QUANTIFYING THE EFFECT OF CLIMATE ON RESPIRATORY VIRUSES: FROM SEASONALITY TO INTERANNUAL VARIABILITY

BY

ADRIANA MORALES MIRANDA

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Mathematics
in the Graduate College of the
University of Illinois Urbana-Champaign, 2023

Urbana, Illinois

Doctoral Committee:

Professor Lee DeVille, Chair
Assistant Professor Pamela P. Martinez
Associate Professor Kay Kirkpatrick
Professor Zoi Rapti

Abstract

Environmental factors have been shown to partially explain the seasonality and interannual variability of infectious diseases. Despite the wide amount of research in this area, the link between climatic factors and the transmission of respiratory diseases at different temporal scales is still poorly understood, especially in the Southern Hemisphere. In this thesis, we focus on the transmission dynamics of Respiratory Syncytial Virus (RSV) and Influenza A, two of the leading pathogens responsible for substantial morbidity and mortality worldwide. Recently available data of weekly lab-confirmed cases from Chile provide a unique opportunity to understand the effect of temperature and specific humidity on the seasonality and interannual variability of RSV and Influenza A.

Using statistical models and techniques we perform an in depth data analysis to explore the association of disease incidence and environmental drivers across all of Chile. In addition, we quantify the effect of climate covariates on the transmission dynamics by formulating multiple mechanistic compartmental models that take into account intrinsic and extrinsic factors that may impact disease transmission. Using a partially observed Markov process modeling framework, we construct Susceptible-Infected-Recovered stochastic transmission models that mimic the dynamics of disease spread and perform iterated filtering to calculate the maximum log likelihood estimate of model parameters using data from the capital of Chile.

Our statistical analyses show consistent annual seasonality, with cases prevalent during the winter months across all of Chile. Specific humidity is significantly associated with the mean timing onset of RSV, where a 5 g/kg increase in mean annual specific humidity and a 1 °C increase in mean annual temperature shift the timing of the RSV epidemic back by one week. These associations are weaker for Influenza A. Furthermore, even though both viruses are sensitive to climate their onset patterns are very different, with RSV starting in the North and Influenza A in the South, highlighting the inherent differences in pathogen virology. Moreover, we observe a latitudinal gradient with respect to the climate covariates suggesting that both viruses can survive during very different winter conditions. When looking at the results from both statistical and mechanistic models, we found no significant association

between climate covariates and the year-to-year variation in amplitude of epidemics for RSV and Influenza A. Nevertheless, the mechanistic epidemiological models presented in this study are able to capture the timing and seasonality of disease onset.

This thesis provides a platform for understanding the impact of environmental factors on the transmission dynamics of RSV and Influenza A. Determining the spatiotemporal transmission patterns of infectious diseases and the environmental factors impacting transmission is one of the key steps in infectious disease modeling and control. While both RSV and Influenza A exhibit strong seasonal epidemics during the winter months, more research is needed to better understand the associations between climate drivers, pathogen survival, human mobility, age structure, and disease transmission. Identifying specific factors relevant to a particular virus in a certain geographic location can help public health professionals make informed decisions, effectively respond to outbreaks, and prioritize the allocation of resources.

To my sister Frances, I couldn't have done this without you.

Acknowledgments

I would like to express my deepest gratitude to my parents, sister, partner, and friends for their unwavering support throughout my journey. Your belief in me, your patience and understanding, and the countless hours you devoted to listening to me talk about my research have been a constant source of motivation.

I am deeply grateful to my thesis advisor, Pamela P. Martinez. Her guidance, support, and expertise have been invaluable throughout this process. Her willingness to take a chance on me and invest time in my work has been invaluable, and I am proud of what we have accomplished together.

I would also like to thank my advisor Kay Kirkpatrick for her mentorship, valuable feedback, and support throughout my graduate studies. I am grateful for her belief in me and my abilities, and for always encouraging me in my endeavors.

I want to thank my doctoral committee members, Lee DeVille and Zoi Rapti, for their willingness to take the time to read my work and give me feedback. I want to thank my colleagues in the PaDAS Lab, mentors, and professors who have helped me along the way. I would also like to extend my thanks to the Alfred P. Sloan Foundation's MPHD Program (awarded in 2017) for partial support of my graduate studies.

Finally, I would like to express my appreciation to everyone who has contributed to this thesis in any way, whether directly or indirectly. This journey has been challenging and at times, overwhelming, but with the support of the people around me, I have been able to achieve my goal.

Thank you all.

Table of Contents

Chapter 1

Introduction

1.1 Problem statement

Considerable attention has been put forth to further understand the mechanisms governing infectious disease dynamics worldwide, but some are still poorly understood. One of them is the role that climate factors play in modulating the seasonality and interannual variability of disease incidence in space and time. In particular, while environmental factors have been used to partially explain the seasonality of respiratory infections, a better understanding of the link between climatic factors and the transmission of infectious diseases is needed, with potential implications for our ability to predict the timing and magnitude of outbreaks.

Recent studies have shown that the transmission of the Influenza A virus increases in drier and colder conditions, and this effect is likely due to the impact of humidity and temperature on the stability of the virus within aerosols and the stability of respiratory droplets in the air [LS14, SK09]. In a study analyzing data from 85 countries, Azziz-Baumgartner *et al.* observed that the timing of influenza epidemics in temperate climates, correlates with low temperatures [ABDN+12]. Peak infection rates were typically observed during or immediately following the coldest month of the year when mean monthly sunshine, precipitation, absolute humidity and latitude were used as parameters. In contrast, influenza viruses in the tropics tended to circulate year-round or to appear in multiple epidemics in a given year. In [YKL+21], the authors developed a model capable of capturing the diverse seasonal pattern in tropical and subtropical climates. Both very low and very high humidity levels were found to facilitate transmission of Influenza A in Hong Kong. This relationship is also modified by temperature such that, when temperature is above some cutoff value, transmissibility is reduced.

Furthermore, in the United States, absolute humidity has been associated with the seasonal onset of influenza virus infection [SK09, SPV+10]. To better understand viral seasonality, animal models have been developed to assess how not only temperature, but also humidity, play a role in infection of influenza virus. Lowen *et al.* were able to demonstrate the effects of temperature and humidity on influenza virus transmission among experimental guinea pigs

[LMSP07, LS14]. Multiple mathematical and statistical models have been implemented to try and understand the impact of climatic factors on the onset and year-to-year variability of Influenza A, but there is no conclusion on whether or not they are the leading factor [SPV+10, SGL11, BAC+13, PB14, YLS15, OA15], and whether similar effects can be found in the Southern Hemisphere.

Compared to Influenza A, Respiratory Syncytial Virus (RSV) is consistently less studied and understood with respect to the role climatic factors play in modulating the seasonality and interannual variability of epidemics. In fact, no laboratory experiments have been done to understand the influence of climatic factors (e.g. temperature, absolute humidity, precipitation) on the survival and transmission capabilities of RSV. Existing studies rely on observational data to understand the role of climatic factors on the transmission of RSV [WWM+98, OJR+18, BMW+19, BAC+13, PVA+15, TL14, WPG+10]. When considering the seasonality of RSV in various climates, studies have shown that, similarly to Influenza A, RSV peak infection varies amongst different locations [OJR+18, BMW+19]. For example, peaks of infection for RSV have been found to correlate with colder weather in temperate, Mediterranean, and desert regions, while precipitation has been identified as a more accurate predictor of infection in tropical and subtropical regions [WWM+98]. In a study comparing RSV infection in temperate, tropical and subtropical climates, RSV infection was found to display seasonality much like the Influenza virus in temperate climates, with infection rates rising during cold and dry winter months, and a wide range of variability in the timing and duration of epidemics in the tropics [BAC+13]. Another review study found similar patterns, with temperate locations of the Northern and Southern Hemispheres characterized by focused peaks of activity during their respective winters [OJR+18]. Although a general pattern exists for RSV, the described variations from season to season within countries suggest that multiple environmental factors may modulate the transmission of the pathogen.

1.2 Motivation

Combining statistical analysis and mathematical modeling can provide a more comprehensive understanding of the transmission dynamics of RSV and Influenza A and how these two pathogens might differ in their sensitivity to climate variability. With the increasing use and complexity of mathematical models, comes the need for better tools to analyze them. The recent advances in high powered computing have made it possible to develop methods that allow us to perform statistically rigorous analysis on mechanistic models of infectious disease dynamics [FK20]. More specifically, the development of sophisticated inference tools

has made it possible to make progress in estimating key of epidemiological parameters that improve the modeling of infectious diseases [Bre18, NKd+23, FPW16, INA+15, DH18].

Models that allow for epidemiological forecasting and the understanding of the fundamental mechanisms that influence the spread of epidemics have become a critical part of the development of public health policies [ADL+20, Pan20, HB20]. For example, such models can help predict the timing and severity of seasonal outbreaks, estimate relevant parameters, identify vulnerable populations, and evaluate the potential impact of interventions. This information can then be used to inform decisions on public health measures such as vaccination campaigns, administration of prophylactic antibodies, social distancing guidelines, and other interventions aimed at reducing the spread of these viruses and minimizing their impact on public health [HBP+20, LE22]. Effectively estimating the size of an outbreak for a particular year can be essential for hospitals to assess their capacity and prepare for an influx of patients. Hospitals need to know the number of expected cases to ensure they have enough resources, including staff, equipment, and supplies, to provide appropriate care [RN14]. Climate and humidity control in hospital units would perhaps be helpful to minimize the spread of a virus to others. It has also been suggested that better ventilation of indoor environments would be an appropriate preventative measure against infection [LLT+07].

Anticipating the associated changes in the burden of infectious diseases can help develop mitigation strategies and the planning of preparedness and response for a specific infectious disease, especially in the context of climate change. Identification of specific factors that are relevant for a particular virus in a certain geographic location is of critical importance to best prevent viral spread in a particular community [PB14]. In addition, estimating the size of an outbreak can help plan for the allocation of resources between different regions. If some locations are likely to have a bigger than usual epidemic, then they may need additional support and resources, especially in places with lower socioeconomic status [HBP+20].

1.3 Contributions to the literature

Both mathematical models and epidemiological data suggest that environmental factors play a large role in the transmission efficiency of these viruses. Overall, climatic factors together with other factors such as immune response and human behavior, are likely to act together and influence the seasonality and interannual variation of disease incidence observed in common respiratory viruses. As described above, attempts has been made to understand the role of climatic factors on the dynamics of RSV and Influenza A. However, lack of data on disease incidence in different climates makes it hard to reach definitive conclusions on the

effects of climatic factors. For instance, in a recent review [BAC+13], the authors note that understanding the effect of latitudinal gradients on the seasonality of RSV and Influenza A is limited by the access to data from different locations. In order to provide a global picture of the seasonal patterns, the authors needed to be fairly inclusive, and some of the studies included had small sample size and/or short duration, potentially biasing the analysis.

Recently available data of multiple respiratory viruses from Chile provide us with the unique opportunity to understand the effect of climate factors on the seasonality and inter-annual variability of respiratory viruses in the Southern Hemisphere, as well as to compare the similarities and difference of these two viral pathogens with similar seasonality but different life history. The climate of Chile comprises a wide range of weather conditions across a large geographic scale, extending across 39 degrees (2698 miles) in latitude, but approximately 9 degrees (110 miles) in longitude for the continental territory. According to the Köppen system, Chile has at least seven major climatic sub-types, which makes it a suitable place to study how environmental drivers affect the incidence of respiratory viruses like RSV and Influenza A.

Data availability can be limited in certain settings, such as low-income countries or during rapidly evolving epidemics where data collection may be challenging. We have 9 years of weekly data, collected across 16 different regions in Chile. All of the data is collected as part of Chile's respiratory virus surveillance initiative conducted by the Institute of Public Health (ISP) within the Health Ministry of Chile's Government [isp]. Each region has multiple sentinel hospitals, where the cases are laboratory confirmed from samples collected, which makes the data very reliable.

The data includes multiple respiratory viruses, which allows us to compare the diseases across regions. RSV and Influenza A are known for impacting different age groups, having different pathogen evolution, as well as different connectivity across regions. They can also differ in their onset, duration, and complications, highlighting the importance of individualized prevention and management strategies. Using statistical models we assess the observed differences between the two viruses. Furthermore, by implementing distinct mathematical models that mimic the transmission of the viruses we use the data to perform parameter inference. The estimated parameters allow us to compare and contrast recovery period, the time of waning immunity, and the impact environmental factors may have on the dynamics of the viruses.

Our work serves as a framework for understanding the impact of environmental factors on the transmission dynamics of diseases. Even though our models are specifically constructed for RSV and Influenza A they can be modified to mimic the transmission dynamics of other pathogens. The methods below can be used to answer multiple questions about the

transmission dynamics of infectious diseases and perform inference on relevant parameters that are used to understand disease spread.

1.4 Methods

Infectious disease modeling encompasses a wide range of methods, where statistical and mechanistic approaches go hand in hand. By combining statistical analyses with mathematical modeling of the transmission dynamics of RSV and Influenza A we aim to understand the effect of temperature and specific humidity on the transmission dynamics of the diseases. First, we use statistical methods to describe, visualize, and interpret the data. Statistical methods allow us to answer questions about the data, test hypotheses, describe associations (correlations), and model relationships (regression) within the data without incorporating the process that governs the dynamics of the virus. To explore the seasonality we look at the density of cases for each region independently and establish the periodicity of the outbreaks by using wavelet analysis. We also calculate critical climate thresholds using segmented regression to establish the relationship between climate covariates and the increase of disease activity for each region. The timing of onset and amplitude of disease incidence is explored by using linear regression. Furthermore, we use climatic and non-climatic variables to fit and analyze a variety of models, including multivariate logistic regression, linear regression, and generalized additive models, to quantify the relationship between seasonality and climate without incorporating dynamical processes.

To further understand and quantify the effect of environmental drivers on the transmission of respiratory viruses we use partially observed Markov process (POMP) models, also known as Hidden Markov Models. POMP models allow us to model infectious disease time series data as a noisy and partially observed realization of the disease transmission process that is assumed to be Markov. This method can be used to test different hypotheses, to quantify parameters, and to infer unobserved state variables. This idea was first introduced by King *et al.* [KDI16]. POMP models consist of an unobserved Markov state process (that describes the dynamics of disease transmission), connected to the data via an explicit model of the measurement process.

A useful tool to mimic the mechanisms of disease spread between individuals are the classic Susceptible-Infected-Recovered (SIR) compartmental models. These dynamical models allow us to follow the movement of individuals from one compartment to the other (i.e., going from being susceptible to infected and subsequently recovered) through time. The deterministic nature of these models allows us to determine the future state of the system solely by

its current state and the mathematical equations governing its behavior. This means the models are Markov, and thus suitable for the construction of our state process. SIR-type models allow us to understand underlying mechanisms driving disease transmission, as they can capture the overall trends and patterns of disease spread, making them very popular in infectious disease research. However, deterministic models do not capture the inherent stochasticity of infectious disease transmission, which can result in significant variability in disease outcomes. To address this limitation, we can add stochasticity to the models through measurement and process noise (e.g. incorporating a probability distribution to simulate the occurrence of stochastic events), providing a more realistic representation of disease dynamics.

Stochastic SIR models can become quite complex as more factors are incorporated such as multiple compartments, demography (i.e., births/deaths), maternal immunity, immunity due to past infections, and seasonality, in order to make models more realistic. The predictive ability of the models also relies on accurate and timely data on disease incidence. However, data can have underreporting and overdispersion, requiring the addition of measurement stochasticity. Estimating all the parameters accurately and efficiently can be challenging, as they can vary depending on factors such as the reliability and completeness of the data, the amount of parameters, the characteristics of the pathogen, as well as measurement noise and process noise. The increased complexity can make the models more difficult to understand and interpret, and can also require more computational resources to implement. Inference of parameters based on the incidence data becomes harder to do because the likelihood is often a complex and high dimensional integral, with no analytical solution.

In cases where the likelihood function is difficult to evaluate we can use simulation-based methods [KBM+99, ADH10, TWS+09, Bre18]. We use iterated filtering [IBK06, INA+15] for maximum likelihood estimation in partially observed epidemic models. This method generates samples from the density of the Markov process instead of evaluating it. The basic idea of iterated filtering is to apply a particle filter [AdFG01] to a model in which the parameter vector for each particle is following a random walk in time. The particle filter approximates the likelihood of the perturbed model by representing it with a set of particles. The particles are then weighted according to their likelihoods under the current approximation to the likelihood function, and re-sampled to generate a new set of particles for the next iteration. The re-sampling techniques make them more efficient than standard Monte Carlo methods [ADH10], which require a lot of simulations in order to obtain an estimate of the likelihood that is suitable for parameter estimation and model selection [Sto19, SBH20, KIMB+22, Bre18]. The `pomp` package [KIMB+22], available in the R programming language, provides the framework for efficient simulation and inference of POMP

models.

We implement two stochastic SIR-type models. The model for RSV is a MSIRS model, meaning that individuals are born into the M population with protective maternal immunity and after becoming infected and recovered individuals can move back into the susceptible compartment because of waning immunity. The MSIRS model also allows for multiple reinfections, with a reduction in susceptibility following the first infection and progressive build up of immunity. For Influenza A we implement a SIRS model, meaning individuals have waning immunity but no maternal immunity. Note that the Influenza A model does not include reduction in susceptibility or progressive build up because of the rapid antigenic evolution of the virus [YKH+13]. Different models are chosen because of the inherent differences in antigenic evolution and in age of infection of each pathogen. Since both respiratory viruses are highly seasonal, we include seasonality through the transmission term. To assess the effect of climate covariates on the dynamics of disease spread we implement different variations of each model by changing the transmission rate in four different ways:

- the transmission rate includes seasonality through the addition of 4 cubic b-splines, but no inter-annual effect,
- the transmission rate includes both seasonal and inter-annual effect through the addition of temperature,
- the transmission rate includes both seasonal and inter-annual effect through the addition of specific humidity, and
- the transmission rate includes both seasonal and inter-annual effect through the addition of both temperature and specific humidity.

Likelihood-based criteria is used for model selection. We use Akaike Information Criterion scores, which take into account model complexity and penalize the likelihood based on the number of parameters. The likelihood ratio test, also known as Wilks test, is then used to assess the goodness of fit of the selected model based on the ratio of their likelihoods.

1.5 Outline

The thesis is organized as follows. Chapter 2 introduces the main concepts of SIR-type models, POMP models, and iterated filtering. The incidence and climate data used for analysis is described in Chapter 3. We also perform an exploratory analysis of environmental

drivers by describing spatiotemporal patterns using multiple statistical techniques. In Chapter 4 we describe and implement statistical and mechanistic compartmental models for RSV and Influenza A. Using the POMP modeling and iterated filtering framework we find the maximum likelihood estimate of key epidemiological parameters from the models. We also perform model selection based on model performance. Finally, in Chapter 5 we give concluding remarks, potential limitations of the current study and how they can be addressed in future research.

Chapter 2

Background

In this chapter we introduce the framework used to build and analyze the proposed models in Section 4.2. Section 2.1 is centered on the history, description and analysis of mathematical modeling in epidemiology, and the applications on infectious diseases. In Section 2.2 we give an overview of the theoretical background of partially observed Markov process models, and the parameter inference method used to evaluate the likelihood.

2.1 Mathematical modeling of infectious diseases

The history of mathematical modeling in epidemiology goes back to the eighteenth century when famous mathematician Daniel Bernoulli first expressed the proportion of susceptible individuals of an endemic infection (smallpox) in terms of the force of infection and life expectancy [Bra, DH00]. Since the 1900s the use of mathematical models to understand and explain the population dynamics of infectious diseases has been increasing [Bra17]. Mathematical models have been used to assess vaccination strategies for whooping cough and measles [AM82], understand HIV transmission dynamics [AMMJ86], determine the transmission potential of smallpox [GL01], find global patterns of dengue and malaria [WLC+10], as well as analyze seasonal fluctuations of diseases like influenza [LS14, SK09], RSV [BMW+19, WWM+98, PB14], and rotavirus [ASS91, MKY+16]. Most recently, the SARS-CoV-2 (COVID-19) pandemic demonstrated the importance of using mathematical models to understand not just the dynamics of the disease, but the public health and sociological impacts of a pandemic [Pan20, ADL+20, AMY+21, JAB21].

Using the language of mathematics, we can describe how disease incidence varies through space and time, and what are the factors responsible for this variation. As such, models can be used to:

1. predict the long-term behavior by using past trends [WPG+10, SK12, YCLS15],
2. assess the impact of vaccination strategies [HKK97, Hsi10],

3. understand how an infectious disease spreads in the real world [ADL+20],

4. specify how various external factors may affect the dynamics [YKH+13, BMW+19, SBH20, LE22],

5. and most importantly to help create public health interventions [JAB21, PB14, LLT+07].

The implementation of mathematical models have allowed us to account for both intrinsic and extrinsic components of disease transmission. In this section, we describe the simplest of epidemiological models, compartmental models, and go into detail on the mathematical equations describing them and important concepts in epidemiology. The following sections are adapted from the book titled "Introduction to Simple Epidemic Models" [KR08]. We encourage the reader to refer to these references for a more comprehensive and extensive explanation of the topics [KR08, DHM90, Fra80, van17, HHOW19a, Fis07a].

A compartmental model consists of categorizing individuals in the "host" population according to their infection status. We assume individuals are either susceptible to infection, currently infected and infectious, or recovered, and thus we call them Susceptible-Infected-Recovered (SIR) models. This formalism, which was initially studied in depth by Kermack and McKendric in 1927, categorizes hosts within a population as Susceptible (if previously unexposed to the pathogen), Infected (if currently infected by the pathogen and infectious), and Recovered (if they have successfully cleared the infection) [KMW27].

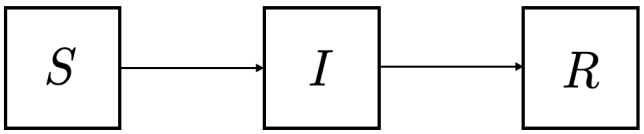

Figure 2.1: Compartmental diagram illustrating the structure of the SIR model.

Each of the compartments (S, I, and R) represent a specific stage of the epidemic (i.e., how hosts move between the susceptible, infected and recovered classes). The model can be represented by a flow diagram where arrows show the movement between the S and I classes and the I and R classes as seen in Figure 2.1 [KR08]. From a modeling perspective, this translates to the process of individuals (hosts) moving between compartments through time where $S(t)$ is the number of individuals not yet infected with the disease, $I(t)$ is the number of individuals who have been infected with the disease, and $R(t)$ is the compartment used for those individuals who have been infected and then recovered from the disease at time t.

There are multiple ways to construct compartmental models. Depending on the host population and pathogen the model could be SI (infected are assumed to remain infectious

for an average period of time after which they die), SIS (individuals can become infected multiple times with no apparent immunity), SEIR (adds an Exposed compartment before becoming infectious and will have a latent period where the pathogen's abundance in the host is not enough to cause transmission), SIRS (there is waning immunity and recovered hosts can move back to susceptible population), SICR (adds a Carrier compartment and susceptible individuals can become infected by either carriers or acutely infectious individuals) and others. For the remainder of this section we will focus on the SIRS deterministic model, giving an extensive overview of the model and its analysis, as well as how it can be improved (i.e., made more realistic) by adding stochasticy and seasonality.

2.1.1 Deterministic SIRS model

The deterministic SIRS model is characterized by having waning immunity, where a recovered individual can move back into the susceptible population (diagram in Figure 2.2).

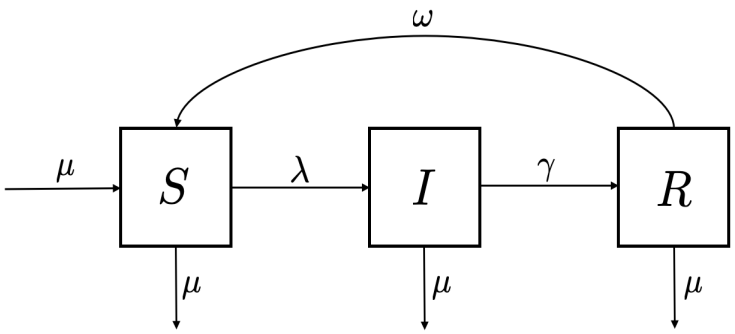

Figure 2.2: Compartmental diagram illustrating the structure of the SIRS model.

For all deterministic compartmental models, the transition rates from one compartment to another are mathematically expressed as derivatives, i.e., ordinary differential equations, in the following way:

$$\frac{dS}{dt} = \mu N - \lambda S - \mu S + \omega R, \tag{2.1}$$

$$\frac{dI}{dt} = \lambda S - \gamma I - \mu I, \tag{2.2}$$

$$\frac{dR}{dt} = \gamma I - \mu R - \omega R, \tag{2.3}$$

where S is the susceptible individuals, I is the infected individuals, and R is the recovered individuals of the population and $N = S + I + R$ at all times t (i.e., we have a closed

population). In the above equations μ is the birth/death rate, $1/\omega$ is the average duration of immunity, and $1/\gamma$ is the average infectious period. The force of infection λ is determined by three distinct factors: the prevalence of infecteds, the underlying population contact structure, and the probability of transmission given contact [KR08].

2.1.2 Force of infection λ

For a directly transmitted pathogen, like Influenza A and RSV, there has to be contact between susceptible and infected individuals, and the probability of this happening is determined by the respective levels of S and I, as well as the inherent contact structure of the host population, and the likelihood that a contact between a susceptible and an infectious person results in transmission [KR08]. Because transmission requires contact between infecteds and susceptibles, two general possibilities exist depending on how the contact structure changes with respect to population size: $\lambda = \beta I/N$ and $\lambda = \beta I$, (where I is the number of infectious individuals, N is the total population size, and β, called the transmission rate, is the product of the contact rates and transmission probability). The first of the two is called a frequency dependent (or mass action) transmission and the second as density dependent (or pseudo mass action) transmission. Frequency dependent transmission assumes the number of contacts is independent of the population size. In contrast, density dependent transmission assumes that as the population size increases, so does the contact rate. We will assume a frequency dependent force of infection and thus $\lambda = \dfrac{\beta I}{N}$ [GBF02].

2.1.3 Derivation of the transmission term β

To derive the frequency dependent transmission term, we assume homogenous mixing in the population, i.e., everyone interacts with equal probability with everyone else. Consider a susceptible individual with an average κ contacts per unit of time [AM91]. Of these, I/N are contacts with infected individuals. During a small time interval $(t,\, \delta t)$, the number of contacts with infecteds is $\kappa\frac{I}{N}\delta t$. If c is defined as the probability of successful disease transmission following a contact, then $1-c$ is the probability of no transmission. By independence of contacts, the probability $1-\delta q$ that a susceptible individual escapes infection following $\kappa\frac{I}{N}\delta t$ contacts is

$$1 - \delta q = (1-c)^{\kappa\frac{I}{N}\delta t}.$$

Taking the logarithm on both sides of the equation we get

$$\begin{aligned}\log(1-\delta q) &= \log\Big((1-c)^{\kappa\frac{I}{N}\delta t}\Big)\\ \log(1-\delta q) &= \kappa\frac{I}{N}\delta t\log(1-c)\\ 1-\delta q &= e^{\kappa\frac{I}{N}\delta t\log(1-c)}\\ \delta q &= 1-e^{\kappa\frac{I}{N}\delta t\log(1-c)}.\end{aligned}$$

Let $\beta = -\kappa\log(1-c)$. Then the probability that the individual is infected following any of these contacts is given by

$$\delta q = 1 - e^{-\beta I/N\delta t}.$$

To translate this probability into the rate at which transmission occurs, first we expand the exponential term using the Taylor series,

$$\delta q = 1 - \left(1 + (-\beta\frac{I}{N}\delta t) + \frac{(-\beta\frac{I}{N}\delta t)^2}{2!} + \frac{((-\beta\frac{I}{N}\delta t)^3}{3!} + \cdots\right).$$

If we simplify and divide both sides by δt we get

$$\frac{\delta q}{\delta t} = \beta\frac{I}{N} - \frac{(-\beta\frac{I}{N})^2}{2!}\delta t - \frac{(-\beta\frac{I}{N})^3}{3!}\delta t^2 - \cdots.$$

Taking the limit of $\delta q/\delta t$ as $\delta t \to 0$ we get

$$\lim_{\delta t\to 0}\frac{\delta q}{\delta t} = \frac{\beta I}{N},$$

which is the transmission rate per susceptible individual, or force of infection λ [AM91]. By extention, the total rate of transmission to the entire susceptible population is given by

$$\frac{d(S/N)}{dt} = -\lambda(S/N) = -\frac{\beta I}{N}\frac{S}{N},$$

which can be re-written as

$$\frac{dS}{dt} = -\lambda(S/N) = -\frac{\beta I}{N}S.$$

2.1.4 Basic reproductive ratio R_0

One of the most important concepts in epidemiology is the basic reproductive ratio, R_0, defined as *the average number of secondary cases arising from an average primary case in an entirely susceptible population* [KR08]. The use of R_0 is heavily influenced by the mathematical application of R_0 done by Diekmann *et al.* in 1990 and then applied by Anderson and May in 1991 in their book "Infectious Diseases of Humans: transmission and control" [DHM90, AM91]. The value of R_0 depends on both the the disease and the host population, and thus the same disease may have different values of R_0 [AM91]. Mathematically R_0 can be calculated as the rate at which new cases are produced by an infectious individual multiplied by the average infectious period, when the entire population is susceptible.

An important use of R_0 is *the threshold phenomenon* which states that the disease can invade if $R_0 > 1$, otherwise the disease dies out. This can be calculated by looking at Equation 2.2 and rewriting it to get

$$\begin{aligned}\frac{dI}{dt} &= \frac{\beta I}{N}S - \gamma I - \mu I \\ &= I(\frac{\beta}{N}S - \gamma - \mu).\end{aligned}$$

Since $\frac{dI}{dt} > 0$ is needed for the pathogen to spread in the population, then

$$\frac{\beta}{N}S - \gamma - \mu > 0 \implies \frac{S}{N} > \frac{\gamma + \mu}{\beta}$$

which means that if the initial fraction of susceptibles $S(0)$ is less than $(\gamma+\mu)/\beta$ the infection dies out, where $(\gamma + \mu)/\beta$ is called the relative removal rate. Note that the inverse of the relative removal rate is the basic reproductive ratio

$$R_0 = \frac{\beta}{\gamma + \mu}.$$

This result was first shown by Kermack and McKendrick in 1927 and is famously referred as the *threshold phenomenon* [KMW27].

2.1.5 Equilibrium analysis

A useful way to study long term behavior of a disease is to explore the system when it is at equilibrium, meaning $\frac{dS}{dt} = \frac{dI}{dt} = \frac{dR}{dt} = 0$. Since $N = S+I+R$, we can re-write $R = N-S-I$ and get

$$\begin{aligned}
\frac{dS}{dt} &= \mu N - \beta\frac{SI}{N} - \mu S + \omega(N - S - I),\\
\frac{dI}{dt} &= \beta\frac{SI}{N} - \gamma I - \mu I,\\
\frac{dR}{dt} &= \gamma I - \mu R - \omega R.
\end{aligned} \tag{2.4}$$

We divide the ODE system in (2.4) by N to obtain the fractions of the population being susceptible, infected and recovered, i.e., $s = \frac{S}{N}$,$i = \frac{I}{N}$, and $r = \frac{R}{N}$, with the sum over all fractions adding up to one then the ODE system becomes

$$\frac{ds}{dt} = \mu - \beta si - \mu s + \omega(1 - s - i), \tag{2.5}$$

$$\frac{di}{dt} = \beta si - \gamma i - \mu i, \tag{2.6}$$

$$\frac{dr}{dt} = \gamma i - \mu r - \omega r. \tag{2.7}$$

Note that (2.5) and (2.6) depend only on s and i and thus it is enough to analyze the system using (2.5) and (2.6). Setting (2.5) and (2.6) to 0 and solving (2.6) we get

$$\begin{aligned}
\frac{di^*}{dt} = \beta i^* s^* - \gamma i^* - \mu i^* &= 0\\
i^*(\beta s^* - \gamma - \mu) &= 0\\
\implies i^* = 0 \ \text{ or } \ s^* &= \frac{\gamma + \mu}{\beta}.
\end{aligned}$$

The two possible equilibrium points are $i^* = 0$ or $s^* = \frac{\gamma+\mu}{\beta}$.

Starting with $i^* = 0$ and substituting into (2.5) we get $s^* = 1$. Thus, the equilibrium point is $(s^*, i^*) = (1, 0)$, and it is known as the *disease free equilibrium*, i.e., the pathogen has suffered extinction and in the long term everyone in the population is susceptible.

Now, consider $s^* = \frac{\gamma+\mu}{\beta}$. Substituting into (2.5) we get

$$\begin{aligned}
-\beta\frac{\gamma + \mu}{\beta}i^* - (\mu + \omega)\frac{\gamma + \mu}{\beta} + \mu + \omega - \omega i^* &= 0\\
-(\gamma + \mu)i^* - \frac{(\mu + \omega)(\gamma + \mu)}{\beta} + \mu + \omega - \omega i^* &= 0
\end{aligned}$$

$$
\begin{aligned}
-(\gamma+\mu+\omega)i^* &= \frac{(\mu+\omega)(\gamma+\mu)}{\beta} - (\mu+\omega) \\
i^* &= \frac{\beta(\mu+\omega) - (\mu+\omega)(\gamma+\mu)}{\beta(\gamma+\mu+\omega)} \\
&= \frac{(\mu+\omega)(\beta-\gamma-\mu)}{\beta(\gamma+\mu+\omega)}.
\end{aligned}
$$

Therefore the equilibrium point is given by

$$
(s^*, i^*) = \left(\frac{\gamma+\mu}{\beta}, \frac{(\mu+\gamma)(\beta-\gamma-\mu)}{\beta(\gamma+\mu+\omega)} \right),
$$

which is called the *endemic equilibrium*, i.e., the disease persists in the population in the long term. Note that the variables s, i, and r cannot be negative since they represent a proportion of the population, and thus i^* is biologically feasible when $R_0 > 1$, which agrees with the *threshold phenomenon*.

With equilibrium states comes the question of *how* did the dynamical systems approach stability. Some trajectories can have oscillatory behavior while others tend to reach the steady state more smoothly. The SIRS model is of the former variety, i.e., the inherent dynamics contain a strong oscillatory behavior. In fact, it has damped oscillations, and thus the amplitude of the fluctuations declines over time as the system reaches the equilibrium.

To explore the equilibrium dynamics we look at the rates of change of the variables (S, I, R) and determine what happens when each one is slightly shifted away from the equilibrium. Let ε be a small perturbation, then by making the substitutions $s = s^* + \varepsilon$, $i = i^* + \varepsilon$, and $r = r^* + \varepsilon$ we can explore the growth or decline of the perturbation term over time. A straightforward way to do this analysis is by looking at the eigenvalues of the Jacobian matrix, J, of the ODE system evaluated at the equilibrium point. For a system of n ODEs, there will be n eigenvalues and stability is ensured if the real part of all eigenvalues is less than zero [Rou19].

Again we can use (2.5) and (2.6) to find the Jacobian matrix

$$
J(s, i) = \begin{pmatrix} -\beta i - \mu - \omega & -\beta s - \omega \\ \beta i & \beta s - \gamma - \mu \end{pmatrix}.
$$

Starting with the disease free equilibrium point $(s^*, i^*) = (1, 0)$, we evaluate the Jacobian and get

$$
J(s^*, i^*) = \begin{pmatrix} -\mu - \omega & -\beta - \omega \\ 0 & \beta - \omega - \mu \end{pmatrix}.
$$

To find the eigenvalues of $J(s^*, i^*)$ we evaluate

$$\det(J - \lambda \mathbf{I}) = (-\mu - \omega - \lambda)(\beta - \gamma - \mu - \lambda) = 0.$$

Therefore it has eigenvalues $\lambda_1 = -\mu - \omega < 0$ and $\lambda_2 = \beta - \gamma - \mu$. If $R_0 < 1$ then $\lambda_2 < 0$ so the disease free equilibrium point is stable. If $R_0 > 1$ then $\lambda_2 > 0$ and thus the disease free equilibrium point is unstable.

For the endemic equilibrium point, the Jacobian matrix is given by

$$J(s^*, i^*) = \begin{pmatrix} -\dfrac{(\mu+\omega)(\beta-\gamma-\mu)}{(\gamma+\mu+\omega)} - \mu - \omega & -\gamma - \mu - \omega \\ \dfrac{(\mu+\omega)(\beta-\gamma-\mu)}{(\gamma+\mu+\omega)} & 0 \end{pmatrix}.$$

The characteristic polynomial is given by

$$\det(J - \lambda \mathbf{I}) = \lambda^2 + \left(\frac{(\mu+\omega)(\beta-\gamma-\mu)}{(\gamma+\mu+\omega)} + \mu + \omega \right)\lambda + (\mu+\omega)(\beta-\gamma-\mu) = 0.$$

Using the quadratic formula we get

$$\lambda = -\frac{1}{2}\left(\frac{(\mu+\omega)(\beta-\gamma-\mu)}{(\gamma+\mu+\omega)} + \mu + \omega \right) \pm \frac{1}{2}\sqrt{\left(\frac{(\mu+\omega)(\beta-\gamma-\mu)}{(\gamma+\mu+\omega)} + \mu + \omega \right)^2 - 4(\mu+\omega)(\beta-\gamma-\mu)}.$$

We can re-write the equations by making the substitution $R_0 = \frac{\beta}{\gamma+\mu}$ to get

$$\lambda = -\frac{1}{2}\left(\frac{(\mu+\omega)(\gamma+\mu)(R_0-1)}{(\gamma+\mu+\omega)} + \mu + \omega \right) \pm \frac{1}{2}\sqrt{\left(\frac{(\mu+\omega)(\gamma+\mu)(R_0-1)}{(\gamma+\mu+\omega)} + \mu + \omega \right)^2 - 4(\mu+\omega)(\gamma+\mu)(R_0-1)}.$$

Let $A = \frac{\gamma+\mu+\omega}{(\mu+\omega)(\beta-\gamma-\mu)}$ be the average age at first infection, $G_I = \frac{1}{\gamma+\mu}$ the average period spent in the infected class, and $G_R = \frac{1}{\mu+\omega}$ the average period spent in the recovered class. We can write

$$\lambda = -\frac{1}{2}\left(\frac{(\mu+\omega)(\gamma+\mu)(R_0-1)}{(\gamma+\mu+\omega)} + \frac{1}{G_R} \right) \pm \frac{1}{2}\sqrt{\left(\frac{1}{A} + \frac{1}{G_R} \right)^2 - 4\frac{1}{G_I}\frac{1}{G_R}(R_0-1)}.$$

Since the endemic equilibrium is only feasible for $R_0 > 1$ and the term $\left(\frac{1}{A} + \frac{1}{G_R}\right)^2$ is sufficiently small such that the discriminant is negative, we will have complex eigenvalues with negative real part making the system locally asymptotically stable. The eigenvalues are the complex conjugates

$$\lambda = -\frac{1}{2}\left(\frac{(\mu+\omega)(\gamma+\mu)(R_0-1)}{(\gamma+\mu+\omega)} + \frac{1}{G_R}\right) \pm \frac{i}{2}\sqrt{4\frac{1}{G_I}\frac{1}{G_R}(R_0-1) - \left(\frac{1}{A}+\frac{1}{G_R}\right)^2},$$

therefore the equilibrium is approached via oscillatory dynamics. The period of the damped oscillations is given by the inverse of the complex part of the eigenvalues multiplied by 2π:

$$T \sim \frac{4\pi}{\sqrt{4\frac{1}{G_I}\frac{1}{G_R}(R_0-1) - \left(\frac{1}{A}+\frac{1}{G_R}\right)^2}}.$$

As we just saw, the SIRS model (with constant parameters) has damped oscillatory dynamics with stable endemicity in the long term. In order to have sustained oscillations and recurrent epidemics, there has to be a forcing mechanism of which there are two forms: seasonal forcing and stochastic forcing, both of which are explained below.

2.1.6 Seasonality

The term seasonality is used to describe the variations in the prevalence of an infectious disease that happen at relatively regular intervals throughout the year. We can better capture and predict the observed patterns of recurrent and seasonal epidemics by adding a seasonally varying parameter that acts as a forcing mechanism [Fis07b]. For example, analysis of measles data showed large amplitude recurrent epidemics with very strong peaks and valleys, a dramatic shift from the simple SIRS model dynamics seen above. In fact, this pattern has been seen for multiple infectious diseases such as chickenpox, mumps, influenza, RSV, and cholera [BFG23, FC82, LY73, YL73, DPLE04, WWC+05, PRE+00, KRP+05]. Seasonality arises from a variety of factors depending on the host and pathogen that is being studied. For instance, measles is commonly associated with increased transmission during school terms [LY73], cholera and rotavirus transmission are thought to increase during monsoon seasons [PRE+00, KRP+05, MKY+16], and the seasonality of influenza and RSV

is linked with the winter season (low humidity and temperature) [SPV+10, LS14, SK09, MKY+16, YCLS15, YLS15, SK12, BMW+19, WWM+98].

The addition of a seasonal forcing term to the simple SIRS model allows for more complex and realistic models. Seasonality has been explored by adding a time dependent function to the transmission rate, instead of having a constant transmission rate. Some common functional forms used are sinusoidal functions, step functions, and splines [PVA+15, LY73, MKY+16]. The choice of functional form used to represent seasonality in the transmission term can cause very different transmission dynamics [KR08]. There is no straightforward way of choosing the forcing term, therefore exploring multiple models with different functional forms and their relative goodness of fit is recommended.

2.1.7 Stochasticity

Another approach to better capture the patterns and features of infectious diseases is using a stochastic framework. So far, we have seen deterministic models where if given the same starting conditions, the same trajectory and cycles will always be observed. This type of model does not account for the real world dynamics of infectious diseases, in which we can observe very different epidemic cycles, particularly for small populations, chance extinctions between epidemics, and early stages of an outbreak when stochastic effects may dominate. Stochastic models allow us to approximate the randomness or probabilistic element of infectious diseases [KR08]. There are two common sources of stochasticity: measurement noise and process noise.

In measurement noise, also known as observational noise, there is an assumed uncertainty in the recorded data. This uncertainty can come from incomplete reporting of true cases, misdiagnosis, and over-reporting [FC82, FG00]. This type of observational error can impact the amplitude of data (interannual variation), but not the periodicity. An important distinction to make is that measurement noise does not impact the epidemiological dynamics and only modifies the reported data [KR08]. Measurement noise is included in the way in which we represent the number of infected individuals and can be done by using distributions (normal, Poisson, binomial, exponential, negative binomial) depending on the type of observational error in the data [KR08, LES03, CRP04, Sto19, SBH20, BHIK09, NMZC16, Bre18].

While measurement noise is included through the data, process noise, also known as environmental noise, is introduced directly into the deterministic equations of the model [BFG23]. This means that the dynamics are subject to some random variability and this variability is propagated forward in time by the underlying equations [KR08]. Adding noise to the model means that the dynamics will deviate from deterministic equilibrium. It is important to ac-

knowledge that the deviations from equilibrium are not completely random, even though the noise term allows trajectories to deviate from the equilibrium, the deterministic component forces them back towards the equilibrium point [KR08].

2.2 Inference for partially observed Markov processes via iterated filtering

Our main goal is to estimate (infer) the parameters that govern the transmission dynamics of respiratory viruses in Chile, and assess how climate influences these dynamics, by testing different hypotheses. To do so, we will use surveillance data from Chile and model it as a partially observed Markov process (POMP). A POMP, also called Hidden Markov Model or state-space model, consists of an unobserved Markov state process, connected to the data via an explicit model of the measurement process. The structure of a POMP model can be seen seen in Figure 2.3 [KDI16].

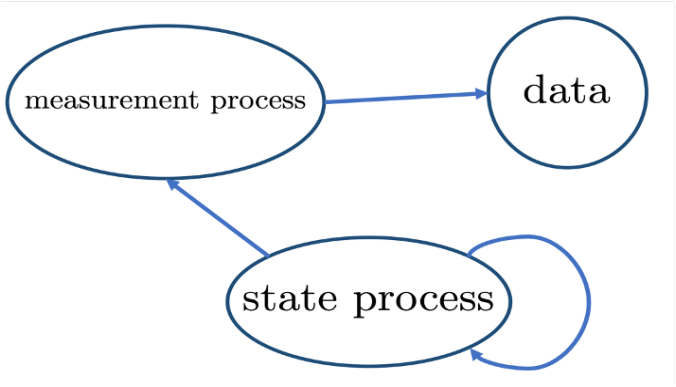

Figure 2.3: POMP structure diagram adapted form the materials and tutorials in [KDI16].

The method of modeling infectious disease time series data as a noisy and partially observed realization of the disease transmission process that is assumed to be Markov was first introduced by King *et al.* [KDI16]. The following sections are adapted from tutorials developed by Aaron A. King [KIMB+22], notes from a short simulation based inference course developed by Aaron A. King and Edward L. Ionides found in [KI22] and Chapter 11 of the book "Handbook of Infectious Disease Data Analysis" written by Theresa Stocks [Sto19]. We encourage the reader to refer to these references for a more comprehensive and extensive explanation of the topics [KIMB+22, KI22, HHOW19b].

In Section 2.2.1 we describe in detail what a POMP is and how to formulate the SIR compartmental model as a Markov transmission model. Section 2.2.2 outlines the important

aspects of likelihood-based inference, some challenges of existing methods. Particle filters and iterated filtering iterated filtering to estimate the likelihood of a POMP are discussed in Sections 2.2.3 and 2.2.4, respectively.

2.2.1 Partially observed Markov process

Mathematically, each model is a probability distribution. Let Y_n be a random variable modeling the observation at time t_n, which denotes the measurement process (or observation process), and X_n the latent state process (or transmission process) where $X_{0:N} = (X_0, \ldots, X_N)$. By definition, the state process model is determined by the density $f_{X_n|X_{n-1}}$ and the initial density f_{X_0}. The measurement process is determined by the density $f_{Y_n|X_n}$. These two submodels determine the full POMP model. If we have a sequence of measurements, y_n^*, made at times t_n, $n \in [1, N]$, then we think of these data, collectively, as a single realization of the Y process. A POMP is characterized by two conditions:

1. the state process, X_n is Markov, i.e.,

$$f_{X_n|X_{n-1},Y_{1:n-1}}(x_n|x_{0:n-1}, y_{1:n-1}) = f_{X_n|X_{n-1}}(x_n|x_{n-1}),$$

   and

2. the measurements, Y_n are conditionally independent and depend only on the state at that time, i.e.,

$$f_{Y_n|X_{0:N},Y_{1:n-1}}(y_n|x_{0:N}, y_{1:n-1}) = f_{Y_n|X_n}(y_n|x_n),$$

   for all t_n, where $n = 1, \cdots, N$.

In other words, the latent state X_n at time t_n is conditionally independent of its history given X_{n-1} and the observation Y_n is conditionally independent of all other variables given X_n. In Figure 2.4 we can see a graph of the conditional independence of POMP models. In general, knowledge of the system's state at any point in time is sufficient to determine the distribution of possible futures.

In the case of SIR models, the state process describes the dynamics of the disease spread, and $X_n = (S(t_n), I(t_n), R(t_n))$ counts the number of susceptible, infected, and recovered individuals at time t_n. The surveillance data is then modeled as a realization of the measurement process, Y_n.

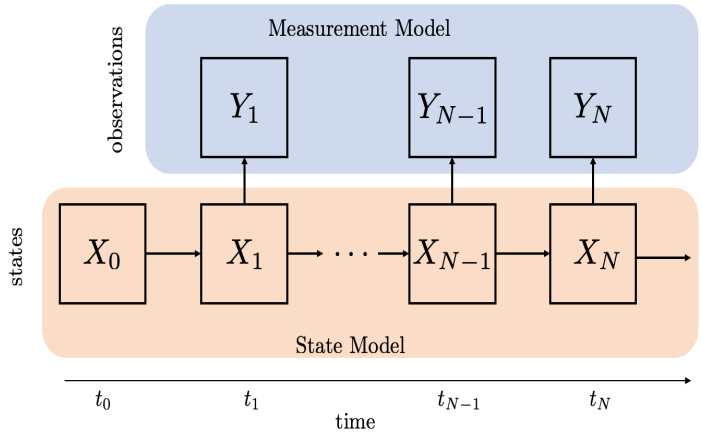

Figure 2.4: POMP conditional independence graph [KDI16]. Here Y_n denotes the observations at time t_n for $n \in [1, N]$, which depend on the state transmission process X_n at that time, and X_n is conditionally independent of its history (i.e., Markov).

2.2.2 Likelihood of a partially observed Markov process

Likelihood-based inference attempts to draw conclusions about a parameter vector by connecting the model to the observations. The idea is to find the parameter values for which the observations are most likely to occur under a chosen probability model. In the case of POMP models, we have a model for the data determined by the measurement density $f_{Y_n|X_n}(y_n|x_{n-1};\theta)$, together with the transition density $f_{X_n|X_{n-1}}(x_n|x_{n-1};\theta)$ and the the initial density $f_{X_0}(x_0;\theta)$, all parameterized by the vector θ, and we want to maximize the joint density function

$$f_{X_{0:N},Y_{1:N}}(x_{0:N}, y_{1:N};\theta) = f_{X_0}(x_0;\theta) \times \prod_{n=1}^{N} f_{X_n|X_{n-1}}(x_n|x_{n-1};\theta) f_{Y_n|X_n}(y_n|x_n;\theta)$$

with respect to θ evaluated at the data $y^*_{1:N}$. Fundamentally, we want to optimize the likelihood function

$$\mathcal{L}(\theta) = f_{Y_{1:N}}(y^*_{1:N};\theta) = \int f_{X_{0:N},Y_{1:N}}(x_{0:N}, y^*_{1:N};\theta) dx_{0:N}. \tag{2.8}$$

The parameter vector which maximizes this function is called the maximum likelihood estimate (MLE) and is given by

$$\hat{\theta} = \arg_{\theta\in\Theta} \max \mathcal{L}(\theta)$$

where Θ is the parameter space containing all possible sets of parameters.

Finding the likelihood can be very challenging. In the case of (2.8), the integral depends

on the number of compartments in the state process (in the case of the SIR models we will implement we have 3 or more compartments), and thus can be high dimensional and cannot be solved analytically. Therefore we will be using the methods described in Sections 2.2.3 and 2.2.4.

2.2.3 Particle filter

A particle filter, also known as a sequential Monte Carlo algorithm, is a well known method used to evaluate the likelihood when there are high dimensional integrals. More about sequential Monte Carlo methods can be found in [AdFG01, AMGC02, ETvdM21, DJ11, NKd+23]. We will give a quick overview of the particle filter method, which is necessary to understand the algorithm described in Section 2.2.4. The general idea is that the likelihood in (2.8) can be factored as

$$\begin{aligned}\mathcal{L}(\theta) &= f_{Y_{1:N}}(y^*_{1:N};\theta)\\ &= \prod_{n=1}^{N} f_{Y_n|Y_{1:n-1}}(y^*_n|y^*_{1:n-1};\theta)\\ &= \prod_{n=1}^{N} \int f_{X_n|Y_n}(y^*_n|x_n;\theta) f_{X_n|Y_{1:n-1}}(x_n|y^*_{1:n-1};\theta)dx_n,\end{aligned}$$

where $f_{X_1|Y_{1:0}} = f_{X_1}$. From the Markov property we have

$$f_{X_{1:N}} = f_{X_1}(X_1) f_{X_2|X_1}(X_2|X_1)\dots f_{X_N|X_{N-1}}(X_n|X_{n-1}).$$

So it follows from the Chapman-Kolmogorov equation that

$$f_{X_n|Y_{1:n-1}}(x_n|y^*_{1:n-1};\theta) = \int f_{X_n|X_{n-1}}(x_n|x_{n-1};\theta) f_{X_{n-1}|Y_{1:n-1}}(x_{n-1}|y^*_{1:n-1};\theta)dx_{n-1},$$

and this is the prediction formula at each time step t_n. By applying Bayes' theorem

$$\begin{aligned}f_{X_n|Y_{1:n}}(x_n|y^*_{1:n};\theta) &= f_{X_n|Y_n,Y_{1:n-1}}(x_n|y^*_n, y^*_{1:n-1};\theta)\\ &= \frac{f_{Y_n|X_n}(y^*_n|x_n;\theta) f_{X_n|Y_{1:n-1}}(x_n|y^*_{1:n-1};\theta)}{\int f_{Y_n|X_n}(y^*_n|x_n;\theta) f_{X_n|Y_{1:n-1}}(x_n|y^*_{1:n-1};\theta)},\end{aligned}$$

we obtain the filtering formula at each time step t_n. Here $f_{X_n|Y_{1:n-1}}(x_n|y^*_{1:n-1})$ is called the prediction distribution and $f_{X_n|Y_{1:n}}(x_n|y^*_{1:n})$ is called the filtering distribution. Note that the prediction formula gives the prediction distribution at time t_n using the filtering distribution

at time t_{n-1}. Subsequently the filtering formula gives the prediction distribution at time t_n using the prediction distribution at time t_n.

More precisely, assume that we have a set $\{X^F_{n-1,j}\}$, , $j = 1, \ldots, J$, where J is the number of points (particles) drawn from the filtering distribution at t_{n-1}. Drawing a sample from a distribution means observing a realization of the random variable which has assigned that distribution to possible outcomes. The prediction formula implies that we obtain a sample $X^P_{n,j}$ of points from the the prediction distribution at time t_n if we simulate from the state process

$$X^P_{n,j} \sim \text{process}(X^F_{n-1,j}, \theta),$$

for $j = 1, \ldots, J$. The filtering formula tells us that resampling from $X^P_{n,j}$ with weights proportional to $w_{n,j} = f_{Y_n|X_n}(y^*_n|X^P_{n,1}; \theta)$ gives us a sample from the filtering distribution at time t_n. Using the Monte Carlo principle it follows that the likelihood in Equation 2.8 can be approximated as

$$\mathcal{L}(\theta) \approx \hat{\mathcal{L}}(\theta) = \frac{1}{J} \sum_{j=1}^{J} f_{Y_n|X_n}(y^*_n|X^P_{n,j}; \theta),$$

since $X^P_{n,j}$ is approximately a sample drawn form $f_{X_n|Y_{1:n-1}}(x_n|y^*_{1:n-1}; \theta)$. A graphical representation of this process can be seen in Figure 2.5. Iterating through the data, for each time step t_n $(n = 0, \ldots, N)$, and doing the recursion of simulating and resampling, as seen in Algorithm 1, gives us the log-likelihood

$$\ell(\theta) = \log(\mathcal{L}(\theta)) = \sum_n \log(\mathcal{L}_n(\theta)) \approx \sum_n \log\Big(\hat{\mathcal{L}}(\theta)\Big).$$

While the particle filter seen in Algorithm 1 can be used to evaluate the likelihood of a POMP model, the method can fail when a very unlikely particle is suggested given the model. If the conditional likelihood of this particle is below a certain tolerance value then the particle is considered to be uninformative. If this situation happens for a lot of the particles, the particle filter is said to suffer from *particle depletion* [KDI16]. There can also be variability in the approximation of the log-likelihood, which can be reduced by increasing the number of particles, but in most cases will remain. This might create problems when using standard optimizers to find the MLE, which assume that the likelihood is evaluated deterministically [Sto19]. A workaround is to use stochastic optimizers such as the iterated filtering method described in Section 2.2.4.

Algorithm 1 Particle filter pseudocode obtained from [KDI16]

Input: Simulator for $f_{X_n|X_{n-1}}(x_n|x_{n-1};\theta)$, evaluator for $f_{Y_n|X_n}(y_n|x_n;\theta)$, simulator for $f_{X_0}(x_0;\theta)$, parameter vector: θ, observation data: $y^*_{1:N}$, number of particles: J.

1: Initialize $X^F_{n-1,j}, j = 1,\ldots,J$, where J is the number of points (particles) drawn from the filtering distribution at t_{n-1}

2: **for** $n = 1,\ldots,N$ **do**

3: Simulate the state process model to obtain a sample $X^P_{n,j}$ of points from the the prediction distribution at time t_n

$$X^P_{n,j} \sim \text{process}(X^F_{n-1,j},\theta), j = 1,\ldots,J$$

4: Obtain a sample of points from the filtering distribution at time t_n, by resampling from $\{X^P_{n,j}, j \in 1:J\}$ with weights

$$w_{n,j} = f_{Y_n|X_n}(y^*_n|X^P_{n,1};\theta)$$

5: Using the Monte Carlo principle the likelihood in Equation 2.8 can be approximated as

$$\mathcal{L}(\theta) \approx \hat{\mathcal{L}}(\theta) = \frac{1}{J}\sum_{j=1}^{J} f_{Y_n|X_n}(y^*_n|X^P_{n,j};\theta),$$

since $X^P_{n,j}$ is approximately a draw form $f_{X_n|Y_{1:n-1}}(x_n|y^*_{1:n-1};\theta)$

6: **end for**

Output: log-likelihood $\ell(\theta) = \log(\mathcal{L}(\theta)) = \sum_n \log(\mathcal{L}_n(\theta)) \approx \sum_n \log\Big(\hat{\mathcal{L}}(\theta)\Big)$

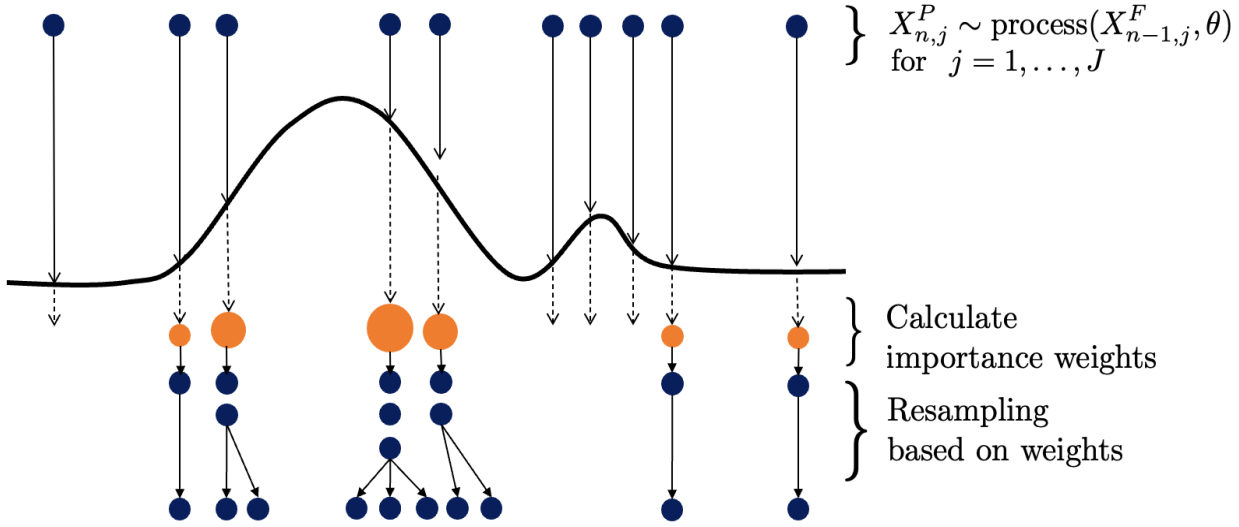

Figure 2.5: Particle filter example adapted from [AdFG01].

2.2.4 Finding the likelihood via iterated filtering

To find the MLE of POMP models Ionides *et al.* developed an iterated filtering method where samples are drawn from the density $f_{X_n|X_{n-1}}$ instead of evaluating it [IBK06, INA+15]. The algorithm applies a particle filter, as seen in Section 2.2.3, to a model in which the parameter vector for each particle is following a random walk in time. After multiple repetitions of the filtering procedure, the intensity of the perturbations (random walk) is reduced. The random walk approaches zero and the modified model approaches the original one. It may seem counterproductive to add more variability to the model, but in fact it combats the particle depletion mentioned in Section 2.8, it smooths the likelihood surface, which facilitates optimization, and it can be shown that the algorithm converges towards the MLE [INA+15].

In the algorithm the initial parameter swarm is perturbed by a perturbation density where its standard deviation σ_m is a decreasing function of m. The state process is initialized as a draw form the initial density f_{X_0}, which is dependent on the newly perturbed parameter vector. We can then apply a particle filter as described in Algorithm 1, where now the parameter vector is stochastically perturbed in every iteration through the data. The particle filter is repeated M times, with decreasing perturbations. The algorithm returns the best guess of the parameter swarm after M iterations.

In recent years iterated filtering has been used to solve likelihood-based inference problems for infectious diseases [MKY+16, IBK06, KIPB08, INA+15, DKWP17, FK20, SBH20, YCLS15]. This has been facilitated by the R package `pomp` [KDI16, KIMB+22]. There are multiple examples and tutorials available online on how to use the `pomp` package, some of

Algorithm 2 Iterated filtering pseudocode obtained from [KDI16]

Input: Simulator for $f_{X_n|X_{n-1}}(x_n|x_{n-1};\theta)$, evaluator for $f_{Y_n|X_n}(y_n|x_n;\theta)$, simulator for $f_{X_0}(x_0;\theta)$, initial parameter swarm: $\{\theta_j^0, j = 1,\dots,J\}$, observation data: $y^*_{1:N}$, number of iterations: M, number of particles: J, perturbation density $h_n(\theta|\phi;\sigma)$, perturbation scale $\sigma_{1:M}$.

for $m = 1,\dots,M$ **do**
  $\theta^{F,m}_{0,j} \sim h_0(\cdot|\theta_j^{m-1};\sigma_m)$ ▷ Perturb initial parameter values
  $X^{F,m}_{0,j} \sim f_{X_0}(\cdot;\theta^{F,m}_{n-1,j})$ for j in $1:J$ ▷ Initialize the state process by drawing from initial density
  for $n = 1,\dots,N$ **do** ▷ Particle filter
    $\theta^{P,m}_{n,j} \sim h_n(\cdot|\theta^{F,m}_{n-1,j},\sigma_m)$ for j in $1:J$
    $X^{P,m}_{n,j} \sim f_{X_n|X_{n-1}}(\cdot|X^{F,m}_{n-1,j};\theta^{P,m}_{n,j})$ for j in $1:J$
    $w^m_{n,j} = f_{Y_n|X_n}(y^*_n|X^{P,m}_{n,j};\theta^{P,m}_{n,j})$ for j in $1:J$
    Draw $k_{1:J}$ with $P(k_j = i) = \dfrac{w^m_{n,i}}{\sum_{u=1}^J w^m_{n,u}}$
    $\theta^{F,m}_{n,j} = \theta^{P,m}_{n,k_j}$ and $X^{F,m}_{n,j} = X^{P,m}_{n,k_j}$ for j in $1:J$
  end for
  Set $\theta^m_j = \theta^{F,m}_{N,j}$
end for

Output: Final parameter swarm, $\{\theta^M_j, j = 1,\dots,J\}$

them can be found in [KI22, KIMB+22]. This package provides the tools needed to perform inferences for nonlinear partially-observed Markov processes as described in Algorithm 2. The user can implement a POMP model by specifying its hidden process and measurement components. After all components of the model are defined iterated filtering can then be performed to estimate the parameters.

We will use iterated filtering methods to perform parameter inference, evaluate the effect of climate covariates, and get a better understanding of disease spread mechanisms of respiratory viruses. In our work, we consider SIRS-type models to describe the transmission dynamics of RSV and Influenza A in Chile. To build the POMP model, the first step is to represent the SIRS models as unobserved Markovian transmission processes, done in Section 4.2, and subsequently connect them to a measurement model, which relates the data to the transmission model. We show the results from our simulations in Section 4.4.

Chapter 3

Data analysis: The association of disease incidence and environmental drivers

3.1 RSV and Influenza A

Viral respiratory infections commonly affect the upper or lower respiratory tract. Some of the most common respiratory virus symptoms include coughing, fever, runny nose, and sore throat. Even with similar symptoms respiratory viruses are highly diverse not only in their viral structure and genome composition but also in their modes of transmission among humans. The severity of viral respiratory illness varies widely. Severe disease is more likely in older patients and infants [WF04, Gle86]. Morbidity may result directly from viral infection or may be indirect, because of pre-existing conditions exacerbated by the disease.

Respiratory viruses also display various transmission patterns among humans; via direct and indirect contact airborne transmission, comprising both air droplets and aerosol. Their transmissibility is influenced by the environment in which pathogen and host meet. Many common respiratory viruses exhibit predictable seasonality, which can be exacerbated by mobility patterns, socioeconomic status, as well as diseases interacting with each other (i.e., being infected with multiple viruses at the same time).

Influenza A and RSV are the leading pathogens responsible for substantial morbidity and mortality due to respiratory tract worldwide [DS, ZTV+12, LWB+22, JSH+07, HCDV22]. Influenza A produces the most severe illnesses in most age groups [Mon95, Sim99]. It is known for providing long-term immunity; however, it has a rapid antigenic evolution, requiring yearly vaccinations [Hsi10, YKH+13, DKWP17]. In contrast, RSV has limited immunity allowing for reinfection throughout a persons life, and for its propensity to produce severe illness in young children, making it less common in older individuals [APC18, MLC+22, CZ22, LWB+22]. Both viruses are known to be highly seasonal, specifically during the winter months in temperate climates and during rainy months in tropical and subtropical regions [BAC+13, WWM+98, SK09, van28, PB12, Mon04, NK22]. Their predictable seasonality allows us to examine data in a manner that helps us define environmental factors that may be influencing differences in seasonality and interannual variability.

It is difficult to count Influenza A and RSV cases because the symptoms associated with

infection are nonspecific and laboratory testing is not routine in most places. Available data can thus be incomplete because of underreporting (e.g. asymptomatic cases, individuals don't go to the doctor/hospital, misdiagnosis). Nevertheless a lot of work has been done to examine environmental factors and their role in disease spread. In the sections below we will describe the data used in this thesis and perform exploratory analysis to better understand patterns seen in the data in the context of seasonality and interannual variability.

3.2 Data

We use a region-level dataset of weekly RSV and Influenza A confirmed cases from Chile spanning 9 years. The spatial extent of our dataset covers a diverse set of climatologies. We combine our incidence data with high resolution climate data (specific humidity and temperature) in order to investigate spatial patterns in dynamics and evaluate the drivers of transmission.

3.2.1 Surveillance data

Surveillance data allows for the continuous examination and analysis of infectious disease incidence, with the goal of early detection of increase in incidence in time and space. Chile has a large surveillance system as part of the Institute of Public Health (ISP) within the Health Ministry of Chile's Government [isp]. Chile's objective is to identify the circulation of influenza-like illness, severe acute respiratory infections, monitoring of emergency room visits because of respiratory viruses, and characterize the antigenic variants. The system is composted of 42 primary care sentinel centers distributed among the 16 regions of the country. It also includes 31 public hospitals that integrate their laboratories for the detection of the respiratory viruses. The sentinel centers and public hospitals are both a part of the larger hospital network across all of Chile [isp]. Furthermore, the ISP processes samples from private centers and hospitals that are not part of the surveillance system.

The complete data set includes the number of weekly respiratory viruses (this includes Adenovirus, Parainfluenza, Influenza A, Influenza B, RSV, and Metapneumovirus) for each region in Chile, as seen in Figure B.2. In this thesis use data of weekly lab-confirmed cases for RSV and Influenza A from 2011 to 2019, as well as the number of detected respiratory viruses (i.e., all cases combined). Each sentinel and laboratory site is designated to one of the 16 regions of Chile, and aggregated to obtain regional weekly confirmed cases to obtain better spatial resolution (Figures 3.1 and 3.2).

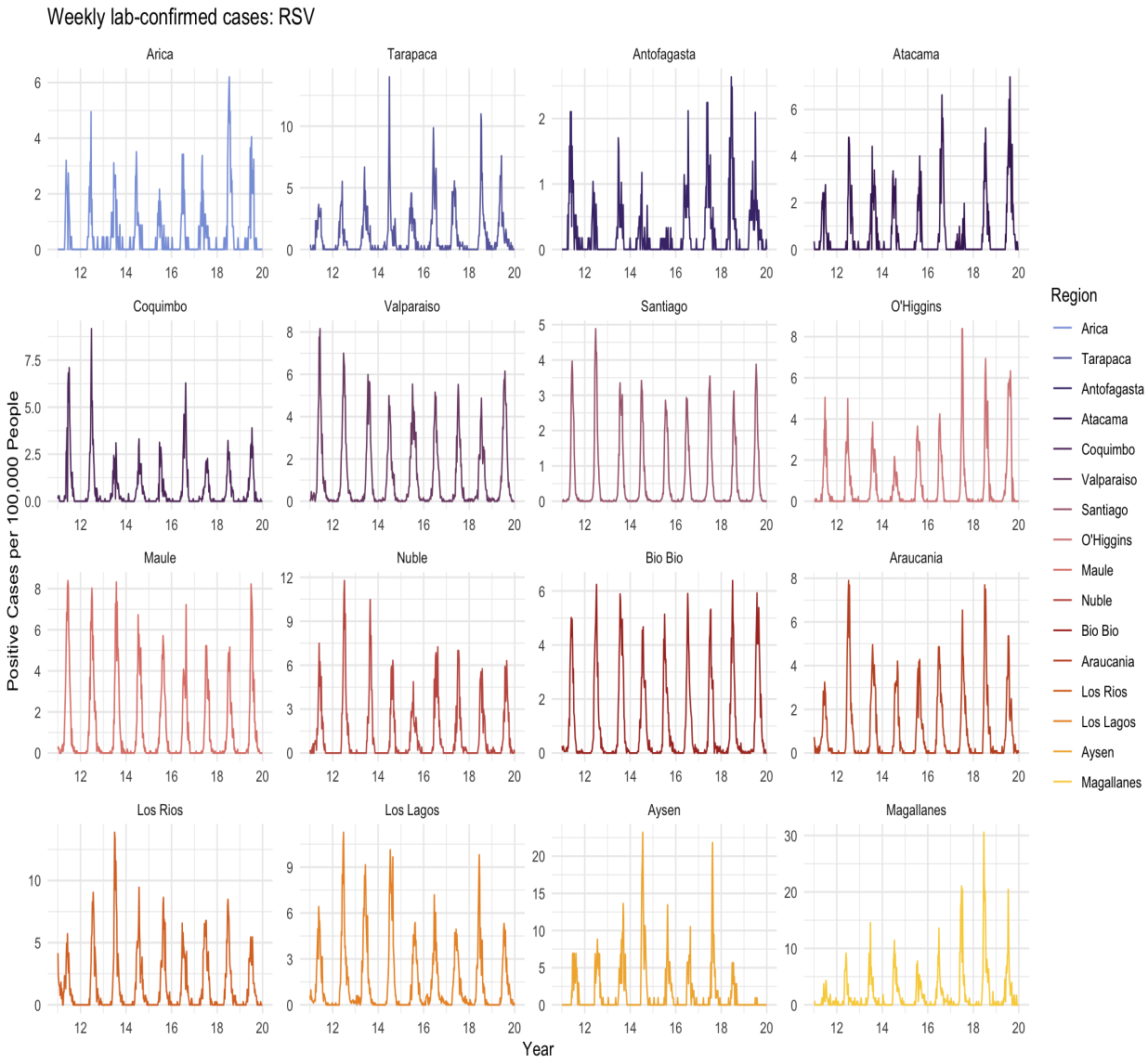

Figure 3.1: Weekly lab-confirmed RSV cases from 2011 to 2019 by region and normalized by population.

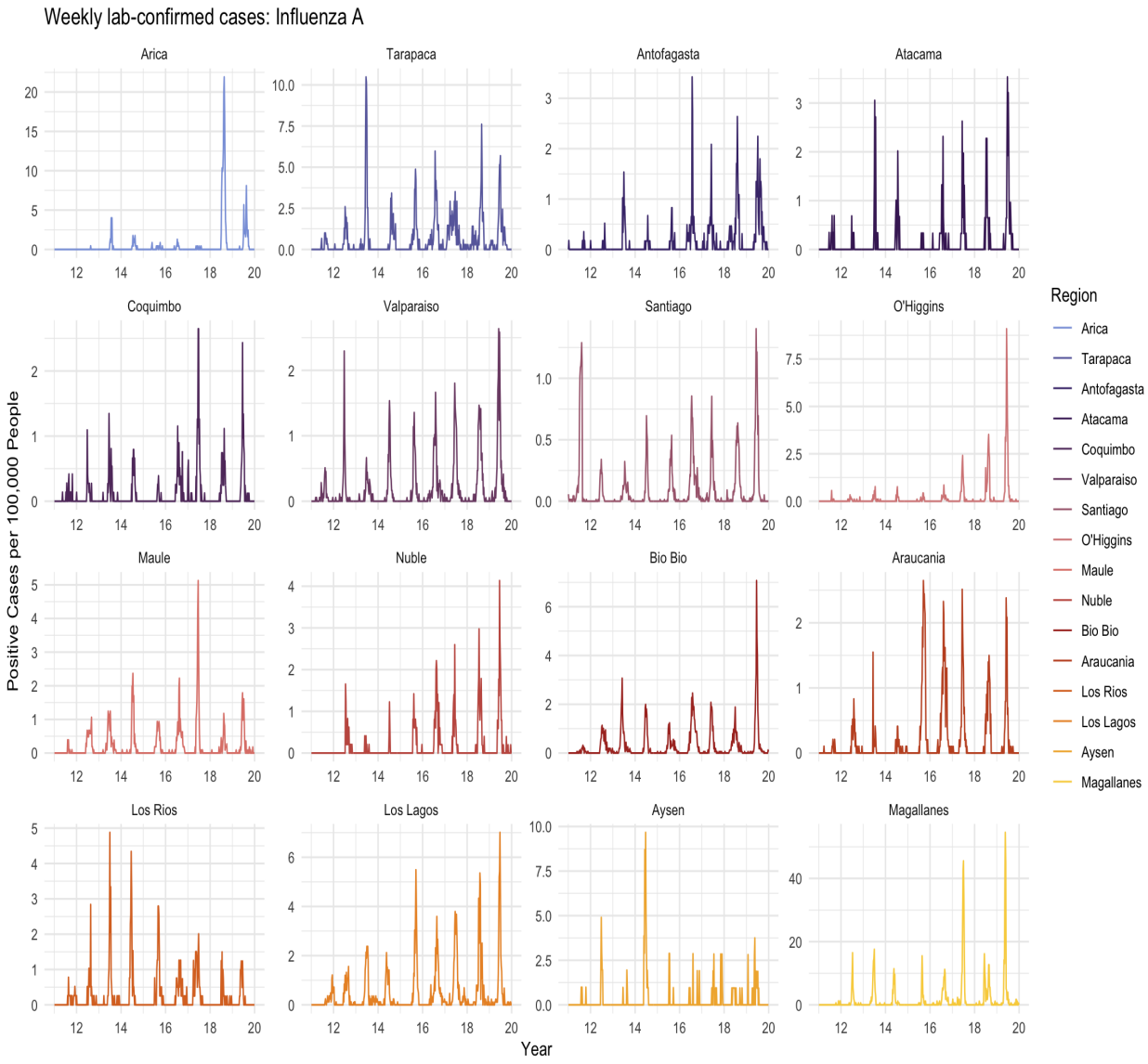

Figure 3.2: Weekly lab-confirmed Influenza A cases from 2011 to 2019 by region and normalized by population.

3.2.2 Climate data

We explore two climate covariates: temperature (°C) and specific humidity (g/kg). Specific humidity refers to the weight (amount) of water vapor contained in a unit weight (amount) of air (expressed as grams of water vapor per kilogram of air) [US]. Absolute humidity (expressed as grams of water vapor per cubic meter volume of air) is a measure of the actual amount of water vapor (moisture) in the air, regardless of the air's temperature [US]. Therefore specific humidity is considered a measure of absolute humidity, and the terms can be interchangeable.

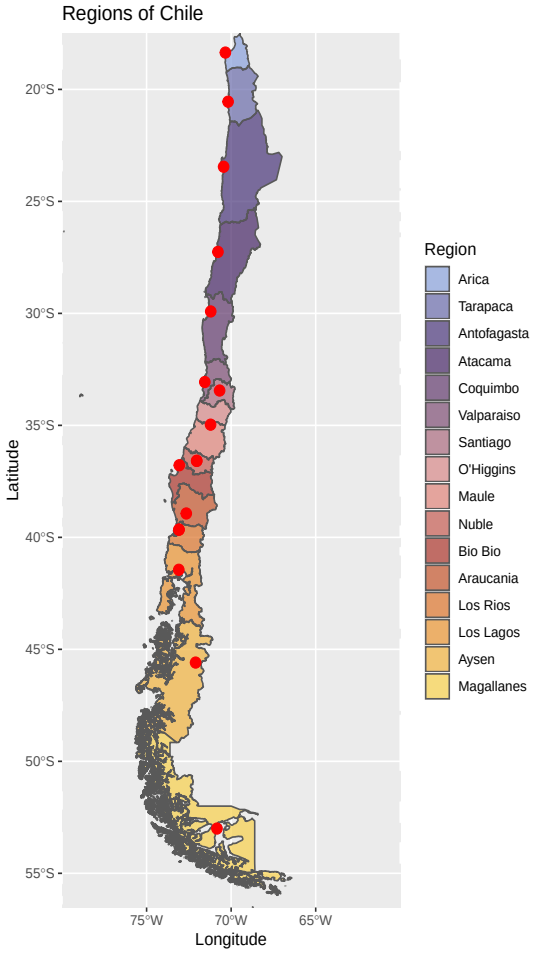

Figure 3.3: Regions of Chile with marked sentinel hospitals (red points) used for data aggregation.

Climate data is often very difficult to obtain at the precise spatial locations of interest. More specifically, we are working with regions and each region may have multiple meteorological sites with vastly different climates (even in the same region depending on elevation). To overcome this issue, meteorological data can be interpolated across multiple spatial lo-

cations. Our data is obtained from the ERA5 a global reanalysis dataset of gridded hourly weather data that combines weather models with satellite observations from across the world into a globally complete and consistent dataset [HBB+18]. The use of multiple data sets and spatial interpolation methods allows for higher accuracy and spatial and temporal resolution in the weather estimates [HBB+20]. We match the gridded climate data from ERA5 based on the latitude and longitude for each region (Figure 3.3). We averaged over the hourly observations to obtain weekly mean values for the temperature and specific humidity as seen in Figure 3.4.

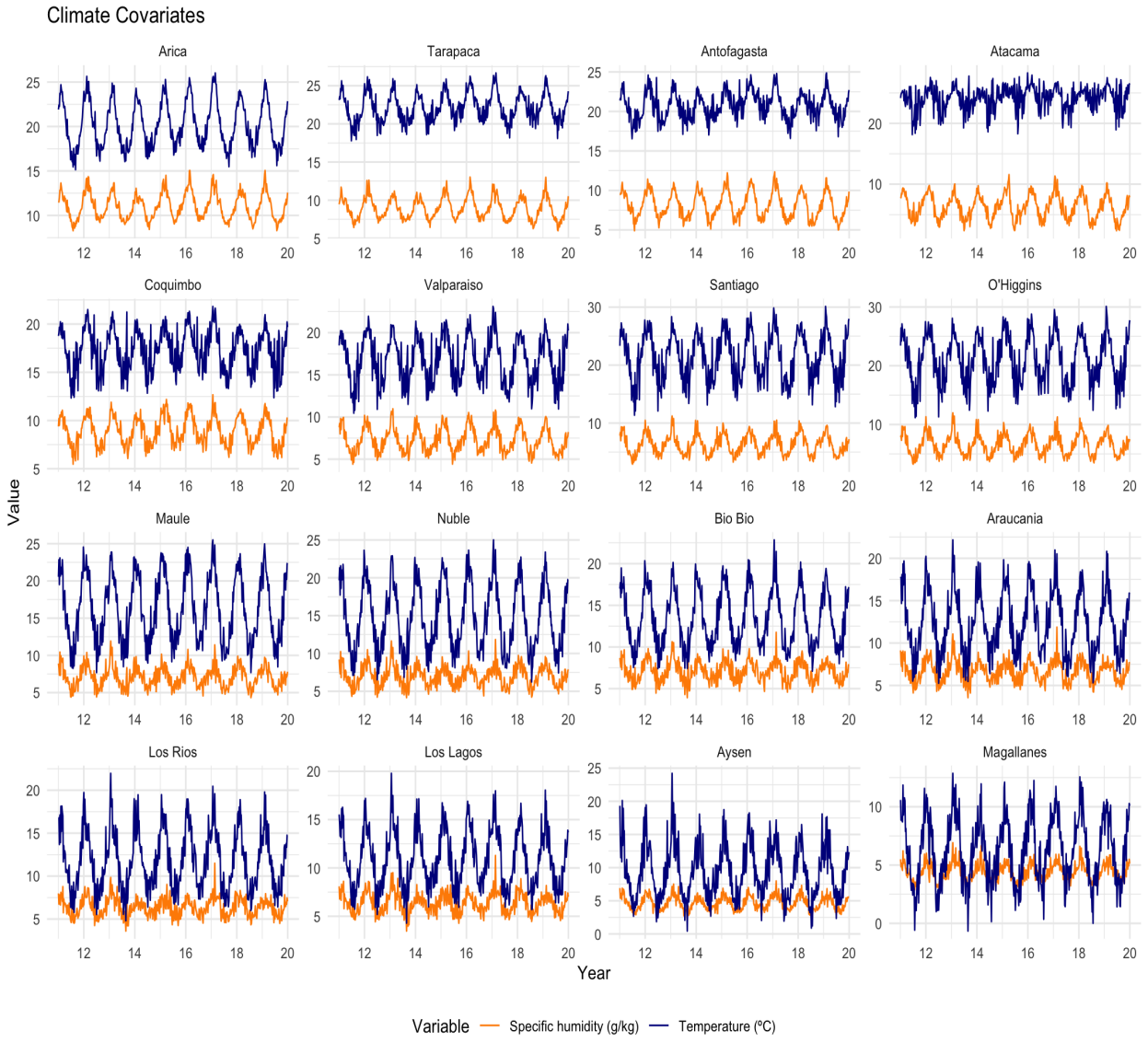

Figure 3.4: Climate covariates for each region.

3.2.3 Demographic data

The yearly population size for each region was obtained from the census international database from 2011 to 2021 [Bur]. We assumed the birth rate varied between regions and over time. We fit a cubic smoothing spline to the demographic data (for more on splines see Appendix C) [H., PSAS19, DM]. Using the fitted cubic spline we predicted the weekly population for each region from the yearly population size, as shown in Figure 3.5**A**. We use the data to normalize the cases by population for each region.

We also calculate the first derivative of the fitted cubic spline to predict the weekly birthrate as shown in Figure 3.5**B**. The models implemented in Section 4.2 consider a non-constant population size, and this metric is incorporated in the models to reproduce the observed population increase over time.

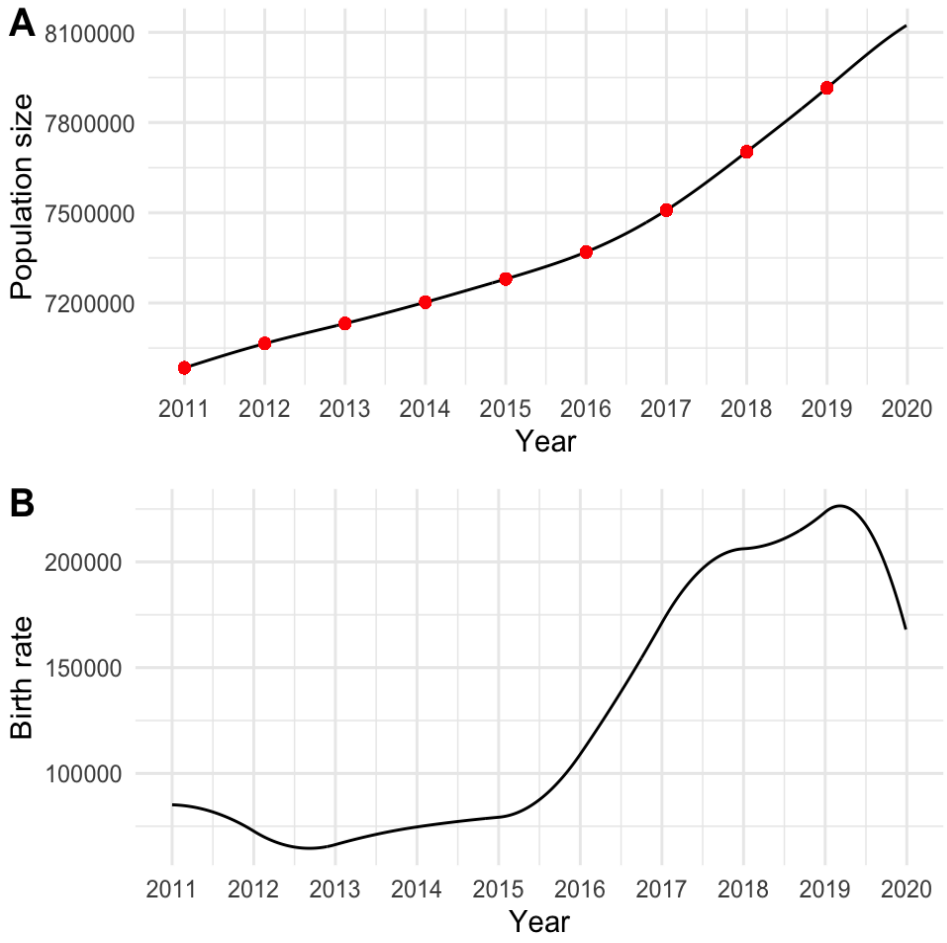

Figure 3.5: **A.** Predicted weekly Santiago population using a smooth cubic spline. **B.** Predicted weekly Santiago birth rate using a smooth cubic spline.

3.3 Impact of environmental drivers on spatiotemporal patterns of disease incidence

In this chapter we use statistical methods to explore, describe, visualize, and interpret the data. The methods used below allow us to answer questions about the data, test hypotheses, describe associations (correlations), and to model relationships (regression) within the data [Ale10]. While these analysis may be indicative of statistical trends, they do not fully capture the intrinsic non-linear epidemic dynamics seen in respiratory viruses like RSV and Influenza A [PVA+15, BMW+19, YCLS15]. We aim to understand our data to make more accurate assumptions about the mechanistic models that are implemented in Chapter 4.

3.3.1 Disease patterns throughout Chile

Respiratory viruses are known to be extremely seasonal [van28, LS14, PB12, Mon04, NK22, SL03, SK09]. In fact, studies have shown that outbreaks of RSV and Influenza A occur in temperate climates during the winter season, whereas low activity is detected during the summer months [NK22, BMW+19, BAC+13, WWM+98, PB14, LS14]. We want to explore the behavior of RSV and Influenza A and determine how seasonal the outbreaks are in Chile. To do so, we first explore the monthly amount of cases across all of Chile, as well as the density of cases for each region independently. To further analyze the seasonality of both viruses, we establish the periodicity of the outbreaks by using wavelet analysis.

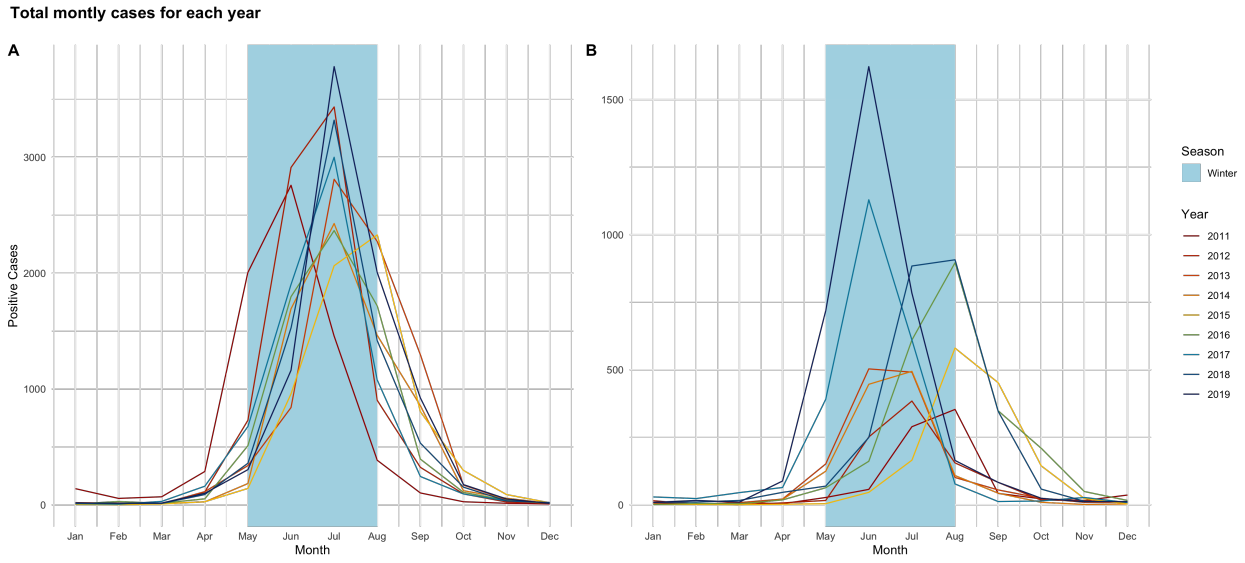

Figure 3.6: Monthly confirmed cases for each year across all of Chile. The shaded region shows the range of months for the winter season in Chile. **A.** RSV. **B.** Influenza A.

Chile is located in the Southern Hemisphere and thus experiences seasons during different months compared to the Northern Hemisphere. Summer is from December to February, autumn is from March to May, winter is from June to August, and spring is from September to November. Both RSV and Influenza A were found to be extremely seasonal across all of Chile, with epidemics occurring during the winter months (see Figure 3.6). This coincides with the seasonal patterns seen in the US [BMW+19, SPV+10, BAC+13, SK09].

We also explore the density of cases for each region. A density plot is a representation of the distribution of a numeric variable. It uses a kernel density estimate to show the probability density function of the variable [Sil18]. A kernel is a special type of probability density function (PDF) with the added property that it must be even. The steps to estimate the density are:

1. Choose a kernel, like Gaussian.

2. At each data point x_i build the scaled kernel function

$$h^{-1}K[(x - x_i)/h]$$

   where K is your chosen kernel function. The parameter h is called the bandwidth, or the smoothing parameter.

3. Add all of the individual scaled kernel functions and divide by n. This places a probability of $1/n$ to each x_i

$$\hat{f}(x) = n^{-1}h^{-1}\sum_{i=1}^{n} K[(x - x_i)/h].$$

   It also ensures that the kernel density estimate integrates to 1 over its support set.

For more on density functions in statistics and data analysis please refer to [Sil18]. In Figure 3.7 we have a density plot of positive RSV and Influenza A cases for each region. Peak timing (max incidence) for Influenza A appears to be more in sync throughout all the regions, whereas RSV has a different peak timing in the first three northern regions and then starts to synchronize as we move towards the south, where it looses the synchrony once again. This pattern has been documented in the Northern Hemisphere, where cases for RSV start in locations closer to the Equator and move North [PVA+15, BMW+19]. To our knowledge this is the first time this pattern has been documented in the Southern Hemisphere and throughout a wide latitudinal range (39 degrees in latitude).

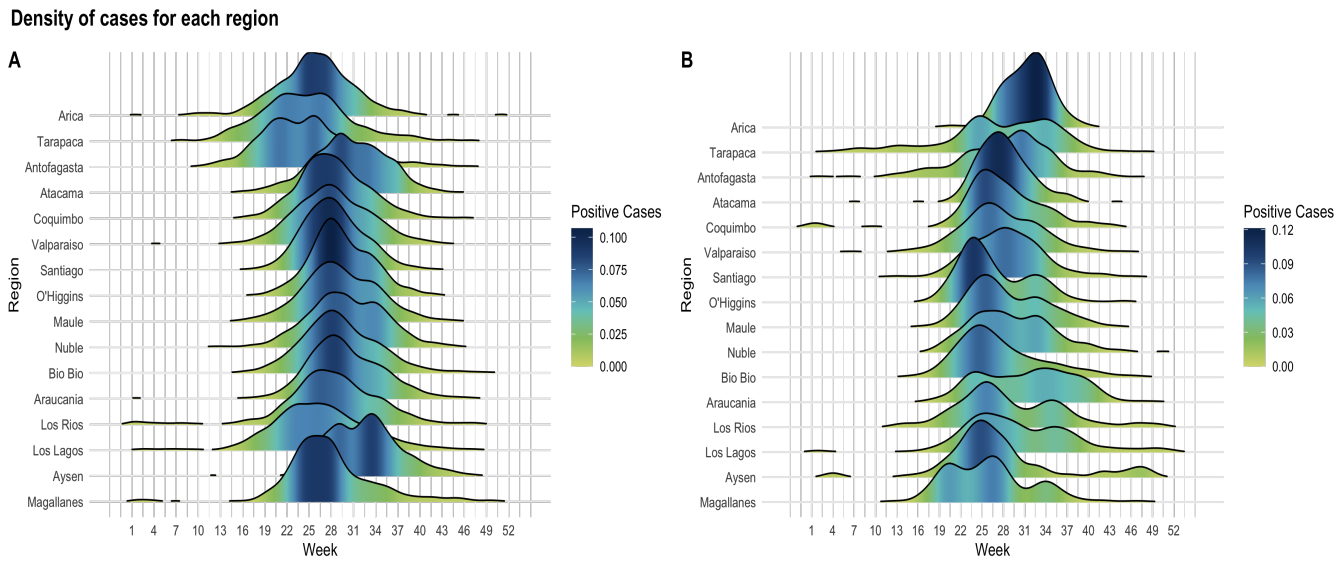

Figure 3.7: Weekly density plots for **A.** RSV. **B.** Influenza A.

An additional wavelet analysis allowed us to look at the periodicity of the incidence data. Wavelet analysis is a method that estimates the periodicity of a time series, and allows changing dynamical patterns over time, and thus is very useful in the study of pathogen dynamics [TC98]. Before wavelets, Fourier transforms (writing any signal as a combination of cosines and sines) were used to extract information from a signal. The main issue with Fourier transforms is that having high resolution in frequency makes it hard to isolate in time, since each frequency exists across all time. The lack of resolution between the frequency and time domain (Heisenberg uncertainty principle [HU16]) in Fourier transforms makes wavelets a better tool for applications in epidemiology [CCC14]. Wavelets can be thought as short "burst" waves that quickly die out. The objective is to use a wavelet transform to deconstruct a signal into multiple wavelets being added together. Wavelets are limited (local) in time and frequency, which allows for more resolution in the time domain.

More concretely, the continuous wavelet transform, $W_n(s)$, of a discrete sequence x_n, with equal spacing δt and $n = 0, \ldots, N-1$, is defined as the convolution of x_n with the scaled and translated version of the normalized wavelet function Ψ,

$$W_n(s) = \sum_{n'=0}^{N-1} x_{n'} \Psi^* \left[\frac{(n'-n)\delta t}{s} \right],$$

where Ψ^* is the complex conjugate of Ψ, and s is the wavelet scale (distance between oscillations in the wavelet), translated along the localized time index n [TC98]. The wavelet function should reflect the type of features present in the time series, and thus it depends

on the type of data to be analyzed. We are primarily interested in the periodicity of our data, and thus the choice of wavelet function is not critical. We will use the continuous complex-valued Morlet wavelet

$$\Psi_0(\eta) = \pi^{-1/4} e^{i\omega_0 \eta} e^{-\eta^2/2},$$

where η is a nondimensional time parameter and ω_0 is the nondimensional frequency and it is set to 6 to satisfy the admissibility condition (implies that the Fourier transform of Ψ vanishes at the zero frequency [Val99]). A graph of the Morlet function can be seen in Figure 3.8.

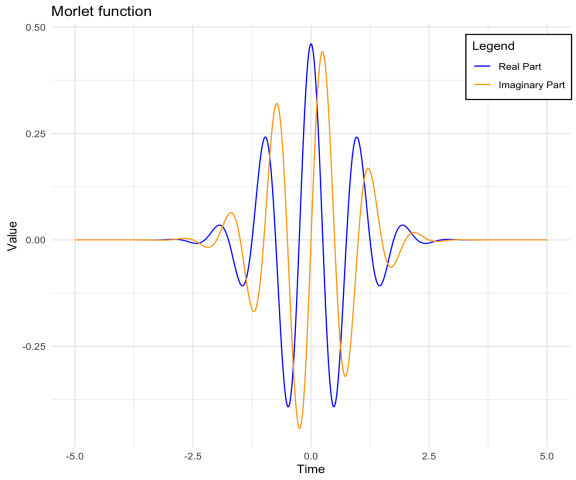

Figure 3.8: Graph of the real and imaginary parts of the Morlet wavelet.

The scale s is chosen arbitrarily and is written as fractional powers of two:

$$s_j = s_0 2^{j\delta j} \quad j = 0, 1, \ldots, J$$
$$J = \frac{1}{\delta j} \log_2(N\delta t / s_0)$$

where s_0 is the smallest scale so that the equivalent Fourier period is approximately $2\delta t$ and J is the largest. For the Morlet wavelet we need to choose the periods by specifying the number of octaves (main periods) and number of voices (number of subdivisions to estimate within each octave), with $1/\delta j$ determining the number of voices per octave and thus it is better to have a large value of voices. Smaller values of δj give finer resolution. Because the Morlet wavelet function is in general complex, the wavelet transform $W_n(s)$ is also complex. The transform can then be divided into the real part, imaginary part, amplitude $|W_n(s)|$, and phase $\arctan(\mathrm{Im}(W_n(s))/\mathrm{Re}(W_n(s)))$. The periodicity of the data is given by the local wavelet power spectrum at time point n and scale s, given by $|W_n(s)|^2$.

To find the periodicity of the Influenza A and RSV data we are using the package `Rwave` in R [CHT98], with the Morlet wavelet to analyze the frequency structure. Choosing eight octaves $\{2^1, 2^2, \ldots, 2^8\}$ with 16 voices gives us 128 periods. We have 9 years of data each with 52 weeks, and thus δt is chosen to be $1/52$ and $N = 468$ for all regions. To calculate the scales we have

$$s_0 = 1/26$$

and

$$\begin{aligned} J &= 16 \log_2((468/52) \cdot 26) \\ &= 128, \end{aligned}$$

giving us

$$s_j = \frac{1}{26} 2^{0.0625}$$

for $j = 0, 1, \ldots, 128$.

The wavelet power spectrum from the wavelet analysis shows annual peaks of power. Figures 3.9 and 3.10 show the power spectrum for all regions. Note that all of the regions have periodicity of 1 for both RSV (Figure 3.9) and Influenza A (Figure 3.10). Power is colour-coded as shown in the legend bars for each region. The peaks in the power spectrum are stronger for RSV than for Influenza A. This can be caused by the sparsity of the Influenza A data in some regions. Statistical methods like wavelet analysis need a lot of data in order to capture all the features at a given time and frequency and thus results vary depending on the amount of data sampled at each region. All regions show the same one year pattern (even if it is not as strong for some) and thus this results are enough for our analysis. Furthermore, Figure 3.11 shows the average power for periods, representing the average of local power from Figures 3.9 and 3.10 for each region. Note that the all regions have a higher average power for a period of 1 year.

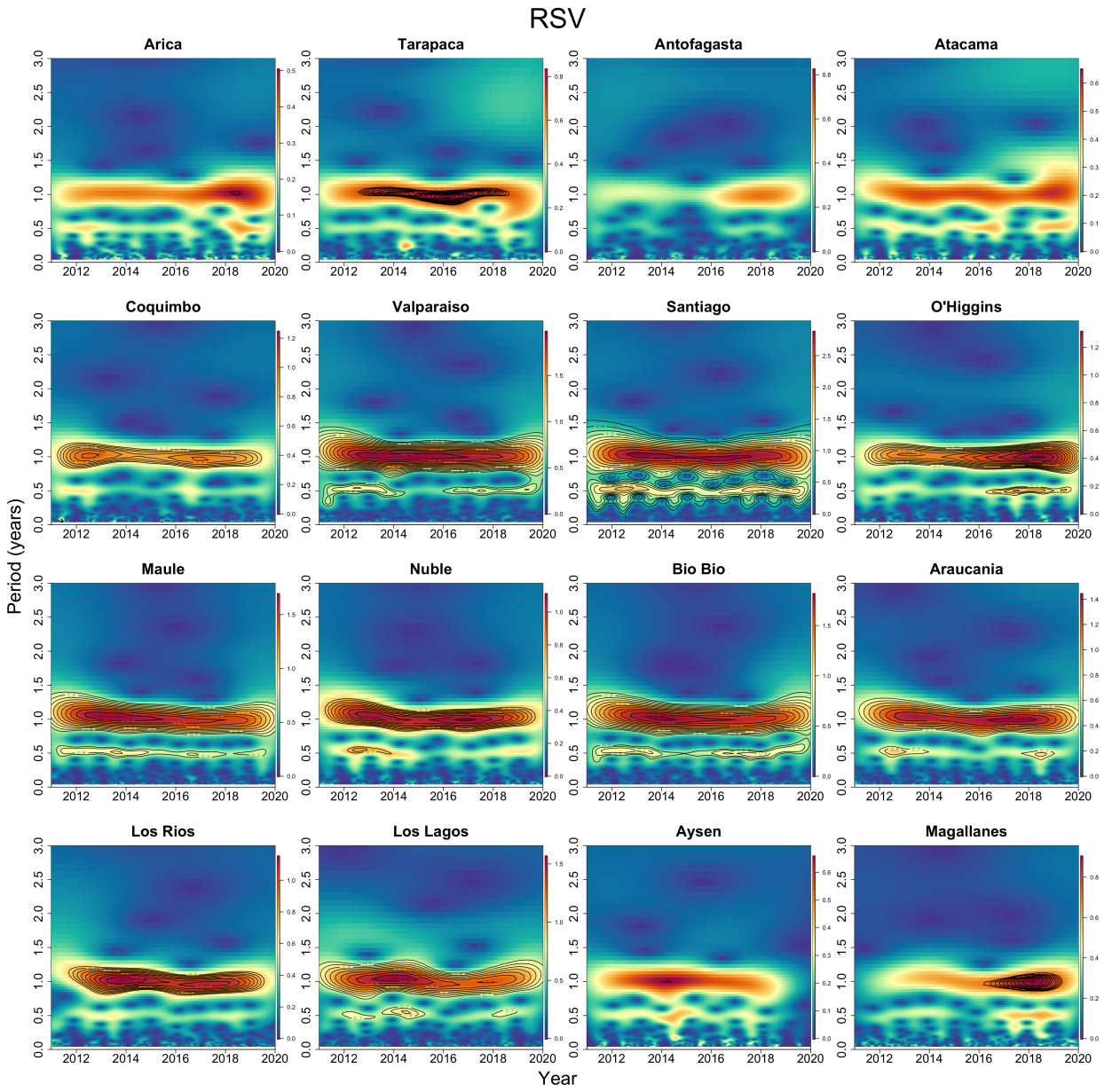

Figure 3.9: Wavelet power spectrum for RSV incidence in all regions of Chile.

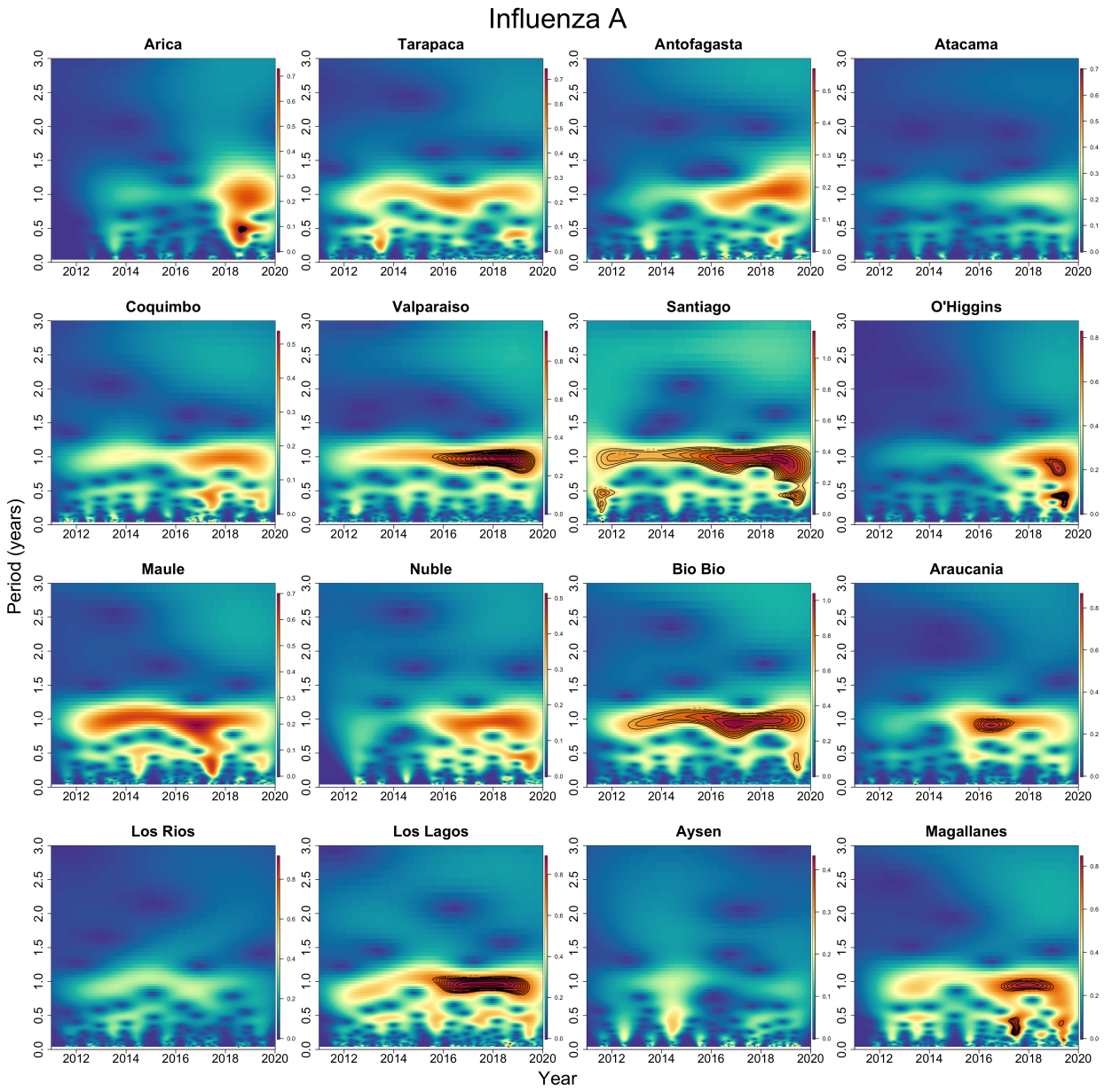

Figure 3.10: Wavelet power spectrum for Influenza A incidence in all regions of Chile.

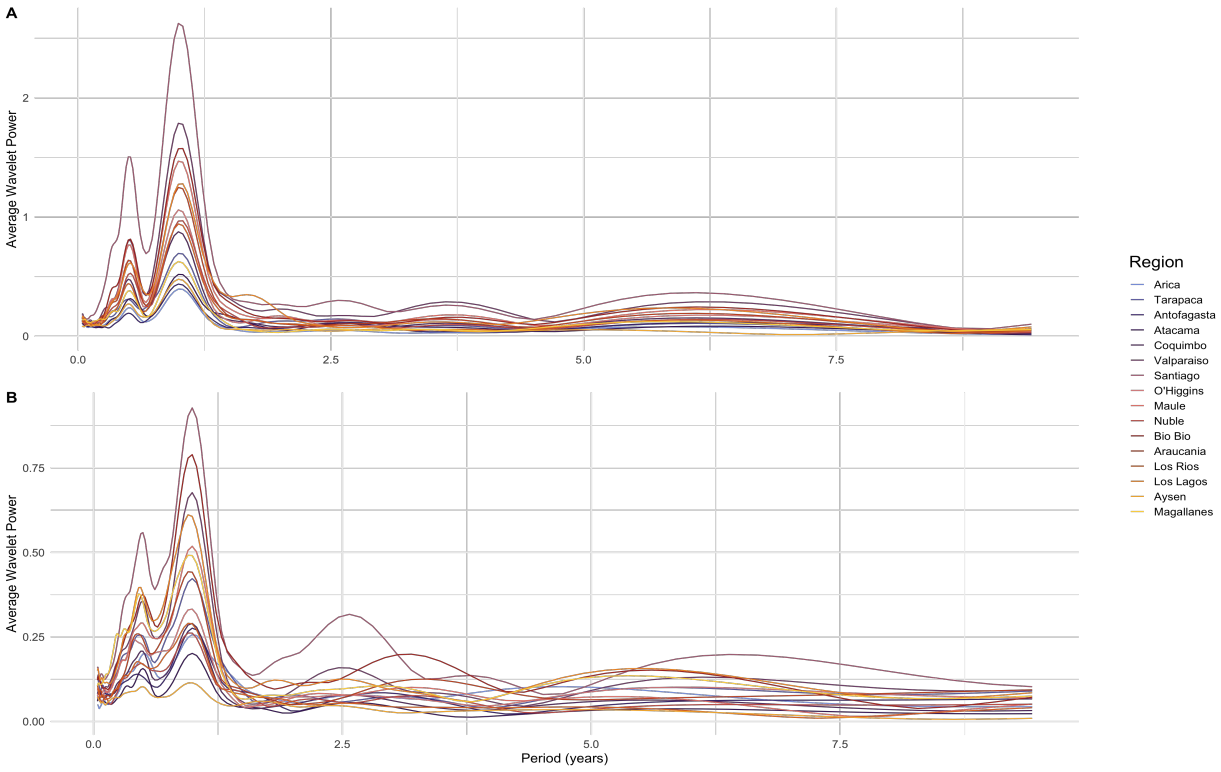

Figure 3.11: Average wavelet

3.3.2 Spatiotemporal associations of climate and disease incidence

3.3.2.1 Critical climate thresholds

In this section we explore the association of climate with activity of disease incidence. In Section 3.3.1 we established that both viruses are highly seasonal, with very high activity during the winter season. To assess the association with climate covariates (temperature and specific humidity) we first look at the proportion of positive cases during each season and compare it with the observed climate. We define the proportion of positive cases as the total amount of positive lab-confirmed cases divided by the total amount of respiratory cases (i.e., the sum of all Influenza A, RSV, Influenza B, Adenovirus, Parainfluenza, and Metapneumovirus cases for a given week). Figure 3.12 shows that high disease activity occurs throughout Chile during very low temperature (0°C- 30°C) and specific humidity (2.5g/kg - 15g/kg) values for both RSV and Influenza A.

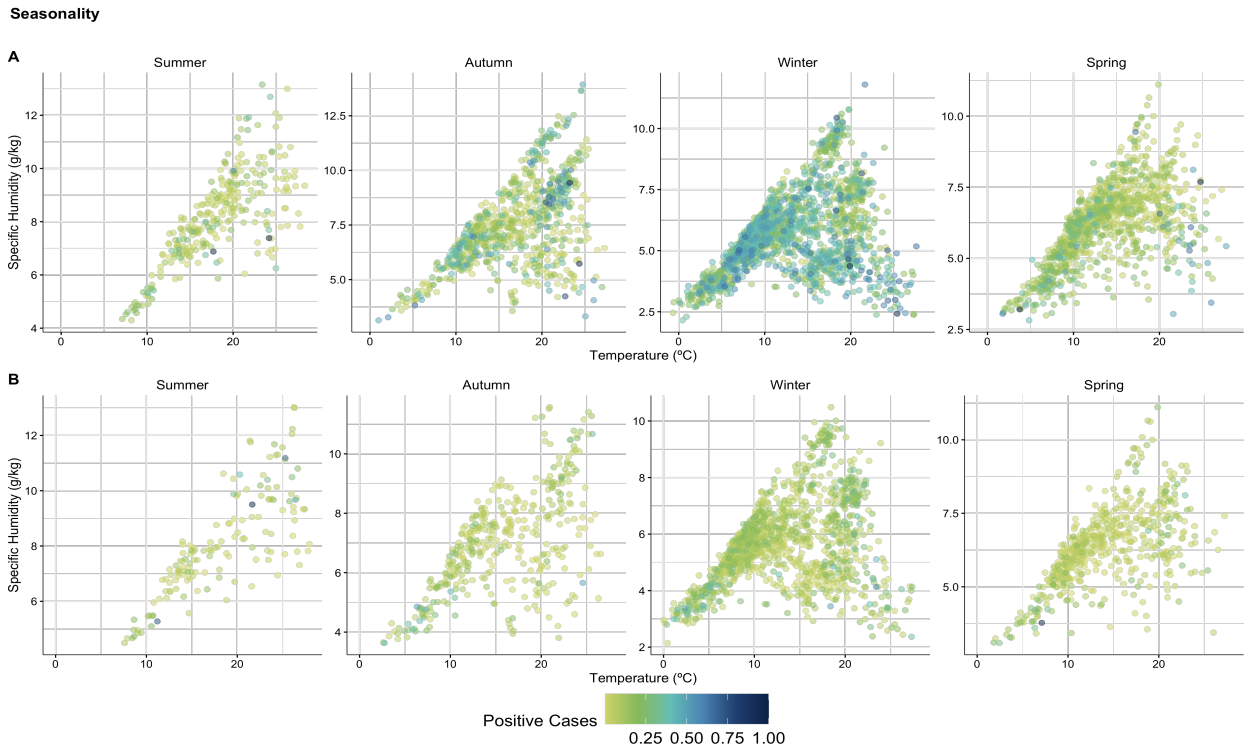

Figure 3.12: **A.** RSV cases by season **B.** Influenza A cases by season.

However, Chile has a wide range of weather conditions across a large geographic scale, extending across 39 degrees in latitude, and approximately 9 degrees in longitude for the continental territory. According to the Köppen system, Chile has at least seven major climatic sub-types, ranging from very cold and dry regions in the North, to Mediterranean

climate (characterized by warm to hot, dry summers and mild, fairly wet winters) in central Chile and temperate regions towards the South. Therefore, each region can have different weather patterns depending on its location. This makes it hard to generalize the impact of climate covariates on disease spread across all of Chile and thus we look at each of the regions individually.

To assess how different the climate covariates are during outbreaks for each of the regions we look at density plots of positive RSV and Influenza A cases but with respect to climate covariates. Figure 3.13 has the density plot of positive RSV and Influenza A cases with respect to temperature.

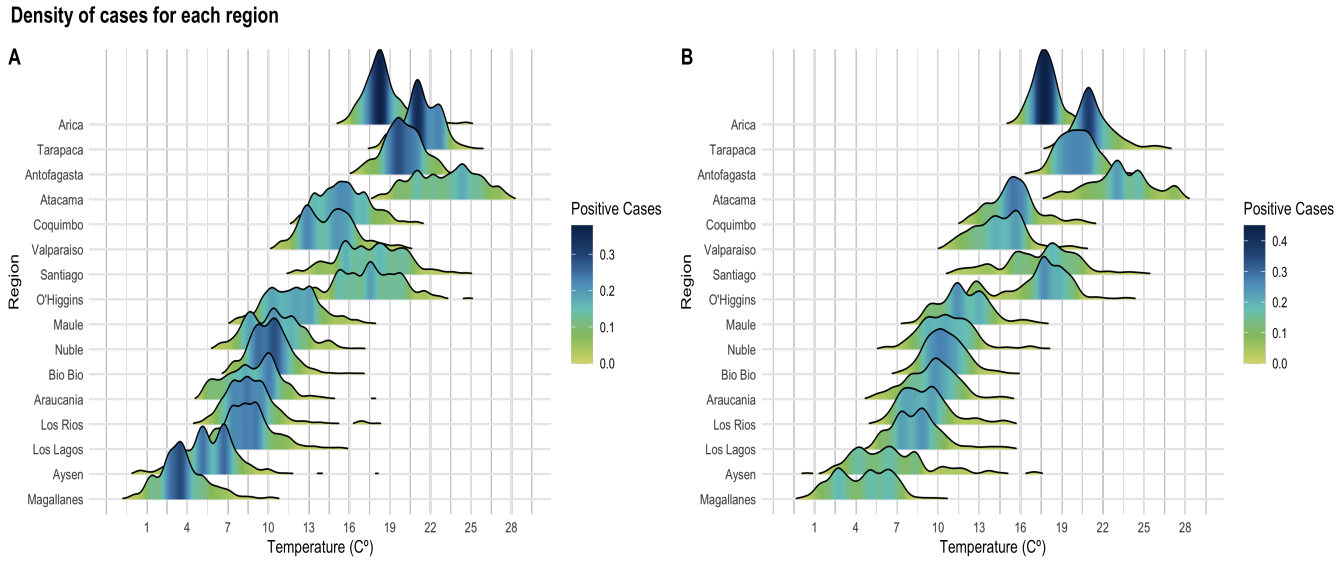

Figure 3.13: **A.** RSV cases with respect to temperature ranges. **B.** Influenza A with respect to temperature ranges.

Note that the ranges of temperature for which cases peak varies widely across regions. There is a latitudinal gradient with respect to temperature, going from North to South. The range of temperatures in Arica (North) is approximately $(16°C - 21°C)$ while in Magallanes (South) it is approximately $(1°C - 7°C)$ for RSV. We observe the same pattern for Influenza A, where Arica has a range of $(14°C - 22°C)$ and Magallanes has a range of $(1°C - 7°C)$. A similar gradient with respect to specific humidity is also observed for both diseases (see Figure B.3 in Appendix B). While it is expected for the climate to be different in each of the regions, these plots suggest that diseases like RSV and Influenza A can survive and thrive during very different cold conditions. In fact, in [LS14], the authors used a guinea pig model of influenza virus transmission to test the impact of ambient temperature and relative humidity on the efficiency of viral spread between hosts and found that transmission was highly efficient at $5°C$ but was blocked or inefficient at $30°C$. Dry conditions were

also found to be more favorable for spread, and several articles have found similar trends [SK09, SPV+10].

We are interested in finding a critical threshold in which positive cases start to sharply increase (i.e., cases start to increase abruptly). Segmented regression is a statistical technique that has been used to model ecological and biomedical thresholds [TL03, SC02, BWBI96, KFFM00]. Segmented regression, also known as piecewise linear regression, is a method in regression analysis in which the independent variable is partitioned into a finite number of equally spaced segments with breakpoints, or knots, at predetermined places [SW89]. The breakpoints can be interpreted as a critical threshold value beyond or below which changes occur. A function can then be fit to model the independent variable in each segment. Assuming a linear relationship between the independent and dependent variable the simplest segmented regression model joins two straight lines at the breakpoint as follows:

$$y_i = \begin{cases} \beta_1 x_i + \beta_0 + e_i, & \text{for } \ x_i \leq \alpha, \\ \beta_1 x_i + \beta_2 (x_i - \alpha) + \beta_0 + e_i, & \text{for } \ x_i > \alpha, \end{cases}$$

where y_i is the dependent variable, x_i is the independent variable, α is the breakpoint (threshold), and e_i are assumed to be independent, additive errors with mean zero, constant variance, and finite absolute moment of order > 2 [SW89, TL03]. The slopes of the lines are given by β_1 and $\beta_1 + \beta_2$. Parametrizing the model in this way forces continuity at the breakpoint.

When applying the segmented linear regression method to data of the form (x, y), in which y is the dependent variable and x the independent variable, the least squares method is applied separately to each segment, by which the regression lines are made to fit the data set as closely as possible while minimizing the sum of squares of the differences between observed and calculated values of the dependent variable.

Using segmented regression, we can find the critical threshold α that a covariate must reach for cases to increase, thus we want to find α such that $y(x)$ sharply increases. We do this by setting $\beta_1 + \beta_2 = 0$, which gives us

$$y(x) = \begin{cases} \beta_1 x + \beta_0 + e & x < \alpha \\ -\beta_2 \alpha + \beta_0 + e & x > \alpha \end{cases}$$

Letting $\beta_3 = -\beta_2 \alpha + \beta_0 + e$ then

$$y(x) = \begin{cases} \beta_1 x + \beta_0 + e & x < \alpha \\ \beta_3 & x > \alpha \end{cases}$$

Therefore we will have one linear equation $\beta_1 x + \beta_0 + e$ and a constant equation β_3. To find these two equations we use the R programming [R C21] package `segmented`, which allows us to estimate linear models having one or more segmented relationships in the linear predictor [Mug08]. It provides estimates of the slopes and breakpoints along with standard errors. The algorithm used by `segmented` is an iterative procedure found in [Mug03]. Table 3.1 shows the temperature and specific humidity thresholds obtained from the segmented regression analysis. Figure 3.14, as well as Figures B.5, B.4, and B.6 in Appendix B show the graphed results of the segmented regression analysis for all 16 regions of Chile, giving us thresholds for both temperature (T_c) and specific humidity (q_c).

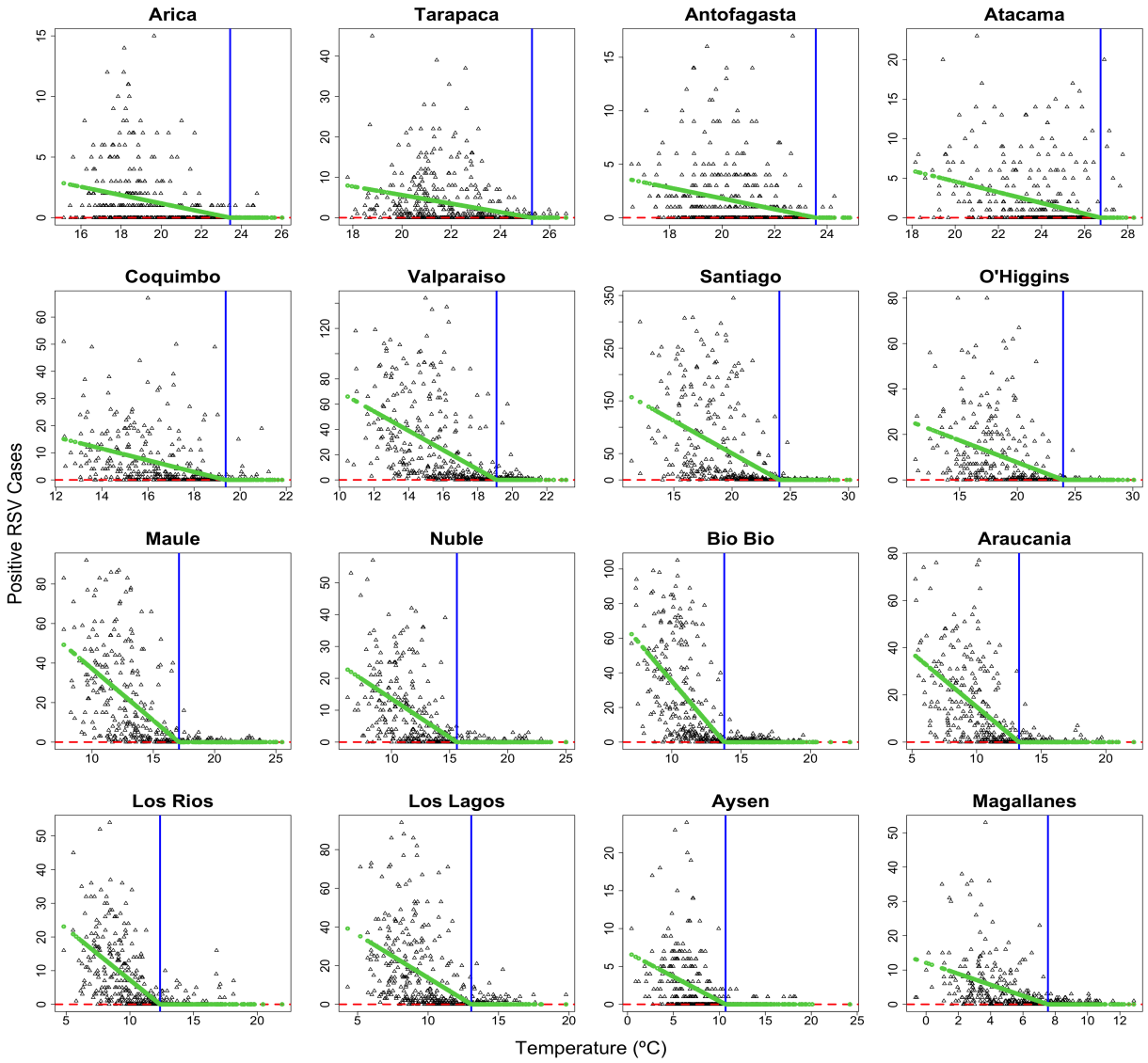

Figure 3.14: Segmented regression for positive RSV cases and temperature (°C).

As expected, similar to Figure 3.13, we find a latitudinal gradient with respect to the climatic thresholds, showing that locations closer to the Equator have higher climatic thresholds for both covariates (Figures 3.15 and 3.16). The critical thresholds from Table 3.1 are very region specific (i.e., vary from region to region) for both RSV and Influenza A. RSV appears to have a stronger latitudinal gradient than Influenza A with respect to specific humidity. This suggests that for Influenza A cases to increase, specific humidity needs to reach a stricter and narrower range of values. Both viruses exhibit similar temperature values, and no meaningful differences were found.

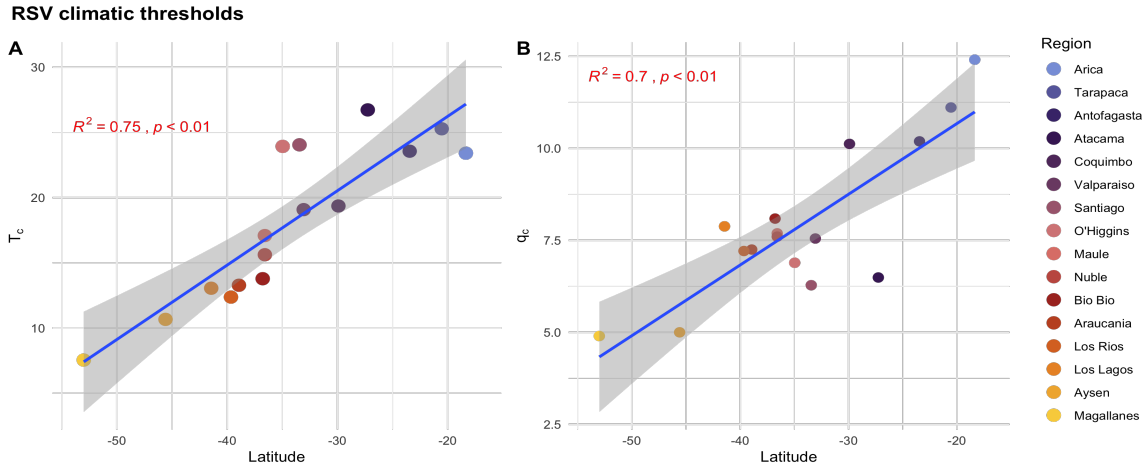

Figure 3.15: Association of climatic thresholds and location for RSV incidence.**A.** Temperature thresholds (° C). **B.** Specific humidity thresholds (g/kg).

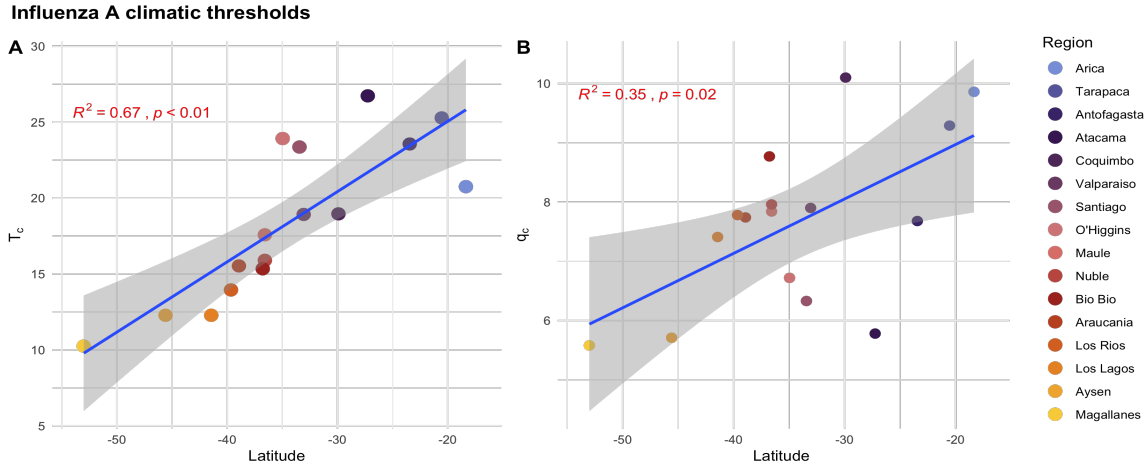

Figure 3.16: Association of climatic thresholds and location for Influenza A incidence. **A.** Temperature thresholds (°C). **B.** Specific humidity thresholds (g/kg).

	RSV		Influenza A	
Region	T_c	q_c	T_c	q_c
Arica	23.41	12.41	20.75	9.86
Tarapaca	25.28	11.11	25.28	9.29
Antofagasta	23.56	10.19	23.56	7.68
Atacama	26.73	6.49	26.73	5.78
Coquimbo	19.36	10.12	18.96	10.10
Valparaiso	19.08	7.55	18.92	7.90
Santiago	24.05	6.28	23.36	6.33
O'Higgins	23.93	6.89	23.92	6.72
Maule	17.09	7.69	17.58	7.84
Nuble	15.62	7.60	15.90	7.96
Bio Bio	13.78	8.09	15.34	8.77
Araucania	13.28	7.25	15.53	7.74
Los Rios	12.37	7.21	13.95	7.78
Los Lagos	13.05	7.88	12.28	7.41
Aysen	10.66	5.00	12.28	5.71
Magallanes	7.54	4.90	10.25	5.58

Table 3.1: Temperature and specific humidity thresholds (T_c and q_c, respectively) obtained from segmented regression analysis for both RSV and Influenza A.

The critical threshold values for Santiago, shown in Table 3.1, will be used in Section 4.2.5 as part of the climate forcing component of the transmission term in our models.

3.3.3 Timing of disease onset

In Section 3.3.1 we established how RSV and Influenza A predominate in the winter season across all of Chile. In other words, both respiratory viruses exhibit a "predictable" seasonality. Therefore, even with a lot of regional and year-to-year variability (see Section 3.3.4), we expect to observe disease spread during the winter. Regardless of this strong seasonal periodicity, it is not enough to clearly pinpoint the timing of disease incidence, i.e., exact week at which we expect cases to start. Knowing the time and place of the occurrence of respiratory viruses can be very useful for prioritizing intervention strategies like vaccination or masking, that could reduce disease frequency and severity. We are interested in assessing the role of environmental covariates on the timing of disease incidence. More precisely we want to answer the following questions:

1. What is the timing of disease onset?

2. How well do covariates explain the variance in onset?

Disease onset can be defined as the time at which disease incidence starts to increase significantly. To calculate mean onset week we first calculate mean incidence per week (i.e., each week averaged over all years for a given location), and then normalize these values between 0 and 1. We define onset as when normalized incidence exceeds 0.2, assuming this value is low enough to constitute onset but high enough to exceed random fluctuations in the data, as calculated in [BMW+19]. We are interested in the effect of covariates on the timing of RSV and Influenza A onset and thus we look at the relationship between the calculated mean onset week and the mean value of the covariates during the mean onset week. In Figure 3.17 we have maps with the onset week for RSV and Influenza A in each region, where lighter colors represent earlier onset and darker colors represent later onset. Figures 3.18 and 3.19 show the correlations between timing of epidemic onset and mean climate covariates, as well as latitude.

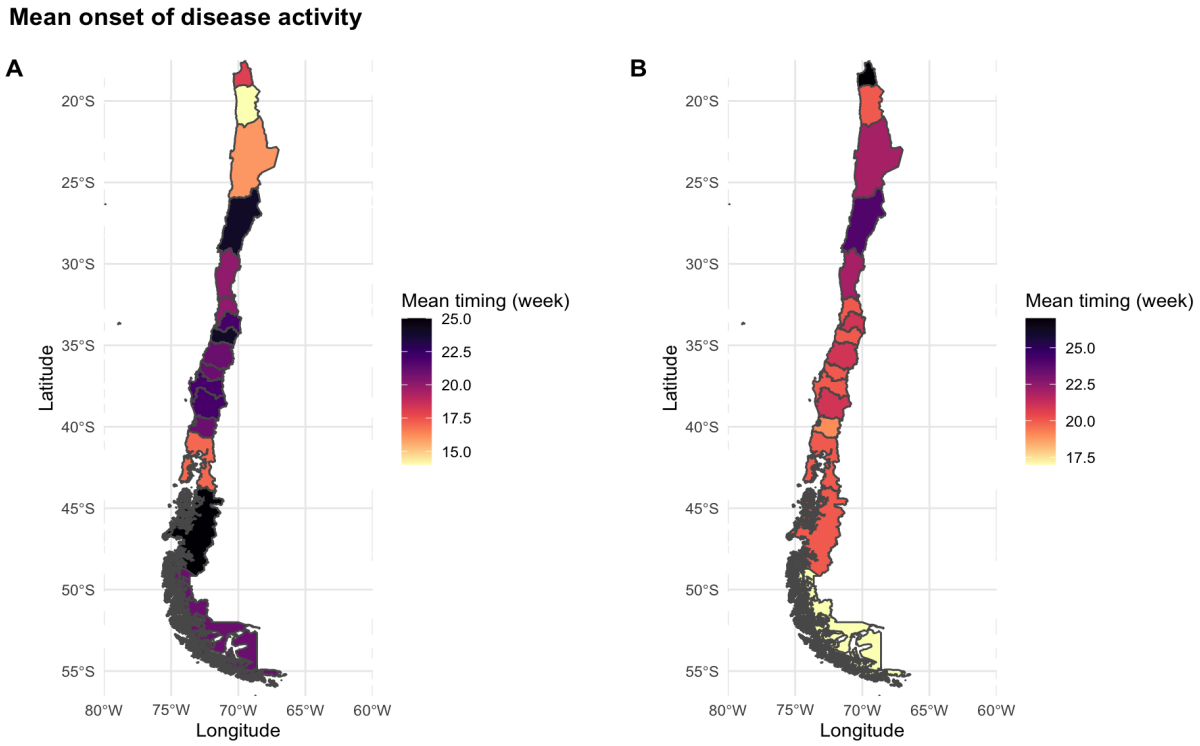

Figure 3.17: **A.** Mean onset week of RSV activity in regions of Chile. **B.** Mean onset week of Influenza A activity in regions of Chile.

Onset analysis for RSV shows similar results to the ones found in [BMW+19]. Specific humidity is significantly associated with the mean timing onset of RSV ($p < 0.01$), and explains 62% of the variance. In the case of Influenza A, this association is weaker, where specific humidity ($p = 0.12$) only explains 12% of the variance. On the other hand, temperature explains 32% of the variance in onset for Influenza A, while for RSV it only explains

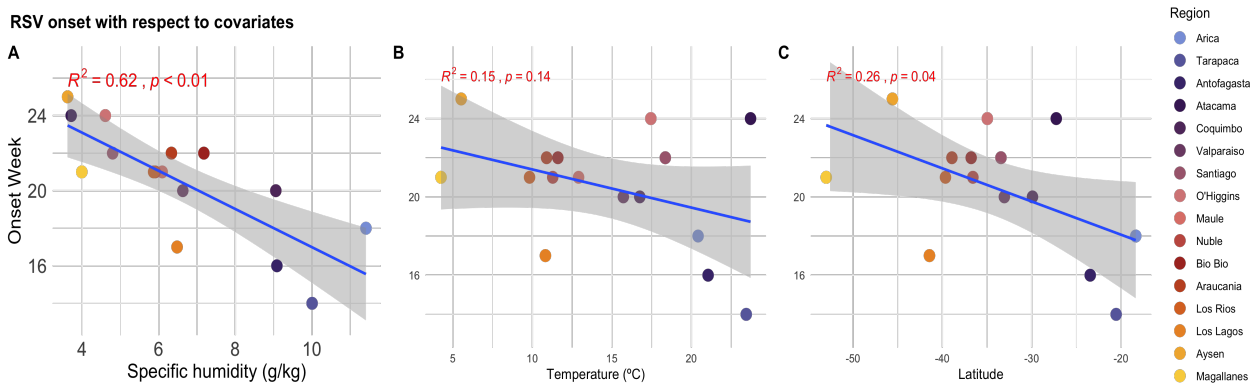

Figure 3.18: **A.** Onset of RSV with respect to specific humidity (g/kg). **B.** Onset of RSV with respect to temperature (°C). **C.** Onset of RSV with respect to latitude. Regions are ordered from North to South.

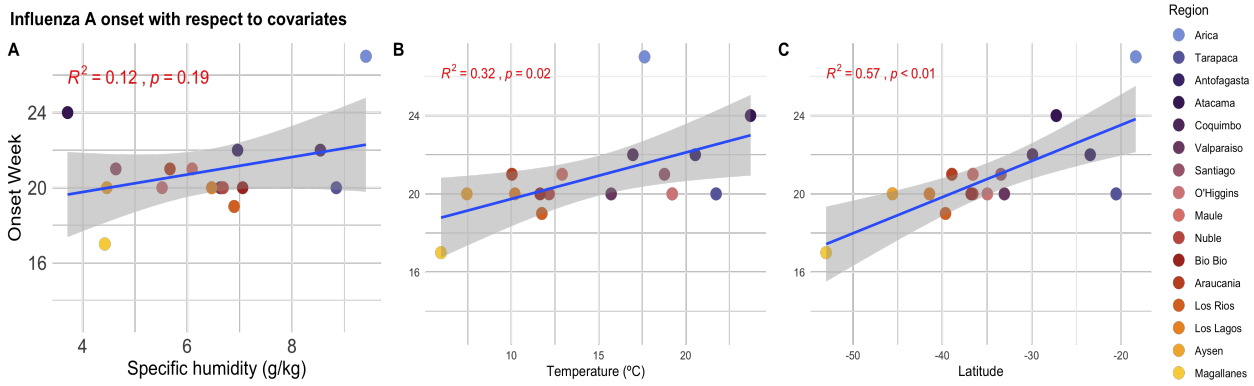

Figure 3.19: **A.** Onset of Influenza A with respect to specific humidity (g/kg). **B.** Onset of Influenza A with respect to temperature (°C). **C.** Onset of Influenza A with respect to latitude. Regions in the legend are ordered from North to South.

15% of the variance.

We also observe a difference in the slope of the linear regression, RSV has a negative slope while Influenza A has a positive slope. This means that RSV incidence starts earlier in the northern part of Chile, while Influenza A starts earlier in the south, which can also be observed in 3.17. This can be seen in Figures 3.18C and 3.19C. Latitude explains 57% and 26% of the variance in Influenza A and RSV onset, respectively.

3.3.4 Interannual variability

The term interannual variability can be defined as the observed year-to-year variation. We are interested in the interannual variability of the timing of onset, as well as the peak size or amplitude of an outbreak or epidemic cycle for a particular year, i.e., year-to-year variation

in the intensity of the spread. In this section we want to answer the following questions:

1. Do year-to-year variations in climate impact the timing of disease onset?
2. Do year-to-year variations in climate impact the timing of the peak?
3. Do year-to-year variations in climate impact the size or amplitude of the outbreak?

The answers to these questions will allow us to determine if environmental factors are driving the interannual variability observed in our data and whether climate covariates are the strongest influencing factor.

3.3.4.1 Onset

To investigate whether year-to-year variations in specific humidity and temperature can alter the timing of onset of the epidemic within a particular location, we calculate the onset week for every year for each location, using the same definition of onset from Section 3.3.3. We then fit a linear regression model with onset week as the dependent variable and climate covariates as the independent variable. The model includes dummies for each year and location to remove mean onset timing. We find that a 5 (g/kg) increase in mean annual specific humidity and a 1°C increase in mean annual temperature shifts the timing of the RSV epidemic back by 1 week ($p < 0.001$), shown in Figure 3.20 and Table 3.2. The effects found for Influenza A were not as significant where 3 (g/kg) increase in mean annual specific humidity ($p < 0.01$) and a 0.1 (°C) increase in mean annual temperature ($p < 0.5$) shifts the timing of the epidemic back by 1 week, as seen in Figure B.7 and Table 3.2.

	RSV	Influenza A
Specific humidity (g/kg)	-5.233***	-3.130*
	(3.144)	(4.687)
Temperature ($°C$)	-1.6213**	-0.1595
	(3.285)	(4.797)
Num. obs.	143	137
Adjusted R^2 specific humidity	0.4855	0.4022
Adjusted R^2 temperature	0.4383	0.3737

Signif. codes: 0 '***'; 0.001 '**'; 0.01 ;'*' 0.05 ;'.' 0.1 ; ' ' 1

Table 3.2: Regression of climate variables on onset week. The model includes location fixed effect to control for spatial heterogeneities in onset determinants. Standard errors are shown in parentheses.

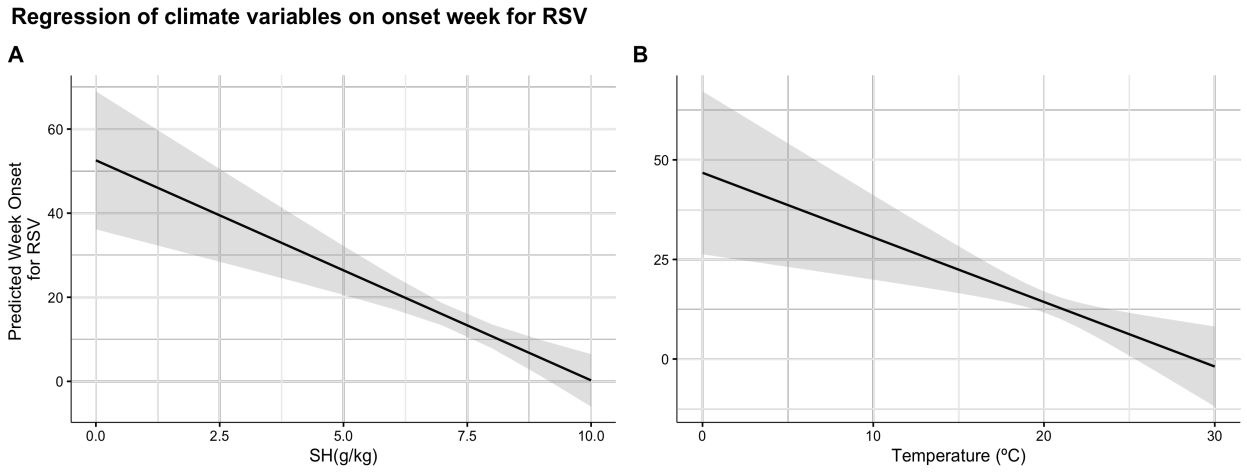

Figure 3.20: **A.** Graph of linear regression model of onset week for RSV with specific humidity (g/kg) as independent variable. **B.** Graph of linear regression model of onset week for RSV with temperature ($^{\circ}C$) as independent variable.

As another measure of mean timing of epidemic onset, we calculate the center of gravity. The timing of disease epidemics is measured by calculating the center of gravity, G, which is defined by the mean week of activity for each season and epidemic year. For each region r and year y, the formula for G is given by

$$G_{r,y} = \frac{\sum_{w \in [1,52]} w * cases_{r,y,w}}{\sum_{w \in [1,52]} cases_{r,y,w}},$$

where $cases_{r,y,w}$ is the number of laboratory confirmed cases reported in region r during epidemic year y and week w, as seen in [PVS+09, PVA+15]. Note that this calculation of timing of onset is less biased than choosing an arbitrary cutoff as done in above and in Section 3.3.3.

Upon inspection of the values obtained from calculating the center of gravity, we note that the onset is occurring at much later times, suggesting that the center of gravity formula is capturing the timing of the peak. Figure 3.21 shows the mean timing of the peak onset for all regions of Chile. We observe the same patterns of onset for RSV and Influenza A from Section 3.3.3. For RSV the regions in the North peak earlier, the central regions seem to peak at the same time, and then the southern regions are not as synchronized. In contrast to RSV, Influenza A is much more in sync throughout all of Chile. When we evaluate the relationship between the environmental factors and timing of peak week, there is no significant association (Figures B.8 and B.9).

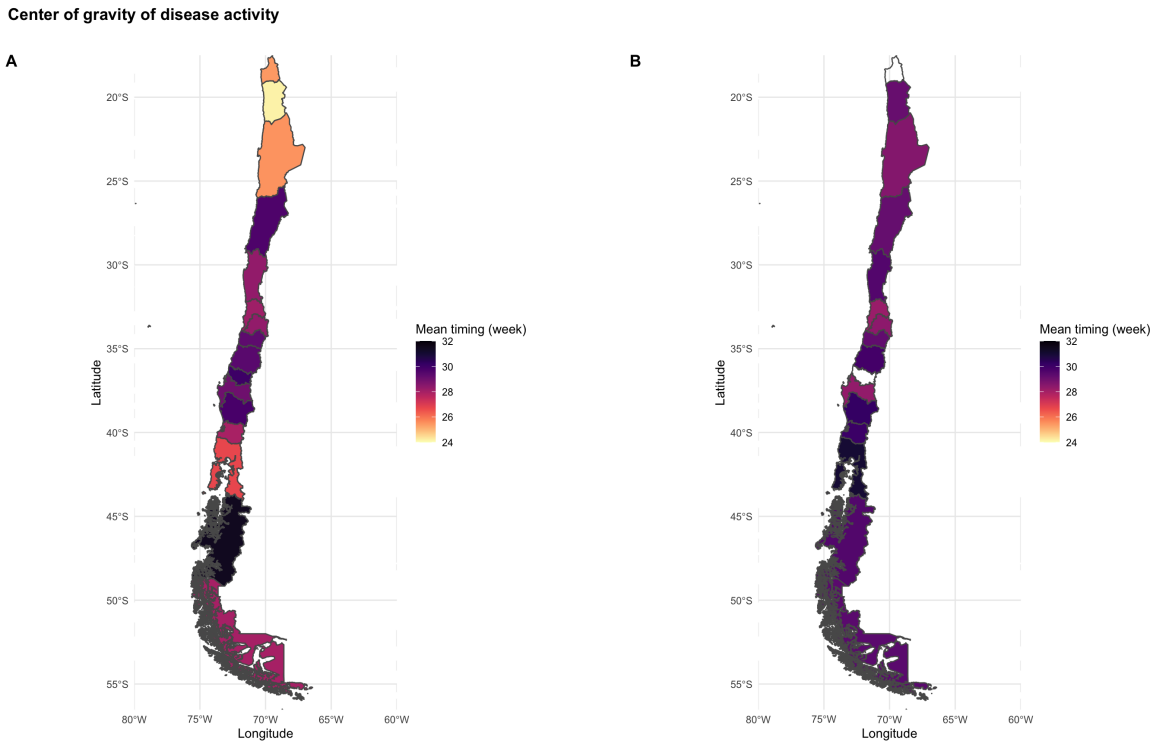

Figure 3.21: **A.** Center of gravity of RSV activity in regions of Chile. **B.** Center of gravity of Influenza A activity in regions of Chile. Note that regions without any color had years with 0 cases and thus would be biased.

3.3.4.2 Size of outbreak

To assess if interannual variability is caused by the climate covariates we look at the amount of cases and the climate covariates during the peak timing (i.e., time at which the maximum amount cases for a particular year is reached). As seen in Figure 3.22 (Figure B.10) there is no significant association between specific humidity and the year-to-year amplitude of the RSV (Influenza A) epidemic. Similar results were found when looking at temperature (see Figures B.11 and B.12 in Appendix B). These results suggest that there are other mechanisms involved in the year-to-year variation of outbreak size for both RSV and Influenza A.

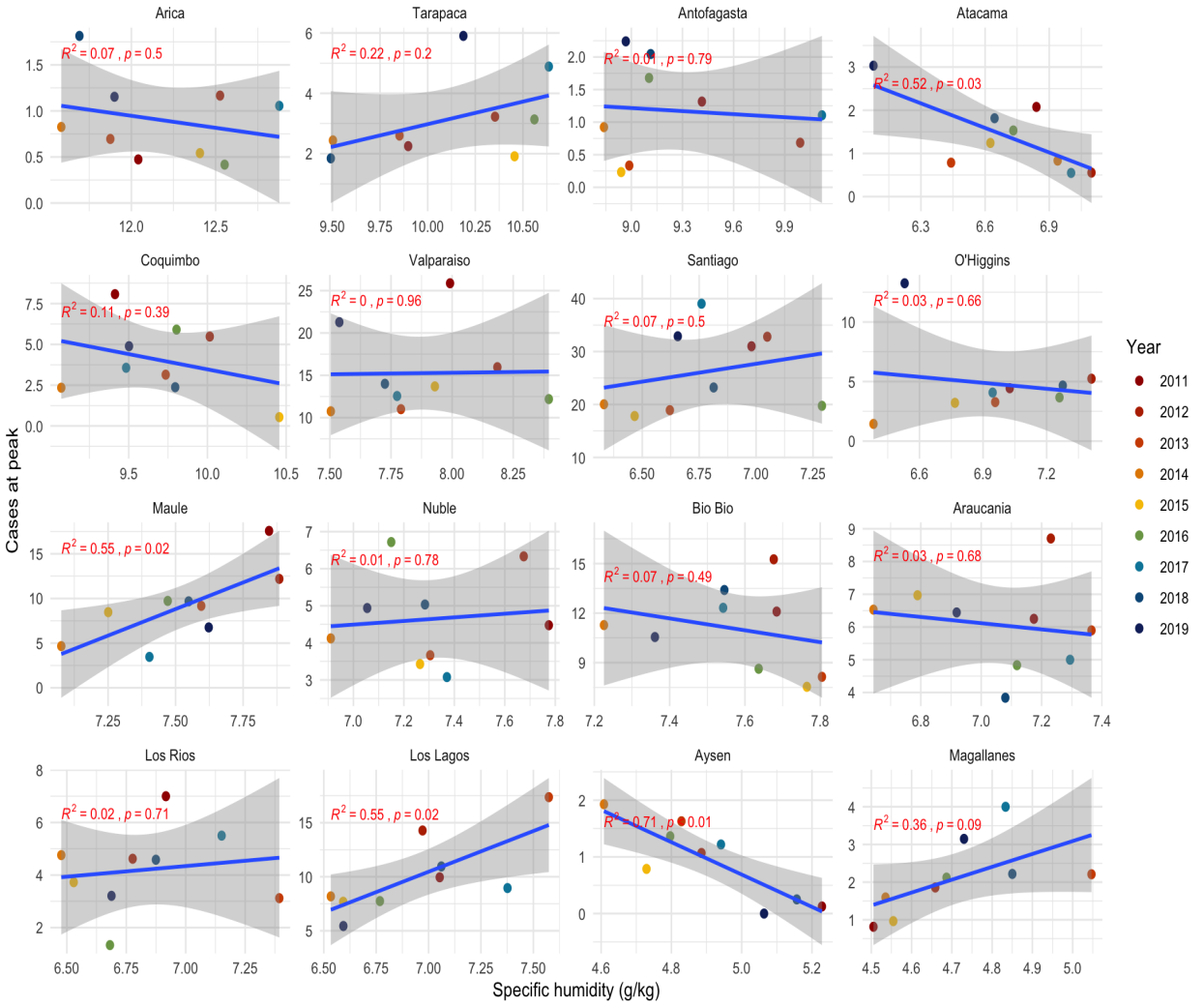

Figure 3.22: Graph of interannual variability in peak size for RSV with respect to specific humidity (g/kg).

3.4 Summary of findings

- There are annual seasonal epidemics during the winter months (May to August) for both RSV and Influenza A.
- There is a latitudinal gradient with respect to both climate covariates, going from north to south, suggesting that diseases like RSV and Influenza A can survive and thrive during very different cold conditions.
- There is a latitudinal gradient with respect to the critical climate thresholds in which

disease incidence starts to sharply increase, and values are region specific. RSV appears to have a stronger latitudinal gradient than Influenza A with respect to specific humidity, suggesting that for Influenza A cases to increase, specific humidity needs to reach a narrower range of values across all of Chile.

- Specific humidity is significantly associated with the mean timing onset of RSV, while for Influenza A this association is weaker.
- RSV onset is earlier in the Northern part of Chile while Influenza A starts earlier in the South. This is similar to the behavior seen in the Northern Hemisphere.
- A 5 (g/kg) increase in mean annual specific humidity and a 1 °C increase in mean annual temperature shifts the timing of the RSV epidemic back by 1 week. The effects found for Influenza A were not as strong where 3 (g/kg) increase in mean annual specific humidity and a 0.1 °C increase in mean annual temperature shifts the timing of the epidemic back by 1 week.
- The timing of peak incidence has a similar pattern to the timing of disease onset, with Influenza A peak timing being more synchronized across all regions.
- There is no significant association between climate covariates and the year-to-year variation in amplitude of epidemics for both RSV and Influenza A.

3.5 Discussion

We observe consistent seasonality, with cases prevalent during the winter across all of Chile. Both diseases exhibit annual (period 1) peaks of power. Specific humidity and temperature are significantly associated with the onset of RSV. Influenza A has a weaker association with respect to mean onset week, with similar mean onset timing for each region. By looking at the year-to-year variations in specific humidity and temperature we find that a 5 (g/kg) increase in mean annual specific humidity and a 1 °C increase in mean annual temperature shifts the timing of the RSV epidemic back by 1 week. The effects found for Influenza A were not as significant where 3 (g/kg) increase in mean annual specific humidity and a 0.1 increase in mean annual temperature ($p < 0.5$) shifts the timing of the epidemic back by 1 week. These differences could be explained by heterogeneity in age of infection and contact patterns across different age groups. RSV is known to be biased towards infants, even though maternal antibodies offer brief protection from infection [OSF+09]. In fact, during the first 2 years of life, all children will have been infected at least once with RSV

[Gle86]. In contrast, because of the rapid antigenic evolution of the different influenza virus types, maternal immunity might not be acquired for influenza [YKH+13]. Influenza A affects more young adults than RSV. However, both RSV and Influenza A cause increased mortality in elderly individuals, which might be due to the decline of immune function as people get older [WF04]. Immune history and connectivity across different regions might also interfere with the role of climate factors in the transmission of respiratory viruses.

There is a latitudinal gradient in connection to temperature, seen in Figure 3.13 (as well as specific humidity), which suggests that the host's response to the environment might be a strong force in the transmission of both RSV and Influenza A. Human behavior changes as temperature changes and thus there might be more indoor crowding, increased number of close contacts, holiday and vacation travel, which can cause increased transmission [ETBE12]. Host defense mechanisms are also impacted by changing environmental conditions [Dow01].

The absence of strong interannual variability indicates that climate covariates do not influence the amplitude of the outbreaks. Some infectious diseases like cholera and dengue exhibit an interannual component that is driven by the dynamics of El Niño Southern Oscillation [PRE+00, CCMH05]. One possible explanation for these differences is the epidemic nature of RSV and Influenza A, and their annual predictability. Diseases like cholera and dengue tend to be unstable because of their regular pandemic behavior. Water-borne and vector-borne diseases also depend on host–pathogen interactions, e.g. drinking contaminated water or the abundance and distribution of mosquitoes (where climate influences the biology of the vector).

The relationship between climate, human behaviour, and infectious disease dynamics is very complex, making it difficult to disentangle the main driving mechanisms of disease spread. To contribute to the understanding of RSV and Influenza A spread mechanisms we will implement a compartmental mechanistic model that takes into account intrinsic and extrinsic factors. We are interested in understanding the processes that generate the observed seasonality and interannual variability of both respiratory viruses. In particular, we wish to quantify the effect of climate covariates on transmission by implementing a stochastic, seasonally forced MSIRS and SIRS modeling framework, for RSV and Influenza A respectively. By incorporating the climate covariates, stochasticity through measurement and process noise, as well as a stochastic observation component, we hope to quantify the effect of climate covariates on the dynamics of the viruses. Using the POMP modeling framework depicted in Section 2.2, we will use iterated filtering to infer initial conditions and parameters of the models and get a better picture of disease spread mechanisms of RSV and Influenza A. The results from this analysis can potentially allow us to identify the role of

climate versus nonlinear disease dynamics [LBI+10] in disease incidence. In the next chapter we test whether the interannual variability observed in RSV and Influenza A is driven by (i) climate, (ii) the intrinsic dynamics of the disease, or (iii) the interactions of these two mechanisms.

Chapter 4

Quantifying the effect of environmental drivers on disease incidence in Santiago, Chile

In this chapter we use both statistical and mathematical modeling to quantify the effect of environmental drivers on disease incidence and and to estimate key epidemiological parameters relevant for the disease transmission dynamics of RSV and Influenza A in Santiago, Chile. Santiago is the capital and largest city of Chile. It is located in central Chile, and thus it is considered to have a Mediterranean climate by the Köppen system, characterized by hot, dry summers and cool, wet winters. Summers have temperatures reaching up to 35 °C on the hottest days and winters reach daily temperatures of up to 14°C and low temperatures near 0 °C. It is considered a metropolitan area, often characterized by high levels of economic activity, cultural diversity, and social interaction. As such, Santiago can play a crucial role in disease spread because of the high population densities and a high degree of connectivity with other regions.

In Section 4.1 we use statistical models to test hypotheses about the relationship between climate covariates and disease activity. More precisely, we explore how timing of disease activity responds to changes in environmental factors by fitting multiple statistical models to the data. In Section 4.2 we dive into the core mathematical models used in the thesis. We construct the deterministic SIR-type models as ordinary differential equations (Sections 4.2.1 and 4.2.3) and their stochastic versions using continuous time Markov chains (Sections 4.2.2 and 4.2.4). The models are then implemented in Section 4.4, using the `pomp` package [KI22, KIMB+22], available in the R programming language. We also give an overview of our model selection process, based on the MLE and provide simulations for the best model in Sections 4.4.2 and 4.4.3. Finally, Section 4.4.5 has a discussion of our results.

4.1 Statistical analysis

Infectious disease modeling encompasses a wide range of methods where each one depends on the research question being answered. Mechanistic models, like the SIR modeling framework discussed in Section 2.1, allow us to study the underlying transmission dynamics of the spread of respiratory viruses. They can be used to forecast or simulate future transmission scenarios

under various assumptions about the parameters governing disease transmission (e.g. waning immunity, natural births and deaths, contact rates, and disease mortality rates) [HB20]. As mentioned in Chapter 2, the addition of seasonality and stochasticity in the model allows for more complex and realistic models. The way in which we include these in the model depends on the knowledge we have of the disease, the type of data we use to fit the parameters of the model, as well as the assumptions we make about the behavior of the disease. Which is why mathematical modeling goes hand in hand with statistical modeling.

In disease analysis, statistical modeling usually formalizes relationships among variables that may influence the spread of disease, describes how one or more variables are related to each other, and tests whether a hypothesis we have on disease spread is true [CMW14]. One type of statistical modeling is exploratory data analysis, as done in Chapter 3. To test hypotheses about the trends observed in the exploratory data analysis we can use confirmatory data analysis. Many traditional statistical analysis and modeling approaches can be found in [Fra80, CMW14, MAA20]. By fitting multiple statistical models to our data we can quantify the relationship between seasonality and climate without incorporating the process that governs the dynamics of the virus. In this section we explore how timing of disease activity responds to changes in environmental factors. Using climatic and non-climatic variables we analyze a variety of models, including multivariate logistic regression, linear regression, and generalized additive models (GAM). All regression models include month dummy variables which allows us to estimate the effect of inter-annual variation in climate conditions while controlling for possible unobserved confounders that vary seasonally.

A GAM with splines provided the best fit to the data than the other models as seen from AIC (information on AIC scores can be found in Section 4.4) scores in Table 4.1. A generalized additive model (GAM) is a Generalized Linear Model in which the expected value of the response variable is related to a nonlinear predictor through a link function. The idea is that the response variable depends linearly on unknown smooth functions of the predictor variable, i.e.,

$$g(\mu) = \beta_0 + f(x_1) + \cdots + f(x_n),$$

allowing us to capture nonlinear relationships. In summary, a GAM is a model which allows the linear model to learn nonlinear relationships by using the sum of arbitrary smooth functions of each variable to model the outcome. There are many smooth functions of the form

$$s(x) = \sum_{i=1}^{k} \beta_i b_i(x)$$

that we can use as basis functions for GAM models. The model can have k weights and

functions per variable in the equation. Bigger k increases the flexibility of the model such that it better captures the behavior of the data, but it has the drawback of increasing computational cost and overfitting. Splines are commonly used in GAMs because of their high degree of flexibility and we chose cubic regression splines as our smoothing functions. Refer to Appendix C for a description of spline functions as well as a brief introduction on Generalized Linear Models. For more on GAMs, [Woo17] is a great introductory reference with practical examples and R programming software implementation.

RSV

Model	AIC	R^2
Linear (specific humidity)	7358.748	0.6733989
Linear (temperature)	7372.274	0.6724792
Linear (specific humidity and temperature)	7359.727	0.6725789
Multivariate logistic model	7316.830	0.6725615
GAM	7092.559	0.6693111
GAM with splines*	6951.189	0.6713978

Influenza A

Model	AIC	R^2
Linear (specific humidity)	4430.835	0.1993223
Linear (temperature)	4381.874	0.2136904
Linear (specific humidity and temperature)	4377.648	0.2139728
Multivariate logistic model	4376.335	0.2157230
GAM	3962.665	0.3336227
GAM with splines*	3911.128	0.350223

* is the chosen model

Table 4.1: Regression of climate variables on onset week. The model includes location fixed effect to control for spatial heterogeneities in onset determinants.

For both RSV and Influenza A the smooth terms of the models were found to be statistically significant ($p < 0.005$). The Influenza A model accounts for 35% of the variance while the RSV model accounts for 67% of the variance. This suggests that climate covariates might be stronger drivers of RSV activity. Figures 4.1 and 4.2 show the marginal effects plots of the predicted proportion of positive RSV and Influenza A cases in Santiago. We find strong evidence of impact of both specific humidity and temperature test-positivity, with highest positivity rates observed for very low values. For both RSV and Influenza A the estimated effect of climate covariates suggests increased activity during cold and dry environments. For Influenza A this is in accordance to laboratory experiments where transmission was more efficient at lower temperature and humidity [LMSP07]. They also found that lowering tem-

perature affects the kinetics of viral shedding in inoculated guinea pigs. In summary, these results suggest that both specific humidity and temperature are important drivers of disease activity in Santiago, with peak activity occurring at low levels.

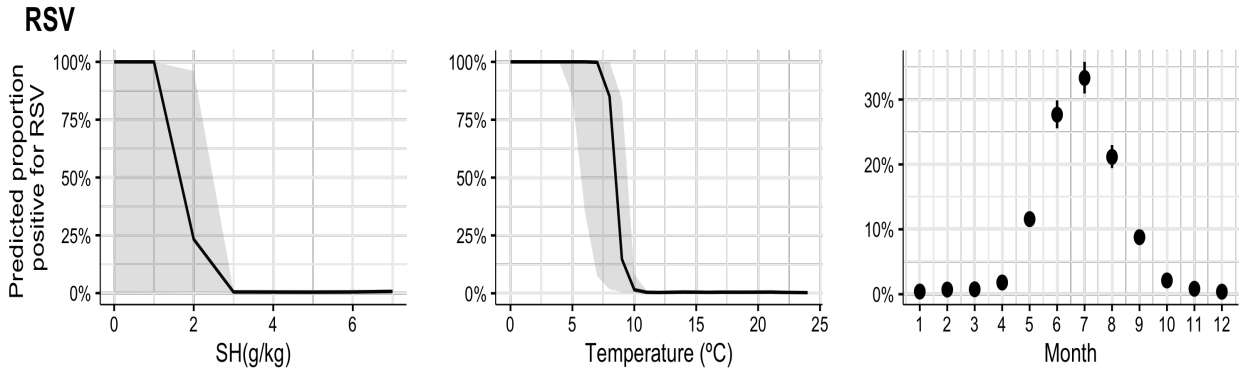

Figure 4.1: Marginal effects plots showing the predicted proportion positive for RSV by **A.** Specific Humidity (g/kg), **B.** Temperature (°C), and **C.** Month.

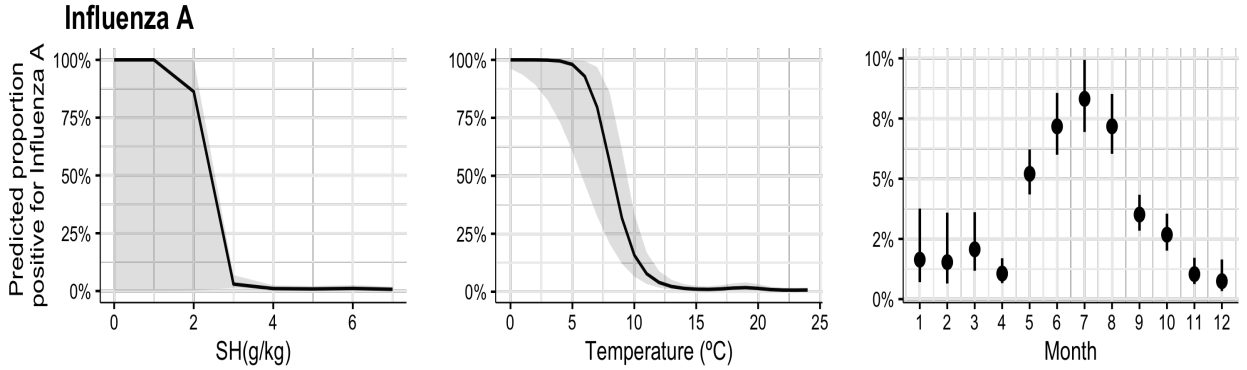

Figure 4.2: Marginal effects plots showing the predicted proportion positive for Influenza A by **A.** Specific Humidity (g/kg), **B.** Temperature (°C), and **C.** Month.

4.2 Implementation of POMP models

To further explore the hypothesis that climate covariates are the main drivers of seasonality and inter-annual variability, we construct mechanistic models that mimic disease transmission with (and without) climate covariates. Using the POMP modeling framework and iterated inference we aim to quantify the effect of climate covariates on disease transmission, and explore the models ability to reproduce the seasonal pattern and variation in the timing of disease activity in Santiago, Chile. Under the current climate change conditions, it is

important to estimate the burden of diseases. These models allow us to understand and project the dynamics of disease transmission under different climate change scenarios (e.g. continuously rising temperatures).

4.2.1 Deterministic MSIRS model: RSV

To explore the mechanistic relationship between climatic factors and seasonal variation of RSV incidence in Santiago, we implement a compartmental Susceptible-Infected-Recovered-Susceptible model with maternal induced immunity (MSIRS) based on [DKWP17, PVA+15]. The model follows the flow of the population in susceptible, infected (and infectious), and recovered compartments for seasonal RSV (Figure 4.3).

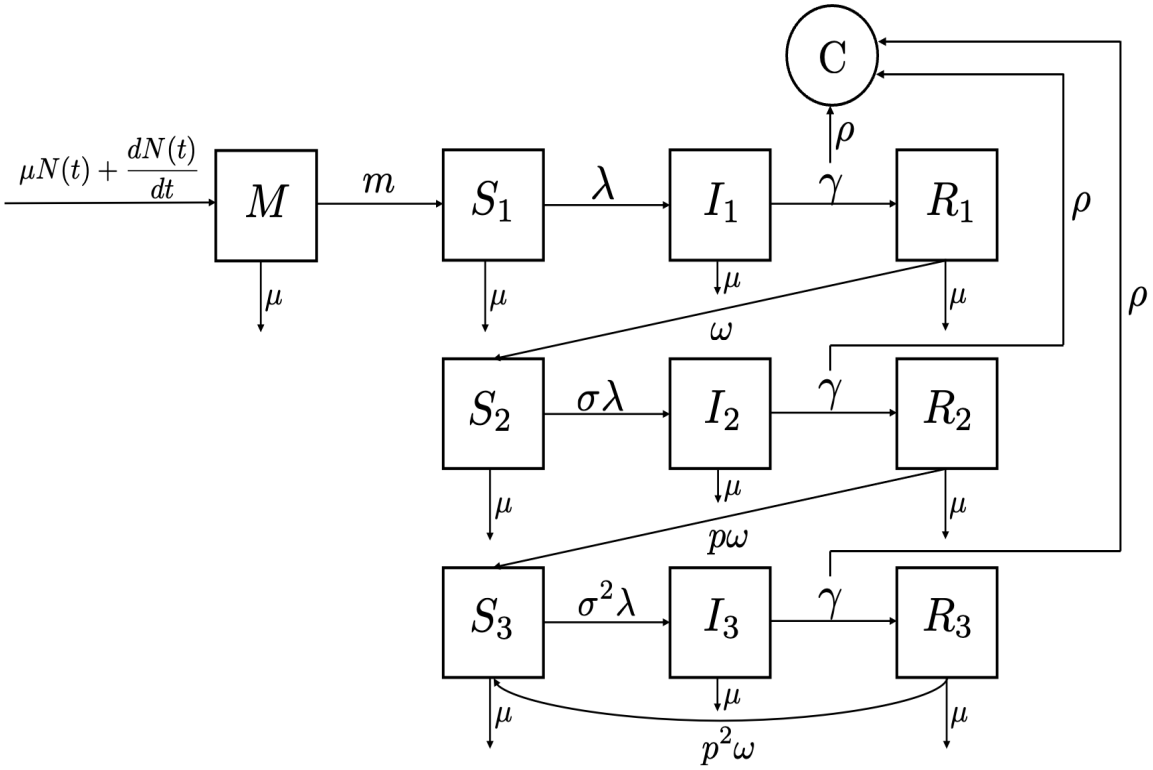

Figure 4.3: Compartmental flow diagram for RSV. The population is divided into four compartments: M, maternally derived immunity; S, susceptible; I, infected; R, recovered. Births enter M at the rate $\mu N(t) + \frac{dN(t)}{dt}$, and all individuals have a mortality rate $\mu = 0.02$, i.e., people live on average 50 years. After first infection the rate of subsequent infection reduces by a factor of σ. There is also progressive build up of immunity following two or more previous infections by a factor of p. Infections are counted upon transitions to the R compartments. The case reports, C, count newly infected individuals (deduced from the total number of infected I) with probability ρ.

We subdivide the population into 10 discrete compartments, stratified by protection from maternal immunity, (M), and repeated infections $(S(t)_{1,2,3}, I(t)_{1,2,3}, R(t)_{1,2,3})$. Maternal antibodies offer brief protection from RSV infection and thus individuals are born into the M population with protective maternal immunity [OSF+09]. Maternal immunity wanes over time at a rate m, leaving the infant susceptible to infection. $S(t)_{1,2,3}$, $I(t)_{1,2,3}$, and $R(t)_{1,2,3}$ denote the number of susceptible, infected, and recovered individuals, respectively, in the population at time $t \in \mathbb{R}^+$. The population size $N(t)$ changes over time and the total birth rate is quantified as $\mu N(t) + \frac{dN(t)}{dt}$ to reproduce the observed population increase over time. People live on average 50 years, and thus the mortality rate μ is fixed and equal to 0.02 for all compartments.

Following the first infection the model takes into account partial immunity from previous exposure to the pathogen through a reduction in susceptibility following the first infection, which reduces the rate of subsequent infection by a factor of σ, consistent with epidemiological studies and previous models of RSV transmission [PVA+15, WMG+07]. We assume that immunity from infection lasts for a limited period of time, $1/\omega$, after which the individual is again fully susceptible. There is also a progressive build up of immunity, by a factor of p, following two or more previous infections [Gle86, HWLS91, PVA+15]. The recovery rate and duration of infection is given by γ and $1/\gamma$, respectively. The period of immunity is given by $1/\omega$ and ω is the rate at which immunity is lost and recovered individuals move into the susceptible class. The deterministic transmission model is given by the following set of differential equations:

$$
\begin{aligned}
\frac{dM}{dt} &= \mu N(t) + \frac{dN(t)}{dt} - mM - \mu M, \\
\frac{dS_1}{dt} &= mM - \lambda S_1 - \mu S_1, \\
\frac{dI_1}{dt} &= \lambda S_1 - \gamma I_1 - \mu I_1, \\
\frac{dR_1}{dt} &= \gamma I_1 - \mu R_1 - \omega R_1, \\
\frac{dS_2}{dt} &= \omega R_1 - \sigma\lambda S_2 - \mu S_2, \\
\frac{dI_2}{dt} &= \sigma\lambda S_2 - \gamma I_2 - \mu I_2, \\
\frac{dR_2}{dt} &= \gamma I_2 - \mu R_2 - p\omega R_2, \\
\frac{dS_3}{dt} &= p\omega R_2 - \sigma^2\lambda S_3 - \mu S_3 + p^2\omega R_3, \\
\frac{dI_3}{dt} &= \sigma^2\lambda S_3 - \gamma I_3 - \mu I_3,
\end{aligned}
\tag{4.1}
$$

$$\frac{dR_3}{dt} = \gamma I_3 - \mu R_3 - p^2 \omega R_3,$$

The force of infection (or per capita rate at which susceptible individuals get infected) is given by

$$\lambda = \beta(t) \frac{I_1(t) + I_2(t) + I_3(t) + \iota}{N(t)},$$

where ι is the mean number of infectives visiting the population at any given time and $\beta(t)$ is the transmission rate. Since both RSV and Influenza A have the same transmission formulation, $\beta(t)$, we describe it in more detail in Section 4.2.5.

Given the equations for the deterministic model in (4.1), the number of newly infected individual accumulated in each observation time unit $[t_n, t_{n+1})$ is then given as

$$H_k(t_n) = \int_{t_n}^{t_{n+1}} \lambda S_k(t) dt.$$

Note this is a high dimensional integral that depends on all state variables M, $S(t)_{1,2,3}$, $I(t)_{1,2,3}$, and $R(t)_{1,2,3}$.

4.2.2 Stochastic MSIRS model: RSV

Recall that we are going to be using the POMP modeling framework and thus our model needs to satisfy the Markov property. While deterministic SIR models, like the one in Section 4.2.1, possess the Markov property the state of the system is fully determined by the initial conditions of the system and thus there is no randomness in the process. This means that deterministic models do not capture the inherent stochasticity of infectious disease transmission, which can result in significant variability in disease outcomes. To address this limitation, we can add stochasticity to the models and build the models by using continuous-time Markov chains. A continuous-time Markov chain (CTMC) can capture the initial disease dynamics and integrate the stochasticity involved in the disease transmission process [TB17]. They are particularly well-suited for modeling epidemics, which are inherently stochastic processes [YC11, MMM19, XZH22].

In a CTMC, the state of the system at any given time is described by a set of possible states, and the transitions between those states are governed by transition rates. These rates can be thought of as the probabilities per unit time that a transition will occur, given that the system is in a particular state. To apply this framework to modeling an epidemic, we define the states of the system to represent the susceptible, infected, and recovered states, and the transition rates to represent the probabilities of moving between those stages (e.g.

transmission rate, the recovery rate, and rate of waning immunity).

In the following the notation is adopted form King *et al.* [KDI16] and Stocks *et al.* [SBH20]. Let $N_{AB}(t)$ denote a stochastic counting process, which counts the number of individuals which have moved from compartment A to compartment B during the time interval $[0, t)$ with $A, B \in \mathcal{X}$, where $\mathcal{X} = \{M, S_1, I_1, R_1, S_2, I_2, R_2, S_3, I_3, R_3\}$ contains all compartments of the model. $N_{\cdot A}(t)$ counts the number of births and $N_{A\cdot}(t)$ the number of deaths in the respective compartment up until time t. The infinitesimal increment probabilities of a move between compartments connected by an arrow in Figure 4.3 fully specify the continuous time Markov process describing disease transmission. The number of individuals changing compartment in an infinitesimal time interval $\tau > 0$ is given by $\Delta N_{AB}(t) = N_{AB}(t + \tau) - N_{AB}(t)$. Then the model in Figure 4.3 can be defined by the following system of transitions rates:

$$
\begin{aligned}
\mathbb{P}[\Delta N_{\cdot M}(t) = 1 | \mathcal{F}_t] &= \mu N(t)\tau + \frac{dN}{dt}\tau + o(\tau) \\
\mathbb{P}[\Delta N_{MS_1}(t) = 1 | \mathcal{F}_t] &= mM(t)\tau + o(\tau) \\
\mathbb{P}[\Delta N_{S_1 I_1}(t) = 1 | \mathcal{F}_t] &= \lambda S_1(t)\tau + o(\tau) \\
\mathbb{P}[\Delta N_{I_1 R_1}(t) = 1 | \mathcal{F}_t] &= \gamma I_1(t)\tau + o(\tau) \\
\mathbb{P}[\Delta N_{R_1 S_2}(t) = 1 | \mathcal{F}_t] &= \omega R_1(t)\tau + o(\tau) \\
\mathbb{P}[\Delta N_{S_2 I_2}(t) = 1 | \mathcal{F}_t] &= \sigma\lambda S_2(t)\tau + o(\tau) \\
\mathbb{P}[\Delta N_{I_2 R_2}(t) = 1 | \mathcal{F}_t] &= \gamma I_2(t)\tau + o(\tau) \\
\mathbb{P}[\Delta N_{R_2 S_3}(t) = 1 | \mathcal{F}_t] &= p\omega R_2(t)\tau + o(\tau) \\
\mathbb{P}[\Delta N_{S_3 I_3}(t) = 1 | \mathcal{F}_t] &= \sigma^2\lambda S_3(t)\tau + o(\tau) \\
\mathbb{P}[\Delta N_{I_3 R_3}(t) = 1 | \mathcal{F}_t] &= \gamma I_3(t)\tau + o(\tau) \\
\mathbb{P}[\Delta N_{R_3 S_3}(t) = 1 | \mathcal{F}_t] &= p^2\omega R_3(t)\tau + o(\tau) \\
\mathbb{P}[\Delta N_{A\cdot}(t) = 1 | \mathcal{F}_t] &= \mu A(t)\tau + o(\tau) \quad \text{for} \quad A(t) \in \mathcal{X}
\end{aligned}
$$

with the filtration $\mathcal{F}_t = \{M(u), S_1(u), I_1(u), R_1(u)(u), S_2(u), I_2(u), R_2(u), S_3(u), I_3(u), R_3(u), \forall 0 \leq u \leq t\}$ denoting the history of the process until time t. The little o notation, $o(\tau)$, is taken to mean $\lim_{\tau \to 0} o(\tau)/\tau = 0$.

Note that the transition rates are related with the state variables in the following way:

$$
\begin{aligned}
\Delta M &= \Delta N_{\cdot M}(t) - \Delta N_{MS_1}(t) - \Delta N_{M\cdot}(t) \\
\Delta S_1 &= \Delta N_{MS_1}(t) - \Delta N_{S_1 I_1}(t) - \Delta N_{S_1\cdot}(t) \\
\Delta I_1 &= \Delta N_{S_1 I_1}(t) - \Delta N_{I_1 R_1}(t) - \Delta N_{I_1\cdot}(t)
\end{aligned}
$$

$$
\begin{aligned}
\Delta R_1 &= \Delta N_{I_1R_1}(t) - \Delta N_{R_1S_2}(t) - \Delta N_{R_1\cdot}(t) \\
\Delta S_2 &= \Delta N_{R_1S_2}(t) - \Delta N_{S_2I_2}(t) - \Delta N_{S_2\cdot}(t) \\
\Delta I_2 &= \Delta N_{S_2I_2}(t) - \Delta N_{I_2R_2}(t) - \Delta N_{I_2\cdot}(t) \\
\Delta R_2 &= \Delta N_{I_2R_2}(t) - \Delta N_{R_2S_3}(t) - \Delta N_{R_2\cdot}(t) \\
\Delta S_3 &= \Delta N_{R_2S_3}(t) + \Delta N_{R_3S_3}(t) - \Delta N_{S_3I_3}(t) - \Delta N_{S_3\cdot}(t) \\
\Delta I_3 &= \Delta N_{S_3I_3}(t) - \Delta N_{I_3R_3}(t) - \Delta N_{I_3\cdot}(t) \\
\Delta R_3 &= \Delta N_{I_3R_3}(t) - \Delta N_{R_3S_3}(t) - \Delta N_{R_3\cdot}(t)
\end{aligned}
$$

If we divide the equations by τ and take the limit as $\tau \to 0$, we obtain the deterministic version of the model seen in (4.1). The number of newly infected individuals accumulated in each observation time period $[t_n, t_{n+1})$ is then given by

$$
H_k(t_n) = N_{S_kI_k}(t_{n+1}) - N_{S_kI_k}(t_n),
$$

for $k \in \{1, 2, 3\}$. We are interested in numerically solving $N_{S_kI_k}$, and the same methods can also be applied to $N_{I_kR_k}$.

4.2.3 Deterministic SIRS model: Influenza A

Naturally occurring infection by influenza viruses is known to induce long-lasting protective immunity. Although it can be strain-specific, mainly because of the emergence of new influenza variants from the rapid antigenic evolution, and therefore recurrent infections can occur. With this in mind we will model the transmission dynamics of Influenza A as an SIRS model (Susceptible-Infected-Recovered-Susceptible), as seen in Figure 4.4.

$S(t)$, $I(t)$, and $R(t)$ denote the number of susceptible, infected, and recovered individuals, respectively, in the population at time $t \in \mathbb{R}^+$. The population size $N(t)$ changes over time and the total birth rate is quantified as $\mu N(t) + \dfrac{dN(t)}{dt}$ to reproduce the observed population increase over time. People live on average 50 years, and thus the mortality rate μ is fixed and equal to 0.02 for all compartments. The deterministic transmission model is given by the following set of differential equations:

$$
\begin{aligned}
\frac{dS}{dt} &= \left[\mu N(t) + \frac{dN(t)}{dt}\right] - \lambda S(t) - \mu S(t) + \omega R(t), \\
\frac{dI}{dt} &= \lambda S(t) - \gamma I(t) - \mu I(t), \\
\frac{dR}{dt} &= \gamma I(t) - \omega R(t) - \mu R(t),
\end{aligned}
\tag{4.2}
$$

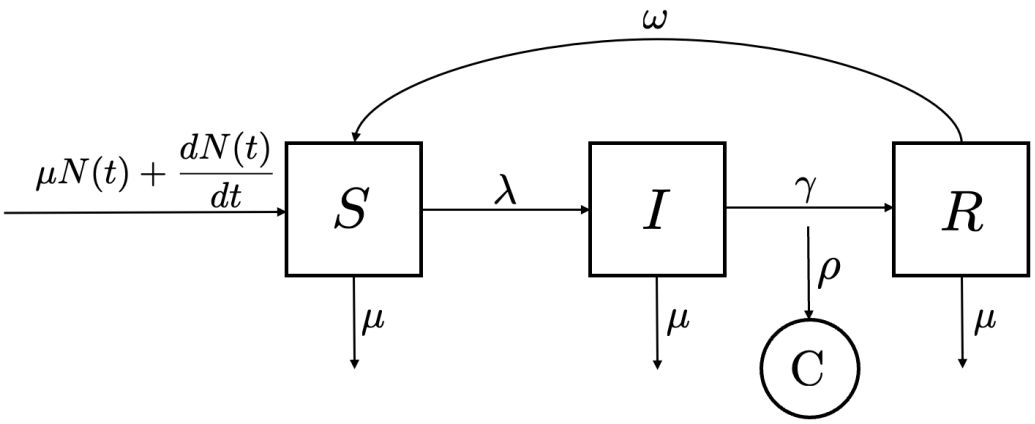

Figure 4.4: Compartmental flow diagram for Influenza A. The population is divided into three compartments: S, susceptible; I, infected; R, recovered. Births enter S at the rate $\mu N(t) + \frac{dN(t)}{dt}$, and all individuals have a mortality rate $\mu = 0.02$, i.e., people live on average 50 years. When immunity decreases, individuals move back to the susceptible population at rate ω. Infections are counted upon transitions to the R compartments.The case reports, C, count newly infected individuals (deduced from the total number of infected I) with probability ρ.

where ω is the rate at which immunity is lost and recovered individuals move into the susceptible class, thus the period of immunity is given by $1/\omega$. The recovery rate and duration of infection are given by γ and $1/\gamma$, respectively.

4.2.4 Stochastic SIRS model: Influenza A

Following the notation from Section 4.2.2, the model in Figure 4.4 can be defined by the following system of transitions rates:

$$\begin{aligned}
\mathbb{P}[\Delta N_{\cdot S}(t) = 1|\mathcal{F}_t] &= \mu N(t)\tau + \frac{dN}{dt}\tau + o(\tau) \\
\mathbb{P}[\Delta N_{SI}(t) = 1|\mathcal{F}_t] &= \lambda S(t)\tau + o(\tau) \\
\mathbb{P}[\Delta N_{IR}(t) = 1|\mathcal{F}_t] &= \gamma I(t)\tau + o(\tau) \\
\mathbb{P}[\Delta N_{RS}(t) = 1|\mathcal{F}_t] &= \omega R(t)\tau + o(\tau) \\
\mathbb{P}[\Delta N_{A\cdot}(t) = 1|\mathcal{F}_t] &= \mu A(t)\tau + o(\tau) \quad \text{for} \quad A(t) \in \mathcal{X}
\end{aligned}$$

where $\mathcal{X} = \{S, I, R\}$, and the filtration $\mathcal{F}_t = \{S(u), I(u), R(u), \forall \ 0 \leq u \leq t\}$ denoting the history of the process until time t. Note that the transmission rates are related with the

state variables in the following way:

$$\begin{aligned}\Delta S &= \Delta N_{\cdot S}(t) - \Delta N_{SI}(t) - \Delta N_{S\cdot}(t) + \Delta N_{RS}(t)\\ \Delta I &= \Delta N_{SI}(t) - \Delta N_{IR}(t) - \Delta N_{I\cdot}(t)\\ \Delta R &= \Delta N_{IR}(t) - \Delta N_{RS}(t) - \Delta N_{R\cdot}(t)\end{aligned}$$

If we divide the equations by τ and take the limit as $\tau \to 0$, we obtain the deterministic version of the model seen in (4.2).

4.2.5 Transmission term $\beta(t)$

As discussed in Section 3.3.1, RSV and Influenza A are highly seasonal and thus we incorporate seasonal periodic forcing through the transmission term $\beta(t)$ as follows

$$\beta(t) = \exp\left[\sum_{i=1}^{4} b_i s_i + b_q \Delta q s_3 + b_T \Delta T s_3\right]\frac{d\Gamma}{dt}.$$

The transmission term $\beta(t)$ includes 4 periodic functions of time to incorporate the seasonality through 4 cubic b-splines s_i (seen in Figure 4.5) and their respective coefficients b_i, as well as the interannual effect of climate through Δq and ΔT with their coefficients b_q and b_T as seen in [MKY+16]. More on splines can be found in Appendix C.

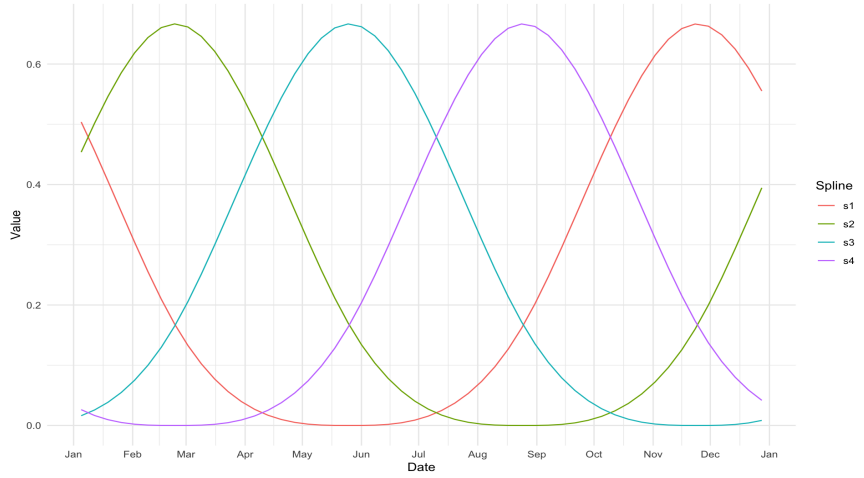

Figure 4.5: Four cubic periodic b-splines. Note the localization of the third spline, s_3 (blue), during the winter season.

The effect of climate forcing enters at two different time scales, seasonal and interannual,

in the expression $b_i s_i + b_q \Delta q s_3 + b_T \Delta T s_3$. The seasonal component quantified by this term can be interpreted as the average influence of the climate driver, which includes specific humidity and temperature. Note that the terms $b_T \Delta T$ and $b_q \Delta q$ are part of the spline s_3, this is because s_3 is considered the winter season spline, i.e., peak of the spline is during the winter months as seen in Figure 4.5. The terms ΔT and Δq are given by

$$\Delta T = \text{pmax}(0, T_c - \bar{T}),$$

and

$$\Delta q = \text{pmax}(0, q_c - \bar{q}),$$

where T_c and q_c are the climate thresholds from Table 3.1. The terms $\bar{T}$ and $\bar{q}$ are the mean monthly values (from May to August) of temperature and specific humidity, respectively. The terms ΔT and Δq are normalized values (by year) and they range from 0 to 1, where values closer to 1 reflect more extreme winter conditions (i.e., lower temperature or specific humidity).

Lastly, we include process noise into $\beta(t)$ in the form of $\frac{d\Gamma}{dt}$. Oftentimes the model does not fully capture the fluctuations seen in the surveillance data. In order to have sufficient stochasticity in the model and capture drivers not covered by the model (i.e., stochastic variability absent in the climate covariate), we add environmental noise through a Gamma distribution Γ, where marginally

$$\Gamma(t+\tau) - \Gamma(t) \sim \text{Gamma}(\frac{\tau}{\alpha^2}, \alpha^2)$$

is the increment of an integrated Gamma white noise process with intensity α. That is, $\Gamma(t+\tau) - \Gamma(t)$ has mean τ and variance $\tau\alpha^2$. The resulting process is overdispersed and converges (as τ goes to zero) to a well-defined process. Note that choosing the noise process $\Gamma(t)$ in a way such that its increments are independent, stationary, non-negative, and unbiased allows us to preserve the Markov property as shown in [BHIK09].

4.2.6 Measurement model

Following the formulation of the process model we need to build the measurement (or observation) model to finish the POMP modeling framework. Many times when dealing with surveillance data there is a lot of variability that cannot be captured under the assumed model. To overcome this, we can introduce overdispersion into the measurement model by choosing a distribution where the variance can be adjusted separately from the mean. With

this in mind we model the reported cases as realizations of the negative binomial distribution

$$C = Y_n \sim \text{NegBin}\left(\rho H(t_n), \frac{1}{\psi}\right),$$

where $H(t_n)$ is the true number of accumulated incidences per time unit $[t_n, t_{n+1})$ and $\text{NegBin}(\nu, \frac{1}{\psi})$ with $\psi > 0$ denotes the negative binomial distribution with mean ν and variance $\nu + \psi\nu^2$. Here, the parameter ρ is the reporting rate. For other types of measurement models for incidence data please refer to [Sto19].

4.3 Model implementation procedure

In this section we aim to give an overview of our model implementation process to find the maximum likelihood estimate (MLE) given a particular set of the parameters. We use the `pomp` package, available in the R programming language, to perform inference on our POMP models. The model is implemented by specifying (i) its state process and (ii) measurement process to then perform maximum likelihood estimation via the iterated filtering algorithm in order to make parameter inference. There are multiple examples and tutorials available online on how to use the `pomp` package, some of them can be found in [KI22, KIMB+22].

The core of the simulation algorithm is to define a POMP object. The POMP object is constructed by defining the state process as our stochastic transmission model and the measurement process as the negative binomial distributed measurement model of the number of newly infected individuals. We use an Euler approximation with a time step size of 1 day to simulate the state process model. We must also define initial values for our parameters (i.e., initial conditions of state variables and transmission rates).

To generate realizations from the state process we use the τ-leap algorithm which is based on the Gillespie algorithm [KR08]. The τ-leap algorithm holds all rates constant in the chosen interval τ and simulates the number of events that will occur in the interval $(t, t+\tau)$. The state variables are the updated and the process is repeated until a stopping time is reached [KDI16]. The number of individuals leaving any of the states in the time interval is then multinomially distributed. The simulation time step τ is chosen to be 1 day.

We begin our simulations with 10,000 values drawn uniformly from a hypercube which contains biologically reasonable values for all parameters and $S(0)+I(0)+R(0) = N(0)$. For the first inference procedure (Algorithm 2) we use `Nmif`=50 iterations, `Np`=1000 particles, a cooling of the perturbations of `cooling.fraction.50`=0.5 (i.e., every 50 iterations we reduce the perturbation by 50%) and random walk standard deviations `rw.sd` of 0.03 for

all the parameters. For each of the outputs we run 5 particle filters (Algorithm 1), each with 1000 particles. From this we calculate the estimated mean of the log-likelihood and the standard error of the Monte Carlo approximation for every parameter set. We then recursively iterate the inference procedure 3 times, each one with the output from the highest 10,000 log-likelihoods from the former inference procedures as the set of initial parameters. This first iterative process is called a global search.

Refinement	Loops	Cooling fraction	# Iterations	# Particles	# Particle filters
1	3	0.7	50	1000	5 with 1000 particles
2	10	0.4	100	2000	10 with 3000 particles
3	10	0.3	150	2000	5 with 2000 particles

Table 4.2: Refinements using the 500 highest log-likelihoods.

As the last step, we perform 3 refinements using the 500 highest log-likelihoods as the initial parameters. For each refinement we decrease the cooling fraction and increase the number of iterations as seen in Table 4.2. We perform multiple loops for each of the starting parameters. Even with advances in power computing this is still a very extensive and time consuming procedure. Because of time constrictions we only performed this procedure on 8 models fitted on the incidence data from Santiago, Chile. The next section provides the models implemented as well as parameter estimates from the iterated filtering procedure.

4.4 Parameter estimates and model selection

4.4.1 Models

In order to better understand the role of climate variables in the transmission of RSV and Influenza A, we tested 4 hypotheses, that can be understood as different variations of the transmission rate $\beta(t)$ described in Section 4.2.5. These models assume the presence/absence of the effect of the climate covariates in the disease transmission during the winter season (i.e., third spline) in the following way:

1. **No climate:** Transmission rate includes seasonality through the addition of 4 splines, but no interannual effect

$$\beta(t) = \exp\left[\sum_{i=1}^{4} b_i s_i\right] \frac{d\Gamma}{dt}.$$

2. **Only temperature:** Transmission rate includes both seasonal and interannual effect through the addition of temperature

$$\beta(t) = \exp\left[\sum_{i=1}^{4} b_i s_i + b_T \Delta T s_3\right] \frac{d\Gamma}{dt}.$$

3. **Only humidity:** Transmission rate includes both seasonal and interannual effect through the addition of specific humidity

$$\beta(t) = \exp\left[\sum_{i=1}^{4} b_i s_i + b_q \Delta q s_3\right] \frac{d\Gamma}{dt}.$$

4. **Mixed:** Transmission rate includes both seasonal and interannual effect through the addition of both temperature and specific humidity

$$\beta(t) = \exp\left[\sum_{i=1}^{4} b_i s_i + b_q \Delta q s_3 + b_T \Delta T s_3\right] \frac{d\Gamma}{dt}.$$

4.4.2 Model selection criteria

We base our criteria for model selection on Akaike Information Criterion (AIC) [Boz87]. It is a statistical measure for the comparative evaluation among models. AIC provides an estimation of the information lost when a specific model is used to represent the process that generated the data [PB19]. With such an approach, we can assess the trade-off between the goodness of the fit and the complexity of the models. A lower AIC indicates the model that fits better to the available data and is the one that explains the greatest amount of variation using the fewest possible independent variables (i.e., parameters).

Mathematically, the AIC is calculated by the following equation:

$$AIC = -2 \cdot \ell + 2 \cdot k$$

where k is number of estimated parameters and ℓ the log likelihood estimate from performing iterated inference on our POMP models. The AIC score takes into account model complexity and penalizes the likelihood based on the number of parameters.

The likelihood ratio test, also known as Wilks test, is then used to assess the goodness of fit of the selected model based on the ratio of their likelihoods [LBG11]. The likelihood-ratio test requires that the models be nested (i.e., model A is a special case of model B). With

a $p < 0.05$ significance level, we would fail to reject the null hypothesis, i.e., we should use model A and not model B.

To perform the likelihood ratio test, we use the log-likelihood of the models obtained from the iterated inference of POMP models. We calculate the likelihood ratio statistic, which is the ratio of the likelihoods of the two models, also expressed as a difference between the log-likelihoods

$$LRT = 2[\ell_{\mathrm{B}} - \ell_{\mathrm{A}}]$$

where ℓ_{B} is the log-likelihood of model B and ℓ_A is the log-likelihood of model A [Woo57]. The value of ℓ_{B} must be larger than or equal to that of ℓ_{A} because model A is a special case of model B. The likelihood ratio statistic follows a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters between the two models, under the null hypothesis.

4.4.3 Results

4.4.3.1 Best models

This section presents the results obtained by using the `pomp` package to perform simulation based inference applied to the incidence data of RSV and Influenza A from Santiago, Chile. Table 4.3 shows the log-likelihood and standard error obtained by using the iterated filtering algorithm, as well as AIC scores and likelihood ratio test significance. For RSV, the model best suited for our data is the one with a transmission rate that includes interannual effect through the addition of specific humidity

$$\beta(t) = \exp\left[\sum_{i=1}^{4} b_i s_i + b_q \Delta q s_3\right] \frac{d\Gamma}{dt}.$$

However, the likelihood ratio test shows a p-value of 0.06 when comparing the humidity and no climate models, indicating that the observed result is likely to occur under both models.

The best model for Influenza A is the one without climate, i.e., the transmission rate includes seasonality through the addition of 4 splines, but no interannual effect

$$\beta(t) = \exp\left[\sum_{i=1}^{4} b_i s_i\right] \frac{d\Gamma}{dt},$$

indicating that climate covariates do not drive the year-to-year variation in the transmission of Influenza A in Santiago.

RSV

Model	Log-likelihood	SE	No. of parameters	AIC	Likelihood ratio test
No climate	-1092.498	0.5249799	23	2230.996	$p = 0.06417736$
Temperature	-1091.672	0.2980107	24	2231.344	
Humidity*	-1090.785	0.4364656	24	2229.570	
Mixed	-1094.403	0.2733507	25	2238.806	$p = 0.0071456$

Influenza A

Model	Log-likelihood	SE	No. of parameters	AIC	Likelihood ratio test
No climate*	-779.1951	1.349879	14	1586.390	
Temperature	-781.7816	0.3777042	15	1593.563	$p = 0.02294057$
Humidity	-782.2934	0.8726984	15	1594.587	$p = 0.0127996$
Mixed	-783.0096	0.4298842	16	1598.019	$p = 0.02204874$

* is the chosen model with lowest AIC score

Table 4.3: Inference results for the four models with maximum likelihood estimate (MLE), Akaike information criterion (AIC), standard error of the Monte Carlo approximations, number of parameters, and likelihood ratio test.

We performed 1000 simulations to obtain a pointwise 95% prediction interval seen in Figures 4.6 and 4.7. The RSV incidence data is covered 75% of the time by the pointwise 95% prediction interval obtained from the simulations, while for Influenza it is covered 89% of the time. This indicates that both models explain the data well, successfully predicting when seasonal outbreaks will start and end. However, we acknowledge that the prediction intervals are larger for Influenza A, and that further research is needed to improve this fitting. One limitation is that all observations from the data are used to fit the model and thus the predictions from the realizations can produce overfitting. In Chapter 5 we give more details on how to improve the models to better capture the interannual variability and better predict the severity of seasonal outbreaks.

4.4.3.2 Parameter estimates

To obtain confidence intervals (CI) for the parameters we look at profiles of the likelihood. Profile likelihood is a statistical method used for estimating parameters in a model by maximizing the likelihood function while holding all other parameters fixed. In other words, one parameter is varied over a range of values, while the likelihood is maximized over all remaining parameters [HIK10]. We calculate a Monte Carlo error adjusted profile log-likelihood of each parameter [Ion05, KIMB+22, KDI16]. For each observation, a CI is estimated based on multiple realizations of the model evaluated at the MLE, where only maximal values are

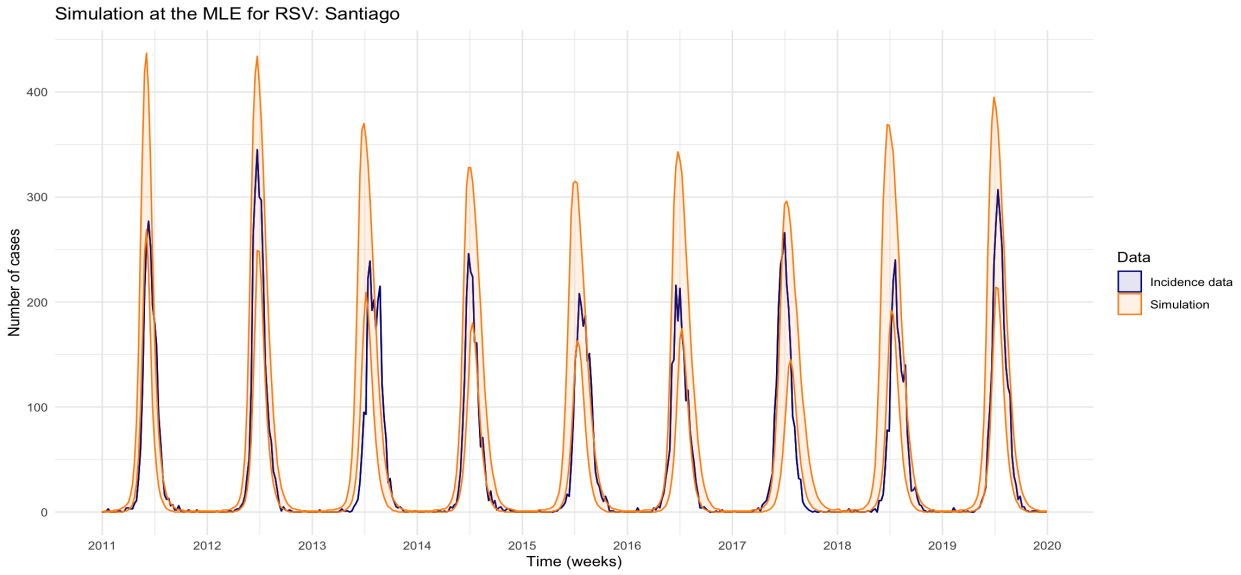

Figure 4.6: The 95% prediction interval (orange shading) for realizations of the MSIRS model (Figure 4.3 with humidity evaluated at the maximum likelihood estimate for the RSV data in Santiago (solid dark blue line). RSV incidence data is covered 75% of the time by the pointwise 95% prediction interval.

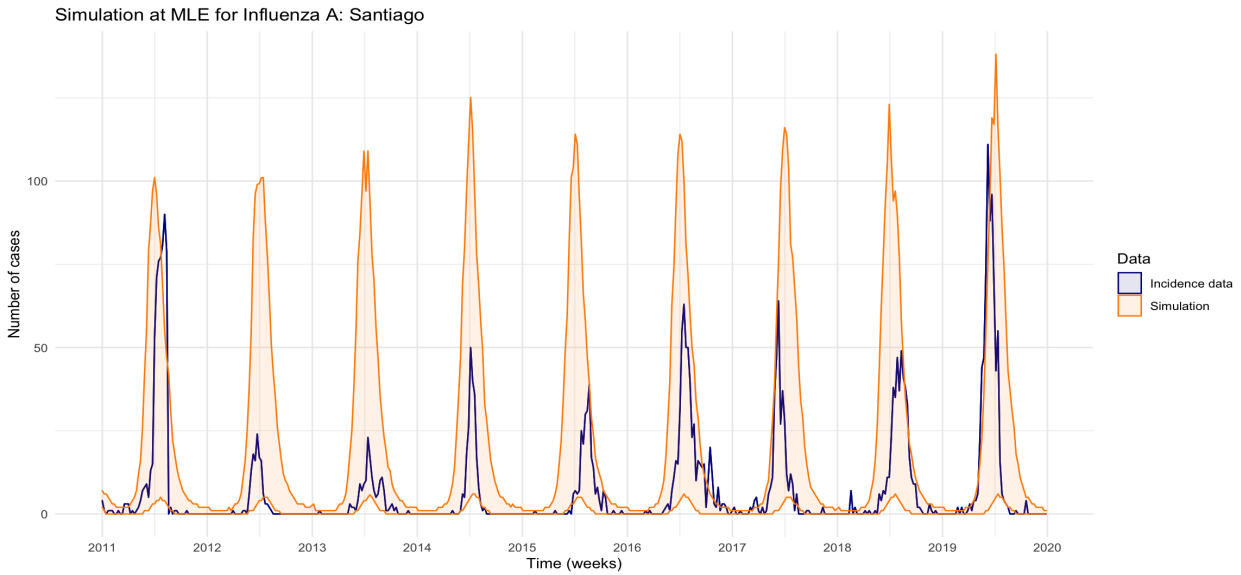

Figure 4.7: The 95% prediction interval (orange shading) for realizations of the SIRS model (Figure 4.4 without climate forcing, evaluated at the maximum likelihood estimate for the Influenza A data in Santiago (solid dark blue line). Influenza A incidence data is covered 89% of the time by the pointwise 95% prediction interval.

kept and plotted.

The graphs in Figures 4.8 and 4.9 show the profile likelihood curves for estimated parameters for RSV and Influenza A, respectively, from which confidence intervals are obtained.

RSV

Parameter	Description	MLE value	Confidence Interval
b_q	Specific humidity coefficient	-0.10	(-0.17, 0.01)
$1/\gamma$ (days)	Infectious period	4.4	(2.8, 4.9)
$1/\omega$ (days)	Waning immunity period	1	(1, 14)
$1/m$ (months)	Maternal immunity period	23.6	(18.1, 37)
$1/p\omega$ (days)	Waning immunity period after 2nd infection	3.7	
$1/p^2\omega$ (days)	Waning immunity period after 3 or more infections	20.9	
σ	Reduction in susceptibility	0.97	

Influenza A

Parameter	Description	MLE value	Confidence Interval
$1/\gamma$ (days)	Infectious period	1	(1, 1.1)
$1/\omega$ (days)	Waning immunity period	53	(21, 53.5)

Table 4.4: MLE value and confidence interval of each parameter from likelihood profiles.

The CI is given by the intersection of the likelihood profile curves (blue curve) and the horizontal line two log-likelihood units below the MLE (orange dashed line) for each parameter. The MLE and confidence interval of each parameter is provided in Table 4.4.

While the MLE value for b_q is -0.10, this estimate is very small and has confidence intervals that cross 0. This result suggests that the impact of the interannual effect of the specific humidity on the transmission is not statistically significant. This assessment coincides with the results from the likelihood ratio test in Section 4.4.3, where the fit from the no climate model can still give us good results.

This approach also allowed us to estimate CIs for key epidemiological parameters that are relevant for the transmission of RSV and Influenza A. During the first 2 years of life, all children will have been infected at least once with RSV [Gle86]. The RSV model was able to capture the period of maternal immunity fairly well with a time of approximately 2 years and a confidence interval going from 1.5 years to 3 years. RSV is also said to provide progressive build up of immunity, following two or more previous infections [Gle86, HWLS91, PVA+15]. The estimate for p, the reduction in waning of immunity, was 0.0005, indicating that immunity is boosted after the second infection and third infection, providing longer periods of immunity (Table 4.4).

We can compare the infectious period ($1/\gamma$) and waning immunity period ($1/\omega$) of the two diseases. Biologically speaking we expect RSV to have a longer infectious period and Influenza A to have a higher waning immunity period. Our models were able to capture this behavior, RSV has an estimated infectious period of approximately 3 to 5 days and Influenza A has an estimated infectious period of 1 to 1.1 days. Note that these are underestimates,

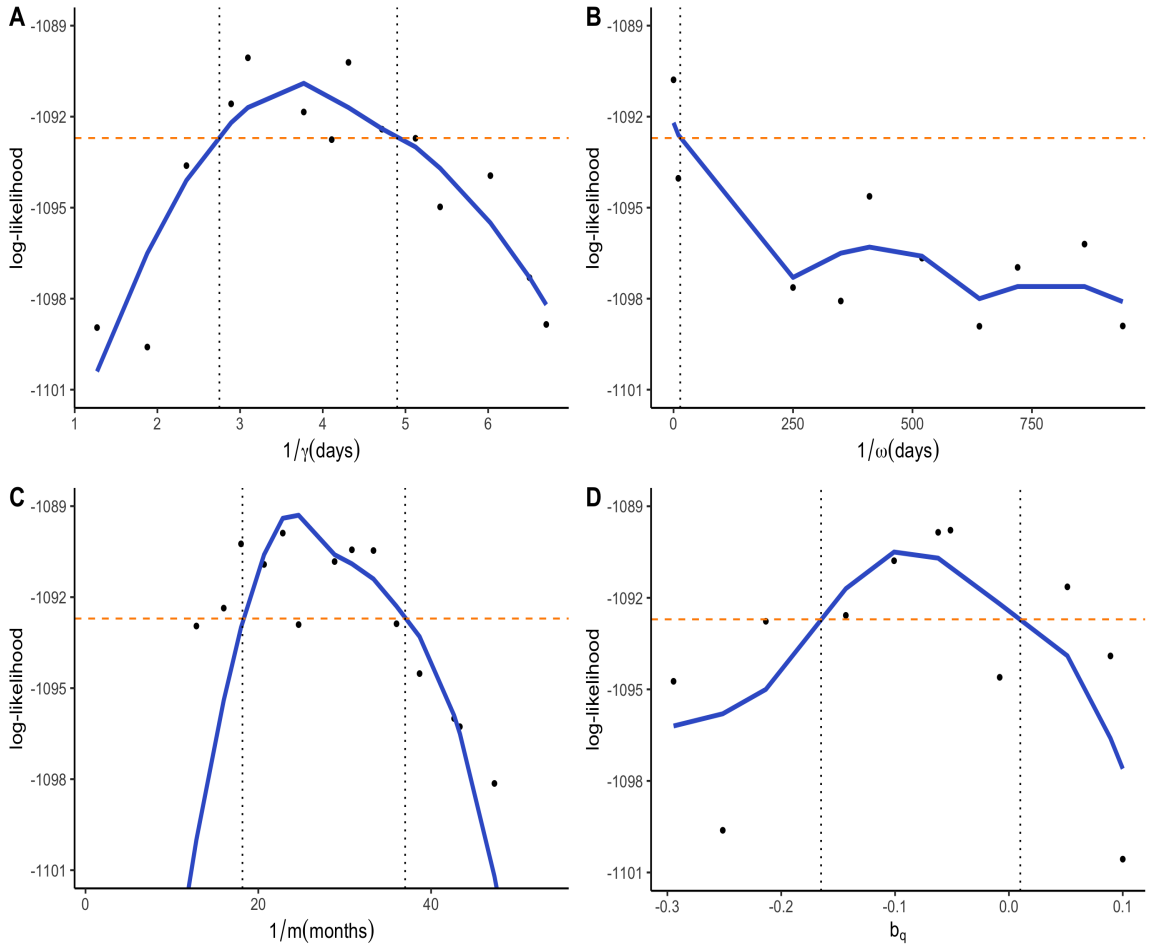

Figure 4.8: Profile likelihood curves for estimated parameters of the transmission model, with corresponding MLE and confidence intervals for RSV. **A.** Infectious period $1/\gamma$. **B.** Period of immunity $1/\omega$. **C.** Duration of maternal immunity. **D.** Specific humidity coefficient b_q.

RSV is said to have an infectious period of 5 to 10 days, and Influenza a period of 3 to 5 days [HDG76, MKA+15, TGW+10, CFR+09]. A possible explanation is that the model is capturing the time it takes for an individual to seek medical help after symptom onset, which may be arising from two different sources. First, we are calculating the newly infected cases based on individuals moving from the infected compartment to the recovered compartment, since it is assumed that individuals remain in isolation once they are confirmed infected. Second, the incidence data is obtained from a surveillance system of 42 primary care sentinel centers distributed among the 16 regions of the country. It also includes 31 public hospitals. These are part of the bigger hospital network across all of Chile. Our data is therefore biased towards people who seek treatment from these specific health centers.

This possible bias towards people who seek treatment is also captured by the estimated

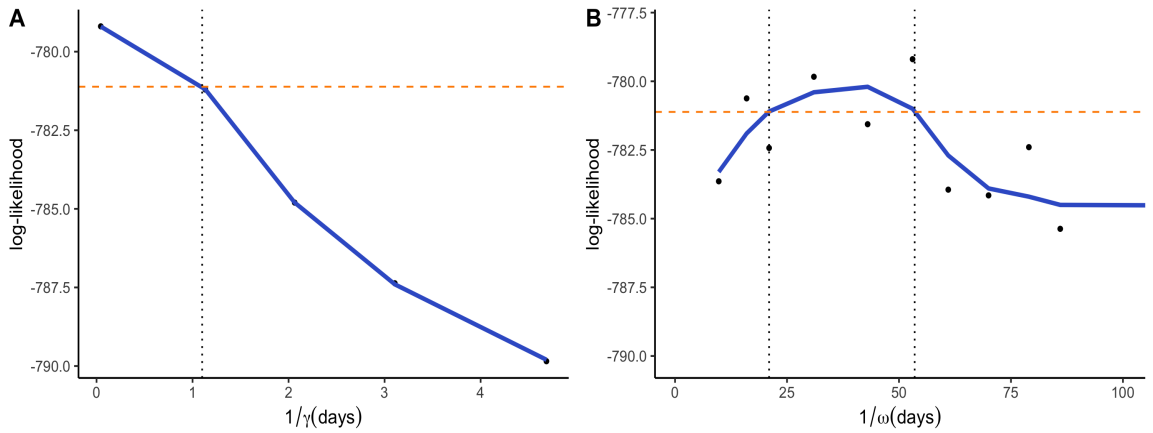

Figure 4.9: Profile likelihood curves for estimated parameters of the transmission model, with corresponding MLE and confidence intervals for Influenza A. **A.** Infectious period $1/\gamma$. **B** Period of immunity $1/\omega$.

values of the the reporting rate ρ. RSV had a ρ of 0.0019 (i.e., 0.19% of cases were reported) and Influenza A an estimated value of 0.0005 (i.e., 0.05% of cases were reported), meaning the data is more likely to be biased towards Influenza A. Because of differences in age of infection it is more probable that parents will seek care for their children than adults seeking care for themselves [Fre21]. Influenza vaccination may also impact the amount of people who get infected and subsequently tested. The quick antigenic changes in Influenza A require annual to semi-annual re-formulation of vaccines and vaccination campaigns[Tre16, HKK97, LC18]. The variable effectiveness of Influenza vaccines means that from year to year there are changes in well they protect against infection versus protection from severity, which depends on the individuals vaccine history, immune history, and age [LC18, PYM+21]. The fast evolution of this pathogen could also explain, at least in part, why interannual changes are not explained by changes in specific humidity or temperature.

In terms of the waning of immunity, this period was higher for Influenza A with a time of approximately 53 days and RSV with approximately 1 day. After the first infection RSV provides very little protection allowing RSV to re-infect the host throughout life [HWLS91, JBMC62, LSOC14]. No studies were found that give estimates for the period of waning immunity for RSV, and to our knowledge this is the first time the period of immunity for RSV is estimated. There are several reasons that could explain such a low estimate for the waning immunity period. First, we are using an MSIRS model when it may be that an MSIS (Susceptible-Infected-Susceptible) model might be more appropriate for modeling the transmission dynamics of RSV as done in [PVA+15]. Second, RSV is mostly considered an infant disease. It remains the most common cause of serious respiratory illness in children under 5 years of age, with an estimated 60,000 deaths worldwide [CZ22, LWB+22]. In fact,

in Chile, approximately 51% of cases are in children between 0 and 4 years of age [MLC+22]. Therefore, an age stratified model might be more appropriate for RSV transmission dynamics.

In contrast, Influenza A is known to provide immunity for approximately 1 to 2 years, but this is still very poorly understood [AT15]. In [PYM+21], a review of multiple studies in humans and animals, the authors argue that immunity from natural infection of Influenza does not prevent infection but reduces the intensity of subsequent infection. Fully estimating the provided immunity from natural infection of Influenza is very difficult, particularly because of frequent virus mutation to evade immunity, difference of immunity in multiple age groups, and dynamic immunity due to previous viral exposures (i.e., an individual may be immune to one strain of influenza for multiple years but not another) [PYM+21].

4.4.3.3 Hidden state variables

We construct stochastic compartmental models for RSV and Influenza A, and perform iterated inference to estimate transition rates only using the available surveillance data. Some of the capabilities of using the iterated inference methodology for partially observed Markov process models, developed by [IBK06], is that we can infer estimates of variables that are not observed. The POMP modeling and iterated filtering framework facilitates the simulation of sample paths that allow us to study the dynamics of the viruses. For example, the models constructed allow us to simulate not just the newly infected individuals, but also all state variables of the transmission process (i.e., $M, S_1, I_1, R_1, S_2, I_2, R_2, S_3, I_3, R_3$ for RSV and S, I, R for Influenza A). Figures 4.10 and 4.11 show the mean of 1000 simulations using the MLE of the parameters. Knowing the values and dynamic behavior of state variables allows us to understand how mitigation strategies (e.g. closing schools) impact the movement of individuals from the susceptible compartment to the infected compartment.

4.4.4 Summary of findings

- The best model for Influenza A is the one without climate indicating that climate covariates do not drive the year-to-year variation in the transmission of Influenza A.
- The model best suited for RSV is the one with a transmission rate that includes interannual effect through the addition of specific humidity, but upon inspection of the parameter estimate of b_q, we determine that the impact of the interannual effect of the specific humidity on the transmission of RSV is not statistically significant.

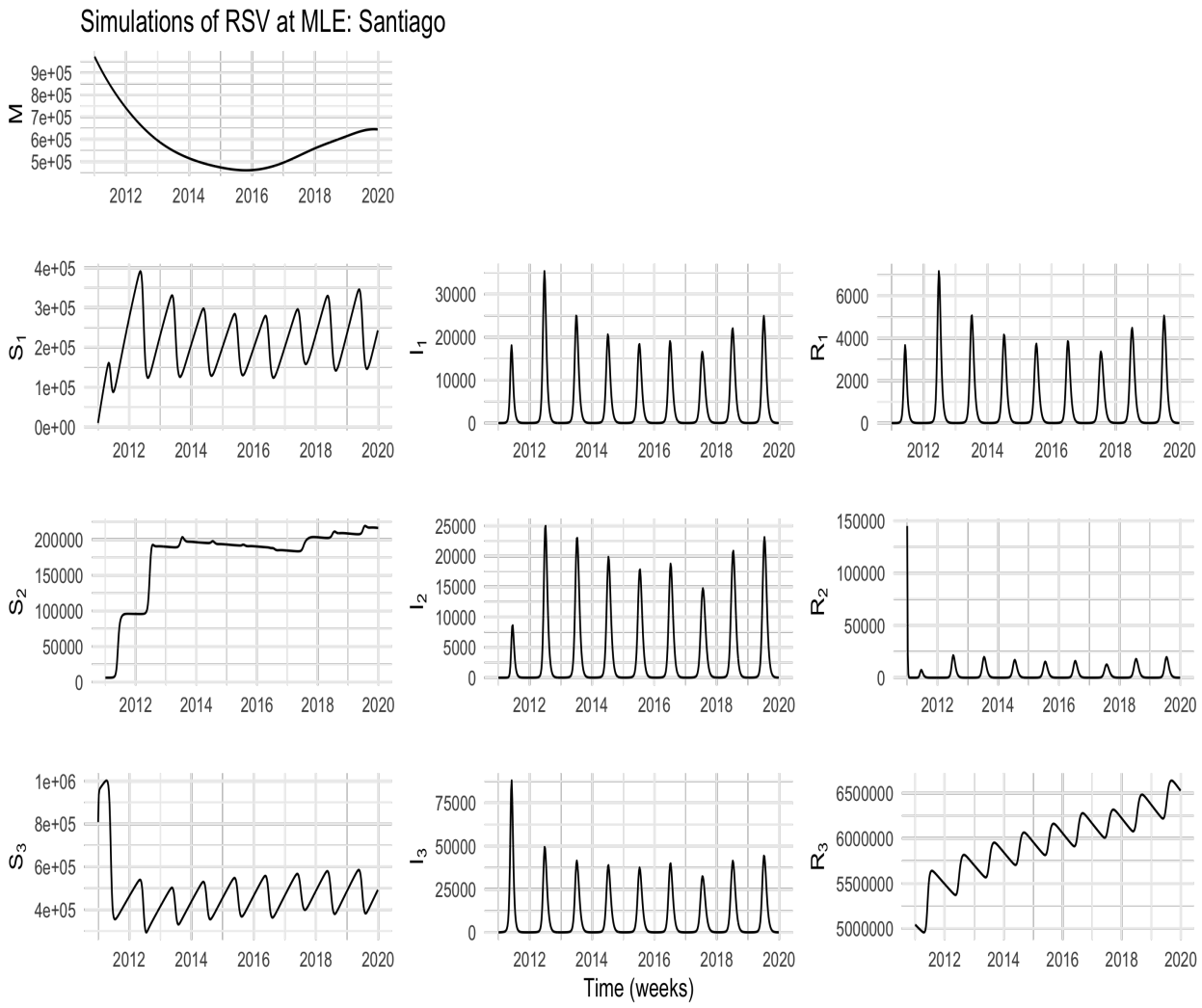

Figure 4.10: Mean of 1000 simulations of RSV MSIRS model without humidity at MLE.

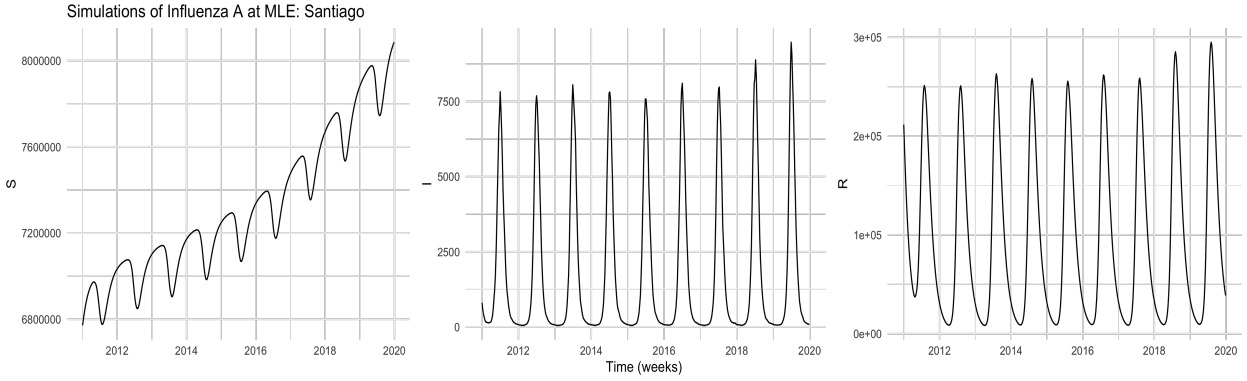

Figure 4.11: Mean of 1000 simulations of Influenza A SIRS model without climate forcing at MLE.

- The models successfully capture the seasonality, timing, and long term dynamics of RSV and Influenza A.
- Using iterated inference, we estimate key epidemiological parameters and find that RSV has a longer infectious period and Influenza A has longer periods of immunity, both of which concur with the differences in pathogen virology.
- No studies were found that give estimates for the period of waning immunity for RSV, and to our knowledge this is the first time the period of immunity for RSV is estimated using surveillance data. Furthermore, the model captures the expected progressive build up of immunity following two or more previous infections, as well as the period of maternal immunity.

4.4.5 Discussion

Modeling the mechanisms of infectious disease transmission can be a challenging process. There is a trade-off between having mathematically tractable models that only explain the general behavior of the disease, and more complex models that can better quantify and predict the dynamics of disease transmission but require heavy computational efforts [DH18]. Using stochastic compartmental epidemiological models we effectively model the transmission of RSV and Influenza A. We also estimate key epidemiological parameters and simulate disease incidence using the POMP modeling and iterated inference framework. While we believe that further research is needed to accurately predict parameter estimates, the model simulations are still able to capture the timing and seasonality of disease onset. Therefore, these results can be useful for making decisions on public health measures such as vaccination campaigns, administration of prophylactic antibodies, introducing social distancing guidelines, and other interventions aimed to reduce the spread of these viruses and minimizing their impact on public health. These models can also be used to forecast the timing of future outbreaks, but more data would be needed to do so.

Furthermore, our findings suggest that more analysis is still needed to disentangle the effect of interannual climate variability on the year-to-year variability in disease incidence. In the next chapter we present some of the modeling limitations that we encountered as well as improvements that can be made to our models to get more accurate and biologically plausible parameter estimates and further understand the impact of environmental drivers on disease transmission.

Chapter 5

Discussion

5.1 Concluding remarks

The hypothesis that climatic factors are the leading drivers responsible for the seasonal and interannual variation observed in respiratory viruses remains unproven. Laboratory experiments have shown that low absolute humidity (of which specific humidity is a measure) constrains both influenza virus survival and transmission [SK09]. Unfortunately for RSV, there are no similar studies, and thus existing models are built upon the assumption that RSV and Influenza A might behave similarly [BMW+19]. While this may be true in a closed laboratory setting, the results of this thesis indicate that while there is a correlation between low specific humidity (and temperature) and the increase of disease activity, it is not enough to determine a causal relationship in terms of the interannual variability observed in the amplitude of outbreaks. Furthermore, our results indicate that the nonlinear dynamics of the disease itself play a role in terms of the seasonality observed, but not the interannual variability of outbreak size. See Table 5.1 for a summary of findings. Nevertheless, our study provides a platform for understanding the impact of environmental factors on the transmission dynamics of RSV and Influenza A. Despite the fact that both viruses exhibit strong seasonal epidemics during the winter months, our results serve as evidence that more research is necessary to disentangle the associations between climate drivers, human behavior, age structure, connectivity across regions, and how they impact infectious disease dynamics.

5.2 Limitations and future directions

It is important to recognize the limitations of our study. First, SIR-type models are considered a simplification of how the transmission of viruses truly behaves. While we did include multiple parameters to make the system more realistic, for model simplicity we did not include the different epidemiological characteristics that different virus strains can have or the interactions between strains. As such some parameter estimates (e.g. recovery rate) may not

RSV	Feature	Influenza A
Chile		
✓	Outbreaks during the winter	✓
✓	Annual outbreaks	✓
✓	Climatic latitudinal gradient	✓
✓	Onset: North to South	–
–	Onset: South to North	✓
–	Synchronous onset	✓
✓	Significant association of onset and SH	–
–	Significant association of onset and temperature	–
✓	Climate factors explain interannual variation of onset	✓
–	Climate factors explain interannual variation of outbreak size	–
Santiago		
✓	Model captured timing of onset	✓
–	Model captures interannual effect of climate on outbreak size	–
✓	Model captures interannual variability of outbreak size	–
✓	Parameter estimates capture disease mechanisms	✓

Table 5.1: **Summary of findings.** A checkmark ✓ indicates the feature is true for the disease, the dash – indicates that the feature was not found. Orange text indicates different outcomes for RSV and Influenza A.

reflect the true epidemiological value, and instead, it might be capturing dynamical processes that were not explicitly modeled. Future research is necessary to explore the impact of co-circulation, specially for Influenza A, which is known to evolve quickly. Similarly, our models did not include several environmental factors that may shape RSV and Influenza A transmission dynamics, in particular, age, vaccination, administration of prophylactic antibodies, and seasonal changes in contact patterns. Because of the simplicity of the model, some values for the parameters found might not be realistic (i.e. biologically possible), because the algorithm is trying to find the optimal solution based on the information given. There can exist multiple maximums because the surface of the likelihood function has ridges, hence we often encountered values that were not biologically plausible, but were a maximum. In this case we need to constrain the starting values to improve the chances that the maximum that has the reasonable parameter values is reached. We overcome this issue by using confidence intervals to quantify the parameter range.

Second, all observations from the data are used to fit the model and thus the predictions from the realizations can produce overfitting. This results in a lack generalizability to other types of data. Cross-validation should be performed on the model to evaluate the model's performance on independent data. Estimates can also be validated by comparing them to other sources of data, such as hospital admission rates or mortality data. In fact, we can use

mortality data to enhance the model by connecting it to the latent state process through a different measurement process. Third, while we account for overdispersion in the data by using a Negative Binomial distribution in our measurement model, there may be other limitations to the data, such as underreporting or incomplete data, which may be biasing the results.

Finally, even though the findings from the exploratory analysis in Chapter 3 regarding climate drivers still hold for all of the regions, we are limited by the fact that the parameters for our model were only estimated for Santiago. Santiago is in the center of the country and has a more Mediterranean climate, characterized by hot, dry summers and cool, wet winters. In contrast, the southernmost region has oceanic climate (i.e. humid temperate climate with high rainfall) while the northern regions are extremely desert like and dry. The difference in climates could allows us to further explore the impact of climate covariates on the transmission. By fitting our model to all the regions in Chile, we can correlate the differences in the estimated climate forcing parameters to climatological differences for each region. Panel models are a useful tool for this scenario. Panel models consist of a collection of independent stochastic processes, generally linked through shared parameters while also having unit specific parameters. Each unit (in our case these are the regions) will have a partially observed Markov process model, as seen in Figure 5.1, and if all regions are modeled as independent, the panel model encompassing all regions is called a PanelPOMP [BIK20].

To calculate the MLE of a PanelPOMP, Bretó *et al.* conveniently developed a framework that is based on the iterated filtering approach of Ionides *et al.* described in Section 2.2.4 [IBK06]. The main idea is that a PanelPOMP model can be represented as a time inhomogeneous POMP model. The time series for each unit, corresponding to a latent POMP process, are concatenated in the following way

$$X(t) = X_u(t_{u,0} + (t - T_{u-1}^{\text{cum}}))$$

for $T_{u-1}^{\text{cum}} \leq t \leq T_u^{\text{cum}}$, where $\{X_u(t), t_{u,0} \leq t \leq t_{u,N_u}\}$ is the latent Markov process for unit u and T_u^{cum} is a cumulative latent POMP process time for all panel units up to unit u, given by

$$T_u^{\text{cum}} = u + \sum_{k=1}^{u} (t_{k,N_k} - t_{k,0})$$

and $T_0^{\text{cum}} = 0$. This concatenation converts the wide panel data into a long format, where every row represents an observation corresponding to a unit. Results from [BIK20] show that a POMP representation using a long format preserves the theoretical justification for iterated filtering. Therefore, similar to iterated filtering [INA+15, IBK06], the panel iterated filtering

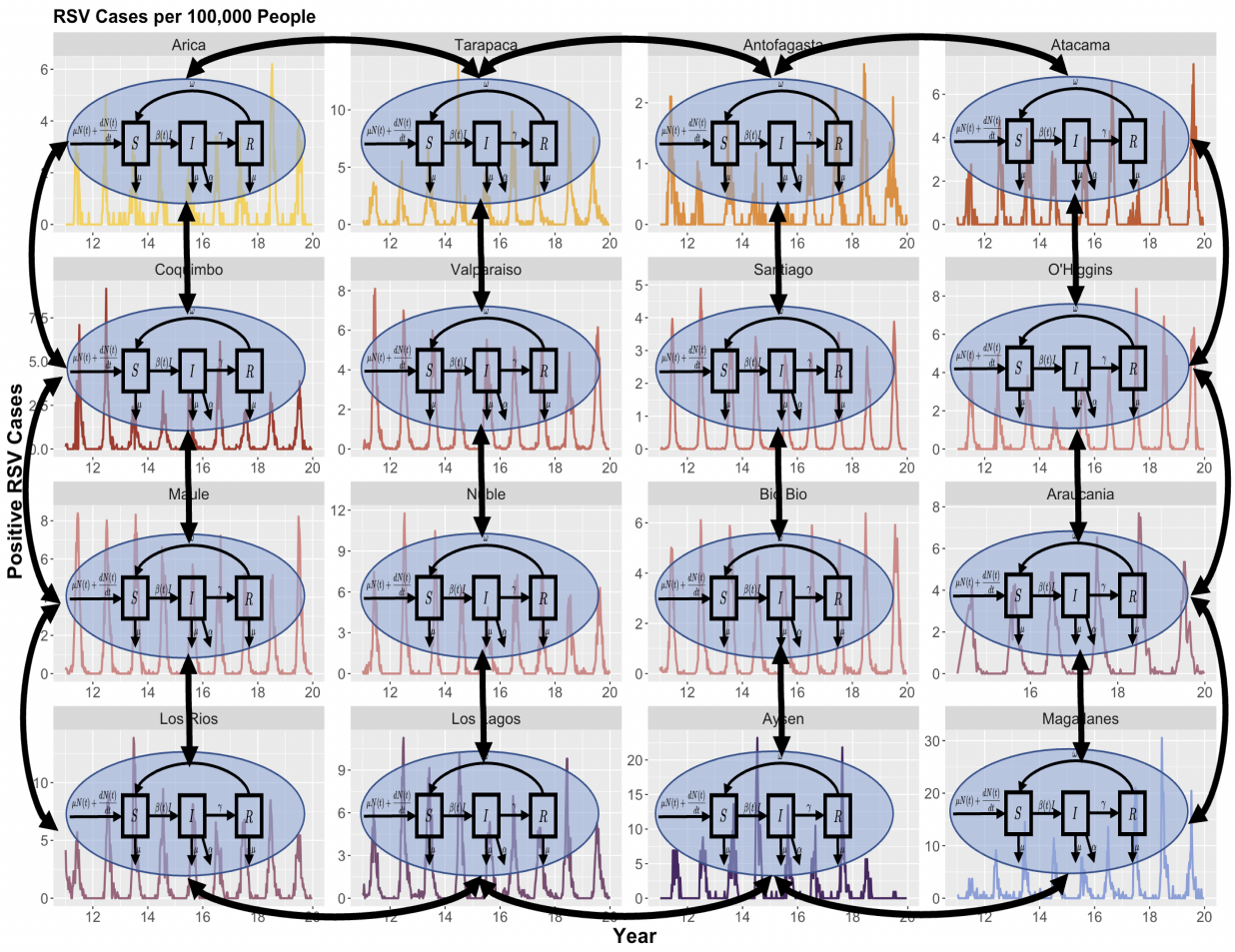

Figure 5.1: Diagram of PanelPOMP concept.

algorithm explores the space of unknown parameters by stochastically perturbing them and applying sequential Monte Carlo (Algorithm 1) to filter the data seeking for parameter values that are concordant with the data. Perturbations are successively diminished over repeated filtering iterations, leading to convergence to a maximum likelihood estimate [BIK20]. The relative performance of the POMP models we implemented tells us that we can use the same stochastic transmission processes from Sections 4.2.2 and 4.2.4 to construct a PanelPOMP model.

Acquiring additional covariate data can also improve our understanding of the seasonal and interannual variability of RSV and Influenza A. A study of RSV in the US, determined that temperature, vapor pressure, precipitation, and potential evapotranspiration were significantly associated with the timing of RSV activity across states [PVA+15]. The model presented in the study was able to replicate biennial patterns of RSV activity, however, they were not able fully explain why RSV activity begins in Florida, one of the warmest states

with long hot and humid summers and mild and wet winters. We observed a similar pattern for RSV in Chile, where timing of onset was earlier in the warmer regions in the North. However, the warmer regions in Chile usually have warm, dry summers and cold, dry winters (a direct contrast to Florida), which might explain why we observe only annual patterns of RSV activity.

As mentioned in the discussion of Chapter 3, RSV can be very age specific, where most children have been infected by 2 years of age. Older people are also known to be impacted by RSV because of their weakened immune response. Our model was fit to data aggregated across all age groups. It is possible to improve the model by incorporating age. In an age stratified SIRS model individuals are divided into different age groups, and the probability of transmission between age groups is calculated based on age-specific contact patterns. The model can be age-stratified as seen in Figure 5.2 and fit with data of hospitalization due to RSV, and related illnesses and patient age. Hospitalization from RSV can be defined by discharge diagnostic codes related to RSV. Fitting our model to the age-specific hospitalization data might give us better parameter estimates. These models can be used to inform public health strategies and interventions, such as vaccination programs targeted at specific age groups, specially when RSV vaccines are in the final stages of approval [HTK16, QXL+22]. Age stratified models can also be used for Influenza A to inform vaccination strategies since morbidity can be very high for very young and old individuals [HKK97, Tre16]. In fact, in [Hsi10] they show that the estimated per contact transmission probability of Influenza A in older people is significantly higher than that of any other age group due to close contacts with individuals from other age groups. However, their age-stratified model found evidence that targeting the very young and the very old for vaccination had very low impact on the overall transmissibility of the disease in the community. This adds to the theory Influenza A transmission is highly driven by human contact networks and connectivity across regions.

Our analysis for Influenza A shows spatial synchrony across regions. This behavior can arise from the connectivity and movement across regions. In fact, Chile displays high levels of internal migration and long-distance commuting attributed to mining and construction activities [RB20]. A study from the US has shown that the regional spread of Influenza A correlates more closely with rates of movement of people to and from their workplaces (workflows) than with geographical distance [VBS+06]. Changing our model to include mobility data and connectivity across regions would allow us to quantify the impact of human behavior in the spread of Influenza A in Chile. SIR-type models that include mobility data are a very useful tool that can help us determine how quickly the disease spreads from one location to another, and to identify areas that are at high risk of becoming infected. As seen with COVID-19, these models can be used to simulate the effects of various control measures,

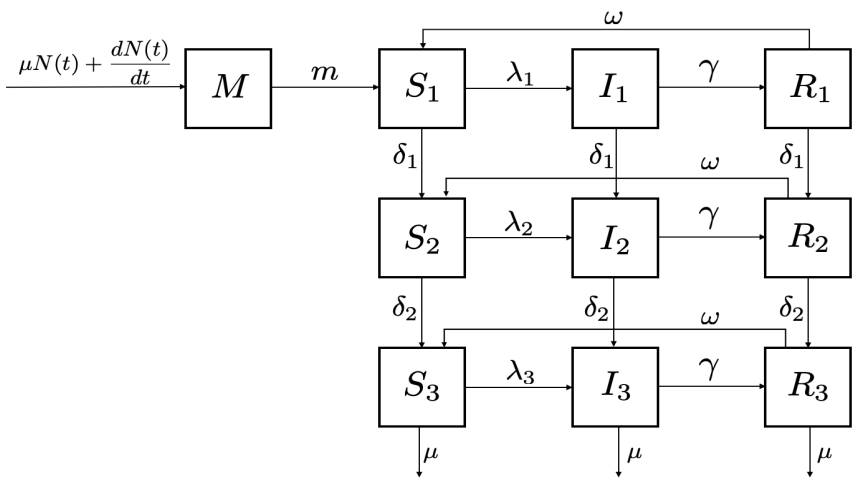

Figure 5.2: Flow diagram of age stratified SIRS model.

such as travel restrictions, social distancing, and lockdowns [KKM+21]. By incorporating these measures into the model, researchers can predict how effective they will be at controlling the spread of the disease. However, obtaining mobility data can be challenging due to privacy concerns, technical limitations, data fragmentation, and data availability [SRR22].

Furthermore, even though both viruses may seem similar in terms of symptoms, the virology and host response to RSV and Influenza A are very distinct. Naturally occurring infection by influenza viruses is known to induce long-lasting protective immunity, although it can be strain-specific, mainly because of the emergence of new influenza variants from the rapid antigenic evolution [NH07]. Recurrent infections therefore occur due to the antigenic variation of strains that spread over time. Predictive models that incorporate evolutionary change, like the one in [DKWP17], allow for accurate prediction ahead of seasonal outbreaks. The purpose of their model is the prediction of interannual disease risk (i.e. anomalously large or small outbreak amplitude) rather than timing of onset. The absence of strong interannual variability in our data exploration suggests, that climate covariates do not influence the amplitude of the outbreaks. By incorporating information on the evolutionary change of the virus we might be able to explain the year-to-year variability observed in our data. In contrast, reinfection with similar strains of RSV can occur multiple times throughout a persons life, especially in young children as discussed above, indicating that the immunity acquired by natural infection is limited and does not provide long-term protection against reinfection [APC18]. Vaccine efforts against Influenza also provide protection, but it is for a shorter amount of time, requiring yearly vaccination campaigns [Tre16]. Even with the improved development of vaccines, there are currently no effective vaccines available against RSV [HTK16, QXL+22]. Nevertheless, vaccine effectiveness can be incorporated into the

model to evaluate how different vaccines and vaccination strategies impact the dynamics of disease transmission when introduced into the population [FTY11, Hsi10, HKK97].

Finally, we could change the form of the transmission term $\beta(t)$. We chose periodic cubic splines because of their inherent flexibility, and included climate forcing through one of the splines. Other studies have used humidity-forced models, where $\beta(t)$ is given by the following expression based on

$$\beta(t) = \left(\frac{e^{-180q(t)}(R_{0_{max}} - R_{0_{min}}) + R_{0_{min}}}{\gamma} \right) \left[\frac{d\Gamma}{dt} \right]$$

where $q(t)$ is the specific humidity at time t, and $R_{0_{max}}$ and $R_{0_{min}}$ denote the maximum and minimum basic reproductive numbers, respectively [SK09, YLS15, YCLS15, DKWP17]. This equation was estimated from influenza virus survival and transmission in laboratory experiments, where it was found that low absolute humidity constrains both transmission efficiency and influenza virus survival [SK09]. This transmission term has been used to model Influenza transmission in temperate regions [SPV+10, SK12, DKWP17].

In tropical or subtropical locations, the annual seasonal pattern is less commonly observed. In [YKL+21], the authors show that modeling Influenza A transmission as monotonically decreasing with increasing absolute humidity is not sufficient to explain patterns in Influenza transmission in the tropics and subtropics. The study models the impact of absolute humidity using a parabola, where transmissibility is highest at very low and very high levels of absolute humidity. This relationship is also modified by temperature such that, when temperature is above some cutoff value, transmissibility is reduced. The transmission term is given by the equation

$$\beta(t) = \frac{\exp\left(aq^2(t) + bq(t) + c \right) \left[\frac{T_c}{T(t)} \right]^{T_{exp}}}{\gamma},$$

where $q(t)$ is specific humidity and $T(t)$ is temperature at time t. The assumption is that if T is below T_c, lower temperatures are able to further increase transmission, whereas temperatures above T_c inhibit Influenza transmission. The strength of this relationship is further determined by the exponent T_{exp}.

In summary, we have combined statistical and mathematical models to disentangle the impact of environmental drivers on the transmission of RSV and Influenza A. Using a wide range of statistical methods, we showed that both RSV and Influenza A have a sensitivity to climate specially in the context of seasonality, timing of onset, and timing in which they peak in Chile. Even though they are both sensitive their onset patterns are very different, with RSV starting in the North and Influenza A in the South of the country, highlighting

the inherent differences in pathogen virology. Furthermore, we observe a latitudinal gradient with respect to the climate covariates, and critical climate thresholds in which disease incidence starts to sharply increase, suggesting that both viruses can survive during very different winter conditions.

We also highlight the usefulness of mechanistic nonlinear dynamical systems that allow us to model the transmission of respiratory viruses. These mathematical models, in combination with the recently developed iterated filtering method, allow us to have a better understanding of the underlying dynamics of the system by estimating key epidemiological parameters. The models implemented in this thesis effectively capture the associations obtained from the statistical analyses and are able to predict the timing of outbreaks. Moreover, the results from our parameter inference and simulations further solidify our hypothesis that even though both viruses are sensitive to climate covariates with respect to seasonality and time of year in which they peak, the interannual variability observed in disease incidence is not modulated by the interannual variability of the climate covariates. Together, these findings suggest that while they do play a role, other factors like host demography, pathogen life history, and connectivity across regions might be stronger drivers of year-to-year variability in the size of outbreaks. Further work is needed to disentangle these associations, nevertheless mathematical models like the ones presented in this thesis can still provide useful information to guide public health policies.

Appendix A

Analysis of MSIRS deterministic model

A.1 Calculating R_0

Below we calculate the basic reproductive number, R_0 for the model in Section 4.2.1. Proposed by van den Driessche and Watmough, the next generation matrix method allows us to determine R_0 for an ODE compartmental model [DHM90, vW02, vW08, van17]. To build the next generation matrix, let $x = (x_1, x_2, \ldots, x_n)^T$ be the number of individuals in each compartment, where the first $m < n$ compartments contain infected individuals. Let $\mathcal{F}_i(x)$ be the rate of influx of new infections in compartment i, and $\mathcal{V}_i(x)$ the net transfer of other transitions between compartment i and other infected compartments, then consider the equations for each x_i to be written in the form

$$\frac{dx_i}{dt} = \mathcal{F}_i(x) - \mathcal{V}_i(x)$$

for $i = 1, 2, \ldots, m$. Define the Jacobian matrix of $\mathcal{F}_i(x)$ and $\mathcal{V}_i(x)$ as

$$F_{i,j} = \left.\frac{\partial \mathcal{F}_i(x)}{\partial x_j}\right|_{x=x_0} \qquad \text{and} \qquad V_{i,j} = \left.\frac{\partial \mathcal{V}_i(x)}{\partial x_j}\right|_{x=x_0}$$

for $1 \leq i,\ j \leq m$, where x_0 is the disease free equilibrium and $\frac{\partial}{\partial x_j}$ is the partial derivative with respect to x_j. Biologically one can think of the entries of matrix F as transmissions and the entries of matrix V as transitions. Then the matrix FV^{-1} has (i, j) entries equal to the expected number of secondary infections in compartment i produced by an infected individual introduced in compartment j. The matrix FV^{-1} is called the next generation matrix and R_0 is given by the spectral radius (i.e. largest eigenvalue) of FV^{-1}.

For our model, we consider the equations of infected compartments I_1, I_2, I_3:

$$\frac{dI_1}{dt} = \lambda S_1 - \gamma I_1 - \mu I_1,$$

$$\frac{dI_2}{dt} = \sigma\lambda S_2 - \gamma I_2 - \mu I_2,$$
$$\frac{dI_3}{dt} = \sigma^2\lambda S_3 - \gamma I_3 - \mu I_3,$$

where $\lambda = \beta(t)\dfrac{I_1 + I_2 + I_3 + \iota}{N}$. Then we have the following transmission matrix

$$F = \begin{pmatrix} \beta & \sigma\beta & \sigma^2\beta \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

and the transition matrix

$$V = \begin{pmatrix} \gamma+\mu & 0 & 0 \\ 0 & \gamma+\mu & 0 \\ 0 & 0 & \gamma+\mu \end{pmatrix}.$$

The inverse of V is given by

$$V^{-1} = \frac{1}{(\gamma+\mu)^3}\begin{pmatrix} \begin{vmatrix} \gamma+\mu & 0 \\ 0 & \gamma+\mu \end{vmatrix} & -\begin{vmatrix} 0 & 0 \\ 0 & \gamma+\mu \end{vmatrix} & \begin{vmatrix} 0 & \gamma+\mu \\ 0 & 0 \end{vmatrix} \\ -\begin{vmatrix} 0 & 0 \\ 0 & \gamma+\mu \end{vmatrix} & \begin{vmatrix} \gamma+\mu & 0 \\ 0 & \gamma+\mu \end{vmatrix} & -\begin{vmatrix} \gamma+\mu & 0 \\ 0 & 0 \end{vmatrix} \\ \begin{vmatrix} 0 & 0 \\ \gamma+\mu & 0 \end{vmatrix} & -\begin{vmatrix} \gamma+\mu & 0 \\ 0 & 0 \end{vmatrix} & \begin{vmatrix} \gamma+\mu & 0 \\ 0 & \gamma+\mu \end{vmatrix} \end{pmatrix} = \begin{pmatrix} \frac{1}{\gamma+\mu} & 0 & 0 \\ 0 & \frac{1}{\gamma+\mu} & 0 \\ 0 & 0 & \frac{1}{\gamma+\mu} \end{pmatrix}.$$

Multiplying F and V^{-1} gives us

$$FV^{-1} = \begin{pmatrix} \frac{\beta}{\gamma+\mu} & \frac{\sigma\beta}{\gamma+\mu} & \frac{\sigma^2\beta}{\gamma+\mu} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Solving the equation $\det(FV^{-1} - yI) = 0$, where I is the identity matrix, to find the eigenvalues of FV^{-1} we get

$$\det(FV^{-1} - yI) = (\frac{\beta}{\gamma+\mu} - y)(y^2) = 0 \implies y = \frac{\beta}{\gamma+\mu}, y = 0.$$

The largest eigenvalue is given by $y = \dfrac{\beta}{\gamma + \mu}$, therefore

$$R_0 = \frac{\beta}{\gamma + \mu}.$$

Appendix B

Extra figures

B.1 Lab-confirmed respiratory viruses in Chile

Figures B.1 and B.2 show the complete data set obtained from the Institute of Public Health (ISP) within the Health Ministry of Chile's Government [isp]. The data includes the number of weekly respiratory viruses for Adenovirus, Parainfluenza, Influenza A, Influenza B, RSV, and Metapneumovirus for each Region in Chile from 2011 to 2019.

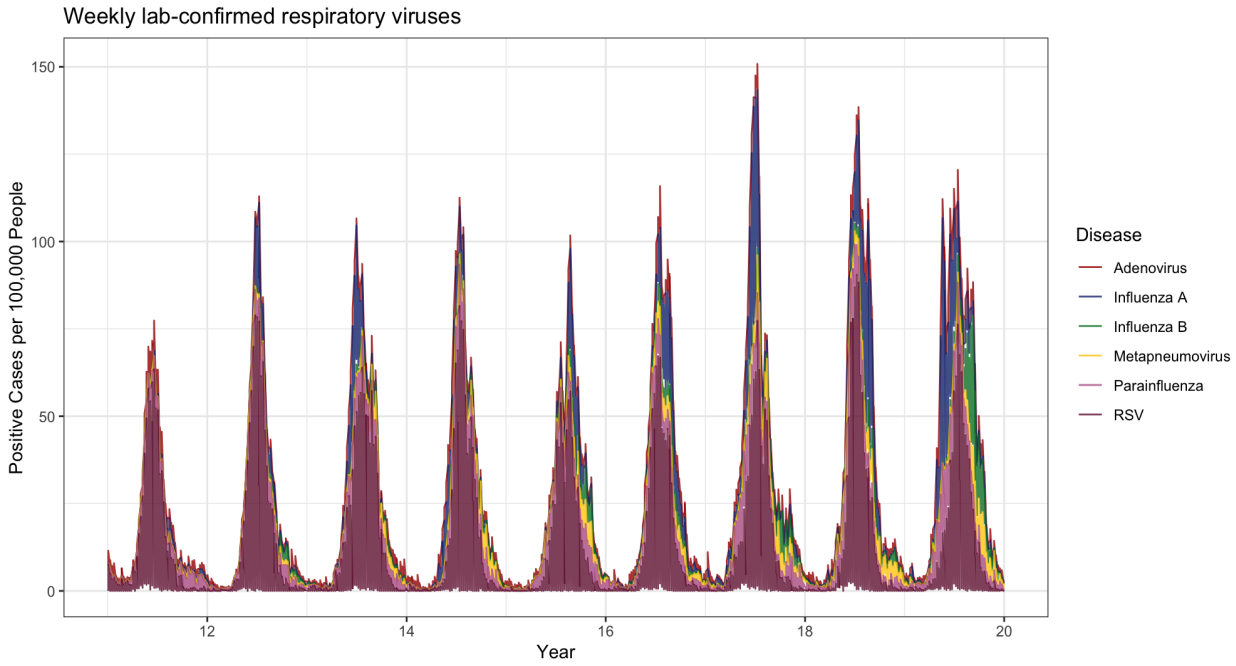

Figure B.1: Weekly lab-confirmed cases of Adenovirus, Influenza A, Influenza B, Metapneumovirus, Parainfluenza, and RSV per 100,000 people aggregated from all regions of Chile.

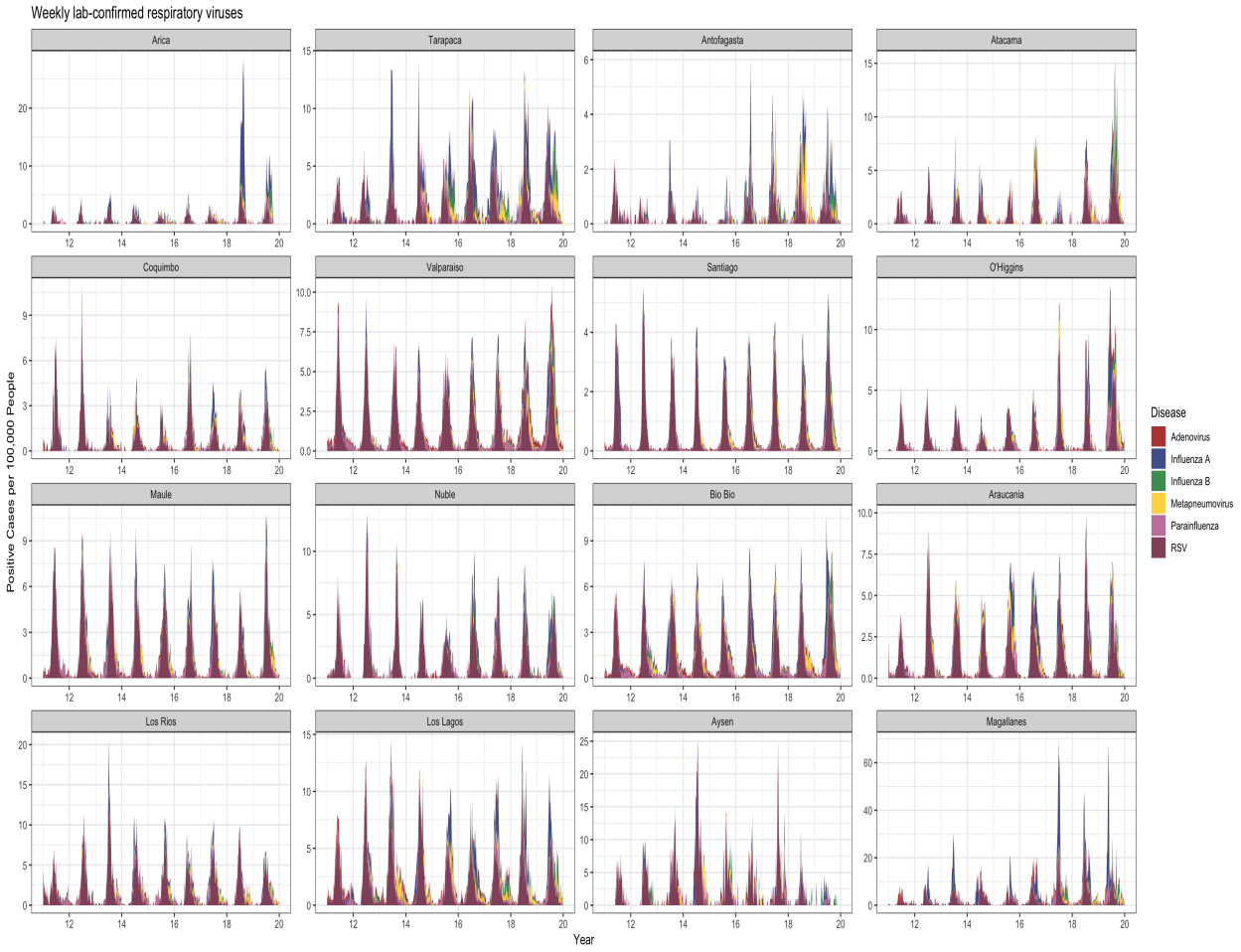

Figure B.2: Weekly lab-confirmed cases Adenovirus, Influenza A, Influenza B, Metapneumovirus, Parainfluenza, and RSV per 100,000 people aggregated by region.

B.2 Density plots: Specific humidity

Figure B.3 shows the density of cases within specific humidity ranges. Note that the ranges of specific humidity for which cases peak varies widely across regions. There is a latitudinal gradient with respect to specific humidity, going from north to south. The range of temperatures in Arica (North) is approximately (8 g/kg−12g/kg) while in Magallanes (South) it is approximately (2 g/kg− 5g/kg) for RSV. We observe the same pattern for Influenza A, where Arica has a range of (8 g/kg− 10g/kg) and Magallanes has a range of (2 g/kg −5g/kg).

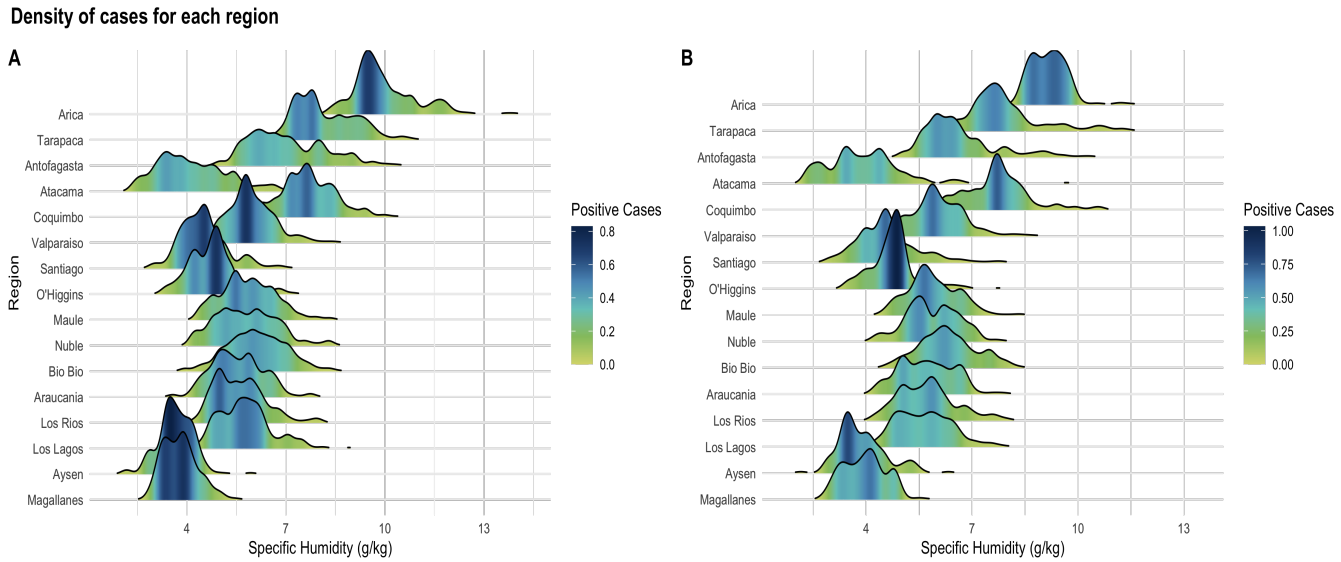

Figure B.3: Density plots with respect to specific humidity ranges **A.** RSV. **B.** Influenza A.

B.3 Segmented regression graphs

Graphed results of the segmented regression analysis for all 16 regions of Chile, giving us thresholds for both temperature (T_c) and specific humidity (q_c).

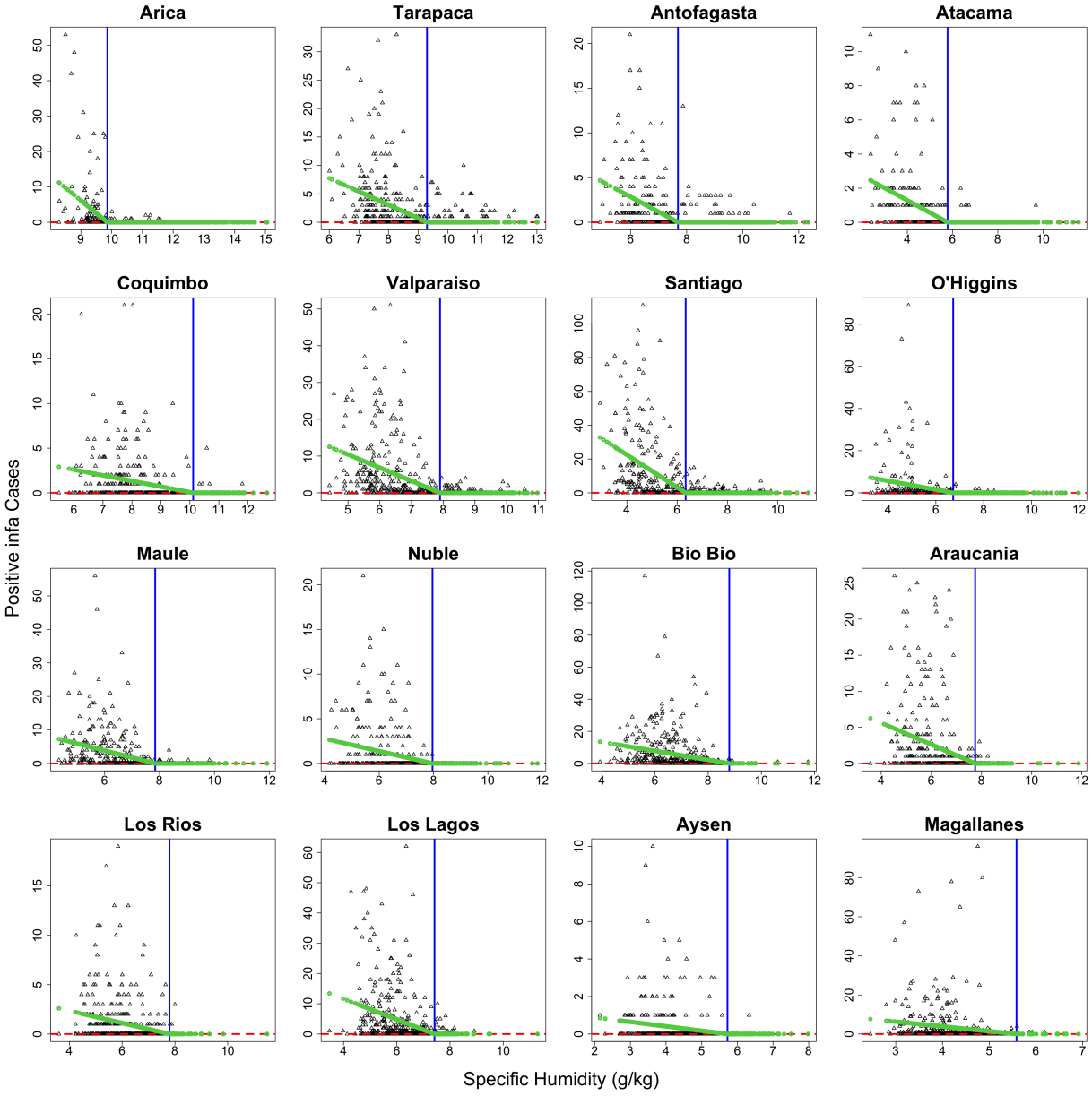

Figure B.4: Segmented regression for positive Influenza A cases and temperature.

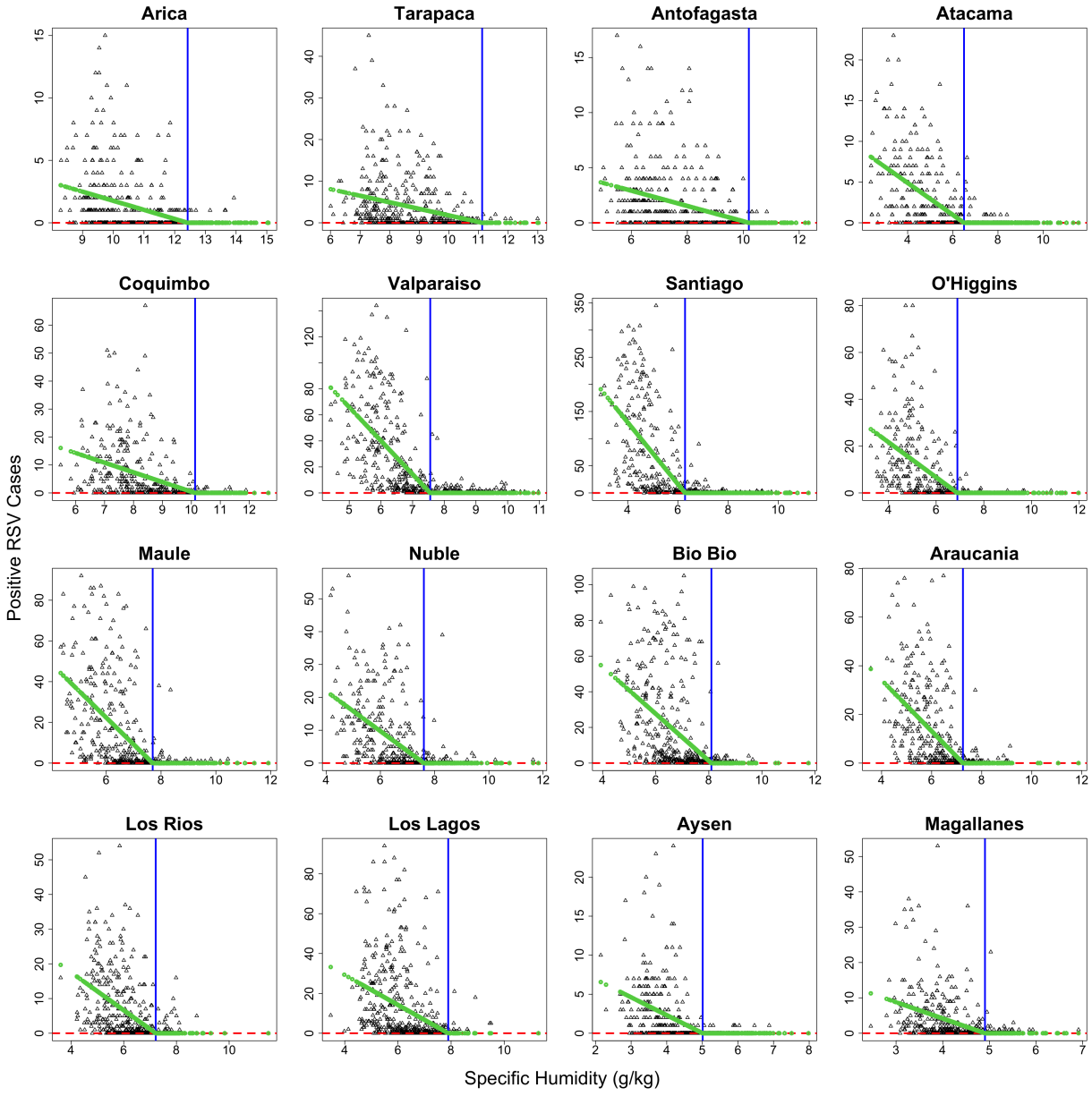

Figure B.5: Segmented regression for positive RSV cases and specific humidity.

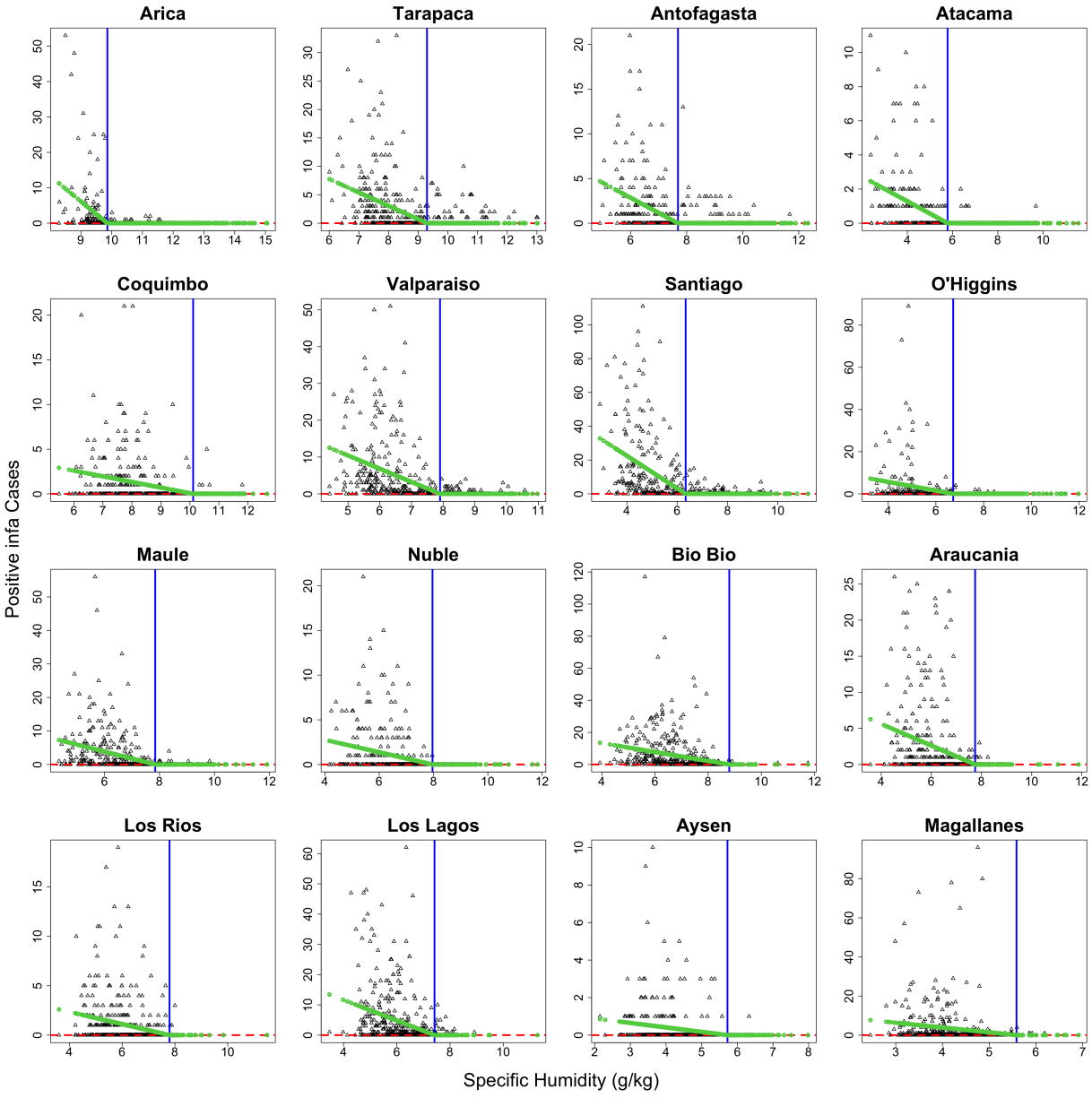

Figure B.6: Segmented regression for positive Influenza A cases and temperature.

B.4 Onset week regression for Influenza A

To investigate whether year-to-year variations in specific humidity and temperature can alter the timing of onset of Influenza A epidemics within a particular location we fit a linear regression model with onset week as the dependent variable and climate covariates as the independent variable, seen in Figure B.7. A 3 (g/kg) increase in mean annual specific humidity ($p < 0.01$) and a 0.1 (°C) increase in mean annual temperature ($p < 0.5$) shifts the timing of the Influenza epidemic back by 1 week, as seen Table 3.2.

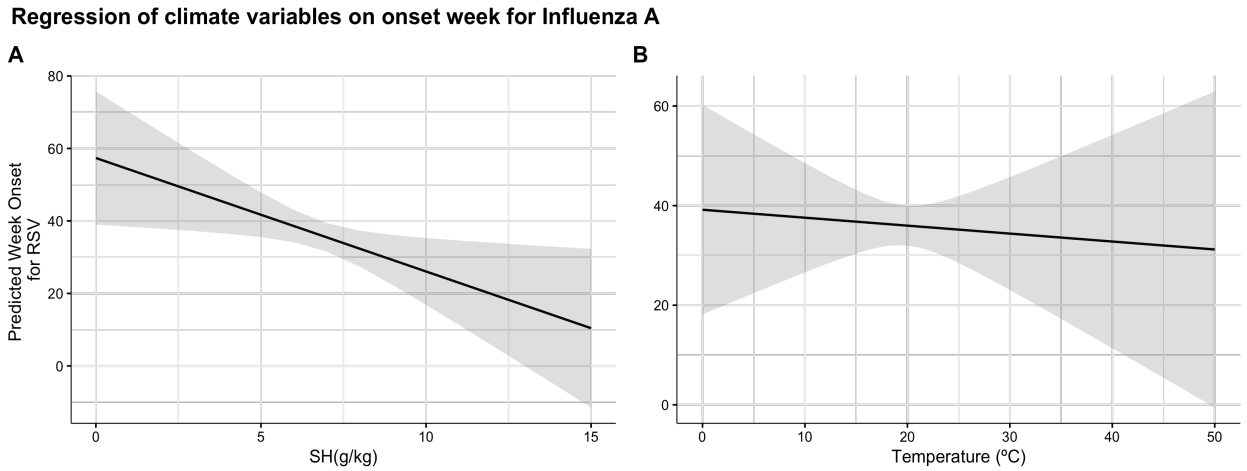

Figure B.7: **A.** Graph of linear regression model of onset week for Influenza A with specific humidity (g/kg) as independent variable. **B.** Graph of linear regression model of onset week for Influenza A with temperature ($°C$) as independent variable.

B.5 Interannual variability

B.5.1 Onset

When we evaluate the relationship between the environmental factors and timing of peak week, there is no significant association. In fact, specific humidity only explains 32% (4%) of the variance in the timing of peak week for RSV (Influenza A). This is 30% (8%) less compared to RSV (Influenza A) onset seen in Figure 3.18 (Figure 3.19). The results for temperature and latitude are similar.

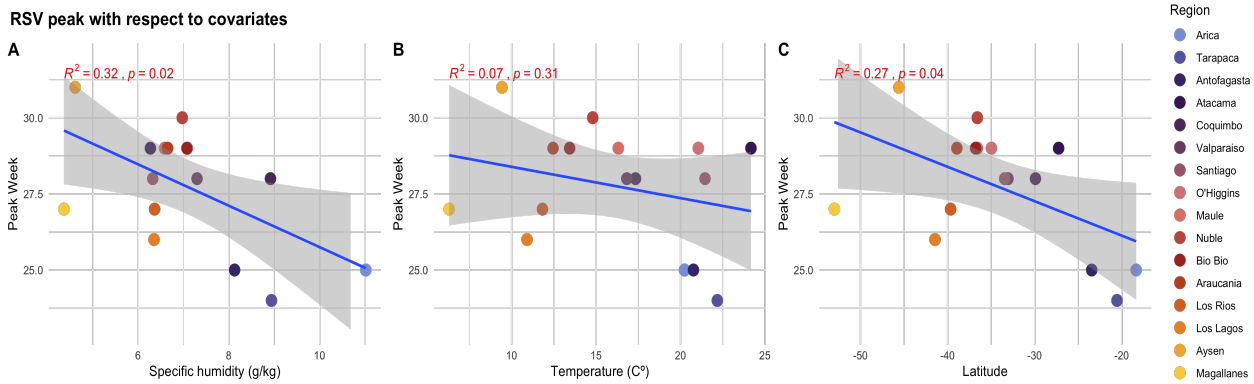

Figure B.8: **A.** Peak week of RSV with respect to specific humidity (g/kg). **B.** Peak week of RSV with respect to temperature (°C). **C.** Peak week of RSV with respect to latitude. Regions are ordered from North to South.

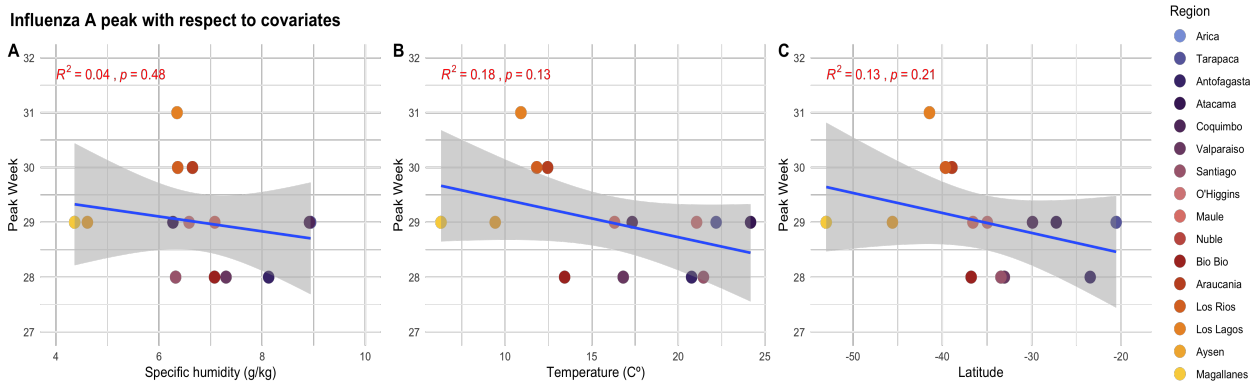

Figure B.9: **A.** Peak week of Influenza A with respect to specific humidity (g/kg). **B.** Peak week of Influenza A with respect to temperature (Cº). **C.** Peak week of Influenza A with respect to latitude. Regions are ordered from North to South.

B.5.2 Size of outbreak

Figures B.10, B.11 and B.12 show the association between climate covariates and interannual variability of outbreak size (i.e., amplitude of epidemic). There is no significant association between the climate covariates and the year-to-year amplitude of the RSV and Influenza A epidemics.

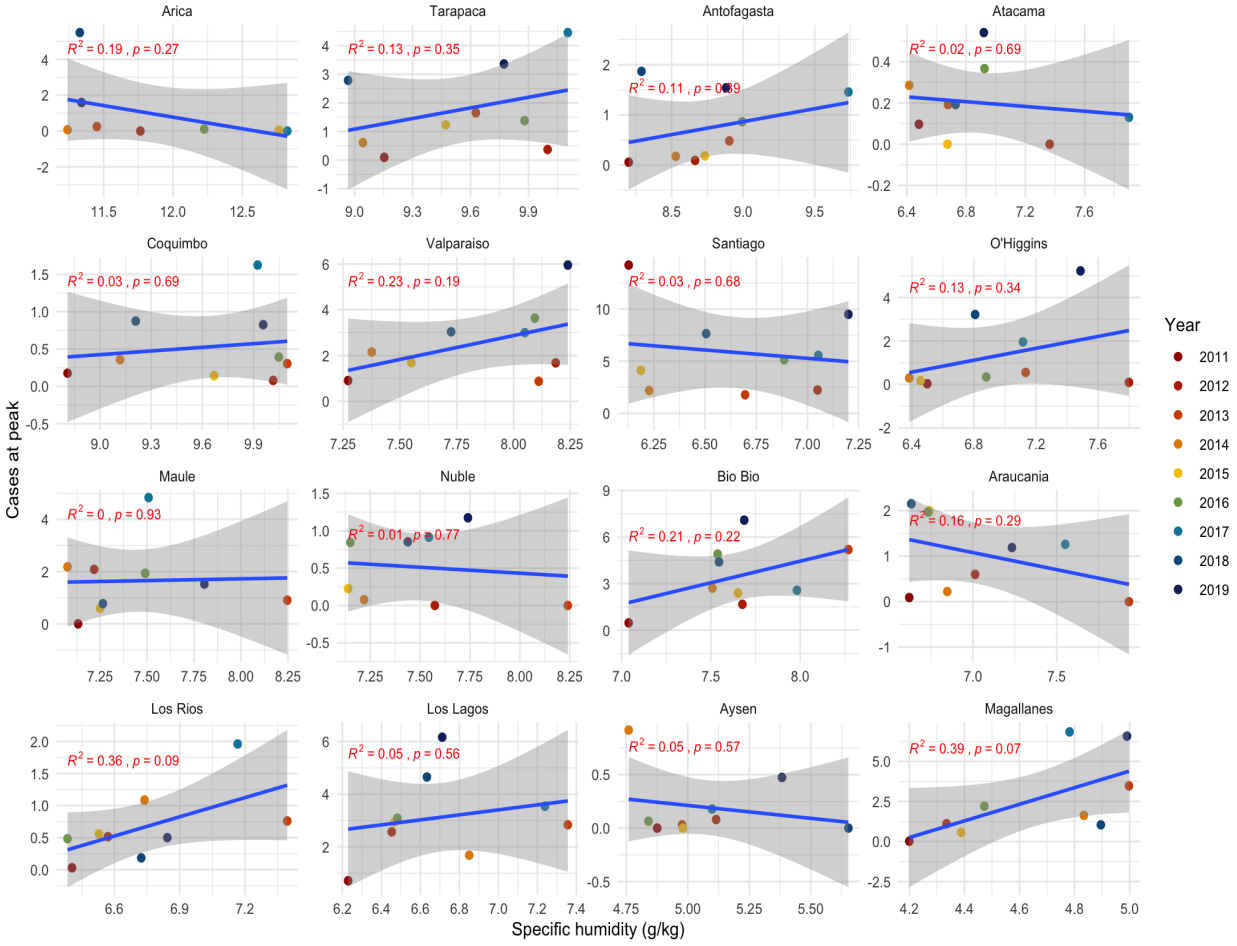

Figure B.10: Graph of interannual variability for Influenza A with respect to specific humidity (g/kg).

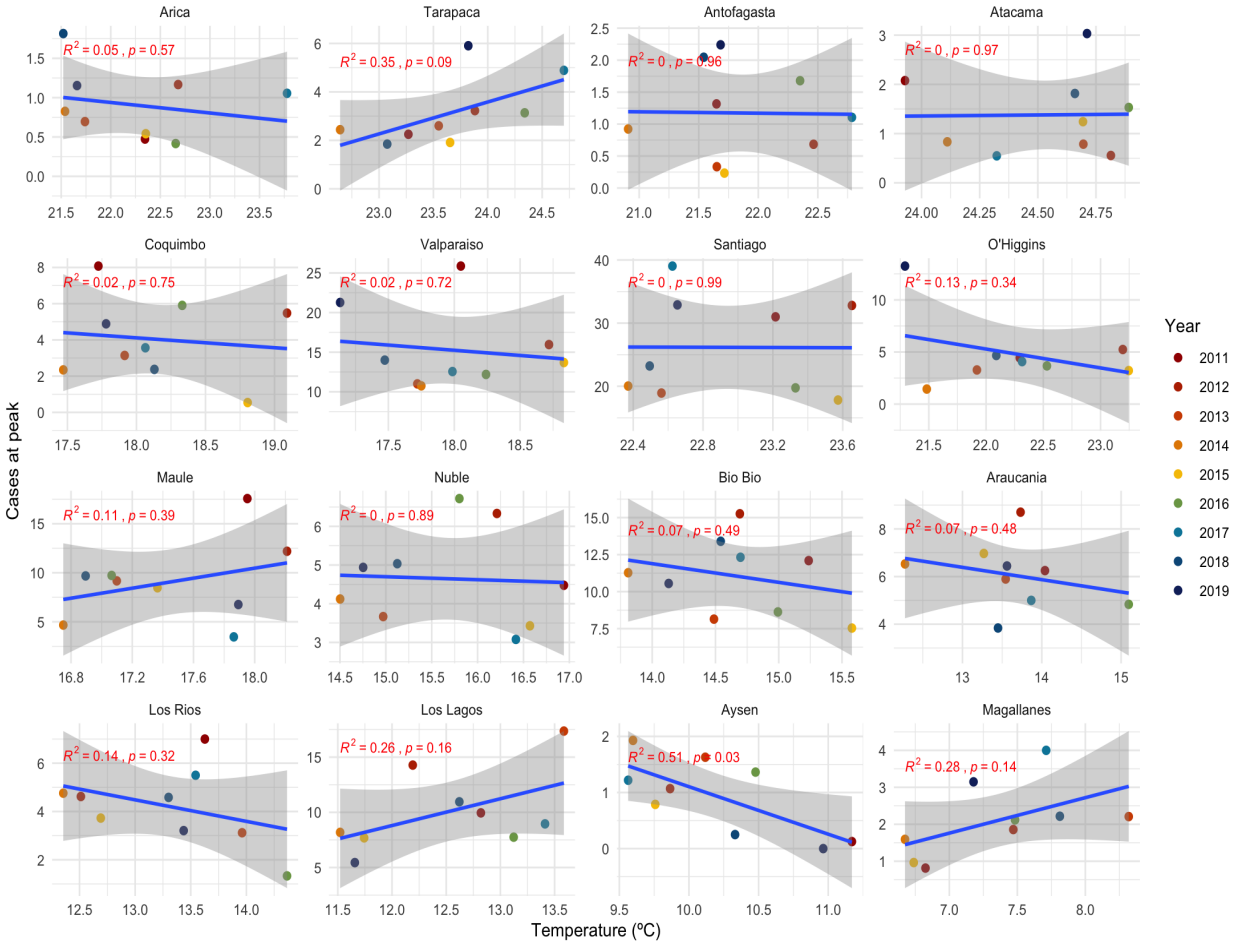

Figure B.11: Graph of interannual variability for RSV with respect to temperature (°C).

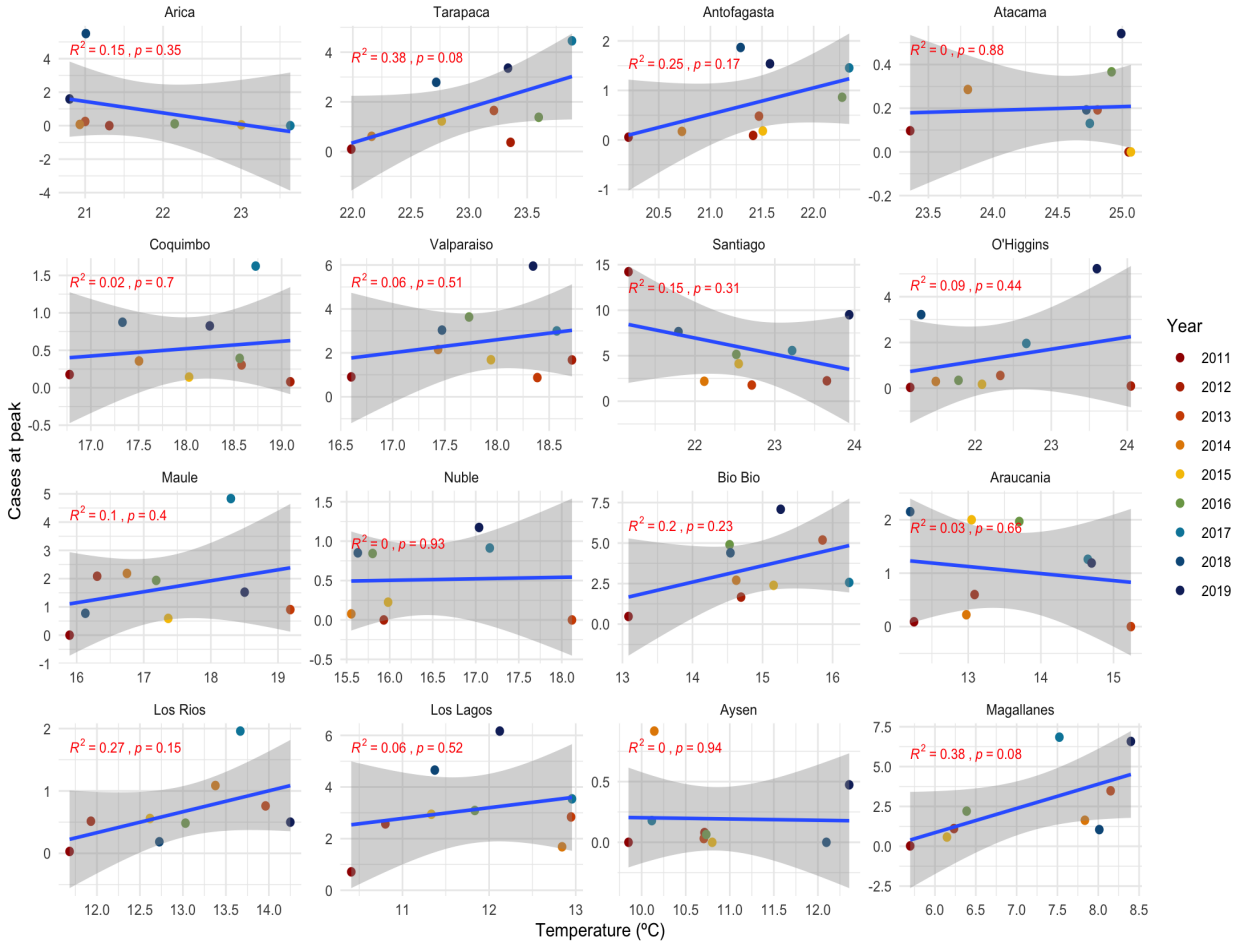

Figure B.12: Graph of interannual variability for Influenza A with respect to temperature (°C).

Appendix C

Statistical methods

C.1 Multivariate logistic regression

Multivariate logistic regression analysis is used to predict the relationships by calculating the probability of something happening depending on multiple variables. Instead of a continuous set of results, a multivariate logistic regression can have multiple independent variables and outcomes.

Mathematically, let $\pi(x)$ represent the probability of an event that depends on n covariates. Then we can estimate the probabilities of outcomes with the following formula

$$\pi(X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}}$$

given the covariates $X_1, X_2, \ldots, X_n$. Note that applying the logit function $\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$ for $p \in (0,1)$, i.e. the inverse of the standard logistic function $\sigma(x) = 1/(1+e^x)$, we get

$$\text{logit}(\pi(X)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n.$$

This shows that multivariate logistic regression is a standard linear regression model, after we transform the outcome using the logit function. Note that the logit function maps probability values from $(0,1)$ to $(-\infty, +\infty)$. For this equation, each data point will have an error term, is often assumed to follow a normal distribution with mean zero and constant variance.

C.2 Generalized Linear Models

In statistics, model selection depends a lot on the data that is being analyzed. Many times, the data does not behave linearly, so a simple linear regression model of the form $y_i = \beta_0 + \beta_1 x_i + \varepsilon$ for data of the form $(x_i, y_i), i = 1, \ldots, n$ would not be able to accurately approximate the observed pattern. Having more flexibility in the model allows for better

predictive power. A generalized linear model (GLM) is a flexible generalization of ordinary linear regression. GLMs can be written in terms of the expected value (mean) of the response variable y in the following way

$$g(\mu) = \sum_{i=1}^{n} \beta_i X$$

where μ is the expected value of the response variable $E(y)$. A *link function* is then used to relate the mean μ to the linear predictor βX. For example, consider a linear regression model, then y has a Normal distribution, i.e. we consider equal variance for all observations, and $\mu = \beta X$, and the link function is called the identity link function. If a Poisson distribution is used as the distribution for y, then the link function $g(\cdot)$ is considered to be the natural logarithm and thus

$$\ln(\mu) = \sum_{i=1}^{n} \beta_i X.$$

There are many ways in which GLMs can be be modified to match the overall pattern of the data that is being modeled by choosing the distribution and link function. To summarize, a GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the variance of each measurement to be a function of the linear predictor. We can incorporate nonlinearity into the model by using generalized additive models. For more on regression methods please refer to the following books [FT94, FKLM13, HHOW19a].

C.3 Splines

A spline is a mathematical representation for fitting flexible shapes and curves. In statistical regression analysis, splines are used to build more flexible models that provide a better fit to the data that is being used. In general, functions (usually polynomials) are joined together through selected points, commonly known as knots, and these functions are chosen to fit the data between two consecutive knots. The type of polynomial and number and placement of knots is what specifies the different types of splines. We give an overview on what splines are and focus on the ones that are used in Sections 3.2.3 and 4.2.5. For more mathematical and technical details, the reader can explore the following sources [dB01, Wan11, Wah90]. We refer to [PSAS19, H.] for a comprehensive review on the implementation of splines in the R programming language.

Mathematically, let $f : \mathbb{R} \to \mathbb{R}$ be a piecewise polynomial function of degree d and define the set of knots $t_1 < t_2 < \ldots < t_k$. The spline f will be a smooth function, satisfying a

smoothness criterion that all derivatives of order less than d are continuous, such that f is a polynomial of degree d.

A spline function f with fixed knots and fixed degree can be written as

$$f(x) = \sum_{i=1}^{k+d+1} b_i s_i(x)$$

where the s_i are a set of truncated power basis functions for d-order splines over the knots $t_1, \ldots, t_k$ and b_i are the associated spline coefficients, with degree d and k knots. This representation is linear in the coefficient vector $b = (b_1, \ldots, b_{k+d+1})$ and thus linear in the transformed variables $s_i, \ldots, s_{k+d+1}$. Therefore the problem is reduced to estimating the coefficients b_i. Assuming data of the form (x_j, y_j), $j = 1, \ldots, n$, a *regression spline* can be defined by

$$\hat{r}(x) = \sum_{i=1}^{k+d+1} \hat{b}_i s_i(x),$$

where $\hat{b}_1 \ldots, \hat{b}_{k+d+1}$ are found by minimizing

$$\sum_{j=1}^{n} \left(y_j - \sum_{i=1}^{k} b_i s_i(x_j) \right)^2.$$

Regression splines are a very good start and can be very useful when using an appropriate amount of knots. To obtain more flexible curves the number of knots or the degree of the polynomial can be increased. It is important to note that increasing the number of knots may overfit the data and increase the variance, on the other hand decreasing the number of knots may give a more restrictive function that has more bias. One issue with regression splines is that the estimates can exhibit high variance at the variance at the boundaries of the domain of $x_1, \ldots, x_n$. The variance increases as the degree of the polynomial gets larger. To avoid this issue, we can impose a further constraint that derivatives of order greater or equal to two are zero at the leftmost and rightmost knots. Splines with this extra constraint are called *natural splines*.

We will focus on the splines used in the thesis: periodic B-splines and cubic smoothing splines. *Cubic splines* are created by using a cubic polynomial in an interval between two successive knots. These are piecewise cubic functions that are continuous, and have a continuous first and second derivatives. Specifically, the knots $t_j = (x_j, y_j)$ and $t_{j+1} = (x_{j+1}, y_{j+1})$ are joined by the cubic polynomial

$$p_j(x) = a_j x^3 + b_j x^2 + c_j x + d_j$$

for $x_j \leq x \leq x_{j+1}$ and $j = 1, \ldots, k-1$. The B-spline basis is a spline basis that is based on a special parametrization of a cubic spline defined on $[a, b]$. The B-spline basis is based on the knot sequence

$$\tau_1 \leq \cdots \leq \tau_d \leq \tau_{d+1} < \tau_{d+2} < \cdots < \tau_{d+k+1} < \tau_{d+k+2} \leq \tau_{d+k+3} \leq \cdots \leq \tau_{2d+k+2}$$

where $\tau_{d+2} := t_1, \ldots, \tau_{d+k+1} : t_k$ are the inner knots and $tau_{d+1} := a, \tau_{d+k+2} := b$ are the boundary knots. For $d > 0$, B-spline basis functions of degree d are defined by the recursive formula

$$B_k^d(c) = \frac{x - \tau_l}{\tau_{l+d} - \tau_l} B_l^{d-1}(x) - \frac{\tau_{l+d+1} - x}{\tau_{l+d+1} - \tau_{l+1}} B_{l+1}^{d-1}(x)$$

for $l = 1, \ldots, k + d + 1$, where

$$B_l^0(x) = \begin{cases} 1, & \tau_l \leq x < \tau_{l+1} \\ 0, & \text{else} \end{cases}$$

Smoothing splines are a special type of spline where the knot positions do not have to be chosen. Instead, they are automatically placed at all points $x_1, \ldots, x_n$, i.e. the inputs are chosen as knots. Cubic smoothing splines are of the form

$$\sum_{i=1}^{n} b_i s_i(x)$$

where $s_1, \ldots, s_n$ are the truncated power basis functions for natural cubic splines with knots at $x_1, \ldots, x_n$. The coefficients are chosen to minimize

$$||y - Bb||_2^2 + \lambda b^T \Omega b, \tag{C.1}$$

where $B = B_{ij} = s_i(x_j)$ is the basis matrix and $\Omega_i j = \int s_i''(t) s_j''(t) dt$ is the penalty matrix. The term $\lambda b^T \Omega b$ is called a regularization term, and it shrinks the components of the optimal coefficients $\hat{b}$ towards zero. The parameter $\lambda \geq 0$ is called the smoothing parameter, and it controls the amount of shrinkage the components will have (higher λ means more shrinkage).

References

[ABDN+12] E. Azziz Baumgartner, C. N. Dao, S. Nasreen, M. U. Bhuiyan, S. Mah-E-Muneer, A. A. Mamun, M. A. Y. Sharker, R. U. Zaman, P. Cheng, A. I. Klimov, M. Widdowson, T. M. Uyeki, S. P. Luby, A. Mounts, and J. Bresee. Seasonality, Timing, and Climate Drivers of Influenza Activity Worldwide. *The Journal of Infectious Diseases*, 206(6), 2012.

[AdFG01] D. Arnaud, N. de. Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Information Science and Statistics. Springer New York, New York, NY, 2001.

[ADH10] C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods: Particle Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3), 2010.

[ADL+20] A. Adiga, D. Dubhashi, B. Lewis, M. Marathe, S. Venkatramanan, and A. Vullikanti. Mathematical Models for COVID-19 Pandemic: A Comparative Analysis. *Journal of the Indian Institute of Science*, 100(4), 2020.

[Ale10] E. C. Alexopoulos. Introduction to multivariate regression analysis. *Hippokratia*, 14(Suppl 1), 2010.

[AM82] R. M. Anderson and R. M. May. Directly Transmitted Infections Diseases: Control by Vaccination. *Science*, 215(4536), 1982.

[AM91] R. M. Anderson and R. M. May. *Infectious diseases of humans: dynamics and control*. Oxford science publications. Oxford University Press, Oxford, 1991.

[AMGC02] M.S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2), 2002.

[AMMJ86] R. M. Anderson, G. F. Medley, R. M. May, and A. M. Johnson. A preliminary study of the transmission dynamics of the human immunodeficiency virus (HIV), the causative agent of AIDS. *IMA journal of mathematics applied in medicine and biology*, 3(4), 1986.

[AMY+21] I. Ahmed, G. U. Modu, A. Yusuf, P. Kumam, and I. Yusuf. A mathematical model of Coronavirus Disease (COVID-19) containing asymptomatic and symptomatic classes. *Results in Physics*, 21, 2021.

[APC18] S. Ascough, S. Paterson, and C. Chiu. Induction and Subversion of Human Protective Immunity: Contrasting Influenza and Respiratory Syncytial Virus. *Frontiers in Immunology*, 9, 2018.

[ASS91] S. A. Ansari, V. S. Springthorpe, and S. A. Sattar. Survival and vehicular spread of human rotaviruses: Possible relation to seasonality of outbreaks. *Reviews of Infectious Diseases*, 13(3), 1991.

[AT15] E. K. Allen and P. G. Thomas. Immunity to Influenza. Preventing Infection and Regulating Disease. *American Journal of Respiratory and Critical Care Medicine*, 191(3), 2015.

[BAC+13] K. Bloom-Feshbach, W. J. Alonso, V. Charu, J. Tamerius, L. Simonsen, M. A. Miller, and C. Viboud. Latitudinal Variations in Seasonal Activity of Influenza and Respiratory Syncytial Virus (RSV): A Global Comparative Review. *PLoS ONE*, 8(2), 2013.

[BFG23] O. N. Bjornstad, B. F. Finkenstadt, and B. T. Grenfell. Dynamics of Measles Epidemics: Estimating Scaling of Transmission Rates Using a Time Series SIR Model. *Ecological Monographs*, 72(2), 2023.

[BHIK09] C. Bretó, D. He, E. L. Ionides, and A. A. King. Time series analysis via mechanistic models. *The Annals of Applied Statistics*, 3(1), 2009.

[BIK20] C. Bretó, E. L. Ionides, and A. A. King. Panel Data Analysis via Mechanistic Models. *Journal of the American Statistical Association*, 115(531), 2020.

[BMW+19] R. E. Baker, A. S. Mahmud, C. E. Wagner, W. Yang, V. E. Pitzer, C. Viboud, G. A. Vecchi, C. J.E. Metcalf, and B. T. Grenfell. Epidemic dynamics of respiratory syncytial virus in current and future climates. *Nature Communications*, 10(1), 2019.

[Boz87] H. Bozdogan. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3), 1987.

[Bra] L. Bradley. Smallpox inoculation : An eighteenth century mathematical controversy / translation and critical commentary by L. Bradley. https://wellcomecollection.org/works/rvydxqky.

[Bra17] F. Brauer. Mathematical epidemiology: Past, present, and future. *Infectious Disease Modelling*, 2(2), 2017.

[Bre18] C. Bretó. Modeling and Inference for Infectious Disease Dynamics: A Likelihood-Based Approach. *Statistical Science*, 33(1), 2018.

[Bur] U.S. Census Bureau. International Database. https://www.census.gov/data-tools/demo/idb/.

[BWBI96] N. G. Berman, W. K. Wong, S. Bhasin, and E. Ipp. Applications of segmented regression models for biomedical studies. *American Journal of Physiology-Endocrinology and Metabolism*, 270(4), 1996. PMID: 8928781.

[CCC14] B. Cazelles, K. Cazelles, and M. Chavez. Wavelet analysis in ecology and epidemiology: Impact of statistical tests. *Journal of the Royal Society Interface*, 11(91), 2014.

[CCMH05] B. Cazelles, M. Chavez, A. J. McMichael, and S. Hales. Nonstationary Influence of El Niño on the Synchronous Dengue Epidemics in Thailand. *PLoS Medicine*, 2(4), 2005.

[CFR+09] B. J. Cowling, V. J. Fang, S. Riley, J S. Malik P., and G. M. Leung. Estimation of the Serial Interval of Influenza. *Epidemiology*, 20(3), 2009.

[CHT98] R. R. Carmona, W. Hwang, and B. Torrésani. *Practical Time-Frequency Analysis : Gabor and Wavelet Transforms with an Implementation in S.* Wavelet Analysis and Its Applications ; v. 9. Academic Press, San Diego, 1998.

[CMW14] D. Chen, B. Moulin, and J. Wu, editors. *Analyzing and Modeling Spatial and Temporal Dynamics of Infectious Diseases: Chen/Analyzing.* Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2014.

[CRP04] T. Coulson, P. Rohani, and M. Pascual. Skeletons, noise and population growth: The end of an old debate? *Trends in Ecology & Evolution*, 19(7), 2004.

[CZ22] C. Cohen and H. J. Zar. Deaths from RSV in young infants—the hidden community burden. *The Lancet Global Health*, 10(2), 2022.

[dB01] C. de Boor. *A practical guide to splines.* Applied Mathematical Sciences, 27. Springer-Verlag, New York, 2001.

[DH00] K. Dietz and J. A. P. Heesterbeek. Bernoulli was ahead of modern epidemiology. *Nature*, 408(6812), 2000.

[DH18] I. Dattner and A. Huppert. Modern statistical tools for inference and prediction of infectious diseases using mathematical models. *Statistical Methods in Medical Research*, 27(7), 2018.

[DHM90] O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz. On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations. *Journal of Mathematical Biology*, 28(4), 1990.

[DJ11] A. Doucet and A. M. Johansen. A tutorial on particle filtering and smoothing: fifteen years later. In *The Oxford handbook of nonlinear filtering*, 2011.

[DKWP17] X. Du, A. A. King, R. J. Woods, and M. Pascual. Evolution-informed forecasting of seasonal influenza A (H3N2). *Science Translational Medicine*, 9(413), 2017.

[DM] Ripley B. D. and M. Maechler. R: Fit a Smoothing Spline. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/smooth.spline.html.

[Dow01] S. F. Dowell. Seasonal Variation in Host Susceptibility and Cycles of Certain Infectious Diseases. *Emerging Infectious Diseases*, 7(3), 2001.

[DPLE04] J. Dushoff, J. B. Plotkin, S. A. Levin, and David J. D. Earn. Dynamical resonance can account for seasonality of influenza epidemics. *Proceedings of the National Academy of Sciences*, 101(48), 2004.

[DS] S. Dattani and F. Spooner. How many people die from the flu? https://ourworldindata.org/influenza-deaths.

[ETBE12] K. T. D. Eames, N. L. Tilston, E. Brooks-Pollock, and W. J. Edmunds. Measured Dynamic Social Contact Patterns Explain the Spread of H1N1v Influenza. *PLoS Computational Biology*, 8(3), 2012.

[ETvdM21] J. Elfring, E. Torta, and R. van de Molengraft. Particle filters: A hands-on tutorial. *Sensors (Basel)*, 21(2), 2021.

[FC82] P. E. M. Fine and J. A. Clarkson. Measles in England and Wales—I: An Analysis of Factors Underlying Seasonal Patterns. *International Journal of Epidemiology*, 11(1), 1982.

[FG00] B. F. Finkenstädt and B. T. Grenfell. Time Series Modelling of Childhood Diseases: A Dynamical Systems Approach. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 49(2), 2000.

[Fis07a] D. N. Fisman. Seasonality of infectious diseases. *Annu Rev Public Health*, 28, 2007.

[Fis07b] D. N. Fisman. Seasonality of Infectious Diseases. *Annual Review of Public Health*, 28(1), 2007.

[FK20] S. Funk and A. A. King. Choices and trade-offs in inference with infectious disease models. *Epidemics*, 30, 2020.

[FKLM13] L. Fahrmeir, T. Kneib, S. Lang, and B. Marx. *Regression: Models, Methods and Applications*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[FPW16] M. Fasiolo, N. Pya, and S. N. Wood. A Comparison of Inferential Methods for Highly Nonlinear State Space Models in Ecology and Epidemiology. *Statistical Science*, 31(1), 2016.

[Fra80] J. C. Frauenthal. *Mathematical modeling in epidemiology*. Universitext. Springer-Verlag, Berlin ;, 1980.

[Fre21] Frequently Asked Questions about Estimated Flu Burden — CDC. https://www.cdc.gov/flu/about/burden/faq.htm, 2021.

[FT94] L. Fahrmeir and G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer Series in Statistics. Springer New York, New York, NY, 1994.

[FTY11] Z. Feng, S. Towers, and Y. Yang. Modeling the Effects of Vaccination and Treatment on Pandemic Influenza. *The AAPS Journal*, 13(3), 2011.

[GBF02] B. T. Grenfell, O. N. Bjørnstad, and B. F. Finkenstadt. Dynamics of Measles Epidemics: Scaling Noise, Determinism, and Predictability with the TSIR Model. 72(2), 2002.

[GL01] R. Gani and S. Leach. Transmission potential of smallpox in contemporary populations. *Nature*, 414(6865), 2001.

[Gle86] W. P. Glezen. Risk of Primary Infection and Reinfection With Respiratory Syncytial Virus. *Archives of Pediatrics & Adolescent Medicine*, 140(6), 1986.

[H.] Nathaniel E. H. Smoothing Spline Regression in R. http://users.stat.umn.edu/~helwig/notes/smooth-spline-notes.html.

[HB20] I. Holmdahl and C. Buckee. Wrong but Useful — What Covid-19 Epidemiologic Models Can and Cannot Tell Us. *New England Journal of Medicine*, 383(4), 2020.

[HBB+18] H. Hersbach, B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, I. Rozum, D. Schepers, A. Simmons, C. Soci, D. Dee, and J-N. Thépaut. ERA5 hourly data on pressure levels from 1959 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS)., 2018.

[HBB+20] H. Hersbach, B. Bell, P. Berrisford, A. Hirahara, S.and Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. Rosnay, I. Rozum, F. Vamborg, S. Villaume, and J. Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730), 2020.

[HBP+20] J. Hess, L. G Boodram, S. Paz, A. M. Stewart Ibarra, J. N. Wasserheit, and R. Lowe. Strengthening the global response to climate change and infectious disease threats. *BMJ*, 2020.

[HCDV22] C. L. Hansen, S. S. Chaves, C. Demont, and C. Viboud. Mortality Associated With Influenza and Respiratory Syncytial Virus in the US, 1999-2018. *JAMA Network Open*, 5(2), 2022.

[HDG76] C. B. Hall, R. G. Jr. Douglas, and J. M. Geiman. Respiratory syncytial virus infections in infants: quantitation and duration of shedding. *J Pediatr*, 89(1), 1976.

[HHOW19a] L. Held, N. Hens, P. D. O'Neill, and J. Wallinga. *Handbook of Infectious Disease Data Analysis*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, 2019.

[HHOW19b] L. Held, N. Hens, P.D. O'Neill, and J. Wallinga. *Handbook of Infectious Disease Data Analysis*. CRC Press, Epidemiology, Biostatistics and Prevention Institute, University of Zurich, 2019.

[HIK10] D. He, E. L. Ionides, and A. A. King. Plug-and-play inference for disease dynamics: Measles in large and small populations as a case study. *Journal of The Royal Society Interface*, 7(43), 2010.

[HKK97] P. O. Honkanen, T. Keistinen, and S. Kivelä. The impact of vaccination strategy and methods of information on influenza and pneumococcal vaccination coverage in the elderly population. *Vaccine*, 15(3), 1997.

[Hsi10] Y. Hsieh. Age groups and spread of influenza: Implications for vaccination strategy. *BMC Infectious Diseases*, 10(1), 2010.

[HTK16] D. Higgins, C. Trujillo, and C. Keech. Advances in RSV vaccine research and development – A global agenda. *Vaccine*, 34(26), 2016.

[HU16] J. Hilgevoord and J. Uffink. The Uncertainty Principle. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2016.

[HWLS91] C. B. Hall, E. E. Walsh, C. E. Long, and K. C. Schnabel. Immunity to and Frequency of Reinfection with Respiratory Syncytial Virus. *Journal of Infectious Diseases*, 163(4), 1991.

[IBK06] E. L. Ionides, C. Bretó, and A. A. King. Inference for nonlinear dynamical systems. *Proc Natl Acad Sci U S A*, 103(49), 2006.

[INA+15] E. L. Ionides, D. Nguyen, Y. Atchadé, S. Stoev, and A. A. King. Inference for dynamic and latent variable models via iterated, perturbed Bayes maps. *Proceedings of the National Academy of Sciences*, 112(3), 2015.

[Ion05] E. L. Ionides. Maximum Smoothed Likelihood Estimation. *Statistica Sinica*, 15(4), 2005.

[isp] Instituto de Salud Pública de Chile. https://www.ispch.gob.cl/virusrespiratorios/.

[JAB21] P. C. Jentsch, M. Anand, and C. T. Bauch. Prioritising COVID-19 vaccination in changing social and epidemiological landscapes: A mathematical modelling study. *The Lancet Infectious Diseases*, 21(8), 2021.

[JBMC62] K. M. Johnson, H. H. Bloom, M. A. Mufson, and R. M. Chanock. Natural reinfection of adults by respiratory syncytial virus. *New England Journal of Medicine*, 267(2), 1962.

[JSH+07] A. G. S. C. Jansen, E. A. M. Sanders, A. W. Hoes, A. M. van Loon, and E. Hak. Influenza- and respiratory syncytial virus-associated mortality and hospitalisations. *European Respiratory Journal*, 30(6), 2007.

[KBM+99] B. E. Kendall, C. J. Briggs, W. W. Murdoch, P. Turchin, S. P. Ellner, E. McCauley, R. M. Nisbet, and S. N. Wood. Why do populations cycle? A synthesis of statistical and mechanistic modeling approaches. *Ecology*, 80(6), 1999.

[KDI16] A. A. King, Nguyen D., and E. L. Ionides. Statistical inference for partially observed Markov processes via the R package pomp. *Journal of Statistical Software, Articles*, 69(12), 2016.

[KFFM00] H. Kim, M. P. Fay, E. J. Feuer, and D. N. Midthune. Permutation tests for joinpoint regression with applications to cancer rates. *Statistics in Medicine*, 19(3), 2000.

[KI22] A. A. King and E. L. Ionides. Simulation-based inference for epidemiological dynamics, 2022.

[KIMB+22] A. A. King, E. L. Ionides, C. Martinez Bretó, S. P. Ellner, M. J. Ferrari, S. Funk, S. G. Johnson, B. E. Kendall, M. Lavine, D. Nguyen, E. B. O'Dea, D.C. Reuman, H. Wearing, and S. M. Wood. *pomp: Statistical Inference for Partially Observed Markov Processes*, 2022. R package, version 4.2.

[KIPB08] A. A. King, E. L. Ionides, M. Pascual, and M. J. Bouma. Inapparent infections and cholera dynamics. *Nature*, 454(7206), 2008.

[KKM+21] N. Kishore, R. Kahn, P. P. Martinez, P. M. D. S., Ayesha S. Mahmud, and C. O. Buckee. Lockdowns result in changes in human mobility which may impact the epidemiologic dynamics of SARS-CoV-2. *Scientific Reports*, 11(1), 2021.

[KMW27] W. O. Kermack, A. G. McKendrick, and G. T. Walker. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115(772), 1927.

[KR08] M. J. Keeling and P. Rohani. *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, 2008.

[KRP+05] K. Koelle, X. Rodó, M. Pascual, Md. Yunus, and G. Mostafa. Refractory periods and climate forcing in cholera dynamics. *Nature*, 436(7051), 2005.

[LBG11] F. Lewis, A. Butler, and L. Gilbert. A unified approach to model selection using the likelihood ratio test: *Unified approach to model selection*. *Methods in Ecology and Evolution*, 2(2), 2011.

[LBI+10] K. Laneri, A. Bhadra, E. L. Ionides, R. C. Bouma, M.and Dhiman, R. S. Yadav, and M. Pascual. Forcing Versus Feedback: Epidemic Malaria and Monsoon Rains in Northwest India. *PLoS Computational Biology*, 6(9), 2010.

[LC18] J. Lewnard and S. Cobey. Immune History and Influenza Vaccine Effectiveness. *Vaccines*, 6(2), 2018.

[LE22] Z. Levine and D. J. D. Earn. Face masking and COVID-19: Potential effects of variolation on transmission dynamics. *Journal of The Royal Society Interface*, 19(190), 2022.

[LES03] R. Lande, S. Engen, and B. Saether. *Stochastic Population Dynamics in Ecology and Conservation*. Oxford University Press, 2003.

[LLT+07] Y. Li, G. M. Leung, J. W. Tang, X. Yang, C. Y. H. Chao, J. Z. Lin, J. W. Lu, P. V. Nielsen, J. Niu, H. Qian, A. C. Sleigh, H.-J. J. Su, J. Sundell, T. W. Wong, and P. L. Yuen. Role of ventilation in airborne transmission of infectious agents in the built environment - a multidisciplinary systematic review. *Indoor Air*, 17(1), 2007.

[LMSP07] A. C. Lowen, S. Mubareka, J. Steel, and P. Palese. Influenza Virus Transmission Is Dependent on Relative Humidity and Temperature. *PLoS Pathogens*, 3(10), 2007.

[LS14] A. C. Lowen and J. Steel. Roles of Humidity and Temperature in Shaping Influenza Seasonality. *Journal of Virology*, 88(14), 2014.

[LSOC14] L. Lambert, A. M. Sagfors, P. J. M. Openshaw, and F. J. Culley. Immunity to RSV in Early-Life. *Frontiers in Immunology*, 5, 2014.

[LWB+22] Y. Li, X. Wang, D. M. Blau, M. T. Caballero, and *et al.* Feikin. Global, regional, and national disease burden estimates of acute lower respiratory infections due to respiratory syncytial virus in children younger than 5 years in 2019: A systematic analysis. *The Lancet*, 399(10340), 2022.

[LY73] W. P. London and J. A. Yorke. Recurrent outbreaks of measles, chickenpox and mumps: I. Seasonal variation in contact rates. *American Journal of Epidemiology*, 98(6), 1973.

[MAA20] J. Mishra, R. Agarwal, and A. Atangana, editors. *Mathematical Modeling and Soft Computing in Epidemiology*. CRC Press, 2020.

[MKA+15] P. K. Munywoki, D. C. Koech, C. N. Agoti, N. Kibirige, J. Kipkoech, P. A. Cane, G. F. Medley, and D. J. Nokes. Influence of age, severity of infection, and co-infection on the duration of respiratory syncytial virus (RSV) shedding. *Epidemiology and Infection*, 143(4), 2015.

[MKY+16] P. P. Martinez, A. A. King, M. Yunus, A. S. G. Faruque, and M. Pascual. Differential and enhanced response to climate forcing in diarrheal disease due to rotavirus across a megacity of the developing world. *Proceedings of the National Academy of Sciences*, 113(15), 2016.

[MLC+22] P. P. Martinez, J. Li, C. P. Cortes, R. E. Baker, and A. S. Mahmud. The Return of Wintertime Respiratory Virus Outbreaks and Shifts in the Age Structure of Incidence in the Southern Hemisphere. *Open Forum Infectious Diseases*, 9(12), 2022.

[MMM19] Y. M. Marwa, I. S. Mbalawata, and S. Mwalili. Continuous Time Markov Chain Model for Cholera Epidemic Transmission Dynamics. *International Journal of Statistics and Probability*, 8(3), 2019.

[Mon95] A. S. Monto. Viral respiratory infections in the community: Epidemiology, agents, and interventions. *The American Journal of Medicine*, 99(6), 1995.

[Mon04] A. S. Monto. Occurrence of respiratory virus: Time, place and person. *Pediatric Infectious Disease Journal*, 23(1), 2004.

[Mug03] V. M. R. Muggeo. Estimating regression models with unknown break-points. *Stat Med*, 22(19), 2003.

[Mug08] V. M. R. Muggeo. segmented: an R Package to Fit Regression Models with Broken-Line Relationships. *R News*, 8(1), 2008.

[NH07] M. I. Nelson and E. C. Holmes. The evolution of epidemic influenza. *Nature Reviews Genetics*, 8(3), 2007.

[NK22] G. Neumann and Y. Kawaoka. Seasonality of influenza and other respiratory viruses. *EMBO Molecular Medicine*, 14(4), 2022.

[NKd+23] K. Newman, V. King, R.and Elvira, P. de Valpine, R. S. McCrea, and B. J. T. Morgan. State-space models for ecological time-series data: Practical model-fitting. *Methods in Ecology and Evolution*, 14(1), 2023.

[NMZC16] K. Nadeem, J. E. Moore, Y. Zhang, and H. Chipman. Integrating population dynamics models and distance sampling data: A spatial hierarchical state-space approach. *Ecology*, 97(7), 2016.

[OA15] O. Olusegun and S. Ayodele. Seasonal Influenza Epidemics and El Niños. *Frontiers in Public Health*, 3, 2015.

[OJR+18] P. Obando-Pacheco, A. J. Justicia-Grande, I. Rivero-Calle, Peter Rodríguez-Tenreiro, C.and Sly, O. Ramilo, A. Mejías, E. Baraldi, N. G. Papadopoulos, H. Nair, M. C. Nunes, L. Kragten-Tabatabaie, T. Heikkinen, A. Greenough, R. T. Stein, P. Manzoni, L. Bont, and F. Martinón-Torres. Respiratory Syncytial Virus Seasonality: A Global Overview. *The Journal of Infectious Diseases*, 217(9), 2018.

[OSF+09] R. Ochola, C. Sande, G. Fegan, P. D. Scott, G. F. Medley, P. A. Cane, and D. J. Nokes. The Level and Duration of RSV-Specific Maternal IgG in Infants in Kilifi Kenya. *PLoS ONE*, 4(12), 2009.

[Pan20] J. Panovska-Griffiths. Can mathematical modelling solve the current Covid-19 crisis? *BMC Public Health*, 20(1), 2020.

[PB12] N. Pica and N. M. Bouvier. Environmental factors affecting the transmission of respiratory viruses. *Current Opinion in Virology*, 2(1), 2012.

[PB14] N. Pica and N. M. Bouvier. Ambient Temperature and Respiratory Virus Infection. *Pediatric Infectious Disease Journal*, 33(3), 2014.

[PB19] V.A. Profillidis and G.N. Botzoris. Trend Projection and Time Series Methods. In *Modeling of Transport Demand*. Elsevier, 2019.

[PRE+00] M. Pascual, X. Rodó, S. P. Ellner, R. Colwell, and M. J. Bouma. Cholera Dynamics and El Niño-Southern Oscillation. *Science*, 289(5485), 2000.

[PSAS19] A. Perperoglou, W. Sauerbrei, M. Abrahamowicz, and M. Schmid. A review of spline function procedures in R. *BMC Medical Research Methodology*, 19(1), 2019.

[PVA+15] V. E. Pitzer, C. Viboud, W. J. Alonso, T. Wilcox, C. J. Metcalf, C. A. Steiner, A. K. Haynes, and B. T. Grenfell. Environmental Drivers of the Spatiotemporal Dynamics of Respiratory Syncytial Virus in the United States. *PLOS Pathogens*, 11(1), 2015.

[PVS+09] V. E. Pitzer, C. Viboud, L. Simonsen, C. Steiner, C. A. Panozzo, W. J. Alonso, M. A. Miller, R. I. Glass, J. W. Glasser, U. D. Parashar, and B. T. Grenfell. Demographic Variability, Vaccination, and the Spatiotemporal Dynamics of Rotavirus Epidemics. *Science*, 325(5938), 2009.

[PYM+21] M. M. Patel, I. A. York, A. S. Monto, M. G. Thompson, and A. M. Fry. Immune-mediated attenuation of influenza illness after infection: Opportunities and challenges. *The Lancet Microbe*, 2(12), 2021.

[QXL+22] X. Qiu, S. Xu, Y. Lu, Z. Luo, Y. Yan, C. Wang, and J. Ji. Development of mRNA vaccines against respiratory syncytial virus (RSV). *Cytokine & Growth Factor Reviews*, 68, 2022.

[R C21] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[RB20] F. Rowe and M. Bell. *The Drivers of Long-Distance Commuting in Chile: The Role of the Spatial Distribution of Economic Activities*. Springer Singapore, Singapore, 2020.

[RN14] S. Rao and A. Nyquist. Respiratory viruses and their impact in healthcare. *Current Opinion in Infectious Diseases*, 27(4), 2014.

[Rou19] M. R. Roussel. *Nonlinear Dynamics*. IOP Publishing, 2019.

[SBH20] T. Stocks, T. Britton, and M. Höhle. Model selection and parameter estimation for dynamic epidemic models via iterated filtering: Application to rotavirus in Germany. *Biostatistics*, 21(3), 2020.

[SC02] Q. Shao and N. A. Campbell. Applications: Modelling trends in groundwater levels by segmented regression with constraints. *Australian & New Zealand Journal of Statistics*, 44(2), 2002.

[SGL11] J. Shaman, E. Goldstein, and M. Lipsitch. Absolute Humidity and Pandemic Versus Epidemic Influenza. *American Journal of Epidemiology*, 173(2), 2011.

[Sil18] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. CRC Press, 2018.

[Sim99] L. Simonsen. The global impact of influenza on morbidity and mortality. *Vaccine*, 17, 1999.

[SK09] J. Shaman and M. Kohn. Absolute humidity modulates influenza survival, transmission, and seasonality. *Proceedings of the National Academy of Sciences*, 106(9), 2009.

[SK12] J. Shaman and A. Karspeck. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences*, 109(50), 2012.

[SL03] L. P. Shek and B. Lee. Epidemiology and seasonality of respiratory tract virus infections in the tropics. *Paediatric Respiratory Reviews*, 4(2), 2003.

[SPV+10] J. Shaman, V. E. Pitzer, C. Viboud, B. T. Grenfell, and M. Lipsitch. Absolute Humidity and the Seasonal Onset of Influenza in the Continental United States. *PLoS Biology*, 8(2), 2010.

[SRR22] M. Sakr, C. Ray, and C. Renso. Big mobility data analytics: Recent advances and open problems. *GeoInformatica*, 26(4), 2022.

[Sto19] T. Stocks. *Iterated filtering methods for Markov process epidemic models*. CRC Press, Department of Mathematics, Stockholm University, 2019.

[SW89] G. A. F. Seber and C. J. Wild. *Nonlinear Regression*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 1989.

[TB17] K. S. Trivedi and A. Bobbio. *Continuous-Time Markov Chain: Availability Models*. Cambridge University Press, 2017.

[TC98] C. Torrence and G. P. Compo. A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 79(1), 1998.

[TGW+10] A. R. Tuite, A. L. Greer, M. Whelan, A.-L. Winter, B. Lee, P. Yan, J. Wu, S. Moghadas, D. Buckeridge, B. Pourbohloul, and D. N. Fisman. Estimated epidemiologic parameters and morbidity associated with pandemic H1N1 influenza. *Canadian Medical Association Journal*, 182(2), 2010.

[TL03] J. D. Toms and M. L. Lesperance. Piecewise regression: A tool for identifying ecological thresholds. *Ecology*, 84(8), 2003.

[TL14] J. W. Tang and T. P. Loh. Correlations between climate factors and incidence-a contributor to RSV seasonality: Climate factors and RSV infections. *Reviews in Medical Virology*, 24(1), 2014.

[Tre16] J. J. Treanor. Influenza Vaccination. *New England Journal of Medicine*, 375(13), 2016.

[TWS+09] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P.H Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface*, 6(31), 2009.

[US] NOAA US Department of Commerce. Discussion on Humidity. https://www.weather.gov/lmk/humidity.

[Val99] C. Valens. A really friendly guide to wavelets. 1999.

[van28] J. J. van Loghem. An Epidemiological Contribution to the Knowledge of the Respiratory Diseases. *Journal of Hygiene*, 28(1), 1928.

[van17] P. van den Driessche. Reproduction numbers of infectious disease models. *Infectious Disease Modelling*, 2(3), 2017.

[VBS+06] C. Viboud, O. N. Bjørnstad, D. L. Smith, L. Simonsen, M. A. Miller, and B. T. Grenfell. Synchrony, Waves, and Spatial Hierarchies in the Spread of Influenza. *Science*, 312(5772), 2006.

[vW02] P. van den Driessche and J. Watmough. Reproduction numbers and subthreshold endemic equilibria for compartmental models of disease transmission. *Mathematical Biosciences*, 180(1), 2002.

[vW08] P. van den Driessche and J. Watmough. Further notes on the basic reproduction number. In Fred Brauer, Pauline van den Driessche, and Jianhong Wu, editors, *Mathematical Epidemiology*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[Wah90] G. Wahba. *Spline models for observational data*. CBMS-NSF Regional Conference series in applied mathematics. Society for Industrial and Applied Mathematics, Philadelphia, Pa, 1990.

[Wan11] Y. Wang. *Smoothing splines : methods and applications*. Monographs on statistics and applied probability. CRC Press, Boca Raton, Fla, 2011.

[WF04] E. E. Walsh and A. R. Falsey. Age related differences in humoral immune response to respiratory syncytial virus infection in adults. *Journal of Medical Virology*, 73(2), 2004.

[WLC+10] T. Wen, N. H. Lin, D. Chao, K. Hwang, C. Kan, K. C. Lin, Joseph T. Wu, S. Y. Huang, I. Fan, and C. King. Spatial–temporal patterns of dengue in areas at risk of dengue hemorrhagic fever in Kaohsiung, Taiwan, 2002. *International Journal of Infectious Diseases*, 14(4), 2010.

[WMG+07] L. J. White, J. N. Mandl, M. G. M. Gomes, A. T. Bodley-Tickell, P. A. Cane, P. Perez-Brena, J. C. Aguilar, M. M. Siqueira, S. A. Portes, S. M. Straliotto, M. Waris, D. J. Nokes, and G. F. Medley. Understanding the transmission dynamics of respiratory syncytial virus using multiple time series and nested models. *Mathematical Biosciences*, 209(1), 2007.

[Woo57] B. Woolf. The log likelihood ratio test (the G-test). *Annals of Human Genetics*, 21(4), 1957.

[Woo17] S. N. Wood. *Generalized Additive Models : An Introduction with R, Second Edition*. CRC Press LLC, Philadelphia, PA, United States, 2017.

[WPG+10] N. A. Walton, M. R. Poynton, P. H. Gesteland, C. Maloney, C. Staes, and J. C. Facelli. Predicting the start week of respiratory syncytial virus outbreaks using real time weather variables. *BMC Medical Informatics and Decision Making*, 10(1), 2010.

[WWC+05] L. J. White, M. Waris, P. A. Cane, D. J. Nokes, and G. F. Medley. The transmission dynamics of groups A and B human respiratory syncytial virus (hRSV) in England & Wales and Finland: Seasonality and cross-protection. *Epidemiology & Infection*, 133(2), 2005.

[WWM+98] M. W. Weber, M. W. Weber, E. K. Mulholland, E. K. Mulholland, and B. M. Greenwood. Respiratory syncytial virus infection in tropical and developing countries. *Tropical Medicine and International Health*, 3(4), 1998.

[XZH22] Z. Xu, H. Zhang, and Z. Huang. A Continuous Markov-Chain Model for the Simulation of COVID-19 Epidemic Dynamics. *Biology*, 11(2), 2022.

[YC11] Reza Yaesoubi and Ted Cohen. Generalized Markov models of infectious disease spread: A novel framework for developing dynamic health policies. *European Journal of Operational Research*, 2011.

[YCLS15] W. Yang, B. J. Cowling, E. H. Y. Lau, and J. Shaman. Forecasting Influenza Epidemics in Hong Kong. *PLOS Computational Biology*, 11(7), 2015.

[YKH+13] R. Yaari, G. Katriel, A. Huppert, J. B. Axelsen, and L. Stone. Modelling seasonal influenza: The role of weather and punctuated antigenic drift. *Journal of The Royal Society Interface*, 10(84), 2013.

[YKL+21] H. Yuan, S. C. Kramer, E. H. Y. Lau, B. J. Cowling, and W. Yang. Modeling influenza seasonality in the tropics and subtropics. *PLOS Computational Biology*, 17(6), 2021.

[YL73] J. A. Yorke and W. P. London. Recurrent outbreaks of measles, chickenpox and mumps: II. Systematic differences in contact rates and stochastic effects. *American Journal of Epidemiology*, 98(6), 1973.

[YLS15] W. Yang, M. Lipsitch, and J. Shaman. Inference of seasonal and pandemic influenza transmission dynamics. *Proceedings of the National Academy of Sciences*, 112(9), 2015.

[ZTV+12] H. Zhou, W. W. Thompson, C. G. Viboud, C. M. Ringholz, P. Cheng, C. Steiner, G. R. Abedi, L. J. Anderson, L. Brammer, and D. K. Shay. Hospitalizations Associated With Influenza and Respiratory Syncytial Virus in the United States, 1993–2008. *Clinical Infectious Diseases*, 54(10), 2012.