Early forecasts of the evolution of the COVID-19 outbreaks and quantitative assessment of the effectiveness of countering measures.

Emanuele Daddi[1*] and Mauro Giavalisco[2†]
1) *CEA, IRFU, DAp, AIM, Université Paris-Saclay, Université Paris Diderot, Sorbonne Paris Cité, CNRS, F-91191 Gif-sur-Yvette, France and*
2) *Department of Astronomy, University of Massachusetts Amherst, 710 N. Pleasant St., Amherst, MA 01003, USA*
(Dated: April 17, 2020)

We discovered that the time evolution of the inverse fractional daily growth of new infections, $N/\Delta N$, in the current outbreak of COVID-19 is accurately described by a universal function, namely the two-parameter Gumbel cumulative function, in all countries that we have investigated. While the two Gumbel parameters, as determined bit fits to the data, vary from country to country (and even within different regions of the same country), reflecting the diversity and efficacy of the adopted containment measures, the functional form of the evolution of $\Delta N/N$ appears to be universal. The result of the fit in a given region or country appears to be stable against variations of the selected time interval. This makes it possible to robustly estimate the two parameters from the data data even over relatively small time periods. In turn, this allows one to predict with large advance and well-controlled confidence levels, the time of the peak in the daily new infections, its magnitude and duration (hence the total infections), as well as the time when the daily new infections decrease to a pre-set value (e.g. less than about 2 new infections per day per million people), which can be very useful for planning the reopening of economic and social activities. We use this formalism to predict and compare these key features of the evolution of the COVID-19 disease in a number of countries and provide a quantitative assessment of the degree of success in in their efforts to countain the outbreak.

I. INTRODUCTION

In absence of containment measures, the early phase of an pandemic outbreak, such as the current COVID-19 (ref. [1–9]) is characterised by an exponential growth with time of the number of infected people, $N(t)$ ([28-33]. In this phase, the rate of new infections as a function of time, $\dot{N}(t) = dN/dt$, such as the number of daily new infected $\Delta N(t)$, is exponential as well. If left unimpeded, the epidemic naturally evolves and slows down, with $N(t)$ departing from the initial exponential growth and ΔN reaching a peak value before starting to decrease. This happens when a sufficiently large fraction of the initial population has become infected that the number of new infections become progressively more inefficient and is matched and eventually surpassed by the rate of deaths and recovered individuals. Such an outcome, however, is not acceptable in the case the current COVID-19 outbreak, since it would result in a very large loss of life. Containment measures are necessary to limit the deaths and avoid saturating health care capabilities, their merit and efficiencies have been discussed in a number of recent studies as soon as the COVID-19 epidemics started spreading (cfr. Refs. [1–9]). Given the relatively long incubation period of the disease, from 2 up to 28 days, and the fact that a large fraction of the infected population appears to be asymptomatic or hypo-symptomatic, however, how can a government monitor the effectiveness of the adopted containment measures and adjust them so that the number of new infected reaches a peak as quickly as possible? How can one predict when ΔN will peak and when its decrease will begin and reach a low, pre-set value of production of new infected to be manageable? A substantial effort is already ongoing to attempt modelling the number of various observables, including newly infected individuals, deaths and recoveries, and using a variety of methods from machine learning to the use of standard SIR approaches (cfr. Refs. [10–18]). The goal of this paper is to introduce and describe a simple *phenomenological* model that can be effectively used to predict the time of arrival of the peak, and its width, solely based on the observed time evolution of the number of known newly infected individuals $\Delta N(t)$. The model is interesting in practice because it is based on the data themselves, without the need of any further assumption. It is also conceptually attractive because, as we shall see, the data show that, in presence of containment measures, the functional dependence with time of the key observable $\Delta N/N$, the fractional number of daily new infected, appears to be universal and characterized by only two parameters, μ and a, as we will describe later. In other words, the presence of containment measures changes the growth of the epidemic from exponential into another functional form, which is universal, independently of the types and magnitude of the measures adopted by various countries to oppose the spread of COVID-19. These measures determine the parameters of the functional form of $\Delta N/N$ but not its general expression. In fact, even the exponential function can be considered as a special case of our general function, in the limit in which one of its parameters (the evolutionary timescale) goes to infinite.

* edaddi@cea.fr
† mauro@umass.edu

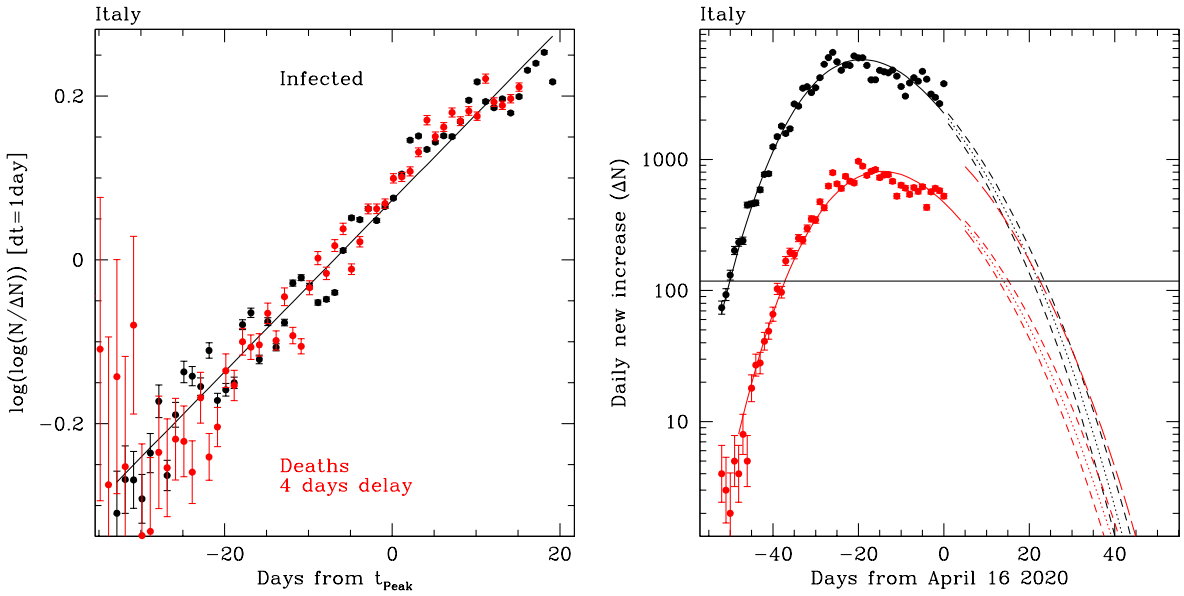

FIG. 1. **Left:** The double log of the ratio of total infected to newly infected is plotted as a function of time (black points). The line is the best fitted Gumbel function, as defined in Section II. In the case of a pure exponential ($N \propto e^{t/\tau}$), this ratio is equivalent to the characteristic evolutionary timescale τ (if the growth from one day to the next were to be considered exponential). The increase in the ratio, as shown in this plot, is thus a direct visualisation of the deceleration in the spread of the disease. Red points shows the same quantities for deaths, delayed by 4 days: it can be seen that deaths follow the same evolutionary rate with typical delays that were of 6-7 days at early times and get to 3-4 days more recently, due to the increasing number of recovered patients (deaths are a more accurate delayed function of $N_{\rm active}$, but exploring this is beyond the scope of this work). **Right:** We show daily values of ΔN for infected (black) and deaths (red, delayed as in the left panel). The solid curves correspond to the Gumbel best fitting function from the left panel (delayed for deaths and scaled down by the integrated CFR); dotted line shows the extrapolation of the same function; dashed lines show the 1σ range around the forecasted trend. The red long-dashed line shows the extrapolation of the sum of deaths and recovered: when this long-dashed line rises above the dotted black line, $\Delta N_{\rm active}$ becomes negative and $N_{\rm active}$ thus starts decreasing, hence corresponding to the peak of $N_{\rm active}$. Throughout this work we refer instead to the peak of ΔN, as shown in this figure. The horizontal line shows the level of 2/day/1M new infected, that we set as the floor to a regime where the outbreak can be controlled as seen in South Korea.

This work is organized as follow: in Section II we introduce the mathematical framework and fitting methods. In Section III we discuss the meaning of the Gumbel parameters and some of its properties. We further verify the forecasting capability of the method in Section IV. In Section V we discuss the quantitative performances in countering the outbreak by various countries. We discuss our results in Section VI and conclude in VII.

II. MATHEMATICAL FRAMEWORK

Let's call $N(t)$ the total number of known infected people at a given time t. We define:

$$dN = rNdt \tag{1}$$

where r, which we call "the reproductive rate per unit time", at any given time, is related to the R_0 parameter (generally introduced in so-called SIR models) by the relation $r = (R_0 - 1)\,\gamma$ (where γ in SIR models is the average duration of the infectious period). In an unchecked epidemic outbreak, r and R_0 are constant and the outbreak grows exponentially until a significant fraction of the population has been infected. If containment measures are present, however, r is not a constant even in the early phases but is a a function of time as a result of the measures put in place to slow and halt the outbreak.

Empirical analysis of the best studied case of the time evolution of r in the COVID-19 outbreak, for example in Italy, Spain, South Korea and the US, shows that a rather accurate description of the observed data is provided by the Gumbel (1935; [19]) cumulative distribution function:

$$dN/dt/N = r = e^{-e^{(t-\mu)/a}} \tag{2}$$

as Figure 1 and 2 illustrate (note that this actually is a reflected Gumbel, given that a is positive). This function, which is specified by only 2 parameters that are determined by fits to the data, provides an excellent representation of the observations in all cases that we have considered, namely countries where the type, implementation and timing of the preventive measures are rather

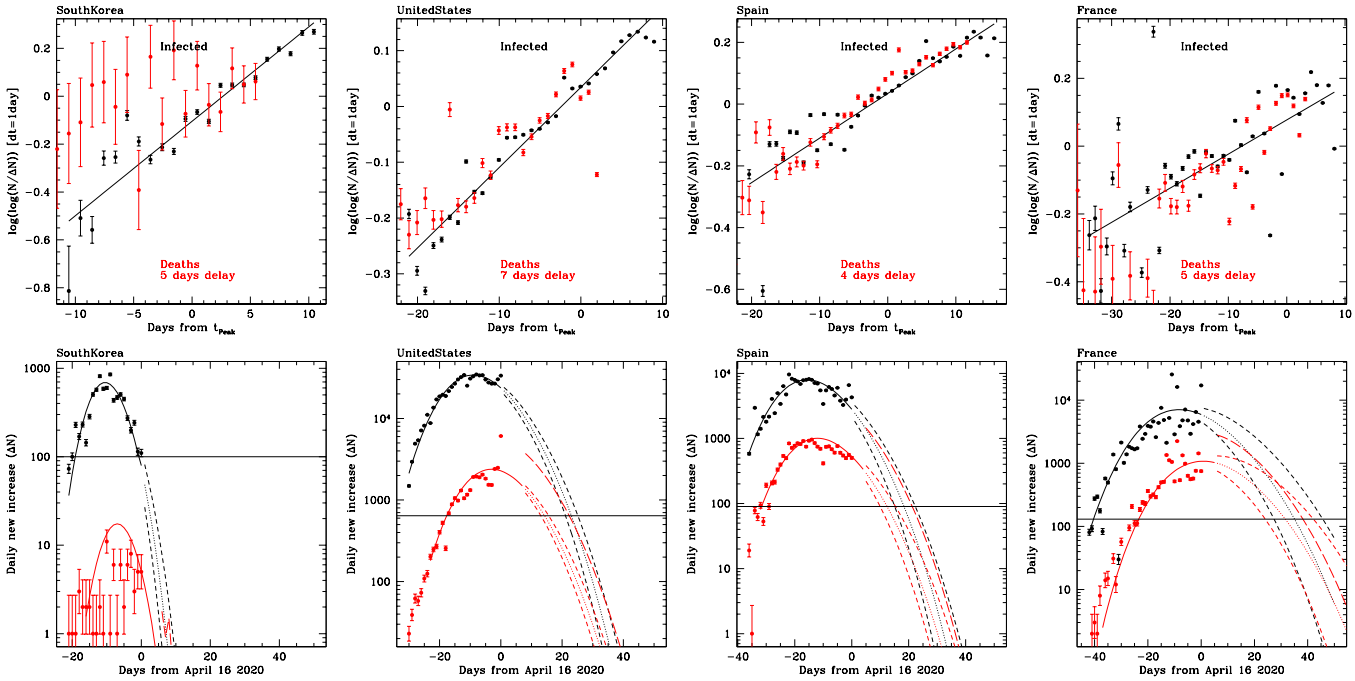

FIG. 2. The same of Fig. 1 but for South Korea, the US, Spain and France, as labeled.

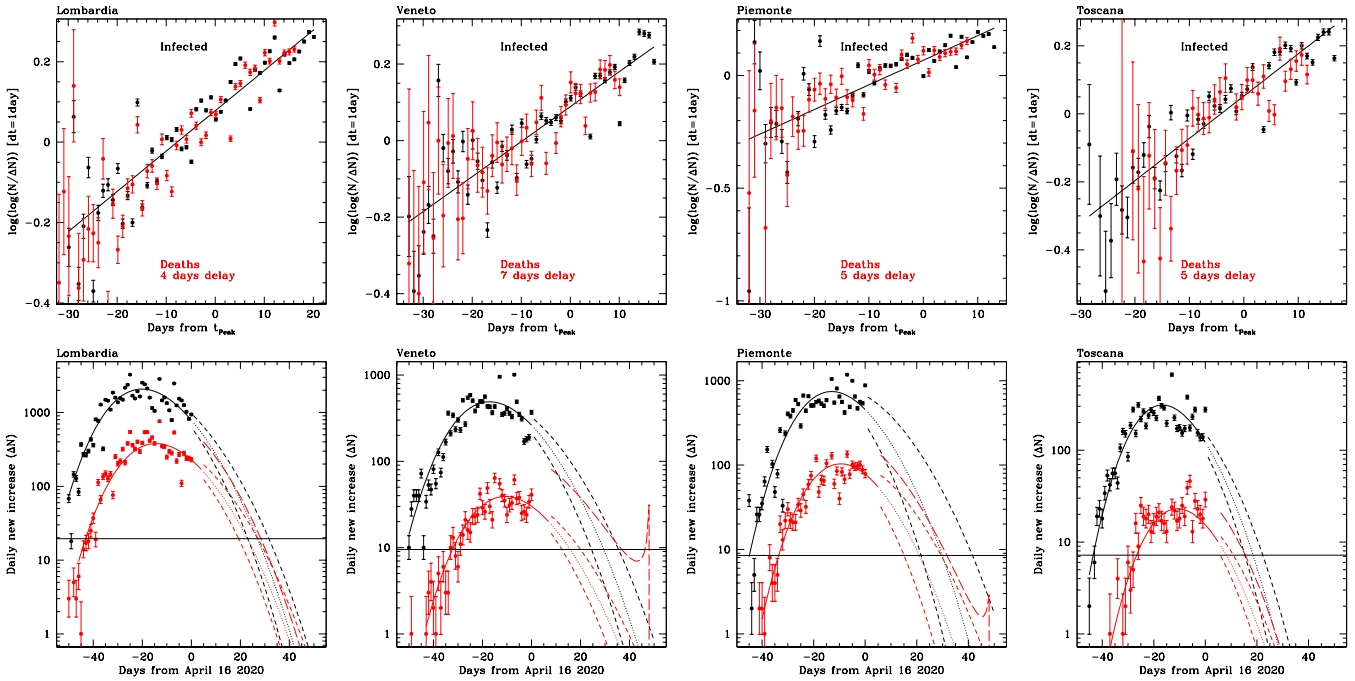

FIG. 3. The same of Fig. 1 but for various regions of Italy, as labeled.

different. As the case of Italy shows, the Gumbel function provides an excellent representation of the data for the country as a whole, and also for the data of individual regions of Italy as well (cfr. Figure 1 and Figure 3). Since the fractional rate of new infected individuals is modeled as a double exponential, the implication is that r has its own "reproductive rate per unit time" which is itself a function of time and decays exponentially, following the equation:

$$dr = -re^{(t-\mu)/a}dt \tag{3}$$

Fig. 1-left shows the Gumbel fit for Italy over the last 45 days (here and in the following we are using $dt = 1$ day). We have collected data from *Worldometer* [21] and official Github repositories [22; 23], and included them in our analysis as they reached a daily production of new infected close to the 1–2/day/M in each country.

To determine the parameters of the Gumbel function we perform a least-square fit of a linear function to the logarithm of the logarithm of $N/\Delta N$ with no weights, thus equally counting each data point, regardless of the Poisson errors, since it can be readily seen from the figure that the scatter of the points around the best-fit function is fairly constant in double-log space as a function of time (except perhaps very early phases that appear to be noisier). This suggests that the uncertainty of each measurement is likely dominated by fluctuations in the reporting and sampling systems rather then by counting (Poisson) errors, and therefore it must be related to the effectiveness and of the testing and reporting capabilities of each countries, as discussed later. We tried different approaches such as fitting the cumulative Gumbel function to the N time series, and the differential Gumbel function to the ΔN time series, but none provided as good a fit as with the one we found with our approach. This conclusion, we believe, is already apparent in the $N/\Delta N$ time series plots in Figures 1–3, where there is no hint that the functional form of the double log of these series ever significantly departs from a linear relation. We notice that a recent study (Bianconi et al 2020; [20]) discuss modeling of the doubling time of newly infected patients as a function of time, which is equivalent to the quantity r in our study (it's inverse). They argue that two asymptotic behaviours are present at early and late phase of the outbreak with different slopes in log space. When using a Gumbel function the need for ad-hoc change of slopes is removed, and the data evolution remain linear in double-log space.

III. THE MEANING AND DETERMINATION OF GUMBEL PARAMETERS

Once the best-fit parameters μ and a have been determined from the data by a fitting procedure, the Gumbel function provides an excellent description of the past time evolution of the COVID-19 disease and allows us to make predictions of the future evolution of the outbreak, including the peak in the number of new infections, as it is illustrated for the predicted evolution of ΔN with time in the case of Italy (Fig. 1-right). The parameter a, which has the physical dimension of time (we have chosen here the day as unit), sets the overall shape of the function, including the width of the peak where larger a value of a implies a longer turnaround time for the outbreak. In particular, we find that the FWHM (Full Width at Half Maximum) of the peak can be approximated as

$$\mathrm{FWHM}_{\mathrm{Peak}} \sim 0.63 \times a + 4.3 \tag{4}$$

The parameter μ sets the timescale of the outbreak and thus allows one to predict the time (day) of the peak in ΔN (t_{peak}, as we shall see below). Finally, there also is an overall normalization coefficient that can be set to reproduce the total number of infected people at the time of the peak, $N(t_{\mathrm{peak}})$. Notice that the time of the peak can be derived from the equation $e^x + x = ln(a)$, where $x = (t_{\mathrm{peak}} - \mu)/a$, which has the solution

$$t_{\mathrm{peak}} = \mu + a(ln(a) - W(a)) \tag{5}$$

where W is the Lambert function. In Table I we report t_{peak}, instead of μ, since they are effectively equivalent parameters but the former has more immediate utility than the latter. Also notice that Eq. 2 cannot be analytically inverted to solve for dN/dt and N, and thus we have proceeded to numerically inverting it, integrating over the natural reporting timescale $\Delta t = 1$ day.

It's interesting to emphasize that in the limit of a purely exponential growth, $N/\Delta N = \tau$ is constant, corresponding to a Gumbel function with a that tends to ∞ and with $\exp(\exp(-a\mu)) = \tau$. The exponential growth is thus a special case in the Gumbel growth we are describing, with infinite timescale a.

IV. VERIFYING THE ACCURACY OF THE MODEL'S PREDICTIONS

If the future evolution continues along the same trend of the best-fit function, i.e. with the same functional form and identical parameters, as the data suggest, the Gumbel function can be used to predict the near future evolution of the outbreak, including providing an estimate of the uncertainty. Such a prediction is shown for the Italian data in Fig. 1-right. Similar predictions for South Korea, Spain, France and the US are also shown in Fig. 2. At the time of this writing, Italy, Spain, France and the US are the countries with the largest number of reported cases. As the plots show (data updated as of April 13), according to our analysis, all of them have already (cumulatively, not necessarily on a region-by-region case) passed the peak of the daily new infected (ΔN). South Korea has actually reached a steady state of daily new infection with $\Delta N \approx 50$–100 (not included in the plots shown; see *worldometer*), thus demonstrating that such a low level can be kept constant over several weeks at least, preventing further exponential diffusion of the infection.

The excellent fit provided by the Gumbel function to the time dependence of $\Delta N/N$, with constant parameters a and μ over a large time interval (Eqs. 2), as shown in left panels of Figures 1–3, suggests that, with adequate sampling, the same best fitting parameters could have been determined significantly earlier on during the diffusion of the COVID-19 infection in each country. If that could indeed be made possible one would have been able to predict key properties of the evolution of the outbreak, such as the timing of the peak in ΔN and its

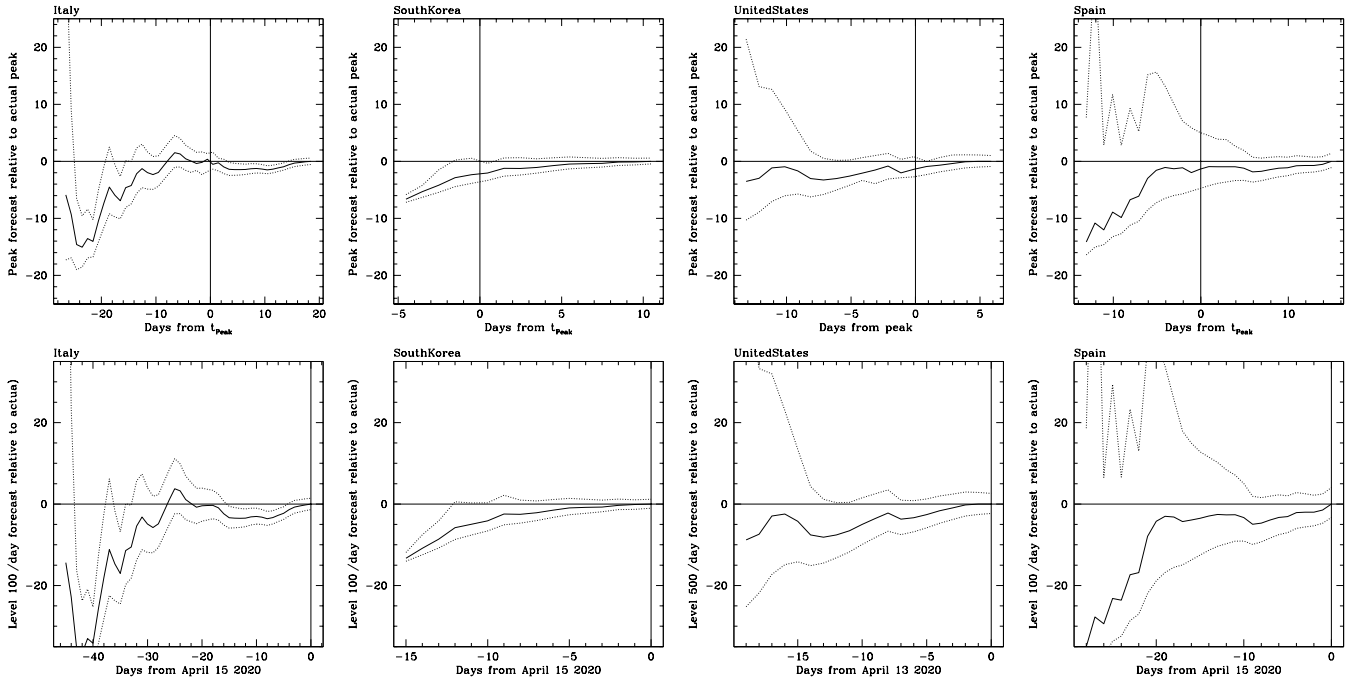

FIG. 4. **Top panels:** Forecasting the peak in ΔN as a function of time for different countries, as labeled. Solid line shows best prediction and dotted line its 1σ range, at each day. For Italy the peak forecast was accurate within 5 days, 20 days in advance of the peak, and about 10 days for the US and Spain. The peak was predicted only a few days in advance for South Korea, that reached the peak very rapidly. **Bottom panels:** Forecasting when various countries (as labeled) will reach a production rate of new infected per day (ΔN) at the level of 2 per milion abitant, post peak. Lines are as in the top panels. For Italy accurate prediction of that regime is obtained with 30 days of advance, about 10-15 days for the US and Spain and 13 days for South Korea. We emphasize that as of the time of writing only South Korea has actually reached this level, and was able to maintain it for at least one month.

width, the point in time when ΔN reaches a pre-set value determined to allow a country to maintain a manageable steady state control of the disease, for example similar to what achieved by South Korea.

To further investigate the feasibility of such predictions in the countries considered here, we have performed the following simulation: we have fit the time series of $\Delta N/N$ only up to a given day in the past, ignoring the additional data points after that, and compared the prediction obtained in this way with the one obtained using the full data set available at this time, including a detailed analysis of the relative uncertainties. Fig. 4 top-panels shows the results of this simulation to illustrate the ability of our methodology to predict the peak in ΔN. The graphs clearly show that the peak could have been predicted as much ahead in time as about 3 weeks for Italy and 1-2 weeks for the other countries. For most values of the time lapse that we have considered, a reasonable 1σ uncertainty for the peak is also predicted, which in fact brackets the true date of the peak. For Italy and South Korea a small underestimate of the peak time is observed using the earliest data points. We also observe a similar behaviour in other countries that we have inspected, a fact that we interprete as due to the progressive settling of the reporting mechanisms and procedures for testing adopted by the various governments rather than a shortcoming of our model.

The bottom panels of Fig.4 show the prediction of the time, after the peak, when the epidemic outbreak can be considered to be under control: in South Korea that time was when it reached $\Delta N \sim 100$ per day, a level that has been so far maintained for about 1 month without the occurrence of a second outbreak wave. This corresponds to approximately 2 new infected per day per million people. Using our model, such crucial moment in time can be predicted with large anticipation, up to about 35 days in the case of the Italian data. Of course, this assumes that the evolution continues with the same Gumbel parameters down to this *floor*: so far, only in the case of South Korea such a level has been convincingly reached.

V. QUANTITATIVE ASSESSMENT OF COUNTRIES PERFORMANCES IN COUNTERING COVID-19

Our quantitative analysis of the evolution of the COVID-19 in various countries allows us to measure their efficiency in countering the virus spread from various parameters. This is based for simplicity in the countries with the largest number of cases, as a reference. We report key quantities for a wider list of countries and re-

TABLE I. We report various estimates from our Gumbel modeling, using data collected until April 16th 2020.

Country/ Region	Pop. 10^6	N final Total	N final per 10^6	ΔN Peak Total	ΔN Peak per 10^6	Peak	Floor[a]	Duration days	a days	rms[b]	PCC[c]
S. Korea	51.4	8093^{+110}_{-58}	157^{+2}_{-1}	690^{+37}_{-39}	$13^{+0.7}_{-0.7}$	March 1^{+1}_{-1}	March 11^{+2}_{-1}	19^{+2}_{-1}	11.0 ± 1.0	0.23	0.93
Austria	9.2	14975^{+244}_{-167}	1618^{+26}_{-18}	625^{+20}_{-22}	67^{+2}_{-2}	March 28^{+1}_{-1}	April 25^{+3}_{-1}	51^{+3}_{-3}	28.2 ± 1.7	0.15	0.94
Switzerland	8.5	28099^{+615}_{-391}	3273^{+71}_{-45}	1142^{+47}_{-52}	133^{+5}_{-6}	March 28^{+2}_{-1}	April 28^{+4}_{-1}	54^{+4}_{-3}	29.2 ± 2.3	0.20	0.90
Spain	46.2	201049^{+8439}_{-5555}	4346^{+182}_{-120}	7986^{+244}_{-307}	172^{+5}_{-6}	March 31^{+2}_{-2}	May 4^{+4}_{-2}	55^{+4}_{-4}	30.1 ± 2.7	0.18	0.89
Germany	82.9	155242^{+8392}_{-5668}	1871^{+101}_{-68}	5430^{+70}_{-127}	65^{+1}_{-1}	March 31^{+2}_{-2}	May 4^{+5}_{-2}	61^{+5}_{-4}	35.4 ± 2.9	0.19	0.89
Chile	18.7	11839^{+1590}_{-996}	632^{+85}_{-53}	416^{+13}_{-1}	$22^{+1}_{-0.1}$	April 7^{+3}_{-2}	May 5^{+6}_{-3}	49^{+6}_{-5}	35.2 ± 4.0	0.15	0.86
Italy	60.6	189404^{+2384}_{-2094}	3122^{+39}_{-34}	5807^{+44}_{-49}	$95^{+0.1}_{-0.1}$	March 29^{+1}_{-1}	May 8^{+2}_{-1}	73^{+2}_{-2}	41.9 ± 1.1	0.07	0.98
US	330.5	858599^{+36521}_{-29961}	2597^{+110}_{-90}	34069^{+179}_{-27}	$103^{+0.1}_{-0.1}$	April 7^{+1}_{-1}	May 9^{+3}_{-1}	54^{+3}_{-2}	30.1 ± 1.4	0.07	0.97
Netherlands	17.2	35958^{+1994}_{-1517}	2082^{+115}_{-87}	1186^{+6}_{-3}	$68^{+0.2}_{-0.1}$	April 3^{+2}_{-2}	May 10^{+4}_{-1}	65^{+4}_{-3}	38.2 ± 2.3	0.12	0.94
Turkey	81.9	118861^{+19162}_{-12243}	1449^{+233}_{-149}	4526^{+282}_{-130}	55^{+3}_{-1}	April 10^{+3}_{-2}	May 11^{+6}_{-3}	51^{+6}_{-5}	31.8 ± 3.4	0.12	0.89
Iran	81.7	99237^{+6011}_{-4511}	1213^{+73}_{-55}	2327^{+6}_{-25}	$28^{+0.1}_{-0.1}$	March 31^{+2}_{-2}	May 14^{+5}_{-3}	93^{+7}_{-7}	58.9 ± 4.4	0.11	0.90
UK	67.1	158479^{+17181}_{-12270}	2360^{+255}_{-182}	5280^{+209}_{-112}	78^{+3}_{-1}	April 9^{+2}_{-2}	May 16^{+5}_{-2}	63^{+5}_{-4}	37.7 ± 3.0	0.11	0.92
France	67.1	236308^{+54763}_{-29334}	3520^{+815}_{-436}	7089^{+704}_{-229}	105^{+10}_{-3}	April 7^{+5}_{-4}	May 21^{+11}_{-6}	76^{+11}_{-8}	43.1 ± 5.8	0.24	0.76
Russia	144.4	$1.7M^{+22M}_{-1.3M}$	$12k^{+154k}_{-9k}$	$28k^{+253k}_{-19k}$	200^{+1756}_{-131}	May 29^{+38}_{-18}	August 27^{+106}_{-45}	151^{+106}_{-47}	94 ± 30	0.07	0.61
Sicilia	5.1	2895^{+198}_{-121}	563^{+38}_{-23}	97^{+2}_{-3}	18^{+1}_{-1}	March 31^{+3}_{-2}	April 27^{+5}_{-2}	51^{+5}_{-4}	37.2 ± 4.0	0.21	0.83
Marche	1.4	5779^{+92}_{-64}	3877^{+61}_{-42}	224^{+7}_{-8}	150^{+4}_{-5}	March 27^{+1}_{-1}	April 29^{+3}_{-1}	59^{+3}_{-3}	31.0 ± 1.9	0.17	0.93
Lazio	5.7	6088^{+310}_{-215}	1057^{+53}_{-37}	210^{+2}_{-4}	36^{+1}_{-1}	March 31^{+2}_{-2}	May 2^{+4}_{-2}	57^{+4}_{-3}	36.0 ± 2.8	0.17	0.90
Toscana	3.7	8768^{+376}_{-259}	2369^{+101}_{-70}	310^{+5}_{-8}	83^{+1}_{-2}	March 31^{+2}_{-2}	May 4^{+4}_{-2}	62^{+4}_{-4}	34.9 ± 2.6	0.18	0.90
Emilia Rom.	4.5	23239^{+651}_{-478}	5080^{+142}_{-104}	772^{+14}_{-18}	168^{+3}_{-4}	March 28^{+2}_{-2}	May 7^{+4}_{-1}	72^{+4}_{-3}	37.8 ± 2.2	0.16	0.93
Campania	5.8	4942^{+974}_{-503}	850^{+167}_{-86}	140^{+7}_{-1}	$24^{+1}_{-0.2}$	April 2^{+5}_{-4}	May 8^{+11}_{-6}	64^{+10}_{-7}	46.2 ± 6.9	0.25	0.73
Lombardia	10	69537^{+2744}_{-1927}	6937^{+273}_{-192}	2077^{+42}_{-57}	207^{+4}_{-5}	March 27^{+2}_{-2}	May 13^{+5}_{-2}	78^{+5}_{-4}	43.3 ± 3.2	0.17	0.89
Piemonte	4.3	23262^{+4605}_{-2129}	5324^{+1054}_{-487}	744^{+32}_{-15}	170^{+7}_{-3}	April 2^{+5}_{-4}	May 15^{+12}_{-7}	75^{+12}_{-9}	39.7 ± 6.4	0.35	0.69
Veneto	4.8	17624^{+1255}_{-840}	3609^{+257}_{-172}	489^{+1}_{-8}	$100^{+0.1}_{-1}$	March 30^{+3}_{-2}	May 15^{+7}_{-3}	80^{+7}_{-5}	47.5 ± 4.1	0.19	0.86

[a] This is defined as reaching the level of 2/day/M new infections
[b] This refer to the fit of $\log(\log(N/\Delta N))$ as in Figs 1–3
[c] Pearson correlation coefficient between $\log(\log(N/\Delta N)$ and time

TABLE II. Estimates of the outbreak diffusion into the population

Country	Fraction of population infected (percent) naive	CFR-corrected	twice corrected[a]
South Korea	0.016	0.016	0.06
Austria	0.17	0.25	0.37
Germany	0.18	0.38	0.79
US	0.24	0.71	2.1
Spain	0.39	2.3	14.0
Italy	0.29	2.2	16.2
France	0.31	2.5	19.8

[a] See Section V D

gions in Table I, so that the reader might extend our considerations elsewhere, if interested.

A. Rapidity in halting and suppressing the infection spread

The parameter a, the slope of the double log of $\Delta N/N$, is the most important parameter because it sets the duration of the outbreak: the *Duration* of the ΔN curve defined as the number of days during which more than 2/day/M new infected are produced. We find that this *Duration* can be estimated as $1.75 \times a$ (Table I).

Figure 5 compares the performance of various countries in containing the disease: in Spain, the US and Austria the virus countering measures seem to be more effective than Italy and Germany, since the former countries have shorter outbreaks and more the reproduction rate of the virus declines faster. By far, South Korea has been overperforming all other countries in terms for its rapidity in suppressing the outbreak.

B. Timeliness of the response

The total number of people infected in a country ($N_{\rm Total}$), depends on the rapidity in containing the spread of the virus (larger a brings more total infections) but also on how early measures are taken, i.e. the rapidity in changing the behaviour of $N/\Delta N$ from almost constant (exponential growth) into a Gumbel-like deceleration. The following formulation is a very good description of our data: $N_{\rm Total} = \Delta N_{\rm Peak} \times (0.66a + 4.8)$. However, $\Delta N_{\rm Peak}$ grows linearly with a at fixed normalization, implying that $N_{\rm Total}$ goes like a^2. However, also the normalization increases $N_{\rm Total}$ linearly. Fig. 5 shows that at fixed outbreak duration, Spain is producing twice the number of infected people per million than Austria and the US: we suggest that the worse performance is due to

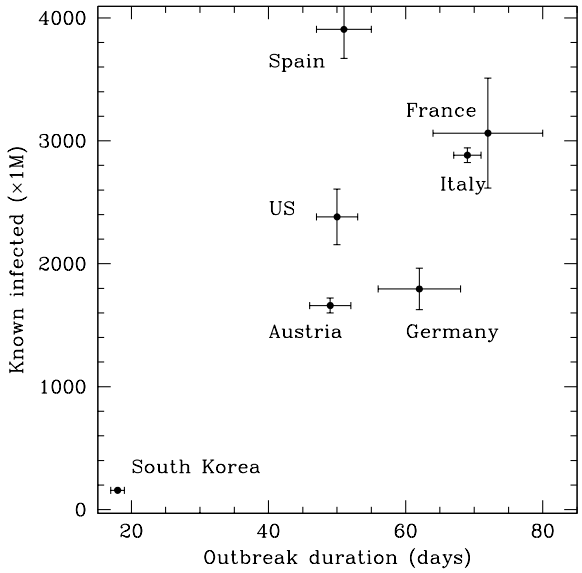

FIG. 5. We compare performances in countering the break from various countries. The x-axis shows the duration of the outbreak, defined as the time during which the number of new infected (ΔN) was higher than about 2/day/M. The y-axis shows the total (at the end of the wave, not just until observed) number of known infected people in the outbreak. Among this four, Italy and France have been the poorest performers in terms of speed in countering the outbreak, probably evidence of lower efficiency in the countermeasures. Spain reached the highest number of infected people per 1M population, suggesting they started countering the virus late. South Korea did outperform the other countries by over one order of magnitude in terms of number of infected per million, and by a factor of 3-4 in terms of duration.

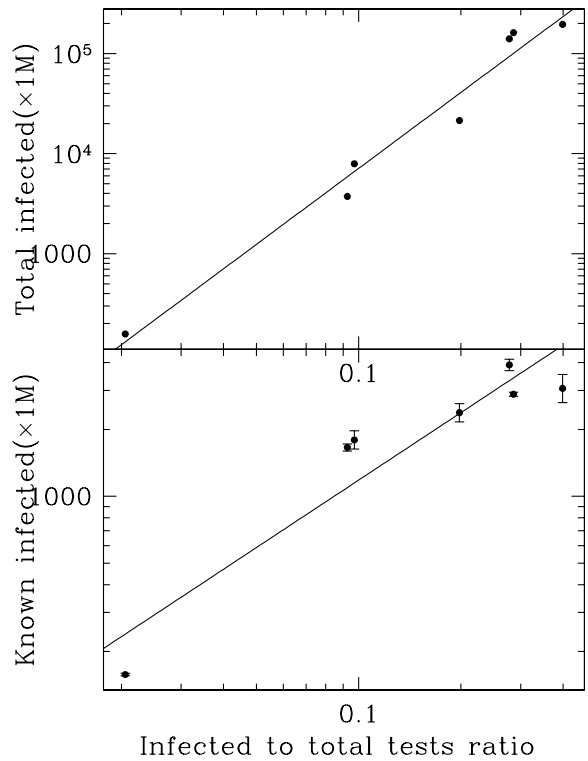

FIG. 6. **Bottom:** Correlation of infected to total test ratios versus total number of infected per milion in selected countries (from left to right points are from South Korea, Austria, Germany, US, Spain and Lombardia – that we use in representation of Italy as it carries the vast majority of its infections). **Top:** Correlation with total infected per 1M, with a correction attempting to account for both the undercounting of known infected and known deaths, assuming a universal CFR for COVID-19.

the longer delay before substantial containment measures have been adopted. Similarly, Italy and France appear to have responded later than Germany.

C. Reliability in reporting

The rms scatter of the residual from the Gumbel fit to the $N/\Delta NN$ time series, and/or the Pearson correlation coefficient, can be related to the day-to-day consistency and reliability in reporting data (Table I). The best performing countries appear to be Italy and the United States, while France appears to be the least well performing. South Korea also did not appear to provide very careful day-to-day reporting, which is also very well illustrated by time series of deaths $D/\Delta D$ that appears to be roughly constant with time and impossible to correlate with the corresponding data for the infections, after accounting for an appropriate delay time. Hence we set the delay to zero for display purposes in Fig.3, but estimate the delay to be actually quite long, as demonstrated that deaths are still counted in significant numbers to today (see *worldometers*).

D. Relation to testing rates

Different values of the ratio of infected to total tests are observed, too. When a large fraction of tested patients is found to be positive for the COVID-19 infection, one can argue not only that a substantial fraction of infected people are not censed, but also that in general the quality of the virus countering efforts are inadequate. This is illustrated by Figure 6, where we show the correlation (Pearson correlation coefficient 0.94) between the number of infections per million versus the fraction of positive tests. Similar levels of correlation are also observed with the parameter a (not surprisingly, given Fig. 5) and, more interestingly, with the integrated case fatality ratio (CFR), see Fig. 7-bottom. We derive the CFR in each country in which the outbreak is still actively occurring by taking into account the delay between the fractional growth rate of infected and deaths. In fact, an excel-

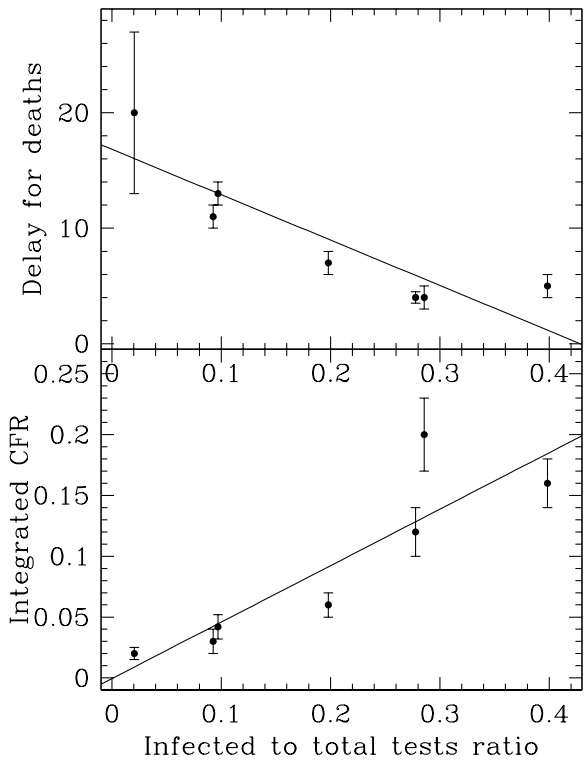

FIG. 7. Correlation of infected to total test ratios versus the integrated CFR (bottom) and the delay between infected and deaths (top), for the same countries as in Fig. 6. The delay is apparent from Figs 1–3 and is measured as the difference in the μ parameters from the Gumbel fit.

lent correlation is observed for CFR with this time delay between $N\Delta N$ of infected and $D/\Delta D$ of deaths (Fig. 7-top). This is obvious when considering that larger testing rates will discover larger numbers of less severely infected patients on average, and with the COVID-19 disease at an earlier stage.

Assuming that the intrinsic fatality of the virus is the same in all countries, Fig. 7 suggests that we can correct for the unaccounted infections by normalizing each country to the same CFR of 2% as observed in South Korea, and consistent to what seen in the Diamond Princess ship outbreak (Russell et al 2020; [27]), where everyone of the 3711 passengers were tested. Using this procedure, we derive the numbers listed in Table II (see CFR-corrected column). This suggests that when the current outbreak waves have subsided, up to about 1-2% of the population of several countries will have been infected, quite larger than the naive estimates of $< 0.3\%$ from the reported infections.

This might still largely underestimate the actual number of infections in a country: when a substantial fraction of COVID-19 patients remain untested, it is reasonable to believe that also deaths are undercounted as well. Estimates based on registries of deaths in various municipalities in Italy have shown [24; 25; 26] that this deaths under-reporting factor can reach up to 4–6×. We make the ad-hoc assumption that this further correction scales with the fraction of positive tests in exactly the same way as for the previous one (ratio of observed CFR to an intrinsic CFR of 2%). We are of course aware that this calculation is affected by uncertainty and a reliable assessment will require a widespread program of population testing to search for the prevalence of antibody to the SARS-CoV-2 virus. It is interesting that the numbers suggest that up to 15-20% of the population will infected (by the end of current outbreak waves) in Spain, Italy and France, a level not too far from the herd immunity (the numbers are much lower in other countries). This implies that containment in these three countries has been the least efficient. It is intriguing that in the Diamond Princess ship, despite the very confined environment, which favored close contacts with infected passengers, only about 20% of the passenger eventually tested positive, a fractions similar to the one inferred for Italy, Spain and France.

VI. DISCUSSION

A. A Universal description of outbreaks evolution

A significant finding of this study is that, in presence of containment measures, the evolution with time of $\Delta N/N$, the fractional rate of new infections in an epidemic outbreak, is remarkably well modeled by the Gumbel function, i.e it has a universal shape regardless of the type of adopted measures. In other words, while the Gumbel parameters a and μ (i.e., $t_{\rm Peak}$) vary from country to country, and from region to region within the same country (e.g. Italy), reflecting the diversity and degree of effectiveness of the adopted containment measures, the time evolution of new infections, which would be exponential in absence of any measure, always converges to the same functional relationship, i.e. the Gumbel function.

It is fundamental to notice that, as a result of the often highly inadequate and biased testing procedure adopted by each country, we do not actually have either complete or unbiassed measures of the true number of infected people (very often only people with severe symptoms of the COVID-19 disease get tested for the virus, which bias the sample towards individual who most susceptible to the disease, and only relatively small numbers of tests are conducted). These systematic effects combine with the varying effectiveness of the containment measures adopted by the specific country (degree of enforcement of social distancing, number and types of allowed essential activities, availability of protective devices) to produce the "effective" metrics that are used in studies such as ours (in this case the variable $N(t)$). The fact that these "effective metrics" all obey the same functional relationship in such different societies as South Korea, Spain, Italy and the US, strongly suggests that

this functional relationship is relatively insensitive to all systematics combined together. These systematics seem to always affect available metrics of the spread of the epidemic in the same way. This is most likely the result of using the ratio of the rate of increase ΔN to N itsef: even if N is affected by a bias, which we can think of as a multiplicative factor, since the variable of relevance is actually $r = \Delta N/N$, to first order the bias is largely eliminated in the ratio. To us, this seems a non-trivial result, reminiscent of a situation where the combined effect of a large number of independent causes result in the well-defined, universal final behavior of a random variate (e.g. normal distribution and the Central Limit Theorem). In this case, dr/r, namely the daily fractional change of r, which is the same as dR_0/R_0, diverges exponentially with time towards more and more negative values (Eq. 3).

B. Relation to the standard SIR model

It is instructive to compare our model to the standard SIR model which is commonly used to describe epidemic outbreaks ([28-33]). In the SIR model a total population T of individuals experience an infectious epidemic is divided into S, I and R groups, namely are the numbers of susceptible, infected and recovered (or killed by the disease) individuals (notice that we have adopted the notation N instead of I, that we keep in the following for clarity), where $T = S + N + R$ (all normalized to the total number of individuals to that T so that $T = 1$). The rates at which people are infected is regulated by the constants β and γ, such that $dS/dt = -\beta\, S\, N$, $dN/dt = \beta\, S\, N - \gamma\, N$ and $dR/dt = \gamma\, N$. In this model, which provides a description of an outbreak in absence of any measure aimed at slowing and reversing the spread of the disease, the quantities γ and β are constant in time. The parameter $R_0 = \beta/\gamma$ ([34]), usually called "the reproduction number", which measures the average number of secondary infected individual created by 1 primary infected primary, is often used as a metric of the effective power of the contagion in the outbreak. The early phase is characterized by $I = I_i \ll 1$ and $S_i = 1 - I_{\approx} 1$ and thus

$$N(t) = N_i\, e^{(R_0-1)\,\gamma\, t}. \tag{6}$$

If $R_0 > 1$ one has an epidemic outbreak. If left unchecked, the epidemic will continue until S substantially decreases and N and ΔN eventually deviates from the exponential growth and ΔN , in particular, reaches a peak: as a result of the decreased efficiency of new infections and the decrease of $N(t)$ as a result of people who recover or die. Letting an epidemic evolve in this way would result in unacceptable loss of life and containment measures must be set in place to break the exponential growth during the early phases. Such containment measures basically aim at breaking the exponential growth during the early phase by making R_0 (r in our model) become a function of time and bring it to $R_0 \leq 1$. It is important to realize that the peak that an unchecked epidemic would reach is qualitatively different from the peak that happens as a result of containment measures. The former basically happens because the population becomes depleted of Susceptible individuals; the latter is the direct result of the containment, which continuously shorten R_0, or r, which in effect is the inverse of the continuously increasing τ time-scale factor of the exponential growth (there is no exponential growth in an unchecked epidemic approaching the peak).

In the previous section we have suggested that, even if the available diagnostics of an epidemic are biased (e.g. the available number of infected individuals is both an incomplete and biased estimator of the true value), a fortuitous property of how all the biases combine together makes the time evolution of the resulting "observed fractional growth rate", $\Delta N/N$, a universal function of time. In addition, the fact of using the ratio of the first derivative of a variable to the variable itself greatly diminishes the effect of any systematic bias. This universal function of time turns out to be very useful to predict both the time occurrence of the peak, its width and the time when the infection is basically under control. That said, it is nonetheless of considerable interest to estimate the total number of infections at a given time during the outbreak (as we have attempted in Section V D), since this can be used to inform the testing strategy, optimize the containment measures, as well as to obtain a fair estimate the CFR of the disease, both cumulative and as a function of the age of the patients, useful to inform the clinical response.

VII. CONCLUSIONS

There are two main conclusions from this work, a conceptual one and a practical one. The conceptual one is that during the exponential phase of a pandemic outbreak and in presence of containment measures, the time evolution of the fractional growth rate of new infections, $\Delta N/N$, follows a universal functional form, which is very well modeled with the two-parameter Gumbel function. Fits to the data show that the two Gumbel parameters vary from country to country (and even from region to region within the same country), reflecting the specifics of the adopted containment measures. Remarkably, however, the functional form, remains the same. Since in absence of the containment measures, the evolution of the pandemic at this stage would still be in the exponential phase (as we have shown, the total number of infected is still small compared to the population (see Table II), it is possible that the containment measures modify the constant time scale of the exponential into a time-variable one, whose time-dependence is described by the Gumbel function. This is specified by two parameters μ and a, which we have fitted from the available data in a number of countries (Table I) and found that these parameters can be robustly constrained from the whole time span,

or subsets of it, from the onset of the outbreak to the current time, and their value remain unchanged in each country.

The main practical message of this work, which stems from the stability of the Gumbel parameters in each country, is that reliable predictions of the future evolution of the COVID-19 outbreak, including key events, can be obtained with substantial advance to inform critical strategic decisions. These events include the time of the peak and the time when the rate of new infections reaches pre-set low level such that the epidemic can be managed and social and economic activities can be resumed. All that is required would be well-controlled and stable strategy of testing and measure of infections and tracking of the evolution of the daily fraction of new infections. This should be possible for most countries, provided that testing is carried out in a stable and well-controlled manner that ensures that the daily rate of new infections is robustly and consistently measured and reported. At the same time, such an analysis provides an effective monitor of the outbreak, helping governments to keep it at a manageable level by strengthening (or softening) the containment measures as needed.

ACKNOWLEDGMENTS

We wish to acknowledge enlightening discussions with many of our colleagues.

[1] L. Tarrataca, C. M. Dias, D. B. Haddad, and E. F. Arruda, arXiv e-prints , arXiv:2004.06916 (2020), arXiv:2004.06916 [q-bio.PE].

[2] J.-T. Brethouwer, A. van de Rijt, R. Lindelauf, and R. Fokkink, arXiv e-prints , arXiv:2004.06891 (2020), arXiv:2004.06891 [cs.SI].

[3] D. Tsiotas and L. Magafas, arXiv e-prints , arXiv:2004.06536 (2020), arXiv:2004.06536 [physics.soc-ph].

[4] J. H. Fowler, S. J. Hill, R. Levin, and N. Obradovich, arXiv e-prints , arXiv:2004.06098 (2020), arXiv:2004.06098 [stat.AP].

[5] A. Senapati, S. Rana, T. Das, and J. Chattopadhyay, arXiv e-prints , arXiv:2004.04950 (2020), arXiv:2004.04950 [q-bio.PE].

[6] M. Chikina and W. Pegden, arXiv e-prints , arXiv:2004.04144 (2020), arXiv:2004.04144 [q-bio.PE].

[7] K. Amla and T. Amla, arXiv e-prints , arXiv:2004.03200 (2020), arXiv:2004.03200 [q-bio.PE].

[8] G. Mohler, F. Schoenberg, M. B. Short, and D. Sledge, arXiv e-prints , arXiv:2004.01714 (2020), arXiv:2004.01714 [q-bio.PE].

[9] R. M. Rizk-Allah and A. E. Hassanien, arXiv e-prints , arXiv:2004.05960 (2020), arXiv:2004.05960 [cs.NE].

[10] A. L. Bertozzi, E. Franco, G. Mohler, M. B. Short, and D. Sledge, arXiv e-prints , arXiv:2004.04741 (2020), arXiv:2004.04741 [q-bio.PE].

[11] M. Villalobos-Arias, arXiv e-prints , arXiv:2004.02406 (2020), arXiv:2004.02406 [q-bio.PE].

[12] B. Mbaye Ndiaye, L. Tendeng, and D. Seck, arXiv e-prints , arXiv:2004.01574 (2020), arXiv:2004.01574 [q-bio.PE].

[13] J. Dehning, J. Zierenberg, F. P. Spitzner, M. Wibral, J. Pinheiro Neto, M. Wilczek, and V. Priesemann, arXiv e-prints , arXiv:2004.01105 (2020), arXiv:2004.01105 [q-bio.PE].

[14] J. Naresh Dhanwant and V. Ramanathan, arXiv e-prints , arXiv:2004.00696 (2020), arXiv:2004.00696 [q-bio.PE].

[15] L. Alvarez, arXiv e-prints , arXiv:2003.10017 (2020), arXiv:2003.10017 [q-bio.PE].

[16] Z. Hu, Q. Ge, S. Li, E. Boerwincle, L. Jin, and M. Xiong, arXiv e-prints , arXiv:2003.09800 (2020), arXiv:2003.09800 [q-bio.PE].

[17] D. Fanelli and F. Piazza, arXiv e-prints , arXiv:2003.06031 (2020), arXiv:2003.06031 [q-bio.PE].

[18] M. e. Prince, arXiv e-prints , arXiv:2020.04.09.20057091 (2020), medRXiv:2020.04.09.20057091.

[19] E. Gumbel, "Les valeurs extremes des distributions statistiques," (1935).

[20] A. Bianconi, A. Marcelli, G. Campi, and A. Perali, arXiv e-prints , arXiv:2004.04604 (2020), arXiv:2004.04604 [q-bio.PE].

[21] *Worldometer*, "www.worldometers.info/coronavirus/," (2020).

[22] pcm-dpc data repository, "github.com/pcm-dpc/covid-19," (2020).

[23] CSS data repository, "github.com/cssegisanddata/covid-19/tree/master/," (2020).

[24] Eco di Bergamo, "www.ecodibergamo.it/stories/bergamo-citta/quasi-mille-morti-nella-bergamascai-sindaci-ma-sono-molti-di-piu_1346006_11/," (2020).

[25] Corriere della Sera, "www.corriere.it/politica/20_marzo_25/numero-vero-morti-covid-19-almeno-4-volte-quello-ufficiale-eebbe3ae-6eb8-11ea-925b-a0c3cdbe1130.shtml," (2020).

[26] Pandemic Substack blog, "pandemic.substack.com/p/the-elephant-in-the-room-undercounting," (2020).

[27] T. W. Russel and al., "cmmid.github.io/topics/covid19/severity/diamond_cruise_cfr_estima (2020).

[28] A. G. Kermack, W. O.; McKendrick, Proceedings of the Royal Society A (1927).

[29] H. Hethcote, SIAM Review **42 (4)**, 599 (2000).

[30] Applied Mathematics and Computation **236**, 184 (2014), arXiv:1403.2160.

[31] N. Bailey, *The mathematical theory of infectious diseases and its applications (2nd ed.)* (London. Griffin, 1975).

[32] C. Altizer, S.; Nunn, *Infectious diseases in primates: behavior, ecology and evolution. Oxford Series in Ecology and Evolution* (2006).

[33] J. Miller, Infectious Disease Modelling **2 (1)**, 35 (2017).

[34] T. G. L. Levin, S.A.; Hallam, ed., *Three Basic Epidemiological Models*, Applied Mathematical Ecology. Biomathematics No. 18 (Berlin: Springer, 1989).