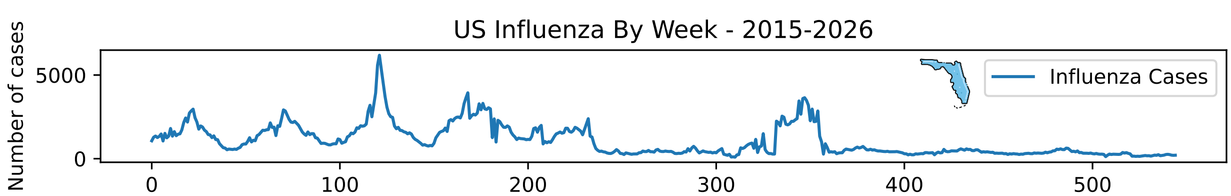
STATS 531 - Final Project

1 Introduction

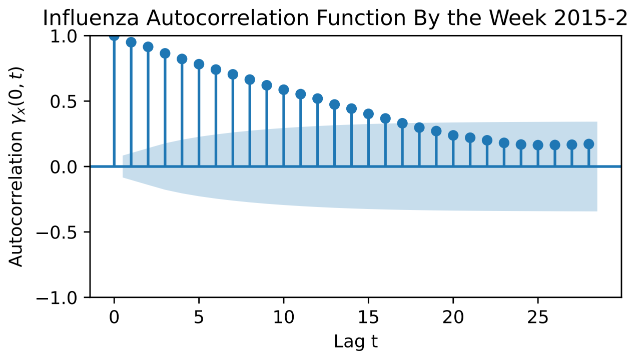
Influenza, particularly with respect to the state of Florida, is a contagious respiratory illness caused by viruses that infect tissues in the nose, throat, and lungs (Centers for Disease Control and Prevention 2026). While preventable by vaccine, it causes mild to severe illness, and even death. Influenza has many unique characteristics for which concepts have been modeled in time series. Adults infect others up to one day before symptoms begin; simulation based inference for epidemiological dynamics have components, such as the Exposed compartment in the SEIR model, that handle this. In this report, we provide a comprehensive overview of time series models, from autoregressive models to state models such as partially observable markov processes, that simulates the spread of influenza in the state of Florida over the weeks of a year from 2015 to 2026; we compare likelihoods of these models and run a likelihood ratio test to determine if complexities significantly improve fit, and determine which time series model is the best for influenza in Florida. Florida sits at a population of approximately 23 million people. We obtain the total number of influenza cases per week from the Florida Influenza Surveillance reports as per FloridaHealth.gov, and conduct analysis from there (Florida Department of Health 2026).

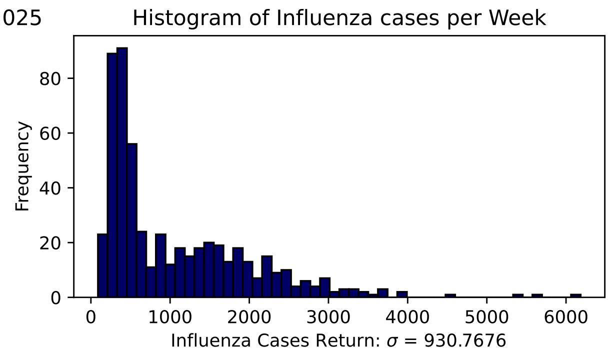
2 Exploratory Data Analysis

The impact of time series on scientific analysis spans a large list of diverse fields in which time series problems arise. Social scientists follow population series, while an epidemiologist studies the number of influenza cases observed over some time period (Shumway and Stoffer 2000). Characteristics of the time and frequency domain approaches provide insight to how the number of influenza cases relate to the previous lags in the series and how the cycles of influenza cases can be represted; these are critical insights. The table below shows time series data of influenza cases in Florida during the 2015 to 2026 time period. Understanding the characteristics of influenza on a time series plot gives us information as how much points depend on each other, the stationarity of the data, and other features.



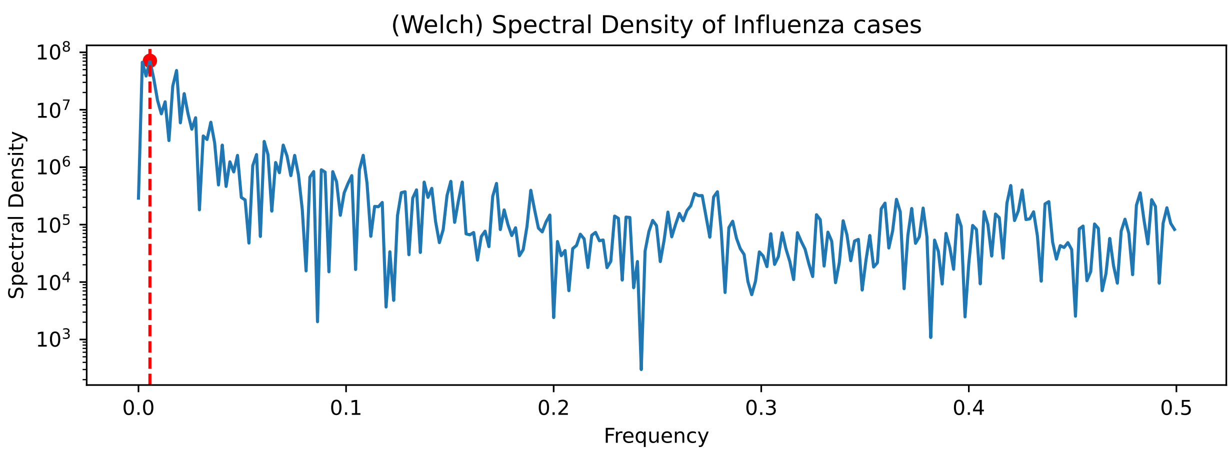
Autocorrelation is defined as $\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}}$ and measures the linear predictability of the time series (Shumway and Stoffer 2000). We see in the Autocorrelation plot below that the predictability of a point in the time series decreases as time goes on; it is easy to obtain a good prediction of the next 5 lags when dealing with influenza cases in Florida, but the subsequent predictions become harder. We show a probability density function of influenza cases in Florida below that reveals most magnitudes per week lie between 0 and 1000.





2.1 Spectral Analysis

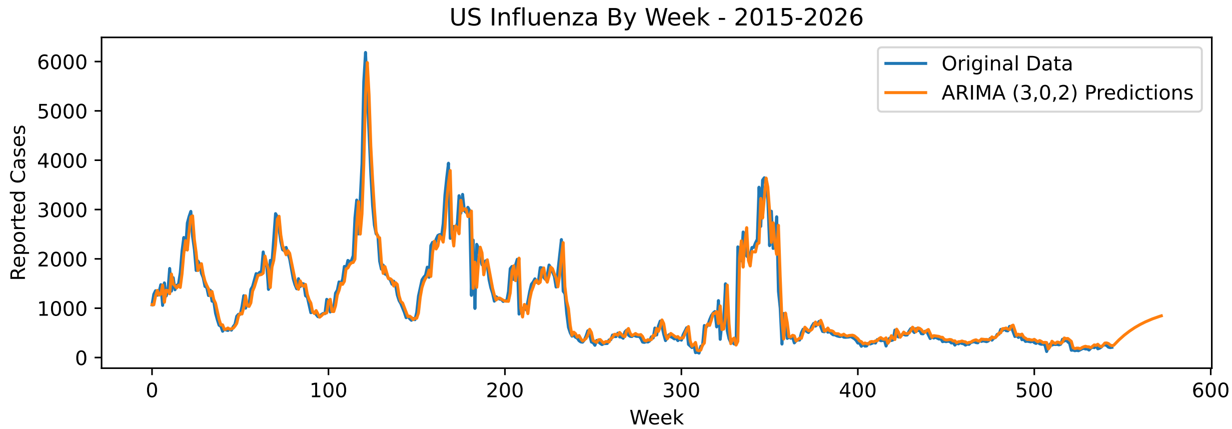
Spectral Analysis looks at the influenza cases problem with respect to the frequency domain (Shumway and Stoffer 2000). How do we represent influenza outbreaks as cycles? What is the dominant frequency of influenza cases per week? We graph a periodogram below that shows the spectral density, measured as $\lambda(\omega) = \int_{\infty}^{-\infty} \gamma(x)e^{-2\pi i\omega x}$ (Ionides 2026c). The most common oscillation of high to low influenza cases is 0.0055, which is a period of 181.67. The Periodogram shows us that this is the most common period, as spectral density falls to only half as much shortly after. The benefit of solving the problem in the frequency domain is that we can see influenza waves oscillate over a period of 181 days.



3 Regression Model with ARMA Errors

An autoregressive model ARMA(p, q) represents the current value in terms of the previous p values and a moving average of order q (Shumway and Stoffer 2000). When autocorrelation of the previous few lags is high, an ARMA model makes it easy to predict the next value in the series, as is the cases with influenza in Florida. It is also easier to We fit an ARIMA(3,0,2) model to the data and overlay the real data with predictions of 11 years of data from 2015. After fitting it, the best model with parameters ϕ and θ filled in can be written as

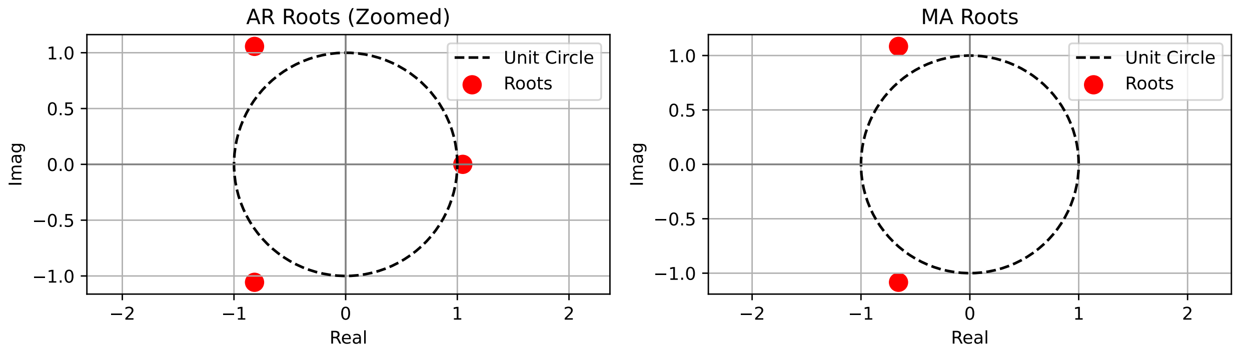
$$Y = 1071.880555 + 0.037676x_1 + 0.311855x_2 + 0.534833x_3 + 0.817041\epsilon_1 + 0.623375\epsilon_2$$



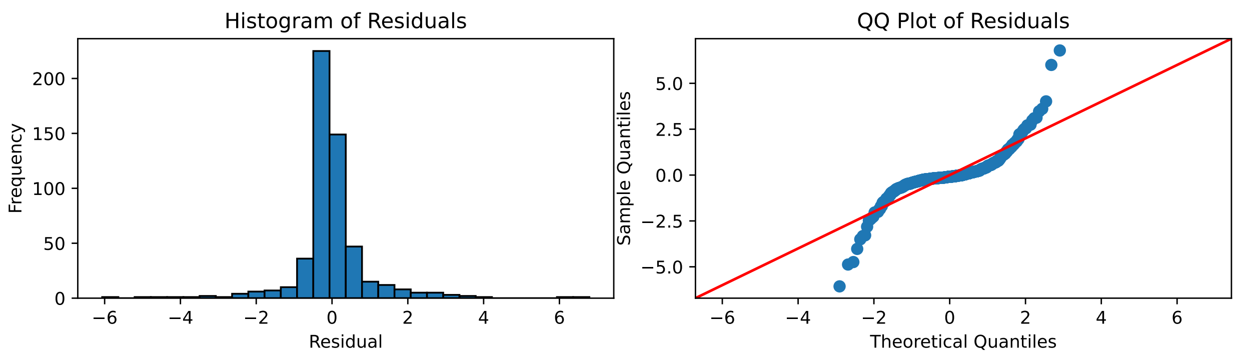
The ARMA (3,2) model pictured above returns an AIC of 7713.54 and a log likelihood of -3849.77. Aikake's Information Criterion (AIC) compares models by their complexity and likelihood; it is defined as $-2\ell(\hat{\theta})+2D$ (Ionides 2026b). ARMA (3,2) has among the lowest AIC's of all configurations of ARMA models; we choose it as it has one of the lowest AIC's, and is low on number of parameters, which limits complexity; this takes Occam's Razor into consideration.

MA(q) AR(p)	0	1	2	3	4	5
0	9000.89	8513.15	8202.48	8039.04	7944.02	7865.15
1	7730.08	7725.11	7713.28	7713.84	7715.68	7717.66
2	7722.84	7717.64	7713.83	7715.72	7717.68	7719.11
3	7712.57	7714.54	7713.54	7715.49	7703.30	7704.99
4	7714.52	7716.58	7715.47	7717.52	7713.08	7706.88

Now we look at the unit roots. The graphs below reveal that our ARMA model is both causal and invertible, as they lie outside the unit circle.



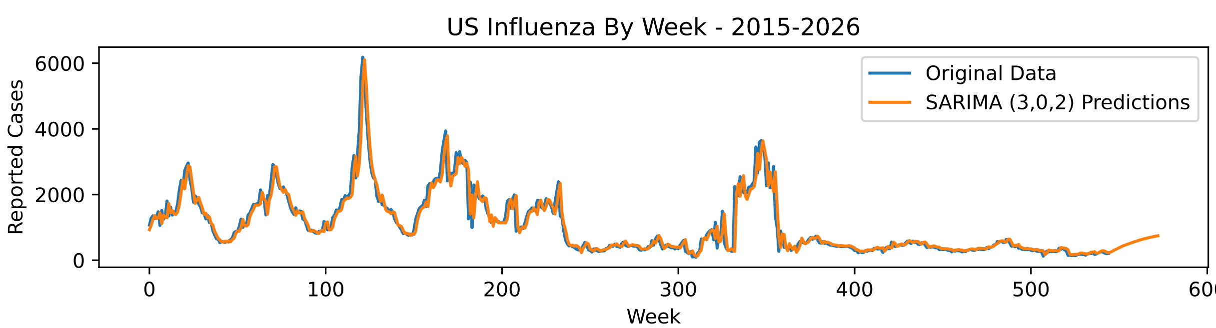
The residuals are normally distributed. We therefore conclude that ARMA (3,2) is a sufficient model for the influenza problem.



4 Regression Model with SARMA Errors

An seasonal autoregressive model SARMA(p, q)(P, Q) represents the current value in terms of the previous p values, a moving average of order q, and a seasonal component. For example, we can represent the current value in terms of lags 1, 12, and 13, instead of all 13 lags. We fit a SARMA(3,0,2)(1, 12) model to the data and overlay the real data with predictions of 11 years of data from 2015; the log likelihood is -3848.14.

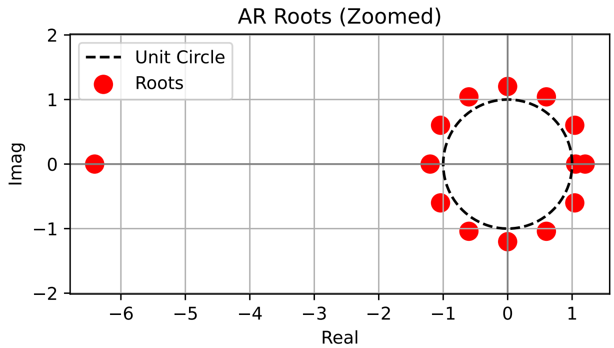
$$y_t = 49.377 + 0.792y_{t-1} + 0.148y_{t-2} + 0.109y_{t-12} + 0.063\epsilon_{t-1} + 0.150\epsilon_{t-2} + \epsilon_t$$

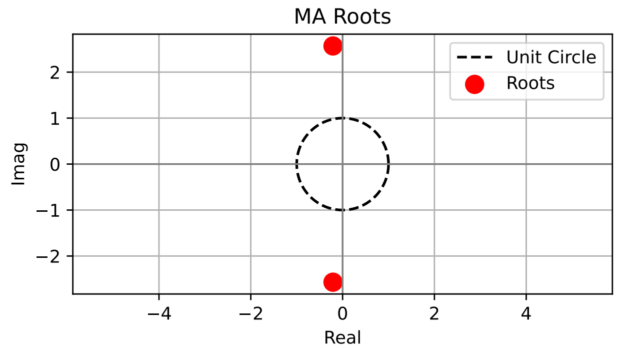


We take the lag every 12th week, to represent a seasonal component. We take the AIC errors for all combinations of p and q that meet this combination, as listed in the table below.

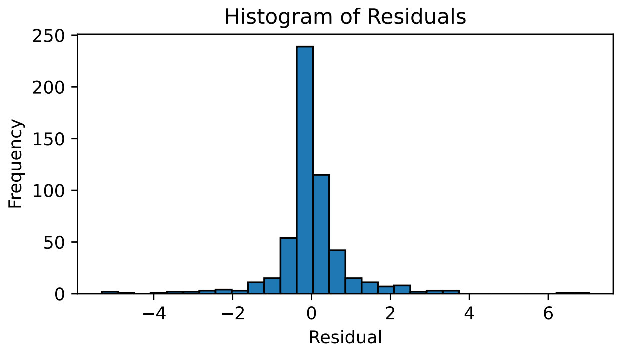
MA(q) AR(p)	0	1	2	3	4	5
0	8831.09	8380.77	8141.65	8005.67	7908.94	7841.03
1	7723.72	7720.42	7708.43	7709.89	7711.76	7713.80
2	7718.59	7714.57	7710.28	7711.84	7713.79	7715.71
3	7708.42	7710.42	7710.28	7712.02	7704.71	7705.18
4	7710.43	7712.43	7712.10	7709.30	7713.06	7718.32

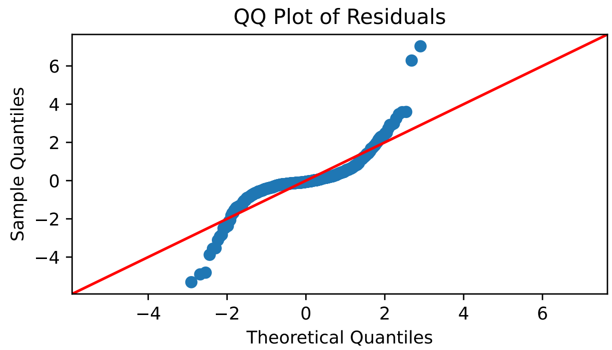
The unit roots for the autoregressive and moving average polynomial roots lie outside of the unit circles. This indicates that the model is causal and invertible.





The residuals are also normally distributed and low. We therefore conclude that the SARMA model handles influenza data effectively.





5 Partially Observable Markov Processes (POMP)

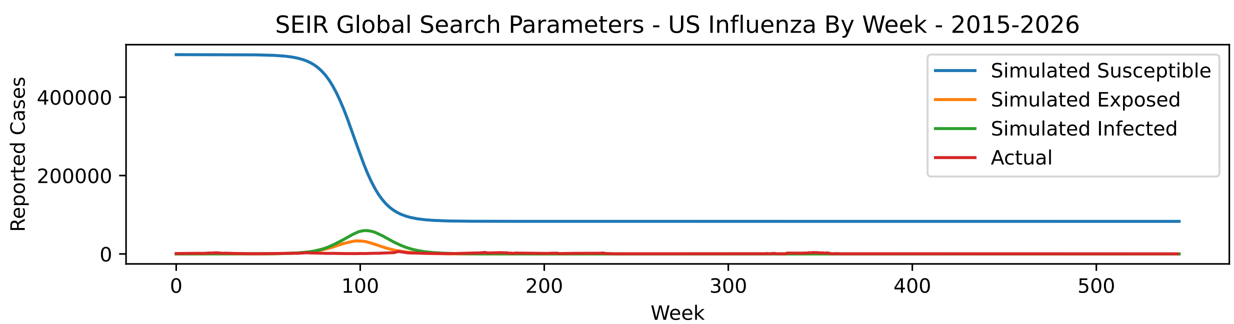
A partially observable markov process is a state model that models real world conditions via unseen states and produces seen results. These models can be replicated through PyPomp. Real world dynamics are simulated through hidden, latent states inside the model, as represented by a hidden markov model. The numer of infected individuals is then observed as y_n, an observable state dependent on the latent state x_n. An SIR model demonstrates how susceptible individuals become infected over time given transmission and recovery rates. The functions below outline an SEIR model, and define the compartments and rates at which individuals leave them (Ionides 2026a).

$$S(t) = S(0)-NSE(t); E(t) = E(0)+NSE(t)-NEI(t); I(t) = I(0)+NEI(t)-NIR(t); R(t) = R(0)+NIR(t)$$

$$NSE(t) \sim Binom(S, 1 - exp(-\beta * I/N * dt); NEI(t) \sim Binom(E, 1 - exp(-\sigma * dt))$$
$$NIR(t) \sim Binom(I, 1 - exp(-\mu_{IR} * dt))$$

Local Search and Global Search determine our parameters such that our NSE(t), NEI(t), and NIR(t) functions, the functions that decide how many individuals move between compartments each turn, are optimized. These functions rely on parameters that give the probability that any individual moves to another component; $1 - exp(-\beta * I/N * dt)$, $1 - exp(-\sigma * dt)$, and $1 - exp(-\mu_{IR} * dt)$ use three parameters, as previously defined, in addition to the number of individuals in the components, to define the probability that any individual moves to another component. The number of individuals that leave each compartment is decided each state update with a binomial distribution.

5.1 Susceptible Exposed Infected Recovered - Unmodified Model Implementation



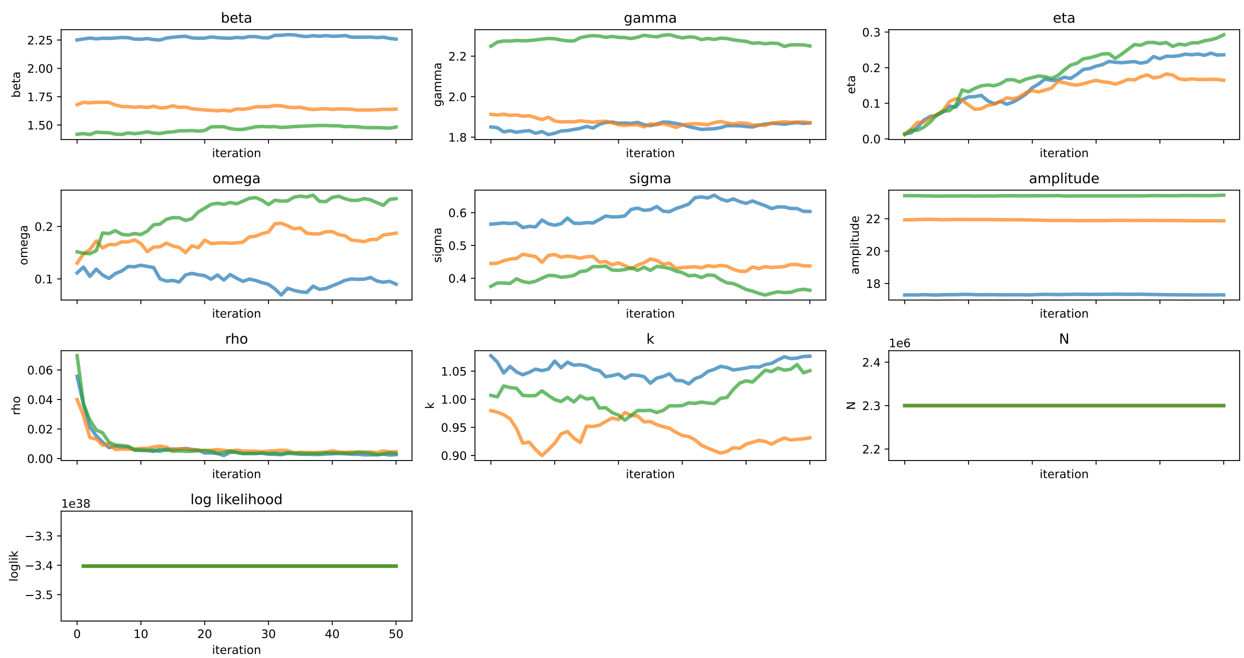
Above, we see the most basic implementation of an SEIR model against actual values. The image above uses the best parameters available as determined by a local search with original parameters, and get the thetas used for the image above. The log likelihood becomes -3.403e+38. Results from global search follow, with a likelihood of -22,588.4. The best parameters, θ, from the global search are as listed below; they ultimately come from natural numbers, using the population of Florida (23 million), a reasonable transmission rate of 2.05 as determined by our local search, a σ of 0.5, and a rate of infection μ_{IR} of 0.234 as determined by global search.

5.2 Susceptible Exposed Infected Recovered - Trigonometric Transmission Rate Implementation

The unmodified SIR model, as presented in lecture slides, reaches a low likelihood with global search, as featured above. However, while the first outbreak matches with the second one, subsequent waves suffer. We introduce a trigonometric component to an SEIR model such that the transmission rate rises and falls over time. Many state models for which parameters are not consistent across the whole time series have been the most optimal and successful (Calafiore, Novara, and Possieri 2020). We similarly define β as a function of time, specifically, through the use of a sine function. After modifying β, the other parameters stay the same (σ defines the rate individuals move from exposed to infected, ω defines the rate individuals move from to recovered to susceptible again, etc.). This model determines the number of individuals leaving a compartment with a poisson distribution.

$$\beta = \beta * (1 + 0.2 * sin(\frac{a * \pi * t}{365})$$

5.2.1 Local Search Parameter Traces



A local search finds that the best likelihood for this model is -3.403e+38 with β as 2.2505, γ as 1.85, eta as 0.015, ω as 0.112, σ as 0.56, a as 17.3, ρ, the reporting rate, as 0.056, and k, the overdispersion parameter, as 1.08. The traces for parameter values during local search per iteration are above; they are mostly all stable or converge. Global search does not have a better likelihood, and, consequentially, is featured in the Appendix.

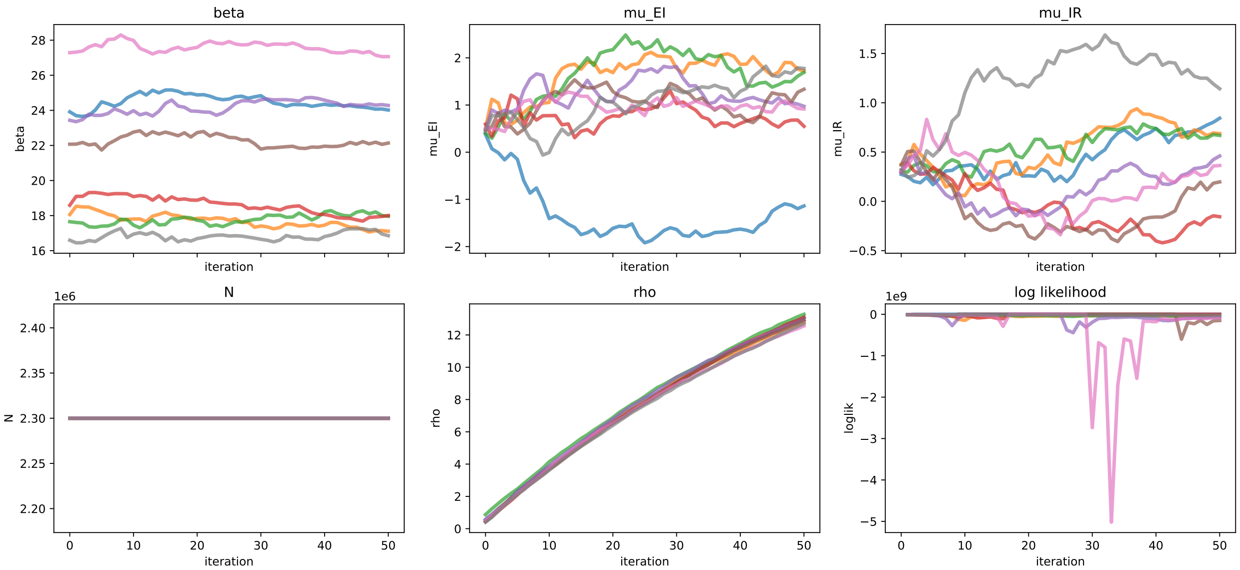
5.3 Susceptible Exposed Infected Recovered - Measured Parameters Implementation

A POMP model can take facts from nature and model many dynamic states inside before producing the observed variables; this makes it an attractive model for influenza. Previous final projects in the University of Michigan's Influenza STATS 531 course have conducted analysis on influenza in other regions. The 2025 Influenza in the Great Lakes project implemented an SEIR model using realistic rates as determined by influenza statistics; models produced with this methodology often have the highest likelihood (531a 2025). While influenza incubates for up to 4 days, an individual may be immune for 3 to 6 months after (Mordant et al. 2023). Adults report influenza at a rate of 0.5% (Mandl et al. 2010). We use the same methodology to define an SEIR model with realistic rates of progression from exposed to infectious, rates of recovery, rate of immunity, reporting rates, and a transmission rate dependent on several variables. We define all four components, S, E, I, and R, and their rates of progression, as listed below.

1. μ_{EI} is the rate of progression of individuals from the exposed compartment to the infectious compartment, and is represented as the reciprocal of the incubation period. The incubation period for influenza is 1 to 4 days (Centers for Disease Control and Prevention, 2025). We therefore start our search for μ_{EI} in the range of $[\frac{1}{1/4}, \frac{1}{1/7}]$, simplified as $[1.75, 7]$ for μ_{EI}.
2. μ_{IR} is the rate of recovery, or rather, how long it takes for an individual to move from infected to recovered. The CDC approximates the time to recover from influenza as 5 to 7 days, with an extra 3 or 4 for recovery (Centers for Disease Control and Prevention, 2025). We start our search for μ_{IR} in the range of $[8, 11]$.
3. μ_{SR} is the rate of immunity before an individual may become infected again. This generally lasts from 3 to 6 months (Mordant et al. 2023). The model was more successful after removing the immunity component μ_{SR}; it is therefore not included in our optimization searches.
4. The reporting rate, ρ. The reporting rate is the fraction of actual cases that are actually reported. Before perturbing parameters, we start with a reporting rate of 0.5%; this number has been used in previous literature (Mandl et al. 2010).
5. Finally, μ_{SE} is the rate at which individuals become infected. It is dependent on the parameter β. The probability of any individual becoming infected during the transition from state x_n to x_{n+1} is $1 - e^{-\beta * I/N * dt}$, or rather, the reciprocal of e raised to the power of the transmission rate times the proportion of infected individuals times the rate of changed.

The values above define the probabilities of individuals to move to another component, and we draw from a binomial distribution each iteration. The results are, overall, very successful, yielding a model with a likelihood of -26,203.87 from local search and a model with a likelihood of -4,687,896.5 from global search. The local search parameter traces is featured below; global search is in the appendix.

5.3.1 Local Search



The log likelihood clearly converges to a high by the 50th iteration. After a local search, we find that the optimal log likelihood is -26,203.87, with N as 2300000, β as 23.92, μ_{EI} as 0.38, μ_{IR} as 0.274, and ρ as 0.57.

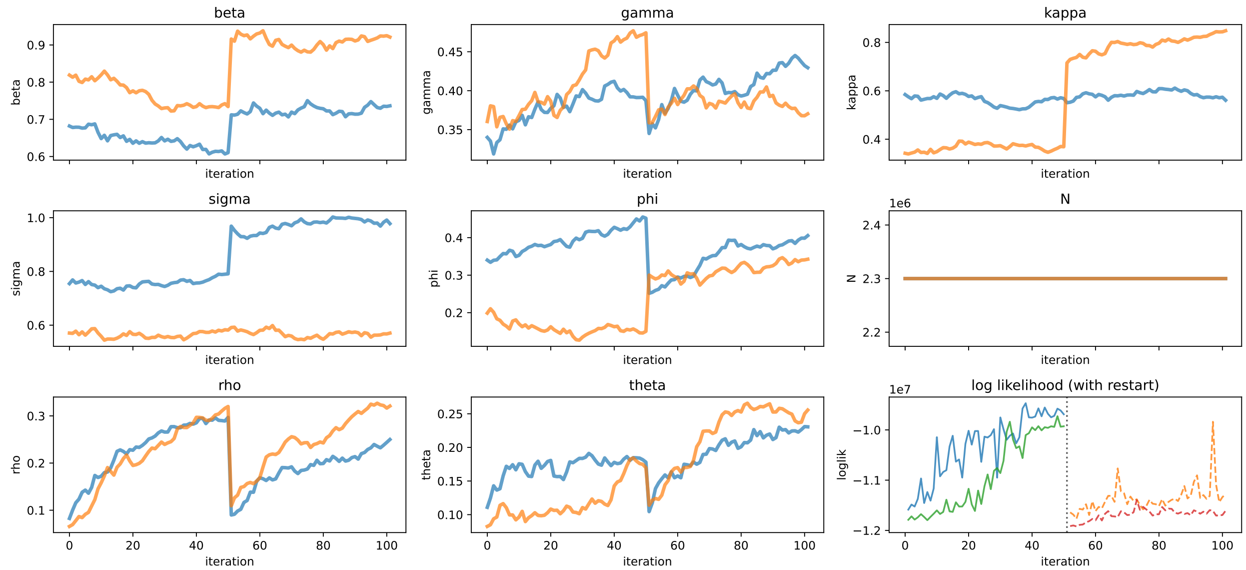
5.4 Susceptible Exposed Infected Hospitalized Dead Recovered Model

The Hong Kong Special Administrative Region of China (HKSAR) experienced a surge of COVID-19 cases in April 2022. The peak of influenza transmission rates favor the form of the COVID-19 transmission rates in a study using state models to graph the outbreaks (Wen et al. 2024). Influenza is seasonal and influenced by weather, school terms, population movement, and air pollutants - transmission shares similarities to that of COVID-19 transmission. We replicate a model used for simulating weekly disease number cases in Hong Kong, using the mechanisms as suggested in the paper. (Wen et al. 2024). The state model features 6 compartments, susceptible, exposed, infected, hospitalized, dead, and recovered; factors such as the percentage of population susceptible to a strain of the virus (ex, Omnicron), also play a role in the model. We define the 6 compartments, and the number of individuals that move between them, below. The parameter β denotes transmission rate, σ denotes the infectiousness onset rate, γ denotes the infectiousness decay rate, ϕ denotes infection-hospitalization-ratio, θ denotes hospitalization-fatality-ratio, k denotes the rate at which individuals leave the hospitalized compartment (Wen et al. 2024).

$$S_{new} = \frac{-\beta}{N}SI; E_{new} = \frac{-\beta}{N}SI - \sigma E; I_{new} = \sigma E - \gamma I; H_{new} = \phi\gamma I - kH; D_{new} = \theta kH; R_{new} = (1-\phi)\gamma I + (1-\theta)kH$$

Local Search produces a model with a likelihood of -8,256,150.0, and global search with a likelihood of -9,467,970.0. We show the parameter trace for local search in the appendix, and global search below.

5.4.1 Global Search



After a global search, we find that the best log likelihood converges to -9,467,970 in all runs. Our best parameters in this configuration include β as 0.712, σ as 0.968, γ as 0.345, κ as 0.55, ϕ as 0.252, θ as 0.104, and ρ as 0.0899.

6 Likelihood Ratio Tests

Table 3 shows log likelihoods of all of each model from each optimization method.

Model	Log Likelihood	Local Search	Global Search
ARMA	-3,849.77	NA	NA
SARMA	-3,848.14	NA	NA
SEIR (Unmodified)	NA	-3.403×10^{38}	-22,588.4
SEIR (Trigonometric β	NA	-3.403×10^{38}	-3.4023×10^{38}
SEIR (Measured)	NA	-26,203.87	-4,687,896.5
SEIHDR	NA	-8,256,150.0	-9,467,970.0

Table 3: Log-likelihood values for all models.

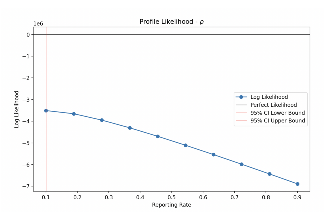
Table 4 shows 3 likelihood ratio tests; we use the likelihood ratio statistic as defined $LRT = -2(L(\hat{\theta_0}) - L(\hat{\theta_A}))$ (GeeksforGeeks 2025). We reject the Global Search results in favor of the Local Search Results for SEIR (Measured), we reject SEIR (Unmodified) in favor of the ARMA model, and we fail to reject SARMA against ARMA. We therefore conclude that, while we may look at the results of the best optimization method and discard the rest, some models, such as ARMA and SARMA need to be compared. Nevertheless, our best state model is rejected in favor of ARMA; we conclude that the autoregressive models are superior for this problem.

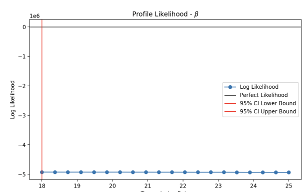
Hypotheses	LRT Equation	LRT Statistic	p-value / Decision
SARMA vs ARMA	$-2(L(\hat{\theta_0}) - L(\hat{\theta_A}))$	3.26	$0.196 > 0.05$ (Fail to reject H_0)
SEIR (Local) vs SEIR (Global)	$-2(L(\hat{\theta_0}) - L(\hat{\theta_A}))$	9,323,385.26	$0 < 0.05$ (Reject H_0)
ARMA vs SEIR (Unmodified)	$-2(L(\hat{\theta_0}) - L(\hat{\theta_A}))$	37,477.26	$0.0001 < 0.05$ (Reject H_0)

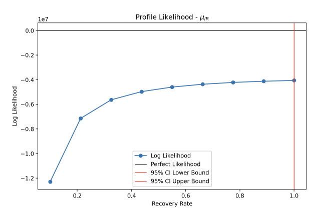
Table 4: Likelihood Ratio Tests

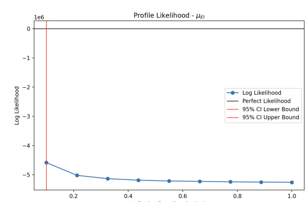
7 Profile Plots

As per the likelihood ratio tests, the SEIR Measured Parameter Implementation features one of the best likelihoods. We therefore take the best parameters, as defined by local search, and determine the highest likelihoods for any value of the transmission rate and the reporting rate. We can stay 95% confident that [18, 18], [0.1, 0.1], [1.0, 1.0], and [0.1, 0.1] are the best parameter value ranges, naively assuming independence, for the transmission rate β, the reporting rate ρ, the recovery rate μ_{IR}, and the incubation infection period μ_{EI} respectively; the code to produce these plots is featured in cdc_seir_implementation.py. We include a two dimensional likelihood surface in the Appendix, and matches our local search results more.









8 Context of Other Projects

Other projects in STATS 531 have analyzed influenza on autoregressive and state models. In this project, we focus on the region of Florida. Influenza cases follow a stationary distribution, and are modeled well on most time series models. Most projects in previous years use real world measurements to model state compartments, such as maintaining an accurate incubation period for influenza. We use the latest information to create our own SEIR Measured Parameter Implementaion, as well as introduce variants of SEIR with components to track states from modern literature, which are on par with our best models.

9 Conclusion

We therefore conclude that SARMA is our best model to measure number of influenza cases a week with; ARMA is not far behind. The SEIR Measured Parameter Implementation is the best state model, with realistic transmission rates and flows; both global and local search optimization methods output models with high likelihoods. Overall, both classes of models are effective at modeling influenza cases, but, as per the likelihood ratio tests, we have statistically significant evidence that a SARMA model has higher likelihood than any of the state models.

10 References

531a, STATS. 2025. "Unveiling the Dynamics of Influenza in the Great Lakes Region." 2025. https://ionides.github.io/531w25/final_project/project01/blinded.html.

Calafiore, Giuseppe C., Carlo Novara, and Corrado Possieri. 2020. "A Time-Varying SIRD Model for the COVID-19 Contagion in Italy." *Annual Reviews in Control* 50: 361–72. https://doi.org/10.1016/j.arcontrol.2020.10.005.

Centers for Disease Control and Prevention. 2026. "About Influenza (Flu)." 2026. https://www.cdc.gov/flu/about/index.html.

Florida Department of Health. 2026. "Influenza Surveillance." 2026. https://www.floridahealth.gov/statistics-data/population-surveillance/influenza-surveillance/.

GeeksforGeeks. 2025. "Likelihood Ratio Test in r." https://www.geeksforgeeks.org/r-language/likelihood-ratio-test/.

Ionides, Edward. 2026a. "Chapter 2: Simulation of Stochastic Dynamic Models." Lecture notes for STATS 531, Winter 2026.

———. 2026b. "Modeling and Analysis of Time Series Data Chapter 5 – Parameter Estimation and Model Identification for ARMA Models." Lecture notes for STATS 531, Winter 2026.

———. 2026c. "Modeling and Analysis of Time Series Data: Chapter 7 – Introduction to Time Series Analysis in the Frequency Domain." Lecture notes for STATS 531, Winter 2026.

Mandl, Kenneth D., J. Marc Overhage, Mark M. Wagner, William B. Lober, Paola Sebastiani, Farzad Mostashari, Julie A. Pavlin, Paul H. Gesteland, Thomas Treadwell, and Eric Koski. 2010. "Rapid Detection of Pandemic Influenza in the Presence of Seasonal Influenza." *BMC Public Health* 10: 726. https://doi.org/10.1186/1471-2458-10-726.

Mordant, FL, OH Price, R Rudraraju, MA Slavin, C Marshall, LJ Worth, H Peck, IG Barr, SG Sullivan, and K Subbarao. 2023. "Antibody Titres Elicited by the 2018 Seasonal Inactivated Influenza Vaccine Decline by 3 Months Post-Vaccination but Persist for at Least 6 Months." *Influenza and Other Respiratory Viruses* 17 (1): e13072. https://doi.org/10.1111/irv.13072.

Shumway, Robert H., and David S. Stoffer. 2000. *Time Series Analysis and Its Applications*. Springer.

Wen, Li, Yi Yin, Qiong Li, Zhihang Peng, and Daihai He. 2024. "Modeling the Co-Circulation of Influenza and COVID-19 in Hong Kong, China." *Advances in Continuous and Discrete Models* 2024 (1): 32. https://doi.org/10.1186/s13662-024-03830-7.

11 Appendix

11.1 2D Likelihood Surface

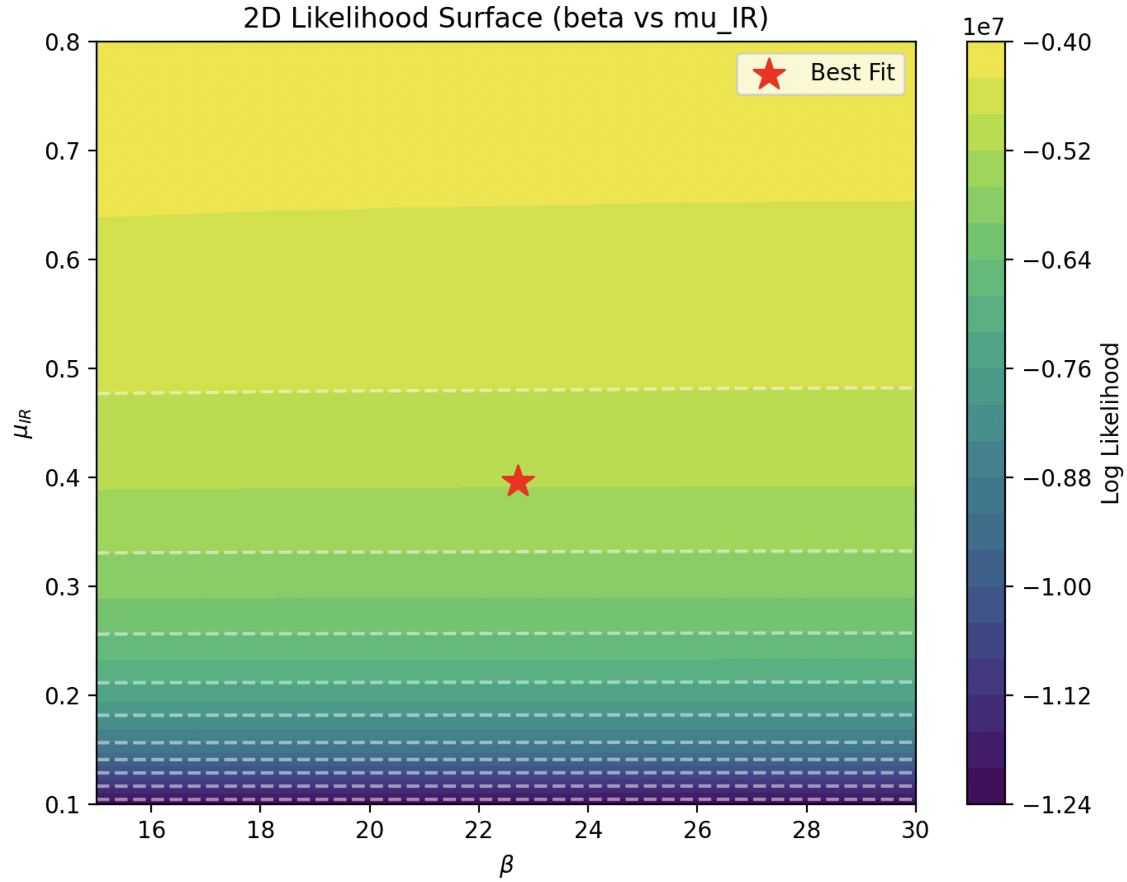


Figure 1: Likelihood Surface Test for the Measured Parameter SEIR Model. The optimal parameter combination involves a β of 22.7 and a μ_{IR} of 0.396, as indicated by the results of the local search.

11.2 Augmented Dickey-Fuller Test

$$H_0 : \text{The process } \{x_t\} \text{ is nonstationary}$$

$$\exists\, t, h, k \text{ such that } (x_t, x_{t+1}, \ldots, x_{t+k}) \overset{d}{\neq} (x_{t+h}, x_{t+1+h}, \ldots, x_{t+k+h})$$

$$H_A : \text{The process } \{x_t\} \text{ is stationary}$$

$$\forall\, t, h, k,\ (x_t, x_{t+1}, \ldots, x_{t+k}) \overset{d}{=} (x_{t+h}, x_{t+1+h}, \ldots, x_{t+k+h})$$

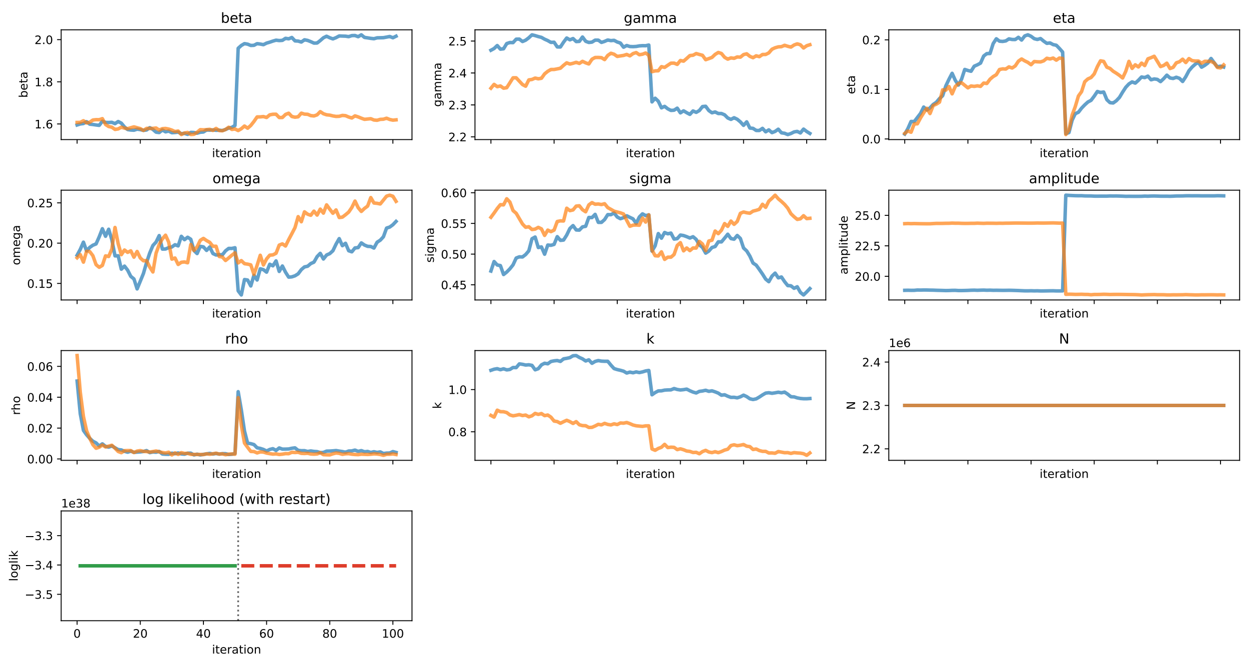
	0
Series	Influenza Cases
ADF Statistic	-3.407463
p-value	0.010707
Lags Used	13
Observations	531
Critical Value (5%)	-2.866998
Reject the H0?	Yes

We reject the null and conclude that the time series is stationary.

11.3 Parameter Traces

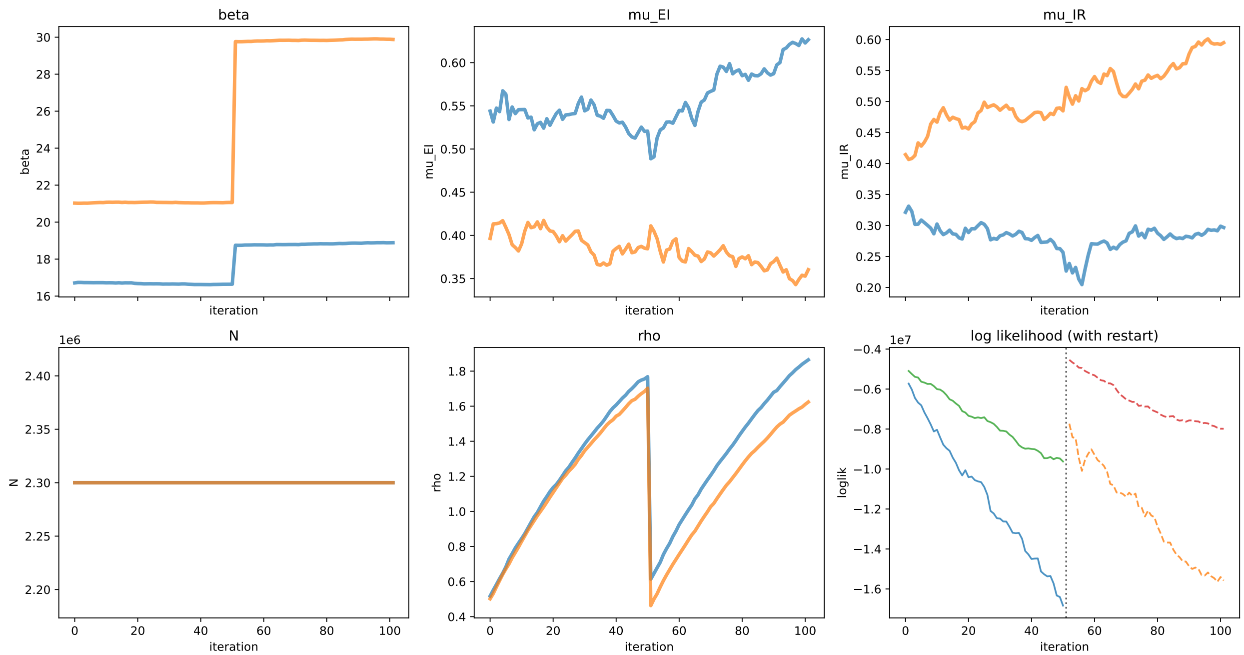
The most successful parameter traces are listed in the main content of the report. Yet, parameter traces for global and local searches with less beneficial results are useful to the reader for purposes of completion; we show them in the Appendix.

11.3.1 SEIR Poisson Global Search



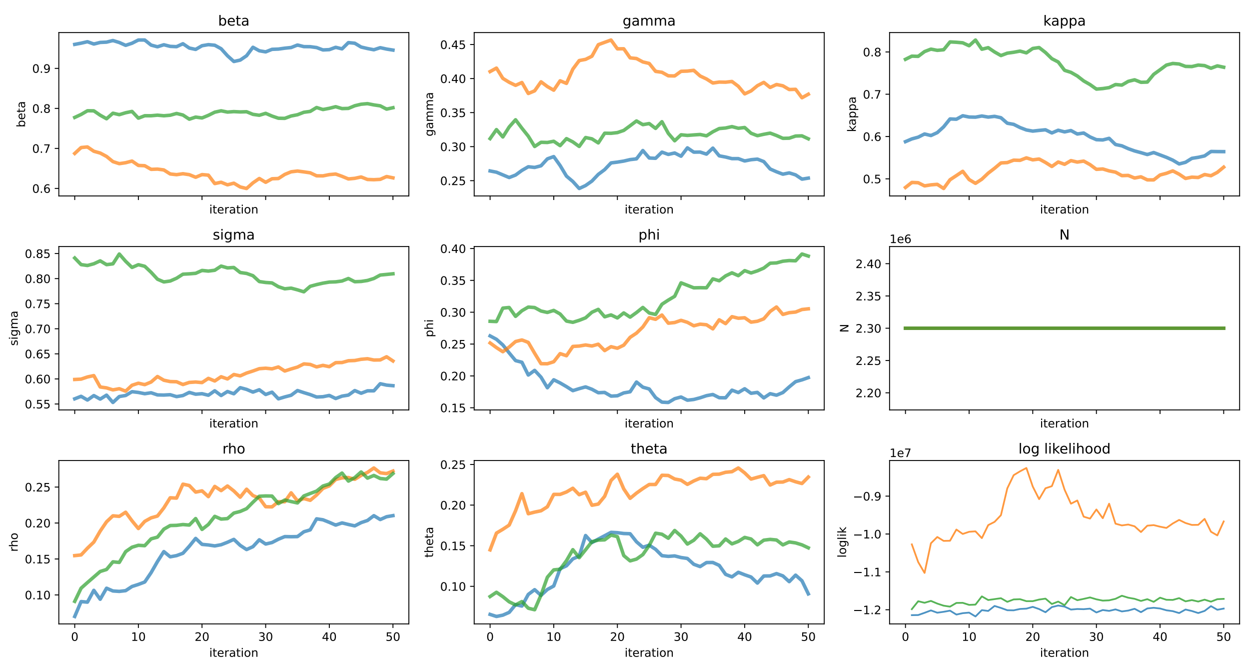
A global search finds that the best likelihood for this model is -3.4023e+38 with β as 1.595, γ as 2.471, eta as 0.00987, ω as 0.185, σ as 0.4724, amplitude as 18.85, ρ as 0.0505, and k as 1.0912.

11.3.2 SEIR CDC Fact Based Implementation Global Search



After a global search, we find that the optimal log likelihood is -4687896.5, with N as 2300000.0, β as 22.70865335380483, μ_{EI} as 0.39714361142110044, μ_{IR} as 0.3961116621569491, and ρ as 0.5058308187492441.

11.3.3 Local Search Susceptible Exposed Infected Hospitalized Dead Recovered Model



After a local search, we find that the best log likelihood is -8256150.0. Our best parameters in this configuration include β as 0.6874075000382809, γ as 0.4102779291683628, σ as 0.5988763233118153, κ as 0.4795870382337614, ϕ as 0.2517532607848387, θ as 0.1448753065414705, and ρ as 0.15465102818467713.

11.3.4 Verhulst Model Application

We model the beginning of an influenza wave in Florida against an SIR model based on the population growth model, the Verhulst Model. The Verhulst is a birth and death process introduced by Pierre Francois Verhulst in 1838 to describe population growth. A verhulst model defines population growth as

$$\frac{dp}{dt} = rP(1 - P/K)$$

, where r is the growth rate, K is the carrying capacity, and P is the current population. We create a POMP model and add $rP(1 - P/K) * \delta t$ each time we update the latent state (rproc). We draw our observations y_n from a poisson distribution with the current population as lambda. While the whole time series is difficult to model with just a Verhulst model, as it is meant to respect carrying capacity, it models an initial outbreak well, with a log likelihood of -178,242.58.