Supplementary Information: Rational evaluation of various epidemic models based on COVID-19 data in China

Wuyue Yang[1a)], Dongyan Zhang[2a)], Liangrong Peng[3], Changjing Zhuge[2b)], Liu Hong[1b)]

[1]Zhou Pei-Yuan Center for Applied Mathematics, Tsinghua University, Beijing, 100084, P.R.C.

[2]Beijing Institute for Scientific and Engineering Computing, College of Applied Sciences, Beijing University of Technology, Beijing, 100124, P.R.C.

[3]College of Mathematics and Data Science, Minjiang University, Fuzhou, 350108, P.R.C.

[a)] Those authors contribute equally to this work.

[b)] Author to whom correspondence should be addressed. Electronic mail: zcamhl@tsinghua.edu.cn(L.Hong), zhuge@bjut.edu.cn(C.Zhuge)

CONTENTS

I. Empirical functions 2

II. Statistical methods for estimating the basic reproduction number 3

A. Exponential growth 3

B. Maximum likelihood estimation 4

C. Sequential Bayesian method 4

D. Estimation of time dependent reproduction numbers 4

E. Forecasting the epidemic trends 6

III. Dynamical equations 7

A. SIR model 7

B. SEIR model 7

C. SEIR-QD model 8

D. SEIR-AHQ model 8

E. SEIR-PO model 9

IV. Evaluation Criteria 10

References 11

I. EMPIRICAL FUNCTIONS

To describe the growth of cumulative number of infected cases due to an infectious disease, like COVID-19, empirical functions in explicit forms are widely used[1]. Here, an incomplete list includes the linear, quadratic, cubic, exponential, Logistic, Hill's, Gompertz's

and Richards' functions, whose explicit forms are listed as follows.

$$Linear\,function: \quad C(t) = a_1 t + a_0, \tag{1}$$

$$Quadratic\,function: \quad C(t) = a_2 t^2 + a_1 t + a_0, \tag{2}$$

$$Cubic\,function: \quad C(t) = a_3 t^3 + a_2 t^2 + a_1 t + a_0, \tag{3}$$

$$Exponential\,function: \quad C(t) = K e^{\gamma t}, \tag{4}$$

$$Logistic\,function: \quad C(t) = \frac{K}{1 + e^{-\gamma(t-t_c)}}, \tag{5}$$

$$Hill's\,function: \quad C(t) = \frac{a_1 - a_2}{1 + (t/t_c)^p} + a_2, \tag{6}$$

$$Gompertz's\,function: \quad C(t) = K e^{-e^{-\gamma(t-t_c)}}, \tag{7}$$

$$Richards'\,function: \quad C(t) = \frac{K}{[1 + \beta e^{-\gamma(t-t_c)}]^{1/\beta}}. \tag{8}$$

II. STATISTICAL METHODS FOR ESTIMATING THE BASIC REPRODUCTION NUMBER

When a population is totally susceptible, the basic reproduction number R_0 is defined as the average number of secondary infectious cases produced by one infectious case during a disease outbreak. The estimation of R_0 plays a key role in the study of epidemics of infectious diseases. In literature, there are many different statistical methods for estimating the basic reproduction number ${R_0}^2$. And some of them were implemented with "R_0 package" in R[3].

A. Exponential growth

Exponential growth estimation method assumes that the number of infected cases increases exponentially in the early stage of infection. In this case, the reproduction number is given by[4]

$$R_0 = \frac{1}{M(-\gamma)} = \frac{1}{\int_0^\infty e^{-\gamma\tau}\omega(\tau)d\tau}, \tag{9}$$

where, r is the growth rate and M is the moment generating function of the generation time distribution $\omega(\tau)$. The latter is generally assumed to follow the Gamma distribution.

B. Maximum likelihood estimation

This method assumes the number of infected cases generated from the first case follows the Poisson distribution, whose mean is directly proportional to the basic reproduction number and can be estimated by using the maximum likelihood method[5].

$$ll\left(R_0\right) = \sum_{i=1}^{T} \log\left(\frac{e^{-\mu_i}\mu_i{}^{dI_i}}{dI_i!}\right), \tag{10}$$

$$\mu_i = R_0 \sum_{k=1}^{i} dI_{i-k}\omega_k, \tag{11}$$

in which $ll(R_0)$ is the likelihood depending on R_0. μ_i and $dI_i = I_i - I_{i-1}$ are the number of daily new infected cases and incident cases at discrete time points i, w_i is the generation time distribution. This method also requires the period during which exponential growth is happening to be identified from the data by statistical tools.

C. Sequential Bayesian method

The sequential Bayesian method is also called real-time Bayesian, which starts with a non-informative prior and tries to predict the posterior distribution of the basic reproduction number R_0 by referring to the Bayesian formula[6].

$$P\left(R_0|dI_0,\cdots,dI_{i+1}\right) = \frac{P\left(dI_{i+1}|R_0,dI_0,\cdots,dI_i\right)P\left(R_0|dI_0,\cdots,dI_i\right)}{P\left(dI_0,\cdots,dI_i\right)}, \tag{12}$$

where $P\left(dI_{i+1}|R_0,dI_0,\cdots,dI_i\right)$ is the likelihood of observing incident cases at time $i+1$ given the value of R_0 and past observations of incident cases from time 0 to i, $P\left(R_0|dI_0,\cdots,dI_i\right)$ is a prior distribution of the basic reproduction number, and $P\left(dI_0,\cdots,dI_i\right)$ is the joint probability of observing the incident cases.

The number of daily new infected cases is also assumed to be Poisson distributed with mean

$$\mu_i = dI_{i-1}e^{\gamma(R_0-1)}. \tag{13}$$

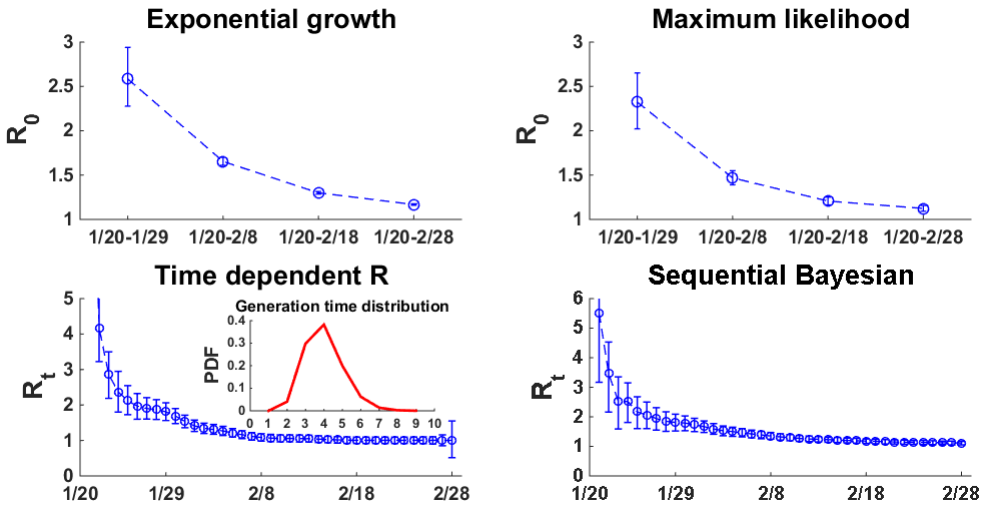
D. Estimation of time dependent reproduction numbers

This method computes reproduction numbers by averaging over all transmission networks compatible with observations[7]. The relative likelihood p_{ij} that a case onset at time i was

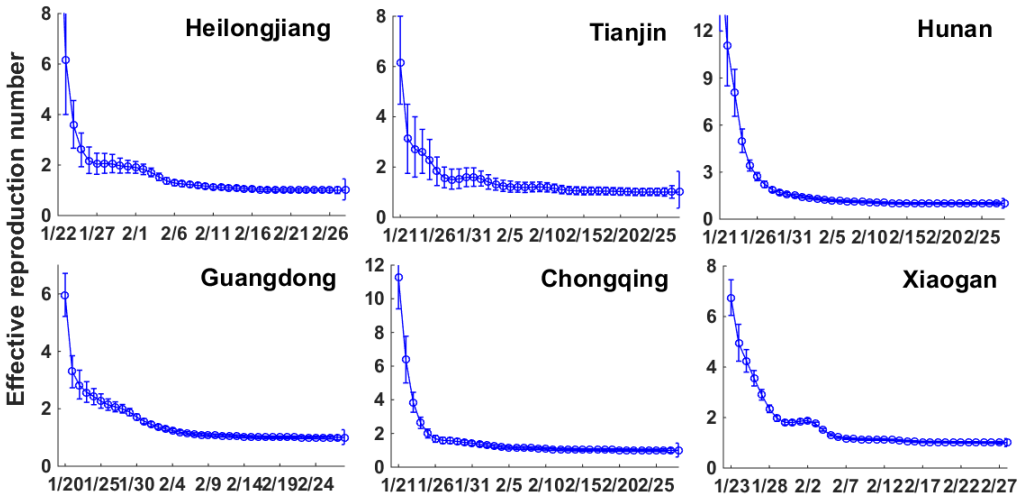
infected by a case onset at time j is given by

$$p_{ij} = \frac{\omega_{i-j}}{\sum_{k=0}^{i-1} \omega_{i-k}}, \tag{14}$$

from which the time-dependent effective reproduction number for case j is defined as $R_j = \sum_i p_{ij}$, and the basic reproduction number is the average of all R_j, i.e. $R_0 = \frac{1}{T}\sum_{j=1}^{T} R_j$.



(a)Shanghai from 01/20/2020 to 02/28/2020.



(b)Effective reproduction number for six provinces/cities in China based on the method of time dependent reproduction number.

FIG. 1. Basic/effective reproduction number for COVID-19 derived from four different statistical methods.

E. Forecasting the epidemic trends

The basic goal of statistical methods is to estimate the basic reproduction number. So when we plan to make predictions on the progression of epidemics, we need to combine them with further assumptions on the dynamics. A most widely adopted one is the exponential growth, which assumes the number of infected populations grows exponentially with the time and the exponent γ can be obtained from the basic reproduction number R_0.

$$R_0 = \frac{1}{M(-\gamma)} = \frac{1}{\int_0^\infty e^{-\gamma\tau}\omega(\tau)d\tau}.$$

In the current study, we assume the generating time distribution $\omega(t)$ obeys the Gamma distribution $\Gamma(k,\theta)$ (see Fig. 1), whose moment generating function is explicitly known as

$$M(t) = (1-t\theta)^{-k}; \forall t < \frac{1}{\theta} \tag{15}$$

From it, we immediately see $\gamma = (R_0^{1/k} - 1)/\theta$. Then inserting γ into either the recurrence formula $\mu_i = dI_{i-1}e^{\gamma(R_0-1)}$ (no free parameter, see Fig. 2) or the Logistic function (two more free parameters, figure in the main text), the progression of epidemics is fitted and predicted.

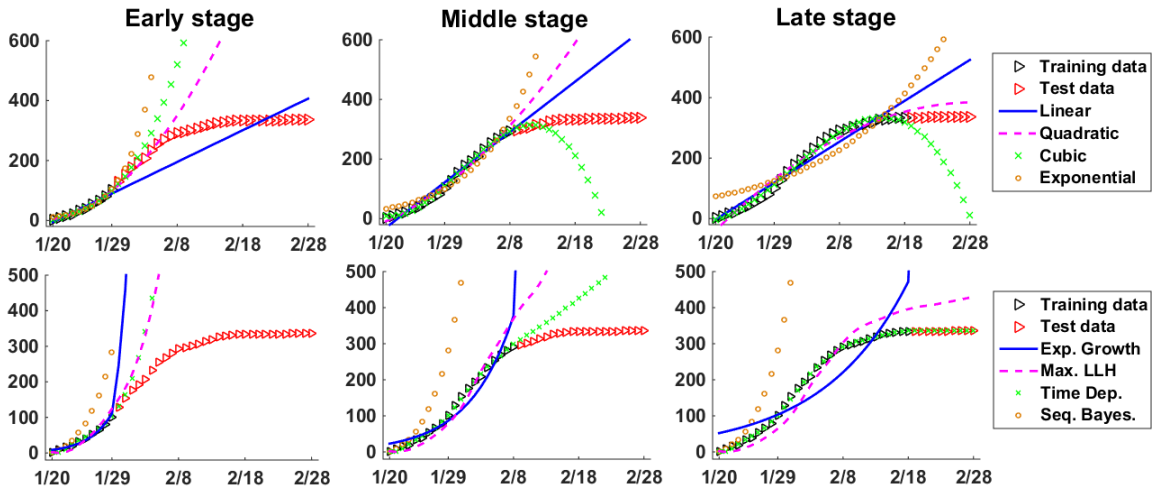


FIG. 2. Forecast of the COVID-19 epidemic in Shanghai from 01/20/2020 to 02/28/2020 based on data of first 10 (early), 20 (middle) and 30 (late) days respectively. The upper column shows the results for linear, quadratic, cubic and exponential functions, while the lower column gives those for four different statistical methods combined with the exponential growth model.

III. DYNAMICAL EQUATIONS

Without considering time delay and spatial inhomogeneity, ordinary differential equations are most widely used models for describing the procedure of epidemics caused by infectious diseases. Here we summarize six different dynamical models reported in literature for studying COVID-19.

A. SIR model

In the classical SIR model, there are three different populations, that are $S(t)$, $I(t)$ and $R(t)$ denoting at time t the respective number of the *susceptible cases*, *infectious cases* (with infectious capacity and not yet recovered) and *recovered cases* (recovered and not be either infectious or infected once again). Their relations are characterized by the following ODEs.

$$\frac{dS(t)}{dt} = -\beta SI, \tag{16}$$

$$\frac{dI(t)}{dt} = \beta SI - \delta I, \tag{17}$$

$$\frac{dR(t)}{dt} = \delta I. \tag{18}$$

Here coefficients β and δ represent the infection rate and recovery rate separately.

B. SEIR model

To account for the infected cases which are still in a latent period and not yet be infectious, a new *exposed* population $E(t)$ is introduced into the SEIR model[8]. Correspondingly, the ODEs for the SIR model are generalized to

$$\frac{dS(t)}{dt} = -\beta SI, \tag{19}$$

$$\frac{dE(t)}{dt} = \beta SI - \gamma E, \tag{20}$$

$$\frac{dI(t)}{dt} = \gamma E - \delta I, \tag{21}$$

$$\frac{dR(t)}{dt} = \delta I, \tag{22}$$

in which the coefficient γ denotes the transition rate of exposed individuals to the infected.

C. SEIR-QD model

To take the effects of quarantine and self-protection into consideration, Peng *et al.*[9] proposed to generalize the classical SEIR model by introducing a new *quarantined* state between *infectious* and *recovery*. The numbers of *death* and *unsusceptible* are denoted as $D(t)$ and $S^A(t)$ separately.

$$\frac{dS(t)}{dt} = -\beta SI - \alpha S, \tag{23}$$
$$\frac{dE(t)}{dt} = \beta SI - \gamma E, \tag{24}$$
$$\frac{dI(t)}{dt} = \gamma E - \lambda I, \tag{25}$$
$$\frac{dQ(t)}{dt} = \lambda I - \delta Q - \kappa Q, \tag{26}$$
$$\frac{dR(t)}{dt} = \delta Q, \tag{27}$$
$$\frac{dD(t)}{dt} = \kappa Q, \tag{28}$$
$$\frac{dS^A(t)}{dt} = \alpha S. \tag{29}$$

in which coefficients $\alpha, \lambda, \delta, \kappa$ denote the protection rate of susceptible individuals, the transition rate of infectious individuals to the quarantined infected class, the recovery rate and the death rate respectively.

D. SEIR-AHQ model

To incorporate appropriate compartments relevant to interventions such as quarantine, isolation and treatment, Tang *et al.*[10] generalized the SEIR model and stratified the populations as *susceptible* (S), *exposed* (E), infectious but not yet symptomatic (*pre-symptomatic*) (A), *infectious* with symptoms (I), *hospitalized* (H) and *recovered* (R) compartments, and further stratified the population to include *quarantined susceptible* (S_q), *isolated exposed*

(E_q) and *isolated infected* (I_q) compartments.

$$\frac{dS(t)}{dt} = -(\beta c + cq(1-\beta))S(I+\theta A) + \lambda S_q, \tag{30}$$
$$\frac{dE(t)}{dt} = \beta c(1-q)S(I+\theta A) - \sigma E, \tag{31}$$
$$\frac{dI(t)}{dt} = \sigma\rho E - (\delta_I + \alpha + \gamma_I)I, \tag{32}$$
$$\frac{dA(t)}{dt} = \sigma(1-\rho)E - \gamma_A A, \tag{33}$$
$$\frac{dS_q(t)}{dt} = cq(1-\beta)S(I+\theta A) - \lambda S_q, \tag{34}$$
$$\frac{dE_q(t)}{dt} = \beta cqS(I+\theta A) - \delta_q E_q, \tag{35}$$
$$\frac{dH(t)}{dt} = \delta_I I + \delta_q E_q - (\alpha + \gamma_H)H, \tag{36}$$
$$\frac{dR(t)}{dt} = \gamma_I I + \gamma_A A + \gamma_H H. \tag{37}$$

In above model, parameters $\{c, \beta, q, \sigma, \lambda, \rho, \delta_I, \delta_q, \gamma_I, \gamma_A, \gamma_H, \alpha\}$ represent the contact rate, probability of transmission per contact, quarantined rate of exposed individuals, transition rate of exposed individuals to the infected, rate at which the quarantined uninfected contacts were released into the wider community class, probability of having symptoms among infected individuals, transition rate of symptomatic infected individuals to the quarantined infected class, transition rate of quarantined exposed individuals to the quarantined infected class, recovery rates of symptomatic infected individuals, asymptomatic infected individuals and quarantined infected individuals, as well as disease-induced death rate.

E. SEIR-PO model

By incorporating the public opinion on COVID-19, Zhang *et al.* further classified the populations of *susceptible* and *exposed* in SEIR model into *unconscious* (S^U, E^U) and *conscious*

(S^A, E^A) based on their different knowledge on epidemics and self-protection.

$$\frac{dS^U(t)}{dt} = -\beta S^U(\eta_1 I + \eta_2 E^A + \eta_3 E^U) - \alpha S^U, \tag{38}$$

$$\frac{dS^A(t)}{dt} = -\eta\beta S^A(\eta_1 I + \eta_2 E^A + \eta_3 E^U) + \alpha S^U, \tag{39}$$

$$\frac{dE^U(t)}{dt} = \beta S^U(\eta_1 I + \eta_2 E^A + \eta_3 E^U) - \alpha E^U - \gamma E^U, \tag{40}$$

$$\frac{dE^A(t)}{dt} = \eta\beta S^A(\eta_1 I + \eta_2 E^A + \eta_3 E^U) + \alpha E^U - \gamma E^A, \tag{41}$$

$$\frac{dI(t)}{dt} = \gamma(E^U + E^A) - \delta I, \tag{42}$$

$$\frac{dR(t)}{dt} = \delta I, \tag{43}$$

where parameters $\{\gamma, \delta, \beta, \eta, \eta_1, \eta_2, \eta_3, \alpha\}$ denote the transition rate of exposed individuals to the infected, recovery rate of infected individuals, infection rate of unconscious susceptible population, reduced infection ratio of conscious susceptible individuals, effective infection factors of infectious individuals, unconscious and conscious exposed individuals, as well as the spreading rate of knowledge about COVID-19 between individuals.

IV. EVALUATION CRITERIA

There are several criteria to evaluate the performance of regression models. Suppose x_i and y_i to be the true values and predicted ones separately.

(1) The root mean square error (RMSE) is defined as

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2}. \tag{44}$$

(2) The Akaike information criterion (AIC) was introduced by Akaike in the early 1970s[11]. It is based on the concept of entropy, and incorporate the model complexity and its goodness of fit together.

$$AIC = 2k - 2\ln(L), \tag{45}$$

in which k is the total number of free parameters in a model, while L is the likelihood function. Under the assumption of Gaussian distributions for residues, the AIC reduces to $AIC = 2k + 2N\ln(RMSE)$.

To eliminate the dependence on sample size, McQuarrie and Tsai[12] introduced AICc as

$$AICc = 2\ln(RMSE) + (N+k)/(N-k-2). \tag{46}$$

(3) In the current study, the robustness index (RB) is defined as ratio as the ratio between the minimum number of confirmed infected cases at the end of prediction time and its maximum within the 95% confidence region.

(4) There are also many other quantities which can be used to characterize the epidemic dynamics. One is the inflection point (IFP)[2], which is defined as the time point when the daily new infected cases reaches its maximum, or when the second order derivative of cumulative infected cases becomes non-positive. The other is the C_{95} point, which is the time point when the cumulative infected cases reaches 95% of its final value.

REFERENCES

[1]Shi Zhao, Salihu S Musa, Hao Fu, Daihai He, and Jing Qin. Simple framework for real-time forecast in a data-limited situation: the zika virus (zikv) outbreaks in brazil from 2015 to 2016 as an example. *Parasites & vectors*, 12(1):344, 2019.

[2]Jinghua Li, Yijing Wang, Stuart Gilmour, Mengying Wang, Daisuke Yoneoka, Ying Wang, Xinyi You, Jing Gu, Chun Hao, Liping Peng, et al. Estimation of the epidemic properties of the 2019 novel coronavirus: A mathematical modeling study. 2020.

[3]Thomas Obadia, Romana Haneef, and Pierre-Yves Boëlle. The r0 package: a toolbox to estimate reproduction numbers for epidemic outbreaks. *BMC medical informatics and decision making*, 12(1):147, 2012.

[4]Jacco Wallinga and Marc Lipsitch. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences*, 274(1609):599–604, 2007.

[5]Laura Forsberg White and Marcello Pagano. A likelihood-based method for real-time estimation of the serial interval and reproductive number of an epidemic. *Statistics in medicine*, 27(16):2999–3016, 2008.

[6]Luis MA Bettencourt and Ruy M Ribeiro. Real time bayesian estimation of the epidemic potential of emerging infectious diseases. *PLoS One*, 3(5), 2008.

[7]Jacco Wallinga and Peter Teunis. Different epidemic curves for severe acute respiratory

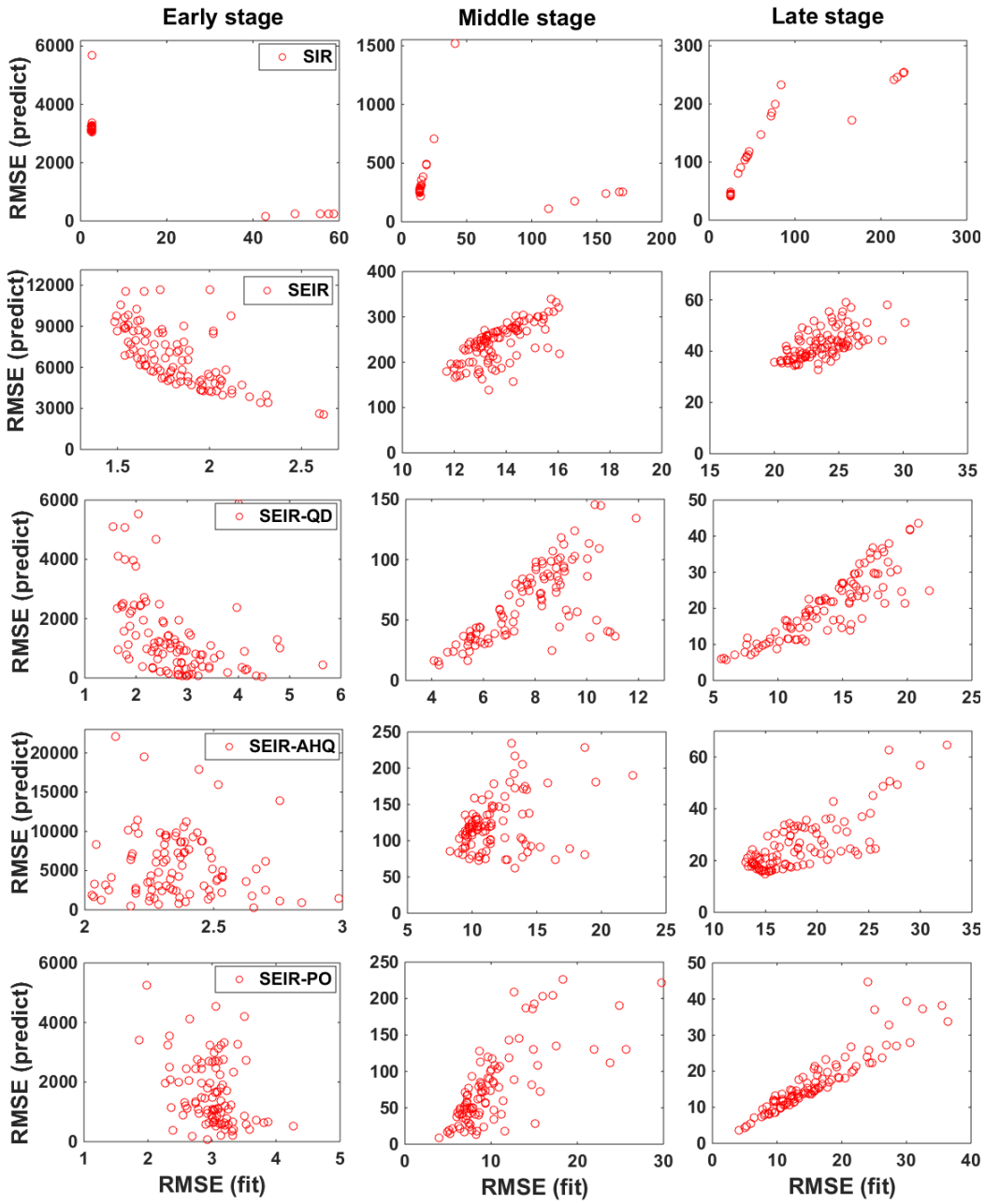


FIG. 3. The correlation between RMSE of training data set and RMSE of testing data set for five dynamical models based on first 10 (early), 20 (middle) and 30 (late) days data of COVID-19 epidemic in Shanghai from 01/20/2020 to 02/28/2020.

syndrome reveal similar impacts of control measures. *American Journal of epidemiology*, 160(6):509–516, 2004.

[8]Huwen Wang, Zezhou Wang, Yinqiao Dong, Ruijie Chang, Chen Xu, Xiaoyue Yu, Shuxian Zhang, Lhakpa Tsamlag, Meili Shang, Jinyan Huang, et al. Phase-adjusted estimation of

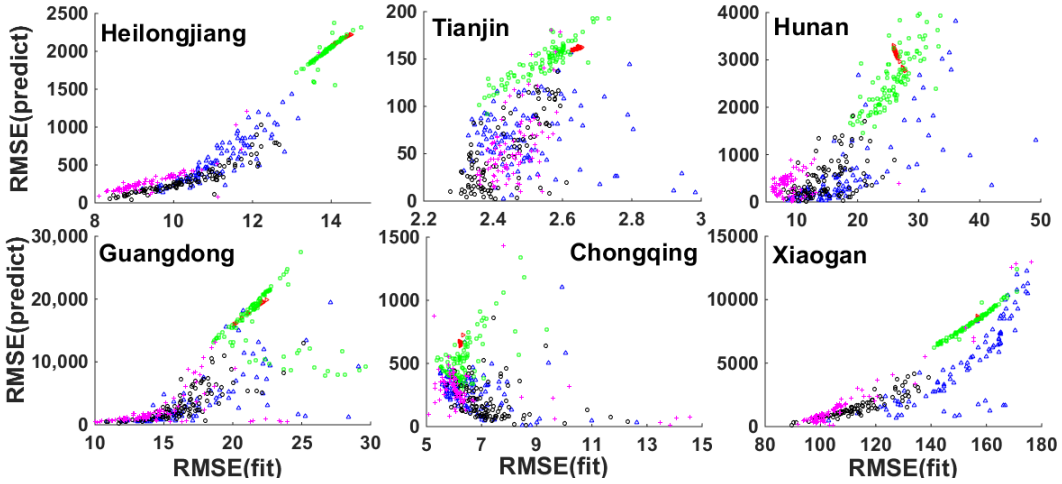


FIG. 4. The correlation between RMSE of training data set and RMSE of testing data set for five dynamical models based on COVID-19 epidemic data from 01/20/2020-02/28/2020 for six provinces/cities mentioned in the main text.

the number of coronavirus disease 2019 cases in wuhan, china. *Cell Discovery*, 6(1):1–8, 2020.

[9]Liangrong Peng, Wuyue Yang, Dongyan Zhang, Changjing Zhuge, and Liu Hong. Epidemic analysis of covid-19 in china by dynamical modeling. *arXiv preprint arXiv:2002.06563*, 2020.

[10]Biao Tang, Xia Wang, Qian Li, Nicola Luigi Bragazzi, Sanyi Tang, Yanni Xiao, and Jianhong Wu. Estimation of the transmission risk of the 2019-ncov and its implication for public health interventions. *Journal of Clinical Medicine*, 9(2), 2020.

[11]Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[12]Allan D. R. Mcquarrie and Chih Ling Tsai. *The Univariate Regression Model*. 1998.