



HAL
open science

Asymptotic estimates of SARS-CoV-2 infection counts and their sensitivity to stochastic perturbation

Davide Faranda, Isaac Pérez Castillo, Oliver Hulme, Aglaé Jézéquel, Jeroen
Lamb, Erica Thompson

► **To cite this version:**

Davide Faranda, Isaac Pérez Castillo, Oliver Hulme, Aglaé Jézéquel, Jeroen Lamb, et al.. Asymptotic estimates of SARS-CoV-2 infection counts and their sensitivity to stochastic perturbation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 2020, 30 (5), pp.051107. 10.1063/5.0008834 . hal-02668288

HAL Id: hal-02668288

<https://hal.science/hal-02668288>

Submitted on 8 Jul 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **Asymptotic estimates of SARS-CoV-2 infection counts and their**
2 **sensitivity to stochastic perturbation**

3 Davide Faranda*

4 *Laboratoire des Sciences du Climat et de l'Environnement,*
5 *CEA Saclay l'Orme des Merisiers, UMR 8212 CEA-CNRS-UVSQ,*
6 *Université Paris-Saclay & IPSL, 91191 Gif-sur-Yvette, France*

7 *London Mathematical Laboratory, 8 Margravine Gardens, London, W6 8RH, UK and*
8 *LMD/IPSL, Ecole Normale Supérieure,*
9 *PSL research University, Paris, France*

10 Isaac Pérez Castillo

11 *Department of Quantum Physics and Photonics, Institute of Physics,*
12 *UNAM, P.O. Box 20-364, 01000 Mexico City, Mexico and*
13 *London Mathematical Laboratory, 8 Margravine Gardens, London, W6 8RH, UK*

14 Oliver Hulme

15 *Danish Research Centre for Magnetic Resonance,*
16 *Centre for Functional and Diagnostic Imaging and Research,*
17 *Copenhagen University Hospital Hvidovre,*
18 *Kettegard Allé 30, 2650, Hvidovre, Denmark and*
19 *London Mathematical Laboratory, 8 Margravine Gardens, London, W6 8RH, UK*

20 Aglaé Jezequel

21 *LMD/IPSL, ENS, PSL Université, École Polytechnique,*
22 *Institut Polytechnique de Paris, Sorbonne Université, CNRS, Paris France and*
23 *Ecole des Ponts, Marne-la-Vallée, France*

24 Jeroen S.W. Lamb

25 *Department of Mathematics, Imperial College London, UK and*
26 *London Mathematical Laboratory, 8 Margravine Gardens, London, W6 8RH, UK*

27 Yuzuru Sato

28
29
30
31
32
33
34
35
36

*RIES/Department of Mathematics, Hokkaido University,
N20 W10, Kita-ku, Sapporo, Hokkaido 001-0020, Japan and
London Mathematical Laboratory, 8 Margravine Gardens, London, W6 8RH, UK*

Erica L. Thompson
*Centre for the Analysis of Time Series,
London School of Economics and Political Science,
Houghton Street, London WC2A 2AE and
London Mathematical Laboratory, 8 Margravine Gardens, London, W6 8RH, UK*
(Dated: March 25, 2020)

Abstract

37

38 Despite the importance of having robust estimates of the time-asymptotic total number of in-
39 fections, early estimates of COVID-19 show enormous fluctuations. Using COVID-19 data for
40 different countries, we show that predictions are extremely sensitive to the reporting protocol and
41 crucially depend on the last available data-point, before the maximum number of daily infections
42 is reached. We propose a physical explanation for this sensitivity, using a Susceptible-Exposed-
43 Infected-Recovered (SEIR) model where the parameters are stochastically perturbed to simulate
44 the difficulty in detecting asymptomatic patients, different confinement measures taken by differ-
45 ent countries, as well as changes in the virus characteristics. Our results suggest that there are
46 physical and statistical reasons to assign low confidence to statistical and dynamical fits, despite
47 their apparently good statistical scores. These considerations are general and can be applied to
48 other epidemics.

49 I. LEAD PARAGRAPH

50 **COVID-19 is currently affecting over 180 countries in the world and poses seri-**
51 **ous threats to public health as well as economic and social stability of many coun-**
52 **tries. Modeling and extrapolating in near real-time the evolution of COVID-19**
53 **epidemics is a scientific challenge, which requires a deep understanding of the**
54 **non-linearities undermining the dynamics of the epidemics. Here we show that**
55 **real-time predictions of COVID-19 infections are extremely sensitive to errors**
56 **in data collection and crucially depend on the last available data-point. We test**
57 **these ideas in both statistical (logistic) and dynamical (Susceptible-Exposed-**
58 **Infected-Recovered) models that are currently used to forecast the evolution of**
59 **the COVID-19 epidemic. Our goal is to show how uncertainties arising from**
60 **both poor data quality and inadequate estimations of model parameters (incu-**
61 **bation, infection and recovery rates) propagate to long term extrapolations of**
62 **infections count. We provide guidelines for reporting those uncertainties to the**
63 **scientific community and the general public.**

* Correspondence to davide.faranda@lsce.ipsl.fr

64 II. INTRODUCTION

65 SARS-CoV-2, a zoonotic virus of the coronavirus family [1], that provokes an infectious
66 disease known as COVID-19, has emerged in China at the end of 2019, affecting first the
67 Hubei province and quickly spreading to all Chinese provinces [2]. The failure of initial
68 containment measures caused the virus to spread internationally, and on March 11th, The
69 World Health Organization (WHO) declared COVID-19 a pandemic [3]. According to the
70 WHO Situation Report-59 released on March 19th [4], the number of countries affected by
71 the pandemic is 176, with 209 839 confirmed infections and 8778 deaths. As this report
72 also notices: *the number of confirmed cases worldwide has exceeded 200 000. It took over*
73 *three months to reach the first 100 000 confirmed cases, and only 12 days to reach the next*
74 *100 000*, an astonishing development, due to the highly contagious character of SARS-CoV-2.

75
76 SARS-CoV-2 causes potentially life-threatening form of pneumonia in a non-negligible
77 patients fraction [5]. Enormous efforts to contain the virus and to not overwhelm inten-
78 sive care facilities are currently taken all over the world. Following the drop in infections
79 observed in the Hubei province, restrictive confinement measures have been taken in many
80 countries [6]. Most of the time, those measures are taken on the basis of epidemics models,
81 which are fitted with dynamical or statistical models on the available data.

82
83 COVID-19 data should be provided daily, following a request of the WHO. To date, the
84 WHO guidelines require countries to report, at each day t , the total number of infected
85 patients $I(t)$ as well as the number of deaths $D(t)$. Unfortunately, there is large variability
86 in the way both $I(t)$ and $D(t)$ are counted. We provide some illustrative example: on the
87 one hand, Italy shows the highest fatality rate:

$$88 \quad f = \sum_{t=1}^{\tau} D(t) / \sum_{t=1}^{\tau} I(t) \simeq 0.07 \quad (1)$$

89 possibly because $D(t)$ includes all deaths who have contracted SARS-CoV-2, indepen-
90 dently on whether the virus is the first cause of death. Moreover, in a recent interview [7],
91 Italian biologist Bucci has stated that $D(t)$ can be underestimated because it does not
92 include those patients who died at home without being tested. On the other hand, in
93 Germany, the fatality rate is extremely low $f \simeq 0.002$. There may be several explanations

94 for this : some query data methodology (e.g. a different method to determine $D(t)$) while
95 others say high testing rates are giving a more accurate picture [8].

96

97 Great uncertainties also exist in the count of $I(t)$. Whereas in the early stage of the epi-
98 demic several countries tested asymptomatic individuals to track back the infection chain,
99 recent policies to estimate $I(t)$ have changed. Most of the western countries now test only
100 patients displaying severe SARS-CoV-2 symptoms. In an effort of tracking all the chain
101 of infections, South Korea has tested many asymptomatic people. This latter strategy has
102 proven effective in supporting actions to reduce the rate of new infections. A recent study [9]
103 has estimated that an enormous part of total infections were undocumented (80% to 90%)
104 and that those undetected infections were the source for 79% of documented cases in China.

105

106 The goal of this paper is to analyse the effect of those large uncertainties in real-time
107 forecasting of the long term behavior of the COVID-19 epidemic [10]. As stated by Polonsky
108 et al [11], there is a need for defining robust methods to assess both the intrinsic errors
109 inherent to fitting procedures as well as those introduced by poor data-quality. Funk et al
110 [12] give a concrete example of this applied to the Ebola epidemics in the Western Area region
111 of Sierra Leone in 2014-15. Classically, epidemiologists rely on Susceptible-Exposed-Infected-
112 Recovered (SEIR) models [13]. These models consist of ordinary differential equations where
113 a population is divided into compartments, with the assumption that every individual in
114 the same compartment has the same characteristics. In SEIR, population is divided into
115 Susceptible, Exposed, Infected and Recovered individuals. Such models predict a sigmoid
116 shape of the total number of infections $C(t) = \sum_{t=1}^{\tau} I(t)$. Using the available national
117 data points $I(t)$ one can obtain long term estimates on the total of COVID-19 infections
118 in each country. This paper focuses on the estimation of the sensitivity of these models to
119 the last available data point, before the inflection point of the $I(t)$ curve is reached. We
120 use SEIR models to show the possible origins of this sensitivity by perturbing the relevant
121 parameters, often assumed deterministic, with a noise that mimics changes in the way the
122 virus is spreading, e.g. as a result of application of confinement measures, or the presence
123 (rate/magnitude) of super-spreaders [14]. The paper is organised as follows: in Section III
124 we discuss the various sources of data for COVID-19 and their shortcomings, and then we
125 discuss in detail the SEIR model and its statistical modelling. In Section IV we discuss the

126 results focusing on the statistical sensitivity of the modelling, and apply it to data from
127 France, UK and Italy. We finish, in Section V, with some remarks and point out some
128 potentially beneficial policy guidelines.

129 III. DATA AND MODELLING

130 A. Data

131 The data repository used in this paper for COVID-19 data is a Visual Dashboard operated
132 by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE).
133 The data repository [15] is also supported by ESRI Living Atlas Team and the Johns Hopkins
134 University Applied Physics Lab (JHU APL). We used datasets of cases confirmed with
135 a laboratory test, irrespective of clinical signs and symptoms [3]. The data contains, as
136 recognized by the public authorities that dispatched them, several inhomogeneities due to
137 the different ways of testing patients with suspicious symptoms. As an example, Italy
138 announced on Feb. 26 that it relaxed testing criteria to the point that contacts linked to
139 confirmed cases or recent travelers to outbreak areas would not be tested anymore, unless
140 they show symptoms. Unlike Italy, South Korea (population of 51 million) is testing 15000
141 to 20000 individuals per day since Feb. 27 with the goal to minimize hospital pressure
142 and stop the epidemics in the early stages [16]. COVID-19 data also suffers from reporting
143 problems due to the local management of health infrastructures. In Italy, healthcare is a
144 regional task and everyday data are collected at a regional level and transmitted to the
145 Protezione Civile, who transfers the data to WHO. Many inconsistencies and delays have
146 been documented in this transfer process [17]. A similar situation occurs in Mexico, in
147 which for instance, private institutions, either hospitals or laboratories, do not possess the
148 necessary national and international certifications given by the *Instituto de Diagnóstico*
149 *y Referencia Epidemiológicos* (InDRE) and therefore their tests are not considered valid
150 and must be redone by certified institutions[18], thus unnecessarily delaying the release
151 of accurate daily reports. COVID-19 data of Mexico was collected from the daily reports
152 generated by Mexico's *Secretaría de Salud*[19]. Our goal is to account for these uncertainties
153 in the modelling of COVID-19 data.

154 **B. An epidemiological Susceptible-Exposed-Infected-Recovered model**

155 The Susceptible-Exposed-Infected-Recovered (SEIR) model [13] is an epidemiological
 156 compartmental model where a total population N is divided into susceptible individuals
 157 S , exposed individuals E , infected individuals I , and the number R of people who have
 158 had the disease and are now either recovered or dead (and assumed not to be susceptible
 159 to reinfection). The model is constructed under the assumption that the total population
 160 $N = S(t) + E(t) + I(t) + R(t)$ does not vary. This implies:

161
$$0 = dN/dt = dS/dt + dE/dt + dI/dt + dR/dt, \quad \forall t \geq 0. \quad (2)$$

162 The model relies on some assumptions. First of all, susceptible individuals end up becom-
 163 ing infected and infected individuals can only recover or die. Individuals who are exposed (E)
 164 have had contact with an infected person, but are not themselves infectious. Furthermore,
 165 those who have recovered or died are forever immune. It is also assumed that susceptibility
 166 is equal for all and that it is proportional to the product of $I(t)$ and $S(t)$ at a time t . These
 167 assumptions lead us to a set of four ordinary differential equations:

168
$$\frac{dS}{dt} = -\lambda S(t)I(t) \quad (3)$$

169
$$\frac{dE}{dt} = \lambda S(t)I(t) - \alpha E(t) \quad (4)$$

170
$$\frac{dI}{dt} = \alpha E(t) - \gamma I(t) \quad (5)$$

171
$$\frac{dR}{dt} = \gamma I(t). \quad (6)$$

172 Here $\gamma > 0$ represents the mean recovery/death rate, or $1/\gamma$ the mean infection period,
 173 $\lambda = \lambda_0/S(0) > 0$ is considered the contact or infection rate of the disease and it is rescaled
 174 by the initial number of susceptible individuals $S(0)$ and α is the inverse of the incubation
 175 period. These expressions satisfy (2) as required. Because data are reported only on a daily
 176 basis, we adopt the discrete SEIR model:

177
$$S(t + 1) = S(t) - \lambda S(t)I(t) \quad (7)$$

178
$$E(t + 1) = (1 - \alpha)E(t) + \lambda S(t)I(t) \quad (8)$$

179
$$I(t + 1) = (1 - \gamma)I(t) + \alpha E(t) \quad (9)$$

180
$$R(t + 1) = R(t) + \gamma I(t). \quad (10)$$

181 This model is obtained rewriting the ordinary differential equations 3-6 with an Euler
182 Scheme and fixing $dt = 1$ day. An important derived quantity of the model is $R_0 = \lambda_0/\gamma$,
183 the average reproduction number of the virus in a population. This quantity represents the
184 number of cases, on average, an infected person will cause during their infectious period.
185 For COVID-19 in Wuhan in January 2020, $R_0 = 2.68$ with 95% CrI 2.47–2.86 according to
186 an estimate performed with Wuhan data [20]. Dynamical modelling of COVID-19 epidemic
187 has been proposed in [21]. In that study, the authors used a Susceptible-Exposed-Infected-
188 Recovered model with delays and performed a sensitivity study on the parameters. Fixing
189 $\lambda \simeq 1$ as in [21] and $\gamma = 0.37$ to recover the value of R_0 found in [20] (assuming that the
190 behavioural elements of viral transmission are consistent in other populations), we are left
191 with the choice of α . The range for incubation period of SARS-CoV-2 has been determined
192 in [22] between 2 and 11 days. As a comparison, this range is estimated to be between 2
193 and 5 days for human coronavirus, and between 2 and 10 days for severe acute respiratory
194 syndrome (SARS) coronavirus [23]. Using a trial and error procedure and a subjective
195 estimation of the quality of the fit, we obtain the best fit when we set $\alpha = 0.27$ (corresponding
196 to an incubation period between 3 and 4 days) and initial conditions $S(0) = 33000$, $I(0) = 2$,
197 $E(0) = R(0) = 0$. The fit against the Chinese data is reported in Figure 1. We are aware
198 that a log-likelihood method with cross-validation would provide a better fit as well as an
199 estimate of the uncertainty. However, we underline that the goal of this work is not to provide
200 the best possible model but rather to explore the sensitivity of it to perturbations. First
201 of all, we note that, despite its simplicity, the model shows qualitatively similar behaviour
202 to the published data. Note that there is a discontinuity in the dataset, which is due to a
203 change in the way infections were counted, introduced on Feb. 12, 2020 [24].

204 This model has also evident deficiencies in representing the COVID-19 infections. First of
205 all the total population N here is to be intended as a number of people who could have been in
206 contact with infected individuals. Furthermore, the population under consideration does not
207 consist of a group of about the same age and general health level, and the the group members
208 do not mix homogeneously. The model does not have any spatial component, nor does it
209 predict the influences of policy and behavioural responses to the progress of the pandemic.
210 More complex models introducing further parameters would likely lead to overfitting and
211 over-confident predictions, due to the limited volume of data currently available. No model
212 will be sufficient to predict the outcome of this pandemic: the outcome depends on our

213 response. Models are presented here with the aim of generating some insight into the overall
214 behaviour and the risks entailed by inaction.

215 C. Statistical Modelling

216 When insight is limited and compartmental models are not suited, phenomenological sta-
217 tistical models provide a starting point for estimation of key transmission parameters, such
218 as the reproduction number, and forecasts of epidemic impact [25]. One of the simplest ways
219 to model the epidemics is to observe that the function $C(t)$ is a sigmoid function and perform
220 a statistical fit of the data to extrapolate the long-term behavior of the epidemics [26, 27].
221 Among all the possible sigmoid functions, two have proven useful in fitting epidemic growth:
222 the generalized logistic distribution [28] and the generalized Gompertz distribution [29]. A
223 complete overview of sigmoid functions is presented in [30], although applied to in a differ-
224 ent context. Since our considerations are independent of the sigmoid function used, we will
225 present results for the generalized logistic model only. The model reads:

$$226 \quad C(t) = a/(1 + b \cdot \exp(-c \cdot t)); \quad (11)$$

227 where a , b and c are parameters of the model. They are linked in a non-explicit way
228 to the solution of the SEIR model. A fit to the Chinese data is presented in Figure 2.
229 Logistic fits are performed with the MATLAB Nonlinear least-squares solver constraining
230 objective function with gradient. At first sight, one can be tempted to use $R^2 \simeq 0.997$ as
231 a quality indicator of the fit. However, we stress that R^2 is not an appropriate measure
232 for nonlinear regression models: given the smoothness of data, there will be lots of models
233 (eg low-order polynomial) which could fit well (get a very good R^2) but would not make
234 credible predictions [31]. These data are however collected at a mature stage of the epidemic
235 and as such the characteristics of the logistic fit to these data can be assigned with greater
236 confidence. In the next section we will discuss the performance of the statistical model in
237 the early stage of the epidemics, where the logistic function can be used to extrapolate the
238 behavior of $C(t)$.

239 IV. RESULTS: STATISTICAL AND DYNAMICAL MODELLING OF EARLY 240 STAGES OF THE EPIDEMICS

241 A. Statistical sensitivity

242 We begin by showing the sensitivity of the logistic extrapolations in the early stage of
243 the epidemics by looking at French data from Mar. 04 to Mar 20. France has previously
244 recorded sporadic cases of SARS-CoV-2 infections but the exponential growth phase started
245 at the beginning of March 2020. To show the high sensitivity to the last point of the
246 datasets we first perform a logistic fit with data starting from different dates and ending
247 Mar. 20 (Figure 3a) and then do the reverse experiment by fitting data starting on Mar.
248 04 but ending at different dates (Figure 3b). Clearly, fits are more stable by removing days
249 from the beginning of the outbreak than from the most recent past. Again, we stress the
250 inadequacy of the R^2 metric as it yields values above $R^2 > 0.99$ for all cases considered in
251 Figure 3. The analysis suggests that, if a large error is presented in the last data point, the
252 extrapolation has less predictive adequacy. This implies very narrow estimates of confidence
253 intervals for $C(t)$: for each fit, confidence intervals are as small as the thickness of the line
254 used in the plots in Figure 3. This prevents a correct evaluation of the confidence interval,
255 which is critical to assess the uncertainties around the future evolution of the epidemics,
256 and to build relevant policies to address the worst case scenario.

257

To further test this concept, we now assume we are uncertain about the magnitude of the last data point $C(t^*)$. To simulate this uncertainty, we replace it with a random number $\xi(t^*)$ drawn from a discrete uniform distribution with mean $C(t^*)$ and standard deviation $0.2C(t^*)$. We therefore construct an ensemble of 100 possible trajectories under this generative process. Results are presented in Figure 4 for UK (a), France (b) and Italy (c). To date, Italy is at a more mature stage of the epidemic, while France and UK face an earlier stage. This is reflected in the spread of the ensemble: for the UK, forecasting the epidemic with a logistic fit is not informative of the course of the epidemic: the ensemble spread just suggests that the current phase is an exponential growth and at best it can inform that worst case scenarios should be considered at this point. The ensemble spread reduces when the epidemics is at a more mature stage (Italy). Indeed, if we set $b = 1$ and

we start the fit from time t_0 then the logistic distribution is written:

$$C(t) = a/(1 + \exp(-c(t - t_0))).$$

258 In the early growth phase, $\exp(-c(t - t_0)) \gg 1$, so:

$$C(t) \sim a \exp(c(t - t_0)) = a \exp(-c \cdot t_0) \exp(-c \cdot t) = A \exp(-c \cdot t)$$

259 .

260 Even though we can fit A and b to data, recalling that $A = a \exp(-c \cdot t_0)$ we have huge
261 freedom over a , the upper asymptote that determines the final count of the epidemics.

262 **B. Dynamical sensitivity in a stochastic SEIR model**

263 Another way to understand the sensitivity in epidemics is to release the assumption
264 that incubation period α , infection rate λ and recovery rate γ are constant through the
265 epidemics [32]. Intrinsically they can vary, because of the presence of individuals with an
266 extremely high transmission rate known as super-spreaders [14], or due to the release or the
267 application of confinement measures, or changes in the SARS-CoV-2 characteristics. They
268 can also display spurious variations due to the way data are reported or collected, for the
269 problems specified above. We explore all these possibilities by considering α , λ , and γ as
270 time varying processes. The idea of using stochastic models to represent epidemics is not new
271 to the literature [33–35]. In the modelling of COVID-19 infections can be further justified
272 by the evidence that $R_0 = \lambda/\gamma$ displays spatial and temporal variability [11]. For example,
273 Wu et al. [20] show fluctuations of R_0 in different Chinese regions. These differences are
274 due to changes in the duration of contagiousness, likelihood of infection per contact and the
275 contact rate [36] which depends on demographic spatial variability [37]. There is however
276 little consensus on which variables or parameters should be perturbed in order to get a
277 realistic behavior. Our goal here is different than obtaining the best possible forecasts of the
278 epidemics as we want to understand which parameter causes a large sensitivity in the final
279 $C(t)$ counts. Let us begin, by alternatively replacing in Equations 7-10 one of the constant
280 parameters $\kappa \in \{\alpha, \lambda, \gamma\}$ with a stochastic process:

$$281 \quad \kappa(t) = \kappa_0 + \sigma \cdot \xi(t) \tag{12}$$

282 where σ is the intensity of the perturbation and $\xi(t)$ a random variable drawn from
 283 a normal distribution $N(0,1)$ at each time. The purpose of equation 12 is to introduce
 284 instantaneous discrete jumps in the values of the daily parameters. This discrete process,
 285 used in [38], is more appropriate than a continuous one (see, e.g. [39]) when observations
 286 are affected by large detection errors, as in the present case. Figure 5 shows an example
 287 of 30 realisations of a stochastic SEIR COVID-19 model, obtained by replacing alternately
 288 α (a,b), λ (b,d) and γ with the stochastic process in Eq 12 and using $\sigma = 0.2\kappa_0$ to get
 289 fluctuations of the order of 20% of each parameter values, in analogy with the statistical
 290 sensitivity studies performed in the previous section. The sensitivity clearly depends on the
 291 perturbed parameter: a perturbation on α mostly implies a different timing of the epidemics
 292 while the final cumulative number of infections $C(t)$ remains unchanged. Perturbations on
 293 λ and γ affect the final $C(t)$ in a deeper way, leading to a total variation in the number of
 294 cases of the order of 20%. Indeed, by changing λ and γ , we also modify the basic reproduc-
 295 tion number R_0 . The idea of having a time-varying reproduction number has been already
 296 exploited in [40], although the authors have directly modelled the dynamics of a dynamic
 297 reproduction number $R(t)$ without introducing a SEIR model.

298

299 As a further step, we add noise simultaneously to all parameters of the SEIR model via
 300 Equation 12. Six realisations of the model are shown in Figure 6. Figure 6-a,b) shows the
 301 evolution of $S(t)$, $R(t)$, $E(t)$ and $C(t)$. We have separated the time evolution of $I(t)$ in
 302 Figure 6-c) to compare it with that of COVID-19 data for China, South Korea and Italy
 303 (Figure 6d). Despite having a quasi-smooth behavior of $C(t)$, we observe a highly non-
 304 smoothness of $I(t)$, which is reflected by the data. The sensitivity of the model is higher
 305 when $I(t)$ is large, because γ and λ directly act on $I(t)$. Therefore, when approaching the
 306 maximum of $I(t)$ ($t \sim 50$ days) small changes in the parameters can greatly affect the final
 307 total count of infections $C(t)$. This implies that mitigation strategies based on the reduction
 308 of λ by self-isolation, social distancing, are way more effective if imposed at the early stage
 309 of the epidemics, as they can suppress positive fluctuations of $I(t)$ and help reducing R_0 .

310 V. DISCUSSION

311 In this work we have discussed the statistical and dynamical sensitivity of asymptotic
312 estimates of COVID-19 infections when performed at the early stages of the epidemics.
313 First of all, we noted that SEIR model, with λ , γ and α inferred from clinical studies,
314 can fit Chinese data with a value of $N \simeq 33000$ that is very different from that of the
315 Chinese, Hubei or Wuhan populations. This enormous discrepancy can be due both to a
316 large underestimation in the number of total cases, or to the effectiveness of confinement
317 measures which results in a smaller exposed population. This estimate should be taken
318 as a first caveat in fitting a SEIR model to infer COVID-19 epidemics evolution in other
319 countries as results may be largely under/over-estimated [11].

320

321 Then, we have shown that statistical fits often used to extrapolate the long term behav-
322 ior of the epidemics are greatly affected by the magnitude of the last data point, despite
323 values of R^2 close to one, leading to unrealistic or over-confident estimates of confidence
324 intervals on the forecast of the total number of infections [41, 42]. In the early stage of the
325 epidemics, we have shown that knowing the last data point with a relative 20% error, can
326 lead to a final extrapolation of infections with an error of several orders of magnitude. In
327 order to improve the estimates of statistical models one should replace R^2 estimates by a
328 formal comparison of model-alternatives using information criteria (e.g. AIC or BIC) or
329 a log-likelihood approach with a leave-one-out cross-validation procedure. A simple cross
330 validation can follow both the approaches described in this paper: i) exclude the last data
331 points and check the stability of the estimates, ii) add noise to the last data point and
332 obtain an ensemble of estimates. Another approach could be based on evaluating every day
333 each model on the performance in predicting the new data point, and then used again with
334 the new data point for an updated estimate.

335

336 Finally, we have investigated whether this statistical sensitivity can be dynamically re-
337 produced with a SEIR model where parameters are considered stochastic processes (Equa-
338 tion 12). We have found that the stochastic dynamics are more sensitive to γ and λ .
339 Perturbations on these parameters are proportional to the number of infected patients $I(t)$
340 and are therefore important in the growth phase of the epidemics. Actual data display fluc-

341 tuations even larger than those simulated in the stochastic models, suggesting that instead
342 of assuming observational Gaussian noise on the parameters, jump processes (e.g. Levy
343 noise) may be more appropriate [43]. Furthermore, we noticed that large fluctuations in
344 the number of detected infections is also due to changes in the testing protocols and avail-
345 ability of tests. All these inconsistencies prevent the possibility of performing meaningful
346 asymptotic statistical or dynamical modelling for COVID-19, or comparing results among
347 different countries. This may be even more problematic in least developed countries, which
348 are just beginning to register cases [44–46].

349

350 Our study suggests that dynamical and statistical modelling should focus on limited
351 stages of the epidemics and restrict the analysis to specific regions, accounting for large un-
352 certainties as done in [47]. Modelling approaches should take into account both statistical
353 uncertainties as well as expert knowledge in a sort of Bayesian framework that allows to
354 guide the choice of prior probabilities [10]. In the interest of preserving the public health of
355 as many individuals as possible, once modelled the uncertainty in the data, the worst case
356 scenarios should always be taken into account very seriously as a guideline to enforce strict
357 confinement measures.

358

359 VI. ACKNOWLEDGMENTS

360 This paper is dedicated to the memory of F Molinari, who recently passed away from
361 COVID-19. DF acknowledges A Adamou, B Dubrulle, F Pons, F Daviaud, P Yiou, M
362 Kagayema, S Fromang and G Ramstein for useful discussions.

363 VII. DATA AVAILABILITY

364 The data that support the findings of this study are openly available in [https://
365 systems.jhu.edu/research/public-health/ncov/](https://systems.jhu.edu/research/public-health/ncov/), maintained by Johns Hopkins Uni-

- 367 [1] Eleanor R Gaunt, Andrew Hardie, Eric CJ Claas, Peter Simmonds, and Kate E Templeton.
368 Epidemiology and clinical presentations of the four human coronaviruses 229e, hku1, nl63, and
369 oc43 detected over 3 years using a novel multiplex real-time pcr method. *Journal of clinical*
370 *microbiology*, 48(8):2940–2947, 2010.
- 371 [2] Jin Wu, Weiyi Cai, Derek Watkins, and James Glanz. How the virus got out. *The New York*
372 *Times*.
- 373 [3] World Health Organization et al. Coronavirus disease 2019 (covid-19): situation report, 51.
374 2020.
- 375 [4] World Health Organization et al. Coronavirus disease 2019 (covid-19): situation report, 59.
376 2020.
- 377 [5] Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang,
378 Guohui Fan, Jiuyang Xu, Xiaoying Gu, et al. Clinical features of patients infected with 2019
379 novel coronavirus in wuhan, china. *The Lancet*, 395(10223):497–506, 2020.
- 380 [6] Roy M Anderson, Hans Heesterbeek, Don Klinkenberg, and T Déirdre Hollingsworth. How
381 will country-based mitigation measures influence the course of the covid-19 epidemic? *The*
382 *Lancet*, 395(10228):931–934, 2020.
- 383 [7] Luca Fraioli. Bucci: “dalla lombardia numeri ormai insensati. i contagiati sono di più”.
384 *Repubblica*, Mar 2020.
- 385 [8] Philip Oltermann. Germany’s low coronavirus mortality rate intrigues experts. *The Guardian*.
- 386 [9] Ruiyun Li, Sen Pei, Bin Chen, Yimeng Song, Tao Zhang, Wan Yang, and Jeffrey Shaman.
387 Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus
388 (sars-cov2). *Science*, 2020.
- 389 [10] Angel N Desai, Moritz UG Kraemer, Sangeeta Bhatia, Anne Cori, Pierre Nouvellet, Mark
390 Herringer, Emily L Cohn, Malwina Carrion, John S Brownstein, Lawrence C Madoff, et al.
391 Real-time epidemic forecasting: Challenges and opportunities. *Health security*, 17(4):268–275,
392 2019.
- 393 [11] Jonathan A Polonsky, Amrish Baidjoe, Zhian N Kamvar, Anne Cori, Kara Durski, W John
394 Edmunds, Rosalind M Eggo, Sebastian Funk, Laurent Kaiser, Patrick Keating, et al. Out-

- 395 break analytics: a developing data science for informing the response to emerging pathogens.
396 *Philosophical Transactions of the Royal Society B*, 374(1776):20180276, 2019.
- 397 [12] Sebastian Funk, Anton Camacho, Adam J Kucharski, Rachel Lowe, Rosalind M Eggo, and
398 W John Edmunds. Assessing the performance of real-time epidemic forecasts: A case study
399 of ebola in the western area region of sierra leone, 2014-15. *PLoS computational biology*,
400 15(2):e1006785, 2019.
- 401 [13] Fred Brauer. Compartmental models in epidemiology. In *Mathematical epidemiology*, pages
402 19–79. Springer, 2008.
- 403 [14] James O Lloyd-Smith, Sebastian J Schreiber, P Ekkehard Kopp, and Wayne M Getz. Super-
404 spreading and the effect of individual variation on disease emergence. *Nature*, 438(7066):355–
405 359, 2005.
- 406 [15] Data last downloaded on Mar. 23 from @x [https://systems.jhu.edu/research/
407 public-health/ncov/](https://systems.jhu.edu/research/public-health/ncov/).
- 408 [16] Kim Arin. Drive-thru clinics, drones: Korea’s new weapons in virus fight. *The Korea Herald*.
- 409 [17] Pagella Politica AGI. Come vanno letti i dati sul coronavirus in italia. *AGI Agenzia Italia*.
- 410 [18] As of 20th of March 2020 only two private hospitals in Mexico have been certified by InDRE
411 to carry out tests.
- 412 [19] There is a delay between the data reported daily by the WHO and that reported by Mexico
413 health authorities.
- 414 [20] Joseph T Wu, Kathy Leung, and Gabriel M Leung. Nowcasting and forecasting the potential
415 domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a
416 modelling study. *The Lancet*, 395(10225):689–697, 2020.
- 417 [21] Liangrong Peng, Wuyue Yang, Dongyan Zhang, Changjing Zhuge, and Liu Hong. Epidemic
418 analysis of covid-19 in china by dynamical modeling. *arXiv preprint arXiv:2002.06563*, 2020.
- 419 [22] Stephen A Lauer, Kyra H Grantz, Qifang Bi, Forrest K Jones, Qulu Zheng, Hannah R Mered-
420 ith, Andrew S Azman, Nicholas G Reich, and Justin Lessler. The incubation period of coron-
421 avirus disease 2019 (covid-19) from publicly reported confirmed cases: Estimation and appli-
422 cation. *Annals of Internal Medicine*, 2020.
- 423 [23] Justin Lessler, Nicholas G Reich, Ron Brookmeyer, Trish M Perl, Kenrad E Nelson, and
424 Derek AT Cummings. Incubation periods of acute respiratory viral infections: a systematic
425 review. *The Lancet infectious diseases*, 9(5):291–300, 2009.

- 426 [24] Amy Gunia and Michael Zennie. China reported a huge increase in new covid-19 cases. here's
427 why it's actually a step in the right direction. *Time*.
- 428 [25] Gerardo Chowell, Doracelly Hincapie-Palacio, Juan Ospina, Bruce Pell, Amna Tariq, Sushma
429 Dahal, Seyed Moghadas, Alexandra Smirnova, Lone Simonsen, and Cécile Viboud. Using phe-
430 nomenological models to characterize transmissibility and forecast patterns and final burden
431 of zika epidemics. *PLoS currents*, 8, 2016.
- 432 [26] Gerardo Chowell. Fitting dynamic models to epidemic outbreaks with quantified uncertainty:
433 a primer for parameter uncertainty, identifiability, and forecasts. *Infectious Disease Modelling*,
434 2(3):379–398, 2017.
- 435 [27] Raimund Bürger, Gerardo Chowell, and Leidy Yissedt Lara-Díaz. Comparative analysis of
436 phenomenological growth models applied to epidemic outbreaks. *Mathematical biosciences
437 and engineering: MBE*, 16(5):4250–4273, 2019.
- 438 [28] Pierre-François Verhulst. Notice sur la loi que la population suit dans son accroissement.
439 *Corresp. Math. Phys.*, 10:113–126, 1838.
- 440 [29] Benjamin Gompertz. Xxiv. on the nature of the function expressive of the law of human
441 mortality, and on a new mode of determining the value of life contingencies. in a letter to
442 francis baily, esq. frs &c. *Philosophical transactions of the Royal Society of London*, (115):513–
443 583, 1825.
- 444 [30] IJ Wellock, GC Emmans, and I Kyriazakis. Describing and predicting potential growth in the
445 pig. *Animal Science*, 78(3):379–388, 2004.
- 446 [31] Andrej-Nikolai Spiess and Natalie Neumeyer. An evaluation of r^2 as an inadequate measure
447 for nonlinear models in pharmacological and biochemical research: a monte carlo approach.
448 *BMC pharmacology*, 10(1):6, 2010.
- 449 [32] Hao Xiong and Huili Yan. Simulating the infected population and spread trend of 2019-ncov
450 under different policy by eir model. *Available at SSRN 3537083*, 2020.
- 451 [33] Lars Folke Olsen and William Morris Schaffer. Chaos versus noisy periodicity: alternative
452 hypotheses for childhood epidemics. *Science*, 249(4968):499–504, 1990.
- 453 [34] Hakan Andersson and Tom Britton. *Stochastic epidemic models and their statistical analysis*,
454 volume 151. Springer Science & Business Media, 2012.
- 455 [35] Joseph Dureau, Konstantinos Kalogeropoulos, and Marc Baguelin. Capturing the time-varying
456 drivers of an epidemic using stochastic dynamical systems. *Biostatistics*, 14(3):541–555, 2013.

- 457 [36] Giulio Viceconte and Nicola Petrosillo. Covid-19 r0: Magic number or conundrum? *Infectious*
458 *Disease Reports*, 12(1), 2020.
- 459 [37] Ilya Kashnitsky. Covid-19 in unequally ageing european regions. 2020.
- 460 [38] Davide Faranda and Sandro Vaienti. Extreme value laws for dynamical systems under obser-
461 vational noise. *Physica D: Nonlinear Phenomena*, 280:86–94, 2014.
- 462 [39] Davide Faranda, Yuzuru Sato, Brice Saint-Michel, Cecile Wiertel, Vincent Padilla, Bérengère
463 Dubrulle, and François Daviaud. Stochastic chaos in a turbulent swirling flow. *Physical review*
464 *letters*, 119(1):014502, 2017.
- 465 [40] Adam J Kucharski, Timothy W Russell, Charlie Diamond, Yang Liu, John Edmunds, Sebas-
466 tian Funk, Rosalind M Eggo, Fiona Sun, Mark Jit, James D Munday, et al. Early dynamics of
467 transmission and control of covid-19: a mathematical modelling study. *The Lancet Infectious*
468 *Diseases*, 2020.
- 469 [41] Andrea Remuzzi and Giuseppe Remuzzi. Covid-19 and italy: what next? *The Lancet*, 2020.
- 470 [42] Choujun Zhan, K Tse Chi, Zhikang Lai, Tianyong Hao, and Jingjing Su. Prediction of covid-19
471 spreading profiles in south korea, italy and iran by data-driven coding. *medRxiv*, 2020.
- 472 [43] Xianghua Zhang and Ke Wang. Stochastic seir model with jumps. *Applied Mathematics and*
473 *Computation*, 239:133–143, 2014.
- 474 [44] Joost Hopman, Benedetta Allegranzi, and Shaheen Mehtar. Managing covid-19 in low-and
475 middle-income countries. *JAMA*, 2020.
- 476 [45] Marius Gilbert, Giulia Pullano, Francesco Pinotti, Eugenio Valdano, Chiara Poletto, Pierre-
477 Yves Boëlle, Eric D’Ortenzio, Yazdan Yazdanpanah, Serge Paul Eholie, Mathias Altmann,
478 et al. Preparedness and vulnerability of african countries against importations of covid-19: a
479 modelling study. *The Lancet*, 2020.
- 480 [46] J Steenhuisen and S Nebehay. Countries rush to build diagnostic capacity as coronavirus
481 spreads. *Reuters*, 2020.
- 482 [47] Jonathan M Read, Jessica RE Bridgen, Derek AT Cummings, Antonia Ho, and Chris P
483 Jewell. Novel coronavirus 2019-ncov: early estimation of epidemiological parameters and
484 epidemic predictions. *MedRxiv*, 2020.

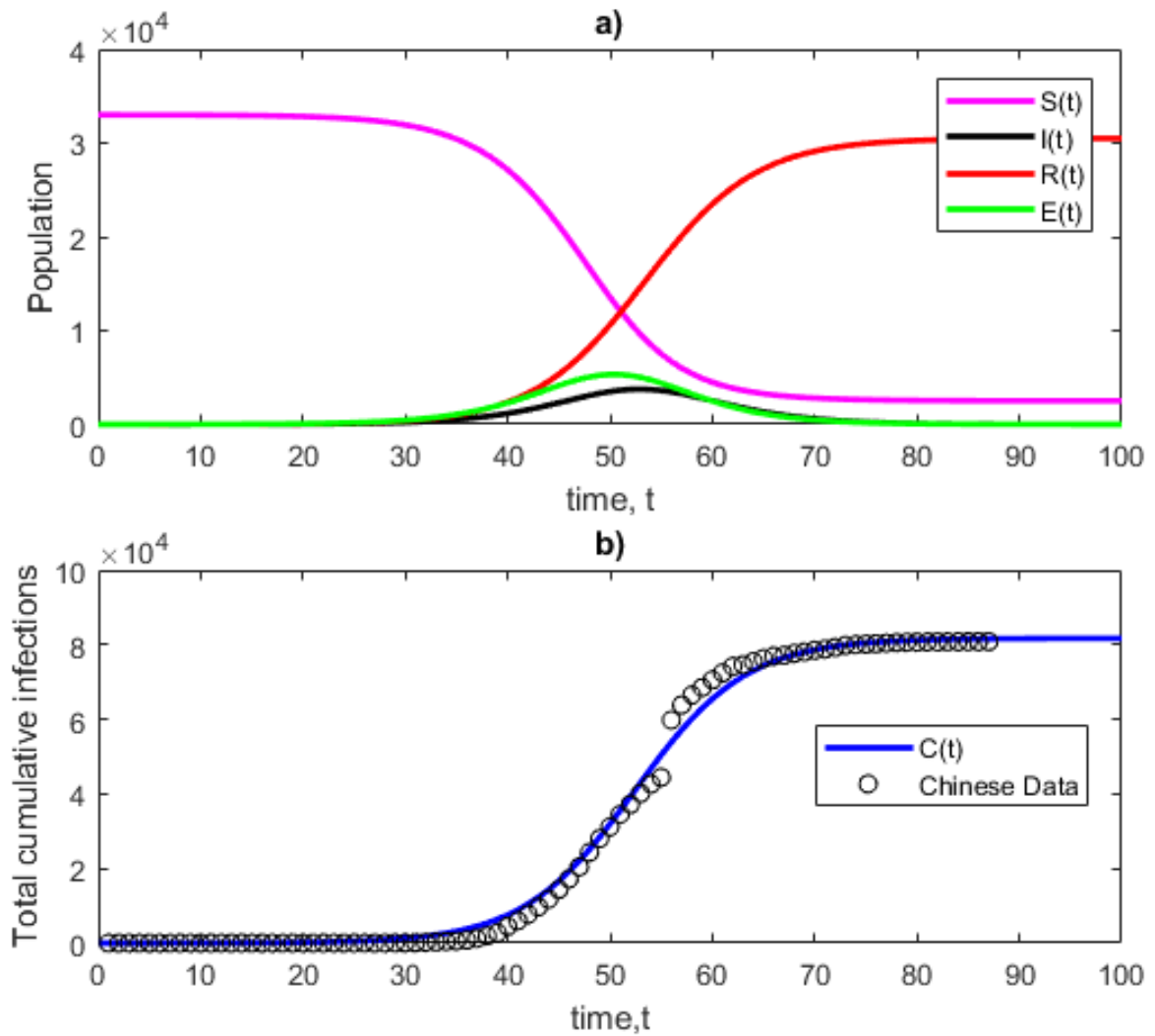


FIG. 1. Example of a Susceptible-Exposed-Infected-Recovered (SEIR) model of COVID-19 (Eqs 7-10) with $\lambda = 1./S(0)$, $\alpha = 0.27$, $\gamma = 0.37$. Initial conditions are set to $I(0) = 2$, $S(0) = 33000$, $E(0) = R(0) = 0$. a) Time evolution for the variables of the system, b) Time evolution for the total number of infections $C(t)$ against the Chinese data with $t=1$ corresponding to Dec 19. 2019.

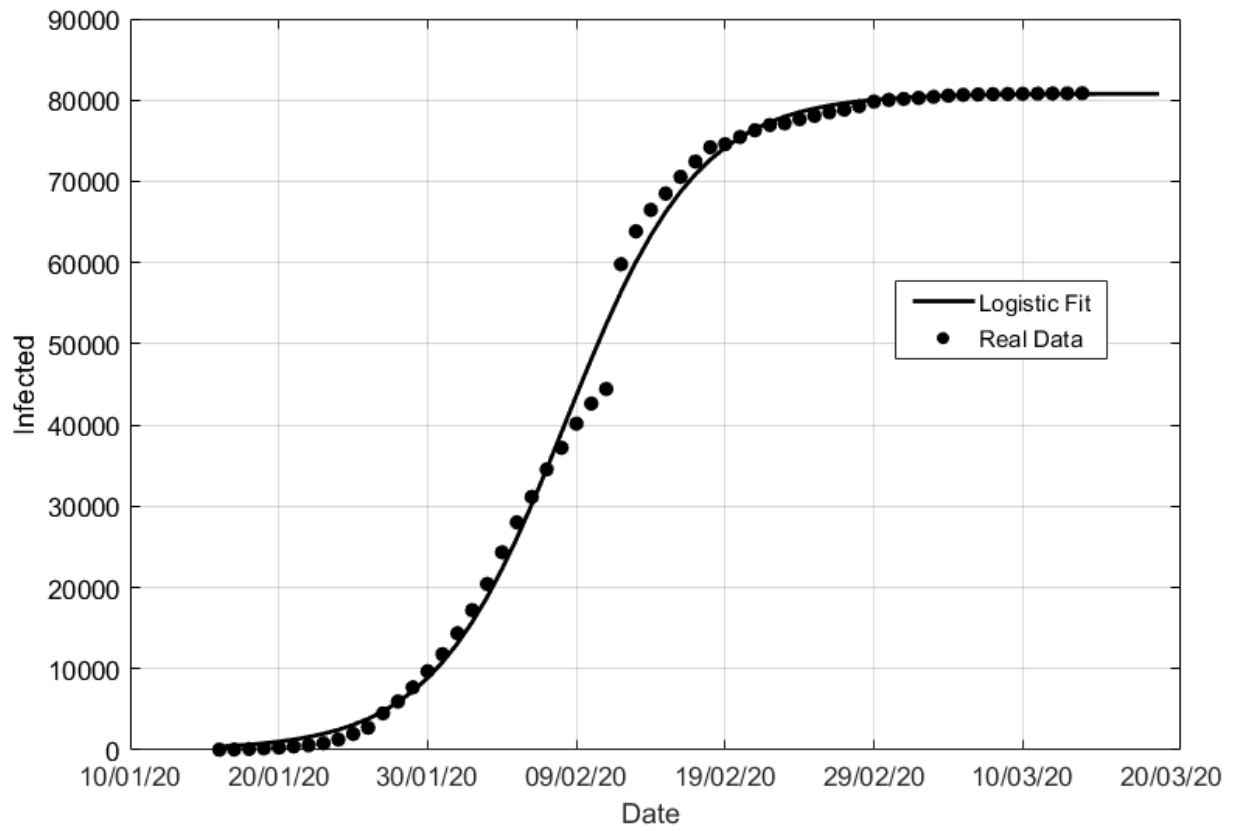


FIG. 2. Logistic (Eq. 11) fit of the Chinese number of infections $C(t)$. The best fit parameters are $a = 80800 \pm 400$, $b = 0.225 \pm 0.005$, $c = 190 \pm 25$.

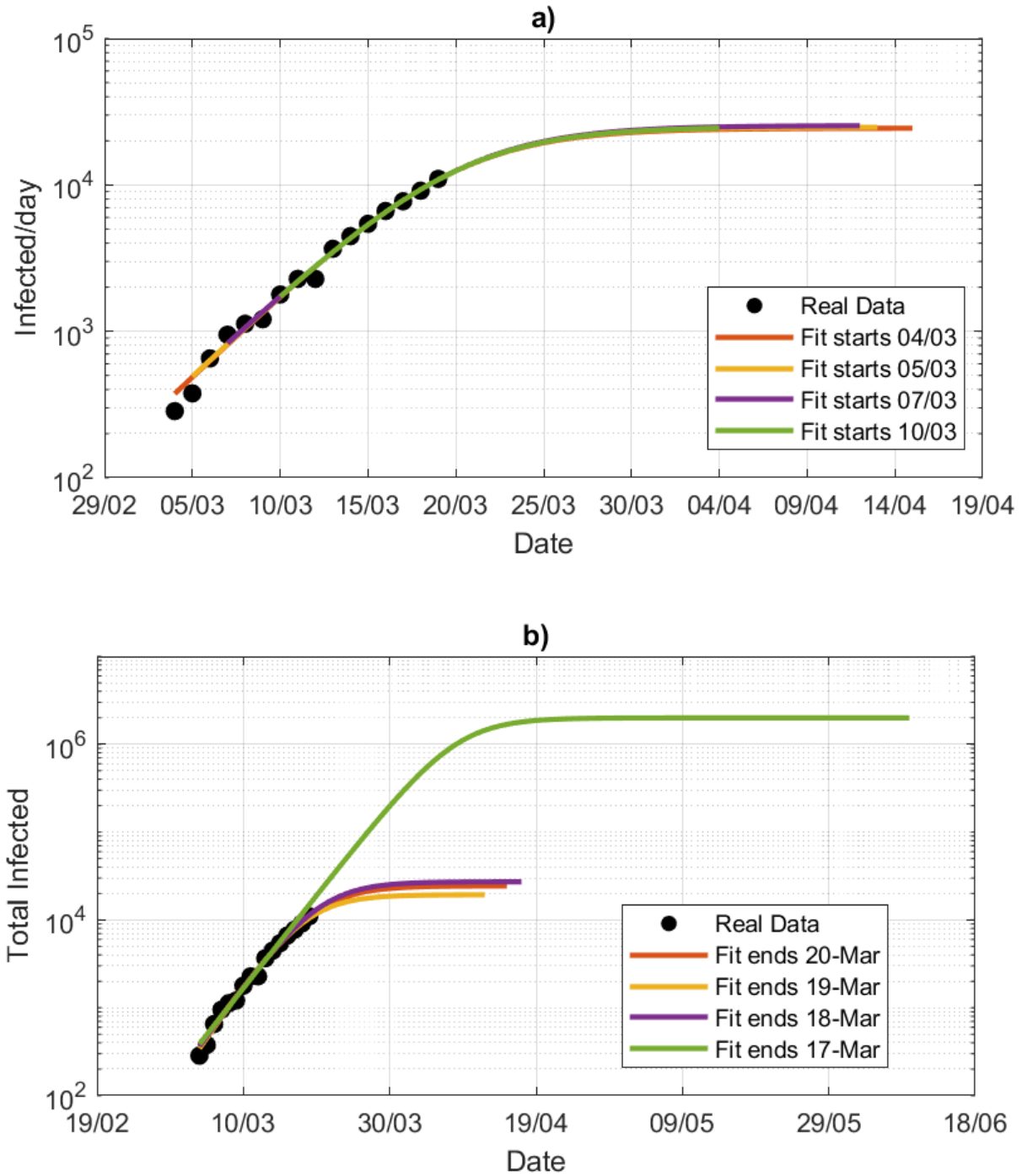


FIG. 3. Logistic distribution fits for the early stages of the epidemic in France. a) Logistic fits with data starting from different dates and ending Mar. 20. b) Logistic fits ending on different dates, but starting Mar. 04.

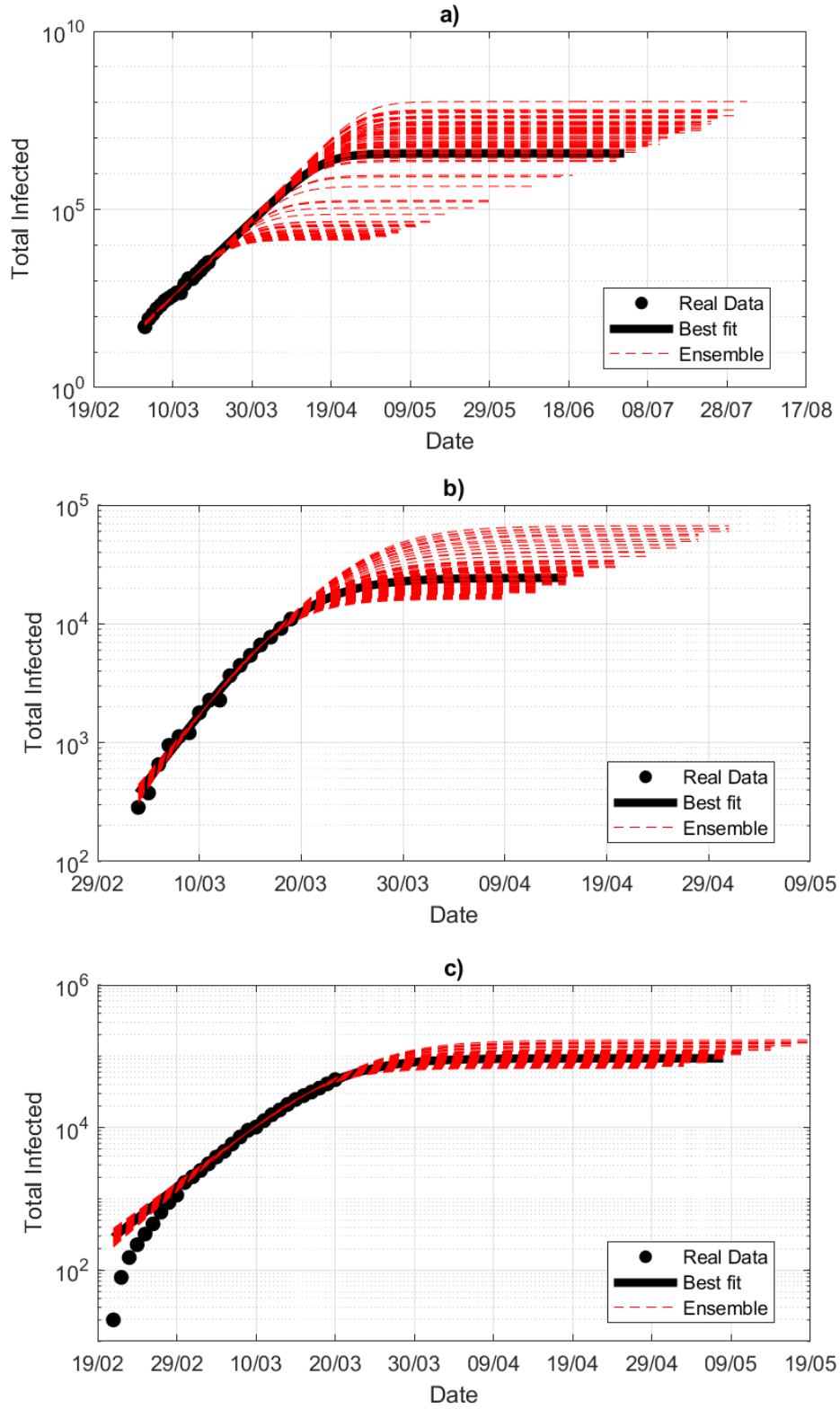


FIG. 4. Logistic distribution substituting the last data point with a random number $\xi(t^*)$ drawn from a uniform distribution with mean $C(t^*)$ and standard deviation $0.2C(t^*)$ for UK (a), France (b) and Italy (c).

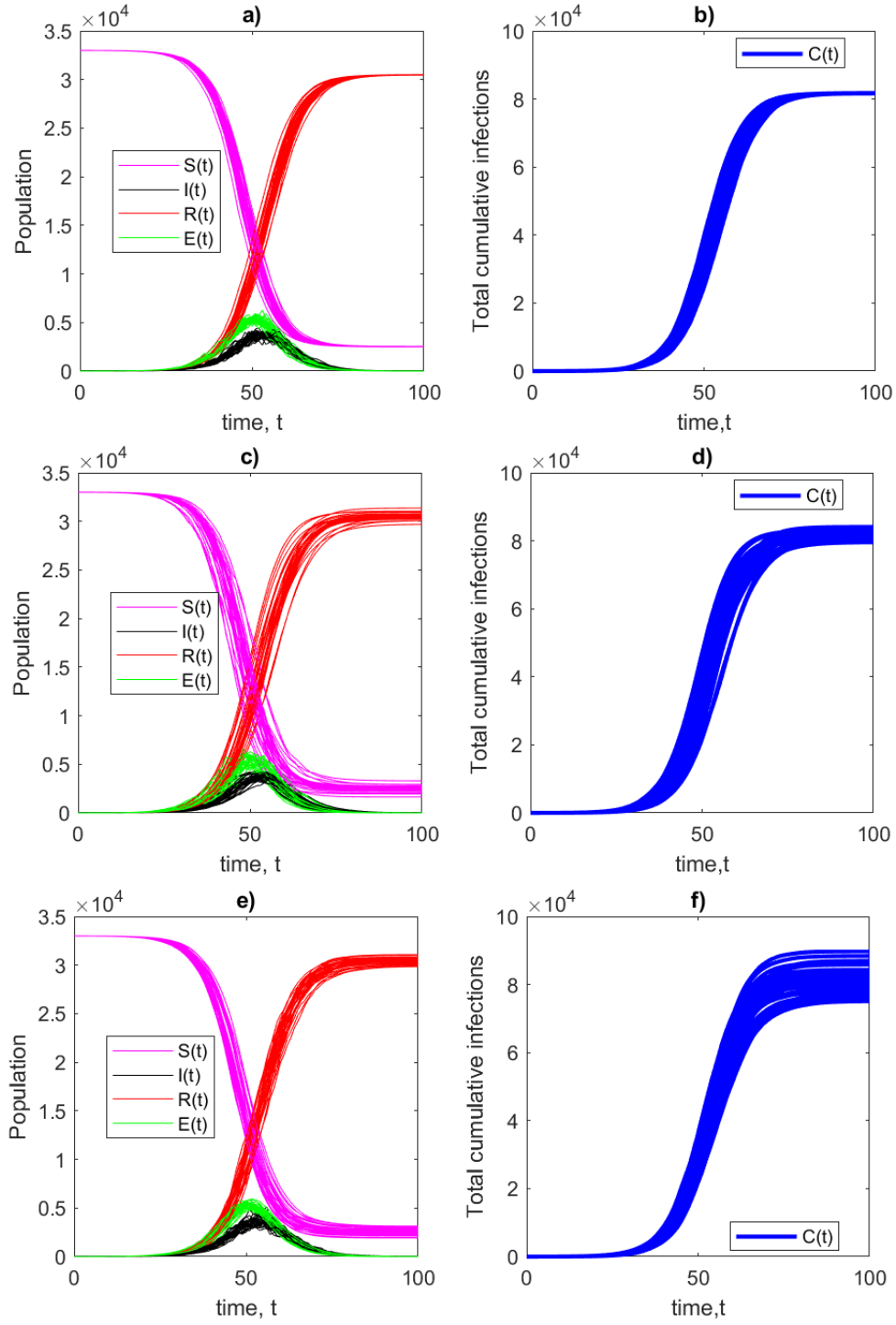


FIG. 5. Example of 30 trajectories of dynamics of stochastic Susceptible-Exposed-Infected-Recovered (SEIR) model for COVID-19, obtained replacing alternately α (a,b), λ (b,d) and γ with the stochastic process with Eq 12. Dynamics are integrated with a fixed initial condition and 30 noise realisations. a,c,e) Time evolution for the variables of the system, b,d,f) Time evolution for the total number of infections $C(t)$.

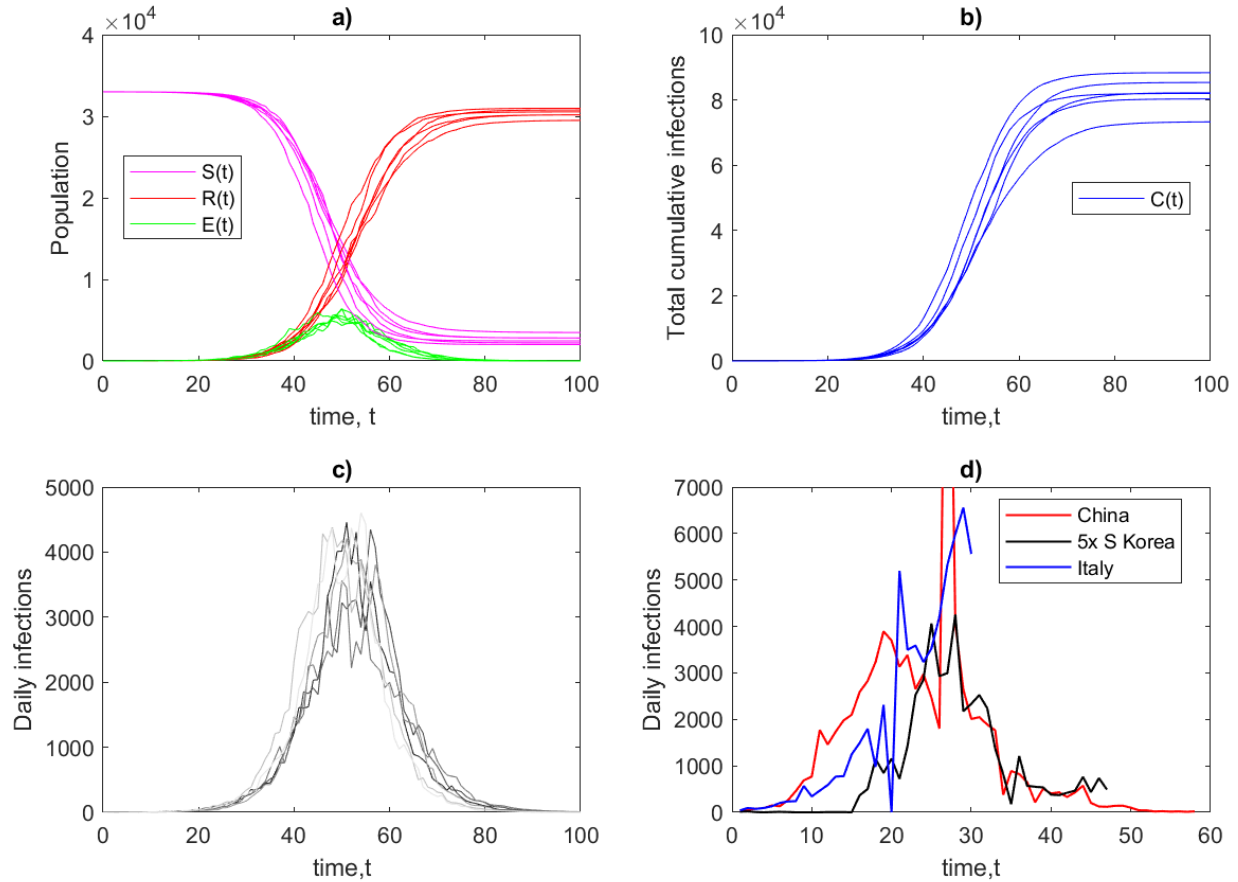


FIG. 6. Example of 6 trajectories of dynamics of stochastic Susceptible-Exposed-Infected-Recovered (SEIR) model for COVID-19, obtained replacing all parameters α , λ and γ with an independent stochastic process as in Eq 12. Dynamics are integrated with a fixed initial condition and 6 noise realisations. a) Time evolution for the variables of the system. b) Time evolution for the total number of infections $C(t)$. c) Time evolution for the daily infections. d) Comparison with daily infections in China (red, starting Dec 19. 2019), South Korea (black, starting Jan 30, 2020), Italy (blue, starting Feb 20, 2020).