Validating Forecasting Strategies of Simple Epidemic Models on the 2015-2016 Zika Epidemic

Nicolas Puglisi

April 26, 2024

Introduction and Background

Epidemic Forecasting

- Forecasting is
- Accurate forecasting of infectious disease outbreaks is vital in safeguarding global health and the well-being of individuals
- Model-based forecasts enable public health officials to:
  - Test what-if scenarios
  - Evaluate control strategies
  - Develop informed policy

VIRGINIA TECH

Epidemic Forecasting - Challenges

- Accurate data availability is critical to inform models and produce reliable forecasts
- Human behavior during an epidemic can significantly impact assumed transmission dynamics
- Forecasting validation*

Zika Virus (ZIKV)

- Zika virus is a mosquito-borne disease characterized by dengue-like symptoms
    - Fever, rash, muscle, and joint pain
- Zika has multiple transmission pathways, including:
    - Mosquito-to-human transmission, human-to-human sexual transmission, and vertical transmission
- Most people who become infected are asymptomatic

2015-2016 Zika Epidemic

- On February 1, 2016, the World Health Organization (WHO) declared Zika-related microcephaly a Public Health Emergency of International Concern (PHEIC)
    - Lasted until November of 2016
- Zika spread throughout South America, heavily affecting Brazil and Colombia

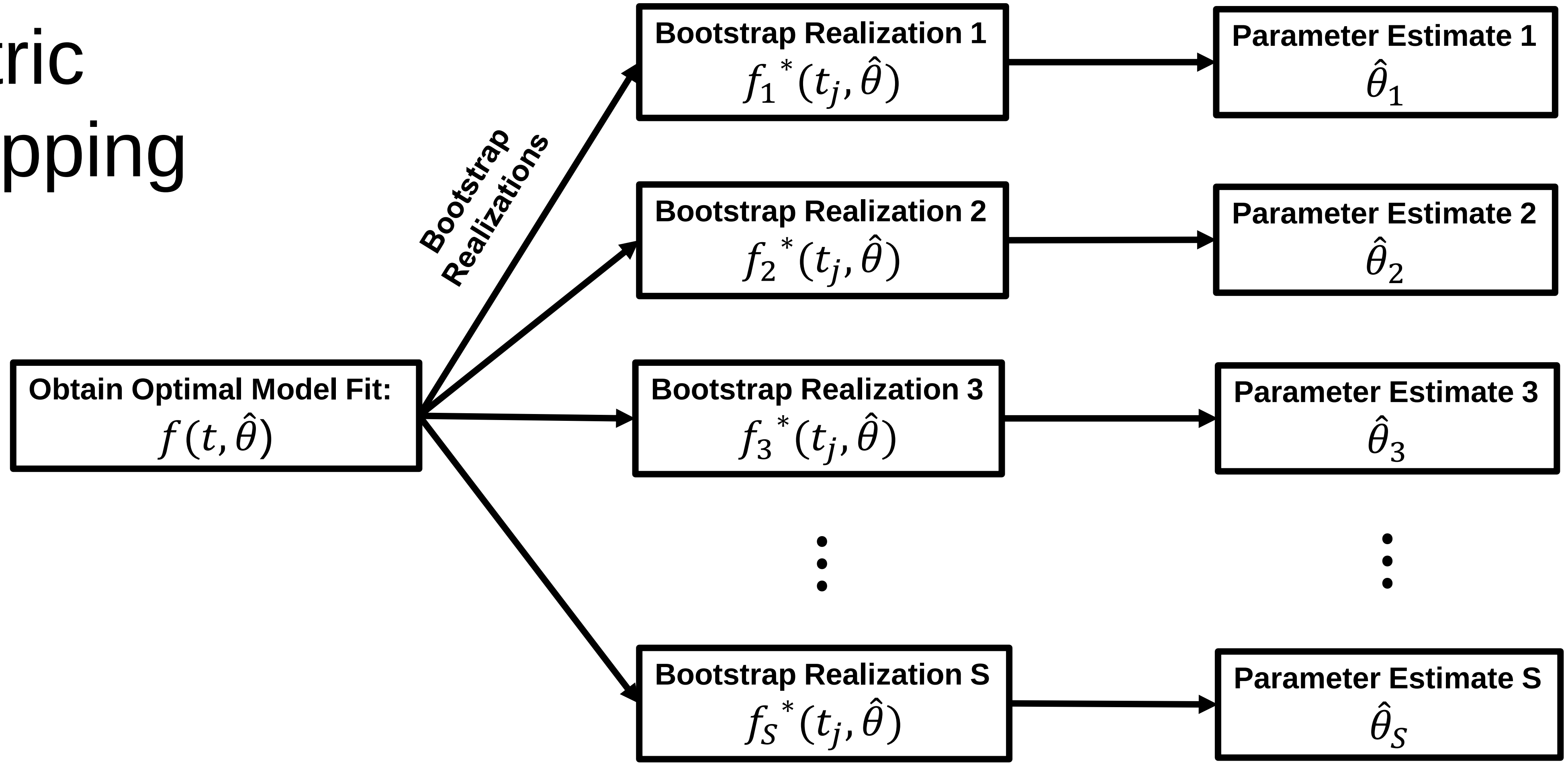
Thesis Question

- How do model-based forecasts of simple epidemic models compare under a Parametric Bootstrapping and Ensemble Kalman Filtering approach?
- Does the top-performing model change as the epidemic progresses, and how do spikes in Zika incidence affect the forecasting performance of each model?

VIRGINIA TECH

Methodology

Parametric Bootstrapping

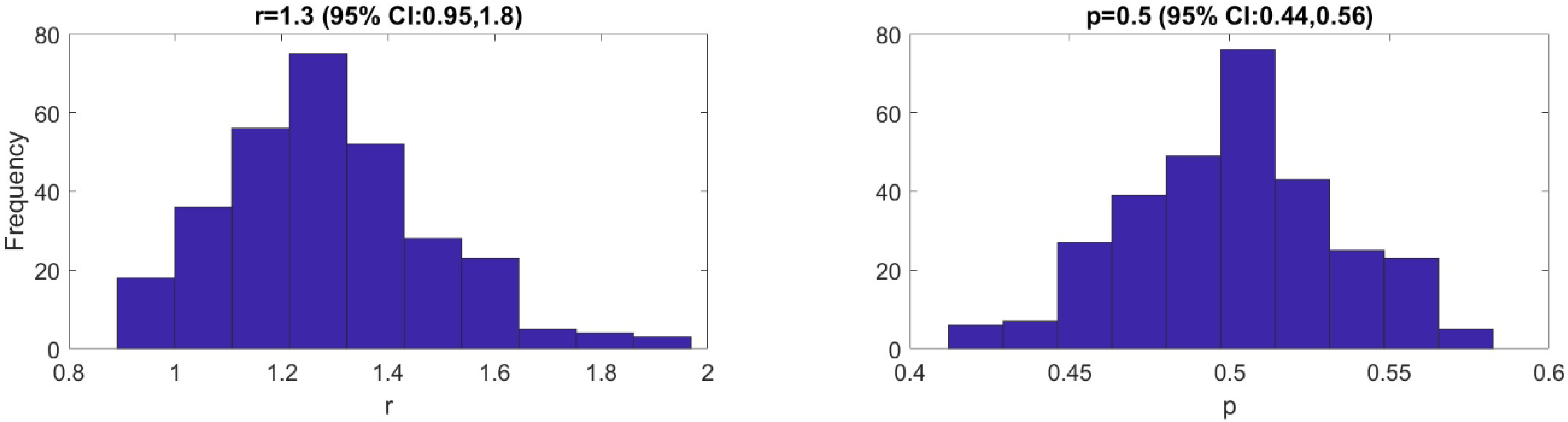


Cumulative Curve Function:

$$F(t_j, \hat{\theta}) = \sum_{l=1}^{j} f(t_l, \hat{\theta})$$

Dataset Resampling:

$$f_k^*(t_j, \hat{\theta}) = \mathrm{Po}(F(t_j, \hat{\theta}) - F(t_{j-1}, \hat{\theta}))$$


r=1.3 (95% CI:0.95,1.8)
Frequency
r
p=0.5 (95% CI:0.44,0.56)
p

Ensemble Kalman Filter (EnKF)

- Forecast Step:

$$S_{j|j} = \{x^1_{j|j}, x^2_{j|j}, ..., x^N_{j|j}\}$$

$$X_{j+1} = F(X_j) + V_{j+1},\ V_{j+1} \sim \mathcal{N}(0, \mathbf{C}_{j+1})$$

$$Y_{j+1} = G(X_{j+1}) + W_{j+1},\ W_{j+1} \sim \mathcal{N}(0, \mathbf{D}_{j+1}),$$

- Analysis Step:

$$x_{j+1|j+1} = x_{j+1|j} + K_{j+1}(y^n_{j+1} - \hat{y}^n_{j+1}), n = 1, ..., N$$

$$y^n_{j+1} = y_{j+1} + w^n_{j+1}, w_{j+1} \sim N(0, \mathbf{D}_{j+1}),$$

$$S_{j+1|j+1} = \{x^1_{j+1|j+1}, x^2_{j+1|j+1}, ..., x^N_{j+1|j+1}\}$$

Models of interest - Growth Models

Generalized Growth Model (GGM):

$$\frac{dC}{dt} = rC^p(t)$$

Generalized Logistic Model (GLM):

$$\frac{dC}{dt} = rC^p[1 - (\frac{C}{K})]$$

Generalized Richards Model (GRM):

$$\frac{dC}{dt} = rC^p[1 - (\frac{C}{K})^a]$$

- $\frac{dC}{dt}$ describes the growth in disease incidence at time t
- C described the cumulative incidence at time t
- Parameters:
- r is a growth parameter
- K represents the size of an epidemic
- p and a are growth scaling parameters, with $0 \leq p \leq 1$

VIRGINIA TECH.

Susceptible-Infected-Recovered (SIR) Model

$$\frac{dS}{dt} = -\frac{\beta S I}{N}$$
$$\frac{dI}{dt} = \frac{\beta S I}{N} - \gamma I$$
$$\frac{dR}{dt} = \gamma I$$
$$\frac{dC}{dt} = \frac{\beta S I}{N}$$

- State variables S, I, and R represent the susceptible, infected, and recovered populations
- C is an auxiliary state variable that tracks cumulative disease incidence
- Parameters:
- β - constant transmission parameter
- $\frac{1}{\gamma}$ - mean infectious period
- N - population size

Susceptible-Exposed-Infected-Recovered (SEIR) Model

$$\frac{dS}{dt} = -\frac{\beta S I}{N}$$
$$\frac{dE}{dt} = \frac{\beta S I}{N} - \kappa E$$
$$\frac{dI}{dt} = \kappa E - \gamma I$$
$$\frac{dR}{dt} = \gamma I$$
$$\frac{dC}{dt} = \kappa E$$

- State variable E represent the exposed population
- C is an auxiliary state variable that tracks cumulative disease incidence
- Parameters:
- β - constant transmission parameter
- $\frac{1}{\kappa}$ - mean latent period
- $\frac{1}{\gamma}$ - mean infectious period
- N - population size

Performance Metrics

- To evaluate forecast prediction error, we used the following three metrics:
    - Mean Absolute Error (MAE): $\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |f(t_i, \hat{\theta}) - y_{t_i}|$
    - Mean Squared Error (MSE): $\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(f\left(t_i, \hat{\theta}\right) - y_{t_i})^2$
    - Root-Mean Squared Error (RMSE): $\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f\left(t_i, \hat{\theta}\right) - y_{t_i})^2}$

VIRGINIA TECH

Performance Metrics

- We use the follow two metrics to evaluate predictive uncertainty:
  - Coverage rate of the $(1 - \alpha)$x100% Prediction Interval:
    - Proportion of observations falling within the PI
  - Weighted Interval Score (WIS):

$$\mathrm{IS}_{\alpha}(F, y) = (u - l) + \frac{2}{\alpha} * (l - y) * \mathbf{I}(y < l) + \frac{2}{\alpha} * (y - u) * \mathbf{I}(y > u)$$

$$\mathrm{WIS}_{\alpha_{0:K}}(F, y) = \frac{1}{K + \frac{1}{2}} \left(\frac{1}{2} |y - m| + \sum_{k=1}^{K} w_k \mathrm{IS}_{\alpha_k}(F, y) \right)$$

VIRGINIA TECH

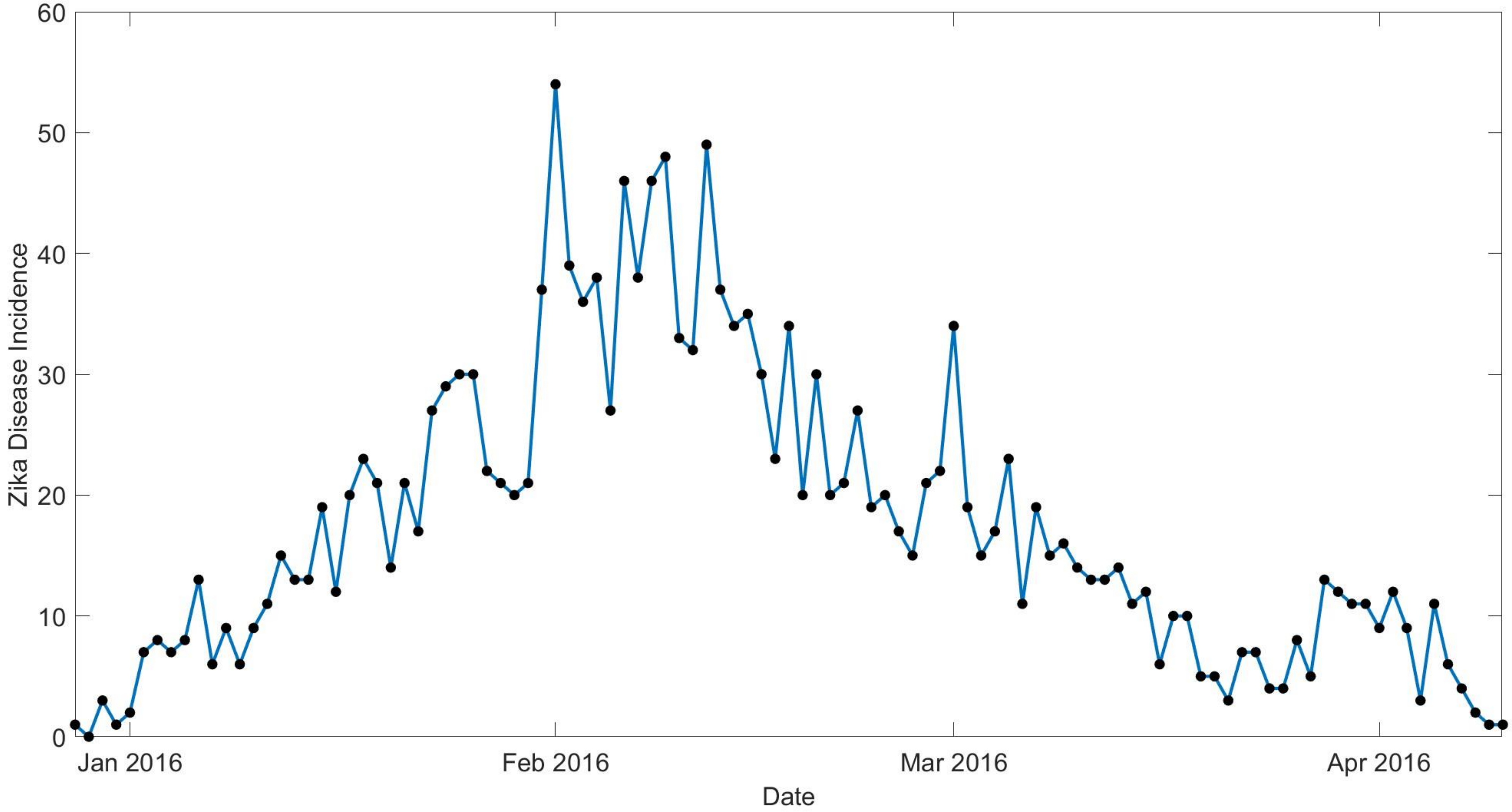
Case Study Construction

- In total, we conduced five case studies
- Case Study 1: Calibrate models on first 35 days of data
  - Generate 14-day-ahead forecast up to day 49
    - Why 14 days?
- Each successive Case Study assimilates data predicted against from previous Case Study, and then refits each model:
  - Case Study 2: Calibrate models on first 49 days of data
    - Forecasts up to day 63 of Zika epidemic

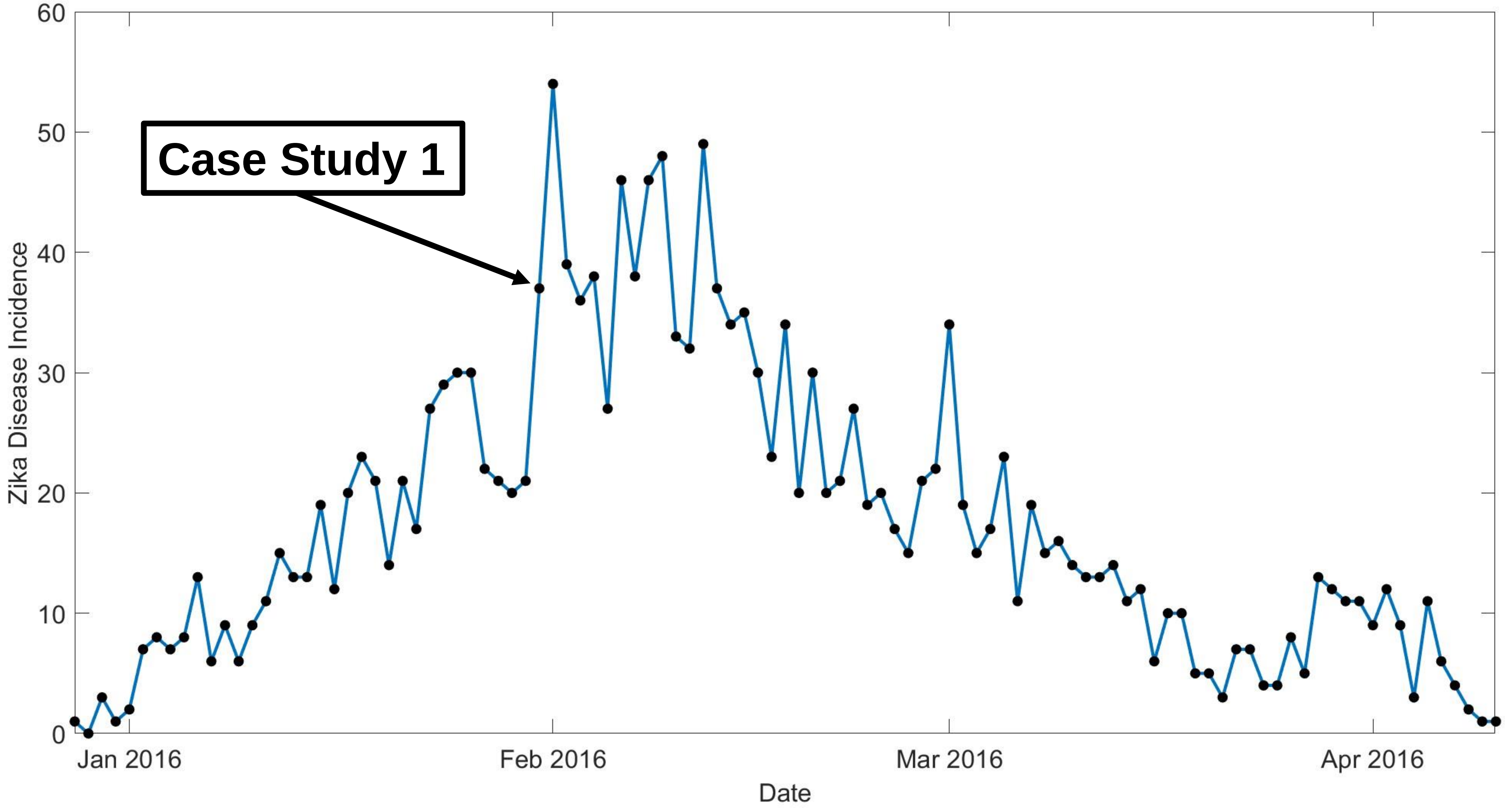
VIRGINIA TECH

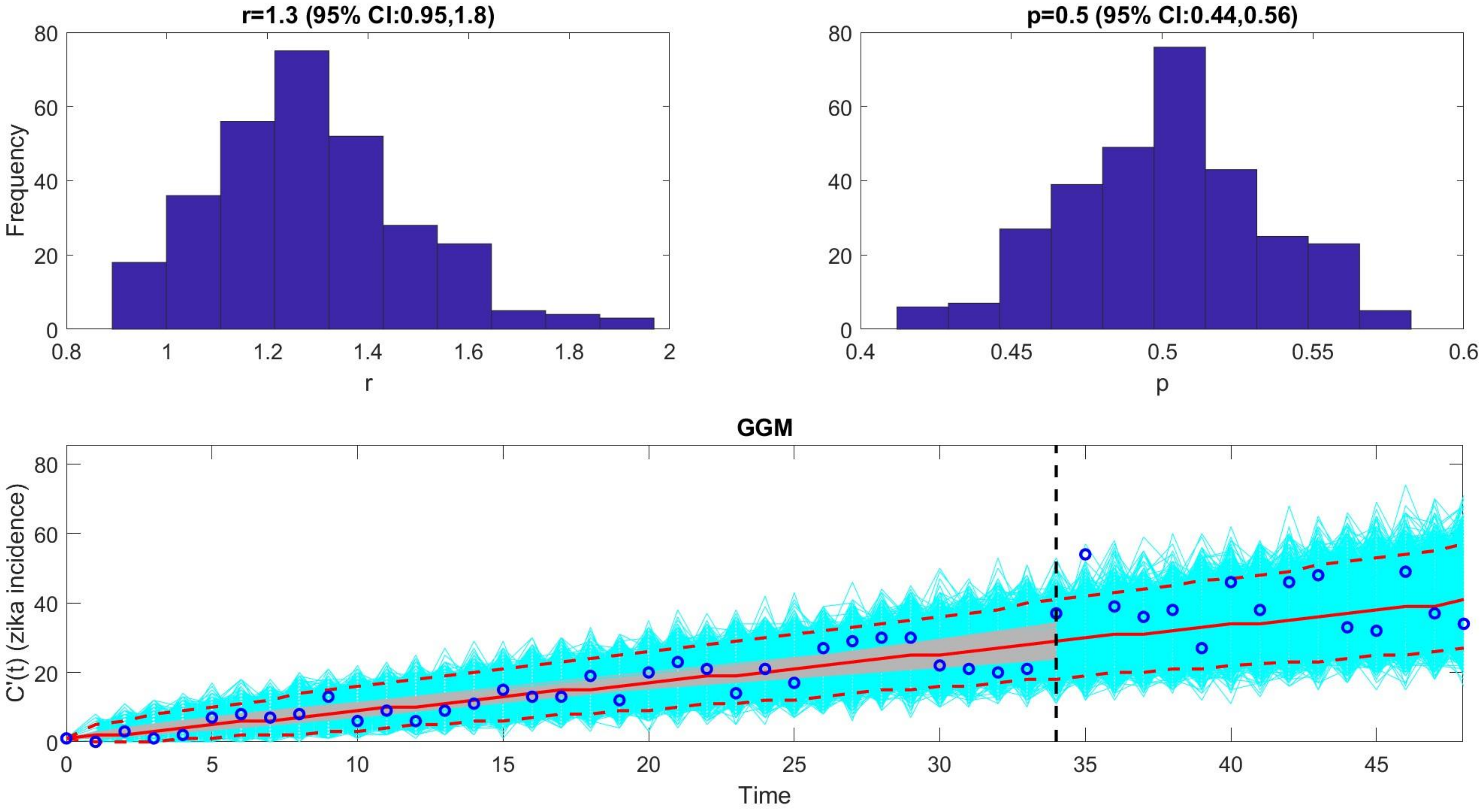
Zika Incidence Data

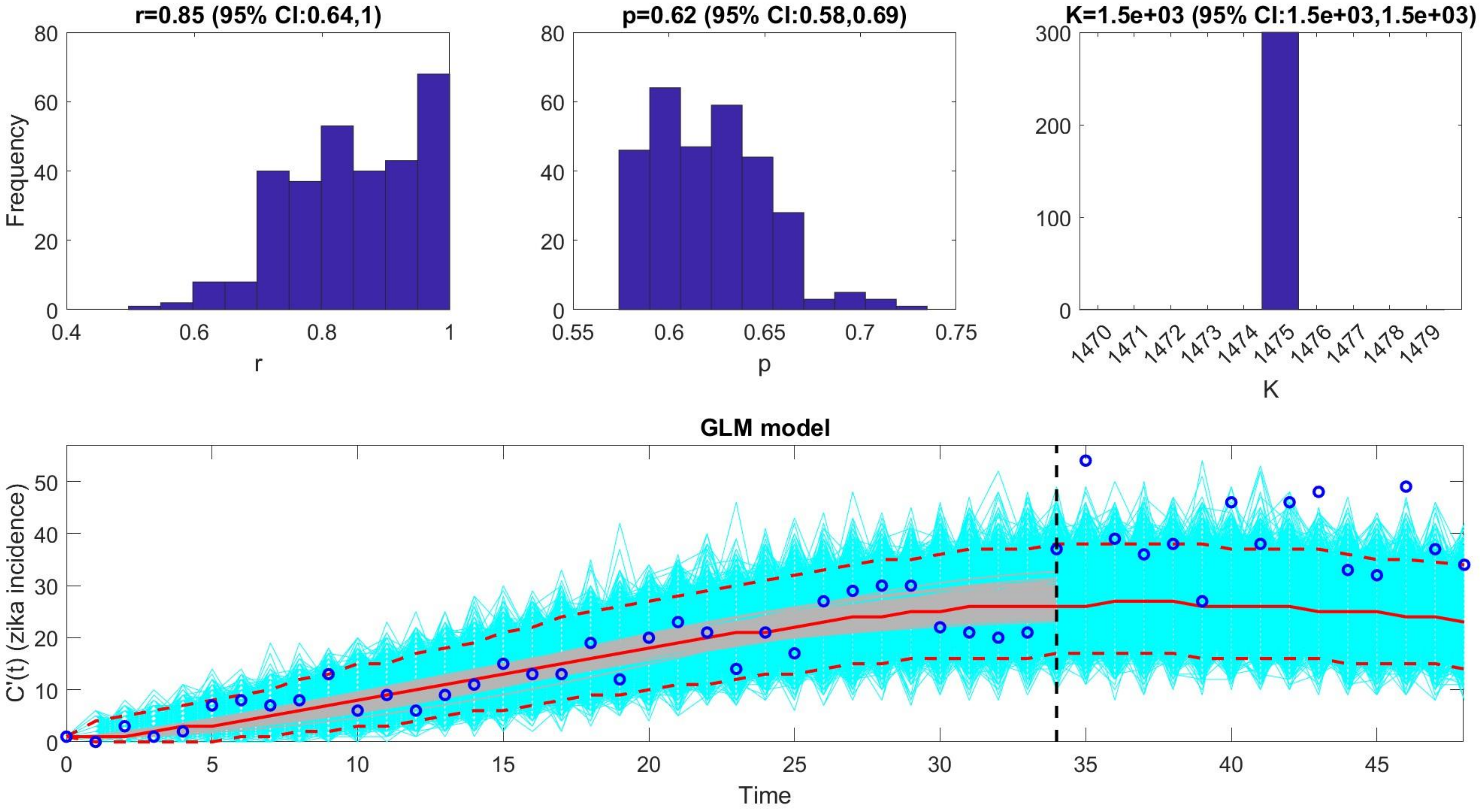
- Zika incidence data was collected from Antioquia, Colombia, from the Ministry of Health of Colombia
    - At the time of collection, Antioquia had an estimated population size of 6.3 million
- Disease reporting is based off onset of symptoms
    - Approximately 5% of cases were laboratory tested

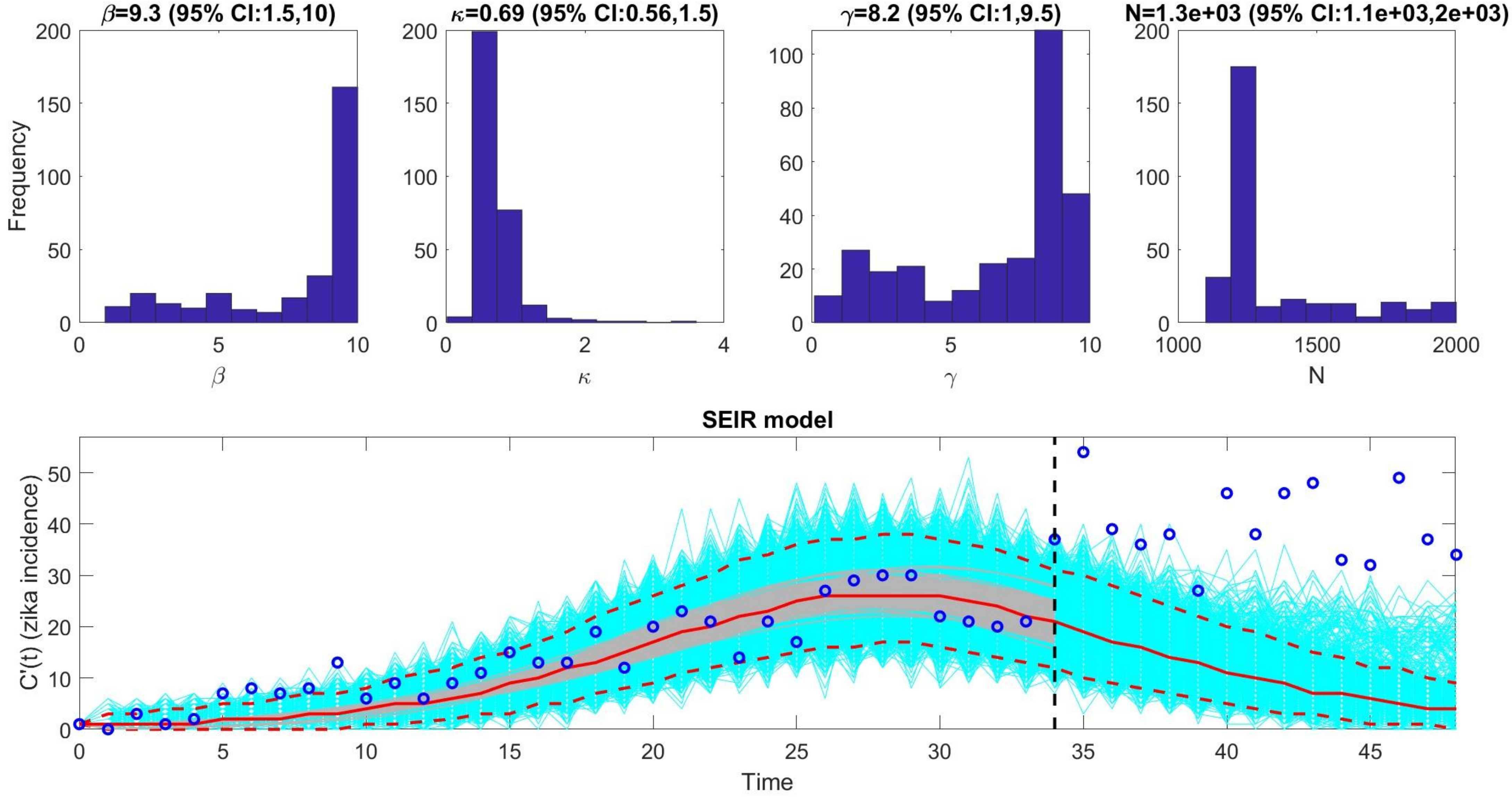

60
50
40
30
20
10
0
Zika Disease Incidence
Jan 2016
Feb 2016
Mar 2016
Apr 2016
Date

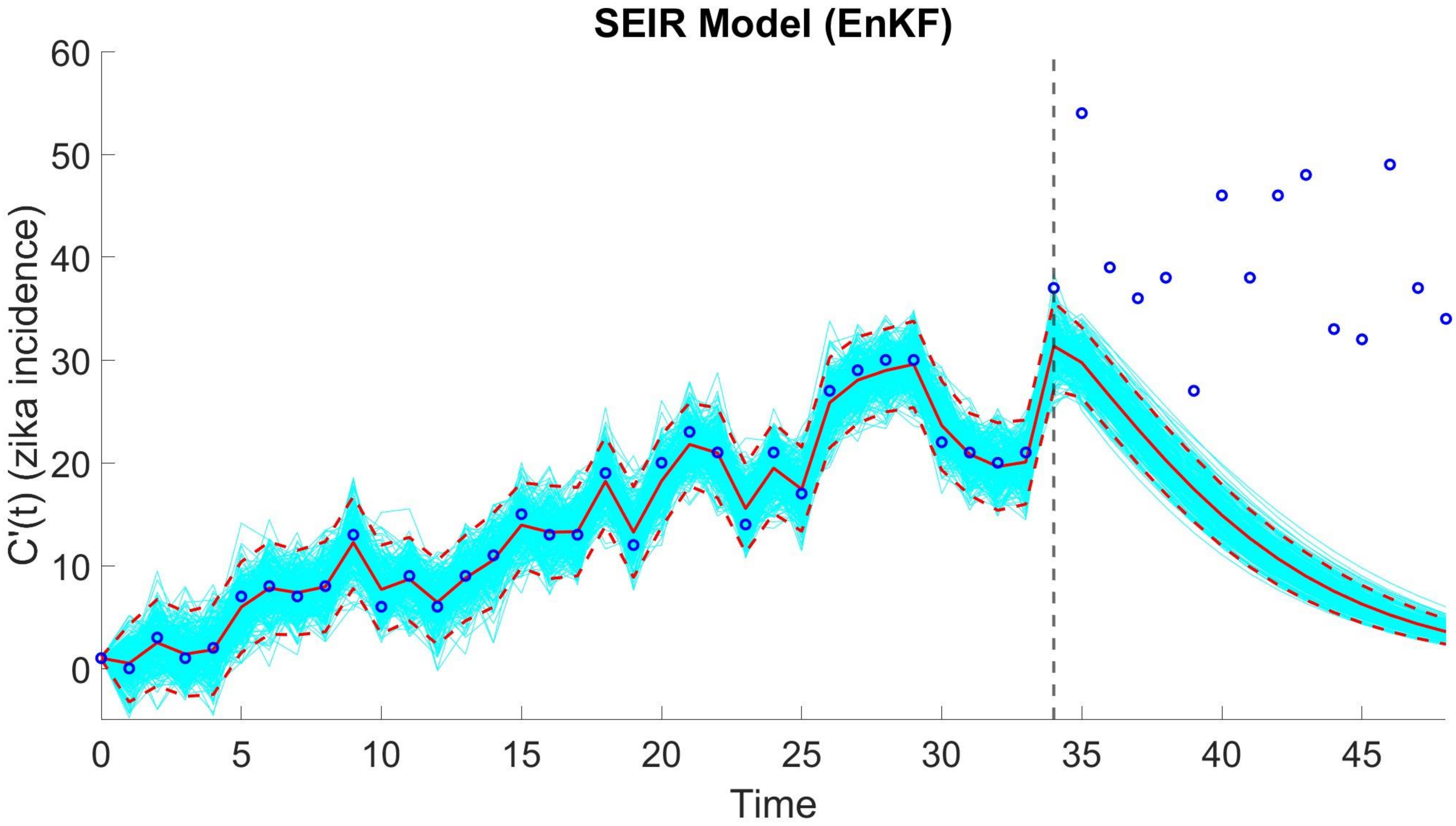
Simulations and Results


Case Study 1
60
50
40
30
20
10
0
Zika Disease Incidence
Jan 2016
Feb 2016
Mar 2016
Apr 2016
Date


r=1.3 (95% CI:0.95,1.8)
Frequency
r
p=0.5 (95% CI:0.44,0.56)
p
GGM
C'(t) (zika incidence)
Time

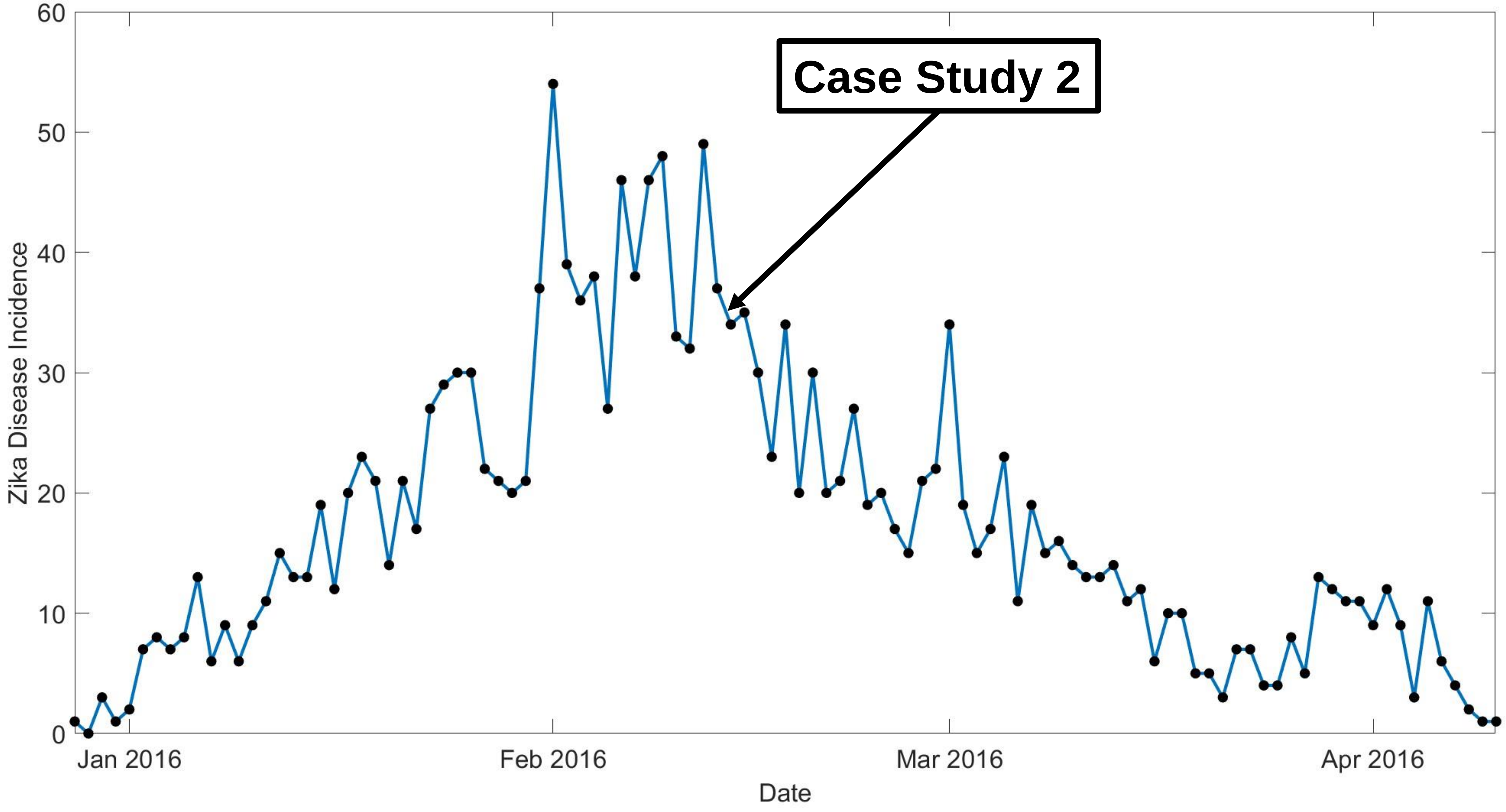

r=0.85 (95% CI:0.64,1)
p=0.62 (95% CI:0.58,0.69)
K=1.5e+03 (95% CI:1.5e+03,1.5e+03)
Frequency
r
p
K
GLM model
C'(t) (zika incidence)
Time

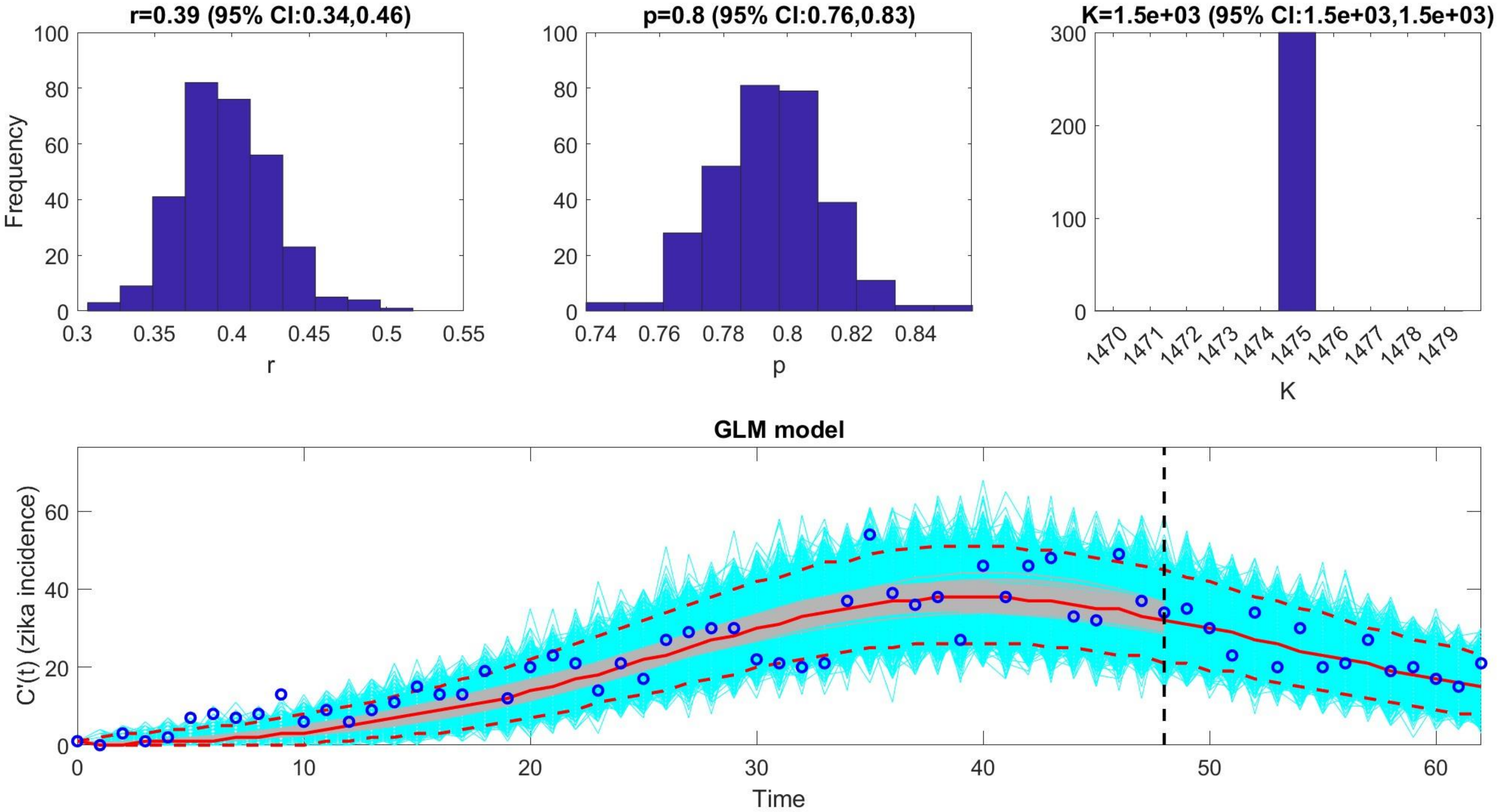

β=9.3 (95% CI:1.5,10)
κ=0.69 (95% CI:0.56,1.5)
γ=8.2 (95% CI:1,9.5)
N=1.3e+03 (95% CI:1.1e+03,2e+03)
Frequency
β
κ
γ
N
SEIR model
C'(t) (zika incidence)
Time

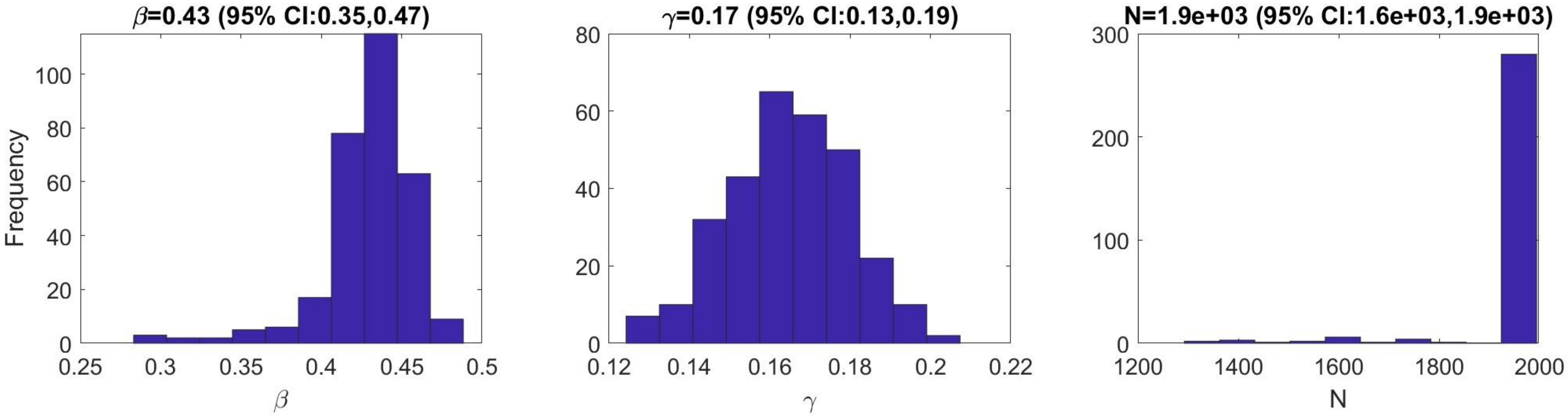

SEIR Model (EnKF)
C'(t) (zika incidence)
Time
60
50
40
30
20
10
0
0
5
10
15
20
25
30
35
40
45

Forecasting Performance Metrics					
Model	MAE	MSE	RMSE	Coverage	WIS
GGM	8.305	97.596	9.879	92.857	5.190
GLM	14.301	257.892	16.059	42.857	10.613
GRM	13.357	230.465	15.181	64.286	8.988
SIR	29.353	925.239	30.418	0	26.604
SEIR	29.716	947.150	30.776	0	26.523
SIR(EnKF)	26.218	773.657	27.815	*	*
SEIR(EnKF)	27.472	850.308	29.160	*	*

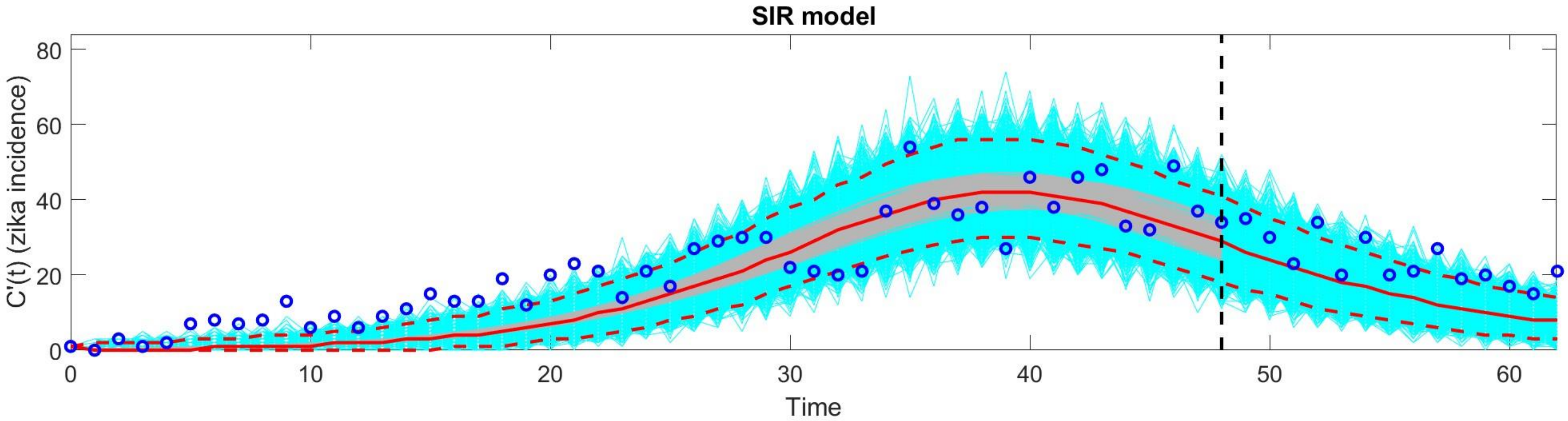
Table 4.1: Performance metrics of model-based forecasts from Case Study 1


Case Study 2
Zika Disease Incidence
Date
0
10
20
30
40
50
60
Jan 2016
Feb 2016
Mar 2016
Apr 2016

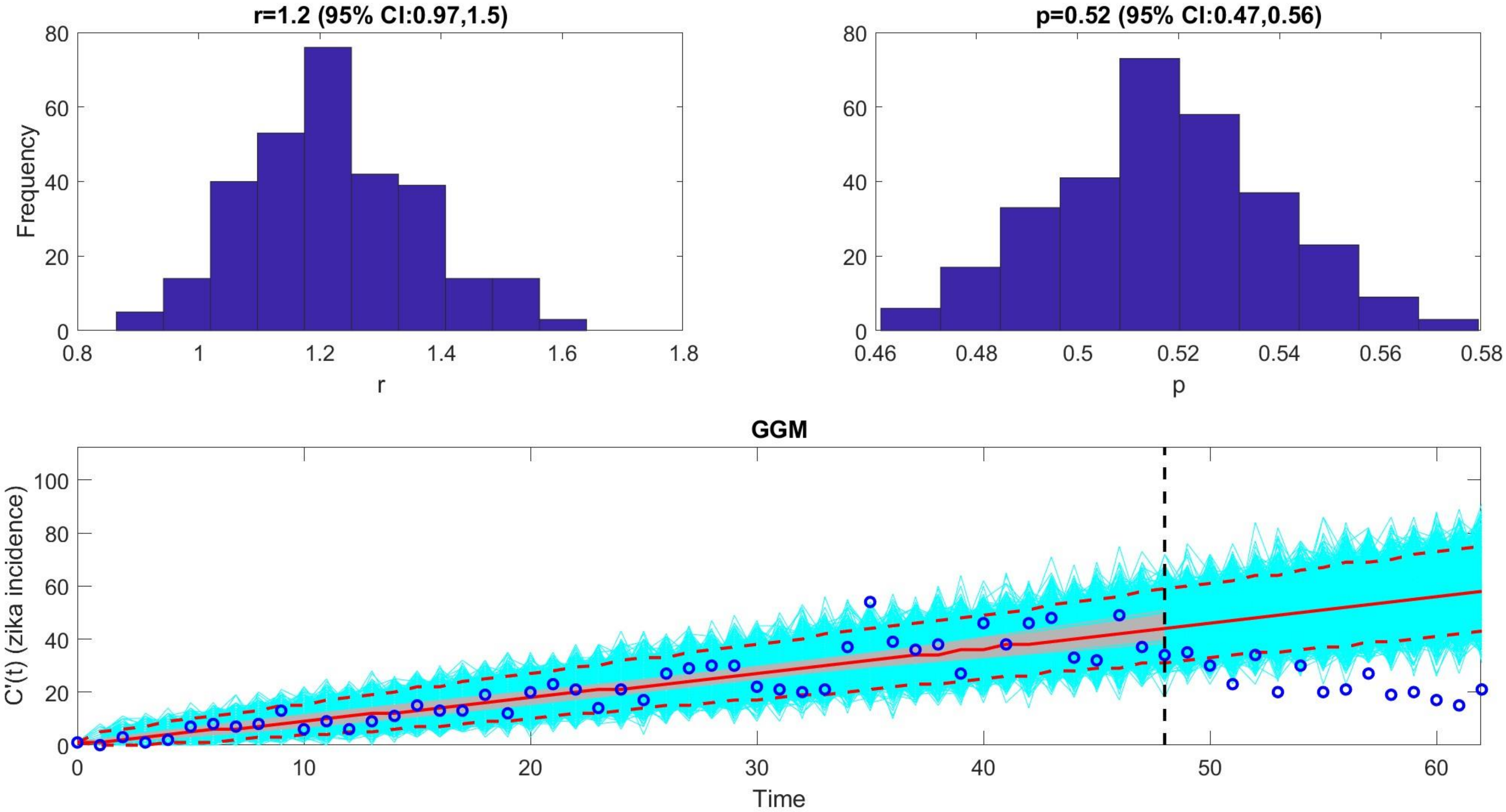

r=0.39 (95% CI:0.34,0.46)
p=0.8 (95% CI:0.76,0.83)
K=1.5e+03 (95% CI:1.5e+03,1.5e+03)
Frequency
r
p
K
GLM model
C'(t) (zika incidence)
Time



β=0.43 (95% CI:0.35,0.47)
γ=0.17 (95% CI:0.13,0.19)
N=1.9e+03 (95% CI:1.6e+03,1.9e+03)
Frequency
β
γ
N

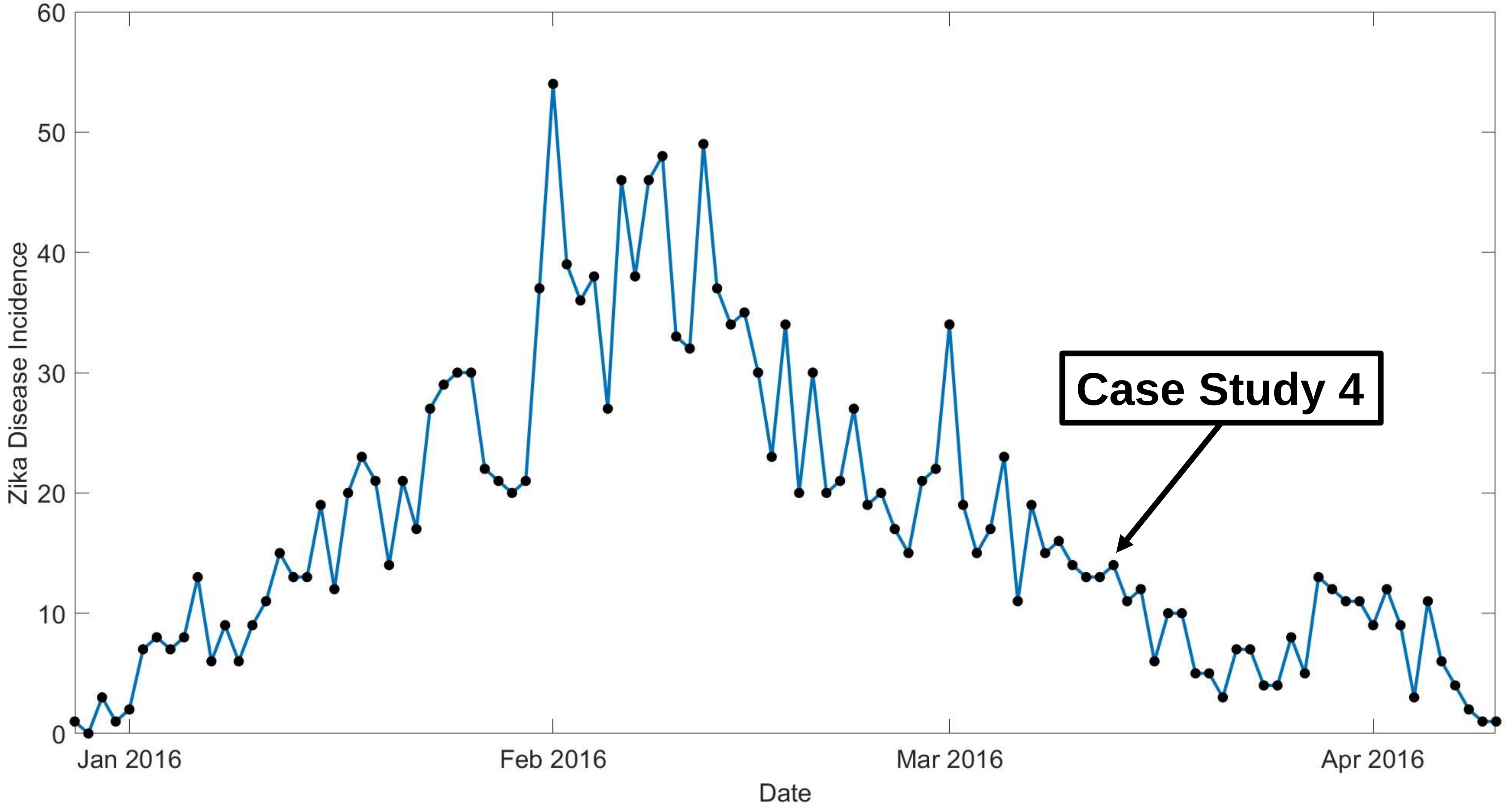


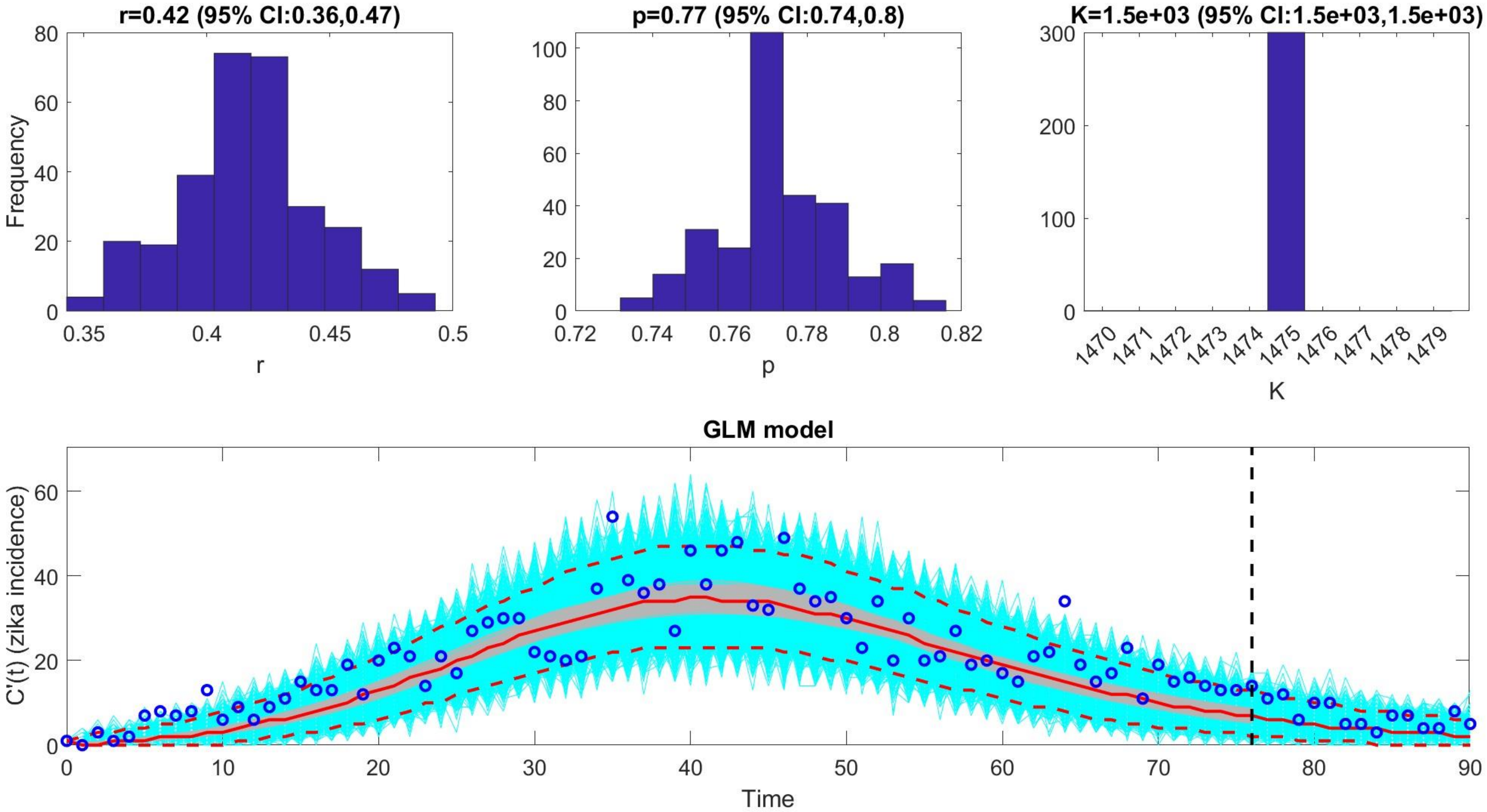
SIR model
C'(t) (zika incidence)
Time


r=1.2 (95% CI:0.97,1.5)
Frequency
r
p=0.52 (95% CI:0.47,0.56)
p
GGM
C'(t) (zika incidence)
Time

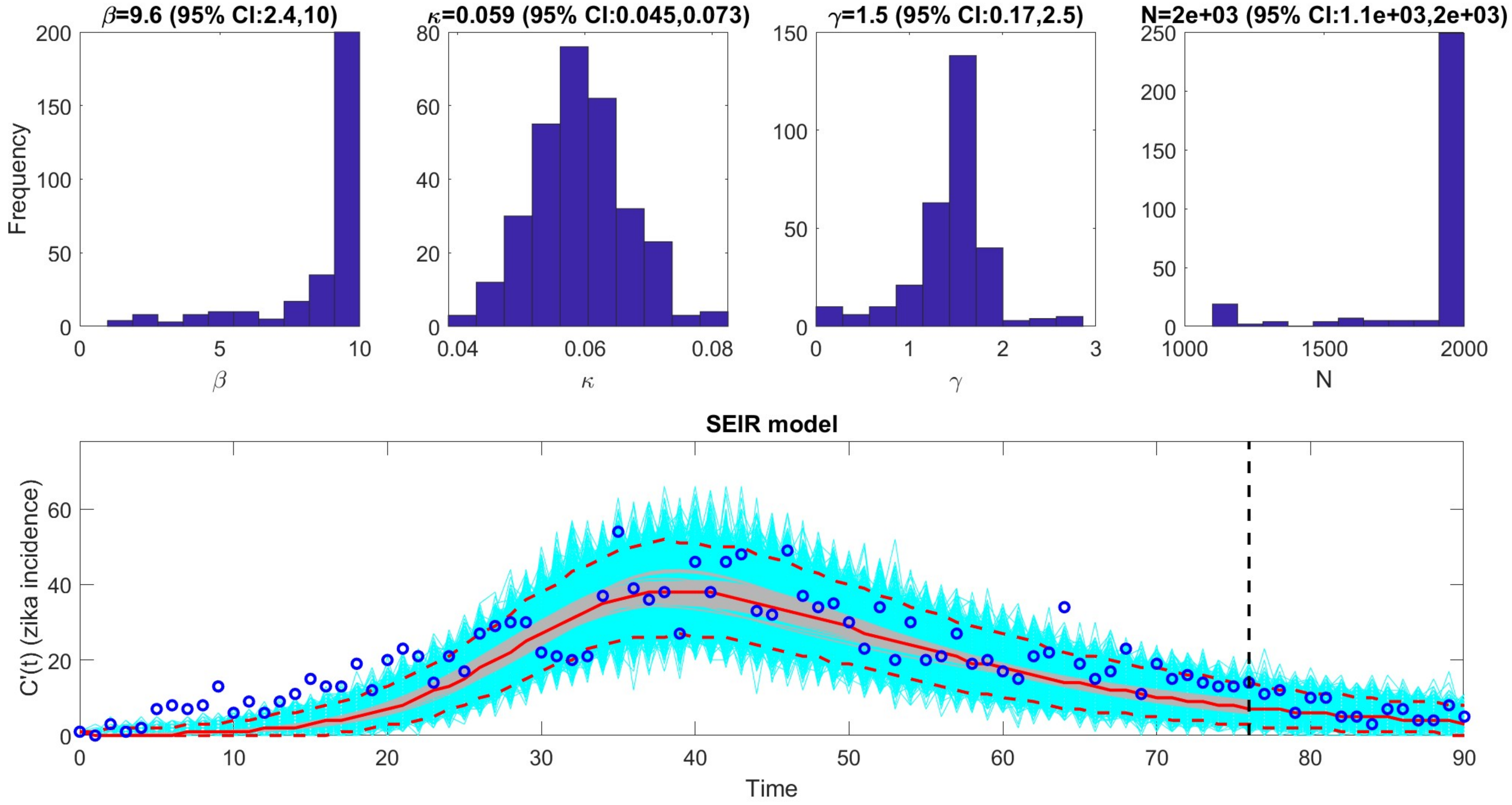
Forecasting Performance Metrics					
Model	MAE	MSE	RMSE	Coverage	WIS
GGM	27.997	871.994	29.530	7.143	22.442
GLM	3.380	17.858	4.226	100	2.180
GRM	3.524	19.240	4.386	100	2.248
SIR	8.182	84.585	9.197	57.143	5.815
SEIR	5.559	41.893	6.472	85.714	3.588
SIR(EnKF)	6.956	64.424	8.027	*	*
SEIR(EnKF)	5.585	42.138	6.491	*	*

Table 4.2: Performance metrics of models-based forecasts from Case Study 2

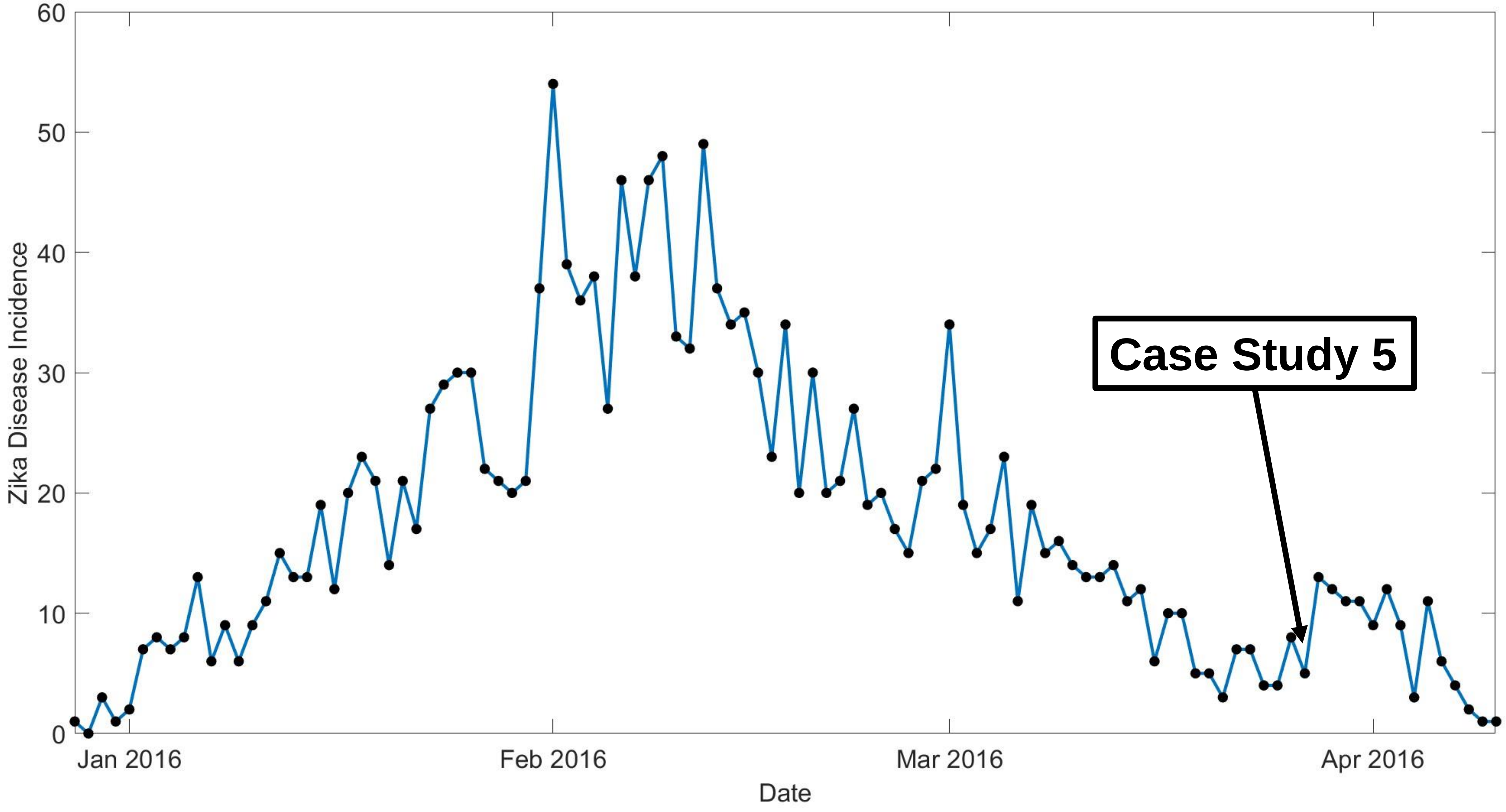

Case Study 4
Zika Disease Incidence
Date
0
10
20
30
40
50
60
Jan 2016
Feb 2016
Mar 2016
Apr 2016

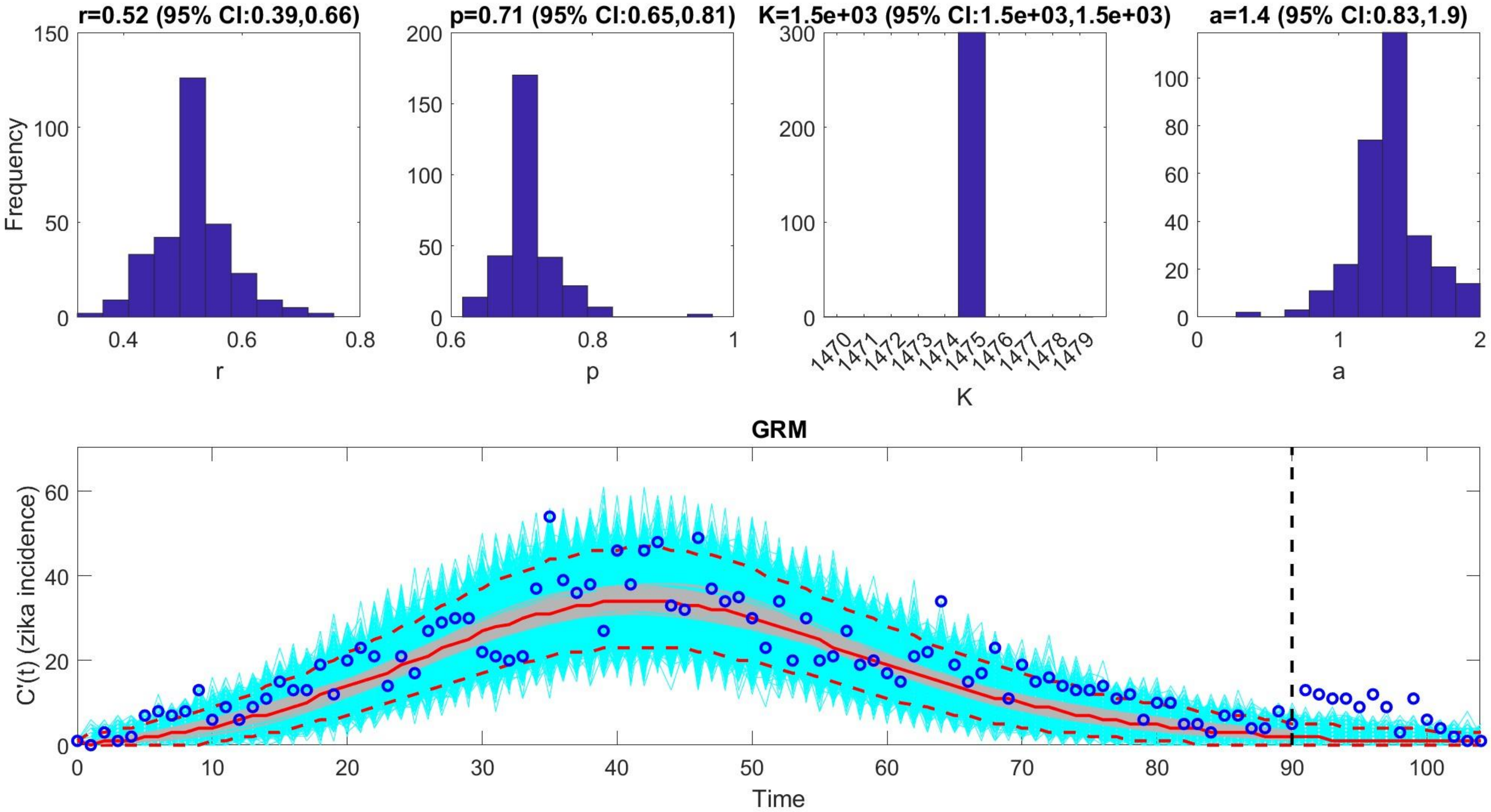


r=0.42 (95% CI:0.36,0.47)
Frequency
r
p=0.77 (95% CI:0.74,0.8)
p
K=1.5e+03 (95% CI:1.5e+03,1.5e+03)
K
GLM model
C'(t) (zika incidence)
Time

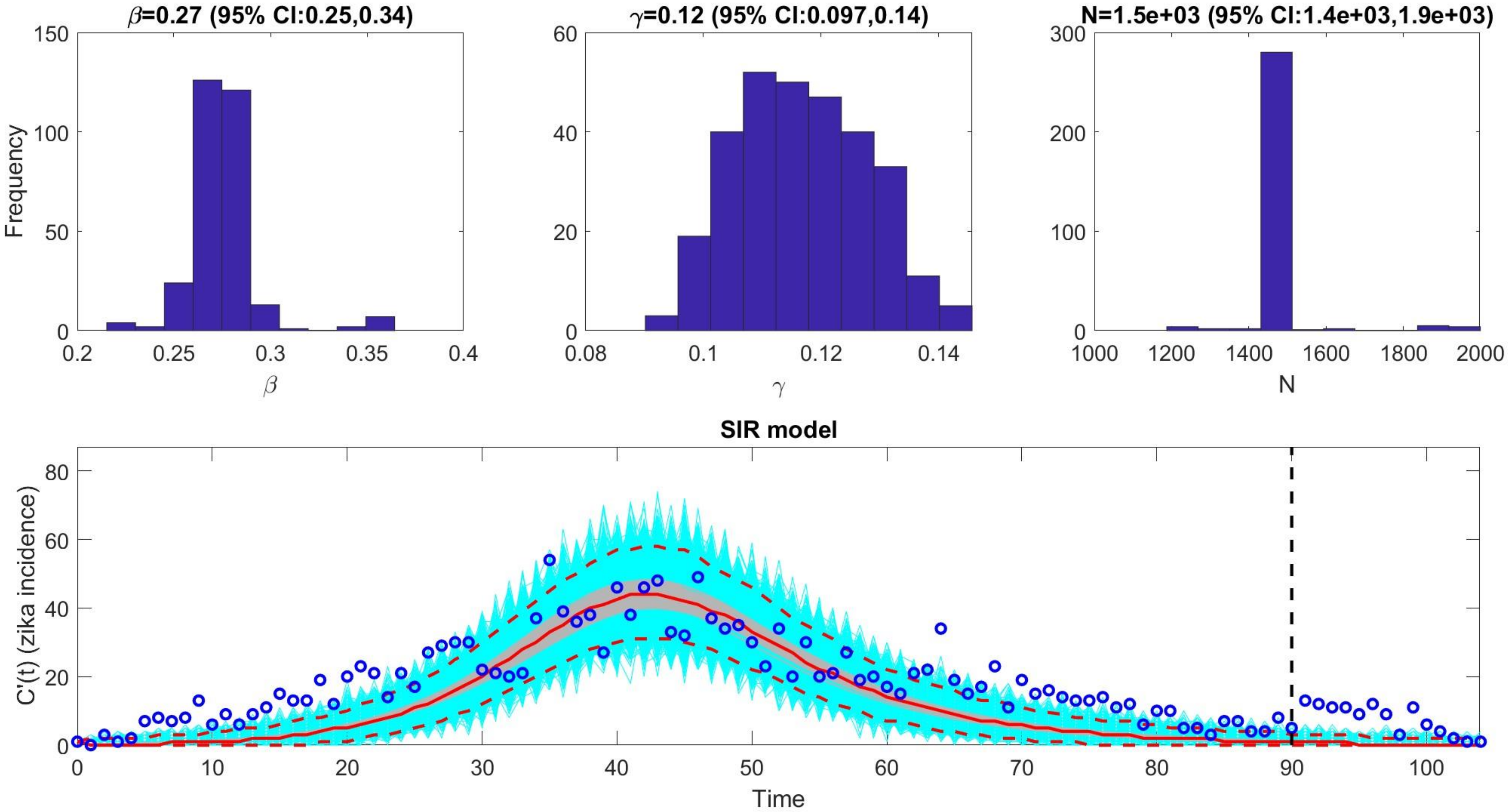


β=9.6 (95% CI:2.4,10)
κ=0.059 (95% CI:0.045,0.073)
γ=1.5 (95% CI:0.17,2.5)
N=2e+03 (95% CI:1.1e+03,2e+03)
Frequency
β
κ
γ
N
SEIR model
C'(t) (zika incidence)
Time

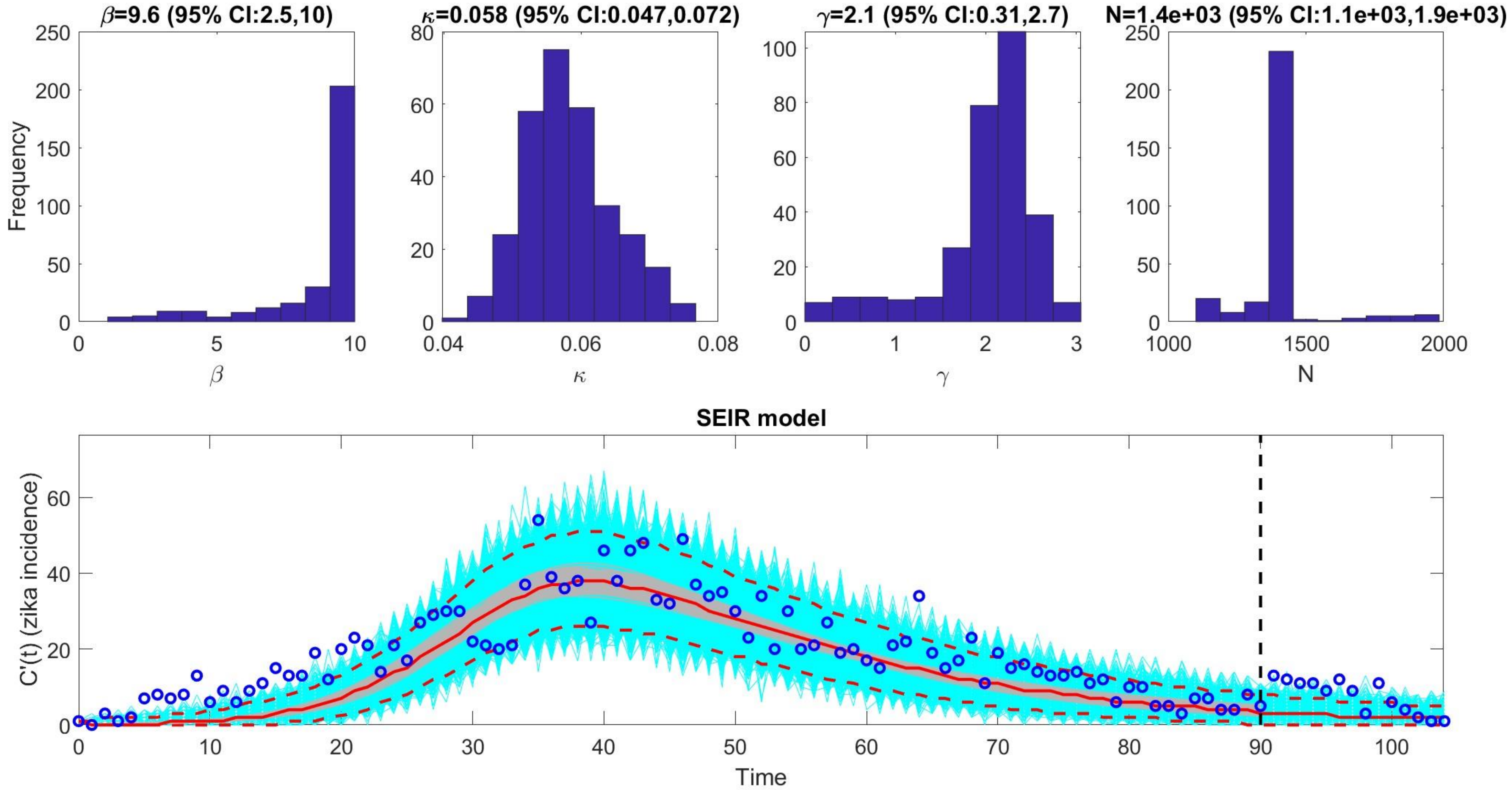

Case Study 5
Zika Disease Incidence
Date
Jan 2016
Feb 2016
Mar 2016
Apr 2016
0
10
20
30
40
50
60



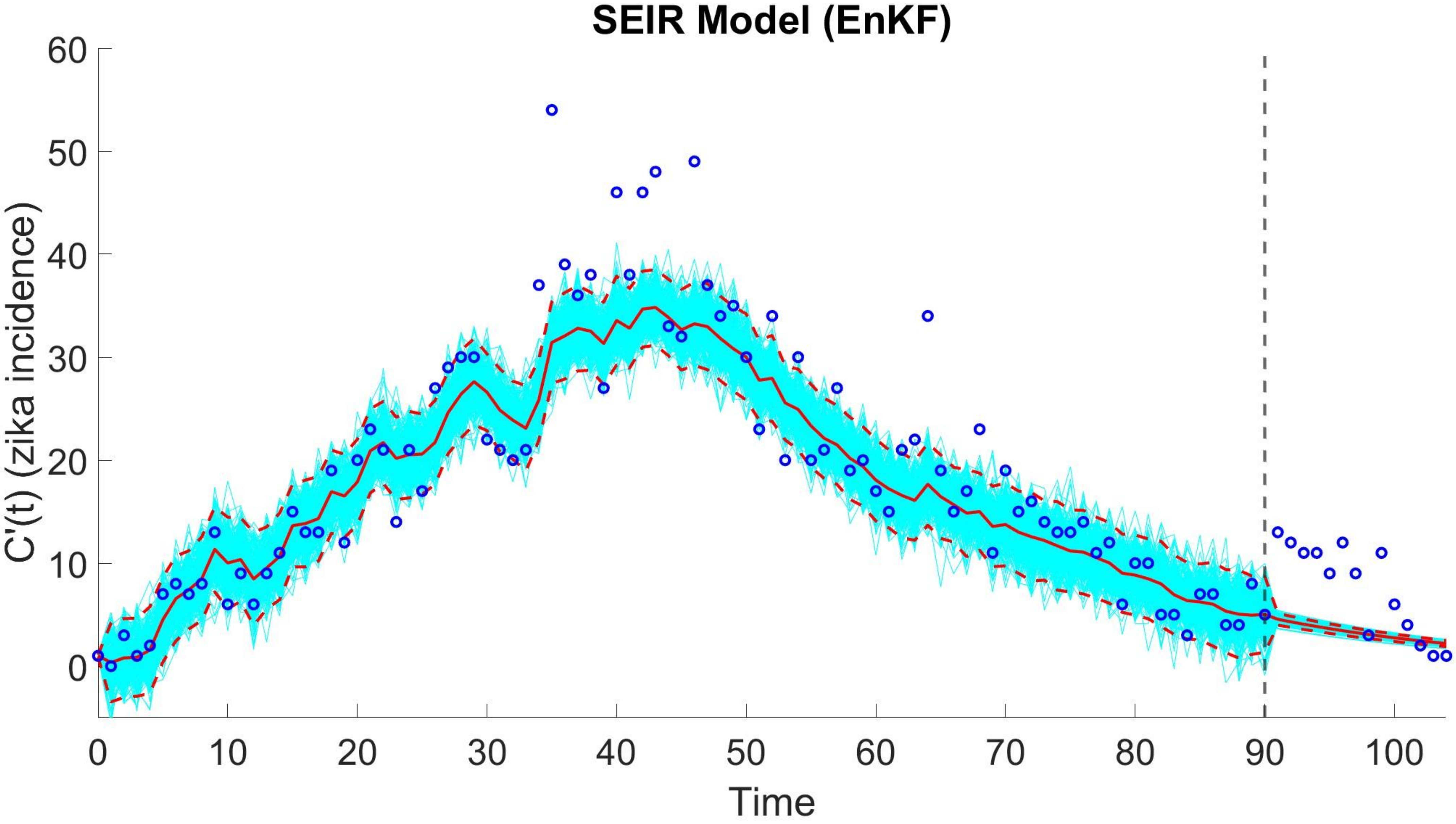
r=0.52 (95% CI:0.39,0.66)
p=0.71 (95% CI:0.65,0.81)
K=1.5e+03 (95% CI:1.5e+03,1.5e+03)
a=1.4 (95% CI:0.83,1.9)
Frequency
r
p
K
a
GRM
C'(t) (zika incidence)
Time



β=0.27 (95% CI:0.25,0.34)
γ=0.12 (95% CI:0.097,0.14)
N=1.5e+03 (95% CI:1.4e+03,1.9e+03)
Frequency
β
γ
N
SIR model
C'(t) (zika incidence)
Time



β=9.6 (95% CI:2.5,10)
κ=0.058 (95% CI:0.047,0.072)
γ=2.1 (95% CI:0.31,2.7)
N=1.4e+03 (95% CI:1.1e+03,1.9e+03)
Frequency
β
κ
γ
N
SEIR model
C'(t) (zika incidence)
Time


SEIR Model (EnKF)
C'(t) (zika incidence)
Time
0
10
20
30
40
50
60
70
80
90
100

Forecasting Performance Metrics					
Model	MAE	MSE	RMSE	Coverage	WIS
GGM	13.961	214.476	14.645	100	4.841
GLM	6.079	52.778	7.265	28.571	5.300
GRM	6.265	55.329	7.438	28.571	5.515
SIR	6.922	65.236	8.077	21.429	6.393
SEIR	5.120	39.638	6.296	42.857	4.236
SIR(EnKF)	5.709	47.053	6.860	*	*
SEIR(EnKF)	4.691	31.860	5.644	*	*

Table 4.5: Performance metrics of model-based forecasts from Case Study 5

Conclusions

- GGM performs well before the peak in Zika incidence
  - The naïve case
- GLM and GRM result in consistent performance across all case studies
- SEIR model demonstrates improvements in forecasting performance as the epidemic progress
    - Including disease mechanism is key
- EnKF technique yields a reduction in prediction errors for the SIR and SEIR models

Limitations

- This study did not include a vector-dynamics into the modeling framework
  - Mosquito transmission is key in understanding Zika
    - However, this study seeks to validate simple models and create a benchmark
- Ensemble Kalman Filter
  - Fixed error structures for state and observational noise

Future Work

- Incorporate vector-borne disease models into forecasting frameworks:
  - Is there an improvement in forecast accuracy when incorporating vector dynamics?
- Investigate observation noise process for the EnKF:
  - Error estimation techniques
- Investigate parameter identifiability of SEIR model
    - Parameter distributions are a diagnostic tool

Acknowledgements

Thank you to my advisor, Dr. Omar Saucedo, for providing mentorship throughout every step of this journey.

Thank you to all the faculty and students of the MathBio lab for creating the most welcoming educational environment.

Thank you to my friends for lifting my spirits when I struggled.

Thank you to Alayna for being patient and understanding.

Finally, thank you to my family for all the love and support.

Thank you!
Any Questions?

References

Forecasting Performance Metrics					
Model	MAE	MSE	RMSE	Coverage	WIS
GGM	21.920	489.729	22.130	0	17.867
GLM	2.989	13.016	3.608	85.714	2.008
GRM	3.801	19.240	4.386	50	2.694
SIR	5.104	31.759	5.636	42.857	4.123
SEIR	2.208	7.775	2.788	100	1.459
SIR(EnKF)	2.589	9.919	3.149	*	*
SEIR(EnKF)	1.846	4.765	2.183	*	*

Table 4.4: Performance metrics of model-based forecasts from Case Study 4

GGM Parameters		
Parameter	Search Interval	Initial Value
r	[0,9]	0.9
p	[0,1]	0.5
GLM Parameters		
Parameter	Search Interval	Initial Value
r	[0,1]	0.5
p	[0,1]	0.5
K	Fixed	1475
GRM Parameters		
Parameter	Search Interval	Initial Value
r	[0,9]	0.9
p	[0,1]	0.5
a	[0,2]	1
K	Fixed	1475

Table A.1: Parameter Search Bounds and Initial Values for GGM, GLM, and GRM Model Fitting

SIR Parameters		
Parameter	Search Interval	Initial Value
β	[0,10]	4
γ	[0,10]	4
N	[1100, 2000]	1475
SEIR Parameters		
Parameter	Search Interval	Initial Value
β	[0,10]	4
κ	[0,4]	1
γ	[0,10]	4
N	[1100, 2000]	1475

Table A.2: Parameter Search Bounds and Initial Conditions for SIR and SEIR Model Fitting

SIR Optimal Parameter Sets			
Case Study	β	γ	N
1	0.78	0.6	1500
2	0.43	0.17	1960
3	0.34	0.13	1689
4	0.31	0.12	1653
5	0.27	0.12	1477

Table A.3: SIR optimal parameter sets for EnKF under each case study

SEIR Optimal Parameter Sets				
Case Study	β	κ	γ	N
1	9.35	0.69	8.17	1260
2	8.79	0.14	4.35	1171
3	9.62	0.07	3.14	1171
4	9.64	0.06	1.51	1987
5	9.62	0.06	2.14	1405

Table A.4: SEIR optimal parameter Sets for EnKF under each case study

VIRGINIA TECH®