THE UNIVERSITY OF CHICAGO

INTERPRETABLE DATA SCIENCE: APPLICATIONS AND THEORY IN CLIMATE SCIENCE AND PUBLIC HEALTH

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
ABBY STEVENS

CHICAGO, ILLINOIS
JUNE 2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

My advisor, Becca Willett, is truly a marvel. Thank you for putting so much effort into finding projects that kept me excited and motivated, for your endless energy and attention to detail, and for your genuine dedication to your craft and collaborations. You have helped me become such a clearer writer and presenter, have helped me navigate many tricky situations (both research-based and not), and I have never doubted that you truly want the best for your research team. Meeting CK and the rest of the team at Argonne was such an unlikely blessing. In addition to introducing me to an entirely new set of research questions and tools, the research atmosphere at Argonne is so collaborative, so engaging, and so motivating. I am endlessly grateful for the hours CK has spent supporting me and my pursuits; his patience, humor, and research perspective always kept me motivated. And to Dan Nicolae - as your eight-time TA, student representative, and very outspoken member of the department, I am so grateful for your leadership and support throughout the years. You've held this department together through some challenging times, and have always been willing to support my various initiatives (and pub nights). I had fun helping build up the data science initiative at UChicago and am grateful for your support of my less traditional research routes.

I owe a huge amount of gratitude to many other faculty, staff and peers at UChicago. Thank you to Mei Wang for always being willing to listen and offer words of encouragement. I am so grateful for the administrative staff in the department, especially Jonathan Rodriguez for handling everything that came his way and Keisha Prowoznik for never losing patience with my constant administrative procrastination. I'm very grateful to Julia Lane and Katie Rosengarten for their support of WiDS and data science at UChicago, and many others for making my time at UChicago so memorable. Many thanks to my cohort Keshav Vemuri, Andrew Goldstein, Jasha Sommer-Simpson, Yi Liu, Lizhen Nie, Jiacheng Wang, Qing Yan, Zhan Lin, and Pinhan Chen for the camaraderie and late nights in Jones. Special shout-out

to Dan Xiang for encouraging me to persevere when things felt futile, Nathan Gill for helping me keep our first-day-of-school promise to finish this thing. Outside my cohort, thanks to Micol Tresoldi, Veena Patel, Melissa Adrian, Irina Cristali, Omar Ghattas, Justin Finkel, Solomon Quinn, Kim Liu, Phillip Lo and others for the friendship and support. I'm so grateful to have met Rebecca Kotsonis and Elysa Strunin through this program and formed an enduring friendship that was solidified at a conference in Paris.

There are of course many to thank outside UChicago as well. First off, thanks to all who played an essential role in me being here in the first place: Iain Charmichael for your relentless encouragement to leave the workforce and give grad school a shot; my wonderful professors at Grinnell, especially Jen Paulhus; and the community I built at the Carleton College Summer Mathematics Program for Women, especially Rina Friedberg, Deanna Haunsperger, and Christina Knudson. Next, I'd like to thank the community I built through WiDS Chicago, especially Abby Smith and Mena Whalen. My thesis is the result of many collaborative projects, some with collaborators outside the university, and I am so grateful for all of them. In particular, I am so grateful for the climate science expertise and insights offered by Antonios Mamalakis and the rest of the TRIPODS team, and for Yue Gao's theoretical brilliance.

Finally, my friends and family have provided me with an overwhelming amount of love and support over the past five years. To my dad, David Stevens, thank you for the professorial advice and calm patience with me as I navigated this situation; to my mom Peg Cashell for always reminding me I have so much more to offer than my academic achievements, and to my brother Michael for the perspective. I couldn't possibly list all the friends who supported me throughout my PhD, but special thanks to Katie, Katie, Emily, Sami, Pat, Kyle, Chris, Julie, Ellie, Andy, and so many more. And to Quinn - I can't possibly express my gratitude to you over these past few years. I deeply truly couldn't have done this without you, and I am so excited for the beautiful life that awaits us. I love you so much.

PREFACE

This thesis is the result of three separate collaborations. Chapter 2 is drawn, with minor modifications, from the publication "Graph-Guided Regularized Regression of Pacific Ocean Climate Variables to Increase Predictive Skill of Southwestern U.S. Winter Precipitation," published in the Journal of Climate in December 2020 (Stevens et al., 2021) and jointly authored with Rebecca Willett at the University of Chicago; Antonios Mamalakis, Efi Foufoula-Georgiou, James Randerson, and Padharic Smyth at UC Irvine; Alejandro Tejedor at the Max Planck Institute, and Stephen Wright from UW Madison. This work was supported by an NSF TRIPODS-CLIMATE Grant (DMS-1839336). I (A. Stevens) was responsible for developing and implementing the algorithms and running all experiments, in addition to creating all tables and figures, while A. Mamalakis contributed the climatological context and interpretation in the manuscript.

Chapter 3 is drawn from a manuscript under review entitled "Lazy Estimation of Variable Importance for Large Neural Networks", jointly authored with Rebecca Willett along with Yue Gao and Garvesh Raskutti at UW Madison. I (A. Stevens) developed and implemented all algorithms, designed and performed all experiments, and wrote most of the manuscript, while Y. Gao was responsible for the theoretical analysis.

Finally, Chapter 4 is part of a large collaboration at Argonne National Lab, drawn from a variety of manuscripts, presentations, and extended abstracts. Sections 4.1-4.4 and 4.6-4.7 are drawn from a manuscript under preparation, "Development of City-Scale Synthetic Population To Simulate COVID-19 Transmission and Response," jointly authored with Chaitanya Kaligotla, Jonathan Ozik, Nicholson Collier, Sara Rimer, and Charles Macal, all part of the Decision and Infrastructure Sciences group at Argonne National Lab, and Anna Hotton and Bogdan Mucenic from the University of Chicago. This work appeared as an extended abstract at the Winter Simulation Conference 2020 (Kaligotla et al., 2020a). Section 4.5 is drawn from the manuscript "Load Balancing Schemes for Large Synthetic Population-Based

Complex Simulators" by the same authors, appearing at the 2021 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW) with many experiments carried out by (Mucenic et al., 2021). Section 4.8 is drawn from an ongoing collaboration entitled "odeling the Impact of Social Determinants of Health on COVID-19 Transmission and Mortality to Understand Health Inequities," funded by the C3.ai Digital Transformation Institute and involving, in addition to the authors already listed, Aditya Khanna, Harold Pollack, and John Schneider from UChicago. In addition to a manuscript by the same name that is in preparation, this section also draws from "Impact of changes in protective behaviors and out-of-household activities by age on COVID-19 transmission and hospitalization in Chicago, Illinois," currently under revision and by the same authors.

ABSTRACT

New statistical and machine learning methods have led to important advances in image and natural language processing, genetics, digital advertising, and other fields where there is an abundance of high-quality digital data and strong market incentives for automating tasks. This thesis intentionally focuses on areas such as climate science and public health, as they have suffered in this space without the same scale of training data and private investment.

Part 1 of this thesis focuses on applications in climate science, specifically forecasting precipitation at seasonal timescales, which is a challenge due to complex dependence structures and a short observational record. To address these challenges, we develop a regularization regression scheme using a graph-guided regularizer that simultaneously promotes sparsity and similar coefficients for highly correlated covariates. We propose a novel way of combining climate model simulations and observations by using large ensemble simulations from a climate model to construct this regularizer, highlighting the potential to combine optimally the space–time structure of predictor variables learned from climate models with new graph-based regularizers to improve seasonal prediction. In Part 2, we develop a fast and flexible method for estimating variable importance (VI) measures with large neural networks. Our VI measure of interest analyzes the difference in predictive power between a full model trained on all variables and a reduced model that excludes the variable(s) of interest, which can be expensive to compute. We replace the need for fully retraining a wide neural network to estimate the reduced model by a linearization initialized at the full model parameters. We provide inferential guarantees for our method and verify its performance on simulated and real data. Part 3 of this thesis describes the development of city-scale synthetic populations for use in an agent-based model (CityCOVID) that simulates the endogenous transmission of COVID-19 and measures the impact of public health interventions.

CHAPTER 1

INTRODUCTION

New statistical and machine learning (ML) methods have led to important advances in image and natural language processing, genetics, digital advertising, and other fields where there is an abundance of high-quality digital data and strong market incentives for automating tasks. ML systems are optimized for making accurate predictions on the training data they are fed. Oftentimes, the systems that do this best are difficult (if not impossible) to interpret, making it difficult to justify their use in the public sphere (e.g. why did this algorithm send me to jail?) The intersection of ML and the social/political domain is fraught, and careful consideration of the types of questions that need to be asked and answered by these ML systems is imperative to their advancement as decision-makers in society. This thesis intentionally focuses on areas such as climate science and public health, as they have suffered in this space without the same scale of training data and private investment.

1.1 Seasonal forecasting of SWUS precipitation

The first part of this thesis focuses on applications in climate science. As the existential threat of climate change looms, research communities are quickly emerging to try to use tools from ML and statistics to understand and mitigate the risks associated with a changing climate. A recent initiative, Climate Change AI[1], seeks to bring together volunteers from academia and industry to pool resources into a more cohesive effort to tackle climate change with machine learning; a comprehensive overview of the landscape of the types of research at this intersection can be found in Rolnick et al. (2019).

This work (Stevens et al., 2021) seeks to develop interpretable methods for predicting precipitation on seasonal timescales (see Figure 1.1 for specific lead times of different fore-

1. https://www.climatechange.ai/

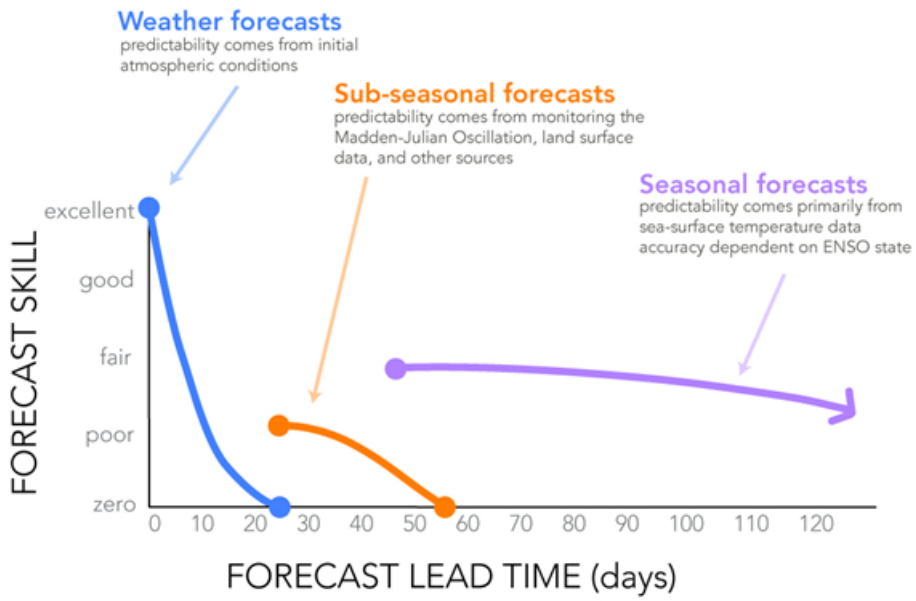


Figure 1.1: Heuristic of accuracy of different climate forecasting timeframes based on different predictors. Source: `https://wpo.noaa.gov/Programs/S2S`

casting timescales). Reliable prediction of seasonal precipitation has important implications for the economies and ecosystems of many regions in the world. Earth observations and climate model outputs are witnessing an unprecedented increase in data volume, from 80 terabytes today to 350 terabytes by 2030 (Overpeck et al., 2011). These unprecedented quantities of high-resolution climate data provide an opportunity to discover previously unknown teleconnections with strong predictive potential to improve seasonal forecasting. However, many statistical prediction schemes which aim to exploit established climate teleconnections between large-scale modes of variability (e.g., the El Niño-Southern Oscillation, the Pacific North America pattern, the Madden-Julian Oscillation, etc.) and regional hydroclimate either fail to capture the highly complex and nonstationary nature of the climate system, or suffer from overparameterization and increased risk of overfitting due to the limited sample size in the observational records. On the other hand, dynamical models show limited predictive skill at lead times longer than two weeks, due to imperfect physical conceptualizations and inaccurate initial conditions. Recently, collaborators on this project discovered a new

teleconnection between sub-tropical sea surface temperatures off the coast of New Zealand (NZI) and regional precipitation in the Southwestern US (SWUS) with stronger and earlier predictive potential than any other known mode of variability, including the El Nino Southern Oscillation (ENSO), which has long been used for SWUS seasonal precipitation forecasting (Mamalakis et al. (2018); Figure 1.2). The tracing of the SST dynamics that led to this discovery was performed in a physically intuitive way and was guided by their specific interest in the precipitation across the SWUS.

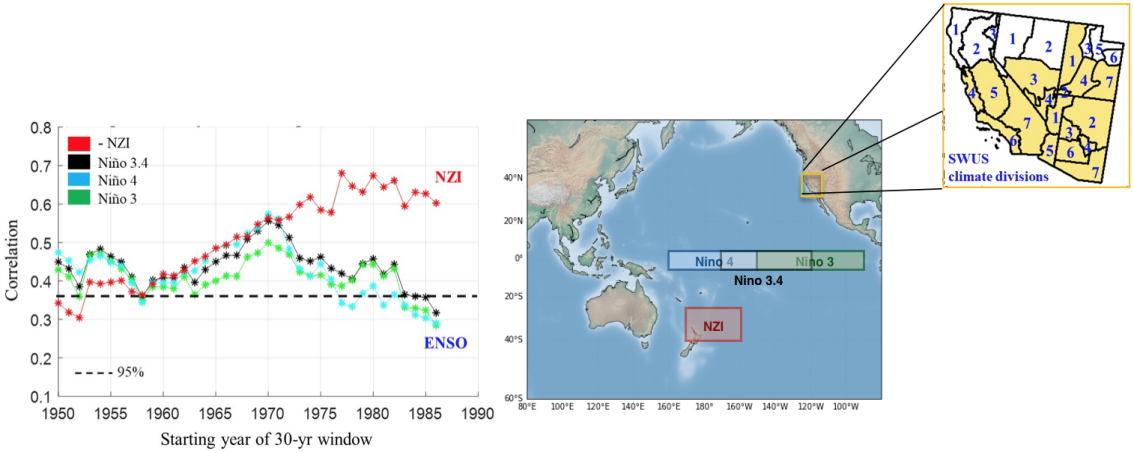


Figure 1.2: Weakening of the ENSO teleconnection after the mid-1970s and the emergence of a new teleconnection. 30-year running averages of cross-correlations between SST anomalies in Sept-Nov in the Niño 3, Niño 3.4 and Niño 4 regions in the equator and winter precipitation (Nov-March) in SWUS show low and decreasing predictability. The newly discovered NZI anomaly exhibits a stronger correlation (up to 0.7 compared to 0.4 for ENSO). Correlations above the dashed lines are statistically significant at the $\alpha = 0.05$ significant level.

We use this discovery as a motivating example of the importance of developing statistical methods which are able to robustly identify important climate dynamics. Rather than relying on ad hoc methods for discovering teleconnections, we seek to cast the forecasting problem as a regression problem in which modes are not specified in advance but rather are allowed to emerge from the data as sources of predictability. Such a method must account for small sample sizes and high dimensionality, strong spatiotemporal dependencies among the predictors, and the need for interpretability in climate science. Our contribution is

regularized regression scheme that reflects structural properties learned through simulated data from physical models. Specifically, we use climate models to estimate a correlation graph among features to form a graph-based regularizer, which estimates coefficients that are well-aligned with the spatiotemporal structure of the features. We apply the learned model to predict winter precipitation in the southwestern United States using sea surface temperatures over the entire Pacific basin, and demonstrate its superiority compared to other regularization approaches and statistical models informed by known teleconnections. Our results highlight the potential to combine optimally the space–time structure of predictor variables learned from climate models with new graph-based regularizers to improve seasonal prediction.

We opt to use a linear model (rather than a potentially more predictive black-box machine learning method) because they are interpretable and well understood by the climate community. When building models in applied scientific domains, we are often not only interested in how well the model predicts, but also *why* the model made the prediction it made. Because of this, many scientific practitioners hesitate to adopt black-box ML methods like neural networks due to the opacity of their decision-making. The next section of this thesis seeks to make neural networks more interpretable by providing a fast, flexible method for estimating *variable importance* measures.

1.2 Lazy estimation of variable importance

As opaque predictive models increasingly impact many areas of modern life, interest in quantifying the importance of a given input variable for making a specific prediction has grown. Recently, there has been a proliferation of model-agnostic methods to measure variable importance (VI) that analyze the difference in predictive power between a full model trained on all variables and a reduced model that excludes the variable of interest. For example, suppose we have data (X, y) and want to learn a regression model. Let

$X_{-j} = (X_1, \ldots, X_{j-1}, \tilde{X}_j, X_{j+1}, \ldots, X_p)$ be a copy of X with the j^{th} variable constructed so its dependence on y is broken (e.g. set to noise). We define the variable importance of variable j as

$$\text{VI}_j := \mathbb{E}\left[(y - \mathbb{E}[y|X_{-j}])^2\right] - \mathbb{E}\left[(y - \mathbb{E}[y|X])^2\right] \tag{1.1}$$

Our goal is to efficiently approximate VI_j when $\mathbb{E}[y|X]$ is estimated with a neural network. As written, computing $\{\text{VI}_j\}_{j=1}^p$ requires training $p+1$ models; the *full* model needs to be trained once, and then a "reduced" model needs to be learned for each of the p variables. This is potentially computationally intractable, particularly for high-dimensional data and complicated training networks. In this work, we propose a fast and flexible method for approximating the reduced model based on a linearization of the neural network around the full model parameters. Let h_θ be a neural network with parameters $\theta \in \mathbb{R}^M$, and suppose we learn the *full* model parameters

$$\theta_0 = \underset{\theta \in \mathbb{R}^M}{\arg\min} \frac{1}{n} \|y - h_\theta(X)\|_2^2 \tag{1.2}$$

In order to estimate VI_j, we need an estimate of what we are calling the *reduced* model $h_{\theta_{-j}}$, where

$$\theta_{-j} = \underset{\theta \in \mathbb{R}^M}{\arg\min} \frac{1}{n} \|y - h_\theta(X_{-j})\|_2^2 \tag{1.3}$$

Leveraging the *lazy training* framework of Chizat et al. (2020), instead of retraining a NN to estimate θ_{-j}, we can instead estimate the difference between the full model parameters θ_0 and θ_{-j} using a linear approximation and simply update the full model parameters with this correction to estimate $h_{\theta_{-j}}$:

$$h_{\theta_{-j}}(X_{-j}) \approx h_{\theta_0}(X_{-j}) + \nabla_\theta h_\theta(X_{-j})^T|_{\theta=\theta_0}(\theta_{-j} - \theta) \tag{1.4}$$

By adding a ridge-like penalty to make the problem convex, we prove that when the ridge penalty parameter is sufficiently large, our method estimates the variable importance measure with an error rate of $O(\frac{1}{\sqrt{n}})$ where n is the number of training samples. We also show that our estimator is asymptotically normal, enabling us to provide confidence bounds for the VI estimates. We demonstrate through simulations that our method is fast and accurate under several data-generating regimes, and we demonstrate its real-world applicability on a seasonal climate forecasting example.

1.3 Synthetic Population Development for COVID-19 Modeling

Part 3 of this thesis describes the development of city-scale synthetic populations for application to an agent-based model (CityCOVID) that simulates the endogenous transmission of Covid-19 and enables computational experiments to measure the impact of public health interventions. CityCOVID is based on the chiSIM framework (Macal et al., 2018) and is a city-scale agent-based model (ABM) of millions of people in a large metropolitan area, currently the Chicago area (Macal et al., 2020). CityCOVID is being used to understand the possible spread of Covid-19 and to model the uncertainties of human behavior in response to public health interventions. Underlying CityCOVID is a synthetic population (Kaligotla et al., 2020b; Macal et al., 2018) that is statistically representative of Chicago's population (2.7 million persons), along with their associated places (1.4 million places) and behaviors (13,000 activity schedules). During a simulated day, agents move from place-to-place, hour-by-hour, engaging in social activities and interactions with other co-located agents, resulting in an endogenous co-location or contact network. Covid-19 transmission is determined via a simulated epidemiological model based on this generated contact network by fitting model parameters that result in simulation output that matches observed daily COVID-19 death and hospitalization data from the City of Chicago.

Our synthetic environment is an augmentation of the RTI synthetic household population

database (Cajka et al., 2010). RTI uses the American Community Survey (ACS) and the Public Use Microdata Sample (PUMS) to generate geolocated households of individuals that statistically represent local demographics (age, sex, race, and household income) at a census block group level. In addition, RTI has generated synthetic workplaces that broadly match localized labor statistics, and has mapped school-aged agents to both private and public schools. To better understand COVID-19 transmission dynamics and outcomes, we expand the RTI synthetic population to include additional attributes like ethnicity and underlying health conditions, develop a new dataset of social places for agents to occupy using cell-phone mobility data, and enable CityCOVID to endogeneously generate contact matrices by assigning schedules to dictate when and where agents go.

With the highly granular synthetic population we develop, we are able to run a wide variety of experiments using the CityCOVID framework. We are able to identify loci of disease transmission, disentangle disparities in COVID-19 transmission and outcomes, and understand the impact of many different public health interventions. CityCOVID has informed policy making in the City of Chicago since the start of the pandemic and is an important development in the realm of highly detailed disease modeling.

Advancing statistical work in these fields at a speed commensurate with their urgency requires allowing applications to guide methodological developments. This thesis evaluates the theoretical and practical needs of the problems that are not currently being met, explores the ways in which these understudied areas can benefit from new applications of existing methodology, and, where this is lacking, proposes novel methods for these pressing issues.

CHAPTER 2

FORECASTING PRECIPITATION AT SEASONAL TIMESCALES

2.1 Introduction

Seasonal prediction of regional hydroclimate is typically based on deterministic physical models or statistical techniques, yet both approaches exhibit limited predictive ability (Wang, 2009; National Academies of Sciences, 2016). Precipitation predictions based on deterministic physical models (regional climate models) exhibit high uncertainty due to imperfect physical conceptualizations, sensitivity to initial and boundary conditions, and variations in model physics and grid resolutions (Chang et al., 2000). On the other hand, predictive statistical approaches (Wu et al., 2009; Schepen et al., 2012; Peng et al., 2014; Tao et al., 2017), which exploit historically and physically established climate teleconnections between regional hydroclimate and large-scale modes of climate variability (*e.g.,* the El Niño-Southern Oscillation, ENSO; see (Ropelewski and Halpert, 1986; Bradley et al., 1987; Redmond and Koch, 1991; McCabe and Dettinger, 1999; Dai, 2013)) also exhibit limited predictive skill. The main reason is that the complex and non-stationary interactions between large scale dynamics and regional hydroclimate cannot be captured sufficiently well with a limited number of pre-specified climate indices (regions used for computing sea surface temperature (SST) anomalies) as predictors, even when sophisticated statistical schemes are used (non-linear statistical schemes, Bayesian techniques, etc).

Recognizing the limitations above, the community has been increasingly embracing the application of methods that aim to learn from both climate models and statistical schemes in order to improve seasonal predictive skill. These methods range from weighted multimodel averaging techniques (Raftery et al., 2005; Luo et al., 2007; Schepen and Wang, 2013; Cheng and AghaKouchak, 2015) or methods that directly combine predictions from climate models

and statistical schemes (Coelho et al., 2004; Schepen et al., 2014; Madadgar et al., 2016), to data-driven approaches based on machine learning, in a setting where predictor variables are not pre-specified but rather are guided by the data or climate model outputs (Quan et al., 2006; DelSole and Banerjee, 2017; Hewitt et al., 2018; Ham et al., 2019; Willard et al., 2021; He et al., 2021). In the former category of methods, the prediction skill depends strongly on the skill of each of the models considered, thus making such techniques prone to all the aforementioned limitations. In contrast, the use of machine learning has potential since climate information from the entire globe can in principle be used to inform the prediction. However, these techniques also face important practical limitations. First, because of the short record of observations and the large number of predictors, the number of degrees of freedom of the problem is vast, significantly increasing the risk of overfitting (Ham et al., 2019). Second, strong spatiotemporal dependences among the predictor variables, which are certainly present in climate applications, need to be taken into consideration for imposing structure in the predictor space, to reduce the dimensionality of the problem and improve physical interpretability.

In this study, we aim to address the challenges discussed above by introducing a regularized regression scheme that accounts explicitly for spatiotemporally correlated predictors. Regularization is an established technique in statistics, machine learning, and signal processing that can mitigate the challenges of many degrees of freedom relative to the amount of data. The key idea is that rather than simply finding the model that best fits the data according to some loss function, we instead minimize the sum of the loss and a regularization function, where the latter reflects some prior belief about which models are better than others. Sparsity regularization (e.g. the least absolute shrinkage and selection operator (LASSO) regularization; Tibshirani and Taylor (2011) has already been explored in the context of precipitation downscaling and data assimilation (Ebtehaj et al., 2012; Ebtehaj and Foufoula-Georgiou, 2013) and climate forecasting (DelSole and Banerjee, 2017; He

et al., 2021), but suffers from ignoring the spatiotemporal dependencies among predictors. In order to respect the embedded space-time structure of the climate system and enforce sparsity, we use a "graph total variation" (GTV) regularizer (i.e. constraint) that promotes similarity of weights (i.e. regression coefficients) for highly correlated predictors. The GTV is a graph-based regularizer, based on the graph formed by the covariance matrix of the predictors, which was recently introduced by Li et al. (2020). To address the issue of the short observational record, and to robustly estimate the covariance matrix of the predictor variables, we make use of a large ensemble of climate model outputs. Using climate model outputs in the training of machine learning (ML) models is a subcase of the general category of techniques that aim to integrate physical knowledge and machine learning Willard et al. (2021), and it has recently been shown to be highly efficient in increasing predictive skill on seasonal to interannual timescales (DelSole and Banerjee, 2017; Ham et al., 2019). Although our study differs from these studies in that it uses the climate model outputs not to train the ML model per se, but only to compute the covariance matrix of the predictors used as a GTV regularizer, it adds to this important new line of research in synergistically leveraging both climate models and observations with the goal of improving prediction skill.

We explore the prediction skill of our methodology for the case study of predicting precipitation over the southwestern US (SWUS), focusing on the winter season (specifically, Nov-Mar), when the majority of precipitation occurs. Despite the increasing attention that it has received over the years (Schonher and Nicholson, 1989; McCabe and Dettinger, 1999; Gershunov and Cayan, 2003; Schubert et al., 2016; Madadgar et al., 2016; Liu et al., 2018; Hao et al., 2018; Zhang, 2018; Mamalakis and Foufoula-Georgiou, 2018; Pan et al., 2019), early and accurate prediction of winter precipitation in SWUS remains a challenge, with significant implications for the region's population and economy (Howitt et al., 2014; Kay et al., 2015; AghaKouchak et al., 2015; Medellín-Azuara et al., 2016). Traditional climatic drivers of SWUS precipitation (e.g. ENSO) explain just a small fraction of the interannual

variability of precipitation totals, which in some cases are determined by a small number of winter storms (Dettinger et al., 2011; Dettinger and Cayan, 2014). Moreover, it is known that the ENSO relationship with SWUS climate undergoes multidecadal fluctuations McCabe and Dettinger (1999); Yu et al. (2012), with many recent studies pointing out that it has been losing strength in the recent decades, while the western Pacific climatic state is gaining in importance (Wang et al., 2014; Baxter and Nigam, 2015; Teng and Branstator, 2017; Seager et al., 2017; Swain et al., 2017; Myoung et al., 2018; Mamalakis et al., 2018; Lee et al., 2018). The special difficulty of this problem also arises because the SWUS lies within a transition zone between the subtropics and the mid latitudes (i.e. 30°- 40° N). In fact, the latter is among the reasons that the effect of climate change on future precipitation trends over the SWUS is highly uncertain, with mid-latitude regions expected to become wetter and subtropical regions drier Allen and Luptowitz (2017). Because of its intrinsic complexity, this region offers an excellent case study for exploring and benchmarking data-driven predictive methods.

As predictor variables, we use late summer and early fall (Jul, Aug, Sep, and Oct) sea surface temperature (SST) over the entire Pacific basin. Note that although there are studies indicating the importance of Atlantic Ocean temperatures as drivers of SWUS precipitation as well (Enfield et al., 2001; McCabe et al., 2004), our focus here is only on the Pacific Ocean, as a first step. We cast the prediction problem as an estimation problem in which predictors are not specified in advance, but rather emerge from the data by minimizing an appropriate loss function. We first demonstrate the increased predictive skill of the proposed GTV model when the covariance matrix that defines the GTV regularizer is computed from a large ensemble of a climate model, rather than the single realization of observations. Second, we benchmark the GTV model against two different classes of predictive models: (1) other regularized regression methods (LASSO and fused LASSO; Tibshirani et al. (2005)) and (2) simple ordinary least squares using known teleconnection indices as predictors. Our analysis

shows that constraining the predictive model by the spatiotemporal covariance of the predictors via a GTV regularization outperforms all the other considered models and substantially increases the seasonal precipitation predictive skill. Lastly, we show that both the GTV performance and the emerged predictors of precipitation are quite robust to perturbations in the covariance matrix that is used to define the GTV regularization term.

The structure of the paper is as follows. In section 2, we describe the prediction problem and the data used. We introduce the proposed methodology in section 3 and discuss the advantages of using a graph-based regularizer in which the graph is based on covariance information from a climate model, instead from the limited observations. In section 4, we present results on the performance of our proposed model and compare its skill to other methods. Moreover, we study the emergent predictors, aiming to gain physical insight about the drivers of SWUS precipitation. We also perform a sensitivity analysis of our results to gain more confidence in the predictive performance and the emergent predictors. Conclusions and directions for future research are discussed in section 5.

2.2 Prediction problem and data/models used

The SWUS (California, Nevada, Utah and Arizona) is composed of 25 climate divisions, for each of which precipitation series are maintained by and publically accessible thanks to the NOAA National Centers for Environmental information[1] (Vose, 2014). The season we aim to predict precipitation is Nov-Mar, when the majority of the annual precipitation occurs; note also that winter precipitation, especially that which is stored as snowpack, is necessary for sustaining the water supply through relatively dry summers (Mote et al., 2005; Shukla et al., 2015; Liu et al., 2018). As shown in Figure 2.1a, the northwestern part of the region receives much higher precipitation than the rest of the SWUS, which is generally considered a fairly dry area. However, the interannual variability in the central and southern

1. `https://www.ncdc.noaa.gov/cag/time-series/us`

part is high, generally higher than 40-50% of the mean precipitation (coefficient of variation of the order of 0.4-0.5 or higher), compared to the northern part of the SWUS, where the coefficient of variation is 0.3 or lower (Figure 2.1b). To distinguish between the two different hydroclimatologies of the northern part and the central/southern part of the SWUS, previous studies have used different approaches, such as focusing on the area below a certain latitude (Liu et al., 2018) or considering only the climate divisions for which a specific predictor (e.g. the Niño 3.4 index; an ENSO index) exhibits a significant relation with precipitation (Mamalakis et al., 2018). Here, we distinguish between the two precipitation regimes based on the area-weighted first principal component (PC1). As can be seen from Figure 2.1c, PC1 (which explains about 64% of the total precipitation variability in the SWUS) is more strongly associated with the central and southern part of the SWUS than with the northern part, thus providing a quantitative way of differentiating between the two regions. The region selected for our analysis, composed of 18 climate divisions, is shown in Figure 2.1d, together with the series of the area-weighted average precipitation.

As predictor variables, we use late summer and early fall (Jul, Aug, Sep, and Oct) SSTs over the Pacific basin, which is defined as the area in 60°S - 60°N and 80°E - 280°E. Historical time series of SST (monthly series on a $1^\circ \times 1^\circ$ grid, see Hirahara et al. (2014)) are made publically available by NOAA/OAR/ESRL PSL[2]. At that resolution, the number of predictor variables is very large (roughly $120 \times 200 \times 4 = 96{,}000$) making the problem highly ill-posed. Thus, we upscale the original SST field by simple areal averaging into grid boxes of size 10° by 10° over the Pacific basin, to reduce the dimensionality of the problem. After removing the boxes over land, we end up with roughly 900 predictor variables in total (i.e. four months over 12 boxes in latitude and 20 boxes in longitude).

The analysis is performed for the years 1940-41 to 2018-19, since SST records are not trustworthy before the 1940s, due to the limited availability of observations over the Pacific

2. https://www.esrl.noaa.gov/psd/data/gridded/data.cobe2.html

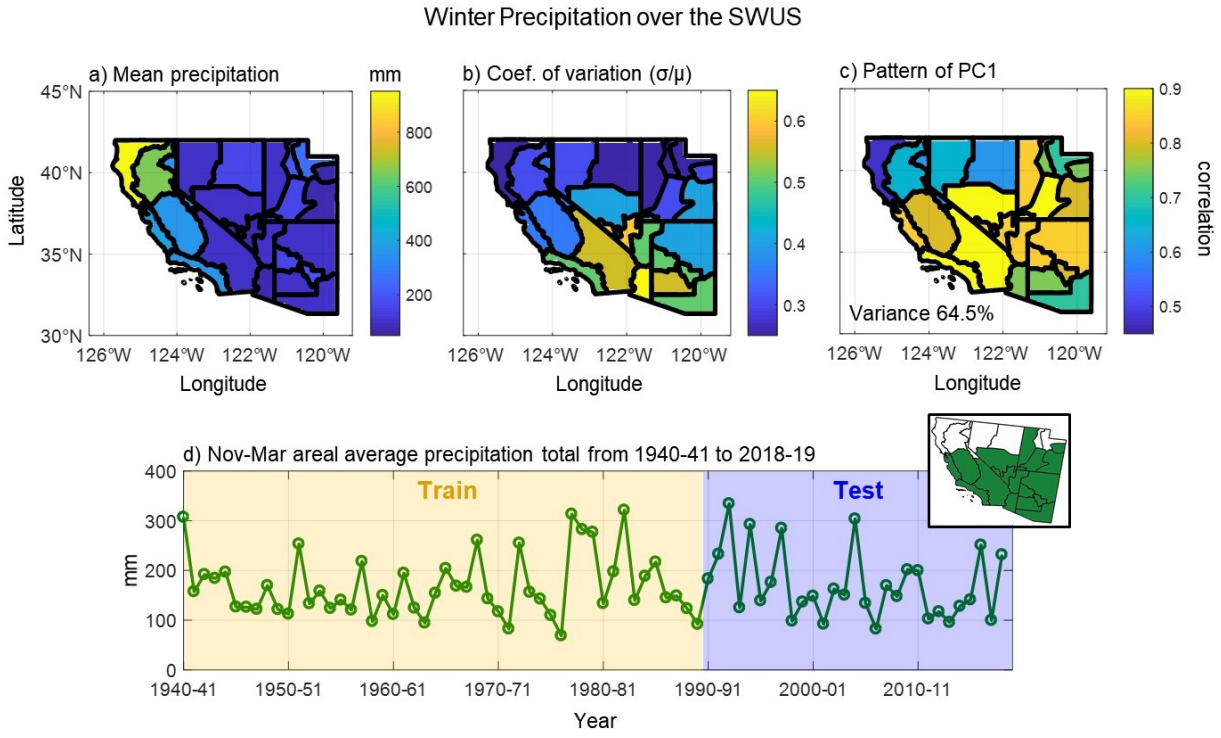


Figure 2.1: Spatial patterns of winter precipitation statistics over the southwestern US. a) Multi-year mean of Nov-Mar precipitation over SWUS, for the period from 1940-41 to 2018-19. b) Coefficient of variation (sample standard deviation divided by sample mean) of Nov-Mar precipitation for the period from 1940-41 to 2018-19. c) Correlation of the first area-weighted principal component (PC1) of the Nov-Mar precipitation over the SWUS and precipitation in each climate division. d) Series of the area-weighted average precipitation over the climate divisions considered in this study (see panel to the right). In our study, we use years 1940-41 to 1989-90 as a training period (for model fitting), and years 1990-91 to 2018-19 as a test period (for model evaluation).

basin and globally (Deser et al., 2010). In particular, we use years 1940-41 to 1989-90 as a training period, and years 1990-91 to 2018-19 as a test period. All individual SST series are linearly detrended and standardized (zero mean and unit variance) before they are used in the analysis.

To reduce the uncertainty in the estimation of the spatio-temporal dependency of the predictors (covariance matrix of SST predictors, used to define the GTV regularizer), we also examined simulations from the Community Earth System Model-Large ENSemble project

(CESM-LENS; Kay et al. (2015))[3]. Specifically, we use monthly series (on a $1.25^\circ \times 0.9^\circ$ grid) of surface temperatures over the Pacific basin, which we also upscale to $10^\circ \times 10^\circ$ grids to match the grids used for the observed SSTs. We note that the CESM-LENS project consists of 40 ensemble members, each one corresponding to the same model physics but different initial conditions in the atmosphere. CESM-LENS relies on historical boundary conditions for the period 1920-2005, and the representative concentration pathway 8.5 (RCP8.5) is applied as forcing for years 2006-2100. Here we used only archives of simulation output from 1940-2005 to build our model covariance matrices, as the focus of our analysis is on improving prediction for the contemporary period. Note that the 40 ensemble members constitute independent, but equally probable trajectories of the Earth system with historical forcing (see Kay et al. (2015) for more information).

2.3 Methodology

Let $y_r^{(i)}$denote the winter precipitation in year i and climate division r. We hypothesize that $y_r^{(i)}$ can be predicted from climate variables at different locations over the Pacific Ocean and different lag times (e.g., months ahead of the winter period), with a model of the form:

$$y_r^{(i)} = \sum_{j=1}^{p} x_j^{(i)} \beta_{j,r} + \epsilon_r^{(i)} \tag{2.1}$$

where $\epsilon_r^{(i)}$ is a Gaussian noise $N(0, \sigma^2)$ term. Writing (2.1) in a matrix form and dropping the index r for convenience results in

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{2.2}$$

3. `http://www.cesm.ucar.edu/projects/community-projects/LENS/data-sets.html`

where $\mathbf{y} = \left(y^{(1)}, y^{(2)}, \ldots, y^{(n)}\right)^T \in \mathbb{R}^n$ is the vector of winter precipitation over n years, $\mathbf{X} = \left[\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}\right]^T \in \mathbb{R}^{n \times p}$ is the matrix of climate variables, i.e., SSTs over the Pacific Ocean and in the four different months preceding the winter season (Jul, Aug, Sep, and Oct), $\beta = \left(\beta_1, \beta_2, \ldots, \beta_p\right)^T \in \mathbb{R}^p$ is the vector of weights corresponding to p predictors, and $\varepsilon = \left(\varepsilon^{(1)}, \varepsilon^{(2)}, \ldots, \varepsilon^{(n)}\right)^T \in \mathbb{R}^n$ is a Gaussian noise vector. We clarify that $\mathbf{x}^{(i)}$is a p-dimensional vector in year i of SSTs arranged by moving along all longitudes and latitudes of the Pacific Ocean and for the four months of July, August, September, and October. Thus, $p = 900$ in our case, while the number of available years is $n = 79$ (i.e. we use 50 years for training and 29 years for testing; from 1940-41 to 1989-90 and from 1990-91 to 2018-19, respectively). Obviously, this problem is highly under-determined, since $n \ll p$. To solve for β, we reduce the effective dimension of the problem by adding regularization terms, leading to the formulation:

$$\hat{\beta} = \arg\min_{\beta} \{\frac{1}{n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda R(\beta)\} \tag{2.3}$$

where $R(\beta)$ is a regularization term, chosen to impose structure and sparsity on β, and $\lambda > 0$ is the regularization parameter. A popular choice for $R(\beta)$ is the LASSO regularizer (Tibshirani, 1996), i.e., $R(\beta) = \|\beta\|_1 = \sum_{i=1}^{p} |\beta_i|$, which yields minimizers of (3) for which β is sparse, i.e., there are only a few spatiotemporal variables that are truly predictive while the rest are conditionally independent of the response $\mathbf{y}$. The value of λ, typically estimated using cross-validation, reflects the weight given to the sparsity constraint. However, the LASSO regularization does not take into account that predictors might have a significant spatiotemporal dependence structure which, if included, might further constrain and improve the prediction.

2.3.1 Graph Total Variation (GTV)

To overcome this problem, we propose a regularizer that accounts explicitly for the spatiotemporal covariance of the predictors. The central idea of this regularizer is that if covariates $\mathbf{x}_j$ and $\mathbf{x}_k$ are highly correlated with one another, then they should receive similar weights $\hat{\beta}_j$ and $\hat{\beta}_k$. This approach helps us select highly correlated *collections* of covariates that serve as strong predictors of precipitation. In contrast, the LASSO estimator would generally select either $\hat{\beta}_j$ or $\hat{\beta}_k$, but not both, and the selected covariate would be very sensitive to any noise in the data. We form a graph to represent the correlations between pairs of covariates, and select a set of weights β that is "aligned" with the graph. This regularization scheme, known as Graph Total Variation (GTV), was introduced in Li et al. (2020). Although graph-based regularizers have been explored before (e.g. fused LASSO, edge LASSO, graph-trend filtering), Li et al. (2020) developed theoretical guarantees for the GTV regularizer for highly correlated covariates, and showed how imposing additional structure on β to encourage "alignment" with the covariance graph can lead to optimal solutions. This property is important in our problem since $\mathbf{X}$ contains highly correlated columns, resulting for example from dependence between SST anomalies at nearby locations for small time lags or at distant locations but lagged in time.

Let$\hat{\mathbf{\Sigma}}$ be an estimate of the covariance matrix of $\mathbf{X}$ and let $\hat{s}_{j,k} = sign\left(\hat{\Sigma}_{j,k}\right)$. The GTV estimator is given by

$$\hat{\beta} = arg\min_{\beta}\left\{\frac{1}{n}\left\|\mathbf{y}-\mathbf{X}\beta\right\|_2^2 + \lambda_{TV}\sum_{j,k}\left|\hat{\Sigma}_{j,k}\right|^{1/2}\left|\beta_j - \hat{s}_{j,k}\beta_k\right| + \lambda_1\left\|\beta\right\|_1\right\} \tag{2.4}$$

where $\lambda_1$and λ_{TV} are regularization parameters chosen through cross validation. Here we use a standard 5-fold cross validation approach applied to the training data (i.e. a total of 50 years) to estimate the optimal $\left(\lambda_1^*, \lambda_{TV}^*\right)$combination. Specifically, we split the training dataset into five non-overlapping, random 10-yr sets; for each of the five sets, we train our

model (i.e. estimate $\hat{\beta}$) using the other four sets and compute the prediction error on the held out fifth set. This is repeated for each candidate tuning parameter pair $(\lambda_1, \lambda_{TV})$. The optimal $\left(\lambda_1^*, \lambda_{TV}^*\right)$combination is the one that, on average, minimizes the prediction error across the five different holdout sets. Note that choosing the value of λ_{TV} that determines the importance of the GTV term via cross-validation can mitigate the effect of any systematic biases reflected in the estimate $\hat{\mathbf{\Sigma}}$, since if $\hat{\mathbf{\Sigma}}$ were not informative at all, the optimal λ_{TV} would be close to zero.

We can interpret the estimator in Equation 2.4 from a graph perspective by defining a covariance graph based on $\hat{\mathbf{\Sigma}}$. Let $\mathbf{G} = (\mathbf{V}, \mathbf{E}, \mathbf{W})$ be an undirected weighted graph with vertices $\mathbf{V} = \{1, 2, \ldots, p\}$, edges $\mathbf{E} := \left\{(j,k) : \left|\hat{\Sigma}_{j,k}\right| > \theta, j \neq k\right\}$, and weight matrix $\mathbf{W}$ with $w_{j,k} = \left|\hat{\Sigma}_{j,k}\right|^{1/2}$. That is, each predictor variable (e.g. SST at a particular place and time) is associated with one of the nodes of the graph, and edges reflect the pairs of predictors that are correlated. A threshold parameter θ can be applied to the covariance matrix (Bickel and Levina, 2008) for assessing with edges (i.e., links between covariates) will be used in the GTV term (see Section 2.3.3 for further discussion about parameter θ).

The expression in (2.4) may be rewritten using new notation that highlights connections with previous methods and known software for solving the optimization problem. Specifically, let $\mathbf{\Gamma} \in \mathbb{R}^{|\mathbf{E}| \times p}$ be the *weighted edge incidence matrix* of $\mathbf{G}$, where each row l represents a pair of connected vertices (j_l, k_l):

$$\begin{aligned} \Gamma_{l,j_l} &= w_{j_l,k_l} \\ \Gamma_{l,k_l} &= -\hat{s}_{j_l,k_l} w_{j_l,k_l} \end{aligned} \tag{2.5}$$

Then, we can write (2.4) as

$$\hat{\beta} = arg\min_{\beta} \left\{ \frac{1}{n} \left\| \mathbf{y} - \mathbf{X}\beta \right\|_2^2 + \lambda_{TV} \left\| \mathbf{\Gamma}\beta \right\|_1 + \lambda_1 \left\| \beta \right\|_1 \right\} \tag{2.6}$$

As mentioned above, GTV promotes estimates of β that contain sparse clusters of coefficients, each cluster corresponding to a highly correlated set of variables. That is, the stronger the correlation between $\mathbf{x}_j$ and $\mathbf{x}_k$, the more similar $\hat{\beta}_j$ and $\hat{\beta}_k$. The $\|\beta\|_1$term promotes overall sparsity. We note that (2.6) can be viewed as a generalized LASSO estimator (Tibshirani and Taylor, 2011), for which a number of efficient implementations exist.

2.3.2 Other regularization methods

This work sits alongside a growing body of literature of structured estimation in high dimensions with a specific application to climate data (e.g., Goncalves et al. (2016)). A variety of regularization schemes have recently shown promise in improving predictive skill by imposing structure and sparsity on the predictors. Chatterjee et al. (212) proposed using the Sparse Group Lasso (SGL), in which the regularizer is given by

$$R(\beta) = f\|\beta\|_1 + (1-f)\|\beta\|_{1,G} \tag{2.7}$$

where $G = \{G_1, G_2, \ldots, G_M\}$ are M groups of variables across multiple locations and times. This scheme yields solutions in which variables at certain locations and times are simultaneously selected or else zeroed out. Using this scheme to predict monthly temperature and precipitation showed significant improvement over LASSO. He et al. (2019) proposed a weighted LASSO given by

$$R(\beta) = \sum_{i=1}^{p} w_i |\beta_i| \tag{2.8}$$

where the weights are chosen to be proportional to the distance between the location of the feature and the location of its response. This penalizes predictors that are far away from the region of interest and it is not appropriate for our problem, in which long-distance climate teleconnections play an important role.

Finally, we note that GTV is a special case of the fused LASSO estimator (Tibshirani

et al., 2005). While these estimators have significant theoretical support, the theory relies on the assumption that $\mathbf{X}$ is full rank and does not consider the role of correlations among columns of $\mathbf{X}$. Furthermore, the edges included in the fusion penalty are assumed to be highly structured (i.e. only direct spatial or temporal neighbors), and the theory does not generalize to the types of unstructured covariance graphs that arise in many applications. In climate and other domains, there are known long-range correlation patterns that would not be captured by a direct neighbor penalty. We will, however, benchmark GTV against the fused LASSO, which has regularization term

$$R(\beta) = \sum_{j,k \in \mathcal{N}} \left|\beta_j - \beta_k\right| + \lambda_1 \|\beta\|_1$$

$$\mathcal{N} := \left\{(j,k) \,|(\mathbf{x}_j, \mathbf{x}_k) \text{ are spatially adjacent}\right\} \cup \left\{(j,k) \,|(\mathbf{x}_j, \mathbf{x}_k) \text{ are temporally adjacent}\right\} \tag{2.9}$$

2.3.3 Using climate model outputs to compute the covariance of the SST predictors

The theoretical guarantees of GTV depend on a sufficiently accurate estimate of the covariance matrix of the Pacific SSTs $\mathbf{\Sigma} := \mathbb{E}(\mathbf{X}^T\mathbf{X})$. However, for our problem where $n \text{ ¡¡ } p$, the sample covariance $\frac{1}{n}\mathbf{X}^T\mathbf{X}$ is a highly uncertain estimate of $\mathbf{\Sigma}$ (Bickel and Levina, 2008; Cai et al., 2016). Thus, we propose to explore the use of ensemble simulations from climate models, the size of which is several times larger than that of observations, in order to reduce the uncertainty of the covariance matrix estimate and improve the performance of the GTV regularized regression. While we acknowledge that climate models might not accurately capture all multi-scale space-time variability of SSTs in the Pacific (e.g. see the studies of (Szoeke and Xie, 2008; Kim et al., 2014; Bellenger et al., 2014; Li and Xie, 2014; Wang and Miao, 2018), we assert that leveraging their information content to improve high dimensional

data-driven predictive methods offers great potential and deserves careful examination. In a recent study by Ham et al. (2019), climate model outputs were used in the context of "physics-guided initialization" (the term is adopted from Willard et al. (2021)). Particularly, the adopted ML model was first trained using climate model outputs, so some initial estimates of the weights were obtained. Then, as a second step, prediction was performed by fine-tuning the weights using historical data (a process known as "transfer learning").

Here, we suggest that defining the GTV regularizer using covariance information from climate models can increase predictive performance. We term this approach "physics-guided regularization". We rigorously demonstrate the merits of this approach using the SST outputs from the 40 ensemble members of CESM-LENS. Since CESM-LENS simulations are produced on a different spatial grid from that of the SST observations, we interpolated late summer and early fall SSTs from CESM-LENS linearly onto the observation grid. We emphasize that the CESM-LENS outputs are used only to estimate the covariance matrix of the predictors, while the training of our model (estimation of the regression parameters and coefficients) and its performance evaluation (see Section 2.4) are always performed using the observed SST and precipitation series in the training and test periods, respectively.

Letting $\mathbf{X_{CL}} \in \mathbb{R}^{40n \times p}$ be the detrended and standardized (zero mean and unit variance) matrix of stacked SST variables from all the CESM-LENS members, we define $\hat{\mathbf{\Sigma}}_{\mathbf{CL}}$ as the sample covariance of $\mathbf{X_{CL}}$. We also define $\hat{\mathbf{\Sigma}}_{\mathbf{obs}}$ as the covariance matrix estimated from the observations. These covariance matrices are $p \times p$ matrices in which all considered variables are ordered by longitude, latitude, and month, resulting in repetitive patterns arising from the spatial and temporal dependencies (see Figure 2.2a for $\hat{\mathbf{\Sigma}}_{\mathbf{obs}}$). Visually, there is no striking difference in the dependence structure of SSTs between different months (Figure 2.2a, left panel). The highest correlations (both positive and negative) are found in the tropics, with strong SST couplings along the eastern and central tropical Pacific basin (high positive correlation) and between the eastern and western tropical Pacific basin (high

negative correlation), features that are a consequence of the ENSO (Wang et al., 2017); e.g., see the zoom-in panels in Figure 2.2a for the October covariance matrix. Figure 2b shows $\hat{\mathbf{\Sigma}}_{\mathbf{obs}}$and $\hat{\mathbf{\Sigma}}_{\mathbf{CL}}$ for the month of October. Although some differences are observed, the CESM-LENS appears to capture well the spatial structure of the observed SST correlations (i.e. tropics vs extratropics etc.), and as demonstrated in section 4.1, the reduced uncertainty of $\hat{\mathbf{\Sigma}}_{\mathbf{CL}}$ adds significant predictive skill and robustness to the GTV model.

To finalize the construction of the graph underlying the GTV regularization, we further process the SST covariance matrix $\hat{\mathbf{\Sigma}}_{\mathbf{CL}}$ estimated from the output of CESM-LENS, using a thresholding procedure with statistical guarantees (see Bickel and Levina (2008); Li et al. (2020)), this method simply sets elements of the sample covariance with absolute value under a certain threshold equal to zero. That is, for a threshold θ, the covariance graph $\mathbf{G}$ has edges $\mathbf{E} := \left\{(j,k) : \left|\hat{\Sigma}_{j,k}\right| > \theta, j \neq k\right\}$. In addition to the statistical advantage of this thresholding in yielding more consistent covariance estimates, thresholding is also useful from a computational perspective, as it drastically limits the number of edges used in the regularization term (i.e., the number of rows in $\mathbf{\Gamma}$ in Equation (6)). We treat the threshold θ as a model parameter and estimate its optimal value in a cross validation training setting (i.e., similarly to λ_1 and λ_{TV}; see section 3.1), which allows us to disregard the smaller, less certain SST correlation values that, if included, would have led to a worse performance (i.e., if $\hat{\mathbf{\Sigma}}_{\mathbf{CL}}$ were not informative at all, the optimal θ would be close to one).

2.3.4 Accounting for non-stationarity in precipitation teleconnections

The last issue that our analysis aims to account for is possible non-stationarities in the strength of the precipitation teleconnections. Traditionally, precipitation in the SWUS has been linked to various large scale modes of climate variability, and more commonly the state of ENSO (Schonher and Nicholson, 1989; Redmond and Koch, 1991; Mo and Higgins, 1998; McCabe and Dettinger, 1999; Cayan et al., 1999). Physically, El Niño (or La Niña)

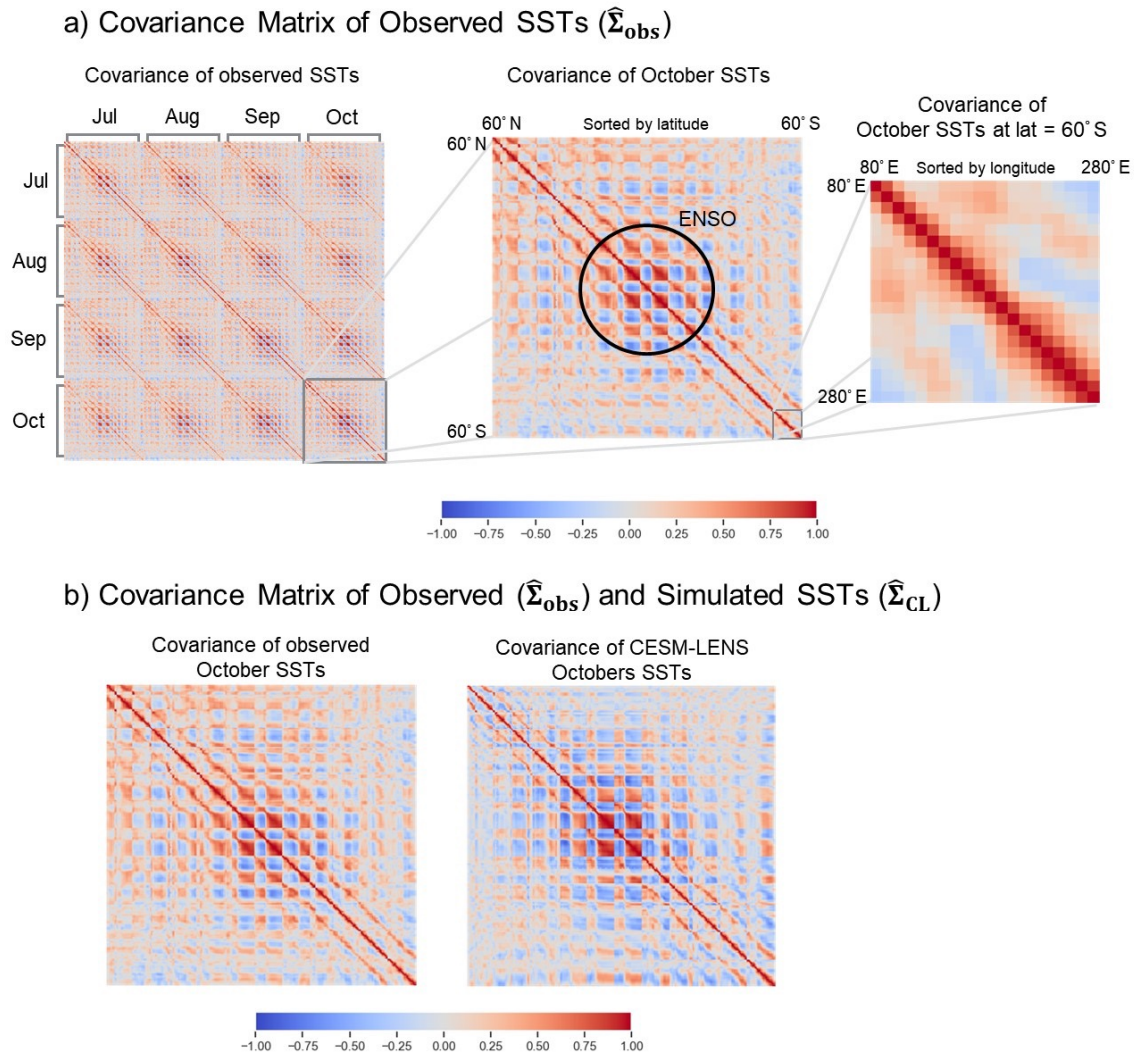


Figure 2.2: Space-time covariance of sea surface temperatures (SSTs) in the Pacific Ocean. a) Sample covariance matrix of the observed Pacific SSTs over longitude-latitude and for four months. Zoom-in covariance in October highlights the spatial extent of the tropical ENSO signal and a further zoom-in at the fixed latitude of 60°S shows the spatial longitudinal dependence. b) Comparison of the sample covariance of October Pacific SSTs as estimated from the observations and the output of CESM-LENS.

events typically associate with persistent low (or high) atmospheric pressure patterns over the northeastern Pacific (a teleconnection which materializes via quasi-stationary Rossby waves (Trenberth et al., 1998; Castello and Shelton, 2004)), and thus disturb the location and strength of the winter time jet stream, which can then bring more (or fewer) winter storms to the SWUS, leading to wet (or dry) conditions over the SWUS and dry (or wet)

conditions over northwestern US. However, recent research shows that the ENSO effect on the atmospheric pressure and (consequently) on precipitation over the eastern Pacific and North America has been decreasing in strength during the last 3-4 decades, while many studies have highlighted to a greater or lesser extent that the western Pacific climatic state (e.g. SSTs) has been a stronger driver of precipitation variability over North America (Wang et al., 2014; Baxter and Nigam, 2015; Teng and Branstator, 2017; Seager et al., 2017; Swain et al., 2017; Myoung et al., 2018; Mamalakis et al., 2018). On a similar note, new research (Johnson et al., 2019) shows that during the last 3-4 decades, western Pacific SSTs have been important players in affecting the connection between the tropical atmospheric circulation and the eastern tropical Pacific SSTs, during weak ENSO events, which highlights changes in the tropical Pacific dynamics (see also Mamalakis et al. (2019). Whether these changes in precipitation teleconnections and Pacific dynamics are a result of internal multidecadal climate variability or anthropogenic forcing is still not clear. However, to acknowledge the non-stationary nature of the prediction problem, we herein use a weighted loss function that gives more weight to the more recent years in the training data set, that is, the period after the 1970-80s. This is roughly the time that most studies have pinpointed as the start of these changes (Wang et al., 2014; Swain et al., 2016; Mamalakis et al., 2019; Johnson et al., 2019), and it is also the period during which the SWUS precipitation variability (inter-annual variance) has started to increase (see Figure 1). As such, with regard to the data-fit term of Equation 2.6, we minimize the weighted loss function

$$\sum_{t=1940}^{1989} \left\{ a^{1989-t} \left(y^{(t)} - \langle \mathbf{x}^{(t)}, \beta \rangle \right)^2 \right\} \tag{2.10}$$

with a being a discount factor set to $a = 0.90$. This simple but effective approach is widely used in the forecasting literature (e.g., Hyndman and Athanasopoulos (2018); Livneh and Badger (2020)) and gives preference to the relationship between Pacific SSTs and SWUS precipitation in the more recent decades, while still retaining some information from earlier

years.

2.4 Results

In Figure 2.3, we summarize schematically the proposed approach based on the GTV described in the previous sections, for the prediction of winter precipitation totals in SWUS. We inform our prediction using observed Pacific SSTs during the late boreal summer and early fall (Jul, Aug, Sep, Oct) and we form a space-time covariance graph with edge weights corresponding to pairwise correlations (normalized covariances) between SST boxes of 10° by 10°, to constrain our regularization scheme, in addition to the traditionally used LASSO term. Correlations are obtained from the output of climate models (i.e. using SST outputs from 40 ensemble members of the CESM-LENS, for the period 1940-2005) to decrease estimation uncertainty and improve prediction in the test period.

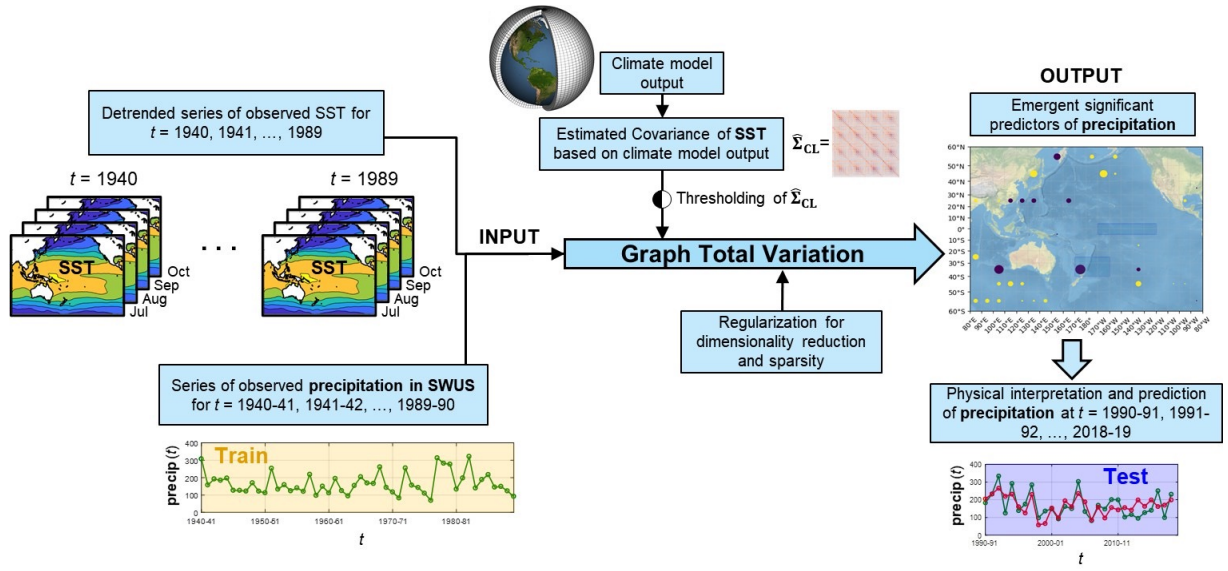


Figure 2.3: Schematic for the graph-guided regularization (Graph Total Variation, GTV) for predicting winter precipitation over SWUS. We use observed Pacific SSTs as input to the predictive model, and we form a space-time covariance graph with edge weights corresponding to pairwise SST covariances to constrain the model via a GTV regularization term. The covariance matrix is estimated based on the output of climate models and subjected to a hard thresholding (see Sec 2.3.3) to increase the consistency of the dependency among predictors for improved performance of the GTV algorithm.

2.4.1 Predictive performance of the GTV model

The GTV model (2.6) was fitted in the training period (from 1940-41 to 1989-90) and the optimal threshold value $\theta^*$and optimal parameter values $\left(\lambda_1^*, \lambda_{TV}^*\right)$ were estimated through a 5-fold cross validation procedure (see section 3.1). For the case of the areal average precipitation over SWUS, this procedure identified the values θ^*= 0.5 and $\left(\lambda_1^*, \lambda_{TV}^*\right)$ = (0.013, 0.0007). To test the sensitivity and robustness of the GTV model to this optimal choice of parameters, and also to showcase the advantage of using the CESM-LENS versus the covariance of observations, we start by presenting and discussing results for three different values θ = (0.35, 0.5, 0.75) and various $(\lambda_1, \lambda_{TV})$ combinations. Figure 4 shows the October SSTs covariance for those different θ thresholds, highlighting the sparseness of the covariance matrix as the threshold increases. Although not shown in Figure 4, the dependency graph formed by the thresholded covariances has, as expected, a decreasing number of links (it is sparser) as θ increases. Specifically, the degree of the GTV graph (defined as the average number of edges each node is connected to) is 80503, 32644, and 5357 for θ = 0.35, 0.5 and 0.75, respectively, highlighting the computational advantages that the reduced-degree graph also offers.

The middle-column panels in Figure 2.4 show the model performance in the test period, measured by the coefficient of determination R^2, as a function of the different combinations of$(\lambda_1, \lambda_{TV})$ parameters. On the same panels, the optimal set of parameters $\left(\lambda_1^*, \lambda_{TV}^*\right)$obtained from the 5-fold cross validation in the training period, conditional on the three values of θ, is also shown. It is observed that for θ^*= 0.5, the optimal parameters $\left(\lambda_1^*, \lambda_{TV}^*\right)$ robustly fall within the region for which the model performance in the test period is also optimal. This illustrates that optimally thresholding the covariance to reduce the spatiotemporally correlated predictors used in the regularization, avoids overfitting, and increases model accuracy. Moreover, our results show that the GTV model explains about R^2 = 40% of precipitation variance in the test period; even for parameter values other than the

optimal $\left(\theta^*, \lambda_1^*, \lambda_{TV}^*\right)$, R^2 is consistently higher than 30%. This is a significant improvement over the R^2 values obtained in prior work, since commonly used teleconnection indices typically result in a much lower fraction of explained variance, on the order of 10-20% (see Lee et al. (2018); Deser et al. (2018), and Figure 2.6 herein). The latter indicates that informing the GTV regularizer based on the covariance matrix $\hat{\mathbf{\Sigma}}_{\mathbf{CL}}$ improves the prediction.

To highlight further the merit of using the climate model covariance $\hat{\mathbf{\Sigma}}_{\mathbf{CL}}$ versus the covariance of the observations$\hat{\mathbf{\Sigma}}_{\mathbf{obs}}$**,** the rightmost column panels in Figure 2.4 present the same analysis as the middle column panels, but using $\hat{\mathbf{\Sigma}}_{\mathbf{obs}}$ instead. It is telling that the performance in the test period in this case is inferior (smaller R^2 values) for all combinations of parameters $(\lambda_1, \lambda_{TV})$ and threshold values θ.

Having established the robustness of the GTV model using $\hat{\mathbf{\Sigma}}_{\mathbf{CL}}$ with the optimal parameters θ^*= 0.5 and $\left(\lambda_1^*, \lambda_{TV}^*\right)$ = (0.013, 0.0007), we compare in Figure 5 the predicted and observed precipitation series for the years from 1990-91 to 2018-1. The predicted series explains about 42% of the precipitation variability and captures adequately many of the extreme precipitation years, i.e. the wet years 1992-93, 1994-95, 1997-98 and 2004-05, and the dry years 1998-99, 2001-02 and 2006-07. Also, the probability of dry/wet hit is high (high chance in predicting dry/wet, when actually dry/wet conditions occur). Specifically, if we define a wet/dry year to be a year that falls above/below the multi-year precipitation average, then our results indicate that our method exhibits a wet hit probability of 64% and dry hit probability of 72%. Moreover, the residuals between the prediction and observations are found to be normally distributed and exhibit insignificant autocorrelation at a 0.05 significance level (see Figure 2.5b,c), consistent with the "white noise" assumption in (1)-(2).

Finally, the residuals of the prediction do not show statistically significant correlation with the Pacific SSTs (Figure 5d), indicating that no information in the Pacific SSTs is left unexplored in the predictive model. Only a few small and incoherent SST patterns are found

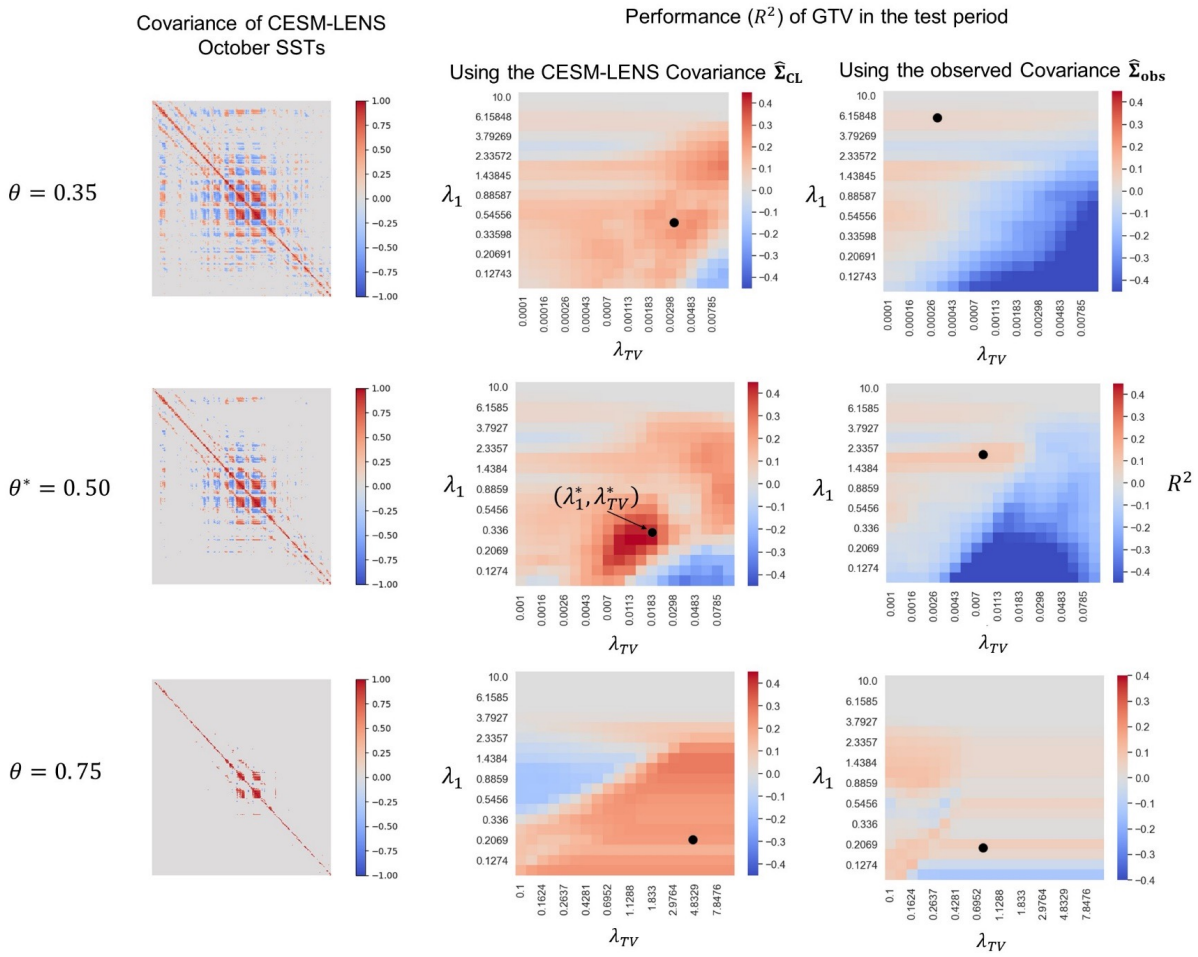


Figure 2.4: Sensitivity analysis of the GTV model performance for a range of covariance thresholds θ and regularization parameters λ_1 and λ_{TV}. Left column panels show the covariance of the October SSTs (as estimated based on the CESM-LENS) for three different thresholds ($\theta = 0.35, 0.5,$ and 0.75). The middle (right) column panels show the coefficient of determination (R^2) between the areal average observed and model predicted precipitation in the test period (1990-91 to 2018-19), when the CESM-LENS covariance (observed SST covariance) is used to define the GTV regularizer. In all panels, and conditional on the corresponding values of θ, the optimal $(\lambda_1^*, \lambda_{TV}^*)$ pair for each model (obtained by using a 5-fold cross validation in the training period 1940-41 to 1989-90) is shown (black dots). These results highlight that (a) the use of the CESM-LENS covariance, instead of the observational covariance, to inform the GTV regularizer leads to a highly robust and improved predictive performance, as judged by the larger domain of regularization parameters with high R^2 values (see middle column plots, as compared to their right counterparts), and (b) our choice to use a threshold of $\theta^* = 0.5$, which was based on a 5-fold cross validation in the training period, shows to yield the most robust and highest predictive performance.

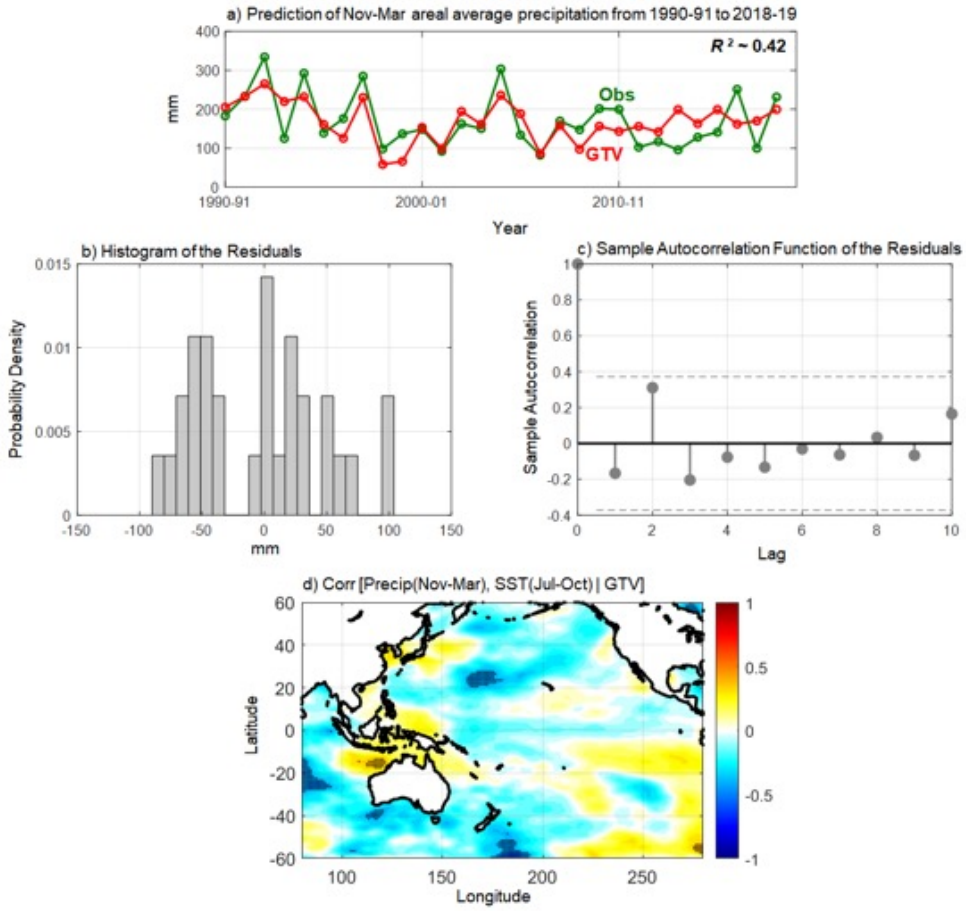


Figure 2.5: Evaluation of the prediction of winter (Nov-Mar) precipitation. a) Series of observed (green) and predicted (red) Nov-Mar areal average SWUS precipitation during the test period from 1990-91 to 2018-19. Prediction is made using the sample covariance from the CESM-LENS output. b-c) Histogram and autocorrelation function of the residual time series during the test period (residuals are between GTV predictions and observations shown in (a)). The null hypothesis that the residuals are normally distributed is not rejected at a 0.05 significance level. Also, the null hypothesis that there is no year to year linear dependence (autocorrelation) in the residual time series is not rejected at the 0.05 significance level. d) Partial correlation between SWUS precipitation in Nov-Mar and linear detrended grid point SSTs in Jul-Oct, after accounting for the GTV prediction. Stippling indicates locally significant correlations. The absence of significant correlation patterns indicates that no more predictive information can be extracted from the Pacific basin SSTs, providing confidence for the fitted model.

to significantly correlate to precipitation at a 0.05 significance level, probably due the fact that in Figure 5d we simultaneously test multiple "local" hypotheses, which increases the chances of a type I error (i.e. rejecting a true null hypothesis; Wilks (2016)). Thus, we

can conclude that our model sufficiently exploits the Pacific SST information, and that any deviation (residual) between our predictions and reality either comes from other forcings not included in our analysis (e.g. SST variability over the Atlantic Ocean; Enfield et al., 2001), or is the result of internal stochastic variability.

2.4.2 Benchmarking against other predictive models

In this section, we compare the prediction skill of the GTV, for all climate divisions over the SWUS and the areal average precipitation, to other regularized regression models and models based on commonly used teleconnections (Figure 2.6). Specifically, we benchmark our results against the following methods:

- LASSO: standard ℓ_1-penalized regression; coefficients are penalized so that the solution is very sparse.
- Fused LASSO: Direct spatial and temporal neighbors are penalized to have similar coefficients (2.9).
- GTV (Obs): GTV with the regularization term defined using the covariance matrix estimated from the observations and thresholded at $\theta^*= 0.5$.
- GTV (CESM-LENS): GTV with the regularization term defined using the covariance matrix estimated from CESM-LENS and thresholded at $\theta^*= 0.5$.
- Ordinary least squares using known teleconnection indices.

We highlight that all methods are trained in years from 1940-41 to 1989-90 using the weighting formula described in Section 2.3.4 to account for non-stationarity and tested in years from 1990-91 to 2018-19. Only observations are used for training and testing; the CESM-LENS output is used simply to define the regularization term in the GTV (CESM-LENS), but not to actually fit or test the model.

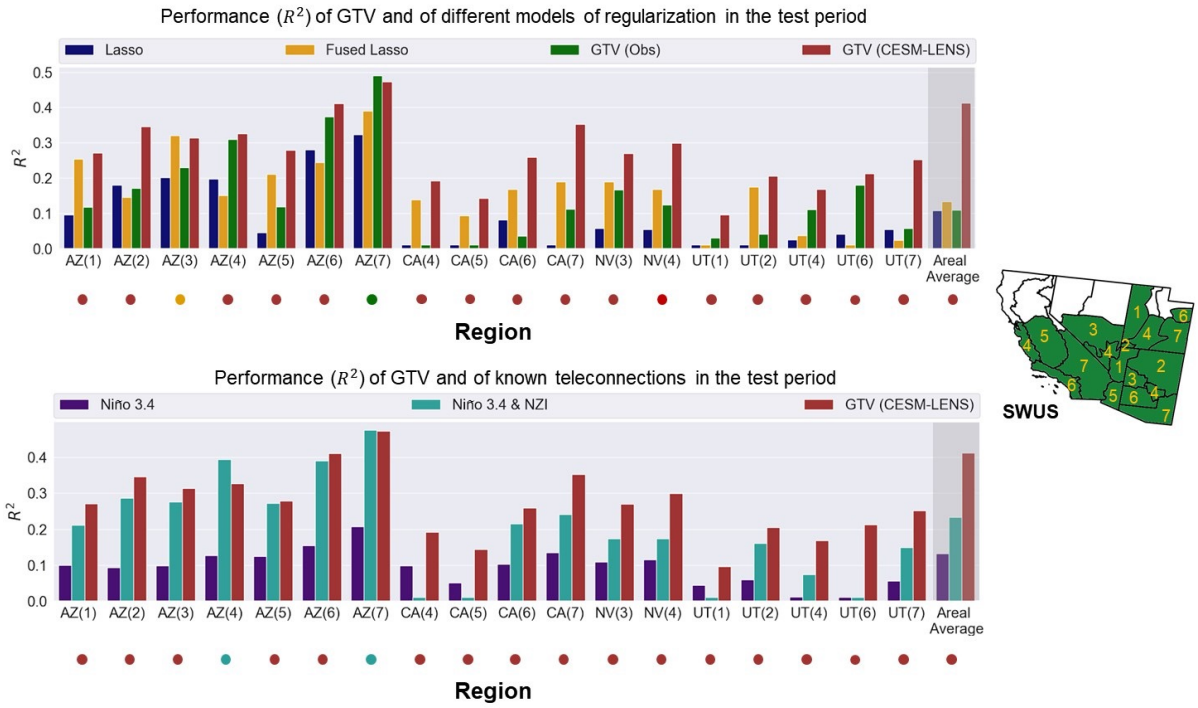


Figure 2.6: Performance of GTV and different methods of regularization (top panel) and of known teleconnections (bottom panel) in predicting precipitation totals over different SWUS divisions in the test period from 1990-91 to 2018-19. The coefficient of determination (R^2) is presented. It is shown that GTV with model-estimated covariance of SSTs outperforms all other regularization methods (top) as well as statistical regression on two known teleconnections (bottom).

First, we find that the prediction accuracy differs significantly among climate divisions of the studied region. Most notably, prediction of the northern climate divisions, CA(4), CA(5), UT(1), UT(6) and UT(7) is poorer relatively to climate divisions over most of Arizona. The fact that this holds for all models indicates that the signal of the Pacific SSTs to precipitation is weaker as one moves to northern California, Nevada, and Utah, which is in accordance with other studies (see Schonher and Nicholson (1989); McCabe and Dettinger (1999); Castello and Shelton (2004); Mamalakis et al. (2018), and our discussion in section 2). With regard to the best performing model, our results show that the proposed GTV model reproduces the highest fraction of precipitation variability over almost all climate divisions, ranging from almost $R^2 = 0.5$ in AZ(7) to $R^2 = 0.1$ in UT(1), and $R^2 = 0.42$ for the areal average

precipitation, when using $\hat{\mathbf{\Sigma}}_{\mathbf{CL}}$. When using$\hat{\mathbf{\Sigma}}_{\mathbf{obs}}$, the performance is poorer and similar to the performance of LASSO and fused LASSO, in terms of the areal average precipitation. Fused LASSO performs slightly better than GTV in AZ(3), but it only slightly exceeds $R^2 = 0.1$ for the areal average precipitation.

As a benchmark, we also compare the prediction performance of GTV with known physical teleconnections. Specifically, we train a weighted (see section 2.3.4) linear regression scheme using the averaged Jul-Oct Niño 3.4 index as our predictor, which captures ENSO variability and is typically associated with SWUS precipitation. We also use a weighted bivariate regression model combining the Niño 3.4 index and the New Zealand Index (NZI) over the same summer months. The NZI has been shown to exhibit high correlation with precipitation over the last four decades (Mamalakis et al., 2018). The latter interhemispheric teleconnection has been suggested to materialize through a western Pacific ocean-atmosphere pathway, whereby SST anomalies in the southwestern Pacific during late boreal summer can modulate time-lagged anomalies of the same sign in the northwestern and central Pacific via perturbation of the regional southern Hadley cell, which in turn affect the jet stream and winter storm tracks to the US west coast. Our results show that ENSO-based predictions explain about 10-15% of the precipitation variability over most climate divisions. When NZI is added, the prediction performance increases significantly and the explained variance is almost twice as high for the areal average SWUS precipitation. However, in almost all climate divisions, the GTV(CESM-LENS) model outperforms all other models. Similar conclusions are reached also based on the mean squared error (see Table 2.1), where the GTV model is *not* the best performing in only three climate divisions out of the 18 (i.e. in AZ(3), AZ(4) and AZ(7)).

Generally, the results described above, and summarized in Table 2.1 and Figure 2.6, show that GTV (CESM-LENS) robustly outperforms the competing regularized regression schemes and known teleconnections, offering promise for increasing the predictive skill of

Region	**Niño 3.4**	**Niño 3.4 & NZI**	**Lasso**	**Fused Lasso**	**GTV (Obs)**	**GTV (LENS)**
Arizona (1)	1.1515	1.0091	1.1556	0.9549	1.1279	**0.9321**
Arizona (2)	1.0219	0.8037	0.9234	0.9631	0.9349	**0.7372**
Arizona (3)	1.1344	0.9102	1.0036	**0.8540**	0.9691	0.8625
Arizona (4)	0.9812	**0.6815**	0.9017	0.9536	0.7758	0.7571
Arizona (5)	1.1927	0.9918	1.3016	1.0754	1.2016	**0.9833**
Arizona (6)	0.9684	0.6985	0.8250	0.8656	0.7178	**0.6744**
Arizona (7)	0.9239	0.6104	0.7885	0.7105	**0.5936**	0.6135
California (4)	0.8560	0.9695	0.9469	0.8174	0.9549	**0.7672**
California (5)	0.9169	1.1322	0.9663	0.8756	0.9663	**0.8270**
California (6)	1.0367	0.9062	1.0598	0.9606	1.1127	**0.8552**
California (7)	1.1252	0.9870	1.3009	1.0544	1.1543	**0.8409**
Nevada (3)	0.9632	0.8927	1.0182	0.8761	0.9006	**0.7889**
Nevada (4)	1.1680	1.0907	1.2489	1.0980	1.1567	**0.9255**
Utah (1)	1.1706	1.2818	1.2238	1.2196	1.1871	**1.1074**
Utah (2)	1.1295	1.0078	1.1973	0.9915	1.1512	**0.9545**
Utah (4)	0.9243	0.8656	0.9108	0.9000	0.8312	**0.7773**
Utah (6)	0.8958	0.9667	0.8537	0.8849	0.7305	**0.7012**
Utah (7)	0.8462	0.7622	0.8470	0.8740	0.8450	**0.6701**
Areal Average	0.9818	0.8671	1.0087	0.9803	1.0074	**0.6652**

Table 2.1: Mean square error (MSE) of different methods of regularization and teleconnections in predicting precipitation totals over different SWUS divisions in the test period from 1990/91 to 2018/19. Precipitation series has been standardized (zero mean and unit variance). For the GTV model the covariance threshold of $\theta^* = 0.5$ has been used. Bold font indicates the method with the lowest MSE for each climate division.

winter precipitation over the SWUS.

2.4.3 Physical interpretation of the predictors

In this section, we seek insight into which SST patterns play an important role in driving winter precipitation variability over the SWUS. In doing so, we seek physical interpretations of the "optimal" solutions of regression coefficients corresponding to each regularization method. We repeat here that we estimate $\hat{\beta}$ by (i) applying a 5-fold cross validation technique to the training data to estimate the regularization parameters of each model, and (ii) minimizing the corresponding loss function using the estimated regularization parameters

from (i). Although this analysis is not suitable to draw rigorous causal inferences, it can highlight important sources of predictability for precipitation, which should be physically interpretable.

The optimal coefficients $\hat{\beta}$ for the LASSO model are presented in Figure 2.7a. Keeping in mind that the LASSO regularization promotes sparsity, this method essentially pinpoints the few regions around the Pacific basin, over which late summer and early fall SSTs contained the highest predictive information for Nov-Mar precipitation, during the training period from 1940-41 to 1989-90. Specifically, negative regression coefficients on the order of -0.3 are found over the tropical and subtropical western Pacific, and positive coefficients of the same order are found over the southern hemisphere mid-latitudes. Fused LASSO (Figure 2.7b), which promotes the assignment of similar weights to neighboring grid boxes, yields a smoother version of the LASSO solution, in which the majority of Pacific participates in the prediction, but many regions (grid boxes) contribute in a negligible way (many coefficients are on the order of 10^{-3}-10^{-4}). There is some contribution by the northern mid-latitude SSTs, but the highest weights are again assigned to the tropical and subtropical western Pacific (negative coefficients), and to the southern hemisphere mid-latitudes, especially over the southeastern Pacific (positive coefficients). Lastly, slightly different solutions are obtained by the GTV model when using $\hat{\mathbf{\Sigma}}_{\mathbf{obs}}$ (Figure 2.7c) or $\hat{\mathbf{\Sigma}}_{\mathbf{CL}}$ (Figure 2.7d). However, in terms of the patterns of the SST predictors (i.e. not in terms of grid by grid comparison), there is some consistency between these two variants over the southwestern Pacific Ocean (where both models exhibit high negative coefficients), and the north eastern and central Pacific basin (positive coefficients).

Although the GTV model (when using $\hat{\mathbf{\Sigma}}_{\mathbf{CL}}$) gives the best prediction performance in the test period, for our physical interpretation, we focus on SST patterns that are consistent across all methods. Specifically, in accordance with recent studies (Wang et al., 2014; Seager et al., 2017; Swain et al., 2017; Myoung et al., 2018; Mamalakis et al., 2018), all models

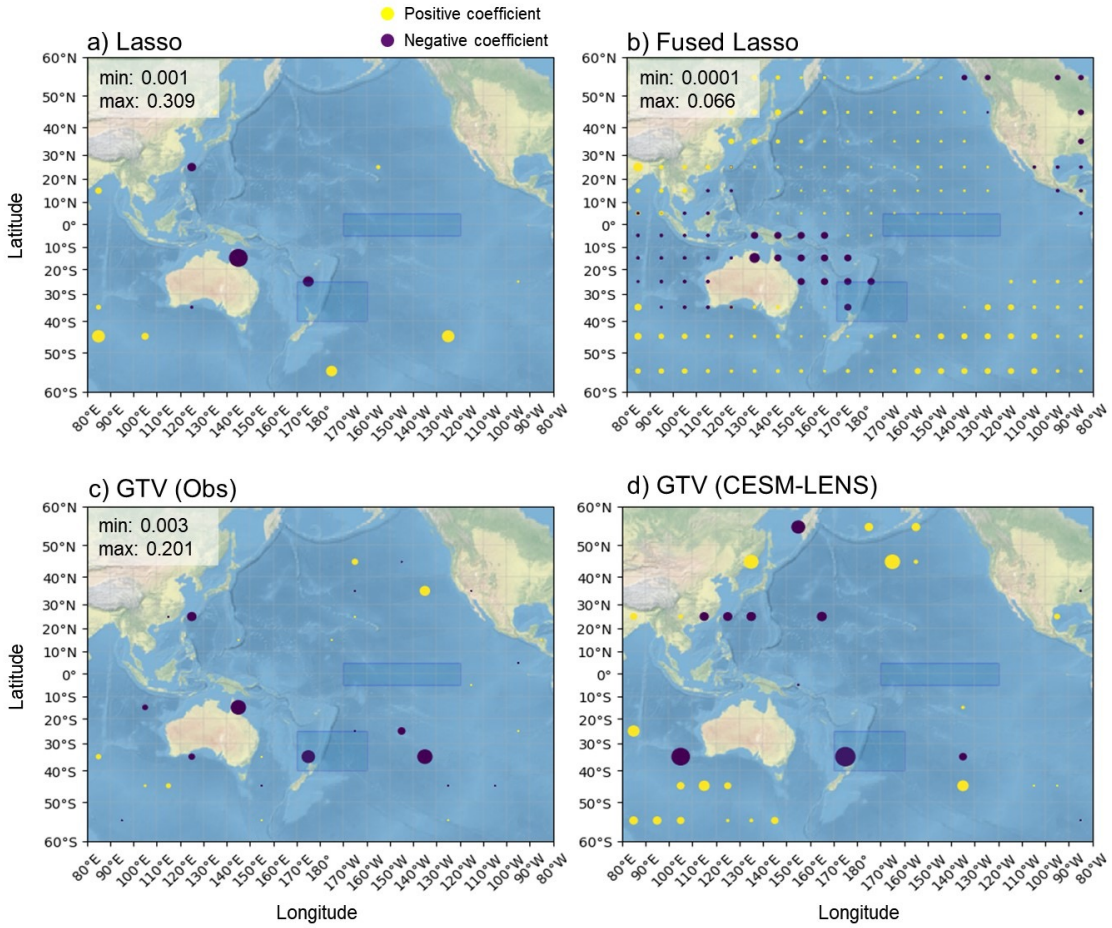


Figure 2.7: The emergent predictors of the areal average SWUS winter precipitation for different models of regularization; a) LASSO, b) Fused LASSO, c) GTV using the sample covariance of the observed SSTs, and d) GTV using the sample covariance from the CESM-LENS output. The $\hat{\beta}$ values are presented (colored circles) after training each method in the training period 1940-1989, using a 5-fold cross validation technique. The color of the circles indicates the sign of the $\hat{\beta}$, values (yellow for positive and purple for negative), while the size of circles is proportional to their magnitude; for each method, the minimum and maximum $\hat{\beta}$, values (in absolute terms) are also given. Niño 3.4 and NZI boxes are also shown. All models highlight to a greater or lesser extent the western and southwestern Pacific SSTs as important predictors of SWUS precipitation.

highlight to a greater or lesser extent the western Pacific SSTs as important predictors of California and SWUS precipitation, rather than the eastern Pacific SSTs. Physically, it has been shown that the western tropical Pacific is a region over which anomalous convection can be an important source of Rossby wave energy, which teleconnects through a quasi-stationary

Rossby wave train with the atmospheric pressure over the northeastern Pacific, affecting the location of the jet stream, and eventually precipitation totals in the north America (Wang et al., 2014). Moreover, the southwestern Pacific (close to New Zealand) has been highlighted in the literature as a special region in leading tropical climate. First, some studies support that climate variability (e.g. SST, sea level pressure, etc.) over the southwestern Pacific leads by a few seasons the ENSO variability (Trenberth and Shea, 1987; van Loon and Shea, 1987; Stephens et al., 2007), and specific indices have been suggested to increase predictive skill of ENSO state (Hamlington et al., 2015). Given that ENSO is known to be related with SWUS precipitation during winter (i.e. for zero lead time), the southwestern Pacific SSTs may provide important predictors of precipitation, by leading the ENSO state, and are highlighted by all models in our analysis. By contrast, eastern tropical Pacific SSTs are not shown to be predictive, since our analysis considers nonzero lead times. More recent studies, however, suggest that western Pacific SSTs can also affect precipitation through a Western Pacific Pathway i.e. not necessarily through ENSO teleconnections (Mamalakis et al., 2018). The latter has been suggested to materialize through the seasonal migration of the intertropical convergence zone and the associated expansion of the southern Hadley cell during late summer (Waliser and Gautier, 1993; Berry and Reeder, 2014; Mamalakis and Foufoula-Georgiou, 2018), which allows for persistent SST anomalies to impact the atmospheric circulation and climate variability in the western tropical Pacific, which as noted earlier is a key region of Rossby wave energy. This teleconnection has been increasing in importance during the last 40 years, which is also the time when new, ENSO-independent SST patterns have been emerging and affecting tropical atmospheric circulation (Johnson et al., 2019).

2.4.4 Sensitivity of the GTV model to uncertainty in the covariance matrix

Finally, to explore the sensitivity of the GTV model to perturbations of the covariance matrix used to define the regularization term, we perform a bootstrap analysis. Namely, rather than stacking all 40 CESM-LENS trajectories to form the covariance matrix, we resample the 40 trajectories (with replacement) and compute the sample covariance of the new sample. Next, we form our GTV regularization term using this resampled covariance matrix, then fit the GTV scheme in the training period, and finally calculate the coefficient of determination (R^2) in the test period. By repeating this procedure 1000 times, we can quantify how the uncertainty in the covariance matrix propagates to uncertainty in the regression coefficients $\hat{\beta}$and model performance. Our results show that the GTV model always captures more than $R^2 = 30\%$ of the variability of the Nov-Mar areal-average precipitation, in some cases reaching $R^2 = 45\%$. The bootstrap average is on the order of $R^2 = 40\%$ and the bootstrap standard deviation is about 5% (see Figure 2.8a). These results indicate that the GTV model is not particularly sensitive to the uncertainty in the covariance matrix and always outperforms all alternative predictive models.

Regarding the propagation of uncertainty to the regression coefficients, the average vector of $\hat{\beta}$ across the 1000 bootstrap realizations (see Figure 2.8b) very closely resembles the results presented in Figure 7d, indicating the importance of the southwestern Pacific Ocean (high negative coefficients), and the northeastern and central Pacific basin (positive coefficients). Moreover, although in some grid points the standard deviation of $\hat{\beta}$ across the 1000 bootstrap realizations is of the same magnitude as the average value, most of the largest coefficients in Figure 2.8b are characterized by small underlying uncertainty in Figure 2.8c, which implies that they are not sensitive to covariance perturbations as quantified here. This provides confidence that these SST features mostly located in the western Pacific are indeed important sources of predictability of the Nov-Mar SWUS precipitation.

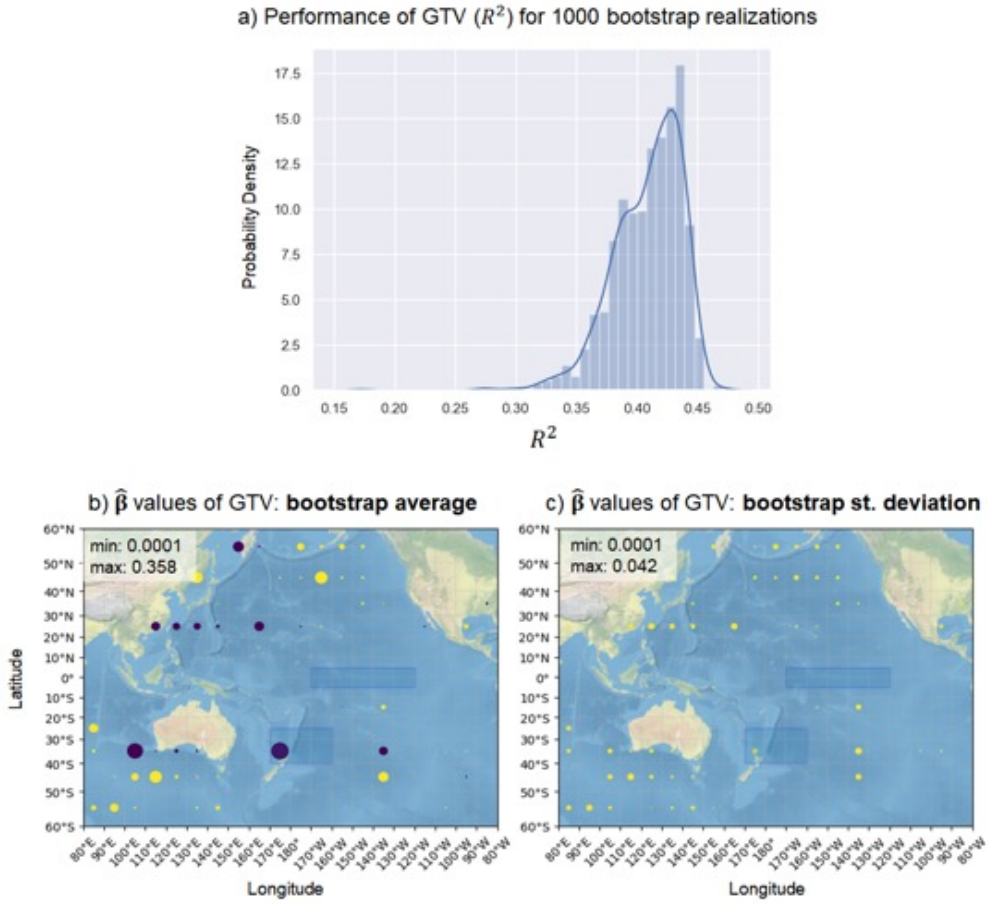


Figure 2.8: Bootstrap investigation of the sensitivity of the GTV to the uncertainty of the covariance estimated from CESM-LENS. a) The histogram of the coefficient of determination (R^2) between the observed Nov-Mar SWUS precipitation and the GTV prediction) across all 1000 bootstrap realizations. b) The vector-average of the 1000 $\hat{\beta}$ vectors from the 1000 bootstrap realizations. For each realization, training is performed in the period 1940-1989, using a 5-fold cross validation technique. c) Same as in (b), but the standard deviation of the 1000 $\hat{\beta}$ vectors is presented. The small uncertainty of the most important predictors (grids with the largest $\hat{\beta}$ values) is noteworthy.

2.4.5 Conclusions and future work

In this study, we approached the problem of early prediction of winter precipitation over the SWUS by using machine learning methodologies to increase predictive skill relative to traditional approaches of utilizing dynamical models or relying on empirically established teleconnections. We use late summer and early fall SST information to predict precipitation based on a newly proposed regularized regression scheme, specifically designed to account for

high dimensionality and high spatiotemporal dependence structure in the predictor variables, making it well suited to climate applications. The proposed predictive model accounts for high spatiotemporal dependence structures in the predictors expressed as a graph, which is then used to define a Graph Total Variation (GTV) regularizer that promotes similar weights for highly correlated predictors. We also address the short observational record and high dimensionality of the problem by using LASSO terms that promote sparsity, as well as by using large-ensemble outputs from climate models to decrease the structural uncertainty in the estimation of the SST covariance matrix.

Our analysis shows that predictive skill for SWUS precipitation can be increased considerably by using our novel regularization methodology, explaining more than 40% of the average precipitation variability over the SWUS. Our model's performance is higher than any other regularized regression model (LASSO and fussed LASSO), and it also outperforms models based on known teleconnection indices. Our results also show that, in accordance with recent literature (DelSole and Banerjee, 2017; Ham et al., 2019), climate models can be used in a non-conventional way (e.g., for training rather than predicting and, in our case, for building the graph-based regularizer) towards increasing prediction accuracy. With regard to important regions/sources of precipitation predictability, our analysis highlights the tropical and subtropical western Pacific SSTs as the most consistently important predictors of precipitation, which have increasingly gained attention in the literature (Wang et al., 2014; Swain et al., 2017; Mamalakis and Foufoula-Georgiou, 2018). Finally, based on a bootstrap analysis, we show that the proposed model is robust to perturbations in the covariance matrix used to form the GTV regularization term.

The results presented herein suggest some further questions and challenges with regard to the exigent task of seasonal SWUS precipitation prediction. For example, future work should address the intricate non-stationarity of the climate system more explicitly by allowing the regression coefficients to vary with time. This property might be especially important

as precipitation variability in the SWUS is expected to increase even more under climate change (Swain et al., 2018). See Appendix A for an initial exploration into accounting for nonstationarities. It should also address quantification of the underlying uncertainty of the regression coefficients (beyond the uncertainty of the covariance matrix explored herein), which can be translated into confidence intervals of the predicted precipitation. Lastly, our approach can be extended by using global information from additional climate variables (e.g. ocean heat content, atmospheric pressure etc.) and using climate model outputs from different projects, like the 6^{th} phase of the Coupled Model Intercomparison Project (Eyring et al., 2016) or the Decadal Prediction Large Ensemble project (Yeager, 2018).

In conclusion, while more complex non-linear models, such as deep neural networks, have been gaining popularity in modeling climate data, our work shows that for high-dimensional problems with limited historical records, sparse linear models with informative regularization can play an important role in building predictive models for climate variability. This is consistent with a recent review paper which focused on seasonal to subseasonal prediction of climate variables over the entire US (He et al., 2021) and which highlighted the success of regularized regression models (such as simple LASSO; see also DelSole and Banerjee (2017)). In addition, an important advantage of sparse linear models in this context is that they are considerably easier to interpret from a physical perspective, compared to non-linear models such as deep neural networks. Our results suggest that a promising direction for future research is the development of new models that can incorporate relevant physical knowledge (e.g., from large ensemble simulations of climate models), that can retain the interpretability of sparse linear models, and that have the flexibility to improve the accuracy of current models for seasonal and subseasonal precipitation prediction.

CHAPTER 3

VARIABLE IMPORTANCE

3.1 Introduction

As predictive modeling becomes ubiquitous across a wide swath of application areas, it is especially critical to understand which variables contribute most to making a particular prediction. Black-box machine learning methods are insufficient in the face of algorithmic decision-making about things like sentencing, healthcare, and education, and working toward developing more interpretable methods is becoming more and more relevant (Rudin and Radin, 2019; Guidotti et al., 2018).

Traditional statistical tools based on parametric models (e.g. p-values, ANOVA) for VI inference are dissatisfying due to restrictive assumptions often violated in modern datasets. Non-parametric extensions thus have been explored (Doksum and Samarov, 1995). In recent decades, many VI methods designed for modern deep learning models have been investigated; most of these methods are gradient-based and depend on the structure and the weights of nodes in a given specific neural network (Shrikumar et al., 2019; Sundararajan et al., 2017; Smilkov et al., 2017; Bach et al., 2015). Few statistically rigorous properties are provided for these methods, and the VI definition is always intimately attached to the network itself, making it hard to interpret in a model-agnostic setting.

In a model-agnostic setting, a natural definition of VI that is independent of the estimation procedure is to measure the loss of predictive power when the variables of interest are deleted. To estimate such model-agnostic VI, *retraining* is the most widely used type of method, which involves training separate models on the reduced data with the variables of interest deleted and assessing the predictive skill difference (Williamson et al., 2021; Lei et al., 2018; Sapp et al., 2014). Retraining often acts as the best benchmark to evaluate other VI estimation methods (Hooker et al., 2019) due to its accuracy, yet it is computationally

infeasible in high-dimensional settings.

Other methods for VI estimation include knockoff methods (Barber and Candes, 2018; Candes et al., 2017) and Floodgate (Zhang and Janson, 2021), which require the co-variate distribution to be known. An alternative approach is to use a *dropout*-type method (Chang et al., 2017). Dropout is best-suited for assessing how much a variable affects a predictive model, as opposed to our goal of assessing how much a variable affects the response. Despite the resulting issues with VI estimation accuracy, it is still widely used in practice as a proxy for VI due to its computational tractability.

In this work, we propose a computationally efficient variable importance estimation procedure for model-agnostic and distribution-free settings with theoretical guarantees that leverages a *lazy retraining* framework inspired by (Chizat et al., 2020). The key idea is to train a new model on the transformed training data, akin to retraining, but on a linearized version of the model centered around model parameters learned from the original (unreduced) training data. We perform ridge regression on this linearized model in the gradient feature space, meaning that our lazy retraining procedure can be computed very quickly. The resulting method, when applied to wide neural network models, admits error bounds that show it is nearly as accurate as full retraining, while computationally it is nearly as fast as dropout. Our theoretical bounds are complemented by a collection of simulations that explore the limitations of dropout and benefits of lazy retraining under a variety of conditions and an application to understanding the importance of various climate indices in a seasonal forecasting task.

In summary, the main contribution of this paper is a **new, computationally efficient VI estimation method with statistical performance guarantees** in a model-agnostic and distribution-free setting when using large neural networks. Our theoretical analysis facilitates statistical inference, and we illustrate our approach on both synthetic and real-world data to support the theoretical claims and demonstrate the utility of our method.

Other empirically-driven VI estimation methods exhibit similarities to our approach; our theoretical analyses may provide new insights into those methods as well as the one we propose in this paper.

3.2 Notation and preliminaries

Suppose we have samples $Z_i = (\mathbf{X}_i, Y_i), i = 1, \ldots, n$ for data $Z = (X, Y) \sim P_0$, where $\mathbf{X}_i \in \mathbb{R}^p$ is the i-th p-dimensional feature vector and Y_i is the i-th observed response. X denotes the multi-variate random variable containing features, Y denotes the response random variable. Let $X_{-j} \in \mathbb{R}^{p-1}$ (resp. $\mathbf{X}_{i,-j}$) denote the features in X (resp. $\mathbf{X}_i$) with the j-th variable removed; on the other hand, if we replace the j-th random variable in X (resp. $\mathbf{X}_i$) by its marginal mean $\mu_j = \mathbb{E}(X_j)$, we denote it as $X^{(j)}$ (resp. $\mathbf{X}_i^{(j)}$), *i.e.*, $X^{(j)} = (X_1, \ldots, X_{j-1}, \mu_j, X_{j+1}, \ldots, X_p)$.

Let $P_0, P_{0,-j}$ be the population distributions for X and X_{-j} and let $P_n, P_{n,-j}$ be the empirical distributions of X and X_{-j} for $j \in [p]$. δ_{Z_i} denotes the point mass probability measure at the i-th observation Z_i. Let f_0 denote the true function mapping X to the expected value of Y conditional on X, and let $f_{0,-j}$ denote the function mapping $X^{(j)}$ to the expected value of Y conditional on $X^{(j)}$:

$$f_0(X) := \mathbb{E}[Y|X]; \tag{3.1}$$

$$f_{0,-j}(X^{(j)}) := \mathbb{E}[Y|X_{-j}]. \tag{3.2}$$

Let f_n be the empirical model trained using all p variables in X within a certain function class $\mathcal{F}$(we refer to this as the *full model*):

$$f_n \in \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} [Y_i - f(\mathbf{X}_i)]^2. \tag{3.3}$$

To measure the accuracy of an approximation $f_n(x)$ to its target function f_0, we use the $L_2(\mu)$-norm

$$\|f_n - f\|^2 = \int |f_n(x) - f(x)|^2 d\mu(x), \tag{3.4}$$

where μ is the probability measure for X. Further, we use ϵ and $\epsilon^{(j)}$ to denote the respective remainder terms:

$$\epsilon := Y - \mathbb{E}[Y|X]; \quad \epsilon^{(j)} := Y - \mathbb{E}[Y|X_{-j}], \ j \in [p]. \tag{3.5}$$

We will define our measure of variable importance (VI) in terms of a *predictive skill measure* $V(f, P)$ (the same measure in Williamson et al. (2021)). Larger values of $V(f, P)$ should indicate better predictive performance. For $Z = (X, Y)$, we denote $\dot{V}(f, P; \delta P)$ as the Gateaux derivative of $V(f, P)$ at P in the direction δP. Specifically, one of the predictive skill measures we consider is the negative mean squared error (MSE):

$$V(f, P) = -\mathbb{E}_{(X,Y)\sim P}[Y - f(X)]^2, \tag{3.6}$$

and the corresponding $\dot{V}(f, P; \delta P)$ is $\dot{V}(f, P; \delta P) = -\int_{Z=(X,Y)} (Y - f(X))^2 d(\delta P)$. Hence, the Gateaux derivative of the negative MSE is $\dot{V}(f, P_0; \delta_{Z_i} - P_0) = -(Y_i - f(X_i))^2 + \mathbb{E}[Y - f(X)]^2$ and $\mathbb{E}[\dot{V}(f_0, P_0; \delta_{Z_i} - P_0)] = 0$.

3.3 Estimating variable importance

The VI measure we consider, which makes no assumptions on the data generating mechanism, is

$$\text{VI}_j := V(f_0, P_0) - V(f_{0,-j}, P_{0,-j}). \tag{3.7}$$

VI_j quantifies the difference in predictive skill between the full model and the reduced model for any $j \in [p]$. Consider the following simple linear model example, where we take

the negative MSE as the predictive skill measure.

Example 3.3.1. *Suppose* $Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon$, *where* $X_i \sim \mathcal{N}(0, \sigma^2), i = 1, 2,$ $Cov(X_1, X_2) = \rho$, *and* ϵ *is a* $\mathcal{N}(0, \sigma_\epsilon^2)$ *noise that is independent of the features. The variable importance of the first variable is*

$$\text{VI}_1 = \beta_1^2 \cdot Var(X_1|X_2) = \beta_1^2(1-\rho^2)\sigma^2$$

due to the fact that $X_1|X_2 \sim \mathcal{N}(\rho X_2, (1-\rho^2)\sigma^2)$

In general, we see from this example that the variable importance measure is determined not only by the relationship between X_j and Y, but also the covariance structure in the features.

Our goal is to estimate VI_j for any variable X_j from data $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ with no assumptions on the relationship between X and Y. For empirical estimators f_n and $f_{n,-j}$ of f_0 and $f_{0,-j}$, a plug-in estimator of our VI measure is

$$\widehat{\text{VI}}_j = V(f_n, P_n) - V(f_{n,-j}, P_{n,-j}). \tag{3.8}$$

The key problem we are concerned with in this paper is how to estimate $f_{n,-j}$ in an accurate and computationally efficient way. Traditionally, people use the following two types of methods to do the estimation: *dropout* and *retraining*.

3.3.1 Dropout

The method we are calling *dropout* estimates $\mathbb{E}(Y|X_{-j})$ by plugging the dropout features $X^{(j)}$ into the full model f_n. In this case, the variable importance measure can be estimated by

$$\widehat{\text{VI}}_j^{(\text{DR})} = V(f_n, P_n) - V(f_n, P_{n,-j}). \tag{3.9}$$

For the negative MSE measure of predictive skill for instance, the dropout estimate measures the difference between the squared error on the *original* training set and the squared error on the training set *after replacing feature j with its mean*. Dropout is superior among all plug-in estimators in terms of computational cost – we only need to train the model once to get f_n. This is desirable, especially when the function class $\mathcal{F}$ is large and complicated, such as with neural networks, and the computational cost for training the model is high. Despite this benefit, dropout is unreliable in many settings, as we will revisit in 3.3.3.

3.3.2 Retrain

An alternative to dropout is what we call *retraining*. Given a function class $\mathcal{F}$, the retraining method estimates VI_j by training separate models

$$f_{n,-j} \in \underset{f \in \mathcal{F}}{\arg\min} [Y_i - f(\mathbf{X}_i^{(j)})]^2 \tag{3.10}$$

for each variable $j \in [p]$ to estimate $f_{0,-j}$. Hence, VI under this framework is estimated via

$$\widehat{\text{VI}}_j^{(\text{RT})} = V(f_n, P_n) - V(f_{n,-j}, P_{n,-j}). \tag{3.11}$$

When taking negative MSE as the predictive skill measure, the retraining estimate in this case measures the difference between the squared error of a model trained *without* feature j and the squared error of a model trained *with* feature j. Retraining is more accurate than dropout as long as the function class $\mathcal{F}$ is large enough, but requires training $p + 1$ models, which can be prohibitively computationally expensive in many settings. In this paper, we are especially interested in the setting when the function class is as large as a wide neural network.

3.3.3 Dropout vs. Retrain for Linear Models

The dropout method is widely used to estimate variable importance due to its efficiency. However, in cases where variables in X are highly correlated, dropout behaves problematically. Below, we will illustrate and quantify the difference of the variable importance estimation in the random design linear model case, where we take the negative MSE as the $V(f, P)$ measure as in (3.6). For simplicity, we restrict the function space $\mathcal{F}$ to the linear function space here.

Suppose $X \in \mathbb{R}^p \sim \mathcal{N}(0, \Sigma)$, $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$. Assume Σ is positive definite. Let $\beta^* := \arg\min_{w \in \mathbb{R}^p} \mathbb{E}[Y - X^\top w]^2$, so $\beta^* = \Sigma^{-1}\mathbb{E}(XY)$. In the population version, the dropout method uses the predictor $X_{-j}^\top \beta_{-j}^*$ (where $\beta_{-j}^* \in \mathbb{R}^{p-1}$ is β^* with its j-th element removed) to estimate $\mathbb{E}(Y|X_{-j})$, while the retraining method uses the predictor $X_{-j}^\top \beta^{(j)}$, where $\beta^{(j)} \in \mathbb{R}^{p-1}$ is $\beta^{(j)} = \arg\min_{w \in \mathbb{R}^{p-1}} \mathbb{E}[Y - X_{-j}^\top w]^2$. The following proposition characterizes the difference between VI estimates corresponding to the retraining and dropout methods.

Proposition 3.3.2. *In the linear function space, the difference between the variable importance estimates for variable j from the population version of the dropout and retraining methods is:*

$$\widehat{\mathrm{VI}}_j^{(\mathrm{DR})} - \widehat{\mathrm{VI}}_j^{(\mathrm{RT})} = \frac{\vec{\boldsymbol{\gamma}}_j{}^\top \Sigma_{(j)}^{-1} \vec{\boldsymbol{\gamma}}_j}{(\Sigma_{jj} - \vec{\boldsymbol{\gamma}}_j{}^\top \Sigma_{(j)}^{-1} \vec{\boldsymbol{\gamma}}_j)^2} \left[\mathbb{E}(X_j Y) - \vec{\boldsymbol{\gamma}}_j{}^\top \Sigma_{(j)}^{-1} \mathbb{E}(X_{-j} Y)\right]^2$$

where $\vec{\boldsymbol{\gamma}}_j = \mathbb{E}(X_j X_{-j}) \in \mathbb{R}^{p-1}$.

If the true model between Y and X is linear, *i.e.,* $Y = X^\top \beta^* + \epsilon$, and $X \perp\!\!\!\perp \epsilon$, the variable importance estimated by retraining linear regression is:

$$\widehat{\mathrm{VI}}_j^{(\mathrm{RT})} = \beta_j^{*2}(\Sigma_{jj} - \vec{\boldsymbol{\gamma}}_j{}^\top \Sigma_{(j)}^{-1} \vec{\boldsymbol{\gamma}}_j); \tag{3.12}$$

furthermore, in this setting $\widehat{\mathrm{VI}}_j^{(\mathrm{RT})}$ is exactly the true variable importance defined in (3.7).

In contrast, the dropout framework will give

$$\widehat{\mathrm{VI}}_j^{(\mathrm{DR})} = \beta_j^{*2} \cdot \Sigma_{jj}. \tag{3.13}$$

If feature j is important and highly correlated with feature k (but independent of all other features), then $\vec{\boldsymbol{\gamma}_j}^\top \Sigma_{(j)}^{-1} \vec{\boldsymbol{\gamma}_j}$ may be very large, making the difference between $\widehat{\mathrm{VI}}_j^{(\mathrm{DR})}$ and $\widehat{\mathrm{VI}}_j^{(\mathrm{RT})}$ similarly large. This example illustrates how dropout can significantly overestimate variable importance, even in simple settings.

3.4 Lazy Training

Our central interest is in inferring VI using complex models that are time-consuming to train, making the baseline retraining method described above computationally infeasible. With this in mind, we turn our attention to neural network (NN) models, a setting in which dropout is widely used.

Motivated by the need for faster and more accurate methods for estimating VI with NN, we propose a computationally efficient VI estimate inspired by the lazy training framework of Chizat et al. (2020) that estimates the difference between the full model parameters and the model parameters when the j-th variable is removed. Like dropout, our procedure only requires us to train the NN once on the full data, and then we solve a linear system to update the full model parameters for each variable $j \in [p]$.

Given the training data $\{(\mathbf{X}_i^{(j)}, Y_i)\}$ sampled from $(X, Y) \sim P_0$ for $i = 1, \ldots, n$ and the underlying function $f_0(X) = \mathbb{E}_{P_0}[Y|X]$, there exists a a neural network function class $\{h_\theta(x) : \mathbb{R}^p \mapsto \mathbb{R} | \theta \in \mathbb{R}^M\}$ that is parameterized by a vector θ, such that when we train the model parameters over this class by

$$\theta_f = \underset{\theta \in \mathbb{R}^M}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} [Y_i - h_\theta(\mathbf{X}_i)]^2, \tag{3.14}$$

the estimation error can be bounded by $\|h_{\theta_f}(x) - f_0(x)\| = O(n^{-1/2})$ up to some log terms (Barron, 1994). To achieve this, the scale of the number of parameters M depends on the complexity of the target function. For very complex functions, we can still achieve this accuracy with $M = O(\sqrt{n})$.

In order to estimate VI_j, we need an estimate of what we are calling the *reduced* model $h_{\theta_{-j}}$, where

$$\theta_{-j} = \underset{\theta \in \mathbb{R}^M}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} [Y_i - h_\theta(\mathbf{X}_i^{(j)})]^2. \tag{3.15}$$

Instead of retraining a NN to estimate θ_{-j}, we can instead estimate the difference between the full model parameters θ_f and θ_{-j} using this linear approximation, and simply update the full model parameters with this correction to estimate $h_{\theta_{-j}}$. We are essentially regressing the error resulting from the dropout estimation against the gradient to estimate this correction, and to do so we solve the following convex problem based on the training data $\{(\mathbf{X}_i^{(j)}, Y_i)\}$ for $i = 1, \ldots, n$ and a 2-norm penalty on the parameters:

$$\begin{aligned} \Delta\theta_j(\lambda, n) = \underset{\omega \in \mathbb{R}^M}{\arg\min} \Big\{ \frac{1}{n} \sum_{i=1}^{n} \big[Y_i - h_{\theta_f}(\mathbf{X}_i^{(j)}) \\ - \omega^\top \nabla_\theta h_\theta(\mathbf{X}_i^{(j)})|_{\theta=\theta_f} \big]^2 + \lambda \|\omega\|_2^2 \Big\}, \end{aligned} \tag{3.16}$$

where $\lambda > 0$ is the penalty parameter.

Accordingly, the reduced neural network parameters are $\Delta\theta_j(\lambda, n) + \theta_f$. For the simplicity of notation, we write $\Delta\theta_j(\lambda, n)$ as $\Delta\theta_j$ for short. Then the reduced model approximation without the j-th feature is $\mathbb{R}^p \mapsto \mathbb{R} : x \mapsto h_{\theta_f + \Delta\theta_j}(x)$. Hence, the variable importance measure under lazy training is

$$\widehat{\text{VI}}_j^{(\text{LAZY})} = V(h_{\theta_f}, P_n) - V(h_{\theta_f + \Delta\theta_j}, P_{n,-j}). \tag{3.17}$$

Under the negative MSE measure $V(f, P)$, we have

$$\widehat{\mathrm{VI}}_j^{(\mathrm{LAZY})} = \frac{1}{n}\sum_{i=1}^{n}\{[Y_i - h_{\theta_f+\Delta\theta_j}(\mathbf{X}_i^{(j)})]^2 - [Y_i - h_{\theta_f}(\mathbf{X}_i)]^2\}.$$

(More precisely, we use data splitting for training and estimating VI as detailed in 1.) Essentially, the linearized approximation of the NN is linear in the gradient feature map $x \mapsto \nabla_\theta h_\theta(x)|_{\theta_f}$. In fact, this gradient feature map induces the Neural Tangent Kernel (NTK, Jacot et al. (2020)): for any $x, x' \in \mathbb{R}^p$,

$$\ker_{\theta_f}(x, x') := \langle \nabla_\theta h_\theta(x)|_{\theta_f}, \nabla_\theta h_\theta(x')|_{\theta_f} \rangle. \tag{3.18}$$

Thus $\Delta\theta_j$ can be viewed as the solution for a kernel ridge regression problem with kernel $\ker_{\theta_f}$.

3.4.1 Theoretical Guarantee

By Williamson et al. (2021), when the empirical estimates for $\mathbb{E}(Y|\mathbf{X})$ and $\mathbb{E}(Y|\mathbf{X}_{-j})$ converge to the target functions f_0 and $f_{0,-j}$ at the rate of $O_P(n^{-1/4})$ in function norm, we achieve an asymptotically normal and efficient estimator for the VI measure. In this section, we give a theoretical guarantee to show that the lazy prediction $h_{\theta_f+\Delta\theta_j}(X^{(j)})$ for the reduced model achieves such convergence rate, so that the lazy training procedure gives an accurate estimate of VI with an error in the order of $O(\frac{1}{\sqrt{n}})$ and we can make inference accordingly.

Let $\mathbf{e}^{(j)}$ denote the difference between the true reduced function $f_{0,-j}(X^{(j)})$ and the corresponding dropout estimation:

$$\mathbf{e}^{(j)} := f_{0,-j}(\mathbf{X}^{(j)}) - h_{\theta_f}(\mathbf{X}^{(j)}) \in \mathbb{R}^n. \tag{3.19}$$

Further, we denote the kernel matrix on $X^{(j)}$ induced by the gradient feature map as $\mathbb{K}^{(j)} \in \mathbb{R}^{n\times n}$, whose elements are defined as:

$$\mathbb{K}^{(j)}_{ik} := \ker_{\theta_f}(\mathbf{X}^{(j)}_i, \mathbf{X}^{(j)}_k),\ i,k \in [n]. \tag{3.20}$$

Assumption 3.4.1. For any $j \in [p]$, there exists a constant $C < \infty$, such that ${\mathbf{e}^{(j)}}^\top [\mathbb{K}^{(j)}]^{-1}\mathbf{e}^{(j)} < C$ with high probability, *i.e.,* ${\mathbf{e}^{(j)}}^\top [\mathbb{K}^{(j)}]^{-1}\mathbf{e}^{(j)} = O_P(1)$. Furthermore, $\mathrm{tr}(\mathbb{K}^{(j)}) = O_P(n)$.

The requirement $\mathrm{tr}(\mathbb{K}^{(j)}) = O_P(n)$ is an assumption commonly used in NTK literature (see e.g. Hu et al. (2019)). Arora et al. (2019, Corollary 6.2) proved that for the NTK induced by an overparametrized two-layer neural network, $y^\top \mathbb{K}(\mathbf{X},\mathbf{X})^{-1}y = O(1)$ as long as $y_i = g(x_i)$ for a certain function g. Based on the definition of $\mathbf{e}^{(j)}$ in (3.19), $\mathbf{e}^{(j)}_i$ is a function of $\mathbf{X}^{(j)}_i$, thus this assumption ${\mathbf{e}^{(j)}}^\top [\mathbb{K}^{(j)}]^{-1}\mathbf{e}^{(j)} = O(1)$ can be satisfied for over-parametrized two-layer neural networks.

Assumption 3.4.2. For the noise term $\epsilon^{(j)}$, we have the following assumption on its conditional tail probability: there exists σ such that for any $j \in [p]$,

$$\mathbb{E}\left[e^{\lambda\epsilon^{(j)}}|X^{(j)}\right] \le e^{\sigma^2\lambda^2/2},\ \text{ for all } \lambda \in \mathbb{R}. \tag{3.21}$$

Assumption 3.4.3. Denote the gradient feature matrix as $\Phi \in \mathbb{R}^{n\times M} = (\nabla_\theta h_\theta(\mathbf{X}_1)|_{\theta=\theta_f}, \ldots, \nabla_\theta h_\theta(\mathbf{X}_n)|_{\theta=\theta_f})^\top$. We assume $\|\Phi^\top \mathbf{e}^{(j)}\|_2 \le O_P(1)$.

This assumption essentially requires that the linear space of neural tangent kernels can well represent $\mathbf{e}^{(j)}$. We know that $\mathbf{e}^{(j)}_i$ is a function of $\mathbf{X}^{(i)}_i$, thus as long as the neural network function class is large enough, this can be satisfied with respect to the sample size n.

Theorem 3.4.4. *Suppose Assumptions 3.4.1, 3.4.2 and 3.4.3 hold. Then, for a neural network structure $h_\theta(\cdot)$ which is L-smooth with respect to its parameters θ, as long as we take the ridge penalty parameter in the order $\lambda = O(n^{1/2})$, then the lazy training method can accurately predict the reduced model without the j-th covariate,* i.e.,

$$\|h_{\theta_f + \Delta\theta_j}(x) - \mathbb{E}(Y|X^{(j)})\|_2 = O_p(n^{-1/4}). \tag{3.22}$$

Therefore our variable importance estimator $\widehat{\text{VI}}_j^{(\text{LAZY})}$ is asymptotically normal and has an error rate $O_p(n^{-1/2})$:

$$\widehat{\text{VI}}_j^{(\text{LAZY})} - \text{VI}_j = \Delta_{n,j} + O_p(n^{-1/2}); \tag{3.23}$$

where

$$\begin{aligned}\Delta_{n,j} = \frac{1}{n}\sum_{i=1}^{n} \big[&\dot{V}(f_0, P_0; \delta_{Z_i} - P_0) \\ &- \dot{V}(f_{0,-j}, P_{0,-j}; \delta_{Z_i} - P_{0,-j})\big] \to_d \mathcal{N}(0, \tau_{n,j}^2);\end{aligned} \tag{3.24}$$

here the variance is $\tau_{n,j}^2 = Var({\epsilon^{(j)}}^2 - \epsilon^2)/n$, where ϵ and $\epsilon^{(j)}$ is defined in 3.5.

This result enables us to construct Wald-type confidence intervals around our LazyVI estimates. In particular, the $\alpha-$level confidence intervals are given by

$$\widehat{\text{VI}}_j^{(\text{LAZY})} \pm z_{\frac{\alpha}{2}} \hat{\tau}_{n,j} \tag{3.25}$$

where $\hat{\tau}_{n,j}$ is the plug-in estimate of $\tau_{n,j}$ in (3.24) and $z_{\frac{\alpha}{2}}$ is the $\alpha/2$ quantile of the standard normal distribution.

3.4.2 Proof Overview

The challenge of proving Theorem 3.4.4 is to bound the error of the lazy neural network trained using data without a certain variable – note that we are bounding the estimation error ($\|h_{\theta_f+\Delta\theta_j} - f_{0,-j}\|$) instead of the prediction error ($\|h_{\theta_f+\Delta\theta_j} - f_0\|$) that is the focus of much of the deep learning community, since the predictive skill of the reduced model is expected to decrease when an important variable is removed. At a high level, our proof reduces the estimation error of the neural network from lazy training to the error between the NTK estimation and the target function, where we use techniques from kernel ridge regression. The difference here is that most NTK papers (see e.g. Jacot et al. (2020)) use random initialization for the parameters and optimization without penalty, while our method starts from a specific initialization (the full model), and requires the penalty parameter λ to be large ($\lambda = O(n^{1/2})$) to ensure convergence.

The following two lemmas give some intuition on how the neural network trained by the lazy procedure can accurately estimate the reduced model. Basically, the bound for the error consists of two parts: the error from the kernel ridge regression (discussed in Lemma 3.4.5), and the error from the linear approximation of the neural network (in Lemma 3.4.6). More proof details are deferred to Appendix B.

Denote the linear approximation of the network as

$$\tilde{h}_{\theta_f+\Delta\theta_j}(x) := h_{\theta_f}(x) + \langle \nabla_\theta h_\theta(x)|_{\theta=\theta_f}, \Delta\theta_j \rangle. \tag{3.26}$$

Lemma 3.4.5. *Letting λ be the penalty parameter in Equation 3.16, we have with probability*

at least $1-\delta$,

$$\begin{aligned}&\|\tilde{h}_{\theta_f+\Delta\theta_j}(X^{(j)}) - f_{0,-j}(X^{(j)})\|_n \\ &\leq \sqrt{\frac{\lambda \mathbf{e}^{(j)\top}[\mathbb{K}^{(j)}]^{-1}\mathbf{e}^{(j)}}{4n}} + \sqrt{\frac{\sigma\, tr[\mathbb{K}^{(j)}]}{4\lambda n}} + \sigma\sqrt{\frac{2\log(1/\delta)}{n}}.\end{aligned} \tag{3.27}$$

Lemma 3.4.5 combined with Assumption 3.4.1 when the penalty parameter is $\lambda = O(n^{1/2})$, yields a bound on the empirical error of the kernel ridge regression component of $O_P(n^{-1/4})$. Based on this empirical bound, we could then further bound the generalization error of the estimated function using function complexity (See Appendix B.2.3).

Lemma 3.4.6. *For a large neural network with width in the order* $O(\sqrt{n})$*, with high probability we have for all* $j \in [p]$*,*

$$\|\tilde{h}_{\theta_f+\Delta\theta_j}(x) - h_{\theta_f+\Delta\theta_j}(x)\| \leq O(n^{-1/4}). \tag{3.28}$$

Lemma 3.4.6 shows that as long as the neural network is sufficiently large, the neural network with updated parameters $\theta_f + \Delta_j$ is close to its linear approximation.

3.5 Implementation

We estimate h_{θ_f} and $h_{\theta_{-j}}$ using $n_1 < n$ samples as training data, and use the remaining $n_2 = n - n_1$ samples to estimate VI. For the dropout method, VI is estimated simply by plugging the modified testing data $\{\mathbf{X}_i^{(j)}\}_{i=n_1+1}^{n}$ into h_{θ_f}. For the retraining method, first $h_{\theta_{-j}}$ is estimated by retraining the NN h with $\{\mathbf{X}_i^{(j)}\}_{i=1}^{n_1}$, and then VI is estimated by plugging the modified testing data into this retrained estimate.

For the lazy training method, which we call LazyVI, we use the training data to estimate the full model parameters, compute the gradient of the network with respect to each model parameter for each training sample, and then regress these gradients against the difference

Algorithm 1 Lazy training for VI

Require: Data: $\{\mathbf{X}_i, Y_i\}_{i=1}^n$; $\lambda > 0$; training size: $0 < n_1 < n$; $n_2 \leftarrow n - n_1$; NN structure: $\theta \in \mathbb{R}^M \mapsto h_\theta(\cdot)$ LazyVI$\{\mathbf{X}_i, Y_i\}_{i=1}^n$; λ, n_1

$\theta_f \leftarrow \arg\min_{\theta \in \mathbb{R}^M} \frac{1}{n_1} \sum_{i=1}^{n_1} [Y_i - h_\theta(\mathbf{X}_i)]^2$

$v_n \leftarrow -\frac{1}{n_2} \sum_{i=n_1+1}^{n} [Y_i - h_{\theta_f}(\mathbf{X}_i)]^2$

for $j \in [p]$ **do**

$\mathbf{X}_i^{(j)} \leftarrow \mathbf{X}_i$; $\mathbf{X}_{ij}^{(j)} \leftarrow \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{X}_{ij}$

$\mathbf{e}_i^{(j)} \leftarrow Y_i - h_{\theta_f}(\mathbf{X}_i^{(j)}),\ i = 1, \ldots, n_1$

$\Phi_i^{(j)} \leftarrow \nabla_\theta h_\theta(\mathbf{X}_i^{(j)})|_{\theta=\theta_f},\ i = 1, \ldots, n_1$

$$\Delta\theta_j \leftarrow \arg\min_{\omega \in \mathbb{R}^M} \frac{1}{n_1} \sum_{i=1}^{n_1} [\mathbf{e}_i^{(j)} - \omega^\top \Phi_i^{(j)}]^2 + \lambda \|\omega\|_2^2$$

$$v_{n,-j} \leftarrow -\frac{1}{n_2} \sum_{i=n_1+1}^{n} [Y_i - h_{\theta_f + \Delta\theta_j}(\mathbf{X}_i^{(j)})]^2$$

$\widehat{\mathrm{VI}}_j \leftarrow v_n - v_{n,-j}$

$t_{i,j} \leftarrow (Y_i - h_{\theta_f + \Delta\theta_j}(\mathbf{X}_i^{(j)}))^2 - (Y_i - h_{\theta_f}(\mathbf{X}_i))^2$

$\hat{\tau}_j \leftarrow \frac{1}{n_2} \sum_{i=1}^{n_2} (t_{i,j} - \bar{t}_j)^2 / n_2$

end for

Ensure: $\widehat{\mathrm{VI}}_j,\ j = 1, \ldots, p.$

between Y the dropout estimates from the training data to estimate the parameter correction $\Delta\theta_j$ for variable j. We then update the full model parameters using this learned correction to compute the VI estimate and its associated standard errors. See Algorithm 1 for full details.

Theorem 3.4.4 makes the assumption that the ridge parameter λ from Equation (3.16) is large. Since we are ultimately interested in estimating $h_{\theta_{-j}}$ and not $\Delta\theta_j$, we evaluate $h_{\theta_f + \Delta\theta_j}(\cdot)$ through K-fold CV to choose $\hat{\lambda}_j$ for each variable (Algorithm 2).

3.6 Experiments

We assess the performance of LazyVI on real and simulated data to highlight key theoretical claims and assumptions and show that our method is empirically practical. For all experi-

Algorithm 2 K-Fold CV for λ_j

Require: $\{\mathbf{X}_i^{(j)}, Y_i, \mathbf{e}_i^{(j)}, \Phi_i^{(j)}\}_{i=1}^{n_1}$ and θ_f from Algorithm 1 in main paper; candidate λ values Λ

Partition $[n_1]$ into K subsets, each denoted S_t

for $\lambda \in \Lambda$ **do**

for $k = 1, \ldots, K$ **do**

$\Delta\theta_j^\lambda = \arg\min_{\omega \in \mathbb{R}^M} \frac{1}{n_1 - |S_k|} \sum_{i \notin S_k} [\mathbf{e}_i^{(j)} - \langle \omega, \Phi_i^{(j)} \rangle]^2 + \lambda \|\omega\|_2^2$

$\hat{Y}_i = h_{\theta_f + \Delta\theta_j^\lambda}(\mathbf{X}_i^{(j)})$ for $i \in S_k$

$\epsilon_{\lambda,k} = \frac{1}{|S_k|} \sum_{i \in S_k} (Y_i - \hat{Y}_i)^2$

end for

$\epsilon_\lambda = \frac{1}{K} \sum_{k=1}^K \epsilon_{\lambda,k}$

end for

Ensure: $\hat{\lambda}_j = \arg\min_\lambda \{\epsilon_\lambda\}_{\lambda \in \Lambda}$

ments, we train a wide, fully connected two-layer neural network with ReLU activation for all simulations. Unless otherwise specified, the width of the hidden layer in the training network is $m = 50$.

3.6.1 Linear data generation

Our first set of simulations serve to support key details of our theoretical analysis. We consider data generated from the linear model $f(X) = 1.5X_1 + 1.2X_2 + X_3 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.1)$ and $X \sim \mathcal{N}(0, \Sigma_{6\times 6})$, so the response only depends on the first three of the six variables. All variables are independent except for X_1 and X_2, whose correlation is ρ. As discussed in Example 3.3.1, the true VI of X_1, X_2, and X_3 are given by $(1.5)^2(1-\rho^2)$, $(1.2)^2(1-\rho^2)$, and 1, respectively, and the VI of the remaining 3 variables is zero. In this simple setting, we find that LazyVI approximates the true VI well with desirable coverage and a considerable speed-up relative to retraining (Figure 3.1).

We show in Prop. 3.3.2 that, when data are generated from a linear model, the difference between the dropout and retraining variable importance estimates is a function of the covariance of X. After training the full model, we use both the dropout and our lazy procedure

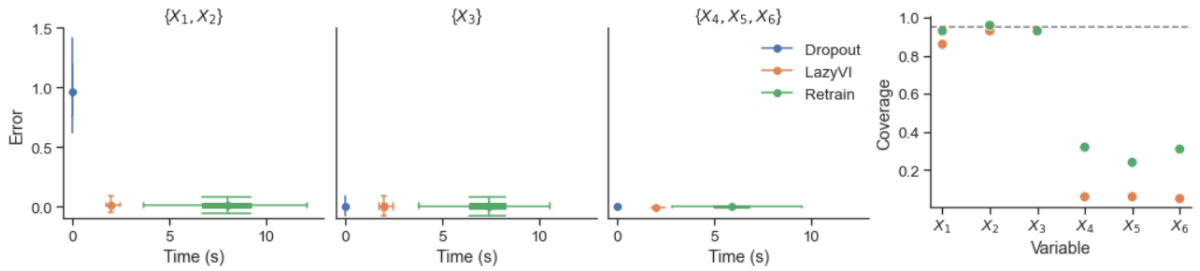


Figure 3.1: Distribution of computation time vs. estimation error relative to retrain ($\widehat{\text{VI}} - \widehat{\text{VI}}^{(\text{RT})}$) for three different groups of variables: important, correlated ($\{X_1, X_2\}$); important, uncorrelated (X_3); and unimportant, uncorrelated ($\{X_4, X_5, X_6\}$). 2D box plots show quantiles across 10 repetitions.

to estimate VI for increasing values of ρ. In Figure 3.4, we show the difference between the dropout and LazyVI estimates for variables X_1 and X_2 alongside the analytic difference between $\widehat{\text{VI}}^{(\text{DR})}$ and VI (dotted line). We see that the gap between LazyVI and dropout evolves with ρ according to the theoretical analysis, providing evidence that LazyVI behaves as expected.

We use this simple linear setting to explore two additional assumptions from our theoretical resuts. First, the linearization in (3.26) is a first order Taylor approximation and assumes the full model parameters are close to the reduced model parameters. If we try to linearize a neural network around a random initialization, our LazyVI estimates are much less accurate and more highly variable (Figure 3.2). Next, our theory assumes that our training network is over-parameterized and sufficiently wide. We compute empirical confidence intervals for LazyVI for increasing network widths and find that coverage increases as the width increases, but at a computational cost (Figure 3.3).

3.6.2 Binary classification

While the simple linear setting is useful in highlighting some key theoretical aspects of our work, a key benefit of LazyVI is its computational efficiency, and we are interested in how well LazyVI can approximate the potentially prohibitively costly retraining method in a variety of simulation settings. Because we borrow much of our theoretical framework from

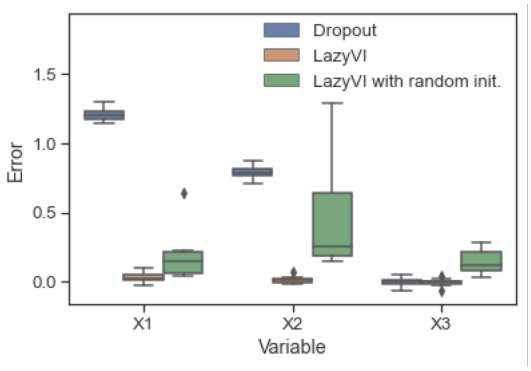


Figure 3.2: Distribution of $\text{VI} - \hat{\text{VI}}$ for the first 3 variables for dropout, LazyVI initialized with the parameters from the full model, and LazyVI with a random initialization.

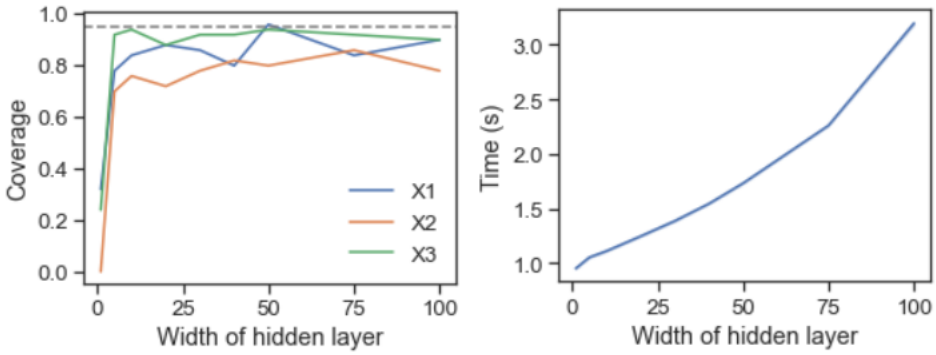


Figure 3.3: Left: empirical coverage of 95% confidence intervals (across 50 repetitions) of LazyVI estimates for increasing widths of the training network for the important variables. Dotted line shows 95% coverage; Right: average computation time for LazyVI with increasing network widths.

Williamson et al. (2021), we also leverage their simulation framework as a useful point of comparison. We draw independent samples $X \sim \mathcal{N}(0, I_{4\times 4})$ and generate a binary outcome $Y \sim \text{Bernoulli}(\Phi(X\beta))$ where $\beta = (2.5, 3.5, 0, 0)$. Because the outcome is binary, we use accuracy as our predictive skill measure, and the true VI values are given by (0.136, 0.236, 0, 0), respectively. We first directly compare the LazyVI and retrain estimators by estimating VI across 100 simulated datasets of sample size $n = 1000$ and computing the empirical 95% confidence intervals. In Figure 3.5, we see that the LazyVI and retrain estimates both achieve the desired level of coverage with low bias. In this simulation, LazyVI took on average 0.6 seconds (including cross-validating to find the optimal ridge parameter), while retraining

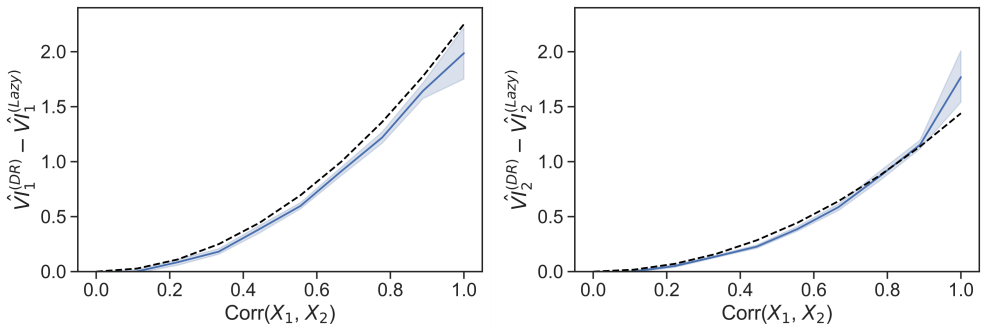


Figure 3.4: Difference between the dropout and LazyVI estimates for X_1 and X_2. Dotted line is theoretical gap and shading shows std. across 10 repetitions.

took 7.5 seconds. In this setting, LazyVI is just as accurate as retraining with a more than 10x speed-up.

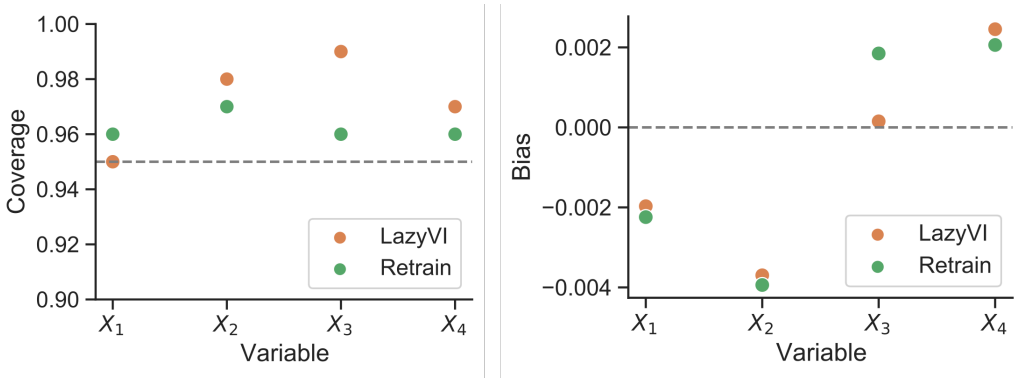


Figure 3.5: Left: Average coverage of empirical 95% confidence intervals from the LazyVI and retrain estimates across 100 simulations. Right: Average empirical bias ($\text{VI} - \hat{\text{VI}}$) of LazyVI and retrain estimates.

3.6.3 Nonlinear, high-dimensional regression

The computational burden of retraining is most pronounced in high-dimensional settings, since estimating $\widehat{\text{VI}}^{(\text{RT})}$ for all variables requires refitting at least p models. For this simulation, we have data $X \sim N(0, \Sigma_{100\times100})$, where variables are independent except $\text{Corr}(X_1, X_2) = 0.5$. Letting $\beta = (5, 4, 3, 2, 1, 0, \ldots, 0)^T \in \mathbb{R}^{100}$, we construct a weight matrix $W \in \mathbb{R}^{m\times p}$ such that the $W_{:,j} \sim \mathcal{N}(\beta_j, \sigma^2)$ (i.e. the weights associated with variable j are centered at β_j). Letting $V \sim \mathcal{N}(0, 1)$, we generate the response $Y_i = V\sigma(W\mathbf{X}_i) + \epsilon_i$ where σ is the ReLU

function. Because the "true" VI values are unknown and difficult to estimate, we present the accuracy of different estimation methods relative to the retraining estimates, which we take as ground truth. We estimate VI for X_1 across 10 simulated datasets ($n = 1000$) and benchmark against retraining using both a linear regression (OLS) and random forest (RF). In Figure 3.6, we show the spread of both the computation time and normalized error (relative to retrain) for all methods. We see that LazyVI is the most accurate method and is substantially faster than retraining, which is especially beneficial in this high-dimensional setting.

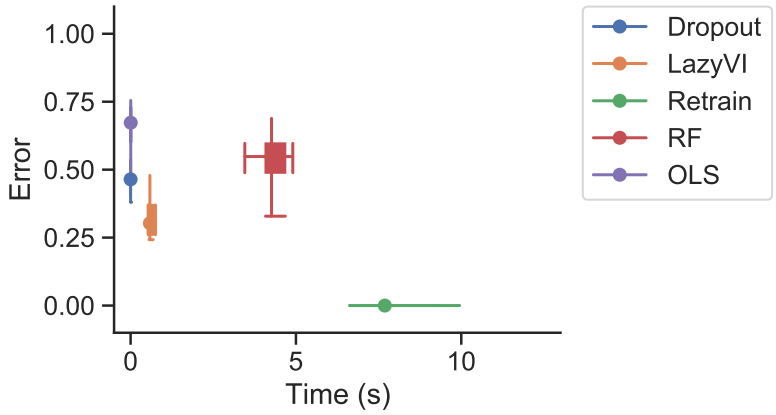


Figure 3.6: Distribution of computation time vs. normalized estimation error relative to retrain for the VI of X_1 $((\hat{\text{VI}} - \widehat{\text{VI}}^{(\text{RT})})/\widehat{\text{VI}}^{(\text{RT})})$ across 10 repetitions.

3.7 Predicting seasonal precipitation

Extreme precipitation events have become more and more common in recent years, and are expected to intensify with climate change (Tabari, 2020; Li et al., 2019). Early and reliable precipitation forecasting is thus critical for regional water resource management, which increasingly impacts large swaths of the population (AghaKouchak et al., 2015). Many studies have shown that the sea surface temperature (SST) over various regions of the ocean, such as the El Niño-Southern Oscillation (ENSO), are predictive of precipitation in the United States (Mamalakis et al., 2018; Dai, 2013; Lenssen et al., 2020). Understanding

which ocean regions are most predictive is challenging, however, due to a short observational record and strong correlations among SSTs (Stevens et al., 2021).

We estimate the importance of different ocean regions for seasonal precipitation forecasting using our lazy training method. The response is the average winter precipitation over the Southwestern US, and as predictors we use 10 ocean climate indices (OCIs), which are defined as the average detrended SST anomalies over different ocean regions (Chen et al., 2016). As data, we use simulations from the Community Earth System Model-Large Ensemble project (CESM-LENS; Kay et al. (2015); de La Beaujardière et al. (2019)). CESM-LENS is a 40-member ensemble of climate simulations, where the ensemble members all have the same physics but different initial conditions. From this dataset, we extracted monthly sea surface temperature (SST) records from 1940-2005 on a $1.25^\circ \times 0.9^\circ$ grid. We compute SST anomalies at each grid point relative to the time period 1950-1989[1] by subtracting the monthly mean and dividing by the monthly standard deviation, and then we linearly detrend each time series.

To compute the 10 ocean climate indices (OCI) used in our experiment, we find the average summer (July-October) monthly SST values of these detrended SST anomalies over specified ocean regions. These regions are well established in the literature; we refer to the supplement from Chen et al. (2016) to define the boundaries of all OCIs besides NZI, for which we use Mamalakis et al. (2018). See 3.1 for the specific boundaries. As a response, we use the average winter (November-March) precipitation over part of the southwestern US (see Stevens et al. (2021)). We are interested in predicting winter precipitation from the previous summer's SSTs.

1. `https://climatedataguide.ucar.edu/climate-data/nino-sst-indices-nino-12-3-34-4-oni-and-tni`

Ocean	OCI	Latitude	Longitude
Pacific	Niño1+2	10°S - 0°	90°W - 80°W
	Niño3	5°S - 5°N	150°W - 90°W
	Niño3.4	5°S - 5°N	170°W - 120°W
	Niño4	5°S - 5°N	160°E - 150°W
	NZI	40°S - 25°S	170°E - 160°W
Atlantic	TNA	5°N - 25°N	55°W - 15°W
	TSA	20°S- 0°	30°W - 10°E
Indian	SWIO	32°S - 25°S	31°E - 45°E
	WTIO	10°S- 10°N	50°E - 70°E
	SETIO	10°S- 0°	90°E - 110°E

Table 3.1: Ocean climate indices (OCIs) are defined as the average of the detrended SST anomalies across the regions indicated above.

To build more intuition as to why dropout behaves problematically when trying to predict precipitation, we attempt this analysis using linear regression and find that the coefficient estimates are highly unstable. Figure 3.7 shows the estimated coefficients with their 95% confidence intervals from a multiple linear regression including all variables, along with estimated coefficients from separate simple linear regressions. Note that the coefficient for Niño 3 is highly negative in the full regression, which is offsetting the highly positive coefficient on Niño 3.4; when separated, both of these indices receive much smaller positive coefficients.

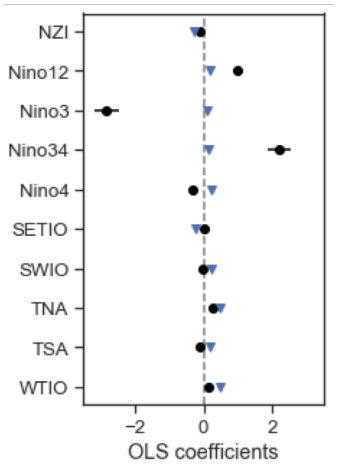


Figure 3.7: OLS coefficients and standard errors for multiple linear regression with all OCIs (black dots) and simple linear regression with each OCI separately (blue triangles).

There are strong correlations among the various OCIs (3.8)) — in particular, the various Niño indices appear to be nearly collinear. Because of this, we would expect methods like linear regression to inaccurately estimate coefficients and their importance (see appendix for more discussion).

We apply LazyVI to this problem and compare with the dropout and retraining VI estimates. We see that dropout drastically overestimates VI of Niño 3 and Niño 3.4 relative to retraining, and that LazyVI results in estimates much closer to the retraining estimates. These results are consistent with recent literature indicating that the predictive ability of Niño is often overstated relative to other OCIs (Mamalakis et al., 2018), suggesting that LazyVI could potentially help us better understand the relative importance of different climate mechanisms.

3.8 Discussion

Assessing variable importance in machine learning is a vital task as learned-based tools are increasingly integrated into societally-impactful systems, including autonomous vehicles, financial and healthcare decision-making, and social and criminal justice. In this work,

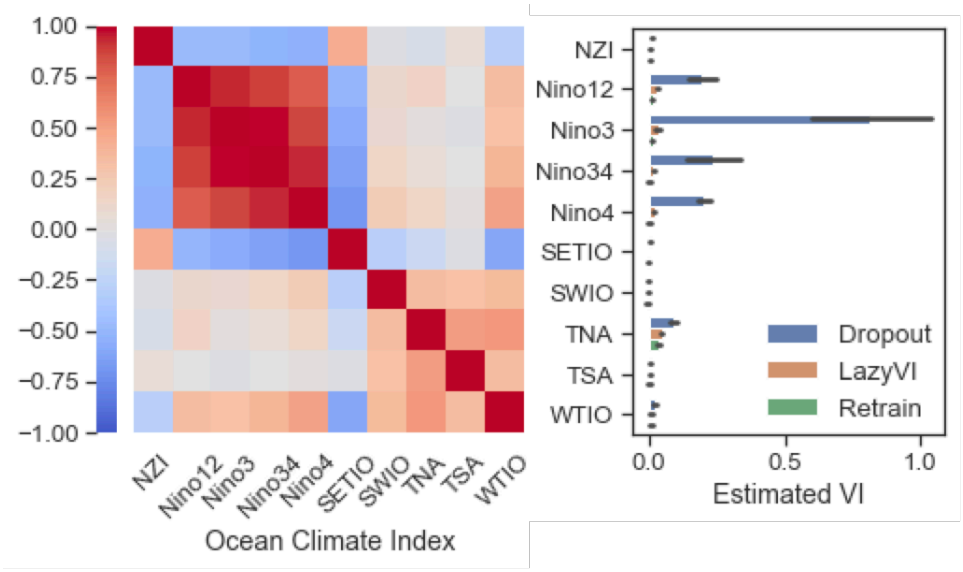


Figure 3.8: Left: sample covariance matrix of the OCIs across the 40 LENS ensemble members; Right: estimated VI for each OCI across 10 different train/test splits.

we propose a method, LazyVI, for efficiently estimating variable importance based on a linearization of a fully trained neural network. We prove that our method provides an accurate estimate of VI and can achieve the same rate of accuracy as a computationally expensive retraining method nearly as quickly as the inaccurate dropout method. We further show how to construct confidence intervals around these estimates.

When features are correlated, the quantity VI defined in (3.7) tends to zero. Recent work proposes using Shapley values to measure variable importance, arguing that their handling of correlated variables, which assigns similar positive weights to correlated important variables, is desirable (Owen and Prieur, 2016; Williamson and Feng, 2020). These papers also note that Shapley values are prohibitively expensive to compute, as they require fitting a new model for each of the 2^p possible subsets of variables. However, we note that computing the Shapley values requires many calculations of the quantity in (3.7); an important avenue of future work is investigating the use of our LazyVI framework to accelerate the computation of Shapley values.

Finally, a potential alternative to our proposed LazyVI method is to first train a full

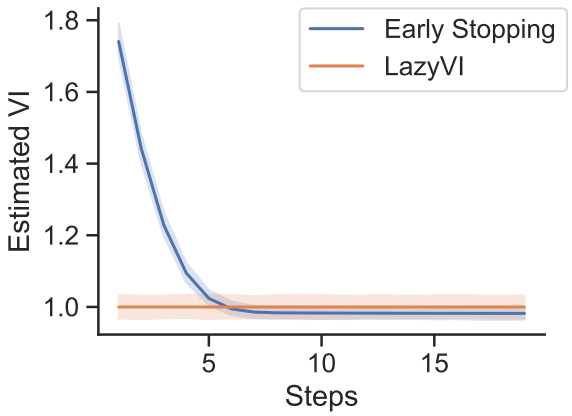


Figure 3.9: Estimated VI with increasing number of steps toward retraining. Shading represents std. across 10 repetitions.

model (as we do) and then train the reduced model using a gradient-based method initialized with the full model parameters and stopped early. Empirical evidence suggests that this approach would have similar speed and accuracy to our LazyVI approach due to the implicit regularization associated with early stopping. 3.9 compares the accuracy of the two methods at different numbers of steps. However, we currently lack theoretical guarantees for early stopping in this setting. It is possible that our theoretical results could lead to new insights into early stopping for assessing VI due to the intimate connection between kernel ridge regression and early stopping algorithms Raskutti et al. (2014). In fact, if the eigenvalues of the NTK matrix at the full model initialization decay in a sufficiently fast rate, early stopping of the reduced model training could give an as good estimate of the reduced model (see 3.9). However, analyzing early stopping in this setting requires characterizing the spectrum of the NTK with *the full model initialization*, whereas most spectral properties of the NTK have been developed under the assumption of a *random initialization* Nguyen et al. (2021); Montanari and Zhong (2020). Better understanding the NTK spectrum after full-model initialization in the future could provide new insights into fast algorithms for VI estimation.

CHAPTER 4

DEVELOPMENT OF CITY-SCALE SYNTHETIC POPULATION TO SIMULATE COVID-19 TRANSMISSION AND RESPONSE

4.1 Introduction

At the time of writing, the COVID-19 pandemic is two years old. We've learned a few things in that time – we probably don't need to disinfect our groceries, we should probably pass on going to the club in the middle of a variant surge, and we should definitely get vaccinated. Arising at a particularly fraught moment in U.S. politics, we've been a nation sharply divided over public health interventions; with the scientific process playing out in real-time and more access to information than ever before in human history, this pandemic has made armchair epidemiologists of us all.[1] A true collaboration between data scientists and epidemiologists was necessary in order to leverage data appropriately to aid in pandemic decision-making, and this chapter describes work that was used by the Chicago Department of Public health, the City of Chicago Mayor's Office, and the Governor of IL's COVID task force to inform policy.

This pandemic has interrupted our daily lives and changed the landscape of social interaction, with many office jobs still remote, schools intermittently closing, conferences and gatherings shifting online, and a varying set of restrictions on how we move about our cities. Global outcomes of COVID-19 have varied drastically based on locally enacted public health interventions, household structures, and lifestyles. In addition, disease outcomes have disparately impacted different populations; age, race, and underlying health conditions are all shown to play a role in disease severity, and disparate access and uptake of vaccines remains

1. This very much includes me.

a problem. A common pitfall of many existing epidemiologic models is a homogeneity assumption; in most models, diseases are assumed to progress in the same way for an entire population. However, due to its social complexities, attempts to model COVID-19 transmission and outcomes must take into account not only the highly social and complicated nature of its transmission, but also the heterogeneity of its progression. These dynamics are complicated and not easily reducible to standard, equations-based epidemiologic models (Thompson and Wattam, 2021).

There are two main classes of models to help predict and make inferences about disease progression. The first and oldest of these, compartment models (Kermack et al., 1927), partition the population into discrete compartments, each describing a different disease state (e.g. Susceptible, Exposed, Infections, or Recovered (SEIR)). Importantly, these models assume homogeneity within each disease state; individuals are assumed to all have an equally likely chance of becoming infected, and transitions between states are governed by a set of differential equations. These types of models are fast and flexible, but the homogeneity assumption makes them a less desirable choice to model COVID-19. With the emergence of accessible high-powered computing over the past several decades (Collier et al., 2015; Ozik et al., 2021a), a new class of simulation-based models have emerged that are able to integrate population heterogeneity in their forecasting. Agent-based models (ABMs) consist of individual agents, bestowed with individual attributes (e.g. location, age, race, sex, occupation, income), interacting (read: infecting) one another in real-time (Macal, 2016; Macal et al., 2018). At a high level, ABMs capture the dynamics of human and social behavior at an individual level in order to study population-level outcomes as complex emergent phenomena. By simulating agent-to-agent and agent-to-environment interactions, epidemiologists are able to simulate disease transmission at the individual level, and policy-makers are able to test public health interventions like stay-at-home orders or closures, making ABMs particularly well-suited for modeling pandemics like COVID-19 (Macal et al.,

2014a).

4.1.1 Large-scale simulations as interpretable models

With COVID-19 at the forefront of the public imagination for so many months, thousands and thousands[2] of papers have been published studying various aspects of its spread and impacts. As mentioned above, the majority of the epidemiological models are either compartmental or agent-based, and since ours is agent-based, this literature review will focus on work using these types of models. For a comprehensive overview of the compartmental COVID-19 transmission models, see Cao and Liu (2021).

In the context of COVID-19 decision-making, understanding heterogeneous disease transmission is key to model interpretability. By endowing individuals in a population to have different attributes and different behaviors, policy-makers are able to better understand who the policies they enact impact the most and why. Because of this, simulation-based agent-based models are far more interpretable than related methods. There are far too many excellent COVID-19-related agent-based models to exhaustively list; see Hinch et al. (2021); Kerr et al. (2021); Gaudou et al. (2020); Milne and Xie (2020); Árbol and Iglesias (2020); Milne et al. (2020); Alagoz et al. (2021) and Cuevas (2020) for some examples. A few works in particular were especially influential in the development of our synthetic population. The first of these, Chang et al. (2021), inspired our use of SafeGraph data (see Section 4.3.1; SafeGraph is a company that collects anonymized location data from mobile phones to points of interest) to help us expand our *mobility network*, which describes the patterns of movement throughout the simulation and forms the basis for disease transmission. While the model developed in Chang et al. (2021) is not explicitly an agent-based model, this work overlaid an SEIR compartmental model on a mobility network derived from SafeGraph data. From SafeGraph, they were able to construct a mapping between each Census Block Groups

2. A quick Google Scholar search of "COVID-19" returns 1.9 million results, in fact!

(CBG) and SafeGraph POI, so each CBG maintained its own disease state. This work was then able to study transmission patterns based on place type and broad demographics of each CBG. Prior to this paper, we were using OpenStreetMap to identify places of interest and assigned places randomly to agents based on location; discovering SafeGraph enabled us to embed actual mobility patterns into our model for a much more realistic view of how COVID-19 spread.

The next two works I discuss here take similar approaches to CityCOVID, with some key differences, and were developed in parallel, lending support to our general approach. The first of these, JUNE (Aylett-Bullock et al., 2021), is an open-source framework for running an agent-based simulation of disease spread, with overall goals and attributes quite similar to ours. JUNE allows for flexible input data to define how granular its synthetic population is (as published, they only consider age and sex), and probabilistically assigns activities at each time step according to age sex, largely based on social contact matrices. Then, JUNE overlays an epidemiology model on top its synthetic population and presents options for testing policy interventions. A key distinction between JUNE and CityCOVID is their use of pre-defined contact matrices in dictating disease transmission, rather than overlaying activity schedules to dictate movement throughout the simulation. Thompson and Wattam (2021) present a Luxembourg-based ABM that is much more detailed with the inclusion of a wide variety of place types and behavioral patterns. Notably, like our approach but unlike most, Thompson and Wattam (2021) use time-use surveys to define interaction patterns; rather than pre-specifying contact patterns as model inputs, incorporating time-use surveys allows contact patterns to *emerge* from the model, enabling closer investigation of sociodemographic transmission disparities. The Luxembourg model is built to enable the investigation of highly-granular transmission vectors, but does not concern itself as much with granularity within the population demographics, and only considers age and household structure.

Our contribution is the development of a highly granular and flexible synthetic population to be used as the input to CityCOVID. With an eye toward better understanding health disparities and the social nature of COVID-19 transmission, we use a wide variety of statistically representative data sources to generate a population of synthetic agents, activity schedules, and places that are then combined in a way to ensure the synthetic population mimics the demographic profile and behavioral patterns of the area under study as closely as possible. At a high level, this work builds upon years of expertise at Argonne National Lab and employs a suite of statistical methods to use micro-level population data to build a large-scale simulation of COVID-19 spread as a function of human and social behavior.

4.2 Large-scale agent-based models and synthetic populations

4.2.1 chiSIM

The Chicago Social Interaction Model (chiSIM, Macal et al. (2018)) is a highly-granular city-scale agent-based model of Chicago, parallelized for running on high-performance computers (HPCs) (Collier et al., 2015). chiSIM was developed at Argonne National Laboratory and has been widely used to inform policy-makers in Chicago on topics like the spread of Ebola and MRSA (Macal et al., 2014b) and the impact of community health interventions (Kaligotla et al., 2018). With a robust simulation framework in place, the chiSIM framework was quickly extended to model the spread of COVID-19 in Chicago at the beginning of March 2020, a project known as CityCOVID[3].

CityCOVID is a city-scale agent-based model (ABM) of millions of people in a large metropolitan area, currently the Chicago area (Macal et al., 2020). CityCOVID is being used to understand the possible spread of COVID-19 and to model the uncertainties of human behavior in response to public health interventions. Underlying CityCOVID is a

3. `https://www.anl.gov/dis/CityCOVID`

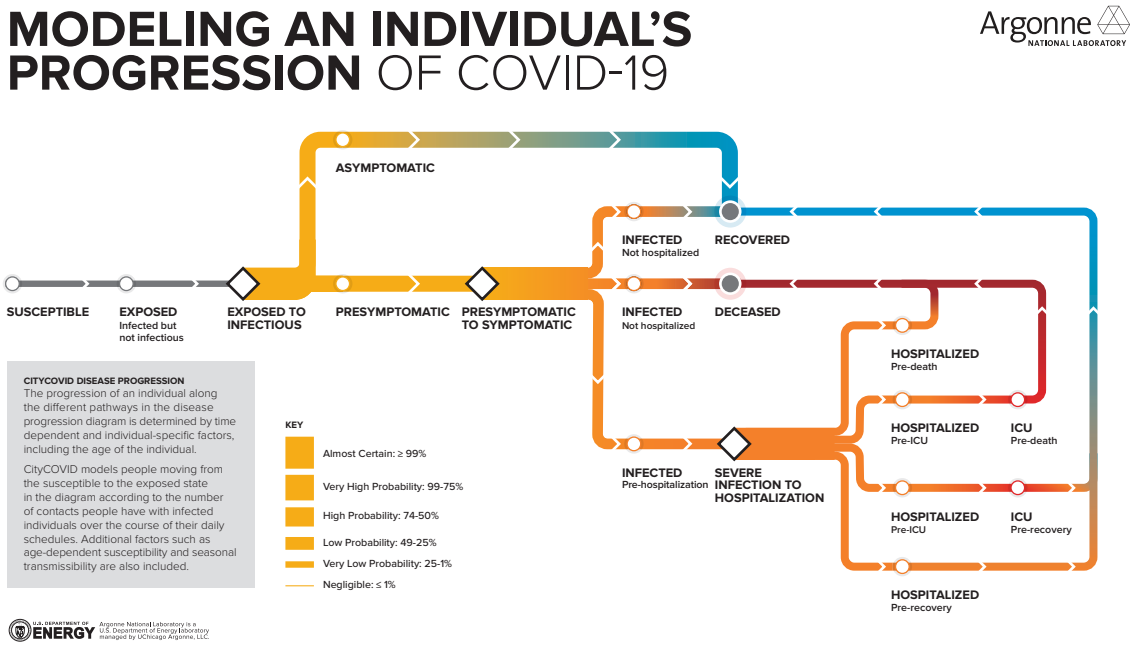


Figure 4.1: Disease progression pathways used to define the CityCOVID epidemiology[3]

synthetic population (Kaligotla et al., 2020b) that is statistically representative of Chicago's population (2.7 million persons), along with their associated places (1.4 million places) and behaviors (13,000 activity schedules). During a simulated day, agents move from place-to-place, hour-by-hour, engaging in social activities and interactions with other co-located agents, resulting in an endogenous co-location or contact network. COVID-19 transmission is determined via a simulated epidemiological model (Figure 4.1) based on this generated contact network. Model parameters are fit to empirical COVID-19 health and hospitalization data, and modeling assumptions about agent behavior and government interventions are informed by the latest scientific findings. Since March 2020, CityCOVID has informed and supported decision-making by the Chicago Department of Public Health (CDPH), the City of Chicago Mayor's office, and the Governor's task force regarding school closures and other non-pharmaceutical interventions (NPI) (Hotton et al., 2022).

4.2.2 Synthetic populations

Synthetic populations act as the primary input to ABMs; it is through their individual characteristics and interactions with their environments that the complex phenomena under study can emerge. When ABMs are being used to understand disease transmission within certain communities and to test policy interventions, it is important for the underlying population to be as realistic as possible in order to capture population heterogeneity. The first large-scale synthetic population used to power an ABM is described in Beckman et al. (1996) for use in a travel forecasting model. This paper was the first of many to use Iterative Proportional Fitting (IPF), and algorithm introduced by Deming and Stephan (1940) and shown to converge by Fienberg (1970), to combine census summary tables and microdata to estimate the joint distribution of population characteristics. Given marginal distributions of population characteristics (such as those provided through census reports), IPF estimates a joint distribution across characteristics by estimating cell values in a contingency table such that the odd ratios of the marginals are held constant. The contingency table is seeded using microdata such as the Public Use Microdata Sample (PUMS) to specify a covariance structure among characteristics, and then the IPF procedure updates the contingency table until the pre-specified marginal totals are matched. The estimated joint distribution defines the weights with which to sample synthetic agents from the microdata. This method was used to develop the first large-scale synthetic population specifically built for disease modeling - the RTI synthetic household population database (Cajka et al., 2010). The RTI synthetic population was created as part of an NIH infectious disease modeling initiative and is discussed in greater detail in Section 4.3.1. IPF has been studied extensively, and a variety of extensions have been proposed to address key issues such as convergence in the presence of missing or incomplete data, fidelity across population axes, and error propagation (Arentze et al., 2007; Guo and Bhat, 2007; Ramadan and Sisiopiku, 2020).

The data requirements for IPF are strict - aggregated data (like the census) is used

to fix the marginals, and disaggregated microdata (like PUMS) are used to intialize the contingency table. Farooq et al. (2013) presents a simulation-based alternative to IPF that enables all partial views of the joint distribution (marginals, conditionals, and samples) and uses Gibbs sampling to draw from the estimated joint distribution. This paper shows that their method outperforms IPF in terms of fit to the joint distribution and holds the marginals just as well. This method is restricted to nonhierarchical data, and Casati et al. (2015) extend the method to hierarchical populations (e.g. individuals with roles inside a household). Recently, Gallagher et al. (2018) developed a flexible open-source framework for generating synthetic ecosystems that allows the user to specify their sampling methodology based on the availability and quality of the input data.

For the purpose of understanding COVID-19 transmission vectors, we require a more granular synthetic population than what is provided by existing synthetic ecosystems. We need agents to have behaviors, the ability to visit and interact at a wide variety of locations, underlying, co-morbid health conditions, and other indicators of vulnerability. The synthetic population we use for CityCOVID is an augmentation of a 2018 version of the RTI synthetic population. This is a slightly different task than what the literature described above is doing - in those cases, the joint distributions that are estimated and then sampled from to fully create a new population, but in our case, we are estimating conditional distributions that we will then assign to our population (e.g. given a young Black male in Hyde Park, Chicago, how likely are they to have a co-morbid health condition?). This chapter describes the development of this augmented synthetic population, designed specifically to enable a highly granular modeling of COVID-19 transmission in Chicago.

4.3 Development of the core synthetic population

The synthetic environment required to run CityCOVID consists of 3 elements - a population of synthetic agents **P**, activity schedules **A**, and places **L**, constructed from various data

sources. Together, they represent *in silico* a statistically representative population of the area under study, which for the purpose of this thesis, is Chicago, Illinois. The development of the synthetic population described here is an expansion of Macal et al. (2018), which has been used to study disease transmission and other social processes (Macal et al., 2014a; Kaligotla et al., 2018; Ozik et al., 2018).

Each agent $p_i \in \mathbf{P}$ is represented by a vector of characteristics $p_i := (i_1, i_2, \ldots, i_n)$ (e.g., age, race, gender, household location) that are statistically representative of the area under study. An activity schedule $A_j \in \mathbf{A}$ contains a vector of individual activities $A_j := (a_1, a_2, \ldots a_K)$ where each a_k :=(start time, stop time, location). Activities are time-continuous (i.e. a_i[stop time] = a_{i+1}[start time]) across a 24-hour day. Each A_j is constructed from a time-use survey (see next section for more details) and is associated with the demographics of the survey responder. Finally, each location $l_t \in \mathbf{L}$ is represented by $l_t :=$ (geolocation, place type). Synthetic population development involves first constructing each of $\mathbf{P}, \mathbf{A}$, and $\mathbf{L}$ using methods described below, and then developing a series of algorithms to integrate the datasets to be used by the ABM.

In this section, we describe the method used to construct the core synthetic population. These methods are flexible and generalizable; while the focus of this thesis is on the Chicago metropolitan area, the data sources are available for all regions of the United States and can easily be filtered for other regions. As a proof of concept for the generalizability of our framework, we developed a similar synthetic population of Northern Idaho for use by the Idaho National Lab.

4.3.1 Data sources

The RTI synthetic household population databases (Cajka et al., 2010) lay the foundation for the CityCOVID synthetic population. The RTI synthetic ecosystem consists of agents with demographic characteristics (age, sex, race, household income/structure, and industry)

and a limited set of locations associated with the individuals (living quarters, schools, and workplaces). In our work, we extend this RTI population to include:

- Activity schedules to dictate agent movement throughout the synthetic environment
- Additional place types to better understand transmission vectors, built from actual mobility data
- Additional attributes for each agent, such as ethnicity, underlying co-morbid health conditions and occupation

In this section, we walk through the various data sources we used to develop the City-COVID synthetic population and the algorithms and methods we used to integrate them all into a cohesive population. The workflow consists of first gathering and processing data from various census and other sources, and then combining them using a variety of statistical methods to create a population that realistically reflects the demographics of the area under consideration. For reference, a very high-level view of our development workflow can be found in Figure 4.2.

RTI U.S. Synthetic Household Population

The RTI U.S. Synthetic Household Population[4] was developed to support the NIH Models of Infection Disease Agent Study (MIDAS [5]) to enable agent-based disease modeling. We use a version of the RTI synthetic population generated from 2018 U.S. Census data to generate our synthetic agents and households. At a high-level, RTI first generates synthetic households that match the American Community Survey (ACS) estimates for household size, race, income, and age of the head of household at the Census Block Group (CBG) level. Then, using Public Use Microdata (PUMS), individuals are sampled and assigned

4. https://www.rti.org/impact/rti-us-synthetic-household-population

5. https://midasnetwork.us/

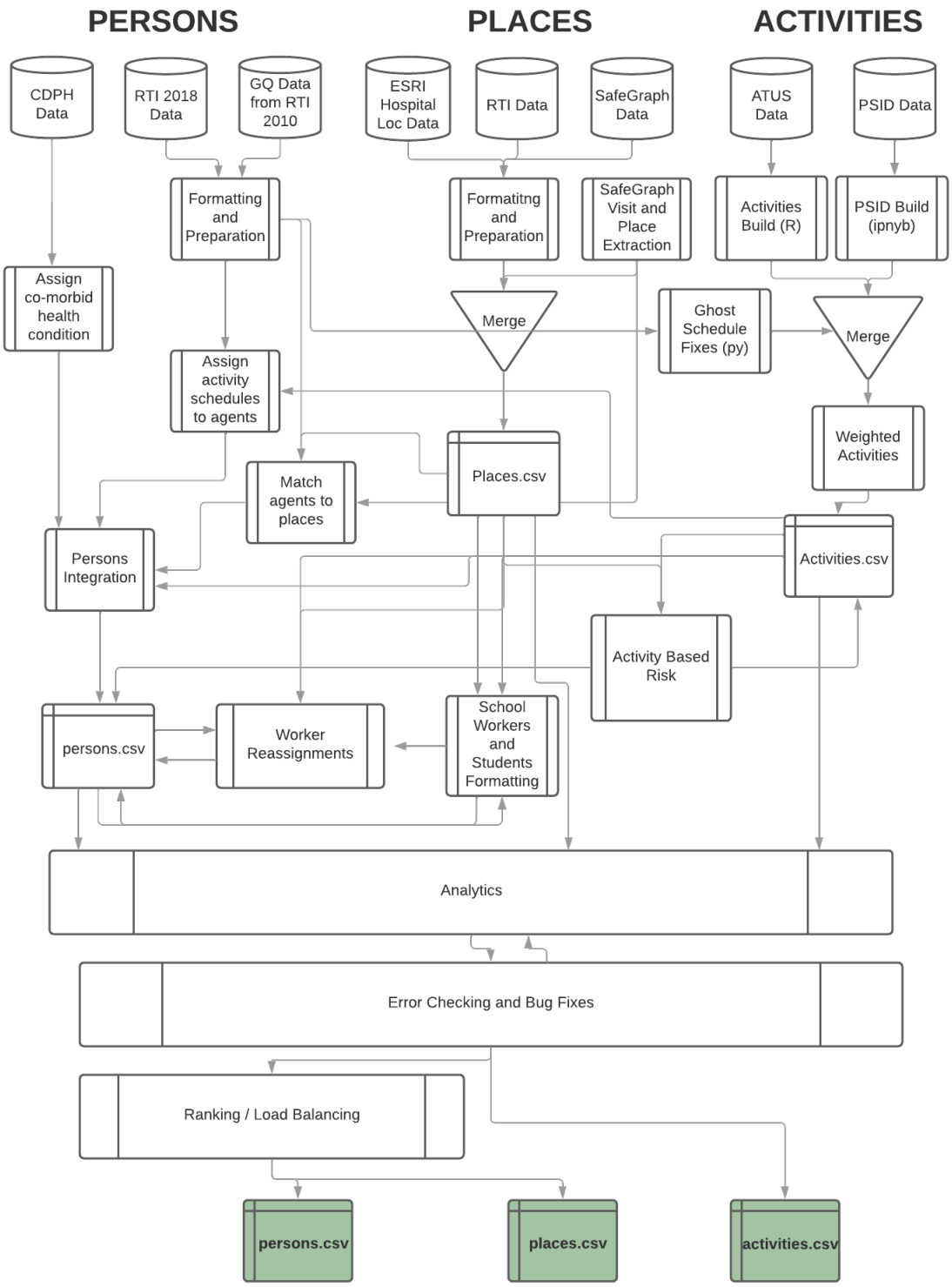


Figure 4.2: Heuristic for the development of the CityCOVID synthetic population

to households such that the population in each CBG aligns with 2018 census estimates of age, race, sex, and relation of person to the head of the household. RTI also maintains a database of actual K-12 schools that eligible synthetic youths are assigned to based on location and capacity, along with a database of synthetic workplaces that are appropriately matched to the population. See Cajka et al. (2010) for full information on the creation of the RTI synthetic ecosystem. Figure 4.3 shows the population density at the CBG level for Chicago.

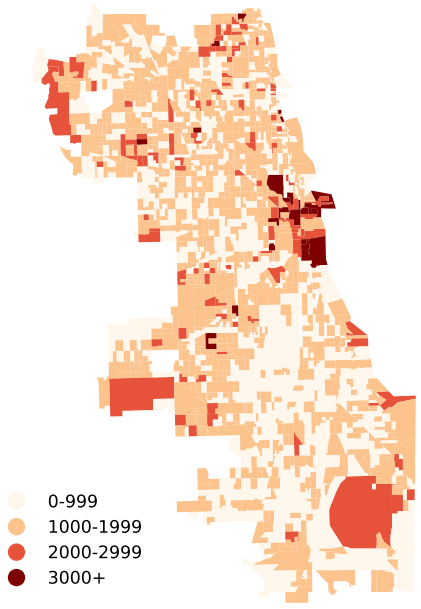


Figure 4.3: Population by CBG

Time-use surveys

Agents' daily activity schedules are constructed from two comprehensive activity surveys of the general US population – the American Time Use Survey (ATUS) and the Panel Study of Income Dynamics (PSID). Specifically, we use ATUS 2018[6] for agents aged 18 and up and PSID[7] data for the under-18 population, as this source has proved more reliable for younger

6. https://www.bls.gov/tus/datafiles-2018.htm

7. https://simba.isr.umich.edu

schedules in the past (Macal et al., 2018). These are both well-established surveys that asks a representative sample of individuals in the United States to track their daily activities for a random weekday or weekend and record the start- and end-time for each activity, along with the location and activity performed. Each record is associated with a detailed demographic profile and associated survey weights, included to correct for oversampling of certain demographics and non-response bias.[8]. In all, we construct $12,419$ unique activity profiles.

SafeGraph

The additional social activity locations not included in the base RTI dataset were extracted from SafeGraph, a data company that aggregates anonymized location data from numerous applications in order to provide insights about physical places, via the SafeGraph Community. To enhance privacy, SafeGraph excludes census block group information if fewer than two devices visited an establishment in a month from a given census block group.[9] The SafeGraph data universe consists of points of interest (POI) and visitor traffic patterns between them. For each POI, we are given geographic information like location and geometry, a categorization of place type given by US Census Bureau industry codes, and additional information like open hours and branding. Associated with POIs are traffic patterns, which consist of information about visits to that POI (total monthly visits, median dwell time, etc) along with information about the demographics of visitors to that POI, such as each visitor's census block group (CBG).

For our Chicago-based synthetic population, we extract Illinois POIs and visit patterns from 2019, which was chosen as the most recent year with "normal" mobility patterns; establishing a baseline from a time before lockdowns enables us to more accurately test policy

8. `https://www.bls.gov/tus/atususersguide.pdf`

9. SafeGraph asked that this phrase be included to cite their product

interventions using CityCOVID (e.g. how will gym closures impact COVID transmission, relative to regular usage?). An entry in the visits data is a monthly aggregation of visit information to a single POI; in total, there were 1,931,186 measurements covering 172,249 distinct locations in the 2019 Illinois data. We use this data to augment our synthetic population with a wide variety of social locations not included in the existing population, with agent visits to these places dictated by the visit patterns in the data. See Table 4.2 for a full list of the place types extracted from SafeGraph.

Source	**Agent attribute**	**Possible values**
RTI 2018	Age	0-99
	Sex	Male, Female
	Race	Black, White, Asian, Other
	Household income	$\leq$10K, 10-15K, 15-25K, 25-35K, 35-50K, 50-100K, >100K
	Household size	1-6
	Workplace industry	2-digit NAICS codes (Table 4.3)
RTI 2010	General quarters	Prison, College Dorm, Nursing Home, Military Barracks
ACS	Occupation	SOC codes (Table 4.4)
ACS	Ethnicity	Hispanic/Latino or not
CDPH	Co-morbidities	Diabetes, Hypertension, Obesity, Asthma
SafeGraph	Social locations	Place types in Table 4.2

Table 4.1: Agent attributes and sources

4.3.2 Integration algorithms

As shown in Figure 4.2, all the different data sources need to be joined together in such a way that the population represents the demographics and behaviors of the population of Chicago. Recall that our population consists of a population of synthetic agents $\mathbf{P}$, activity schedules $\mathbf{A}$, and locations $\mathbf{L}$. For the population to be a valid input to CityCOVID, we need to map agents to activity schedules through a mapping function $f_a : \mathbf{A} \rightarrow \mathbf{P}$, and agents to locations with $f_l : \mathbf{L} \rightarrow \mathbf{P}$. Once these datasets are combined, the simulation is able to detect where an individual should be at a given time of day through their activity schedule, and then send them to an appropriate location based on their available places.

Source	Place Type	Total	Per agent
RTI	Household	1,164,722	1
	School	2,259	1
	Workplace	152,340	1
	Prison	75	1
	College Dorm	47	1
SafeGraph	Restaurants and bars	20,183	5
	Recreation	7,881	5
	Stores	7,778	5
	Convenience stores	525	3
	Supermarket	2,210	3
	Museums	157	3
	Places of worship	3,202	1
	Library	273	1
	Daycare	1,961	1
Hand Curated	Hospital	122	1
	Long-term care facility	113	1
	Urgent care clinic	14,187	1

Table 4.2: Types of places in the synthetic population, along with their sources, the total number of each place type, and the number assigned to each individual agent

Mapping agents to schedules

CityCOVID (and the chiSIM framework in general), unlike many similar COVID-19 ABMs (see e.g., Aylett-Bullock et al. (2021)), endogenously generates contact networks (Macal et al., 2018) as an output of the simulation rather than using an imposed pre-generated contact network (Mossong et al., 2008) to determine transmission vectors and dynamics. Contact occurs between agents when they are at the same place at the same time, and mixing dynamics emerge after the simulation has played out over an extended period of time. For this to happen, the ABM must send agents to different places at certain times of day in a realistically stochastic manner; to achieve this, each agent is assigned a list of 10 weekday and weekend activity profiles that are then randomly chosen on a given day of the simulation. Essentially, different people have different schedules, and we use simulations to investigate at mixing as a result of this heterogeneous behavior. This enables us to generate

individual-level transmission vectors based on co-locations that capture the heterogeneity of the population.

Activity profiles are probabilistically assigned to agents through a mapping function $f_a : \mathbf{A} \rightarrow \mathbf{P}$ based on overlapping demographics and sample weights. For each agent, we select all activity profiles that are associated with an ATUS or PSID respondent of the same age, race, and sex. These demographics are chosen to be granular enough to capture important transmission patterns while still ensuring random mixing. We are able to find at least one perfectly matched schedule for 99% of agents; the remaining are assigned schedules based on age. Schedules are separated by weekend and weekday, and we create an empirical probability distribution over each group based on the survey weights. We use these probabilities to randomly sample (with replacement) ten weekday and ten weekend schedules for each agent, taking into consideration whether an agent is associated with each place a given activity will send them to (e.g. we will not assign an agent a "school" activity if they do not have a school assigned to them).

Formally, for each activity profile $A_j \in \mathbf{A}$, let $d_{A_j} := \left(\text{sex}_j, \text{age}_j, \text{race}_j\right)$ be the demographic attributes associated with A_j, and let w_j be its corresponding survey weight. Separately for weekdays and weekends, each agent $p_i \in \mathbf{P}$ is assigned a list of ten schedules according to Algorithm 3. Figure 4.4 shows the distribution of agent locations by hour on a randomly selected weekday and weekend for each individual. We see that during the week, many agents are in school or at work, while on the weekend we see a larger variety of social mixing occurring.

Matching agents to places

The base synthetic population from RTI assigns each agent to either a household or group quarters (like nursing homes, college dorms, or prisons; there was significant concern early in the pandemic over rapid spread in these types of places due to close proximity) and then,

Algorithm 3 Schedule mapping ($f_a : \mathbf{A} \to \mathbf{P}$)

Require: Activity schedules $\mathbf{A}$, Agents $\mathbf{P}$

for $p_i \in \mathbf{P}$ **do**

$\mathbf{A}_i = \{A_j \in \mathbf{A}$ s.t. p_i and d_{A_j} overlap on attributes and have compatible place types$\}$

$W_i = \sum_{j \in [\mathbf{A}_i]} w_j$ where $[\mathbf{A}_i]$ is the list of indices in A_i

$P_i = (w_j / W_i)_{j \in [\mathbf{A}_i]}$

p_i[schedules] $\sim$ Multinomial(10, P_i)

end for

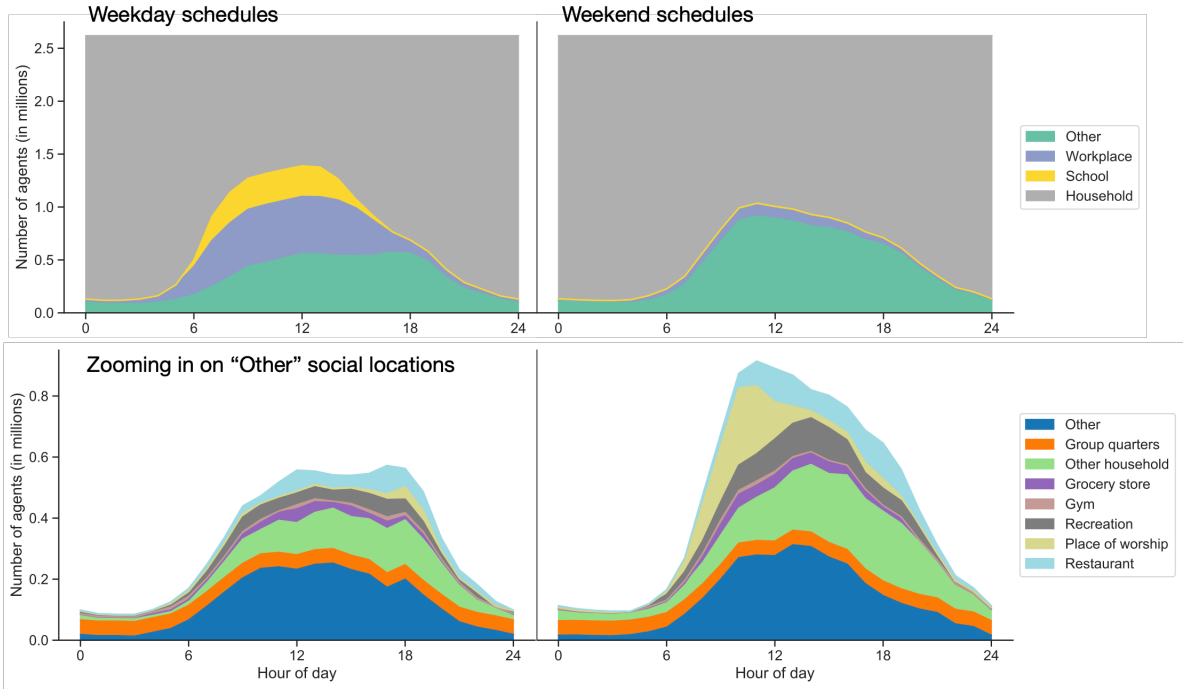


Figure 4.4: Distribution of agent schedule locations for randomly selected weekday and weekend schedules.

depending on age, schools and workplaces. Because so much of the early-pandemic debate around COVID-19 policy interventions revolved around the accessibility of social spaces, we substantially increase the number of place types agents can occupy using POIs from SafeGraph. Agents are assigned places through a matching function $f_l := \mathbf{L} \to \mathbf{P}$. Each agent is assigned a certain number of each category of SafeGraph place type (see Table 4.2 for a comprehensive list) that the CityCOVID ABM randomly chooses from when an agent's schedule sends them to that location category. For example, each agent is assigned five

different possible restaurants or bars; if their schedule says to go to a restaurant at 7pm on a Friday, CityCOVID will randomly choose one of these five places to send them to. In making these assignments, we seek to mimic, as realistically as possible, the actual flow of traffic through these places. For each place, we extract the monthly number of visits by the *visitor home CBG* field in the 2019 SafeGraph patterns data. This field is determined by analyzing six weeks of data during nighttime hours, and for increased anonymity, is only populated if there are at least two devices visiting a POI from a given origin CBG. Because not all visitors are assigned a home CBG, we estimate the proportion of visitors from each CBG across the entire year and then distribute the average *total number of monthly visits* field, which is more accurately reported, according to these proportions. This level of detail in combination with demographic-based schedule assignments allows us to realistically understand how and where COVID-19 transmission might occur.

Formally, we let T store the total monthly visitors to each place such that $T_{i,j} := \{$Raw number of visitors to place i during month $j\}$, and C store the monthly visits by home CBG such that $C_{i,j,k} := \{$Raw number of visitors from home CBG k to place i during month $j\}$. Using these inputs, we construct an estimate V of the monthly visitors from each CBG such that $V_{i,k} := \{$Estimated monthly visitors from home CBG k to place $i\}$ using Algorithm 4.

Algorithm 4 Visit pattern extraction

Require: SafeGraph raw visit patterns T and C
  for $i \in$ SafeGraph places **do**
    $\tilde{T}_i = \frac{1}{12}\sum_{j=1}^{12} T_{i,j}$ ▷ Average total monthly visits
    for $k \in CBGs$ **do**
      $\tilde{C}_{i,k} = \frac{1}{12}\sum_{j=1}^{12} C_{i,j,k}$ ▷ Average monthly visits by home CBG
    end for
    for $k \in CBGs$ **do**
      $V_{i,k} := \tilde{T}_i \left(\frac{\tilde{C}_{i,k}}{\sum_k \tilde{C}_{i,k}}\right)$ ▷ Distribute average total visits according to distribution from estimated CBG visit patterns
    end for
  end for

Then, for each agent and each type of SafeGraph POI, we have an estimate of which

specific places that agent is likely to visit based on their home CBG. For a given CBG, we extract all SafeGraph places of a particular type (e.g. restaurants) and the corresponding estimates of the number of monthly visits from that CBG. We then build an empirical probability distribution across these places, which we use as weights in a multinomial draw to assign each agent in that CBG a set of places for CityCOVID to choose (Algorithm 5). See Figure 4.5 for an example of the assigned locations for a handful of place types relative to three distinct agent households.

Algorithm 5 Place matching ($f_l := \mathbf{L} \rightarrow \mathbf{P}$)

Require: Estimated CBG visits V, SafeGraph place type p, number of assignments n
  for $k \in$ *agent household CBGs* **do**
    $v_k = \sum_{i \in [p]} V_{i,k}$ where $[p]$ is the set of places of type p
    $P_k = (V_{i,k}/v_k)_{i \in [p]}$ ▷ Probability vector across all places of type p
    for $a \in$ Agents with residing in CBG k **do**
      $p_a \sim$ Multinomial(n, P_k)
      Randomly choose from p_a when agent's schedule indicates they go to p
    end for
  end for

4.3.3 Core population overview

To recap, the core synthetic population necessary for running CityCOVID consists of agents $\mathbf{P}$, activity schedules $\mathbf{A}$, location $\mathbf{L}$, a schedule mapping $f_a := \mathbf{A} \rightarrow \mathbf{P}$, and a place matching $f_l := \mathbf{L} \rightarrow \mathbf{P}$. All required input datasets are open-source and available for the entire United States, and the integration algorithms are efficient and easy to implement. With this core population, CityCOVID is well calibrated and can begin to answer important questions about COVID-19 transmission and outcomes.

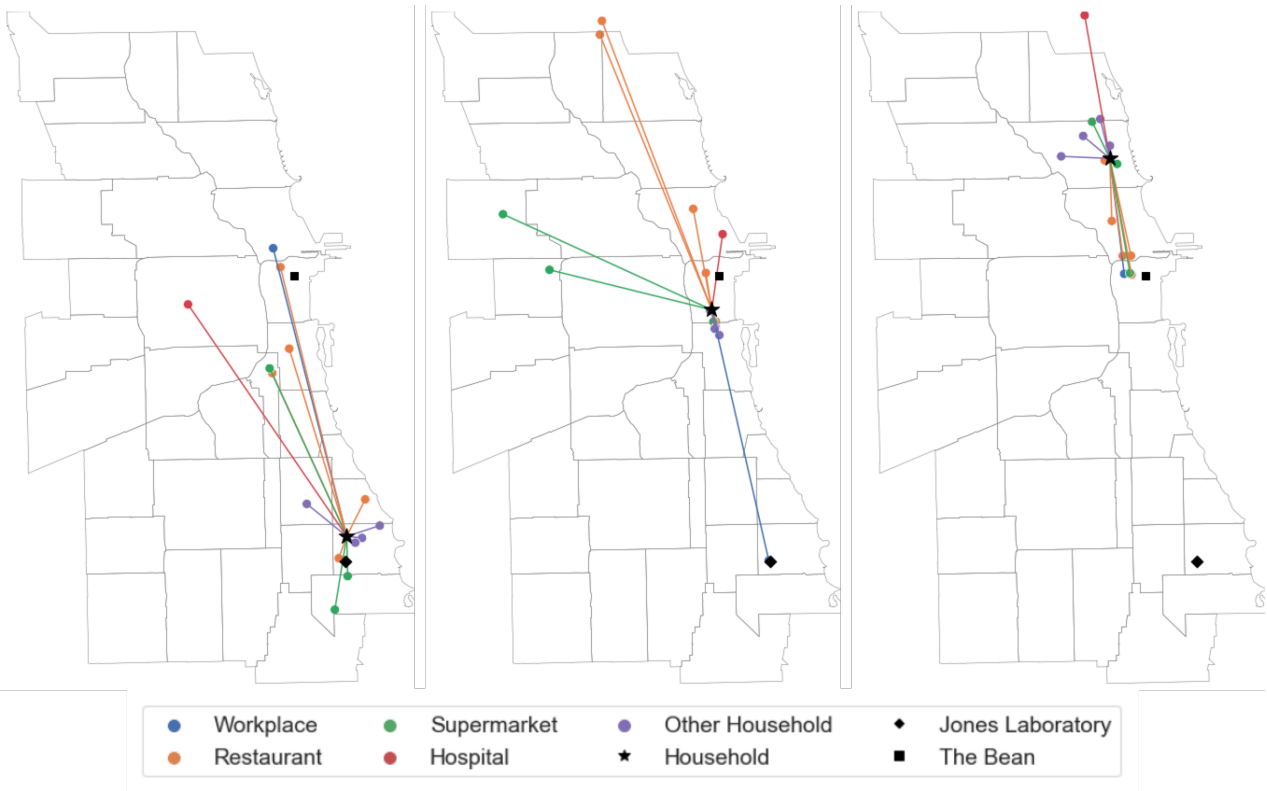


Figure 4.5: A sample of the places assigned to three agents representing different demographic groups. Lines connect places to the agent's household, and colors represent different place types. The agent demographics (Age, Sex, Race, Income) are as follows - Left: (30, Female, White, $\leq$ \$10k); Middle: (46, Female, White, $\geq$ \$100k); Right: (34, Male, Asian, $\geq$ \$100k). The Bean and Jones Laboratory are shown for spatial context.

4.4 Expansion of synthetic populations for detailed experimentation

Our core synthetic population is granular and robust, enabling highly detailed modeling of COVID-19 transmission dynamics in Chicago and elsewhere (Ozik et al., 2021b). With such a framework in place, public health researchers have the opportunity to explore detailed hypothesis about disease transmission and impacts. Some hypotheses can be answered using the core synthetic population, but others require the addition of new population attributes using localized data sources. The methods used to build the core population offer a glimpse into a larger body of work known as *population synthesis* (Ramadan and Sisiopiku, 2020),

which describes the integration of multiple partial views of population attributes.

4.4.1 Population synthesis

A realistic synthetic population is a necessary component of an ABM to provide decision-making support, especially for COVID interventions and in the absence of other methods with similar levels of fidelity (e.g., field experiments). The census and other surveys provide the most complete views of population attributes available to researchers, but important privacy concerns limit access to census surveys to high-level aggregates and small microdata samples (PUMS). As discussed in Section 4.2.2, the typical procedure for synthesizing aggregated and disaggregated data is IPF, which involves fitting a contingency table to the available data. IPF has a number of known pitfalls - for example, if there is a zero in the initialization of the contingency table (i.e. the PUMS sample is missing some combination of attributes), IPF fixes the cell value at zero. For a complete discussion of the pitfalls of IPF and a proposed MCMC-based alternative, see Farooq et al. (2013).

In this section, we describe three population synthesis methods to add the following new attributes to the core synthetic population: ethnicity, essential worker status, and underlying health conditions, all of which are critical for understanding disparities in COVID-19 transmission and outcomes. A variety of data sources are used to integrate these attributes and each require a different set of constraints, highlighting some key insights and challenges in synthetic population development that are broadly applicable to different types of attributes.

4.4.2 Assigning ethnicity

The first new attribute we assign to the core population is ethnicity, and its assignment procedure is straightforward. Using the American Community Survey (ACS) 2019 census estimates, we estimate the proportion of the population who are Hispanic or Latino by race at the census tract level. This process involves aggregating all races besides White, Black,

and Asian into an "Other" category and appropriately aggregating the associated margins of error. To incorporate the appropriate uncertainty into our estimates, for each census tract c and race r we draw an estimated proportion $p_{cr} \sim N(\text{ACS proportion, ACS margin of error})$ (clipping at zero if necessary). Then, for agent A_i in census tract c_i and of race r_i, $Pr(A_i \text{ is Hispanic or Latino}) = p_{c_i r_i}$.

4.4.3 Identifying essential workers through occupations

We next determine whether each agent is an essential worker, which enables experimentation on workplace vulnerability.In Chicago, Illinois Governor Pritzker's Executive Order 20-10[10], introduced on March 20, 2020, required non-essential workers to stay home and provided general guidelines for who was and was not considered an essential worker. Understanding who was working from home during the different phases of Chicago's reopening strategy is critical for understanding how occupational risk factored into peak COVID-19 transmission.

As we can see in Table 4.1, the baseline RTI 2018 synthetic population mapped agents to broad workplace industries, which are represented by two-digit codes from the North American Industry Classification System (NAICS)[11], which is the Federal standard used by statistical agencies in classifying businesses; Table 4.3 gives descriptions of the 20 high-level industries represented by NAICS. The NAICS system sub-divides these high-level industries up to six digits of granularity; the RTI 2018 synthetic population does not include that information. An agents' workplace industry provides insight into whether they are an essential worker, but industry alone cannot differentiate between those within an industry who were under the stay at home order and those who were still required to go to work (e.g. a corporate worker for a large retail brand might be working from home while a store employee might not). Occupations, not high-level industries, likely provide a more accurate picture of

10. https://www2.illinois.gov/Pages/Executive-Orders/ExecutiveOrder2020-10.aspx

11. https://www.census.gov/naics/

essential workers.

Following the guidelines provided by Governor Pritzker's executive order, the Chicago Metropolitan Agency for Planning (CMAP)[12] defined 11 broad occupations (from the *Standard Occupational Classification* (SOC) system[13]) as essential (Table 4.4). These occupational classifications paint a clearer picture of the essential workforce than NAICS industries alone, and thus integrating occupations into the synthetic population will enable less ambiguous differentiation between essential and non-essential workers.

Table 4.3: NAICS industries

NAICS code	**Industry Title**
11	Agriculture, Forestry, Fishing and Hunting
21	Mining
22	Utilities
23	Construction
31-33	Manufacturing
42	Wholesale Trade
44-45	Retail Trade
48-49	Transportation and Warehousing
51	Information
52	Finance and Insurance
53	Real Estate Rental and Leasing
54	Professional, Scientific, and Technical Services
55	Management of Companies and Enterprises
56	Administrative and Support and Waste Management and Remediation Services
61	Educational Services
62	Health Care and Social Assistance
71	Arts, Entertainment, and Recreation
72	Accommodation and Food Services
81	Other Services (except Public Administration)
92	Public Administration

12. https://www.cmap.illinois.gov/updates/all/-/asset_publisher/UIMfSLnFfMB6/content/metropolitan-chicago-s-essential-workers-disproportionately-low-income-people-of-color

13. https://www.bls.gov/soc/home.htm

Table 4.4: Essential SOC occupations

SOC Code	**Occupation description**
21-0000	Community and social services
29-0000	Health diagnosing and treating practitioners, health technologists and technicians, and other technical occupations
31-0000	Healthcare support
33-0000	Protective service
35-0000	Food preparation and service
37-0000	Building and grounds cleaning and maintenance
45-0000	Farming, fishing, and forestry
47-0000	Construction and extraction
49-0000	Installation, maintenance, and repair
51-0000	Production
53-0000	Transportation and material moving

The U.S. Census Bureau's American Community Survey (ACS)[14] is an annual survey that collects detailed socio-economic information from a large sample of U.S. households. This survey includes things like employment, education, housing, and more, and is published both as aggregate summary statistics and Public Use Microdata Samples (PUMS). We use the 2018 one-year ACS PUMS[15], filtered to only include PUMAs (the smallest identifiable geographic unit in the data) in the City of Chicago. PUMS contains a record for each person in a random sample of approximately 1% of the U.S. population. These records include demographic information like age, race-ethnicity, and gender; employment information like income, industry, and occupation; and other important information like household structure and PUMS. Each record is associated with a weight that indicates the number of people in the U.S. population represented by that record[16].

IPUMS classifies occupation and industry using the OCC and IND codes, respectively,

14. https://www.census.gov/programs-surveys/acs/about.html

15. https://usa.ipums.org/usa/sampdesc.shtml#us2018a

16. https://usa.ipums.org/usa/intro.shtml#weights

while the essential occupations are defined by SOC codes and the synthetic population is equipped with NAICS codes for industry. IPUMS provides crosswalks to map between the OCC and SOC, as well as IND and NAICS[17]. Similarly, we need to map the PUMA geographies from the IPUMS surveys to the synthetic population, which are specified down to the Census Block Group (CBG) level. CBGs are nested within Census Tracts, which are then nested within PUMAs. A crosswalk is provided by the Census to map between the two[18].

To assign occupation (and thus essential worker status) to the synthetic population, we develop a methodology to probabilistically assign an IPUMS record to each agent in the synthetic population based on overlapping attributes.

Let $\mathbf{R} := \{r_1, \ldots, r_K\}$ be the set of available records in IPUMS, where record r_j and agent p_i are defined by the following vectors of attributes:

$$\begin{aligned} r_j &:= \big(\text{sex}_j, \text{age}_j, \text{race}_j, \text{ethnicity}_j, \text{income}_j, \text{PUMA}_j, \text{NAICS}_j, \text{SOC}_j, w_j\big) \\ p_i &:= \{\text{sex}_i, \text{age}_i, \text{race}_i, \text{ethnicity}_i, \text{income}_i, \text{PUMA}_i, \text{NAICS}_i\} \end{aligned} \tag{4.1}$$

where w_j is the IPUMS sample weight. For each agent we extract a subset of records R_i so that the attributes of the records in R_i suitably overlap with p_i, and then an occupation is drawn from the possible records in R_i. Formally, let $r_j^{(t)}$ represent the t^{th} element of vector r_j (e.g. $r_j^{(1)}$ corresponds to the sex associated with record r_j). If agent p_i has all attributes in common with record r_j, then p_i overlaps with r_j on $\{r_j^{(1)}, \ldots, r_j^{(7)}\}$ with $|p_i \cap r_j| = 7$. Note that the ordering of the vector r_j is specified from most to least general in order ensure close record matches. Formally,

$$R_i := \{r_j \ : \ |p_i \cap r_j| = \underset{t=1,\ldots,7}{\arg\max}\{r_j^{(k)} \in p_i \ \forall k \leq t\}\}$$

17. https://usa.ipums.org/usa/volii/occ_ind.shtml

18. https://www.census.gov/programs-surveys/geography/guidance/geo-areas/pumas.html

Because all records are not equally representative of the true population, we select an occupation from the records in R_i with probabilities proportionate to the sample weights as follows. Suppose agent p_i has M matching records in R_i, so $R_i = \{r_{j_1}, r_{j_2}, \ldots, r_{j_M}\}$ and let $W_i = \{w_{j_1}, w_{j_2}, \ldots, w_{j_M}\}$ be their associated weights. Then, for record r_{jk}, the probability that agent p_i is assigned the occupation associated with record r_{jk} is given by

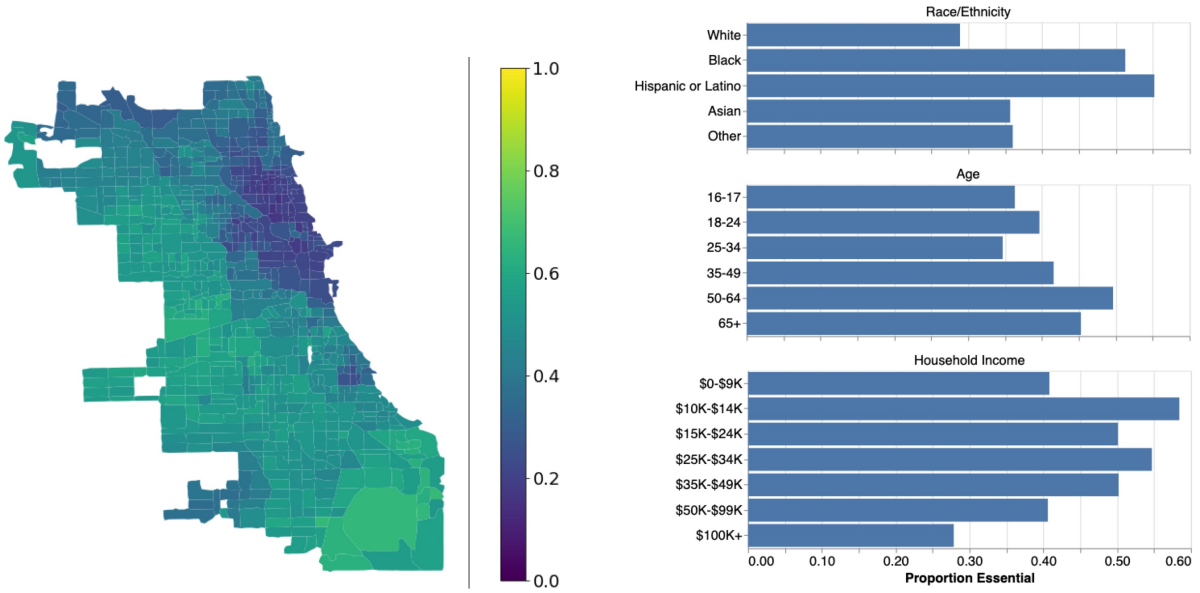
$$Pr(p_i \text{ is assigned } r_{jk}) = \frac{w_{jk}}{\sum_{t=1}^{M} w_{jt}}$$

All agents are thus assigned an occupation through a multinomial draw from R_i with these probabilities. All agents who are assigned one of the essential occupations from Table 4.4 are designated as essential workers.

As mentioned in Section 4.3.1, the workplaces mapped to agents in the RTI databases are purely synthetic. They first used a commercial data product called InfoUSA to estimate the number of "firms" by size in each CBG, and then used Census data to estimate the number of workers commuting between census tracts. Agents of an appropriate age were then randomly assigned a workplace according to these guidelines. As part of our experimental framework, we want to simulate worker-visitor interactions in a workplace setting by remapping workers in relevant industries to SafeGraph social locations.

Overall, this assignment procedure designates approximately 42% of Chicago workers as essential, consistent with an analysis performed by the Chicago Metropolitan Agency for Planning[19]. Also consistent with this analysis, we find large disparities in the spatial and demographic patterns of essential workers (Figure 4.6). In particular, we notice that far more Black and Hispanic/Latino workers are essential than White/Asian workers, and in general essential workers are part of lower-income households.

19. https://www.cmap.illinois.gov/updates/all/-/asset_publisher/UIMfSLnFfMB6/content/metropolitan-chicago-s-essential-workers-disproportionately-low-income-people-of-color



(a) By Census Tract

(b) By race/ethnicity, age, and household income

Figure 4.6: Proportion of workers our method assigned essential status broken down by different demographic variables.

4.4.4 Underlying health conditions

Chronic, underlying health conditions like diabetes, hypertension, obesity, and asthma are known to significantly impact COVID-19 outcomes(Sanyaolu et al., 2020). In this section, we describe the use of a Chicago-specific data source for assigning underlying health conditions to individuals; however, the methodology is general enough that other data sources could be easily swapped in. Once incorporated, we can update the disease pathways in the CityCOVID epidemiological model to better predict hospitalization and mortality outcomes based on these underlying conditions.

Chicago Department of Public Health - Healthy Chicago Survey

Since 2014, the Chicago Department of Public Health (CDPH) had conducted an annual Healthy Chicago Survey[20] (HCS) to better understand the state of public health in Chicago. This survey is conducted through randomized phone calls and collects a wide variety of variables, including demographics and health outcomes. The results of these surveys are reported through the Chicago Health Atlas website, which includes useful visualizations and data downloads for a variety of aggregations of the survey. Among major cities in the United States, Chicago is particularly segregated, and there exist significant health disparities along racial lines. For example, a recent report produced using data from the HCS states that there is a 9.2 year life expectancy gap between Blacks and non-Blacks in Chicago, with the gap largely attributable to chronic health conditions (Aikens et al., 2021).

For a wide range of health indicators, including those we are considering for this work, the Chicago Health Atlas website provides data summaries for individual demographic groups (age, sex, race-ethnicity, and poverty), along with Chicago community areas. Table4.5 gives a breakdown of the percent of different demographics with each condition; we notice that there are wide disparities in disease incidence along racial lines.

Assignment methodology

Assigning these underlying health conditions to agents is less straightforward than ethnicity or essential worker status. There are known associations between conditions (e.g., obesity and diabetes or hypertension); independent assignments of each condition may be inaccurate. Furthermore, we only have a partial view of this data; we don't have access to any joint distributions among demographics (e.g. age and gender). This population synthesis task is slightly different from most described in the literature; typically, the goal is to estimate joint

20. `https://www.chicago.gov/city/en/depts/cdph/supp_info/healthy-communities/healthy-chicago-survey.html`

Table 4.5: Percent of each demographic group with underlying health condition from the 2019 CDPH Healthy Chicago Survey reports. Uncertainty represents a 95% confidence interval.

		Obesity	**Asthma**	**Diabetes**	**Hypertension**
Age	18-29	21.2±4.1	9.7±3.1	2.3±1.6	6.1±2.5
	30-44	31.9±3.7	9.0±2.2	4.3±1.6	15.2±3.0
	45-64	37.0±3.5	11.0±2.2	15.6±2.6	40.8±3.7
	65+	32.8±4.8	7.6±2.3	22.1±4.1	63.3±4.9
Gender	Female	33.4±2.7	11.2±1.8	9.4±1.5	26.9±2.5
	Male	27.9±3.0	7.7±1.7	10.3±1.8	28.6±2.9
Race/ ethnicity	Hispanic or Latino	37.5±4.1	7.4±2.4	12.0±2.5	19.7±3.2
	Black	39.3±3.5	13.9±2.4	11.4±2.1	36.9±3.3
	White	23.7±3.3	8.4±2.1	8.3±2.2	29.5±3.5
	Other	9.8±6.2	N/A	N/A	12.4±6.7

distributions from which a synthetic population can be sampled; here, we need to assign a new attribute to an existing synthetic population, so we need to estimate conditional distributions. To that end, we assume each agent $A_i \in \mathbf{A}$ is drawn from a joint distribution of demographics D_i and underlying health conditions C_i. In other words, we assume

$$A_i \sim P(D_i, C_i)$$

With joint information from the CDPH survey and RTI 2018, the demographics currently under consideration are age, gender, race, and location (we map households to community areas using their coordinates) and we encode health conditions as a binary vector that indicate which of the conditions an agent has (i.e., $C_i := \{\mathbb{1}_{Diabetes}, \mathbb{1}_{Obesity}, \mathbb{1}_{Hypertension}, \mathbb{1}_{Asthma}\}$). Using the aggregate information from CDPH and relevant PUMS microdata, our goal was to estimate the conditional distribution $P(C_i|D_i)$ for each agent A_i.

We ultimately realized that the data sources required to assign a vector of conditions to each agent, even through partial views, were not easy to access. Instead, we opted to assign each health condition independently using a similar procedure to assigning essential workers; we hope to develop a method to account for condition dependencies as future work.

4.5 Load balancing and other computational considerations

CityCOVID is an MPI-distributed ABM (Collier et al., 2015) based on the chiSIM framework (Macal et al., 2018). As part of this work, we developed a *load balancing* algorithm to ensure efficient computation at scale to enable usable in silico experimentation towards informing decision-making, where each unique place is assigned to a specific computing rank. Rank assignments are based on both balancing the number of agents across computing processes and minimizing cross-process communication. To do this, we form a network where each place is a vertex and edge weights correspond to the number of agents moving between those places at a given time-step. We then use the Metis[21] graph partitioning software to optimally allocate vertices to a specified number of nodes. In each time step $t = 1, 2, \ldots, \mathbf{T}$ of the simulation, as agents in the simulation move from one place to another (e.g., from home to school) according to their assigned schedule, they may move among ranks (if the new place is on a different rank). Figure 4.7 illustrates this cross-rank movement.

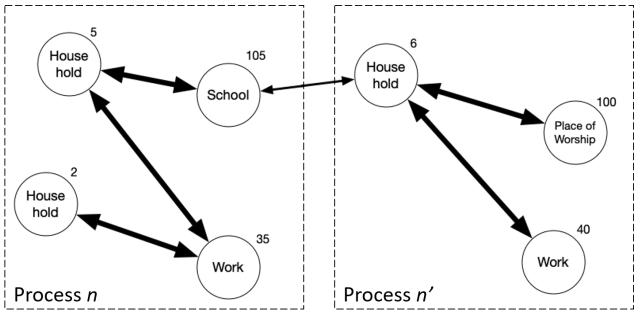


Figure 4.7: Visual Representation of Multi-Rank Distribution of Synthetic Population. Places are assigned to a specific rank (process) and agents move across ranks based on their activity schedules and assigned places. Cross-rank movement between ranks n and n' is shown.

In CityCOVID, agents are assigned 10 possible weekday and weekend demographically appropriate schedules that are randomly sampled on a given simulation day. In order to estimate the mobility network, we must run the CityCOVID model for an extended period

21. `http://glaros.dtc.umn.edu/gkhome/metis/metis/overview`

of time to capture the stochastic variation in the network (Kaligotla et al., 2020a). CityCOVID's agents are also assigned a handful of possible social places to go to, like restaurants and stores, to encourage realistic mixing patterns. By increasing the number of place types and adding randomness to the place assignment procedure, estimating the network to partition becomes very computationally expensive, particularly when we are running experiments to determine optimal ways to structure the network.

At each time step, the CityCOVID simulation:

1. selects and moves agents that need to go to a different rank or are coming from another rank (based on their daily activity schedules)

2. adds moved agents to their respective places on their destination ranks

3. runs a transmission model for agents co-located in each place; and iv) runs a disease progression model for each agent

Two factors impact the run-time performance of this model. First, an uneven distribution of persons among ranks results in longer run-times. With uneven workloads, ranks with relatively lesser loads (i.e., those with fewer people for which to run the disease progression model) would be idle while waiting on ranks with greater loads to finish processing. Second, an MPI-related overhead in transferring persons between ranks (including each person's state) causes increased run-times.

An efficient parallel implementation of CityCOVID requires us to balance the rank load and inter-rank movement. In other wwords, the algorithm for assigning places to ranks in our synthetic population must account for an even distribution of agents among the ranks while minimizing the movement of agents between these ranks. We define a *load balancing* algorithm $f\{\mathbf{L}, \mathbf{A}, \mathbf{P}\} : \mathbf{L} \rightarrow n$, where each unique place in $\mathbf{L}$ is assigned to a specific computing rank $n \in N$, based on the synthetic population ($\mathbf{P}$) and their activities ($\mathbf{A}$). The total number of ranks N is set based on the model's scaling characteristics on the specific

HPC resource (Collier et al., 2015).

4.5.1 Graph Partitioning Problem

We map the load balancing objective to a graph partitioning problem. In this graph, each place in $\mathbf{L}$ is represented as a vertex, and edges connect places between which agents can move according to their assigned activity profiles. Both vertices and edges have weights, each encapsulating one of the two factors affecting load balancing and run-time performance.

Vertex weights quantify the corresponding place's agent load, and edge weights quantify agent movement between pairs of places. The load balancing objective is to partition the graph into N parts (corresponding to the N ranks) such that vertex weights are evenly distributed between parts, and the sum of the weights of edges crossing two distinct parts is minimized. Figure 4.7 shows a representative example of vertex and edge weights for two connected ranks.

Computing weights

Each agent is assigned ten possible weekday and weekend schedules, matched across different census and survey-based data sets by agents' demographic profile (Kaligotla et al., 2020b). These schedules consist of hourly activities at one of $m \subset \mathbf{L}$ possible places. For each hour of each possible schedule $k \subset \mathbf{A}$, we extract the vertices $v^k = \{v_0^k, v_1^k, ..., v_m^k\}$ corresponding to the m possible places an agent could be. Then, for each $v_i^k \in v^k$, we increment the weight of v_i^k by d_s/m where $d_s = 5$ if k is a weekday schedule and $d_s = 2$ if k is a weekend schedule. Then, to compute edge weights, we extract the list $u^k = \{u_0^k, u_1^k, ..., u_{m'}^k\}$ of $m' \subset \mathbf{L}$ vertices corresponding to the places an agent has associated with the activity type for the subsequent hour, and we increment the weight of each edge (v^k, u^k) by $d_s/(m \times m')$. Figure 4.8 shows a sample subgraph of the weighted network generated through this process for 3163 vertices and 3573 edges. The graph generated for the entire Chicago synthetic population

has approximately 1.38 million vertices and 167 million edges.

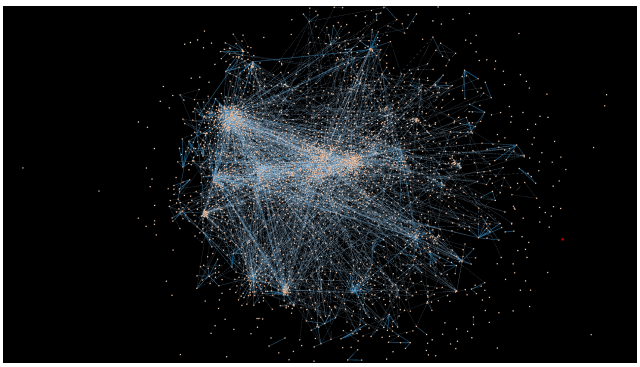


Figure 4.8: Sample subgraph network (3163 vertices and 3573 edges) for a subset of places.Vertex and edge colors represent weights, with lighter colors signifying larger weights.

We use the METIS graph partitioning software as part of our load balancing strategy (Collier et al., 2015; Karypis and Kumar, 1998). METIS is a set of applications for partitioning graphs and finite element meshes; specifically, the incorporated *gpmetis* tool splits weighted vertices into parts while minimizing the cross-part communication cost, making it suitable for balancing our places among ranks. Using *gpmetis*, we partition the graph into $N = 256$ ranks, and assign ranks to each place accordingly.

4.5.2 Load Balancing Experiments

While the graph partitioning is automated through METIS, the question that remains to be addressed is how to best construct the graph itself – that is, whether different vertex and edge weighting methods affect load balancing, and by extension, CityCOVID run-time performance.

We designed several load balancing schemes with varying edge and vertex weighting methods to investigate their effect on run-time performance. We then load balanced our synthetic population for each scheme detailed in Table 4.6 (resulting in different sets of place-rank assignment), ran the CityCOVID simulation on HPC resources for each scheme, and measured the run-time performance. Table 4.6 also indicates which HPC cluster is used

for each scheme.

Table 4.6: Weighting Schemes for Load Balancing Experiments

Exp#	Scheme Description	HPC
1	24h vertex weighting + No edge weighting	Bebop
2A	24h vertex weighting + 24h edge weighting	Bebop
2B	Daytime vertex weighting + 24h edge weighting	Bebop
2B'	Daytime vertex weighting + 24h edge weighting	Theta
3	No vertex weighting + 24h edge weighting	Bebop
4	Hierarchical Scheme with vertex weighting	Theta
5	Hierarchical Scheme with vertex + edge weighting	Theta

Scheme 1 uses only vertex weights so that the partitioning ignores cross-rank movement. In contrast, Scheme 3 uses only edge weights, thereby minimizing only the cross-rank movement costs. Scheme 2A uses both vertex and edge weights over a 24 hr day. Scheme 2B is similar but only considers vertex weights for daytime agent activity (between 9 am and 6 pm in each schedule) to prioritize the busiest portion of agent schedules. Scheme 2B' is similar to 2B but is run on a different HPC cluster. Finally, we generated hierarchical load balancing schemes (4 and 5), which consisted of partitioning the graph with METIS on two levels - nodes, and ranks per node. CityCOVID was distributed across four computing nodes (with either 36 cores [Bebop] or 64 cores [Theta] per node). We first split the graph into four parts to distribute across nodes and then partitioned the resulting subgraphs following the specified schemes. We used schemes 4 and 5 to determine whether cross-node movement disproportionately contributes to the total computational cost, compared to cross-rank movement.

The scale of generated co-location networks in CityCOVID is larger than previous work (Macal et al., 2014a), which used vertex and edge partitioning procedures (scheme $2A$) to balance the co-location network. CityCOVID thus represents an extended case study to

investigate the performance of different graph partitioning schemes, providing a basis for further work to generalize load balancing algorithms in large-scale ABMs.

4.5.3 Performance of Load Balancing Experiments

Each of the schemes in Table 4.6 was run on the indicated HPC resource for a full simulated year ($T = 8736$) on an identical synthetic population. The runs were done on the Bebop cluster and the Theta supercomputer, both hosted at Argonne National Laboratory. Bebop is a 1024 node cluster with two different types of compute nodes. We used the Intel Broadwell nodes, which have 36 cores and 96GB of memory per node. Theta has 4392 Intel Knights Landing computer nodes with 64 cores and 192GB per node. For each HPC run, the following metrics of interest were recorded at each time step t and rank $n \in N$ to compare run-time performance: the total number of persons received from other ranks at a given rank (p_{recv}), the total number of persons on a given rank at at the end of the time step (p_{load}), and the computational run-time for each rank for each time step (r).

Comparing Run-Times Across Schemes

We compare the performance of our load balancing schemes based on two run-time metrics. The per-step run-time, $\rho(t)$, represents how long it takes to run a given time step t, and is calculated as $\rho(t) = \max_{\{\forall n \in N\}}(r_n^t)$. Figure 4.9 shows the distribution of $\rho(t)$ for each scheme in Table 4.6.

The total run-time metric, Γ, represents the total time to run CityCOVID for a year-long simulation and is calculated by summing the per-step run-time ($\rho(t)$) across all time steps, i.e., $\Gamma = \sum_{t=1}^{t=8736} \rho(t)$. The comparative performance for our experiments in terms of Γ (in seconds) are: 325 (Exp 1); 401 (Exp 2A); 378 (Exp 2B); 1260 (Exp 2B'); 1639 (Exp 3); 1705 (Exp 4); and 2018 (Exp 5). We see schemes 1, $2B$, and $2A$ have the lowest average $\rho(t)$ and Γ run-time metrics, while schemes $2B'$, 3, 4, and 5 perform comparatively

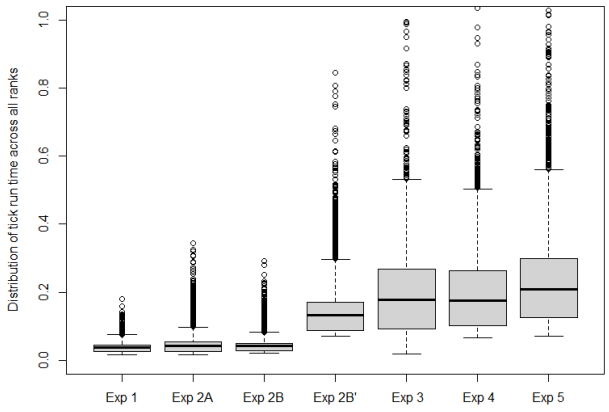


Figure 4.9: Distribution of per-step run-time $\rho(t)$ for schemes in Table 4.6.

worse. Figure 4.9 indicates that scheme 1 has a smaller $\rho(t)$ IQR range than $2B$ and $2A$, indicating consistently better performance across all time steps and ranks, making it our best load balancing scheme. Comparing schemes $1, 2A, 2B$, and 3, we notice that schemes considering vertex weighting performed considerably better than when only edge weighting was considered. This observation leads us to take a closer look at the relationship between rank load (p_{load}) and cross-rank movement (p_{recv}) on run-times (r). Figure 4.12 illustrates this comparison at each time step t and rank n for our four best-performing schemes.

From Figure 4.12, we observe that a) greater rank load alone does not imply poorer performance - scheme 1 and $2B$ show a cluster of high-load ranks compared to $2A$, and b) greater cross-rank movement (in $2A$ and $2B$, relative to 1) does incur a performance penalty, albeit small. Together with our comparison between vertex and edge weighting schemes, these observations indicate that cross-rank movement in the presence of vertex-weighting is relatively more costly than rank-load in terms of performance. Characterizing the relationship between rank load and cross-rank movement alongside vertex weighting is a promising starting point towards developing a generalized load balancing model to predict and optimize simulation run-times.

A comparison of run-time performance metrics for schemes 1, $2A$, $2B$, $2B'$, and 3, against

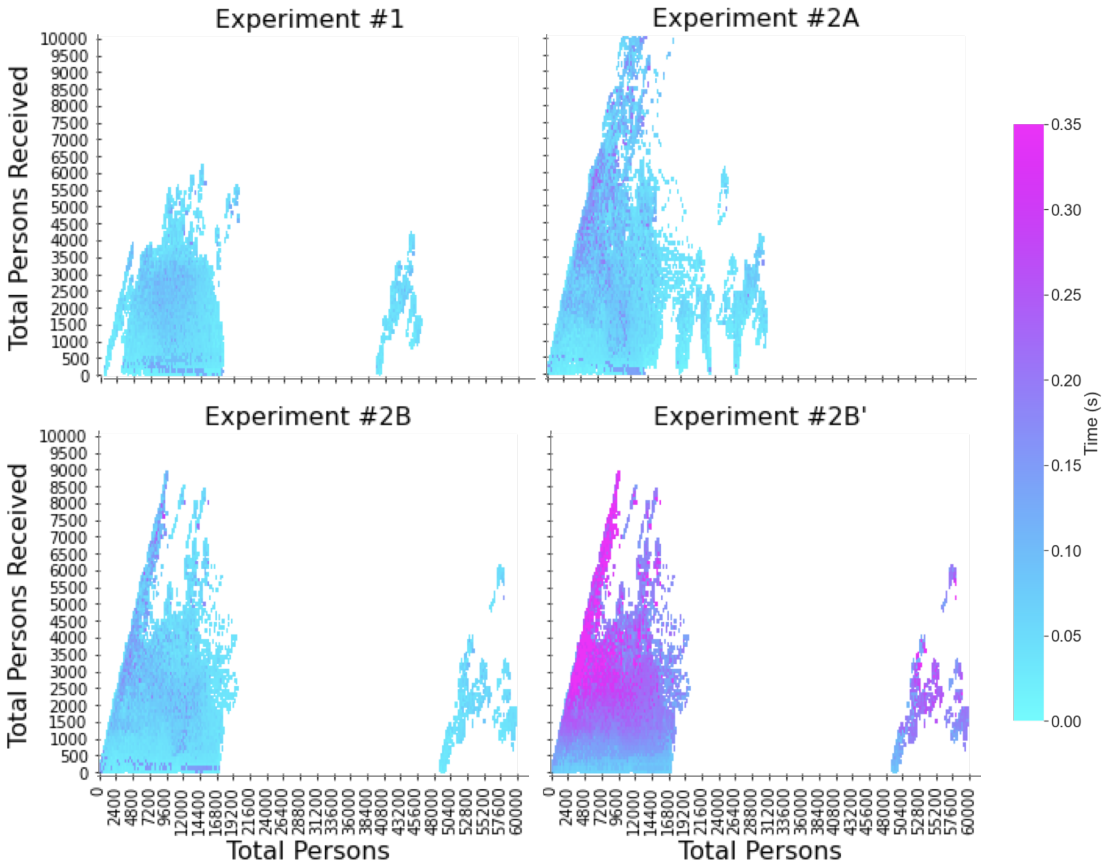


Figure 4.10: Run-time by total persons (p_{load}) and total persons received (p_{recv}) for schemes $\{1, 2A, 2B\}$ (all run on Bebop for consistent comparison) and $2B'$ (to show differences between HPC resources). Results are compiled across all simulation ticks and ranks, and plots are truncated to only show $p_{load} \leq 60,000$ and $p_{recv} \leq 10,000$ for consistency.

hierarchical schemes 4 and 5, indicates that, at least for the message passing patterns within CityCOVID, cross-node movement does not disproportionately contribute to overall performance cost compared to cross-rank movement. Investigations into how this result translates to more general ABM message passing patterns are topics of future work.

For completeness, we include a comparison of the same scheme across different HPC resources, where we utilize the same number of ranks for each (256) to aid in the comparison. We see in Figure 4.12 that scheme $2B$, run on Bebop, demonstrates different run-time characteristics, and generally outperforms $2B'$, run on Theta. Factors affecting the outcome include differences in CPU performance, cores per CPU, network infrastructure, and MPI

implementations. HPC resource-specific details have the potential to significantly affect overall performance, indicating that tunable parameters are likely to be needed for any generalized load balancing scheme. Future work will focus on characterizing specific HPC infrastructure variables on run-time performance.

4.6 Validation of synthetic population

With so many datasets and algorithms required to develop a synthetic population, it is imperative that we are able to easily verify the accuracy of the population at all stages of development. To do so, we built an interactive analytics suite to generate analyses and visualizations to be shared with domain experts for validation. In addition to verifying that the population demographics matched known regional statistics, we needed to ensure that the activity schedules and place assignments were reasonable, as these are what determine contact between agents and ultimately disease transmission. Figre 4.11 provides some examples of the interactive analytics built for synthetic pouplation validation; the interactivity enables easy comparison across a wide variety of axes. For example - we see that, on average, men in the 25-34 age group are spending more time at work than their female counterparts.

Easy generation of these types of visuazliation allowed our team to quickly debug make necessary updates when something was flagged as unexpected, lending confidence to the accuracy of the synthetic population.

4.7 CityCOVID model calibration

The synthetic population described in this chapter is the primary input to the CityCOVID ABM. Each agent in the synthetic population has an individualized COVID-19 disease progression pathway (see Figure 4.1 for the general framework) based on heterogeneous agent attributes, exposure through co-location over time with infected individuals, and external

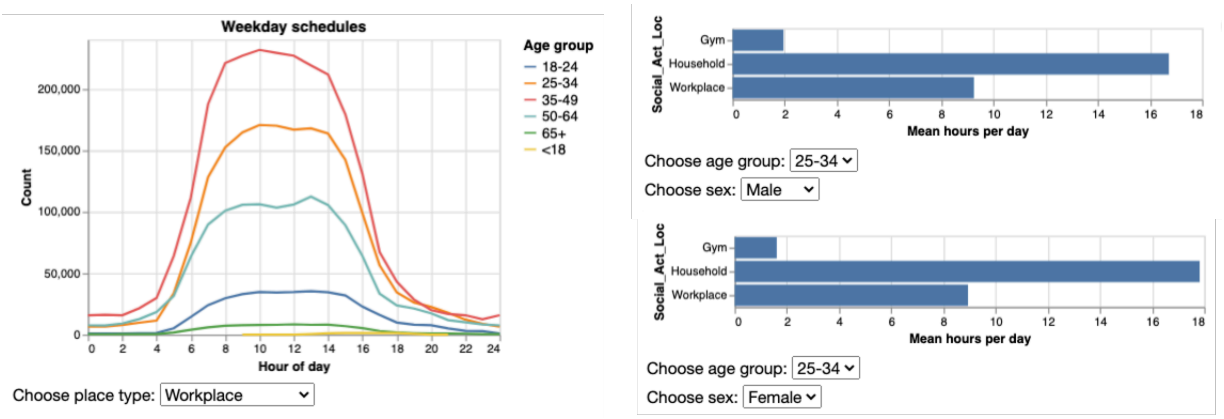


Figure 4.11: Example of the interactive analytics tools built to validate the synthetic population. On the left, we see the agent gradient of work activities (note that the drop-down enables generation of this plot for any place type), and on the right we see, on average, how long agents spend in different place types on weekdays, with the ability to filter by age and gender.

factors such as seasonality of viral transmissibility. Agent attributes are fixed by the synthetic population and exposures are dictated by activity schedules and place assignments, but the external factors such as the number of initial exposures, transmission probability, and seasonality are model parameters that need to be tuned.

Calibrating model parameters requires repeatedly running the CityCOVID simulation and comparing the output with empirical hospitalization and deaths data for Chicago, both of which are readily available on the City of Chicago data portal[22]. In all, CityCOVID needs to tune 9 parameters, and which is extremely computationally expensive. In recent years, ML-based approaches have been developed to speed the calibration process, and as part of the broader ecosystem of this work, the Extreme-scale Model Exploration with Swift/T (EMEWS) (Ozik et al., 2016) framework was developed to efficiently implement these systems. For CityCOVID, the model calibration process employed an approximate Bayesian computation (ABC) framework to more efficiently calibrate model parameters (Ozik et al., 2021b), which enable generating simulated posterior distributions of the model input pa-

22. `https://data.cityofchicago.org/`

rameters (the priors are constructed from the scientific literature), and as a result we can generate posterior model outputs that efficiently propagate model uncertainties (Beaumont, 2019).

Once calibrated, CityCOVID is run with the optimal model parameters to simulate Chicago COVID-19 deaths and hospitalizations under a wide variety of experimental scenarios. Figure 4.12 shows the optimally calibrated forecasts for daily deaths and hospitalizations before and after the inclusion of SafeGraph place types. In both cases, we see that the median model prediction (black line) matches the empirical data (black dots) well, but the inclusion of SafeGraph places (hence a more realistically mobile synthetic population) decreases the uncertainty of the estimates, as evidenced by tighter confidence bands.

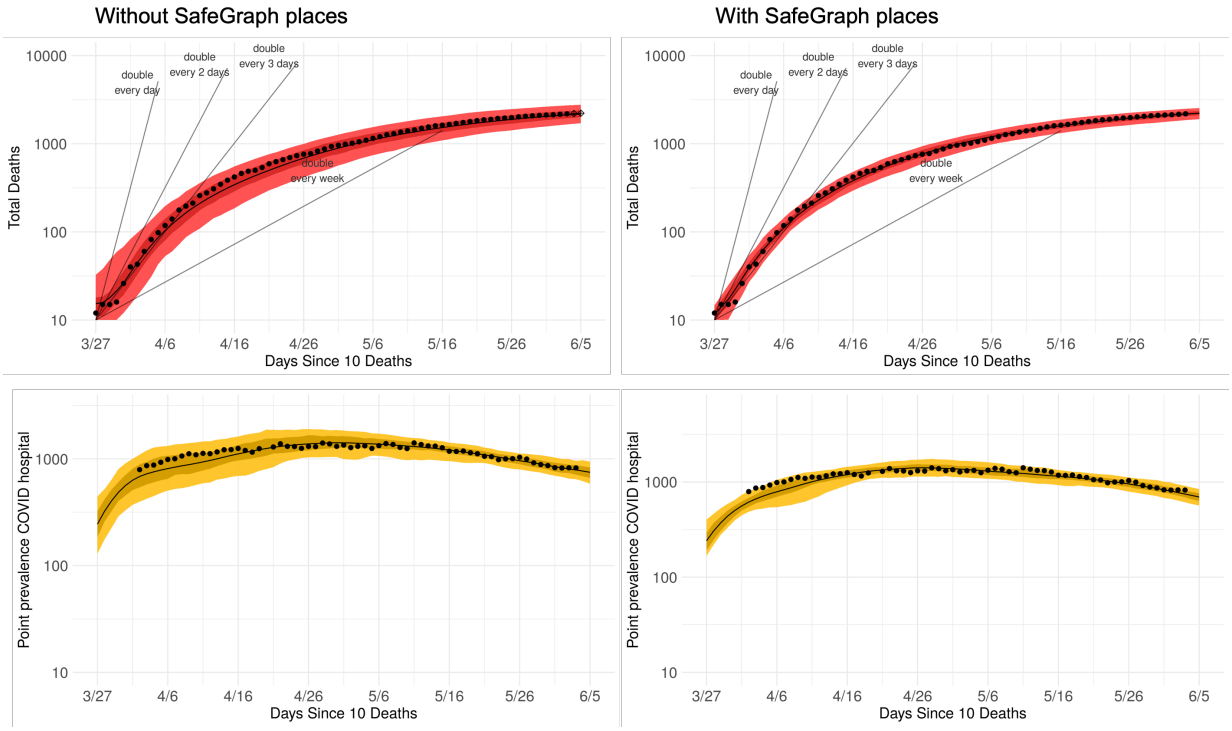


Figure 4.12: COVID-19 attributed hospitalization and death outputs from CityCOVID before and after the inclusion of SafeGraph places in the synthetic population. The black dots show the empirical Chicago data; the black line is the median simulation output, and the shading represents confidence bands (50% simulation bands in the dark region and 95% in the lighter region.)

4.8 Experimentation with synthetic populations and agent-based models

With the highly granular synthetic population built and CityCOVID calibrated, we can run a wide variety of experiments. One important use-case for CityCOVID is studying the impact of social determinants of health (SDOH) on COVID-19 transmission and outcomes, and the additional updates to the core synthetic population (Section 4.4) were made in support of this. The COVID-19 pandemic has highlighted drastic health inequities, particularly in cities such as Chicago, Detroit, New Orleans, and New York City. Reducing COVID-19 morbidity and mortality will likely require an increased focus on SDOH, given their disproportionate impact on populations most heavily affected by COVID-19. A better understanding of how factors such as household income, housing location, health care access, and incarceration contribute to COVID-19 transmission and mortality is needed to inform policies around social distancing and testing and vaccination scale-up.

4.8.1 Contact matrices for occupational risk and racial disparities

With an eye toward understanding the observed racial/ethnic disparities in COVID-19 outcomes, we hypothesize that occupational risk may be partially true to blame. This may be due to an overrepresentation of Blacks and Hispanics in the essential workforce, service industry and public facing jobs that offer limited opportunities to work from home, resulting in more mobility and more potential for exposure at work.

The first part of this analysis seeks to understand the patterns of interactions between different subgroups of our population. We do this by generating *co-location* or *contact matrices*, which enumerate the contacts (agents in the same place at the same time) between pre-specified demographic groups. These matrices are often used to understand population mobility dynamics (Iozzi et al., 2010a) and can be generated without fully deploying the

ABM. Contact matrices have long been an important piece of the epidemiological modeling process.

Pseudo-simulation

Contact matrices can be directly generated using the synthetic population underlying CityCOVID (in contrast to fully running CityCOVID on an HPC). We call this a *pseudo-simulation* to differentiate it from the full CityCOVID simulation; it is able to estimate contact patterns but not disease transmission. The psuedo-simulation consists of randomly sending agents to one of their assigned locations for each hour of each day of week. We then keep track of who was at each location during that time in order to estimate contact matrices. See Algorithm 6 for details.

Algorithm 6 Weekly pseudo-simulation

```
for p_i ∈ P do
    for day d = 1, 2, ..., 7 do
        if day ≤ 5 then
            Randomly select a_i ∈ A_weekday
        end if
        if day ∈ {6, 7} then
            Randomly select a_i ∈ A_weekend
        end if
        for hour h = 1, 2, ..., 24 do
            Randomly choose location l from a_i[h]
            Record that agent p_i was at location l during hour h on day d
        end for
    end for
end for
```

Our contact matrices are generated from the output of the pseudo-simulation. Following Iozzi et al. (2010b), we define two agents to be in contact with one another if they are in the same place at the same time (here, that means within the same hour). Typically, contact matrices are generated by investigating the patterns of contacts between different age groups. Due to the highly granular nature of the synthetic population underlying CityCOVID, we

are able to generate contact matrices between a variety of different sub-populations (e.g. race, sex, income levels, etc).

From the pseudo-simulation, we know exactly who was at each location during a given hour of day and which sub-populations they belong to. For each location, we want to compute, on average, how often each group within the sub-population was in contact with every other sub-population group. For the sake of exposition, suppose there are 10 agents sharing a location during one day/hour; 4 belong to group A and 6 belong to group B. Each agent in group A encounters 3 other agents in group A and 6 other agents in group B. Summing across agents, there are $4 \times 3 = 12$ contacts in total among group A and $4 \times 6 = 24$ between group A and group B.

Recall that each agent $p_i \in \mathbf{P}$ is represented by a set of m socio-demographic characteristics $p_i = \{\text{Age}_i, \text{Race-Ethnicity}_i \ldots, \text{Sex}_i\}$. See Table 4.1 for all possible characteristics and their associated categories. Let S represent a subset of these characteristics (e.g. $S = \{\text{Age}\}$ or $S = \{\text{Age}, \text{Race-Ethnicity}\}$) and $S_i \in S$ be agent p_i's realization of those categories (e.g. $S_i = \{\text{Young}\}$ or $S_i = \{\text{Young}, \text{Black}\}$). We are interested in quantifying contacts between all possible realizations of S (i.e. the Cartesian product of the sets of characteristics in S). Let C denote the number of possible realizations of S.

Following Klepac et al. (2020), we define our contact matrices $M \in \mathbb{R}^{C \times C}$ such that

$$M_{A,B} = \frac{\text{Number of contacts between group A and group B}}{\text{Total population of group A}}$$

This value represents the average number of contacts an agent in group A has with agents in group B on average. Because contact patterns are often computed by age, we first validate our contact networks by investigating age group mixing patterns overall, and then separated by household and non-household contacts (top row of Figure 4.13). These are expected patterns based on related literature (see e.g., Iozzi et al. (2010a)); off-diagonal age contacts within a household are representative of parent/child interactions, while out-of-household

interactions more commonly occur within similar age groups. The bottom row of Figure 4.13 shows the interactions between "workers" and "visitors" at essential workplaces as an attempt to investigate the hypothesis that occupational risk influence disparities in COVID-19 outcomes. The age patterns are expected; essential workers are likely to be in the 20-50 age range, while visitors are more evenly distributed. In general we see that workers are in a lower income bracket than visitors, again not surprising. The asymmetry of the race matrix provides insight into workplace-based racial disparities; we see that Black workers interact with White visitors more often than White workers interacting with Black visitors.

Algorithm 7 Contact counting

for $l_i \in \mathbf{L}$ **do**
  for day $d = 1, 2, \ldots, 7$ **do**
    for hour $h = 1, 2, \ldots, 24$ **do**
      $M_{i,i} += \binom{|G_i|}{2}$
      $M_{i,j} += |G_i||G_j|$
    end for
  end for
end for

4.8.2 Impact of changes in protective behaviors and out-of-household activities by age on COVID-19

A wide variety of scenarios can be explored using the CityCOVID framework. While we have many experiments in the pipeline, results from an experiment varying different levels of out-of-household activities are particularly illuminating (Hotton et al., 2022). Even with the vaccine widely administers and largely effective, protective behaviors like social distancing and masking remain an important public health tool for mitigating the spread of COVID-19. Using CityCOVID, we are able to vary the level of protective behavior and out-of-household activities (OOHA) relative to the calibrated model and quantify the changes in COVID-19 outcomes under different behavioral scenarios. We find that decreasing adult protective

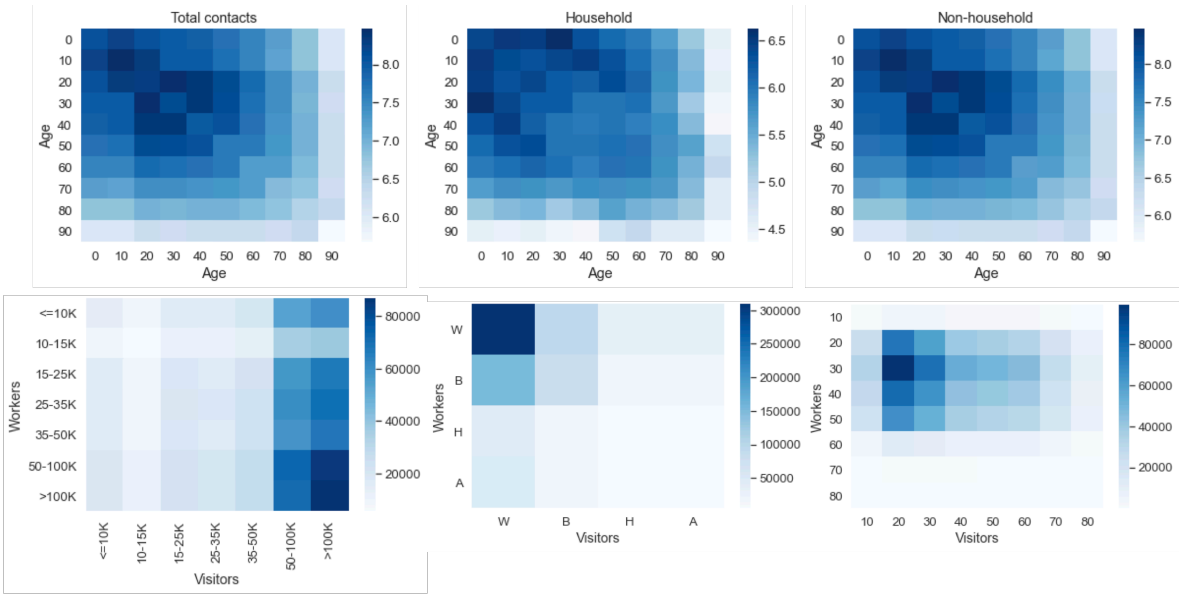


Figure 4.13: Sampling of contact matrices generated from the pseudo-simulation. Top: average daily contacts between age groups overall, within-household, and outside of household. Bottom: investigation of contact patterns between workers and visitors at essential workplaces for different demographic groups.

behaviors and increasing adult OOHA both substantially impacted COVID-19 outcomes; school reopening had relatively little impact when adult protective behaviors and OOHA were maintained. As of January 1, 2021, a 50% reduction in young adult (age 18-40) protective behaviors resulted in increased latent infection prevalence per 100,000 from 3.39 to 28.32 and 5.51 to 32.9 with 15% and 45% school reopening. Increasing adult (age $\geq$ 18) OOHA from 65% to 80% of pre-pandemic levels resulted in increased latent infection prevalence per 100,000 from 18.06 to 93.14 and 20.36 to 117.06 with 15% and 45% school reopening. Similar patterns were observed for hospitalizations. Figure 4.14 shows the hospitalizations and deaths predicted by CityCOVID under different OOHA and school reopening levels.

4.9 Challenges and future work

CityCOVID is a massive undertaking with many stakeholders and moving pieces. The scale of the model is unprecedented and requires specialized implementations and oversight to

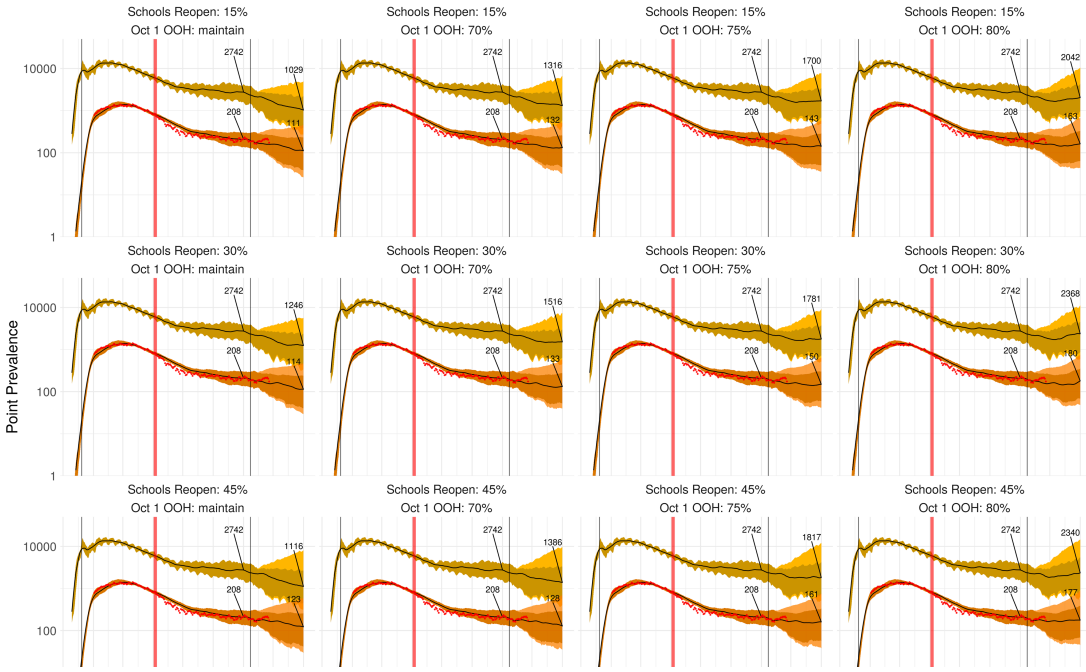


Figure 4.14: Impact of increases in adult out-of-household activities (OOHA) on COVID-19 infections under various school reopening scenarios. Vertically from top to bottom, effect of increasing school reopening for a given level of adult OOHA. Yellow plots indicate point prevalence of latent infections and red plots indicate point prevalence of hospitalizations. Horizontally from left to right, effect of increasing adult OOHA (65%, 70%, 75%, and 80% of pre-pandemic levels) or a given level of school reopening, from March 2020 to November 2020.

run; as such, any downstream updates made to the synthetic population can propagate unexpected outcomes in a full model run, and with each run a significant time and resource committment, it is important that each update is intentional and accurate. Due to the resource constraints inherent to this work, setting up a new experimental framework to answer additional question with the full CityCOVID model is a lengthy process involving many steps, and so we found that, in practice, the psuedo-simulation framework was able to act as a reasonable proxy and still leverage most of the potential of the synthetic population.

That said - with new COVID-19 variants always emerging, mandates for protective behavior changing, and vaccine efficacy dwindling, running experiments with CityCOVID remains relevant. When resources free up, we plan to fully deploy SDOH experiments on CityCOVID

to validate findings from the contact matrices. We are also interested in studying vaccine distribution, uptake, and efficiacy, as well as a more concerted effort to understand the interplay between SDOH, chronic health conditions, and COVID-19 outcomes.

In this work, we've shown how using statistical methods to use micro-level population data to build large-scale simulations is an effective tool for modeling epidemics like COVID-19. We have built a generalizable and flexible synthetic population development framework that is significantly more granular than other similar populations, enabling a deep investigation of heterogeneity in disease transmission. This is an important step toward building more accurate, interpretable models to understand highly complex systems.

CHAPTER 5

CONCLUSIONS

Anchored by the idea that a prediction made without context is not only unhelpful but ultimately dangerous when modeling systems that impact society, my thesis is an effort toward making data science more interpretable, reliable, and application-driven.

The first section of this thesis discussed the use of statistics and machine learning in climate science, particularly around making seasonal forecasts. In this setting, we were faced with a short observational record, a high-dimensional dataset, and complex dependencies among the covariates. However, rather than attempting to use an uninterpretable but potentially more predictive model, we developed a suitable linear model that was not only predictive but also easily understood by climate scientists. In doing so, we uncovered further challenges, such as nonstationarities, and opportunities, such as leveraging climate model simulations to augment the available data. By carefully listening to our collaborators rather than attempting to use the most exciting new tools at our disposable, we developed a solution that was well-received, easy to implement, and can be built upon for future work.

Next, I developed a new method to make neural networks, the textbook definition of "black box" models, easier to understand. Using an established measure of variable importance (VI) that is prohibitively computationally burdensome for large neural networks, we first show that a common approximation called *dropout* is problematic under certain settings. We then develop a method that uses a linear approximation of the neural network around a specific initialization to approximate the VI measure in a significantly more efficient manner. We provide theoretical guarantees for our method, which is rare for these types of approximations and enable us to provide confidence intervals for our estimates. We demonstrate through simulations and real data that our method is fast and accurate, and argue that it is flexible enough to extend to a variety of other settings.

The final section in my thesis is quite different from the first two; rather than working with

statistical prediction models, I focus on an agent-based simulation model for understanding the transmission and outcomes of COVID-19 in Chicago. My specific contribution was the development of a robust and highly granular synthetic population that enable models to capture population heterogeneity to better understand the social structures guiding macro-level outcomes, something that has been widely observed but is difficult to model. We develop a wide variety of tools to synthesis different datasets so that, together, they statistically represent the population of Chicago along with its mobility patterns. We then show that we can use this highly detailed population to understand the disparate impact of COVID-19 on different populations, and test the impact of non-pharmaceutical interventions. An agent-based model is highly interpretable, in that we can understand the specific population traits that lead to different outcomes, and our model has been widely used by governmental decision-makers to inform COVID-19 policymaking.

While specific future directions have been noted throughout the thesis, I want to emphasize here that we are still in the very early stages of making statistics and machine learning widely applicable to climate science, public health, and other application areas with significant human impacts. The dangers of unchecked algorithmic decision-making are widely documented, and continued efforts to make predictive models more transparent and reliable should be at the center of every machine learning application that materially affects peoples' lives. My intention in writing this thesis is to shed light on both the possibilities and limitations of the use of data science in systems that impact society, and I am eager to continue to work toward building responsible, transparent, and reliable data systems for the things that matter most.

APPENDIX A

ADDITIONAL TOPICS IN SEASONAL FORECASTING

A.1 MultiGTV

We know that precipitation patterns across the entire SWUS are tied to similar summer atmospheric events, but possibly to a different degree for each region. To that end, rather than treating individual regions (or their weighted average) as independent prediction problems, we extended our method to a multitask setting that retains the GTV regularizer for within-region structure and makes the additional assumption that there is an unknown subset of covariates that are relevant for prediction, and this subset is preserved across the m regions of interest.

Let $Y = [y^{(1)}, y^{(2)}, \dots y^{(m)}] \in \mathbb{R}^{n \times m}$ be the matrix of the m related response vectors and $B = [\beta^{(1)}, \beta^{(2)}, \dots \beta^{(m)}] \in \mathbb{R}^{p \times m}$ be the matrix of the corresponding m coefficient vectors. The multitask extension of GTV, which we refer to as *MultiGTV*, is defined by the objective

$$\hat{B} = \arg\min_{B} \left\{ \|Y - XB\|_F^2 + \lambda_1 \sum_{r=1}^{m} \|\Gamma \beta^{(r)}\|_1 + \lambda_2 \sum_{j=1}^{p} \|\beta_j^{(:)}\|_2 \right\} \tag{A.1}$$

where Γ is the edge-incidence matrix from (2.5). The first regularization term promotes similarity of coefficients for highly correlated predictors within each region, and the second regularization term promotes shared support across regions.

Note that (A.1) is convex but non-differentiable and non-separable. The first regularization term is a sum of non-differentiable terms over the columns of B, while the second is the sum of non-differentiable terms over the rows. Standard gradient descent methods will not work in this setting, so we develop a new algorithm that incorporates ideas from proximal gradient methods and ADMM to solve this problem (Algorithm 8).

Algorithm 8 MultiGTV

Choose $B_0, \lambda_1, \lambda_2$
for $k = 1, 2, \ldots$ **do**
  for $r = 1, 2, \ldots, m$ **do**
    $\beta_k^{(r)} = B_{k,r}$
    $z_k^{(r)} \leftarrow \beta_k^{(r)} + \eta X^T \left(y^{(r)} - X\beta_k^{(r)} \right)$ ▷ Gradient descent step
    $w_k^{(r)} \leftarrow \arg\min_\beta \|\beta - w_k^{(r)}\| + \eta\lambda_1 \|D\beta\|_1$ ▷ Within-region GTV
  end for
  $W_k = \left[w_k^{(1)}, w_k^{(2)}, \ldots, w_k^{(m)} \right]$
  $B_{k+1} = W_k$
  for $j = 1, 2, \ldots, p$ **do** ▷ Group soft-thresholding
    $r_j = W_k - \sum_{k \neq j} e_k \beta_k^T$
    $B_{k+1}^{(j)} = \left(1 - \frac{\eta\lambda_2}{\|e_j^T r_j\|_2} \right)_+ e_j^T r_j$
  end for
end for

A.2 Accounting for non-stationarities

It is abundantly clear in the climate literature that anthropogenic forcings in the climate system (global warming) has a significant impact on the Earth's coupled ocean-atmosphere-land dynamics (Mamalakis and Foufoula-Georgiou, 2018). In the data we have studied so far, we see a clear shift in the mean trend of SSTs before and after the 1970s, the inter-annual variance of SWUS precipitation has increased, and the strength of established teleconnections (large scale climate anomalies that are linked to one another) have weakened over time (Johnson et al., 2019). That is - there are complicated trends in each of y, X, and β from (2.4) that might be impacting our ability to effectively apply statistical and machine learning methods.

In our published work (Stevens et al., 2021), we apply more weight to more recent observations in our training period in the data-fit term, a common practice (Hyndman and Athanasopoulos, 2018). We may find greater success with models in which regression coefficients may change as a function of time, i.e., for $t = 1, \ldots n$,

$$y_t = X_t \beta_t^* + \epsilon_t$$

Understanding the best way to do this in a seasonal forecasting setting is non-trivial, as the underlying regression problem is already high-dimensional ($p \gg n$) and the columns of X are highly correlated. Allowing our coefficients to vary with time add additional complexity to an already highly underdetermined and complex system.

To demonstrate the necessity of accounting for the nonstationarity of the climate system in our predictive procedures, we briefly explore its impact on the problem setting from the above section (i.e. forecasting winter precipitation over the southwestern US using sea-surface temperature (SST) over the Pacific basin from the previous summer). In this problem setting, we considered 900 covariates (monthly (July-Oct) SST on a $10^\circ \times 10^\circ$ spatial grid) and have just 80 years worth of observational data.

For this demonstration, however, we consider just 2 covariates: averaged summer SSTs over the NZI region and the Nino 3.4 region, both of which are climate indices known to be predictive. To understand in what sense the dynamics are changing, we perform a number of analyses.

A.2.1 Coefficients

First, we investigate how and if the regression coefficients change over time. We do this through two different methods:

1. A growing window approach: starting with a 20-year window, simple linear regression is used to find the regression coefficient for each of the two climate indices. Then, we increase the window by one year, refit the model, and compute the new coefficient. This is repeated until the model is fit on all available data.

   The coefficients estimated using this approach are as follows (note we show the absolute

value of the NZI coefficient for easier comparison). We also plot the coefficients with their standard errors (separated so the scales are more informative) in Figure A.1.

Figure A.1: (Top) Regression coefficients for the growing window approach (Bottom) Coefficients with 95% included (rescaled)

2. A sliding window approach: beginning with the 1940-1969 time period, we compute regression coefficients for a sliding window of 30 years for every window through 1990-2019. Again, we plot the coefficients and their standard errors in Figure A.2.

In both cases, we certainly see some temporal dependence of the regression coefficients, further confirmed by looking at their standard errors.

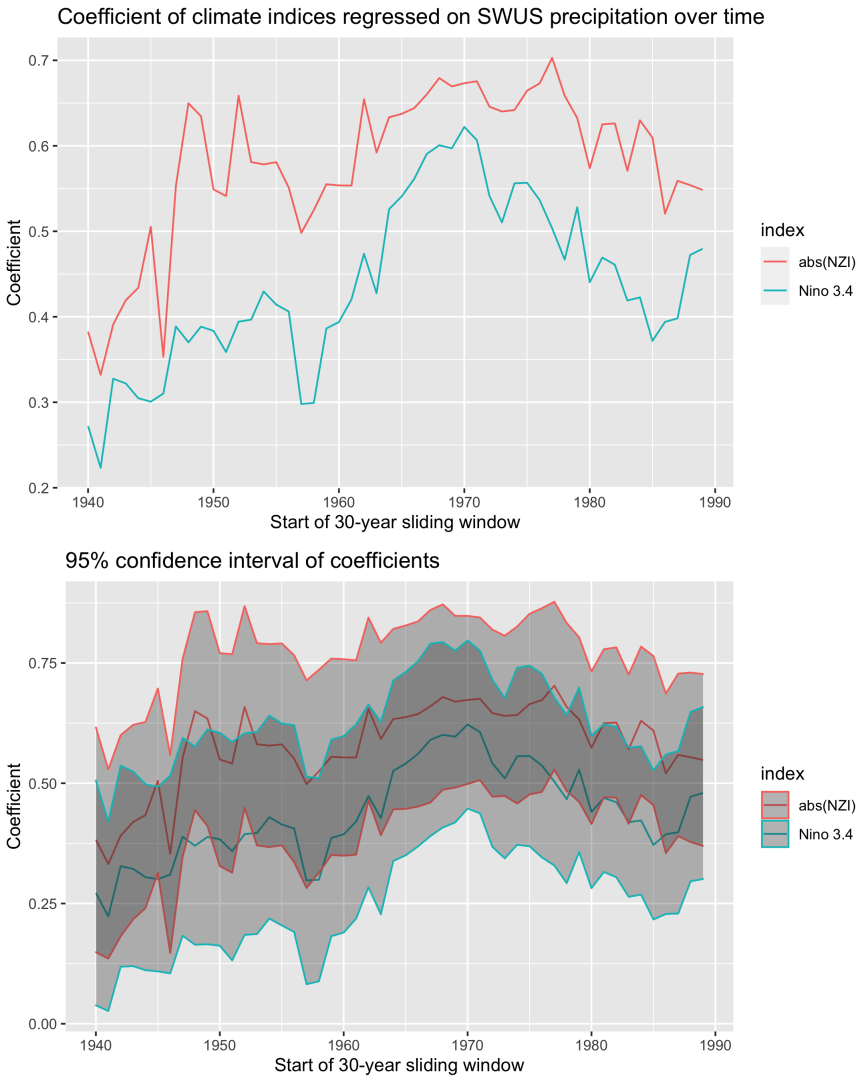


Figure A.2: (Top) Regression coefficients for the sliding approach (Bottom) Coefficients with 95% included (rescaled)

A.2.2 Prediction

Next, we investigate the impact of time on these models' predictive ability. Following a similar procedure as above, we perform two analyses:

1. We fit a "growing" training period, starting with 20 years and increasing year by year, and compute the R^2 value on the data from the 20 years following the training period.

2. We fit a "sliding" training period, where we train the models on 30 years of data,

compute R^2 on the next 20, and then slide the entire procedure up one year.

The results of these analyses are shown in Figure A.3.

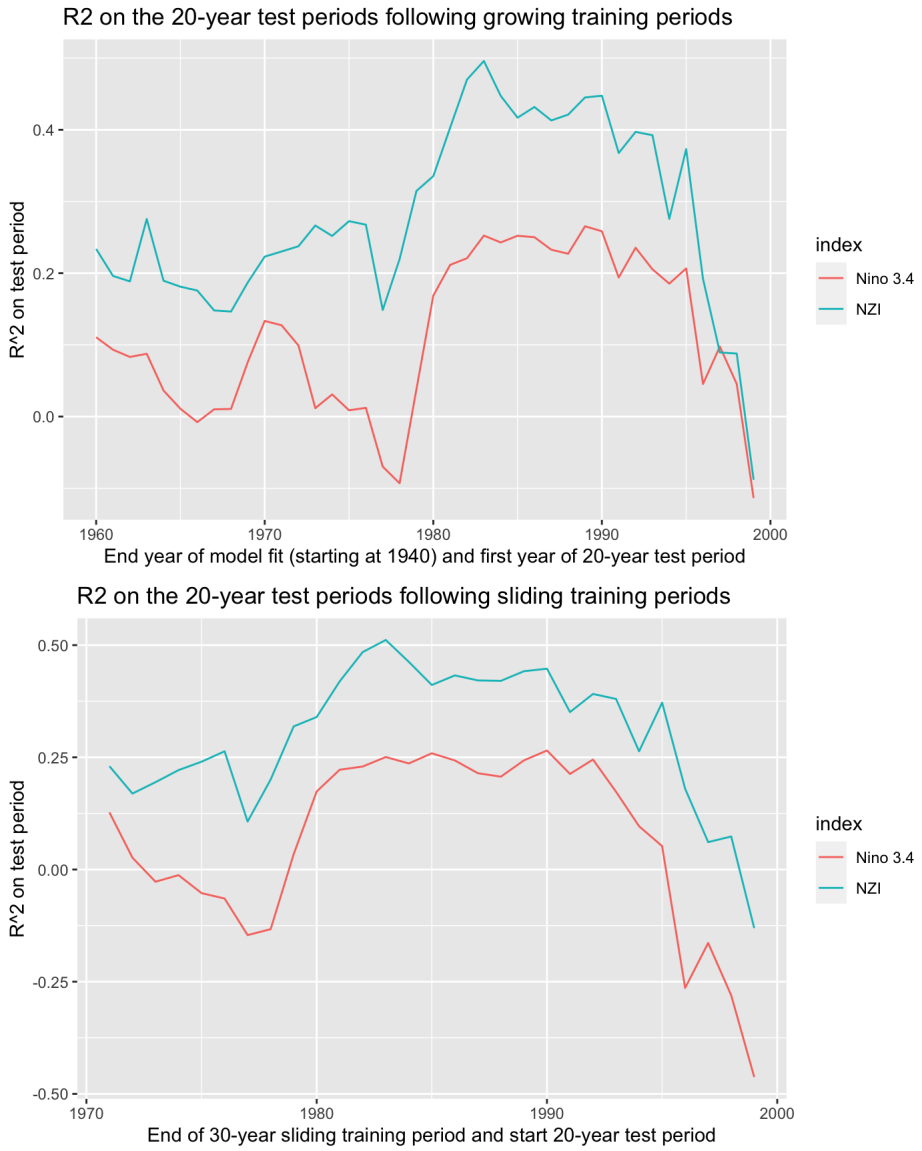


Figure A.3: (Top) Test R^2 for the "growing" training period scenario from 1960 - 2000. (Bottom) Test R^2 for the "sliding" training period scenario from 1970 - 2000.

In both settings, the predictive performance is highly dependent on time. We see a stark decline over the past 10-20 years in predictability using these indices, corresponding somewhat to the patterns seen in the estimated coefficients.

This analysis suggests that we are dealing with coefficients that are in some way time-dependent - in this note we explore a few approaches in trying to extend high-dimensional predictive methods to a time-dependent setting. However, it is also worth considering that there is nonstationarity coming from either the predictors or the response - this would assume either covariate shift or the need for some type of auto-regressive regime. We will also briefly explore some possibilities in that domain that could be applicable to our problem setting.

A.2.3 Time-dependent coefficients

In this section, we consider regression problems of the form

$$y_t = X_t \beta_t^* + \epsilon_t$$

where $\beta_t^* \in \mathbb{R}^p$ depends on time, reflecting the nonstationarity of the system. Estimating β_t^* without further structural assumptions would be impossible; in the climate data described above, this would require estimating $pn \approx 70000$ parameters with only 80 years of observations. Some structural assumption that might be reasonable to make are that the climate dynamics are changing slowly over time, or that there exist subsets of the data that can each be well-modeled by the same set of coefficients.

Fused Lasso

Under the assumption that climate dynamics are changing slowly over time, we can formulate this as a Fused Lasso Tibshirani et al. (2005) problem

$$\{\hat{\beta}_1, \ldots, \hat{\beta}_n\} = \text{argmin}_{\beta_1,\ldots,\beta_n} \left\{ \sum_{t=1}^{n} (y_t - X_t \beta_t)^2 + \lambda_1 \sum_{t=1}^{n} \|\beta_t\|_1 + \lambda_2 \sum_{t=2}^{n} \|\beta_t - \beta_{t-1}\|_1 \right\} \tag{A.2}$$

In our previous work, we used the Fused Lasso-like GTV term (Li et al., 2020) promote adherence of our estimator to underlying spatiotemporal structure of the covariates; here it is used to promote regression coefficients that adapt to significant changes in the underlying linear model. To solve this problem using existing tools, we rewrite A.2 as

$$\hat{\beta} = \text{argmin}_{\beta} \left\{ \|y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\Gamma\beta\|_1 \right\} \tag{A.3}$$

where

$$X = \begin{bmatrix} X_1 & 0 & \dots & 0 \\ 0 & X_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & X_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \text{and} \quad \Gamma = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & -1 & 1 \end{bmatrix}$$

Here, we have $X \in \mathbb{R}^{n \times pn}$, so in settings where $p \gg n$, the problem becomes quite ill-posed. There is an extensive body of literature on various extensions of A.3 with a variety of regularization terms. Degras (2019) specifically considers the problem of *model segmentation* in high-dimensional time series settings and proposes replacing the ℓ_1 regularization on the fused term with an ℓ_2-norm and calls it Sparse Group Fused Lasso. The author suggests that this formulation is more suitable for segmenting multivariate models than the ℓ_1-norm, which affects each of the p predictors separately and thus produces change points only shared by few predictor variables. The author further provides a helpful overview of related methods (pages 2-5); another closely related method is the Group Fused Lasso (Alaíz et al., 2013), which replaces the ℓ_1 norms with ℓ_2 norms and provide efficient iterative implementations, although their work focuses on image denoising rather than temporal smoothing.

A.3 Combining climate models and observations

While recent years have seen an explosion in the amount high-resolution climate data available for use in machine learning methods, the observational record is still very short - we have fewer than 100 years of reliable data for tasks such as seasonal-to-subseasonal forecasting. However, there is a large supply of output data from dynamical model simulations of many climatological processes. While these simulations are imperfect and contain many biases, we believe that they contain useful information that could improve models built using the observational record. Our published work shows that we can leverage CESM-LENS large ensembles to help impose structure on our estimated coefficients, but there is much to be done to better understand when and why including climate models in a machine learning system might help or hurt the prediction task.

Suppose we observe data $X_{obs} \in \mathbb{R}^{n \times p}$ and $y \in \mathbb{R}^n$, $p \gg n$. Further suppose that we have simulated data $X_{sim} \in \mathbb{R}^{m \times p}$ and $y_{sim} \in \mathbb{R}^m$, $m \asymp p$. We are interested in the conditions under which using the simulated data improves the estimation of β in the linear model

$$y_{obs} = X_{obs}\beta + \epsilon \qquad \epsilon \sim N(0, \sigma^2 I_n)$$

when we assume that

$$y_{sim} = X_{sim}(\beta + \Delta) + \epsilon_{sim} \qquad \epsilon_{sim} \sim N(0, \sigma^2_{sim}), \quad \Delta \sim N(0, \Sigma_\Delta)$$

That is - the regression coefficients for our simulated system are centered at β but shifted by some random Δ. Equivalently, we have

$$y_{sim} \sim N(X_{sim}\beta, \tilde{\Sigma})$$
$$\tilde{\Sigma} = X_{sim}\Sigma_{\Delta}X_{sim}^T + \sigma_{sim}^2 I_m$$

A.3.1 Models

We consider three different linear systems we can solve to recover β:

$$y_{obs} = X_{obs}\beta + \epsilon_1, \qquad \epsilon_1 \sim N(0, \sigma^2 I_n) \tag{A.4}$$

$$\begin{bmatrix} y_{obs} \\ y_{sim} \end{bmatrix} = \begin{bmatrix} X_{obs} \\ X_{sim} \end{bmatrix} \beta + \epsilon_2, \qquad \epsilon_2 \sim N\left(0, \begin{bmatrix} \sigma^2 I_n & 0 \\ 0 & \tilde{\Sigma} \end{bmatrix}\right) \tag{A.5}$$

$$\begin{bmatrix} \frac{1}{\sigma} y_{obs} \\ \tilde{\Sigma}^{-1/2} y_{sim} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma} X_{obs} \\ \tilde{\Sigma}^{-1/2} X_{sim} \end{bmatrix} \beta + \epsilon_3, \qquad \epsilon_3 \sim N(0, I_{n+m}) \tag{A.6}$$

Because we are assuming $p \gg n$, we will use ridge regression to estimate β for each of the models.

A.3.2 Risk of estimators

For $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$ and linear system $y = X\beta + \epsilon$, $\epsilon \sim N(0, \Sigma)$, the ridge estimator is given by

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T y$$

The risk of $\hat{\beta}$ is

$$\mathbb{E}\|\hat{\beta} - \beta\|_2^2 = \text{var}(\hat{\beta}) + \text{bias}^2(\hat{\beta})$$

Letting $W_\lambda = (X^T X + \lambda I_p)^{-1} X^T$, we can show that

$$\text{var}(\hat{\beta}) = \frac{1}{n}\text{tr}\left(W_\lambda \Sigma W_\lambda^T\right)$$
$$\text{bias}^2(\hat{\beta}) = \beta^T (W_\lambda X - I_p)^T (W_\lambda X - I_p)\beta$$

We now compute the closed-form ridge estimator and associated risk for each of the three systems. We make the following simplifying assumptions:

- X_{obs} is rotated so that $X_{obs}^T X_{obs} = \text{diag}(\lambda_1, \ldots, \lambda_p)$
- X_{sim} is rotated so that $X_{sim}^T X_{sim} = \text{diag}(\delta_1, \ldots, \delta_p)$
- $\Sigma_\Delta = \text{diag}(\alpha_1, \ldots, \alpha_p)$.

Case 1: Observations Only

$$\hat{\beta}_1 = (X_{obs}^T X_{obs} + \lambda_p I)^{-1} X_{obs}^T y_{obs}$$
$$\text{risk}(\hat{\beta}_1) = \frac{\sigma^2}{n} \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + \lambda)^2} + \sum_{j=1}^{p} \beta_j^2 \left(\frac{\lambda}{\lambda_j + \lambda}\right)^2$$

Case 2: Observations and Simulations Naively Combined

$$\hat{\beta}_2 = (X_{obs}^T X_{obs} + X_{sim}^T X_{sim} \lambda_p I)^{-1} (X_{obs}^T y_{obs} + X_{sim}^T y_{sim})$$

$$\text{risk}(\hat{\beta}_2) = \frac{\sigma^2}{n+m} \sum_{j=1}^{p} \frac{\lambda_j + \delta_j + \frac{1}{\sigma^2}\delta_j^2 \alpha_j}{(\lambda_j + \delta_j + \lambda)^2} + \sum_{j=1}^{p} \beta_j^2 \left(\frac{\lambda}{\lambda_j + \delta_j + \lambda}\right)^2$$

Case 3: MAP Estimator

$$\hat{\beta}_3 = \left(\frac{1}{\sigma^2} X_{obs}^T X_{obs} + X_{sim}^T \tilde{\Sigma}^{-1} X_{sim} + \lambda_p I \right)^{-1} \left(\frac{1}{\sigma^2} X_{obs}^T y_{obs} + X_{sim}^T \tilde{\Sigma}^{-1} y_{sim} \right)$$

$$\text{risk}(\hat{\beta}_3) = \frac{\sigma^2}{n+m} \sum_{j=1}^{p} \frac{\lambda_j + \xi_j}{(\lambda_j + \xi_j + \lambda)^2} + \sum_{j=1}^{p} \beta_j^2 \left(\frac{\lambda}{\lambda_j + \xi_j + \lambda} \right)^2$$

$$\xi_j = \delta_j \left(1 - \frac{\delta_j \alpha_j}{\sigma^2 + \delta_j \alpha_j} \right)$$

APPENDIX B

TECHNICAL DETAILS FOR LAZYVI

B.1 Supporting lemmas

Conditions

(A1) There exists some constant $C > 0$ such that, for each sequence $f_1, f_2, \cdots \in \mathcal{F}$ such that $\|f_i - f_0\|_{\mathcal{F}} \to 0$, $|V(f_j, P_0) - V(f_0, P_0)| \leq C\|f_j - f_0\|_{\mathcal{F}}^2$ for each j large enough;

(A2) There exists some constant $\delta > 0$ such that for each sequence $\epsilon_1, \epsilon_2, \cdots \in \mathbb{R}$ and $h, h_1, h_2, \cdots \in \mathbb{R}$ satisfying that $\epsilon_j \to 0$ and $\|h_j - h\|_\infty \to 0$, it holds that

$$\sup_{f \in \mathcal{F}: \|f - f_0\|_{\mathcal{F}} < \delta} \left| \frac{V(f, P_0 + \epsilon_j h_j) - V(f, P_0)}{\epsilon_j} - \dot{V}(f, P_0; h_j) \right| \to 0;$$

(B2) $\int [g_n(z)]^2 dP_0(z) = o_P(1)$;

Lemma B.1.1. *(Williamson et al. (2021)) Suppose (A1-A2, B2) regularity conditions hold. Denote $f_n(X)$ and $f_{n,-j}(X_{-j})$ as the estimate for f_0 and $f_{0,-j}$, Then for a predictive skill measure $V(f, P)$ satisfying conditions (A1)-(A2), (B2) in Appendix, as long as the estimators satisfy the following condition:*

$$\|f_n - f_0\|_{\mathcal{F}} = O_P(n^{-\frac{1}{4}}), \quad \|f_{n,-j} - f_{0,-j}\|_{\mathcal{F}} = O_P(n^{-\frac{1}{4}}), \tag{B.1}$$

for all $j \in [p]$, then we have

$$\begin{aligned} v_n - v_0 =& \frac{1}{n} \sum_{i=1}^{n} \dot{V}(f_0, P_0; \delta_{Z_i} - P_0) + O_P(\frac{1}{\sqrt{n}}), \\ v_{n,-j} - v_{0,-j} =& \frac{1}{n} \sum_{i=1}^{n} \dot{V}(f_{0,-j}, P_{0,-j}; \delta_{Z_i} - P_{0,-j}) + O_P(\frac{1}{\sqrt{n}}), \end{aligned} \tag{B.2}$$

where $v_n = V(f_n, P_n)$ *and* $v_{n,-j} = V(f_{n,-j}, P_{n,-j}),\ \forall j \in [p]$.

B.2 Missing Proofs

B.2.1 Proof of Lemma 3.4.5

In this section, we present the detailed proof of Lemma 3.4.5, which gives the empirical estimation error bounds for the NTK kernel ridge regression estimation. The proof follows the proof framework provided in Hu et al. (2019).

[Proof.]

First of all, according to kernel ridge regression, denote

- $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$;
- $\mathbf{h}_{\theta_f}^{(j)} = (h_{\theta_f}(\mathbf{X}_1^{(j)}), \ldots, h_{\theta_f}(\mathbf{X}_n^{(j)}))^T$;
- $\mathbf{f}_{0,-j} = (f_{0,-j}(\mathbf{X}_1^{(j)}), \ldots, f_{0,-j}(\mathbf{X}_n^{(j)}))^T$.

we have

$$(\tilde{h}_{\theta_f+\Delta\theta_j}(\mathbf{X}_1^{(j)}), \ldots, \tilde{h}_{\theta_f+\Delta\theta_j}(\mathbf{X}_n^{(j)}))^T = \mathbb{K}^{(j)}(\mathbb{K}^{(j)} + \lambda I_n)^{-1}(\mathbf{Y} - \mathbf{h}_{\theta_f}^{(j)}) + \mathbf{h}_{\theta_f}^{(j)}. \tag{B.3}$$

Recall that $\epsilon^{(j)} = Y - \mathbb{E}(Y|X_{-j}) = Y - f_{0,-j}(\mathbf{X}^{(j)})$, we define its observed samples as

$$\boldsymbol{\epsilon}^{(j)} = \Big(Y_1 - f_{0,-j}(\mathbf{X}_1^{(j)}), \ldots, Y_n - f_{0,-j}(\mathbf{X}_n^{(j)})\Big)^T = \mathbf{Y} - \mathbf{f}_{0,-j};$$

Recall the definition of $\mathbf{e}^{(j)}$, we have

$$\mathbf{e}^{(j)} = \mathbf{f}_{0,-j} - \mathbf{h}_{\theta_f}^{(j)}.$$

Hence we have

$$
\begin{aligned}
&\sqrt{n}\|\tilde{h}_{\theta_f+\Delta\theta_j}(X^{(j)}) - f_{0,-j}(X^{(j)})\|_n \\
=&\sqrt{\sum_{i=1}^{n}\left[\tilde{h}_{\theta_f+\Delta\theta_j}(\mathbf{X}_i^{(j)}) - f_{0,-j}(\mathbf{X}_i^{(j)})\right]^2} \\
=&\|\mathbb{K}^{(j)}(\mathbb{K}^{(j)}+\lambda I_n)^{-1}(\mathbf{Y}-\mathbf{h}_{\theta_f}^{(j)}) + \mathbf{h}_{\theta_f}^{(j)} - \mathbf{f}_{0,-j}\| \\
=&\|\mathbb{K}^{(j)}(\mathbb{K}^{(j)}+\lambda I_n)^{-1}(\mathbf{f}_{0,-j}+\boldsymbol{\epsilon}^{(j)}-\mathbf{h}_{\theta_f}^{(j)}) + \mathbf{h}_{\theta_f}^{(j)} - \mathbf{f}_{0,-j}\| \\
=&\|\mathbb{K}^{(j)}(\mathbb{K}^{(j)}+\lambda I_n)^{-1}(\mathbf{e}^{(j)}+\boldsymbol{\epsilon}^{(j)}) - \mathbf{e}^{(j)}\| \\
=&\|\mathbb{K}^{(j)}(\mathbb{K}^{(j)}+\lambda I_n)^{-1}\boldsymbol{\epsilon}^{(j)} - \lambda(\mathbb{K}^{(j)}+\lambda I_n)^{-1}\mathbf{e}^{(j)}\| \\
\leq&\|\mathbb{K}^{(j)}(\mathbb{K}^{(j)}+\lambda I_n)^{-1}\boldsymbol{\epsilon}^{(j)}\| + \|\lambda(\mathbb{K}^{(j)}+\lambda I_n)^{-1}\mathbf{e}^{(j)}\|.
\end{aligned}
\tag{B.4}
$$

According to 3.4.2 and Hsu et al. (2012), we have

$$
P\left(\|A\boldsymbol{\epsilon}^{(j)}\|^2/\sigma^2 > \mathrm{tr}(\Sigma) + 2\sqrt{\mathrm{tr}(\Sigma^2)t} + 2\|\Sigma\|t \mid \mathbf{X}^{(j)}\right) \leq e^{-t}, \tag{B.5}
$$

where $A = \mathbb{K}^{(j)}(\mathbb{K}^{(j)}+\lambda I_n)^{-1}$, and $\Sigma = A^\top A$. Hence we have with probability at least $1-\delta$ for any $\delta > 0$, we have

$$
\|\mathbb{K}^{(j)}(\mathbb{K}^{(j)}+\lambda I_n)^{-1}\boldsymbol{\epsilon}^{(j)}\| \leq \sigma\sqrt{\mathrm{tr}(\Sigma) + 2\sqrt{\mathrm{tr}(\Sigma^2)\log(\frac{1}{\delta})} + 2\|\Sigma\|\log(\frac{1}{\delta})}. \tag{B.6}
$$

Let $\lambda_1, \ldots, \lambda_n > 0$ be the eigenvalues of $\mathbb{K}^{(j)}$, we then have

$$
\begin{aligned}
&\mathrm{tr}[\Sigma] = \mathrm{tr}[A^\top A] = \sum_{i=1}^{n}\frac{\lambda_i^2}{(\lambda_i+\lambda)^2} \leq \sum_{i=1}^{n}\frac{\lambda_i^2}{4\lambda_i\cdot\lambda} = \frac{\mathrm{tr}[\mathbb{K}^{(j)}]}{4\lambda}, \\
&\mathrm{tr}[\Sigma^2] = \mathrm{tr}[A^\top A^2 A^\top] = \sum_{i=1}^{n}\frac{\lambda_i^4}{(\lambda_i+\lambda)^4} \leq \sum_{i=1}^{n}\frac{\lambda_i^4}{4^4\lambda(\frac{\lambda_i}{3})^3} = \frac{3^3\mathrm{tr}[\mathbb{K}^{(j)}]}{4^4\lambda} \leq \frac{\mathrm{tr}[\mathbb{K}^{(j)}]}{4\lambda}, \\
&\|\Sigma\| = \|\mathbb{K}^{(j)}(\mathbb{K}^{(j)}+\lambda I_n)^{-2}\mathbb{K}^{(j)}\| \leq 1.
\end{aligned}
\tag{B.7}
$$

Hence we have with probability at least $1-\delta$,

$$\begin{aligned}
&\|\mathbb{K}^{(j)}(\mathbb{K}^{(j)}+\lambda I_n)^{-1}\boldsymbol{\epsilon}^{(j)}\| \\
\leq&\sigma\sqrt{\frac{\mathrm{tr}[\mathbb{K}^{(j)}]}{4\lambda}+2\sqrt{\frac{\mathrm{tr}[\mathbb{K}^{(j)}]}{4\lambda}}+2\log(\frac{1}{\delta})} \\
\leq&\sigma\sqrt{\frac{\mathrm{tr}[\mathbb{K}^{(j)}]}{4\lambda}+2\sqrt{\frac{\mathrm{tr}[\mathbb{K}^{(j)}]}{2\lambda}}+2\log(\frac{1}{\delta})} \\
=&\sigma\sqrt{\frac{\mathrm{tr}[\mathbb{K}^{(j)}]}{4\lambda}}+\sigma\sqrt{2\log(\frac{1}{\delta})}.
\end{aligned} \tag{B.8}$$

By the fact that

$$\|\lambda(\mathbb{K}^{(j)}+\lambda I_n)^{-1}\mathbf{e}^{(j)}\|=\lambda\sqrt{(\mathbf{e}^{(j)})^\top(\mathbb{K}^{(j)}+\lambda I_n)^{-2}\mathbf{e}^{(j)}}, \tag{B.9}$$

we have

$$\|\tilde{h}_{\theta_f+\Delta\theta_j}(X^{(j)})-f_{0,-j}(X^{(j)})\|_n\leq\frac{\lambda}{\sqrt{n}}\sqrt{(\mathbf{e}^{(j)})^\top[\mathbb{K}^{(j)}+\lambda I_n]^{-2}\mathbf{e}^{(j)}}+\sigma\sqrt{\frac{\mathrm{tr}[\mathbb{K}^{(j)}]}{4n\lambda}}+\sigma\sqrt{\frac{2}{n}\log(\frac{1}{\delta})}. \tag{B.10}$$

□

B.2.2 Lemma B.2.1 and its proof

Define the Hilbert norm for a function $f(x)=\alpha^T K(x,\mathbf{X}^{(j)}),\ \forall\alpha\in\mathbb{R}^n$ in the NTK kernel space is: $\|f\|_{\mathcal{H}}=\sqrt{\alpha^T\mathbb{K}^{(j)}\alpha}$. The following lemma is to bound the Hilbert norm for $\tilde{h}_{\theta_f+\Delta\theta_j}-h_{\theta_f}$ so that we could bound the complexity of the function class it lies in.

Lemma B.2.1. *With probability at least $1-\delta$, for any $j\in[p]$ we have*

$$\|\tilde{h}_{\theta_f+\Delta\theta_j}(x)-h_{\theta_f}(x)\|_{\mathcal{H}}\leq\sqrt{(\mathbf{e}^{(j)})^T(\mathbb{K}^{(j)}+\lambda I_n)^{-1}\mathbf{e}^{(j)}}+\frac{\sigma}{\sqrt{\lambda}}\left(\sqrt{n}+\sqrt{2\log(\frac{1}{\delta})}\right). \tag{B.11}$$

[Proof.] Recall that $\tilde{h}_{\theta_f+\Delta\theta_j}(x) = \ker_{\theta_f}(x, \mathbf{X}^{(j)})(\mathbb{K}^{(j)} + \lambda I_n)^{-1}(\mathbf{Y} - \mathbf{h}_{\theta_f}^{(j)}) + h_{\theta_f}(x)$. Based on the fact that $\mathbf{Y} - \mathbf{h}_{\theta_f}^{(j)} = \mathbf{e}^{(j)} + \boldsymbol{\epsilon}^{(j)}$, we have

$$
\begin{aligned}
&\|\tilde{h}_{\theta_f+\Delta\theta_j}(x) - h_{\theta_f}(x)\|_{\mathcal{H}} \\
=&\|(\mathbf{Y} - \mathbf{h}_{\theta_f}^{(j)})^T(\mathbb{K}^{(j)} + \lambda I_n)^{-1}\ker_{\theta_f}(\mathbf{X}^{(j)}, x)\|_{\mathcal{H}} \\
=&\sqrt{(\mathbf{e}^{(j)} + \boldsymbol{\epsilon}^{(j)})^T(\mathbb{K}^{(j)} + \lambda I_n)^{-1}\mathbb{K}^{(j)}(\mathbb{K}^{(j)} + \lambda I_n)^{-1}(\mathbf{e}^{(j)} + \boldsymbol{\epsilon}^{(j)})} \\
\leq&\sqrt{(\mathbf{e}^{(j)} + \boldsymbol{\epsilon}^{(j)})^T(\mathbb{K}^{(j)} + \lambda I_n)^{-1}(\mathbf{e}^{(j)} + \boldsymbol{\epsilon}^{(j)})} \\
\leq&\sqrt{(\mathbf{e}^{(j)})^T(\mathbb{K}^{(j)} + \lambda I_n)^{-1}\mathbf{e}^{(j)}} + \sqrt{(\boldsymbol{\epsilon}^{(j)})^T(\mathbb{K}^{(j)} + \lambda I_n)^{-1}\boldsymbol{\epsilon}^{(j)}} \\
\leq&\sqrt{(\mathbf{e}^{(j)})^T(\mathbb{K}^{(j)} + \lambda I_n)^{-1}\mathbf{e}^{(j)}} + \sqrt{\frac{(\boldsymbol{\epsilon}^{(j)})^T\boldsymbol{\epsilon}^{(j)}}{\lambda}}.
\end{aligned}
\tag{B.12}
$$

Using the concentration inequality in Hsu et al. (2012) again, we have with probability at least $1 - \delta$, we have

$$
\sqrt{(\boldsymbol{\epsilon}^{(j)})^T\boldsymbol{\epsilon}^{(j)}} \leq \sigma\sqrt{n + 2\sqrt{n\log(\frac{1}{\delta})} + 2\log(\frac{1}{\delta})} \leq \sigma\left(\sqrt{n} + \sqrt{2\log(\frac{1}{\delta})}\right). \tag{B.13}
$$

Hence we prove B.2.1 by combining B.12 and B.13.

B.2.3 Generalization error bound and its proof

In the following, we will bound the generalization error based on the above empirical error bound.

Lemma B.2.2. *For any $j \in [p]$, let $\|\cdot\|$ be the $L_2(P_0)$ norm defined as $\|f\| = \sqrt{\int |f(x^{(j)})|dP_0(x)}$,*

then we have with probability at least $1-\delta$ *for any* $\delta > 0$*,*

$$\|\tilde{h}_{\theta_f+\Delta\theta_j} - f_{0,-j}\| \leq \left\{ \frac{\lambda}{\sqrt{n}} \sqrt{(\mathbf{e}^{(j)})^\top [\mathbb{K}^{(j)} + \lambda I_n]^{-2} \mathbf{e}^{(j)}} + \sigma \sqrt{\frac{tr[\mathbb{K}^{(j)}]}{4n\lambda}} + \sigma \sqrt{\frac{2}{n} \log(\frac{3}{\delta})} \right\} + \frac{2\sqrt{tr[\mathbb{K}^{(j)}]}}{n} \left[O(1) + \frac{\sigma}{\sqrt{\lambda}} (\sqrt{n} + \sqrt{2\log(3/\delta)}) \right] + \sqrt{\frac{\log(3/\delta)}{2n}}. \tag{B.14}$$

Under Assumptions 3.4.1 and 3.4.2, when we take the penalty parameter in the rate $\lambda = O(\sqrt{n})$*, we have with high probability that* $\|\tilde{h}_{\theta_f+\Delta\theta_j} - f_{0,-j}\| \leq O(n^{-1/4})$*.*

[Proof.] According to 3.4.5, we know that with probability at least $1-\delta/3$,

$$\|\tilde{h}_{\theta_f+\Delta\theta_j} - f_{0,-j}\|_n \leq \frac{\lambda}{\sqrt{n}} \sqrt{(\mathbf{e}^{(j)})^\top [\mathbb{K}^{(j)} + \lambda I_n]^{-2} \mathbf{e}^{(j)}} + \sigma \sqrt{\frac{\mathrm{tr}[\mathbb{K}^{(j)}]}{4n\lambda}} + \sigma \sqrt{\frac{2}{n} \log(\frac{3}{\delta})}. \tag{B.15}$$

By Bartlett and Mendelson (2002), we know that the empirical Rademacher complexity for a function class $\mathcal{F}_B = \{f(x) = \alpha^T \mathrm{ker}_{\theta_f}(\mathbf{X}^{(j)}, x) : \|f\|_{\mathcal{H}} \leq B\}$ is bounded as

$$\hat{\mathcal{R}}_S(\mathcal{F}_B) \leq \frac{B\sqrt{\mathrm{tr}[\mathbb{K}^{(j)}]}}{n}.$$

According to Mohri et al. (2018), with probability at least $1-\delta/3$, we have

$$\begin{aligned} \sup_{\tilde{h}_{\theta_f+\Delta\theta_j} - h_{\theta_f} \in \mathcal{F}} & \left\{ \left\| \tilde{h}_{\theta_f+\Delta\theta_j}(x^{(j)}) - h_{\theta_f}(x^{(j)}) - \left(f_{0,-j}(x^{(j)}) - h_{\theta_f}(x^{(j)}) \right) \right\| \right. \\ & \left. - \|\tilde{h}_{\theta_f+\Delta\theta_j}(x) - f_{0,-j}(x^{(j)})\|_n \right. \\ \leq & 2\hat{\mathcal{R}}_S(\mathcal{F}) + \sqrt{\frac{\log(3/\delta)}{2n}}. \end{aligned} \tag{B.16}$$

From 3.4.1 and B.2.1, we have with probability $1-\delta/3$

$$\|\tilde{h}_{\theta_f+\Delta\theta_j}(x) - h_{\theta_f}(x)\|_{\mathcal{H}} := B' \leq O(1) + \frac{\sigma}{\sqrt{\lambda}}\left(\sqrt{n} + \sqrt{2\log(\frac{3}{\delta})}\right). \tag{B.17}$$

Then we have with probability $1-\delta$,

$$\begin{aligned}
&\|\tilde{h}_{\theta_f+\Delta\theta_j} - f_{0,-j}\| \\
\leq &\|\tilde{h}_{\theta_f+\Delta\theta_j} - f_{0,-j}\|_n + 2\hat{\mathcal{R}}_S(\mathcal{F}) + \sqrt{\frac{\log(3/\delta)}{2n}} \\
\leq &\left\{\frac{\lambda}{\sqrt{n}}\sqrt{(\mathbf{e}^{(j)})^\top[\mathbb{K}^{(j)}+\lambda I_n]^{-2}\mathbf{e}^{(j)}} + \sigma\sqrt{\frac{\mathrm{tr}[\mathbb{K}^{(j)}]}{4n\lambda}} + \sigma\sqrt{\frac{2}{n}\log(\frac{3}{\delta})}\right\} + \frac{2B'\sqrt{\mathrm{tr}[\mathbb{K}^{(j)}]}}{n} + \sqrt{\frac{\log(3/\delta)}{2n}} \\
\leq &\left\{\frac{\lambda}{\sqrt{n}}\sqrt{(\mathbf{e}^{(j)})^\top[\mathbb{K}^{(j)}+\lambda I_n]^{-2}\mathbf{e}^{(j)}} + \sigma\sqrt{\frac{\mathrm{tr}[\mathbb{K}^{(j)}]}{4n\lambda}} + \sigma\sqrt{\frac{2}{n}\log(\frac{3}{\delta})}\right\} \\
&+ \frac{2\sqrt{\mathrm{tr}[\mathbb{K}^{(j)}]}}{n}\left[O(1) + \frac{\sigma}{\sqrt{\lambda}}(\sqrt{n} + \sqrt{2\log(3/\delta)})\right] + \sqrt{\frac{\log(3/\delta)}{2n}}
\end{aligned} \tag{B.18}$$

By the assumptions that $\|[\mathbb{K}^{(j)}+\lambda I_n]^{-1}\mathbf{e}^{(j)}\|^2 = O_P(1/\sqrt{n})$ and $\mathrm{tr}[\mathbb{K}^{(j)}] = O(n)$ in 3.4.1, when we take $\lambda = O(\sqrt{n})$, we have

$$\|\tilde{h}_{\theta_f+\Delta\theta_j} - f_{0,-j}\| \leq O(n^{-1/4}). \tag{B.19}$$

B.2.4 Proof of 3.4.6

3.4.6 *For a large neural network whose width is in the order of $O(\sqrt{n})$ where n is the training sample size, our lazy trained neural network is close to its linearization with high probability:*

$$\|\tilde{h}_{\theta_f+\Delta\theta_j}(x) - h_{\theta_f+\Delta\theta_j}(x)\| \leq O(n^{-1/4}). \tag{B.20}$$

[Proof.] Since $\tilde{h}_{\theta_f+\Delta\theta_j}(x) = h_{\theta_f} + \Delta\theta_j^T \nabla_\theta h_\theta(x)|_{\theta=\theta_f}$ is a linearization of $h_{\theta_f+\Delta\theta_j}(x)$ around the initialization θ_f, according to Theorem 2.1 in Lee et al. (2020), when the neural network has a width M, the neural network is close to its linearization with probability arbitrarily close to 1:

$$\|\tilde{h}_{\theta_f+\Delta\theta_j}(x) - h_{\theta_f+\Delta\theta_j}(x)\|_2 = O(\frac{1}{\sqrt{M}}). \tag{B.21}$$

Specifically, when the neural network M takes the order of $O(\sqrt{n})$, we have $\|\tilde{h}_{\theta_f+\Delta\theta_j}(x) - h_{\theta_f+\Delta\theta_j}(x)\|_2 = O(n^{-1/4})$.

B.2.5 Proof of the Main Theorem (3.4.4)

Based on Lemmas B.2.2 and 3.4.6, for a neural network with width at least $M = O(\sqrt{n})$ when the assumptions hold true, by triangle inequality we have

$$\|h_{\theta_f+\Delta\theta_j} - f_{0,-j}\| \leq \|h_{\theta_f+\Delta\theta_j} - \tilde{h}_{\theta_f+\Delta\theta_j}\| + \|\tilde{h}_{\theta_f+\Delta\theta_j} - f_{0,-j}\| = O_p(n^{-1/4}). \tag{B.22}$$

This holds true for any $j \in [p]$. Then by B.1.1, we finish the proof for 3.4.4.

B.2.6 Proof of 3.3.1

The density of X_1 given X_2 in the setting of 3.3.1 is:

$$f(x_1|x_2;\rho,\sigma) = \frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)\sigma^2}(x_1^2 - 2\rho x_1 x_2 + x_2^2)\right\}, \tag{B.23}$$

thus we have $X_1|X_2 \sim \mathcal{N}(\rho X_2, (1-\rho^2)\sigma^2)$.

REFERENCES

AghaKouchak, A., Feldman, D., Hoerling, M., Huxman, T., and Lund, J. (2015). Water and climate: Recognize anthropogenic drought. *Nature*, 524:409–411.

Aikens, B., Harper, D., Paul-Brutus, R., Scrutchins, D., and Simpson, Y. (2021). The State of Health for Blacks in Chicago. Technical report, Chicago Department of Public Health, Health Equity Index Committee.

Alagoz, O., Sethi, A. K., Patterson, B. W., Churpek, M., and Safdar, N. (2021). Effect of Timing of and Adherence to Social Distancing Measures on COVID-19 Burden in the United States. *Annals of Internal Medicine*, 174(1):50–57. Publisher: American College of Physicians.

Alaíz, C. M., Barbero, A., and Dorronsoro, J. R. (2013). Group fused lasso. In *Proceedings of the 23rd International Conference on Artificial Neural Networks and Machine Learning: ICANN 2013 - Volume 8131*, page 66–73.

Allen, R. and Luptowitz, R. (2017). El Niño–like teleconnection increases California precipitation in response to warming. *Nat. Commun.*, 8:16055.

Arentze, T., Timmermans, H., and Hofman, F. (2007). Creating synthetic household populations: Problems and approach. *Transportation Research Record*, 2014(1):85–91.

Arora, S., Du, S. S., Hu, W., Li, Z., and Wang, R. (2019). Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks.

Aylett-Bullock, J., Cuesta-Lazaro, C., Quera-Bofarull, A., Icaza-Lizaola, M., Sedgewick, A., Truong, H., Curran, A., Elliott, E., Caulfield, T., Fong, K., Vernon, I., Williams, J., Bower, R., and Krauss, F. (2021). June: open-source individual-based epidemiology simulation. *Royal Society Open Science*, 8(7).

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Barber, R. F. and Candes, E. J. (2018). A knockoff filter for high-dimensional selective inference.

Barron, A. R. (1994). Approximation and estimation bounds for artificial neural networks. *Machine learning*, 14(1):115–133.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482.

Baxter, S. and Nigam, S. (2015). Key role of the North Pacific Oscillation–west Pacific pattern in generating the extreme 2013/14 North American winter. *J. Climate*, 28:8109–8117.

Beaumont, M. A. (2019). Approximate Bayesian Computation. *Annual Review of Statistics and Its Application*, 6(1):379–403.

Beckman, R. J., Baggerly, K. A., and McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, 30(6):415–429.

Bellenger, H., Guilyardi, E., Leloup, J., Lengaigne, M., and Vialard, J. (2014). ENSO representation in climate models: From CMIP3 to CMIP5. *Climate Dyn.*, 42:1999–2018.

Berry, G. and Reeder, M. J. (2014). Objective identification of the intertropical convergence zone: Climatology and trends from the ERA-Interim. *J. Climate*, 27:1894–1909.

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Stat.*, 36:2577–2604.

Bradley, R. S., Diaz, H. F., Kiladis, G. N., and Eischied, J. K. (1987). ENSO signal in continental temperature and precipitation records. *Nature*, 327:497–501.

Cai, T. T., Zhao, R., and Zhou, H. H. (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electron. J. Stat.*, 10:1–59.

Cajka, J. C., Cooley, P. C., and Wheaton, W. D. (2010). Attribute assignment to a synthetic population in support of agent-based disease modeling. Technical report, RTI Press publication No. MR-0019-1009, Research Triangle Park, North Carolina.

Candes, E., Fan, Y., Janson, L., and Lv, J. (2017). Panning for Gold: Model-X Knockoffs for High-dimensional Controlled Variable Selection. *arXiv:1610.02351 [math, stat]*. arXiv: 1610.02351.

Cao, L. and Liu, Q. (2021). COVID-19 Modeling: A Review. SSRN Scholarly Paper ID 3899127, Social Science Research Network, Rochester, NY.

Casati, D., Müller, K., Fourie, P. J., Erath, A., and Axhausen, K. W. (2015). Synthetic population generation by combining a hierarchical, simulation-based approach with reweighting by generalized raking. *Transportation Research Record*, 2493(1):107–116.

Castello, A. F. and Shelton, M. L. (2004). Winter precipitation on the US Pacific coast and El Niño–southern oscillation events. *Int. J. Climatol.*, 24:481–497.

Cayan, D. R., Redmond, K. T., and Riddle, L. G. (1999). ENSO and hydrologic extremes in the western United States. *J. Climate*, 12:2881–2893.

Chang, C.-H., Rampasek, L., and Goldenberg, A. (2017). Dropout feature ranking for deep learning models. *arXiv preprint arXiv:1712.08645*.

Chang, S., Pierson, E., Koh, P. W., Gerardin, J., Redbird, B., Grusky, D., and Leskovec, J. (2021). Mobility network models of COVID-19 explain inequities and inform reopening. *Nature*, 589(7840):82–87. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7840 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computational science;Epidemiology;SARS-CoV-2;Society Subject_term_id: computational-science;epidemiology;sars-cov-2;society.

Chang, Y., Schubert, S. D., and Suarez, M. J. (2000). Boreal winter predictions with the GEOS-2 GCM: The role of boundary forcing and initial conditions. *Quart. J. Roy. Meteor. Soc.*, 126:2293–2321.

Chatterjee, S., Steinhaeuser, K., Banerjee, A., Chatterjee, S., and Ganguly, A. (212). Sparse group lasso: Consistency and climate applications. In *Proceedings of the 12th SIAM International Conference on Data Mining*, pages 47–58.

Chen, Y., Morton, D. C., Andela, N., Giglio, L., and Randerson, J. T. (2016). How much global burned area can be forecast on seasonal time scales using sea surface temperatures? *Environmental Research Letters*, 11(4):045001.

Cheng, L. and AghaKouchak, A. (2015). A methodology for deriving ensemble response from multimodel simulations. *J. Hydrol.*, 522:49–57.

Chizat, L., Oyallon, E., and Bach, F. (2020). On Lazy Training in Differentiable Programming. *arXiv:1812.07956*.

Coelho, C. A. S., Pezzulli, S., Balmaseda, M., Doblas-Reyes, F. J., and Stephenson, D. B. (2004). Forecast calibration and combination: A simple Bayesian approach for ENSO. *J. Climate*, 17:1504–1516.

Collier, N., Ozik, J., and Macal, C. M. (2015). Large-Scale Agent-Based Modeling with Repast HPC: A Case Study in Parallelizing an Agent-Based Model. In *Euro-Par 2015: Parallel Processing Workshops*, number 9523 in Lecture Notes in Computer Science, pages 454–465. Springer International Publishing, Vienna, Austria.

Cuevas, E. (2020). An agent-based model to evaluate the COVID-19 transmission risks in facilities. *Computers in Biology and Medicine*, 121:103827.

Dai, A. (2013). The influence of the inter-decadal Pacific oscillation on US precipitation during 1923–2010. *Climate Dyn.*, 41:633–646.

de La Beaujardière, J., Banihirwe, A., Shih, C. F. G., Paul, K., and Hamman, J. (2019). Ncar cesm lens cloud-optimized subset. *UCAR/NCAR Computational and Informations Systems Lab*.

Degras, D. (2019). Sparse Group Fused Lasso for Model Segmentation. *arXiv e-prints*, page arXiv:1912.07761.

DelSole, T. and Banerjee, A. (2017). Statistical seasonal prediction based on regularized regression. *J. Climate*, 30:1345–1361.

Deming, W. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444.

Deser, C., Alexander, M. A., Xie, S. P., and Phillips, A. S. (2010). Sea surface temperature variability: Patterns and mechanisms. *Annu. Rev. Mar. Sci.*, 2:115–143.

Deser, C., Simpson, I. R., Phillips, A. S., and McKinnon, K. A. (2018). How well do we know ENSO's climate impacts over North America, and how do we evaluate models accordingly? *J. Climate*, 31:4991–5014.

Dettinger, M. and Cayan, D. (2014). Drought and the California Delta—A matter of extremes. *San Franc. Estuary Watershed Sci.*, 12(2).

Dettinger, M., Ralph, F. M., Das, T., Neiman, P. J., and Cayan, D. (2011). Atmospheric rivers, floods, and the water resources of California. *Water*, 3:445–478.

Doksum, K. and Samarov, A. (1995). Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. *The Annals of Statistics*, pages 1443–1473.

Ebtehaj, A. M. and Foufoula-Georgiou, E. (2013). On variational downscaling, fusion, and assimilation of hydrometeorological states: A unified framework via regularization. *Water Resour. Res.*, 49:5944–5963.

Ebtehaj, A. M., Foufoula-Georgiou, E., and Lerman, G. (2012). Sparse regularization for precipitation downscaling. *J. Geophys. Res.*, 117:D08107.

Enfield, D. B., Mestas-Nuñez, A. M., and Trimble, P. J. (2001). The Atlantic multidecadal oscillation and its relation to rainfall and river flows in the continental U.S. *Geophys. Res. Lett.*, 28:2077–2080.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E. (2016). An overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.*, 9:1937–1958.

Farooq, B., Bierlaire, M., Hurtubia, R., and Flötteröd, G. (2013). Simulation based population synthesis. *Transportation Research Part B: Methodological*, 58(C):243–263.

Fienberg, S. E. (1970). An iterative procedure for estimation in contingency tables. *Ann. Math. Statist.*, 41(3):907–917.

Gallagher, S., Richardson, L. F., Ventura, S. L., and Eddy, W. F. (2018). Spew: Synthetic populations and ecosystems of the world. *Journal of Computational and Graphical Statistics*, 27(4):773–784.

Gaudou, B., Huynh, N. Q., Philippon, D., Brugière, A., Chapuis, K., Taillandier, P., Larmande, P., and Drogoul, A. (2020). COMOKIT: A Modeling Kit to Understand, Analyze, and Compare the Impacts of Mitigation Policies Against the COVID-19 Epidemic at the Scale of a City. *Frontiers in Public Health*, 8.

Gershunov, A. and Cayan, D. R. (2003). Heavy daily precipitation frequency over the contiguous United States: Sources of climatic variability and seasonal predictability. *J. Climate*, 16:2752–2765.

Goncalves, A. R., Banerjee, A., Sivakumar, V., and Chatterjee, S. (2016). Structured Estimation in High Dimensions: Applications in Climate. In *Large-Scale Machine Learning in the Earth Sciences*, pages 13–32. Chapman and Hall/CRC.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5).

Guo, J. Y. and Bhat, C. R. (2007). Population synthesis for microsimulating travel behavior. *Transportation Research Record*, 2014(1):92–101.

Ham, Y.-G., Kim, J.-H., and Luo, J.-J. (2019). Deep learning for multi-year ENSO forecasts. *Nature*, 573:568–572.

Hamlington, B. D., Milliff, R. F., van Loon, H., and Kim, K.-Y. (2015). A Southern Hemisphere sea level pressure-based precursor for ENSO warm and cold events. *J. Geophys. Res. Atmos.*, 120:2280–2292.

Hao, Z., Singh, V. P., and Xia, Y. (2018). Seasonal drought prediction: Advances, challenges, and future prospects. *Rev. Geophys.*, 56:108–141.

He, S., Li, X., DelSole, T., Ravikumar, P., and Banerjee, A. (2021). Sub-seasonal climate forecasting via machine learning: challenges, analysis, and advances. In *AAAI Conference on Artificial Intelligence*.

He, S., Li, X., Sivakumar, V., and Banerjee, A. (2019). Interpretable predictive modeling for climate variables with weighted lasso. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1385–1392.

Hewitt, J., Hoeting, J. A., Done, J. M., and Towler, E. (2018). Remote effects spatial process models for modeling teleconnections. *Environmetrics*, 29:e2523.

Hinch, R., Probert, W. J. M., Nurtay, A., Kendall, M., Wymant, C., Hall, M., Lythgoe, K., Cruz, A. B., Zhao, L., Stewart, A., Ferretti, L., Montero, D., Warren, J., Mather, N., Abueg, M., Wu, N., Legat, O., Bentley, K., Mead, T., Van-Vuuren, K., Feldner-Busztin, D., Ristori, T., Finkelstein, A., Bonsall, D. G., Abeler-Dörner, L., and Fraser, C. (2021). OpenABM-Covid19—An agent-based model for non-pharmaceutical interventions against COVID-19 including contact tracing. *PLOS Computational Biology*, 17(7):e1009146. Publisher: Public Library of Science.

Hirahara, S., Ishii, M., and Fukuda, Y. (2014). Centennial-scale sea surface temperature analysis and its uncertainty. *J. Climate*, 27:57–75.

Hooker, S. et al. (2019). A benchmark for interpretability methods in deep neural networks. In *NeurIPS*.

Hotton, A., Ozik, J., Kaligotla, C., Collier, N., Stevens, A., Khanna, A., MacDonnell, M., Wang, C., LePoire, D., Chang, Y., Martinez-Moyano, I., Mucenic, B., Pollack, H., Schneider, J., and Macal, C. (2022). Impact of changes in protective behaviors and out-of-household activities by age on COVID-19 transmission and hospitalization in Chicago, Illinois. *Under review*.

Howitt, R. E., Medellín-Azuara, J., MacEwan, D., Lund, J. R., and Sumner, D. A. (2014). Economic Analysis of the 2014 Drought for California Agriculture. Technical report, University of California, Davis, Center for Watershed Sciences.

Hsu, D., Kakade, S., and Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17:1–6.

Hu, W., Li, Z., and Yu, D. (2019). Simple and effective regularization methods for training on noisily labeled data with generalization guarantee. *arXiv preprint arXiv:1905.11368*.

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: Principles and Practice (2nd ed)*. OTexts: Melbourne, Australia.

Iozzi, F., Trusiano, F., Chinazzi, M., Billari, F. C., Zagheni, E., Merler, S., Ajelli, M., Del Fava, E., and Manfredi, P. (2010a). Little italy: An agent-based approach to the estimation of contact patterns- fitting predicted matrices to serological data. *PLOS Computational Biology*, 6.

Iozzi, F., Trusiano, F., Chinazzi, M., Billari, F. C., Zagheni, E., Merler, S., Ajelli, M., Fava, E. D., and Manfredi, P. (2010b). Little Italy: An Agent-Based Approach to the Estimation of Contact Patterns- Fitting Predicted Matrices to Serological Data. *PLOS Computational Biology*, 6(12):e1001021. Publisher: Public Library of Science.

Jacot, A., Gabriel, F., and Hongler, C. (2020). Neural tangent kernel: Convergence and generalization in neural networks.

Johnson, N. C., L'Heureux, M. L., Chang, C.-H., and Hu, Z.-Z. (2019). On the delayed coupling between ocean and atmosphere in recent weak El Niño episodes. *Geophys. Res. Lett.*, 46:11 416–11 425.

Kaligotla, C., Ozik, J., Collier, N., Macal, C. M., Boyd, K., Makelarski, J., Huang, E. S., and Lindau, S. T. (2020a). Model exploration of an information-based healthcare intervention using parallelization and active learning. *Journal of Artificial Societies and Social Simulation*, 23(4):1.

Kaligotla, C., Ozik, J., Collier, N., Macal, C. M., Lindau, S., Abramsohn, E., and Huang, E. (2018). Modeling an information-based community health intervention on the south side of chicago. In *Proceedings of the 2018 Winter Simulation Conference*, pages 2600–2611, Piscataway, New Jersey. Institute of Electrical and Electronics Engineers, Inc.

Kaligotla, C., Stevens, A., Ozik, J., Collier, N., Macal, C., Martinez-Moyano, I. J., Stevens, A., Mucenic, B., Hotton, A., and Choe, K. W. (2020b). Development of a large-scale synthetic population to simulate covid-19 transmission and response. *Proceedings of the 2020 Winter Simulation Conference*.

Karypis, G. and Kumar, V. (1998). A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 20(1):359–392.

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J. M., Bates, S. C., Danabasoglu, G., Edwards, J., Holland, M., Kushner, P., Lamarque, J.-F., Lawrence, D., Lindsay, K., Middleton, A., Munoz, E., Neale, R., Oleson, K., Polvani, L., and Vertenstein, M. (2015). 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*, 96(8):1333 – 1349.

Kermack, W. O., McKendrick, A. G., and Walker, G. T. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115(772):700–721. Publisher: Royal Society.

Kerr, C. C., Stuart, R. M., Mistry, D., Abeysuriya, R. G., Rosenfeld, K., Hart, G. R., Núñez, R. C., Cohen, J. A., Selvaraj, P., Hagedorn, B., George, L., Jastrzebski, M., Izzo, A. S., Fowler, G., Palmer, A., Delport, D., Scott, N., Kelly, S. L., Bennette, C. S., Wagner, B. G., Chang, S. T., Oron, A. P., Wenger, E. A., Panovska-Griffiths, J., Famulare, M., and Klein, D. J. (2021). Covasim: An agent-based model of COVID-19 dynamics and interventions. *PLOS Computational Biology*, 17(7):e1009149. Publisher: Public Library of Science.

Kim, S. T., Cai, W., Jin, F.-F., and Yu, J.-Y. (2014). ENSO stability in coupled climate models and its association with mean state. *Climate Dyn.*, 42:3313–3321.

Klepac, P., Kucharski, A. J., Conlan, A. J., Kissler, S., Tang, M. L., Fry, H., and Gog, J. R. (2020). Contacts in context: large-scale setting-specific social mixing matrices from the BBC Pandemic project. preprint, Epidemiology.

Lee, J., Xiao, L., Schoenholz, S. S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. (2020). Wide neural networks of any depth evolve as linear models under gradient descent. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124002.

Lee, S.-K., Lopez, H., Chung, E.-S., DiNezio, P., Yeh, S.-W., and Wittenberg, A. T. (2018). On the fragile relationship between El Niño and California rainfall. *Geophys. Res. Lett.*, 45:907–915.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Lenssen, N. J. L., Goddard, L., and Mason, S. (2020). Seasonal Forecast Skill of ENSO Teleconnection Maps. *Weather and Forecasting*, 35(6):2387–2406. Publisher: American Meteorological Society Section: Weather and Forecasting.

Li, C., Zwiers, F., Zhang, X., Chen, G., Lu, J., Li, G., Norris, J., Tan, Y., Sun, Y., and Liu, M. (2019). Larger Increases in More Extreme Local Precipitation Events as Climate Warms. *Geophysical Research Letters*, 46(12):6885–6891.

Li, G. and Xie, S. (2014). Tropical biases in CMIP5 multimodel ensemble: The excessive equatorial Pacific cold tongue and double ITCZ problems. *J. Climate*, 27:1765–1780.

Li, Y., Mark, B., Raskutti, G., Willett, R., Song, H., and Neiman, D. (2020). Graph-Based Regularization for Regression Problems with Alignment and Highly Correlated Designs. *SIAM Journal on Mathematics of Data Science*, 2(2):480–504. Publisher: Society for Industrial and Applied Mathematics.

Liu, T., Schmitt, R. W., and Li, L. (2018). Global search for autumn-lead sea surface salinity predictors of winter precipitation in southwestern United States. *Geophys. Res. Lett.*, 45:8445–8454.

Livneh, B. and Badger, A. M. (2020). Drought less predictable under declining future snowpack. *Nat. Climate Change*, 10:452–458.

Luo, L., Wood, E. F., and Pan, M. (2007). Bayesian merging of multiple climate model forecasts for seasonal hydrological predictions. *J. Geophys. Res.*, 112:D10102.

Macal, C. (2016). Everything you need to know about agent-based modelling and simulation. *Journal of Simulation*, 10(2).

Macal, C. M., Collier, N. T., Ozik, J., Tatara, E. R., and Murphy, J. T. (2018). ChiSIM: An Agent-Based Simulation Model of Social Interactions in a Large Urban Area. In *Proceedings of the 2018 Winter Simulation Conference*, pages 810–820, Piscataway, New Jersey. Institute of Electrical and Electronics Engineers, Inc.

Macal, C. M., North, M. J., Collier, N., Dukic, V. M., Wegener, D. T., David, M. Z., Daum, R. S., Schumm, P., Evans, J. A., Wilder, J. R., Eells, S. J., and Lauderdale, D. S. (2014a). Modeling the transmission of community-associated methicillin-resistant staphylococcus aureus: A dynamic agent-based simulation. *Journal of Translational Medicine*, 12(1):1–12.

Macal, C. M., North, M. J., Collier, N., Dukic, V. M., Wegener, D. T., David, M. Z., Daum, R. S., Schumm, P., Evans, J. A., Wilder, J. R., Miller, L. G., Eells, S. J., and Lauderdale, D. S. (2014b). Modeling the transmission of community-associated methicillin-resistant Staphylococcus aureus: a dynamic agent-based simulation. *Journal of Translational Medicine*, 12:124.

Macal, C. M., Ozik, J., Collier, N. T., Kaligotla, C., MacDonell, M. M., Wang, C., LePoire, D. J., Chang, Y., and Martinez-Moyano, I. J. (2020). Citycovid: A computer simulation of covid-19 spread in a large urban area. *Proceedings of the 2020 Winter Simulation Conference*.

Madadgar, S., AghaKouchak, A., Shukla, S., Wood, A. W., Cheng, L., Hsu, K.-L., and Svoboda, M. (2016). A hybrid statistical-dynamical framework for meteorological drought prediction: Application to southwestern United States. *Water Resour. Res.*, 52:5095–5110.

Mamalakis, A. and Foufoula-Georgiou, E. (2018). A multivariate probabilistic framework for tracking the intertropical convergence zone: Analysis of recent climatology and past trends. *Geophys. Res. Lett.*, 45:13 080–13 089.

Mamalakis, A., Yu, J.-Y., Randerson, J. T., AghaKouchak, A., and Foufoula-Georgiou, E. (2018). A new interhemispheric teleconnection increases predictability of winter precipitation in southwestern US. *Nat. Commun.*, 9:2332.

Mamalakis, A., Yu, J.-Y., Randerson, J. T., AghaKouchak, A., and Foufoula-Georgiou, E. (2019). Reply: A critical examination of a newly proposed interhemispheric teleconnection to southwestern US winter precipitation. *Nat. Commun.*, 10:2918.

McCabe, G. J. and Dettinger, M. D. (1999). Decadal variations in the strength of ENSO teleconnections with precipitation in the western United States. *Int. J. Climatol.*, 19:1399–1410.

McCabe, G. J., Palecki, M. A., and Betancourt, J. L. (2004). Pacific and Atlantic Ocean influences on multidecadal drought frequence in the United States. *Proc. Natl. Acad. Sci. USA*, 101:4136–4141.

Medellín-Azuara, J., MacEwan, D., Howitt, R. E., Sumner, D. A., and Lund, J. R. (2016). Economic Analysis of the 2015 Drought for California Agriculture. Technical report, University of California, Davis, Center for Watershed Sciences.

Milne, G. J. and Xie, S. (2020). The Effectiveness of Social Distancing in Mitigating COVID-19 Spread: a modelling analysis. Technical report, medRxiv. Type: article.

Milne, G. J., Xie, S., and Poklepovich, D. (2020). A Modelling Analysis of Strategies for Relaxing COVID-19 Social Distancing. Technical report, medRxiv. Type: article.

Mo, K. C. and Higgins, R. W. (1998). Tropical influences on California precipitation. *J. Climate*, 11:412–430.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of machine learning*.

Montanari, A. and Zhong, Y. (2020). The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *arXiv preprint arXiv:2007.12826*.

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G. S., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M., and Edmunds, W. J. (2008). Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases. *PLOS Medicine*, 5(3):e74. Publisher: Public Library of Science.

Mote, P. W., Hamlet, A. F., Clark, M. P., and Lettenmaier, D. P. (2005). Declining mountain snowpack in western North America. *Bull. Amer. Meteor. Soc.*, 86:39–50.

Mucenic, B., Kaligotla, C., Stevens, A., Ozik, J., Collier, N., and Macal, C. (2021). Load Balancing Schemes for Large Synthetic Population-Based Complex Simulators. In *2021 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 985–988.

Myoung, B., Yeh, S.-W., Kim, J., and Kafatos, M. C. (2018). Impacts of Pacific SSTs on atmospheric circulations leading to California winter precipitation variability: A diagnostic modeling. *Atmosphere*, 9:455.

National Academies of Sciences, Engineering, a. M. (2016). *Next Generation Earth System Prediction: Strategies for Subseasonal to Seasonal Forecasts*. The National Academies Press.

Nguyen, Q., Mondelli, M., and Montufar, G. F. (2021). Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep relu networks. In *International Conference on Machine Learning*, pages 8119–8129. PMLR.

Overpeck, J. T., Meehl, G. A., Bony, S., and Easterling, D. R. (2011). Climate data challenges in the 21st century. *Science*, 331:700–702.

Owen, A. B. and Prieur, C. (2016). On Shapley value for measuring importance of dependent inputs.

Ozik, J., Collier, N. T., and Macal, C. M. (2021a). The Chicago Social Interaction Model (ChiSIM). Technical report, Argonne National Lab. (ANL), Argonne, IL (United States).

Ozik, J., Collier, N. T., Wozniak, J. M., Macal, C. M., and An, G. (2018). Extreme-Scale Dynamic Exploration of a Distributed Agent-Based Model With the EMEWS Framework. *IEEE Transactions on Computational Social Systems*, 5(3):884–895.

Ozik, J., Collier, N. T., Wozniak, J. M., and Spagnuolo, C. (2016). From desktop to Large-Scale Model Exploration with Swift/T. In *2016 Winter Simulation Conference (WSC)*, pages 206–220. ISSN: 1558-4305.

Ozik, J., Wozniak, J. M., Collier, N., Macal, C. M., and Binois, M. (2021b). A population data-driven workflow for COVID-19 modeling and learning. *The International Journal of High Performance Computing Applications*, 35(5):483–499. Publisher: SAGE Publications Ltd STM.

Pan, B., Hsu, K., AghaKouchak, A., Sorooshian, S., and Higgins, W. (2019). Precipitation prediction skill for the west coast United States: From short to extended range. *J. Climate*, 32:161–182.

Peng, Z., Wang, Q. J., Bennett, J. C., Pokhrel, P., and Wang, Z. (2014). Seasonal precipitation forecasts over China using monthly large-scale oceanic–atmospheric indices. *J. Hydrol.*, 519:792–802.

Quan, X., Hoerling, M., Whitaker, J., Bates, G., and Xu, T. (2006). Diagnosing sources of U.S. seasonal forecast skill. *J. Climate*, 19:3279–3293.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, 133:1155–1174.

Ramadan, O. E. and Sisiopiku, V. P. (2020). A critical review on population synthesis for activity- and agent-based transportation models. In Luca, S. D., Pace, R. D., and Djordjevic, B., editors, *Transportation Systems Analysis and Assessment*, chapter 1. IntechOpen.

Raskutti, G., Wainwright, M. J., and Yu, B. (2014). Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *Journal of Machine Learning Research*.

Redmond, K. T. and Koch, R. W. (1991). Surface climate and streamflow variability in the western United States and their relationship to large-scale circulation indices. *Water Resour. Res.*, 27:2381–2399.

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C., Ng, A. Y., Hassabis, D., Platt, J. C., Creutzig, F., Chayes, J., and Bengio, Y. (2019). Tackling Climate Change with Machine Learning. *arXiv:1906.05433 [cs, stat]*. arXiv: 1906.05433.

Ropelewski, C. F. and Halpert, M. S. (1986). North American precipitation and temperature patterns associated with the El Niño/Southern Oscillation (ENSO). *Mon. Wea. Rev.*, 114:2352–2362.

Rudin, C. and Radin, J. (2019). Why are we using black box models in ai when we don't need to? a lesson from an explainable ai competition. *Harvard Data Science Review*, 1(2). https://hdsr.mitpress.mit.edu/pub/f9kuryi8.

Sanyaolu, A., Okorie, C., Marinkovic, A., Patidar, R., Younis, K., Desai, P., Hosein, Z., Padda, I., Mangat, J., and Altaf, M. (2020). Comorbidity and its Impact on Patients with COVID-19. *Sn Comprehensive Clinical Medicine*, 2(8):1069–1076.

Sapp, S., van der Laan, M. J., and Page, K. (2014). Targeted estimation of binary variable importance measures with interval-censored outcomes. *The international journal of biostatistics*, 10(1):77–97.

Schepen, A. and Wang, Q. J. (2013). Towards accurate and reliable forecasts of Australian seasonal rainfall by calibrating and merging multiple coupled GCMs. *Mon. Wea. Rev.*, 141:4554–4563.

Schepen, A., Wang, Q. J., and Robertson, D. (2012). Evidence for using lagged climate indices to forecast Australian seasonal rainfall. *J. Climate*, 25:1230–1246.

Schepen, A., Wang, Q. J., and Robertson, D. (2014). Seasonal forecasts of Australian rainfall through calibration and bridging of coupled GCM outputs. *Mon. Wea. Rev.*, 142:1758–1770.

Schonher, T. and Nicholson, S. E. (1989). The relationship between rainfall and ENSO events. *J. Climate*, 2:1258–1269.

Schubert, S., Chang, Y., Wang, H., Koster, R., and Suarez, M. (2016). A modeling study of the causes and predictability of the spring 2011 extreme US weather activity. *J. Climate*, 29:7869–7887.

Seager, R., Henderson, N., Cane, M. A., Liu, H., and Nakamura, J. (2017). Is there a role for human-induced climate change in the precipitation decline that drove the California drought? *J. Climate*, 30:10 237–10 258.

Shrikumar, A., Greenside, P., and Kundaje, A. (2019). Learning Important Features Through Propagating Activation Differences. *arXiv:1704.02685 [cs]*. arXiv: 1704.02685.

Shukla, S., Steinemann, A., Iacobellis, S. F., and Cayan, D. R. (2015). Annual drought in California: Association with monthly precipitation and climate phases. *J. Appl. Meteor. Climatol.*, 54:2273–2281.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

Stephens, D. J., Meuleners, M. J., van Loon, H., Lamond, M. H., and Telcik, N. P. (2007). Differences in atmospheric circulation between the development of weak and strong warm events in the Southern Oscillation. *J. Climate*, 20:2191–2209.

Stevens, A., Willett, R., Mamalakis, A., Foufoula-Georgiou, E., Tejedor, A., Randerson, J. T., Smyth, P., and Wright, S. (2021). Graph-Guided Regularized Regression of Pacific Ocean Climate Variables to Increase Predictive Skill of Southwestern U.S. Winter Precipitation. *Journal of Climate*, 34(2):737–754.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR.

Swain, D. L., Horton, D. E., Singh, D., and Diffenbaugh, N. S. (2016). Trends in atmospheric patterns conducive to seasonal precipitation and temperature extremes in California. *Sci. Adv.*, 2:e1501344.

Swain, D. L., Langenbrunner, B., Neelin, J. D., and Hall, A. (2018). Increasing precipitation volatility in twenty-first-century California. *Nat. Climate Change*, 8:427–433.

Swain, D. L., Singh, D., Horton, D. E., Mankin, J. S., Ballard, T. C., and Diffenbaugh, N. S. (2017). Remote linkages to anomalous winter atmospheric ridging over the northeastern Pacific. *J. Geophys. Res. Atmos.*, 122:12 194–12 209.

Szoeke, S. P. d. and Xie, S. (2008). The tropical eastern Pacific seasonal cycle: Assessment of errors and mechanisms in IPCC AR4 coupled ocean–atmosphere general circulation models. *J. Climate*, 21:2573–2590.

Tabari, H. (2020). Climate change impact on flood and extreme precipitation increases with water availability. *Scientific Reports*, 10(1).

Tao, L., He, X., and Wang, R. (2017). A hybrid LSSVM model with empirical mode decomposition and differential evolution for forecasting monthly precipitation. *J. Hydrometeor.*, 18:159–176.

Teng, H. and Branstator, G. (2017). Causes of extreme ridges that induce California droughts. *J. Climate*, 30:1477–1492.

Thompson, J. and Wattam, S. (2021). Estimating the impact of interventions against COVID-19: From lockdown to vaccination. *PLOS ONE*, 16(12):e0261330. Publisher: Public Library of Science.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *J. Roy. Stat. Soc.*, 67B:91–108.

Tibshirani, R. and Taylor, J. (2011). The solution path of the generalized lasso. *Ann. Stat.*, 39:1335–1371.

Trenberth, K. E., Branstator, G. W., Karoly, D., Kumar, A., Lau, N.-C., and Ropelewski, C. (1998). Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures. *J. Geophys. Res.*, 103:14 291–14 324.

Trenberth, K. E. and Shea, D. J. (1987). On the evolution of the Southern Oscillation. *Mon. Wea. Rev.*, 115:3078–3096.

van Loon, H. and Shea, D. J. (1987). The Southern Oscillation. Part VI: Anomalies of sea level pressure on the Southern Hemisphere and of Pacific sea surface temperature during the development of a warm event. *Mon. Wea. Rev.*, 115:370–379.

Vose, R. S. (2014). Improved historical temperature and precipitation time series for U.S. climate divisions. *J. Appl. Meteor. Climatol.*, 53:1232–1251.

Waliser, D. E. and Gautier, C. (1993). A satellite-derived climatology of the ITCZ. *J. Climate*, 6:2162–2174.

Wang, B. (2009). Advance and prospectus of seasonal prediction: Assessment of the APCC/CliPAS 14-model ensemble retrospective seasonal prediction (1980–2004). *Climate Dyn.*, 33:93–117.

Wang, C., Deser, C., Yu, J.-Y., DiNezio, P., and Clement, A. (2017). El Niño and Southern Oscillation (ENSO): A Review. In Glynn, P. W., Manzello, D. P., and Enochs, I. C., editors, *Coral Reefs of the Eastern Tropical Pacific*, volume 8, pages 85–106. Springer Netherlands, Dordrecht.

Wang, S. Y., Hipps, L., Gillies, R. R., and Yoon, J. H. (2014). Probable causes of the abnormal ridge accompanying the 2013–2014 California drought: ENSO precursor and anthropogenic warming footprint. *Geophys. Res. Lett.*, 41:3220–3226.

Wang, T. and Miao, J.-P. (2018). Twentieth-century Pacific decadal oscillation simulated by CMIP5 coupled models. *Atmos. Ocean. Sci. Lett.*, 11:94–101.

Wilks, D. S. (2016). "The stippling shows statistically significant grid points": How research results are routinely overstated and overinterpreted, and what to do about it. *Bull. Amer. Meteor. Soc.*, 97:2263–2273.

Willard, J., Jia, X., Xu, S., Steinbach, M., and Kumar, V. (2021). Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems. *arXiv:2003.04919 [physics, stat]*. arXiv: 2003.04919.

Williamson, B. D. and Feng, J. (2020). Efficient nonparametric statistical inference on population feature importance using Shapley values. *arXiv:2006.09481 [stat]*. arXiv: 2006.09481.

Williamson, B. D., Gilbert, P. B., Simon, N. R., and Carone, M. (2021). A general framework for inference on algorithm-agnostic variable importance. *arXiv:2004.03683*.

Wu, Z., Wang, B., Li, J., and Jin, F.-F. (2009). An empirical seasonal prediction model of the East Asian summer monsoon using ENSO and NAO. *J. Geophys. Res.*, 114:D18120.

Yeager, S. G. (2018). Predicting near-term changes in the Earth system: A large ensemble of initialized decadal prediction simulations using the Community Earth System Model. *Bull. Amer. Meteor. Soc.*, 99:1867–1886.

Yu, J.-Y., Zou, Y., Kim, S. T., and Lee, T. (2012). The changing impact of El Niño on US winter temperatures. *Geophys. Res. Lett.*, 39:L1570.

Zhang, L. and Janson, L. (2021). Floodgate: inference for model-free variable importance. *arXiv:2007.01283 [stat]*. arXiv: 2007.01283.

Zhang, T. (2018). Predictability and prediction of Southern California rains during strong El Niño events: A focus on the failed 2016 winter rains. *J. Climate*, 31:555–574.

Árbol, P. M. R. d. and Iglesias, L. L. (2020). Comparison of epidemic control strategies using agent-based simulations. Technical report, medRxiv. Type: article.