InnoVationinfo

Journal of
Health Science and Development

ISSN: 2581-7310
Volume 3: 2
J Health Sci Dev 2020

Pitting the Gumbel and logistic growth models against one another to model COVID-19 spread

Keunyoung Yoo[1]

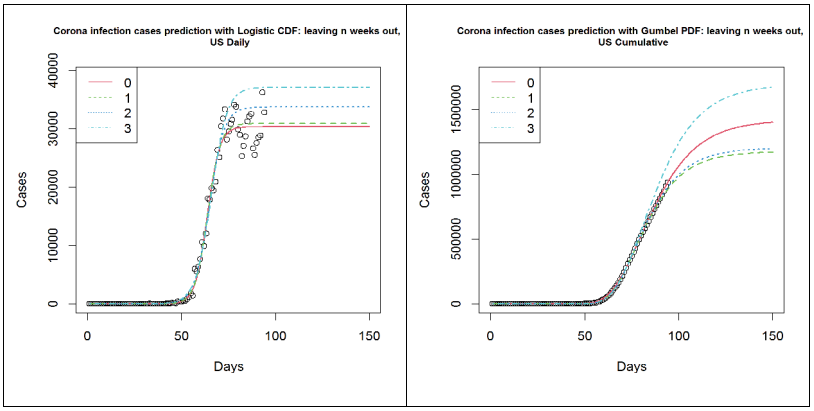
Mohammad Arashi[*1,2]

Andriette Bekker[1]

[1]*Department of Statistics, Faculty of Natural and Agricultural Sciences, University of Pretoria, South Africa*
[2]*Department of Statistics, Faculty of Mathematical Sciences, Shahrood University of Technology, Shahrood, Iran*

Abstract

In this paper, we investigate briefly the appropriateness of the widely used logistic growth curve modeling with focus on COVID-19 spread, from a data-driven perspective. Specifically, we suggest the Gumbel growth model for behaviour of COVID-19 cases in several countries in addition to the United States of America (US), for better detecting the growth and prediction. We provide a suitable fit and predict the growth of cases for some selected countries as illustration. Our contribution will stimulate the correct growth spread modeling for this pandemic outbreak.



Graphical Abstract

Highlights

Exploring logistic curve modeling for COVID-19 data and illustrating the shortcomings.

Proposing the modeling of COVID-19 spread with the Gumbel growth curve.

Fitting of COVID-19 data from different countries to strongly support the Gumbel model choice.

Keywords: Asymmetric; Logistic growth model; Non-linear regression; Prediction.

Introduction

Nowadays, the Coronavirus pandemic, known as COVID-19, caused by a novel pathogen named Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV2), has shown that in early stages of infection, symptoms of severe acute respiratory infection can occur and it is rapidly spreading across the globe. Since we have limited knowledge

Article Information

Article Type: Analysis Article

Article Number: JHSD 125

Received Date: 25 August, 2020

Accepted Date: 19 October, 2020

Published Date: 26 October, 2020

[*]**Corresponding author:** Mohammad Arashi, Department of Statistics, Faculty of Mathematical Sciences, Shahrood University of Technology, Shahrood, Iran Tel: +27 12 420 3774; Email: m_arashi_stat@yahoo.com

Citation: Yoo K, Arashi M and Bekker A (2020) Pitting the Gumbel and logistic growth models against one another to model COVID-19 spread. J Health Sci Dev Vol: 3, Issu: 2 (17-30).

about COVID-19, epidemiological modeling is still under development and modeling the ecological growth based on the population demographic information is feasible for reporting. It is to support the shaping of decisions around different non-pharmaceutical interventions.

The logistic function/curve is commonly used for dynamic modeling in many branches of science including chemistry, physics, material science, forestry, disease progression, sociology, etc. But, the question is whether it is also suitable for COVID-19 spread modeling from the available data viewpoint. The principle of exponential growth can be applied to the transmission of COVID-19 (see Little, for a web based dashboard [1]). It is known that the exponential model is adequate to describe for a short period and in general it will quickly deviate from actual numbers as time passes. The logistic growth curve was successful in modeling some epidemics [2-6]. Our primary goal is to see whether the logistic function can suitably predict the spread. Some endeavors have been made to predict and forecast the future trajectory of the COVID-19 outbreak. We refer to Cohen, Bastista, Roser et al., Hsu, Anastassopoulou et al., Maier and Brockmann, Cassaro and Pires, Ceylan, Salehi et al, Sauer and Petropoulos and Makridakis to mention a few related studies [7-17].

In none of the above mentioned studies, the Gumbel function is applied for predicting the growth of COVID-19. Hence, in this contribution, a dynamic Gumbel model is used to track the coronavirus COVID-19 outbreak. We organize the rest of this work as follows. In the forthcoming section, we provide the source of data and software used for comparison and fitting purposes. Section 5 includes the analysis of logistic modeling, outlines the shortcomings, proposes the Gumbel model as the suitable candidate; followed by comparison with the Logistic model. Section 7 illustrates the potential of the Gumbel model with the analysis of the COVID-19 data for selected countries. We conclude our contribution in Section 8.

Experimental data

There are a number of sources on the web that provide data on COVID-19 cases. One such site is "The Humanitarian Data Exchange" and one can find daily cumulative cases of COVID-19 per country. (https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases) has a downloadable "time_series_covid19_confirmed_global.csv" starting from 2020-01-22, and for some countries, it even has the data broken down into different states or provinces. In order to perform the desired analysis, daily cases for each country had to be obtained, but some countries, such as the US and Australia, had the data broken down to state or provincial level. Since the focus of this research was per country, R open source software was used to sum along the unique values of Country, appropriately transforming the data for our analysis, then and non-linear regression was performed using the nls function.

Methodology and Results

Preliminary insight by using the logistic growth model

In this section, we conduct data analysis using the commonly used logistic curve modeling. A logistic function is a common sigmoid curve with the following functional form for the dynamic model of population at time t

$$P(t) = \frac{L}{1 + e^{-\alpha\,(t - t)}} \tag{1}$$

with initial condition $P(to) = Po$, L is the carrying capacity, the maximum capacity of the environment here, $\alpha > 0$. Here, Eq. (1) divided by L corresponds to the cumulative distribution function (CDF) of a logistic distribution at point t. The probability density function (PDF) is simply obtained by differentiating the latter with respect to t.

This is useful because the difference of two Gumbel-distributed random variables has a logistic distribution. The seemingly exponential growth of COVID-19 cases across the globe is typically the lower half of a logistic curve during the early stage.

Results and discussion-Logistic model: The analysis is data-driven, and therefore, the focus of the paper is not from an epidemiological perspective. Nevertheless, the parameter estimates are relatable to the real world. L represents how many cases we expect to see in the end, α is how quickly the virus has spread/cleared and to is where the peak increase in cases was observed. To illustrate the failing of the logistic model, the US data was the focus here.

Modeling the US cases, based on data until 28 March, the following results were obtained for regression. The model was highly significant with a p-value less than 0.0001 and this is shown in the plot as well, where the actual US data and the model are almost indistinguishable. This data suggests the total number of COVID-19 cases will be approximately between 226,000 and 265,000. The number of cases for the next 7 weeks was forecasted using these estimates. However, when data until 4 April is subsequently used, parameter L, which represents the final number of cases (477922) is far beyond what was predicted using data until the previous week (upper bound forthe

confidence limit of a was 265206). The slope parameter, α, decreased while the location parameter, to increased. (Figures 1,2 and Tables 1,2,3).

Using data until 2020-04-25, the new estimate for L once again exceeds what was predicted using previous data and the slope parameter, α, decreased while the location parameter, to increased. (Figure 3)

Modelling the cumulative cases can be viewed as trying to model the forest as a whole, as opposed to looking at each tree. Even if a particular tree is twice as tall as most other trees in the forest, it will not make a big impact on the whole when all the heights are summed up. Therefore, to introduce more variability to the data, the next approach was to analyse the daily new cases instead, by taking the difference of the cumulative data. This way, the magnitude of daily cases will not be reduced as more data is acquired, and it will capture the effect of large spikes. In other words, we are zooming into the data to give more weight to the daily number of cases.

The parameter estimates below are based on the same

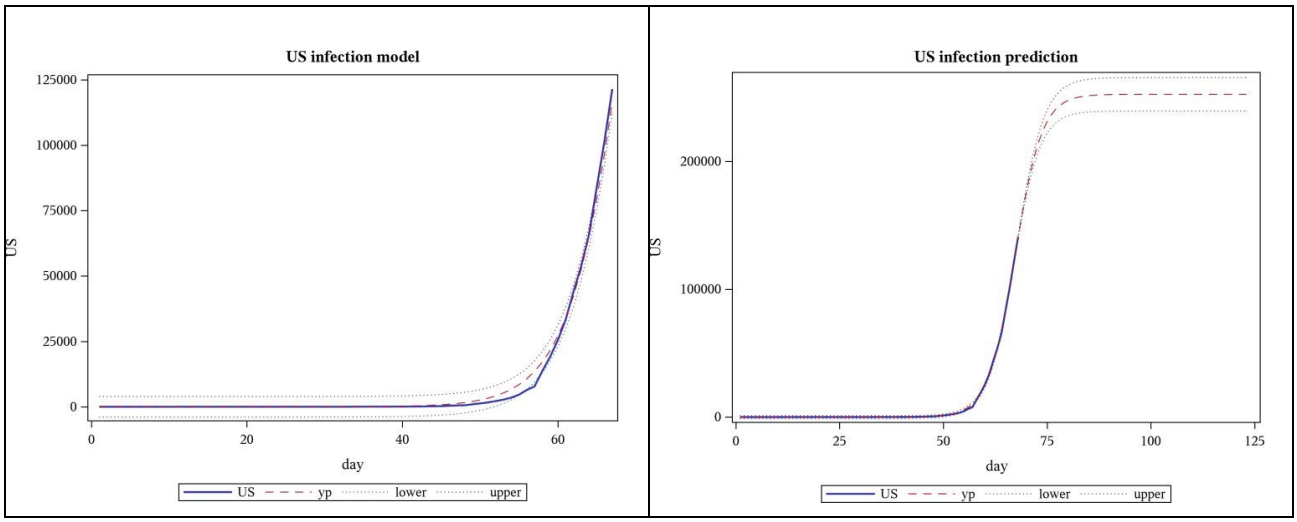


Figure 1: Observed US cases from 2020-01-22 to 03-28 and forecast for 7 weeks, using logistic function.

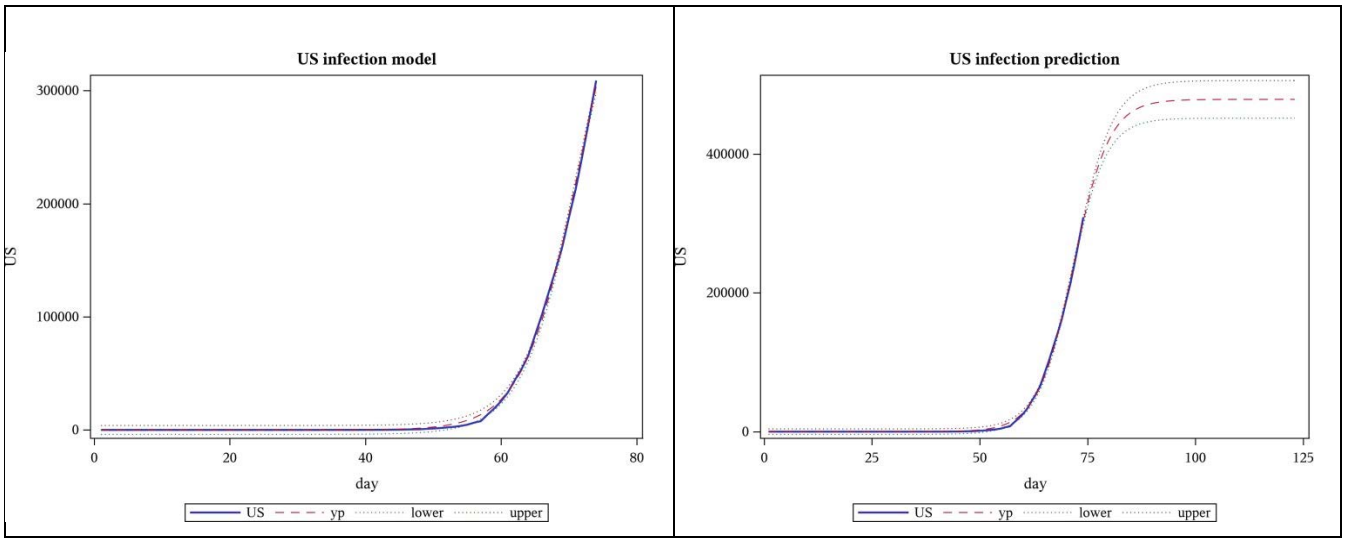


Figure 2: Observed US cases from 2020-01-22 to 04-04 and forecast for 7 weeks, using logistic function.

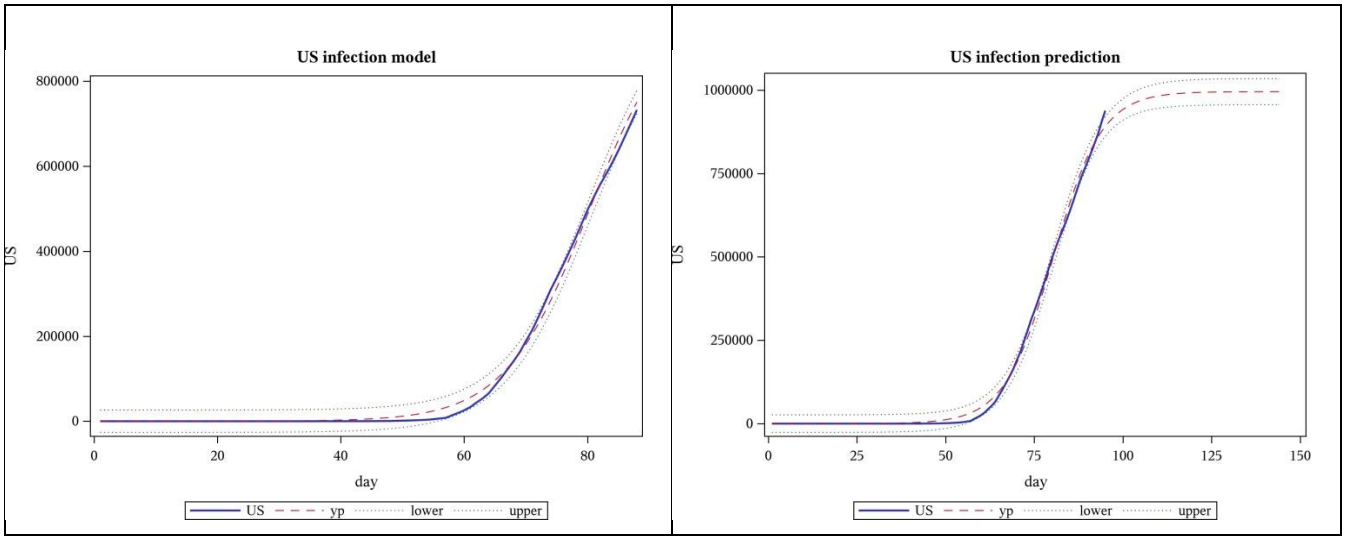


Figure 3: Observed US cases from 2020-01-22 to 04-25 and forecast for 7 weeks, using logistic function.

Parameter	Estimate	Approx Standard Error	Approx 95% Confidence Limits	
L	245898	9653.8	226590	265206
α	0.3098	0.00483	0.3001	0.3195
t_o	67.1124	0.2341	66.6442	67.5806

Table 1: US- logistic function based parameter estimates from 2020-01-22 to 03-28.

Parameter	Estimate	Approx Standard Error	Approx 95% Confidence Limits	
L	477922	13319.9	451282	504562
α	0.2403	0.00405	0.2322	0.2484
t_o	71.7055	0.2409	71.2236	72.1874

Table 2: US- logistic function based parameter estimates from 2020-01-22 to 04-04.

Parameter	Estimate	Approx Standard Error	Approx 95% Confidence Limits	
L	995370	14689.3	966196	1024544
α	0.1459	0.00321	0.1395	0.1523
t_o	80.3266	0.2772	79.7761	80.8772

Table 3: US- logistic function based parameter estimates from 2020-01-22 to 04-25.

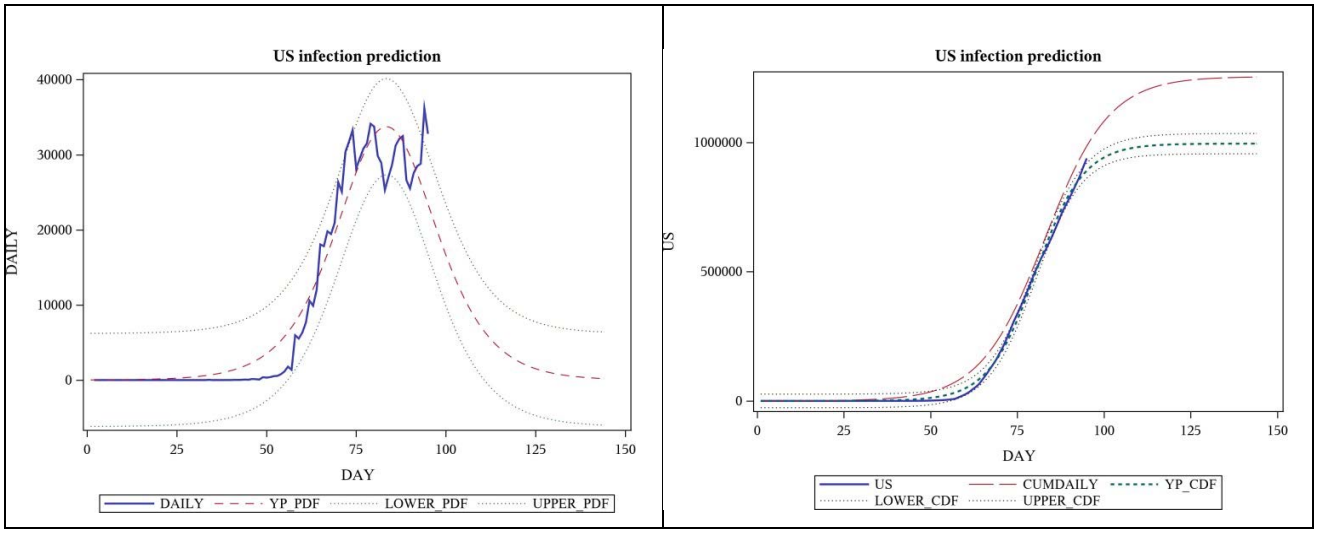


Figure 4: Observed daily US cases from 2020-01-22 to 04-25 and forecast for 7 weeks, using logistic function's derivative.

data as the one above. The p-value for the model was still <0.0001, suggesting that its significance was not lost in the new approach. One observation was lost in the process of taking the difference, but by looking at the daily cases and trying to fit a PDF instead of a CDF, we can get a much detailed view of the situation, and it has increased the estimate for the total number of cases. See figure 4; the view by focusing on daily case modelling using the first derivative of the logistic function (Table 4).

Shortcomings: In using a sigmoid function to model the data, an implicit assumption was made that it will take the same length of time for the spread of virus to "rise" as it will to "fall." This comes from the fact that the Logistic function is symmetrical about the inflection point. The bar charts (Figure 5) show the daily new cases for Spain, Italy and the US. Just looking at the charts below is enough to question whether trying to fit a symmetrical shaped curve will provide a good fit or predictability. Hence the next step was to find a distribution whose CDF appears to have the general "S" shape which has the characteristics of a sigmoid function, yet possesses some skewness built into it such that when modelling the daily new cases, it fits the asymmetrical data well. After looking at numerous distributions that meet all criteria, the Gumbel distribution seemed to possess promising properties.

Figure 6 suggests how easily our eyes can deceive us. The red lines (CDF and PDF) are the Logistic distribution and the blue lines (CDF and PDF) are the Gumbel distribution. (The dashed lines are the PDF and the solid lines are the CDF.) If we were to just view the CDFs in isolation, there is no way that a human will be able to tell whether the curve is symmetric or not. Even with the x and y axis drawn, merely shifting the Gumbel CDF to the left slightly will be enough to fool the viewer that the distribution is convincingly symmetric. On the other hand, detecting symmetry (or lack thereof) using a PDF is visually clear, and it does not require an expert to determine that while the dashed red curve (of the Logistic function) is symmetric, the dashed blue curve (of Gumbel) is not. Hence looking at the daily data and detecting this skewness was crucial in suggesting an alternative model.

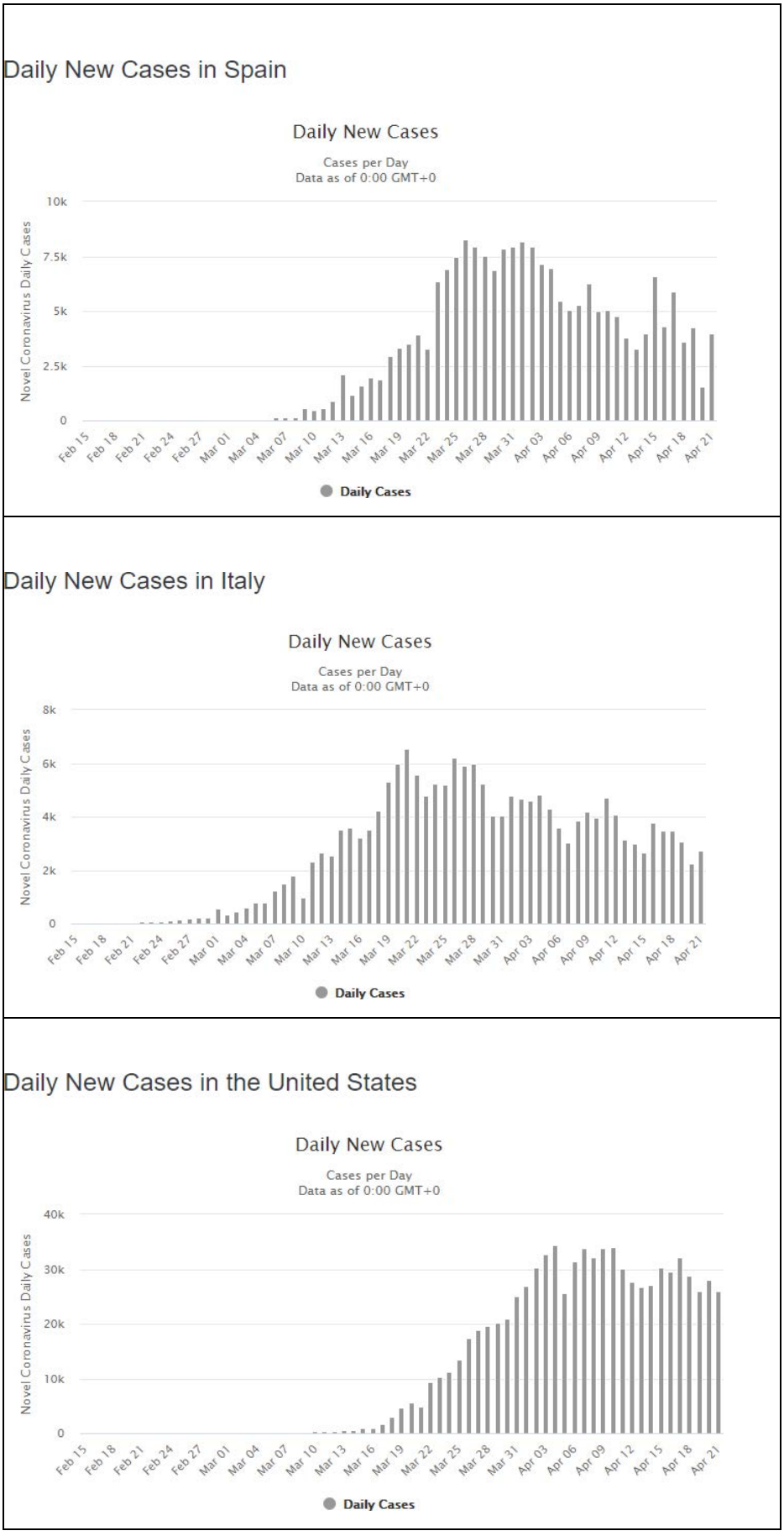


Figure 5: Bar chart of Daily new cases in Spain, Italy and the US (extracted from https://www.worldometers.info/coronavirus/country/us/).

Gumbel growth modeling

The Gumbel distribution has been frequently used for practical probabilistic modeling. Gumbel (Anderson and Daniewicz, Gomez et al., Hyun et al., Huang et al.) presents a model as an extension of the exponential distribution with the feature that it can be used to fit extreme datasets [18-21]. A Gumbel dynamic model of population at time t is defined by

$$P(t) = Ae^{-e^{\frac{t-t_o}{\beta}}} \quad (2)$$

with initial condition $P(to)$ = Po, A is the carrying capacity, the maximum capacity of the environment here, $\beta > 0$. Here, Eq. (2) divided by L corresponds to the CDF of the Gumbel distribution at point t. The PDF is simply obtained by differentiating the latter with respect to t.

Overall, the same process as the logistic function was performed with the Gumbel distribution's PDF and CDF. The results indicate that using Gumbel is strongly preferred over the logistic, regardless of whether the Gumbel PDF (daily) or CDF (cumulative) is used. The parameter estimates for the total number of cases are no longer caught up within a week and even visually, the trajectory of the graph suggests paths for each country that are smoother and more accommodating towards future outcomes.

Regarding the parameter estimates, while the roles of "A" and "to" are analogous to those of "L" and "to" from the logistic function, respectively, the parameter "beta" plays a somewhat different role- as a slope/duration dual-function parameter which shrinks or stretches the curve. The Gumbel model incorporates some level of skewness which allows it to pick up broader variation in the data. Note that the standard errors are larger for the PDF

based estimates, which is to be expected since it uses the volatile daily data as opposed to rather-stable cumulative data .(Tables 5,6 and Figure 7)

Comparison and Discussion

The following plots (Figure 8) summarise the key difference in using Gumbel distribution over Logistic distribution for the modelling of COVID-19 infection cases. The data used here is the number of cases in the US until 2020-04-25, where the circles represent the number of cumulative cases. The left panel shows the different models based on the Logistic function and the right panel shows the different models based on the Gumbel distribution's CDF. The different lines indicate how many weeks' worth of observations have been left out to simulate the results that were obtained in the past. On the left panel, it is clear that the Logistic model fails to capture an important trait in the data, hence it fails to keep up with the data. This is, as argued above, due to the asymmetric nature of the data. On the right panel, however, the Gumbel model is much more robust in picking up such trends. Though the prediction from 3 weeks ago has overestimated the number of cases, thereafter the estimates have remained rather stable and appear to be converging for the past 2 weeks.

Gumbel Modelling for Some Selected Countries

In this section, we analyse the dynamics of the coronavirus disease COVID-19 for some selected countries to show the potential of the Gumbel model (Figure 9). The time frame window is from 2019-12-31 to 2020-10-12, except for Turkey (from 2020-03-12) and Peru (from

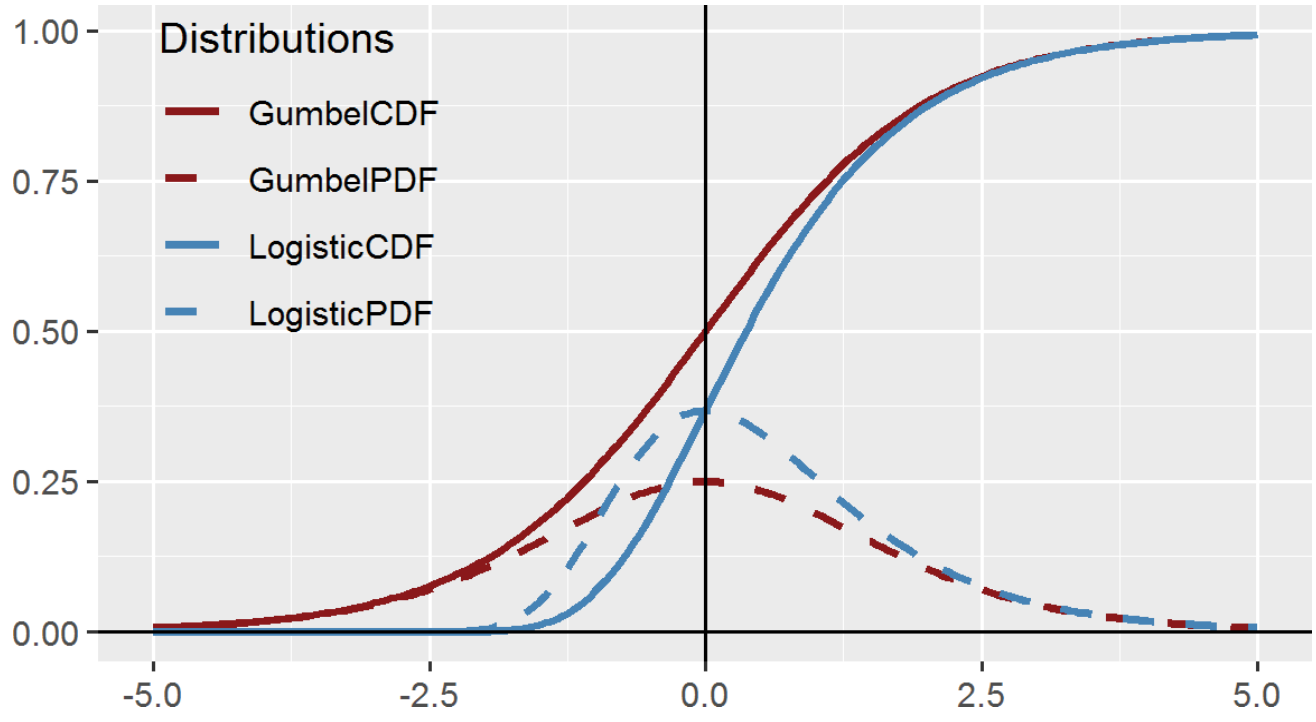


Figure 6: PDF and CDF of Logistic distribution and Gumbel distribution.

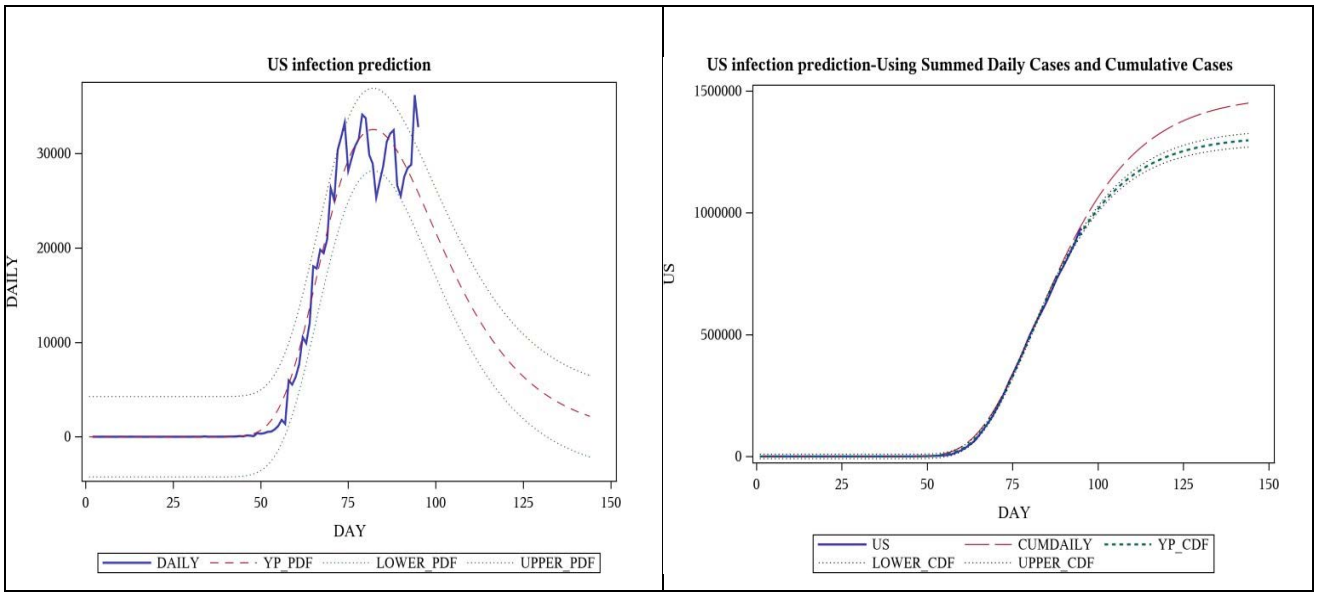


Figure 7: Observed daily US cases from 2020-01-22 to 04-25 and forecast for 7 weeks, using Gumbel PDF.

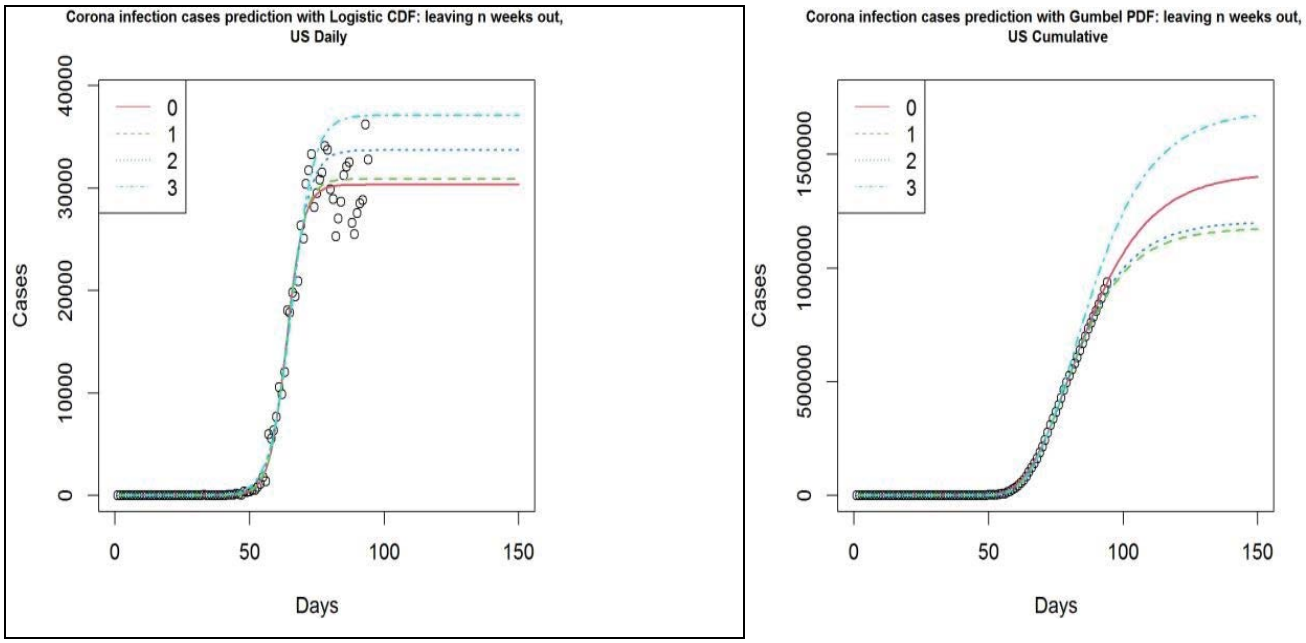


Figure 8: US- Logistic function (left panel) and Gumbel PDF (right panel) based forecasts from 2020-01-22 to 04-04 (3), 04-11 (2), 04-18 (1) and 04-25 (0).

2020-02-28). Additionally, only the Gumbel PDF model ran without singularity or iteration issues with some countries, which is also evidence that speaks to its robustness. Further, for practical purposes, we provided a month prediction for November given in table 7.

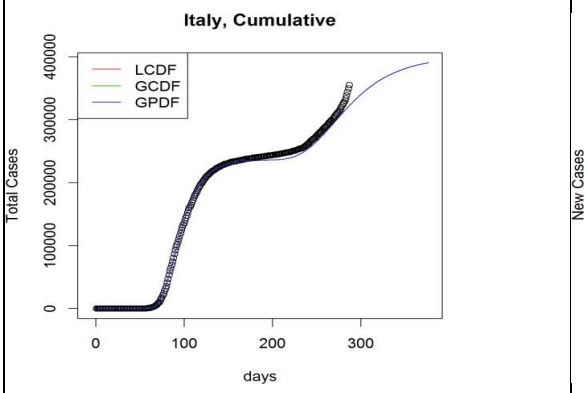
Conclusion

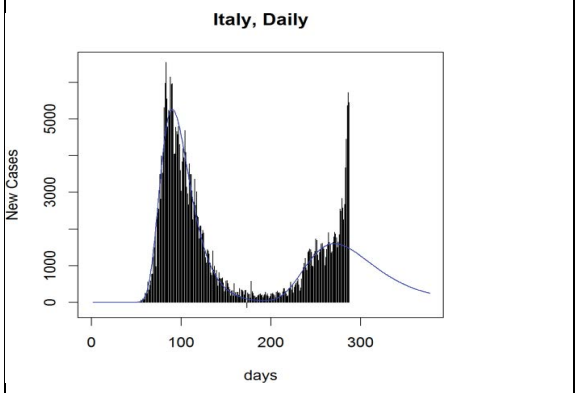
In this paper, we have investigated the logistic growth model. The shortcomings were shown. We guided the reader to the solution of the use of the Gumbel model as an appropriate choice and completed the prediction for several countries. As Panovska-Griffiths pointed out one model cannot answer all the questions [22]. We hope this contribution can be a part of the set of solutions. The authors hope that this model will be of assistance for decision makers. This paper is part of an ongoing project related to modeling and prediction of COVID-19 spread.

Declaration of Competing Interest

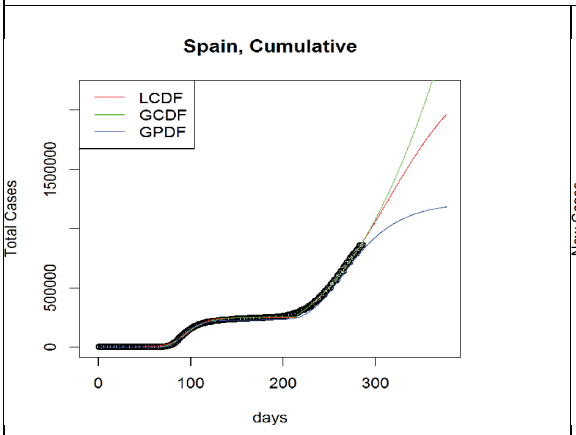
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

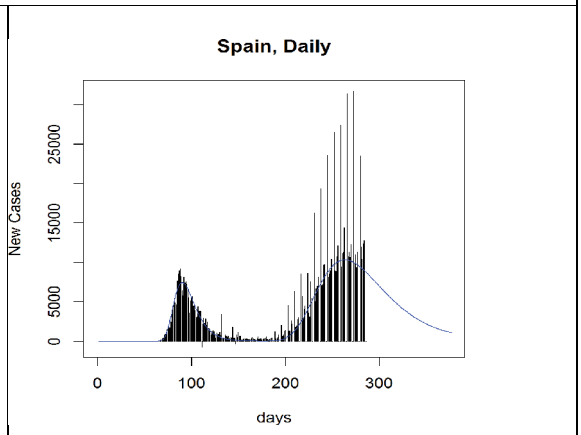
Italy



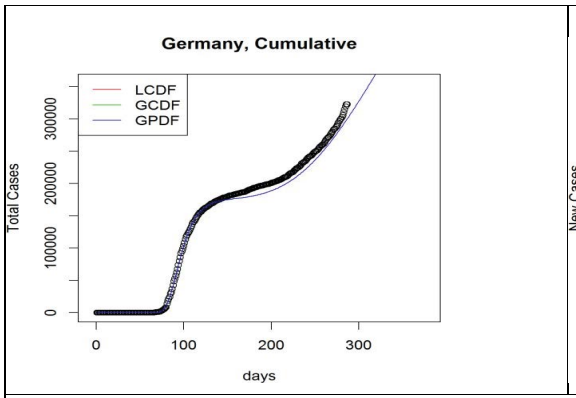


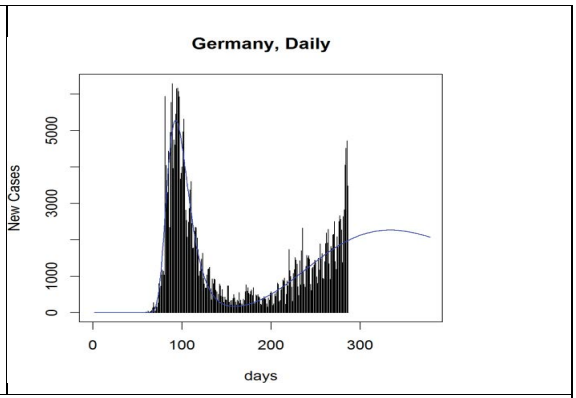
Spain



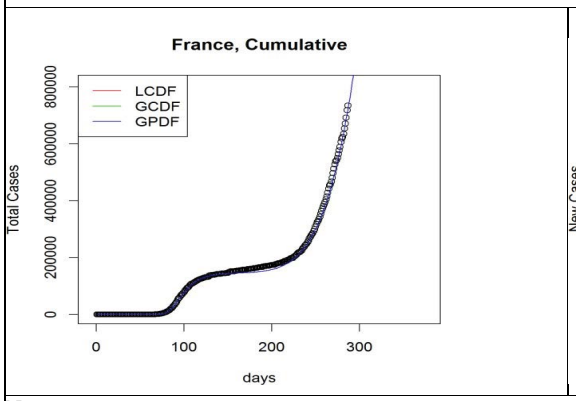


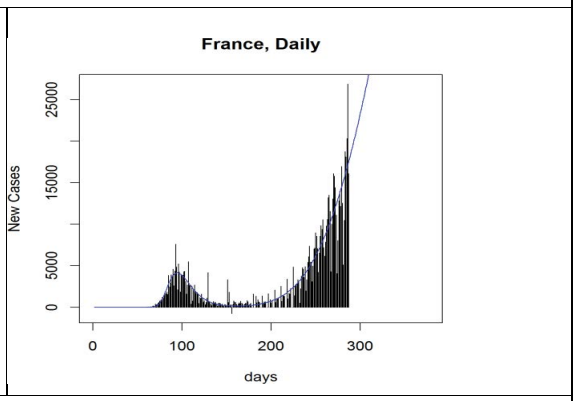
Germany



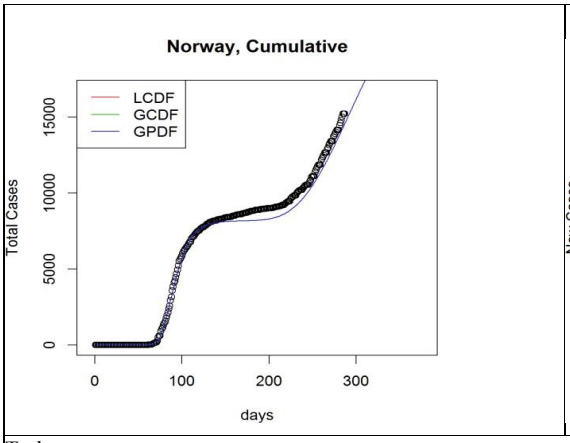


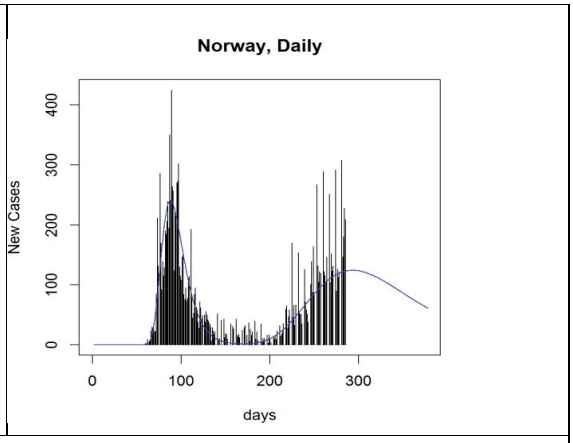
France



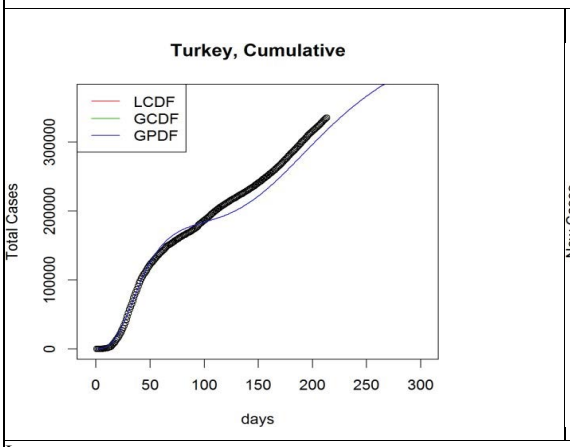


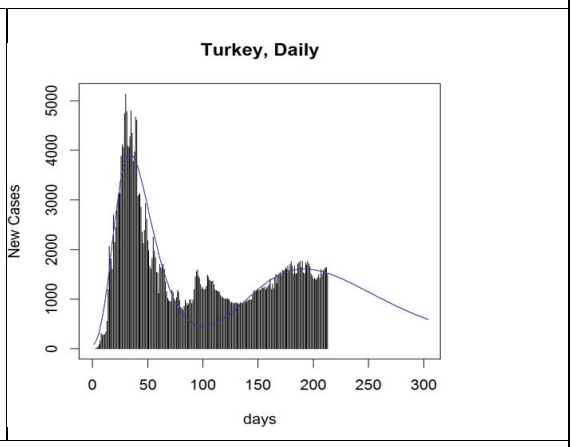
Norway



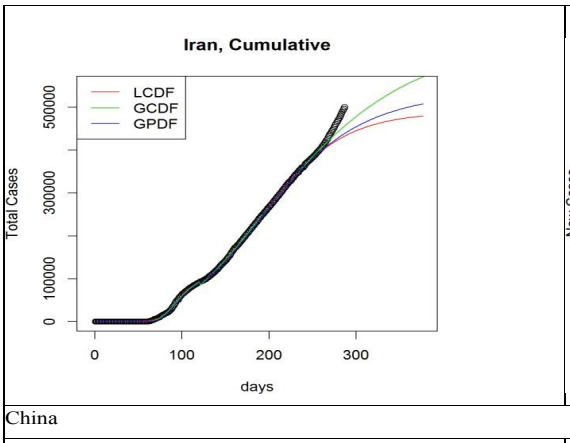


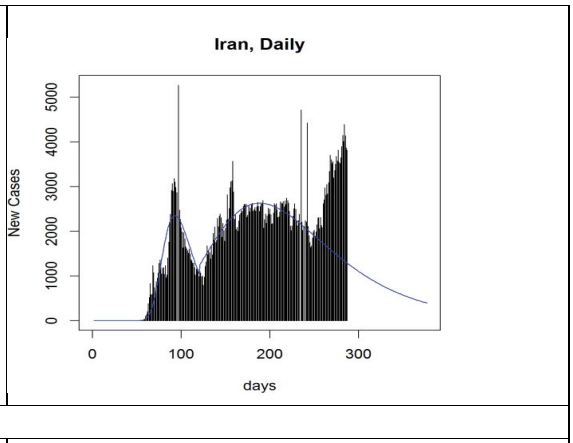
Turkey



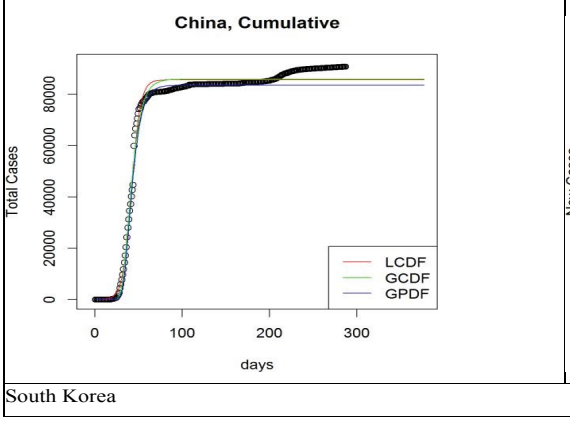


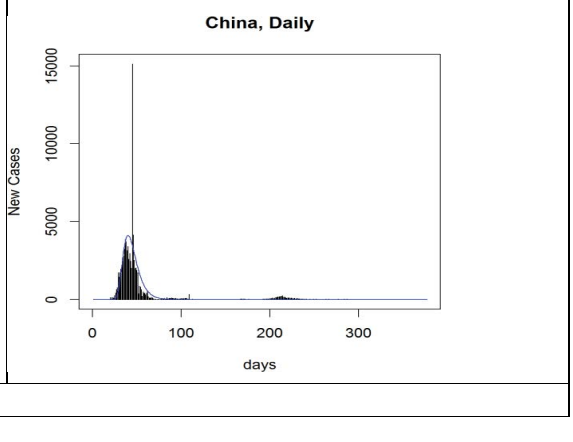
Iran



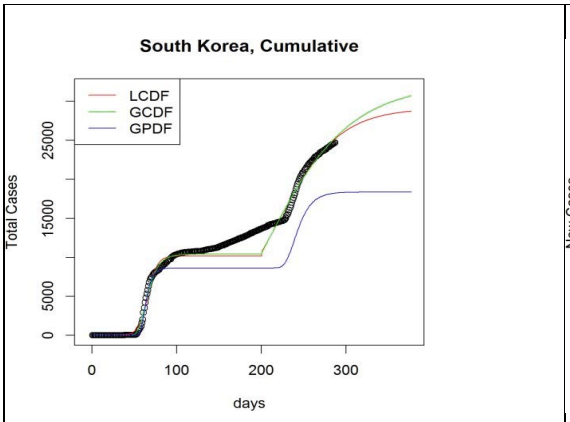


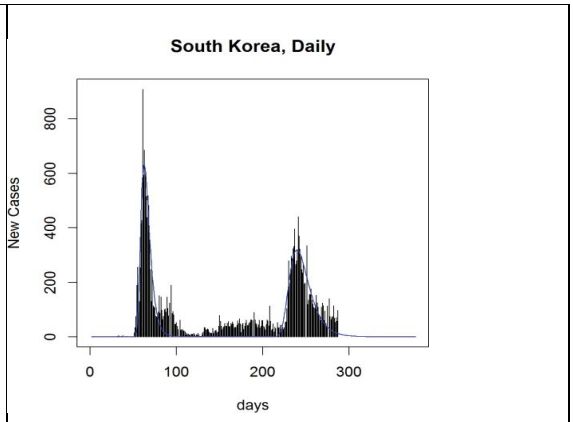
China



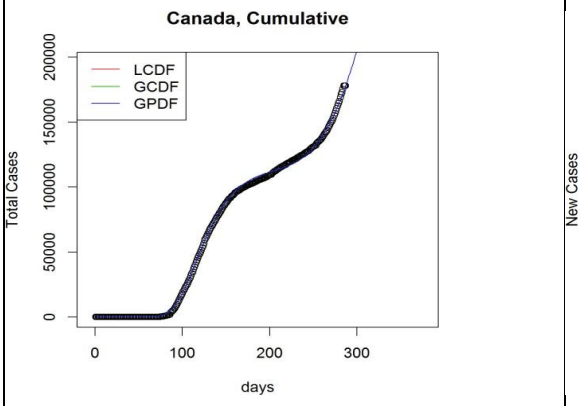


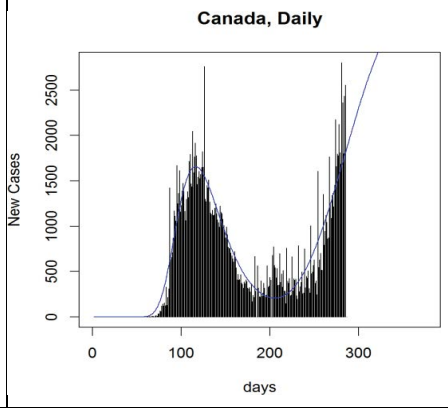
South Korea



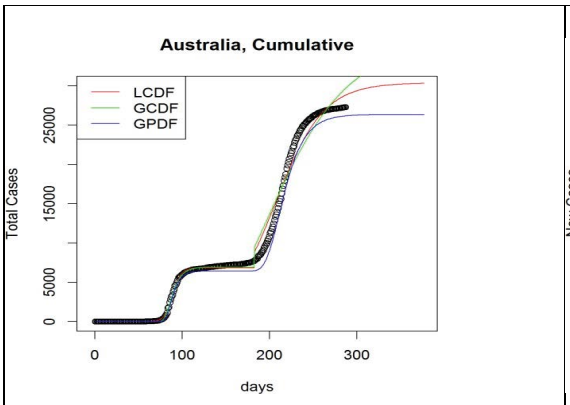


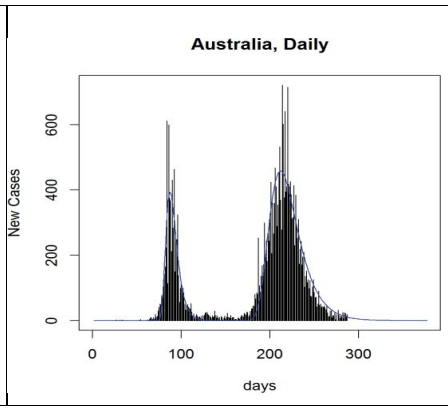
Canada



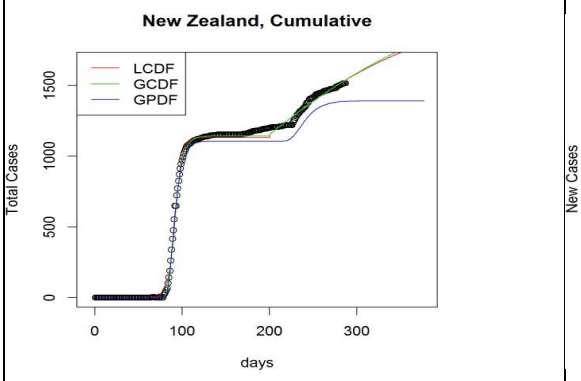


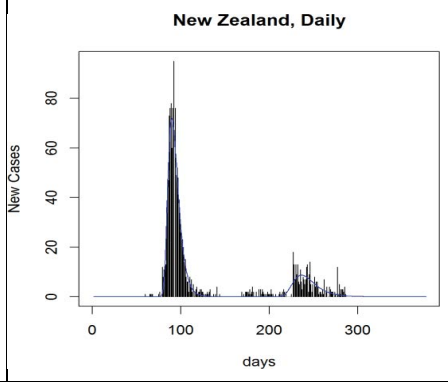
Australia





New Zealand





India

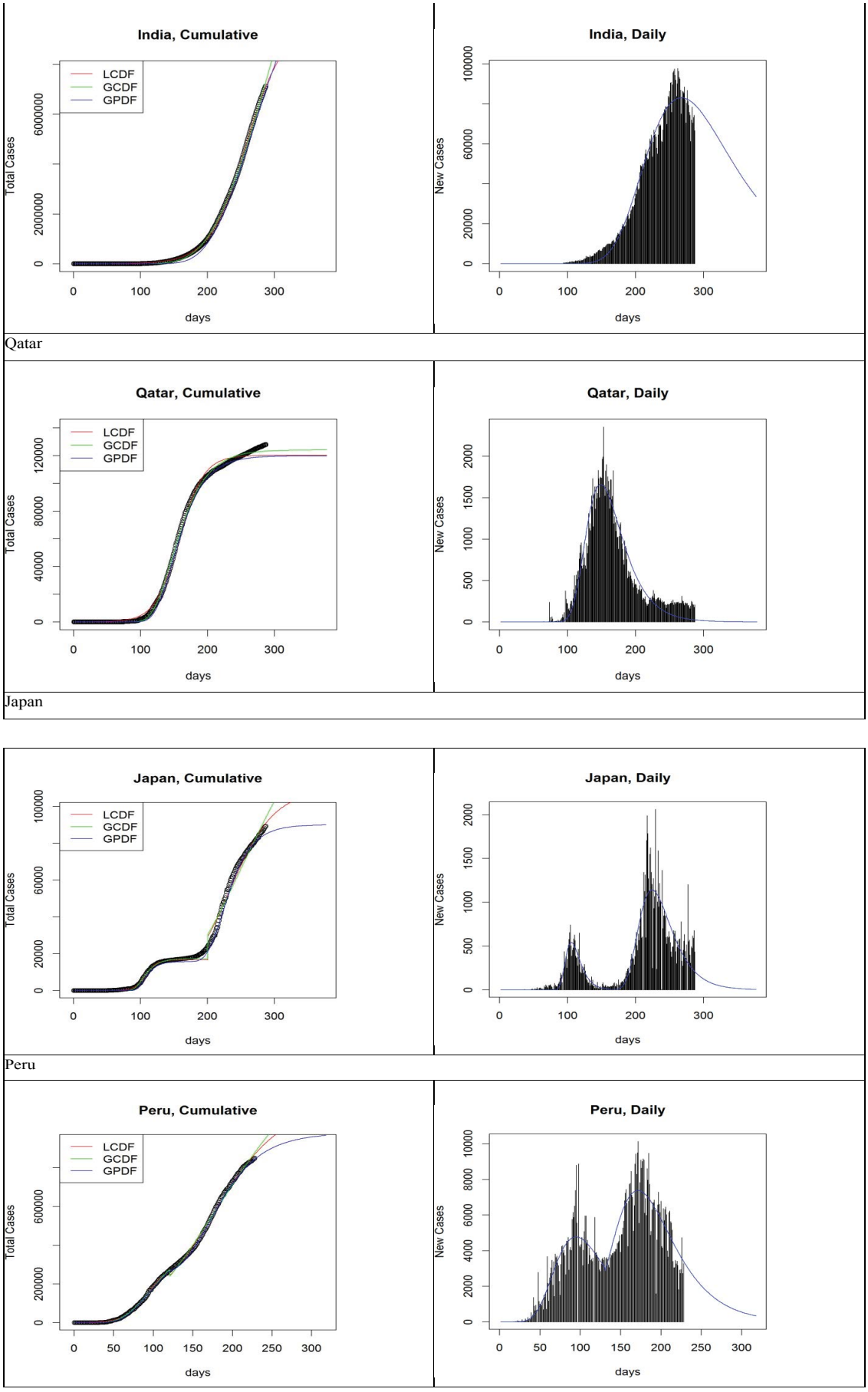


Figure 9: Logistic CDF, Gumbel CDF and PDF based forecasts from 2020-03-12 (Turkey), 2020-02-28(Peru), 2019- 12-31(all other countries) to 2020-10-12, for selected countries.

Parameter	Estimate	Approx Standard Error	Approx 95% Confidence Limits	
L	1254903	47672.7	1160207	1349599
α	0.1076	0.00493	0.0978	0.1174
t_o	83.4375	0.5791	82.2871	84.5879

Table 4: US- logistic function's derivative based parameter estimates from 2020-01-22 to 04-25.

Parameter	Estimate	Approx Standard Error	Approx 95% Confidence Limits	
A	1485959	50413.7	1385819	1586100
β	16.7931	0.6251	15.5514	18.0349
t_o	82.1391	0.5729	81.0010	83.2771

Table 5: US- Gumbel PDF based parameter estimates.

Parameter	Estimate	Approx Standard Error	Approx 95% Confidence Limits	
L	1315121	14613.8	1286097	1344146
α	14.8945	0.1874	14.5224	15.2666
to	79.8614	0.1936	79.4769	80.2458

Table 6: US- Gumbel CDF based parameter estimates.

Date	Italy	Spain	Germany	France	Norway	Turkey	Iran	China
1-Nov-2020	1132	6118	2182	30934	121	1316	1011	0
2-Nov-2020	1114	5995	2188	31531	120	1305	998	0
3-Nov-2020	1095	5874	2195	32133	120	1294	986	0
4-Nov-2020	1076	5755	2201	32738	119	1283	973	0
5-Nov-2020	1058	5636	2206	33347	119	1272	961	0
6-Nov-2020	1040	5519	2212	33959	118	1261	949	0
7-Nov-2020	1021	5403	2217	34575	118	1249	937	0
8-Nov-2020	1003	5288	2222	35193	117	1238	925	0
9-Nov-2020	985	5175	2226	35815	116	1227	913	0
10-Nov-2020	967	5064	2231	36439	116	1215	902	0
11-Nov-2020	949	4953	2235	37066	115	1204	890	0
12-Nov-2020	931	4845	2239	37696	114	1192	879	0
13-Nov-2020	914	4738	2242	38327	114	1181	867	0
14-Nov-2020	896	4633	2246	38961	113	1169	856	0
15-Nov-2020	879	4529	2249	39597	112	1158	845	0
16-Nov-2020	862	4427	2252	40235	111	1146	834	0
17-Nov-2020	845	4326	2254	40874	111	1135	823	0
18-Nov-2020	829	4227	2257	41514	110	1123	813	0
19-Nov-2020	812	4130	2259	42156	109	1111	802	0
20-Nov-2020	796	4035	2261	42799	108	1100	792	0
21-Nov-2020	780	3941	2262	43443	107	1088	781	0
22-Nov-2020	764	3849	2264	44088	107	1077	771	0
23-Nov-2020	748	3758	2265	44733	106	1065	761	0
24-Nov-2020	732	3669	2265	45379	105	1054	751	0
25-Nov-2020	717	3582	2266	46024	104	1043	741	0
26-Nov-2020	702	3497	2267	46670	103	1031	731	0

Date	South Korea	Canada	Australia	New Zealand	India	Qatar	Japan	Peru
1-Nov-2020	2	2526.641	3	0	70744	12	94	2141
2-Nov-2020	2	2556.059	3	0	70223	11	90	2090
3-Nov-2020	2	2585.1	3	0	69697	11	86	2039
4-Nov-2020	2	2613.751	3	0	69166	10	83	1990
5-Nov-2020	1	2641.999	3	0	68631	10	80	1942
6-Nov-2020	1	2669.835	2	0	68091	10	76	1894
7-Nov-2020	1	2697.245	2	0	67547	9	73	1848
8-Nov-2020	1	2724.221	2	0	66999	9	70	1803
9-Nov-2020	1	2750.752	2	0	66448	9	68	1758
10-Nov-2020	1	2776.827	2	0	65894	8	65	1715
11-Nov-2020	1	2802.437	2	0	65337	8	62	1672
12-Nov-2020	1	2827.574	2	0	64778	8	60	1631
13-Nov-2020	1	2852.229	2	0	64216	7	57	1590
14-Nov-2020	1	2876.394	1	0	63652	7	55	1550
15-Nov-2020	1	2900.06	1	0	63086	7	53	1512
16-Nov-2020	1	2923.221	1	0	62518	7	51	1474
17-Nov-2020	1	2945.869	1	0	61949	6	49	1437
18-Nov-2020	0	2967.999	1	0	61379	6	47	1401
19-Nov-2020	0	2989.603	1	0	60808	6	45	1365
20-Nov-2020	0	3010.677	1	0	60237	6	43	1331
21-Nov-2020	0	3031.215	1	0	59665	5	41	1297
22-Nov-2020	0	3051.212	1	0	59092	5	40	1264
23-Nov-2020	0	3070.663	1	0	58520	5	38	1232
24-Nov-2020	0	3089.565	1	0	57948	5	37	1200
25-Nov-2020	0	3107.914	1	0	57376	5	35	1170
26-Nov-2020	0	3125.707	1	0	56804	5	34	1140
27-Nov-2020	0	3142.94	1	0	56233	4	32	1110
28-Nov-2020	0	3159.612	1	0	55663	4	31	1082
29-Nov-2020	0	3175.719	1	0	55094	4	30	1054
30-Nov-2020	0	3191.261	1	0	54527	4	28	1027

Table 7: Predictions for November 2020.

References

1. Little N (2020) COVID-19 Tracker Canada. COVID19Tracker.ca.
2. Chowell G (2017) Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts. Infect Dis Model 2: 379-98.
3. Viboud C, Simonsen L, Chowell GA (2016) Generalized-growth model to characterize the early ascending phase of infectious disease outbreaks. Epidemics 15: 27-37.
4. Chowell G, Hincapie-Palacio D, Ospina J, Pell B, Tariq A, et al. (2016) Using Phenomenological Models to Characterize Transmissibility and Forecast Patterns and Final Burden of Zika Epidemics. PLoS Curr 31: 8.
5. Chowell G, Tariq A, Hyman JM (2019) A novel sub-epidemic modeling framework for short-term forecasting epidemic waves. BMC Med 17: 1-18.
6. Chowell G, Luo R, Sun K, Roosa K, Tariq A, et al. (2020) Real-time forecasting of epidemic trajectories using computational dynamic ensembles. Epidemics 30: 100379.
7. Cohen J (2020) Scientists are racing to model the next moves of a coronavirus that's still hard to predict. Science.
8. Batista M (2020) Estimation of the final size of the COVID-19 epidemic. medRxiv.
9. Roser M, Ritchie H, Ortiz-Ospina E, Hasell J (2020) Coronavirus Pandemic (COVID-19). OurWorldInData.org.
10. Hsu J (2020) Here's how computer models simulate the future spread of new coronavirus. Sci Am.
11. Anastasopoulou C, Russo L, Tsakris A, Siettos C (2020) Data-based analysis, modelling and forecasting of the COVID-19 outbreak. PLOS ONE 15: 0230405.
12. Maier BF, Brockmann D (2020) Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China. Science. 368: 742-746.
13. Cassaro FAM, Pires LF (2020) Can we predict the occurrence of COVID-19 cases? Considerations using a simple model of growth. Sci Total Env 728: 138834.
14. Ceylan Z (2020) Estimation of COVID-19 prevalence in Italy, Spain, and France. Science of the Total Environment 729: 138817.
15. Salehi M, Arashi M, Bekker A, Johan Ferreira JT, Chen D, et al. (2020) A synergetic R Shiny portal to track COVID-19 demographic information. Data Science Journal.

16. Sauer N (2020) Logistic growth and immunity.

17. Petropoulos F, Makridakis S (2020) Forecasting the novel coronavirus COVID-19. PLOS ONE 15: 0231236.

18. Anderson KV, Daniewicz SR (2018) Statistical analysis of the influence of defects on fatigue life using a Gumbel distribution. Int J Fatigue 112: 78-83.

19. Gomez YM, Bolfarine H, Gomez HW (2019) Gumbel distribution with heavy tails and applications to Environmental data. Math Comp Sim 157: 115-129.

20. Hyun N, Couper DJ, Zeng D (2019) Gumbel regression models for a monotone increasing continuous biomarker subject to measurement error. J Stat Plann Inf 203: 160-168.

21. Huang P, Hu F, Dong F (2019) Parameter estimation of Gumbel distribution and its application to pitting corrosion depth of concrete girder bridges. Cluster Comp 22: S3405-S3411.

22. Panovska-Griffiths J (2020) Can mathematical modelling solve the current Covid-19 crisis? BMC Public Heath 20: 551.

Citation: Yoo K, Arashi M and Bekker A (2020) Pitting the Gumbel and logistic growth models against one another to model COVID-19 spread. J Health Sci Dev Vol: 3, Issu: 2 (17-30).