Frequentist approach to fitting and forecasting epidemics

Gerardo Chowell, PhD
Professor of Epidemiology and Biostatistics

SISMID 2024

Georgia State University | SCHOOL OF PUBLIC HEALTH

Compartmental ODE model

An ODE model comprised of a system of h ordinary-differential equations is given by:

$$\dot{x_1}(t) = g_1(x_1, x_2, \dots x_h, \Theta)$$

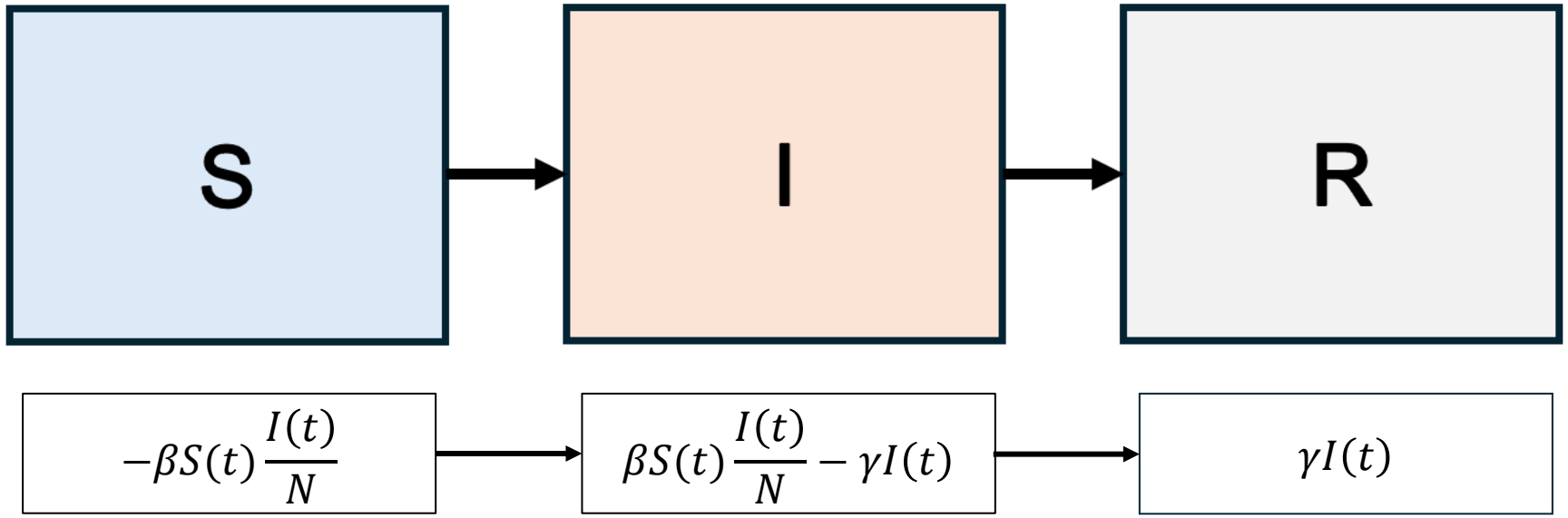
$$\dot{x_2}(t) = g_2(x_1, x_2, \dots x_h, \Theta)$$

$$\dots$$

$$\dot{x_h}(t) = g_h(x_1, x_2, \dots x_h, \Theta).$$

- $\dot{x_i}$: Rate of change of the system state x_i $(i = 1,2, \dots, h)$
- $\Theta = (\theta_1, \theta_2, \dots, \theta_m)$: Set of model parameters.
- $f(t, \Theta)$: Expected temporal trajectory of the *observed state* of the system.
  - Observed state: the specific state variable of the ODE system that has been observed or measured
  - Latent states: the ODE states that are not directly observed but are inferred from the mathematical modeling of the observed variables.

Example: SIR Model



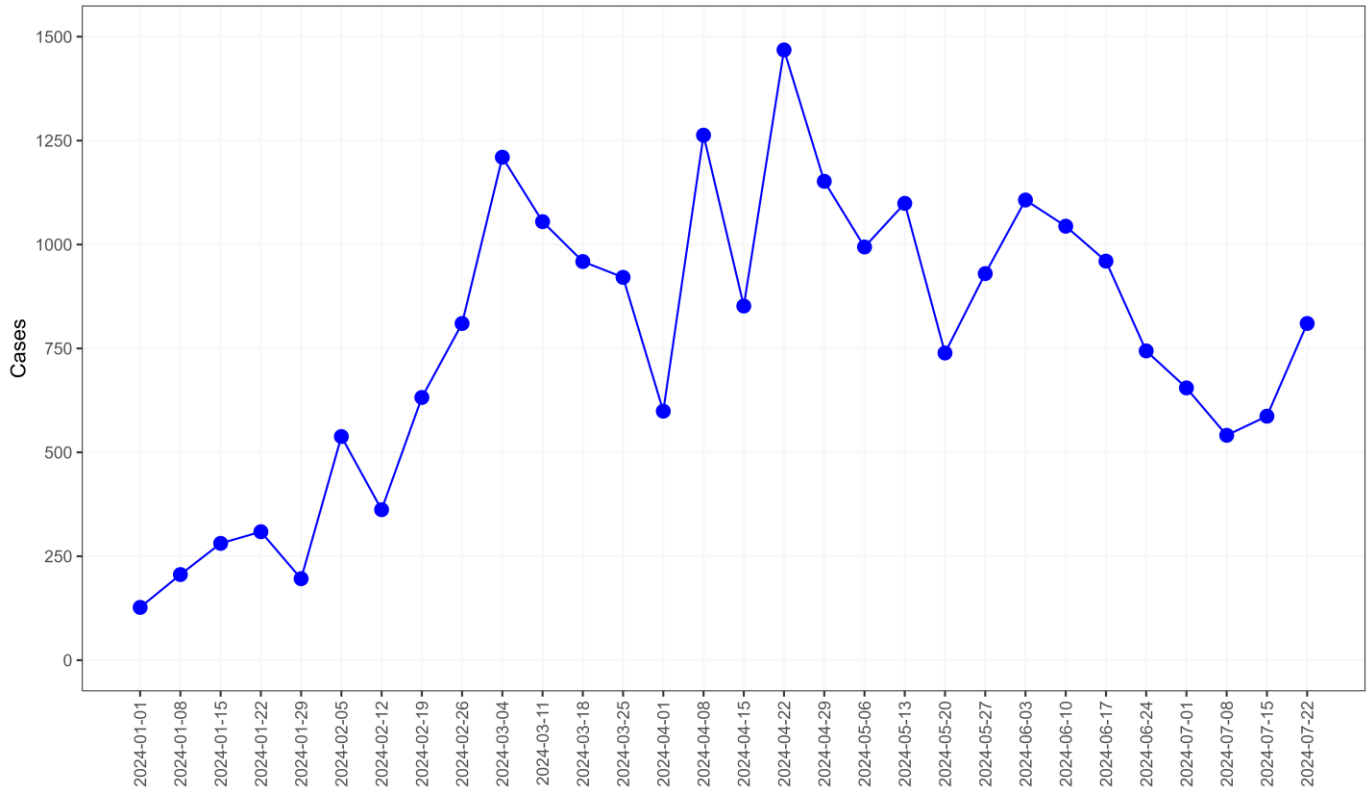
Parameters (Θ): β (transmission rate), N (population size), γ (removal rate)

State Variables ($\dot{x}_i$): $\boldsymbol{S}$usceptible, $\boldsymbol{I}$nfectious, $\boldsymbol{R}$ecovered

Note: Negative values indicate moving out of a compartment and positive values indicate inward movement

Model Calibration: Observed time series data

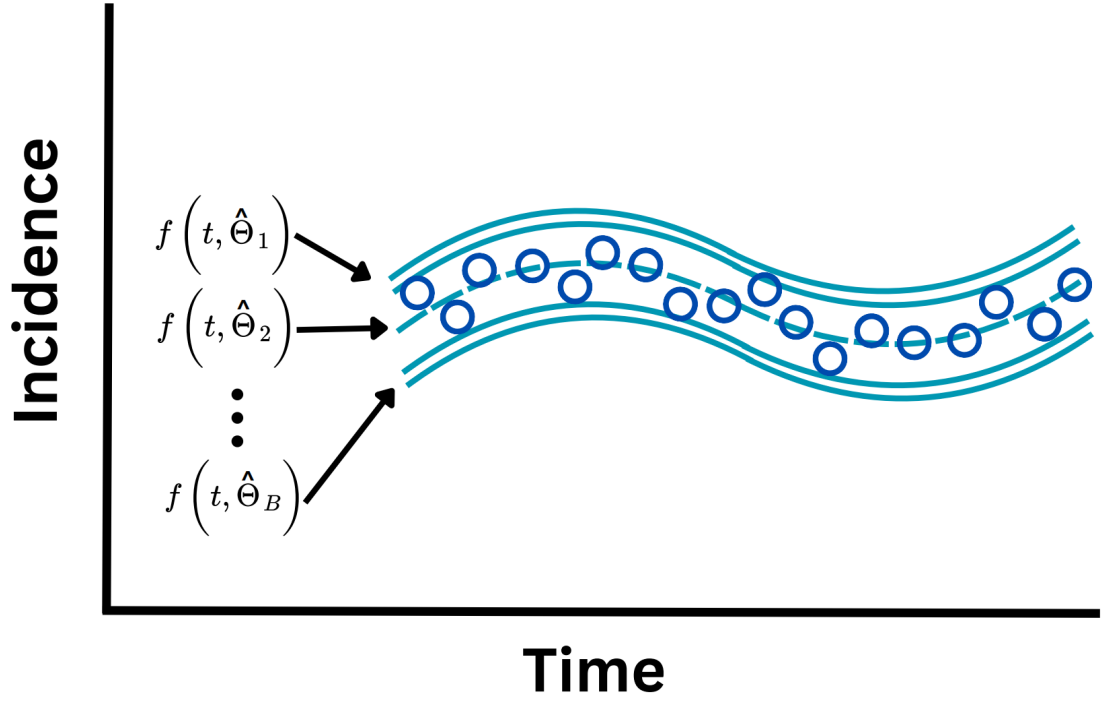
Assuming that there is a single observed state (i.e., recorded infections) let $y_{t_1}, y_{t_2}, \ldots, y_{t_n}$ denote the time series of the observed state of the system used to calibrate the model.



Here, t_j, $j = 1,2, \ldots, n$, are the time points for the time series data (i.e., dates), and n is the number of time points (etc., 29 in the above figure).

Parameter Inference

- Let $f(t, \Theta)$ denote the expected temporal trajectory of the observed state of the system.
- We can estimate the set of model parameters, denoted by Θ, by fitting the model solution to the observed data via **nonlinear least squares** or **maximum likelihood** estimation.



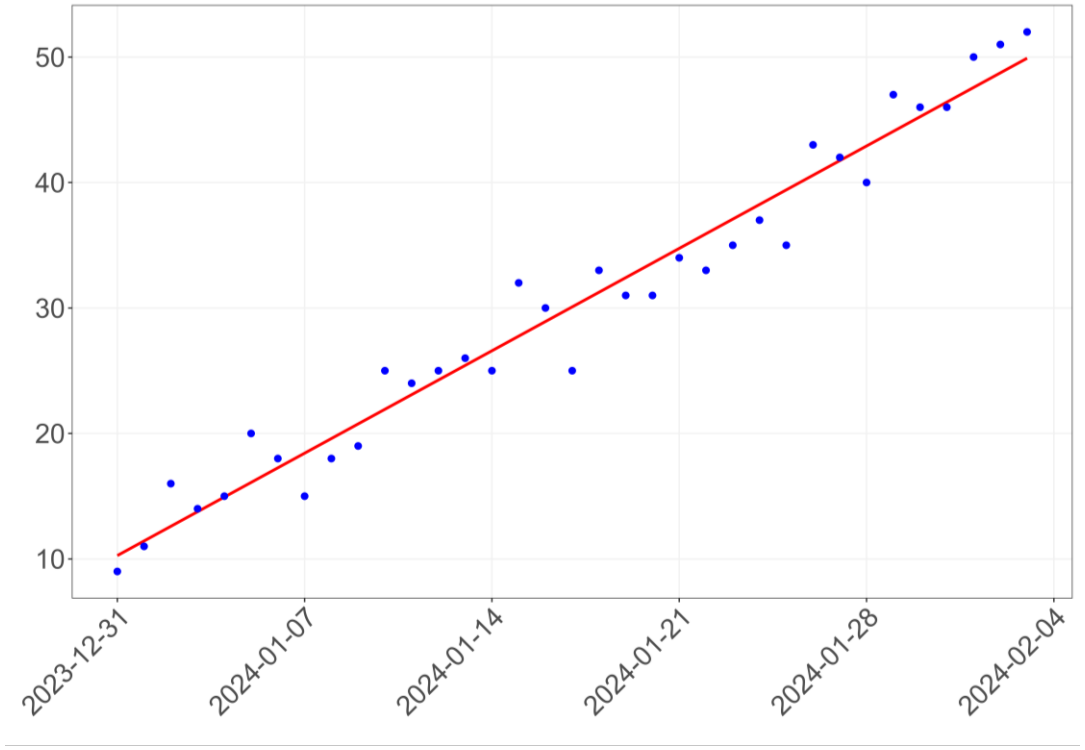
- This is the model calibration step that consists of searching for a match between observed and simulated model solutions via statistical inference.

Adapted From Chowell et al. Infectious Disease Modeling. 2017

Nonlinear least squares fitting (NLSQ)

- Nonlinear least squares estimation is achieved by searching for the set of parameters $\hat{\theta}$ that minimizes the sum of squared differences between the observed data $y_{t_1}, y_{t_2}, \ldots, y_{t_n}$ and the best fit of the model (model mean) which corresponds to $f(t, \Theta)$.

$$\widehat{\Theta} = \arg\min \sum_{j=1}^{n} (f(t_j, \Theta) - y_{t_j})^2$$

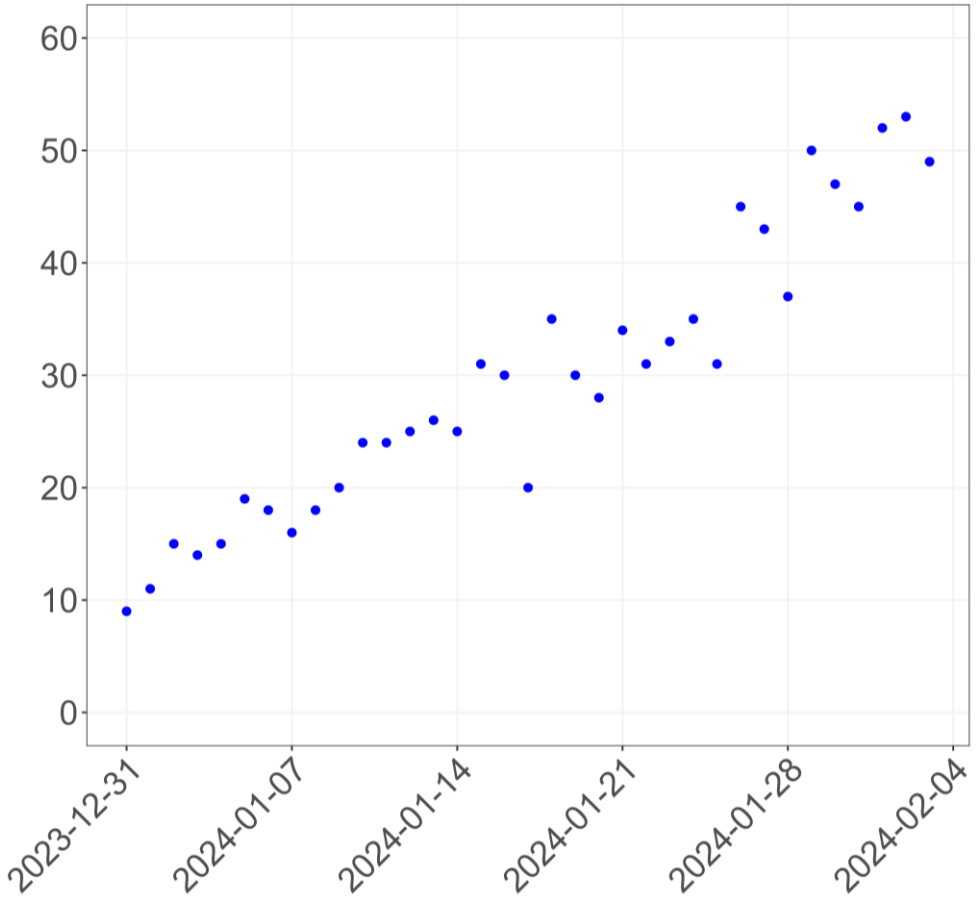


When to **NOT** use NLSQ: Errors have non-constant variance (heteroscedasticity); it can lead to inefficiency in the NLS estimates.

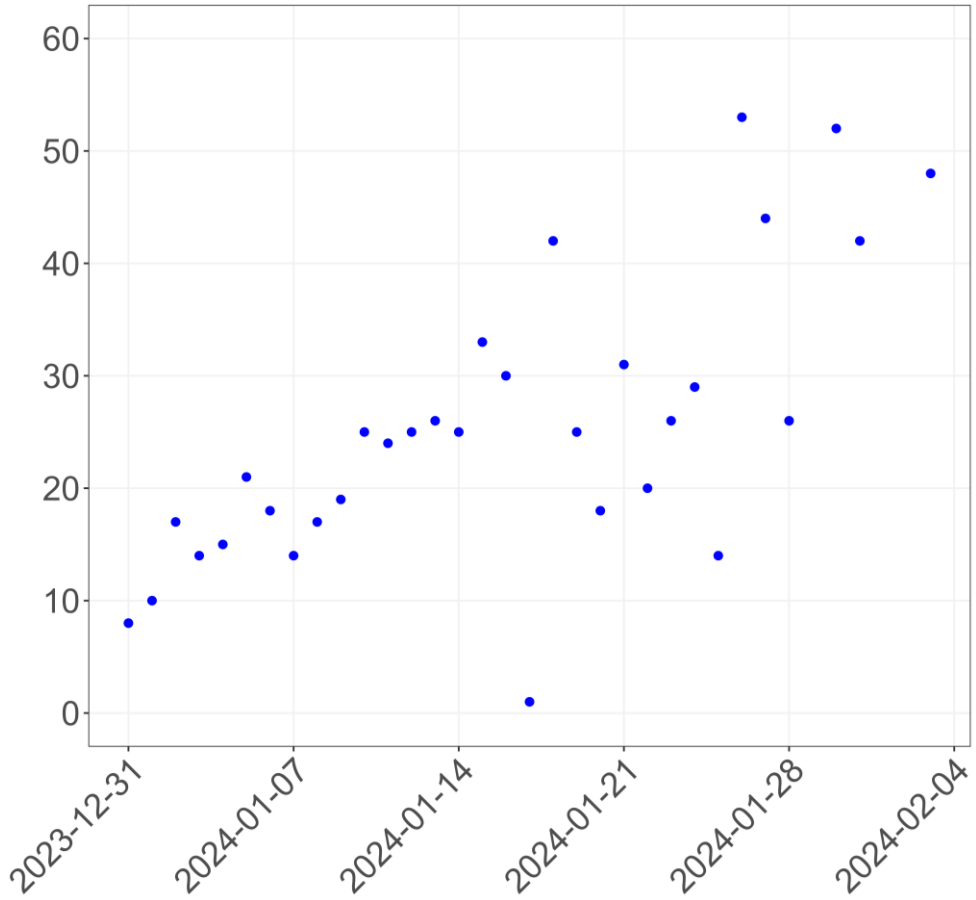
Solution: Maximum Likelihood Estimation (MLE)

Estimate parameters assuming a **specific error** structures in the data.

Poisson



Negative Binomial (Overdispersion)



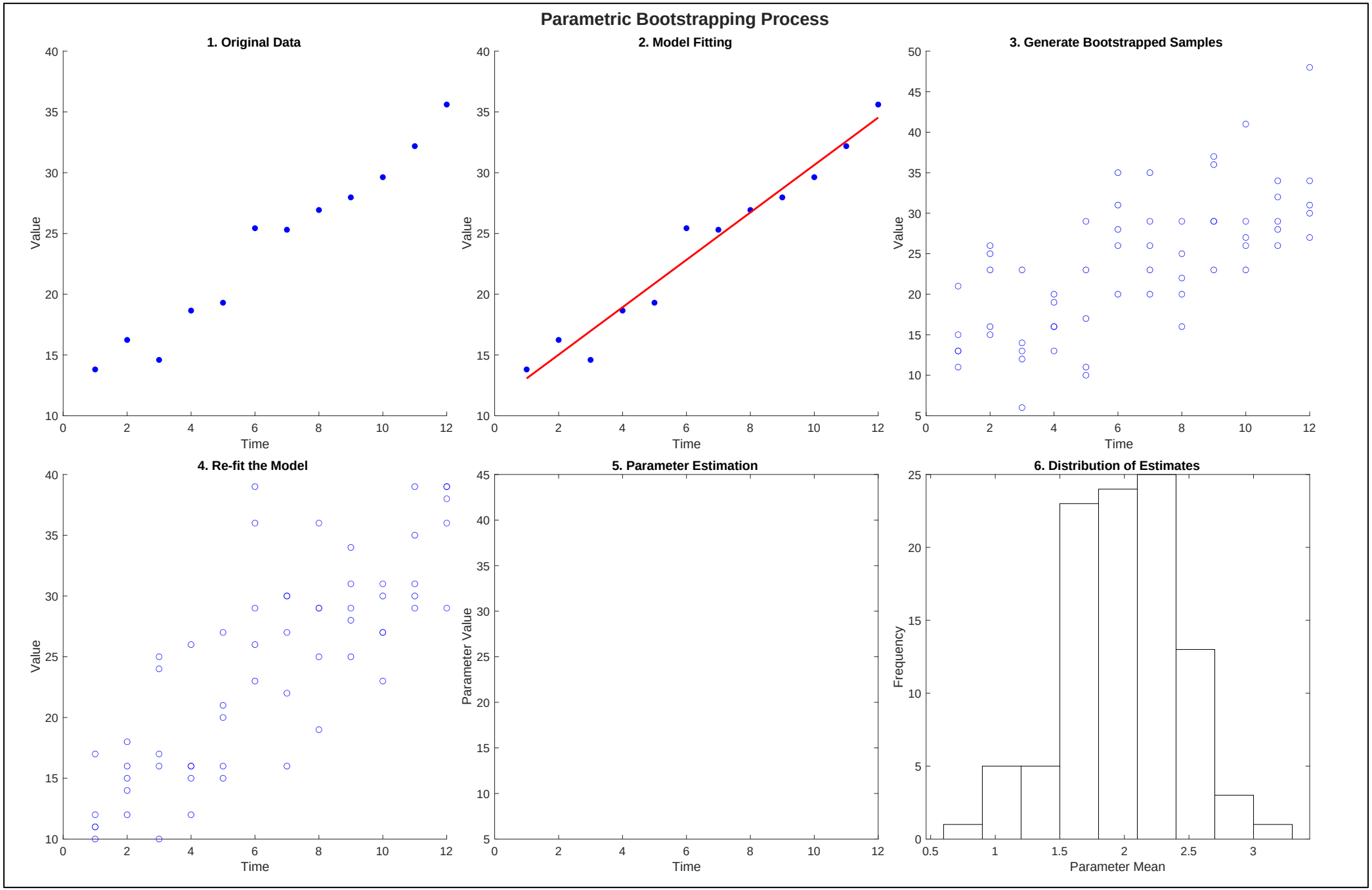
Parametric bootstrapping to quantify uncertainty

Using the best-fit model $f(t, \hat{\Theta})$, generate *B*-times replicated simulated datasets of size *n*, where the observation at time t_j is sampled from the corresponding distribution.

Refit the model to each of the *B* simulated datasets to re-estimate the parameters using the same estimation method for the bootstrap sample as for the original data. The new parameter estimates are $\widehat{\Theta}_b$, where $b = 1,2,\dots,B$.

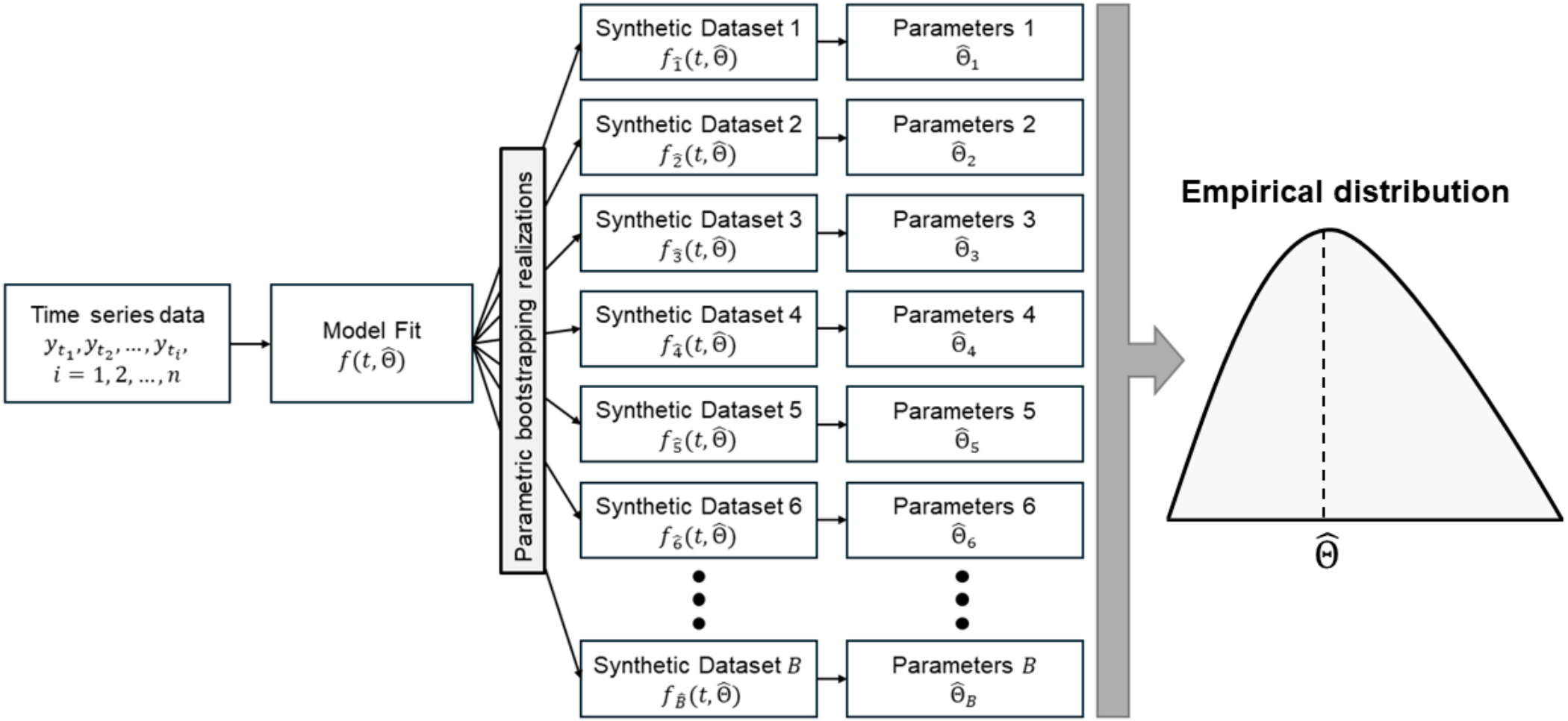
Using $(\widehat{\Theta}_b)$ it is possible to:

1. Characterize the empirical distribution of each parameter estimate
2. Calculate the variance
3. Construct confidence intervals for each parameter
4. Obtain the uncertainty around the model fit from $f(t, \widehat{\Theta}_1)$, $f(t, \widehat{\Theta}_2), \dots, f(t, \widehat{\Theta}_B)$.


Parametric Bootstrapping Process
1. Original Data
2. Model Fitting
3. Generate Bootstrapped Samples
4. Re-fit the Model
5. Parameter Estimation
6. Distribution of Estimates
Value
Time
Parameter Value
Frequency
Parameter Mean

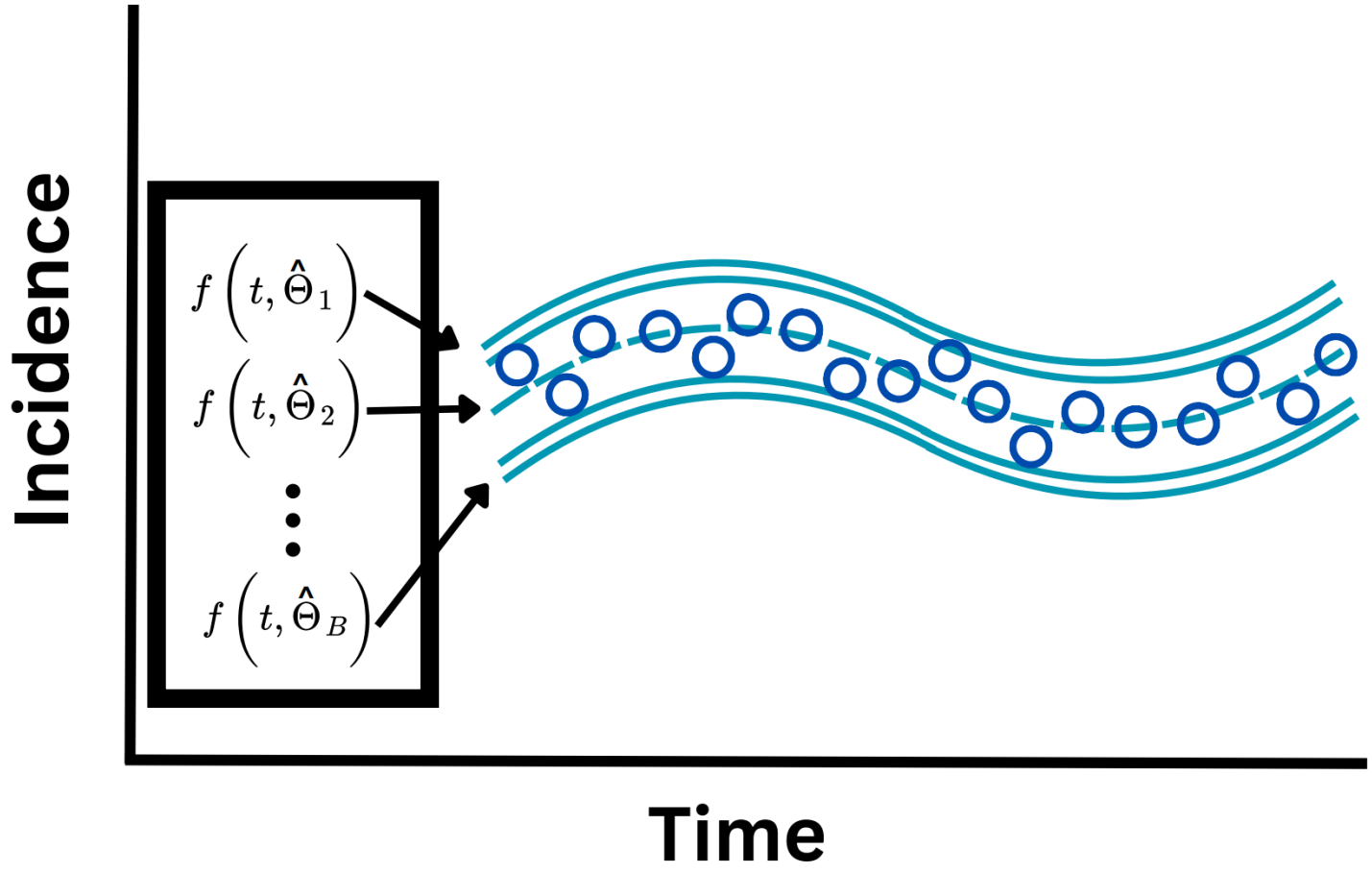
Bootstrapping: Quantifying Parameter Estimation

Model parameters and their confidence intervals are estimated by fitting the model to the aggregated incidence curve.



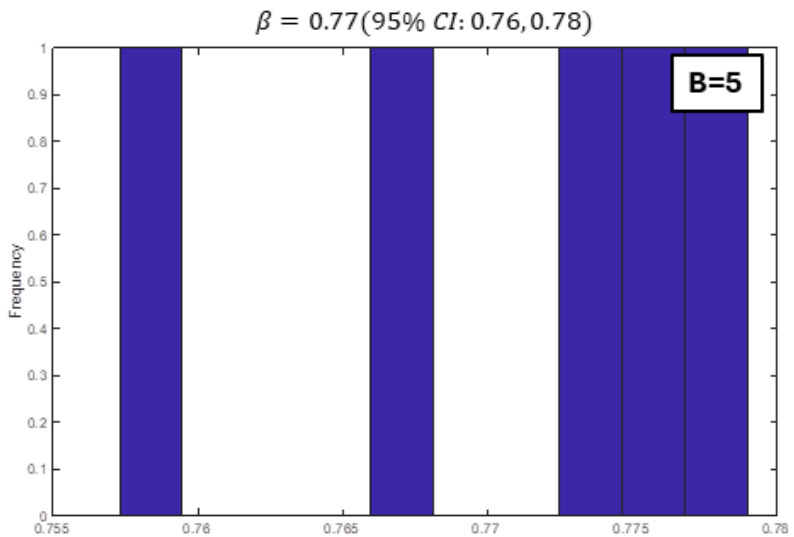
Bootstrapping: Uncertainty of the model best fit

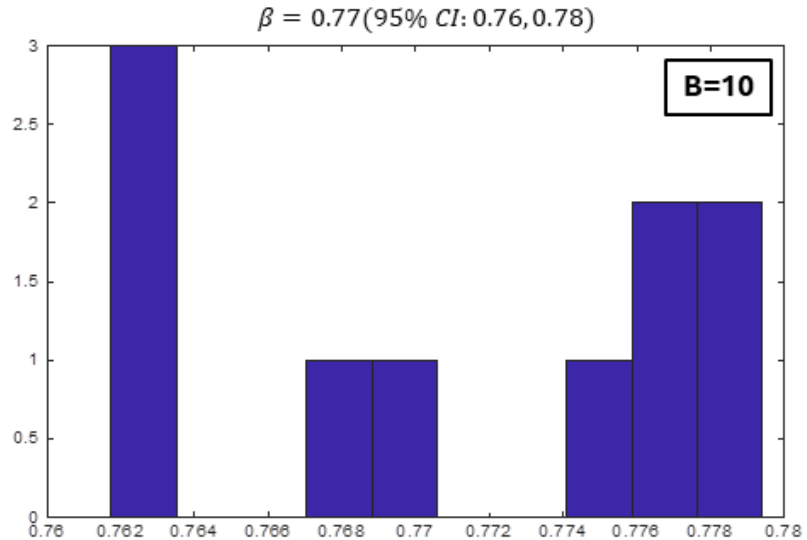
95% prediction intervals of model fits are obtained using parametric bootstrapping.

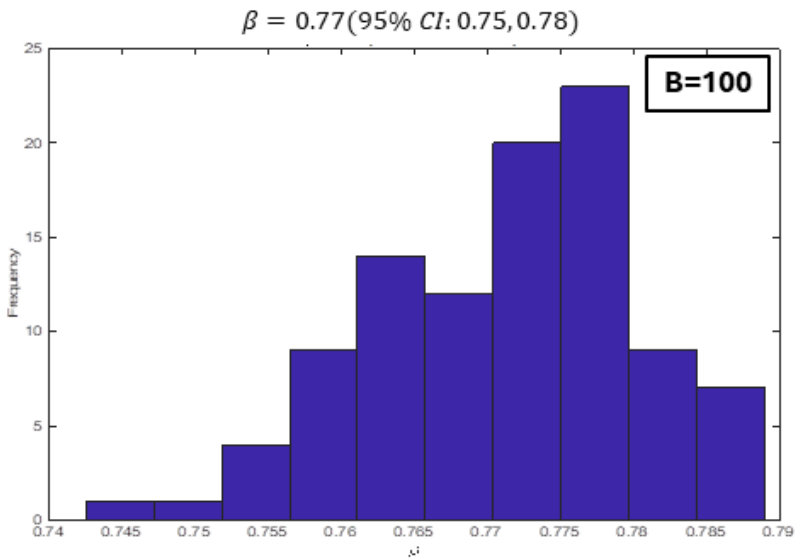


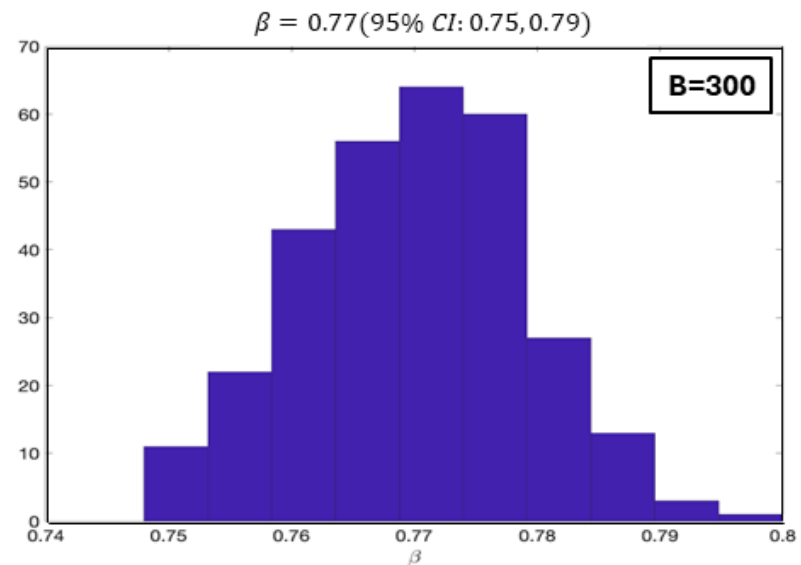
Adapted From Chowell et al. Infectious Disease Modeling. 2017

Bootstrap realizations









Model-based forecasts with quantified uncertainty

Based on the best-fit model $f(t, \widehat{\Theta})$, we can make h ahead forecasts using the estimate $f(t+h, \widehat{\Theta})$:

$$f(t+h, \widehat{\Theta}_1),$$

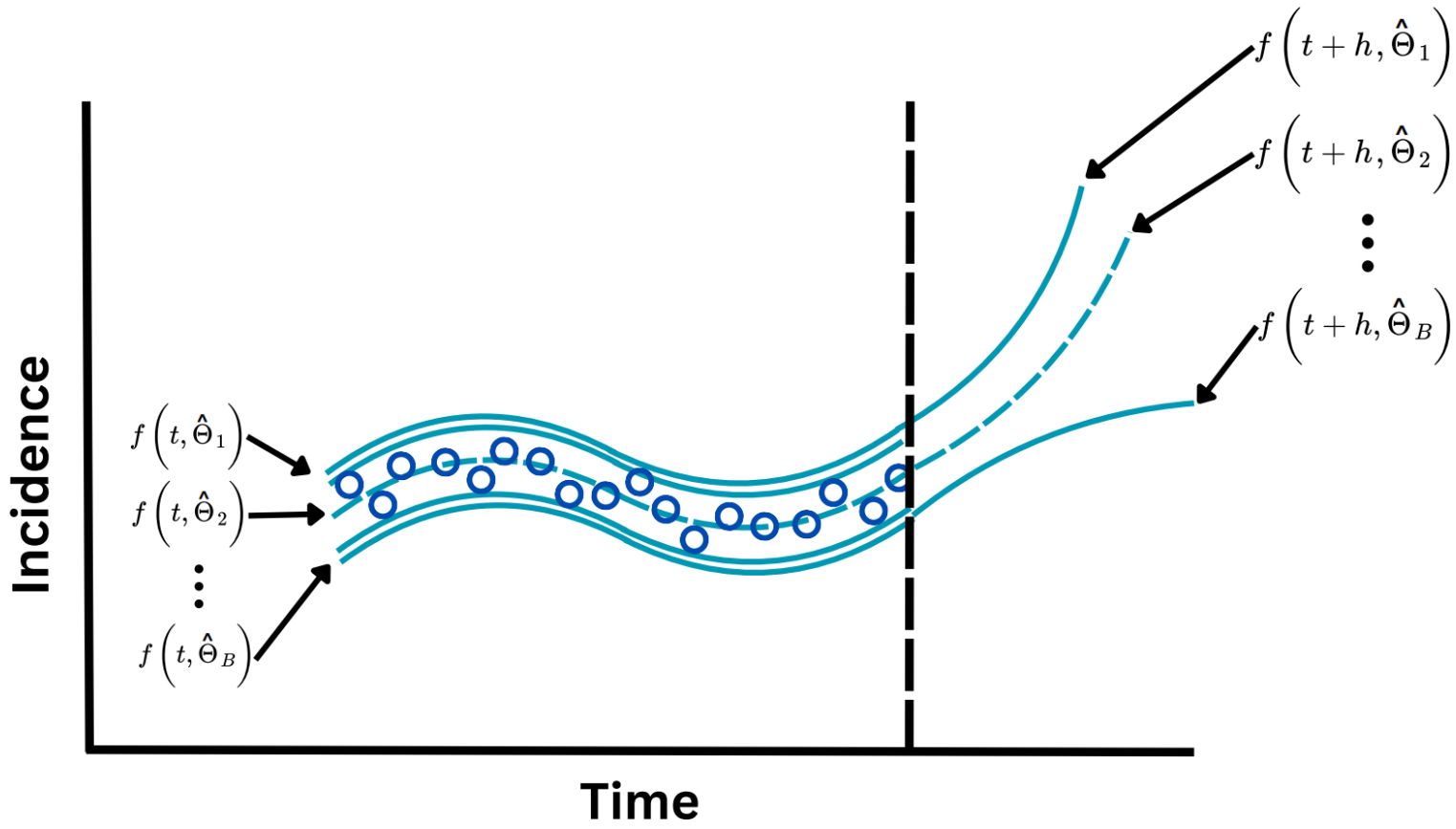
$$f(t+h, \widehat{\Theta}_2),$$

...

$$f(t+h, \widehat{\Theta}_B)$$

- $f(t+h, \widehat{\Theta}_B)$: Forecasted value of the current state of the system
- h: Forecasting horizon
- $\widehat{\Theta}_B$: Estimation of parameter set Θ from the b_{th} bootstrap sample

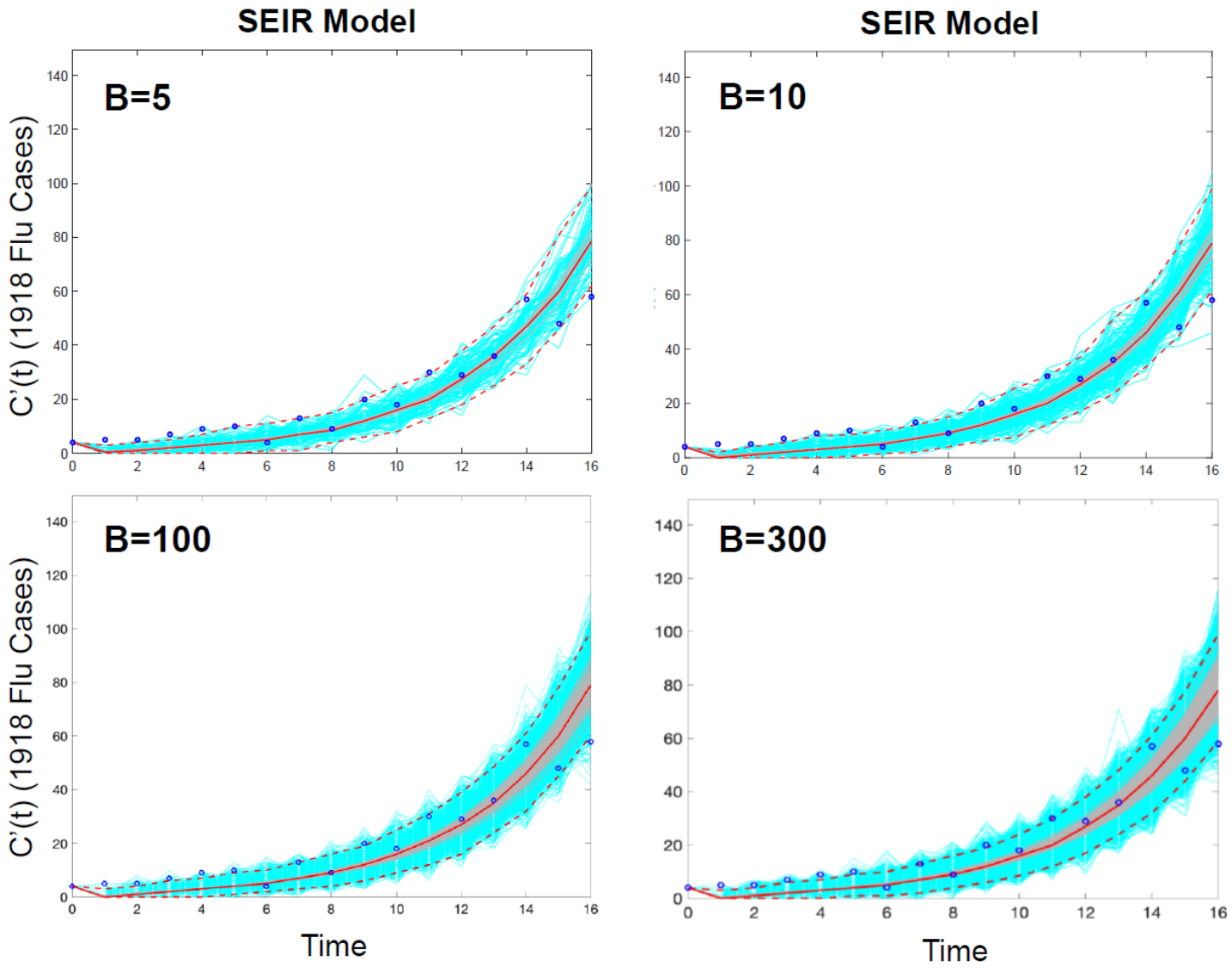
Bootstrapping: Model-based forecasts with quantified uncertainty



- We can $f(t+h,\widehat{\Theta}_B)$ to calculate the bootstrap variance to measure the uncertainty of the forecasts and use the 2.5% and 97.5% percentiles to construct the 95% prediction intervals (PI), with the assumed error structure.

Adapted From Chowell et al. Infectious Disease Modeling. 2017

Example: Model fit with the Poisson error structure



Reported cases of the 1918 influenza pandemic in San Francisco (*Chowell et al. JR Soc. Interface 2017)*

Performance Metrics: Assessing mean trajectory compared to observed data

Mean Absolute Error

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \left| f\left(t_i, \widehat{\Theta}\right) - y_{t_i} \right|$$

Mean Squared Error

$$MSE = \frac{1}{n}\sum_{i=1}^{n} \left(f\left(t_i, \widehat{\Theta}\right) - y_{t_i} \right)^2$$

Performance Metrics: Assessing model fit uncertainty

Weighted Interval Score

$$WIS_{\alpha_{0:K}}(F,y) = \frac{1}{K+\frac{1}{2}} . (w_0 . |y-m| + \sum_{k=1}^{K} w_k . IS_{\alpha_k}(F,y))$$

y_{t_i}: Time series of incident cases describing epidemic wave

t_i: Time points of time series data

$f\left(t_i, \widehat{\Theta}\right)$: Model fit

95% Prediction Interval Coverage

$$\frac{1}{n}\sum_{t=1}^{n} 1\{y_t > L_t \cap\ y_t < U_t\}$$

L_t: Lower bound of 95% prediction interval

U_t: Upper bound of 95% prediction interval

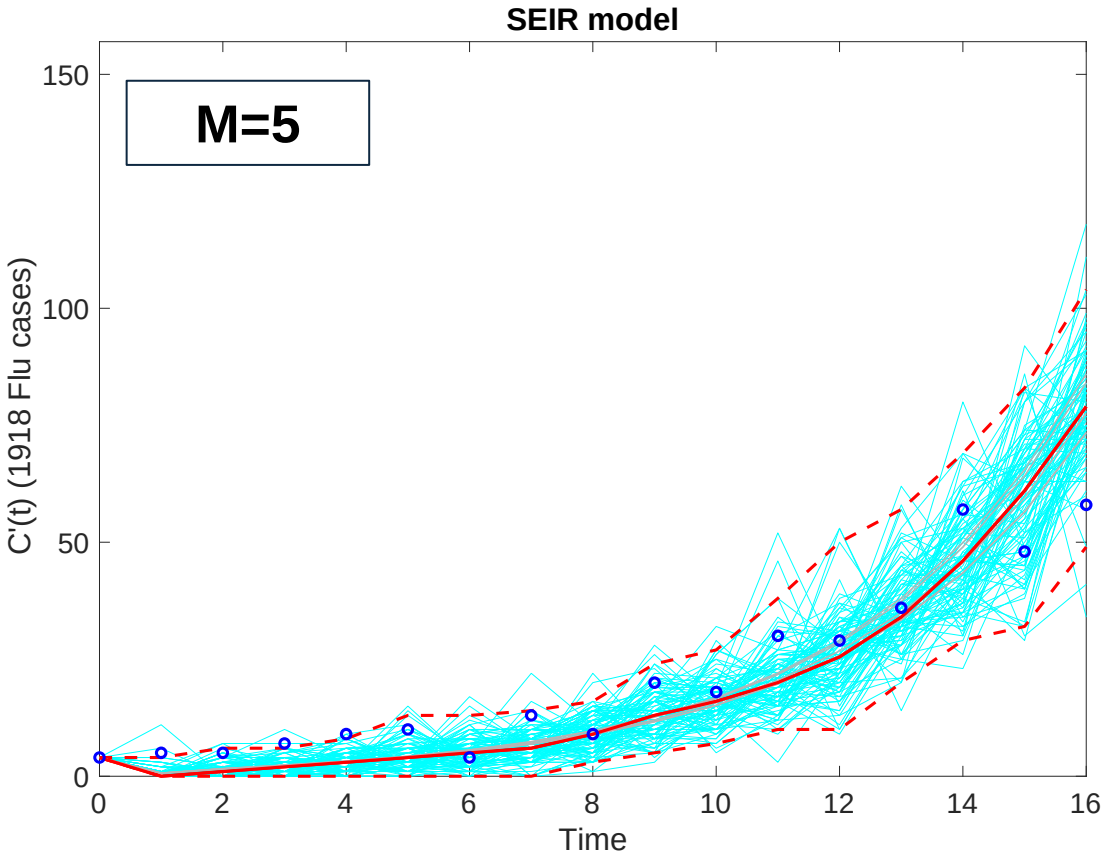
Tilmann Gneiting & Adrian E Raftery (2007)

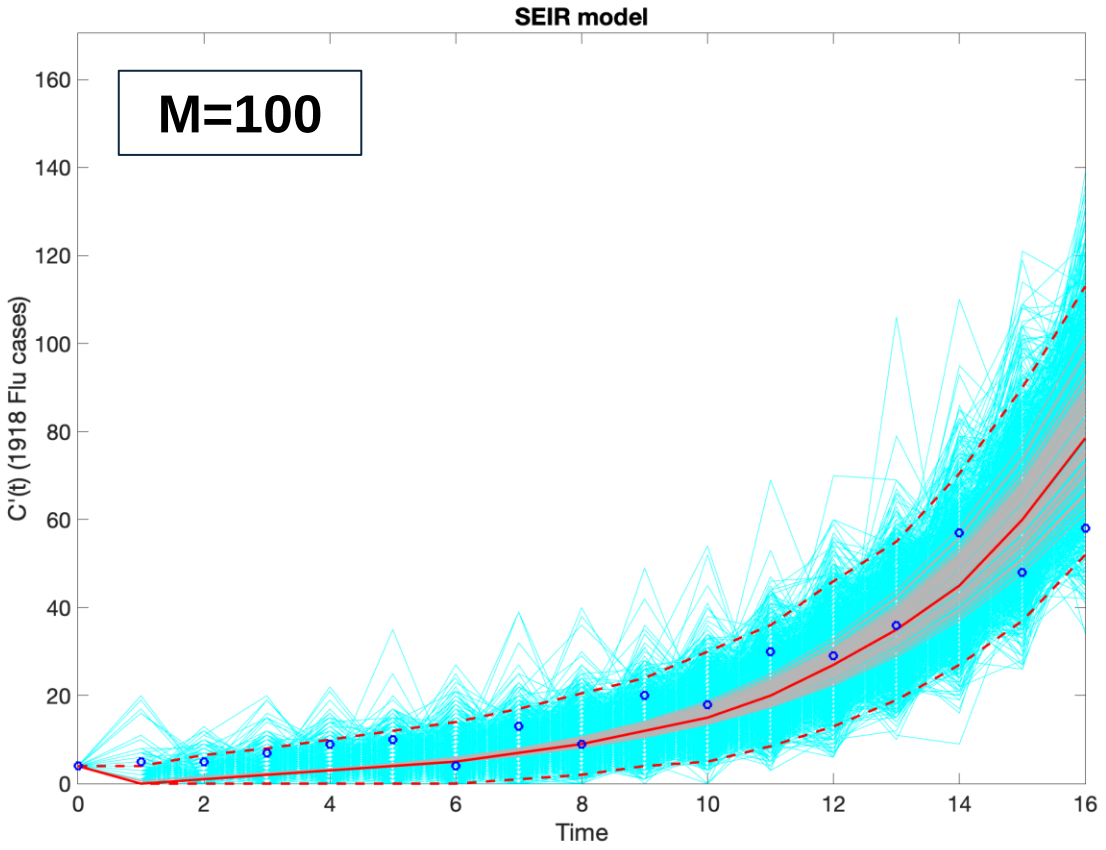
Monte Carlo Standard Errors (MCSEs)

Bootstrap realizations	MCSE (β)	β	β 95% CI LB	β 95% CI UB	MAE	MSE	Coverage 95% PI	WIS	R0	R0 95% CI LB	R0 95% CI UB
5	0.0042	0.78	0.77	0.79	5.76	64.03	64.71	3.95	3.17	3.10	3.19
10	0.0023	0.77	0.76	0.78	5.74	61.21	52.94	3.96	3.16	3.12	3.20
100	0.0009	0.77	0.75	0.79	5.74	62.22	58.82	3.89	3.17	3.08	3.23
300	0.0005	0.77	0.75	0.79	5.73	59.28	58.82	3.85	3.16	3.08	3.23

- Measures variability of simulation estimates.
- Involves Monte Carlo mean and standard deviation.
- Smaller values indicate higher precision.
- Balance between computational cost and desired precision.

Model fit (Negative Binomial Error Structure)





10-day Forecast and performance metrics