Spatio-Temporal Methods for Analysis of Implications of Natural Hazard Risk

by

Esther Boyle

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved September 2023 by the
Graduate Supervisory Committee:

Petar Jevtić, Co-Chair
Nicolas Lanchier, Co-Chair
Shiwei Lan
John Fricks
Melanie Gall
Susan Cutter
Dan Cheng
Paul McNicholas

ARIZONA STATE UNIVERSITY

December 2023

ABSTRACT

As the impacts of climate change worsen in the coming decades, natural hazards are expected to increase in frequency and intensity, leading to increased loss and risk to human livelihood. The spatio-temporal statistical approaches developed and applied in this dissertation highlight the ways in which hazard data can be leveraged to understand loss trends, build forecasts, and study societal impacts of losses. Specifically, this work makes use of the Spatial Hazard Events and Losses Database which is an unparalleled source of loss data for the United States.

The first portion of this dissertation develops accurate loss baselines that are crucial for mitigation planning, infrastructure investment, and risk communication. This is accomplished thorough a stationarity analysis of county level losses following a normalization procedure. A wide variety of studies employ loss data without addressing stationarity assumptions or the possibility for spurious regression. This work enables the statistically rigorous application of such loss time series to modeling applications.

The second portion of this work develops a novel matrix variate dynamic factor model for spatio-temporal loss data stratified across multiple correlated hazards or perils. The developed model is employed to analyze and forecast losses from convective storms, which constitute some of the highest losses covered by insurers. Adopting factor-based approach, forecasts are achieved despite the complex and often unobserved underlying drivers of these losses. The developed methodology extends the literature on dynamic factor models to matrix variate time series. Specifically, a covariance structure is imposed that is well suited to spatio-temporal problems while significantly reducing model complexity. The model is fit via the EM algorithm and Kalman filter.

The third and final part of this dissertation investigates the impact of compounding hazard events on state and regional migration in the United States. Any attempt to capture trends in climate related migration must account for the inherent uncertainties surrounding

climate change, natural hazard occurrences, and socioeconomic factors. For this reason, I adopt a Bayesian modeling approach that enables the explicit estimation of the inherent uncertainty. This work can provide decision-makers with greater clarity regarding the extent of knowledge on climate trends.

DEDICATION

I am extremely grateful to my advisors Dr. Petar Jevtić and Nicolas Lanchier, along with the members of my committee for their support and feedback.

Special thanks also to my my friends and family near and far for their patience and unwavering support and encouragement. You know who you are.

For Bryce and Neo: you are my whole world and I wouldn't be here without you.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Background

Severe disaster events resulting from natural hazards pose a great risk to life and communities' well-being worldwide. Over the last decade, a multitude of studies have investigated trends in losses from natural hazard events, together with their societal implications. Disaster related losses have multifaceted ramifications, for example on infrastructure (Boyle, Inanlouganji, et al., 2022), national security (Boyle, Chiaradonna, et al., 2022), criminal activity (Jevtic and Gall, 2023), and migration patterns (Dodman et al., 2022). Thus far, despite the high variance and uncertainty of disaster loss datasets, studies agree that losses have increased over time (Cutter and Emrich, 2005; Dodman et al., 2022; Gall et al., 2011). There is also strong scientific evidence that climate change will continue to have significant effects on the frequency and severity of extreme events (Abatzoglou and Williams, 2016; Knutson et al., 2019). Yet, it is difficult to discern to what degree climate change has already impacted hazard related losses and their subsequent societal implications (R. Pielke, 2020; R. A. Pielke and Sarewitz, 2005). There is additional uncertainty regarding and how these patterns will evolve in the future (R. Pielke, 2020; R. A. Pielke and Sarewitz, 2005).

Identifying the driving forces of these losses, forecasting, and estimating societal impacts are ongoing statistical endeavors. Disaster loss data is spatio-temporal in nature, and thus studies that make use of it must apply appropriate spatio-temporal statistical techniques. Misunderstanding of the data nuances, biased modeling assumptions, and inappropriate choices of statistical methodology can all lead to inappropriate application of

these foundational datasets (Gall et al., 2009). Thus, the work in this dissertation aims to illustrate effective spatio-temporal approaches to model disaster losses and their societal impacts.

In light of the evolving spatio-temporal patterns of losses resulting from natural hazards over the last few decades, what statistical methodologies are most effective for forecasting losses and understanding societal impacts? This dissertation is organized into three chapters that together aim to answer this question. The first section provides background by conducting a thorough investigation into hazard trends in the United States at the county level, stratified by hazard type. Normalization techniques and statistical tests are applied to analyze statistical properties of the observed losses and detect underlying trends, which can have ramifications for forecasting and application studies. In the second section, a novel matrix variate dynamic factor model is developed for forecasting losses that is able to account for both the unobserved underlying drivers of losses, and the high dimensionality of the data. The final chapter illustrates the strength of Bayesian approaches for application studies in the presence of uncertainty regarding natural hazards. Specifically, a carefully constructed Bayesian regression model is chosen to study the impacts of compounding hazard events on state and regional migration in the United States. Together, these three avenues of research combine rigorous statistical analysis with advanced forecasting and modeling methodologies to provide valuable insights into the evolving landscape of natural hazard losses and their societal impacts. By shedding light on historical trends, current realities, and potential future scenarios of losses and their societal repercussions, this dissertation provides practical statistical tools for making well-informed decisions regarding natural hazard preparedness.

1.2 Data

Throughout this research, the primary data source is The Spatial Hazard Events and Losses Database Version 21 (SHELDUS V. 21.0) (Cutter et al., 2008; SHELDUS, 2023), which is the premier database on U.S. hazard losses. SHELDUS V. 21.0 contains hazard data for the United States from 1960 through 2021, and covers a variety of natural hazards, such as hurricanes, wildfires, and thunderstorms. SHELDUS contains records of events that had resulting property losses, crop losses, injuries, and fatalities.

Data in SHELDUS is available for download in two different forms: raw or aggregated. Raw data records in SHELUDS are county level event records, and give the associated losses, injuries, and fatalities from a specific hazard event in a given county, along with the type of hazard(s) that occurred, and the beginning and ending date of the event. Other details are also available, such as associated perils, whether the event was given a Presidential Disaster Declaration, and other specific notes regard the event. SHELDUS data is also available in spatio-temporal aggregated form, where records are aggregated spatially to the county or state level, or to entire U.S. level, for a given time period, such as for a given month or year. One can also form, for example, a time series of total property losses associated with hurricanes for each state each year from 1960-2021.

The information in SHELDUS can be leveraged to understand the intensity and evolution of disaster events. For this reason, SHELDUS is a comprehensive and valuable resource for researchers, policymakers, and emergency management professionals interested in understanding and analyzing the impact of natural hazard events in the United States. SHELDUS is maintained by the ASU Center for Emergency Management and Homeland Security. Throughout these projects, the data is used according to the terms of the End User Licence Agreement ("Spatial hazard events and losses database for the United States User Agreement", 2023).

This dissertation exemplifies the usefulness of research databases such as SHELDUS in a variety of statistical contexts. In particular, this data can be used to understand hazard trends, create comprehensive predictive models of loss, and discern relationships between hazards events and other social phenomena. The findings in this work also highlight the data limitations posed by disaster loss databases, and how to address these complications.

While the SHELDUS database remains among the most comprehensive data sets available to researchers, there is a great deal of uncertainty surrounding loss estimates in general, even in SHELDUS (Cutter et al., 2008). Failing to address these limitations can lead to biased or misinterpreted results (Gall et al., 2009). Some hazards have more consistent loss reporting practices than others. For example, prior to 1996, reports from the National Oceanic and Atmospheric Administration were reported on a logarithmic scale. In SHELDUS, these records were then translated in a conservative manner, likely leading to an underestimation in losses (Gall et al., 2009). In addition, loss values may be broken down into multiple separate loss events, even if they resulted from the same overall hazard event ("Storm Data FAQ Page", 2023). This suggests that studies relying on raw counts of disaster loss records or that attempt to use these datasets for frequency analysis suffer from severe biases.

Incorrect assumptions about the impact of climate change on events are also implicitly included in many models, or at least erroneously implied by the results of such models. Despite the expectation for increased hazard events, it is difficult to establish if changes in any of these phenomena have yet occurred (Kossin, 2018; R. Pielke, 2020). Attribution is a difficult and ongoing endeavor. Thus when it comes to loss estimation, increases cannot be justifiably attributed solely to climate change. In fact, many studies exemplify that adjustments of loss estimates to societal changes such as population and wealth eliminate the statistical significance of the observed loss increases (R. Pielke, 2020; Weinkle et al., 2018).

From a modeling perspective, if it is determined that a given time series of losses are stationary following a normalization procedure, then the dataset can be harnessed with reduced possibility for spurious regression (Everitt et al., 2010). However, a wide variety of studies employ these loss datasets without addressing stationarity assumptions or the possibility for spurious regression.

Thus it is important to develop both appropriate time series of losses, and accurate loss baselines. One product that aims to estimate baseline losses is the National Risk Index produced by the Federal Emergency Management Agency (FEMA). This data product is designed to assess and quantify disaster risk at the national, regional, state, and community levels in the United States by incorporating information from the SHELDUS database together with infrastructure, socioeconomic, and demographic variables (Zuzak et al., 2022). Response to the FEMA risk index is largely positive due to its accessibility and ease of use, however, it is not static and thus cannot be used to understand underlying trends in the composite data sources (Zuzak et al., 2022). In addition, as the methodology is not probabilistic in nature, it cannot be used to make future statements or forecasts. Another recent publication proposed a disasters index based on the National Oceanic and Atmospheric Administration's "Storm Data" product (Mahanama et al., 2021). These types of tools are crucial for mitigation planning, infrastructure investment, and risk communication (FEMA, 2023). In fact, in December 2022, U.S. Congress passed The Community Disaster Resilience Zones Act of 2022, which requires the annual calculation of a national risk index (Community Disaster Resilience Zones Act, 2022).

To aid in these efforts, the first portion of this dissertation aims to rigorously develop a variety of adjusted loss datasets that have been appropriately normalized and assessed for stationarity. This advancement improves on the developed loss baselines, however, it also extend beyond simple baselines by allowing a more confident application of such loss time series to modeling applications.

1.3 Modeling and Forecasting Disaster Losses

Accurate forecasts of expected hazard losses are necessary to develop appropriate investment in mitigation strategies. In fact, this is the backbone of a billion dollar industry; companies such as AIR, Verisk, and Aon make use of intricate proprietary datasets and climate models for use by the insurance and reinsurance industries. These models inform insurers about the risk of their portfolios based on increased risk to property and life in certain areas. They also can be used to build specialized insurance products and price catastrophe bonds ("The Verisk Severe Thunderstorm Model for the United States", 2022).

Alongside the nuances of employing disaster loss data, a complicating factor in loss forecasting is the multitude of interdependent driving forces, some of which are unobservable. Such factors include changes in detection and reporting, population and population density, wealth, building codes, infrastructure resiliency, demographics, social vulnerability, and climatic changes (Gall et al., 2009; Martinez, 2020; Weinkle et al., 2018).

Thus the second portion of this dissertation is the development of a factor based forecasting model which does not require the explicit modeling of underlying drivers of loss. As a particular use case, I developed a specialized matrix variate factor analysis model for time series to forecast losses from severe convective storms. Joint insured losses resulting from convective storms are nearly as high as those from hurricanes ("Convective Storms: State of the Risk Triple-I Issues Brief", 2022). As these storms are becoming more frequent and severe (NASA, 2023; Sander et al., 2013), insurers should be keenly interested to understand the joint distribution of losses across space and time.

The comprehensive modeling of severe convective storms requires a uniquely tailored approach to account for complexity of the intertwined perils associated with convective storms, such as hail, high wind, tornadoes, and lightning. Each of these perils can cause significant damage, yet they are certainly correlated as they arise from the same phenom-

ena. To the author's knowledge, there have been no models in the public literature considering losses from convective storms stratified by peril. Understanding the emerging spatio-temporal distributions of losses is necessary for predicting and pricing the risks related to such events.

Thus, this research aids the understanding of this important phenomena for both the academic community and practitioners. The developed approach allows for simultaneous modeling of losses associated with different perils across space and time, while also allowing for correlation in losses. The fit values from this model may be used to identify overlapping spatial regions with similar underlying drivers of loss behavior.

1.4 Societal Implications

The final project is intended to exemplify the possibilities of using hazard loss data to study potential societal impacts. This application applies hazard data from SHELDUS to study the impact of compound hazard events on migration. Compound events consist of multiple hazards that occur simultaneously or in short succession (Hillier and Dixon, 2020; Zscheischler et al., 2018), and can lead to disproportionately extreme outcomes (AghaKouchak et al., 2020; Vahedifard et al., 2016).

This study considers some of the most pressing risk categories for compounding disaster losses, such as compounding heat, drought, and wildfire, and compounding wind and flooding (Zscheischler et al., 2018). Careful selection of appropriate metrics is necessary to avoid definitional concerns often present in disaster loss databases. Thus, we specifically consider events in the same year and state that align with the compound hazard categories defined by (Zscheischler et al., 2018). From this framework, two types of measures are constructed that attempt to capture the impact of compounding hazard events over time from the SHELDUS Version 21.0 database: dollar losses and counts. We also incorporate a minimum cutoff of disaster losses to ensures that the count of loss events is not inflated

by small entries, but rather focuses on more substantial disasters. This approach serves as a significant improvement over previous studies of migration and repeated events, which have simply relied on decadal aggregated data (Boustan et al., 2020; Saldaña-Zorrilla and Sandberg, 2009).

This analysis takes a Bayesian approach while making use of spatio-temporal panel data. Previous studies into climate related migration, particularly those aimed at detecting and forecasting temporal trends, suffer from widely varying results and scientific uncertainty (Beyer et al., 2022). Enhancing comprehension of these trends to the fullest extent possible would prove immensely valuable from a policy perspective. However, any attempt to capture such trends must account for the inherent uncertainties surrounding climate change, natural hazard occurrences, and other contributing factors to migration (Azose and Raftery, 2015).

By adopting a Bayesian modeling approach, this study enables the estimation of the inherent uncertainty, providing decision-makers with greater clarity regarding the extent of knowledge on climate trends. The ability to account for uncertainty is a powerful strength of Bayesian models, making them an invaluable tool for improving decision-making in the face of complex, uncertain systems like climate change (Gelman et al., 2013; Gelman and Hill, 2006).

In the context of hazard related migration, these findings can assist in emergency response and disaster management planning for displaced populations, encompassing both preparation for changes in population as well as provision of aid and resources to affected communities. This model can additionally be leveraged to identify areas that are particularly vulnerable to compound hazards. This is especially important given that migration can be a crucial driver of the nonuniform distribution of climate vulnerability (Cai et al., 2016; Cattaneo and Peri, 2016; Kaczan and Orgill-Meyer, 2020).

Chapter 2

DEVELOPMENT AND STATIONARITY ASSESSMENT OF NORMALIZED COUNTY LEVEL NATURAL HAZARD LOSSES

As climate change unfolds, scientific consensus projects increased frequency and severity of natural hazards such as hurricanes (Knutson et al., 2019), drought (Spinoni et al., 2013), and wildfires (Abatzoglou and Williams, 2016; NASA, 2023). In addition, there is ample evidence that losses from natural hazards have been increasing both globally and in the US in recent decades (Botzen et al., 2019; Gall et al., 2011). However, determining whether detectable changes in physical hazard phenomena have already occurred is challenging (Kossin, 2018), making it difficult to attribute losses to climate change (R. Pielke, 2020).

Despite the frequent reports linking increased disasters to climate change (Gramling, 2022; Lopez, 2022; Milman et al., 2021), some studies suggest that thus far, increased disaster losses are primarily or even wholly the result of increased exposure due to societal changes such as increased population, wealth, and infrastructure (R. Pielke, 2020; Weinkle et al., 2018). It has also been evidenced that the varied vulnerability of communities over space and time can significantly impact losses from these types of events (Mechler and Bouwer, 2014).

These underlying factors complicate attribution research aiming to link hazard losses to changes in climate. However, these confounding factors do not necessarily mean it is impossible to detect loss trends due to changes in hazard phenomena, whether now or in the future (R. Pielke, 2020; Sander et al., 2013). The ability to detect these trends will likely continue to be hazard dependent, as projections for changes in hazard frequency and severity are not uniform (NASA, 2023; Visser and Petersen, 2012). As climate change does

begin to have a detectable impact on disaster loss data, there will be strong implications for risk to human life and property. Thus it is worth continual investigation to assess whether changes in losses are associated with detected changes in natural hazard phenomena.

This study conducts a thorough investigation into hazard trends in the United States at the county level, stratified by hazard type. This is the first such study comprehensively investigating different types of hazards at a local scale across the US. In this analysis, a spatially disaggregated time series of county level losses is normalized to 2022 U.S. dollars (USD) by adjusting for inflation and changes in population and wealth. Statistical tests are then applied to analyze stationarity and detect residual trends in hazard losses at the local level. This analysis highlights the underlying autocorrelation and growth in variance that can impact the appropriate use of these time series in future studies.

2.1 Literature Review

The primary methodology that has emerged to investigate trends in hazard losses is the so called "normalization" approach, which aims at estimating the loss that a historical event would cause if it occurred today. This is accomplished by adjusting historical loss values for changes in inflation, population growth, and wealth (Gall et al., 2009; R. A. Pielke et al., 2008; Weinkle et al., 2018). After accounting for societal changes driving losses via normalization, any remaining trend in losses could potentially be driven by changes in disaster occurrences resulting from climate change. Viewed in this light, any such trends in loss data should be consistent with trends in related weather extremes (R. Pielke, 2020).

Loss normalization studies thus far have been concentrated on a few types of hazards such as floods (Downton et al., 2005) and hurricanes (Martinez, 2020; Weinkle et al., 2018), with very little information available about trends from other hazards, such as wildfires (Doerr and Santın, 2016). These studies are also typically carried out for a specific region or country (R. Pielke, 2020), rather than analyzing spatio-temporally stratified data. This is

in part due to the data limitations posed by many of the available datasets, such as the lack of spatially disaggregated economic measures (Botzen et al., 2019).

Findings thus far have suggested no upward trend in hurricane losses over the previous decades (Weinkle et al., 2018). However, there has been a detectable increase in losses related to convective storms, even after normalization (Sander et al., 2013). Meanwhile, one study of SHELDUS data from 1960-2009 exemplified that even after adjusting for inflation, population growth, and wealth, per capita direct losses showed a clear upward trend (Gall et al., 2011). These varying results, which were also highlighted in a 2020 review of the normalization literature (R. Pielke, 2020), suggest that trends in hazard losses are not uniform spatio-temporally or across hazard types.

The majority of publications that adopt a normalization approach have the intent of determining stationarity or trends, which can then be used in future attribution studies regarding climate change or vulnerability. However outside of the normalization literature, there are many studies that make use of hazard loss data without any investigation into their stationarity. From a modeling perspective, if it is determined that a given time series of losses are stationary following a normalization procedure, the dataset can be harnessed with reduced possibility for spurious regression (Everitt et al., 2010). Unfortunately these datasets are often utilized inappropriately due to a misunderstanding of the data nuances and inappropriate choices of statistical methodology (Gall et al., 2009).

2.2 Methodology

Past research has typically standardized losses on a per-event basis or on an annual basis (R. Pielke, 2020; Weinkle et al., 2018). However, using disaster loss databases like SHELDUS can pose a challenge when it comes to defining what exactly constitutes an ”event.” This is because some loss events are broken down into multiple records in the database (Gall et al., 2009; “Storm Data FAQ Page”, 2023). For example, if a tornado

crosses county lines or lifts off the ground for more than a few minutes before touching back down, it is considered a new tornado ("Storm Data FAQ Page", 2023). This feature of the data can make it difficult to determine whether changes in event size are simply due to changes in reporting. For this reason, this stationary analysis examines losses at the annual aggregate scale instead of scrutinizing events at the individual level.

Thus in this study, multiple time series of total annual property damage segmented by location and hazard type are constructed from records in the SHELDUS Version 21.0 database (SHELDUS, 2023). These data series are subsequently normalized to 2022 U.S. Dollars by adjusting for changes in population and wealth using data from the Bureau of Economic Analysis and the U.S. Census Bureau. Luckily, increased data availability in recent years makes localized studies of both hazard trends and socioeconomic trends possible.

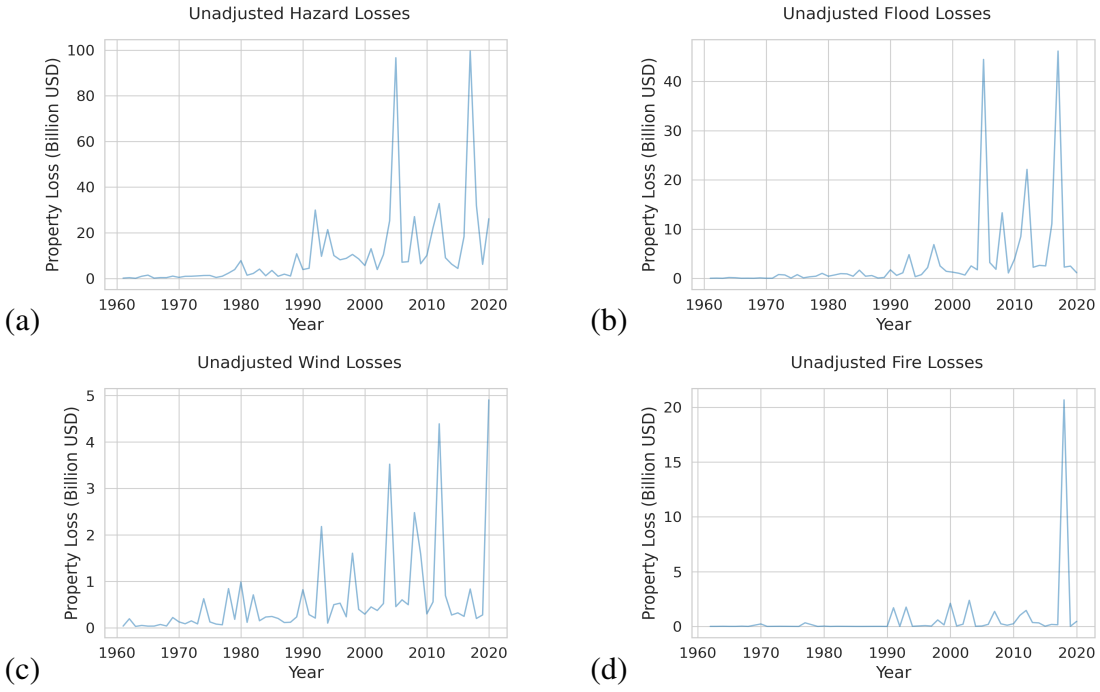

Figure 2.1: Property losses in unadjusted U.S. Dollars resulting from (a) all hazards in SHELDUS (b) floods (c) wind (d) wildfires. Data from SHELDUS V. 21.0 database (SHELDUS, 2023).

2.2.1 Normalization

We adopt a well established normalization technique (R. Pielke, 2020; Weinkle et al., 2018). The method was posed in 2005 by Pielke and Landsea, and adjusts loss values for inflation, changes in real wealth per capita, and changes in population:

$$D_{2022} = D_y \times I_y \times RWPC_y \times P_{2022/y} \tag{2.1}$$

Here, D_y is the total damage in year y, and D_{2022} represents the damage adjusted to 2022 U.S. Dollars. I_y is an inflation adjustment, here implemented using the gross domestic product price deflator produced by the Bureau of Economic Analysis. The GDP deflator measures changes in prices paid for goods and services produced in the United States ("Prices amp; inflation", n.d.). The $RPCW_y$ term stands for "Real wealth per capita", and is intended to adjust loss values to account for the growth in exposure due to increased material possessions. To represent this wealth, we adopt the current-cost net stock of fixed assets and consumer durable goods produced by the Bureau of Economic Analysis ("National Data Fixed Assets Accounts Tables", n.d.). Finally, $P_{2022/y}$ adjusts for the change in population within a given county of interest, using population data from the U.S. Census Bureau.

2.2.2 Data Selection

The SHELDUS database goes back to 1960, however reporting in some counties is more consistent than others. Thus for each county and hazard type, a dynamic subset of the time series is chosen based on data availability. We filtered out all counties that were formed or dissolved over the course of the study period. Additionally, losses in counties of Alaska, Hawaii, and Virginia are omitted due to consistent data quality and frequently

changing county lines. This filtering resulted in 2983 considered counties. All 2983 of these counties were considered for stationarity tests for the case of all hazards combined.

It is important to note that removing the observations from Alaska, Hawaii, and Virginia can introduce bias, particularly if the loss trends are very different in those states than in the rest of the U.S. If these states were able to be including in this stage of the the analysis, it would increase the number of tests performed. As further detailed in the next section, performing additional tests could affect statistical significance when adjustments for multiple comparisons are made. Nevertheless, this work can provide valuable insights and initial evidence to loss trends across much of the U.S.

Not all hazards are observed in all counties, however, it can be difficult to discern when zero losses result from truly no events, or from a lack of reporting. For this reason, time series that began with a series of zeros (leading nulls) were subset to begin at the year of first reported losses, under the assumption that the given county did not begin reporting until later in the study period. When the leading nulls were dropped, all counties still had decades of sufficient data to run the tests for overall losses.

Data for specific hazards is more sparse, in particular because most counties do not observe all hazard types. For this reason, only regions with least 10 data points of that particular hazard type were considered for the analysis. In addition, a series of robustness test were conducted. Data was considered with both dropping all nulls, and setting nulls equal to zero. In addition, the analysis was repeated with restricted data from 1995 to present in order to reduce the effects of known historical changes in reporting.

2.2.3 Stationarity Analysis

After normalizing each county level hazard specific time series to current U.S. Dollars, stationarity of the resulting normalized time series is assessed. There are a number of statistical tests that can be used so assess time series stationarity under different assump-

tions, such as the Durbin-Watson Test and the Augmented Dickey Fuller Tests (Palma, 2016). However, it is often not realistic to assume normality of hazard loss data. Thus as in (Weinkle et al., 2018), we first adopt the non-parametric Mann-Kendall statistic (Hollander et al., 2015) to test for monotonic trend. The Mann-Kendall test is a variant of Kendall's Sign test for independence of two random variables. The test statistic based upon pairwise comparison of each point in the time series to those after it. Consider:

$$K = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{sign}(Y_j - Y_i)$$

where

$$\text{sign}(Y_j - Y_i) = \begin{cases} 1 & \text{if } Y_j > Y_i \\ 0 & \text{if } Y_j = Y_i \\ -1 & \text{if } Y_j < Y_i \end{cases}$$

If K is positive, this suggests that the later observations of the time series Y_j tend to be larger than the earlier time periods, Y_i, and thus a positive trend could be present. Meanwhile, if K is significantly negative, this suggests the sequence is decreasing (Hollander et al., 2015).

While trend detection has been the primary focus of previous normalization studies (R. Pielke, 2020; Weinkle et al., 2018), the presence of autocorrelation or heteroskedastic errors has consequences for studies that aim to make use of SHELDUS derived time series. Thus, we next fit a simple linear model, with or without trend as suggested by the Mann-Kendall test. Autocorrelation tests follow from the Durbin-Watson statistic (Durbin and Watson, 1950):

$$d = \frac{\sum_{t=2}^{T} (e_t - e_{t-1})^2}{\sum_{t-1}^{T} e_t^2} \tag{2.2}$$

Here, e_i are the model residuals and T is the number of observations in the time series. The statistic d lies between 0 and 4, with a value near 2 suggesting no autocorrelation. The test against positive serial correlation is most common. In the case of positive serial correlation, error terms propagate into future observations. Following the Durbin-Watson test, I analyzed heteroskedasticity using the nonparametric Goldfeld-Quandt test (Goldfeld and Quandt, 1965). In the nonparametric version of the Goldfeld-Quandt test, "peaks" in the model residuals are counted, where a residual e_j defined as a peak if $e_j \geq e_i$ for all $i < j$. It is a form of permutation test, thus the number of peaks is compared to the number of peaks that would occur under different permutations of the residuals. All of the discussed tests are implemented in R.

Due to the large number of tests that are run, the significance level for each set of tests was adjusted for Type 1 error using the Holm–Bonferroni method. In this instance, the p-values from m tests are sorted in ascending order and then tested against the adjusted significance level given by:

$$P_k = \frac{\alpha}{m + 1 - k}. \tag{2.3}$$

Depending on the application of these results, a more conservative or more strict version of the tests may be preferred. For example, previous studies regarding climate change have sometimes prioritized reducing type 1 error in order to illustrate the strong evidence for changes in climate (Knutson et al., 2019). However some studies instead opt to reduce type 2 error in order to provide a worse case scenario type analysis that may be of interest to policy makers (Knutson et al., 2019). For this reason, results for the both the un-adjusted $\alpha < .05$ confidence level and adjusted confidence levels are reported. In instance where the un-adjusted tests detected significance but the adjusted tests did not, we report the findings as *weak* evidence. In cases where both tests results are significant, we report *strong* evidence of the phenomena.

2.3 Results

Figure 2.2 summarizes the results of the Mann Kendall tests for trend in normalized losses, where all hazard types in SHELDUS are considered. There were 2983 considered for analysis, and 823 counties with detected monotonic trends. However after adjusting for multiple comparisons, only 19 counties remained significant. All of these counties, reported in Table 2.1, were detected to have a decreased trend in losses.

It is important to note that a detected *negative* trend in normalized losses does not necessarily imply decreasing overall dollar losses in an area when adjusted for inflation only. Rather, a negative trend can result when the growth in population and wealth in an area outpaces the increased or stationary losses observed. Thus these results reveal that normalization techniques were able to account for the vast majority of growth in losses across the counties considered in the U.S over the last 80 years.

Note however the counties in Virginia and elsewhere such as Maimi-Dade that were excluded from the study shown in white in figure 2.2. These excluded counties may be a source of unintended bias. In the future, further methodology could be developed that provides rules for allocating losses in counties with a history of changing boundaries. For example, losses could be disaggregated to the historical census block level before re-aggregating to the new county lines.

Figure 2.3 depicts the counties that detected significant autocorrelation in normalized losses. However, no counties showed evidence of autocorrelation when adjusting for multiple comparisons. Heteroskedasticity is revealed to be more of a concern, as shown in figure 2.4. Here, 281 counties showed evidence of heteroskedasticity ($p < .05$), and 41 counties having strong evidence of heteroskedasticity after adjusting for multiple comparisons.

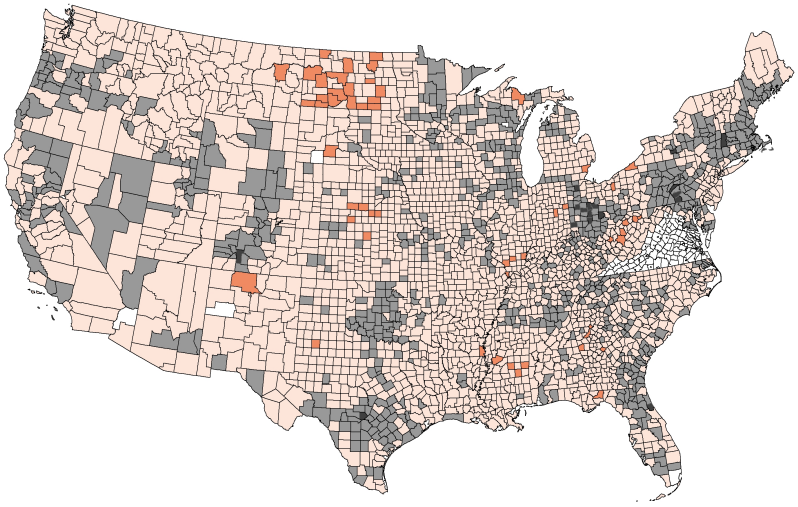

Figure 2.2: Detected monotonic trends in normalized losses.
White: not considered. Pink: considered for analysis, no evidence of trend.
Grey: weak evidence of decreased trend. Dark grey: strong evidence of decreased trend.
Orange: weak evidence of increased trend.

Table 2.1: Counties with detected decreased trend in normalized losses after multiple comparisons procedure.

State	County	State	County
CO	Hinsdale	OH	Fairfield
FL	Flagler	OH	Fayette
MA	Berkshire	OH	Henry
MA	Dukes	OH	Madison
MA	Nantucket	OH	Shelby
NY	Schuyler	OH	Union
OH	Butler	PA	Adams
OH	Champaign	PA	Cumberland
OH	Delaware	PA	Juniata
		TX	Kendall

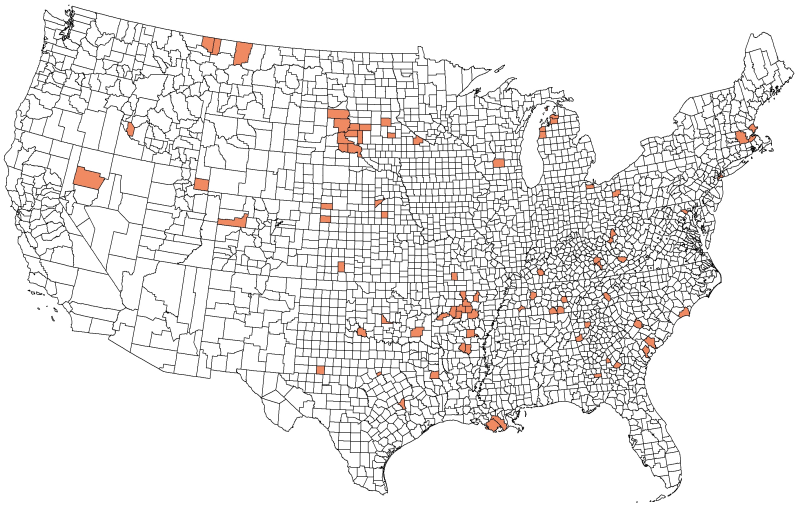

Figure 2.3: Detected autocorrelation in normalized losses.
Orange: weak evidence of autocorrelation.
White: no evidence of autocorrelation.

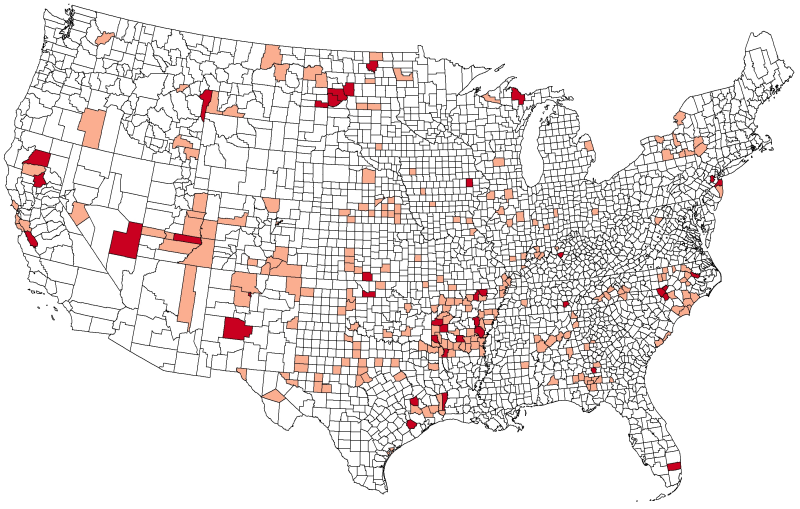

Figure 2.4: Detected heteroskedasticity in normalized losses.
Red: strong evidence of heteroskedasticity.
Orange: weak evidence of heteroskedasticity.
White: no evidence of heteroskedasticity.

2.3.1 Hazard Specific Results

At the individual hazard level, significant decreasing trends were detected for many of the hazards, including hail and coastal events. However, after robustness checks and subsetting the data from 1995 onward, no strong evidence of trend was detected in any counties for coastal events, heat, drought or floods. This suggests not only that the normalization techniques were able to adequately account for the changes in losses, but that they may be over-emphasizing the trends in population and wealth in light of the conservative estimates of loss reported in SHELDUS (Gall et al., 2009). The results from the robust tests are given in tables 2.5 and 2.3.

Table 2.2: Counties with detected trends, autocorrelation, and heteroskedasticity in normalized losses from specific hazards.

Hazard	Coastal	Drought	Flooding	Fog	Hail	Heat
Total Counties	56	397	2,867	202	2,206	171
Negative Trend ($p < .05$)	16.07%	15.87%	14.79%	8.42%	23.07%	0.00%
Negative Trend ($p_k < .05$)	1.79%	0.00%	0.07%	0.00%	0.45%	0.00%
Postitive Trend ($p < .05$)	0.00%	0.00%	0.00%	0.00%	0.18%	0.00%
Positive Trend ($p_k < .05$)	0.00%	0.00%	0.42%	0.00%	0.00%	0.00%
Autocorr. ($p < .05$)	3.57%	7.56%	6.00%	6.44%	6.39%	2.34%
Autocorr. ($p_k < .05$)	0.00%	4.53%	0.91%	0.99%	1.41%	1.17%
Hetero. ($p < .05$)	7.14%	0.00%	14.4%	0.00%	7.48%	0.58%
Hetero. ($p_k < .05$)	3.57%	0.50%	13.67%	1.49%	6.53%	1.17%

Table 2.3: Counties with detected trends, autocorrelation, and heteroskedasticity in normalized losses from specific hazards.

Hazard	Hurricane	Landslide	Lightning	Severe Storms	Tornado	Wildfire
Total Counties Considered	797	303	2,038	2,873	2,516	8.96%
Negative Trend ($p < .05$)	6.65%	5.94%	19.63%	27.98%	12.04%	0.00%
Negative Trend ($p_k < .05$)	0.00%	0.00%	0.05%	0.42%	0.00%	0.00%
Postitive Trend ($p < .05$)	0.00%	0.00%	0.05%	2.58%	0.20%	0.00%
Positive Trend ($p_k < .05$)	0.00%	0.00%	0.00%	0.00%	0.00%	1.85%
Autocorr. ($p < .05$)	1.00%	4.29%	6.28%	7.90%	4.81%	0.28%
Autocorr. ($p_k < .05$)	1.00%	1.32%	1.18%	0.52%	0.87%	0.00%
Hetero. ($p < .05$)	4.14%	3.30%	5.45%	9.29%	11.25%	0.00%
Hetero. ($p_k < .05$)	5.52%	5.94%	1.82%	1.84%	6.40%	0.00%

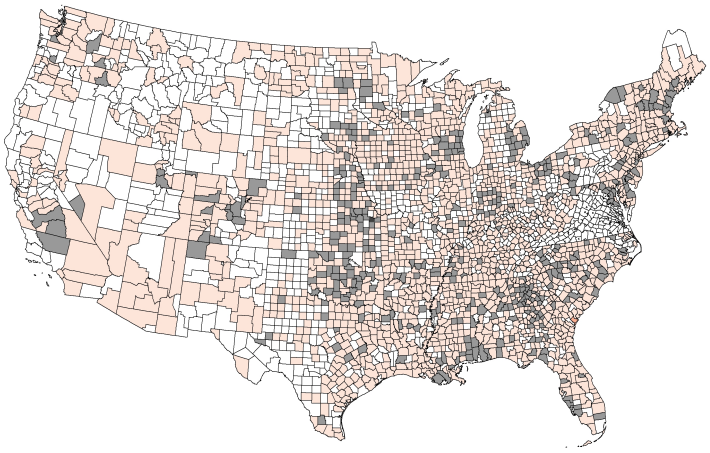

Figure 2.5: Detected monotonic trends in normalized lightning losses.
Light orange: Considered county. Light Grey: Weak evidence of decreased trend.

Autocorrelation and heteroskedasticity continue to have weak evidence at the hazard level. While these effects are far from universal, the results still suggest that studies making use of SHELDUS data should take precaution and adjust for these possibilities. For example, when analyzing losses from lightning, 11.25% of the counties considered had weak evidence of heteroskedasticity and 6.4% had strong evidence of heteroskedasticity. In Chapter 3, this heteroskedasticity in losses from lightning will be adjusted for before forecast modeling.

2.4 Discussion

The outcomes of our normalization procedure corroborate conclusions drawn from prior research (R. Pielke, 2020), emphasizing that the increase in losses can predominantly be ascribed to increases in population and economic affluence. However, our investigation also highlights the potential for certain normalization procedures to be overly stringent. Notably, in instances where heightened exposure arises from factors such as population and wealth, there might simultaneously be a mitigation in exposure due to reinforced structures and infrastructure (Mechler and Bouwer, 2014). Thus our resulting data and findings can serve as a building block for more elaborate studies into the role of vulnerability in hazard trends over time (Mechler and Bouwer, 2014). Another avenue for future exploration might make use of indirect measures of loss such as impacts on the labor market and loss of ecosystem services (Barbier, 2012; Botzen et al., 2019).

These possibilities highlight that the choice of which measures to include in normalization may have a significant impact on trend detection. In particular, an alternative normalization approach was posed by Collins and Lowe (Collins, 2001; Weinkle et al., 2018), and adopts an alternate metric to account for changing wealth and population:

$$D_{2022} = D_y \times I_y \times RWPHU_y \times HU_{2022/y} \tag{2.4}$$

Here, $RWPHU_y$ stands for the real wealth per housing unit, and $HU_{2022/y}$ is an adjustment for the change in the number of housing units. In this approach, the current-cost net stock of fixed assets and consumer durable goods is still used as a base measure of wealth, but it is adjusted to wealth per-housing unit rather than per-individual. Then, the damages are adjusted to the current number of housing units as estimated by the U.S. Census Bureau. This approach was originally suggested because the change in exposed property has exceeded population growth in certain areas, particularly along coastlines (R. A. Pielke et al., 2008). However, our results suggest no trend in coastal related losses after normalization for population alone. Future exploration of exposed property would likely benefit from the inclusion of other metrics to control for vulnerability.

One potential source of bias in this study is the removal of counties that were created, dissolved, or altered over the course of the study. If these counties were included in the analysis as they are currently available in SHELDUS, there would likely be significant bias manifested as idiosyncratic shocks in population, wealth and and losses in the affected location. However, if these states were able to be including in this stage of the the analysis, it would increase the number of tests performed, affecting the significance level when adjusting for multiple comparisons. In future work, further methodology could be developed that provides rules for allocating losses in geographical areas with a history of changing boundaries in a manner that preserves continuity of the population. For example, losses could be disaggregated to the historical census block level before re-aggregating to the new county lines, allowing for analysis over the entire study period within the same geography. Future studies could also aggregate to a higher level, such as state or U.S. level to see if similar patterns persist and various levels of analysis.

One specific application of this research is to attribution studies, which aim to link residual loss trends to changing hazard occurrences. A limitation of these types of studies is the relatively limited record of loss data on the temporal scale that can be leveraged to

detect climatic effects (R. Pielke, 2020). A strength of this study is that the loss reports go back to 1960 for much of the U.S. However this analysis reveals that data quality can still be a significant concern when trying to detect underlying trends. Robustness checks can prove beneficial in cases with many zero or missing observations.

This research has important implications for future studies relying on hazard loss data. Regression type studies that aim to link hazard losses to other phenomena face the possibility for spurious regression if the underlying time series are not stationary (Everitt et al., 2010). In addition, loss datasets may be highly correlated with socioeconomic variables corresponding to loss exposure and vulnerability, making it difficult to interpret results when multiple of such metrics are included (Hoffmann et al., 2021; R. Pielke, 2020). This analysis suggests that the vast amount of trends in hazard loss can be accounted for via growing population and wealth. Thus to avoid spurious regression, application studies need to account for these factors.

However even after accounting for trends in population and wealth, hazard loss datasets are not necessarily stationary. While autocorrelation and heteroskedasticity were not a significant issue for every county, these data anomalies do occur for many subsets of the data. For this reason, spatio-temporal methodologies adopted in future studies need to account for these possibilities. In the following two chapters, spatio-temporal models are developed to forecast trends in losses and study societal impacts. In both chapters, the models developed accommodate for the underlying socioeconomic factors driving the increase in losses, while addressing issues of autocorrelation and heteroskedasticity.

Chapter 3

JOINT LOSS MODEL FOR CONVECTIVE STORM PROPERTY DAMAGE: MATRIX VARIATE TIME SERIES BILINEAR FACTOR ANALYSIS APPROACH

Property losses associated with natural hazards such as wildfires, hurricanes, and thunderstorms are among the most common and expensive losses considered by insurers ("Convective Storms: State of the Risk Triple-I Issues Brief", 2022; Lyubchich et al., 2019). In particular, severe convective storms (also known as severe thunderstorms) may result in a combination of perils at once, such as tornadoes, hail, high wind, and lightning (NOAA, 2022; Thorson, 2020). As figure 3.1 illustrates, in many years, joint insured losses resulting from convective storms are nearly as high as those from hurricanes ("Convective Storms: State of the Risk Triple-I Issues Brief", 2022; Lyubchich et al., 2019). Exposure to these natural hazards is only expected to increase with climate change (Brooks, 2013; NASA, 2023; Pinto et al., 2012), and with socioeconomic changes (R. Pielke, 2020; Pinto et al., 2012). Thus, understanding the emerging spatio-temporal distributions of losses will be

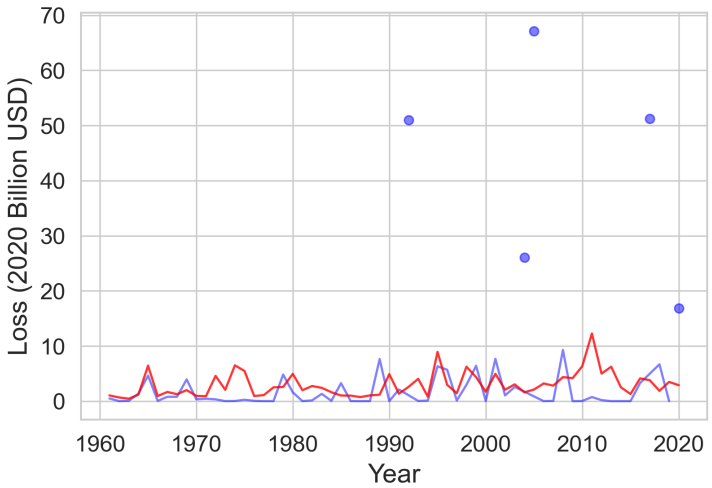

Figure 3.1: Blue: property losses from hurricanes.
Red: property losses from convective storms.

necessary for predicting and pricing the risks related to such events. Yet to the author's knowledge, there have been no studies in the public literature considering losses from convective storms stratified by perils such as high wind, tornadoes, lightning, and hail using one unified model.

The comprehensive modeling of severe convective storm losses requires a uniquely tailored approach to account for complexity of the intertwined perils. Such loss datasets are stratified in three dimensions: spatially, temporally, and by peril type. Thus, this data form a so-called *three-way* dataset, also referred to as a tensor dataset or matrix variate dataset. Three-way data occurs when variables can be stratified across three different dimensions, meaning each observation constitutes a realization of a random matrix (Coppi et al., 1989). For example in this context, losses during a given time period can be aggregated and organized into a data table with rows representing losses from different perils, and columns representing spatial regions. The collection of these annual matrix observations form a matrix variate time series.

In such matrix variate datasets, there is often a complex dependency structure between the rows and columns. In our application, there may be correlation in losses between the n spatial locations (rows) and p perils (columns), and T time points. Most univariate and

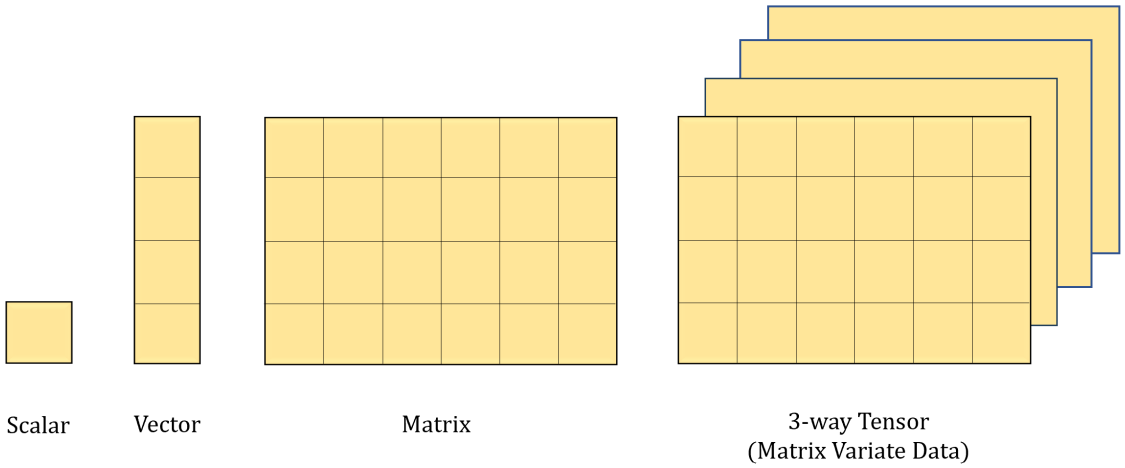

Figure 3.2: Matrix variate data structure.

multivariate models are not able to easily capture such a rich dependence structure accompanying these datasets, as a common modeling assumption in the univarite and multivarite context is independent errors. Matrix variate models are well suited to this type of problem because they are able to account for these correlated responses. The usefulness of matrix variate models in the insurance sector has recently been demonstrated (Boyle et al., in press, 2023).

The study of this complex phenomena is also impacted by a combination of observed and unobserved factors, such as climate change and economic factors (R. Pielke, 2020). Any given model of these losses could attempt to account for some of these complexities using proxy covariates, such as temperature or GDP. However, the evolution of these convective storm losses over time are also driven by underlying processes that are difficult to account for, such as changes in loss reporting (Gall et al., 2009) and the state of local infrastructure (Watson et al., 2021).

The developed methodology in this research addresses these issues through a novel matrix variate (bilinear) factor analysis model that uses dynamic factor analysis. Such an approach allows for simultaneous modeling of losses associated with different perils across space and time, thus allowing for correlation in losses. By adopting a factor based approach, the model is able to account for unforeseen forces that drive joint losses. By

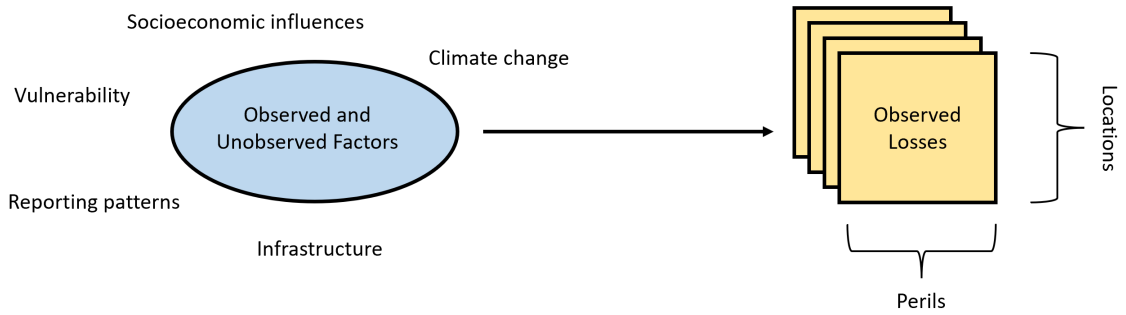

Figure 3.3: Underlying processes impacting observed matrix variate losses resulting from convective storms.

drawing on county level spatio-temporal disaster data going back to 1970, the model is be able to capture the evolving joint loss patterns over time and forecast future losses.

This approach extends the emerging matrix variate modeling literature (Chen and Lee, 2022; Ding and Dennis Cook, 2018; Viroli, 2012) by developing a dynamic factor analysis model in the matrix variate context. The developed approach gives significant gains in fit time by reducing model complexity through parameter count, while maintaining an appropriate covariance and factor loading structure. This approach is very flexible, and can be leveraged in a variety of econometric settings. The specific findings from this research also increase the understanding of the convective storm phenomena for the academic community and actuarial practitioners, with further implications for infrastructure development and mitigation planning.

3.1 Literature Review

3.1.1 Convective Storms

Severe thunderstorms occur when air quickly moves upward into the atmosphere, driven either by rising heat, colliding warm and cool air, or the interaction of the atmosphere with the surrounding terrain ("How Thunderstorms Form", 2022). While all thunderstorms result from convection and have a potential for loss, severe convective storms are significantly more dangerous with potential for much higher damage, thus we refer here to these phenomena as severe convective storms. Severe convective storms are categorized as "Severe Thunderstorms" by the National Oceanic and Atmospheric Administration, and are defined as such when hail is one inch or greater, wind gusts are above 58 mph, and/or a tornado occurs (NOAA, 2022). Although the presence of lightning is not necessary for a convective storm to be categorized as a Severe Thunderstorm, lightning may also occur resulting in direct damage.

Models that attempt to predict the meteorological aspects of thunderstorms and their related perils, such as lightning, hail, and heavy rain, are plentiful (Healy et al., 2022; Schmeits et al., 2008). Such studies of the physical phenomena sometimes rely on insurance loss data to aid their development. For example, both (Kapsch et al., 2012; Kunz and Kugel, 2015) made use of insurance loss data to verify their models of hailstorms in southwest Germany.

Alongside these investigations of the physical phenomena are studies of direct losses from individual perils related to convective storms, typically in an insurance context. For example (Changnon, 2009) demonstrated a recent increase in losses from major hail events in the U.S., although the study did not attempt to attribute these losses to any driving factors such as increased population and wealth or changes to the frequency and severity hail storms. Meanwhile, (Barredo, 2010) found no trend in windstorm losses in Europe after adjusting for changes in socioeconomic variables. In another study, wind speed data was transformed into a time series of losses from windstorms in the Netherlands (Cusack, 2012). They found that while windstorm losses were at a low, the driving climate forces of these losses had changed over the last few decades. Regarding tornadoes, (Refan et al., 2020) provided a simulation based study that incorporated tornado climatology, tornado intensity and path of travel, together with an exposure map, fragility and vulnerability functions, and a financial module. While there are a multitude of models in the literature to quantify losses due to individual perils, studies focused on joint losses from convective storms are less common. One study considered census-block-level losses from extreme cold, hail, lightning, and tornadoes in Louisiana (Mostafiz et al., 2020), albeit the analysis was statistically independent for each hazard. Many modeling approaches that jointly consider convective storm perils simultaneously while addressing correlation are proprietary ("The Verisk Severe Thunderstorm Model for the United States", 2022). This study aims to provide both a model and results that are publicly available for the research community.

3.1.2 Matrix Variate Data and Modeling

Here we give background into the literature regarding the matrix variate modeling paradigm. Consider a dataset of convective storm losses stratified across n geographical areas with p different associated perils. This data can naturally be arranged in a single $n \times p$ matrix. Now, if we observe losses stratified in this manner for T consecutive years, the result is a length T time series of $n \times p$ matrices. To account for this three-way stratification of convective storm loss data, this study makes use of a matrix-variate based model.

A straightforward distributional assumption for matrix variate data is the matrix variate normal distribution Gupta and Nagar, 2018. A $n \times p$ matrix $\mathbf{X}$ has a matrix variate normal distribution given that its probability density function is specified by:

$$f_{n,p}(\mathbf{X}|\mathbf{M}, \mathbf{\Sigma}, \mathbf{\Psi}) = (2\pi)^{-np/2}|\mathbf{\Sigma}|^{-p/2}|\mathbf{\Psi}|^{-n/2}\text{etr}\left(-\frac{1}{2}\mathbf{\Psi}^{-1}(\mathbf{X}-\mathbf{M})'\mathbf{\Sigma}^{-1}(\mathbf{X}-\mathbf{M})\right),$$

where $etr(\cdot) = e^{tr(\cdot)}$, $\mathbf{M}$ is the mean matrix, and the two matrices $\mathbf{\Sigma}$ and $\mathbf{\Psi}$ represent the within-row covariance and within-column covariance, respectively. This matrix normal distribution can be seen as a natural extension of the multivariate formulation, or as a special case of a multivariate normal distribution with a specific added covariance structure. The relationship between the multivariate and matrix variate normal distributions can be seen further through the vectorization function, which transforms a matrix into a vector. Applying the fact that $|\mathbf{\Psi} \otimes \mathbf{\Sigma}| = |\mathbf{\Sigma}|^{-p}|\mathbf{\Psi}|^{-n}$ and

$$\frac{1}{2}vec(\mathbf{X}-\mathbf{M})'(\mathbf{\Psi}\otimes\mathbf{\Sigma})^{-1}vec(\mathbf{X}-\mathbf{M}) = \text{tr}(-\frac{1}{2}\mathbf{\Sigma}^{-1}(\mathbf{X}-\mathbf{M})'\mathbf{\Psi}^{-1}(\mathbf{X}-\mathbf{M})),$$

we have the result that $vec(\mathbf{X}) \sim N_{np}(vec(\mathbf{M}), \mathbf{\Psi} \otimes \mathbf{\Sigma})$ where $\otimes$ represents the Kronecker product. Thus, if a given matrix $\mathbf{X}$ is matrix variate normally distributed, then its vectorized form is follows the multivariate normal distribution.

One primary benefit of employing the matrix variate distribution within a modeling context is the imposed structure of the covariance matrix. In this way, separate estimates

are obtained for $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$ rather than solely an estimate of the Kronecker product $\boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}$. This allows the modeler to decompose the associated correlation structure row-wise and column-wise. In this application, this corresponds to identifying the correlation structure across the intertwined perils, and the spatial correlation structure. In addition, the imposed structure reduces the number of parameters from the unrestricted multivariate form.

3.1.3 Dynamic Factor Analysis

By adopting a matrix variate framework, the developed model is better able to account for the high dimensionality of convective storm data in comparison to its multivariate counterparts, while maintaining the complex correlation structure between losses in all three dimensions. However when it comes to modeling losses from any natural hazards, including convective storms, there are also complex and interrelated underlying socioeconomic and climatic factors. Socioeconomic forces such as population, wealth, and vulnerability heavily impact losses, and may even be the primary or sole driver of losses (R. Pielke, 2020). However unlike many hazards for which losses are stationary after normalizing for socioeconomic factors (Weinkle et al., 2018), this is not the case with convective storms (Sander et al., 2013). Considering cumulative losses from all perils related to convective storms, a landmark study coupled a time series of thunderstorm potential with a time series of losses from Munich Re's NatCatService database (Sander et al., 2013). Notably, they found that recent changing patterns in thunderstorm occurrences can indeed be detected in loss data, even after accounting for socioeconomic factors.

To account for the complex underlying driving forces of these losses, we adopt a factor analysis approach. Factor analysis allows for modeling the relationship between a response variable and unobserved variables called *factors*. A typical factor analysis model for multivariate data has the following form:

$$y_i = ZX_i + v_i \qquad \text{where} \quad v_i \sim N(0, R).$$

Here, y_i is a vector of observations, X_i is a vector of q unobserved latent factors, Z is a matrix of factor loadings, and v_i is the vector of model error, which is considered to be multivariate normally distributed with covariance matrix R (Holmes et al., 2012). The factor loadings act similarly to coefficients in a regression model, as they quantify the relative importance of the unseen factors. This approach is particularly well suited to high dimensional problems, as the underlying factors can be used to extract useful lower dimensional patterns in large datasets. For example, the fit values of factor loadings may be used to form overlapping spatial regions with similar underlying drivers of loss behavior. This can allow insurers to better understand the spatial distribution of risk present in their current portfolios. Identification of regions with similar drivers of risk may also be used in the future to aid supplementary models by providing a baseline assumption of similar loss behavior in those areas.

While factor analysis is a helpful tool for dealing with high dimensional data that has unseen underlying processes, alone it is not able to account for potential correlation over time, as seen in time series data. This is because the factor analysis model assumes independent error terms v_i. Dynamic factor analysis is an extension of factor analysis that explicitly allows for these unseen factors to be evolving over time. The dynamic factor analysis model is as follows (Holmes et al., 2012):

$$y_t = ZX_t + v_t \qquad \text{where} \quad v_t \sim N(0, R)$$

$$X_t = X_{t-1} + w_t \qquad \text{where} \quad w_t \sim N(0, Q)$$

Here, we start with the same basic factor analysis model as before, where each observation comes from a different point in the time series. However, we now allow for the underlying unseen factors to be evolving in time by following a *non-stationary* random walk with error term w_t, taken to be multivariate normally distributed with mean vector 0 and covariance matrix Q. Parametric fitting of the dynamic factor analysis model follows

from the maximum likelihood framework under the multivariate normal assumption. These parameter estimates are computed using the EM algorithm, making use of the Kalman filter and smoother common to multivariate state space models (Dempster et al., 1977; Holmes et al., 2012).

Dynamic factor analysis is a special case of multivariate autoregressive state-space modeling, and has been an active area of study for many years (Ensor, 2013; Stock and Watson, 2011), however this approach has been restricted to multivariate datasets. In order to study losses from convective storms, we could simply vectorize our matrix variate data, and use the already existing dynamic factor analysis models available. However, this approach suffers in two ways. First, important structural information is held within the matrix format, as there are latent factors impacting the various perils (rows) and spatial locations (columns). By using the vector form, all of this structural information is lost. Yet this model can be viewed as form of the multivariate dynamic factor model with novel non-linear constraints.

Secondly, the dynamic factor analysis model suffers from the curse of dimensionality (Holmes et al., 2012; Stock and Watson, 2011), as there are a large number of parameters to be fit. The matrix variate approach greatly reduces the number of parameters, thus speeding up time needed to fit the model and stabilizing the estimates. For this reason, this work extends the dynamic factor analysis approach to allow for modeling of matrix variate data.

3.2 Methodology

3.2.1 Bilinear Factor Analysis

While factor analysis has been widely used in many applications involving high-dimensional vector data, matrix variate, or *bilinear* factor analysis is best suited to handle factor analysis

of matrix variate data (Xie et al., 2008). Bilinear factor analysis for matrix variate data has been growing in popularity (Gallaugher and McNicholas, 2017). While factor analysis has been considered for vector time series data (Ensor, 2013), bilinear factor analysis has not yet been considered in a time series context.

Our Dynamic Bilinear Factor Analysis model for time series has the form:

$$Y_t = AX_tB + V_t \qquad \text{where} \quad V_t \sim MVN(0, \Sigma, \Psi) \tag{3.1}$$

$$X_t = X_{t-1} + \text{diag}(w_t) \qquad \text{where} \quad w_t \sim N(0, Q) \tag{3.2}$$

Here, Y_t is an $n \times p$ matrix of loss observations in time period t from convective storms stratified across geographic regions and perils. X_t is a $q \times q$ diagonal matrix of unobserved factors, while A and B are matrices of factor loadings, and V_t is the matrix variate normally distributed model error with covariance matrices Σ and Ψ. $\text{diag}(w_t)$ is the error of the random walk taken by the unseen factors. This error is taken to be multivariate normally distributed, with mean matrix 0 and covariance matrix Q. Note that in this formulation, there are two factor loading matrices, A and B. A acts as a front loading matrix that contains factor loadings onto the rows, and thus contains information about the relative importance of each unseen factor on the different spatial locations. Similarly, B contains the factor loadings onto the columns, and thus contains information about the relative importance of the unseen factors on each peril associated with convective storms. To develop

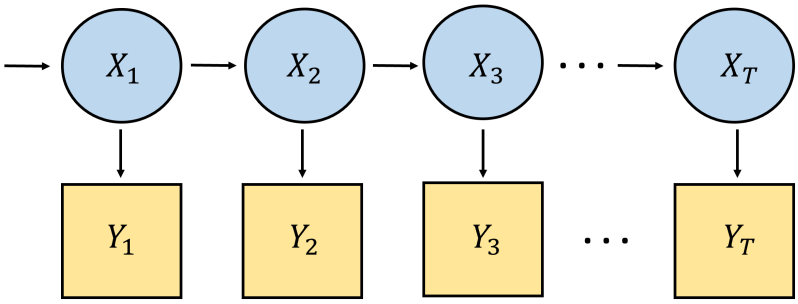

Figure 3.4: The matrix variate dynamic factor model as a state space process. Underlying unobserved forces evolve in time, but only the resulting losses are observed.

this model, maximum likelihood parameter estimates are here derived. Implementation of the EM algorithm is given, which can be implemented in any programming language of choice (Dempster et al., 1977).

The primary intent of the factor analysis approach is to account for the unseen underlying forces driving losses. Thus, these factors are extracted from the fit models to reveal spatio-temporal patterns of losses. Thus the fit factor loading matrices are used to identify regions and associated perils that have similar unobserved underlying patterns.

3.2.2 Identifiablity

In the multivariate setting, the dynamic factor model is not identifiable without constraints. This is because for any non-singular matrix F, the estimates $y_t = (ZF^{-1})(Fx_t)+v_t$ give the same model fit as $y_t = Zx_t + v_t$. It has been shown that q^2 total constraints are needed to ensure identifiability (Bai and Wang, 2014). The typical approach taken in multivariate dynamic factor analysis is to set the covariance matrix of the factors to the identity matrix ($Q = I_q$), and additionally set the ijth element of Z to zero for $j > i$ (Bai and Wang, 2014; Zuur et al., 2003). However, the bilinear dynamic factor model imposes an alternative parsimonious structure on the multivariate model. The matrix variate loading matrices A and B have nq and qp free variables, respectively. The corresponding vectorized multivariate model has a loading matrix Z with dimensions $np \times q$. Thus $npq - nq - pq = q(np - n - p)$ constraints are imposed on the multivariate model by adopting the matrix variate structure. This means as long as $2np - 2n - 2p + 1 \geq q$ with $Q = I_q$, the model is identified. Alternatively, Q could be allowed to be unconstrained with $np - n - p \geq q$.

In the factor analysis literature, it is a common approach to perform a rotation to the factor and loading matrices after model fitting to simplify interpretation of the factors. The most popular choice of rotation is the varimax, which aims to express the factor loadings on each output variable in terms of as few factors as possible (Kaiser, 1958). This rotation

approach assumes that the factors are orthogonal. To make use of the varimax rotation here we adopt the identifiability constraint $Q = I_q$ as it implies that the factors are orthogonal. Setting $Q = I_q$ also allows for more factors to be estimated, which can be helpful for datasets with smaller response matrices.

One final identifiablity concern comes from the fact that A and B are only defined to a multiplicative constant, as are Σ and Ψ. This is because if $\hat{A}$ and $\hat{B}$ are estimates of A and B, then $B' \otimes A = \frac{1}{a}B' \otimes aA$, meaning $a\hat{A}$ and $\frac{1}{a}\hat{B}$ are also estimates. The same relationship is true between Σ and Ψ. Here we impose that the first entries in A and Σ are set to 1. Note that this implies the scaled relationships between A and B and Σ and Ψ are of interest, rather than the absolute values. The nuances of interpreting these values are further discussed alongside the results in section 3.3.

3.2.3 Model Fitting via EM Algorithm and Kalman Filter

Assume $Y_1, Y_2, ..., Y_T$ form a time series of $n \times p$ observed random matrices arising from the factor relationships in (3.1) and (3.2). The complete-data likelihood is given by:

$$\prod_{t=1}^{T} f(Y_t|\mathbf{X}, \mathbf{\Theta}) \times f(x_0) \prod_{t=1}^{T} f(x_t|x_{t-1}, \mathbf{\Theta}) \tag{3.3}$$

Under the assumption of normality, the complete data log likelihood is given by:

$$\begin{aligned} L(\mathbf{\Theta}|Y_1, ..., Y_T, \mathbf{X}) = &-\frac{(np+q)T}{2}\log(2\pi) - \frac{nT}{2}\log|\Psi| - \frac{pT}{2}\log|\Sigma| - \frac{T}{2}\log|Q| \\ &- \frac{1}{2}\sum_{t=1}^{T}\mathrm{tr}(\Psi^{-1}(Y_t - AX_tB)'\Sigma^{-1}(Y_t - AX_tB)) - \frac{1}{2}\sum_{t=1}^{T}(x_t - x_{t-1})'Q^{-1}(x_t - x_{t-1}) \end{aligned} \tag{3.4}$$

Expectation Step

The complete-data expected log-likelihood is given by taking the expected value of the above:

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}') =& E_X(L(\boldsymbol{\Theta}|Y_1,...,Y_T,X_1,...X_T)) = \\ & -\frac{(np+q)T}{2}\log(2\pi) - \frac{nT}{2}\log|\Psi| - \frac{pT}{2}\log|\Sigma| - \frac{T}{2}\log|Q| \\ & -\frac{1}{2}\sum_{t=1}^{T} E_X\Big[\mathrm{tr}(\Psi^{-1}(Y_t - AX_tB)'\Sigma^{-1}(Y_t - AX_tB))\Big] \\ & -\frac{1}{2}\sum_{t=1}^{T} E_X\Big[(x_t - x_{t-1})'Q^{-1}(x_t - x_{t-1})\Big] \end{aligned} \tag{3.5}$$

Expanding, the expected log-likelihood function depends on the following terms:

$$\begin{aligned} & -\frac{1}{2}\sum_{t=1}^{T}\mathrm{tr}(\Psi^{-1}Y_t'\Sigma^{-1}Y_t) \\ & \frac{1}{2}\sum_{t=1}^{T}\mathrm{tr}(B\Psi^{-1}Y_t'\Sigma^{-1}AE_X[X_t]) \\ & \frac{1}{2}\sum_{t=1}^{T}\mathrm{tr}(A'\Sigma^{-1}Y_t\Psi^{-1}B'E_X[X_t]) \\ & -\frac{1}{2}\sum_{t=1}^{T}\mathrm{tr}(B\Psi^{-1}B'E_X[X_tA'\Sigma^{-1}AX_t]) \\ & -\frac{1}{2}\sum_{t=1}^{T}\mathrm{tr}(E_X[x_tx_t']Q^{-1}) \\ & \frac{1}{2}\sum_{t=1}^{T}\mathrm{tr}(E_X[x_{t-1}x_t']Q^{-1}) \\ & \frac{1}{2}\sum_{t=1}^{T}\mathrm{tr}(E_X[x_tx_{t-1}']Q^{-1}) \\ & -\frac{1}{2}\sum_{t=1}^{T}\mathrm{tr}(E_X[x_{t-1}x_{t-1}']Q^{-1}) \end{aligned} \tag{3.6}$$

Now note that for a diagonal matrix $X = diag(x)$, we have that $XUX = U \circ xx'$, where $\circ$ denotes the Hadamard (element-wise) product. This means that the term $E_X[X_tA'\Sigma^{-1}AX_t]$ can be rewritten as $(A'\Sigma^{-1}A) \circ E_X[x_tx_t']$. Thus in the E step, we need to calculate

the expected values $\tilde{x}_t = E_X[x_t]$ with $\tilde{X}_t = E_X[X_t] = diag(\tilde{x}_t)$, $\tilde{P}_t = E_X[x_t x_t']$, and $\tilde{P}_{t,t-1} = E_X[x_t x_{t-1}']$ with $\tilde{P}'_{t,t-1} = E_X[x_{t-1} x_t']$ for $t = 1, ..., T$.

These terms are estimated using the established Kalman filter and smoother for vector state space models (Holmes et al., 2012; Shumway and Stoffer, 1982). We adopt the following notation common in the literature (Holmes et al., 2012):

$$\tilde{x}_t = x_t^T = E_X(x_t | Y_1, ..., Y_T, \boldsymbol{\Theta}) \tag{3.7}$$

$$\tilde{V}_t = V_t^T = \mathrm{var}_X(x_t | Y_1, ..., Y_T, \boldsymbol{\Theta}) \tag{3.8}$$

$$\tilde{V}_{t,t-1} = V_{t,t-1}^T = \mathrm{cov}_X(x_t x_{t-1}' | Y_1, ..., Y_T, \boldsymbol{\Theta}) \tag{3.9}$$

$$\tilde{P}_t = P_t^T = E_X(x_t x_t' | Y_1, ..., Y_T, \boldsymbol{\Theta}) = \tilde{V}_t + x_t x_t' \tag{3.10}$$

$$\tilde{P}_{t,t-1} = P_{t,t-1}^T = E_X(x_t x_{t-1}' | Y_1, ..., Y_T, \boldsymbol{\Theta}) = \tilde{V}_{t,t-1} + x_t x_{t-1}' \tag{3.11}$$

To make use of the Kalman filter and smoother for vector state space models, the model in equations (3.1) and (3.2) can be expressed in the vectorized form:

$$\begin{aligned} \mathrm{vec}(Y_t) &= \mathrm{vec}(AX_tB) + \mathrm{vec}(V_t) \\ &= (B' \otimes A)\mathrm{vec}(X_t) + \mathrm{vec}(V_t) = \\ &= (B' \otimes A)Dx_t + \mathrm{vec}(V_t) \\ \text{where} \quad D_{ij} &= \begin{cases} 1 & i + q = j + qj \\ 0 & \text{otherwise} \end{cases} \\ \text{and} \quad \mathrm{vec}(V_t) &\sim N(0, \Psi \otimes \Sigma) \end{aligned} \tag{3.12}$$

Letting let $Z_t = (B' \otimes A)D$, the Kalman filter estimates are:

$$x_t^{t-1} = x_{t-1}^{t-1} \tag{3.13}$$

$$V_t^{t-1} = V_{t-1}^{t-1} + Q \tag{3.14}$$

$$x_t^t = x_t^{t-1} + K_t(\mathrm{vec}(Y_t) - Z_t x_t^{t-1}) \tag{3.15}$$

$$V_t^t = (I_q - K_t Z_t) V_t^{t-1} \tag{3.16}$$

$$K_t = V_t^{t-1} Z_t' (Z_t V_t^{t-1} Z_t' + \Psi \otimes \Sigma)^{-1} \tag{3.17}$$

The expectations are then computed from the Kalman smoother and lag-1 covariance smoother:

$$x_{t-1}^T = x_{t-1}^{t-1} + J_{t-1}(x_t^T - x_t^{t-1}) \tag{3.18}$$

$$V_{t-1}^T = V_{t-1}^{t-1} + J_{t-1}(V_t^T - V_t^{t-1}) J_{t-1}' \tag{3.19}$$

$$J_{t-1} = V_{t-1}^{t-1} (V_t^{t-1})^{-1} \tag{3.20}$$

$$V_{T,T-1}^T = (I - K_T Z_T) V_{T-1}^{T-1} \tag{3.21}$$

$$V_{t-1,t-2}^T = V_{t-1}^{t-1} J_{t-2}' + J_{t-1}(V_{t,t-1}^T - V_{t-1}^{t-1}) J_{t-2}' \tag{3.22}$$

Our model fitting and analysis is implemented in R, in order to make use of the Kalman Filter and smoother provided by the KFAS and MARSS packages (Holmes et al., 2012). This particular software was chosen due to the speedy implementation of the Kalman filter in C.

Maximization Step

Next, we seek to maximize the expected log-likelihood with respect to $\boldsymbol{\theta}$. Applying matrix calculus we arrive at the following estimators:

$$\hat{A} = \Big[\sum_{i=1}^{T} Y_t \hat{\Psi}^{-1} \hat{B}' \tilde{X}_t \Big] \Big[\sum_{i=1}^{T} (\hat{B} \hat{\Psi}^{-1} \hat{B}') \circ \tilde{P}_t \Big]^{-1} \tag{3.23}$$

$$\hat{B} = \Big[\sum_{i=1}^{T} (\hat{A}' \Sigma^{-1} \hat{A}) \circ \tilde{P}_t \Big]^{-1} \Big[\sum_{i=1}^{T} \tilde{X}_t \hat{A}' \hat{\Sigma}^{-1} Y_t \Big] \tag{3.24}$$

$$\hat{\Sigma} = \frac{1}{pT} \sum_{i=1}^{T} \Big[(Y_t - \hat{A} \tilde{X}_t \hat{B}) \hat{\Psi}^{-1} Y_t' - Y_t \hat{\Psi}^{-1} \hat{B}' \tilde{X}_t \hat{A}' + \hat{A} [(\hat{B} \hat{\Psi}^{-1} \hat{B}') \circ \tilde{P}_t] \hat{A}' \Big] \tag{3.25}$$

$$\hat{\Psi} = \frac{1}{nT}\sum_{i=1}^{T}\Big[(Y_t - \hat{A}\tilde{X}_t\hat{B})'\hat{\Sigma}^{-1}Y_t - Y_t'\hat{\Sigma}^{-1}\hat{A}\tilde{X}_t\hat{B} + \hat{B}'[(\hat{A}'\hat{\Psi}^{-1}\hat{A}) \circ \tilde{P}_t]\hat{B}\Big] \tag{3.26}$$

In order to obtain these estimates within the maximization step, we adopt two conditional maximization steps, resulting in an ECM algorithm. Here we partition $\boldsymbol{\theta}$ into $\boldsymbol{\theta}_1 = (A, \Sigma)$ and $\boldsymbol{\theta}_2 = (B, \Psi)$, which are updated iteratively. The E step and two conditional M steps are repeated until a convergence criteria is reached. Here, we we take the model as converged when the change in log-likelihood is less than 0.001.

Algorithm 1: Matrix Variate Dynamic Factor Model ECM

Result: Obtain estimates for X_t, A, B, Σ and Ψ for $t = 1, ..., T$

Initialize $\boldsymbol{\theta}$;

while *convergence criteria not met* **do**

E-Step: Compute $Q(\boldsymbol{\theta}|\boldsymbol{\theta}') = E(log(f(x|\boldsymbol{\theta}))|X, \boldsymbol{\theta}')$ via the Kalman Filter and Smoother;

M1-Step: Update $\hat{A}$, and $\hat{\Sigma}$ according to (3.23) and (3.25) with $\hat{B}$ and $\hat{\Psi}$ held fixed;

M2-Step: Update $\hat{B}$ and $\hat{\Psi}$ according to (3.24) and (3.26) using $\hat{A}$, and $\hat{\Sigma}$ from the M1-Step;

end

3.2.4 Data

Loss data for convective storms stratified by peril came from the SHELDUS Version 21.0 database (SHELDUS, 2023). SHELDUS draws its information about severe thunderstorms from National Climatic Data Center "Storm Data and Unusual Weather Phenomena" reports. Thus we queried records from SHELDUS with a hazard type of "Severe Thunderstorm". These records follow the definitions as laid out by the National Weather Service:

the occurrence of hail at least 1 inch in diameter, wind gusts of at least 58 mph, and/or a tornado (NOAA, 2022).

It is important to note that these definitions have changed slightly over the last century. In particular, there were competing definitions for severe weather prior to 1970, before settling on specific wind and hail requirements (Galway, 1989). As such, SHELDUS records for this study are restricted to 1970 onward. In addition, it is difficult to discern at what date the NWS began including all tornadoes in reports of severe thunderstorms. Since the current definition includes the presence of a tornado as criteria for a severe convective storm, all records of tornado losses from 1970 onward are also be included in the dataset.

One final data complication is that from 1970 until April 1, 2009, the hail criterion was set at 3/4 inch hail, as opposed to the current requirement of 1 inch hail. In practice, these cutoffs are used to specify that the hail was large enough to cause typical damage to property, as laboratory impact testing for various materials indicate a minimum threshold for damage of 1 inch (Brown et al., 2015). Unfortunately, hail size in not always reported along with loss estimates in SHELDUS. However, while smaller hail may be capable of causing damage to delicate crops, such as wine grapes (Brown et al., 2015), these forms of loss are not reflected in the property loss data from SHELDUS, but rather in the agricultural loss data in SHELDUS. Thus we chose to also include all records that result in property loss from hail.

Note that records labeled only as wind events or flood events were not included, as these events can be associated with other phenomena, such as hurricanes. However, the queried Severe Storm events may still result in wind or flood losses. Records associated only with lightning but not Severe Thunderstorm were also not included, as lightning damage can result from smaller thunderstorms that are not categorized as Severe. In addition, it is worth noting that Derechos are included in SHELDUS as both ”Severe Thunderstorm” and ”Wind” events, and thus are included in this study. Derechos are widespread and long-

lived windstorms resulting from a thunderstorm and are characterized by a line of intense, damaging straight-line winds NOAA, 2022.

After the event records were selected from SHELDUS, the losses were attributed to one of the following perils: wind, hail, lightning, tornado, and heavy rain/flooding. If a given record was associated with more than one peril, losses were split evenly among those categories. If no details were given other than "Severe Thunderstorm", the losses were split evenly across all 5 categories. After this disaggregation, the losses were aggregated up to the annual level for the spatial region of interest for each hazard type. Finally, the losses attributed to heavy rain and flooding as a result of thunderstorms was dropped. Drivers of flood loss are much more complex and thus are not considered in this study.

3.3 Results

3.3.1 Case Study: State Level Analysis for Southern Climatic Region

Here we apply the matrix variate dynamic factor approach to the three way dataset of convective storm losses obtained from SHELDUS Version 21.0 stratified by hail, wind, lightning, and tornado. First, a small observation matrix is considered by aggregating to the annual state level for the Upper Midwest Climate Region as defined by the U.S. National Center for Environmental Information. This includes Iowa, Minnesota, Michigan and Wisconsin. Thus each observation matrix has dimension 4×4, where the 4 rows correspond to the 4 spatial locations, and the 4 columns correspond to the 4 perils. Losses were log transformed and mean centered before model fitting, as dollar losses typically are heavy tailed.

A series of models were fit with varying numbers of underlying factors and different random initializations of the parameter estimates. In all iterations, the initial values for the unobserved factors were fixed at 0. The use of random initializations can help prevent the

EM algorithm from converging to a local maximum of the likelihood function rather than the global maximum. Thus 50 random initializations were considered for each number of factors. Convergence was excepted after the change in absolute log-likelihood was less than 0.001.

In line with previous studies of bilinear factor analysis (Gallaugher and McNicholas, 2017), we select the model with the lowest Bayesian Information Criterion (BIC) for comparison. The BIC is preferred in this context as it gives a higher penalty to models with a large number of parameters, while both AIC and AICc can still give too much preference overly complex models in the state space modeling context (Cavanaugh and Shumway, 1997). This is in part due to the small sample size relative to the number of parameters.

The BIC is given by:

$$\text{BIC} = k\log(\text{N}) - 2L(\hat{\boldsymbol{\Theta}}) \tag{3.27}$$

where k is the number of parameters in the model and N is the number of observations (in this case, $N = n \times p \times T$). A summary of each optimal model fit is given in Table 3.1. The BIC indicates that the model with 2 factors is preferred to reduce complexity. The resulting parameter estimates for this model are given in Tables 3.2 - 3.5, both before and after performing a varimax rotation. The varimax rotated factor loadings can be used to identify which variables each factor is most associated with.

Table 3.1: Convective storms model comparison.

Factors	Iterations	Parameters	Log-likelihood	BIC
1	32	28	-1256	2701
2	63	48	-1250	2639
3	77	56	-1250	2663
4	60	64	-1246	2678
5	189	72	-1244	2676

Table 3.2: Matrix variate factor loading matrix A, before performing a varimax rotation.

State	Loadings on Factor 1	Loadings on Factor 2
Iowa	1.00	0.33
Michigan	0.02	0.39
Minnesota	-0.07	-0.08
Wisconsin	-0.27	0.31

Table 3.3: Matrix variate factor loading matrix A, after performing a varimax rotation. Gives the effects of factors on spatial locations.

State	Loadings on Factor 1	Loadings on Factor 2
Iowa	1.032	-0.221
Michigan	0.216	0.323
Minnesota	-0.103	0.00
Wisconsin	0.00	0.403

The fits suggest that the first factor is most associated with losses in Iowa (1.032), with some effect on Michigan and Minnesota as well. The second factor is most associated with losses in Wisconsin, with smaller impacts on Iowa and Michigan. Note that in this model, the two factors are able to detect different driving forces impacting Minnesota and Wisconsin independently. When it comes to the different perils considered, after rotating the factors, the first factor is found to be most associated with hail and wind. Meanwhile, the second factor has no effect on wind, and is most closely associated with lightning. In this instance, the two factors were able to separate out underlying factors driving wind completely from the second factor. Together, these fits suggest that the two factors are picking up two major underlying trends: underlying unobserved factors driving losses from hail and wind in Iowa, and factors driving losses lighting in Iowa, Michigan and Wisconsin.

In factor analysis studies, using the varimax rotation both imposes necessary identifiability constraints, and also aids in model interpretability. However, recall that in addition in

Table 3.4: Matrix variate factor loading matrix B, before performing a varimax rotation.

Factor	Hail	Lightning	Tornado	Wind
1	0.68	0.42	-0.44	0.48
2	-0.09	0.37	0.03	-0.19

Table 3.5: Matrix variate factor loading matrix B after performing a varimax rotation. Gives the effects of factors on perils.

Factor	Hail	Lightning	Tornado	Wind
1	0.624	0.166	-0.395	0.510
2	0.277	0.540	-0.208	0.00

this application, the first entry of A was set to 1 in order to impose identifiability between the A and B matrices. The solution $\hat{A}$ and $\hat{B}$ can be re-scaled to any equally valid solution $a\hat{A}$ and $\frac{1}{a}\hat{B}$. If we chose to do so for any a, the magnitude of the entries of would change, but the interpreted relative importance of the locations for each of the factors would not. Similarly, any alternative solution would preserve the relative importance of each of the hazard types for each of the factors as given in 3.5.

However, because the identifiability constraint inherently affects the magnitude of the loadings in both the A and B matrices, care must be taken when comparing the entries *between* $\hat{A}$ and $\hat{B}$. For example, it might be tempting to conclude from Tables 3.3 and 3.5 that Iowa is more important to factor 1 than hail is. However, consider the fact that $B' \otimes A = \frac{1}{\sqrt{.624/1.032}} B' \otimes \sqrt{.624/1.032} A$. Thus $\sqrt{.624/1.032}\hat{A}$ and $\frac{1}{\sqrt{.624/1.032}}\hat{B}$ are an equally valid solution, and under this solution the factor loadings on Factor 1 for Iowa and Hail are identical. For this reason, the obtained $\hat{A}$ and $\hat{B}$ matrices are most useful in separately providing information about the importance of each location relative to the others, and each hazard relative to the others. If a specific combination of hazard and location is of interest, the entries of the matrix $\hat{B} \otimes \hat{A}$ can easily be calculated through

simple multiplication. For example, we could compute the loading on first factor for hail in Iowa using $1.032 \times 0.624 = 0.644$. The location-hazard pair loadings are not impacted by the identifiability constraint, as the choice of scaling factor a is moot after multiplication.

Aside from the magnitude of the values, the sign of each factor loading does not have inherent meaning on its own, due to the identifiability constraint, transformation of the data, and changing sign of the evolving factor states. However, they can be used to compare effects within the factors. Thus, for example, both factors indicate inverse relationships between hail and tornadoes. Meanwhile, the first and second factors have inverse effects on losses in Iowa.

Tables 3.6 and 3.7 give the estimates for the fit covariance matrices. Note that Σ and Ψ provide estimates of the remaining model error after accounting for the factors. For example, comparing $\Sigma(1,2) = 0.25$ to $\Sigma(1,3) = 0.43$, we see that the residual covariance between Iowa and Michigan is lower than the residual covariance between Iowa and Minnesota. The fact that all signs are positive across both matrices indicates that the residual errors are all positively correlated with each other. This structure suggests that for future analysis, it may be worth considering additional factors in order to reduce the residual correlations. This would however come at the price of increasing model complexity.

Next we illustrate the use of the matrix variate dynamic factor model in the context of forecasting. Using the fit parameters, 50,000 random walks of the factors were initialized

Table 3.6: Covariance matrix of observation error Σ.
Gives the variance and covariances of spatial locations.

State	Iowa	Michigan	Minnesota	Wisconsin
Iowa	1.00	0.25	0.43	0.40
Michigan	0.25	2.02	0.40	0.44
Minnesota	0.43	0.40	3.41	0.63
Wisconsin	0.40	0.44	0.63	2.29

Table 3.7: Covariance matrix of observation error Ψ.
Gives the variance and covariances of perils.

Hazard	Hail	Lightning	Tornado	Wind
Hail	1.39	0.69	0.54	0.67
Lightning	0.69	0.96	0.44	0.82
Tornado	0.54	0.44	1.00	0.43
Wind	0.67	0.82	0.43	0.87

starting at the time T smoothed estimate of the factors. These simulated factor states were then used to create simulated loss states at time $T + h$ given by $AX_{T+h}B$. These estimates were then used to construct a 95% bootstrap confidence interval. Figures 3.5 -3.8 illustrates these forecast intervals for the four perils in Iowa. Notice that the correlations between hazard types are positive, and thus forecasts for the different hazards are heavily correlated as well. In particular, these results suggest a similar pattern of losses from hail, wind, and lightning, but a slight increase in losses from tornado. This type of analysis can be used by insurers to understand future losses of a portfolio containing multiple related hazards.

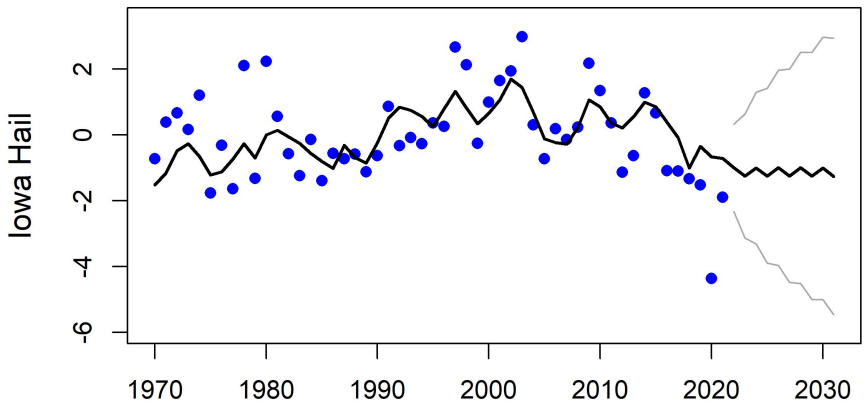

Figure 3.5: Forecast of loss from hail in Iowa.

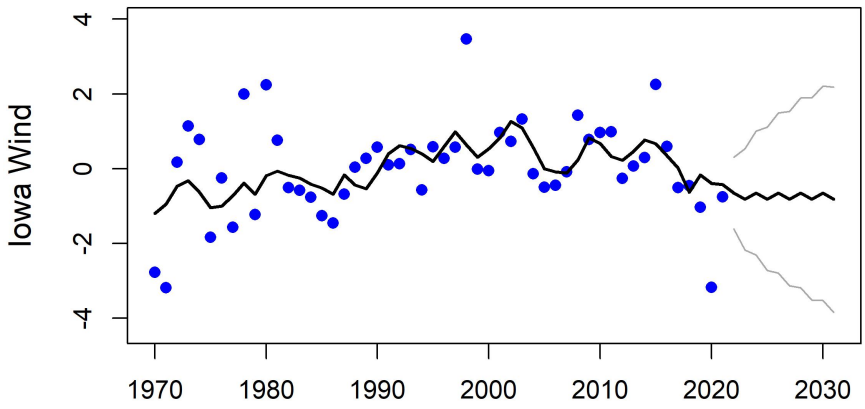

Figure 3.6: Forecast of loss from wind in Iowa.

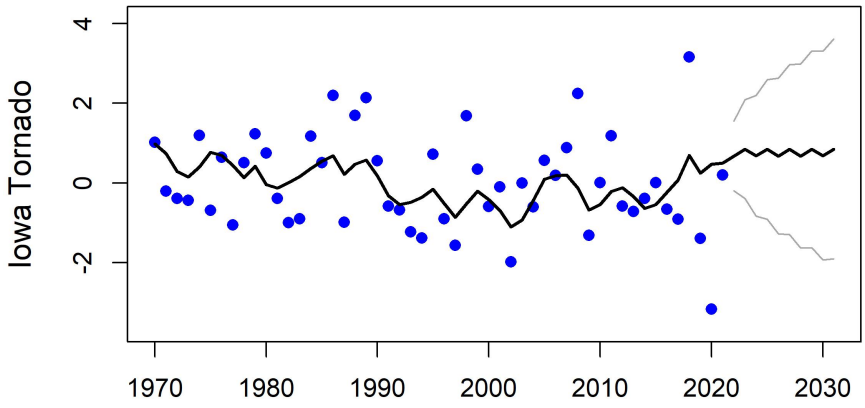

Figure 3.7: Forecast of loss from tornadoes in Iowa.

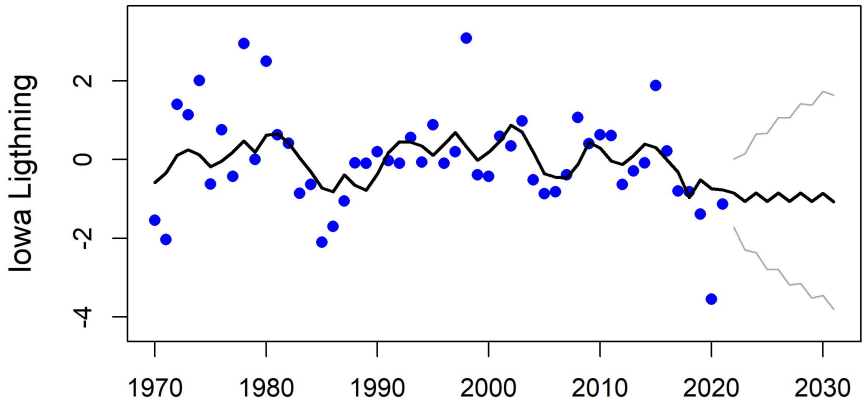

Figure 3.8: Forecast of loss from lightning in Iowa.

3.3.2 Case Study: County Level Analysis in Ohio

As the number of spatial locations increases, the feasibility of the multivariate approach reduces drastically due to the extended compute time. To illustrate this, a matrix variate model was fit for the 88 counties in Ohio. In this instance, the multivariate model would have 62,128 covariance parameters alone, with an additional $352q$ factor loading parameters for a model with q factors. The MARSS package estimated 57.4 Gb of memory needed to fit the model. However, the matrix variate model with 2 factors was able to converge in 903 iterations.

The raw loss data for the beginning and end of the study period are given in Figures 1 and 2. Early on in the study period, losses are reported in more regional patterns, while more recent years of data have more fine grained reports of losses. In addition, the variance of annual estimates are much larger toward the end of the study in 2020 than at the beginning of the study in 1970. Dollar losses were log transformed to reduce this heteroskedasticity.

This case study illustrates the manner in which the factor loadings can detect underlying patterns as losses evolve over time. The resulting factor loadings, following a varimax rotation, are presenting in Figure 3.9 and Table 3.8. The factor matrix A gives the loadings onto the rows, which represent the 88 spatial locations shown in Figure 3.9. Each factor identifies clustering regions with similar underlying structure. The first factor focuses on the northeast and southwest corners of the state. The second factor is most associated with a cluster in the southeast. This diverging pattern is due to the varimax rotation, which aims to associate as many of the loadings with one factor or the other as possible.

To emphasize that the factor loadings onto the first factor are all positive, and the loadings onto the second factor are all negative, the two maps are depicted in different color schemes. However, is important to note that the sign of these factor loadings do not have meaning on their own due to the identifiability constraint, the centering of the data, and the sign of the evolving factors. They do however imply an inverse relationship with the other factor effects in the matrix.

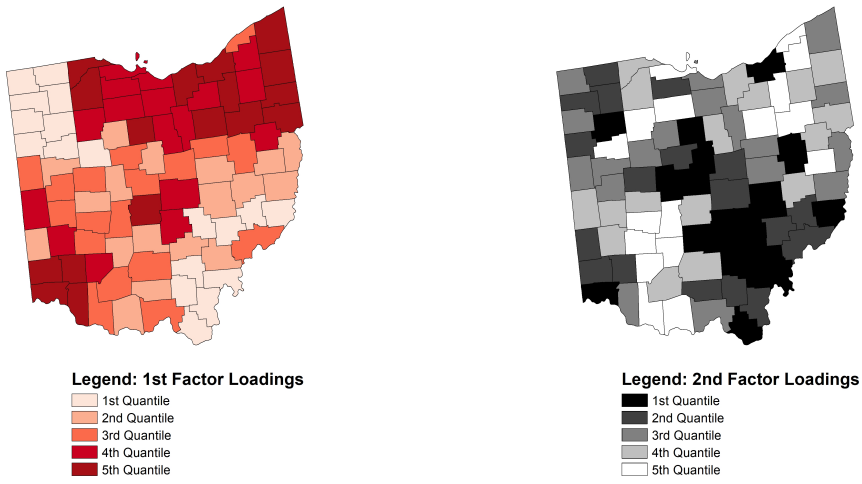

Figure 3.9: Factor loadings (A matrix) fit from the matrix variate dynamic factor model. Factors have been rotated using varimax.

Table 3.8: Factor loading matrix B fit from the matrix variate dynamic factor model. Factors have been rotated using varimax.

Factor	Hail	Lightning	Tornado	Wind
1	-2.0600	-2.044	-2.190	-2.407
2	-2.0165	-2.033	-1.839	-2.010

Table 3.8 gives the loadings of these same factors on the different hazard types. The first factor is evenly spread among the 4 hazard types. Thus the first factor seems to suggest that the losses within the first spatial cluster are fairly evenly spread among the hazard types. The second factor in Table 3.8 is also fairly evenly spread, except with a lower impact on losses from Tornadoes. These results can be used by insurers to understand the spatio-temporal patterns of losses in their portfolio, and form insurance products that adequately account for the correlation in losses from multiple perils.

In future work, it would be of interest to refit this model for varying numbers of underlying factors, which would lead to detection of more spatial clusters. The ideal number of clusters could then be chosen using methods such as AIC or BIC. However, for each additional factor included in the model with a large number of spatial locations, the computation time increases significantly. For this reason, an area of future work would be to adapt the EM algorithm to a quasi-maximum likelihood approach (Doz et al., 2012) to reduce fit time.

3.4 Discussion

From a technological standpoint, this modeling approach expands upon recent developments in the statistical literature (Ensor, 2013; Gallaugher and McNicholas, 2017; Xie et al., 2008) to develop a bilinear factor analysis model for matrix variate data. The developed methodology allows for sophisticated modeling and forecasting of convective storm losses

by simultaneously accounting for spatial correlation, correlation across different perils, and the evolution of unobserved latent variables over time. The fit values from this model may be used to identify overlapping spatial regions with similar underlying drivers of loss behavior.

The likelihood approach to factor analysis, while computationally intensive, has some benefits over component based methods, such as interpretability of the model and the ability to impose various constraints and derive estimates for missing data. This matrix variate dynamic factor model is a flexible framework and could be adapted in many ways for future use. For example, a loading matrix could be imposed on the evolving factors in order to account for known changes in the underlying factors, or expected structural breaks. The model could also be adapted to allow for maximum likelihood fitting of missing observation data. One final area of future work would be to adapt the EM algorithm to a quasi-maximum likelihood approach (Doz et al., 2012) in order to reduce fit time. Such an approach would be able to retain the benefits and flexibility of this likelihood approach, while speeding up fit time.

This work has value for policy makers, risk managers, and insurers in their development of new approaches for convective storm resilience. For insurers, the detected spatial patterns can bring insight to the spatial distribution of risk present across their portfolios. This may lead to new products such as parametric insurance and increased risk-sharing through deductibles. For policy makers, these findings will aid in development of effective response strategies to allocate disaster resilience and recovery resources. Identification of regions with similar drivers of risk may also be used to aid supplementary models by providing a baseline assumption of similar loss behavior in those areas.

Chapter 4

NATURAL HAZARDS AND MIGRATION IN THE UNITED STATES

The complex yet well established relationship between climate and migration was recently highlighted in the 2022 Assessment Report from the Intergovernmental Panel on Climate Change (Dodman et al., 2022). In particular, there is an overwhelming consensus that both long run climatic conditions (Backhaus et al., 2015; Cattaneo and Peri, 2016) as well as acute hazard events (Gröschl and Steinwachs, 2017) are important factors for migration worldwide (Dodman et al., 2022; Hoffmann et al., 2020; Šedová et al., 2021). Empirical research into the evolution of these effects in the United States is sparse (Hoffmann et al., 2020; Hoffmann et al., 2021; Šedová et al., 2021). In addition, there are few studies into the impacts of accumulated shocks over time (Hoffmann et al., 2021).

Events that consist of multiple hazards that occur simultaneously or in short succession are referred to as compound hazards (Hillier and Dixon, 2020; Zscheischler et al., 2018). Often times, the underlying hazards may have a relatively low impact if occurring in isolation, however the interaction of these events can create extreme outcomes (AghaKouchak et al., 2020; Vahedifard et al., 2016). One example of this phenomena was the 2010 heatwave in Russia that was intensified by co-occurring drought, heat, fire and air pollution (Zscheischler et al., 2018), resulting in nearly 5,000 deaths (Met Office, 2023). Another example was the deadly 2018 debris flow event in California which killed 23 people and damaged over 400 homes (Kean et al., 2019). The event was heavily impacted by preceding drought, wildfire, and heavy precipitation (AghaKouchak et al., 2020).

There is reason to suspect that repeated hazards over time could have a different impact on migration patterns than isolated acute events. For example, some studies suggest that

psychological drivers, such as fear and perceived risk, can greatly influence individual's migration choices (Shukla et al., 2018).

This study investigates the effect of compound hazard events on regional migration patterns in the U.S., adopting the established and well-suited panel approach (Hoffmann et al., 2020; Hoffmann et al., 2021). We hypothesize persistent, cumulative, and compounding hazards may instigate migration into different areas of the country with varying climate and hazard exposure. In addition, our choice of study design is able to effectively address some of the data and modeling complications that have affected climate and migration papers in the past, which have lead to highly variable results (Beyer et al., 2022; Hoffmann et al., 2021).

Specifically, this study investigates two different categories of hazard events: the occurrence of significant large events, and compounding hazard events over time. Databases of disaster losses often have definitional concerns, whereby the same climate-related or geophysical event is entered as multiple separate loss events (Gall et al., 2009). By carefully choosing spatio-temporal co-occurrence of events in time, we develop a data-driven definition of what constitutes a compound hazard event.

From a methodological perspective, in this work, we adopt a Bayesian spatio-temporal panel approach, with the varying measures of hazard events as regressors. Bayesian models have recently been suggested as a potential solution to the shortcomings of previous migration modeling attempts (Beyer et al., 2022). Previous studies have emphasized detection of temporal trends, but the chosen models have failed to adequately capture the temporal variation in the data (Beyer et al., 2022), leading to widely varying results (Hoffmann et al., 2021). In the Bayesian modeling approach, rather than attempting to discern the true value of the parameter estimates with an accompanying confidence interval, a probability interval for the unknown quantity is calculated. This interval thus has a clear interpretation

as having a high probability of containing the parameter value of interest (Gelman et al., 2013).

As there are still many aspects of climate-related migration that are not well understood, many authors have called for more nuanced research into the effects in different spatial regions over time (Hoffmann et al., 2020; Kaczan and Orgill-Meyer, 2020; Klepp, 2017). Thus, this study extends current findings of climate-related migration to address the varying effects of acute and compounding events on regional migration, while also exemplifying use of Bayesian models for future migration related studies.

4.1 Literature Review

Broad view studies of migration patterns typically consider migration rates or counts between geographical regions, such as countries or states. In these instances, the data forms a spatio-temporal panel, with migration counts or rates observed annually. Thus, panel data approaches are typically employed (Hoffmann et al., 2020; Hoffmann et al., 2021; Šedová et al., 2021), with the aim to estimate the impact of changes in climatic variables over time (Beyer et al., 2022). These types of longitudinal studies can be broken down into those that consider slow-onset events, such as drought (Cattaneo and Peri, 2016), and acute events such as severe storms or wildfires (Gröschl and Steinwachs, 2017). Overall, previous meta-analyses suggest that climatic events are more likely to lead to internal rather than international migration (Hoffmann et al., 2020; Šedová et al., 2021).

The majority of literature has focused on slow onset events, such as sea level rise and desertification (Hoffmann et al., 2021). Such studies tend to make use of long term climate variables, such as precipitation levels and temperature (Hoffmann et al., 2020). In the studies looking at quick onset events, the choice of metric to describe events vary widely (Hoffmann et al., 2021). Some studies include metrics to estimate severity of events in a region. For example, (Gröschl and Steinwachs, 2017) considered maximum observed

values, such as maximum earthquake intensity and maximum wind speed. More commonly considered is a simple count of the number of events during a given time period (Beine and Parsons, 2014; Cattaneo and Peri, 2016; Neumann et al., 2015; Saldaña-Zorrilla and Sandberg, 2009). One study in Mexico considered the number of regional disaster events between 1990 and 2000 as a predictor in a spatial regression model (Saldaña-Zorrilla and Sandberg, 2009). Their model suggested that regions with a higher frequency of disasters have higher rates of out-migration. Another study of international migration at the decade level considered the count of total disasters in a given decade (Beine and Parsons, 2014), and found no evidence of impact between 1960 and 2000. While these types of studies do consider metrics of repeated events, it is difficult to discern how truly compounding these events are at such a large spatio-temporal scale.

Studies of climate-related migration specifically in the U.S. have not addressed the year to year impacts of natural hazards in the United States. Studies that have focused on climate-related variables, such as temperature and precipitation, have found significant effects of climate on U.S. migration patterns (Poston et al., 2009). Studies regarding acute events have been limited in scope. In one study (Boustan et al., 2020), the authors analyzed federally designated disasters on the decadal level in the United States from 1920 to 2010 to see impacts to the economy. This study was thus conducted at a fine spatial resolution, but a broad temporal resolution. Over the span of the study, they found that severe disasters increased out-migration rates at the county level by 1.5 percentage points. The migration response to milder disasters is smaller but has been increasing over time.

Another approach at the county level is to fit separate bilateral models for migration each year, with a focus on understanding income inequality in tandem with climate-related migration (Chen and Lee, 2022). This study focused on detecting spatial patterns, rather than temporal patterns. Other studies have focused solely on migration patterns following specific large events, such Hurricanes Katrina and Rita (Myers et al., 2008).

Unfortunately, the use of raw disaster counts in migration studies without careful data cleaning is not ideal. This is because many databases of disaster events have definitions that make the count of disasters difficult to interpret (Gall et al., 2009). For example, the U.S. Storm Data product, maintained by the National Oceanic and Atmospheric Administration, distinguishes between episodes, which includes an entire storm system, and events, which may occur as part of the same storm system ("Storm Data FAQ Page", 2023). In addition, in this database, a single tornado is counted as multiple separate events if it lifts off the ground for more than 4 minutes or 2 miles ("Storm Data FAQ Page", 2023). Any attempt to catalogue disaster occurrences must address such definitions in one way or another. Thus when aggregating the count of events over a time period that come from such databases, it may not be clear what the count truly represents.

It is not surprising then that migration studies that include metrics of event intensity, instead of simple event occurrence, are more likely to detect a signal (Hoffmann et al., 2021; Šedová et al., 2021). For example, these studies might include dollar losses associated with an event, the length of the event, or the number of people effected as a measure of intensity. When it comes to compound events, the associated hazards can be interrelated. This implies that assessing the risk of each event individually may not provide the same level of information regarding the compounded catastrophic impacts (Zscheischler et al., 2018).

In many studies, long run climatalogical variables like temperature and precipitation are considered alongside measurements of hazard occurrence (Beine and Parsons, 2014; Cattaneo and Peri, 2016). In addition, these studies often include other control variables, such as GDP. Overall, the choice of which variables to include in migration models is highly dependent on the spatio-temporal area of interest, and the intention of the study (Hoffmann et al., 2021). Because climate metrics such as precipitation, temperature, and disaster occurrence are not independent, there is the possibility for omitted variable bias. In addition, previous studies have also suggested that climate and hazards can impact migration not only

directly, but through socioeconomic or agricultural channels (Beine and Parsons, 2014; Cai et al., 2016). Thus, it is common to include dummy variable fixed-effect terms to control for time-invariant factors that affect migration flows, which may be difficult to explicitly measure (Cai et al., 2016; Hoffmann et al., 2021).

One specific type of panel data model employed is the so-called gravity model, which assumes that bilateral migration is proportional the population sizes of the starting and ending location, and the distance between them (Backhaus et al., 2015; Beine and Parsons, 2014; Beyer et al., 2022). They often take on a log-log form, and are fit using OLS (Beyer et al., 2022; Hoffmann et al., 2021), although some studies have used Poisson regression or negative binomial regression to account for zero-inflated data (Beyer et al., 2022).

Gravity models attempt to deal with the spatial nature of the data by considering physical distance between locations, or if two locations share a common border (Beine and Parsons, 2014). However, these studies rarely take an overtly spatial approach, or test for spatial auto-correlation in the model residuals (Hoffmann et al., 2021; Saldaña-Zorrilla and Sandberg, 2009). Due to the large panel sizes, these types of models typically are able to capture the vast majority of spatial variation (Beyer et al., 2022). However, these models do not typically allow for temporal auto-correlation, and have recently been criticized for their inability to explain temporal variation, leading to highly variable results (Beyer et al., 2022; Hoffmann et al., 2020). In addition, climactic shocks may have longer term impacts on migration, necessitating temporal lags (Hoffmann et al., 2021).

Bayesian models have been proposed as an alternative approach to effectively model temporal variation in a stochastic manner (Abel et al., 2013; Azose and Raftery, 2015; Beyer et al., 2022; Bijak, 2011). Despite the complexity and uncertainty of climate, hazard, and migration patterns, Bayesian approaches can be used to identify climate-related trends and make reasonable projections while acknowledging the inherent uncertainty (Abel et al., 2013).

4.2 Methodology

4.2.1 Data

This study makes use of the Census Bureau's state to state migration flow datasets (Be, 2021) from state to state. These are annual bilateral flows both to and from a given state and are estimated from the American Community Survey. Estimates of GDP come from the Bureau of Economic Analysis (Bureau of Economic Analysis, 2022). In addition, we consider the climate regions defined by NOAA (Karl and Koss, 1984) to investigate regional trends.

We adopt the definition of compound events as events that are co-occurring in space and time. We specifically consider events in the same year and state that align with the compound hazard categories defined by (Zscheischler et al., 2018), given in Table 4.1. From this framework, two types of measures are constructed that attempt to capture the impact of compounding hazard events over time from the SHELDUS Version 21.0 database: dollar losses and counts. First, we consider a time series of total dollar losses arising from each compound hazard category in each state and region. Next we count the number of county level events in SHELDUS causing over a specific threshold of losses. Here, we consider a cutoff of $10,000. Filtering through a minimum cutoff of disaster losses ensures that the count of loss events is not inflated by small entries, but rather focuses on more substantial disasters.

Note that the category "Flooding" is included, despite only consisting of a single hazard. This is because flooding is inherently a compound phenomena, impacted by features such as river flow and coastal water level Zscheischler et al., 2018. Flood events are may subsequently result from heavy precipitation or storm surge. For this reason, we consider both flooding as an independent category, as well as flooding that occurs together with high wind.

4.2.2 Bayesian Panel Model

In this study, a series of Bayesian spatio-temporal panel models are considered with different metrics for the number and severity of compound hazard events as covariates. Macro studies of migration tend to adopt either raw migration counts or migration rates as the outcome variable of interest for modeling purposes (Beyer et al., 2022; Hoffmann et al., 2021). Because the variability in migration counts grows proportionally with population size, dividing by population in some instances can help to stabilize the variance of the estimates (Azose and Raftery, 2015). In this application, the count of migrants c_{ijt} from location i to location j in year t are be considered. We also investigate the migration rates, where the migration count is divided by the population of the originating country during the year of outward migration p_{it}. The resulting rate $m_{ijt} = c_{ijt}/p_{it}$ in effect gives the fraction of the population in location i that moved from location i to location j that year.

By definition, observable migration rates are bound in the interval $[0, 1]$, which has implications for the distributional assumptions of regression approaches. In addition, migration data are often zero inflated. Despite these data properties, many studies still adopt a standard OLS model for bilateral migration with normal errors, although, Poisson and Negative Binomial models are at times adopted for migration count data (Hoffmann et al.,

Table 4.1: Compound hazard categories considered.

Compound Hazard Category	SHELDUS Records	Model Alias
Wildfire	Heat, Drought, Wildfire, Lightning	fire
Convective Storm	Severe Storms, Lightning, Hail, Wind	storm
Wind with Severe Precipitation	Wind, Flooding	wind_flood
Flooding	Flooding	flood
Coastal	Coastal, Hurricane	coastal
Heat	Drought, Heat	heat_drought

2021). In this work, we address the distributional concerns by considering a panel of state and region level aggregates. Because we are interested in detecting regional trends in migration, our primary model aggregates migration flows to the state-region pair. This forms a bilateral dataset capturing information about state of origin and region of destination. After assessing the distribution of the aggregated data, we decided not to form a migration rate, but to normalize using the log transform.

We also considered for each state the out migration rate, in migration rate, and net migration rate. These choice of models are well suited to our key interest of understanding the effects of compound hazard events on in and out migration. In addition, this choice greatly stabilizes the dataset and speeds fitting via Markov Chain Monte Carlo.

At the suggestion of Beyer et al., 2022 and Hoffmann et al., 2021, we begin with a simple regression structure with few controls taking the form:

$$\begin{aligned} ln(m_{it}) =& \alpha_{ij} + \beta_0 + \beta_1 S_{it} + \beta_2 Pop_{it} + \beta_3 Pop_{jt} \\ &+ \beta_4 GDP_{jt}/Pop_{it} + \beta_5 GDP_{jt}/Pop_{jt} + \epsilon_{it} \end{aligned} \tag{4.1}$$

Here, S_{it} is a measure of the severity of compound hazard category in location i. Meanwhile, GDP_{it}/Pop_{it} gross domestic product per-capita of the originating state, and GDP_{jt}/Pop_{jt} is the gross domestic product per-capita of the destination region. GDP per capita as a metric of wealth is a well established confounding factor in studies of hazards and migration (Hoffmann et al., 2021). The α_{ij} term controls for other unobserved time-invariant state confounding affects. In Bayesian regression approach, each of the parameters are assigned a prior distribution. As a first step, we assume normal priors for all regression variables. The model error ϵ_{it} is also taken to be normally distributed, with mean 0. Each model was fit using the *brms* package in R, which fits via Markov Chain Monte Carlo sampling using the NUTS sampler in Stan.

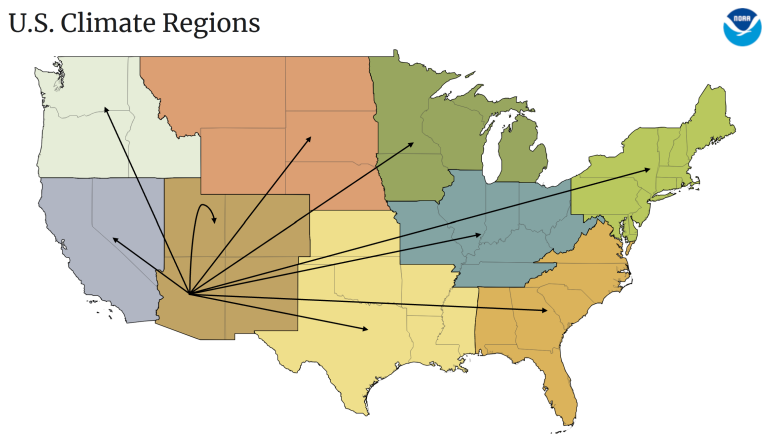

Figure 4.1: Migration flows considered between originating state and destination region. Climate regions defined by NOAA (NCEI, 2023).

4.3 Results

4.3.1 Bilateral Model

First we consider the bilateral model, which facilitates the identification of spatial patterns and effects for both origins and destinations of migration. In this instance, we consider migration from each state to each climate region of the U.S., as defined by NOAA (Karl and Koss, 1984) and as seen in figure 4.1. Typically in the non-Bayesian approach, fixed effects are included for spatial location pairs. In our Bayesian approach, we allow for varying intercepts. These varying intercepts capture the effects of time-invariant factors specific to each state-region pair, such as their geographical distance, political leaning, climate, and culture (Beyer et al., 2022).

A second approach sometimes taken in the literature is to allow for two separate fixed effect terms, one to account for effect of the origin location, and one for destination location. These fixed effects are more general and capture the standard migration patterns to and from each location. However, they are not able to account for ongoing migration relationships between the regions.

First, a null model was fit only employing the varying intercept terms. The Bayesian model was fit using 4 chains, and all parameters converged with an $\hat{R}$ of 1. The findings in Table 4.2 reveal that the model utilizing paired effects comprehensively captures the majority of the variability. We report the Bayesian R^2 metric, which attempts to measure the portion of the variability accounted for by the model. This finding is in line with previous studies containing pairwise fixed effects (Beyer et al., 2022). Meanwhile, the model employing separate fixed effects fails to adequately explain much of the observed variability in the data. This suggests the presence of latent, consistent factors influencing migration across each state-region pair over time. Models without these state-region pair effects might suffer from omitted variable bias due to the many complicating factors impacting migration, unless these factors are accounted for explicitly in the model.

Next we considered a baseline model incorporating the two most commonly considered covariates in migration studies: population and GDP per capita (Beyer et al., 2022). Including just these two additional covariates raised the R^2 significantly to 0.94 for the pairwise model. However, the R^2 estimate hardly changed for the separate state and region varying effect model. These results are confirmed by their density vs posterior fit plots, shown in Figure 4.2. This suggests that the vast majority of the variation in the data is due to state-region pair effects, population and wealth.

Next, we fit a series of models for each considered hazard metric with varying intercepts for each state pair. For each model fit, none of the metrics for hazards were significant, both in the compound models and general hazard metric models. In part this is not surprising, as the varying intercepts explain the majority of the observed variance. These results suggest minimal to no impact of hazard losses over the pairwise effects. However, if the level of hazards are fairly stable over the study period, it is possible that any effects of hazards on migration are being accounted for in the pairwise effects. Thus we next considered models with varying effects for state and region only, in a cross study design.

Table 4.2: Fit of baseline bilateral models with varying intercepts.

Varying Intercepts	Model Terms	R^2
Pairwise	$\alpha_{ij} + \beta_0$	0.78
Pairwise	$\alpha_{ij} + \beta_0 + \beta_1 Pop_{it} + \beta_2 Pop_{jt} + \beta_3 GDP_{it}/Pop_{it} + \beta_4 GDP_{jt}/Pop_{jt}$	0.94
State and Region	$\alpha_i + \gamma_j + \beta_0$	0.39
State and Region	$\alpha_i + \gamma_j + \beta_0 + \beta_1 Pop_{it} + \beta_2 Pop_{jt} + \beta_3 GDP_{it}/Pop_{it} + \beta_4 GDP_{jt}/Pop_{jt}$	0.39

Results from the cross study design again suggested no significant effect of hazards above the pairwise effects. Moreover, these models were not able to account for the majority of the variation in the data, with R^2 remaining below 0.4 in all instances. Thus for a final comparison, the models were refit using more general metrics of hazards. These metrics included a time series of total dollar losses in each state and region, as well as counts of reported hazard events with losses above various thresholds: \$100K, \$1M,\$10M, \$100M, and \$1B. As one final alternative, the count of Presidential Disaster Declarations in each state was considered, as these events are likely to cause significant hardship to the effected communities. None of these approaches showed significant effects.

These comprehensive results suggest that at this time, there is little detectable affect of hazard events on regional migration. Previous studies have demonstrated the impact of hazards on migration at a local level (Boustan et al., 2020; Chen and Lee, 2022; Myers et al., 2008), however we notably find these patterns do not persist to regional migration levels. These results suggest that individuals and households are not choosing migration as their primary adaptation strategy to deal with increased hazards in their climate region and the imposing psychological concerns regarding climate change. Instead, families in exposed areas may simply choose to invest in hardening their home in preparation for inclement hazard events (Donner and Rodriguez, 2008).

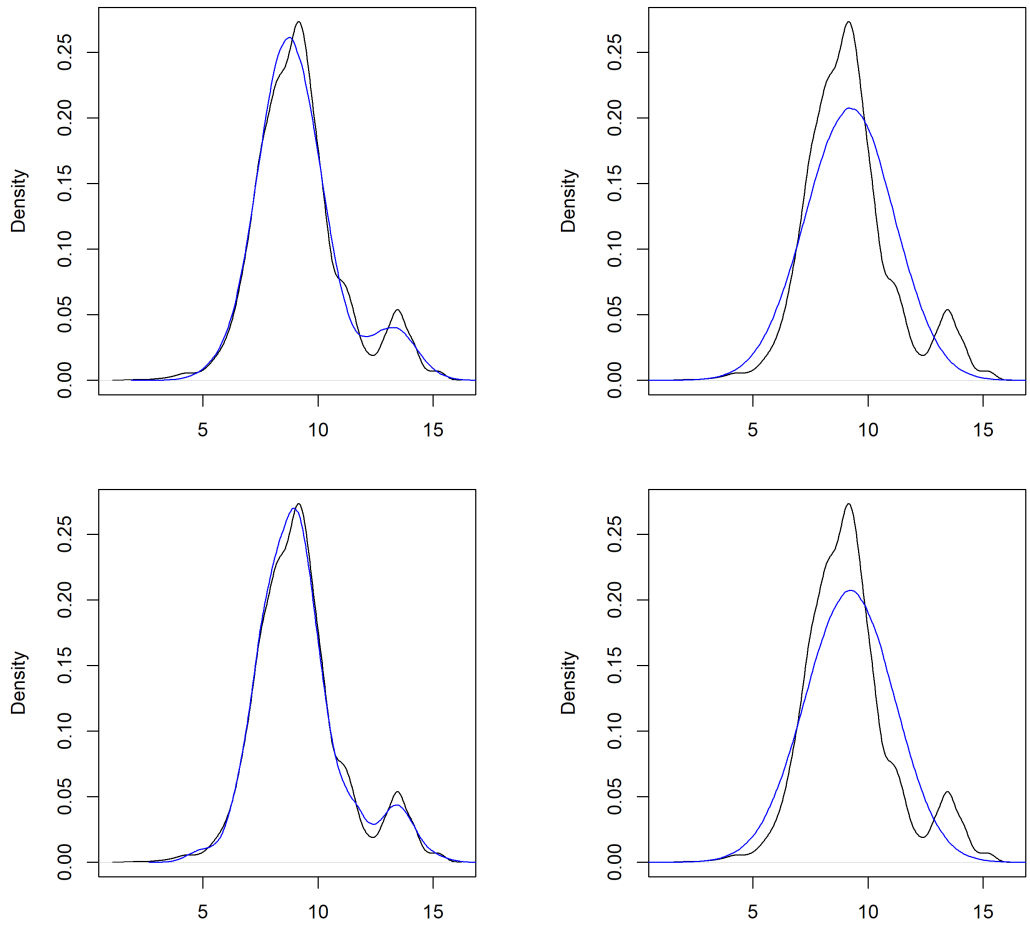

Figure 4.2: Data density (black) and model posterior distribution (blue).
Top left: State-region pair varying intercept.
Bottom left: State-region pair varying intercept with population and GDP per capita.
Top right: Separate origin state and destination region varying intercepts.
Bottom right: Separate state and region varying intercepts with covariates.

4.3.2 Bayesian Perspective of Temporal Trends in U.S. Migration

One advantage of the Bayesian approach is the formation of a probability distribution for the fit parameters. While our migration models did not find significance of the climate-related variables, we illustrate here how the Bayesian modeling approach can be used to understand temporal trends in migration.

We refit the basic migration model that incorporated pairwise fixed effects, population, and GDP, but included a trend term. The results showed a small but statistically significant

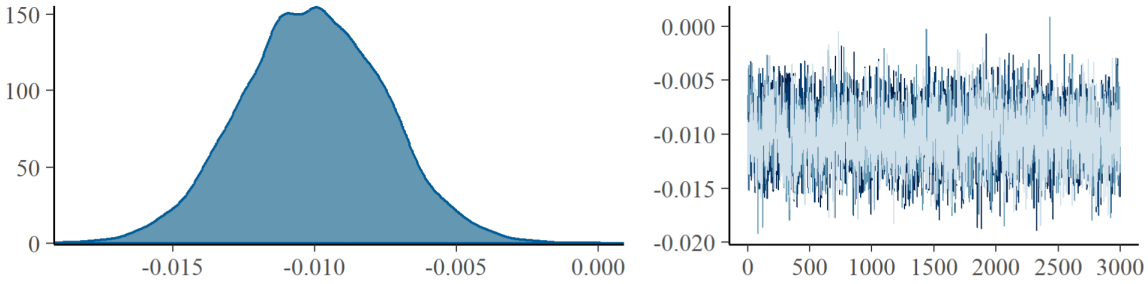

Figure 4.3: Left: Posterior distribution of the trend term.
Right: Trace of the MCMC chains, indicating convergence.

decrease in out of state migration overall after accounting for pairwise effects, population, and GDP. In contrast, the state and region varying intercepts model did not detect significant trend.

The posterior distribution for the identified trend term in the pairwise varying intercept model is given in figure 4.3. To anticipate forthcoming trends, future projections of population and GDP can be incorporated along with the trend to generate the marginal posterior in distribution for migration between each state region pair.

4.3.3 State Level Models

With the limited findings in the bilateral model, we subsequently explored distinct models for aggregate inflow, aggregate outflow, and net flow for each state. In each model, varying intercepts were included for the state in order to account for the baseline migration rate within the respective state over the 15 year study period. These models give surprising results. The resulting coefficients for compound hazards in Table 4.3 measure the nationwide effect of each compound hazard category above baselines.

After inspecting the outflow and inflow data, migration rate per 10,000 was chosen as the outcome variable for each. However, the net flow data had a different structure, being inherently centered around zero. Normalization by population would result in a more

complex bimodal distribution. Thus instead, the raw net flow data was assumed to come from a shifted log normal distribution.

For all models among the count related measures of hazards, the convective storm category was the only compound hazard type that was significant. For every additional convective storm related event the migration rate increased by about 0.02%.

The most substantial effect seen from dollar losses is from coastal events. Every additional million dollar in losses from coastal related events was associated with a decrease in the outward migration rate by approximately 1.2%. This finding, while counter-intuitive, is actually supported by previous studies (Boustan et al., 2020; Donner and Rodriguez, 2008). The population in coastal regions have increased dramatically over the last few decades (Nicholls, 1995), and while these populations are at an increased risk to coastal related hazards and flooding (Klein et al., 2003), previous experts indicate that homeowners in these areas are reluctant to relocate. This is largely due to strong individual, market, and regulatory incentives that encourage oceanfront property owners to stay in place (Donner and Rodriguez, 2008).

Furthermore, movement of displaced populations following the largest coastal hazard events, such as hurricanes Katrina and Rita, are not spatially random (Myers et al., 2008). Migration in these instances tends to be to neighboring counties, reducing the effect of state to state migration (Myers et al., 2008). Interestingly, while coastal events were found to be associated with decreased out-migration, they also were found to be associated with decreased in-migration. The suggests that overall mobility following severe coastal events is reduced.

Meanwhile, each additional million dollar in fire related hazards was found to correspond to a modest 0.02% increase in outward migration rate, and an increase of one million dollar in losses from combined wind and floods was associated with 0.07% increase in out migration rate.

Table 4.3: Marginal effects of compound hazards.

Covariate	Outflow	Inflow	Net Outflow
count_fire	0.0003	0.0003	0.0427
count_storm	0.0003*	0.0002	-0.0131
count_windflood	-0.0003	-0.0003	0.0264
count_flood	0.0003	0.0002	-0.0316
count_coastal	-0.0004	-0.0003	-0.0692
count_heatdrought	0.0004	0.0004	-0.0397
loss_fire	0.0021*	0.0010	1.2700
loss_storm	0.0012	0.0013	-1.9700
loss_windflood	0.0007*	0.0005	-0.2430
loss_flood	-0.0002	-0.0002	-0.0815
loss_coastal	-0.0118*	-0.0121*	0.2400
loss_heatdrought	0.0811	0.1380	-0.9950

Losses are reported at the million dollar scale.

* Indicates that the 95% probability interval for the parameter did not contain zero.

4.4 Discussion

Overall, these results suggest that at this time, there is little detectable affect of hazard events on regional migration. This is an unexpected finding in the context of climate change. While previous studies have demonstrated the impact of hazards on migration at a local level (Boustan et al., 2020; Chen and Lee, 2022; Myers et al., 2008), we find these patterns do not persist to interstate and inter regional migration levels.

Various additional factors could contribute to these observed outcomes. As observed in previous studies, the limited time series duration of the available migration data may hinder the identification of long term trends. In addition, the choice of dataset could have significant impacts on results. For example, the Census Bureau solely provides county-to-

county migration data at five-year intervals, thus making the choice of dataset a trade off between temporal and spatial scale. For this reason, it may prove valuable in future work to make use of data over a longer time period at varying levels of spatial granularity. The inclusion of supplementary datasets, such as IRS migration data (IRS, 2023), could aid in this type of future analysis.

The evidence provided from this study is that individuals and households are not choosing migration as their primary adaptation strategy to deal with increased hazards and the imposing psychological concerns regarding climate change. For example, families in exposed areas may simply choose to invest in hardening their home in preparation for inclement hazard events (Donner and Rodriguez, 2008). However, many experts question the sustainability of remaining in areas particularly prone to future hazards, such as coastal regions (Donner and Rodriguez, 2008). Previous studies suggests that involuntary migration is both economically and psychologically costlier than voluntary migration (Partridge et al., 2017). Future research in this regard might explore questions such as who among a population are more likely to move or stay, and how moving might affect migrants' well-being (Belasen and Polachek, n.d.).

Studies into climate-related migration, particularly those aimed at detecting and forecasting temporal trends, suffer from widely varying results and a great deal of scientific uncertainty (Beyer et al., 2022). Enhancing comprehension of these trends to the fullest extent possible would prove immensely valuable from a policy perspective. However, any attempt to capture such trends must account for the inherent data limitations and uncertainties surrounding climate change, natural hazard occurrences, and other contributing factors to migration (Azose and Raftery, 2015).

This study illustrates the manner in which a Bayesian modeling approach can estimate the inherent uncertainty in migration modeling, providing decision-makers with greater clarity regarding the extent of knowledge on migration trends. The ability to account for

uncertainty is a powerful strength of Bayesian models, making them an invaluable tool for improving decision-making in the face of complex, uncertain systems like climate change (Gelman et al., 2013; Gelman and Hill, 2006).

Chapter 5

CONCLUDING REMARKS

As the impacts of climate change worsen in the coming decades, natural hazards are expected to increase in frequency and intensity, leading to increased loss (Gall et al., 2011) and risk to human livelihood (Devkota et al., 2016). While exact estimates are unclear, climate events are projected to lead to the displacement of a significant number of people in the coming decades (Hoffmann et al., 2021; Islam and Khan, 2018). This dissertation aimed to develop effective statistical methodologies for forecasting losses from natural hazards and understanding their societal impacts in light of the evolving spatio-temporal patterns of losses over the last few decades. The developed spatio-temporal statistical approaches highlight the ways in which hazard databases can be leveraged in these endeavours.

The analysis carried out in Chapter 2 provides relevant background information for modeling hazard losses by suggesting that trends in hazard losses can predominantly be ascribed to growing population and wealth. These findings corroborate conclusions drawn from prior research (R. Pielke, 2020). Importantly, these results indicate that to avoid spurious regression, application studies must account for these socioeconomic factors. However, even after accounting for trends in population and wealth, our results show that hazard loss datasets are not necessarily stationary. While autocorrelation and heteroskedasticity were not a significant issue for every county in the study, these data anomalies did occur for many subsets of the data. Thus studies making use of hazard loss data must apply appropriate spatio-temporal techniques to account for these possible data features.

In light of these features of the data, Chapter 3 introduces a novel bilinear factor analysis model for matrix variate data that allows for sophisticated modeling and forecasting of

convective storm losses by simultaneously accounting for spatial correlation, correlation across different perils, and the evolution of unobserved latent variables over time. Two case studies were considered to exemplify use of the model. The first study was for the 4 states in the Upper Midwest climate region, and the second for the 88 counties in the state of Ohio. In the first case study, the model highlighted two major underlying factors driving losses in the Upper Midwest climate region: one most associated with hail and wind in Iowa, and one most associated with lightning in Iowa, Michigan, and Wisconsin. These estimated factors were then used to forecast by creating confidence intervals of future losses. In the second case study modeling losses from convective storms in the counties of Ohio, the model identified two major clusters with different underlying loss structures: one in the northeast part of Ohio, and one in the southwest portion. These findings highlight the value of the developed model to identify overlapping spatial regions with similar underlying drivers of loss behavior, and leverage these for forecasting.

Finally, Chapter 4 turns to the study of societal impacts of natural hazards by investigating their relationship with migration in the United States. In particular, this research illustrates how Bayesian approaches can be used to estimate the inherent uncertainty in such application studies. Overall, these results suggest that at this time, there is little detectable affect of hazard events on regional migration. While previous studies have demonstrated the impact of hazards on migration at a local level (Boustan et al., 2020; Chen and Lee, 2022; Myers et al., 2008), we find these patterns do not persist to interstate and inter regional migration levels.

From a methodological standpoint, the models developed in Chapters 3 and 4 make significant contributions to the relevant established literature. The model developed in Chapter 3 expands upon recent statistical developments (Ensor, 2013; Gallaugher and McNicholas, 2017; Xie et al., 2008) to derive a bilinear factor analysis model for matrix variate data. This developed methodology allows for sophisticated modeling and forecasting of convec-

tive storm losses by simultaneously accounting for spatial correlation, correlation across different perils, and the evolution of unobserved latent variables over time. Further, Chapter 4 extends the methodological approaches taken in the literature on migration and hazards by adopting a Bayesian modeling framework, which has recently been suggested as a potential solution to the shortcomings of previous migration modeling attempts (Beyer et al., 2022). The resulting model can be used to estimate the inherent uncertainty present in natural hazard related migration studies.

Together, these three avenues of research combine rigorous statistical analysis with advanced forecasting and modeling methodologies to provide valuable insights into the evolving landscape of natural hazard losses and their societal impacts. For example, each of the proposed spatio-temporal strategies can be used to identify areas that are particularly vulnerable to specific types of hazards. On its own, this information can be used to inform mitigation strategies through insurance products, infrastructure development, and allocation of disaster response resources. Appropriate forecasting models can particularly influence development of insurance products to account for both spatial heterogeneity and loss trends. In the context of hazard related migration, these findings can assist in emergency response and disaster management planning for displaced populations, encompassing both preparation for changes in population as well as provision of aid and resources to affected communities (Cai et al., 2016; Cattaneo and Peri, 2016; Kaczan and Orgill-Meyer, 2020).

Bibliography

Abatzoglou, J. T., & Williams, A. P. (2016). Impact of anthropogenic climate change on wildfire across western US forests. *Proceedings of the National Academy of Sciences*, *113*(42), 11770–11775. https://doi.org/10.1073/pnas.1607171113

Abel, G., Bijak, J., Findlay, A., McCollum, D., & Wiśniowski, A. (2013). Forecasting environmental migration to the united kingdom: An exploration using bayesian models. *Population and Environment*, *35*(2), 183–203. https://doi.org/10.1007/s11111-013-0186-8

AghaKouchak, A., Chiang, F., Huning, L. S., Love, C. A., Mallakpour, I., Mazdiyasni, O., Moftakhari, H., Papalexiou, S. M., Ragno, E., & Sadegh, M. (2020). Climate extremes and compound hazards in a warming world. *Annual Review of Earth and Planetary Sciences*, *48*(1), 519–548. https://doi.org/10.1146/annurev-earth-071719-055228

Azose, J. J., & Raftery, A. E. (2015). Bayesian probabilistic projection of international migration. *Demography*, *52*(5), 1627–1650. https://doi.org/10.1007/s13524-015-0415-0

Backhaus, A., Martinez-Zarzoso, I., & Muris, C. (2015). Do climate variations explain bilateral migration? a gravity model analysis. *IZA Journal of Migration*, *4*(1). https://doi.org/10.1186/s40176-014-0026-3

Bai, J., & Wang, P. (2014). Identification and bayesian estimation of dynamic factor models. *Journal of Business &amp Economic Statistics*, *33*(2), 221–240. https://doi.org/10.1080/07350015.2014.941467

Barbier, E. B. (2012). Progress and challenges in valuing coastal and marine ecosystem services. *Review of Environmental Economics and Policy*, *6*(1), 1–19.

Barredo, J. I. (2010). No upward trend in normalised windstorm losses in europe: 1970–2008. *Natural Hazards and Earth System Sciences*, *10*(1), 97–104. https://doi.org/10.5194/nhess-10-97-2010

Be, U. C. (2021). State-to-state migration flows. https://www.census.gov/data/tables/time-series/demo/geographic-mobility/state-to-state-migration.html

Beine, M., & Parsons, C. (2014). Climatic factors as determinants of international migration. *The Scandinavian Journal of Economics*, *117*(2), 723–767. https://doi.org/10.1111/sjoe.12098

Belasen, A. R., & Polachek, S. W. (n.d.). Natural disasters and migration. https://doi.org/10.4337/9781782546078.00026

Beyer, R. M., Schewe, J., & Lotze-Campen, H. (2022). Gravity models do not explain, and cannot predict, international migration dynamics. *Humanities and Social Sciences Communications*, *9*(1). https://doi.org/10.1057/s41599-022-01067-x

Bijak, J. (2011). *Forecasting international migration in europe: A bayesian view*. Springer Netherlands. https://doi.org/10.1007/978-90-481-8897-0

Botzen, W. J. W., Deschenes, O., & Sanders, M. (2019). The economic impacts of natural disasters: A review of models and empirical studies. *Review of Environmental Economics and Policy*, *13*(2), 167–188. https://doi.org/10.1093/reep/rez004

Boustan, L. P., Kahn, M. E., Rhode, P. W., & Yanguas, M. L. (2020). The effect of natural disasters on economic activity in US counties: A century of data. *Journal of Urban Economics*, *118*, 103257. https://doi.org/10.1016/j.jue.2020.103257

Boyle, E., Chiaradonna, S., & Jevtic, P. (2022). A stochastic assessment of service loss due to cyber vulnerabilities of power network infrastructure: A case study of puerto rico. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4234702

Boyle, E., Inanlouganji, A., Carvalhaes, T., Jevtić, P., Pedrielli, G., & Reddy, T. A. (2022). Social vulnerability and power loss mitigation: A case study of puerto rico. *International Journal of Disaster Risk Reduction*, *82*, 103357. https://doi.org/10.1016/j.ijdrr.2022.103357

Brooks, H. (2013). Severe thunderstorms and climate change. *Atmospheric Research*, *123*, 129–138. https://doi.org/10.1016/j.atmosres.2012.04.002

Brown, T. M., Pogorzelski, W. H., & Giammanco, I. M. (2015). Evaluating hail damage using property insurance claims data. *Weather, Climate, and Society*, *7*(3), 197–210. https://doi.org/10.1175/wcas-d-15-0011.1

Bureau of Economic Analysis. (2022). Gross domestic product by county, metro, and other areas.

Cai, R., Feng, S., Oppenheimer, M., & Pytlikova, M. (2016). Climate variability and international migration: The importance of the agricultural linkage. *Journal of Environmental Economics and Management*, *79*, 135–151. https://doi.org/10.1016/j.jeem.2016.06.005

Cattaneo, C., & Peri, G. (2016). The migration response to increasing temperatures. *Journal of development economics*, *122*, 127–146.

Cavanaugh, J. E., & Shumway, R. (1997). A bootstrap variant of aic for state-space model selection. https://api.semanticscholar.org/CorpusID:16239552

Changnon, S. A. (2009). Increasing major hail losses in the u.s. *Climatic Change*, *96*(1-2), 161–166. https://doi.org/10.1007/s10584-009-9597-z

Chen, T. H. Y., & Lee, B. (2022). Income-based inequality in post-disaster migration is lower in high resilience areas: Evidence from u.s. internal migration. *Environmental Research Letters*, *17*(3), 034043. https://doi.org/10.1088/1748-9326/ac5692

Collins, S. P., D. Lowe. (2001). A macro validation dataset for us hurricane models. http://www.casact.org/pubs/forum/01wforum/01wf217.pdf

Convective storms: State of the risk triple-i issues brief. (2022). *Triple-I Issues Brief*. https://www.iii.org/sites/default/files/docs/pdf/triple-i%5C_state%5C_of%5C_the%5C_risk%5C_convective%5C_storms%5C_04212022.pdf

Coppi, R., Bolasco, S., Critchley, F., & Escoufier, Y. (1989). Multiway data analysis.

Cusack, S. (2012). A 101 year record of windstorms in the netherlands. *Climatic Change*, *116*(3-4), 693–704. https://doi.org/10.1007/s10584-012-0527-0

Cutter, S. L., & Emrich, C. (2005). Are natural hazards and disaster losses in the u.s. increasing? *Eos, Transactions American Geophysical Union*, *86*(41), 381. https://doi.org/10.1029/2005eo410001

Cutter, S. L., Gall, M., & Emrich, C. T. (2008). Toward a comprehensive loss inventory of weather and climate hazards. *Climate extremes and society*, 279–295.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Devkota, R. P., Pandey, V. P., Bhattarai, U., Shrestha, H., Adhikari, S., & Dulal, K. N. (2016). Climate change and adaptation strategies in budhi gandaki river basin, nepal: A perception-based analysis. *Climatic Change*, *140*(2), 195–208. https://doi.org/10.1007/s10584-016-1836-5

Ding, S., & Dennis Cook, R. (2018). Matrix variate regressions and envelope models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *80*(2), 387–408.

Dodman, D., Hayward, B., Pelling, M., Broto, V. C., W. Chow, E., Chu, R. D., Khirfan, L., McPhearson, T., Prakash, A., Zheng, Y., & Ziervogel, G. (2022). Cities, settlements and key infrastructure. *Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, *90*(3), 907–1040. https://doi.org/doi:10.1017/9781009325844.008

Doerr, S. H., & Santın, C. (2016). Global trends in wildfire and its impacts: Perceptions versus realities in a changing world. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1696), 20150345. https://doi.org/10.1098/rstb.2015.0345

Donner, W., & Rodriguez, H. (2008). Population composition, migration and inequality: The influence of demographic changes on disaster risk and vulnerability. *Social Forces*, *87*(2), 1089–1114. https://doi.org/10.1353/sof.0.0141

Downton, M. W., Miller, J. Z. B., & Pielke, R. A. (2005). Reanalysis of u.s. national weather service flood loss database. *Natural Hazards Review*, *6*(1), 13–22. https://doi.org/10.1061/(asce)1527-6988(2005)6:1(13)

Doz, C., Giannone, D., & Reichlin, L. (2012). A quasi–maximum likelihood approach for large, approximate dynamic factor models. *Review of Economics and Statistics*, *94*(4), 1014–1024. https://doi.org/10.1162/rest_a_00225

Durbin, J., & Watson, G. S. (1950). Testing for serial correlation in least squares regression: I. *Biometrika*, *37*(3/4), 409. https://doi.org/10.2307/2332391

Ensor, K. B. (2013). Time series factor models. *Computational Statistics*, *5*, 97–104. https://doi.org/https://doi.org/10.1002/wics.1245

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2010). *Cluster analysis* (5th ed.). John Wiley & Sons.

FEMA. (2023). National Risk Index for Natural Hazards [[Accessed March 2023]].

Gall, M., Borden, K. A., & Cutter, S. L. (2009). When do losses count? six fallacies of natural hazards loss data. *Bulletin of the American Meteorological Society*, *90*, 799–809.

Gall, M., Borden, K. A., Emrich, C. T., & Cutter, S. L. (2011). The unsustainable trend of natural hazard losses in the united states. *Sustainability*, *3*(11), 2157–2181. https://doi.org/10.3390/su3112157

Gallaugher, M. P. B., & McNicholas, P. D. (2017). A mixture of matrix variate bilinear factor analyzers. https://doi.org/10.48550/ARXIV.1712.08664

Galway, J. G. (1989). The evolution of severe thunderstorm criteria within the weather service. *Weather and Forecasting*, *4*(4), 585–592. https://doi.org/10.1175/1520-0434(1989)004⟨0585:teostc⟩2.0.co;2

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman & Hall/CRC.

Gelman, A., & Hill, J. (2006). *Analytical methods for social research: Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Goldfeld, S. M., & Quandt, R. E. (1965). Some tests for homoscedasticity. *Journal of the American Statistical Association*, *60*(310), 539–547. https://doi.org/10.1080/01621459.1965.10480811

Gramling, C. (2022). Extreme weather in 2022 showed the global impact of climate change. https://www.sciencenews.org/article/extreme-weather-climate-change-2022

Gröschl, J., & Steinwachs, T. (2017). Do natural hazards cause international migration?∗. *CESifo Economic Studies*, *63*(4), 445–480. https://doi.org/10.1093/cesifo/ifx005

Gupta, A. K., & Nagar, D. K. (2018). *Matrix variate distributions*. Chapman; Hall/CRC.

Healy, D., Mohammed, Z., Kanwal, N., Asghar, M. N., & Ansari, M. S. (2022). Deep learning model for thunderstorm prediction with class imbalance data. In *Lecture notes in networks and systems* (pp. 195–205). Springer Nature Singapore. https://doi.org/10.1007/978-981-16-7618-5_17

Hillier, J. K., & Dixon, R. S. (2020). Seasonal impact-based mapping of compound hazards. *Environmental Research Letters*, *15*(11), 114013. https://doi.org/10.1088/1748-9326/abbc3d

Hoffmann, R., Dimitrova, A., Muttarak, R., Cuaresma, J. C., & Peisker, J. (2020). A meta-analysis of country-level studies on environmental change and migration. *Nature Climate Change*, *10*(10), 904–912. https://doi.org/10.1038/s41558-020-0898-6

Hoffmann, R., Šedová, B., & Vinke, K. (2021). Improving the evidence base: A methodological review of the quantitative climate migration literature. *Global Environmental Change*, *71*, 102367. https://doi.org/10.1016/j.gloenvcha.2021.102367

Hollander, M., Wolfe, D. A., & Chicken, E. (2015). *Nonparametric statistical methods*. Wiley. https://doi.org/10.1002/9781119196037

Holmes, E. E., Ward, E. J., & Kellie, W. (2012). Marss: Multivariate autoregressive state-space models for analyzing time-series data. *R J.*, *4*(1), 11.

How thunderstorms form. (2022). https://scied.ucar.edu/learning-zone/storms/how-thunderstorms-form

IRS. (2023). https://www.irs.gov/statistics/soi-tax-stats-migration-data

Islam, M. R., & Khan, N. A. (2018). Threats, vulnerability, resilience and displacement among the climate change and natural disaster-affected people in south-east asia: An overview. *Journal of the Asia Pacific Economy*, *23*(2), 297–323. https://doi.org/10.1080/13547860.2018.1442153

Jevtic, P., & Gall, M. (2023). Effects of natural disasters on spatio-temporal patterns of crime types in the united states. https://cina.gmu.edu/projects/effects-of-natural-disasters-on-spatio-temporal-patterns-of-crime-types-in-the-united-states/

Kaczan, D. J., & Orgill-Meyer, J. (2020). The impact of climate change on migration: A synthesis of recent empirical insights. *Climatic Change*, *158*(3), 281–300. https://doi.org/10.1007/s10584-019-02560-0

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*(3), 187–200. https://doi.org/10.1007/bf02289233

Kapsch, M.-L., Kunz, M., Vitolo, R., & Economou, T. (2012). Long-term trends of hail-related weather types in an ensemble of regional climate models using a bayesian approach. *Journal of Geophysical Research: Atmospheres*, *117*(D15), n/a–n/a. https://doi.org/10.1029/2011jd017185

Karl, T. R., & Koss, W. J. (1984). Regional and national monthly, seasonal, and annual temperature weighted by area, 1895-1983. *istorical Climatology Series 4-3*, 38.

Kean, J. W., Staley, D. M., Rengers, F. K., Allstadt, K. E., Coe, J. A., Sigman, A. J., & Hernandez, J. (2019). Debris-flow inundation and damage data from the 9 january 2018 montecito debris-flow event. https://doi.org/10.5066/P9JQJU0E

Klein, R., Nicholls, R., & Thomalla, F. (2003). The resilience of coastal megacities to weather-related hazards. In A. Kreimer, M. Arnold, & A. Carlin (Eds.), *Building safer cities* (pp. 101–120). World Bank.

Klepp, S. (2017). Climate change and migration - an overview, oxford research encyclopedia of climate science. *Oxford Research Encyclopedia of Climate Science*.

Knutson, T., Camargo, S. J., Chan, J. C. L., Emanuel, K., Ho, C.-H., Kossin, J., Mohapatra, M., Satoh, M., Sugi, M., Walsh, K., & Wu, L. (2019). Tropical cyclones and climate change assessment: Part i: Detection and attribution. *Bulletin of the American Meteorological Society*, *100*(10), 1987–2007. https://doi.org/10.1175/bams-d-18-0189.1

Kossin, J. P. (2018). A global slowdown of tropical-cyclone translation speed. *Nature*, *558*(7708), 104–107. https://doi.org/10.1038/s41586-018-0158-3

Kunz, M., & Kugel, P. I. (2015). Detection of hail signatures from single-polarization c-band radar reflectivity. *Atmospheric Research*, *153*, 565–577. https://doi.org/10.1016/j.atmosres.2014.09.010

Lopez, G. (2022). A summer of climate disasters. https://www.nytimes.com/2022/09/07/briefing/climate-change-heat-waves-us-europe.html

Lyubchich, V., Newlands, N. K., Ghahari, A., Mahdi, T., & Gel, Y. R. (2019). Insurance risk assessment in the face of climate change: Integrating data science and statistics. *WIREs Computational Statistics*, *11*(4). https://doi.org/10.1002/wics.1462

Mahanama, T., Shirvani, A., & Rachev, S. (2021). A natural disasters index. *Environmental Economics and Policy Studies*, *24*(2), 263–284. https://doi.org/10.1007/s10018-021-00321-x

Martinez, A. B. (2020). Improving normalized hurricane damages. *Nature Sustainability*, *3*(7), 517–518. https://doi.org/10.1038/s41893-020-0550-5

Mechler, R., & Bouwer, L. M. (2014). Understanding trends and projections of disaster losses and climate change: Is vulnerability the missing link? *Climatic Change*, *133*(1), 23–35. https://doi.org/10.1007/s10584-014-1141-0

Met Office. (2023). The russian heatwave of summer 2010. https://www.metoffice.gov.uk/weather/learn-about/weather/case-studies/russian-heatwave

Milman, O., Witherspoon, A., Liu, R., & Chang, A. (2021). The climate disaster is here – this is what the future looks like. https://www.theguardian.com/environment/ng-interactive/2021/oct/14/climate-change-happening-now-stats-graphs-maps-cop26

Mostafiz, R. B., Friedland, C. J., Rohli, R. V., Gall, M., Bushra, N., & Gilliland, J. M. (2020). Census-block-level property risk estimation due to extreme cold temperature, hail, lightning, and tornadoes in louisiana, united states. *Frontiers in Earth Science*, *8*. https://doi.org/10.3389/feart.2020.601624

Myers, C. A., Slack, T., & Singelmann, J. (2008). Social vulnerability and migration in the wake of disaster: The case of hurricanes katrina and rita. *Population and Environment*, *29*(6), 271–291. https://doi.org/10.1007/s11111-008-0072-y

NASA. (2023). The effects of climate change. https://climate.nasa.gov/effects/

National data fixed assets accounts tables. (n.d.). https://apps.bea.gov/iTable/?ReqID=10&step=2

NCEI. (2023). Geographical Reference Maps — National Centers for Environmental Information (NCEI) — ncei.noaa.gov [[Accessed 02-11-2023]].

Neumann, K., Sietz, D., Hilderink, H., Janssen, P., Kok, M., & van Dijk, H. (2015). Environmental drivers of human migration in drylands – a spatial picture. *Applied Geography*, *56*, 116–126. https://doi.org/10.1016/j.apgeog.2014.11.021

Nicholls, R. J. (1995). Coastal megacities and climate change. *GeoJournal*, *37*(3), 369–379. https://doi.org/10.1007/bf00814018

NOAA. (2022). Severe weather 101: Thunderstorm basics. https://www.nssl.noaa.gov/education/svrwx101/thunderstorms/

Palma, W. (2016). *Time series analysis*. Wiley-Blackwell.

Partridge, M. D., Feng, B., & Rembert, M. (2017). Improving climate-change modeling of US migration. *American Economic Review*, *107*(5), 451–455. https://doi.org/10.1257/aer.p20171054

Pielke, R. (2020). Economic 'normalisation' of disaster losses 1998–2020: A literature review and assessment. *Environmental Hazards*, *20*(2), 93–111. https://doi.org/10.1080/17477891.2020.1800440

Pielke, R. A., Gratz, J., Landsea, C. W., Collins, D., Saunders, M. A., & Musulin, R. (2008). Normalized hurricane damage in the united states: 1900–2005. *Natural Hazards Review*, *9*(1), 29–42. https://doi.org/10.1061/(asce)1527-6988(2008)9:1(29)

Pielke, R. A., & Sarewitz, D. (2005). Bringing society back into the climate debate. *Population and Environment*, *26*(3), 255–268. https://doi.org/10.1007/s11111-005-1877-6

Pinto, J., Karremann, M., Born, K., Della-Marta, P., & Klawa, M. (2012). Loss potentials associated with european windstorms under future climate conditions. *Climate Research*, *54*(1), 1–20. https://doi.org/10.3354/cr01111

Poston, D. L., Zhang, L., Gotcher, D. J., & Gu, Y. (2009). The effect of climate on migration: United states, 1995–2000. *Social Science Research*, *38*(3), 743–753. https://doi.org/10.1016/j.ssresearch.2008.10.003

Prices amp; inflation. (n.d.). https://www.bea.gov/resources/learning-center/what-to-know-prices-inflation

Refan, M., Romanic, D., Parvu, D., & Michel, G. (2020). Tornado loss model of oklahoma and kansas, united states, based on the historical tornado data and monte carlo simulation. *International Journal of Disaster Risk Reduction*, *43*, 101369. https://doi.org/10.1016/j.ijdrr.2019.101369

Community Disaster Resilience Zones Act. (2022). Pub. l. no. 17–255, 136 stat. 2363.

Saldaña-Zorrilla, S. O., & Sandberg, K. (2009). Spatial econometric model of natural disaster impacts on human migration in vulnerable regions of mexico. *Disasters*, *33*(4), 591–607. https://doi.org/10.1111/j.1467-7717.2008.01089.x

Sander, J., Eichner, J. F., Faust, E., & Steuer, M. (2013). Rising variability in thunderstorm-related u.s. losses as a reflection of changes in large-scale thunderstorm forcing$*$. *Weather, Climate, and Society*, *5*(4), 317–331. https://doi.org/10.1175/wcas-d-12-00023.1

Schmeits, M. J., Kok, K. J., Vogelezang, D. H. P., & van Westrhenen, R. M. (2008). Probabilistic forecasts of (severe) thunderstorms for the purpose of issuing a weather alarm in the netherlands. *Weather and Forecasting*, *23*(6), 1253–1267. https://doi.org/10.1175/2008waf2007102.1

Šedová, B., Čizmaziová, L., & Cook, A. (2021). A meta-analysis of climate migration literature. *CEPA Discussion Papers; 29*. https://doi.org/10.25932/PUBLISHUP-49982

SHELDUS. (2023). Spatial hazard events and losses database for the united states. https://cemhs.asu.edu/sheldus

Shukla, R., Agarwal, A., Sachdeva, K., Kurths, J., & Joshi, P. K. (2018). Climate change perception: An analysis of climate change and risk perceptions among farmer types of indian western himalayas. *Climatic Change*, *152*(1), 103–119. https://doi.org/10.1007/s10584-018-2314-z

Shumway, R. H., & Stoffer, D. S. (1982). AN APPROACH TO TIME SERIES SMOOTHING AND FORECASTING USING THE EM ALGORITHM. *Journal of Time Series Analysis*, *3*(4), 253–264. https://doi.org/10.1111/j.1467-9892.1982.tb00349.x

Spatial hazard events and losses database for the united states user agreement. (2023). https://sheldus.asu.edu/SHELDUS/docs/END_USER_LICENSE_AGREEMENT.pdf

Spinoni, J., Naumann, G., Carrao, H., Barbosa, P., & Vogt, J. (2013). World drought frequency, duration, and severity for 1951-2010. *International Journal of Climatology*, *34*(8), 2792–2804. https://doi.org/10.1002/joc.3875

Stock, J. H., & Watson, M. (2011). Dynamic factor models. *Oxford Handbooks Online*.

Storm data faq page. (2023). https://www.ncdc.noaa.gov/stormevents/faq.jsp

Thorson, J. (2020). Convective storm modeling: The difference is in the details. https://www.assetworks.com/convective-storm-modeling-details-rm20/

Vahedifard, F., AghaKouchak, A., & Jafari, N. H. (2016). Compound hazards yield louisiana flood. *Science*, *353*(6306), 1374–1374. https://doi.org/10.1126/science.aai8579

The verisk severe thunderstorm model for the united states. (2022). *Verisk*. https://doi.org/ https://www.air-worldwide.com/siteassets/Publications/Brochures/documents/us-severe-thunderstorm-brochure

Viroli, C. (2012). On matrix-variate regression analysis. *Journal of Multivariate Analysis*, *111*, 296–309.

Visser, H., & Petersen, A. C. (2012). Inferences on weather extremes and weather-related disasters: A review of statistical methods. *Climate of the Past*, *8*(1), 265–286. https://doi.org/10.5194/cp-8-265-2012

Watson, P. L., Koukoula, M., & Anagnostou, E. (2021). Influence of the characteristics of weather information in a thunderstorm-related power outage prediction system. *Forecasting*, *3*(3), 541–560. https://doi.org/10.3390/forecast3030034

Weinkle, J., Landsea, C., Collins, D., Musulin, R., Crompton, R. P., Klotzbach, P. J., & Pielke, R. (2018). Normalized hurricane damage in the continental united states 1900–2017. *Nature Sustainability*, *1*(12), 808–813. https://doi.org/10.1038/s41893-018-0165-2

Xie, X., Yan, S., Kwok, J. T., & Huang, T. S. (2008). Matrix-variate factor analysis and its applications. *IEEE Transactions on Neural Networks*, *19*(10), 1821–1826. https://doi.org/10.1109/TNN.2008.2004963

Zscheischler, J., Westra, S., van den Hurk, B. J. J. M., Seneviratne, S. I., Ward, P. J., Pitman, A., AghaKouchak, A., Bresch, D. N., Leonard, M., Wahl, T., & Zhang, X. (2018). Future climate risk from compound events. *Nature Climate Change*, *8*(6), 469–477. https://doi.org/10.1038/s41558-018-0156-3

Zuur, A. F., Fryer, R. J., Jolliffe, I. T., Dekker, R., & Beukema, J. J. (2003). Estimating common trends in multivariate time series using dynamic factor analysis. *Environmetrics*, *14*(7), 665–685. https://doi.org/10.1002/env.611

Zuzak, C., Mowrer, M., Goodenough, E., Burns, J., Ranalli, N., & Rozelle, J. (2022). The national risk index: Establishing a nationwide baseline for natural hazard risk in the US. *Natural Hazards*, *114*(2), 2331–2355. https://doi.org/10.1007/s11069-022-05474-w

APPENDIX A

CONVECTIVE STORM LOSSES 1970 AND 2020 BY HAZARD TYPE

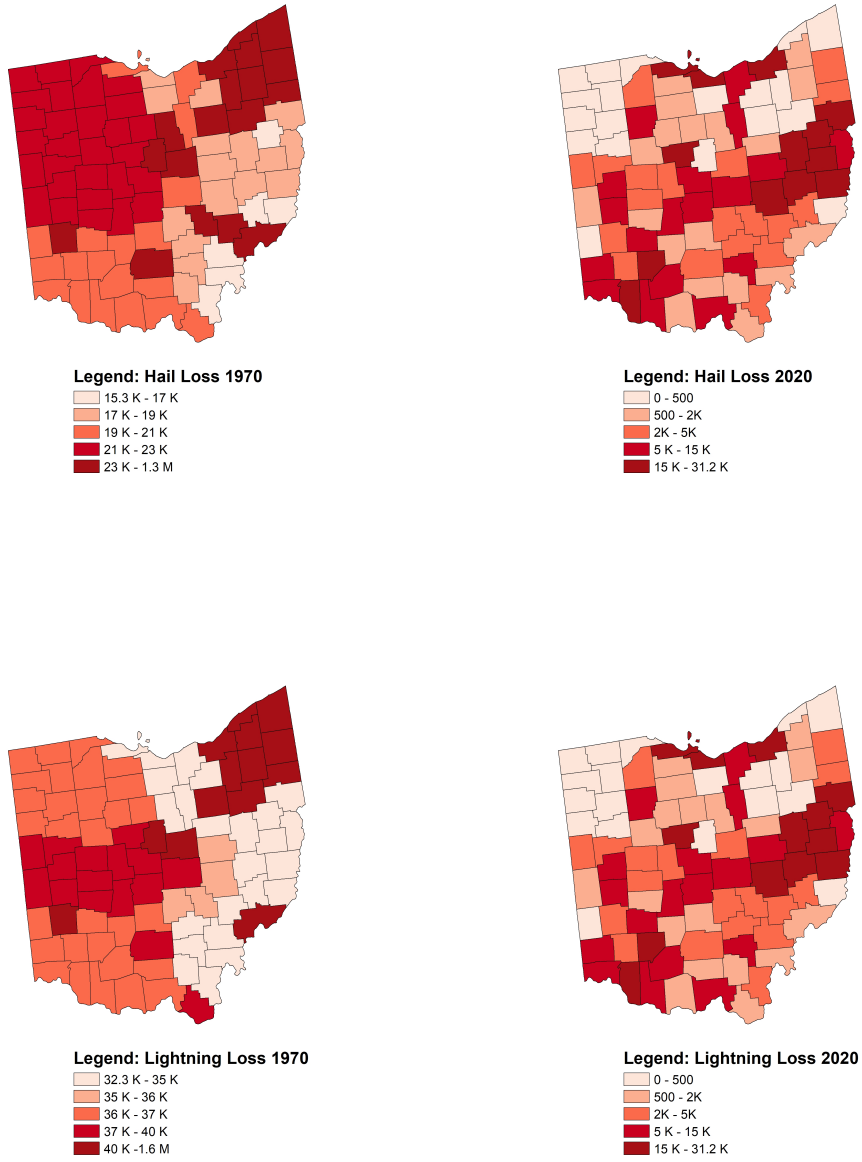

Figure 1: Convective storm losses in Ohio, reported in 2020 U.S. Dollars (SHELDUS, 2023 Version 21).

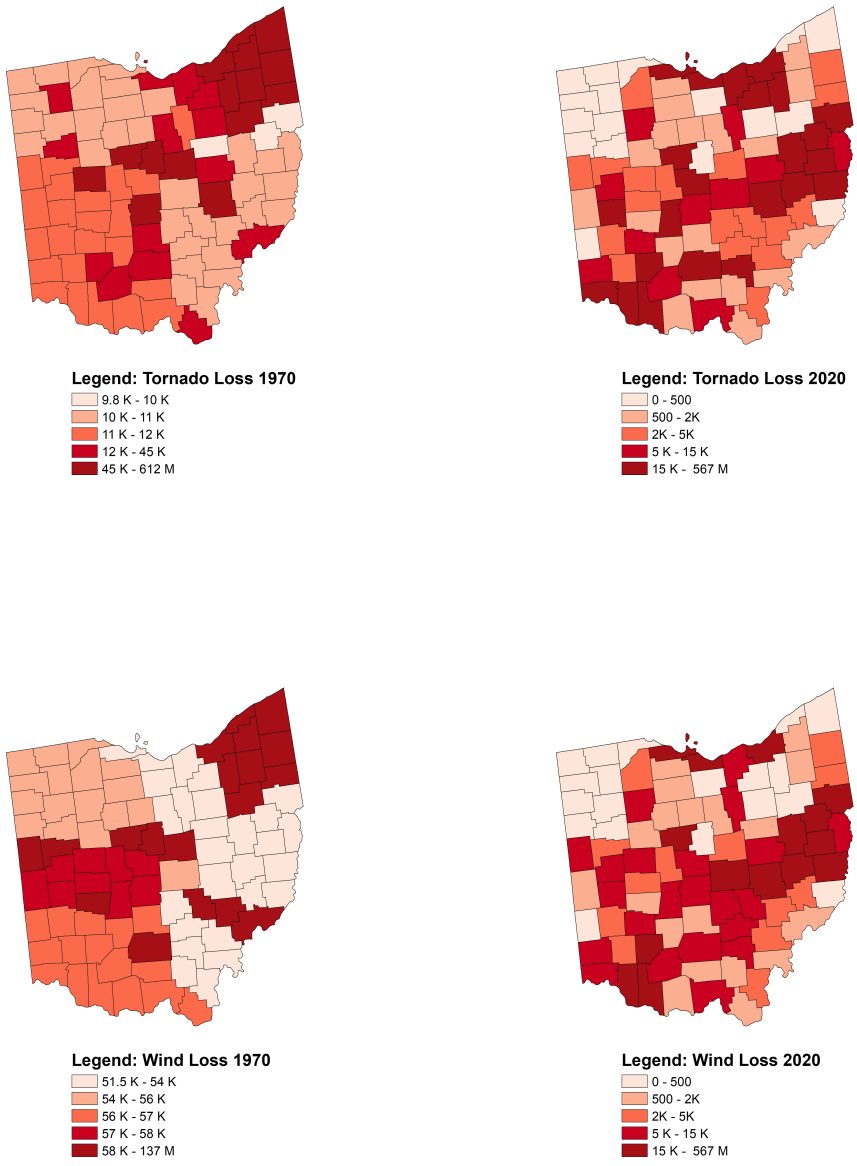

Figure 2: Convective storm related losses in 2020 U.S. Dollars SHELDUS, 2023.