Vining, Akil Andrews, Humayra Tasnim, Sarah Ackerman, Antonio Griego, and the rest of the Moses Biological Computation lab members over the past 11 years. A special thank you to Zach Abrams, Jayson Grace, and Troy Squillaci for your friendship and the countless hours studying together to get through courses over the years. I could not have passed some of those courses without you.

Finally, thank you to my dissertation committee, Melanie Moses, Matthew Fricke, Laura Swiler, Abdullah Mueen, and Tobias Fischer, for their time and dedication to guiding me and providing feedback to build this dissertation.

Seeking Structure in Complex Systems: From Feature Analysis to Space-Time Causal Discovery with Earth Science Applications

by

J. Jake Nichol

Bachelor of Science, Computer Science, University of New Mexico, 2016

Master of Business Administration, University of New Mexico, 2017

Doctor of Philosophy, Computer Science, University of New Mexico, 2025

ABSTRACT

Complex systems are difficult to study because of their many interacting parts, emergent phenomena, and feedback loops. These systems underpin all life on Earth. We need improved tools for seeking an understanding of them. This body of research presents my investigations into data-driven methods for understanding complex systems, including my invention of a novel causal discovery meta-algorithm for space-time gridded data. I demonstrated machine learning feature importance and causal discovery capabilities for comparing simulated and observed climate data. I developed a new benchmark for modeling space-time dynamics of locally driven phenomena and examined a prominent causal discovery algorithm. Finding that contemporary causal discovery struggles with the high-dimensionality of space-time gridded data, I developed Causal Space-Time Stencil Learning (CaStLe), a causal discovery meta-algorithm for recovering the space-time evolution of advective phenomena. Finally, I extended CaStLe to recover multivariate space-time dynamics. This research enhances scientists' capabilities to explore and understand complex systems in our universe.

Contents

1 Introduction **1**

1.1 The Pursuit of Causal Discovery 3

1.2 Statistical Learning 6

1.2.1 Explainability in Machine Learning 7

1.2.2 Bayesian Networks 9

1.3 Causal Network Learning 11

1.3.1 Definitions, Notations, and Key Causal Assumptions 11

1.3.2 Consistency 16

1.3.3 Validation and Falsifiability 17

1.3.4 Time Series Causal Discovery 19

1.4 Earth Science Challenges 20

1.4.1 Earth Science Data 22

2 Related Work **24**

2.1 Causal Discovery 25

2.1.1 Causal Network Learning 27

2.1.2 Structural Causal Models 30

2.2 Attribution in Climate Science 31

2.3 Causal Discovery for Earth Systems Science 32

2.3.1 Specific Application Challenges 34

2.3.2 Recent Efforts to Overcome Application Challenges 38

2.4 Applications of Causal Network Discovery for Climate Science . . 40

Part I Foundations of Structure Learning for Earth Systems 46

3 Machine Learning Feature Analysis Illuminates Disparity Between E3SM Climate Models and Observed Climate Change 47

3.1 Publication Notes 47

3.2 Abstract 48

3.3 Introduction 48

3.4 Related Work 51

3.5 Data and Methods 52

3.5.1 Data 53

3.5.2 Random Forests 56

3.5.3 Pre-Processing 58

3.5.4 Model Training and Hyper-Parameter Tuning 59

3.5.5 Feature Importance Measurement 60

3.5.6 Model Evaluation 61

3.6 Results 62

3.7 Discussion 65

3.8 Conclusions 68

4 Causal Discovery for Climate Model Evaluation **71**

4.1 Publication Notes 71

4.2 Introduction 72

4.3 Methods 73

4.4 Results 74

4.5 Discussion 74

Part II Local Causal Discovery in High-Dimensional Gridded Data **75**

5 Benchmarking the PCMCI Causal Discovery Algorithm for Spatiotemporal Systems **76**

5.1 Publication Notes 76

5.2 Introduction 77

5.3 Methods 77

5.4 Results and Discussion 78

6 Space-Time Causal Discovery in Earth System Science: A Local Stencil Learning Approach **80**

6.1 Publication Notes 80

6.2 Abstract 81

6.3 Introduction 83

6.4 Background: Causal Discovery and Formal Mathematical Scope . . 90

6.4.1 Related Work: Causal Structure Learning 92
6.4.2 PDE-Like Systems 99
6.4.3 Causal Discovery of Physical Dynamics: Dynamical Constraints 101
6.5 Data: The 1991 Mt. Pinatubo Eruption 104
6.5.1 Held-Suarez-Williamson-Volcanic 106
6.5.2 Mt. Pinatubo in E3SMv2-SPA 108
6.6 Methodology: Causal Discovery with CaStLe 108
6.6.1 Notation 108
6.6.2 Causal Space-Time Stencil Learning 109
6.6.3 The CaStLe Meta-Algorithm 110
6.6.4 Theoretical Properties 114
6.6.5 Methodological Limitations 116
6.6.6 Strategies for Addressing Limitations 117
6.7 Results: Discovering Atmospheric Dynamics in Global Climate Models 119
6.7.1 Validation with HSW-V 120
6.7.2 Extending to More Complexity: E3SMv2-SPA Modeled Aerosols 125
6.8 Validation and Benchmarking 128
6.8.1 Evaluating CaStLe: A Comparative Analysis 128
6.9 Discussion 132

7 M-CaStLe: Uncovering Local Causal Structures in Multivariate Space-Time Gridded Data **143**

7.1 Publication Notes 143

7.2 Abstract 143

7.3 Introduction 145

7.3.1 Background and Motivation 148

7.3.2 Foundations of the CaStLe Framework 149

7.3.3 Theoretical Properties and Empirical Validation of CaStLe . . 153

7.3.4 Research Gap and Motivation for Multivariate Extension . . . 156

7.3.5 Contributions 158

7.3.6 Paper Organization 159

7.4 Methods 159

7.4.1 Phase 1: The Locally Encoded Neighborhood Structure (LENS) 160

7.4.2 Phase 2: The Parent-Identification Phase (PIP) 161

7.4.3 Interpretability: Decomposing the Multivariate Stencil 162

7.5 Experiments 163

7.5.1 VAR Dynamics Benchmarking 164

7.5.2 Advection-Diffusion-Reaction Dynamics 173

7.5.3 Atmospheric Chemistry 183

7.6 Discussion 187

8 Conclusion **193**

8.1 Part I: Synthesis of Foundations Work 194

8.1.1 Machine Learning Feature Importance for Climate Models . . 194
8.1.2 Causal Discovery for Climate Model Evaluation 197
8.2 Part II: Discovery of Local Dynamics 197
8.2.1 Grid-Level Benchmarking of PCMCI 197
8.2.2 CaStLe: Grid-Level Causal Discovery 197
8.2.3 M-CaStLe: Multivariate Grid-Level Causal Discovery 199
8.3 Connections and Research Frontiers 201

APPENDICES **205**

A Understanding Assumptions 205
A.1 CaStLe Assumptions 206
A.2 Causal Discovery Assumptions 207
A.3 Relationship Between Assumption Sets 209
A.4 Potential Violations and their Manifestations 210
B Statistical and Time Complexity 212
B.1 Time Complexity 212
B.2 Statistical Consistency 216
C Asymptotic Consistency 217
D Application to Non-Linear Dynamics: Continuous Systems via Burgers' Equation 223
D.1 Burgers' Equation: Model and Parameters 224
D.2 Advection Angle Estimation 225
D.3 Experimental Setup 226

E Proposed Modification of Statistical Methods for CaStLed Data . . 228
F Limitations of Dimensionality Reduction for Space-Time Causal Discovery 229
G Additional experimental details for Section 3.6 238
H Analysis of Spatial Blocking 238
I Analysis of Assumption Violation Examples 242
I.1 Time Resolution is Too Coarse (Assumption T1) 242
I.2 Time Interval is Too Long (Assumption T2) 242
I.3 Grid Resolution is Too Coarse (Assumption S1) 245
I.4 Block Sizes are Too Large (Assumption S2) 246
J Additional GCM Results 249
K Additional VAR Results 250
L PC-Stable-Single 253
M Completed Data Generation Parameters 253
N Additional vector autoregression model (VAR) Results 255
N.1 PC Comparison Results 256
N.2 Hyperparameter Testing 257

List of Figures

3.1 Comparison of observed, pan-Arctic mean September sea ice extent with predictions from Energy Exascale Earth System Model (E3SM)'s historical ensembles 1-5. The mean of Energy Exascale Earth System Model (E3SM) simulations is shown with 95% confidence interval (shaded). 56

3.2 June feature importance. Standard box-and-whisker plot (McGill et al., 1978) of values for 13 predictions generated by 385 models. The average R^2, anomaly correlation coefficient (ACC), and mean absolute error (MAE) are displayed in the gray boxes. The blue line in each dataset is the mean importance of a random variable in each feature set. 62

3.3 July feature importance. Standard box-and-whisker plot (McGill et al., 1978) of values for 13 predictions generated by 385 models. The average R^2, anomaly correlation coefficient (ACC), and mean absolute error (MAE) are displayed in the gray boxes. The blue line in each dataset is the mean importance of a random variable in each feature set. 63

3.4 August feature importance. Standard box-and-whisker plot (McGill et al., 1978) of values for 13 predictions generated by 385 models. The average R^2, anomaly correlation coefficient (ACC), and mean absolute error (MAE) are displayed in the gray boxes. The blue line in each dataset is the mean importance of a random variable in each feature set. 65

6.1 Schematic overview of the key elements of CaStLe and the process followed in its application to Mount Pinatubo's eruption of stratospheric aerosols. Beginning with Earth system model output, Step 1. is to collect stratospheric wind and aerosol data. Step 2. is to apply our novel CaStLe meta-algorithm to the aerosol data to obtain a causal graph describing the space-time evolution of the aerosols. Finally, we use the wind fields to help validate the causal graph results in Step 3. 105

6.2 Illustration of CaStLe (Algorithm 1) as applied to space-time data on a 4×4 grid. *Step A* (§6.6.3): for every interior grid cell, its 3×3 (Moore) neighborhood is selected. (Note, all four 4×4 grids in the second panel are identical.) *Step B* (§6.6.3): Data are represented in a reduced coordinate space obtained by appending time series from each neighborhood according to its position relative to the neighborhood's center. *Step C* (§6.6.3): during the Parent Identification Phase (PIP), a causal discovery algorithm is used to estimate the parents of the center time series; the resulting graph forms the causal stencil. *Step D* (§6.6.3): the estimated stencil is expanded to its equivalent representation in the original space. Note that each *time chunk* (colored intervals in the center panel) in the reduced space corresponds to an interior grid cell of the original data, and that each edge in the final causal graph reflects to a stencil edge learned during the PIP. See §6.6.3 for details. 110

6.3 Application of CaStLe-PC-Stable to HSW-V simulation of the 1991 Mt. Pinatubo eruption. The stencils estimated by CaStLe (white) capture the underlying high-altitude wind fields (green) using only satellite-measured AOD, with near perfect accuracy in high aerosol regions (red-orange). Autodependencies are shown with black nodes where grid cells cause themselves, and gray nodes where there is no autodependence. All links represent a six hour time lag, the time resolution of the HSW-V dataset. On longer horizons (bottom row), CaStLe is able to recover equatorial wind currents as far away as South America, half-way around the world from Mt. Pinatubo (white triangle). CaStLe accurately identifies the prevailing westerly atmospheric winds because it was able to identify the space-time dependence between neighboring grid cells. Additional details are given in Section 6.7. 139

6.4 Causal maps inferred from the PC algorithm applied naively to all grid cells and CaStLe's equivalent results immediately to the west of Mt. Pinatubo; a 35×35 grid between $-20.00°$ to $50.00°$N and $55.00°$ to $125.00°$E in a 8.5 day span after the eruption. All links represent a six hour time lag, the time resolution of the HSW-V dataset. As expected, PC struggled with the high dimensionality and the discovered dependencies do not conform to the ground-truth understanding that aerosols advected towards the west. It also fails to identify local dynamics, instead drawing most connections over great distances. The PC analysis was computed in 729 minutes on 1,600 grid cells, while the CaStLe analysis was computed in 0.46 seconds. 140

6.5 Application of CaStLe-PC-Stable to E3SMv2-SPA simulation of the 1991 Mt. Pinatubo eruption. The stencils estimated by CaStLe (white) capture the underlying high-altitude wind fields (green) using only total aerosol optical depth (AOD). Autodependencies are shown with black nodes where grid cells cause themselves, and gray nodes where there is no autodependence. All links represent a one day time lag, the time resolution of the E3SMv2-SPA dataset. The heatmap depicts AOD from any source at 50 hPa. The top panel depicts learning from the first 20 days after eruption, which began on day 15. The bottom panel depicts learning approx 6 months after the eruption over a 20-day time period. In the more challenging setting of the fully-coupled E3SMv2-SPA model, our results in the first weeks are still generally consistent with those in HSW-V presented in Section 6.7.1, showing that CaStLe is largely robust to greater complexity. In the bottom panel, the aerosols and winds are in a different regime. CaStLe stencils are still consistent in the tropics and now begin to recover dynamics pushing aerosols northwards above central Asia and southwards through western North America. A more complex model and smaller block sizes illustrate more nuanced dynamics, and there is more to learn from these, however, we leave deeper atmospheric dynamics analysis to future work. 141

6.6 Comparison of CaStLed and non-CaStLed causal discovery approaches on linear-Gaussian dynamics, including Granger causality or FullCI (orange), PC (green), PCMCI (red), and DYNOTEARS (purple), as well as a statistical model of the data generating process (blue) presented with both MCC and F_1 metrics. In the low-sample size regime (T=10, left) CaStLed approaches can accurately recover the underlying causal graph, with performance increasing on larger grid sizes (solid lines); by contrast, non-CaStLed approaches are unable to perform better than mere chance (dashed lines). Even a model based on the underlying data generating process (Sparse VAR, blue) is significantly outperformed by its CaStLed counterpart. In the high-sample size regime (T=150, right), non-CaStLe approaches have improved performance but still compare unfavorably with their CaStLed counterparts. 142

7.1 A conceptual diagram of the Locally Encoded Neighborhood Structure (LENS) that CaStLe constructs for learning underlying local causal dynamics in gridded data. This encoding transforms the original grid space into a local neighborhood structure without marginalization, preserving all of the local relationships in the gridded time series data. 153

7.2 A demonstration of the full CaStLe process to produce a causal stencil graph on an example input 4×4 gridded space-time system. In the Locally Encoded Neighborhood Structure (LENS) phase, neighborhood information is collected from each of the interior grid cells, which are then concatenated to form the Locally Encoded Neighborhood Structure (LENS). Finally, the PIP phase applies an adapted time series causal discovery algorithm to learn the space-time parents of the center node. The learned stencil depicts the underlying space-time structure of each grid cell in the original data. 154

7.3 A schematic diagram of the input, computational phases, and output of Multivariate Causal Space-Time Stencil Learning (M-CaStLe). Similar to CaStLe's procedure (c.f. Figure 7.2), the first phase collects local neighborhood information into the Locally Encoded Neighborhood Structure (LENS), which now collects information for each variable's time series in each grid cell. The second phase applies the Parent-Identification Phase (PIP) to every variable at every position in the Locally Encoded Neighborhood Structure (LENS) to determine which variables cause the center variables from each location in the Locally Encoded Neighborhood Structure (LENS). Finally, the resulting multivariate stencil graph can be decomposed into the *spatial graph* and *reaction graph* for improved interpretability and potential analysis. 158

7.4 Showing precision and recall alongside predicted positive rate, a measure of how often a positive is predicted among all other predictions. As variables increase, the predicted positive rate decreases, which diminishes recall. 169

7.5 A comparison between Multivariate Causal Space-Time Stencil Learning (M-CaStLe)-PC and PC considering the F_1 score for $V = 4$ as the number of links increases on a 4×4 grid. Multivariate Causal Space-Time Stencil Learning (M-CaStLe)-PC outperforms PC in every case because PC struggles with the very high dimensionality of the systems since it is naive to the spatial and variable structures. . . 171

7.6 In simple chains of multivariate stencils, even with an extremely large number of variables, recall can be captured perfectly if the signal strength is large enough. 173

7.7 Time series causal graph for the advection-diffusion-reaction (ADR) model. Nodes represent species u_1 and u_2 at each consecutive time step (t and $t+1$). Straight arrows indicate intra-species relationships across time, capturing diffusion and advection dynamics. The curved arrow represents inter-species causal interactions governed by reaction terms $R_1(u_1,u_2)$ and $R_2(u_1,u_2)$. This graph structure models the temporal and causal dependencies inherent in the advection-diffusion-reaction (ADR) system, providing a framework for evaluating the species interactions found in causal stencil graphs. 177

7.8 Histogram showing the frequency distribution of F_1 scores for reaction graph estimations ($n = 672$). The distribution has a mean of 0.912 (red dashed line) and median of 1.000 (orange dashed line), indicating that the majority of reaction graphs achieved perfect F_1 scores, with a smaller subset showing lower performance. Rare cases with $F_1 = 0$ are when no links were identified in the graph. . . 180

7.9 The distribution of angle estimation errors across diffusion coefficients using boxplots, with black markers representing median errors and green markers indicating the proportion of cases where errors exceed 45°. The global median error is 4.76°, and median errors are generally constant before increasing to 12.36° at 0.40. The proportion of large errors ($> 45°$) starts at 0% for 0.005, increases steadily to 14.3% for 0.40, and shows notable jumps at 0.03 and 0.20. While smaller diffusion coefficients are associated with lower median errors and fewer extreme cases, higher coefficients exhibit a higher likelihood of large errors, with maximum errors reaching 180° for coefficients ≥ 0.03. 181

7.10 The distribution of angle estimation errors across advection velocities using boxplots, with black markers representing median errors and green markers indicating the proportion of cases where errors exceed 45°. The global median error is 4.76°, but median errors vary across velocities, decreasing from 23.19° for $v = 1.00$ to 2° for $v \geq 2.70$, before rising again to 9.00° at $v = 4.00$. The proportion of large errors ($> 45°$) decreases consistently with increasing velocity, starting at 33.3% for $v = 1.00$ and reaching 0% for $v \geq 2.70$, with a slight increase to 9.5% at $v = 4.00$. The increase in errors at $v = 4.00$ may be attributed to species advecting beyond the analysis window too quickly, reducing the accuracy of angle estimation. 182

7.11 The reaction graph decomposed from Multivariate Causal Space-Time Stencil Learning (M-CaStLe)'s multivariate causal stencil graph estimated from Mt. Pinatubo's chemical and radiation pathway computed in Energy Exascale Earth System Model version 2 with Stratospheric Prognostic Aerosols (E3SMv2-SPA). Edges and nodes are colored by the identified strength of cross or auto dependence, with positive dependence depicted in reds and negative dependence in blues. Multivariate Causal Space-Time Stencil Learning (M-CaStLe) correctly identifies the general pathway $SO_2 \rightarrow H_2SO_4 \rightarrow SO_4$ and $\{SO_2, H_2SO_4, SO_4\} \rightarrow$ FSDS. SO_2 has one false positive, appearing to cause both SO_4 directly. 186

D1 Application of CaStLe-PC to advection estimation from non-linear PDE dynamics. In the left panel, the first three columns depict realizations of Burgers' equation under different advection-to-diffusion regimes; the fourth column depicts the causal stencil identified by CaStLe-PC; and the final column compares the estimated advection angle with the true advection angle. The right panel depicts the accuracy of CaStLe-PC under various signal-to-noise conditions. Each combination of advection and diffusion rates were tested with 500 angles sampled uniformly from $[0°, 360°)$. In low-diffusion (high SNR) scenarios, CaStLe-PC can identify the underlying advection clearly (top row of left panel and yellow-green columns in right panel). By contrast, in low-advection (low SNR) scenarios, CaStLe-PC struggles to accurately identify the underlying advective dynamics (bottom row of left panel and blue bars in right panel). Even in highly diffusive scenarios, CaStLe-PC is able to accurately estimate the underlying advection when it is sufficiently strong (around $M/c \geq 20$) as shown in the middle row of the left panel. Additional details are given in D. 227

F1 PCA study of Burgers' equation solution ($\theta = 45°$, $M = 6$, $c = 0.05$). Four empirical orthogonal functions (EOFs) capture $\approx 91\%$ of variance, with spatial patterns (left) and temporal evolution (right). The bottom panels show explained variance distribution and PCMCI causal graph, which fails to accurately represent the known directional advection process in the underlying PDE, highlighting limitations of this approach for local causal structures in space-time systems. 232

F2 PCA-Varimax study of Burgers' equation solution ($\theta = 45°$, $M = 6$, $c = 0.05$). Four empirical orthogonal functions (EOFs) capture $\approx 91\%$ of variance, with spatial patterns (left) and temporal evolution (right). The bottom panels show explained variance distribution and PCMCI causal graph, which fails to accurately represent the known directional advection process in the underlying PDE, highlighting limitations of this approach for local causal structures in space-time systems. 233

F3 PCA study of the HSW-V dataset, in the time interval 21 days post-eruption. Four empirical orthogonal functions (EOFs) capture $\approx$85% of variance, with spatial patterns (left) and temporal evolution (right). The bottom panels show explained variance distribution and PCMCI causal graph, which fails to accurately represent the known directional advection process in the underlying system, highlighting limitations of this approach for local causal structures in space-time systems. 234

F4 PCA-Varimax study of the HSW-V dataset, in the time interval 21 days post-eruption. Four empirical orthogonal functions (EOFs) capture $\approx$85% of variance, with spatial patterns (left) and temporal evolution (right). Since varimax rotation does not preserve the explained variance ordering, we reordered EOFs according to the identified centroid's longitude. The bottom panels show explained variance distribution and PCMCI causal graph, which fails to accurately represent the known directional advection process in the underlying system, highlighting limitations of this approach for local causal structures in space-time systems. 235

F5 PCA study of the E3SMv2-SPA dataset, in the time interval of days 15-35. Nine empirical orthogonal functions (EOFs) capture $\approx$87% of variance, with spatial patterns (left) and temporal evolution (right). The bottom panels show explained variance distribution and PCMCI causal graph, which fails to accurately represent the known directional advection process in the underlying system, highlighting limitations of this approach for local causal structures in space-time systems. 236

F6 PCA-Varimax study of the E3SMv2-SPA dataset, in the time interval of days 15-35. Nine empirical orthogonal functions (EOFs) capture $\approx$87% of variance, with spatial patterns (left) and temporal evolution (right). Since varimax rotation does not preserve the explained variance ordering, we reordered EOFs according to the identified centroid's longitude. The bottom panels show explained variance distribution and PCMCI causal graph, which fails to accurately represent the known directional advection process in the underlying system, highlighting limitations of this approach for local causal structures in space-time systems. 237

H1 Results of CaStLe applied to HSW-V 21 days after the Mt. Pinatubo eruption with three different block sizes, $12° \times 12°$, $20° \times 20°$, and $60° \times 60°$. We find that results are generally consistent over the same area for each block size, with smaller block sizes allowing for additional nuance in some areas. Note that the $20° \times 20°$ block panel is similar to the results shown in Figure 3, but more longitudes were added to get a space factorable by more integers, such as 12, 20, and 60. 240
H2 The PC algorithm and CaStLe applied to E3SMv2-SPA in the $15° \times 15°$ block between 15.00° to 30.00°N and 75° to 90°E. from the day of the eruption to 20 days later. PC struggles to estimate an interpretable and physically meaningful graph of the dependence structure in this area. In contrast, CaStLe is able to identify an interpretable dependence structure that represents the local dynamics within the space. 241
I1 Results of using a coarsened temporal resolution (two-daily) in the E3SMv2-SPA study. CaStLe finds many fewer links in this setting. It is clear that when time is too coarse, causal structures fail to be detected. However, the remaining links that are found are largely true positives, suggesting that CaStLe is relatively robust to coarser time sampling. 243

I2 Results of applying CaStLe to a longer time interval from day 15 to 65. CaStLe identifies more links, indicating it is learning too many causal structures in the data, but still finds many of the true positives we found in our initial study. This indicates that many of the blocks in this interval have temporal causal stationarity, leading CaStLe to perform adequately. 244

I3 Results of applying CaStLe to a time interval that is too long and contains too many causal structures, day 15 to 200. We see that CaStLe identifies many links in each block. Comparing them to the winds is ineffective because the wind arrows are averages over the whole period rather than reflections of how they change in time, which CaStLe is learning from. With such a density of links, it is further challenging to know which are correct and which are spurious. 244

I4 Results of using a coarse grid (9°) in the E3SMv2-SPA study. We find that CaStLe performs very well overall. There are few false positives and it clearly captures the overall advection dynamics of the system. 246

I5 Results of using a coarse grid (18°) in the E3SMv2-SPA study. CaStLe performs well in the early time interval, clearly identifying the east-to-west advection pattern. However, in the later time interval, it finds no spatial structures apart from autodependencies in each block. This is likely because the east-to-west advection is weaker in this period and the grid is too coarse to capture the narrower bands of northward advection that dominates the interval. 247

I6 Results of using block sizes too large in the E3SMv2-SPA study. We see that many true positives are found, but many false positives as well. CaStLe seems to identify multiple contradictory causal structures within many cells, which may lead to more spurious links discovered. Even where links appear correct, they are largely uninterpretable in the presence of contradictions. 248

J1 Application of CaStLe-DYNOTEARS to HSW-V simulation of the 1991 Mt. Pinatubo eruption. The stencils estimated by CaStLe (white) capture the underlying high-altitude wind fields (green) using only satellite-measured AOD, with near perfect accuracy in high aerosol regions (red-orange). On longer horizons (bottom row), CaStLe is able to recover equatorial wind currents as far away as South America, half-way around the world from Mt. Pinatubo (white triangle). CaStLe accurately identifies the prevailing westerly atmospheric winds because it was able to identify the space-time dependence between neighboring grid cells. 249

K1 Matthews correlation coefficient (MCC) comparison between CaStLed and non-CaStLed causal discovery approaches on 2D VAR dynamics for each sparsity level, including Granger causality (orange), PC (green), PC-Stable-Single (cyan), PCMCI (red), DYNOTEARS (purple), and a statistical model of the data generating process (blue). See Section 6.8.1 for experimental details. 251

K2 F_1 score comparison between CaStLed and non-CaStLed causal discovery approaches on 2D VAR dynamics for each sparsity level, including Granger causality (orange), PC (green), PC-Stable-Single (cyan), PCMCI (red), DYNOTEARS (purple), and a statistical model of the data generating process (blue). See Section 6.8.1 for experimental details. 252

M1 Parameter ranges used in our experimental design, showing the link count distribution for each grid size and variable count combination. Each horizontal line represents the span of network links tested, with each parameter combination having at least 30 replicate experiments (n values shown). Our experiments covered grid sizes from 4×4 to 10×10 and 1-6 variables per grid. All experiments used 1000 time samples and coefficient values between 0.1 and 1.0. The network density, d, defined as the ratio of actual links, L, to maximum possible links $d = \frac{L}{(3\times3\times V^2)}$, where $d \in (0, \ldots 0.5]$. Not all density values produced 30 stable systems within our computational constraints, particularly at higher densities. This visualization shows which parameter combinations successfully generated sufficient replicates for statistical analysis. 255

M2 The relationship between link coefficients and the number of links present. As the number of links increases, maximum (blue) and minimum (green) link coefficients show a clear decreasing trend, with their distribution becoming narrower and centered around lower values. This reveals that networks with more links have weaker signals, suggesting that highly interconnected systems cannot be stable with large dependencies. 256

N3 Comparisons between Multivariate Causal Space-Time Stencil Learning (M-CaStLe)-PC and PC considering the F_1 score, precision, and recall for all V as the number of links increases on a 4×4 grid. Multivariate Causal Space-Time Stencil Learning (M-CaStLe)-PC outperforms PC in every case because PC struggles with the very high dimensionality of the systems since it is naive to the spatial and variable structures. 258

N4 Shows that Multivariate Causal Space-Time Stencil Learning (M-CaStLe)'s hyperparameters influence all metrics except recall. This indicates that the poorer recall performance cannot be explained by Multivariate Causal Space-Time Stencil Learning (M-CaStLe)'s hyperparameters. 259

List of Tables

3.1 **Training Features and June Data Excerpt**: total cloud cover percentage (CLT), downward longwave flux at surface (FLWS), pressure at the surface (PS), sea ice extent (SIE), sea ice volume (SIV), near-surface specific humidity (SSH), sea surface temperature (SST), temperature at the surface (TS), wind u component/zonal (uwind), and wind v component/meridional (vwind). Values listed are means over the pan-Arctic grid for each day of the month, rounded to two-decimal places for display only. 55

A Capabilities of CaStLe for Earth science applications. This table summarizes the key methodological advantages of CaStLe and their relevance to specific Earth science phenomena, highlighting applications where grid-level causal discovery enables analyses that were previously infeasible with prior causal discovery approaches. 205

1 Introduction

The principal function of science is to explore and explain our universe. To fulfill this charge, scientists seek to answer the questions of *'how?'* and *'why?'* In this pursuit, we strive to expand human knowledge, improve the well-being of all life, and develop practical applications that transform our world. Complex systems are fundamental to science because they represent the intricate reality of our world. By their nature, complex systems are difficult to study because of their many interacting parts, emergent phenomena, feedback loops, and tipping points. While many complex systems underpin life on Earth, our tools for studying them are limited.

This dissertation investigates the state of the art in data-driven *structure learning* methodologies for explaining and understanding complex systems, particularly for space-time Earth systems. As I use it in this work, structure learning is a class of methods that identify underlying dynamics, or structure, from data. In Part I, I outline the basics of the structure learning task and study how machine learning feature importance and causal discovery can be used to estimate structure in the Earth system.

Finding that the state-of-the-art primarily tackles global-scale emergent structures, Part II focuses on identifying local-scale structures in gridded datasets. This work begins with benchmarking causal discovery algorithms for learning grid-level space-time dynamics. It corroborates that causal discovery algorithms struggle with datasets containing hundreds of thousands of grid cells, each with several orders of magnitude fewer observations in time. This imbalance is one aspect of the *curse of dimensionality* (Bellman, 1957; Bühlmann and Geer, 2011), where many variables relative to sample size limits conventional statistical methods and renders many forms of inference, including causal discovery, unreliable without dimensionality reduction.

To resolve that challenge, I developed a novel method, Causal Space-Time Stencil Learning (CaStLe), that significantly improves the performance and efficiency of causal discovery in local space-time dynamics. It does so via two parts: (i) the Locally Encoded Neighborhood Structure (LENS) reorganizes the given data such that the high-dimensionality of gridded data is eliminated and the sample complexity of the underlying grid-level structure is maximized; and (ii) the Parent-Identification Phase (PIP), which selectively applies causal discovery to minimize the search space while side-stepping spatial confounding. The initial implementation of CaStLe was univariate, in that it could only identify the space-time structure of a single quantity of interest. This work concludes with extending CaStLe to Multivariate Causal Space-Time Stencil Learning (M-CaStLe), which adapts the LENS and PIP to capture space-time structure between multiple quantities of interest.

1.1 The Pursuit of Causal Discovery

The scientific method provides consistent rigor to answer the *'how?'* and *'why?'* questions. With it, we design experiments, collect data on what we observe, and determine what we can learn from those data. Causal inference is the process of answering these questions and determining when such an answer is attainable. Pearl and Mackenzie (2018) suggest that causal inference is conducted via three operations, which he calls the *Ladder of Causation*:

rung one: seeing (observing and collecting information)

rung two: doing (intervention and experimentation)

rung three: imagining alternatives (counterfactual analysis)

Causal discovery is an algorithmic methodology for finding causal hypotheses and eliminating spurious correlations in data, grounded in strict assumptions that represent domain expertise. Machine learning is typically classified as rung one, seeing; it produces observational distributions from which predictions of future states can be made. Causal graph discovery is rung two, doing; it produces interventional distributions in the form of causal models. These can be used to reason about the effects of intervention. Finally, structural causal models and digital twins are examples of rung three, because they enable one to reason about the implications of alternative scenarios. (Peters et al., 2017)

Statistical and machine learning are standard toolsets to quantify and predict relationships when only observational (non-manipulated) data is available. Statistics can describe data and inform us of the underlying distribution, but it generally defers further inference (Pearl and Mackenzie, 2018). Correlated relationships between variables are bidirectional and often ambiguous. Since correlation does not imply causation, one can only make stronger inferences with stronger assumptions.

Machine learning models capture patterns rather than learn to understand underlying mechanisms by computing statistics and fitting functions that separate data. The algorithms learn functions that map input to output, predicting a probabilistic distribution. Its primary goal is to model the given data to predict the classification or future values, i.e., regression. Machine learning has proven to be an informative and useful tool, but prediction is only correlation and, thus, also does not imply causation. The nascent field of explainable machine learning is bearing fruit in some domains. However, it is also limited to descriptive statistics and correlated information. Using explainable machine learning for elucidating the dynamics in a system may be a promising starting point towards finding causality when ground truth is nebulous. Later, in Chapter 3, I will discuss an analysis with random forest feature importance (Nichol et al., 2021a).

The most reliable, though still imperfect, method of estimating causal relationships is with the randomized control trial (RCT) framework. In conducting an RCT, scientists make tacit assumptions called identifiability conditions: the causal Markov condition, ignorability/exchangeability, positivity/overlap, no interference, and consistency. Ideal RCTs meet these assumptions by design; however, errors or biases, such as selection bias, may break identifiability. Hernán and Robins (2020) explain the remaining assumptions for causal inference in detail. I define the causal assumptions important for causal discovery in Section 1.3.1.

The RCT is a powerful tool, but conducting them is not feasible in many cases, such as when randomizing treatment is unethical, impossible, or too expensive or inconvenient. One such example is the Earth science domain. In geophysics, many natural events are impossible to conduct ourselves, i.e., we cannot induce tectonic earthquakes. In other fields, such as atmospheric science, we cannot ethically intervene randomly without fully understanding the downstream impacts of each intervention, e.g., stratospheric aerosol injection for solar radiation management. We have one Earth, and we cannot afford to disrupt it carelessly.

In some cases where RCTs are infeasible, we can conduct observational studies with frameworks like the target trial (Rubin, 1974; Robins, 1986; Dorn, 1953; Feinstein, 1970; Dawid, 2000). However, this relies upon enough sampling to measure a representative distribution of possible outcomes, posing another challenge for causal inference in Earth sciences: we can only observe one instance of the possible outcomes of the Earth's dynamics. One solution may lie in simulations, and numerical Earth system models (ESMs) are an ongoing research area. However, their complexity makes models imperfect, computationally expensive, and challenging to evaluate.

Founded on principles from path analysis (Wright, 1921), contemporary causal discovery is developing into a rigorous mathematical framework, primarily due to work by Rubin (1974); Spirtes, Glymour, and Scheines (1993); Pearl (1995a); Peters, Janzing, and Schlkopf (2017). This framework can mathematically describe the causal questions asked, counterfactuals, interventions, relevant variables to measure, and potential answers to the causal questions. In the past three decades, algorithms have been designed to leverage this framework for reconstructing causal graphs or, interchangeably, causal networks. We can compute statistical relationships and make strict assumptions with observational data and the true underlying causal structure to reconstruct the causal structure that generated the observed data. These assumptions are also known as the identifiability conditions in causal inference. Algorithmically reconstructing causal graphs is called causal discovery, causal network learning, or causal learning.

1.2 Statistical Learning

Peters et al. (2017, p.46) write that "formally, learning causal models is substantially different from the [statistical] learning scenario because it aims at inferring a model that describes the behavior of the system under interventions and not just observations taken from the same distribution. Therefore, there is no straightforward way to adopt arguments from statistical learning theory, to obtain a learning theory for causal relations." Statistical machine learning generally aims to learn a function that fits given data, and we hope it can extrapolate from unseen data. Explainability tools, either derived directly from the model (e.g., decision trees and random forest Gini importance) or many models trained on permuted data, fundamentally describe the models *alone*, rather than the true underlying dynamics in the data.

Tautologically, if the goal is to identify and learn about the dynamics in a system, then causality is fundamentally the only way to reason about those dynamics. As Pearl and Mackenzie (2018) state, contemporary machine learning fundamentally cannot consider the causality in a system because it lacks a language for causality, i.e., counterfactuals and interventions. While we hope a trained model has learned some true underlying function in the data's generating process, it is causally unverifiable. Showing that a model consistently handles new data well increases the confidence that the model has generalized the true causal process, but the error in a model is merely a correlated observation; it does not verify causality.

1.2.1 Explainability in Machine Learning

The black-box nature of most machine learning models poses a big challenge for interpreting and validating their results. *Trustworthy machine learning* and *fairness in machine learning* efforts have turned to uncertainty quantification and explainability methods to validate further, and to understand how and why a particular model has been fit. Some machine learning methods have an inherent explainability, such as decision trees and random forests (Breiman, 2001; Nembrini et al., 2018). Because these models are built iteratively, Gini impurity, the probability of misclassifying an observation, is computed for every node split in the trees. Gini impurities can be aggregated after learning to produce a Gini importance, or feature importance, for each feature. These importance values measure how much each feature contributed to reducing the model's error on average.

Other machine learning models must use ad hoc and post hoc methods to measure the importance of features for model learning. Examples include Shapley values (Lundberg and Lee, 2017), Locally Interpretable Model-Agnostic Explanations (LIME) (Krishnapuram et al., 2016), SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017), and DeepLIFT (Shrikumar et al., 2017). Shapley values, LIME, and SHAP, are model agnostic methods, so they can be applied to support vector machines, random forests, neural networks, etc. DeepLIFT is a member of a class of methods specifically for neural networks. All of these function by measuring the contribution of each feature to the model or a specific prediction. They train many models and vary whether each feature is included by permuting each feature.

Explainability may illuminate causality with respect to the model, but it cannot illuminate causality within the studied system. That is evident because explainability methods make no assumptions about the system itself, nor the data observed. The methods and the models have no way of knowing whether the data and features are representative of the system. They are only aware of the models and data given. Because of this, inferences from these will always fail to rise above making purely associational observations of the given data.

In general, it is acceptable that explainability methods fail to elucidate causality within a system because they make no claims beyond a rigorous attempt at explaining the given model. To these methods, the generating process is not what created the input data but the model itself. They fundamentally address a different question from causal discovery.

1.2.2 Bayesian Networks

Judea Pearl wrote in his book, *The Book of Why*, that he initially made the same mistake as many philosophers and economists, and that I would suggest is made by many in machine learning now: putting probability first and causality second (Pearl and Mackenzie, 2018, p.50). He thought that uncertainty was the most important thing missing from artificial intelligence and insisted that uncertainty be represented by probabilities. With that in mind, he developed Bayesian networks to reason under uncertainty.

Bayesian networks encode conditional probabilities between events. Given that we observe certain probabilities of events, Bayesian networks can compute the likelihood of other events or whether certain facts are true or false. This computation is called *belief propagation*.

Pearl says that while Bayesian networks are still popular for reasoning under uncertainty, they fail to accomplish what he was after: identifying and quantifying causality. Bayesian networks fail to climb beyond rung one in his Ladder of Causation. He says, "Bayesian networks inhabit a world where all questions are reducible to probabilities, or degrees of association between variables..." (Pearl and Mackenzie, 2018, p.51). Pearl solved this problem after putting aside Bayesian networks to develop structural causal models (SCM) and the *Do* notation which provide a mathematical language for writing down *what we know* and *what we want to know*. His *Do-Calculus* (Pearl, 2012) enables us to compute counterfactual and interventional distributions from observational data, as opposed to probabilistic distributions alone.

Bayesian networks are quite similar to causal networks, however. Pearl (1995b) and Pearl and Mackenzie (2018) write about how to transition from a Bayesian network to a causal network in. Bayesian networks' probabilistic and belief propagation properties are still valid in causal networks. The main difference is in how they are constructed. Bayesian networks are a graphical form of conditional probability tables.

A causal network changes the language of the relationships between nodes; the meaning of their construction and interpretation change. Rather than a relationship between nodes indicating that they probabilistically coincide, in a causal network, it indicates which node another node "'listens' to before choosing its value," (Pearl and Mackenzie, 2018, p.129). The *listening* analogy describes the causal assumptions, i.e., the knowledge one has of the system. A missing link between nodes denotes that the two are independent in both Bayesian networks and causal networks. Though, in a causal network, a missing link may also indicate two nodes are indirectly independent. As Pearl notes, this implies that causal assumptions cannot be made-up and can be falsified against the observed data. Pearl's transition from Bayesian networks to causal networks coincided with the work of Spirtes and Glymour's (1991) development of causal discovery, which is the reconstruction of the causal network from observational data.

1.3 Causal Network Learning

In this work, I will focus on conditional independence-based causal network learning[1] (Spirtes et al., 1993; Runge et al., 2019a) for reconstructing causal graphs. Time series adaptations are well suited for the stochastic, highly autocorrelated, and high-dimensional data in Earth science (Runge, 2018a; Runge et al., 2019a). Other forms of causal discovery include nonlinear state-space methods (Arnhold et al., 1999; Sugihara et al., 2012), and structural causal models (Spirtes and Zhang, 2016; Peters et al., 2017).

1.3.1 Definitions, Notations, and Key Causal Assumptions

Causal Graphs

For a multivariate time series $\mathbf{X}$, X^i denotes the time series of the i^{th} variable, $X^i_{t-\tau}$ denotes the time series lagged by τ time steps, and $\mathbf{X}_t^- = (\mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \ldots)$ are lagged time series of $\mathbf{X}$, representing its temporal parents.

A causal graph, or a causal network, is a directed acyclic graph (DAG) or partially-directed acyclic graph (PDAG) that encodes the causal structure between variables in a system. Using DAGs to represent causal relationships is credited to Pearl (1995a, 1998). A causal time series graph adapts the causal DAG to incorpo-

[1] Causal network learning is also known as "causal discovery," "causal graph discovery," and "structure learning," and I will use these terms interchangeably throughout this dissertation.

rate time lags. Each variable has a node for the original, present time t, and every time lag, $t-\tau$. This is theoretically an infinite graph, but in practice, we truncate the graph to a maximum time lag, τ_{max}.

A link between variables in a causal graph, G, marks a dependence between two variables. Variables $X^i_{t-\tau}$ and X^j_t are connected by a lag-specific directed link, $X^i_{t-\tau} \to X^j_t \in G$ for $\tau > 0$, if and only if

$$X^i_{t-\tau} \not\perp\!\!\!\perp X^j_t \mid (\mathbf{X}^-_t \setminus \{X^i_{t-\tau}\}), \tag{1.1}$$

where $\not\perp\!\!\!\perp$ denotes statistical dependence ($\perp\!\!\!\perp$ would denote independence). Thus, Equation 1.1 can be read as *"X^j_t is dependent on $X^i_{t-\tau}$, conditional on [$\boldsymbol{X}^-_t$, excluding the set $\{X^i_{t-\tau}\}$]."* Autodependencies are links where $i = j$. Links from X^i_t to X^j_t are called contemporaneous links. Some algorithms represent these with an undirected edge in the graph, others can use collider rules to possibly orient these (Runge, 2020; Spirtes et al., 1993).

The parents of a node, X^i_t, in G, are mathematically written as

$$\mathscr{P}(X^i_t) = \{X^k_{t-\tau} : X^k \in \mathbf{X}, \tau > 0, X^k_{t-\tau} \to X^i_t\}. \tag{1.2}$$

D-separation

Independence between nodes within a graph is called *d-separation*, for directed-separation, or sometimes just *separation*. It tells us where and when association can *flow*, or be measured, between two nodes. If two nodes are not d-separated, then their data will be correlated. This is an important property for interpreting graphs, but with the assumptions detailed in the following section, we can infer the graph from measured dependencies in data.

To explain d-separation, we first need to explain how association flows between variables in a causal graph. The rules of d-separation operate on the three main components in a causal DAG: the *chain*, $(X \to Y \to Z \to ...)$ and $(... \leftarrow X \leftarrow Y \leftarrow Z)$; the *fork*, $(X \leftarrow Y \to Z)$; and the *collider* or *V-structure*, $(X \to Y \leftarrow Z)$. In chains and forks, association flows between all variables. For the chains/fork example above, $X \not\perp\!\!\!\perp Y \not\perp\!\!\!\perp Z$. Note that the two chains and the fork all have the same independence relationships. This set of independence relationships represent a *Markov equivalence class* of causal graphs.

Dependence is transitive, so we also have that $X \not\perp\!\!\!\perp Z$. In colliders, association flows only between the parents (i.e., X and Z here) and the child (i.e., Y) node. Thus, in the collider example above, $X \perp\!\!\!\perp Z$, but $X \not\perp\!\!\!\perp Z$ and $Z \not\perp\!\!\!\perp Y$.

When we condition on variables, we say they are blocking variables because they may block the flow of association. When a variable is conditioned on, or blocked, in chains and forks, they no longer allow the flow of association between the variables. In this way, we can "close" chains and forks. In the case of the chains and fork above, if we condition on Y, then Y is blocked, and we get the dependence relation $X \perp\!\!\!\perp Z \mid Y$. On the other hand, when the child node in a collider is conditioned upon, we have the opposite; colliders "open," and association flows between parents. In the example above, when we condition on Y, we get the relationship $X \not\perp\!\!\!\perp Z \mid Y$.

From this, the definition of d-separation is as follows:

Nodes X and Y are d-separated given a conditioning set S, with $X, Y \notin S$, if and only if all paths between X and Y are blocked, denoted

$$X \bowtie Y \mid S, \tag{1.3}$$

where S may be empty. D-separation applies to the children of nodes as well. If Z in the collider above had a child node, W, then Z and W would be d-separated just as X and Z are d-separated.

Causal Assumptions

Like many statistical machine learning approaches, causal discovery has specific assumptions, some that depend on the algorithm and the data. In addition, there are three untestable assumptions and require domain expertise to safely assume: the causal Markov condition, faithfulness, and causal sufficiency. These assumptions represent the domain expertise required to infer beyond mere statistical inference to answer causal questions. They are summarized below, and are detailed further in Runge (2018a), which includes clear examples for each assumption that illustrate how algorithms can infer incorrect links when assumptions are not met.

The **causal Markov condition** is necessary for all independence-based methods. It states that if and only if the joint distribution of a time series process, **X**, with the true causal graph G,

$$X_t^- \setminus \mathscr{P}_{Y_t} \bowtie Y_t \mid \mathscr{P}_{Y_t} \implies \mathbf{X}_t^- \setminus \mathscr{P}_{Y_t} \perp\!\!\!\perp Y_t \mid \mathscr{P}_{Y_t}, \tag{1.4}$$

for all $Y_t \in \mathbf{X}_t$, with parents $\mathscr{P}_{Y_t}$. Essentially, this states that d-separation in the graph implies independence in the data. The contraposition is implied:

$$\mathbf{X}_t^- \setminus \mathscr{P}_{Y_t} \not\perp\!\!\!\perp Y_t \mid \mathscr{P}_{Y_t} \implies X_t^- \setminus \mathscr{P}_{Y_t} \not\bowtie Y_t \mid \mathscr{P}_{Y_t} \tag{1.5}$$

The **faithfulness** assumption guarantees that the graph contains all conditional independence relationships that the causal Markov condition implies. A causal graph is faithful if and only if for the joint distribution of a time series process, $\mathbf{X}$, with the true causal graph G, for all disjoint subsets of nodes $Y, Z, S \subset G$

$$X_Y \perp\!\!\!\perp X_Z \mid X_S \implies Y \bowtie Z \mid S. \tag{1.6}$$

This states that d-separation is implied by independence. The contraposition is also implied,

$$Y \not\bowtie Z \mid S \implies X_Y \not\perp\!\!\!\perp X_Z \mid X_S. \tag{1.7}$$

Causal sufficiency is often the more difficult to assume in open and complex systems. It assumes that all common causes of two or more variables are included in the analysis. Formally, a set of variables, S, is causally sufficient for a process, $\mathbf{X}$, if and only if every common cause, or parent, of any two or more variables in W, is included in W, or has some value for all units in the population.

In this work, we are primarily interested in time series data and time-lagged relationships, and these methods require the **time-order assumption**: that the past causes the future, causality cannot travel faster than the speed of light, and that the future cannot cause effects in the past. Depending on the algorithm, assumptions for **stationarity** and **dependency type** are necessary. Glymour et al. (2019) argue that assuming nonstationarity may be allowed in some cases and could even be leveraged as more information. However, as Runge (2018a) notes, stationarity may be indicative of a confounding variable that violates causal sufficiency.

1.3.2 Consistency

Consistency is an important trait of a causal discovery algorithm. If an algorithm is consistent, it has been proven to converge to the true causal graph in the limit of infinite sample sizes. Each algorithm will be defined in part by a set of causal assumptions that are integral to the proof. A common set of those assumptions are described in Section 1.3.1.

Some algorithms, such as conditional independence-based approaches, require additional statistical assumptions. For example, conditional independence-based algorithms testing with a non-parametric regression independence test will need to assume that the function estimator converges correctly, that the noise in the model is additive and independent, and that the unconditional independence test of the residuals converges.

Universal consistency is defined for iterative causal algorithms (Runge, 2018a):

Denoted by $\hat{G}_n$, the estimated graph of some causal estimator from a sample of distribution P, with sample size n, and by the true causal graph G. A causal estimator is said to be universally consistent if $\hat{G}_n$ converges in probability to G for every distribution P,

$$\lim_{n\to\infty} Pr(\hat{G}_n \neq G) = 0. \tag{1.8}$$

This says that the probability of misestimating the true graph becomes arbitrarily small for large sample sizes for any distribution P.

Universal consistency is weaker than *uniform consistency*, which "bounds the error probability as a function of the sample size, giving a rate of convergence" (Runge, 2018a). For a merely universally consistent algorithm, the sample size required for a given error threshold will be different for each distribution, P. Runge (2018a) notes that uniformly consistent conditional independence-based algorithms can only exist under additional assumptions.

1.3.3 Validation and Falsifiability

As discussed by Runge et al. (2019b), method development in causal discovery requires benchmark datasets with ground truth causal structures. CauseMe.net is a website the authors have made for collecting benchmarking datasets for validating causal discovery algorithms. Ground truth for those data sets must come from expert knowledge or randomized experiments. Observational causal networks can be falsified with experimental results. Unfortunately, much of the motivation to use causal discovery is in cases where experimental results do not exist, when randomized control trials are infeasible. When expert knowledge of a causal structure and experimental results do not exist, causal models must be validated by validating each of the causal assumptions made by the algorithm. Since causal discovery algorithms can be proven to be *consistent*, as defined in Section 1.3.2, validating the assumptions can show that the resulting causal network is asymptotically correct to infinitely large sample sizes.

Peters et al. (2017, p.120) also discuss the falsifiability of causal models. They state that traditional machine learning algorithms build probabilistic models, structural causal models can be used for counterfactual models, and causal graphical models can be used for interventional models. They write that two models are equivalent if their corresponding predictions agree. Likewise, we can falsify a probabilistic or interventional model if the corresponding distributions disagree with the observed data. In the case of traditional machine learning, this is commonly computed with validation datasets to ensure that prediction distributions agree with unseen data. In the case of interventional, causal graphical models, if a model correctly predicts the observational distribution but fails to predict the interventional distribution, from a randomized trial, for example, then the model is falsified. Peters et al. (2017, p.120) state that falsifying counterfactual models is difficult in general.

1.3.4 Time Series Causal Discovery

Temporal information is critical to inferring the Earth system's dynamics because the Earth system is a temporal process. Many causal discovery methods imply the inherent temporal aspects of causality without representing it explicitly. Peters et al. (2017, p.10) note that although it is sometimes said that causality discussions must account for time, usually time is not necessary to discuss the effect of interventions. They write that both statistical learning and causal learning can be thought of as "abstractions of an underlying more accurate physical model that describes reality more fully." This is quite obviously true for numerical Earth system models in which differential equations define the dynamics of hundreds of quantities around the globe. It is even more so for the natural system that Earth system models attempt to estimate.

Peters et al. (2017, p.26) note that "an event can only influence events lying in its light cone, since no signal can travel faster than the speed of light in a vacuum." That is, physics explicitly excludes causation from the future to the past. They explain that although this is true, it is widely believed that microscopic and quantum mechanical systems are invertible. They say that the asymmetry of time-order is less critical for describing a causal relationship than the asymmetry of the information carried causal function between events. This is why time is not included in descriptions of physical laws, such as $F = m \times a$. However, the time-order asymmetry is sometimes essential for inferring the direction of causal dependence from data alone (Runge, 2018a).

The consequence of discarding time-order asymmetries in data is that temporal information for interpreting dependence relationships is lost and cannot inform inference. If no other asymmetry is captured in the data, then we need temporal information to elucidate. Some systems, such as climatological processes, are often best summarized in data by time series. Rather than a set of independent samples, a summary in time is necessary to describe Earth system dynamics accurately. Conditional independence-based causal discovery is flexible enough to be adapted for time series input (Runge et al., 2019a).

Many causal discovery algorithms are designed for independent and identically distributed samples. The causal graph can include temporally lagged variables to capture temporal relationships between variables. Each node is multiplied into nodes for each time step. Theoretically, this creates an infinitely large *time series graph*, which each variable, X, is represented as many nodes, $\{X_t, X_{t-1}, X_{t-2}, X_{t-3}, \ldots\}$. In practice, we limit the number of lags to a time step that is large enough to capture the theoretical temporal dependence between the variables of interest.

1.4 Earth Science Challenges

A critical problem in Earth science is identifying the causal pathways from an intervening Earth system event, such as a wildfire, volcanic eruption, or atmospheric injection, to the many impacts on climate, weather, ecology, and human livelihoods in various places on Earth. Causal pathways are paths through a graph of nodes representing various climate impacts or quantities of interest. There is a critical need for analyses that trace the causal path from an intervention, through mediating effects, to impacts that affect life, economic systems, natural resources, and more.

Climate interventions of interest include anthropogenic climate change and natural and artificial stratospheric aerosol injection (SAI). Volcanoes are an occasional source of natural interventions in the climate, injecting teragrams of gases into the stratosphere (Guo et al., 2004a); though eruptions of that magnitude are rare, only occurring every 50 to 100 years. Artificial SAI events are manufactured efforts to change climate regionally or globally. Examples include geoengineering ideas, such as reducing global mean temperatures with sulfuric gas injection. A related example is weather interventions, such as China's rain-making effort, Sky River (Gimeno et al., 2014; Wang et al., 2018), which attempts to bring more rain to a historically arid region. Understanding the downstream impacts of these interventions is vital for evaluating the risks of geoengineering and predicting the impact on neighboring regions.

Reconstructing the causal space-time pathways from intervention to impact will provide critical insights to understand intentional and unintentional interventions. If the global community decides to attempt geoengineering to mitigate climate change impacts, experiments may start small and localized. We need tools to understand the effects of the experiments. If another country decides to implement geoengineering for itself, perhaps at the expense of its neighbors' moisture, then causal analysis will be critical for understanding those impacts.

Many Earth science problems, particularly those considering a relatively short time window, are very data-sparse. Measurement frequency can vary depending on the variable, quality needs, and equipment. Sometimes daily or sub-daily observations are available, but not for very far into the past, often weekly or monthly data is most abundant. The dataset may contain hundreds of variables on millions of grid cells. Frequently, one may want to understand the interdependence of a few variables in several hundred positions with an order of magnitude fewer observations per variable/position pair. This presents a high-dimensional problem, posing poor statistical power and high sample complexity for statistical methods.

1.4.1 Earth Science Data

Earth system data is obtained from several different sources, such as station measurements, satellites, data-fused reanalysis products, and Earth system model output. The data is multimodal and can have a large variety of spatial and temporal resolutions. Station measurements can poll a quantity very often, but only provides data for a point in space. Satellites cover large strips of space over the globe, but measurements can be less frequent, particularly in a specific area of interest, and still often have missing data due to cloud cover. Reanalysis products combine station measurements, satellite data, and weather or climate modeling to produce a hybrid of fused, interpolated data that generally covers all space on the globe.

Reanalysis products and Earth system model output are most convenient because they are spatially complete and temporally consistent, but come with more assumptions than raw measurements. Spatially, the data from these sources is generally arranged on a discrete 3D grid. Grids can take many forms, most common are cubed latitude-longitude grids. Geodesic grids are also used in order to achieve better geometric regularity between cells. (Ebert-Uphoff and Deng, 2014). Earth system model output is frequently analyzed on a per-run basis, a per-model basis with ensembles of runs, or with Coupled Model Intercomparison Project (CMIP) output. CMIP is a collaboration project that combines output from over 100 models, sourced from over 50 modeling centers.

The research in this dissertation addresses many of the ideas and challenges above. It examines the capabilities and limitations of contemporary data-driven modeling. After identifying a key research gap in grid-level causal discovery, this work introduces two novel methodologies for causal discovery of local grid-level dynamics, CaStLe and M-CaStLe, that advance the state-of-the-art in performance and efficiency. I demonstrate these advances with new benchmarking approaches and realistic applications in the Earth sciences. The following chapters detail the path from explainable machine learning for Earth system model evaluation to causal discovery of Earth system dynamics to novel causal discovery approaches for gridded space-time data. With these advances, this work contributes to toolsets for further scientific discovery.

2 Related Work

The RCT was the first innovation to measure causal effects in experiments directly. Ronald A. Fisher is credited with first using randomization for experiments in 1925 (Fisher, 1925; Hall, 2007). Around the same time, Wright (1921) wrote about using what he called path analysis to evaluate and represent directed statistical dependencies. According to Pearl and Mackenzie (2018), path analysis is a direct ancestor to modern causal inference techniques, though it was not recognized as such until the 1950s. Splawa-Neyman et al. (1923) were the first to publish on a potential outcomes framework, providing a notation for causal effects in a randomized setting (Rubin, 2005).

In the 1970s, Donald Rubin's potential outcomes framework opened the door to causal inference in non-randomized observational studies (Rubin, 1974). Potential outcomes try to address the *fundamental problem of causal inference*: once treatment is given to an individual, we can no longer observe what could have occurred had the individual not received treatment. More specifically, as Holland (1986) writes, "it is impossible to observe the value of $Y_t(u)$ and $Y_c(u)$ on the same unit and, therefore, it is impossible to observe the effect of t on u" for potential outcomes, Y, of treatment, t, and control, c, on the individual unit, u (Holland, 1986). While these quantities cannot be observed or computed, this framework allows us to compute other causal quantities based on certain assumptions in non-randomized studies.

Pearl (2012) added to Rubin's potential outcomes notation with the *do-calculus*, a way of clarifying the notation for describing the change in probability distributions of a given quantity from *doing* an intervention on that quantity. In 2000, Pearl presented the structural causal model (SCM), which is a nonparametric form of structural equation models (SEM) (Pearl, 2000, 2001). Economists and sociologists have used SEMs for decades, and they trace their conceptions to Spearson (Tarka, 2018).

Among many other contributions, Robins (1986) introduced a graphical approach to causal inference, the finest fully randomized, causally interpretable structure tree graph. Pearl (1995b) improved on this approach by introducing directed acyclic graphs (DAGs) from computer science and graph theory to causal inference. In that work, Pearl shows how independencies can be described in a Bayesian network graph and how we can similarly represent causal relationships.

2.1 Causal Discovery

Causal discovery, or causal learning, is the pursuit of computing the causal structure from observational data. It intends to outline when an association is causal or merely correlated (Peters et al., 2017). Many algorithms do this by detecting spurious correlations in data, and after making strict assumptions, the causal structure can be found (Runge et al., 2019b). The necessary assumptions are derived from decades of previous causal inference literature. They can be used to prove *consistency*, which is the property that an algorithm converges to the true causal graph in the limit of infinite sample data (Runge, 2018a).

Wiener (1956) published the idea that a variable could be considered causal to another if the ability to predict the second is improved by including information about the first. Granger (1969) later published a practical method for computing on this notion, now known as *Granger causality*. Typically, Granger causality refers to linear bivariate analysis using linear regression models (Peters et al., 2017) or vector autoregressive models (Runge, 2018a). Granger causality has several limitations, outlined in Peters et al. (2017), including an inability to detect indirect causes, failure in the presence of deterministic dependencies, a limitation to only detecting lagged dependencies, and problems with sub-sampled time series (Runge et al., 2019b).

A nonlinear, multivariate approach to Granger causality is called transfer entropy (Peters et al., 2017; Runge, 2018a; Runge et al., 2019b). Peters et al. (2017) state that transfer entropy fails in many of the same scenarios as Granger causality. However, they write, "we emphasize that the qualitative statement about presence or absence of causal inference in the case of two causally sufficient time series only fails for a rather artificial scenario, while quantifying the causal influence via transfer entropy can be problematic also in less artificial scenarios," (Peters et al., 2017, p.207).

2.1.1 Causal Network Learning

In the 1990s, Peter Spirtes, Clark Glymour, and Richard Scheines developed graphical causal discovery, also known as causal network learning (Spirtes et al., 1993). Spirtes and Glymour invented the PC algorithm, named for their first names (Spirtes and Glymour, 1991). This algorithmically attempts to reconstruct the causal structure from observational data. The main underlying idea stems from Reichenbach's Common Cause Principle (Reichenbach, 1956): that if two variables are statistically dependent, there must be a causal relationship between the two or a third common driver of the two.

The full description and pseudocode for PC can be found in Spirtes et al. (1993), and I will provide a brief outline here. It begins with a fully connected graph in which each node is assigned a variable. To leverage Reichenbach's principle, PC iteratively tests each pair of variables, X and Y, for independence, conditioned on a set of one or more variables, $\mathbf{Z}$, denoted $X \perp\!\!\!\perp Y \mid \mathbf{Z}$, while dependence would be denoted $X \not\perp\!\!\!\perp Y$. If two variables are conditionally independent, their link is removed. This first phase results in an undirected *skeleton* graph. In short, the second and third phases use rule sets to orient edges based on principles of how association flows between nodes in a graph. See *d-separation*, detailed here in Section 1.3.1 and in Spirtes et al. (1993). To accurately estimate causal effects, PC relies on strict assumptions, including faithfulness, the causal Markov condition, and causal sufficiency.

Causal sufficiency is one of the more challenging and commonly violated assumptions in causal inference. Spirtes, Glymour, and Scheines' fast causal inference algorithm (FCI) does not require the causal sufficiency assumption (Spirtes et al., 1993). This algorithm does not require the causal sufficiency assumption and, as a consequence, will only produce a Markov equivalence class of partially directed acyclic graphs. The consistency of PC and FCI is shown in Spirtes et al. (1993).

Runge et al. (2019a) published an adaptation to the PC algorithm called PC momentary conditional independence (PCMCI). PCMCI is specifically written for reconstructing lagged-causal time-series graphs (Runge, 2018a). This two-phase algorithm first uses a modified PC algorithm adapted for time series, called PC_1, which attempts to construct a sparse partially directed graph. In the second phase, momentary conditional independence (MCI) is computed for each connected variable pair to reduce the graph further to converge on the estimated causal graph. MCI conditions on both the parents of a given variable, X, as well as the lagged, or time-shifted, parents of X.

Each phase of PCMCI serves a specific purpose in identifying the causal structure. PC_1 removes irrelevant conditions of each variable via iterative conditional independence tests. PC_1 tests only the condition subset with the largest association instead of testing all possible combinations like PC (Runge, 2018a). The MCI phase then controls the relatively high false-positive rate for highly interdependent time series. Conditioning on lagged parents of each variable controls for highly autocorrelated time series data and makes MCI an estimator of causal strength. Both PC_1 and MCI can be implemented with any conditional independence test. Tests for linear models, nonlinear additive noise models, and nonparametric models exist (Peters et al., 2017; Runge, 2018a; Runge et al., 2019b).

Runge (2018a); Runge et al. (2019a) show empirical results from tests on synthetic data to benchmark PCMCI against several other algorithms, including PC, FCI, convergent cross-mapping, LiNGAM (Shimizu et al., 2006), and Granger-causality. They show that PCMCI performs best or above average in terms of high true positive rates and low false positive rates on time series data in several tests with dynamical noise, autocorrelation, and high dimensionality. After identifying the graph, PCMCI was also able to compute true causal effects well (Runge et al., 2019a).

The PC, FCI, and PCMCI algorithms are examples of causal discovery's conditional independence (CI) based causal network learning pillar. These are highly adaptable algorithms because they can be implemented with any conditional independence test. Choosing the correct one depends on specific assumptions about the data and the functional form of the dependencies within. These range from the linear partial correlation test, nonparametric residual-based tests for nonlinear dependencies with additive Gaussian noise (Ramsey, 2014; Runge et al., 2019a), kernel-based approaches (Zhang et al., 2011a), information-theoretic conditional mutual information (Runge, 2018b), and neural networks (Sen et al., 2017).

2.1.2 Structural Causal Models

Because Granger causality and many causal network learning algorithms require a time delay between cause and effect, they cannot easily determine contemporaneous dependencies (Runge et al., 2019b). Contemporaneous dependencies primarily exist when causation occurs faster than the available time-sampling interval. SCMs typically ignore the time-order of causal dynamics; instead, they operate on the assumption that the past is already coded into covariates (Peters et al., 2017). They can estimate contemporaneous effects because they make additional assumptions about the functional forms between dependencies (Runge et al., 2019b).

As SEM's causal-descendant, SCMs are used to model nonlinear causal relationships and require added assumptions for correct estimation (Peters et al., 2017). These allow for the estimation of direct and indirect causal effect, a quantitative estimate of causal strength, without further assumptions on the functional forms interdependencies or distribution of error terms in the data (Tarka, 2018). Peters et al. (2017) overview SCMs in the bivariate and multivariate cases. They describe SCMs' uses for causal discovery and applications to machine learning. Despite their advantages, SCMs have not yet been applied to Earth system sciences (Runge et al., 2019b).

2.2 Attribution in Climate Science

While the climate science literature does not broadly use causal discovery or causal inference techniques explicitly, a primary interest in climate science is detecting and attributing changes in our climate. *Detection* and *attribution* have precise definitions in climate science. Detecting a signal change requires demonstrating that the observed signal differs in a statistically significant way from natural variability. Detection does not imply an attribution of that change. Attribution requires (1) showing that an observed signal is unlikely in natural variability, (2) consistent with estimated changes to the signal given anthropogenic and natural forcing, and (3) not consistent with alternative, plausible explanations of the observed signal (Houghton et al., 2001).

In 1996, Klaus Hasselmann published one of the first attempts to quantitatively attribute climate changes (Hasselmann, 1997). Until then, there was mounting evidence that global warming could be attributed to anthropogenic forcing, but it was largely qualitative or circumstantial. He provides a multi-pattern fingerprinting framework for statistically attributing climate signals.

Hasselmann states that for the attribution problem, further hypotheses regarding the cause of a detected change need to be considered. This demonstrates the counterfactual theory required for causal inference (Pearl and Mackenzie, 2018). He further writes that an obstacle for quantitative signal-to-noise analyses is that they require information on the space-time structure of the predicted climate signal and the climate variability. This implies the same expert provided causal structural knowledge that Runge, Pearl, Peters, and others suggest is critical for effective causal inference (Runge et al., 2019b; Pearl and Mackenzie, 2018; Peters et al., 2017). He then describes an idea similar to causal sufficiency: "A discrimination between competing forcing mechanisms can clearly be meaningfully attempted only if all candidate mechanisms and their associated climate change signals are specified."

Finally, because of the finite nature of real data, Hasselmann states that it can never be ruled out that there may be other overlooked forcing mechanisms that would generate the observed signal. The consequences of this fact are "unequivocal attribution is achieved only in the hypothetical infinite-sequence limit ... We must, therefore, restrict ourselves in principal to a statistical definition of attribution that applies only in the limited sense of establishing a ranking within a given finite set of candidate forcing mechanisms." This essentially iterates the same limitations of finite data in causal discovery (see *consistency* in Section 2.1), detailed by Runge (2018a) and Peters et al. (2017) and described in Section 2.3.1.

2.3 Causal Discovery for Earth Systems Science

Causal discovery has been applied to Earth systems science several times recently. Runge et al. (2019b) cite several papers in which Granger causality, causal network learning algorithms, and nonlinear state-space methods have been applied to climate science problems. Causal network learning applications are relatively recent and primarily focused on climate science (Ebert-Uphoff and Deng, 2012; Kaufman et al., 2020; Kretschmer et al., 2016; Nowack et al., 2020; Runge et al., 2014). These will be detailed further in Section 2.4.

Runge et al. (2015a) present a framework for identifying gateways, mediators, and causal effects in Earth systems. First, they use varimax-rotated principal component analysis on gridded sea-level pressure data to identify localized areas of variability, such as the El Niño Southern Oscillation and the Quasi-biennial Oscillation climate modes, as described in Vejmelka et al. (2014). With those, they can project the original data onto the selected components to create a time series signal for the given quantity in several regions. They then use the regions as nodes in their time series causal discovery algorithm, which identifies the causal relationship between nodes and removes spurious associations found in the data. With that, they are able to identify teleconnections between climate modes and sea level pressure components. Beyond that, they use their established causal networks to compute causal effect metrics for how much a component impacts others in the space-time system.

Runge et al. (2019b) give an overview of causal discovery methods for Earth systems science problems. They identify several classes of causal discovery methods suited for several classes of problems. The classes of problems they list are causal hypothesis testing, complex network analysis, analysis of the causes of extreme events, and causal model comparisons. They also provide examples of these methods used to solve various space-time problems, including an Arctic climate problem, an ecology problem, and a cardiology problem. They discuss the many challenges in applying causal discovery to Earth systems science, from methodological to data to computational and statistical challenges. These are discussed in detail in 2.3.1. Finally, they present future research directions for causal discovery and call for more scientists to work on using causal discovery to solve the challenges in Earth systems science.

Eyring et al. (2019) published a perspective paper on climate model evaluation tools. In it, they say that better tools are required to effectively evaluate the quality of climate models. Climate models are our primary means of studying and experimenting with climate dynamics, and understanding how well they perform is critical to that research. They say, "other promising diagnostic developments on the horizon that should be further advanced include studies that assess responses to perturbations rather than mean climate, and the application of innovative data science methods in Earth system science such as neural networks, machine learning-based anomaly detection techniques, graphical models and causal discovery."

2.3.1 Specific Application Challenges

Runge et al. (2019b) overview the process, data, and computational and statistical challenges faced in applying causal discovery to Earth sciences. The following is a recapitulation of the relevant challenges in that overview.

Process Challenges

The time-dependent processes in Earth systems give rise to strong **autocorrelation** and **time delays** for processes to act on one another. **Nonlinearities** and **state-dependencies and synergies** make selecting an estimation method critical. The wrong method may struggle to disentangle the autocorrelation and internal dynamics and thus fail to achieve the correct causal structure. Various geoscience time series may be acting on **different time scales**, which can be separated to incorporate and interpret different relationships. Many statistical methods make assumptions about the **noise distribution** in the data. Many methods assume additive Gaussian noise, but nonlinear and model-free solutions exist (Peters et al., 2017; Runge, 2018a; Runge et al., 2019b). Processes with heavy tails and extreme outliers may violate linearity and normality assumptions.

Data Challenges

Climate data is space-time, meaning it is measured and computed on a 3-dimensional grid over the Earth's land, oceans, and atmosphere. Hundreds of individual quantities can be collected, leading to a very high-dimensional problem. Extracting features from this data can be a big challenge.

Observational data is incomplete; some processes cannot be adequately measured and quantified. It comes from satellite and station measurements and can include several forms of **measurement error**, such as measurement noise, instrumental biases, and missing data. Often, observational data comes in the form of reanalysis data. Reanalysis is a data assimilation effort to fill data gaps and meaningfully represent quantities of interest via observed data and model output. Finally, satellite measurements only date to 1979, so observational time series are often short. If problems with the observational data are directly related to the processes of interest, then selection bias may be a problem.

Simulation data is vast. Although its spatial resolution is generally smaller than observational data, it is typically 0.5 degrees to 1.0 degrees latitude and longitude. The temporal resolution and timescales of simulations are often higher than observational datasets. They can span hundreds of years and include hourly data.

Because of the high-dimensional and complex data, **variable extraction** is difficult. Time series variables need to be extracted from space-time data; sometimes, feature construction techniques are necessary to form causally relevant features. To do this, fingerprinting (Hasselmann, 1997) and dimensionality reduction techniques (Vejmelka et al., 2014), such as empirical orthogonal functions (EOF)[1] (Hannachi et al., 2007) and varimax-rotated principal component analysis (PCA) (Hannachi et al., 2007), are often necessary. Additionally, these features should be interpretable, representing physical processes in the system.

Often, causal drivers cannot be measured, which leads to **latent, or unobserved, variables** in the analysis. The absence of common causes, or a variable that causes two or more other variables, can lead to spurious links detected in the causal discovery algorithm. Runge notes that failing to account for important

[1] The climate community refers to principal components as EOFs.

drivers, such as anthropogenic climate forcings, may render time series stationary.

Like latent variables, **subsampling** is when a time series is too infrequently sampled. If the causal mechanism acts on a smaller time scale than measured, the mechanism may not be detectable. On the other hand, **Time-aggregation** may reduce the data size and algorithm's computational complexity, but it can make relationships appear contemporaneous or cyclic.

Computational and Statistical Challenges

Sample size and dimensionality is an issue for the scalability and time complexity of many causal discovery methods. While many methods are proven to be correct in the limit of unlimited data by consistency, they are typically relatively slow, some polynomially and some cubically (Runge et al., 2019b). The opposite problem is more likely in observational climate science because, as mentioned earlier, the observed record is still short. When sample sizes are too small, causal relationships may not be reliably estimated. In the case of PC and related methods, conditional independence tests may produce incorrect results, and orientation rules may contradict each other if sample sizes are too small. If dimensionality is high and the sample size is small, conditional independence tests may be underpowered. Lastly, **uncertainty quantification**, which includes statistical test uncertainties and data measurement uncertainties, is an ongoing research challenge for causal inference.

Rejoinder to the Challenges

Most of these challenges discussed are also challenges for traditional correlation, regression, and machine learning methods. However, interpretation of those remains nebulous and often leads to incorrect conclusions. The assumptions made by causal inference and causal discovery merely require subject matter expertise; they encode the domain knowledge to infer causal dependencies and reject spurious association from observational data. Likewise, it is a mistake to embark on traditional statistical and machine learning endeavors without subject matter expertise because of the propensity to mishandle data and make spurious inferences. Because of these factors, Runge et al. (2019b) note that there is "no strong reason to avoid adoption and exploration of modern causal inference techniques."

It seems clear that climate attribution, described in Section 2.2, and causal discovery are fundamentally equivalent endeavors, from intent to results and limitations. Given that both are approached correctly, they are equally valid in asserting the causal dependence between climatological processes. This further iterates Runge's assertion that there is no reason to avoid the exploration of modern causal inference for learning about the Earth's climate.

2.3.2 Recent Efforts to Overcome Application Challenges

As described above, one of the challenges in statistical and causal inference in climate science is the amount of data available. It is common for observational and simulated datasets to be available on a coarse temporal resolution, such as monthly.

When we seek to discover causal dependencies that occur on a finer resolution than measured, we may only find contemporaneous or undirected dependencies. In fact, one of the basic assumptions of the PC, FCI, and PCMCI algorithms is *no instantaneous effects* (Spirtes et al., 1993; Peters et al., 2017; Runge, 2018a). That is, no two variables may act on one another instantly or, practically speaking, within one observed timestep.

To detect contemporaneous links, rather than assume they do not exist, Runge published an adapted version of his PCMCI algorithm, which he calls PCMCI+ (Runge, 2020). Runge notes that autocorrelation is key to increasing contemporaneous link orientation recall. PCMCI+ also "improves the reliability of CI tests by optimizing the choice of conditioning sets and yields much higher recall, well-controlled false positives, and faster runtime than the original PC algorithm for highly autocorrelated time series." Empirically, it maintains performance for time series with low autocorrelation.

Similar to FCI, Runge's Latent PCMCI (LPCMCI) is an implementation of PCMCI to handle the case in which causal sufficiency cannot be assumed, when latent variables exist (Gerhardus and Runge, 2020). This algorithm is tolerant of latent variables while possibly illuminating their existence. Tolerating latent confounding is critical in many open systems in which it is impossible to observe and account for all confounding. The downside of these methods is that they can only estimate the causal structure up to a Markov equivalence class.

2.4 Applications of Causal Network Discovery for Climate Science

Ebert-Uphoff and Deng (2012) may have been the first to apply causal networks to climate science in 2012. They cite inspiration from seminal papers from Tsonis and Roebber (2004) and Tsonis et al. (2006) for their initial work on correlated climate teleconnections, and from Pearl and Mackenzie (2018) and Spirtes et al. (1993), for their causal discovery work. Ebert-Uphoff and Deng apply the PC algorithm to 500 millibars geopotential height at individual grid cell locations. Geopotential height is the height above sea level of a specific pressure level in a specific location, adjusted for the variations in gravity due to changes in latitude.

Ebert-Uphoff and Deng's work is similar to previous work identifying correlated teleconnections by creating a network of dependencies between grid cells on the globe of one variable. Their contribution is to apply causal inference to those teleconnections, removing spurious relationships and identifying a causal network. There are a couple of limitations to their approach. Without including a time series implementation of the PC algorithm, their method treats each day's observation as an independent sample rather than a time-dependent process. They also use neighboring grid cells in the network, possibly violating independence assumptions in the conditional independence tests. Major modes of climate variability may not be adequately captured in single grid cells either, so a weaker signal may lead to undetected links. Still, grid cell level nodes may increase the total captured spatial variability because spatial aggregation and dimensionality reduction techniques can reduce variance.

Kretschmer et al. (2016) applied causal discovery to detect causal effects in Arctic midlatitude winter circulation. They apply a version of the PC algorithm adapted for time series and use seven different variables in the Arctic. They are regional ice, ocean, and atmospheric quantities. They aggregated daily data into monthly means because they were specifically looking for processes acting on a monthly scale. Finally, they used weighted spatial averaging to convert the data into 1-dimensional time series. They validated their findings by careful analysis of the variable selections. They selected variables from work conducted in previous Arctic climatological studies and included proxies for some unmeasurable complex phenomena.

Nowack et al. (2020) used PCMCI to evaluate how similar climate model runs were to observed dynamics. Specifically, they developed graphs depicting how sea level pressure in 50 regions on the globe relates to each other region. In the correlation setting, a relation between variables across space on the globe is called a teleconnection. They discovered causal graphs for 20 models in the Coupled Model Intercomparison Project Phase 5 (CMIP5). Each model was represented by several simulation runs, each used to generate their own graph. They used the F_1 score to measure the similarity between graphs.

Using spatial sea level pressure data, Nowack et al. (2020) detected 50 regions of interest using a common technique in climate science. First, PCA identified the first 100 orthogonal components. Then, they use the varimax rotation algorithm, which has been found to increase the interpretability of components and localize them in space. They note, "principal components without rotation consecutively maximize variance and therefore often mix contributions of physically defined modes such as the El Niño Southern Oscillation, Pacific Decadal Oscillation, or the North Atlantic Oscillation, whose time-behavior is not orthogonal, making patterns more difficult to interpret."

Finally, they select 50 of the 100 components based on spatial separability and frequency spectra. Resulting are 50 discrete regions with high variability and independent patterns. They used the 50 components for each node in the causal discovery analysis. Lastly, they note that "the selection of components defining the network nodes will typically be guided by expert knowledge in conjunction with dimension reduction techniques."

Tibau et al. (2022) built on the dimensionality reduction approach, augmenting it to output grid-cell-level networks. They specifically delineate *mode-level* (dimensionality reduction or cell aggregation) and grid-level causal discovery. Their augmentation is called Mapped-PCMCI, which first applies dimensionality reduction, then computes a mode-level causal network with PCMCI, and finally maps the grid cells within the modes to each other using the network previously constructed. Their resulting network consists of edges between grid cells, but the method assumes that cells within modes are fully connected, i.e., each cell is dependent on all of its neighbors. In contrast, our work specifically seeks inter-cell spatial relationships. Finally, they also describe the failure of a traditional causal discovery approach for grid-cell-level data, "[if] we apply PCMCI directly at the grid-level, the low power of this high-dimensional and redundant estimation problem (see Section 2.2.2) leads to most links being missing."

Recently, a new tradition, causal representation learning, developed out of machine learning to leverage causal reasoning for their models (Schölkopf et al., 2021). While still a developing field, it shows particular promise for estimating relationships in the presence of latent confounding. Boussard et al. (2023) and Brouillard et al. (2024) developed the Causal Discovery with Single-parent Decoding (CDSD) algorithm within the causal representation learning framework and applied it to the climate science field. CDSD performs well in high-dimensional data settings but through a different mechanism. It performs dimensionality reduction by learning latent variables and enforcing a "single-parent" constraint where each grid cell belongs to exactly one latent factor. This naturally clusters grid cells into coherent, often contiguous regions and enables the discovery of causal relationships between these larger-scale patterns. In contrast to grid-level structure learning, CDSD identifies broader teleconnection pathways between regional climate modes. Thus, CDSD abstracts to a higher level by mapping the native grid space to an identifiable latent representation before performing causal discovery.

Several studies have addressed local-scale phenomena. Pfleiderer et al. (2020) applied causal discovery to identify precursors to seasonal hurricane frequency. They utilized the precursors to inform a predictive model. Polkova et al. (2021)

identified local drivers of marine cold-air outbreaks in the Barents Sea. These demonstrate that existing causal discovery approaches can be valuable for seasonal and sub-seasonal phenomena. However, both marginalized large regions prior to analysis, reducing the space's dimensionality, and did not evaluate the space-time evolution of phenomena nor grid-level dynamics.

There are some examples of causal discovery algorithms leveraging spatial information. Zhu et al. (2016) developed pg-Causality that applies space-time pattern mining and a Gaussian Bayesian Network to seek local dependencies in the space-time propagation of air quality data. Sheth et al. (2022) developed STCD for understanding hydrological systems. They constrained the discovery of spatial structures by only allowing higher elevation nodes to be parents of lower elevation nodes because water follows the gravity gradient. While both cleverly use mined or known spatial structure to inform their causal discovery, they are both limited to use in sparse point-measured data from static base stations rather than gridded data. Further, these methods enforce constraints as filtering mechanisms. Neither address the scalability challenges in high-dimensional gridded data.

Parallel Approaches in Neuroscience: Causal Discovery for High-Dimensional Spatial-Temporal Data

Other scientific domains face similar challenges with high-dimensional space-time data. Neuroscience, for example, needs to study mechanisms in brain interactions, and fMRI images may contain thousands to millions of pixels. The anatomy of the brain also exhibits locality constraints. Ramsey (2014) made computational optimizations to the Greedy Equivalence Search algorithm, including sparsity constraints and limiting the distance of potential parents, to recover graphs with millions of nodes. Saetia et al. (2021) marginalized regions of interest in the brain using spatial averaging and then applied the PCMCI algorithm to construct causal graphs. There is a common interest in recovering graphs of high-dimensional grid-level data throughout the sciences. Developing more tools that enhance the estimation and interpretability of causal graphs in these spaces will help advance our understanding of space-time structures across the sciences.

What is clear from prior work is that grid-level analyses are challenging, both statistically and computationally, due to how many grid cell dependencies need to be estimated, the enormous number of observations needed, and the redundant information content of nearby cells.

Part I

Foundations of Structure Learning for Earth Systems

3 Machine Learning Feature Analysis Illuminates Disparity Between E3SM Climate Models and Observed Climate Change

3.1 Publication Notes

Citation: Nichol, J. Jake, et al. "Machine learning feature analysis illuminates disparity between Energy Exascale Earth System Model (E3SM) (E3SM Project, 2018) climate models and observed climate change." Journal of Computational and Applied Mathematics, vol. 395, 2021, p. 113451.

Publication date: October 2021

Publisher: Journal of Computational and Applied Mathematics

Formatting: The original published text has been preserved as much as possible while still adhering to the formatting requirements of this dissertation.

Data and Software Availability: The paper is available at `https://www.osti.gov/biblio/1782577`.

Funding: This work is supported by Sandia Earth Science Investment Area Laboratory Directed Research and Development funding. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and

Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525.

3.2 Abstract

In September of 2020, Arctic sea ice extent was the second-lowest on record. State of the art climate prediction uses Earth system models (ESMs), driven by systems of differential equations representing the laws of physics. Previously, these models have tended to underestimate Arctic sea ice loss. The issue is grave because accurate modeling is critical for economic, ecological, and geopolitical planning. We use machine learning techniques, including random forest regression and Gini importance, to show that the Energy Exascale Earth System Model (E3SM) (E3SM Project, 2018) relies too heavily on just one of the ten chosen climatological quantities to predict September sea ice averages. Furthermore, E3SM gives too much importance to six of those quantities when compared to observed data. Identifying the features that climate models incorrectly rely on should allow climatologists to improve prediction accuracy.

3.3 Introduction

We have observed dramatic declines in Arctic sea ice since the advent of satellite imaging (Stroeve and Notz, 2018). This change is of critical importance to global economic, social, political, and ecological landscapes, not least because of the opening of new navigable sea routes and the impact on wildlife (Arc, 2019; Smith and Stephenson, 2013). As an essential component of the Earth's climate, sea ice loss drives the positive feedback between surface albedo and Arctic warming and may contribute to changes in ocean circulation and mid-latitude weather (Goosse et al., 2018; Sevellec et al., 2017; Cohen et al., 2018; Cvijanovic et al., 2017).

Earth system models (ESMs) provide state of the art simulations of the global climate. They include general circulation and thermodynamic models for ocean and atmosphere, and models for land, sea ice, and land ice processes. Collecting an ensemble of parameterized ESM runs produces a distribution of forecasts that provide bounds on predictions. Simulations of Arctic sea ice in these models include complex interactions between the ice, ocean, and atmosphere. However, limitations in ESMs, such as the inability to resolve critical small-scale processes, can lead to biases when compared to observations. It is, therefore, critical to identify sources of bias.

Previous generations of ESMs have, on average, underestimated the rate of sea ice loss in the Arctic (Rosenblum and Eisenman, 2017). This is apparent in data from the Coupled Model Intercomparison Project (CMIP), which includes simulation results from a broad array of ESMs from modeling centers around the globe. CMIP *phases* mark improvements in the state of the art. The extent of sea ice loss has been a consistent problem, first identified in phase 3 (Meehl et al., 2007; Stroeve et al., 2007). By phase 5 (CMIP5), overall model bias had improved (Taylor Karl E., Stouffer Ronald J., 2012). However, Rosenblum and Eisenman

(Rosenblum and Eisenman, 2017), in an analysis of 118 simulation runs from 40 CMIP5 simulations, found that 89% of CMIP5 model runs underpredicted the rate at which sea ice extent is lost (km^2/decade) by more than a standard deviation; and 2014 loss by an average of 2 million km^2. The disagreement with observation may imply that ESMs' parameters are not well-tuned. Stroeve et al. (Stroeve et al., 2007) suggest this discrepancy is due to missing key causal mechanisms or represent a misunderstanding of underlying physical processes.

The Energy Exascale Earth System Model (E3SM) (E3SM Project, 2018), developed by the United States Department of Energy (DOE), is included in phase 6 (CMIP6) (Eyring et al., 2016) (March 2019). E3SM is a new state of the science climate modeling and prediction project. In CMIP5 and E3SM, the rates of pan-Arctic sea ice change are similar to observation before 1996 but deviate from observation afterward. In CMIP5's case, the rate of loss is less than observed (Rosenblum and Eisenman, 2017), while E3SM's is greater than observed (Section 3.5.1: Data). These differences in sea ice loss rates lead to inaccurate long term predictions about absolute sea ice extent in the Arctic. To our knowledge, our work is the first mechanistic analysis of E3SM accuracy.

We use random forest regression (RFR) (Breiman, 2001) and Gini importance (Nembrini et al., 2018) to determine which E3SM features drive climate predictions. We perform an identical study of historical observations to identify the features that are most influential on prediction of actual sea ice loss. By comparing the two, we determined that E3SM relies too heavily on some features, to the detriment of others, resulting in a divergence from observation. This work elucidates differences in sea ice response between observational data and E3SM simulations and can help improve sea ice prediction.

3.4 Related Work

Stroeve et al. (Stroeve et al., 2012) analyze the agreement between simulated Arctic models, CMIP3 and CMIP5, and observed data. They report that while phase 5 models are an improvement over phase 3 they consistently overestimate forecasted ice extent in the Arctic. The authors suggest that modeling may be improved by including more complex mechanisms such as sea ice albedo parameterization, thickness distributions, and melt ponds.

Rosenblum and Eisenman (Rosenblum and Eisenman, 2017) examined CMIP5's sea ice extent predictions in the Arctic and found overprediction of sea ice extent. Correcting the models required an increase in warming well above observed rates, leading the authors to conclude that the current methods were systematically flawed.

Ionita et al. (Ionita et al., 2018) presented a method for using multiple linear regression to predict the September sea ice extent minimums in the pan-Arctic region and the East Siberian Sea. Notably, they used step-wise regression because it may highlight the underlying coupled physical mechanisms between factors. For the pan-Arctic region, their model was able to predict sea ice extent anomalies for May, June, and July fairly accurately (reporting r-values between 0.84 and 0.9).

Although they found a “skillful” model could be built from their list of Arctic features, they did not analyze the relative importance of those features for their models.

Reid and Tarantino used support vector regression (SVR) to predict the Arctic sea ice extent (Reid and Tarantino, 2014). SVRs were able to construct predictive models, but they only considered sea ice extent as a predictor and could not analyze any other features for their importance. They chose SVRs because they are successful in predicting complex dynamical systems such as climate. The authors reported the comparative results of tuning the SVR, and compared them to CMIP5 ensembles but not to observation.

3.5 Data and Methods

Our methods were able to account for discrepancies in climate simulations and observations. Like multiple linear regression and its associated term-weights, random forests are a machine learning method that is wholly transparent (Breiman, 2001), unlike many other so-called “black box” methods, such as SVRs. We used RFRs and their corresponding Gini importance measure to determine how much influence each input feature has on E3SM predictions. With those tools, we analyzed each feature’s impact on historical sea ice extent and used that information to highlight discrepancies with E3SM.

3.5.1 Data

Our machine learning (ML) models used monthly averages of June, July, and August data from the atmosphere, ocean, and sea ice to predict September sea ice extent for a given year. Results from observational and reanalysis data products are then compared against results from five ensemble members of the E3SM *historical* dataset. The features our ML models are trained on are a subset of physical quantities simulated by E3SM in the Arctic. We chose these features because they match observable features in nature and that we hypothesized would be good predictors of sea ice loss. Each feature of each dataset is a time series beginning with the start of the satellite era in 1979 and ending with the last year of available E3SM output, 2014.

The observational data included monthly sea ice extent computed from gridded, daily, passive-microwave satellite observations of sea ice concentration provided by the National Snow & Ice Data Center (Peng et al., 2013). Sea ice concentration is a percentage value of ice in each grid cell, and sea ice extent (SIE) is computed as the total area of cells containing more than 15% ice. Sea ice volume reanalysis data were provided by the Pan-Arctic Ice Ocean Modeling and Assimilation System (Schweiger et al., 2011). Atmospheric data (total cloud cover percentage (CLT), downward longwave flux at surface (FLWS), pressure at the surface (PS), near-surface specific humidity (SSH), temperature at the surface (TS), wind u component/zonal (uwind), and wind v component/meridional (vwind)) were from an atmosphere reanalysis provided by the National Centers for Environmental Prediction (NOAA et al., 2019a). Sea surface temperature (SST) was provided by the National Oceanic and Atmospheric Administration (NOAA et al., 2019b). For each of the atmospheric data variables, as well as SST, monthly Arctic area averages were computed from the global gridded fields.

We used the DOE's E3SM for climate simulation data in this work (E3SM Project, 2018; Golaz et al., 2019). E3SM version 1 was a fork of the community Earth system model (Kay et al., 2015), which was a part of the CMIP5 collection analyzed by Rosenblum and Eisenman (Rosenblum and Eisenman, 2017). E3SM is a global model comprised of submodels for land, atmosphere, land ice, sea ice, oceans, and rivers. Specifically, we used data from E3SM's *historical* ensembles 1-5 at one-degree global resolution.

E3SM published five historical ensemble runs to offer a distribution of forecasts. The runs were initialized from different years of a 500-year pre-industrial control simulation. The historical runs start in 1850, running for 165 years to 2014. The final 36 years, 1979 to 2014, were used in our analysis to match the years of observed data. Small differences in each run's initial conditions can significantly impact long-term results, though average behavior between runs is expected to be consistent.

Table 3.1 summarizes the observed features we collected; an excerpt of June values is included. Each feature is a time series of the feature's mean in a given month from 1979 to 2014. Values in the time series are an area-sum over the

Table 3.1: **Training Features and June Data Excerpt**: total cloud cover percentage (CLT), downward longwave flux at surface (FLWS), pressure at the surface (PS), sea ice extent (SIE), sea ice volume (SIV), near-surface specific humidity (SSH), sea surface temperature (SST), temperature at the surface (TS), wind u component/zonal (uwind), and wind v component/meridional (vwind). Values listed are means over the pan-Arctic grid for each day of the month, rounded to two-decimal places for display only.

	June										Sept.
Year	CLT (%)	FLWS (W/m^2)	PS (Pa)	SIE (10^6km^2)	SIV (10^6km^3)	SSH (mg/kg)	SST (°C)	TS (°C)	uwind (m/s)	uwind (m/s)	SIE (10^6km^2)
1979	42.08	256.56	97 930.00	12.53	29.79	4.31	0.56	273.46	0.94	0.48	5.90
1980	40.89	259.51	97 901.00	12.20	29.15	4.44	0.68	274.67	0.99	0.47	6.83
1981	40.47	258.13	98098	12.43	26.82	4.27	0.65	274.27	0.06	0.06	6.40
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2012	40.36	271.60	98 105.00	10.67	16.00	5.12	1.39	277.28	−0.03	−0.06	3.55
2013	40.66	266.93	97 989.00	11.36	17.54	4.98	1.26	276.50	0.93	0.42	5.27
2014	39.84	263.94	98.19	11.03	17.68	4.72	1.47	275.67	0.00	0.04	5.38

pan-Arctic oceanic region. Each feature's monthly data is a mean of every Arctic sample in the given month, resulting in a single value per month. Generally, the observational and reanalysis datasets have similar magnitudes to the simulation data. However, for CLT, the NCEP reanalysis is significantly lower than the E3SM data. This is a known bias in the NCEP reanalysis data, and future work could investigate feature analyses of alternative reanalysis datasets (Zib et al., 2012).

The data used in this work is publicly available on the E3SM website. The five historical ensemble runs were retrieved from the v1 one-degree data CMIP6 release. To disambiguate them from our machine learning models and observed data, we will refer to E3SM's *historical ensembles 1-5* as *simulations 1-5*, simulation runs, or simply E3SM runs for the remainder of this paper. Figure 3.1 shows a comparison of the observed and simulation datasets evaluated in this work.

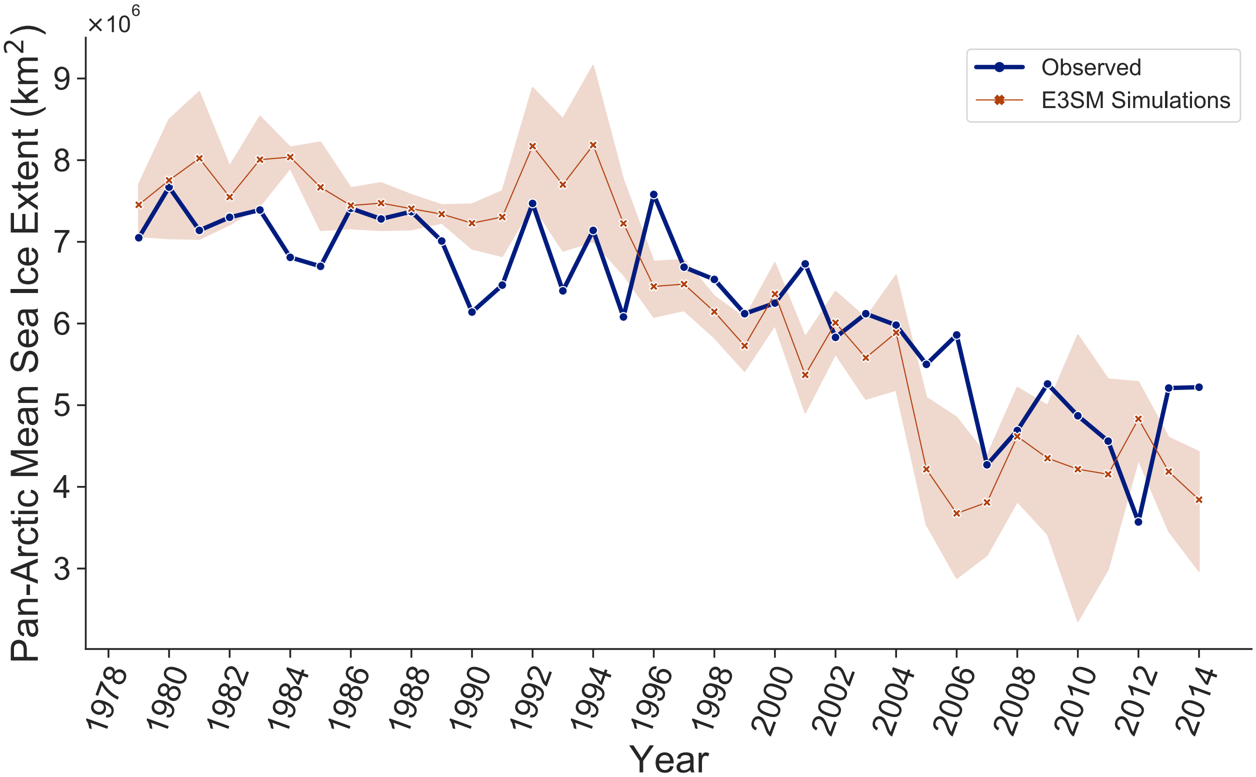


Figure 3.1: Comparison of observed, pan-Arctic mean September sea ice extent with predictions from E3SM's historical ensembles 1-5. The mean of E3SM simulations is shown with 95% confidence interval (shaded).

3.5.2 Random Forests

We found that linear models performed poorly on our data. For this work, we used RFR models because they are relatively simple, intuitive models that can learn nonlinear relationships between features. As a part of their training, the decision trees in random forests generate Gini impurity measures. These measures are aggregated after training to determine the Gini importance of each feature. In our case, we computed importance as the total reduction in mean absolute error (MAE) caused by each feature.

RFR is an ensemble learning technique, similar to a combination of bootstrap aggregation (bagging (Breiman, 1996)) and decision tree regression. Bagging is a method to combine the knowledge of many naive estimators, or trees in our case, by providing a subset of the full sample set to each estimator. The result is the average of many noisy, but unbiased, estimators, reducing overall variance. Random forests improve the bagging method by choosing random subsets of the feature set for each node split in every tree (Banfield et al., 2007). The number of random features each node considers, and when to split are tuned hyper-parameters. The final forest's estimate is the average prediction from the random trees.

For N trees, $T_1, ..., T_N$, random forest regression prediction is computed as follows:

$$RF(N) = \frac{1}{N} \sum_{n=1}^{N} T_n(x)$$

given the training sample, x.

The random forest implementation we used was the random forest regressor from Python's sci-kit learn package (Pedregosa et al., 2011). The implementation uses a perturb and combine technique (Breiman, 1998a) made for tree regressors. Perturb and combine reduces test set error by introducing a diverse set of regressors via randomized regressor construction. For the rest of the data analysis, we used Python's Numpy package (Van Der Walt et al., 2011). We utilized Python's Seaborn package (Waskom and the seaborn development team, 2020) for data visualization.

3.5.3 Pre-Processing

To prepare the data for training, we split it into training and testing years. Our goal was not to develop predictive models for next year's sea ice extent. We were more interested in finding models that have learned the data well that we then used for feature analysis. Thus, we split the training and testing data randomly.

Because some years are easier to forecast than others, we should model every combination of training and testing years. For 36 total years and 18 testing years, we computed $\binom{36}{18} = 9\,075\,135\,300.00$ total combinations of training and testing years. Since it is infeasible to train that many models and evaluate each feature's importance, we used this standard method to compute a sample size:

$$\frac{(z\text{-}score)^2 \times \sigma \times (1-\sigma)}{e^2}$$

with a *z-score* computed with 95% confidence, $e = 5\%$ margin of error, and standard deviation σ, which yielded 385 sample sets on which to train and test our models. We illustrate with 18 testing years because it is the maximum value of $\binom{36}{X}, X \in [1, 36]$.

Decision trees, and thus random forests, are scale-invariant (Breiman et al., 1984). This means that although our data varies greatly in scale between, for example, sea ice extent, in millions of km^2, and wind speeds, less than $1.00\,\mathrm{m/s}$, the models' accuracy is unaffected. This is an advantage over many other ML models, and we can leave the data generally untouched. However, random forests extrapolate poorly for data outside of their training's minimum and maximum values (Hengl et al., 2018). This presented a problem for our analysis of the dataset because, as shown in Figure 3.1, the latter third of the data has values generally lower than any in the first two thirds. We detrended training and testing data separately to mitigate that problem by forcing the data to have a zero mean. After training and fitting our models, we retrended the data and the model's predictions to evaluate their error.

3.5.4 Model Training and Hyper-Parameter Tuning

Finally, we trained RFR models on the data the training splits provided. Note that the trees in our forests were allowed to grow until all leaves were pure, even if they contained a single sample. Decision trees are often pruned to reduce overfitting, but Breiman (Breiman, 1998b) suggests letting trees grow fully in random forests to boost accuracy and increase ensemble diversity. Banfield et al. (Banfield et al., 2003, 2007) also discuss ensemble size in random forests and conclude that many more trees are necessary than are typically used. Ensemble size is an important hyper-parameter to tune because the number of trees in the forest directly impacts the possible feature sets the forest can explore, and too many trees can reduce a random forest's performance while also sacrificing run-time. Our forests comprised 250 decision trees. The number of trees was determined empirically. Forests of size 10, 50, 100, 250, 500, and 1000 trees were evaluated and their performance was measured on the basis of the test $\overline{R^2}$ (average R^2) and average test anomaly correlation coefficient (ACC), which are detailed in Section 3.5.6. We found that 250 tree models maximized $\overline{R^2}$ and $\overline{ACC}$. Lastly, the trees in each forest used mean squared error as their nodes' splitting criterion.

3.5.5 Feature Importance Measurement

We used Gini importance because of the non-linearities in climate data; in particular, Gini importance is not susceptible to data multicollinearities. Given that all of our features come from the same complex system, it would be difficult to eliminate features by simple correlation measures. In standard usage, Gini importance is normalized to compare relative importance within a single dataset. We chose to preserve the absolute importance values, letting us compare across datasets.

We also considered drop-column and permutation importance methods (Breiman, 2001). However, we found them to be unsuitable because they are highly susceptible to multicollinearity. Because many physical processes are directly acting on each other, Arctic features are inherently correlated, and any leave-one-out importance method will highlight that correlation. We found that the correlation leads these methods to attribute more importance to the least correlated feature, and it becomes difficult to glean meaningful insights.

3.5.6 Model Evaluation

We used the R^2 (coefficient of determination) from the Nash-Sutcliffe efficiency definition, given by:

$$R^2(\hat{y}, y) = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \overline{y})^2},$$

where y are the true values, $\hat{y}$ are the predicted values, and $\overline{y}$ is the mean of y. This definition has a range of $(-\inf, 1]$ where 1 is the best possible score.

In addition to $\overline{R^2}$, we evaluated model performance with average MAE ($\overline{MAE}$) and $\overline{ACC}$. Again, average here means the mean value measured in 385 models with random training and testing year splits. Since $\overline{MAE}$ is in millions of km^2, we took the Sea Ice Outlook's 2019 season report (Bhatt et al., 2020) as a baseline. This report includes several different types of data-driven models and presents one-year forecasts. These should have less error than ours, given how many more years we forecasted at once. With the exception of a few outliers between 2008 and 2019, sea ice forecast error was between -0.4 and 0.6 million km^2.

ACC is the Pearson's correlation coefficient (r-value) of sea ice extent anomalies. A time series' anomaly is a measure of the data's deviation from its *climatology*. In our case, the climatology is the mean value of the true values the models are attempting to forecast. This function is defined by:

$$ACC(\hat{y}, y) = \frac{\sum[(\hat{y} - \overline{y})(y - \overline{y})]}{M \times \sigma_{\hat{y}} \times \sigma_y}$$

where y are the true values, $\hat{y}$ are the predicted values, M is the number of samples

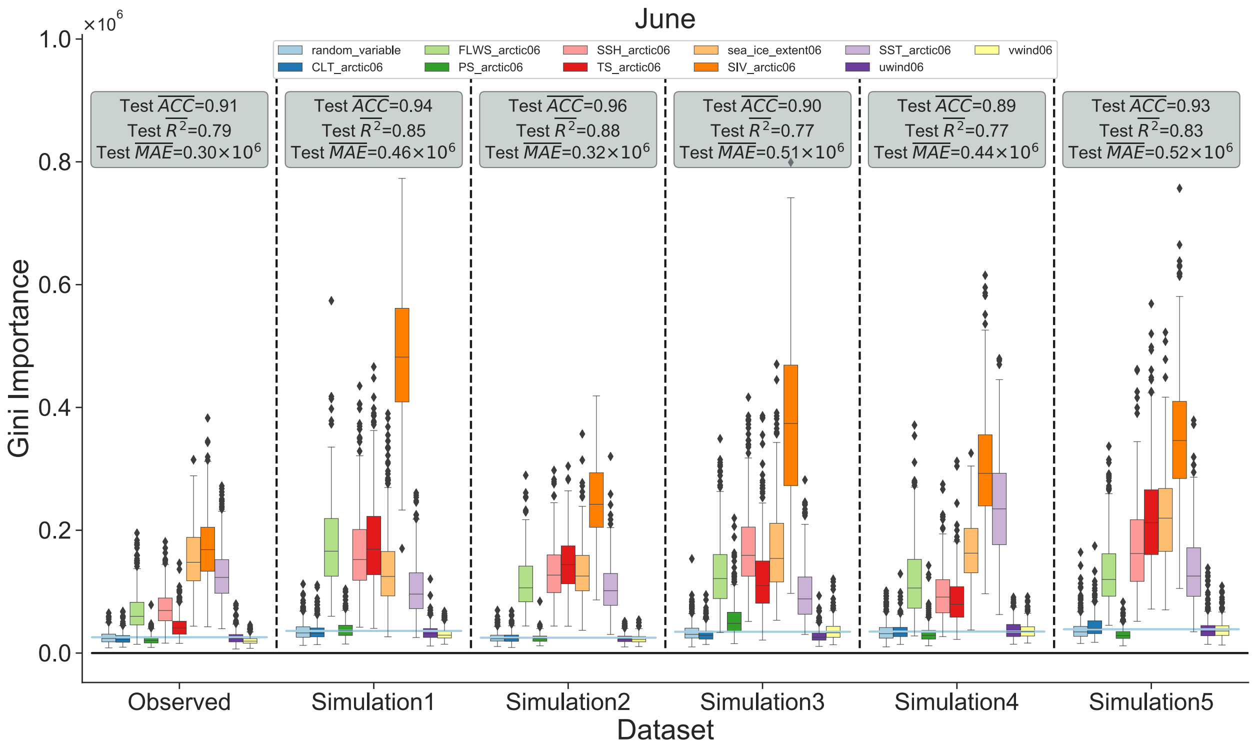


Figure 3.2: June feature importance. Standard box-and-whisker plot (McGill et al., 1978) of values for 13 predictions generated by 385 models. The average R^2, anomaly correlation coefficient (ACC), and mean absolute error (MAE) are displayed in the gray boxes. The blue line in each dataset is the mean importance of a random variable in each feature set.

in y and $\hat{y}$, $\bar{y}$ is the mean or climatology of y, $\sigma_{\hat{y}}$ is the standard deviation of the predicted values, and σ_y is the standard deviation of the true values.

3.6 Results

Our goal is to learn the importance of climate features on the predictions made by E3SM and compare that to the actual importance of those features on observed sea ice extent. We found that was best accomplished by training RFRs on 23 uniformly randomly chosen years and testing with the remaining 13. Our performance measure was based on the mean of $\overline{R^2}$ scores among datasets for the June

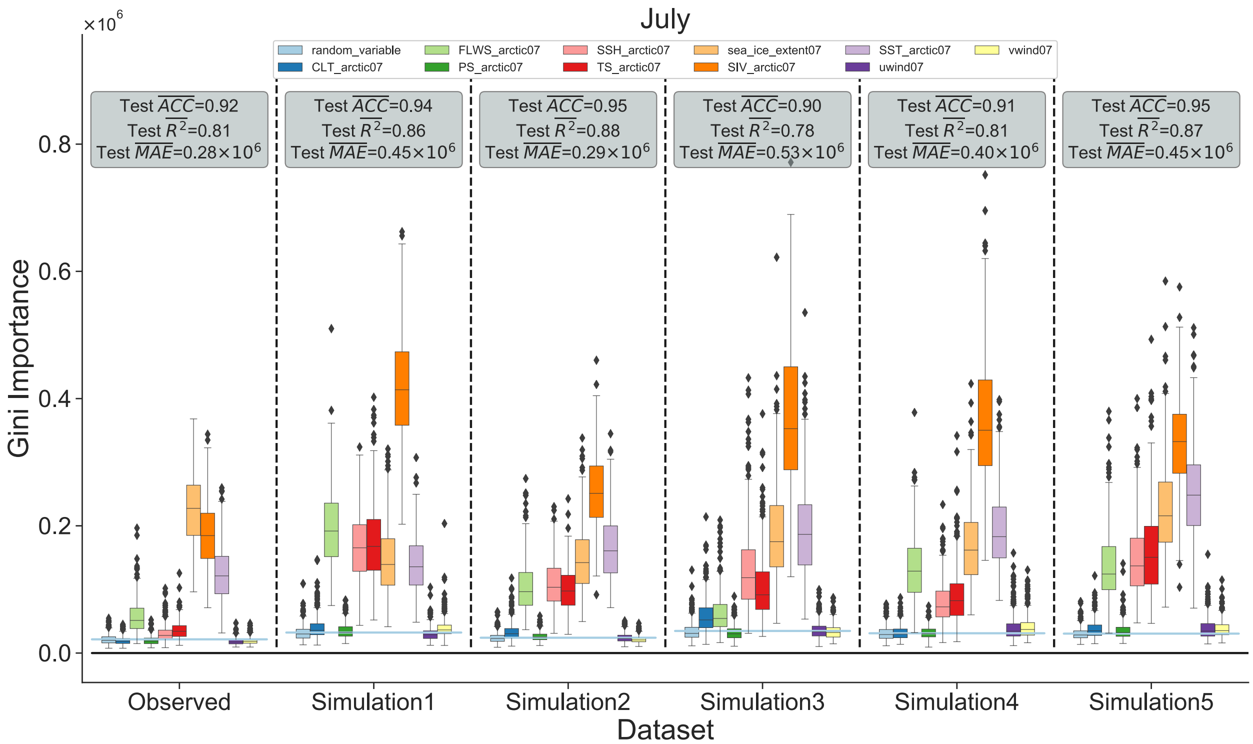


Figure 3.3: July feature importance. Standard box-and-whisker plot (McGill et al., 1978) of values for 13 predictions generated by 385 models. The average R^2, anomaly correlation coefficient (ACC), and mean absolute error (MAE) are displayed in the gray boxes. The blue line in each dataset is the mean importance of a random variable in each feature set.

input data. This train-test-split resulted in maximum and minimum $\overline{R^2}$ scores of 0.88 and 0.77, respectively, yielding a measure of 0.83. $\overline{R^2}$ denotes the average R^2 of the 385 models.

We replicated our analysis for each month between June and August, predicting September SIE. Each subsequent month generates less error. Within each dataset, each feature's relative importance changes. Some features' importance is correlated with the progression of months, while others appear to change randomly.

Figure 3.2 shows June's feature importance values. The average train and test error values indicate that the models generally learn the data well. The blue line shows the mean feature importance of a random variable included in each model's feature set. The random variable indicates a lower bound on importance; any feature with an importance value near this line has virtually no importance. We found that adding a random variable decreases individual model performance, but the effect is minimized when taking the mean over every model.

There are some similarities between each dataset. They share the same list of six important features, though their order and magnitudes differ. sea ice volume (SIV) is consistently the most important, though the degree of absolute importance varies. SIV, TS, SSH, SIE, FLWS, and SST are important in each dataset. The datasets, except for simulation 3, share the same list of unimportant features as well. These are CLT, PS, uwind, and vwind. One apparent exception is June's PS in Figure 3.2: simulation 3; however, excluding PS from the training data, results in a negligible difference in $\overline{R^2}$ (0.7681 vs. 0.7682).

July features, shown in Figure 3.3, predicted as well or better than June in each of our error metrics; simulation 3 had the lowest $\overline{R^2}$, 0.78, and simulation 2 had the highest, 0.88. The same features were important in July as in June, but the relative importance values changed. June's sea ice extent became more important in the observed dataset, surpassing the importance of SIV. SSH became less important in the observed dataset, too, settling just above the random variable. SSH remained as important in the simulation datasets.

The most dramatic change in importance occurs in August. These results are in Figure 3.4. Error was significantly better with simulations 3 and 4 having the minimum $\overline{R^2}$, 0.87, and simulations 1 and 2 having the maximum, 0.91. In August, sea

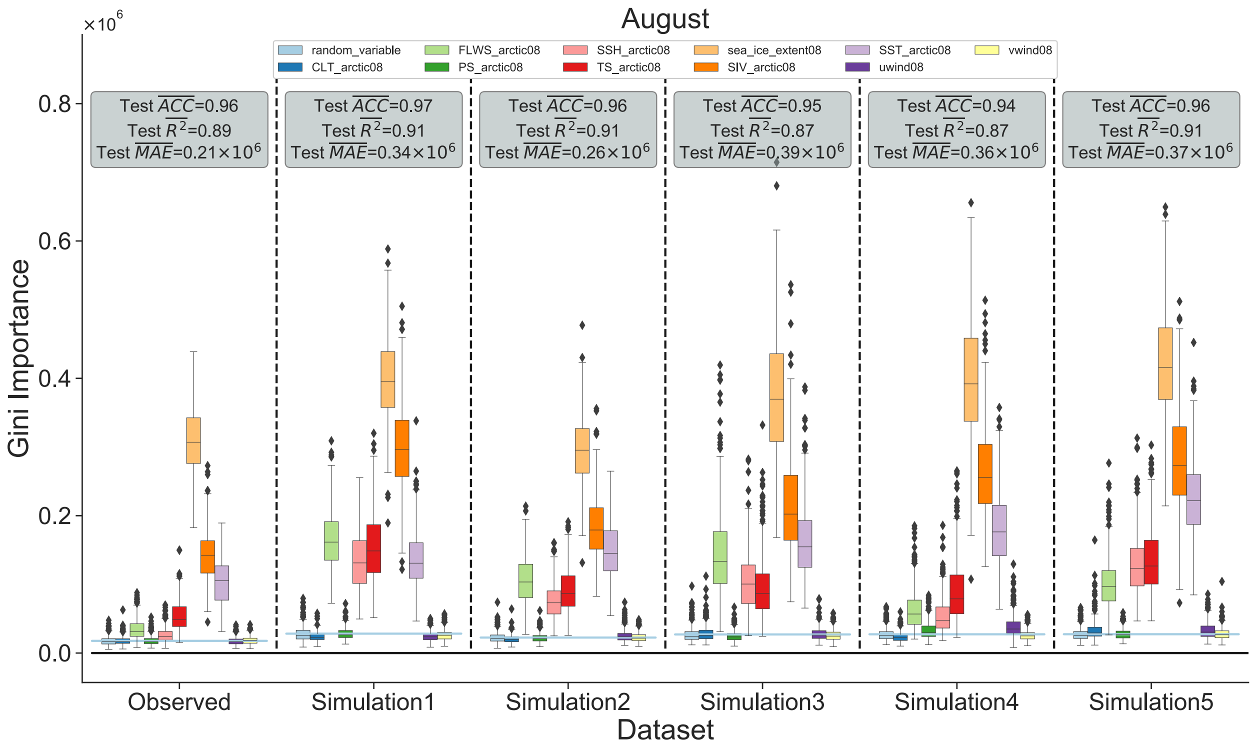


Figure 3.4: August feature importance. Standard box-and-whisker plot (McGill et al., 1978) of values for 13 predictions generated by 385 models. The average R^2, anomaly correlation coefficient (ACC), and mean absolute error (MAE) are displayed in the gray boxes. The blue line in each dataset is the mean importance of a random variable in each feature set.

ice extent was always the most important. The importance values of the remaining features generally changed very little throughout datasets.

3.7 Discussion

We found that our RFR ML models were able to accurately learn each of the datasets. After examining the Gini importances computed within each model, we discovered some key differences in how each dataset relates to September pan-Arctic sea ice extent.

A problem with our dataset is that the satellite record only goes back to 1979.

One solution is to adapt the models to forecast sea ice extent continuously throughout each year. This is in line with Reid and Tarantino's approach (Reid and Tarantino, 2014) (see Section 3.4), but with random forests instead of support vector machines and including many features instead of only sea ice extent. The models would train on the full year of data and see 432 data points rather than 36 in the time series. Several observed features are measured more frequently than monthly, some every few hours of every day, so a means to incorporate inconsistent sampling resolutions of features should be investigated to leverage all of the data available. Another solution could be to use a surrogate model to generate more data that is similar to the first 15 years of observed data, which have a much flatter trend. The surrogate model would let the new data agree with what the model learns about input features.

The combined error metrics and general consistency of results between each dataset suggests that our models have learned the data well, and the feature analysis can identify key patterns. It is meaningful that the same six features are considered important across datasets and input-months. Since our analysis is of the pan-Arctic region, it is possible that the set of unimportant features would be more important in specific subregions of the Arctic.

Though the most important feature in June and August is consistent between simulation and observation, the absolute importance differs markedly. One clear pattern is that June shows an acute reliance on sea ice volume for both observations and simulations. By August the reliance is traded for sea ice extent. This finding is consistent with earlier studies evaluating sea ice predictability using lag-correlation analyses with ESM ensemble data (Ordonez et al., 2018; Blanchard-Wrigglesworth et al., 2011).

Although the observed and simulated data share patterns, there is a clear difference between them. In July, simulations and observed data do not agree on the most important feature. In June, July, and August, simulated data relies too heavily on almost all the important features. In each dataset, importance values diminish for the remaining features in June and July, and their distributions overlap more than they did in June, but the observed dataset still shows the least importance in FLWS, SSH, TS, and SST.

Interestingly, simulations 1 and 2 forecasted with the highest $\overline{R^2}$ each input month, and simulations 3 and 4 had the lowest $\overline{R^2}$ in each input month. Simulations 1 and 2 have the lowest $\overline{MAE}$ and highest $\overline{ACC}$ among the simulation runs, and 3 and 4 have the highest $\overline{MAE}$ and lowest $\overline{ACC}$ among the simulations runs. Although the differences are small, these consistencies may indicate some commonality between these simulation runs.

Our ML models performed better on the observed data than on the simulations as measured by $\overline{MAE}$ and $\overline{ACC}$, but is not reflected in $\overline{R^2}$. That suggests that the mean value, or the trend after retrending, was very predictable, but its intervariability, which R^2 explains, was less predictable. The likely explanation is in the difference in the complexity of the systems. Observed features of the continuous Earth system are artificially discretized. In any complex system, intervariability is difficult to forecast. However, because we chose largely relevant features as predictors, we could capture the macro-level patterns, as evidenced by the macro-level error measures: $\overline{MAE}$ and $\overline{ACC}$.

3.8 Conclusions

We demonstrated that random forest regression and the associated Gini importance measure can provide insight into why ESMs incorrectly estimate sea ice extent in recent decades. We found a discrepancy in the feature importance between observed and simulation datasets. In particular, the discrepancy between E3SM and observation appear to be due to an over-reliance on June sea ice extent and August sea ice volume. The order of feature importance was also different between E3SM and observation, and the ordering was not consistent within E3SM ensemble members. In all cases, E3SM over-relies on six features compared to observed data. Machine learning allows us to fill the gaps in the underlying physics of ESMs, providing a metric for Stroeve et al.'s (Stroeve et al., 2012) hypothesis that ESMs are missing complex relations and causal mechanisms.

In the future, we can evaluate more features that can be measured or constructed in each dataset. An analysis, including all months of the year in each model will be elucidating as well. Sea ice extent is measured daily via satellite imagery. We can understand how each dataset explains sea ice extent at a higher resolution every month of the year.

We can repeat our analysis on other regions, including Antarctica, where there are also problematic disagreements with observations (Rosenblum and Eisenman, 2017). An analysis like this of other climate models could be insightful too. It would be particularly interesting to compare simulations in which there few to no correlated features. That would allow for variations on the analysis, such as more modeling approaches, which require linearly independent features, and more feature analysis methods, such as drop-column importance, which would otherwise struggle with multicollinearities.

Further insight could be gained by repeating our analysis with a machine learning method other than RFR, however the following methods have their own challenges. Most neural network models would need more observed data than is available to converge. We found that multiple linear regression cannot learn the data well because the relationships between features are nonlinear. Reid and Tarantino (Reid and Tarantino, 2014) found that SVR can forecast the data well, but it is unclear what the best feature analysis method would be.

Given the discoveries in this paper, we can run experiments with E3SM to determine how reducing feature disagreements between the observed and simulation datasets impact E3SM's forecasts. That process may not yield results for several reasons, including that E3SM's real feature set is large and complex, focusing analysis on the Arctic region is too restricting to estimate the effects of the global Earth model, or our ML models are too limited by small datasets. Despite these challenges, our results can potentially guide climate modelers as they develop the next generation of ESMs.

Acknowledgments

This work is supported by Sandia Earth Science Investment Area Laboratory Directed Research and Development funding. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525.

This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

4 Causal Discovery for Climate Model Evaluation

4.1 Publication Notes

This chapter summarizes material from two complementary publications:

Citations:

Nichol, J. Jake, et al. “Learning Why: Data-Driven Causal Evaluations of Climate Models.” ICML 2021 Workshop Tackling Climate Change with Machine Learning, 2021

Nichol, J. Jake, et al. “Causal Evaluations for Identifying Differences between Observations and Earth System Models.” Technical Report, OSTI, 2021.

Publication date: 2021

Conference: International Conference on Machine Learning 2021 Workshop Tackling Climate Change with Machine Learning

Technical Report, U.S. Department of Energy Office of Scientific and Technical Information (OSTI)

Formatting: The original published texts have been combined and adapted while adhering to the formatting requirements of this dissertation. The chapter draws from both the ICML workshop paper (framework) and the OSTI technical report (implementation and results).

Data and Software Availability: The workshop paper is available at https://www.climatechange.ai/papers/icml2021/53. The technical report is available at https://www.osti.gov/servlets/purl/1888471.

Funding: This work is supported by Sandia Earth Science Investment Area Laboratory Directed Research and Development funding. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525.

4.2 Introduction

Chapter 3 is a valuable start to using data-driven approaches to learn more about the structure underlying observational data. However, machine learning algorithms and the resulting feature importance values are correlational and cannot be readily interpreted causally. Furthermore, they do not enable us to evaluate the system holistically to understand the relationships between all features. In this chapter, I evaluate the same problem with causal discovery. Causal discovery is an algorithmic causal inference framework that can produce causal graphs depicting the estimated causal relationships between given features. While the random forest regression (RFR) (Breiman, 2001) feature importance values described *what* features are important, causal discovery can help answer *why* relationships exist between features.

Nichol et al. (2021b) and Nichol et al. (2021c) are complementary works where I explored applying a state-of-the-art causal discovery algorithm for the Arctic sea ice system. Nichol et al. (2021b) discusses the research framework and methodology, and Nichol et al. (2021c) discusses the implementation and results. This work extends the RFR feature importance approach in Chapter 3 to a causal discovery framework.

The PCMCI (Runge et al., 2019a) time series causal discovery algorithm was applied to Arctic climate features that may explain sea ice extent. PCMCI produces causal directed acyclic graphs (DAGs) that, if its assumptions are satisfied, represent the estimated causal relationships between given features. Comparing causal graphs estimated from different datasets and data sources enables a more mechanistic comparison. It can answer whether two data sources are *structurally* similar.

4.3 Methods

For this work, I evaluated Arctic features including surface pressure, precipitation, solar radiation, cloud cover, sea surface temperature, sea ice volume, sea ice extent, wind speeds, surface air temperature, and sea surface humidity. I applied PCMCI to each dataset, observed and five Energy Exascale Earth System Model (E3SM) (E3SM Project, 2018) datasets, and compared the resulting causal graphs. I used the F_1 score, the harmonic mean of precision and recall, as a similarity metric for comparing estimated causal graphs.

4.4 Results

All data sources (observed and E3SM simulated) had similar graphs. However, the E3SM graphs were more dense, implying that more features were interconnected. This seems to corroborate our RFR feature importance finding that E3SM feature importances were over-weighting features, but more rigorous analysis is needed to confirm that connection more generally.

4.5 Discussion

While the F_1 score is a good starting point for graph similarity, it cannot pinpoint where the graph differences are. I proposed further work to develop more node-level comparison metrics to understand structural similarities and differences better. I additionally recommended more subregional analyses (both the RFR and causal analyses evaluated quantities spanning the entire Arctic) because more meaningful insights may be gleaned from relating smaller regions within the Arctic. This limitation of regional aggregation obscures the spatial dynamics that define many Earth systems and highlights the need for methods that could preserve spatial information while enabling causal discovery. Developing a better understanding of the smaller-scale processes that accumulate to produce emergent phenomena in the Earth system was the impetus for the work in Part II.

Part II

Local Causal Discovery in High-Dimensional Gridded Data

5 Benchmarking the PCMCI Causal Discovery Algorithm for Spatiotemporal Systems

5.1 Publication Notes

Citation: Nichol, J. Jake, et al. "Benchmarking the PCMCI Causal Discovery Algorithm for Spatiotemporal Systems." [Report No. 1991387]. 2023. U.S. Department of Energy, Office of Scientific and Technical Information.

Publication date: 2023

Publisher: U.S. Department of Energy, Office of Scientific and Technical Information.

Formatting: The original published text has been summarized and adapted to adhere to the formatting requirements of this dissertation.

Data and Software Availability: The paper is available at https://www.osti.gov/biblio/1991387.

Funding: This work is supported by Sandia Earth Science Investment Area Laboratory Directed Research and Development funding. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security

Administration under contract DE-NA-0003525.

5.2 Introduction

Grid-level causal discovery is an enticing research avenue because it has the potential to reveal local-level dynamics that give rise to global complexity. However, it has been challenging to achieve. Ebert-Uphoff and Deng (2012) applied grid-level causal discovery in a climate setting. However, they required significant grid coarsening to obtain more time samples per grid cell than grid cells. They cited computational complexity as the primary constraint on grid resolution. Nowack et al. (2020) applied PCMCI to dimension-reduced climate data but stated that a grid-level approach would be interesting. However, the high-dimensionality and redundancy of neighboring grid cells present challenges for the statistical power of estimation methods. Finally, Tibau et al. (2022) repeated the above concerns and presented Mapped-PCMCI, which maps grid cells between dimension-reduced regions that PCMCI connects. While their approach does create a graph of grid cells, it does not provide a means for understanding inter-grid cell relationships. To begin exploring this domain, I needed to start by developing a controllable, scalable benchmark for evaluating grid-level causal discovery.

5.3 Methods

To systematically investigate the limitations of existing causal discovery methods for grid-level data, I developed grid-level space-time benchmarks for causal discovery methods and evaluated the PCMCI (Runge et al., 2019a) time series causal discovery algorithm (Nichol et al., 2023). PCMCI was developed for highly autocorrelated time series data. It has been applied extensively in the Earth sciences (Runge et al., 2019c). However, its application methodology has been limited to regional analyses in which Earth science time series are obtained from dimensionality reduction methods such as weighted averages, principal component analysis (PCA), and related methods (Runge et al., 2015b; Tibau et al., 2022).

Our grid-level benchmark began with a 1D spatial grid and was extended to a 2D grid, for which each grid cell contained a time series with defined dependencies on its immediate neighbors. Both were structured as vector autoregression models (VARs), which enables a mathematically defined model that generates data and maps directly to a ground-truth causal graph.

5.4 Results and Discussion

Using graph similarity metrics to compare PCMCI's estimated causal graph with each dataset's underlying VAR, I found PCMCI struggled to estimate the graphs reasonably, except when it had unrealistically high amounts of time samples per grid cell. In particular, for a given number of time samples per grid cell, performance declined as grid sizes got larger. In short, I determined that significant algorithmic advances would be needed to apply causal discovery like PCMCI at the grid-level.

The work presented computational advances as well. While using VARs for systems modeling and causal discovery benchmarking is not new (Runge et al., 2019d), my innovation was using them to model stable space-time dynamical systems with locally dependent grid cells. I produced gridded space-time data using a sliding dot product with a local neighborhood dependence matrix (NDM). In that way, they are similar to how cellular automata are defined, in which a single grid-level rule determines complex global behavior.

These benchmark results demonstrated that conventional causal discovery algorithms cannot scale to grid-level applications without fundamentally modifying their approach. This motivated the development of Causal Space-Time Stencil Learning (CaStLe), a meta-algorithm specifically designed to overcome these limitations through novel use of spatial replicates and local structure learning, which will be presented in Chapter 6.

6 Space-Time Causal Discovery in Earth System Science: A Local Stencil Learning Approach

6.1 Publication Notes

Citation: Nichol, J. Jake, et al. "Space-Time Causal Discovery in Earth System Science: A Local Stencil Learning Approach." Journal of Geophysics Research: Machine Learning and Computation, under review.

Publication date: N/A

Conference: N/A

Formatting: The original text has been preserved as much as possible while still adhering to the formatting requirements of this dissertation.

Data and Software Availability: The paper is currently under review and not yet publicly available.

Funding: This work is supported by Sandia Earth Science Investment Area Laboratory Directed Research and Development funding. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell

International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525.

6.2 Abstract

Causal discovery tools enable scientists to infer meaningful relationships from observational data, spurring advances in fields as diverse as biology, economics, and climate science. Despite these successes, the application of causal discovery to space-time systems remains immensely challenging due to the high-dimensional nature of the data. For example, in climate sciences, modern observational temperature records over the past few decades regularly measure thousands of locations around the globe. To address these challenges, we introduce **Ca**usal **S**pace-Time S**t**encil **Le**arning (Causal Space-Time Stencil Learning (CaStLe)), a novel meta-algorithm for discovering causal structures in complex space-time systems. CaStLe leverages regularities in local space-time dependencies to learn governing global dynamics. This local perspective eliminates spurious confounding and drastically reduces sample complexity, making space-time causal discovery practical and effective. For causal discovery, CaStLe flexibly accepts any appropriately adapted time series causal discovery algorithm to recover local causal structures. These advances enable causal discovery of geophysical phenomena that were previously unapproachable, including non-periodic, transient phenomena such as volcanic eruption plumes. Regularities in local space-time dependencies are transformed into *informative spatial replicates*, which actually improves CaStLe's per-

formance when applied to ever-larger spatial grids. We successfully apply CaStLe to discover the atmospheric dynamics governing the climate response to the 1991 Mount Pinatubo volcanic eruption. We provide validation experiments to demonstrate the effectiveness of CaStLe over existing causal-discovery frameworks on a range of geophysics-inspired benchmarks while identifying the method's limitations and domains where its assumptions may not hold.

Plain Language Summary

We introduce a new method for learning the dynamics of causal systems, that is, the physical rules that define a system's behavior. While this task, *causal discovery*, is not new, existing tools are ill-suited for many large geophysics datasets. Current state-of-the-art approaches use statistical techniques to search for causal relationships between all aspects of a system, examining billions of possible causal effects, or simplifying the data by focusing on the most important variables. Instead of an exhaustive search or oversimplifying the data, we incorporate basic physical principles—requiring effects to be "local" and "uniform"—to massively simplify the causal discovery problem. We demonstrate that our approach can recover known geophysical dynamics by applying it to the 1991 Mt. Pinatubo eruption, validating its ability to uncover space-time causal structure from observational data.

6.3 Introduction

Explaining the causal dynamics that govern geophysical phenomena is paramount in the Earth sciences. Climate models, for example, critically depend on understanding both local and global causal pathways to model the complex Earth system. Understanding short- and long-term consequences of the Earth system's behavior is essential for future model development, our scientific knowledge, and preparing for the future. More specifically, in atmospheric science, we know the initial state of specific wind modes, such as the quasi-biennial oscillation or the Brewer-Dobson circulation, dramatically affects the later evolution and impact of volcanic eruptions, major wildfires, or geoengineering efforts such as stratospheric aerosol injection (Hitchman et al., 1994; Jones et al., 1998; Aquila et al., 2014; Gray et al., 2018).

Traditional statistical methodologies, while providing valuable insights, often fall short of capturing the complex causal relationships inherent in geophysical systems. Causal models are hard-won and often represent the culmination of many decades of research. Causal discovery tools aim to accelerate the discovery of these relationships using statistically-rigorous techniques to separate predictable, but indirect, statistical relationships from direct causal connections. Causal discovery has been successful across the sciences, providing new understandings of climate, biological, genetic, neural, and other dynamical systems (Ebert-Uphoff and Deng, 2012; Sugihara et al., 2012; Neto et al., 2010; Zhang et al., 2011b; Kamiński et al., 2001; Tsonis et al., 2017).However, applying existing causal methods to space and time structured data remains limited due to the complexity and scale of such systems.

This work presents a novel causal discovery methodology that overcomes these challenges to recover networks describing local causal structures from gridded data. A fundamental insight driving the present work is that in many complex systems, global phenomena—whether climate teleconnections, brain functional networks, or ecosystem dynamics—emerge from countless repeated and structured local interactions. We can better understand how complex global patterns arise by accurately capturing these foundational local structures.

Today's Earth science measurement and modeling capabilities provide a wealth of data for studying our planet's complex dynamics. However, due to the immense complexity of these dynamics, simple analyses provide only a limited understanding of the data. Causal discovery tools offer the ability to understand finer mechanistic details via causal graphs' simplicity, interpretability, and flexibility. causal discovery is a field that utilizes algorithmic causal inference to identify causal models as dependencies between fields of interest, which are often represented as a directed acyclic graph. Causal graphs let us analyze the space-time evolution of fields of interest and causal discovery can estimate them without requiring hypothesized physical models. Insights gleaned from causal discovery can further inform physical models, validate simulations against observational data, and identify future research questions.

While causal discovery show considerable promise for addressing problems in the Earth sciences, the enormous size and scope of Earth science data have limited its applications. For example, atmospheric data often contains hundreds of thousands of grid cells, each with several orders of magnitude fewer observations in time. That imbalance is one aspect of the *curse of dimensionality* (Bellman, 1957; Bühlmann and Geer, 2011), where high dimensionality relative to sample size challenges conventional statistical methods and renders many forms of inference, including causal discovery, unreliable without dimensionality reduction. Despite these obstacles, causal discovery has been successfully applied in Earth science (Deng and Ebert-Uphoff, 2014; Runge et al., 2015b; Capua et al., 2019, 2020; Nowack et al., 2020; Krich et al., 2020; Galytska et al., 2022; Tibau et al., 2022; O'Kane et al., 2024; Zhao et al., 2024), primarily via dimensionality reduction techniques to reduce the number of relationships to estimate. Those contributions identified teleconnection pathways to recover large, periodic climate modes and their effects. While a dimensionality reduction approaches can be practical, the analysis of local effects has been considered challenging and generally avoided due to the curse of dimensionality (Ebert-Uphoff and Deng, 2012; Runge et al., 2015b; Nowack et al., 2020).

In contrast to dimensionality reduction methods that marginalize large amounts of information, our work leverages the known locality in space-time systems to harness *informative spatial replicates*, i.e., repeating space-time relationships, without loss of local structural information, to identify local causal graphs. These advances enables us to approach problem classes in space-time systems that are typically intractable with prior art—both in terms of performance and algorithmic efficiency. We highlight two features of Causal Space-Time Stencil Learning (CaStLe) that are useful contributions to causal discovery for geoscience problems: the ability to learn grid-level relationships instead of regional relationships from reduced dimensional data (e.g. principal components or modes) and the ability to handle dynamic, advective processes.

Prior causal discovery work in Earth science has primarily focused on large-scale regional phenomena, such as the El Niño Southern Oscillation. These patterns——generally consistent in their spatial distribution and periodic in nature——are well suited to global dimensionality reduction techniques, which project fields onto a small number of modes. While global teleconnections are crucial research areas, they ultimately emerge from local causal interactions. However, dimensionality reduction sacrifices critical local information, making it impossible to see how local structures give rise to global patterns. CaStLe reduces problem complexity in a fundamentally different way: By identifying and leveraging the repeating local structures, it preserves the relationships at the grid level while remaining applicable to spacetime systems that exhibit multiscale organization.

Typical dimensionality reduction approaches to causal discovery decrease the data space from many grid cells to a few regional modes and uses many observations, resulting in a *little p, large n* problem, where p is the number of variables and n is the number of data points. In contrast, phenomena that evolve dynamically in space or occur rarely, like volcanic plumes, are harder to analyze and often have few data points. Such problems are *large p, little n*. CaStLe makes causal discovery of the space-time evolution of these phenomena tractable for the first time by leveraging the gridded sample space, avoiding the marginalization that reduces many grid cells into a single time series per regional mode, and recovering interpretable space-time causal structures.

This work's primary case study is the 1991 Mount Pinatubo eruption. It injected a plume of aerosols into the stratosphere, which then advected around the tropical zone before dispersing northward and eventually diffusing around the globe. This example demonstrates the characteristics of the unique, transient problem class, has an established research history, and exhibits dynamics verifiable with a known causal driver: stratospheric wind.

We introduce a new Earth system causal network, the *causal stencil graph*, which describes local space-time causal structures between adjacent locations, and a new estimation methodology, **Ca**usal **S**pace-Time **St**encil **Le**arning (CaStLe), that is capable of describing local mechanistic pathways in space and time between grid cells. Grid-level causal discovery in high dimensional space-time data has been previously considered intractable due to the curse of dimensionality (Nowack et al., 2020; Tibau et al., 2022). Though demonstrated with climate model output, our methodology applies to any space-time system where local physical interactions drive global behavior, including fluid dynamics, biological pattern formation, or material transport processes.

CaStLe combines modern causal discovery with classical physics-based principles, namely spatial and temporal locality, to accurately perform causal discovery on large spatial domains. Our novel local-coordinate-space projection does not marginalize any data points, such that local causal information is lost, which is a common sacrifice of other space-time dimension reduction techniques such as weighted averaging or principal component analysis (PCA). This preservation of local information is crucial because global-scale phenomena in complex systems emerge from interactions at smaller scales. By mapping these foundational causal pathways, CaStLe provides insights not just into immediate local effects but also into how these effects propagate and combine to create larger-scale patterns.

With these advances, CaStLe achieves remarkable improvements over state-of-the-art space-time causal discovery approaches. CaStLe is a flexible framework that can be implemented by adapting any given time series causal discovery algorithm to the stencil approach. Our approach performs excellently in high-dimensional data regimes, making it capable of describing the local space-time evolution of transient phenomena transporting over many grid cells.

The Earth system is rich with transient phenomena examples including forest fires, monsoons, coastal erosion, salt or freshwater incursions, inter-tropical convergence zone shifts, and atmospheric rivers. Aside from elucidating underlying dynamics, CaStLe can be used to identify and characterize causal change points, such as polar vortex disruption and ocean current disruptions. Additionally, understanding these local dynamic structures can give further insights into the construction and evolution of important macro phenomena such as the El Niño Southern Oscillation, the Quasi-Biennial Oscillation, and the Madden-Julian Oscillation. Table A in the Appendix summarizes the capabilities of CaStLe and their relevance to specific Earth science applications. These capabilities address analytical needs that have been challenging or infeasible with previous causal discovery approaches.

The remainder of this paper is organized as follows: Section 6.4 provides a brief background on causal discovery and its use in Earth science; Section 6.5 describes our case studies in the HSW-V and E3SMv2-SPA models and available data; Section 6.6 explains our novel CaStLe methodology; Section 6.7 demonstrates CaStLe's ability to recover known volcanic aerosol evolution in climate models of different resolution; and finally, Section 6.8 illustrates CaStLe's computational, and performance improvements over the state-of-the-art methods with synthetic data experiments.

Contributions

We introduce the CaStLe approach to causal discovery from space-time data. CaStLe allows the discovery of causal structures in high-dimensional spatial data, avoiding the need for dimension reduction techniques that dominate causal discovery of space-time data, e.g., the work by Nowack et al. (2020). By working in the raw data space, CaStLe's causal graphs are *inherently interpretable* and do not require mapping structures from the dimension-reduced space back onto the original data. We provide a theoretical analysis of CaStLe, showing that it has attractive computational and statistical properties and, rather remarkably, that CaStLe's accuracy actually increases on larger spatial domains. We apply CaStLe to two simulations of a major volcanic eruption and demonstrate how it can be used to better understand how stratospheric winds mediate the climate response to volcanic activity. Our first study is of a relatively simplified model to validate the methodology with proxy ground-truth. In our second study, we consider a more realistic model and find that CaStLe still provides consistent and valuable results, demonstrating its value for realistic atmospheric dynamics. Finally, extensive numerical experiments measure the advantages of CaStLe and demonstrate: i) significantly improved performance over existing causal discovery methods on a set of vector autoregressive (VAR) benchmarks; and ii) the use of CaStLe to identify the governing dynamics of Burgers' non-linear partial differential equation (PDE). While our case studies utilize climate model data, the methodology is domain-agnostic and can be applied to any high-dimensional space-time system meeting our locality and stationarity assumptions.

6.4 Background: Causal Discovery and Formal Mathematical Scope

Here, we provide a brief overview of the causal discovery field and the mathematical scope of our contributions. For a broader overview of causal discovery and its applications to Earth science, see the reviews by Glymour et al. (2019), Runge et al. (2019b), and Runge et al. (2023), and the book by Peters et al. (2017). Additionally, we outline the mathematical constraints and assumptions that define where our methodology can be applied in the class of space-time systems.

Causal discovery is a field of causal inference that seeks to recover causal dynamics from observational data. In the parlance of causal inference, *observational data* is data that is passively observed rather than data to which treatments (e.g. manipulations) have been applied. Observational data can be natural (e.g. physical observations) or synthetic (e.g. simulations). The present work exclusively pertains to untreated data, so we will use *observational* in this way.

While correlation does not imply causation, causal discovery is built upon Reichenbach's common cause principle (Reichenbach, 1956): if two quantities are correlated then one must cause the other or there is a third causal driver of the two. causal discovery generally has two output classes: a causal graph/network (Pearl, 1995a) or a structural causal model (Pearl, 1998). We focus on causal graphs, which are networks of variables (nodes) connected by edges that denote a causal dependence. Causal graphs can be more appealing than structural equation models because they are human-interpretable and do not require prior knowledge of the underlying causal function. In the study of Earth science, causal graphs may often be preferred to visually describe space-time relationships on the globe. Our contribution produces a novel type of causal graph, the causal space-time stencil, which is detailed in Section 6.6 and an example of which is in panel 4 of Figure 6.2.

6.4.1 Related Work: Causal Structure Learning

In recent decades, causal inference has been developed into a rigorous mathematical framework (Rubin, 1974; Pearl, 2000; Pearl et al., 2016). These developments made algorithmic discovery of causal structures from observational data possible (Spirtes et al., 1993; Peters et al., 2017; Glymour et al., 2019). Causal structures can be modeled with two common forms: structural causal models (SCMs) and causal graphs. Both describe a functional relationship between a variable X_j and its causal parents, denoted $\mathscr{P}(j)$.

For example, if X_i causes X_j, then it is said X_i is a parent of X_j and $i \in \mathscr{P}(j)$. Formally, Peters et al. (2017, p.83) defines an SCM as follows:

> *A **structural causal model (SCM)** consists of a collection of d **(structural) assignments***
>
> $$X_j := f_j(\boldsymbol{X}_{\mathscr{P}(j)}, \eta_j), \qquad j = 1, \ldots, d,$$
>
> *where $\mathscr{P}(j) \subseteq \{1, \ldots, d\} \backslash \{j\}$ are called **parents of** X_j: and a joint distribution $\mathbf{P}_\eta = P_{\eta_1, \ldots, \eta_d}$ over the noise variables, which we require to be jointly independent; that is $\mathbf{P}_\eta$ is a product distribution* [in our notation].

An SCM admits a unique causal graph, where $X_j \to X_i$ if $j \in \mathscr{P}(i)$ and $j \not\to X_i$ if $j \notin \mathscr{P}(i)$. While discovery of an SCM requires hypothesizing all f_j's, discovering a causal graph can be done without knowing the exact functions. Because a causal graph does not imply a specific function between variables, each may imply multiple SCMs. This does limit some of the inferential power of causal graphs, in exchange for more versatility.

Algorithms for discovering causal graphs have two primary classes: constraint-based and score-based algorithms. Constraint-based methods use statistical tests to compute conditional independence relationships between sets of variables. Once a set of independence relationships is established, it utilizes causal assumptions and reasoning to connect the variables with directed links. Score-based approaches are similar but use score optimization to determine causal dependence between variables. Both constraint-based and score-based algorithms produce causal graphs because they operate on graphical structures and independence relations rather than the explicit parametric relationships between variables required to specify a complete SCM.

Early causal discovery algorithms developed as two parallel traditions. The temporal Granger causality (Granger, 1969) methodology was an early innovation using time series data to determine if the past history of X aids the prediction of Y better than Y's history alone. If so, then X *Granger causes* Y. Independently, the constraint-based PC algorithm (named for its authors Peter and Clark) (Glymour and Scheines, 1986) and FCI (Spirtes and Glymour, 1991) developed out of the inductive causation (Pearl and Verma, 1992) framework and the earlier SGS algorithm (Spirtes and Glymour, 1991), significantly improving the efficiency of causal discovery using statistical structures in observed data. In time, other structural algorithms developed, such as LiNGAM (Shimizu et al., 2006), utilizing asymmetries in non-linear and non-Gaussian data for inferences, and NOTEARS (Zheng et al., 2018), a graph score-optimization-based method. Eventually, these two traditions converged as structural methods were developed to take advantage of temporally ordered data. Key advances included: hMRF (Liu et al., 2010), which uses hidden Markov models for estimation and is grounded in Granger causal structures, PCMCI (Runge et al., 2019a) (and related PCMCI+ and LPCMCI), which improves PC to handle autocorrelated dependencies better, and DYNOTEARS (Pamfil et al., 2020), which extends the NOTEARS method to time series. More recently, a third tradition, causal representation learning, developed out of machine learning (ML) to leverage causal reasoning in ML models (Schölkopf et al., 2021). While still a developing field, it shows particular promise for estimating relationships in the presence of latent confounding.

The directed nature of time provides a powerful asymmetry to leverage, often sufficient to overcome the difficulties of autocorrelation, automatically orienting discovered relationships in time. In contrast, spatial data lacks an obvious uniform directional structure and poses challenges for causal discovery. As discussed in Section 6.3, while some approaches have incorporated domain-specific spatial constraints for point-measurement networks, none have developed a generalizable framework that leverages fundamental physical principles of locality to enable scalable causal discovery in high-dimensional gridded space-time systems.

Causal Discovery in Earth Science

We present a brief review of causal discovery for Earth science to position CaStLe within the literature. Please also see the extensive reviews by Runge et al. (2023) and Ali et al. (2024).

Ebert-Uphoff and Deng (2012) were the first to apply a causal discovery algorithm, PC-stable (Colombo and Maathuis, 2014), to the climate science domain. They were able to find a grid-cell-level causal teleconnection network in 50 year daily geopotential height data using the PC algorithm. Ebert-Uphoff and Deng (2014); Deng and Ebert-Uphoff (2014) further explored application requirements and climatological interpretations of the geopotential height analysis. In each paper, they note grid challenges related to the high expense of many grid cells, aggregation effects, and cell spacing. The first paper limits the number of grid cells to 800, while the subsequent analyses limited grid cells to 200 to minimize computational costs. While their results are compelling, they use extensive decadal data and recover patterns common to all 50 years. The fundamental difference between our work and Ebert-Uphoff and Deng's work is that they recover causal graphs from recurring atmospheric phenomena with sufficiently large datasets on relatively coarse-grained grids, whereas CaStLe is recovers networks of isolated phenomena with many more grid cells and many fewer time samples per cell.

Runge et al. (2015b) introduced an alternative approach to causal discovery of space-time Earth science data. They reduced the dimensionality with varimax-rotated principal component analysis prior to applying the causal discovery algorithm, producing a graph relating discrete, potentially remote, regions. Their causal graph is most similar to a teleconnection network between large areas on the globe. Nowack et al. (2020) utilized that framework to evaluate CMIP5 models. Particularly of note, they point out the challenges and strengths of Ebert-Uphoff and Deng (2012)'s grid-cell-level approach, "... while an analysis at the grid-cell-level is more granular which, however, carries the challenges of higher dimensionality, will have a strong redundancy among neighbouring grid cells, and grid-level metrics will require handling varying spatial resolution among data sets."

Tibau et al. (2022) built on the dimensionality reduction approach, augmenting it to output grid-cell-level networks. They specifically delineate *mode-level* (dimensionality reduction or cell aggregation) and grid-level causal discovery. Their augmentation is called Mapped-PCMCI, which first applies dimensionality reduction, then computes a mode-level causal network with PCMCI, and finally maps the grid cells within the modes to each other using the network previously constructed. Their resulting network is one consisting of edges between grid cells, but the method assumes that cells within modes are fully connected, i.e., each cell is dependent on all of its neighbors. In contrast, our work specifically seeks inter-cell spatial relationships. Finally, they also describe the failure of a traditional causal discovery approach for grid-cell-level data, "[if] we apply PCMCI directly at the grid-level, the low power of this high-dimensional and redundant estimation problem (see Section 2.2.2) leads to most links being missing."

Boussard et al. (2023) and Brouillard et al. (2024) developed the Causal Discovery with Single-parent Decoding (CDSD) algorithm within the causal representation learning framework and applied it to the climate science field. Like CaStLe, CDSD performs well in high-dimensional data settings but through a different mechanism. It performs dimensionality reduction by learning latent variables and enforcing a "single-parent" constraint where each grid cell belongs to exactly one latent factor. This naturally clusters grid cells into coherent, often contiguous regions and enables the discovery of causal relationships between these larger-scale patterns. In contrast to CaStLe's grid-level structure learning, CDSD identifies broader teleconnection pathways between regional climate modes. Thus, while CaStLe preserves the original grid structure to capture fine-grained causal dynamics, CDSD abstracts to a higher level by mapping the native grid space to an identifiable latent representation before performing causal discovery.

Several studies have addressed local-scale phenomena. Pfleiderer et al. (2020) applied causal discovery to identify precursors to seasonal hurricane frequency. They utilized the precursors to inform a predictive model. Polkova et al. (2021) identified local drivers of marine cold-air outbreaks in the Barents Sea. These demonstrate that existing causal discovery approaches can be valuable for seasonal and sub-seasonal phenomena. However, both marginalized large regions prior to analysis, reducing the space's dimensionality, and did not evaluate the space-time evolution of phenomena nor grid-level dynamics.

There are some examples of causal discovery algorithms leveraging spatial information. Zhu et al. (2016) developed pg-Causality that applies space-time pattern mining and a Gaussian Bayesian Network to seek local dependencies in the space-time propagation of air quality data. Sheth et al. (2022) developed STCD for understanding hydrological systems. They constrained the discovery of spatial structures by only allowing higher elevation nodes to be parents of lower elevation nodes because water follows the gravity gradient. While both cleverly use mined or known spatial structure to inform their causal discovery, they are both limited to use in sparse point-measured data from static base stations rather than gridded data. Further, these methods enforce constraints as filtering mechanisms, whereas CaStLe actively leverages spatial structure to enhance statistical power. Neither address the scalability challenges in high-dimensional gridded data.

Parallel Approaches in Neuroscience: Causal Discovery for High-Dimensional Spatial-Temporal Data

Other scientific domains face similar challenges with high-dimensional space-time data. Neuroscience, for example, needs to study mechanisms in brain interactions, and fMRI images may contain thousands to millions of pixels. The anatomy of the brain also exhibits locality constraints. Ramsey (2014) made computational optimizations to the Greedy Equivalence Search algorithm, including sparsity constraints and limiting the distance of potential parents, to recover graphs with millions of nodes. Saetia et al. (2021) marginalized regions of interest in the brain using spatial averaging and then applied the PCMCI algorithm to construct causal graphs. There is a common interest in recovering graphs of high-dimensional grid-level data throughout the sciences. Developing more tools that enhance the estimation and interpretability of causal graphs in these spaces will help advance our understanding of space-time structures across the sciences.

What is clear from prior work is that grid-level analyses are challenging, both statistically and computationally, due to how many grid cell dependencies need to be estimated, the enormous number of observations needed, and the redundant information content of nearby cells. As we present in the following sections, CaStLe adds to the literature as it overcomes the statistical and computational limitations of grid-level analysis by leveraging the known physical structure of spatial information to produce interpretable graphs describing local causal structures.

6.4.2 PDE-Like Systems

We seek to perform causal discovery from space-time data governed by consistent physical laws. As detailed in Section 6.6, CaStLe operates via two phases. The first restructures the given space-time data into a lower-dimensional local neighborhood space without marginalization or loss of any data points; the second is the causal discovery step. This section details the assumptions required for efficient use of spatial replicates that enable CaStLe's first phase, scalability properties, performance in high-dimensional settings, and interpretability. We note that the assumptions necessary for the second phase will be inherited from our meta-algorithm's chosen causal discovery method. In general, they will be the causal Markov condition, faithfulness, and often causal sufficiency, which we define formally in Appendix A.2.

We take PDE-like models as our starting point, and assume that all behavior in the given space are driven by a fixed set of dynamics that apply at infinitesimal time and spatial scales. Specifically, we assume that, for data observed in discrete space and time, the evolution of a single grid cell is controlled only by the values of its immediate spatial neighbors at the previous time step. Using causal discovery, we seek to determine which neighbors have a causal impact on a given grid cell and the direction of that relationship. Our analytical framework has similarities to the sparse identification framework initially developed by Brunton et al. (2016), though our approach builds upon causal discovery rather than sparse regression. Because our approach can use non-linear conditional independence tests, we can avoid the difficult dictionary construction step associated with sparse regression methods.

In contrast to causal discovery methods, other current research also focuses on approximating ordinary differential equations or PDE-like systems with operator learning approaches, such as operator neural networks (Li et al., 2020; Pathak et al., 2022; Hart et al., 2023). These Fourier Neural Operators (FNO) focus on generating accurate models of the PDE-like evolution of key variables over time and space. Their assumptions are rooted in several of the same fundamental physical principles of how PDEs propagate effects in space and time as CaStLe: locality in space and time and spatial stationarity. While CaStLe is not meant to be a predictive model, it captures important relationships between grid cells in an interpretable fashion, providing insights into the underlying causal structures.

6.4.3 Causal Discovery of Physical Dynamics: Dynamical Constraints

We state here four key assumptions that capture what we describe as a PDE-like system $\boldsymbol{X}_t$:

T1) Temporal Locality: for any $\tau \neq 1$, $X_{i,t-\tau} \not\rightarrow X_{j,t}$ for any spatial coordinates (i, j)

T2) Temporal Causal Stationarity: the dynamics governing the evolution of $\boldsymbol{X}_t$ do not change over time. That is, $X_{i,t-1} \rightarrow X_{j,t} \Leftrightarrow X_{i,t-1+\tau} \rightarrow X_{j,t+\tau}$ for any time offset τ.

S1) Spatial Locality: if (i, j) are not neighbors (in a problem-specific sense) then $X_{i,t_1} \not\rightarrow X_{j,t_2}$ for any t_1, t_2.

S2) Spatial Causal Stationarity: the dynamics governing the evolution of $\boldsymbol{X}_t$ do not change over space. That is, $X_{i,t-1} \rightarrow X_{j,t} \Leftrightarrow X_{i+s,t-1} \rightarrow X_{j+s,t}$ for any spatial offset s.

Here, $\not\rightarrow$ denotes the absence of a direct causal relationship between two variables.

Therefore, if an SCM exists for a given system, then it will have a functional shape constrained by our assumptions: $X_t = f(X_{t-1}, \eta_t)$, for some vector of noise, η_t. In the context of an SCM, the constraints are: temporal locality (T1) adds lagged relationships between parent and child variables; spatial locality (S1) restricts possible parents to those in the spatial neighborhood of each variable (grid cell), that is, f_i is only a function of the neighborhood of i (f_i depends only on $\boldsymbol{X}_{\mathscr{P}(i)}$); and temporal/spatial causal stationarity (T2 & S2) require that there be only one function, f, for all space and time in the window/region of analysis.

Building on physical principles, Assumption T1 implies that causal dependencies follow the "arrow of time" while S1 disallows "action at a distance." Assumptions T2 and S2 serve to ensure that there is a consistent causal structure to target. Assumption S1 further requires that f_i is only a function of the neighborhood of i (f_i depends only on $\boldsymbol{X}_{\mathscr{P}(i)}$). We refer the reader to the book by Peters et al. (2017) for a more detailed discussion of how SCMs can be used to model physical systems.

We deliberately chose lag-1 temporal relationships in assumption T1 because they reflect fundamental physical principles: In the discretized form of PDEs, each element depends on the future state of the immediate past of its neighboring elements. The symmetry of the radius-1 neighborhood in assumption S1 and the single lag constraint in assumption T1 captures the essential causal dynamics in physical processes when temporal and spatial data resolutions are appropriately balanced.

While not descriptive of all possible systems, we assert these locality and stationarity assumptions are descriptive of any system governed or modeled after PDEs, cellular automata (Bhattacharjee et al., 2020), or Tobler's First Law of Geography (Miller, 2004; Walker, 2022). These assumptions reflect fundamental principles of locality and consistency that apply across numerous domains, from fluid dynamics to reaction-diffusion systems. However, for these to hold in practicality, one must also assume sufficient data is available to characterize locality and dynamics are smooth and non-turbulent, relative to the analysis frame. These assumptions imply that there is an optimal balance between temporal and spatial resolution sufficient to impose space-time locality. The exact value of this scaling is problem-dependent, as more rapidly evolving systems require higher temporal resolution, and we do not explore it further here. However, we note that similar concerns are well-studied in the design of numerical differential equation solvers where spatial and temporal discretizations must be chosen suitably consistently.

Section 6.6 and A detail how these assumptions are essential for our methodology, CaStLe, and discuss their limitations. Section 6.6.6 discusses strategies for managing those limitations. While CaStLe's framework assumptions (T1, S1, T2, S2) enable efficient use of space-time samples, the algorithm adapted for CaStLe's parent-identification phase will have additional causal assumptions.

Interestingly, CaStLe's spatial locality assumption (S1) creates an environment where, when properly implemented, causal sufficiency can be satisfied by construction. When we focus on learning only the parents of the center cell while including all potential spatial neighbors in the analysis, we automatically satisfy causal sufficiency for that specific node if S1 holds. While reliant on S1 holding, this is significant because causal discovery is notoriously the most challenging causal discovery assumption to ensure in real-world settings (Spirtes et al., 1993; Raghu et al., 2018). As we discuss in Section 6.6.5, sufficiency may be relaxed depending on which causal discovery algorithm is adapted for the parent-identification phase. However, satisfying it by construction may enable implementation choices with fewer compromises.

In the following sections, we discover grid-cell-level causal graphs under these five assumptions. Assumptions T1 and S1 allow us to significantly reduce the scope of the problem, as there are only 9 possible parents of a grid cell in 2D (8 neighbors and itself). Assumptions T2 and S2 suggest that we only need to determine a single local causal graph, because spatial stationarity allows us to extend it to the entire domain.

6.5 Data: The 1991 Mt. Pinatubo Eruption

Mount Pinatubo's eruption in 1991 was a massive, natural intervention in the climate, with effects that had a relatively high signal-to-noise ratio. The event launched 20 Tg of SO_2 gas into the atmosphere (Guo et al., 2004a,b; Kremser et al., 2016). The sulfate aerosols that resulted from these gases remained in the stratosphere for approximately two years, leading to stratospheric warming of $\sim$ 1.5K and surface cooling of 0.2-0.5K (Dutton and Christy, 1992; Labitzke and McCormick, 1992; Parker et al., 1996a; Soden et al., 2002). This aerosol injection has recently been the object of much study, with some authors suggesting it as a natural proxy for proposed stratospheric aerosol injection (SAI) responses to global climate change (Trenberth and Dai, 2007). Recent work continues to characterize the nature of the response to the Pinatubo eruption, with the timing and

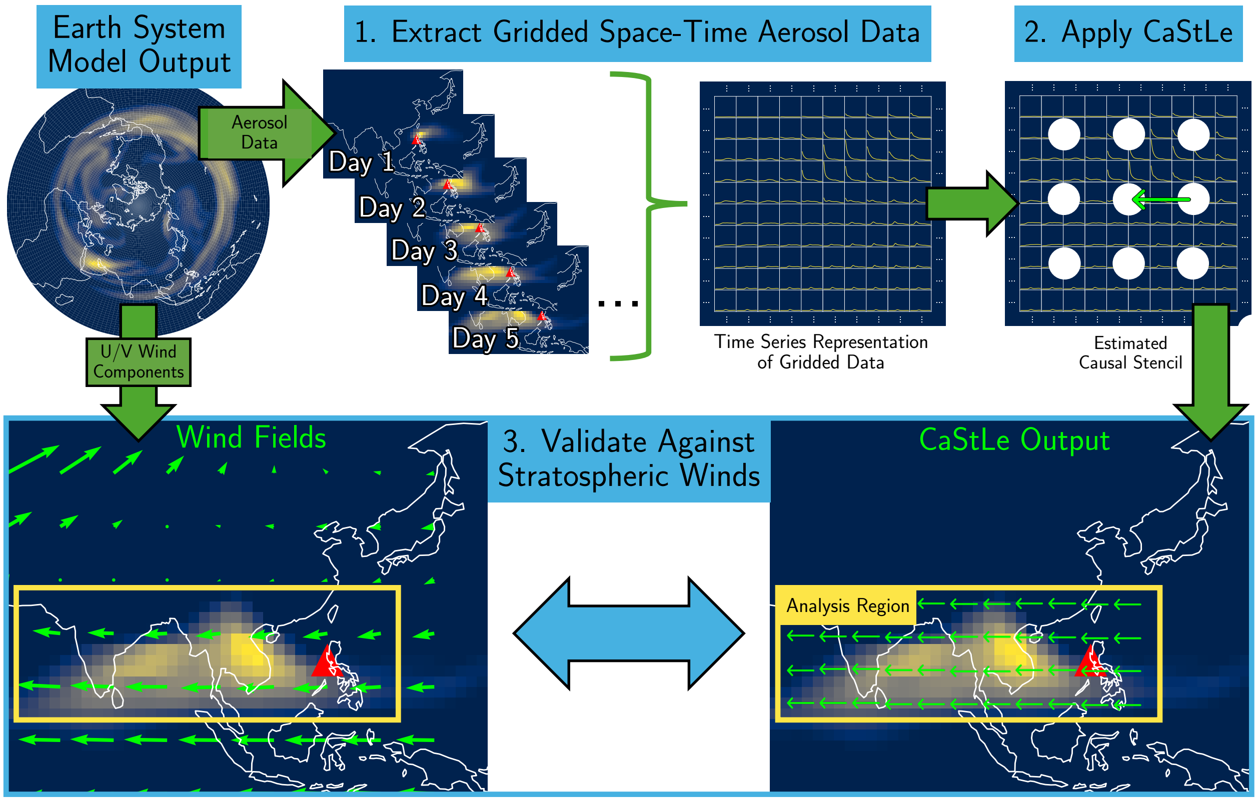


Figure 6.1: Schematic overview of the key elements of CaStLe and the process followed in its application to Mount Pinatubo's eruption of stratospheric aerosols. Beginning with Earth system model output, Step 1. is to collect stratospheric wind and aerosol data. Step 2. is to apply our novel CaStLe meta-algorithm to the aerosol data to obtain a causal graph describing the space-time evolution of the aerosols. Finally, we use the wind fields to help validate the causal graph results in Step 3.

spatial structure of the surface response being essential factors to inform policy decisions (Weylandt and Swiler, 2024).

Large volcanoes can impact climate quantities, such as surface temperatures, on timescales from months to years (Parker et al., 1996b; Robock, 2000; Timmreck, 2012; Marshall et al., 2022). However, to evaluate whether CaStLe could recover the initial advection dynamics of volcanic aerosols, we focused on the period shortly after the eruption that includes stratospheric aerosol transport. The recent paper by Marshall et al. (2022) indicates: "Although global-scale climatic impacts following the formation of stratospheric sulfate aerosol are well understood, many aspects of the evolution of the early volcanic aerosol cloud and regional impacts are uncertain." This initial spread of aerosols in the stratosphere is a geophysical process, falling between synoptic weather patterns and longer-term impacts.

We utilized models of the event, combining stratospheric aerosol and wind data, as case study to illustrate the analysis possible with CaStLe. Figure 6.1 is a high-level illustrative schematic of the this work's key ideas: We collect gridded space-time data, e.g. aerosol optical depth (AOD) measurements, and apply it to CaStLe to learn a causal stencil graph. We then map the stencil to the original grid space. Finally, we compare the data to ground-truth. o be clear, the ground-truth in our later case studies is a proxy, referring to the models' understood underlying dynamics, not the true realization of AOD in Earth's atmosphere or a mathematical representation of the dynamics. In Section 6.7, we compare to the wind fields carrying AOD as a proxy ground-truth. In Section 6.8, we compare CaStLe results from synthetic data to mathematically-known ground-truth.

6.5.1 Held-Suarez-Williamson-Volcanic

For our first case study, we utilized the limited-variability ensemble approach of the Held-Suarez-Williamson-Volcanic (HSW-V) model (Hollowed et al., 2024). HSW-V is an atmosphere-only model built in the Department of Energy's Energy Exascale Earth System Model version 2 (E3SMv2) (Golaz et al., 2022). HSW-V does not set out to replicate the historical Mt. Pinatubo eruption or any other, but uses the Mt. Pinatubo's eruption characteristics "to produce a plausible realization of a volcanic event, simulated with a minimal forcing set" (Hollowed et al., 2024). The model was developed specifically to facilitate basic research of attribution methodologies by providing realistic source-to-impact pathways of eruption quantities. We use this model to create a realistically complex dataset of stratospheric aerosol and wind dynamics with a clear ground-truth to demonstrate the capabilities of CaStLe and the correctness of its results.

We gathered aerosol optical depth (AOD), sulfate, and zonal (U) and meridional (V) wind fields for analysis. Only AOD is provided to CaStLe, while the sulfate, U, and V wind components are used for validating results, as detailed in Section 6.7. AOD is a derived quantity that measures the extinction of a beam of light through the atmosphere by atmospheric aerosols, i.e., it describes the amount of light occluded by atmospheric particles. One of the simplifying aspects of HSW-V is that all aerosol particles originate from SO_2 gas ejected by the volcano; this avoids confusing signals from other sources, such as smoke and dust, in the atmosphere.

The data collected from the HSW-V ensemble run are on a 2° grid with 6-hourly average observations. We selected AOD in grid cells between $-20.00°$ to 40.00°N and $-120.00°$ to 140.00°E, comprising 3,900 grid cells. We used the first three weeks post-eruption for our analysis.

6.5.2 Mt. Pinatubo in E3SMv2-SPA

For our second case study, we considered a simulation of the Mt. Pinatubo eruption in the fully coupled E3SMv2 model augmented with Stratospheric Prognostic Aerosol capability (E3SMv2-SPA) as detailed and validated by Brown et al. (2024). E3SMv2-SPA includes atmosphere, land, ocean, sea ice, land ice, and river components. AOD, U, and V wind fields are analogously collected from this dataset. However, in this model, aerosols are a natural feature, thus complicating the analysis of aerosol optical depth.

Data were collected on a daily temporal resolution for a 1° spatial grid. We selected grid cells between -30.00° to 60.00°N and -180.00° to 180.00°E. Analysis covered the first six months. Because this data has a coarser temporal resolution and finer spatial resolution than our study of HSW-V, we coarsened the CaStLe spatial grid to a 3° grid, resulting in 3,600 total grid cells. This helps ensure that the motion of aerosol particles between grid cells is measured within the one-day sample period.

6.6 Methodology: Causal Discovery with CaStLe

6.6.1 Notation

We first introduce notation used in the remainder of this paper. Data is observed on a spatial domain $\mathscr{D}$, which we typically take to be a finite subset of the real plane, $\mathbb{R}^2$. The causal structure generating this data can be represented by a directed acyclic graph $\mathscr{G} = (\mathscr{V}, \mathscr{E})$, where $\mathscr{V} = \mathscr{D}$. CaStLe represents local causal structure with a *stencil*, which we identify as a graph $\tilde{\mathscr{G}} = (\tilde{\mathscr{V}}, \tilde{\mathscr{E}})$ in a reduced coordinate space ($|\tilde{\mathscr{V}} = 9|$). In both the original and reduced spaces, let $\mathscr{P}(v)$ be the *potential* causal parents of v and let $\mathscr{P}(v)$ be the *actual* causal parents of v. We take $\mathscr{D}$ to be points on a regular grid of size $N \times N$, observed over T time steps, giving data $\boldsymbol{X} \in \mathbb{R}^{N^2 \times T}$. When transformed to the reduced space used by CaStLe, the resulting data matrix will be denoted $\tilde{\boldsymbol{X}} \in \mathbb{R}^{T(N-2)^2 \times 9}$. Quantities estimated from data are denoted with a hat, e.g., $\hat{\mathscr{P}}(v)$. We provide additional background on the interpretation of the causal graphs $\mathscr{G}, \tilde{\mathscr{G}}$ in Section 6.4.1 and formally specify the mapping between $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$, or equivalently, between $\mathscr{V}$ and $\tilde{\mathscr{V}}$, in Section 6.6.3.

6.6.2 Causal Space-Time Stencil Learning

We now introduce the CaStLe paradigm for the causal discovery of local space-time dynamics. Under our assumptions, CaStLe identifies a *sketch* of the local causal dynamics, which we call a stencil. This stencil can then be used to construct the causal graph for the entire system (S2). The stencil is estimated in a reduced coordinate space, where we only examine the direct neighbors of a given grid cell (S1). We can pool information across time (T2) and space (S2) in order to estimate the stencil accurately, and the problem is tractable because we only seek causal parents which are local in time (T1). As we will see, this combination of reduced search space and pooled information provides a powerful approach to causal discovery and enables accurate causal discovery from high-dimensional grid-cell-level data.

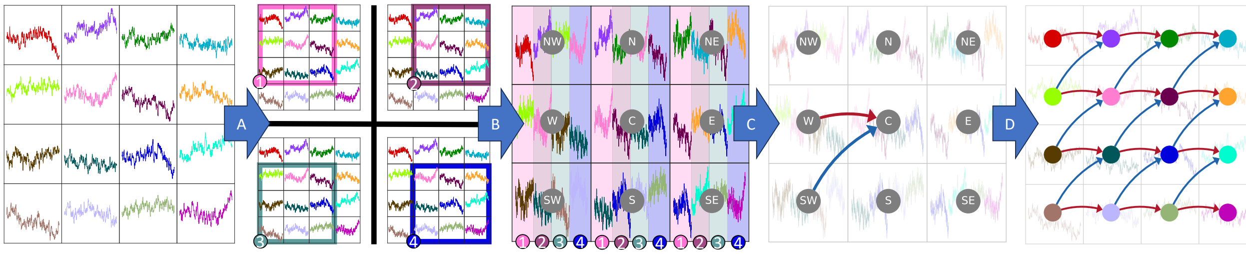


Figure 6.2: Illustration of CaStLe (Algorithm 1) as applied to space-time data on a 4×4 grid. *Step A* (§6.6.3): for every interior grid cell, its 3×3 (Moore) neighborhood is selected. (Note, all four 4×4 grids in the second panel are identical.) *Step B* (§6.6.3): Data are represented in a reduced coordinate space obtained by appending time series from each neighborhood according to its position relative to the neighborhood's center. *Step C* (§6.6.3): during the Parent Identification Phase (PIP), a causal discovery algorithm is used to estimate the parents of the center time series; the resulting graph forms the causal stencil. *Step D* (§6.6.3): the estimated stencil is expanded to its equivalent representation in the original space. Note that each *time chunk* (colored intervals in the center panel) in the reduced space corresponds to an interior grid cell of the original data, and that each edge in the final causal graph reflects to a stencil edge learned during the PIP. See §6.6.3 for details.

Having motivated the CaStLe approach to causal discovery from space-time data in Section 6.4.2, we now state it formally as Algorithm 1, describe its computational steps, and then analyze its statistical and computational properties.

6.6.3 The CaStLe Meta-Algorithm

Steps A-B: Projection to a Reduced Coordinate Space

CaStLe begins by transforming the given data from its original domain into a reduced coordinate space that captures the underlying causal dynamics' locality and spatial homogeneity. In this transformation, all data points are preserved, i.e., no marginalization or truncation occurs. This process is represented as Steps A and B in Figure 6.2 and Algorithm 1. In Step A, the local 3×3 (Moore) neighborhood of

Algorithm 1 CaStLe for Space-Time Data in 2D ($\mathscr{D} \subseteq \mathbb{R}^2$)

Inputs:

- Parent-Identification Phase subroutine PIP
- Gridded space-time data $\boldsymbol{X} \in \mathbb{R}^{T\times N^2}$

---

1. Step A: Extract 3×3 Moore Neighborhoods
   - For each interior point in the original space, construct local view of the data $\boldsymbol{X}_i = [X_{\cdot\mathscr{P}(i)}] \in \mathbb{R}^{T\times 9}$
2. Step B: Construct Reduced Space Data Matrix
$$\tilde{\boldsymbol{X}} = [\boldsymbol{X}_1^\top \boldsymbol{X}_2^\top \ldots \boldsymbol{X}_{(N-2)^2}^\top]^\top \in \mathbb{R}^{T(N-2)^2\times 9}$$
3. Step C: Perform Parent-Identification in Reduced Space
$$\mathsf{PIP}(\tilde{\boldsymbol{X}}) = \tilde{\mathscr{E}} = (\hat{\mathscr{P}}(\mathtt{C}) \times \mathbb{R}^9) \subseteq \mathscr{P}(\mathtt{C}) \times \mathbb{R}^9$$
4. Step D: Expand Stencil Graph to Original Coordinate Space:
   - $\mathscr{E} = \emptyset \subseteq \mathscr{V}^2 \times \mathbb{R}$
   - For each $(p,w) \in \hat{\mathscr{E}}$:
$$\mathscr{E} = \mathscr{E} \cup \{(p(v), v, w) \text{ for } v \in \mathscr{V}\}$$

Outputs:

- Graph Stencil, $\tilde{\mathscr{E}}$
- Estimated Causal Graph, $\mathscr{G} = (\mathscr{V}, \mathscr{E})$

each interior cell is selected, and each cell is labeled by its location relative to the center cell (`S`, `NW`, `E`, etc.). This process creates $(N-2)^2$ sub-views in $\boldsymbol{X}_i \in \mathbb{R}^{T\times 9}$.

In Step B, these views are concatenated along the time dimension to create a reduced coordinate space data matrix $\tilde{\boldsymbol{X}} \in \mathbb{R}^{T(N-2)^2\times 9}$. Note, when concatenating the subviews, data are aligned by their coordinates relative to the neighborhood center so that, e.g., data from all `NW` cells are aligned upon concatenation, even though they originally come from different spatial locations. Although this transformation results in specific time series segments appearing in multiple reduced space cells, these repetitions do not eventually create spurious dependencies in the causal stencil, as CaStLe only seeks lag-1 dependencies. The repeated segments are well-separated in the temporal dimension, and no chunks appear in different cells in the same interval.

We depict this process on a 4×4 grid in the first half of Figure 6.2. In Step A, the four interior cells are sequentially highlighted, and their local neighborhoods are extracted, which are depicted in boxes colored according to the center used. In Step B, the local data views are concatenated to form one set of time series, with each temporal *chunk* reflecting the color of the center cell of the underlying data view.

Step C: Parent-Identification Phase

CaStLe next examines the reduced coordinate space data representation, $\tilde{\boldsymbol{X}}$, to identify the stencil of the local causal dynamics. This is done by applying an augmentation of an arbitrary time series causal discovery algorithm to identify the parents of the center cell, `C`. We emphasize that we only seek the parents of `C`, not a full causal structure, in this step and refer to it as the *Parent Identification Phase* (PIP). Under assumption S1 (locality), all parents of `C` are present at this step, satisfying causal sufficiency, ensuring more accurate estimation of the causal stencil. By contrast, while the data of the parents for the exterior cells, e.g. `W`, is included in the reduced data space matrix, $\tilde{\boldsymbol{X}}$, it spreads across multiple columns, and accurate parent identification is not possible. The output of this process is a set of (up to) 9 weighted edges, corresponding to the parents of `C` (the eight neighboring cells and `C` itself).

We depict the PIP in Step C of Figure 6.2, where two parents of `C` are identified: `W`, which has a positive dependence on `C`, and `SW`, exhibiting negative dependence. Note that while the PIPs we implemented in testing—see Section 6.8.1—had no trouble with the *seams* connecting each time *chunk* in the reduced space, we propose an improved testing implementation in E to alleviate potential statistical testing issues.

Step D: Graph Reconstruction in the Original Space

Finally, CaStLe uses the stencil constructed in Step C to reconstruct the causal graph in the original data space, in a process that essentially reverses Steps A and B. Specifically, for each edge identified in $\tilde{\mathscr{E}}$, corresponding edges are added to grid cell in the original domain. We depict this in the final step of Figure 6.2 where the stencil is repeated throughout the entire 4×4 space, copying the two parents of `C` identified in Step C, to create a causal graph in the original space. Note also that we use the stencil to identify parents for both interior and boundary cells, omitting edges that go "off-grid" when applying the stencil to boundary cells.

6.6.4 Theoretical Properties

CaStLe has many advantages over classical causal discovery algorithms in gridded space-time settings. By reducing the causal discovery problem to identifying the causal parents of the center cell (`C`) in the reduced space, CaStLe achieves significant improvements in both the computation necessary to infer the causal graph and the statistical quality of that graph. As previewed in Section 6.4.2, the PIP's focus on identifying only the parents of the center cell creates an important connection to the causal discovery assumption of causal sufficiency. Because we include all spatial neighbors (as defined by our locality assumption S1) in the conditioning set, all potential parents of the center cell are present in the analysis. If our spatial locality assumption holds, causal sufficiency is automatically satisfied within each local stencil analysis. This represents a key advantage of the CaStLe framework - while the Markov condition and faithfulness remain necessary assumptions for the PIP algorithm, our implementation leverages spatial structure to ensure causal sufficiency by construction.

Below, we briefly outline the theoretical implications and their contributions to CaStLe's remarkable performance and algorithmic improvements. Their derivation, a deeper analysis, and a discussion on graph estimation asymptotic consistency are provided in B. We discuss CaStLe's asymptotic consistency in C, which shows that CaStLe converges on the correct causal stencil as grid size increases, given a PIP consistent in increasing time samples. These properties illustrate the mathematical justification for CaStLe's empirical correctness and improvement over the state of the art shown in the following sections.

CaStLe yields significant improvements to both *time complexity*, a measure of an algorithm's computation time as it scales with input size (e.g., number of time steps, graph nodes), and *statistical complexity*, a measure of estimation performance given larger sample sizes. Following the complexity analysis of Kalisch and Bühlmann (2007), we show that traditional causal discovery approaches are bounded by $\mathscr{O}(np^3 2^p) = \mathscr{O}(T(N^2)^3 2^{N^2}) = \mathscr{O}(TN^6 2^{N^2})$, for T time samples and $N \times N = N^2$ grid cells. Since CaStLe computes on the smaller *reduced coordinate space*, and only seeks causal parents of one node, rather than parents of all nodes, several terms become constants, resulting in $\mathscr{O}(np^3 2^p) = \mathscr{O}(T(N-2)^2 \times 9^3 \times 2^9) = \mathscr{O}(TN^2)$. CaStLe's computational complexity is $\mathscr{O}(TN^2)$, a major improvement over existing approaches. For more details on this derivation, see Appendix B.1. By leveraging locality and spatial replicates, CaStLe identifies causal structure for the entire graph ($\mathscr{O}(N^4)$ possible edges) in N^2 time. Kalisch and Bühlmann (2007, Appendix B) show that the probability of the PC algorithm incorrectly estimating the true graph is bounded by $\approx \mathscr{O}(N^{2N^2})$, whereas we find that CaStLe's error probability scales as $\approx \mathscr{O}\left(\frac{N^2 T}{e^{N^2 T}}\right)$. From this, as the grid size grows larger, we see PC is less likely to estimate the correct causal graph, while CaStLe is more likely to estimate the correct graph. Furthermore, both of these effects are exponential, implying significant performance differences even on moderately sized graphs; this change from a regime of exponential decay to super-exponential growth in graph recovery performance makes local causal graph recovery feasible, finally enabling the tools of causal discovery to scalably explore grid-level Earth science dynamics in commonly high-dimensional settings.

6.6.5 Methodological Limitations

CaStLe's assumptions may pose challenges in some domains of interest, and violations of these assumptions can affect the CaStLe output. For example, large-scale homogeneity can be difficult to achieve in geosciences, which is the primary rationale for the spatial-blocking strategy that we implement for our application in Section 6.7. Locality assumptions (T1 & S1) create a framework where the causal Markov condition can be effectively applied to local structures, while causal stationarity assumptions (T2 & S2) create consistency in these structures across space and time. However, the PIP algorithm we use within CaStLe additionally requires standard causal discovery assumptions, particularly the causal Markov condition and faithfulness, which is a separate non-trivial assumption. We list causal sufficiency as an assumption, however, if the others hold then it follows that all of the causal parents of the stencil's center are in its immediate neighborhood, so sufficiency is satisfied by construction. Alternatively, causal sufficiency may be relaxed if the chosen PIP is an algorithm that does not rely on sufficiency, such as the FCI algorithm (Glymour et al., 2019). As such, violations of CaStLe's assumptions relate directly to violations of the causal Markov condition, faithfulness, and causal sufficiency. Both Spirtes et al. (1993, p. 29) and Runge (2018a) discuss assumption violations in causal discovery and some examples of how they manifest in resulting graphs. We have included a more detailed discussion on each assumption and their limitations in A.

6.6.6 Strategies for Addressing Limitations

To address the limitations of CaStLe's assumptions, several practical strategies can be employed. One effective approach is the use of spatial blocking to create subdivisions where dynamics are more uniform, thus mitigating the violation of spatial causal stationarity (S2). The selection and size of these blocks are highly domain-dependent and can be guided by subject matter expertise. An automated approach may be sufficient for certain dynamics, such as stratospheric dynamics, but more manual approaches may be necessary for surface-level dynamics where blocks are chosen based on topological assumptions. In specific areas of interest, blocks can be manually created to avoid topological boundaries such as coastlines, rivers, and mountain ranges, ensuring that the assumptions of spatial homogeneity are better satisfied.

Additionally, strategies such as variograms can be used to test for spatial statistical stationarity, providing heuristics for effective blocking. In future work, an iterative block size estimation approach could be considered. Varying the block size serves as a form of *stability check*, a technique widely applied in ML to ensure robustness of discoveries to parameter choices and modeling assumptions (Allen et al., 2023). However, it is important to note that there may not always be a single optimal block size due to the complex nature of spatial dynamics. Instead, there may be a range of valuable block sizes depending on the needs for analysis and the limitations of the setting. Because CaStLe is data efficient, it may be better to tend towards smaller blocks, which are more likely to be homogeneous, but possibly at the cost of some interpretability.

Deep learning and space-time feature engineering approaches may be fruitful directions for future research on automated block-identification. Methods such as δ-MAPS (Fountalis et al., 2018), feature extraction with convolutional neural networks (Nukavarapu et al., 2023), and spatiotemporal cluster analysis (Davis et al., 2025) are strong starting points. These computational approaches could automate the identification of optimal spatial blocks, reducing reliance on manual delineation and subject matter expertise while preserving the statistical properties necessary for valid causal discovery with CaStLe.

By employing these strategies and acknowledging their limitations, the robustness and applicability of CaStLe in various domains can be significantly enhanced, allowing for more accurate causal discovery in complex space-time systems. In general, more data at higher spatial and temporal resolutions will make satisfying the assumptions easier. The appeal of CaStLe is when one is interested in small-scale local dynamics, it is preferable to analyze raw gridded data directly, because marginalization can introduce statistical artifacts.

I provides an empirical investigation of how violations of each assumption affect CaStLe's performance when applied to our E3SMv2-SPA case study. Our analysis reveals that CaStLe is surprisingly robust to moderate assumption violations. While violations of spatial and temporal causal stationarity (particularly with overly large blocks or extended time intervals) introduce more false positives and reduce interpretability, CaStLe often still identifies key true causal pathways. This robustness to moderate assumption violations further expands the practical utility of CaStLe in realistic Earth science applications where perfect adherence to assumptions is rarely possible.

6.7 Results: Discovering Atmospheric Dynamics in Global Climate Models

As described in Section 6.5, we applied CaStLe to output of the Held-Suarez-Williamson-Volcanic atmosphere model, tuned to accurately reproduce the observed Pinatubo response (Hollowed et al., 2024), and the E3SMv2-SPA model including the eruption. In this section, we describe how we applied CaStLe to these case studies and present the results.

6.7.1 Validation with HSW-V

We first note important implementation considerations, particularly how CaStLe's assumptions are satisfied. In general, if assumptions T1, T2, S1, and S2 are uncertain, either because of data availability or dynamical instability, then assumptions can be verified using subject matter expertise. In this study of Mt. Pinatubo, we describe how we carefully managed each assumption prior to applying CaStLe.

In order to be sure CaStLe's assumptions of temporal locality, temporal causal stationarity, and spatial locality (T1, T2, and S1) held in the dataset's $2°$ grid resolution (corresponding to approximately 214 km at 15 degrees N), we used atmospheric wind speeds at the time of the eruption, which were recorded at 25 m/s on average at 30 hPa; cf. Figure 1 in Thomas et al. (2009). That speed translates to a theoretical maximal aerosol travel distance of 540 km over a 6-hour period, meaning aerosols should move fast enough to traverse one $2°$ grid cell per time step.

Spatial causal stationarity, assumption S2, is indeed violated considering the globe holistically. We resolved this challenge by using a spatial blocking strategy to create subdivisions in which dynamics were more uniform, and applied CaStLe within each separately. As noted in Section 6.6.6, the selection of blocks and their size is a potential challenge and is highly domain-dependent. We conducted a sensitivity analysis of block sizes, which is presented in H, and determined that dynamics were consistent in various of block sizes. We chose a middle size, $20° \times$

20°, for this analysis to balance more nuanced outputs (smaller sizes) with less risk of false positives (larger sizes). This case study was selected for its relatively simple advective dynamics to clearly validate CaStLe and demonstrate its results in an atmospheric setting. We observe that stratospheric winds vary smoothly and slowly, without hard boundaries, which enables us to use a regular grid of blocks. Other settings, such as surface level analyses, the blocking strategy will certainly require special treatment to avoid analysis across hard dynamical boundaries, such as coastlines and mountain ranges. In H, we also demonstrate that blocking alone is not sufficient for non-CaStLed approaches to succeed.

We chose CaStLe's PIP to be the PC-Stable-Single algorithm because in our validation experiments in Section 6.8.1, we found it to be the marginally more effective PIP. However, those experiments showed any tested PIP algorithm is effective. PC-Stable-Single is the PC-Stable causal discovery algorithm (Colombo and Maathuis, 2014) adapted to find the causal parents of only one node; its pseudocode is provided in L. Specific CaStLe parameterizations are given in G. In J, we present similar results using DYNOTEARS for CaStLe's PIP.

Our proxy ground-truth in this case study was stratospheric winds that cause suspended aerosols to advect through space. We display dominant wind fields throughout the space to validate the resulting graphs. Our dataset included wind components in 72 pressure levels in the HSW-V dataset, so we display column-averages of the levels at the levels where volcanic sulfate was most prevalent. Specifically, we chose pressure levels containing more than 5.00 µg of sulphate

Kg air, which were between $\sim$ 6-114.00 hPa. With this, we effectively captured the stratosphere and 56% of all sulfate aerosols in all atmosphere levels. By comparing winds in at the stratospheric levels where most of the sulfur was present, we can directly compare CaStLe's discovery of AOD's space-time evolution to wind data in the same locations.

Comparing the wind and recovered stencils in Figure 6.3, it is clear to see that CaStLe is able to accurately reconstruct the prevailing stratospheric winds using only AOD observations. As these wind fields are the key drivers of aerosol dispersal, it is clear that CaStLe can accurately capture the dynamics dictating the spatial pattern of the Pinatubo response. The CaStLe stencils best capture the underlying wind fields when AOD levels are high. When there are few particles in a region, it is challenging to determine wind by solely observing dispersal patterns. We also observe a zonal (East-West) pattern driving the aerosol dispersion, with Pinatubo aerosols transported nearly fully around the equator within 3 weeks, while meridional (North-South) dispersion taking much longer. This alignment between CaStLe-derived causal structures and observed wind patterns demonstrates the method's effectiveness in reconstructing the physical mechanisms driving aerosol transport, particularly in regions with sufficient particle density to enable clear detection of dispersal trajectories.

Comparative Analysis of CaStLe Versus Traditional Approaches on HSW-V

The current state-of-the-art causal discovery methods cannot tractably approach this study of Mt. Pinatubo's aerosol short-term evolution. As described in Section 6.3, dimensionality reduction techniques commonly used to make them tractable are suitable for spatially static, periodic space-time patterns. However, they are not good solutions for studying a dynamic, transient pattern because modes derived from those techniques are space-timely invariant. Moreover, they are meant to capture large-scale teleconnections, rather than local dynamics that eventually give rise to global phenomena such as teleconnections. For a detailed demonstration of why dimensionality reduction approaches, such as PCA and PCA-varimax, are insufficient for capturing local causal structures in space-time systems like volcanic eruption plumes, see F.

Traditional approaches attempted without dimensionality reduction suffer from the *curse of dimensionality* when applied to short-term global-scale phenomena because there are more grid cells than temporal observations. They also struggle to identify local connections in the massive search space they seek, where every grid cell may be dependent on any other grid cell; i.e., they are not constrained by local causal structure. Finally, their efficiency scales poorly as the grid size gets larger, requiring a lot of time to execute on relatively small grids. We present specifics below and discuss time complexity in depth in Section 6.6.4 and Appendix B.1.

Here, we demonstrate the disparity in performance between traditional approaches and CaStLe for our HSW-V case study using the PC algorithm. The reasons for the disparity are explored in Sections 6.3 and 6.4. Because PC did not terminate within 48 hours on the full spatial region studied in Section 6.7.1, we restricted the analysis space the area between 20.00° to 50.00°N and 55°W to 120°E in the first 8.5 days after the eruption. On the 2° grid, the given space is equivalent to a 35×35 grid, or 1,225 grid cells. Since temporal observations were 6-hourly, there were 34 time series samples per grid cell.

Figure 6.4 shows the results of the PC causal algorithm and CaStLe-PC-Stable applied to a large section of grid cells for the HSW-V problem. Figure 6.4a illustrates that PC is incapable of reconstructing a graph with any meaningful physical interpretation. There are some local dynamics found, but they are dominated by the many links across disparate locations. PC was implemented here with the partial correlation conditional independence test, a test alpha-value of 0.00001, and a p-value threshold of 0.05 to remove links below that threshold in the final graph. P-values were corrected using the Benjamini-Hochberg procedure prior to final thresholding.

In Figure 6.4b, CaStLe was applied to 10°-by-10° blocks, rather than the 20°-by-20° blocks in Figure 6.3. The smaller block size enables more link density and nuanced results, with the possibility of more mistakes. In this illustration, we chose to display the stencils mapped back to the original space for each block to compare to PC more fairly and demonstrate how much more sparse CaStLe's results are. We found that CaStLe was again able to recover the westward aerosol transport from Mt. Pinatubo. Because HSW-V only models aerosols from the volcano, there is little to no aerosol signal outside the plume, and results in these areas will be less reliable.

Additionally, the run-time of the PC algorithm is demonstrably poorer than CaStLe. The PC algorithm experiment in Figure 6.4a PC took 65 minutes to execute for a 35×35 grid size. In contrast, the CaStLe experiment in Figure 6.4b completed all blocks serially in 0.46 seconds on the same data. Further, for each of the panels in Figure 6.3, CaStLe computed the 39 stencils for the 3,900 grid cells in a total of 10 seconds. These empirical data points are explained by CaStLe's improved theoretical properties, as detailed in Section 6.6.4 and B.

6.7.2 Extending to More Complexity: E3SMv2-SPA Modeled Aerosols

Given the intended simplicity of the HSW-V model, we also evaluated a simulation of the Mt. Pinatubo eruption in E3SMv2-SPA. More complex graphs arise with a more complex model, providing an opportunity for more nuanced analysis and discovery, but with a higher chance of false positives and false negatives. E3SMv2-SPA is a fully coupled model, so AOD results from many sources including the volcanic eruption and Saharan dust. As such, we expect results to be somewhat noisier, however, as we demonstrate below, CaStLe is still able to identify important features of transport. Because of this additional complexity, we focus on CaStLe as an exploratory tool and leave additional analysis to future work. However, even with the added complexity, CaStLe can obtain compelling results consistent with dominant stratospheric winds as well as the dynamics discovered in our study of HSW-V.

We used 15° spatial blocks so that CaStLe operates on a 5×5 grid space per block. This size strikes a balance in the trade-off that a smaller block-grid enables more nuance in the final output, and larger block-grids take advantage of more spatial replicates to multiply sample size. We chose to study the eruption in two distinct 20-day intervals spanning a six month period to understand the changing evolution of the plume.

Similarly to HSW-V, we utilize the U and V wind fields to visually validate the CaStLe results. In this case, we did not average over multiple altitudes, instead opting to simply use the 50 hPa wind fields; this altitude was shown in Brown et al. (2024, Figure S6) to contain significant levels of the sulfate aerosols.

Figure 6.5 depicts the results of our experiment on E3SM. Again, we applied CaStLe-PC-Stable to construct causal stencils for each given spatial block. We selected two intervals of interest from our results to show here. Day 15 is June 15, 1991, the day of the eruption, so the top row of Figure 6.5 is the first 20 days after the eruption. The bottom row was selected to illustrate later dynamics when aerosols have circumnavigated the tropical zone and more northward advection is present. Days 175-195 are November 22 to December 12, 1991, a little over six months after the eruption.

In the more challenging setting of the fully-coupled E3SMv2-SPA model, our results in the first weeks are still generally consistent with those in HSW-V presented in Section 6.7.1, showing that CaStLe is largely robust to greater complexity. We note that visually identifying the sulfate aerosol plume is much more difficult in this case as the background AOD is quite strong. A solution may be to apply CaStLe to AOD anomalies (computed by subtracting grid cell long-term AOD means from the signal in the analysis period), thus potentially removing background variability from the analysis. However, our goal in this work is to present CaStLe as applied to raw data to illustrate what it can and cannot accomplish in complex, heterogeneous settings.

Regardless, we observe that tropical westward advection is present throughout both studied time periods, but more complexity is present in other regions, in part due to the background AOD. Six months later, the aerosols and winds are in a different regime. We observe northward and southward causal structures in the northern latitudes matching dominant wind fields in the area, with CaStLe stencils still consistent in the tropics. Additionally, CaStLe recovers dynamics moving aerosols northwards above central Asia and southwards through western North America. Causal structures are recovered more often and more accurately where stronger winds coincide with more aerosol presence, building a map of significant aerosol movement. A more complex model and smaller block sizes illustrate more nuanced dynamics, and there is more to learn from these; however, we leave deeper atmospheric dynamics analysis to future work.

6.8 Validation and Benchmarking

In this section, we demonstrate the effectiveness of the CaStLe approach to space-time causal discovery, highlighting its ability to identify structure in low-signal and data-sparse regimes. We first demonstrate the benefits the CaStLe approach can provide to *any* causal discovery algorithm using a synthetic linear-Gaussian dynamics benchmark; we then apply CaStLe to an important non-linear PDE problem, showing that we can determine the underlying advective forcing.

6.8.1 Evaluating CaStLe: A Comparative Analysis

We demonstrate the effectiveness of CaStLe using a set of local interaction models (LIMs), building upon the comparison framework introduced by Nichol et al. (2023). In summary, we defined a stencil for each experiment that dictates how each grid cell depends on its nine neighbors (including itself). A LIM is a special case of an SCM, which simulates the evolution of a gridded space by computing the current state of each grid cell based on a predefined function of the historical states of its neighbors. In the linear case, this is most simply accomplished with vector autoregression (VAR) models, where the coefficient is sparse, only containing nonzero entries where a desired dependence exists between neighbors. The function is defined by a linear function of coefficients in the given stencil. Our results appear in Figure 6.6, which shows that CaStLe provides significant improvements in graph recovery regardless of the causal discovery algorithm used in the parent identification phase.

Data: Benchmark Construction

In order to compare different causal discovery algorithms with a common set of benchmarks, we begin by generating coefficient matrices parameterizing spatially homogeneous and statistically stationary VAR(1)s that satisfy our key assumptions S1 and S2. We generate coefficient matrices for these VARs, $\tilde{M}$, using the following sampling scheme:

1. Generate a random 3×3 *local dynamics matrix*, M, with d non-zero elements, one of which is the central element (autocorrelation). Each of the d non-zero elements, $\{a_i\}_{i=1}^{d}$, have a random value $1.0 \geq \text{coefficient}_i \geq s_*$.

2. Expand M to $\tilde{M}$ on a grid of size $N \times N$ (cf. Step D of Algorithm 1 or Figure 2-2 of Nichol et al. (2023))

3. If $|\lambda_{\max}(\tilde{M})| \geq 1$, scale $\tilde{M}$ by $|\lambda_{\max}(\tilde{M})|$.

4. If $m < s_* \; \forall m \in \tilde{M}$, reject, else accept.

where $|\lambda_{\max}(\tilde{M})|$ is the maximum absolute eigenvalue of $\tilde{M}$, which when above 1.0 indicates the system is numerically unstable (Strang, 2016, p.307). We note that this process is essentially an accept-reject scheme used to sample from the set of statistically stationary & spatially homogeneous VARs on a 2D grid with minimum signal strengths $s_* \geq 0.1$ and fixed sparsity levels in the range $d \in \{1, 2, \ldots, 9\}$.

After each $\tilde{M}$ is generated, we create a single realization, using standard Gaussian noise applied independently, cell-wise at each time step.

Method Comparison: Highlighting CaStLe's Strengths

On each realization, we apply one of three causal discovery algorithms, in both CaStLed and non-CaStLed form: i) the `PC` algorithm of Spirtes and Glymour (1991) as adapted to time series by Runge et al. (2019a, Algorithm S1 with $q = 1$); ii) `PCMCI`, an autocorrelated time series extension of `PC` developed by Runge et al. (2019a); and iii) the `DYNOTEARS` approach of Pamfil et al. (2020), itself a time series adaption of the `NOTEARS` approach of Zheng et al. (2018). We additionally compare each of these against a simple sparse VAR approach, where we estimate VAR coefficients directly using ordinary least squares (OLS) and truncate coefficients with magnitude less than s_*; this approach is not necessarily causal, but it is the exact model of our data generating process and provides a useful point of comparison.

We compare the estimated graph structure with the true graph derived from the sparsity pattern of $\tilde{M}$ and report the average Matthews' Correlation Coefficient (MCC) (Matthews, 1975) and F_1 score over 30 replicates. We used an adapted MCC formula derived by Nichol et al. (2023), which accounts for edge cases in which the denominator would be zero, but is otherwise defined as:

$$\text{MCC} = \frac{(\text{TP} \times \text{TN} - \text{FP} \times \text{FN})}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \tag{6.1}$$

where TP, FP, TN, and FN are true positive count, false positive count, true negative count, and false negative count, respectively. Here, a positive is a graph edge that exists, and a negative is a graph edge that does not exist. The MCC graph similarity measure is sometimes preferable to the more common F_β Score (β is chosen such that recall is considered β times as important as precision), which is dependent on the ratio of positive to negative test cases; we treat link positives equally to link negatives, hence our preference for MCC. Figure 6.6 includes the F_1 score due it its common use in causal discovery, but results are similar.

In Figure 6.6, we depict CaStLe performance results on a 2D VAR with ground-truth link density $d = \frac{4}{9}$. We show two extremes of sample size: a low-sample regime of $T = 10$, which is barely enough to identify the local dynamics of 9 cells, and a high-sample regime of $T = 150$. Our results are quite striking: in the low-sample regime, the CaStLed versions of each algorithm can accurately infer graph structure, with near-perfect performance on grids of size 10×10. By contrast, the performance of the non-CaStLed versions is essentially no better than random guessing, with only the sparse VAR able to exhibit any skill, and then only on small grids. In the high-sample regime, the CaStLed variants perform well on all grid sizes, with CaStLe-PC consistently achieving perfect recovery; the non-CaStLed variants perform better, as expected, but their performance still decays quickly as the spatial grid grows.

While the stronger performance of the CaStLed variants is noteworthy, the exhibited trends are even more important and highlight the true strength of the CaStLe approach: CaStLed approaches *improve* on larger grids while traditional approaches suffer. While Figure 6.6 shows results for the fixed link density $d = \frac{4}{9}$, we present results for all other link densities in K.

Having established CaStLe's strong performance on linear dynamics, we also validated its effectiveness on non-linear systems that more closely resemble realistic physical processes in Earth science. Specifically, we applied CaStLe to the advection-diffusion dynamics of Burgers' equation, a fundamental non-linear PDE that models a combination of advective and diffusive processes. Unlike our VAR benchmarks, which are discrete linear models with random initializations, Burgers' equation presents continuous non-linear dynamics that allow us to evaluate CaStLe's ability to recover spatial propagation patterns under controlled conditions. Our analysis demonstrates that CaStLe successfully identifies the underlying advection angle across a range of diffusion conditions, further supporting its applicability to complex space-time systems. This non-linear validation's complete methodology and results are presented in D.

6.9 Discussion

We have introduced CaStLe, a novel causal discovery meta-algorithm tailored for analyzing grid-level space-time data sets arising in Earth science. CaStLe can be directly applied to grid-level data and does not require pre-processing and spatial dimension reduction, allowing it to capture dynamics in the natural domain of the data rather than a derived (PCA-type) space. This distinction is crucial because global-scale phenomena across many complex systems—whether climate teleconnections, ecological patterns, or fluid dynamics—emerge from networks of local causal interactions that are often lost in dimensionality reduction approaches. While demonstrated with Earth science case studies, CaStLe is fundamentally domain-agnostic, applicable to any space-time system governed by local physical interactions, from fluid dynamics and heat transfer to biological pattern formation.

CaStLe can overcome the limitations of existing causal discovery approaches in Earth science's space-time data, filling a significant gap. By leveraging realistic assumptions of locality and homogeneity, CaStLe creates "spatial replicates" to substitute large observational domains for lengthy time series. This process transforms the spatial causal discovery problem from the high-dimensional (many variables, few observations) to the low-dimensional (few variables, many observations) regime, allowing accurate and efficient discovery of underlying causal dynamics. A key aspect of CaStLe is the causal *stencil* graph, a simplified representation of the local dynamics driving larger global behaviors. This notion of a stencil is particularly well-suited for systems able to be modeled by PDEs, as PDE-type dynamics inherently enforce both locality and homogeneity, as well as the sufficiency assumptions necessary for causal discovery to be *truly causal*.

We used these insights to identify the space-time evolution of volcanic aerosols that erupted from Mount Pinatubo in the HSW-V and E3SMv2-SPA models. We found that CaStLe found the expected path of advection in both models and more nuanced dynamics, including northward and southward dispersion, in E3SMv2-

SPA. We showed that CaStLe outperforms its peers in the causal discovery of synthetic benchmarks generated by vector autoregressive structural causal models. Additionally, as detailed in D, we found that CaStLe could accurately identify the advection angle in our Burgers' equation benchmark, demonstrating that it can filter out the "noise" of diffusion.

Our brief theoretical analysis of CaStLe in Section 6.6.4 and in B, demonstrates two regimes of consistent estimation for CaStLe, i.e., CaStLe recovers the true causal dynamics: long time series ($T \to \infty$) or large grid sizes ($N \to \infty$). This starkly contrasts existing approaches, whose performance rapidly deteriorates as $N \to \infty$. Several other important theoretical questions remain open, including the optimal relationship between sampling rates and grid resolution, behavior under mild violation of the key assumptions, and the correct target of inference for systems without clear advective dynamics (e.g., the chemical evolution of atmospheric aerosols).

We have focused on space-time data observed on regular 2D grids, but we believe that this assumption can be relaxed to adapt CaStLe for a broader range of observational structures. CaStLe can also be adapted to multivariate space-time data (more than one observation at each point) by including more comeasured variables in CaStLe's transformation of the region to the reduced coordinate space, enabling causal discovery of the space-time interactions of multiple species on the grid-level, which is a particularly exciting avenue of future research and application to Earth system dynamics. Developing data-driven methods for evaluating block sizes based on output robustness will enable more automatic application of CaStLe, requiring less subject matter expertise. Finally, causal representation learning is a nascent field combining the estimation power of machine learning with the strength of inference of causal discovery. Applying these techniques in CaStLe's parent-identification phase or for discovering spatial embeddings for regional block analysis is an exciting potential direction for future work.

Because our assumptions are readily satisfied by many physical systems, CaStLe can be applied quite broadly in the physical sciences. It may find value in any space-time system in which quantities at every point in space impact their adjacent spatial neighbors. In the Earth system, it may be of particular interest for studying forest fires, ocean dynamics, salt/fresh water incursions, and coastal erosion, for example. For atmospheric rivers, CaStLe could identify pathways of moisture transport and evolution; for wildfire spread, it could reveal causal relationships between local weather conditions and fire behavior; for drought propagation, it could track how soil moisture deficits spread across regions. CaStLe's preservation of local causal structures while efficiently handling high-dimensional data offers advantages over approaches requiring dimension reduction. For datasets where the temporal sampling is too coarse relative to the spatial resolution, extending to a radius-2 neighborhood might be appropriate while still maintaining our core assumption of locality. This extension would preserve the fundamental CaStLe methodology—only the dimensionality of the reduced coordinate space would increase. Additionally, CaStLe provides a promising framework for Earth system model evaluation (Nowack et al., 2020; Nichol et al., 2021a), potentially identifying where models produce correct outcomes through incorrect causal mechanisms.

While climate science typically studies large, long-term phenomena, the community increasingly recognizes the importance of understanding multi-scale interactions (Diffenbaugh et al., 2005; Palu, 2019; Agarwal et al., 2019; Zhang et al., 2022a). Teleconnections present an exciting challenge for future applications of CaStLe. These statistical dependencies between distant regions appear to violate locality but physically result from countless local interactions that are often unobserved or unmodeled. A two-stage methodology could be effective for tackling this challenge. First, apply CaStLe to discover local causal stencils, and then apply a complementary causal discovery technique to connect the discovered local processes across scales. This approach could bridge the gap between local and global causal discovery in climate science.

Complex space-time systems present apex challenges for causal discovery, combining chaotic dynamics, high dimensionality, noisy observational records, and complex underlying physical processes. CaStLe represents the first successful application of causal graph discovery to learn grid-cell-level causal structures in Earth systems. By preserving local causal structures while efficiently handling high-dimensional data, CaStLe presents a path toward connecting micro-scale interactions with macro-scale phenomena, potentially offering new insights into how global patterns emerge from local causal mechanisms. There are rich future research directions, including multivariate analysis and automated block size selection. The feasible discovery of local causal stencils presents an exciting new frontier for causal discovery of space-time data, particularly in the Earth sciences.

Open Research Section

The data generated and used for our HSW-V, VAR, and PDE experiments in Sections 6.7.1, 6.8.1, and D are available on Zenodo via https://doi.org/10.5281/zenodo.12701546 with GNU Lesser General Public License v3.0 or later (Nichol, 2024). The data used for the E3SMv2-SPA experiments in Section 6.7.2 can be found in Brown et al. (2024). The code for generating data, running experiments, and generating figures can be found here https://github.com/jjakenichol/CaStLe.

Acknowledgments

We thank Kara Peterson, the Deputy Principal Investigator of the CLDERA (CLimate impact: Determining Etiology thRough pAthways) project at Sandia National Laboratories (SNL), for helping to make this work possible. We also thank Joey Hart at SNL for helping with 2D Burgers' equation modeling and Tom Ehrmann at SNL for his help in understanding the Atmospheric dynamics we sought to capture. We thank everyone on CLDERA's simulation team, especially Benj Wagman, Hunter Brown, and Joe Hollowed, for developing the E3SMv2-SPA and HSW-V models, preparing the data, and providing their expertise. Finally, we thank the reviewers who devoted their time to helping us significantly improve the communication of this work.

This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multi-mission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC (NTESS), a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration (DOE/NNSA) under contract DE-NA0003525. This written work is authored by employees of NTESS. The employees, not NTESS, own the right, title, and interest in and to the written work and is responsible for its contents. Any subjective views or opinions that might be expressed in the written work do not necessarily represent the views of the U.S. Government. The publisher acknowledges that the U.S. Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this written work or allow others to do so, for U.S. Government purposes. The DOE will provide public access to results of federally sponsored research in accordance with the DOE Public Access Plan.

This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

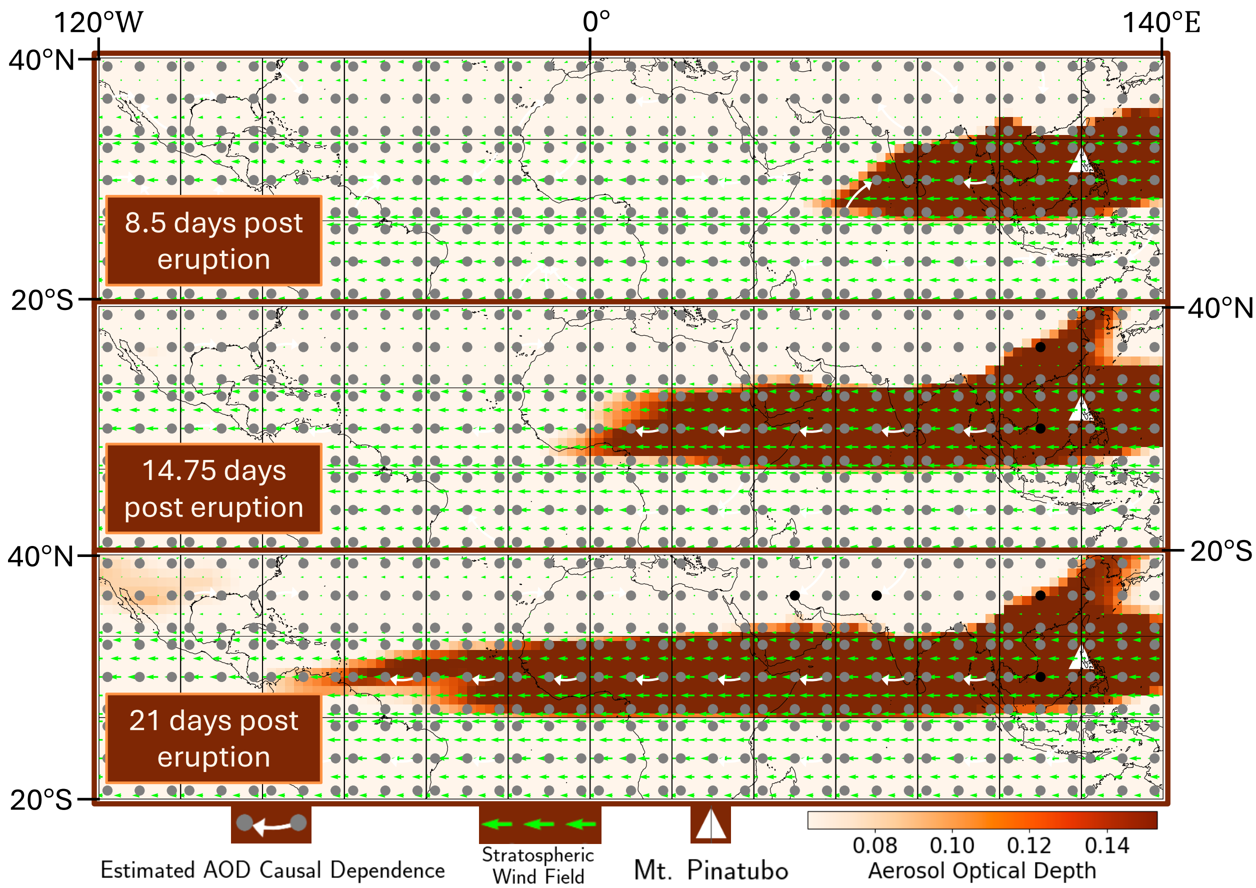
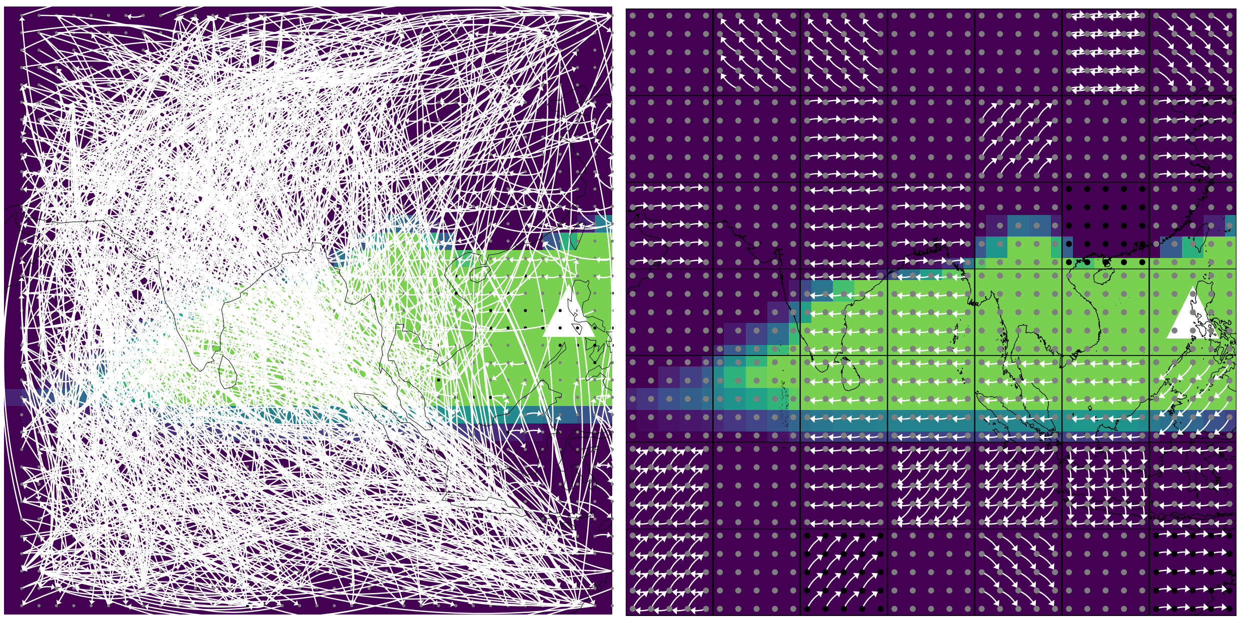


Figure 6.3: Application of CaStLe-PC-Stable to HSW-V simulation of the 1991 Mt. Pinatubo eruption. The stencils estimated by CaStLe (white) capture the underlying high-altitude wind fields (green) using only satellite-measured AOD, with near perfect accuracy in high aerosol regions (red-orange). Autodependencies are shown with black nodes where grid cells cause themselves, and gray nodes where there is no autodependence. All links represent a six hour time lag, the time resolution of the HSW-V dataset. On longer horizons (bottom row), CaStLe is able to recover equatorial wind currents as far away as South America, half-way around the world from Mt. Pinatubo (white triangle). CaStLe accurately identifies the prevailing westerly atmospheric winds because it was able to identify the space-time dependence between neighboring grid cells. Additional details are given in Section 6.7.



(a) PC algorithm results　　　　(b) CaStLe results

Figure 6.4: Causal maps inferred from the PC algorithm applied naively to all grid cells and CaStLe's equivalent results immediately to the west of Mt. Pinatubo; a 35×35 grid between $-20.00°$ to $50.00°$N and $55.00°$ to $125.00°$E in a 8.5 day span after the eruption. All links represent a six hour time lag, the time resolution of the HSW-V dataset. As expected, PC struggled with the high dimensionality and the discovered dependencies do not conform to the ground-truth understanding that aerosols advected towards the west. It also fails to identify local dynamics, instead drawing most connections over great distances. The PC analysis was computed in 729 minutes on 1,600 grid cells, while the CaStLe analysis was computed in 0.46 seconds.

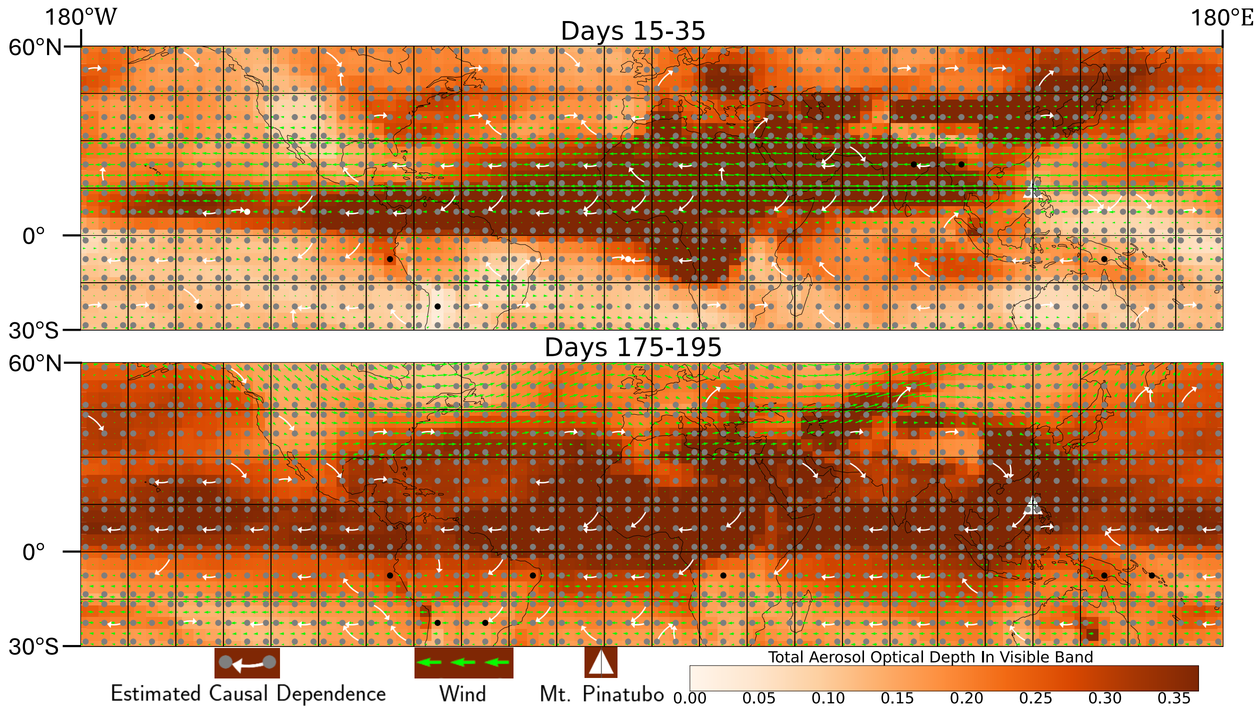


Figure 6.5: Application of CaStLe-PC-Stable to E3SMv2-SPA simulation of the 1991 Mt. Pinatubo eruption. The stencils estimated by CaStLe (white) capture the underlying high-altitude wind fields (green) using only total aerosol optical depth (AOD). Autodependencies are shown with black nodes where grid cells cause themselves, and gray nodes where there is no autodependence. All links represent a one day time lag, the time resolution of the E3SMv2-SPA dataset. The heatmap depicts AOD from any source at 50 hPa. The top panel depicts learning from the first 20 days after eruption, which began on day 15. The bottom panel depicts learning approx 6 months after the eruption over a 20-day time period. In the more challenging setting of the fully-coupled E3SMv2-SPA model, our results in the first weeks are still generally consistent with those in HSW-V presented in Section 6.7.1, showing that CaStLe is largely robust to greater complexity. In the bottom panel, the aerosols and winds are in a different regime. CaStLe stencils are still consistent in the tropics and now begin to recover dynamics pushing aerosols northwards above central Asia and southwards through western North America. A more complex model and smaller block sizes illustrate more nuanced dynamics, and there is more to learn from these, however, we leave deeper atmospheric dynamics analysis to future work.

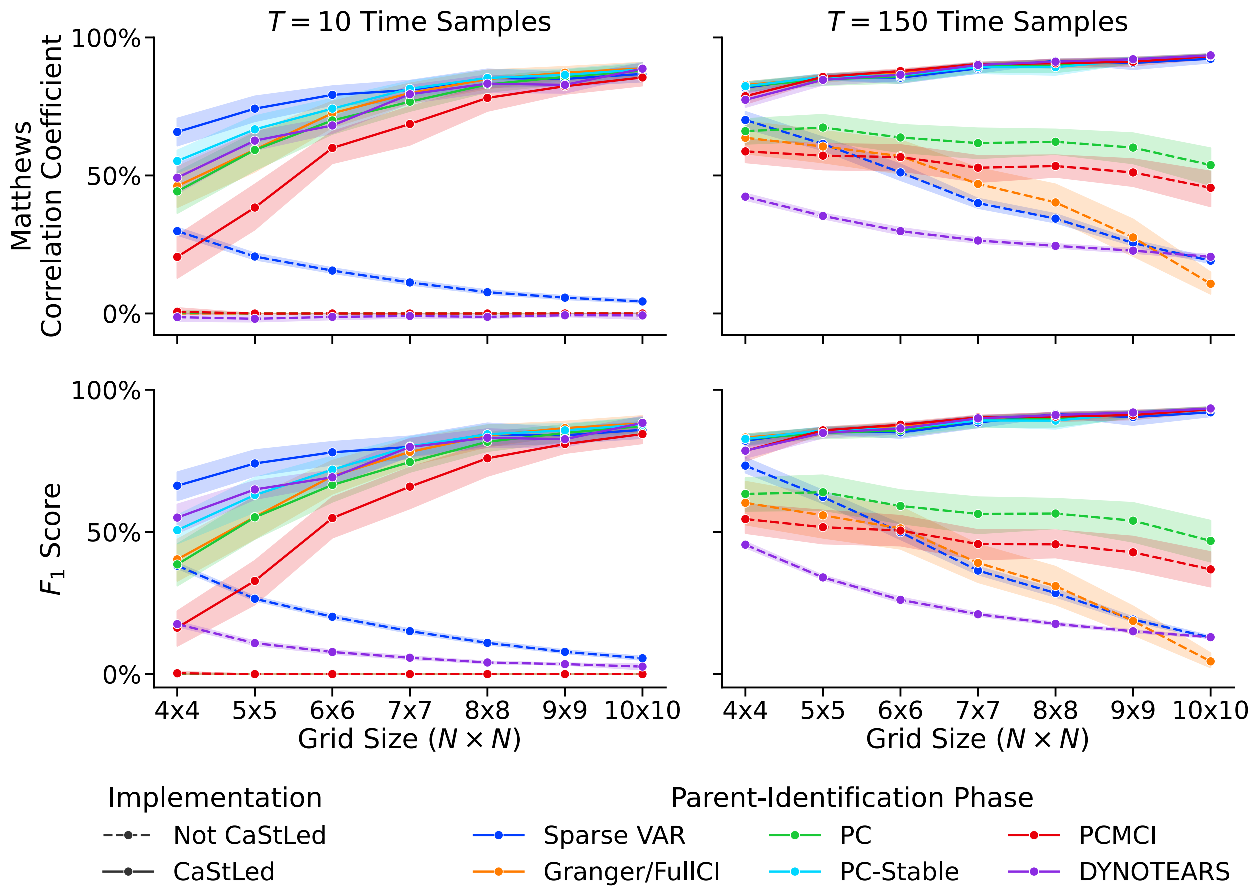


Figure 6.6: Comparison of CaStLed and non-CaStLed causal discovery approaches on linear-Gaussian dynamics, including Granger causality or FullCI (orange), PC (green), PCMCI (red), and DYNOTEARS (purple), as well as a statistical model of the data generating process (blue) presented with both MCC and F_1 metrics. In the low-sample size regime (T=10, left) CaStLed approaches can accurately recover the underlying causal graph, with performance increasing on larger grid sizes (solid lines); by contrast, non-CaStLed approaches are unable to perform better than mere chance (dashed lines). Even a model based on the underlying data generating process (Sparse VAR, blue) is significantly outperformed by its CaStLed counterpart. In the high-sample size regime (T=150, right), non-CaStLe approaches have improved performance but still compare unfavorably with their CaStLed counterparts.

7 M-CaStLe: Uncovering Local Causal Structures in Multivariate Space-Time Gridded Data

7.1 Publication Notes

Citation: Nichol, J. Jake, et al. "Space-Time Causal Discovery in Earth System Science: A Local Stencil Learning Approach." Submitted to the 32nd SIGKDD Conference on Knowledge Discovery and Data Mining, 2026.

Publication date: N/A

Conference: N/A

Formatting: The original text has been preserved as much as possible while still adhering to the formatting requirements of this dissertation.

Data and Software Availability: The paper is currently being prepared for submission and is not yet publicly available.

7.2 Abstract

Causal discovery tools propose to solve one of science's most important and challenging problems, the identification of underlying structure from observed phe-

nomena. Many systems prohibit the feasible or ethical application of more robust methods, such as randomized control trials. In particular, space-time systems, such as the Earth system or ecological systems, are attractive for causal discovery because they could suffer costly alterations if they are manipulated haphazardly. However, space-time systems are challenging to evaluate because their discretized representation as gridded space-time data is often very high-dimensional—possessing many more grid cells than temporal observations. The CaStLe meta-algorithm introduced by Nichol et al. (2025) proposed to solve that problem in scenarios satisfying their assumptions. However, it is limited to univariate analysis, identifying the space-time structure underlying a single quantity.

In this work, we present Multivariate Causal Space-Time Stencil Learning (M-CaStLe), a multivariate extension to CaStLe. We adapt the two phases of CaStLe to first collect the multiple variables in the repeating local neighborhood information in space-time gridded data, and second evaluate the causal parents of variables in the local neighborhood structure. M-CaStLe produces a multivariate causal stencil graph, which extends the CaStLe stencil to represent each variable at each location of the Moore neighborhood. We've added a decomposition method for interpreting the multivariate stencil in terms of just spatial dynamics or just inter-variable dynamics with the spatial graph and reaction graph, respectively. To evaluate M-CaStLe, we developed a multivariate space-time vector autoregression model (VAR) benchmark methodology. The multivariate space-time VARs provide data generation and ground truth causal stencils for direct evaluation of M-CaStLe.

Our experiments demonstrate that M-CaStLe achieves high precision across varying numbers of variables and grid sizes, indicating reliable identification of true positive links. However, recall decreases with an increasing number of variables, suggesting more complex systems have more challenging signals to identify. Further analysis shows that recall improves with stronger signal strengths, even in systems with up to 200 variables, indicating good performance in very high variable regimes. Comparisons with the PC algorithm reveal that M-CaStLe-PC consistently outperforms PC in high-dimensional settings, highlighting M-CaStLe's robustness in complex multivariate systems.

7.3 Introduction

Causal discovery is a set of causal inference tools for estimating the underlying structure in observed phenomena. While optimal causal estimation requires randomization, in many settings it is infeasible or unethical to apply (Runge et al., 2019b; Glymour et al., 2019). Thus, causal discovery for space-time systems is critical for scientific inquiry of complex emergent phenomena in physical systems because they often present challenges for randomization. For example, we have one Earth and randomly intervening in its systems is both prohibitively expensive and unethical due to unknown downstream effects. Likewise, neuroscience and ecology are prohibitive to random intervention.

Since the advent of Granger causality (Granger, 1969), the Rubin causal model (Rubin, 2019), causal graphs (Pearl et al., 2016), and the PC algorithm (Spirtes et al., 1993) (named for its authors, Peter and Clark), causal inference and causal discovery of observed data have developed into a rigorous mathematical framework. Today, causal discovery has become a rich literature with many algorithms and applications throughout the sciences (Glymour et al., 2019; Runge et al., 2023), including the health, Earth, and social sciences (Ebert-Uphoff and Deng, 2012; Cooper et al., 2015; Runge et al., 2019b; Nowack et al., 2020; Feder et al., 2022; Zanga et al., 2022; Sadeghi et al., 2023). Finally, causal representation learning is an exciting nascent field is developing that merges the flexibility and predictive power of machine learning with causal discovery techniques (Schölkopf et al., 2021).

This work presents a causal discovery approach for space-time systems with gridded data. Unlike space-time systems with point data, such as city-level data, gridded datasets generally enable the analysis of continuous effects over space, since they are regular and complete throughout the grid. However, such systems come with dimensionality challenges. Frequently, the number of grid cells scales faster than the number of temporal samples per grid cell (Runge et al., 2019b). Further challenging their analysis, such systems usually have multiple interacting variables per grid cell that are of scientific interest.

For example, in the Earth system, several interacting quantities may be measured over tens of thousands of grid cells, with hundreds of observations per variable in each grid cell. Atmospheric data often contains hundreds of thousands of grid cells, each with several orders of magnitude fewer observations in time. That imbalance is one aspect of the *curse of dimensionality* (Bellman, 1957; Bühlmann and Geer, 2011), where high dimensionality relative to sample size challenges conventional statistical methods and renders many forms of inference, including causal discovery, unreliable without dimensionality reduction.

Dimensionality reduction, such as principal component analysis (PCA) (Greenacre et al., 2022; Weylandt and Swiler, 2024), marginalizes large regions of grid cells into several one-dimensional time series. Each time series is then used for individual variables in the chosen causal discovery algorithm (Runge et al., 2015b). This procedure is effective for identifying large-scale patterns such as climate teleconnections (Tibau et al., 2022), but eliminates local grid-level interactions by construction. While large-scale patterns are important aspects of study in complex systems, the nature of their emergence is also important to understand. Local interactions determine the location and magnitude of larger patterns and other mid-scale phenomena, such as weather and seasonal patterns in atmospheric sciences.

Nichol et al. (2025) developed Causal Space-Time Stencil Learning (CaStLe), which is capable of grid-level causal discovery of high-dimensional space-time data. CaStLe can efficiently identify local causal relationships of a given quantity in space-time systems where traditional approaches fail. However, many scientific questions in complex space-time systems require analysis of multiple quantities per grid cell, such as temperature and soil moisture in Earth system monitoring of drought conditions (Sun et al., 2021) or infection dynamics in epidemiological modeling using infection severity, duration of infection, and population age (Ganesan and Subramani, 2021; Paul et al., 2021).

In this work, we propose an extension to the CaStLe meta-algorithm enabling multivariate space-time causal discovery of grid-level data. We show that Multivariate Causal Space-Time Stencil Learning (M-CaStLe) can effectively capture the causal relationships in multivariate space-time systems. Our results demonstrate that M-CaStLe is capable of accurately estimating local multivariate space-time structures from gridded data, outperforming the PC algorithm, especially in high-dimensional settings. This suggests that M-CaStLe is a robust tool for causal discovery in complex multivariate systems, providing valuable insights into the underlying dynamics of such systems.

7.3.1 Background and Motivation

CaStLe is a meta-algorithm for causal discovery in high-dimensional space-time systems. By leveraging local causal regularities, CaStLe transforms the causal discovery problem from a high-dimensional space with many variables and limited observations to a low-dimensional embedding with fewer variables and more abundant observations. This transformation enhances the efficiency and accuracy of causal discovery, facilitating the identification of causal relationships in their natural context. The present work extends of CaStLe, aiming to broaden its applicability to multivariate space-time dynamics, making it a versatile tool for analyzing various space-time systems in the physical sciences.

7.3.2 Foundations of the CaStLe Framework

In many natural and engineered systems, complex global behaviors emerge from simple local interactions that follow consistent physical dynamics. Nichol et al. (2025) called such systems *partial differential equation (PDE)-like* because they exhibit consistent dynamics defined by interactions between adjacent points in space, with smooth transitions between dynamical boundaries and equilibria. These are characterized by a set of fundamental assumptions that constrain their dynamics:

T1) Temporal Locality: for any $\tau \neq 1$, $X_{i,t-\tau} \not\to X_{j,t}$ for any spatial coordinates (i, j)

T2) Temporal Causal Stationarity: the dynamics governing the evolution of $\boldsymbol{X}_t$ do not change over time. That is, $X_{i,t-1} \to X_{j,t} \Leftrightarrow X_{i,t-1+\tau} \to X_{j,t+\tau}$ for any time offset τ.

S1) Spatial Locality: if (i, j) are not neighbors (in a problem-specific sense) then $X_{i,t_1} \not\to X_{j,t_2}$ for any t_1, t_2.

S2) Spatial Causal Stationarity: the dynamics governing the evolution of $\boldsymbol{X}_t$ do not change over space. That is, $X_{i,t-1} \to X_{j,t} \Leftrightarrow X_{i+s,t-1} \to X_{j+s,t}$ for any spatial offset s.

Here, $\not\to$ denotes the absence of a direct causal relationship between two variables. Nichol et al. (2025, Appendix A) describes these assumptions in detail, including ways they may be violated and their Appendix I demonstrates some examples of their violations and CaStLe's robustness to them. To apply causal discovery, the causal assumptions *the causal Markov condition* and *faithfulness* (Spirtes et al., 1993) must be additionally assumed. Because of the locality assumptions, the commonly required *causal sufficiency* assumption may be relaxed (Nichol et al., 2025).

Such systems exhibit both temporal and spatial locality. Temporal locality (T1) dictates that state transitions depend only on the immediate past, preventing "backward causation" and respecting the arrow of time. Spatial locality (S1) ensures that interactions occur only between proximate elements, eliminating action at a distance.

The governing dynamics in these systems demonstrate invariance across both time and space. Temporal causal stationarity (T2) means the rules of evolution remain constant throughout the analysis period—the same causes produce the same effects regardless of when they occur. Spatial causal stationarity (S2) implies that these rules apply uniformly across the domain—the physical location of an element does not alter how it responds to its neighbors. While many macro-scale spaces contain multiple sets of equilibrium dynamics, there are typically micro-scale regions containing stationary spatial causality.

These systems can be represented through structural causal models (SCM) of the form:

$$X_{i,t} = f_i(X_{\mathcal{N}(i),t-1}, \eta_{i,t}) \tag{7.1}$$

Where $X_{\mathcal{N}(i),t-1}$ represents the states of elements in the neighborhood of i at the previous time step, and $\eta_{i,t}$ captures stochastic innovations. Under spatial causal stationarity, the functional form f_i is identical for all i, reducing to a single function f that applies throughout the domain. In short, this space-time SCM implies grid cells exhibit Granger-causal dynamics, which imply that each grid cell's temporal information content encodes the past-history of itself and its immediate neighbors.

This framework encompasses numerous well-studied systems including those governed by partial differential equations, cellular automata, and various lattice models in statistical physics. The approach provides a powerful foundation for both forward simulation and inverse problems—identifying the underlying causal structure from observed spatiotemporal data.

CaStLe not only seeks to identify local causal dynamics but also to do so for high-dimensional systems. In some cases, it may be enough to apply causal discovery independently to small groups of local grid cells; however, in many systems of study, more grid cells are present than observations within each. To accomplish discovery in this regime, we need to efficiently use all the dynamical information in a system.

CaStLe leverages the inherent locality and stationarity to collect time series representing the space-time replicates in such systems. Every grid cell's time series encodes the causal influence of its neighbors, and they can be used as informative replicates of the system's local dynamics. CaStLe processes a set of grid cells, collecting each one's data on its local dependence, then learns the causal structure of the grid cells and their neighborhoods.

CaStLe's first phase is to form the Locally Encoded Neighborhood Structure (LENS), an embedding representing the Moore neighborhood–a 3×3 matrix of a grid cell and its eight immediate neighbors. The LENS contains concatenated time series from each grid cell's Moore neighborhood so that the local dynamics from each neighbor is repeated. The embedding is a 3×3 matrix, with each entry representing the North West, North, North East, West, center, East, South West, South, and South East grid positions of the Moore neighborhood. Each entry of the embedding contains long concatenated time series collected from throughout the original grid space. Each time series is of length $T \times (N-2)^2$, for the grid's dimension N and T time samples per grid cell. The embedding does not marginalize any data, so no information loss occurs, as would happen during other dimensionality reduction techniques. Figure 7.1 is a conceptual diagram depicting using the local Moore neighborhood to construct the LENS.

Once the embedding is constructed, CaStLe's second phase, the Parent-Identification Phase (PIP) applies an adapted causal discovery algorithm to the embedding. Any time series causal discovery algorithm may be adapted by requiring it to treat the embedding's center grid cell as special: it may be the only child in the resulting causal graph; parents are unrestricted. This adaptation has multiple effects: it creates a graph of the generalized ancestry for each grid cell, eliminates would-be

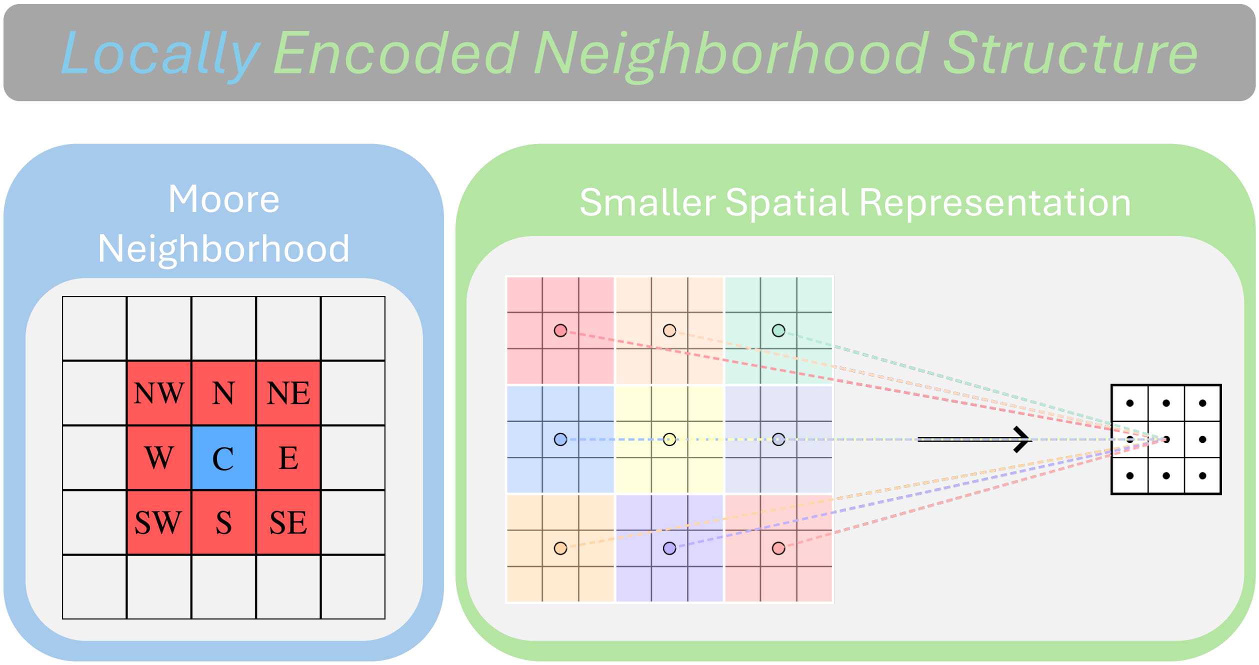


Figure 7.1: A conceptual diagram of the LENS that CaStLe constructs for learning underlying local causal dynamics in gridded data. This encoding transforms the original grid space into a local neighborhood structure without marginalization, preserving all of the local relationships in the gridded time series data.

unobserved confounding between the embedding's outer grid cells and their neighbors beyond the embedding, and increases computational and statistical efficiency, which is detailed below. The result of the PIP on the embedding is the *causal stencil graph*, a representation of the local causal dynamics between all grid cells in the system.

7.3.3 Theoretical Properties and Empirical Validation of CaStLe

Nichol et al. (2025) showed that CaStLe exhibits significant performance and efficiency improvements for grid-level causal discovery. It successfully reconstructed known volcanic aerosol dynamics, driven by stratospheric winds, in the weeks after the Mount Pinatubo eruption of 1991. We demonstrated its general performance

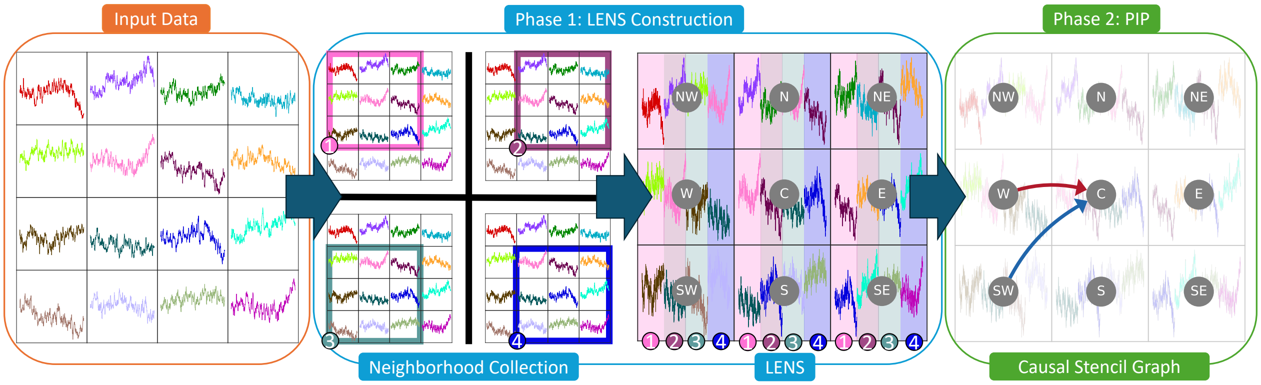


Figure 7.2: A demonstration of the full CaStLe process to produce a causal stencil graph on an example input 4×4 gridded space-time system. In the LENS phase, neighborhood information is collected from each of the interior grid cells, which are then concatenated to form the LENS. Finally, the PIP phase applies an adapted time series causal discovery algorithm to learn the space-time parents of the center node. The learned stencil depicts the underlying space-time structure of each grid cell in the original data.

on advective-diffusive dynamics with a Burgers' equation simulation study. We compared it to existing causal discovery algorithms with ground truth defined by space-time vector autoregression model (VAR) models.

Because CaStLe constructs the LENS, a lower spatial-dimension embedding, and the PIP limits potential causal children to only the center node, the number of variables and possible links are both fixed to nine. That property enables much more efficient causal discovery. Computational complexity is a measurement of the asymptotic bounding on how many computational resources are required for increasingly large input sizes. The PC algorithm has a computational complexity bounded by $\mathcal{O}(Tp^3 2^p)$, when applied to an p grid cells, with T time samples per cell. We showed that CaStLe is bounded by $\mathcal{O}(Tp)$.

CaStLe also exhibits improved sample complexity, which measures the asymptotic bounds on how many samples are required to ensure correct graph estimation. The probability of the PC algorithm incorrectly estimating the true graph is bounded by $\approx \mathcal{O}(p^p)$. In contrast, we find that CaStLe's error probability scales as $\approx \mathcal{O}\left(\frac{pT}{e^{pT}}\right)$. From this, as the grid size grows larger, we find that PC is less likely to estimate the correct causal graph, while CaStLe is more likely to estimate the correct graph.

Nichol et al. (2025) also demonstrated several empirical results of CaStLe with benchmarks and realistic climate model output studies. It was shown that CaStLe can robustly capture the transport patterns of volcanic aerosols emitted by the 1991 Mount Pinatubo eruption. It outperformed the PC algorithm in terms of accuracy and execution time, largely because PC naively sought causal relationships between all grid cells without the benefits of the LENS. CaStLe was also robust to moderate assumption violations. The VAR benchmark study compared CaStLe to popular time series causal discovery methods, including the PC algorithm (Spirtes and Glymour, 1991), PCMCI (Runge et al., 2019a), and DYNOTEARS (Pamfil et al., 2020). They found that CaStLe variants performed well, with better results on larger grids, while non-CaStLe algorithms struggled and performed more poorly on larger grids. The Burgers' equation study evaluated CaStLe's performance in different advection speed and diffusivity regimes via advection-diffusion partial differential equation (PDE) model output. CaStLe performed well except in settings where diffusion dominated, making advection signals unrecoverable.

7.3.4 Research Gap and Motivation for Multivariate Extension

Nichol et al. (2025) showed that CaStLe can reconstruct the local space-time causal structure between grid cells of one quantity, e.g., atmospheric aerosols. While helpful in understanding the underlying dynamics of a species transporting or propagating in a complex environment, it leaves learning impacts of that transport to later inference and analysis. Such a manual or post-hoc multivariate inference becomes complex as the number of variables increases.

For example, in Nichol et al. (2025), CaStLe identified the space-time evolution of volcanic aerosols in the stratosphere from the Mt. Pinatubo eruption. Given the rich literature of that eruption, we know that the volcano's SO_2 output increased stratospheric temperatures and decreased tropospheric temperatures for two-to-five years (Dutton and Christy, 1992; Labitzke and McCormick, 1992; Parker et al., 1996a; Soden et al., 2002). The eruption's SO_2 did not directly impact temperature, the plume of gas underwent chemical and physical evolutions, forming H_2SO_4 and advecting and diffusing around the globe. However, univariate CaStLe needs to analyze each chemical species separately and cannot determine interactions between species.

To estimate the space-time dynamics of each variable separately and then infer variable interactions afterward potentially introduces errors and does not have the benefit of joint estimation, which is available in time series causal discovery, such as PCMCI (Runge et al., 2019a). Furthermore, learning space-time causal structures from each variable independently may miss cross-variable confounding, leading to space-time estimation errors and incorrect inference of the underlying physical process.

Joint estimation of space-time dynamics and variable interactions can enable more complex analyses. For example, SO_2 follows a chemical and physical causal pathway to mediate temperature. SO_2 reacts with water molecules to become H_2SO_4. Finally, H_2SO_4 interacts with incoming solar radiation, which impacts temperatures. Understanding the local space-time dynamics of these aerosol species as they transport around the globe may help explain local temperature impacts. Domains outside of atmospheric chemistry and Earth systems science where estimating grid-level multivariate interactions in space-time systems (MacEachren et al., 1999; Haas, 2002) would be valuable are computational fluid dynamics (Wimer et al., 2023), spatiotemporal pharmacokinetics (Guarin et al., 2021; Klingelhuber et al., 2024), and computational chemistry (Higham, 2008; Owen et al., 2024).

Multivariate interactions are challenging to estimate at the grid-level because the high-dimensionality of datasets observed from space-time systems becomes more challenging with more variables because each variable entails p more grid cells to estimate for the same T observations per grid cell per variable. CaStLe solves the high-dimensional challenge in many univariate space-time systems. Extending its capabilities to discover variable interactions simultaneously with space-time dynamics for each variable enables robust discovery of how they interact in space and time. Doing so while maintaining the interpretability of the graphs at

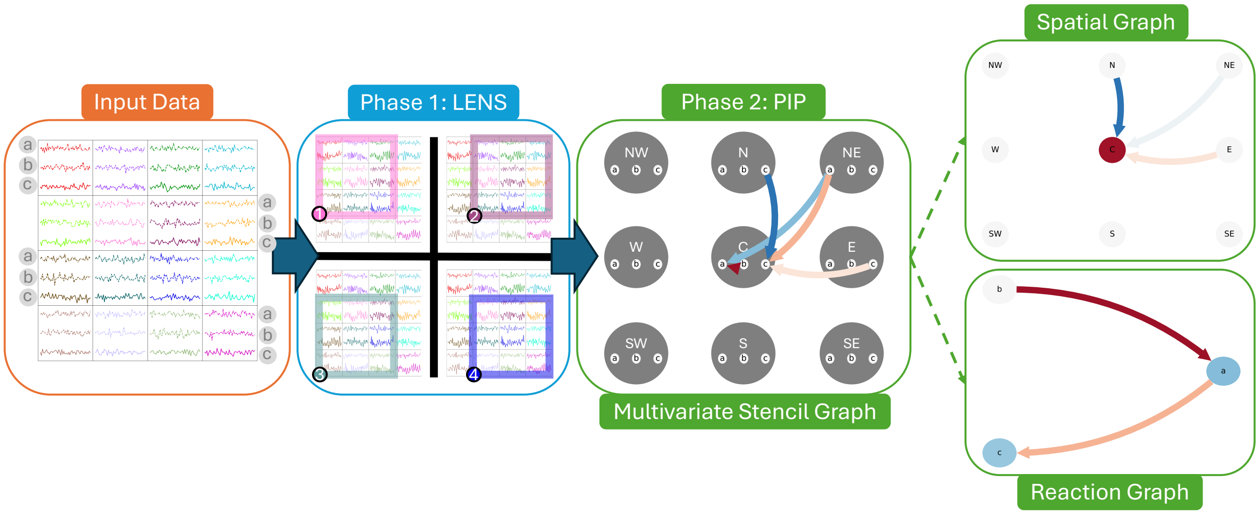


Figure 7.3: A schematic diagram of the input, computational phases, and output of M-CaStLe. Similar to CaStLe's procedure (c.f. Figure 7.2), the first phase collects local neighborhood information into the LENS, which now collects information for each variable's time series in each grid cell. The second phase applies the PIP to every variable at every position in the LENS to determine which variables cause the center variables from each location in the LENS. Finally, the resulting multivariate stencil graph can be decomposed into the *spatial graph* and *reaction graph* for improved interpretability and potential analysis.

scale is also challenging. Multivariate stencil graphs need to contain many more nodes for each variable and still describe local structures.

7.3.5 Contributions

M-CaStLe solves these challenges by adapting both phases of the original CaStLe meta-algorithm. The first phase, which restructures the given gridded data into the LENS, is adapted to restructure multivariate data to preserve space-time and inter-variable relationships. The univariate PIP sought causal relationships terminating in only one node (the center). The multivariate PIP is adapted to find parents and children of multiple sets of nodes for each variable.

These advances enable the simultaneous estimation of space-time dynamics and inter-variable interactions, providing a more comprehensive understanding of complex systems. This capability is particularly valuable in fields such as atmospheric science, where understanding the interplay between different chemical species and their impact on climate is crucial. We validated M-CaStLe through extensive experiments on synthetic benchmarks. Our results demonstrate that M-CaStLe outperforms existing methods in terms of accuracy and computational efficiency, particularly in high-dimensional settings. The empirical validation shows that M-CaStLe can robustly capture the causal structure of multivariate space-time systems, making it a powerful tool for scientific discovery and analysis.

7.3.6 Paper Organization

The remainder of this paper is organized as follows: in Section 7.4 we introduce M-CaStLe, our multivariate extension to CaStLe; Section 7.5.1 discusses our benchmark's experimental setup with VARs; Section 7.5.1 presents a rigorous analysis of the multivariate results of M-CaStLe benchmarked on multivariate models of space-time dynamics; and finally we discuss the presented work and future directions in Section 7.6.

7.4 Methods

Multivariate CaStLe (M-CaStLe) extends CaStLe's capabilities to discover local space-time causal structures in multivariate data. M-CaStLe produces the multivariate causal space-time stencil graph, which describes how a *set of variables* interact within their Moore neighborhood over time. The multivariate stencil is often challenging to interpret immediately due to its complexity. To improve interpretability, we present the *reaction graph* and the *spatial graph*, which decompose from the multivariate stencil output of M-CaStLe into a graph of inter-variable relationships (without a spatial aspect) and a graph of spatial relationships (without variable relationships).

M-CaStLe adapts both phases of the CaStLe meta-algorithm to enable construction of a LENS containing multiple variables and successful causal discovery of space-time and inter-variable dependencies within the LENS. Input data consists of V variables measured on an $N \times N$ grid over T time steps, yielding a tensor $\mathbf{X} \in \mathbb{R}^{N \times N \times V \times T}$. Figure 7.3 depicts each step of M-CaStLe. In this example, we illustrate a simple 4×4 original grid space, G, which has $V = 3$ locally interacting variables, a, b, and c, with $T = 500$ time samples.

7.4.1 Phase 1: The Locally Encoded Neighborhood Structure (LENS)

Phase 1 collects neighborhoods in the same fashion as the univariate CaStLe, but it now collects multiple time series per spatial location in the Moore neighborhood for each variable. The univariate LENS is a 3×3 matrix where each element contains one time series of length $T \times (N-2)^2$. Since M-CaStLe has V variables, the multivariate LENS is a 3×3 matrix where each element contains V time series of length $T \times (N-2)^2$. In short, it is a tensor in $\mathbb{R}^{3 \times 3 \times V \times L}$, where $L = T \times (N-2)^2$ is the length of each concatenated time series. In Figure 7.3, Phase 1 depicts the process the LENS construction follows to collect time series from each Moore neighborhood as the window slides across G. It collects all three variables from each grid cell within the neighborhood window and concatenates them to the LENS, according to their position relative to the center of the neighborhood window and the respective variable in each position. Like the univariate LENS, there is no marginalization or loss of data, and its structure allows it to be fully invertible. We do not have a reason to invert the procedure in this analysis, but it illustrates that no information loss occurs.

7.4.2 Phase 2: The Parent-Identification Phase (PIP)

In univariate CaStLe, the PIP adapts a given time series causal discovery algorithm, such as DYNOTEARS (Pamfil et al., 2020), to seek the parents of only the center node in the LENS. To adapt this approach to M-CaStLe, we do the same for each variable in the center node. Rather than allowing one child in the discovery process, we now allow V children. This has the effect of every variable in every position in the LENS having a potential causal effect on every variable in the center position. Resulting is a multivariate stencil, such as the one depicted in the third panel of Figure 7.3. This example illustrates a stencil of three variables with dependencies between each over space and time.

7.4.3 Interpretability: Decomposing the Multivariate Stencil

While the stencil in Figure 7.3 may be interpretable after careful viewing, multivariate stencils of more variables or with more dependencies can be challenging to parse visually. For that reason, we have developed a decomposition scheme to analyze the variable interactions and the spatial structure of all variables separately. The far right of Figure 7.3 illustrates the spatial graph and reaction graph corresponding to the stencil to their left.

Computing the stencil decomposition is straightforward and similar for both the spatial and reaction graphs. To compute the spatial graph, the stencil links are aggregated along the variable dimension, and the location from which they originate is preserved. For example, in Figure 7.3, two links are coming from the NE position to the center, a negative dependence (light blue) via $a \rightarrow a$ and a positive dependence (orange) via $a \rightarrow c$, and both of those are aggregated to find one weakly negative link NE $\rightarrow$ C in the spatial graph. Note that there is a $b_{center} \rightarrow a_{center}$ link in the stencil and that it is represented as an autodependence link in the spatial graph, illustrated by the center node's coloring. The node and link colors directly associate with the continuous link dependence strength output by M-CaStLe, where -1 represents the strongest negative dependence (dark blue) and $+1$ represents the strongest positive dependence (red).

The reaction graph is computed by aggregating stencil links along the spatial dimension while preserving the variable dimension. For example, in Figure 7.3, there are two links $c \to c$ in both the N and E locations, strongly negative (blue) from the N and weakly positive (red) from the E. Those are aggregated to form the light-blue c node in the reaction graph. Resulting is a graph of variables that represents the aggregate strengths of dependencies from any direction.

To aggregate the stencil link coefficients, we use Fisher's z-transformation. It stabilizes the variance of the correlation coefficients, making them more suitable for averaging. The process involves converting each coefficient into a z-score, computing the arithmetic mean of the z-scores, and then converting the average z-score back to a correlation coefficient using the inverse Fisher's z-transformation. This method ensures that the combined value accurately reflects the underlying dependencies between variables.

7.5 Experiments

We present three sets of experiments to demonstrate M-CaStLe's capabilities. They progress from an idealized synthetic benchmark with analytically known ground truth, where performance can be evaluated exactly, to a PDE study with derived ground truth to a study of atmospheric chemistry from a complex, fully-coupled Earth system model. The idealized benchmark uses randomly generated VARs to where assumptions are perfectly satisfied. The PDE study uses an advection-diffusion-reaction (ADR) model to simulate different advection and diffusion settings in a more realistic, continuous setting. The atmospheric chemistry study uses the Energy Exascale Earth System Model version 2 with Stratospheric Prognostic Aerosols (E3SMv2-SPA) (Brown et al., 2024) to demonstrate reconstructing chemical pathways in a real-world application setting.

7.5.1 VAR Dynamics Benchmarking

We developed random and stable multivariate space-time systems with two spatial dimensions. They mathematically map to ground truth causal stencil graphs for evaluating the performance of M-CaStLe with a variety of system parameters.

Our methodology for generating data builds upon the work used by Nichol et al. (2025), which is fully detailed by Nichol et al. (2023). They developed a procedure for generating benchmark datasets of stable 2D space-time systems through the systematic construction of coefficient matrices parameterizing VARs of order 1 (VAR(1)s). Causal graphs have a direct mapping from VARs (Peters et al., 2017; Runge et al., 2019a), which enables precise benchmark comparisons between VAR modeled data and causal discovery estimated graphs.

A system on an $N \times M$ grid with T time samples, $\boldsymbol{X} \in \mathbb{R}^{N \times M \times T}$ with elements $X_{i,j,t}$, can be modeled by a VAR(1) with

$$\boldsymbol{X}_t = \boldsymbol{A}\boldsymbol{X}_{t-1} + \boldsymbol{\eta}_t, \tag{7.2}$$

where $\boldsymbol{A}$ is the coefficient matrix encoding linear dependencies between all variables in the system and $\boldsymbol{\eta}$ represents independent *innovations* on $\boldsymbol{X}$ for each variable at each time step. In this case, innovations are modeled with a unit normal distribution.

The space-time VAR methodology initializes a 3$\times$3 matrix defining local grid-level dynamics between neighbors, called the neighborhood dependence matrix (NDM). Random NDMs of predetermined sparsity, d, are generated to describe how every grid cell in the space is dependent on the grid cells in its Moore neighborhood. To simulate an entire grid, the NDM can be structurally mapped to an $\boldsymbol{A}$ matrix for the entire grid. For an $N \times M$ grid space, $\boldsymbol{A} \in \mathbb{R}^{NM \times NM}$. Finally, most 2D VARs are not numerically stable. To ensure stability, $\rho(\boldsymbol{A}) < 1.0$, where $\rho(\boldsymbol{A})$ is the spectral radius of $\boldsymbol{A}$ (Strang, 2016, p.307). Through the NDM definition, VARs can simulate locality in physical systems.

Experimental Setup: Multivariate Space-Time VARs

To adapt the space-time VAR procedure for multivariate systems, we grow the NDM in a new variable dimension, which gets mapped to a larger, flat, $\boldsymbol{A}$ matrix. The multivariate NDM describes interactions between multiple variables at the local grid-level, enabling VAR modeling of multivariate space-time dynamics.

For a system of $\boldsymbol{V}$ variables, the multivariate dynamics are represented by set of $V \times V$ 3$\times$3 matrices. Each 3$\times$3 matrix corresponds to the space-time dependence structure of a particular pair of parent and child variables. Like the univariate NDM, each entry in each 3$\times$3 matrix is a coefficient value representing the influence of the entry's spatial location in the Moore neighborhood on the center location.

The NDM is mapped to an $\boldsymbol{A}$ matrix, which represents the interactions of every grid cell-variable on every other grid cell-variable. For a grid of size $N \times M$ spatial dimensions and V variables, the matrix $\boldsymbol{A} \in \mathbb{R}^{NMV \times NMV}$. With the computed $\boldsymbol{A}$ matrix, we again enforce stability by ensuring $\rho(\boldsymbol{A}) < 1.0$, where $\rho(\boldsymbol{A})$ is the spectral radius of $\boldsymbol{A}$.

With a stable $\boldsymbol{A}$ matrix, experimental data can be generated for any number of grid cells, time samples, local dependencies, and variables. Although $\boldsymbol{A}$ is larger, the VARs still have the form of Equation 7.2. Since most $\boldsymbol{A}$ matrices will be unstable, our implementation uses an accept-reject scheme similar to the univariate approach of Nichol et al. (2025) to generate stable $\boldsymbol{A}$ matrices:

1. Generate a random set of 3×3 *local dynamics matrices*, $\{C_{ij}\}$, for each pair of child and parent variables, resulting in $V \times V$ matrices. Each C_{ij} has d non-zero elements, including the central element (autocorrelation), where $1 \leq d \leq 9$. Each of the d non-zero elements, $\{a_i\}_{i=1}^{d}$, have a random value $1.0 \geq \text{coefficient}_i \geq s_*$.

2. Expand $\{C_{ij}\}$ to form the matrix $\boldsymbol{A}$ for a grid of size $N \times M$, resulting in $\boldsymbol{A} \in \mathbb{R}^{NMV \times NMV}$.

3. If $|\lambda_{\max}(\boldsymbol{A})| \geq 1$, scale $\boldsymbol{A}$ by $|\lambda_{\max}(\boldsymbol{A})|$.

4. If $c < s_* \; \forall c \in \boldsymbol{A}$, reject, else accept.

where $|\lambda_{\max}(\boldsymbol{A})|$ is the maximum absolute eigenvalue of $\boldsymbol{A}$. This is used to sample from the set of statistically stationary & spatially homogeneous VARs on a 2D grid with minimum signal strengths $s_* \geq 0.1$ and fixed sparsity levels in the range $d \in \{1, 2, \ldots, 9\}$.

Metrics

Since VARs map directly to ground truth causal graphs, we measured M-CaStLe's performance using binary classification measures. Let $G = (V, E)$ be the ground truth graph where V is the set of nodes and $E \subseteq V \times V$ is the set of edges. For any node pair $(i, j) \in V \times V$, a positive instance is defined as $(i, j) \in E$ and a negative instance as $(i, j) \notin E$. This enables our usage of precision, recall, and F_1 score, defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7.3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7.4}$$

$$\text{F}_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7.5}$$

where TP, FP, TN, and FN denote true positives, false positives, true negatives, and false negatives, respectively. Put simply, precision is the proportion of correctly detected positives to all detected positives, with a range of $[0, 1]$, where 1 is a perfect precision; recall is the proportion of correctly detected positives to how many positives should have been detected, with a range of $[0, 1]$, where 1 is a perfect recall; and the F_1 score is the harmonic mean of precision and recall, with a range of $[0, 1]$, where 1 indicates perfect graph estimation.

Data Generation

We used the following data generation parameter ranges with 30 replicates each:

- Time samples $T = 1000$
- $N \times N$ grid sizes where $N \in [4,5,6,7,8,9,10]$
- Number of variables $V \in [1,2,3,4,5,6]$
- Density $d \in (0,\dots 0.5]$
- Coefficients $c \in [0.1, 1.0]$

where density is relative to the stencil graph density: $d = \frac{L}{(3\times 3\times V^2)}$ with L links, such that $d \leq 1$. Since a $V = 1$ system can have up to $L = 9$, the most allowable here are $L = 4$. A $V = 6$ system may have $L \in [1,\dots 162]$. However, not all densities produced 30 stable systems after 48 hours of the accept-reject scheme described in Section 7.5.1. It is clear that there are zero systems in the limit of increasing density with a given minimum coefficient size. Appendix M details which of the above combinations successfully produced 30 systems for analysis. In total, 56,283 experiments were generated, with more experiments for systems of more variables.

Benchmark Results

We present empirical results of M-CaStLe's performance on our VAR benchmarks varying: the number of variables, grid sizes, the number of graph dependencies

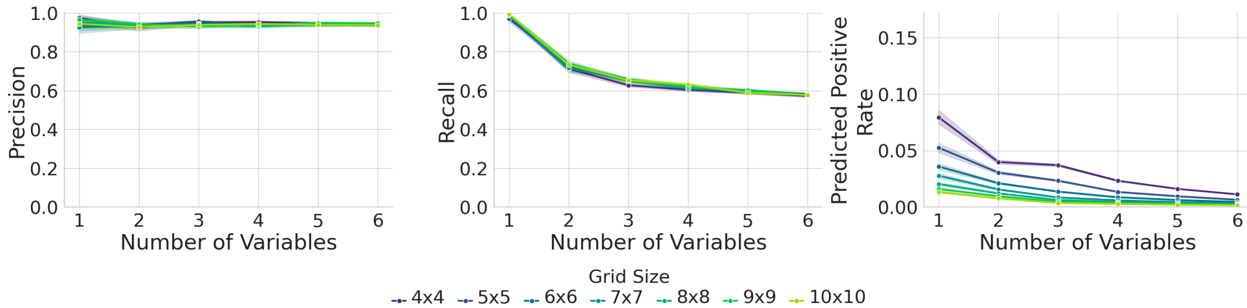


Figure 7.4: Showing precision and recall alongside predicted positive rate, a measure of how often a positive is predicted among all other predictions. As variables increase, the predicted positive rate decreases, which diminishes recall.

(graph edges), and the magnitude of coefficients. These demonstrate that M-CaStLe is suitable for estimation of local multivariate space-time structures from gridded data. Additionally, we compare M-CaStLe's performance to the popular PC algorithm for causal discovery.

Multivariate Performance Figure 7.4 illustrates precision, recall, and positive prediction rate (PPR) in our experiments as the number of variables increases, with individual lines for each of the grid sizes. All available densities are marginalized in each line, with 95% confidence intervals. We found that precision is very high in all cases, regardless of the number of variables or grid size, with an average value of ≈ 0.94. This indicates that when M-CaStLe identifies a positive link, it is likely to be a true positive. We found that recall is very high for $V = 1$ and decreases as the number of variables increase, with the mean value ≈ 0.62. This indicates that M-CaStLe may be relatively conservative, identifying a little more than half of the true links in the systems with more variables. However, it may also indicate limitations of the synthetic data model.

We considered PPR to better understand what is limiting recall performance. PPR is the fraction of all possible connections that were predicted as positive, regardless of correctness, given by

$$\mathrm{PPR} = \frac{\mathrm{TP}+\mathrm{FP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}, \tag{7.6}$$

with a range of $[0,1]$, where 1 indicates all possible edges were estimated (a fully dense graph). No particular PPR value necessarily indicates good performance, because it is a measure of the estimated graph's density.

In Figure 7.4, we see that recall and PPR are both decreasing as graph size increases (larger grid size and more variables). This possibly indicates that as the graphs are getting larger, signals are more challenging to detect. We investigated the data generation model's apparent limitations in Appendix M. Figure M1 demonstrates that fewer stable systems could be generated for larger graphs, relative to their potential. Figure M2 demonstrates that as the number of links increases among all systems, the maximum and minimum coefficients in each system quickly decrease. This indicates that the systems may be more challenging to correctly estimate, suggesting that M-CaStLe's recall may be more reflective of the data generating model than being a conservative estimator.

Comparison to the PC Algorithm Nichol et al. (2025) compared CaStLe to several prior causal discovery methods and found CaStLe outperformed the others, particularly has the data dimensionality increased. In the multivariate regime, the data's dimensionality is multiplied by the number of variables. Multivariate systems should be far more challenging for causal discovery without dimensionality reduction. Here, we compare M-CaStLe to the PC algorithm, which is still in popular (Glymour et al., 2019) use and is the predecessor to most constraint-based causal discovery algorithms.

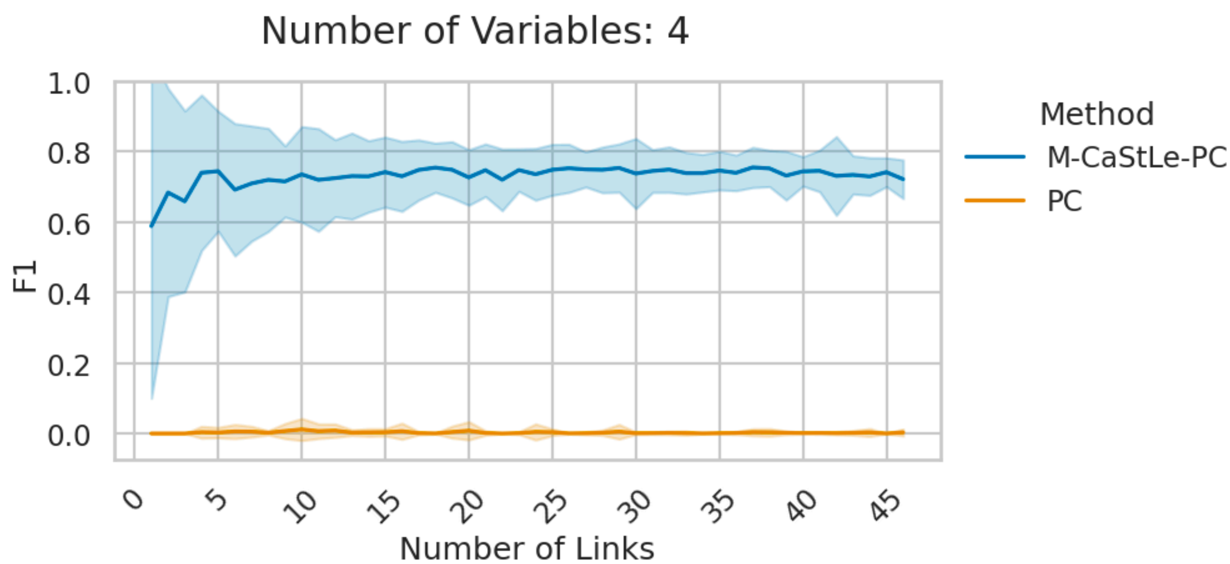


Figure 7.5: A comparison between M-CaStLe-PC and PC considering the F_1 score for $V = 4$ as the number of links increases on a 4×4 grid. M-CaStLe-PC outperforms PC in every case because PC struggles with the very high dimensionality of the systems since it is naive to the spatial and variable structures.

Figure 7.5 shows the F_1 score of M-CaStLe-PC and PC with increasing links, with the number of variables held constant to $V = 4$. The remaining V variables are given in Appendix N.1. We see that M-CaStLe-PC's F_1 score is consistently much higher than PC's. PC struggles with the very high dimensionality of the system since it is naive to the spatial and variable structure. Given that F_1 score is the harmonic mean of precision and recall, we can see that M-CaStLe's aggregate performance is between the very high precision and relatively low recall described above.

Exploring Recall To better evaluate the reason for M-CaStLe's relatively low recall, we tested it on a separate set of benchmark systems. In these, we constructed simple systems with many more variables and a range of coefficient magnitudes. The systems model a chain of dependence between each variable where there is one link per variable. The link is assigned a random parent location in the Moore neighborhood, and points to the center of the next variable. With this, we model different spatial relationships between variables but only one between variables. We explored $V \in \{10, 50, 100, 200\}$ and a set of 20 coefficients $\{c_i\}_{i=0}^{19}$ logarithmically spaced from 0.01 to 2.0, where $c_i = 0.01 \times 10^{2i/19}$. Every link had the same coefficient for each realization. Each realization had $T = 1000$ time samples and we restricted the grid size to 4×4, which is the most challenging for M-CaStLe because there are fewer spatial replicates to leverage.

Figure 7.6 illustrates that recall increases proportionately with coefficient magnitude for all numbers of variables. Recall is 0 when coefficients are too small and 1 when they are large enough. There is an inflection interval in the coefficient magnitudes in which recall increases sharply. The three-parameter sigmoid functions fit to each set set of Vs shows that recall is ordered by V. That means that, while high recall is achievable for up to 200 variables, systems with more variables are marginally more challenging to estimate, which conforms to our expectations. These results show that high recall is possible in high variable regimes if signals

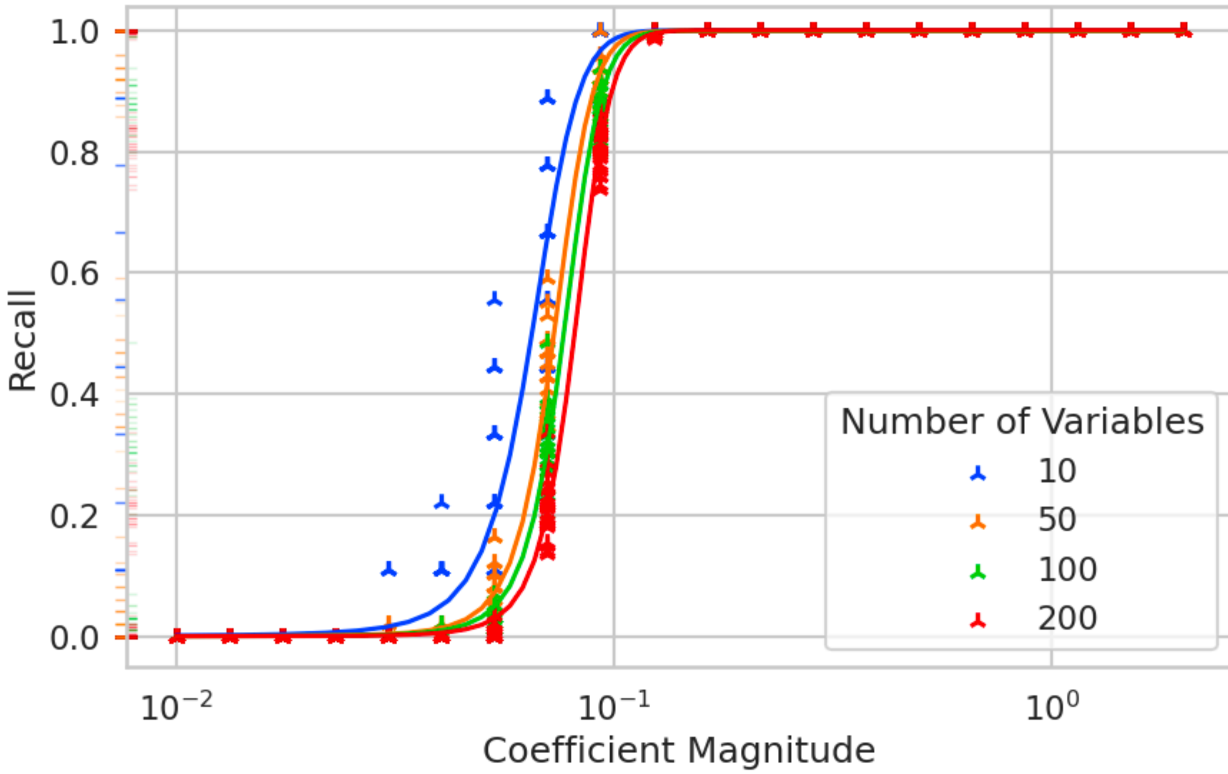


Figure 7.6: In simple chains of multivariate stencils, even with an extremely large number of variables, recall can be captured perfectly if the signal strength is large enough.

are strong enough.

7.5.2 Advection-Diffusion-Reaction Dynamics

While VAR benchmark is useful to understand how M-CaStLe performs in controllable, idealized settings and fairly compare methodologies, it cannot capture more realistic physical dynamics. They are continuous and often include complex local relationships that include transport and diffusion. To examine M-CaStLe's performance in more realistic settings, we implemented an ADR PDE model that describes the transport and interaction of two chemical species.

Model Definition

The two species interacting in the model are denoted as u_1 and u_2. Their dynamics are defined by:

$$\frac{\partial u_1}{\partial t} + \mathbf{v} \cdot \nabla u_1 = D_1 \nabla^2 u_1 + R_1(u_1, u_2) \tag{7.7}$$

$$\frac{\partial u_2}{\partial t} + \mathbf{v} \cdot \nabla u_2 = D_2 \nabla^2 u_2 + R_2(u_1, u_2) \tag{7.8}$$

where

$$\mathbf{v} = (v_x, v_y) = (v \cos(\theta), v \sin(\theta)) \tag{7.9}$$

In our experiments, the concentration of the first species, u_1, is initialized with a non-zero concentration, and u_2 begins at zero concentration. The interaction between the two species is characterized by a decay relationship, where species u_1 converts into species u_2 over time. Decay is captured in the reaction terms $R_1(u_1, u_2)$ and $R_2(u_1, u_2)$, which dictate the rate of conversion from u_1 to u_2.

The system is influenced by several key parameters that are set for the duration of each experiment:

- **Advection Velocity** (v): The speed at which the species are transported through the medium.
- **Advection Angle** (θ): Determines the direction of transport for both species.
- **Diffusion Rate** (D_1, D_2): Quantifies the tendency of the species to spread out due to concentration gradients. A higher diffusion rate results in a more rapid

dispersion of the species in the spatial domain.

- **Reaction Rate** (R_1, R_2): Govern the rates at which species u_1 decays into species u_2.

The reaction terms $R_1(u_1, u_2)$ and $R_2(u_1, u_2)$ govern the interaction between the two chemical species. Specifically, species u_1 undergoes a linear decay, converting into species u_2 at a rate controlled by two parameters: the decay rate (α) and a conversion factor (β) that determines the proportion of u_1 converted into u_2. The reaction terms are defined as:

$$R_1(u_1, u_2) = -\alpha u_1, \quad R_2(u_1, u_2) = \beta \alpha u_1 \tag{7.10}$$

This formulation ensures that the decay of u_1 is proportional to its concentration, while the production of u_2 is scaled accordingly.

Neumann boundary conditions with zero flux were applied along all edges of the spatial domain. To avoid boundary effects influencing the dynamics, the spatial domain was chosen to be significantly larger than the region of interest. Only the interior region was analyzed, where species concentrations remain unaffected by the boundaries.

Ground Truth, Estimation, and Metrics

As Rubenstein et al. (2018) demonstrate, there is no direct one-to-one mapping between causal graphs and differential equation systems, because the same underlying dynamics can be represented by multiple valid causal structures depending on variable choice, time scales, and levels of abstraction. That fundamental ambiguity poses a significant challenge for causal discovery algorithms applied to PDE-governed systems, since there is no explicitly defined ground truth causal graph inherent to the differential equation specification.

Thus, to assess the quality of causal stencil graphs estimated by M-CaStLe, we derive the ground truth from properties of the PDE definition and a means of comparing the stencils to the ground truth. The two primary dynamical elements captured by the causal stencil graph are spatial dependence and interactions between species. Spatial dynamics can be modeled in the ADR model by defining advection angle and velocity in $\mathbf{v}$ and species interactions can be modeled in the reaction terms, R_1, and R_2.

Estimation To derive advection angle and species interaction properties from the causal stencil graph output by M-CaStLe, we follow the decomposition procedure outlined in Section 7.4.3 to obtain spatial and reaction graphs. In this decomposed state, the reaction graph can be directly compared to a causal graph representing the species interaction dynamics. Figure 7.7 shows the ground truth causal graph for two species in our ADR model. Both species persist from one time step to the next as they advect and diffuse in the space, denoted by the autodependency straight arrows, and u_1 decays into u_2, denoted by the curved arrow. The reaction graph from the multivariate stencil graph decomposition has the same form, enabling a one-to-one comparison.

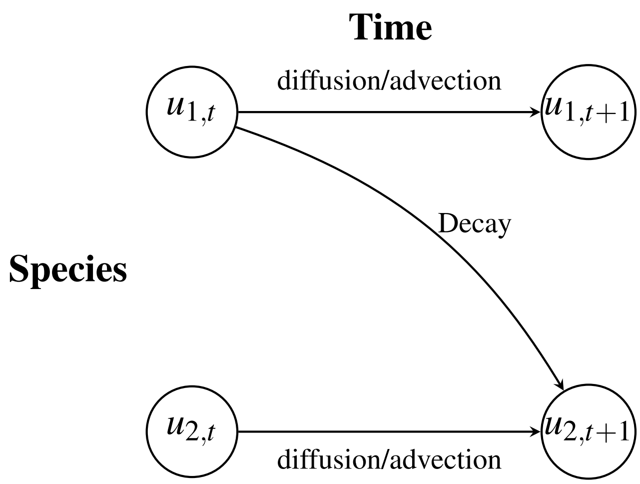


Figure 7.7: Time series causal graph for the ADR model. Nodes represent species u_1 and u_2 at each consecutive time step (t and $t+1$). Straight arrows indicate intra-species relationships across time, capturing diffusion and advection dynamics. The curved arrow represents inter-species causal interactions governed by reaction terms $R_1(u_1,u_2)$ and $R_2(u_1,u_2)$. This graph structure models the temporal and causal dependencies inherent in the ADR system, providing a framework for evaluating the species interactions found in causal stencil graphs.

To compare the discovered spatial dynamics to the ADR advection angle, we derived an angle from the computed multivariate stencil graph. Nichol et al. (2025, Appendix D) present a methodology for extracting an angle from the eight potential edges from exterior nodes of a univariate stencil. The approach involves computing a weighted circular mean of the angles represented by each edge, with weights determined by the correlation coefficients associated with each edge. In this work, we computed a non-negative weighted circular mean of the 16 potential exterior edges in the multivariate stencil to estimate the angle of advection in the ADR model. Mathematically, this can be described in the general case as

$$\hat{\theta} = \mathtt{atan2}\left(\sum_{l\in\mathscr{P}(\mathtt{C})} e_l \sin\theta_l, \sum_{l\in\mathscr{P}(\mathtt{C})} e_l \cos\theta_l\right).$$

where atan2 is the signed arctangent function; $\mathscr{P}(\text{Parents}(C_1, C_2, \ldots, C_k)) = \mathscr{P}(\{d_i : d \in \{\mathtt{NW}, \mathtt{N}, \mathtt{NE}, \mathtt{E}, \mathtt{SE}, \mathtt{S}, \mathtt{SW}, \mathtt{W}\}, i \in \{1, 2, \ldots, k\}\})$, for i species, represents all potential parents of all center cells across all species; e_l represents the strength of that edge (0 for non-present edges); and θ_l represents the angle of that edge $(135°, 90°, \ldots, 180°)$.

Metrics To measure the error of the reaction dynamics, we used the F_1 Score, defined as follows:

$$F_1 = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

where TP, FP, TN, and FN are true positive count, false positive count, true negative count, and false negative count, respectively. Here, a positive is a graph edge that exists, and a negative is a graph edge that does not exist.

To compute the angle estimation error, we find the minimum angle difference between the true and estimated angles (Nichol et al., 2025, Appendix D). It is described as

$$\Delta\theta = \min\left(|\theta_1 - \theta_2|, 360 - |\theta_1 - \theta_2|\right)$$

Experimental Parameters

To evaluate the performance of M-CaStLe on realistic physical dynamics, we conducted a series of ADR model experiments. Diffusion coefficients for the two species (D_1 and D_2) varied together between 0.005 and 0.4, advection velocities (v) ranged from 0.5 to 4.0, with angles (θ) from 0° to 90°. Reaction rates (α) were set to 1.0 with a scaling factor (β) of 1.0, which capture the decay of u_1 into u_2. We explored ranges of α and β, but found they only scaled the modeled data and M-CaStLe results were not affected. The initial concentration of u_1 was set to 50 and u_2's was set to zero. Varying initial concentration also had no effect on M-CaStLe results, so it is omitted. These experiments systematically explore the interplay between advection, diffusion, and reaction processes under diverse conditions to assess the method's ability to infer causal dynamics.

For all ADR experiments, we implemented M-CaStLe's PIP with the PC-stable (Colombo and Maathuis, 2014) causal discovery algorithm. M-CaStLe inherits two hyperparameters from PC-stable. First is a p-value threshold for individual independence tests, which we refer to as PC_α. Second is a p-value threshold used in the Benjamini–Hochberg false discovery rate correction (Benjamini and Hochberg, 1995), applied after the initial graph estimation, which we call the *graph-threshold*. In the experiments presented, we set both to 0.01, however, we found that results were not sensitive to these parameters.

ADR Results

Figure 7.8 shows the frequency distribution of F_1 scores of the reaction graph estimates for the ADR model experiments. These results show that reaction graph performance is very high, with a median F_1 score of 1.0, which means perfect graph estimation. A small subset of experiments have F_1 scores lower than 0.8. There were two experiments with $F_1 = 0.0$ because no links were estimated. These results highlight M-CaStLe's effectiveness at recovering reaction dynamics in recoverable regimes. Cases with lower F_1 scores underscore the inherent challenges of causal inference in regimes where reaction signals are weak or absent.

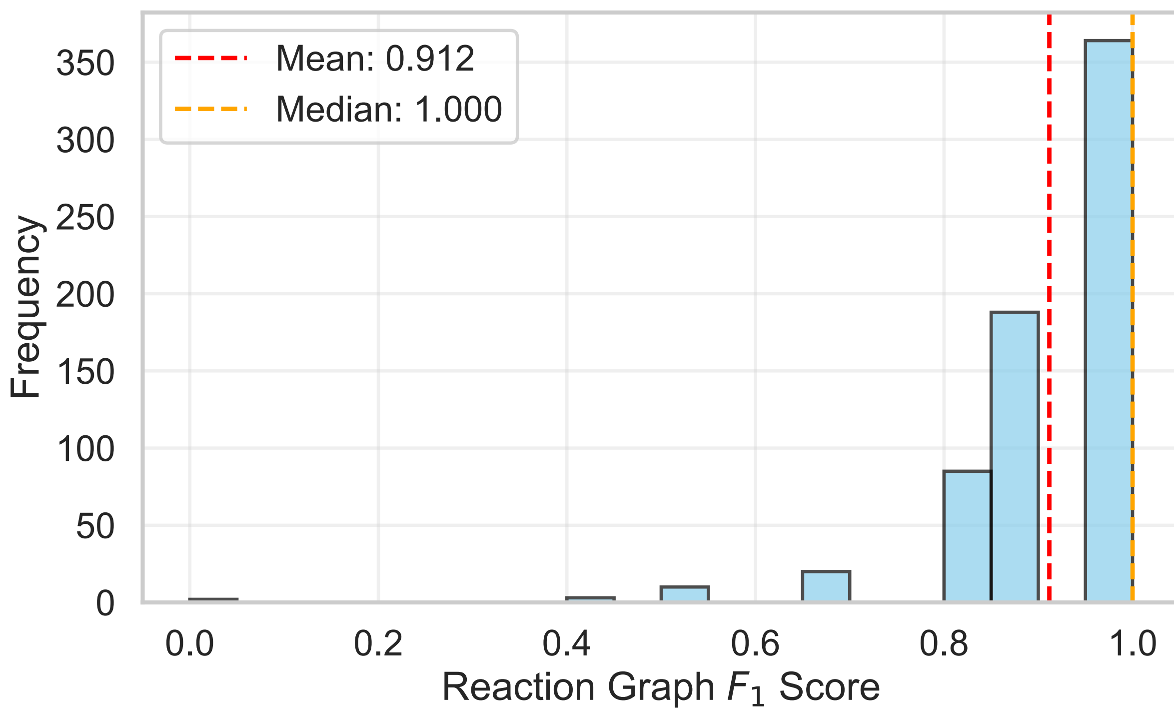


Figure 7.8: Histogram showing the frequency distribution of F_1 scores for reaction graph estimations ($n = 672$). The distribution has a mean of 0.912 (red dashed line) and median of 1.000 (orange dashed line), indicating that the majority of reaction graphs achieved perfect F_1 scores, with a smaller subset showing lower performance. Rare cases with $F_1 = 0$ are when no links were identified in the graph.

Figures 7.9 and 7.10 present angle estimation performance using M-CaStLe's estimated causal stencil graphs. The median error across all experiments was 4.76°. Given that stencil links are arranged discretely every 45°, and maximum error is 180°, it is clear M-CaStLe generally has very good performance.

In Figure 7.9, we show how angle estimation error varies with diffusion rate. Because stencil links have a 45° resolution, we also plot the proportion of experiments with angle errors exceeding this threshold, considering these as large errors. Median errors are low for smaller diffusion rates, but increase as diffusion

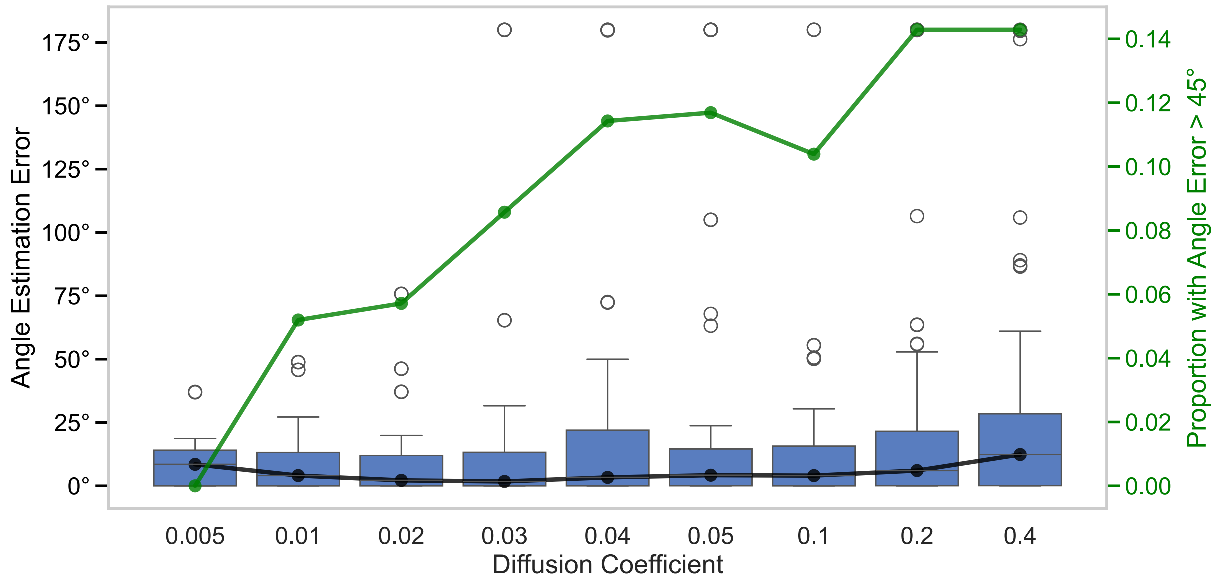


Figure 7.9: The distribution of angle estimation errors across diffusion coefficients using boxplots, with black markers representing median errors and green markers indicating the proportion of cases where errors exceed 45°. The global median error is 4.76°, and median errors are generally constant before increasing to 12.36° at 0.40. The proportion of large errors ($> 45°$) starts at 0% for 0.005, increases steadily to 14.3% for 0.40, and shows notable jumps at 0.03 and 0.20. While smaller diffusion coefficients are associated with lower median errors and fewer extreme cases, higher coefficients exhibit a higher likelihood of large errors, with maximum errors reaching 180° for coefficients ≥ 0.03.

increases to 0.4. While even the worst median error is 12.36°, the proportion of large errors increases steadily to 0.14 with increasing diffusion. When diffusion is greater than 0.02, there are extreme outliers where error reaches 180°. These extreme errors are consistent with the expectation that advection dynamics become increasingly ambiguous or absent as diffusion dominates.

In Figure 7.10, we illustrate how angle estimation error changes with advection velocity. Error decreases significantly from its peak, 23.19°, as velocity increases. Median error increases again when velocity reaches its maximum at $v = 4.0$. That is likely because the species advect outside the analysis window too quickly. The

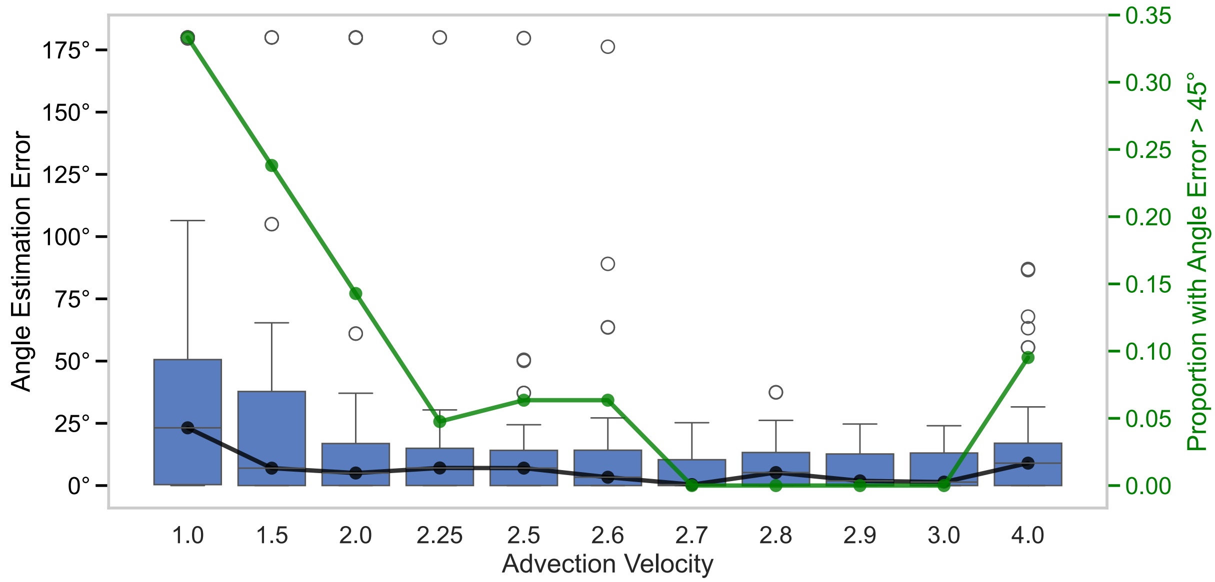


Figure 7.10: The distribution of angle estimation errors across advection velocities using boxplots, with black markers representing median errors and green markers indicating the proportion of cases where errors exceed 45°. The global median error is 4.76°, but median errors vary across velocities, decreasing from 23.19° for $v = 1.00$ to 2° for $v \geq 2.70$, before rising again to 9.00° at $v = 4.00$. The proportion of large errors ($> 45°$) decreases consistently with increasing velocity, starting at 33.3% for $v = 1.00$ and reaching 0% for $v \geq 2.70$, with a slight increase to 9.5% at $v = 4.00$. The increase in errors at $v = 4.00$ may be attributed to species advecting beyond the analysis window too quickly, reducing the accuracy of angle estimation.

same pattern appears in the green line, as the proportion of large errors decreases significantly from 0.35 when $v = 1.0$ to 0.0, before increasing again when $v = 4.0$.

M-CaStLe demonstrates robust performance across a range of diffusion and advection regimes. While the most difficult settings increase the rate of large errors, results show that error is generally low in all cases. The observed increase in errors in certain regimes (e.g., high diffusion or extreme advection velocities) is consistent with the expected behavior of the system, where dynamics become less recoverable due to the fundamental nature of the underlying processes. The interplay between diffusion, advection, and reaction significantly impacts causal inference accuracy. Strong advection improves directional clarity, while high diffusion introduces ambiguity that makes advection dynamics less distinguishable or absent. Reaction dynamics are recoverable in most cases, with rare exceptions where signals are weak or absent.

7.5.3 Atmospheric Chemistry

To explore realistic physical dynamics with well-understood ground truth, we examine atmospheric aerosol chemistry following the 1991 Mount Pinatubo eruption. This volcanic event injected approximately 20 Tg of SO_2 into the atmosphere (Guo et al., 2004a; Kremser et al., 2016), forming stratospheric sulfate aerosols that persisted for approximately two years. The resulting climate perturbation produced stratospheric warming of $\sim$1.5K and surface cooling of 0.2–0.5K (Dutton and Christy, 1992; Parker et al., 1996a; Soden et al., 2002), providing a natural analogue for stratospheric aerosol injection climate strategies (Trenberth and Dai, 2007; Weylandt and Swiler, 2024). Ongoing research efforts focus on characterizing the detailed mechanisms of the Pinatubo response, particularly emphasizing the temporal evolution and geographic patterns of surface climate impacts, which provide crucial insights for climate intervention policy development (Weylandt and Swiler, 2024).

We analyze a Mount Pinatubo eruption simulation using the fully coupled E3SMv2-SPA model, which incorporates atmosphere, land, oceanic, sea ice, land ice, and river components (Brown et al., 2024). Particularly of note is the model's prognostic sulfate chemistry including SO_2, OH, H_2SO_4, and SO_4 species. We extracted the total mass of H_2SO_4 and SO_2 through the atmospheric column, aerosol optical depth from SO_4 (AODSO4), and surface shortwave radiative flux data (FSDS). The intrinsic representation of aerosols within this modeling framework introduces additional complexity to the chemical pathway interpretation.

Experimental Setup

A large number of Earth system simulations were conducted for the CLimate Impact: Determining Etiology thRough pAthways (CLDERA) project at Sandia National Laboratories (Bull et al., 2024). Experiments varied levels of Mt. Pinatubo prognostic aerosols in the E3SMv2-SPA model to study with different levels of climate variability. Some sets of experiments used tagged aerosols from the eruption to track which aerosol particles originated from eruption as they spread around the globe and underwent chemistry into different species.

In this work, we examine the chemistry pathway of tagged Mt. Pinatubo-born SO_2 aerosols. The primary pathway of these aerosols undergo the following pathway as they have global temperature impacts: $SO_2 \rightarrow H_2SO_4 \rightarrow SO_4$ and $\{SO_2, H_2SO_4, SO_4\} \rightarrow \text{FSDS}$ (Hu et al., 2024). The impacts to incoming solar radiative flux mediated atmospheric temperature, which then had countless climate impacts around the globe (Thomas et al., 2009; Kremser et al., 2016; Mills et al., 2017). Here, we seek to reconstruct this pathway over a large region of the Earth using M-CaStLe, demonstrating that physical mechanistic space-time pathways can be recovered from complex real-world systems.

During the early months of the Mt. Pinatubo eruption, stratospheric winds were consistently moving westward in the tropical latitudes, where Pinatubo resides (Thomas et al., 2009). We captured gridded data in the region between 0.00° to 30.00°N and 60.00° to 90.00°E, a small distance west of Mt. Pinatubo. The dataset has a 1° spatial resolution (corresponding to approximately 107 km at 15 degrees N), with daily time samples per grid cell. We captured seven days of data after the eruption on June 15, 1991.

Despite such coarse and sparse temporal information, we show that much of the chemical pathway is recoverable because of M-CaStLe's use of spatial replicates. $N \times N = 30 \times 30$ grid cells with $T = 7$ time samples per species per grid cell, M-CaStLe's LENS accumulates a total of $(N-2)^2 T = 5,488$ samples per species per LENS node.

Results: Mt. Pinatubo Chemical and Radiation Impact Pathway Recovery

Because atmospheric aerosols mediate radiative flux at rates far exceeding the temporal resolution of our daily sampled data (Zhang et al., 2022b), we applied M-CaStLe *link assumptions* to constrain the parent sets specifically for radiative flux. Similar to the link assumptions functionality in the Tigramite causal discovery Python package (Runge, 2024), these specified assumptions embed subject matter expertise into the causal discovery algorithm. In this case, we stipulated that FSDS cannot be a causal parent of any other species. We did not specify whether it could be a child of any species, nor did we constrain any other relationships. To be clear, chemical species estimation was not given any link assumptions.

We implemented M-CaStLe's PIP using the PCMCI algorithm (Runge et al., 2019d). We set PC_α and the graph-threshold hyperparameters to 0.01. Finally, we applied a coefficient threshold of 0.125 to the graph as a regularizer to remove some particularly weak dependencies. Reaction graph dependencies may have a strength below that threshold due to aggregation effects in the decomposition process.

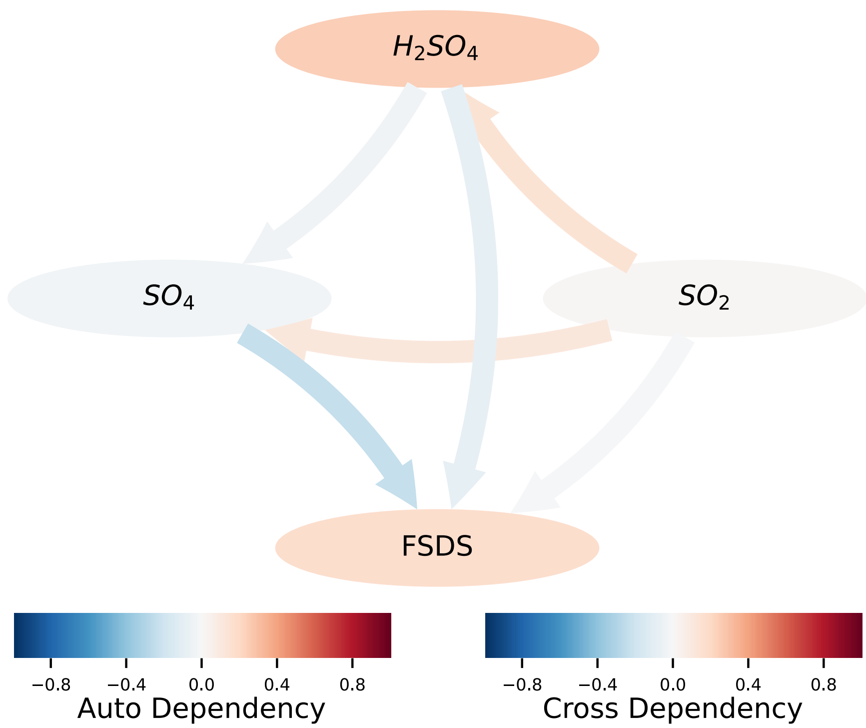


Figure 7.11: The reaction graph decomposed from M-CaStLe's multivariate causal stencil graph estimated from Mt. Pinatubo's chemical and radiation pathway computed in E3SMv2-SPA. Edges and nodes are colored by the identified strength of cross or auto dependence, with positive dependence depicted in reds and negative dependence in blues. M-CaStLe correctly identifies the general pathway $SO_2 \rightarrow H_2SO_4 \rightarrow SO_4$ and $\{SO_2, H_2SO_4, SO_4\} \rightarrow$ FSDS. SO_2 has one false positive, appearing to cause both SO_4 directly.

Figure 7.11 depicts the reaction graph derived from the estimated M-CaStLe stencil graph. The pathway $SO_2 \to H_2SO_4 \to SO_4$ and $\{SO_2, H_2SO_4, SO_4\} \to$ FSDS is identified. The polarity of each link is correct too, though the links $H_2SO_4 \to SO_4$ and $SO_2 \to$ FSDS are weak. There is one false positive, $SO_2 \to$ FSDS and $SO_2 \to SO_4$, which seems to shortcut the mediation through H_2SO_4.

SO_2 is not highly auto-dependent, which is consistent with knowledge that most of it converts to other species. H_2SO_4 and FSDS are positively auto-dependent, which is consistent with knowledge that they persist for long periods of time; FSDS being an external forcing from the sun. SO_4 is negatively auto-dependent, which is consistent with knowledge that it quickly falls out of the atmosphere via precipitation and dry deposition.

There are two major limitations of the data that may explain the false positive. First is the relatively coarse time sampling, which was also problematic for estimating the dependence of FSDS without link assumptions. The second is all of the chemical species are column-integrated quantities, meaning that they are 2D fields per time step, representing the total mass or aerosol optical depth of the quantity through the entire atmosphere. This limitation is common in simulation datasets and observed satellite data.

7.6 Discussion

We have proposed M-CaStLe, a multivariate extension to space-time grid-level causal discovery with CaStLe (Nichol et al., 2025). M-CaStLe adapts both the

Locally Encoded Neighborhood Structure construction and Parent-Identification Phase to learn inter-variable relationships in gridded space-time data. To represent these complex relationships, M-CaStLe produces a multivariate causal stencil graph that depicts which variable at each location in a Moore neighborhood causes each variable. To aid interpretation of the multivariate stencil, we introduced a decomposition method to extract spatial relationships and inter-variable relationships separately with the *spatial graph* and *reaction graph*.

Like CaStLe, M-CaStLe overcomes the limitations of high-dimensional gridded space-time systems, where there are more grid cells to estimate that time series samples in each. The inclusion of multiple variables exacerbates the high-dimensional challenge, but M-CaStLe includes variable structures in the spatial replicates it leverages to form the LENS. The LENS collects repeating multivariate spatial structures to form a 3×3 spatial data representation of the underlying dynamics. With that, the PIP recovers the multivariate stencil describing the underlying causal relationships that define the system's grid-level behavior.

We developed a multivariate gridded space-time benchmark framework, building upon the work by Nichol et al. (2023). The benchmark defines mathematical structures (VAR models) representing the space-time relationships between grid cells with multiple variables per grid cell. The structures directly translate to causal graphs for ground truth evaluation.

M-CaStLe performed well in the benchmark experiments. Its precision and recall were near 1 in systems with multiple variables when signal strengths were large enough. We applied the time series adapted PC causal discovery algorithm to the same benchmarks. We found that M-CaStLe had much better performance on multivariate systems than the PC causal discovery algorithm.

Recall suffered in highly complex systems cases because more complex systems exhibited smaller signal strengths per interaction. This supports our hypothesis that larger and more complex systems with many interacting components have fewer stable parameterizations. That is additionally supported by recent work investigating the *piranha problem* (Tosh et al., 2025), which describes the inevitable consequence that large complex systems will converge to weaker signals to maintain stability.

While the VAR benchmarks are valuable for understanding M-CaStLe with exactly satisfied assumptions and exactly known ground truth, it is a linear discrete system that is not particularly realistic. We tested M-CaStLe on realistic continuous physical dynamics with the advection-diffusion-reaction (ADR) PDE model to determine how well it generalizes to that domain. We found that M-CaStLe exhibits high performance in recovering inter-species dynamics. At the same time, we found that it was also able to recover the spatial evolution of the species with high accuracy in a variety of advection and diffusion environments.

The ADR experiments illustrated that M-CaStLe performs well on realistic physical dynamics, but it was only a two-species system with a simple inter-species relationship. We tested M-CaStLe on realistic atmospheric chemistry dynamics in the Earth system with E3SMv2-SPA model output. In this case, we considered four species as they advected, diffused, and reacted after their atmospheric injection from the Mt. Pinatubo eruption. We found that M-CaStLe successfully recovered the chemical pathway from SO_2 to mediating solar radiative flux. After decades of research, we know this pathway led to significant atmospheric temperature effects with global climate impacts. This study demonstrates that M-CaStLe can identify multivariate space-time dynamics in real-world complex settings.

This work shows that M-CaStLe is robust across synthetic benchmarks, realistic physics models, and real-world atmospheric dynamics. Other application domains, such as computational chemistry, fluid dynamics, and spatiotemporal pharmacokinetics can modeled or observed at sufficiently high resolutions given their smaller scale in comparison to the Earth system. While some dataset limitations still exist, Nichol et al. (2023) proposed other future research directions that may yield value in spite of those limitations. In particular, where spatial resolution is insufficiently matched temporal resolution, extending CaStLe and M-CaStLe to collect and evaluate larger neighborhoods, such as a radius-2 Moore neighborhood, could enable finding causal relationships that skip over immediately adjacent grid cells.

In this work, we have introduced M-CaStLe, a multivariate extension to the grid-level space-time causal discovery meta-algorithm, CaStLe. M-CaStLe addresses the significant challenge of estimating causal relationships in high-dimensional space-time systems with multiple interacting variables, which traditional approaches struggle to handle effectively. By enabling the simultaneous estimation of space-time dynamics and inter-variable interactions, M-CaStLe can enable advances in our understanding of complex systems, particularly in fields such as atmospheric science, computational fluid dynamics, computational chemistry, spatiotemporal pharmacokinetics, and epidemiological modeling. Our benchmark experiments demonstrate that M-CaStLe outperforms existing methods in accuracy, making it a robust and valuable tool for scientific discovery and analysis. Our PDE and atmospheric chemistry experiments demonstrate that M-CaStLe's capabilities can be valuable in real-world physiccal systems. Univariate CaStLe was a significant step for the analysis of high-dimensional grid-level dynamics and M-CaStLe makes multivariate space-time analysis possible. As a powerful tool for uncovering intricate causal relationships, M-CaStLe paves the way for more informed decision-making and deeper insights into the underlying mechanisms of complex phenomena.

Acknowledgment

We thank Diana Bull, the Principal Investigator of the CLDERA (CLimate impact: Determining Etiology thRough pAthways) project at Sandia National Laboratories, and the entire CLDERA team, for helping to make this work possible. We thank Joey Hart and Shane McQuarrie for developing the initial ADR model. Finally, we thank Laura P. Swiler for significant guidance and feedback throughout the development of this paper.

This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multi-mission laboratory man-

aged and operated by National Technology & Engineering Solutions of Sandia, LLC (NTESS), a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration (DOE/NNSA) under contract DE-NA0003525. This written work is authored by employees of NTESS. The employees, not NTESS, own the right, title, and interest in and to the written work and is responsible for its contents. Any subjective views or opinions that might be expressed in the written work do not necessarily represent the views of the U.S. Government. The publisher acknowledges that the U.S. Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this written work or allow others to do so, for U.S. Government purposes. The DOE will provide public access to results of federally sponsored research in accordance with the DOE Public Access Plan.

8 Conclusion

This dissertation's research rests upon the shoulders of 100 years of data-driven knowledge discovery. It does so by advancing our understanding of what contemporary methods are capable of for complex systems and filling a critical research gap in the discovery of underlying local dynamics. The Causal Space-Time Stencil Learning (CaStLe) meta-algorithm developed here enables scalable causal discovery of grid-level dynamics in multiple variables for high-dimensional data—an important and elusive advancement in causal discovery research, particularly for the Earth sciences. These contributions equip scientists to approach more nuanced problems to explain the complex systems that rule our environment.

This chapter first summarizes Parts I and II of the work detailed in this dissertation and then explores exciting future avenues of research. Part I explored the foundational work I completed in exploring the capabilities of machine learning feature importance and state-of-the-art causal discovery for structure learning in the Earth sciences. Part II described my contributions to grid-level causal discovery with CaStLe and M-CaStLe.

8.1 Part I: Synthesis of Foundations Work

8.1.1 Machine Learning Feature Importance for Climate Models

Chapter 3 sought to learn if machine learning (ML) feature importance can be used to identify differences between climate model ensemble members' output data and observed data from satellite reanalysis products. In particular, I wanted to understand if I could predict and explain the Arctic's minimum yearly sea ice extent. Sea ice extent measures the square area of sea covered by ice, an important factor in Arctic life and trade vessel navigation. I trained ML models on 10 Arctic features that predict yearly sea ice extent minimums. Comparing ML model outputs between Arctic datasets gave us an understanding of their differences.

My methodology used separate random forest regression (RFR) (Breiman, 2001) models to learn from an observational dataset and five Energy Exascale Earth System Model (E3SM) (E3SM Project, 2018) simulation ensemble members. Random forests are ML models formed from aggregated decision trees. As RFR models train, they simultaneously build Gini importance values as part of the tree structures. It determines which features provide the most predictive power and encodes them in its Gini importance values. Thus, Gini importance describes how important each training feature is for the model's predictive power.

With the six trained models, I compared the calculated feature importance values to understand differences in the datasets. The baseline was data collected from satellite reanalysis products, which are observational datasets that use sophisticated models to interpolate missing data where clouds obstructed satellites. With that, I could compare its feature importance values with those of RFR models trained on the climate model simulation runs. I found important similarities between the datasets, suggesting that the models captured some fundamental dynamics in the Arctic climate. The E3SM model runs were the most similar to each other and had some noticeable differences with the observational dataset. While both datasets identified the same six important features, the E3SM datasets consistently over-weighted these features, with both ranking and magnitude discrepancies.

This work contributes to the broader climate analysis toolset by demonstrating how explainable machine learning can be used to learn about complex datasets. The work shows how physics-based models and ML can be used in tandem to learn more about critical systems in the Earth's climate. ML analyses like this can enhance climate model evaluation to improve existing model development and tuning practices. While more complex ML models proliferate, this work illustrates one important reason to maintain interpretability and explainability. Rather than simply demonstrating that discrepancies exist, analyses like this can help pinpoint potential sources of the discrepancies and lead climate model developers to the right place for refinement.

However, ML feature importance metrics are limited (Mandler and Weigand, 2024). The models themselves are subject to critical failures, such as various biases (Mehrabi et al., 2021), Simpson's paradox (Selvitella, 2017), and the Clever Hans effect (Lapuschkin et al., 2019), which can harm prediction performance or even make them appear to have high predictive skill, whereas it performs poorly outside the given training and testing datasets (Lee and Chen, 2025). Feature importance itself can be misleading and failure-prone due to issues such as multicollinearity between features (Cammarota and Pinto, 2021). Even when everything works as intended, it is important to know that ML feature importance is not a causal description of the data's underlying generating process. However, it is rather a description of the trained model itself. (Parr and Wilson, 2021; Parr et al., 2024)

This research has been significantly extended and advanced with follow-on work by Brown et al. (2025), where several coauthors from our original study developed a novel pathway detection methodology. They went beyond a comparative analysis to create networks of connected features based on random forest feature importance to relate climate quantities. Their work builds on our initial claim that ML feature importance can be used to obtain insights into systems' underlying structure. The progression from feature importance comparisons to network construction demonstrates the continued impact of our initial insights.

The work in this chapter has become a part of a broader literature on machine learning for the Earth sciences (Labe and Barnes, 2022; Konya and Nematzadeh, 2024; Lao et al., 2024). In the subsequent chapters, I investigated causal inference frameworks to understand underlying dynamics better.

8.1.2 Causal Discovery for Climate Model Evaluation

Chapter 4 explored causal discovery and what it can offer for climate model evaluation. This work finds complimentary patterns with Chapter 3 and demonstrates the limitations of regionally averaged climate quantities. In order to find more physically meaningful structure, local-level relationships need to be sought because this is where real interactions occur to give rise to global or regional patterns. These findings motivated our later work to discover grid-level causal relationships.

8.2 Part II: Discovery of Local Dynamics

8.2.1 Grid-Level Benchmarking of PCMCI

Chapter 5's benchmark evaluations of PCMCI revealed fundamental limitations of causal discovery algorithms that are naive of spatial structure. As suggested by prior work (Ebert-Uphoff and Deng, 2012; Nowack et al., 2020; Tibau et al., 2022), grid-level causal discovery is a significant challenge due to the high-dimensionality of gridded data. While PCMCI has been successful in recovering regional causal teleconnection networks, learning grid-level structures requires additional advancements. This presented a significant gap in the literature that I filled with CaStLe.

8.2.2 CaStLe: Grid-Level Causal Discovery

In Chapter 6, I introduced CaStLe, a grid-level causal discovery meta-algorithm. CaStLe addresses the fundamental challenge of causal discovery that I identified previously: many space-time gridded datasets are high-dimensional in practice. High sample complexity reduces the power of causal discovery's statistical estimators. CaStLe remedies this by two central premises: underlying dynamics act locally, each grid cell influences only its neighbors, and neighboring grid cells generally exhibit similar dynamics. Through these, CaStLe leverages locality and stationarity to collect *informative spatial replicates* for local causal structures, which boosts efficiency and efficacy of the causal discovery task.

CaStLe produces a novel causal graph type, the *causal stencil graph*, which is a spatially structured graph representing a Moore neighborhood of nodes, which represent grid cells. The Moore neighborhood is a grid cell and its eight immediate neighbors. The stencil graph describes which neighbors are causal parents of the center node, enabling full representation of local causal structure.

CaStLe is generally applicable to physics-governed space-time systems that satisfy the locality and stationarity assumptions. These include many processes in Earth science, fluid dynamics, and other fields where effects propagate locally through space and exhibit consistent or smoothly changing dynamics across regions. It can be applied in extremely data-poor settings, where only short time intervals are observed. It is especially valuable in settings in which grid-level dynamics define the phenomena under study and marginalization would destroy that information. These include advective, transient, and non-periodic phenomena such as volcanic eruptions, wildfires, and traveling weather fronts. CaStLe is a flexible meta-algorithm, enabling implementation with today's best causal discovery algorithms and those of the future, including causal representation learning. It is highly extensible, being adaptable to multiple variables, more than two spatial dimensions, longer time lags, and larger local neighborhoods.

CaStLe provides another path for physical model evaluation by elucidating where and why behavior does not match intended dynamics. For the first time, grid-level processes are recoverable with causal discovery, which opens the door to future multi-scale analyses to determine how local structures give rise to emergent global patterns. However, this initial version of CaStLe is univariate: it can only estimate space-time dynamics of one quantity, such as aerosols. It would be significantly more valuable estimating the space-time dynamics of multiple variables and their interactions. This is precisely what Chapter 7 addresses.

8.2.3 M-CaStLe: Multivariate Grid-Level Causal Discovery

I followed the development of CaStLe with an extension enabling multivariate analyses simultaneously with space-time structure discovery. Chapter 7 details the methodological innovations making that possible. We adapted both phases of CaStLe, developed a method for interpretability, and benchmarked M-CaStLe.

CaStLe's Locally Encoded Neighborhood Structure (LENS) was adapted to include multiple variables per time series. The mapping from the given gridded space to the multivariate LENS is represented by the transformation $\mathbb{R}^{N\times N\times V\times T} \rightarrow \mathbb{R}^{3\times 3\times V\times L}$, where $L = T(M-1)(N-1)$ denoted the length of each concatenated time series. With this, multiple variables' space-time structures are captured. CaStLe's Parent-Identification Phase (PIP) was adapted by allowing each of the variables in the center grid cell of the LENS to be children and no other grid cells. That allows for an adapted time series causal discovery algorithm to estimate the multivariate space-time dynamics underlying the given data.

M-CaStLe demonstrated robust performance across diverse systems, including synthetic benchmarks, physical models, and real-world applications. In multivariate vector autoregression model (VAR) benchmark experiments, M-CaStLe achieved near-perfect precision and recall in systems with strong signal strengths, outperforming traditional PC causal discovery algorithms, particularly in complex systems with multiple interacting variables. In the advection-diffusion-reaction (ADR) partial differential equation (PDE) model, M-CaStLe successfully recovered inter-species reaction dynamics with a median F_1 score of 1.0 and estimated spatial advection angles with a median error of 4.76°, showcasing its ability to infer causal relationships in realistic continuous physical systems. Finally, in the Earth system model simulation of the Mt. Pinatubo eruption, M-CaStLe accurately reconstructed the chemical pathway ($SO_2 \rightarrow H_2SO_4 \rightarrow SO_4$) and its mediation of solar radiative flux, despite challenges posed by coarse temporal sampling and column-integrated quantities. These results highlight M-CaStLe's versatility and effectiveness in uncovering multivariate space-time causal dynamics across synthetic, physical, and real-world domains.

M-CaStLe is the first causal discovery approach to enable grid-level causal discovery of multiple variables. This new capability can facilitate new research directions in physical systems such as the Earth sciences, computational chemistry, ecology, fluid dynamics, and pharmacokinetics. It presents many opportunities for interdisciplinary collaborations to analyze systems in a new way.

8.3 Connections and Research Frontiers

The research detailed in this dissertation traces a methodological journey from correlative machine learning approaches to mechanistic causal discovery frameworks for complex physical systems, with an emphasis on Earth science. The work spans multiple scales, progressing from regional analyses to tackling high-dimensional grid-level dynamics. The primary contribution, CaStLe, accomplished grid-level discovery for the first time by leveraging locality and stationarity principles: simplifying the causal discovery task without sacrificing spatial information through dimensionality reduction. Instead, CaStLe maintains critical spatial structure by collecting informative spatial replicates. The resulting causal stencil graph describes local causal structures between grid cells in a highly interpretable format. M-CaStLe enables a more comprehensive system understanding by extending capabilities to multiple variables. This work provides scientists with new tools to discover how local dynamics give rise to emergent global phenomena by bridging statistical learning with physical interpretation.

Several promising avenues emerge for extending this work. The most immediate opportunities lie in expanding CaStLe's applications within Earth sciences, particularly to atmospheric rivers, wildfire spread, drought propagation, and ocean dynamics—phenomena where local causal structures drive global patterns. A particularly valuable application would be Earth system model evaluation, where CaStLe could identify discrepancies between modeled and observed causal mechanisms at the grid level, potentially revealing cases where models produce correct outcomes through incorrect causal pathways. Beyond Earth sciences, CaStLe's applicability extends to any space-time system governed by local physical interactions, including fluid dynamics, biological pattern formation, and material transport processes.

Technical improvements should focus on developing data-driven methods for automated parameter selection, reducing the current reliance on domain expertise for block sizing and other parameters. For multi-scale analysis, a two-stage methodology appears promising: first applying CaStLe to discover local causal stencils, then employing complementary causal discovery techniques to connect these local processes across scales, potentially bridging local and global causal discovery in climate science. Integration with emerging causal AI methods, particularly causal representation learning for enhanced parent-identification and automated spatial embedding discovery, offers additional opportunities to improve CaStLe's performance and reduce its dependence on manual parameter tuning. Finally, where spatial resolution is insufficiently matched temporal resolution, extending CaStLe and M-CaStLe to collect and evaluate larger neighborhoods, such as a radius-2 Moore neighborhood, could enable finding causal relationships that skip over immediately adjacent grid cells.

This dissertation contributes to the century-long pursuit of understanding complex physical systems by advancing data-driven methodologies that bridge statistical learning and physical interpretation. It begins with machine learning-based analyses, demonstrating how feature importance metrics can provide insights into climate model discrepancies and enhance model evaluation practices. Building on this foundation, the work transitions to causal discovery, addressing the limitations of regional analyses and pioneering grid-level approaches to uncover local dynamics. The development of CaStLe and its multivariate extension, M-CaStLe, represents a significant advance in causal discovery, enabling scalable, interpretable analyses of gridded high-dimensional space-time systems. These tools illuminate how local interactions give rise to emergent global phenomena, offering new pathways for interdisciplinary research and model refinement. By equipping scientists with methods to uncover nuanced causal structures, this work contributes to the broader challenge of understanding and predicting the behavior of complex systems that govern our environment.

APPENDICES

Table A: Capabilities of CaStLe for Earth science applications. This table summarizes the key methodological advantages of CaStLe and their relevance to specific Earth science phenomena, highlighting applications where grid-level causal discovery enables analyses that were previously infeasible with prior causal discovery approaches.

Capability	**Description**	**Relevant Applications**
Local mechanism discovery	Global phenomena emerge from local causal interactions. Previous approaches use dimensionality reduction, losing this local information.	Volcanic plume transport (Sjolte et al., 2021), wildfire propagation & plume transport (Baranowski et al., 2021), atmospheric rivers (Payne et al., 2020; Baño-Medina et al., 2025; Higgins et al., 2025)
Transient, non-periodic phenomena	CaStLe effectively identifies grid-level causal pathways.	Volcanic eruptions, heat waves (Keellings and Moradkhani, 2020), wildfires (Driscoll et al., 2024)
High-dimensional data settings	CaStLe leverages spatial replicates to make high-dimensional problems tractable.	Gridded Earth science data from: regional climate modeling, satellite observation analysis, climate reanalysis products (Ali et al., 2024, Table 3)
Earth system model evaluation and comparison	CaStLe enables comparison of causal mechanisms between models and observations at the grid level, potentially identifying where models produce correct outcomes through incorrect causal mechanisms.	Grid-level causal model evaluation that identifies local mechanism differences between models and observations, extending beyond previous approaches that were limited to regional-scale analysis (Nowack et al., 2020; Nichol et al., 2021a)

A Understanding Assumptions

In this section, we outline the key assumptions underpinning the CaStLe framework and their relationship to causal discovery assumptions.

A.1 CaStLe Assumptions

CaStLe operates via two complementary sets of assumptions:

1. **CaStLe Framework Assumptions (T1, S1, T2, S2)**: These enable efficient use of spatiotemporal data by leveraging locality and stationarity to transform a high-dimensional problem into a tractable one.

2. **Causal Discovery Assumptions**: The causal discovery algorithm used within CaStLe's Parent Identification Phase requires its own set of assumptions - typically the Causal Markov Condition, Faithfulness, and Causal Sufficiency.

While these assumption sets are conceptually distinct and serve different purposes, they work together to enable scalable causal discovery in high-dimensional space-time systems.

In review, our framework introduces four key assumptions to capture a "PDE-like" system $\boldsymbol{X}_t$, creating an environment where local space-time dynamics can be efficiently learned:

T1) Temporal Locality: restricts causal influence to the most recent past state, one time lag, aligning with how PDEs are discretized.

T2) Temporal Causal Stationarity: ensures consistent causal structure over time.

S1) Spatial Locality: limits causal influence to immediate spatial neighbors.

S2) Spatial Causal Stationarity: ensures consistent causal structure across space.

These assumptions enable CaStLe to leverage “spatial replicates”—treating each local neighborhood as providing information about the same underlying causal process. This transforms what would be a high-dimensional, data-sparse problem (many variables, few observations) into a data-rich problem (few variables, many observations).

A.2 Causal Discovery Assumptions

Separately, the causal discovery algorithm used within CaStLe's PIP require its own assumptions. The three foundational assumptions of causal discovery are provided below, verbatim from Runge (2018a). In depth discussion of each is discussed by Spirtes et al. (1993, Ch. 3), and Peters et al. (2017, Ch. 6.5). They are discussed in terms of directed graph *separation* ($\bowtie$), where variables are separated when all causal paths between them are “blocked” by conditioning variables, preventing information flow through the graph structure. Separation is detailed more thoroughly by Runge (2018a, Section III B.).

- **Causal Markov condition**:

  The joint distribution of a time series process $\boldsymbol{X}$ with graph $\mathscr{G}$ fulfills the Causal Markov Condition if and only if for all $Y_t \in \boldsymbol{X}_t$ with parents $\mathscr{P}_{Y_t}$ in the graph

  $$\boldsymbol{X}_t^- \setminus \mathscr{P}_{Y_t} \bowtie Y_t \mid \mathscr{P}_{Y_t} \implies \boldsymbol{X}_t^- \setminus \mathscr{P}_{Y_t} \perp\!\!\!\perp Y_t \mid \mathscr{P}_{Y_t}, \tag{1}$$

  that is, from separation in the graph (since the parents $\mathscr{P}_{Y_t}$ separate Y_t from $\boldsymbol{X}_t^- \setminus \mathscr{P}_{Y_t}$ in the graph) follows independence.

This includes its contraposition

$$\boldsymbol{X}_t^- \setminus \mathscr{P}_{Y_t} \not\perp\!\!\!\perp Y_t \mid \mathscr{P}_{Y_t} \implies \boldsymbol{X}_t^- \setminus \mathscr{P}_{Y_t} \not\bowtie Y_t \mid \mathscr{P}_{Y_t}, \tag{2}$$

from dependence follows connectedness.

– A variable is conditionally independent of its non-effects given its direct causes.

- **Faithfulness**:

The joint distribution of a time series process $\boldsymbol{X}$ with graph $\mathscr{G}$ fulfills the Faithfulness condition if and only if for all disjoint subsets of nodes (or single nodes) $A, B, S \subset \mathscr{G}$ it holds that

$$X_Y \perp\!\!\!\perp X_Z \mid X_S \implies Y \bowtie Z \mid S, \tag{3}$$

that is, from independence follows separation, which includes its logical contraposition

$$Y \not\bowtie Z \mid S \implies X_Y \not\perp\!\!\!\perp X_Z \mid X_S, \tag{4}$$

from connectedness follows dependence.

– Every conditional independence in the data must correspond to a separation in the causal graph (no accidental cancellations).

- **Causal sufficiency**:

A set $W \subset V \times \mathbb{Z}$ of variables is causally sufficient for a process $\boldsymbol{X}$ if and only if in the process every common cause of any two or more variables in W is in W or has the same value for all units in the population.

A.3 Relationship Between Assumption Sets

While CaStLe assumptions (T1-S2) and causal discovery assumptions serve different purposes, there are important interactions between them:

- CaStLe assumptions create an environment where causal discovery becomes tractable in some high-dimensional gridded settings.
- CaStLe assumptions do not guarantee causal discovery assumptions will be satisfied.
- For example, even in perfectly stationary systems (T2, S2 satisfied), faithfulness can be violated through counteracting mechanisms, as demonstrated in Runge (2018a).
- Similarly, the Causal Markov Condition is a property of the joint distribution that cannot be derived from locality assumptions.

Instead of replacing causal discovery assumptions, CaStLe's assumptions create a context where causal discovery methods can be applied efficiently to high-dimensional space-time data.

CaStLe's Implementation and Causal Sufficiency

One meaningful connection exists between CaStLe's implementation and causal discovery assumptions: When CaStLe focuses on identifying only the parents of the center cell while including all potential spatial neighbors (per assumption S1), causal sufficiency is automatically satisfied for that specific node by construction - assuming S1 holds true.

This is a significant benefit, as causal sufficiency is typically the most difficult assumption to guarantee in practice (Spirtes et al., 1993; Raghu et al., 2018). While CaStLe cannot guarantee faithfulness or the Markov condition holds, its design cleverly leverages spatial structure to address causal sufficiency within each local analysis.

A.4 Potential Violations and their Manifestations

Violations of CaStLe's assumptions can occur in various ways, leading to different manifestations in the causal discovery process. Violations of CaStLe's assumptions can affect results in different ways:

1. Violations of Temporal/Spatial Locality (T1, S1): If causal effects extend beyond immediate neighbors, CaStLe will miss these connections, creating false negatives.

2. Violations of Stationarity (T2, S2): If dynamics change across space or time, CaStLe's stencil will represent only an average pattern, potentially creating

both false positives and negatives.

3. Even with CaStLe assumptions holding, traditional faithfulness violations can occur through cancellation effects or deterministic relationships.

Below, we provide examples of how these assumptions can be violated and their potential impacts, drawing on the discussion by Runge (2018a).

Temporal and Spatial Locality (T1, S1)

- *General Violation*: These assumptions can be violated by any process that introduces dependencies beyond immediate temporal or spatial neighbors.
- *Example – Time Aggregation*: Time aggregation can violate temporal locality by introducing dependencies across multiple time steps. Runge (2018a) discusses how time aggregation can cause such violations (Section IV.B, Example 4). Figure 5 in Runge (2018a) illustrates the impact of time aggregation on causal inference.
- *Example - Spatial Aggregation*: Similarly, spatial aggregation can violate spatial locality by introducing dependencies across non-neighboring spatial units.

Temporal and Spatial Causal Stationarity (T2, S2)

- *General Violation*: These assumptions can be violated by any process that introduces changes in the causal relationships over time or space.

- *Example – Counteracting Mechanisms*: Counteracting mechanisms or heterogeneous processes can violate these stationarity assumptions. If the data contains opposing generating processes (e.g., different hemispheres in climate data), the faithfulness assumption may be violated. This results in unstable and inconsistent causal relationships. Runge (2018a) discusses such violations in Section IV.C, Example 5, and provides an illustration in Figure 6.

Understanding potential violations and their manifestations is crucial for applying our framework effectively in realistic scenarios. Section 6.6.6 outlines practical strategies to mitigate these violations.

B Statistical and Time Complexity

In this section, we elaborate on Section 6.6.4 and provide a more detailed discussion of the time-complexity (Appendix B.1) and statistical (Appendix B.2) properties of CaStLe. Additionally, we provide analyses giving conditions under which CaStLe is (asymptotically) guaranteed to recover the true causal graph, independent of the specific PIP used.

B.1 Time Complexity

Steps A, B, and D of CaStLe consist primarily of copying and rearranging of data, so we focus our analysis on the complexity of Step C, which dominates the runtime of CaStLe. Because CaStLe can use a variety of PIPs within Step C, we begin with a general analysis of the worst-case time complexity of causal discovery algorithms. Throughout, recall that a runtime complexity $\mathscr{O}(f(n))$ implies there exists a fixed constant $C \geq 0$ such that that the algorithm terminates in at most $Cf(n)$ steps *for any input of size n*.

Kalisch and Bühlmann (2007) and Runge (2018a) discuss the time complexity of causal discovery, particularly the PC algorithm. Much of constraint-based causal discovery is descendant of PC, and it represents a valuable baseline for comparing the computational complexity of CaStLe and prior work. Causal discovery is largely bounded by how long it requires to determine independence between nodes (bounded by samples and size of conditioning sets of nodes) and how many times it needs to do so (generally bounded by the number of nodes). Runge (2018a) cite the time complexity of a single conditional independence test using ordinary least squares (linear partial correlation), while Kalisch and Bühlmann (2007) explore bounds on the number of tests in PC. Our analysis is consistent with theirs, which we derive from first principles.

Consider causal discovery in p-dimensions (p measured variables) with n samples; suppose further that it is known, *a priori*, that any node in the causal graph has at most degree q: that is, no element has more than q causal parents. An exhaustive search for the causal parents of a *single node* will require evaluating $\sum_{i=0}^{q} \binom{p}{i} = \mathscr{O}(2^p)$ possible sets of parents; repeating this process for all p nodes evaluation of up to $\mathscr{O}(p2^p)$ possible causal graphs. If we construct graphs using statistical tests for linear partial (conditional) correlation, each test can be performed in $\mathcal{O}(n\,p\min\{n,p\}) = \mathcal{O}(np^2)$ time (the time required to fit an OLS regression to n observations and p variables using a direct method such as an SVD or QR factorization), yielding an overall runtime of

$$\mathcal{O}(np^2 * p2^p) = \mathcal{O}(np^3 2^p).$$

This analysis is quite loose, and as Runge (2018a) notes, the complexity of a *single* linear conditional independence test can be reduced to $\mathcal{O}(np^2q^2)$ when efficient algorithms are used. Far stronger guarantees can be provided for specific causal discovery algorithms that more efficiently search the space of possible graphs. Regardless, even this rough analysis will be sufficient to demonstrate the algorithmic improvements attained by CaStLe.

We now consider the specific context of causal discovery from gridded time series data. Here, we have $n = T$ total observations and have $p = N^2$ features of our data. Direct application of causal discovery to this data gives a worst-case complexity of

$$\mathcal{O}(np^3 2^p) = \mathcal{O}(T(N^2)^3 2^{N^2}) = \mathcal{O}(TN^6 2^{N^2}),$$

so the complexity of standard causal discovery methods grows *super-exponentially* with the size of the grid. For the purposes of direct comparison to CaStLe, where $p = N^2$, we assume PC's $\tau_{max} = 1$. By contrast, the reduced space where CaStLe's PIP operates has $T(N-2)^2$ observations and only $p = 9$ features, yielding a *poly-*

nomial worst-case runtime of

$$\mathscr{O}(np^3 2^p) = \mathscr{O}(T(N-2)^2 * 9^3 * 2^9) = \mathscr{O}(TN^2).$$

Even for grids of relatively modest size, this improvement can be significant: consider a small 30×30 grid; at 1° resolution, this covers approximately 1.5% of the globe. Unstructured causal discovery methods need to consider approximately $30^6 * 2^{30}$ possible graphs, while CaStLe needs to evaluate only $9^3 * 2^9 = 373,248$ graphs, representing an improvement of approximately 2×10^{12}-fold. Specific PIPs may provide less dramatic improvements, but it is clear that CaStLe can be expected to be millions-if not billions-of times more efficient than existing approaches.

Note that in our application scenarios, CaStLe is always applied to a square $N \times N$ grid. However, more generally we can consider p grid cells. Traditional causal discovery will be bounded by

$$\mathscr{O}(Tp^3 2^p),$$

while CaStLe will be bounded by

$$\mathscr{O}(Tp).$$

Thus, if grid cells scale linearly, CaStLe scales linearly in both samples and grid cells.

B.2 Statistical Consistency

Statistically, we see that CaStLe can achieve significantly improved estimation performance compared to a full graph inference approach. Rather than give a general analysis, we rely on the prior work of Kalisch and Bühlmann (2007) to compare CaStLe-PC with the standard PC algorithm. Using the same definitions of n, p, q as in our previous analysis, Kalisch and Bühlmann (2007, Appendix B) show that the probability of the PC algorithm incorrectly estimating the causal graph incorrectly is bounded above by

$$P[\hat{\mathscr{G}} \neq \mathscr{G}] = \mathscr{O}\left(p^{q+2}(n-q)e^{-c(n-q)}\right).$$

In our setting, this gives an error probability of

$$\mathscr{O}\left(p^{q+2}(n-q)e^{-c(n-q)}\right) = \mathscr{O}\left((N^2)^{N^2+2}(T-N^2)e^{-c(T-N^2)}\right) = \mathscr{O}\left(N^{2N^2}e^{cN^2} * Te^{-cT}\right)$$

for PC applied in the original data space. It is clear that this quantity grows rapidly in N, consistent with the intuition that causal discovery algorithms struggle when applied to larger spatial domains. By contrast, this analysis implies that the error probability of CaStLe-PC scales as

$$\mathscr{O}\left(p^{q+2}(n-q)e^{-c(n-q)}\right) = \mathscr{O}\left(9^{9+2}(T(N-2)^2-9)e^{-c(T(N-2)^2-9)}\right) = \mathscr{O}\left(\frac{TN^2}{e^{TN^2}}\right)$$

Quite surprisingly, this *decreases* with the graph size (N), implying that CaStLe actually achieves *better performance* when applied to larger spatial domains. We demonstrate the remarkable practical effect of this scaling in Section 6.8.1. Similar improvements can be shown for any base causal discovery algorithm (and associated PIP) for which precise estimates of statistical convergence rates are available.

C Asymptotic Consistency

We examine the asymptotic consistency of CaStLe, with a particular focus on the Parent Identification Phase (PIP). Asymptotic consistency is a fundamental property that ensures the accuracy of causal graph estimates as the number of observations increases. We begin by establishing the technical assumptions necessary for our analysis, specifically those related to the p-values generated by the PIP for edge existence. These assumptions are critical for maintaining control over both false positive and false negative rates, thereby ensuring the reliability of our causal inferences. The central theorem we present demonstrates that, under these conditions, CaStLe achieves asymptotic consistency as the number of nodes approaches infinity. In the case of Bayesian score optimization causal discovery, such as DYNOTEARS, Bayesian posterior probabilities can be used in lieu of p-values with suitable minor modifications to the combination procedure. The proof is structured into three parts, addressing the independence of observations, the application of Fisher's method for combining p-values, and the implications of using overlapping regions. Through this analysis, we aim to reinforce the validity of our algorithm and its effectiveness in uncovering causal relationships in gridded space-time data structures.

Technical Assumption (P1):

- The Parent Identification Phase, PIP$(\cdot)$, produces p-values for edge existence, which satisfy the following:
  - For every non-edge (i,j) $(j \notin \mathscr{P}(i))$, $\mathbb{P}(p_{\text{PIP}}^{(i,j)} \leq u) = u$ for all $u \in [0,1]$; that is $p_{\text{PIP}}^{(i,j)} \sim \mathscr{U}([0,1])$ is uniformly distributed.
  - For every edge (i,j) $(j \notin \mathscr{P}(i))$ and every $T > T_0$, there exists $\pi_{(i,j)}^{T}(u) > 0$ such that $\mathbb{P}(p_{\text{PIP}}^{(i,j)} \leq u) \leq \max\{0, u - \pi_{(i,j)}^{T}(u)\} < u$ for all $u \in [0,1]$.

  Taken together, these require that the PIP$(\cdot)$ control the false positive rate at the nominal significance level used and that the false negative rate is less than the false positive rate.

Here, T_0 is a minor technical assumption to allow the PIP to have non-trivial accuracy: we use it to exclude trivial cases like $T = 1$, in which no time series causal discovery mechanism can be accurate.

Additionally, note that we typically assume that the PIP$(\cdot)$ is asymptotically consistent, so that $\pi_{(i,j)}^{T}(u)$ is bounded above 0 for all u as $T \to \infty$. This can be used to prove T-asymptotic consistency of CaStLe, but in this section we aim only to prove N-asymptotic consistency.

Theorem: Suppose $\mathscr{D}$ is an $\mathbb{R}^{T\times N\times N}$ realization of a data-generating process satisfying T1-S2. Suppose also that $\mathsf{PIP}(\cdot)$ is a parent-identification-phase satisfying P1. Then, there exists a T_0 such that for any $T \geq T_0$, CaStLe is asymptotically consistent as $N \to \infty$; that is, the causal graph estimated by CaStLe converges to the true causal graph generating $\mathscr{D}$ with probability 1.

Proof. This proof proceeds in three parts:

- First, we argue that, for large N, well-separated (non-overlapping) spatial regions can be considered IID realizations.
- Next, we argue that the application of Fisher's method leads to asymptotic consistency of CaStLe.
- Finally, we argue that "infill" of the overlapping regions does not invalidate the asymptotic consistency.

At a high level, we argue that, because it is T-asymptotically consistent, there exists some T_0 where the PIP has non-trivial power. We then apply standard statistical methods for combining several weak p-values to obtain a global strong p-value. The technical bookkeeping of our argument serves primarily to deal with the fact that we use overlapping spatial regions and cannot assume independence of the individual p-values; we overcome this by selecting regions that are sufficiently spatially separated to be statistically independent on the time scale considered.

Without loss of generality, we focus on asymptotically consistent estimation of a single edge, say $(\mathsf{East}, \mathsf{Center})$. Extension to all 9 stencil edges follows immediately by a standard union bound argument.

Part I: For analytical simplicity, we divide the spatial region into square regions of size $(5+2T)\times(5+2T)$. On a grid of size $N\times N$, there are $B_{N,T} = \lfloor N/(5+2T)\rfloor$ such regions. We apply the PIP$(\cdot)$ to the center 3×3 region of each region separately, obtaining $B_{N,T}$ p-values for the existence of the edge. Because these central regions are separated by (at least) $2T+2$ grid cells and causal effects exist at a distance of at most $2T$ under our data generating model, these p-values can be treated as statistically independent. (This is essentially the same argument used by Goerg and Shalizi (2013), though their application is quite different.)

Part II: Given $B_{N,T}$ independent p-values, we then apply Fisher's method for combining p-values. Specifically, given a set of p-values for edge non-existence, Fisher's method controls the *familywise error-rate*, rejecting the global null (no edges anywhere). By our assumption of spatial homogeneity, if an edge exists in at least one region, it must exist everywhere, so Fisher's method precisely tests for edge existence in the stencil.

Recall that Fisher's method constructs a test statistic $T = -2\sum_{b=1}^{B}\log p_b$ and tests it against a null χ^2_B distribution. We consider two cases:

1. If the edge does not exist, each p-value is $\mathscr{U}([0,1])$ by construction and the test statistic T follows its null distribution. So long as the global significance level used for Fisher's test α_{Fisher} is converging to 0 as $N\to\infty$, we have asymptotic consistency for edge absence.

2. If the edge does exist, each p-value is less than α with probability $(1+c)\alpha$ for some c strictly positive. We then have that T has a *non-central* χ^2 distribution,

which is asymptotically distinguishable from a (central) χ^2 at all significance levels as $N \propto B \to \infty$.

Taken together, these guarantee the the output of Fisher's method is asymptotically consistent for both edge presence and edge absence.

Part III: In practice, we apply CaStLe not to disjoint regions but to overlapping regions. As discussed elsewhere, the region-discretization strategy and the use of Fisher's method are such that this does not cause "cross-contamination" or invalid tests of edge existence. We note here that this strategy also does not invalidate asymptotic consistency of CaStLe. Specifically, we note that, with overlapping regions, the p-values used in Fisher's method may no longer be assumed independent.

In this case, however, this is not an issue as they exhibit positive dependence (as they are taken from overlapping data). As such, the true degrees of freedom of T under the null are less than the nominal degrees of freedom; this leads Fisher's method to be (if anything) overly conservative in finite samples. Hence, for the case of edge absence, the nominal significance level is understated and we retain consistency as long as we take $\alpha_{\text{Fisher}} \xrightarrow{N\to\infty} 0$; for the case of edge presence, it suffices to note that the true sampling distribution is still asymptotically distinguishable from the null (since each individual p-value is powerful), so we retain consistency. $\square$

We note that Fisher's method may not be the optimal method for combining p-values. In particular, Holm's method allows for arbitrary dependence of the p-

values, likely yielding better performance at finite N, but we do not pursue this approach here as the implementation and theoretical analysis are somewhat more difficult. As with Fisher's method, Holm's method controls the error rate of the global null which, under our assumptions of causal stationarity, is precisely the correct null for accurate stencil estimation.

Additionally, we note that the p-values produced by the PIP under the null do not need to precisely satisfy a uniform distribution; conservative p-values decrease the value of Fisher's statistic T, thereby lowering the rate of false positives.

Remark: If PIP$(\cdot)$ is strongly asymptotically consistent as $T \to \infty$, it must satisfy assumption P1.

Proof. We argue by contradiction. Suppose that PIP$(\cdot)$ were not asymptotically consistent and that the false positive rates and false negative rates of the PIP were equal (or worse, the false negative rate was greater than the false positive rate). Specifically, assume that there exists a true edge (i, j) and some $\pi_- > 0$ such that $\mathbb{P}(p_{\text{PIP}}^{(i,j)} \leq u) > \pi_- + u$ for all T and all u. For the PIP to guarantee no false positives, we must take $\alpha \to 0$ as $T \to \infty$. But this would imply that there remains an asymptotic π_- probability of a false negative ($\mathbb{P}(p_{\text{PIP}}^{(i,j)} \leq \alpha) > \alpha + \pi_i \geq \pi_- > 0$), contradicting our assumption of asymptotic consistency.

$\square$

D Application to Non-Linear Dynamics: Continuous Systems via Burgers' Equation

This appendix extends our validation of CaStLe to non-linear dynamical systems through application to Burgers' equation, demonstrating the method's effectiveness beyond the linear systems discussed in the main text.

Having established the strong performance of CaStLe on discrete models of linear dynamics, we turn to a far more challenging domain: continuous models with non-linear PDEs. Specifically, motivated by our interest in turbulent atmospheric dynamics, we consider Burgers' equation, a PDE used to model a combination of advective (directed flow) and diffusive processes (Burgers, 1948). While initially developed to model fluid flows, Burgers' equation has been successfully applied to a variety of fields, such as turbulence, non-linear wave propagation, traffic flow, cosmology, gas dynamics, and more (Bonkile et al., 2018). In the following experiments, we again implemented CaStLe's PIP with the PC-Stable-Single algorithm.

We note that the interaction of PDE dynamics with causal language is rather subtle: while PDEs are imbued with a "forward" direction in time, the actual numerical methods used to solve them include "forward" and "backward" steps in the underlying integrators as well as sophisticated interpolation schemes. Our focus here is not on finding a causal model for the PDE solution per se, but on identifying the structure of the underlying advection. This choice is motivated in part by the results of Rubenstein et al. (2018), who explored the related problem of identifying causal models from deterministic ordinary differential equations (ODEs). As they note, there is not generally a single causal graph corresponding to an ODE, with different models being appropriate at equilibrium or under various interventions. Given the additional complexity of PDEs, we believe that identifying the underlying advection angle provides the most meaningful causal representation of Burgers-type dynamics, particularly as it relates to our volcanic eruption aerosol case study.

D.1 Burgers' Equation: Model and Parameters

In two dimensions, Burgers' equation can be written as:

$$\frac{\partial u}{\partial t} + \underbrace{u\left(\alpha\frac{\partial u}{\partial x} + \beta\frac{\partial u}{\partial y}\right)}_{\text{Advective Dynamics}} = \underbrace{c\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right)}_{\text{Diffusive Dynamics}} + f \tag{5}$$

where α, β are the advection coefficients in the x, y directions, capturing directed flow dynamics; c is the diffusion coefficient; and f is a forcing term representing additional mass being injected into the system. In order to create a closed system with no exogenous forcings, we take $f = 0$ uniformly throughout this section.

The left panel of Figure D1 shows three different solutions to Burgers' equation at different advection angles (θ), advection strength ($M = \sqrt{\alpha^2 + \beta^2}$), and diffusivities ($c$), each with the same initial conditions. Examining the time evolution of these solutions (left to right), we see that the high-advection low-diffusion systems (top) exhibit a clear direction of flow, while it is far more difficult to find direction in low-advection high-diffusion systems (bottom). We take inferring the angle of advection as our principal task: given an observed solution u to Equation (5), can we determine the angle of the underlying advective dynamics?

D.2 Advection Angle Estimation

Given a CaStLe-estimated stencil, we infer the angle of underlying advection in the following manner: i) identify each potential parent edge of $\mathtt{C}$ with a vector, taking the angle of the underlying edge in the reduced space as direction and the (signed) strength of the underlying relationship as magnitude; ii) sum these vectors to obtain an aggregate estimate of the advective dynamics; iii) take the angle of the vector sum as an estimate of the underlying advection angle. In pseudo-code, we can write this as

$$\hat{\theta} = \mathtt{atan2}\left(\sum_{l \in \mathscr{P}(\mathtt{C})} e_l \sin\theta_l, \sum_{l \in \mathscr{P}(\mathtt{C})} e_l \cos\theta_l\right).$$

Here atan2 is the signed arctangent function, $\mathscr{P}(\mathtt{C}) = \{\mathtt{NW}, \mathtt{N}, \ldots, \mathtt{W}\}$ represents all potential parents the center cell, e_l represents the strength of that edge (0 for non-present edges), and θ_l represents the angle of that edge $(135^\circ, 90^\circ, \ldots, 180^\circ)$. This process allows us to estimate all angles instead of just the eight angles present in the stencil structure.

D.3 Experimental Setup

In order to assess the effectiveness of CaStLe-PC in a variety of regimes, we generate (approximate) solutions to Equation (5) with 500 angles sampled uniformly from $[0°, 360°)$, advection magnitudes varying from 1 to 10 and diffusion coefficients from 0.05 to 0.5. The diffusion-free ("noiseless") case of $c = 0$ is numerically unstable. To compute the simulated Burgers' dynamics, we use MATLAB's default PDE solver (`pdesolve`) on a circular mesh of radius 3 and 100 time steps equally spaced between $t = 0$ and $t = 1$. Then we interpolated the finite-element solution onto a grid of size 25×25, covering the square $[-1, 1]^2$, yielding spatial points that are approximately 0.1 units apart. We restrict our solution to avoid any boundary conditions. Finally, we apply CaStLe-PC and the aforementioned advection angle estimation method, and compare the estimated angle to the true angle. We demonstrate three realizations of this process in the left-hand panel of Figure D1.

Angle Estimation Results

Our results appear in the right panel of Figure D1, where we plot the difference in the true and estimated angle, taking care to account for the "wrapping" behavior of angle-valued data. We see that stronger advection (higher SNR) consistently leads to improved estimation (downward trend within each group), with estimated angles consistently within $10°$ for advection magnitude 5 or greater. Comparing across different levels of the diffusion coefficient c, we note that higher c increases

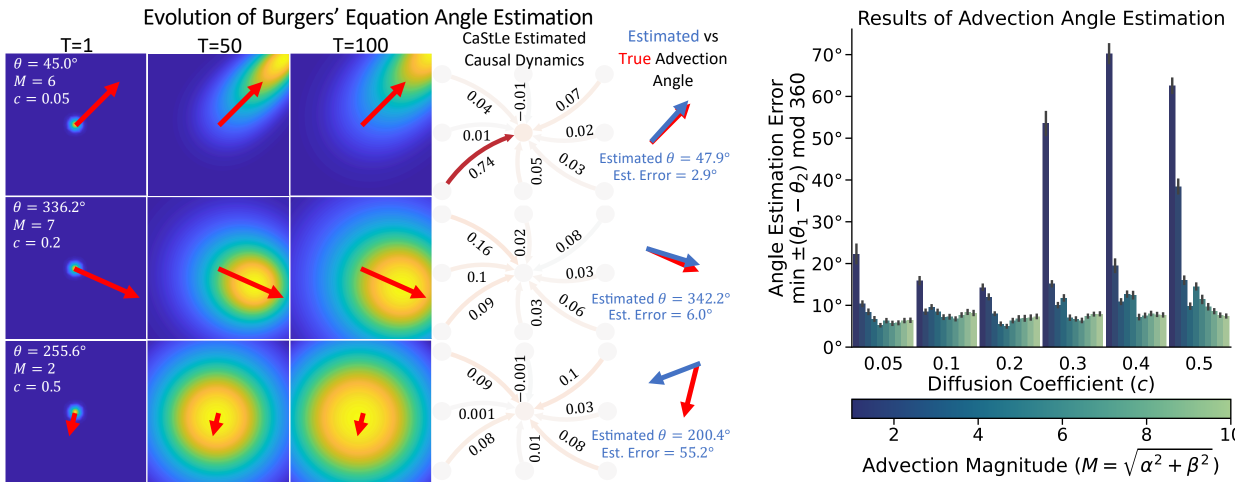


Figure D1: Application of CaStLe-PC to advection estimation from non-linear PDE dynamics. In the left panel, the first three columns depict realizations of Burgers' equation under different advection-to-diffusion regimes; the fourth column depicts the causal stencil identified by CaStLe-PC; and the final column compares the estimated advection angle with the true advection angle. The right panel depicts the accuracy of CaStLe-PC under various signal-to-noise conditions. Each combination of advection and diffusion rates were tested with 500 angles sampled uniformly from $[0°, 360°)$. In low-diffusion (high SNR) scenarios, CaStLe-PC can identify the underlying advection clearly (top row of left panel and yellow-green columns in right panel). By contrast, in low-advection (low SNR) scenarios, CaStLe-PC struggles to accurately identify the underlying advective dynamics (bottom row of left panel and blue bars in right panel). Even in highly diffusive scenarios, CaStLe-PC is able to accurately estimate the underlying advection when it is sufficiently strong (around $M/c \geq 20$) as shown in the middle row of the left panel. Additional details are given in D.

the angle estimation error, as we would expect in the higher-noise regimes. For low advection magnitude and $c \geq 0.3$, we see an average error approaching the "pure guessing" value of 90°. Even at high diffusion levels ($c = 0.5$), moderate advection magnitudes of 5-6 are sufficient to ensure accurate estimation. From these, we see that CaStLe-PC is able to consistently recover advection structure across a wide range of SNR regimes. As demonstrated in F, traditional dimension reduction approaches such as PCA and PCA-varimax, when combined with standard causal discovery methods, fail to accurately capture the advection dynamics in Burgers' equation, particularly in identifying the correct advection angle. This highlights

CaStLe's unique ability to preserve and extract meaningful causal structures from nonlinear PDE systems that would otherwise be lost through dimensionality reduction.

The takeaway from these results is that CaStLe can not only generalize to continuous, non-linear models of advection and diffusion, but it can successfully infer the direction of causality in any advective-diffusive system, given that the diffusion is not so large as to dominate advection. Further, each simulation has only one signal surrounded by large areas without data or causal information. Despite this sparsity and the presence of regions where diffusive information flow might suggest incorrect advection angles, CaStLe successfully identifies the correct advection angle when analyzing the full space. CaStLe is asked to learn from the full space, but successfully hones in on the correct advection angle. With these results, we believe CaStLe can be applied to a broad range of space-time systems with advective-diffusive properties to better understand their dynamics.

E Proposed Modification of Statistical Methods for CaStLed Data

While essentially any consistent PIP may be used in Step C, we anticipate that most PIPs will be derived from already existing causal discovery algorithms. Often, these algorithms are statistical in nature and it may be inappropriate to apply them directly to $\tilde{\boldsymbol{X}}$ due to the *seams* connecting each time *chunk*. For a statistical method, which computes a p-value for each potential edge (smaller p-values leading to present edges), we suggest the following *chunk testing* modification:

1. For each chunk $b \in \{1, \ldots, (N-1)^2\}$, let p_b be the p-value resulting from the PIP applied to that chunk.

2. Compute $T = -2\sum_b \ln p_b$

3. Let $p_{\text{agg}} = 1 - \chi^2_{2(N-1)^2}(T)$ where $\chi^2_k(x)$ is the cumulative distribution function (CDF) of a χ^2 random variable with k degrees of freedom evaluate at x.

4. If $p_{\text{agg}} < p_*$, identify a parent.

This method adapts Fisher's classical method for combining independent p-values to our setting. In practice, however, we have found that for sufficiently large T, this *chunking* is unnecessary as the proportion of *seams* in $\tilde{\boldsymbol{X}}$ goes to zero, and the PIP identifies the correct causal structure despite the small fraction of points of misspecification $(1/T)$.

F Limitations of Dimensionality Reduction for Space-Time Causal Discovery

We demonstrate the limitations of dimensionality reduction approaches such as PCA and PCA-varimax when applied to space-time causal discovery of advective-diffusive processes. Causal discovery methods in Earth science often employ these techniques to reduce the high dimensionality of gridded data before applying causal discovery algorithms. While effective for identifying large-scale teleconnections, we show that these approaches fail to capture the local causal structures that are essential for understanding space-time dynamics at the grid-cell level.

To illustrate these limitations, we apply PCA and PCA-varimax dimension reduction followed by PCMCI causal discovery——the procedure described by Runge et al. (2015b), Nowack et al. (2020), and Tibau et al. (2022) and employed in subsequent work—to each of our case studies: Burgers' equation, HSW-V, and E3SMv2-SPA. Our analysis reveals that while dimensionality reduction techniques can identify dominant modes of variability, they struggle to preserve the spatial relationships between neighboring grid cells, thus obscuring the local causal pathways that CaStLe is specifically designed to recover.

For the PCMCI step, we explored multiple lag values in our experiments and found that the results were consistently unable to capture the directional advection structure regardless of lag parameter choice. This suggests that the limitation is a fundamental constraint of the dimensionality reduction approach. In the results below, we show the simplest case with a maximum lag of 1.

Figure F1 shows the PCA analysis of Burgers' equation, where four EOFs capture approximately 91% of variance but the resulting PCMCI causal graph fails to recover the directional advection process, demonstrating PCA's inability to preserve local causal structures. Figure F2 shows similar limitations with PCA-Varimax applied to the same Burgers' equation data, where despite the rotation enhancing spatial localization of patterns, the causal graph still cannot represent the known directional advection dynamics. Figure F3 illustrates PCA applied to the HSW-V volcanic aerosol dataset, where four EOFs explain 85% of variance but produce a causal graph that misrepresents the known transport mechanisms. Figure F4 demonstrates that even with varimax rotation, which provides more spatially distinct patterns in the HSW-V dataset, the resulting causal graph cannot capture the directional flow of volcanic aerosols. The EOFs were reordered according to the identified centroids' longitude to improve interpretability. Figure F5 shows the application of PCA to the E3SMv2-SPA climate model data, where nine EOFs account for 87% of variance, yet the PCMCI causal graph fails to detect the underlying atmospheric circulation patterns. Figure F6 reveals that PCA-Varimax rotation of the E3SMv2-SPA data, with EOFs similarly reordered by longitudinal position for interpretability, still fails to recover the known directional transport processes, further confirming the limitations of dimensionality reduction for space-time causal discovery.

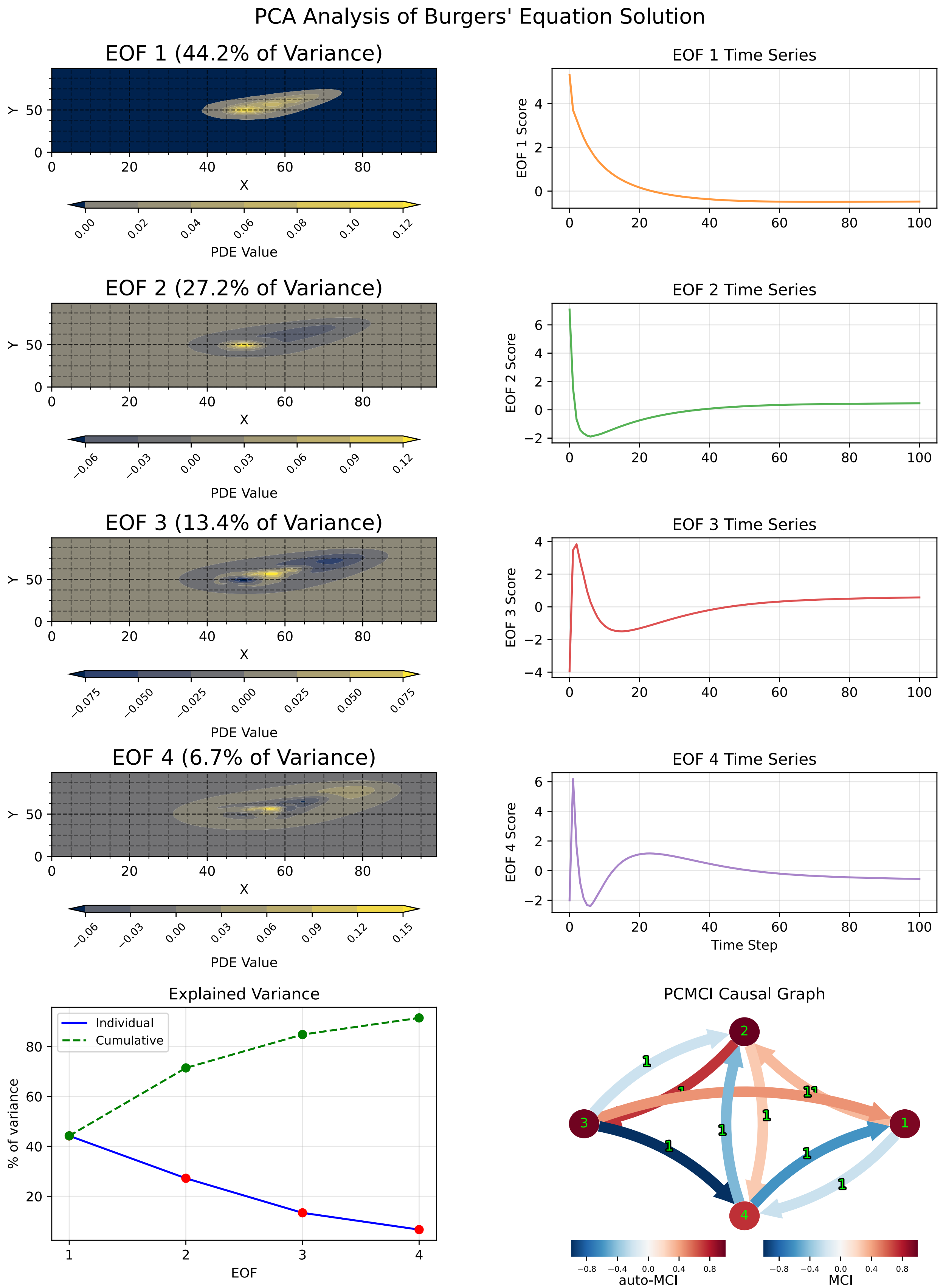


Figure F1: PCA study of Burgers' equation solution ($\theta = 45^\circ$, $M = 6$, $c = 0.05$). Four empirical orthogonal functions (EOFs) capture $\approx$91% of variance, with spatial patterns (left) and temporal evolution (right). The bottom panels show explained variance distribution and PCMCI causal graph, which fails to accurately represent the known directional advection process in the underlying PDE, highlighting limitations of this approach for local causal structures in space-time systems.

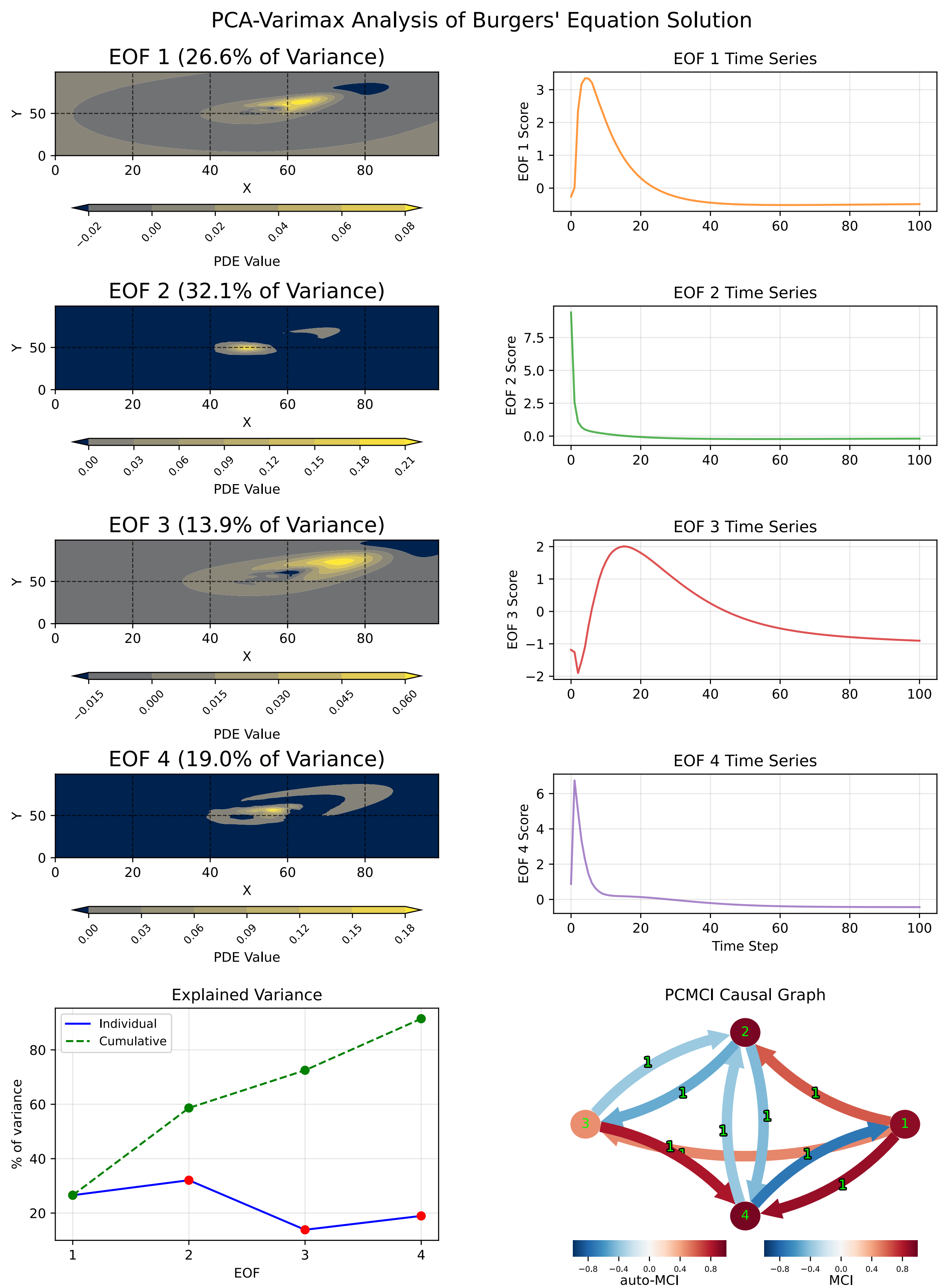


Figure F2: PCA-Varimax study of Burgers' equation solution ($\theta = 45^\circ$, $M = 6$, $c = 0.05$). Four empirical orthogonal functions (EOFs) capture 91% of variance, with spatial patterns (left) and temporal evolution (right). The bottom panels show explained variance distribution and PCMCI causal graph, which fails to accurately represent the known directional advection process in the underlying PDE, highlighting limitations of this approach for local causal structures in space-time

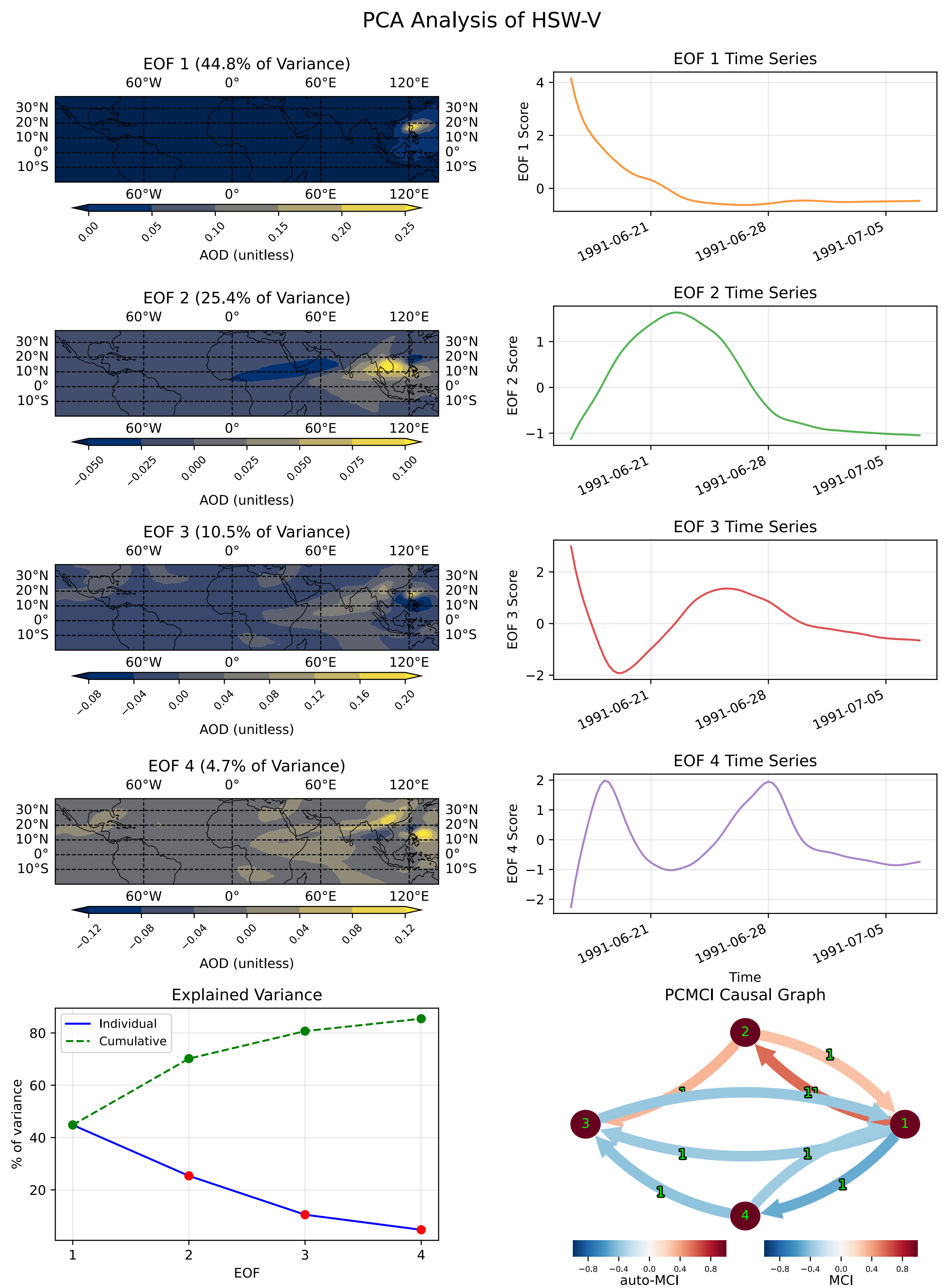


Figure F3: PCA study of the HSW-V dataset, in the time interval 21 days post-eruption. Four empirical orthogonal functions (EOFs) capture ≈85% of variance, with spatial patterns (left) and temporal evolution (right). The bottom panels show explained variance distribution and PCMCI causal graph, which fails to accurately represent the known directional advection process in the underlying system, highlighting limitations of this approach for local causal structures in space-

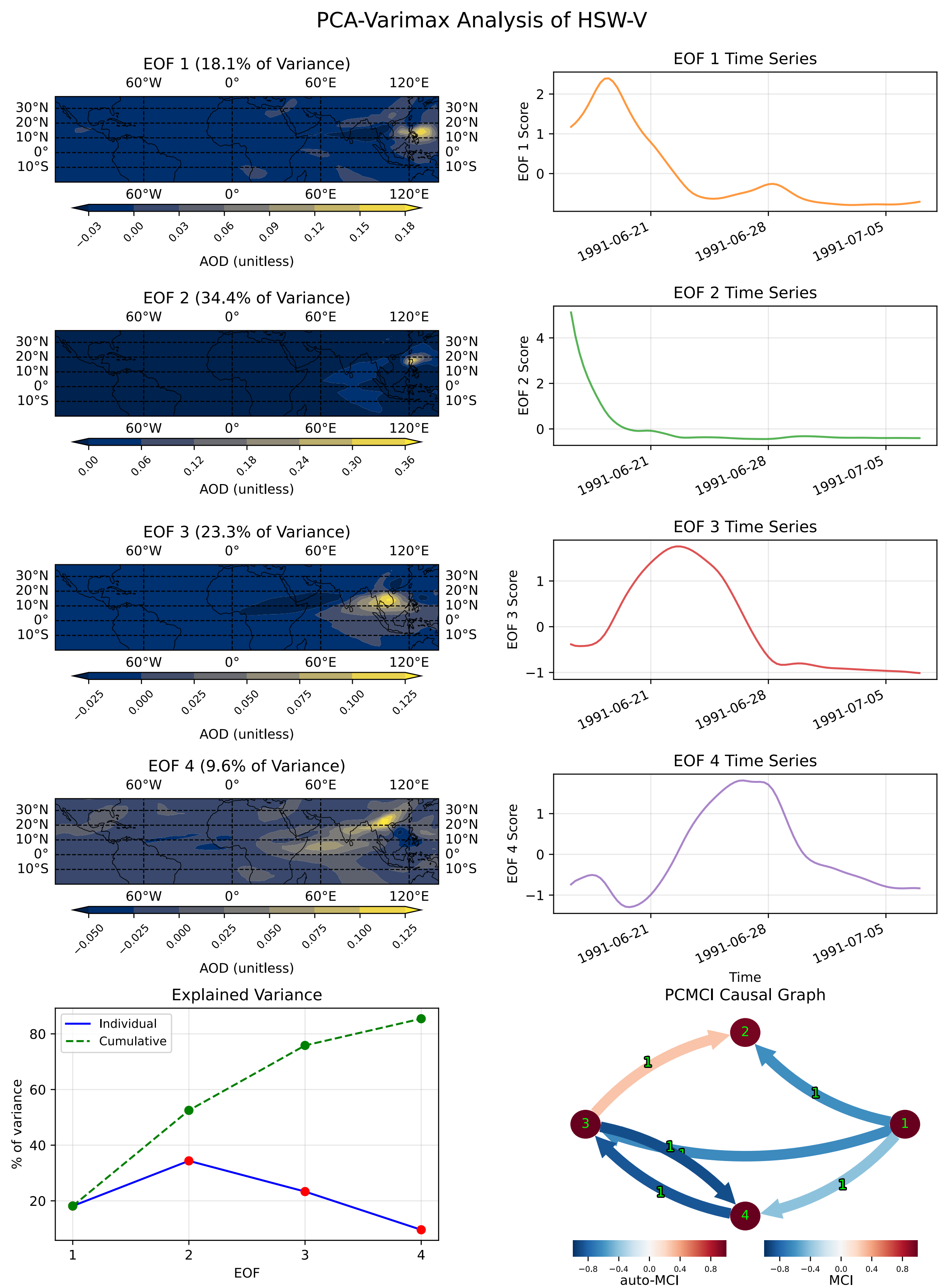


Figure F4: PCA-Varimax study of the HSW-V dataset, in the time interval 21 days post-eruption. Four empirical orthogonal functions (EOFs) capture ≈85% of variance, with spatial patterns (left) and temporal evolution (right). Since varimax rotation does not preserve the explained variance ordering, we reordered EOFs according to the identified centroid's longitude. The bottom panels show explained variance distribution and PCMCI causal graph, which fails to accurately represent

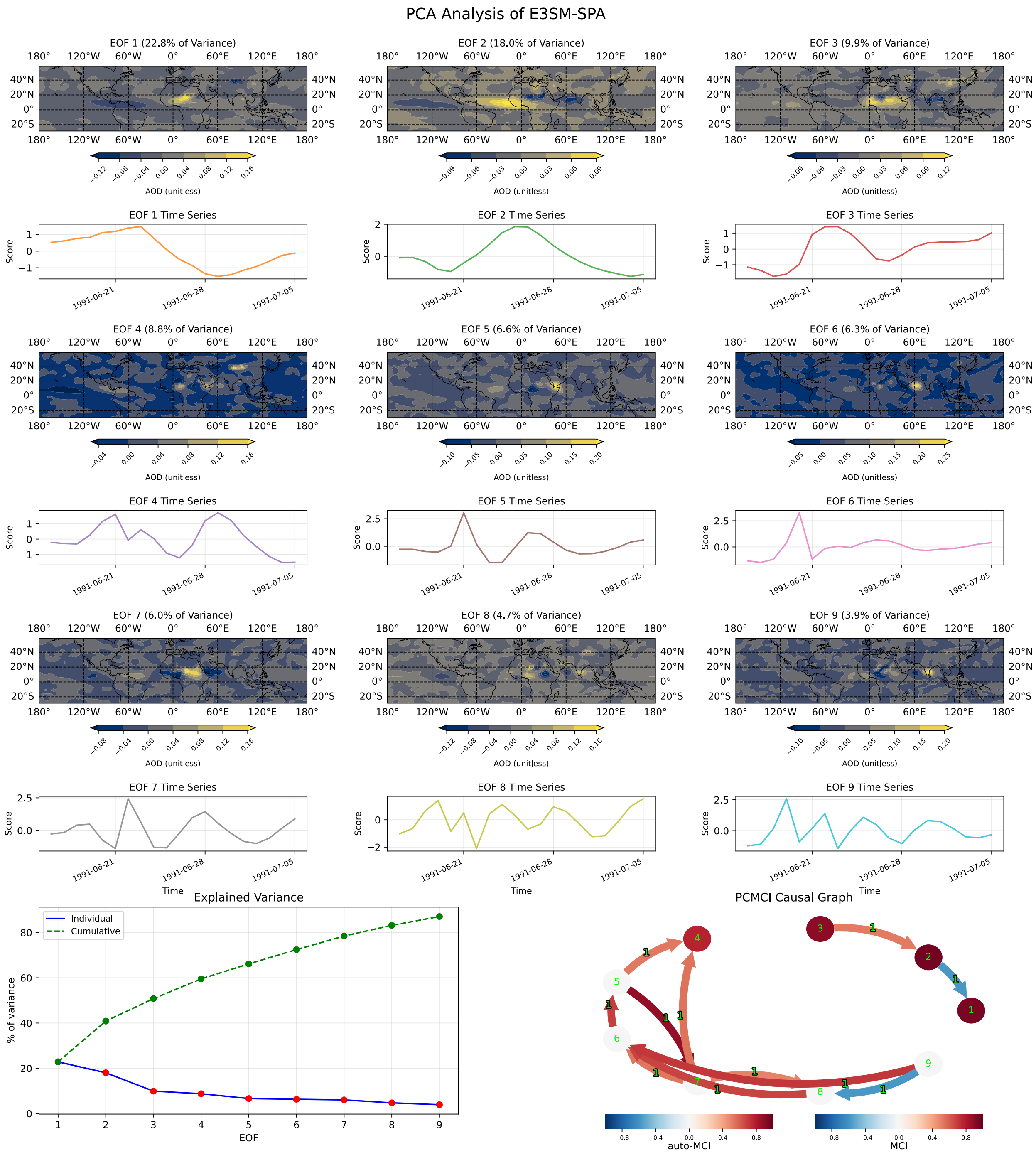


Figure F5: PCA study of the E3SMv2-SPA dataset, in the time interval of days 15-35. Nine empirical orthogonal functions (EOFs) capture ≈87% of variance, with spatial patterns (left) and temporal evolution (right). The bottom panels show explained variance distribution and PCMCI causal graph, which fails to accurately represent the known directional advection process in the underlying system, highlighting limitations of this approach for local causal structures in space-time systems.

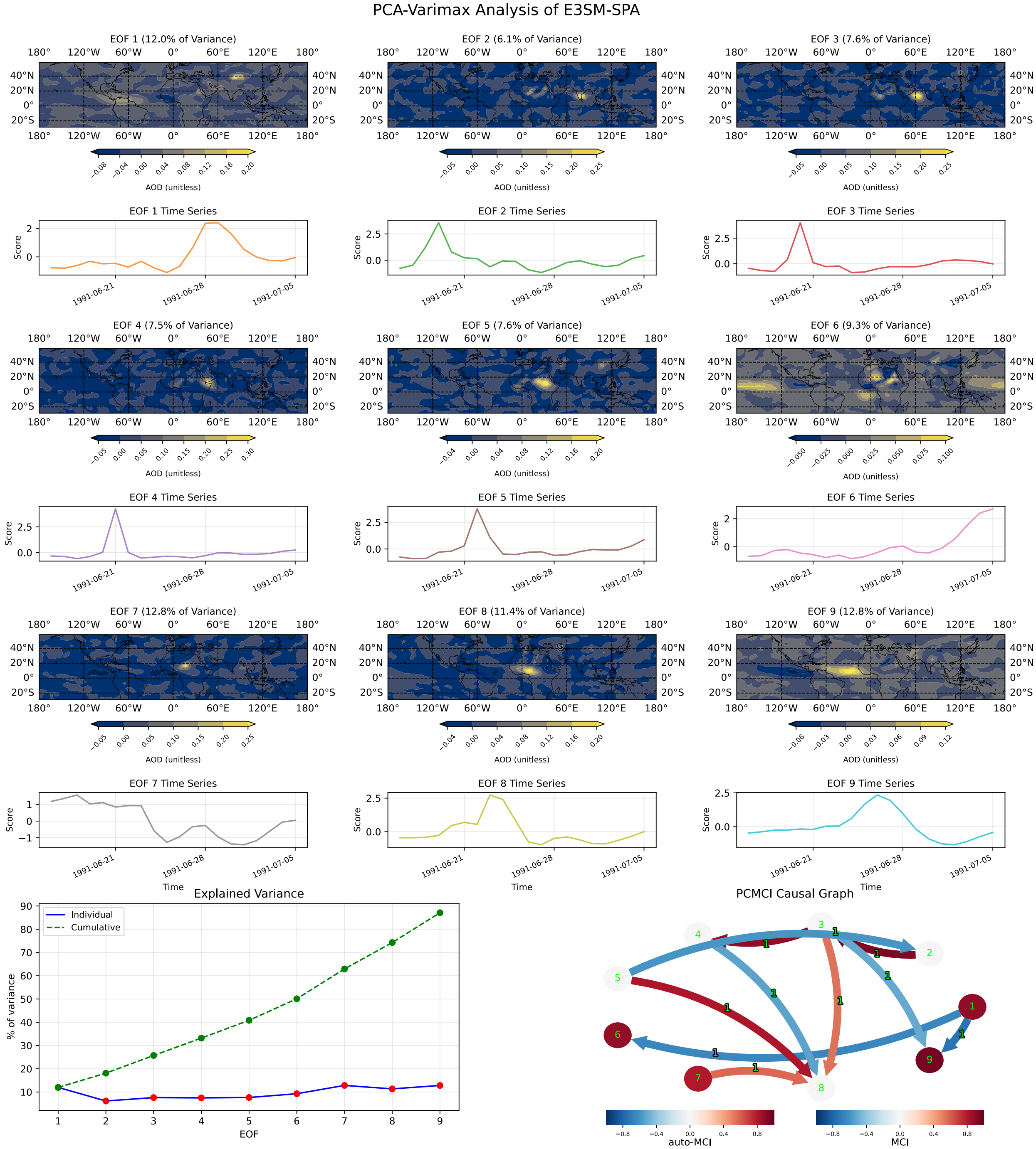


Figure F6: PCA-Varimax study of the E3SMv2-SPA dataset, in the time interval of days 15-35. Nine empirical orthogonal functions (EOFs) capture ≈87% of variance, with spatial patterns (left) and temporal evolution (right). Since varimax rotation does not preserve the explained variance ordering, we reordered EOFs according to the identified centroid's longitude. The bottom panels show explained variance distribution and PCMCI causal graph, which fails to accurately represent the known directional advection process in the underlying system, highlighting limitations of this approach for local causal structures in space-time systems.

G Additional experimental details for Section 3.6

CaStLe inherits several of the runtime parameters of the underlying PIP used. In Section 3.6, we set these values at relatively stringent threshold to highlight the most robust and important dynamics and to yield a highly interpretable graph; additional weaker dynamics can be recovered by relaxing these choices at the (potential) cost of additional false positive edges and less interpretability. Data-driven optimization of these parameters is difficult, though the validation strategies suggested by Allen et al. (2023) may be useful here. Specifically, we set a p-value threshold of 1×10^{-5} and removed estimated partial correlations of magnitude less than 0.35; we note here that, due to the adaptive search heuristics used by the PIP, the p-value threshold applied here is not a proper measure of statistical significance, but only a *heuristic* measure of estimated strength. We note that our resulting interpretations are generally quite robust to specific choices of these values.

H Analysis of Spatial Blocking

Here, we briefly investigate two impacts of spatial blocking, of the kind used in Section 3.6. Spatial blocking is a process in which regions of the global space are separated into blocks where CaStLe is applied individually and independently. This can be done for the sake of interpretability and to help ensure the spatial causal structure is uniform and homogeneous in the blocked space, satisfying Assumption

S2.

First, we consider the impact of block size on the HSW-V case study. In our demonstration in Section 6.7.1, we approached block size heuristically, and we chose a relatively large block size to demonstrate correctness saliently. We found that results are generally robust to larger and smaller block sizes in the HSW-V case. In Figure H1, we show that the recovered dynamics in each stencil are generally the same over space for each block size. We see that larger block sizes are easier to interpret at a glance, while smaller sizes describe more nuance. We also found that results were generally robust to block size in the E3SMv2-SPA case.

Second, we consider the impact of a blocking strategy for causal discovery generally by comparing results of the PC algorithm to one block in E3SMv2-SPA to CaStLe-PC's results from the same data. Our comparison of CaStLe and the PC algorithm in Figure 4 make it clear that CaStLe captures the spatial evolution of Mt. Pinatubo's plume much more effectively and about 80,000 times faster. However, one may be concerned that sparsity and correctness could be achieved with blocking alone. In Figure H2a, PC struggles to estimate an interpretable and physically meaningful graph of the dependence structure in this area because of the signal redundancy between nonadjacent grid cells and that there are only 20 observations per grid cell and 25 grid cells. Figure H2b illustrates much better performance from CaStLe, in which CaStLe learns a stencil from the region and projects it back into the original grid space.

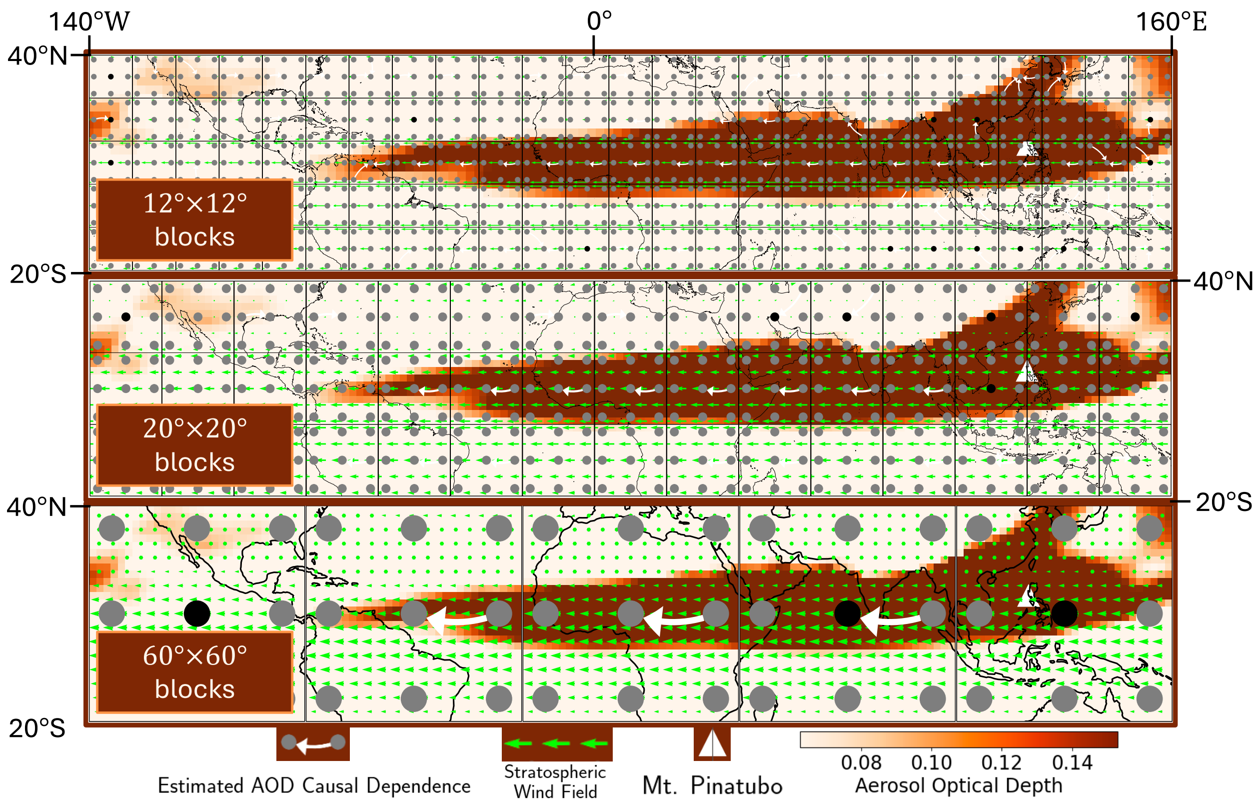
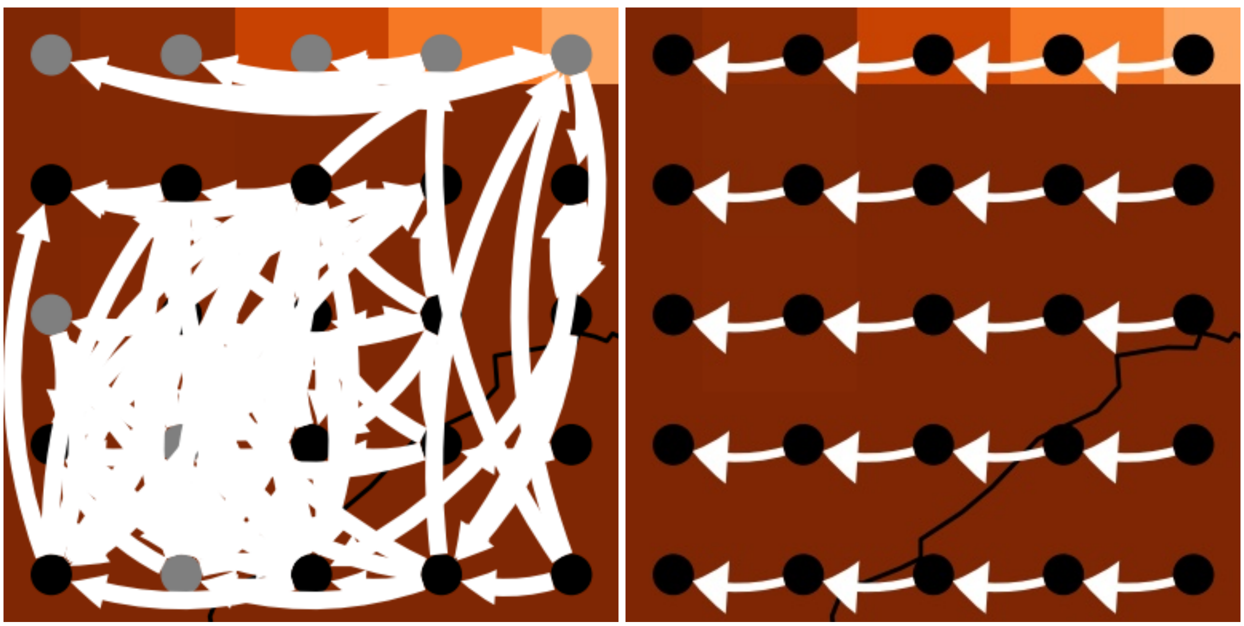


Figure H1: Results of CaStLe applied to HSW-V 21 days after the Mt. Pinatubo eruption with three different block sizes, $12^\circ \times 12^\circ$, $20^\circ \times 20^\circ$, and $60^\circ \times 60^\circ$. We find that results are generally consistent over the same area for each block size, with smaller block sizes allowing for additional nuance in some areas. Note that the $20^\circ \times 20^\circ$ block panel is similar to the results shown in Figure 3, but more longitudes were added to get a space factorable by more integers, such as 12, 20, and 60.



(a) PC algorithm results (b) CaStLe results

Figure H2: The PC algorithm and CaStLe applied to E3SMv2-SPA in the $15° \times 15°$ block between 15.00° to 30.00°N and 75° to 90°E. from the day of the eruption to 20 days later. PC struggles to estimate an interpretable and physically meaningful graph of the dependence structure in this area. In contrast, CaStLe is able to identify an interpretable dependence structure that represents the local dynamics within the space.

I Analysis of Assumption Violation Examples

Here, we evaluate the impacts of potential violations of CaStLe's assumptions in our study of E3SMv2-SPA from Section 6.7.2.

I.1 Time Resolution is Too Coarse (Assumption T1)

The dataset's time resolution can determine if the temporal locality assumption (T1) holds. If the time resolution is too coarse, the temporal causal structures may be marginalized out or unmeasured. Dependencies between neighboring grid cells may not be manifested in the sparse time sampling. Here, we explore how our study of E3SMv2-SPA from Section 6.7.2 changes after coarsening the temporal resolution.

We coarsened the time resolution by two, from a daily to a two-daily resolution.

Figure I1 demonstrates that CaStLe finds much fewer links when the time resolution is too coarse. However, the links that are detected are mostly consistent with known advective processes.

I.2 Time Interval is Too Long (Assumption T2)

When the time interval is too long, there may be too many causal structures in the data. This violates temporal causal stationarity (T2). Here, we investigate such a scenario.

We first computed causal stencils for an extended period, between day 15, the day of the eruption, to day 65. This is 30 days longer than our initial analysis from

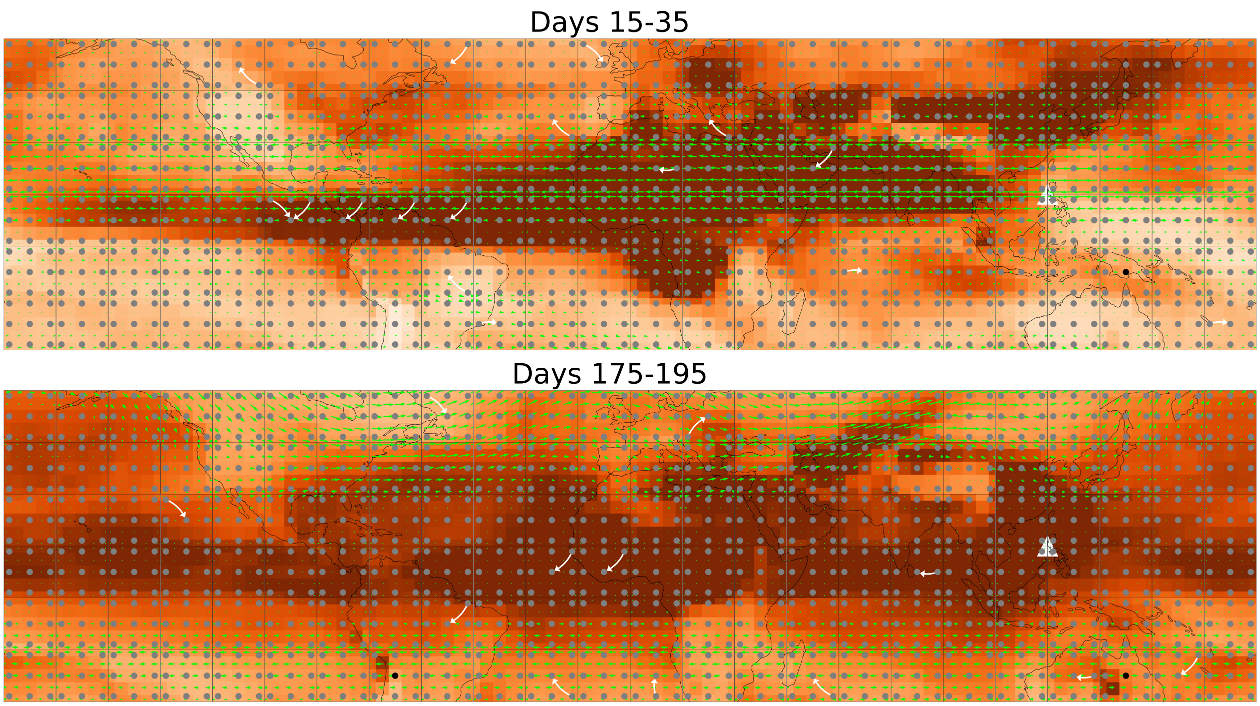


Figure I1: Results of using a coarsened temporal resolution (two-daily) in the E3SMv2-SPA study. CaStLe finds many fewer links in this setting. It is clear that when time is too coarse, causal structures fail to be detected. However, the remaining links that are found are largely true positives, suggesting that CaStLe is relatively robust to coarser time sampling.

the start of the eruption.

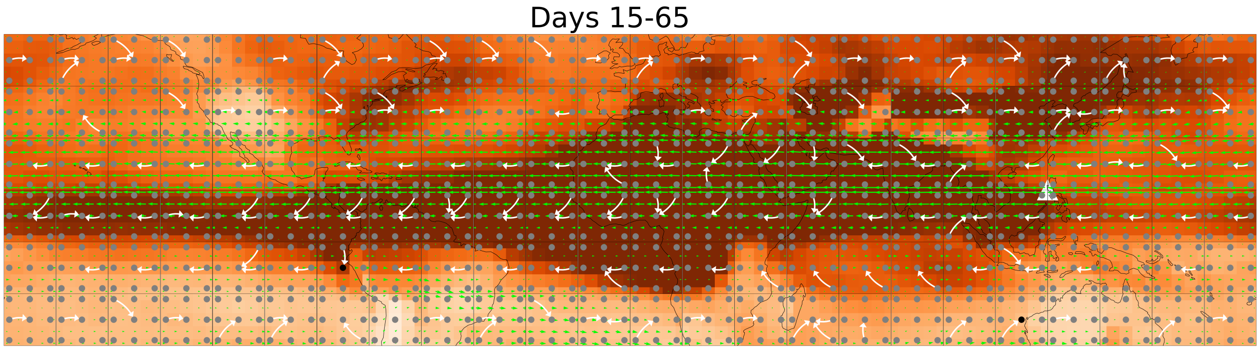


Figure I2: Results of applying CaStLe to a longer time interval from day 15 to 65. CaStLe identifies more links, indicating it is learning too many causal structures in the data, but still finds many of the true positives we found in our initial study. This indicates that many of the blocks in this interval have temporal causal stationarity, leading CaStLe to perform adequately.

We then computed causal stencils for the entire period between day 15 to day 215, roughly six months later.

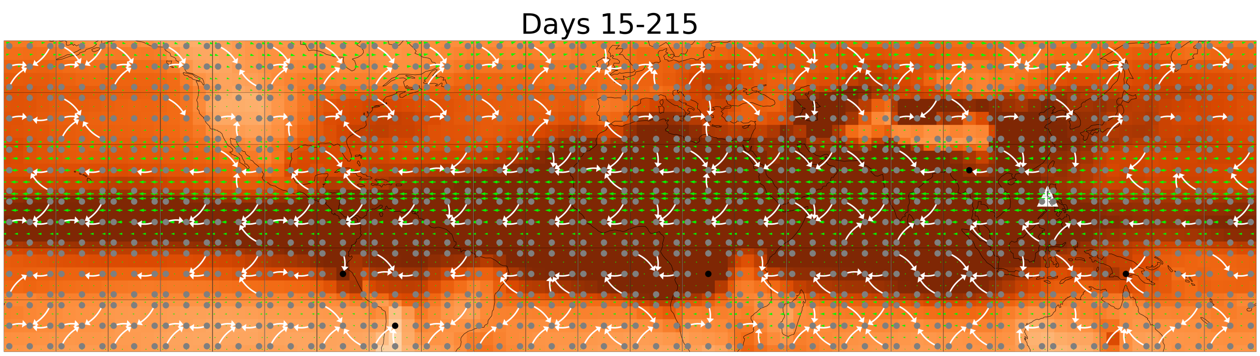


Figure I3: Results of applying CaStLe to a time interval that is too long and contains too many causal structures, day 15 to 200. We see that CaStLe identifies many links in each block. Comparing them to the winds is ineffective because the wind arrows are averages over the whole period rather than reflections of how they change in time, which CaStLe is learning from. With such a density of links, it is further challenging to know which are correct and which are spurious.

Figure I2 shows that when the time interval is longer, CaStLe identifies more links, indicating it is learning too many causal structures in the data, but still finds many of the true positives we found in our initial study. Figure I3 demonstrates the challenges of applying CaStLe to a time interval that contains too many difference causal structures. CaStLe identifies many links, creating uninterpretable stencils. The winds are a poor comparison because each arrow is a temporal average for that location, which is not representative over the entire interval. CaStLe may be capturing many spurious links or capturing all of the many fluctuating dynamics over the interval. Resulting is are uninterpretable stencils with unknown true and false positives. However, there are some blocks in the equatorial regions with sparse stencils. That indicates that dynamics were relatively stationary over the period.

I.3 Grid Resolution is Too Coarse (Assumption S1)

An appropriate grid resolution is important for satisfying the spatial locality assumption (S1). If the grid is too coarse then the underlying spatial structure may be marginalized out or unmeasured. If it is too small, causal relationships may appear outside the stencil neighborhood, requiring a radius-2 neighborhood implementation. Here, we investigate a grid resolution that is too coarse.

We coarsened the grid to 9°, rather than the 3° used in Section 6.7.2. Given that, to maintain 5×5 grid cells per block, each block is again $45^\circ \times 45^\circ$.

In Figure I4, we see that CaStLe performs very well overall. There are few false positives and it clearly captures the overall advection dynamics of the system.

We also coarsened the grid to 18°, resulting in $90^\circ \times 90^\circ$ blocks. In Figure I5, CaStLe performs well in the early time interval, clearly identifying the east-to-west advection pattern. However, in the later time interval, it finds no spatial structures

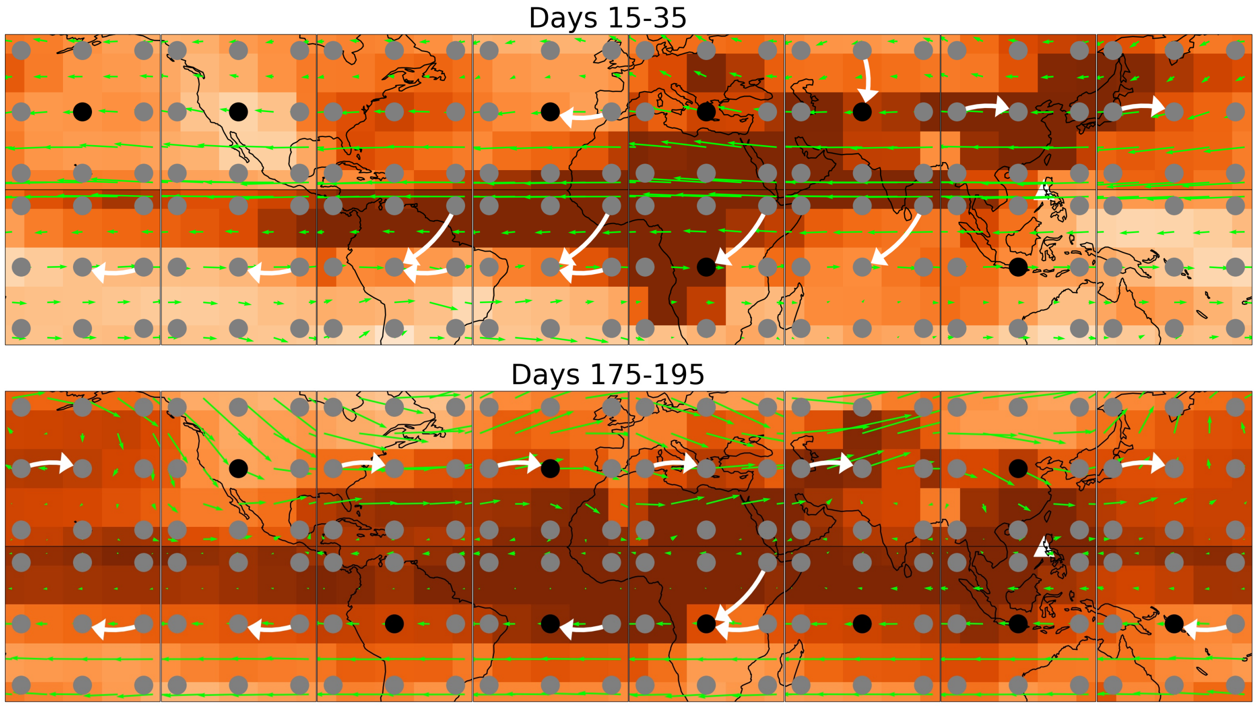


Figure I4: Results of using a coarse grid (9°) in the E3SMv2-SPA study. We find that CaStLe performs very well overall. There are few false positives and it clearly captures the overall advection dynamics of the system.

apart from autodependencies in each block. This is likely because the east-to-west advection is weaker in this period and the grid is too coarse to capture the narrower bands of northward advection that dominates the interval.

We find that CaStLe is very robust to this assumption violation. It captures all of the most dominant advection patterns, while struggling to find smaller, weaker ones.

I.4 Block Sizes are Too Large (Assumption S2)

In H, we found that CaStLe's output was robust to very large and very small block sizes. Spatial blocks are intended to isolate regions such that only one underlying spatial causal structure exists in the block. If the blocks are too large, then As-

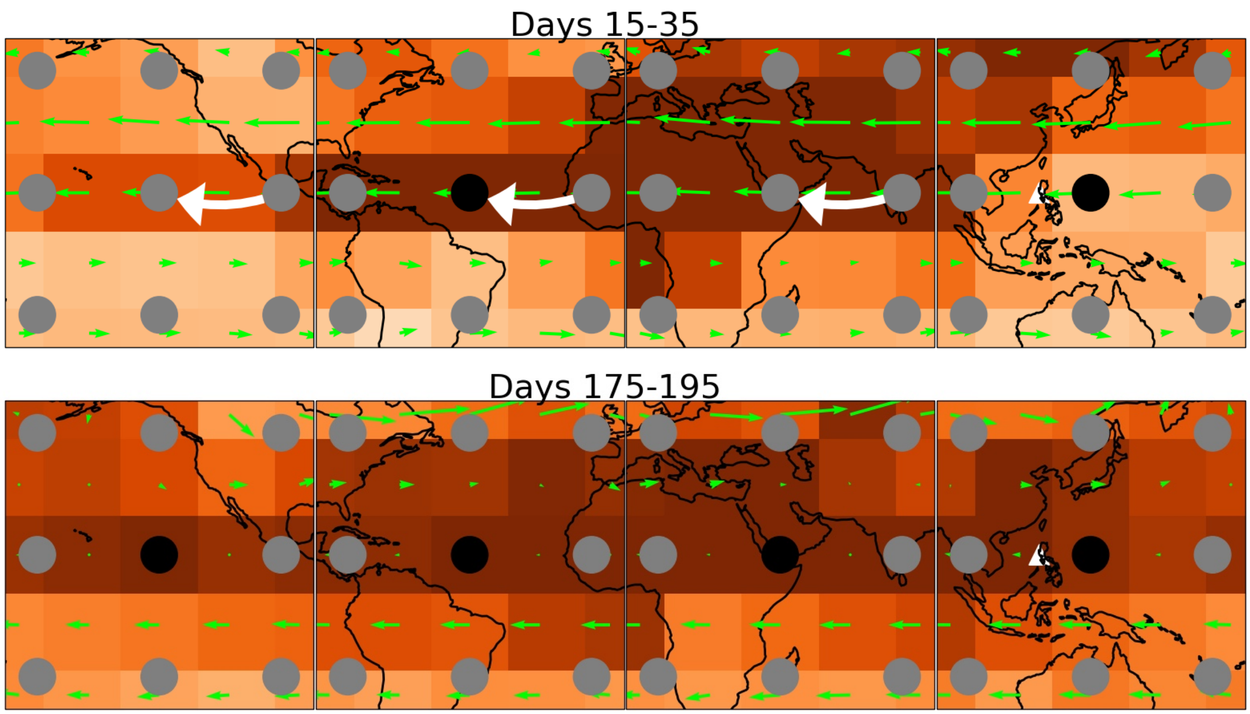


Figure I5: Results of using a coarse grid (18°) in the E3SMv2-SPA study. CaStLe performs well in the early time interval, clearly identifying the east-to-west advection pattern. However, in the later time interval, it finds no spatial structures apart from autodependencies in each block. This is likely because the east-to-west advection is weaker in this period and the grid is too coarse to capture the narrower bands of northward advection that dominates the interval.

sumption S2 may be violated.

In Figure I6, we used block sizes equal to $45^\circ \times 45^\circ$. Here, each block has 15×15 grid cells. This is in contrast to the 5×5 grid cell, $15^\circ \times 15^\circ$ blocks used in Section 6.7.2.

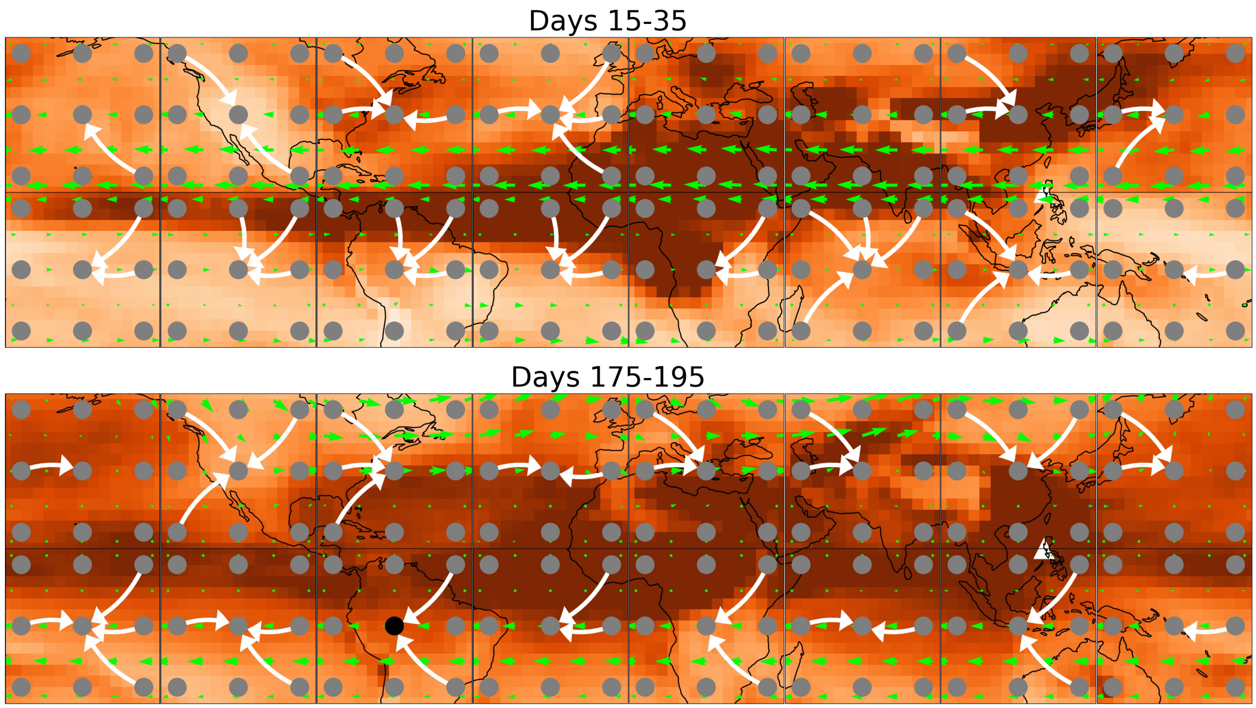


Figure I6: Results of using block sizes too large in the E3SMv2-SPA study. We see that many true positives are found, but many false positives as well. CaStLe seems to identify multiple contradictory causal structures within many cells, which may lead to more spurious links discovered. Even where links appear correct, they are largely uninterpretable in the presence of contradictions.

We find that while true positives remain, several false positives appear. Some positives may be the result of identifying multiple causal structures correctly within the space, while others may be confused results found because of the high density of links. In further testing with intermediate block sizes, we found that CaStLe is moderately robust to this assumption violation. As block sizes approach a more appropriate size, false positives diminish and true positives remain.

J Additional GCM Results

Figure J1 depicts results of implementing CaStLe with the Bayesian score optimization causal discovery algorithm, DYNOTEARS. We also presented results of DYNOTEARS applied to our VAR benchmark in Section 6.8.1. Here, we show that CaStLe-DYNOTEARS is able to recover comparable results to the CaStLe-PC-stable results shown in Section 6.7.1.

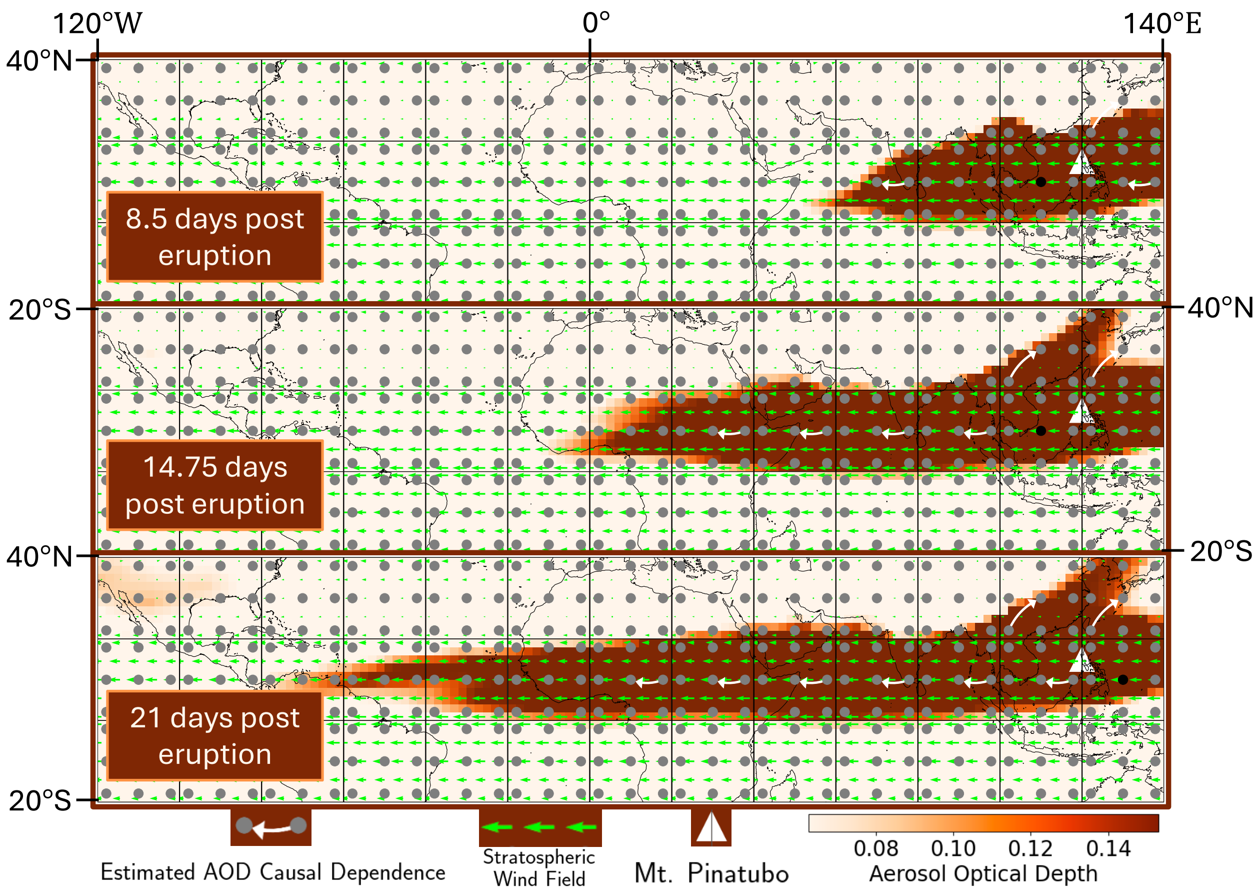


Figure J1: Application of CaStLe-DYNOTEARS to HSW-V simulation of the 1991 Mt. Pinatubo eruption. The stencils estimated by CaStLe (white) capture the underlying high-altitude wind fields (green) using only satellite-measured AOD, with near perfect accuracy in high aerosol regions (red-orange). On longer horizons (bottom row), CaStLe is able to recover equatorial wind currents as far away as South America, half-way around the world from Mt. Pinatubo (white triangle). CaStLe accurately identifies the prevailing westerly atmospheric winds because it was able to identify the space-time dependence between neighboring grid cells.

K Additional VAR Results

In Section 6.8.1, we demonstrated the strong performance of CaStLe on VAR-generated space-time data with fixed sparsity level $d = 4$; in particular, CaStLed variants uniformly improve over the performance of equivalent unstructured causal discovery algorithms. We repeat this analysis for a variety of sparsity levels in Figures K1 and K2 for the MCC and F_1 score similarity metrics, respectively. As in Figure 6.6, the CaStLed variants continue to significantly outperform across all sparsity levels, d; furthermore, as noted above, we observe that CaStLe can correctly estimate the underlying grid even on as few as $T = 10$ time samples when a sufficiently large grid is observed; non-CaStLe methods struggle on larger grid sizes, consistent with our analyses in the previous section. A time limit of 48 hours of wall-clock time was applied for each individual graph estimation: performance properties of methods that did not terminate during this window are not shown (*e.g.*, DYNOTEARS with $d = 6; T = 10; N = 10$).

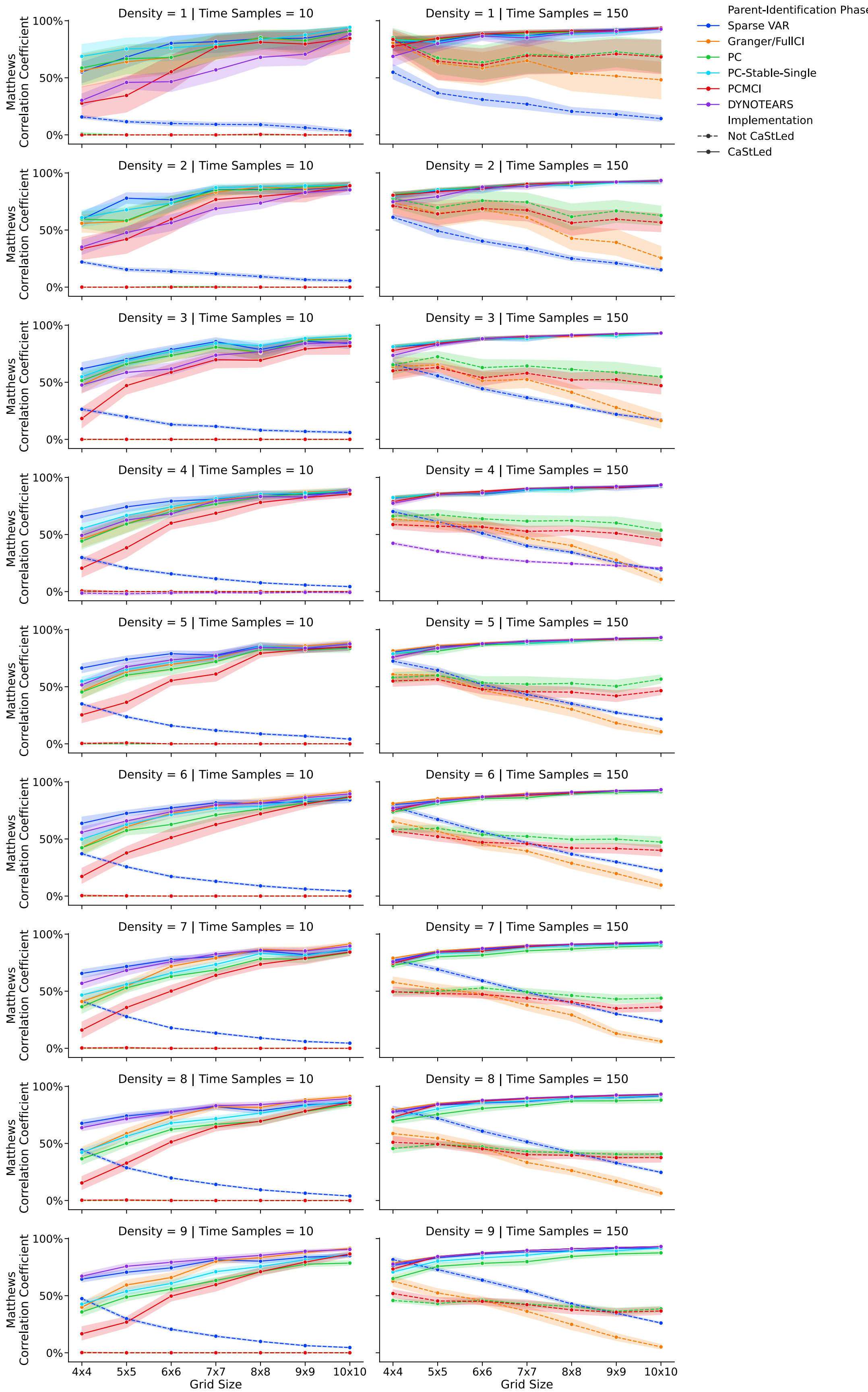


Figure K1: Matthews correlation coefficient (MCC) comparison between CaStLed and non-CaStLed causal discovery approaches on 2D VAR dynamics for each sparsity level, including Granger causality (orange), PC (green), PC-Stable-Single (cyan), PCMCI (red), DYNOTEARS (purple), and a statistical model of the data generating process (blue). See Section 6.8.1 for experi-

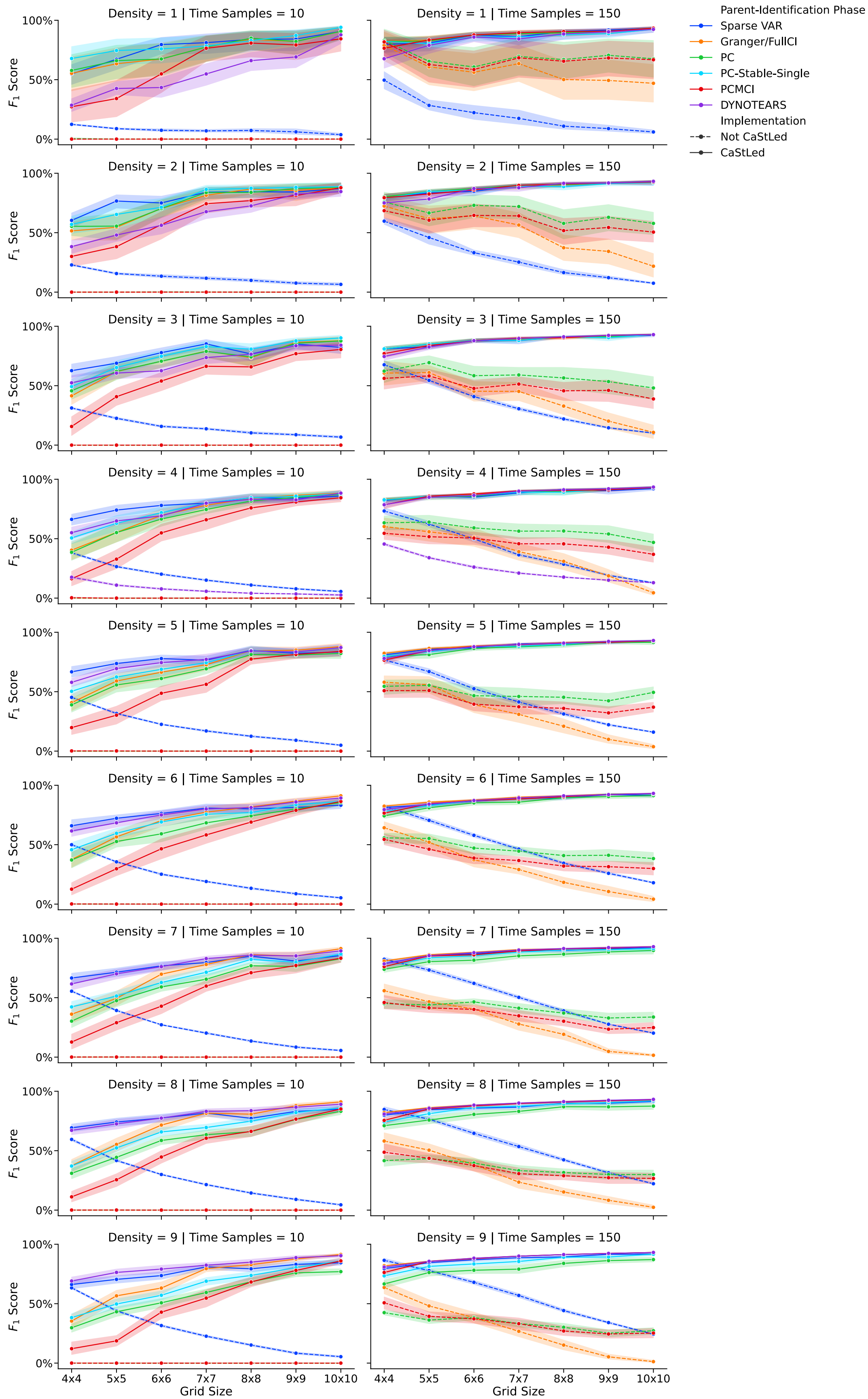


Figure K2: F_1 score comparison between CaStLed and non-CaStLed causal discovery approaches on 2D VAR dynamics for each sparsity level, including Granger causality (orange), PC (green), PC-Stable-Single (cyan), PCMCI (red), DYNOTEARS (purple), and a statistical model of the data generating process (blue). See Section 6.8.1 for experimental details.

L PC-Stable-Single

For the convenience of the reader, we include pseudo-code for the PC-Stable-Single algorithm of Runge et al. (2019a), itself an adaptation of the PC-Stable algorithm of Colombo and Maathuis (2014). We use this as the PIP used for the CaStLe-based analyses shown in Sections 6.7.1, 6.7.2, and D. As our experiments in the proceeding section show, PC-Stable-Single exhibits small, but consistent improvements over alternative PIP choices.

M Completed Data Generation Parameters

As noted in Section 7.5.1, not all parameter combinations generated stable systems. Here, we present the parameter ranges that did successfully generate 30 replicates to produce out results. We additionally evaluate the range of coefficient sizes generated, demonstrating the difficulty of creating complex systems with strong signals and many interdependencies.

Parameter ranges used in our experimental design, showing the link count distribution for each grid size and variable count combination. Each horizontal line represents the span of network links tested, with each parameter combination having at least 30 replicate experiments (n values shown). Our experiments covered grid sizes from 4×4 to 10×10 and 1-6 variables per grid. All experiments used 1000 time samples and coefficient values between 0.1 and 1.0. The network density, defined as the ratio of actual links (L) to maximum possible links in a 3×3

Algorithm 2 PC-stable-single

Precondition: Time series dataset $\mathbf{X} = \{X^1, X^2, ..., X^N\}$, selected variable X^j, maximum time lag τ_{max} (default $\tau_{max} = 1$), significance threshold α_{PC}, maximum condition dimension p_{max} (default $p_{max} = N\tau_{max}$), maximum number of combinations q_{max} (default $q_{max} = 1$), conditional independence test function I.

function CI($X, Y, \mathbf{Z}$)
    Test $X \perp\!\!\!\perp Y | \mathbf{Z}$ using test statistic measure I
    return p-value, test statistic value I
Initialize set of parents $\widehat{\mathscr{P}}(X_t^j) = \{X_{t-\tau}^i : i \in \{1, ..., N\}, \tau \in \{1, ..., \tau_{max}\}\}$
Initialize dictionary of test statistic values $I^{min}(X_{t-\tau}^i \to X_t^i) = \infty \; \forall X_{t-\tau}^i \in \widehat{\mathscr{P}}(X_t^j)$
for $p = 0, ..., p_{max}$ **do**
    if $|\widehat{\mathscr{P}}(X_t^j)| - 1 < p$ **then**
        Break for-loop ▷ Algorithm has converged
    for all $X_{t-\tau}^i$ in $\widehat{\mathscr{P}}(X_t^j)$ **do**
        $q = -1$
        for all lexicographically chosen subsets $\mathscr{S} \subseteq \widehat{\mathscr{P}}(X_t^j) \setminus \{X_{t-\tau}^i\}$, with $|\mathscr{S}| = p$ **do**
            $q = q + 1$
            if $q >= q_{max}$ **then**
                Break from inner for-loop
            Run CI test to obtain (p-value, I) $\leftarrow CI(X_{t-\tau}^i, X_t^i, \mathscr{S})$
            if $|I| < I^{min}(X_{t-\tau}^i \to X_t^i)$ **then** ▷ Store min. I of parent among all tests
                $I^{min}(X_{t-\tau}^i \to X_t^i) = I$
            if p-value $> \alpha_{PC}$ **then** ▷ Removed only after all $X_{t-\tau}^i$ have been tested
                Mark $X_{t-\tau}^i$ for removal from $\widehat{\mathscr{P}}(X_t^i)$
                Break from inner loop
    Remove non-significant parents from $\widehat{\mathscr{P}}(X_t^i)$
    Sort parents in $\widehat{\mathscr{P}}(X_t^i)$ by $I^{min}(X_{t-\tau}^i \to X_t^i)$ from largest to smallest
return $\widehat{\mathscr{P}}(X_t^i)$

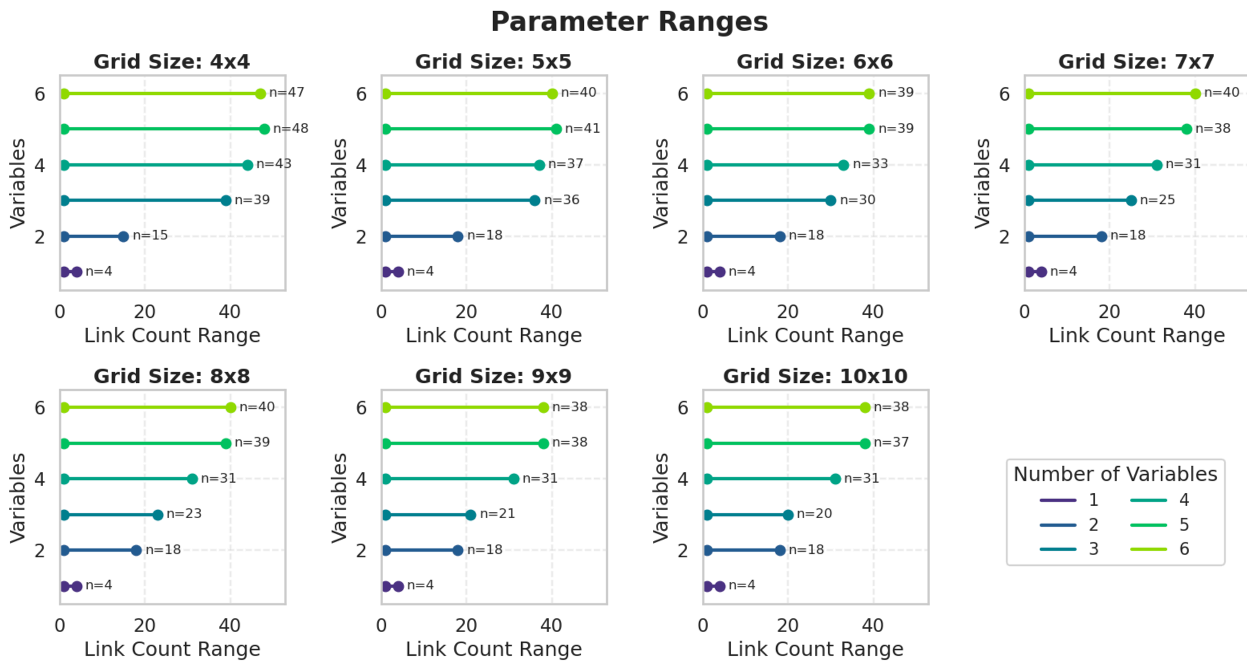


Figure M1: Parameter ranges used in our experimental design, showing the link count distribution for each grid size and variable count combination. Each horizontal line represents the span of network links tested, with each parameter combination having at least 30 replicate experiments (n values shown). Our experiments covered grid sizes from 4×4 to 10×10 and 1-6 variables per grid. All experiments used 1000 time samples and coefficient values between 0.1 and 1.0. The network density, d, defined as the ratio of actual links, L, to maximum possible links $d = \frac{L}{(3\times3\times V^2)}$, where $d \in (0, \ldots 0.5]$. Not all density values produced 30 stable systems within our computational constraints, particularly at higher densities. This visualization shows which parameter combinations successfully generated sufficient replicates for statistical analysis.

stencil graph (d = L/(3×3×V²)), ranged from near zero to 0.5. Not all theoretical density values produced 30 stable systems within our computational constraints, particularly at higher densities. This visualization shows which parameter combinations successfully generated sufficient replicates for statistical analysis.

N Additional VAR Results

In this appendix, we present additional results related to the performance of our proposed method, M-CaStLe, with VAR benchmarks. We delve into various met-

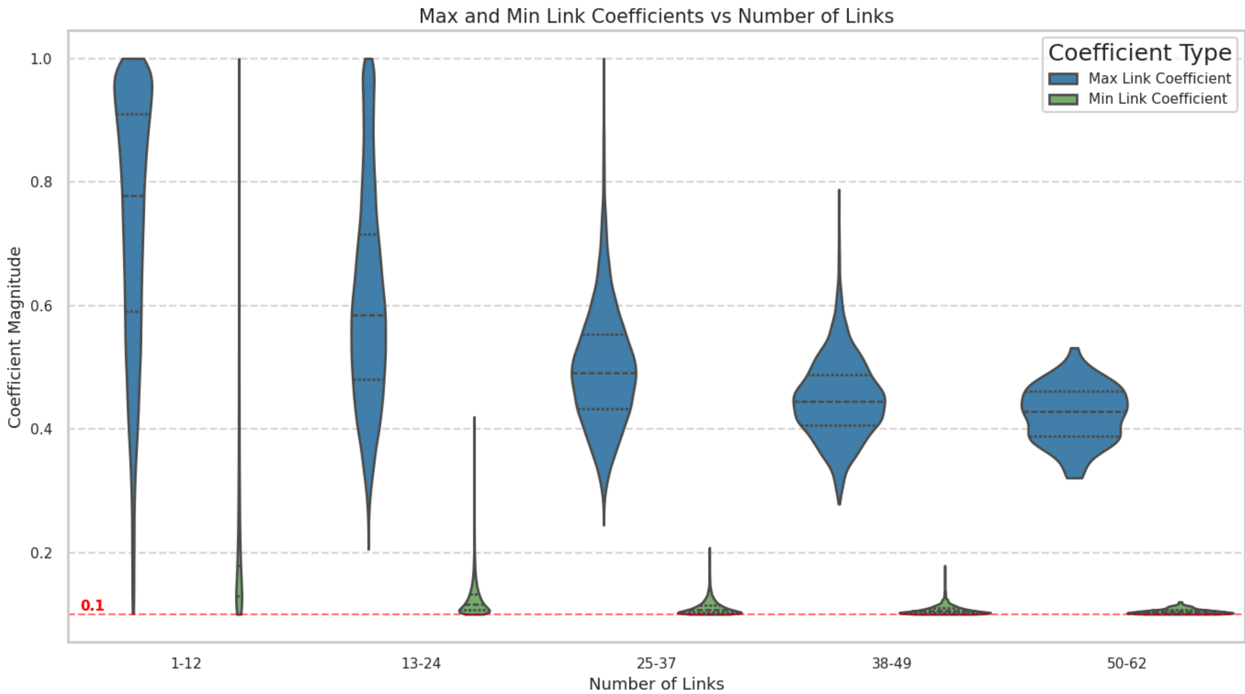


Figure M2: The relationship between link coefficients and the number of links present. As the number of links increases, maximum (blue) and minimum (green) link coefficients show a clear decreasing trend, with their distribution becoming narrower and centered around lower values. This reveals that networks with more links have weaker signals, suggesting that highly interconnected systems cannot be stable with large dependencies.

rics that evaluate the effectiveness of M-CaStLe.

N.1 PC Comparison Results

We examined the impact of the number of variables on key performance indicators such as F_1 score, precision, and recall. We provide a comparison between M-CaStLe-PC and the time series PC algorithm. This analysis, illustrated in Figure N3, emphasizes how M-CaStLe-PC consistently outperforms PC across various scenarios, particularly as the number of links increases in a 4×4 grid. The results underscore the challenges faced by PC in high-dimensional environments, where its naive approach to spatial and variable structures limits its effectiveness.

N.2 Hyperparameter Testing

To better understand the effect of M-CaStLe's hyperparameters, we conducted a sensitivity analysis, which is presented in Figure N4. We find that most metrics are moderately impacted by both hyperparameters, except for recall. It appears that recall reaches a limit, which is consistent with our findings in Section 7.5.1.

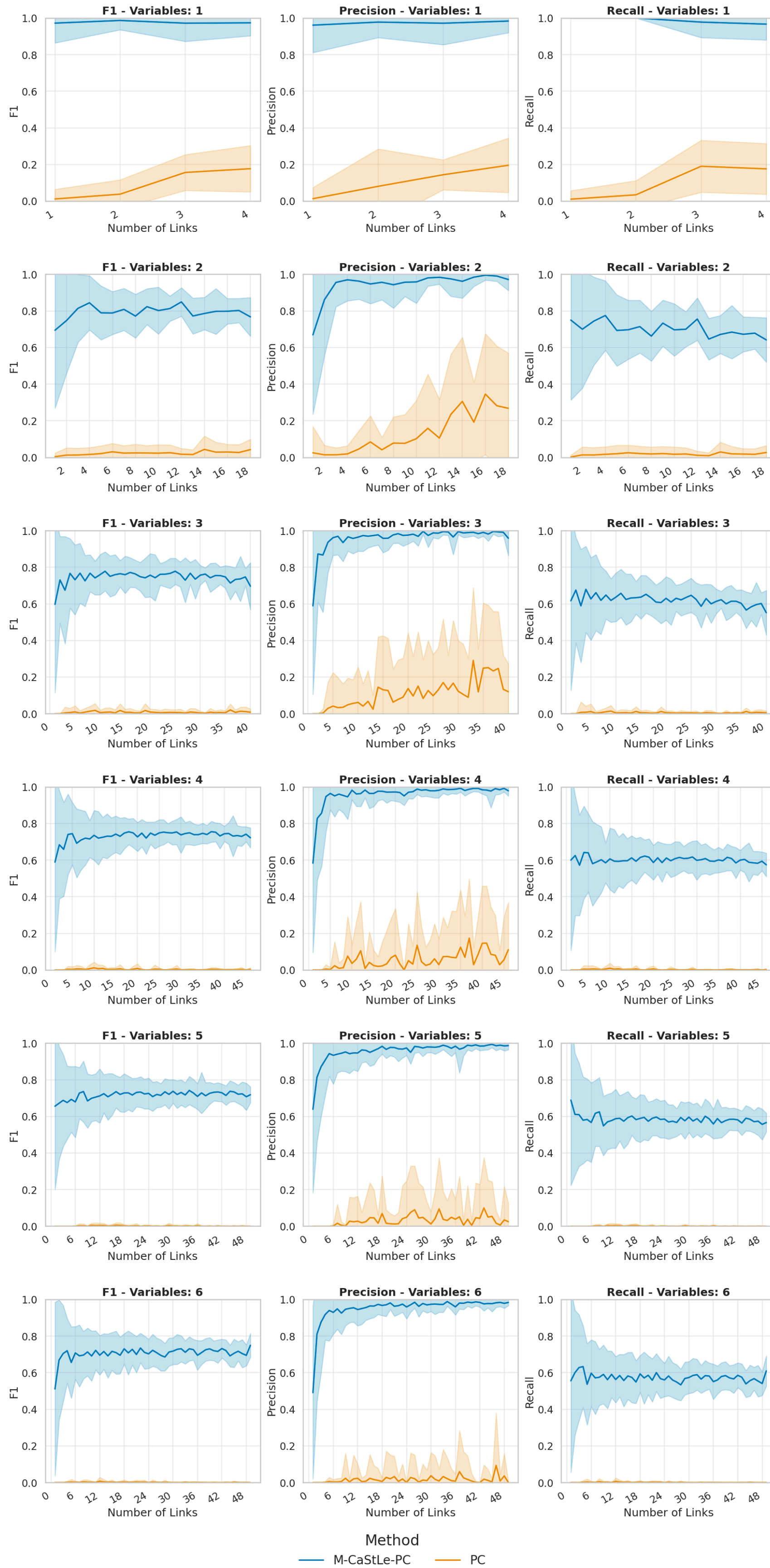


Figure N3: Comparisons between M-CaStLe-PC and PC considering the F_1 score, precision, and recall for all V as the number of links increases on a 4×4 grid. M-CaStLe-PC outperforms PC in every case because PC struggles with the very high dimensionality of the systems since it is naive to the spatial and variable structures.

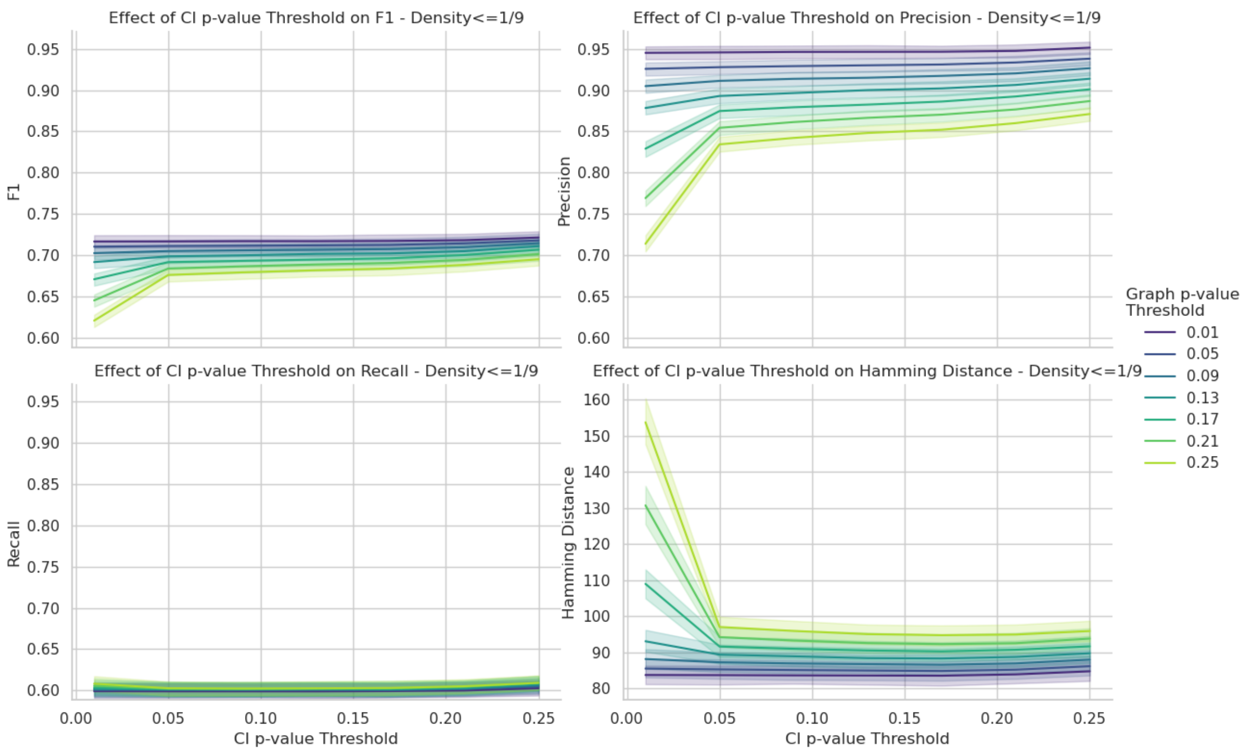


Figure N4: Shows that M-CaStLe's hyperparameters influence all metrics except recall. This indicates that the poorer recall performance cannot be explained by M-CaStLe's hyperparameters.

Bibliography

R. McGill, J. W. Tukey, and W. A. Larsen, “Variations of box plots,” *The American Statistician*, vol. 32, no. 1, pp. 12–16, 1978.

R. E. Bellman, *Dynamic Programming*. Princeton, NJ: Princeton University Press, 1957, introduced the term “curse of dimensionality” in the preface: “All this may be subsumed under the heading ’the curse of dimensionality.’ Since this is a curse which has hung over the head of the physicist and astronomer for many a year...”.

P. Bühlmann and S. v. d. Geer, “Statistics for High-Dimensional Data, Methods, Theory and Applications,” *Springer Series in Statistics*, pp. 99–182, 2011.

J. Pearl and D. Mackenzie, *The Book of Why*. New York: Basic Books, 2018.

J. Peters, D. Janzing, and B. Schlkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, Massachusetts: The MIT Press, 2017.

J. J. Nichol, M. G. Peterson, K. J. Peterson, G. M. Fricke, and M. E. Moses, “Machine learning feature analysis illuminates disparity between E3SM climate models and observed climate change,” *Journal of Computational and Applied Mathematics*, vol. 395, p. 113451, 10 2021.

M. A. Hernán and J. M. Robins, *Causal Inference: What if*. Crc Press, 2020.

D. B. Rubin, “Estimating causal effects of treatments in randomized and nonrandomized studies,” *Journal of Educational Psychology*, vol. 66, no. 5, pp. 688–701, 1974.

J. Robins, “A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect,” *Mathematical Modelling*, vol. 7, no. 9, pp. 1393–1512, 1986. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0270025586900886

H. F. Dorn, “Philosophy of Inferences from Retrospective Studies,” *American Journal of Public Health and the Nations Health*, vol. 43, no. 6_Pt_1, pp. 677–683, 1953.

A. R. Feinstein, “Clinical biostatistics,” *Clinical Pharmacology & Therapeutics*, vol. 11, no. 2, pp. 282–292, 1970.

A. P. Dawid, “Causal Inference Without Counterfactuals,” *Journal of the American Statistical Association*, vol. 95, no. 450, p. 407, 2000.

S. Wright, “Correlation and Causation,” *Journal of Agricultural Research*, vol. 20, 1921. [Online]. Available: https://books.google.co.ao/books/about/Journal_of_Agricultural_Research.html?hl=pt-PT&id=lNNdIV_qpwIC&utm_source=gb-gplus-shareJournal

P. Spirtes, C. Glymour, and R. Scheines, "Causation, Prediction, and Search," *Lecture Notes in Statistics*, 1993.

J. Pearl, "Causal Diagrams for Empirical Research," *Biometrika*, vol. 82, no. 4, p. 669, 1995.

L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: http://dx.doi.org/10.1023/A%3A1010933404324

S. Nembrini, I. R. König, and M. N. Wright, "The revival of the Gini importance?" *Bioinformatics*, vol. 34, no. 21, pp. 3711–3718, 2018.

S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 31*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: https://dl.acm.org/doi/10.5555/3295222.3295230

B. Krishnapuram, M. Shah, A. Smola, C. Aggarwal, D. Shen, R. Rastogi, M. T. Ribeiro, S. Singh, and C. Guestrin, "“Why Should I Trust You?”," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, ser. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

A. Shrikumar, P. Greenside, and A. Kundaje, "Learning Important Features through Propagating Activation Differences," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17. JMLR.org, 2017, p. 3145–3153.

J. Pearl, "The Do-Calculus Revisited," *arXiv*, 2012.

——, "FROM BAYESIAN NETWORKS TO CAUSAL NETWORKS," *Mathematical Models for Handling Partial Knowledge in Artificial Intelligence*, pp. 157–182, 1995.

P. Spirtes and C. Glymour, "An Algorithm for Fast Recovery of Sparse Causal Graphs," *Social Science Computer Review*, vol. 9, no. 1, pp. 62–72, 1991.

J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, "Detecting and quantifying causal associations in large nonlinear time series datasets," *Science Advances*, vol. 5, no. 11, pp. 4996—5023, 2019. [Online]. Available: http://advances.sciencemag.org/

J. Runge, "Causal network reconstruction from time series: From theoretical assumptions to practical estimation," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 28, no. 7, p. 075310, 2018.

J. Arnhold, P. Grassberger, K. Lehnertz, and C. Elger, "A robust method for detecting interdependences: application to intracranially recorded EEG," *Physica D: Nonlinear Phenomena*, vol. 134, no. 4, pp. 419–430, 1999.

G. Sugihara, R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, and S. Munch, "Detecting Causality in Complex Ecosystems," *Science*, vol. 338, no. 6106, pp. 496–500, 2012. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1227079

P. Spirtes and K. Zhang, "Causal discovery and inference: concepts and recent methodological advances," *Applied Informatics*, vol. 3, no. 1, p. 3, 2016.

J. Pearl, "Graphs, Causality, and Structural Equation Models," *Sociological Methods & Research*, vol. 27, no. 2, pp. 226–284, 1998. [Online]. Available: https://doi.org/10.1177/0049124198027002004

J. Runge, "Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets," in *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, ser. Proceedings of Machine Learning Research, vol. 124. PMLR, 2020, pp. 1388–1397. [Online]. Available: https://proceedings.mlr.press/v124/runge20a.html

C. Glymour, K. Zhang, and P. Spirtes, "Review of Causal Discovery Methods Based on Graphical Models," *Frontiers in Genetics*, vol. 10, p. 524, 2019.

J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, J. Munoz-Mari, E. H. v. Nes, J. Peters, R. Quax, M. Reichstein, M. Scheffer, B. Scholkopf, P. Spirtes, G. Sugihara, J. Sun, K. Zhang, and J. Zscheischler, "Inferring causation from time series in Earth system sciences," *Nature Communications*, vol. 10, no. 2553, 2019.

S. Guo, G. J. S. Bluth, W. I. Rose, I. M. Watson, and A. J. Prata, "Re-evaluation of SO2 release of the 15 June 1991 Pinatubo eruption using ultraviolet and infrared satellite sensors," *Geochemistry, Geophysics, Geosystems*, vol. 5, no. 4, 2004.

L. Gimeno, R. Nieto, M. Vázquez, and D. A. Lavers, "Atmospheric rivers: a mini-review," *Frontiers in Earth Science*, vol. 2, p. 2, 2014.

G. Wang, D. Zhong, T. Li, Y. Zhang, C. Meng, M. Zhang, X. Song, J. Wei, and Y. Huang, "Study on sky rivers: Concept, theory, and implications," *Journal of Hydro-environment Research*, vol. 21, pp. 109–117, 2018.

I. Ebert-Uphoff and Y. Deng, "Causal Discovery from Spatio-Temporal Data with Applications to Climate Science," *2014 13th International Conference on Machine Learning and Applications*, pp. 606–613, 2014.

R. Fisher, *Statistical Methods for Research Workers*, ser. Biological monographs and manuals. Oliver and Boyd, 1925. [Online]. Available: https://books.google.com/books?id=I0NBAAAAIAAJ

N. S. Hall, "R. A. Fisher and his advocacy of randomization," *Journal of the History of Biology*, vol. 40, no. 2, pp. 295–325, 2007.

J. Splawa-Neyman, D. M. Dabrowska, and T. P. Speed, "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," *Roczniki Nauk Rolniczych Tom X [in Polish]; translated in Statistical*

Science, vol. 5, no. 4, pp. 465–472, 1923. [Online]. Available: http://www.jstor.org/stable/2245382

D. B. Rubin, “Causal Inference Using Potential Outcomes: Design, Modeling, Decisions,” *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 322–331, 2005.

P. W. Holland, “Statistics and Causal Inference,” *Journal of the American Statistical Association*, vol. 81, no. 396, p. 945, 1986.

J. Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000. [Online]. Available: https://books.google.com/books?id=wnGU_TsW3BQC

——, “Causal inference in the health sciences: a conceptual introduction,” *Health services and outcomes research methodology*, vol. 2, no. 3, pp. 189–220, 2001.

P. Tarka, “An overview of structural equation modeling: its beginnings, historical development, usefulness and controversies in the social sciences,” *Quality & Quantity*, vol. 52, no. 1, pp. 313–354, 2018.

N. Wiener, “The theory of prediction,” *Modern mathematics for engineers*, 1956.

C. W. J. Granger, “Investigating Causal Relations by Econometric Models and Cross-spectral Methods,” *Econometrica*, vol. 37, no. 3, p. 424, 1969, granger Causality seminal paper.

H. Reichenbach, *The Direction of Time*, ser. Dover books on physics, M. Reichenbach, Ed. Dover Publications, 1956.

S. Shimizu, P. O. Hoyer, A. Hyvarinen, and A. Kerminen, “A Linear Non-Gaussian Acyclic Model for Causal Discovery,” *Journal of Machine Learning Research*, vol. 7, no. 72, p. 2003-2030, 2006. [Online]. Available: https://www.jmlr.org/papers/volume7/shimizu06a/shimizu06a.pdf

J. D. Ramsey, “A Scalable Conditional Independence Test for Nonlinear, Non-Gaussian Data,” *arXiv*, vol. abs/1401.5031, 2014. [Online]. Available: http://arxiv.org/abs/1401.5031

K. Zhang, J. Peters, D. Janzing, and B. Schölkopf, “Kernel-Based Conditional Independence Test and Application in Causal Discovery,” in *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, ser. UAI’11. Arlington, Virginia, USA: AUAI Press, 2011, p. 804–813.

J. Runge, “Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information,” vol. 84, pp. 938–947, 09–11 Apr 2018. [Online]. Available: https://proceedings.mlr.press/v84/runge18a.html

R. Sen, A. T. Suresh, K. Shanmugam, A. G. Dimakis, and S. Shakkettai, “Model-Powered Conditional Independence Test,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS’17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 2955–2965.

J. Houghton, Y. Ding, D. Griggs, M. Noguer, P. v. d. Linden, X. Dai, M. Maskell, and C. Johnson, "Climate Change 2001: The Scientific Basis," *Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change (IPCC)*, vol. 881., p. 881, 2001.

K. Hasselmann, "Multi-pattern fingerprint method for detection and attribution of climate change," *Climate Dynamics*, vol. 13, no. 9, pp. 601–611, 1997.

I. Ebert-Uphoff and Y. Deng, "A new type of climate network based on probabilistic graphical models: Results of boreal winter versus summer," *Geophysical Research Letters*, vol. 39, no. 19, 2012.

Z. S. Kaufman, N. Feldl, W. Weijer, and M. Veneziani, "Causal Interactions Between Southern Ocean Polynyas and High-Latitude Atmosphere-Ocean Variability," *Journal of Climate*, vol. 33, no. 11, pp. 4891–4905, 2020.

M. Kretschmer, D. Coumou, J. F. Donges, and J. Runge, "Using Causal Effect Networks to Analyze Different Arctic Drivers of Midlatitude Winter Circulation," *Journal of Climate*, vol. 29, no. 11, pp. 4069–4081, 2016.

P. Nowack, J. Runge, V. Eyring, and J. D. Haigh, "Causal networks for climate model evaluation and constrained projections," *Nature Communications 2020 11:1*, vol. 11, no. 1, pp. 1—11, 2020. [Online]. Available: http://www.nature.com/articles/s41467-020-15195-y

J. Runge, V. Petoukhov, and J. Kurths, "Quantifying the Strength and Delay of Climatic Interactions: The Ambiguities of Cross Correlation and a Novel Measure Based on Graphical Models," *Journal of Climate*, vol. 27, no. 2, pp. 720–739, 2014.

J. Runge, R. V. Donner, and J. Kurths, "Optimal model-free prediction from multivariate time series," *Physical Review E*, vol. 91, no. 5, p. 052909, 2015.

M. Vejmelka, L. Pokorná, J. Hlinka, D. Hartman, N. Jajcay, and M. Paluš, "Nonrandom correlation structures and dimensionality reduction in multivariate climate data," *Climate Dynamics*, vol. 44, no. 9-10, pp. 2663–2682, 2014.

V. Eyring, P. M. Cox, G. M. Flato, P. J. Gleckler, G. Abramowitz, P. Caldwell, W. D. Collins, B. K. Gier, A. D. Hall, F. M. Hoffman, G. C. Hurtt, A. Jahn, C. D. Jones, S. A. Klein, J. P. Krasting, L. Kwiatkowski, R. Lorenz, E. Maloney, G. A. Meehl, A. G. Pendergrass, R. Pincus, A. C. Ruane, J. L. Russell, B. M. Sanderson, B. D. Santer, S. C. Sherwood, I. R. Simpson, R. J. Stouffer, and M. S. Williamson, "Taking climate model evaluation to the next level," *Nature Climate Change*, vol. 9, no. 2, pp. 102–110, 2019, "Other promising diagnostic developments on the horizon that should be further advanced include studies that assess responses to perturbations rather than mean climate95, and the application of innovative data science methods in Earth system science96 such as neural networks97, machine learning-based anomaly detection techniques98, graphical models and causal discovery99.".

A. Hannachi, I. T. Jolliffe, and D. B. Stephenson, "Empirical orthogonal functions and related techniques in atmospheric science: A review," *International Journal of Climatology*, vol. 27, no. 9, pp. 1119–1152, 2007.

A. Gerhardus and J. Runge, "LPCMCI: Causal Discovery in Time Series with Latent Confounders," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 12 615–12 625. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/94e70705efae423efda1088614128d0b-Paper.pdf

A. Tsonis and P. Roebber, "The architecture of the climate network," *Physica A: Statistical Mechanics and its Applications*, vol. 333, pp. 497–504, 2004.

A. A. Tsonis, K. L. Swanson, and P. J. Roebber, "What Do Networks Have to Do with Climate?" *Bulletin of the American Meteorological Society*, vol. 87, no. 5, pp. 585–595, 2006.

X.-A. Tibau, C. Reimers, A. Gerhardus, J. Denzler, V. Eyring, and J. Runge, "A spatiotemporal stochastic climate model for benchmarking causal discovery methods for teleconnections," *Environmental Data Science*, vol. 1, p. e12, 2022.

B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward Causal Representation Learning," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 612–634, 2021.

J. Boussard, C. Nagda, J. Kaltenborn, C. E. E. Lange, P. Brouillard, Y. Gurwicz,

P. Nowack, and D. Rolnick, “Towards Causal Representations of Climate Model Data,” *arXiv*, 2023.

P. Brouillard, S. Lachapelle, J. Kaltenborn, Y. Gurwicz, D. Sridhar, A. Drouin, P. Nowack, J. Runge, and D. Rolnick, “Causal Representation Learning in Temporal Data via Single-Parent Decoding,” *arXiv*, 2024.

P. Pfleiderer, C.-F. Schleussner, T. Geiger, and M. Kretschmer, “Robust predictors for seasonal Atlantic hurricane activity identified with causal effect networks,” *Weather and Climate Dynamics*, vol. 1, no. 2, pp. 313–324, 2020.

I. Polkova, H. Afargan-Gerstman, D. I. V. Domeisen, M. P. King, P. Ruggieri, P. Athanasiadis, M. Dobrynin, O. Aarnes, M. Kretschmer, and J. Baehr, “Predictors and prediction skill for marine cold-air outbreaks over the Barents Sea,” *Quarterly Journal of the Royal Meteorological Society*, vol. 147, no. 738, pp. 2638–2656, 2021.

J. Y. Zhu, C. Zhang, H. Zhang, S. Zhi, V. O. Li, J. Han, and Y. Zheng, “pg-Causality: Identifying Spatiotemporal Causal Pathways for Air Pollutants with Urban Big Data,” *IEEE Transactions on Big Data*, vol. 4, no. 4, pp. 571–585, 2016.

P. Sheth, R. Shah, J. Sabo, K. S. Candan, and H. Liu, “STCD: A Spatio-Temporal Causal Discovery Framework for Hydrological Systems,” *2022 IEEE International Conference on Big Data (Big Data)*, vol. 00, pp. 5578–5583, 2022.

S. Saetia, N. Yoshimura, and Y. Koike, “Constructing Brain Connectivity Model Using Causal Network Reconstruction Approach,” *Frontiers in Neuroinformatics*, vol. 15, p. 619557, 2021.

E3SM Project, “Energy Exascale Earth System Model (E3SM),” [Computer Software] https://dx.doi.org/10.11578/E3SM/dc.20180418.36, Apr. 2018. [Online]. Available: https://dx.doi.org/10.11578/E3SM/dc.20180418.36

J. Stroeve and D. Notz, “Changing state of Arctic sea ice across all seasons,” sep 2018.

“Arctic report card 2019,” National Oceanic and Atmospheric Administration, Washington, DC, Tech. Rep., December 2019, released at the American Geophysical Union Fall Meeting in San Francisco, California, December 10, 2019. [Online]. Available: https://www.arctic.noaa.gov/Report-Card

L. C. Smith and S. R. Stephenson, “New trans-Arctic shipping routes navigable by midcentury,” *PNAS*, vol. 110, no. 13, pp. 4871–4872, 2013.

H. Goosse, J. E. Kay, K. C. Armour, A. Bodas-Salcedo, H. Chepfer, D. Docquier *et al.*, “Quantifying climate feedbacks in polar regions,” *Nature Communications*, vol. 9, no. 1919, 2018.

F. Sevellec, A. V. Fedorov, and W. Liu, “Arctic sea-ice decline weakens the atlantic meridional overturning circulation,” *Nature Climate Change*, vol. 7, pp. 604–610, 2017.

J. Cohen, K. Pfeiffer, and J. A. Francis, “Warm Arctic episodes linked with increased frequency of extreme winter weather in the United States,” *Nature Communications*, vol. 9, no. 869, 2018.

I. Cvijanovic, B. D. Santer, C. Bonfils, D. D. Lucas, J. C. H. Chiang, and S. Zimmerman, “Future loss of Arctic sea-ice cover could drive a substantial decrease in california’s rainfall,” *Nature Communications*, vol. 8, no. 1947, 2017.

E. Rosenblum and I. Eisenman, “Sea ice trends in climate models only accurate in runs with biased global warming,” *Journal of Climate*, vol. 30, no. 16, pp. 6265–6278, aug 2017.

A. G. Meehl, C. Covey, T. Delworth, M. Latif, B. Mcavaney, J. F. B. Mitchell *et al.*, “THE WCRP CMIP3 Multimodel Dataset: A New Era in Climate Change Research,” *American Meteorological Society*, no. September, 2007.

J. Stroeve, M. M. Holland, W. Meier, T. Scambos, and M. Serreze, “Arctic sea ice decline: Faster than forecast,” *Geophysical Research Letters*, vol. 34, no. 9, 2007.

M. G. A. Taylor Karl E., Stouffer Ronald J., “An Overview of CMIP5 and the Experiment Design,” *American Meteorological Society*, vol. 3, no. april, pp. 485–498, 2012.

V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer *et al.*, “Overview of the Coupled Model Intercomparison Project

Phase 6 (CMIP6) experimental design and organization," *Geoscientific Model Development*, vol. 9, no. 5, pp. 1937–1958, 2016. [Online]. Available: https://www.geosci-model-dev.net/9/1937/2016/

J. C. Stroeve, V. Kattsov, A. Barrett, M. Serreze, T. Pavlova, M. Holland *et al.*, "Trends in Arctic sea ice extent from CMIP5, CMIP3 and observations," *Geophysical Research Letters*, vol. 39, no. 16, pp. 1–7, 2012.

M. Ionita, K. Grosfeld, P. Scholz, R. Treffeisen, and G. Lohmann, "September Arctic Sea Ice minimum prediction - a new skillful statistical approach," *Earth System Dynamics Discussions*, pp. 1–23, sep 2018. [Online]. Available: https://www.earth-syst-dynam-discuss.net/esd-2018-61/

T. G. Reid and P. M. Tarantino, "Arctic sea ice extent forecasting using support vector regression," in *Proceedings - 2014 13th International Conference on Machine Learning and Applications, ICMLA 2014*. Institute of Electrical and Electronics Engineers Inc., feb 2014, pp. 1–6.

G. Peng, W. N. Meier, D. J. Scott, M. H. Savoie, and N. Snow, "A long-term and reproducible passive microwave sea ice concentration data record for climate studies and monitoring," *Earth System Science Data*, pp. 311–318, 2013.

A. Schweiger, R. Lindsay, J. Zhang, M. Steele, H. Stern, and R. Kwok, "Uncertainty in modeled Arctic sea ice volume," *Journal of Geophysical Research: Oceans*, vol. 116, no. 9, pp. 1–21, 2011.

NOAA, OAR, and ESRL-PSD, "Ncep-doe reanalysis 2," 2019, nCEP_Reanalysis 2 data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA. [Online]. Available: https://www.esrl.noaa.gov/psd/

——, "Noaa extended reconstructed sea surface temperature," 2019, NOAA_ERSST_V4 data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA. [Online]. Available: https://www.esrl.noaa.gov/psd/

J.-C. Golaz, P. M. Caldwell, L. P. Van Roekel, M. R. Petersen, Q. Tang, J. D. Wolfe, G. Abeshu, V. Anantharaj, X. S. Asay-Davis, D. C. Bader, S. A. Baldwin, G. Bisht, P. A. Bogenschutz, M. Branstetter, M. A. Brunke, S. R. Brus, S. M. Burrows, P. J. Cameron-Smith, A. S. Donahue, M. Deakin, R. C. Easter, K. J. Evans, Y. Feng, M. Flanner, J. G. Foucar, J. G. Fyke, B. M. Griffin, C. Hannay, B. E. Harrop, M. J. Hoffman, E. C. Hunke, R. L. Jacob, D. W. Jacobsen, N. Jeffery, P. W. Jones, N. D. Keen, S. A. Klein, V. E. Larson, L. R. Leung, H.-Y. Li, W. Lin, W. H. Lipscomb, P.-L. Ma, S. Mahajan, M. E. Maltrud, A. Mametjanov, J. L. McClean, R. B. McCoy, R. B. Neale, S. F. Price, Y. Qian, P. J. Rasch, J. E. J. Reeves Eyre, W. J. Riley, T. D. Ringler, A. F. Roberts, E. L. Roesler, A. G. Salinger, Z. Shaheen, X. Shi, B. Singh, J. Tang, M. A. Taylor, P. E. Thornton, A. K. Turner, M. Veneziani, H. Wan, H. Wang, S. Wang, D. N. Williams, P. J. Wolfram, P. H. Worley, S. Xie, Y. Yang, J.-H. Yoon, M. D. Zelinka, C. S. Zender, X. Zeng, C. Zhang, K. Zhang, Y. Zhang, X. Zheng, T. Zhou, and Q. Zhu, "The DOE E3SM Coupled Model Version 1 : Overview and Evaluation at Standard

Resolution," *Journal of Advances in Modeling Earth Systems*, vol. 11, no. 7, pp. 2089–2129, 2019.

J. E. Kay, C. Deser, A. Phillips, A. Mai, C. Hannay, G. Strand *et al.*, "The community earth system model (cesm) large ensemble project: A community resource for studying climate change in the presence of internal climate variability," *Bulletin of the American Meteorological Society*, vol. 96, no. 8, pp. 1333–1349, 2015. [Online]. Available: https://doi.org/10.1175/BAMS-D-13-00255.1

B. J. Zib, X. Dong, B. Xi, and A. Kennedy, "Evaluation and intercomparison of cloud fraction and radiative fluxes in recent reanalyses over the arctic using BSRN surface observations," *Journal of Climate*, vol. 25, no. 7, pp. 2291–2305, 2012.

L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

R. E. Banfield, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer, "A Comparison of Decision Tree Ensemble Creation Techniques," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 173–180, Jan. 2007.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

L. Breiman, “Arcing classifiers,” *Ann. Statist.*, vol. 26, no. 3, pp. 801–849, 06 1998. [Online]. Available: https://doi.org/10.1214/aos/1024691079

S. Van Der Walt, S. C. Colbert, and G. Varoquaux, “The numpy array: a structure for efficient numerical computation,” *Computing in Science & Engineering*, vol. 13, no. 2, p. 22, 2011.

M. Waskom and the seaborn development team, “mwaskom/seaborn,” Sep. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.592845

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Wadsworth International Group, Belmont, California, 1984.

T. Hengl, M. Nussbaum, M. N. Wright, G. Heuvelink, and B. Gräler, “Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables,” *PeerJ*, vol. 6, 2018. [Online]. Available: https://doi.org/10.7717/peerj.5518

L. Breiman, “Rejoinder: Arcing classifiers,” *The Annals of Statistics*, vol. 26, no. 3, pp. 841–849, 1998. [Online]. Available: http://www.jstor.org/stable/120059

R. E. Banfield, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer, “A new ensemble diversity measure applied to thinning ensembles,” in *Multiple Classifier Systems*, T. Windeatt and F. Roli, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 306–316.

U. S. Bhatt, P. Bieniek, C. Bitz, E. Blanchard-Wrigglesworth, H. Eicken,

H. Goessling *et al.*, “2019 sea ice outlook full post-season report,” February 2020, editors: Turner-Bogren, B. and H. V. Wiggins. [Online]. Available: https://www.arcus.org/sipn/sea-ice-outlook/2019/post-season

A. C. Ordonez, C. M. Bitz, and E. Blanchard-Wrigglesworth, “Processes controlling Arctic and Antarctic sea ice predictability in the Community Earth System Model,” *Journal of Climate*, vol. 31, pp. 9771–9786, 2018.

E. Blanchard-Wrigglesworth, K. C. Armour, C. M. Bitz, and E. Deweaver, “Persistence and inherent predictability of arctic sea ice in a GCM ensemble and observations,” *Journal of Climate*, vol. 24, no. 1, pp. 231–250, jan 2011.

J. J. Nichol, M. Peterson, G. M. Fricke, and K. Peterson, “Learning Why: Data-Driven Causal Evaluations of Climate Models.” *ICML 2021 Workshop Tackling Climate Change with Machine Learning*, 2021.

J. J. Nichol, M. Peterson, and K. Peterson, “Causal Evaluations for Identifying Differences between Observations and Earth System Models,” 2021. [Online]. Available: https://www.osti.gov/biblio/1820528

J. J. Nichol, M. Weylandt, M. Smith, and L. Swiler, “Benchmarking the PCMCI Causal Discovery Algorithm for Spatiotemporal Systems,” Sandia National Laboratories, Tech. Rep., 2023. [Online]. Available: https://www.osti.gov/biblio/1991387

J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, J. Munoz-Mari, E. H. v. Nes, J. Peters, R. Quax, M. Reichstein, M. Scheffer, B. Scholkopf, P. Spirtes, G. Sugihara, J. Sun, K. Zhang, and J. Zscheischler, "Inferring causation from time series in Earth system sciences," *Nature Communications*, vol. 10, no. 1, 2019.

J. Runge, V. Petoukhov, J. F. Donges, J. Hlinka, N. Jajcay, M. Vejmelka, D. Hartman, N. Marwan, M. Paluš, and J. Kurths, "Identifying causal gateways and mediators in complex spatio-temporal systems," *Nature Communications*, vol. 6, no. 1, p. 8502, 2015.

J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, "Detecting and quantifying causal associations in large nonlinear time series datasets," *Science Advances*, vol. 5, no. 11, p. eaau4996, Nov 2019. [Online]. Available: https://www.science.org/doi/10.1126/sciadv.aau4996

M. H. Hitchman, M. McKay, and C. R. Trepte, "A climatology of stratospheric aerosol," *Journal of Geophysical Research: Atmospheres*, vol. 99, no. D10, pp. 20 689–20 700, 1994. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/94JD01525

D. B. A. Jones, H. R. Schneider, and M. B. McElroy, "Effects of the quasi-biennial oscillation on the zonally averaged transport of tracers," *Journal of Geophysical Research: Atmospheres*, vol. 103, no. D10, pp. 11 235–11 249, 1998.

V. Aquila, C. I. Garfinkel, P. Newman, L. Oman, and D. Waugh, "Modifications of the quasi-biennial oscillation by a geoengineering perturbation of the stratospheric aerosol layer," *Geophysical Research Letters*, vol. 41, no. 5, pp. 1738–1744, 2014.

L. J. Gray, J. A. Anstey, Y. Kawatani, H. Lu, S. Osprey, and V. Schenzinger, "Surface impacts of the Quasi Biennial Oscillation," *Atmospheric Chemistry and Physics*, vol. 18, no. 11, pp. 8227–8247, 2018.

E. C. Neto, M. P. Keller, A. D. Attie, and B. S. Yandell, "Causal graphical models in systems genetics: A unified framework for joint inference of causal network and genetic architecture for correlated phenotypes," *The Annals of Applied Statistics*, vol. 4, no. 1, pp. 320–339, 2010.

X. Zhang, X.-M. Zhao, K. He, L. Lu, Y. Cao, J. Liu, J.-K. Hao, Z.-P. Liu, and L. Chen, "Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information," *Bioinformatics*, vol. 28, no. 1, pp. 98–104, 2011.

M. Kamiński, M. Ding, W. A. Truccolo, and S. L. Bressler, "Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance," *Biological Cybernetics*, vol. 85, no. 2, pp. 145–157, 2001.

A. A. Tsonis, E. R. Deyle, H. Ye, and G. Sugihara, "Convergent Cross Mapping: Theory and an Example," *Advances in Nonlinear Geosciences*, pp. 587–600, 2017.

Y. Deng and I. Ebert-Uphoff, "Weakening of atmospheric information flow in a warming climate in the Community Climate System Model," *Geophysical Research Letters*, vol. 41, no. 1, pp. 193–200, 2014.

G. D. Capua, M. Kretschmer, R. V. Donner, B. v. d. Hurk, R. Vellore, R. Krishnan, and D. Coumou, "Tropical and mid-latitude teleconnections interacting with the Indian summer monsoon rainfall: a theory-guided causal effect network approach," *Earth System Dynamics*, vol. 11, no. 1, pp. 17–34, 2019.

G. D. Capua, J. Runge, R. V. Donner, B. v. d. Hurk, A. G. Turner, R. Vellore, R. Krishnan, and D. Coumou, "Dominant patterns of interaction between the tropics and mid-latitudes in boreal summer: causal relationships and the role of timescales," *Weather and Climate Dynamics*, vol. 1, no. 2, pp. 519–539, 2020.

C. Krich, J. Runge, D. G. Miralles, M. Migliavacca, O. Perez-Priego, T. El-Madany, A. Carrara, and M. D. Mahecha, "Estimating causal networks in biosphere–atmosphere interaction with the PCMCI approach," *Biogeosciences*, vol. 17, no. 4, pp. 1033–1061, 2020.

E. Galytska, K. Weigel, D. Handorf, R. Jaiser, R. H. Köhler, J. Runge, and V. Eyring, "Causal model evaluation of Arctic-midlatitude teleconnections in CMIP6," *Journal of Geophysical Research: Atmospheres*, vol. 128, no. 17, 2022.

T. J. O'Kane, D. Harries, and M. A. Collier, "Bayesian Structure Learning for Climate Model Evaluation," *Journal of Advances in Modeling Earth Systems*, vol. 16, no. 5, 2024.

H. Zhao, V. Kitsios, T. J. O'Kane, and E. V. Bonilla, "Bayesian Factorised Granger-Causal Graphs For Multivariate Time-series Data," *arXiv*, 2024.

J. Runge, A. Gerhardus, G. Varando, V. Eyring, and G. Camps-Valls, "Causal inference for time series," *Nature Reviews Earth & Environment*, vol. 4, no. 7, pp. 487–505, 2023.

J. Pearl, M. Glymour, and N. Jewell, *Causal Inference in Statistics: A Primer*. Wiley, 2016. [Online]. Available: https://books.google.com/books?id=L3G-CgAAQBAJ

C. Glymour and R. Scheines, "Causal modeling with the TETRAD program," *Synthese*, vol. 68, no. 1, pp. 37–63, 1986.

J. Pearl and T. S. Verma, "A statistical semantics for causation," *Statistics and Computing*, vol. 2, no. 2, pp. 91–95, 1992.

X. Zheng, B. Aragam, P. Ravikumar, and E. P. Xing, "DAGs with NO TEARS: Continuous Optimization for Structure Learning," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 9492–9503.

Y. Liu, A. Niculescu-Mizil, A. Lozano, and Y. Lu, "Learning Temporal Causal Graphs for Relational Time-Series Analysis," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ser. ICML'10. Madison, WI, USA: Omnipress, 2010, p. 687–694.

R. Pamfil, N. Sriwattanaworachai, S. Desai, P. Pilgerstorfer, K. Georgatzis, P. Beaumont, and B. Aragam, "DYNOTEARS: Structure Learning from Time-Series Data," *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, vol. 108, pp. 1595–1605, 2020. [Online]. Available: https://proceedings.mlr.press/v108/pamfil20a.html

S. Ali, U. Hasan, X. Li, O. Faruque, A. Sampath, Y. Huang, M. O. Gani, and J. Wang, "Causality for Earth Science – A Review on Time-series and Spatiotemporal Causality Methods," *arXiv*, 2024.

D. Colombo and M. H. Maathuis, "Order-independent constraint-based causal structure learning," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3741–3782, 2014.

S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, no. 15, pp. 3932–3937, 2016.

Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar, "Fourier Neural Operator for Parametric Partial Differential Equations," *arXiv*, 2020.

J. Pathak, S. Subramanian, P. Harrington, S. Raja, A. Chattopadhyay, M. Mardani, T. Kurth, D. Hall, Z. Li, K. Azizzadenesheli, P. Hassanzadeh, K. Kashinath, and

A. Anandkumar, “FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators,” *arXiv*, 2022.

J. Hart, M. Gulian, I. Manickam, and L. P. Swiler, “Solving High-Dimensional Inverse Problems with Auxiliary Uncertainty via Operator Learning with Limited Data,” *Journal of Machine Learning for Modeling and Computing*, vol. 4, no. 2, pp. 105–133, 2023.

K. Bhattacharjee, N. Naskar, S. Roy, and S. Das, “A survey of cellular automata: types, dynamics, non-uniformity and applications,” *Natural Computing*, vol. 19, no. 2, pp. 433–461, 2020.

H. J. Miller, “Tobler’s First Law and Spatial Analysis,” *Annals of the Association of American Geographers*, vol. 94, no. 2, pp. 284–289, 2004.

R. T. Walker, “GEOGRAPHY, VON THÜNEN, AND TOBLER’S FIRST LAW: TRACING THE EVOLUTION OF A CONCEPT,” *Geographical Review*, vol. 112, no. 4, pp. 591–607, 2022.

V. K. Raghu, J. D. Ramsey, A. Morris, D. V. Manatakis, P. Sprites, P. K. Chrysanthis, C. Glymour, and P. V. Benos, “Comparison of strategies for scalable causal discovery of latent variable models from mixed data,” *International Journal of Data Science and Analytics*, vol. 6, no. 1, pp. 33–45, 2018.

S. Guo, W. I. Rose, G. J. S. Bluth, and I. M. Watson, “Particles in the great

Pinatubo volcanic cloud of June 1991: The role of ice," *Geochemistry, Geophysics, Geosystems*, vol. 5, no. 5, 2004.

S. Kremser, L. W. Thomason, M. v. Hobe, M. Hermann, T. Deshler, C. Timmreck, M. Toohey, A. Stenke, J. P. Schwarz, R. Weigel, S. Fueglistaler, F. J. Prata, J. Vernier, H. Schlager, J. E. Barnes, J. Antuña-Marrero, D. Fairlie, M. Palm, E. Mahieu, J. Notholt, M. Rex, C. Bingen, F. Vanhellemont, A. Bourassa, J. M. C. Plane, D. Klocke, S. A. Carn, L. Clarisse, T. Trickl, R. Neely, A. D. James, L. Rieger, J. C. Wilson, and B. Meland, "Stratospheric aerosol—Observations, processes, and impact on climate," *Reviews of Geophysics*, vol. 54, no. 2, pp. 278–335, 2016.

E. G. Dutton and J. R. Christy, "Solar radiative forcing at selected locations and evidence for global lower tropospheric cooling following the eruptions of El Chichón and Pinatubo," *Geophysical Research Letters*, vol. 19, no. 23, pp. 2313–2316, 1992.

K. Labitzke and M. P. McCormick, "Stratospheric temperature increases due to Pinatubo aerosols," *Geophysical Research Letters*, vol. 19, no. 2, pp. 207–210, 1992.

D. E. Parker, H. Wilson, P. D. Jones, J. R. Christy, and C. K. FOLLAND, "The impact of Mount Pinatubo on world-wide temperatures," *International Journal of Climatology*, vol. 16, no. 5, pp. 487–497, 1996.

B. J. Soden, R. T. Wetherald, G. L. Stenchikov, and A. Robock, "Global Cooling

After the Eruption of Mount Pinatubo: A Test of Climate Feedback by Water Vapor," *Science*, vol. 296, no. 5568, pp. 727–730, 2002.

K. E. Trenberth and A. Dai, "Effects of Mount Pinatubo volcanic eruption on the hydrological cycle as an analog of geoengineering," *Geophysical Research Letters*, vol. 34, no. 15, 2007.

M. Weylandt and L. P. Swiler, "Beyond PCA: Additional Dimension Reduction Techniques to Consider in the Development of Climate Fingerprints," *Journal of Climate*, vol. 37, no. 5, pp. 1723–1735, 2024.

D. E. Parker, H. Wilson, P. D. Jones, J. R. Christy, and C. K. Folland, "The impact of mount pinatubo on world-wide temperatures," *International Journal of Climatology*, vol. 16, no. 5, pp. 487–497, 1996. [Online]. Available: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0088%28199605%2916%3A5%3C487%3A%3AAID-JOC39%3E3.0.CO%3B2-J

A. Robock, "Volcanic eruptions and climate," *Reviews of Geophysics*, vol. 38, no. 2, pp. 191–219, 2000. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/1998RG000054

C. Timmreck, "Modeling the climatic effects of large explosive volcanic eruptions," *Wiley Interdisciplinary Reviews: Climate Change*, vol. 3, no. 6, pp. 545–564, 2012.

L. R. Marshall, E. C. Maters, A. Schmidt, C. Timmreck, A. Robock, and

M. Toohey, "Volcanic effects on climate: recent advances and future avenues," *Bulletin of Volcanology*, vol. 84, no. 5, p. 54, 2022.

J. P. Hollowed, C. Jablonowski, H. Y. Brown, B. R. Hillman, D. L. Bull, and J. L. Hart, "Localized injections of interactive volcanic aerosols and their climate impacts in a simple general circulation model," *EGUsphere*, vol. 2024, pp. 1–38, 2024.

J. Golaz, L. P. V. Roekel, X. Zheng, A. F. Roberts, J. D. Wolfe, W. Lin, A. M. Bradley, Q. Tang, M. E. Maltrud, R. M. Forsyth, C. Zhang, T. Zhou, K. Zhang, C. S. Zender, M. Wu, H. Wang, A. K. Turner, B. Singh, J. H. Richter, Y. Qin, M. R. Petersen, A. Mametjanov, P. Ma, V. E. Larson, J. Krishna, N. D. Keen, N. Jeffery, E. C. Hunke, W. M. Hannah, O. Guba, B. M. Griffin, Y. Feng, D. Engwirda, A. V. D. Vittorio, C. Dang, L. M. Conlon, C. Chen, M. A. Brunke, G. Bisht, J. J. Benedict, X. S. Asay-Davis, Y. Zhang, M. Zhang, X. Zeng, S. Xie, P. J. Wolfram, T. Vo, M. Veneziani, T. K. Tesfa, S. Sreepathi, A. G. Salinger, J. E. J. R. Eyre, M. J. Prather, S. Mahajan, Q. Li, P. W. Jones, R. L. Jacob, G. W. Huebler, X. Huang, B. R. Hillman, B. E. Harrop, J. G. Foucar, Y. Fang, D. S. Comeau, P. M. Caldwell, T. Bartoletti, K. Balaguru, M. A. Taylor, R. B. McCoy, L. R. Leung, and D. C. Bader, "The DOE E3SM Model Version 2: Overview of the Physical Model and Initial Model Evaluation," *Journal of Advances in Modeling Earth Systems*, vol. 14, no. 12, 2022.

H. Y. Brown, B. Wagman, D. Bull, K. Peterson, B. Hillman, X. Liu, Z. Ke, and

L. Lin, "Validating a microphysical prognostic stratospheric aerosol implementation in E3SMv2 using observations after the Mount Pinatubo eruption," *Geoscientific Model Development*, vol. 17, no. 13, pp. 5087–5121, 2024.

M. Kalisch and P. Bühlmann, "Estimating high-dimensional directed acyclic graphs with the PC-algorithm," *Journal of Machine Learning Research*, vol. 8, pp. 613–636, 2007. [Online]. Available: https://www.jmlr.org/papers/v8/kalisch07a.html

G. I. Allen, L. Gan, and L. Zheng, "Interpretable Machine Learning for Discovery: Statistical Challenges and Opportunities," *Annual Review of Statistics and Its Application*, vol. 11, no. 1, pp. 97–121, 2023.

I. Fountalis, C. Dovrolis, A. Bracco, B. Dilkina, and S. Keilholz, "δ-MAPS: from spatio-temporal data to a weighted and lagged network between functional domains," *Applied Network Science*, vol. 3, no. 1, p. 21, 2018.

N. Nukavarapu, J.-A. Yang, and M. M. Jankowska, "Unsupervised Deep Learning Approach to Analyze Spatio-Temporal Change in Satellite Imagery," *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, vol. 00, pp. 2496–2499, 2023.

W. L. Davis, M. L. Carlson, I. K. Tezaur, D. L. Bull, K. J. Peterson, and L. P. Swiler, "Spatio-temporal multivariate cluster evolution analysis for detecting and tracking climate impacts," *Journal of Computational and Applied Mathematics*, vol. 465, p. 116583, 2025.

M. A. Thomas, M. A. Giorgetta, C. Timmreck, H.-F. Graf, and G. Stenchikov, "Simulation of the climate impact of Mt. Pinatubo eruption using ECHAM5 – Part 2: Sensitivity to the phase of the QBO and ENSO," *Atmospheric Chemistry and Physics*, vol. 9, no. 9, pp. 3001–3009, 2009.

G. Strang, *Introduction to Linear Algebra*, 5th ed. Wellesley, MA: Wellesley-Cambridge Press, 2016.

B. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta (BBA) - Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0005279575901099

N. S. Diffenbaugh, J. S. Pal, R. J. Trapp, and F. Giorgi, "Fine-scale processes regulate the response of extreme events to global climate change," *Proceedings of the National Academy of Sciences*, vol. 102, no. 44, pp. 15 774–15 778, 2005.

M. Palu, "Coupling in complex systems as information transfer across time scales," *Philosophical Transactions of the Royal Society A*, vol. 377, no. 2160, p. 20190094, 2019.

A. Agarwal, L. Caesar, N. Marwan, R. Maheswaran, B. Merz, and J. Kurths, "Network-based identification and characterization of teleconnections on different scales," *Scientific Reports*, vol. 9, no. 1, p. 8808, 2019.

Z. Zhang, G. Li, Y. Cai, X. Cheng, Y. Sun, J. Zhao, P. Shu, L. Ma, and Z. An, "Millennial-Scale Monsoon Variability Modulated by Low-Latitude Insolation During the Last Glaciation," *Geophysical Research Letters*, vol. 49, no. 1, 2022.

J. Nichol, "CaStLe Data Release for JGR MLC 2024," Jul. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.12701546

J. J. Nichol, M. Weylandt, G. M. Fricke, M. E. Moses, D. Bull, and L. P. Swiler, "Space-Time Causal Discovery in Earth System Science: A Local Stencil Learning Approach," *Journal of Geophysical Research: Machine Learning and Computation*, vol. 2, no. 3, 2025.

D. B. Rubin, "Essential concepts of causal inference: a remarkable history and an intriguing future," *Biostatistics & Epidemiology*, vol. 3, no. 1, pp. 140–155, 2019.

I. Ebert-Uphoff and Y. Deng, "Causal Discovery for Climate Research Using Graphical Models," *Journal of Climate*, vol. 25, no. 17, pp. 5648–5665, 2012.

G. F. Cooper, I. Bahar, M. J. Becich, P. V. Benos, J. Berg, J. U. Espino, C. Glymour, R. C. Jacobson, M. Kienholz, A. V. Lee, X. Lu, R. Scheines, and team, and the Center for Causal Discovery, "The center for causal discovery of biomedical knowledge from big data," *Journal of the American Medical Informatics Association*, vol. 22, no. 6, pp. 1132–1136, 2015.

A. Feder, K. A. Keith, E. Manzoor, R. Pryzant, D. Sridhar, Z. Wood-Doughty, J. Eisenstein, J. Grimmer, R. Reichart, M. E. Roberts, B. M. Stewart, V. Veitch, and D. Yang, “Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond,” *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 1138–1158, 2022.

A. Zanga, E. Ozkirimli, and F. Stella, “A Survey on Causal Discovery: Theory and Practice,” *International Journal of Approximate Reasoning*, vol. 151, pp. 101–129, 2022.

A. Sadeghi, A. Gopal, and M. Fesanghary, “Causal Discovery in Financial Markets: A Framework for Nonstationary Time-Series Data,” *arXiv*, 2023.

M. Greenacre, P. J. F. Groenen, T. Hastie, A. I. D’Enza, A. Markos, and E. Tuzhilina, “Principal component analysis,” *Nature Reviews Methods Primers*, vol. 2, no. 1, p. 100, 2022.

H. Sun, H. Liu, Y. Ma, and Q. Xia, “Optical Remote Sensing Indexes of Soil Moisture: Evaluation and Improvement Based on Aircraft Experiment Observations,” *Remote Sensing*, vol. 13, no. 22, p. 4638, 2021.

S. Ganesan and D. Subramani, “Spatio-temporal predictive modeling framework for infectious disease spread,” *Scientific Reports*, vol. 11, no. 1, p. 6741, 2021.

S. K. Paul, S. Jana, and P. Bhaumik, “A Multivariate Spatiotemporal Model of COVID-19 Epidemic Using Ensemble of ConvLSTM Networks,” *Journal of The Institution of Engineers (India): Series B*, vol. 102, no. 6, pp. 1137–1142, 2021.

A. M. MacEachren, M. Wachowicz, R. Edsall, D. Haug, and R. Masters, “Constructing knowledge from multivariate spatiotemporal data: integrating geographical visualization with knowledge discovery in database methods," *International Journal of Geographical Information Science*, vol. 13, no. 4, pp. 311–334, 1999.

T. C. Haas, "New systems for modeling, estimating, and predicting a multivariate spatio–temporal process," *Environmetrics*, vol. 13, no. 4, pp. 311–332, 2002.

N. T. Wimer, L. Esclapez, N. Brunhart-Lupo, M. H. d. Frahan, M. Rahimi, M. Hassanaly, J. Rood, S. Yellapantula, H. Sitaraman, B. Perry, M. Martin, O. Doronina, S. N. Appukuttan, M. Rieth, and M. Day, "Visualizations of a methane/diesel RCCI engine using PeleC and PeleLMeX," *Physical Review Fluids*, vol. 8, no. 11, p. 110511, 2023.

M. Guarin, R. Faelens, A. Giusti, N. D. Croze, M. Léonard, D. Cabooter, P. Annaert, P. d. Witte, and A. Ny, "Spatiotemporal imaging and pharmacokinetics of fluorescent compounds in zebrafish eleuthero-embryos after different routes of administration," *Scientific Reports*, vol. 11, no. 1, p. 12229, 2021.

F. Klingelhuber, S. Frendo-Cumbo, M. Omar-Hmeadi, L. Massier, P. Kakimoto, A. J. Taylor, M. Couchet, S. Ribicic, M. Wabitsch, A. C. Messias, A. Iuso, T. D. Müller, M. Rydén, N. Mejhert, and N. Krahmer, "A spatiotemporal proteomic map of human adipogenesis," *Nature Metabolism*, vol. 6, no. 5, pp. 861–879, 2024.

D. J. Higham, “Modeling and Simulating Chemical Reactions,” *SIAM Review*, vol. 50, no. 2, pp. 347–368, 2008.

L. D. Owen, W. Ge, M. Rieth, M. Arienti, L. Esclapez, B. S. Soriano, M. E. Mueller, M. Day, R. Sankaran, and J. H. Chen, “PeleMP: The Multiphysics Solver for the Combustion Pele Adaptive Mesh Refinement Code Suite,” *Journal of Fluids Engineering*, vol. 146, no. 4, 2024.

P. K. Rubenstein, S. Bongers, B. Schölkopf, and J. M. Mooij, “From Deterministic ODEs to Dynamic Structural Causal Models,” in *UAI’18: Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, ser. UAI’18. AUAI Press, 2018. [Online]. Available: http://auai.org/uai2018/proceedings/papers/43.pdf

Y. Benjamini and Y. Hochberg, “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

D. L. Bull, K. J. Peterson, L. Shand, L. P. Swiler, I. K. Tezaur, B. K. Cook, A. G. Salinger, C. M. Amann, B. M. Watts, R. W. Leland *et al.*, “Climate impact: Determining etiology through pathways (cldera),” Sandia National Lab. (SNL-NM), Albuquerque, NM (United States), Tech. Rep., 09 2024. [Online]. Available: https://www.osti.gov/biblio/2480139

A. Hu, X. Liu, Z. Ke, B. Wagman, H. Brown, Z. Lu, D. Bull, and K. Peterson, “Size-resolved process understanding of stratospheric sulfate aerosol following the pinatubo eruption," *EGUsphere*, vol. 2024, pp. 1–28, 2024. [Online]. Available: https://egusphere.copernicus.org/preprints/2024/egusphere-2024-2227/

M. J. Mills, J. H. Richter, S. Tilmes, B. Kravitz, D. G. MacMartin, A. A. Glanville, J. J. Tribbia, J. Lamarque, F. Vitt, A. Schmidt, A. Gettelman, C. Hannay, J. T. Bacmeister, and D. E. Kinnison, "Radiative and Chemical Response to Interactive Stratospheric Sulfate Aerosols in Fully Coupled CESM1(WACCM)," *Journal of Geophysical Research: Atmospheres*, vol. 122, no. 23, pp. 13,061–13,078, 2017.

K. Zhang, W. Zhang, H. Wan, P. J. Rasch, S. J. Ghan, R. C. Easter, X. Shi, Y. Wang, H. Wang, P.-L. Ma, S. Zhang, J. Sun, S. M. Burrows, M. Shrivastava, B. Singh, Y. Qian, X. Liu, J.-C. Golaz, Q. Tang, X. Zheng, S. Xie, W. Lin, Y. Feng, M. Wang, J.-H. Yoon, and L. R. Leung, "Effective radiative forcing of anthropogenic aerosols in E3SM version 1: historical changes, causality, decomposition, and parameterization sensitivities," *Atmospheric Chemistry and Physics*, vol. 22, no. 13, pp. 9129–9160, 2022.

J. Runge, "Tigramite," https://github.com/jakobrunge/tigramite, 2024.

C. Tosh, P. Greengard, B. Goodrich, A. Gelman, A. Vehtari, and D. Hsu, "The Piranha Problem: Large Effects Swimming in a Small Pond," *Notices of the American Mathematical Society*, vol. 72, no. 01, p. 1, 2025.

H. Mandler and B. Weigand, "A review and benchmark of feature importance methods for neural networks," *ACM Computing Surveys*, vol. 56, no. 12, pp. 1–30, 2024.

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

A. Selvitella, "The ubiquity of the Simpson's Paradox," *Journal of Statistical Distributions and Applications*, vol. 4, no. 1, p. 2, 2017.

S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking Clever Hans predictors and assessing what machines really learn," *Nature Communications*, vol. 10, no. 1, p. 1096, 2019.

H. Lee and S. Chen, "Systematic Bias of Machine Learning Regression Models and Correction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–11, 2025.

C. Cammarota and A. Pinto, "Variable selection and importance in presence of high collinearity: an application to the prediction of lean body mass from multi-frequency bioelectrical impedance," *Journal of Applied Statistics*, vol. 48, no. 9, pp. 1644–1658, 2021.

T. Parr and J. D. Wilson, "Partial dependence through stratification," *Machine Learning with Applications*, vol. 6, p. 100146, 2021.

T. Parr, J. Hamrick, and J. D. Wilson, “Nonparametric feature impact and importance,” *Information Sciences*, vol. 653, p. 119563, 2024.

M. G. Brown, M. G. Peterson, I. K. Tezaur, K. J. .Peterson, and D. L. Bull, “Random forest regression feature importance for climate impact pathway detection,” *Journal of Computational and Applied Mathematics*, vol. 464, p. 116479, 2025.

Z. M. Labe and E. A. Barnes, “Comparison of Climate Model Large Ensembles With Observations in the Arctic Using Simple Neural Networks,” *Earth and Space Science*, vol. 9, no. 7, 2022.

A. Konya and P. Nematzadeh, “Recent applications of AI to environmental disciplines: A review,” *Science of The Total Environment*, vol. 906, p. 167705, 2024.

J. Lao, X. Wang, B. Shi, B. Wang, and Z. Jiao, “Advances in the application of artificial intelligence in environmental science,” *Chinese Journal of Nature*, vol. 46, no. 4, pp. 271–280, 2024.

J. Sjolte, F. Adolphi, H. Guðlaugsdòttir, and R. Muscheler, “Major Differences in Regional Climate Impact Between High- and Low-Latitude Volcanic Eruptions,” *Geophysical Research Letters*, vol. 48, no. 8, 2021.

K. Baranowski, C. Faust, P. Eby, and N. Bharti, “Quantifying the impacts of Australian bushfires on native forests and gray-headed flying foxes,” *Global Ecology and Conservation*, vol. 27, p. e01566, 2021.

A. E. Payne, M.-E. Demory, L. R. Leung, A. M. Ramos, C. A. Shields, J. J. Rutz,

N. Siler, G. Villarini, A. Hall, and F. M. Ralph, “Responses and impacts of atmospheric rivers to climate change,” *Nature Reviews Earth & Environment*, vol. 1, no. 3, pp. 143–157, 2020.

J. Baño-Medina, A. Sengupta, J. D. Doyle, C. A. Reynolds, D. Watson-Parris, and L. D. Monache, “Are AI weather models learning atmospheric physics? A sensitivity analysis of cyclone Xynthia,” *npj Climate and Atmospheric Science*, vol. 8, no. 1, p. 92, 2025.

T. B. Higgins, A. C. Subramanian, P. A. G. Watson, and S. Sparrow, “Changes to Atmospheric River Related Extremes Over the United States West Coast Under Anthropogenic Warming,” *Geophysical Research Letters*, vol. 52, no. 5, 2025.

D. Keellings and H. Moradkhani, “Spatiotemporal Evolution of Heat Wave Severity and Coverage Across the United States,” *Geophysical Research Letters*, vol. 47, no. 9, 2020.

D. A. Driscoll, K. J. Macdonald, R. K. Gibson, T. S. Doherty, D. G. Nimmo, R. H. Nolan, E. G. Ritchie, G. J. Williamson, G. W. Heard, E. M. Tasker, R. Bilney, N. Porch, R. A. Collett, R. A. Crates, A. C. Hewitt, E. Pendall, M. M. Boer, J. Gates, R. L. Boulton, C. M. Mclean, H. Groffen, A. C. Maisey, C. T. Beranek, S. A. Ryan, A. Callen, A. J. Hamer, A. Stauber, G. J. Daly, J. Gould, K. L. Klop-Toker, M. J. Mahony, O. W. Kelly, S. L. Wallace, S. E. Stock, C. J. Weston, L. Volkova, D. Black, H. Gibb, J. J. Grubb, M. A. McGeoch, N. P. Murphy, J. S. Lee, C. R. Dickman, V. J. Neldner, M. R. Ngugi, V. Miritis, F. Köhler, M. Perri,

A. J. Denham, B. D. E. Mackenzie, C. A. M. Reid, J. T. Rayment, A. Arriaga-Jiménez, M. W. Hewins, A. Hicks, B. A. Melbourne, K. F. Davies, M. E. Bitters, G. D. Linley, A. C. Greenville, J. K. Webb, B. Roberts, M. Letnic, O. F. Price, Z. C. Walker, B. R. Murray, E. M. Verhoeven, A. M. Thomsen, D. Keith, J. S. Lemmon, M. K. J. Ooi, V. L. Allen, O. T. Decker, P. T. Green, A. Moussalli, J. K. Foon, D. B. Bryant, K. L. Walker, M. J. Bruce, G. Madani, J. L. Tscharke, B. Wagner, C. R. Nitschke, C. R. Gosper, C. J. Yates, R. Dillon, S. Barrett, E. E. Spencer, G. M. Wardle, T. M. Newsome, S. A. Pulsford, A. Singh, A. Roff, K. J. Marsh, K. Mcdonald, L. G. Howell, M. R. Lane, R. H. Cristescu, R. R. Witt, E. J. Cook, F. Grant, B. S. Law, J. Seddon, K. K. Berris, R. M. Shofner, M. Barth, T. Welz, A. Foster, D. Hancock, M. Beitzel, L. X. L. Tan, N. A. Waddell, P. M. Fallow, L. Schweickle, T. D. L. Breton, C. Dunne, M. Green, A.-M. Gilpin, J. M. Cook, S. A. Power, K. Hogendoorn, R. Brawata, C. J. Jolly, M. Tozer, N. Reiter, and R. D. Phillips, "Biodiversity impacts of the 2019–2020 Australian megafires," *Nature*, vol. 635, no. 8040, pp. 898–905, 2024.

G. Goerg and C. Shalizi, "Mixed licors: A nonparametric algorithm for predictive state reconstruction," in *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, C. M. Carvalho and P. Ravikumar, Eds., vol. 31. Scottsdale, Arizona, USA: PMLR, 29 Apr–01 May 2013, pp. 289–297. [Online]. Available: https://proceedings.mlr.press/v31/goerg13a.html

J. Burgers, “A Mathematical Model Illustrating the Theory of Turbulence,” *Advances in Applied Mechanics*, vol. 1, pp. 171–199, 1948.

M. P. Bonkile, A. Awasthi, C. Lakshmi, V. Mukundan, and V. S. Aswin, “A systematic literature review of Burgers’ equation with recent advances,” *Pramana*, vol. 90, no. 6, p. 69, 2018.