

A Global Scale Estimate of Novel Coronavirus (COVID-19) Cases Using Extreme Value Distributions

M. Aadhityaa¹, K. S. Kasiviswanathan^{2*}, Idhayachandhiran Ilampooranan²,

B. Soundharajan¹, M. Balamurugan³, Jianxun He⁴

¹

² Department of Water Resources Development and Management, Indian Institute of Technology Roorkee, India

³ Department of Civil and Environmental Engineering, National University of Singapore, Singapore

⁴ Department of Civil Engineering, Schulich School of Engineering, University of Calgary, Calgary, Canada

* k.kasiviswanathan@wr.iitr.ac.in (KSK)

Abstract

The COVID-19 pandemic has created a global crisis and the governments are fighting rigorously to control the spread by imposing intervention measures and increasing the medical facilities. In order to tackle the crisis effectively we need to know the trajectories of number of the people infected (i.e. confirmed cases). Such information is crucial to government agencies for developing effective preparedness plans and strategies. We used a statistical modeling approach – extreme value distributions (EVDs) for projecting the future confirmed cases on a global scale. Using the 69 days data (from January 22, 2020 to March 30, 2020), the EVDs model predicted the number of confirmed cases from March 31, 2020 to April 9, 2020 (validation period) with an

23 absolute percentage error < 15 % and then projected the number of confirmed cases until the end
24 of June 2020. Also, we have quantified the uncertainty in the future projections due to the delay
25 in reporting of the confirmed cases on a global scale. Based on the projections, we found that
26 total confirmed cases would reach around 11.4 million globally by the end of June 2020. The
27 USA may have 2.9 million number of confirmed cases followed by Spain-1.52 million and Italy-
28 1.28 million.

29
30 **Keywords:** COVID-19, Statistical modeling; Extreme value distributions; Future projection;
31 Confirmed cases.

32 **Introduction**

33 The first case of respiratory disease, pneumonia, with symptoms similar to the severe acute
34 respiratory syndrome coronavirus (SARS-CoV) was reported in Wuhan City, China in December
35 2019 [1]. A novel strain of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was
36 confirmed on January 7, 2020 [1, 2]. The novel corona virus (COVID-19), which is the seventh
37 member of the coronavirus family, along with the SARS-CoV and the middle east respiratory
38 syndrome coronavirus (MERS-CoV) spread from animals to humans [2]. Since the reporting of
39 the zero-patient in December 2019, COVID-19 has spread dramatically worldwide and the
40 World Health Organization (WHO) has declared the outbreak as Public Health Emergency of
41 International Concern (PHEIC) on January 30, 2020 [2]. As of April 9, 2020, globally the total
42 number of confirmed, recovered, and mortality cases were 1,595,350, 353,975 and 95,455
43 respectively [3]. The reported cases globally are drastically increasing from 4,219/day in
44 February 2020 to 50,784/day in early April 2020. The trajectory of the infection spread during
45 the coming days is important information in order to plan, prepare, and scale up the intervention

46 measures including the medical facilities to meet the increased influx of patients and other
47 governing policies to control the transmission.

48
49 To date (as of April 9, 2020), the United States of America (USA) has the highest number of
50 affected cases (461,437), followed by Spain (153, 222) and Italy (143,626) [3]. The global case-
51 fatality rate (CFR), which is the ratio of the confirmed deaths to the confirmed cases, has
52 increased from 1.37% on January 19, 2020 to 5.95% by April 9, 2020 [4]. In contrast, the current
53 CFR of China, Japan, and Iran has either reduced or remained constant when compared to their
54 initial CFR [4]. The global CFR is however increasing continuously due to the different timing
55 of the onset of the pandemic in different countries. Overall, the number of confirmed cases has
56 been explosively increasing with time so far.

57
58 Modeling tools have been widely used to predict the COVID-19 spread to help the medical
59 professionals, policymakers, and governing bodies for implementing interventions measures to
60 control the pandemic. Since the onset of COVID-19, studies have used different mathematical
61 and dynamic stochastic transmission models to predict the transmission and intervention impacts
62 [5, 6, 7, 8, 9, 10, 11]. However, these epidemiological models involve a number of parameters
63 that are not readily available due to the absence or lack of the data for extracting the knowledge
64 especially during the early period of the outbreak. These parameters have been either assumed or
65 adopted from previous pandemic studies and consequently the performance of these models has
66 been questioned [1, 11].

67

68 To address the above shortcomings of epidemiological models, we proposed a statistical
69 modelling approach, using Extreme Value Distributions (EVD) to describe the evolution of the
70 COVID-19 spread and then to generate the future projections. EVDs are used to fit series of
71 observation mainly to estimate extreme events of future that were not observed in the past.
72 Though application of EVDs are very common in earth sciences to model the unusual events,
73 recently, EVDs are applied in health care sector and shown to produce promising results [12].

74
75 One of the main challenges in simulating and projecting the COVID-19 spread trajectory is the
76 delay in reporting (R_d) which is the lag time between onset of symptoms of the disease and date
77 of reporting [13]. The delay in reporting may vary due to various reasons such as delay in (i)
78 reporting at the hospital, (ii) diagnostics, (iii) reporting the confirmed cases in databases etc.
79 Also, R_d imposes high uncertainty in estimating the spread trajectory and thus excluding R_d in
80 modeling analysis could lead to unrealistic projections with underestimation in the projected
81 confirmed cases [9]. Very few studies have considered R_d in their modeling studies to estimate
82 the transmission dynamics of COVID-19 and reported an average R_d value of 7.6 days and 6.1
83 days [9,13].

84
85 Therefore, for quantifying the uncertainty in projected cases, we have estimated the fold increase
86 in the confirmed cases due to R_d . Thus, the key contributions of the study are (i) using EVD
87 theory for the projection of COVID-19 cases, (ii) incorporating R_d value to estimate the
88 uncertainty in future projections, and (iii) global scale projection of confirmed and death cases.

89

90 **Materials and Method**

91 We collected the daily time series of the number of confirmed and death cases from John
92 Hopkins University Center for Systems Science and Engineering [3] for 177 countries, out of
93 which only 42 countries (refer S1 table) that exceeded 1000 confirmed cases (as on March 30,
94 2020) were considered for the analysis. These 42 countries spread across different continents
95 except Antarctica and accounts for 96.5% of the total confirmed cases globally as on March 30,
96 2020. We observed that majority of the countries with significant number of confirmed cases are
97 from Europe and Asia followed by North America and South America.

98 **S1 Table: Total number of confirmed COVID 19 cases as on March 30, 2020 (List of**
99 **countries short listed based upon a minimum threshold of 1000 confirmed cases)**

100

101 Though the number of COVID-19 infected cases were reported even before January 22, 2020 in
102 China especially in the Hubei province, we have not considered that data in the analysis since the
103 COVID-19 outbreak has been contained and extensive research studies have already been
104 conducted [14, 15, 16, 17].

105

106 Application of extreme value distribution has already been explored to model the mortality and
107 morbidity rate associated with pneumonia, influenza and cardiovascular diseases in the public
108 health planning [12, 18]. Therefore, in this paper, we have explored the applicability of EVDs in
109 modeling the confirmed COVID-19 cases. Initial statistical analysis of the data revealed that the
110 critical stage of COVID-19 outbreak largely has no trend in the number of people being infected
111 and therefore the use of EVDs are justified. Among the EVDs, three-parameter distributions such
112 as Generalized Extreme Value (GEV), Generalized Pareto (GP), and Generalized Likelihood

113 (GL) distributions were explored. The parameters of these distributions mainly define the
114 characteristics such as scale, shape and location of the data being fitted using the EVDs. The tail
115 behaviour of the distribution is described by the shape parameter which is estimated from higher
116 order moments, and precise estimation of shape parameter is often computationally difficult [19,
117 20] and requires suitable moment estimation approaches. Among several methods (i.e. the
118 method of likelihood and the probability weighted moments) for estimating the distribution
119 parameters, the L-moment method has been demonstrated to be more effective in estimating the
120 shape parameters and hence used in this study [21]. As many existing literatures elaborately
121 describe the mathematical description about the extreme value distributions and L-moment
122 methods, the detailed explanations are not provided in this paper.

123
124 Conventionally, EVDs have often been applied in the extreme statistical analysis, which
125 estimates the quantities corresponding to specific return periods or probabilities. In this analysis,
126 the sample data from the population are expected to be independent and identically distributed.
127 In the extreme statistical analysis for natural extreme events such as flooding, earthquake, and
128 tsunami, the annual maximum values are often considered. We observed that in the case of
129 COVID-19, the daily recorded confirmed cases are independent with no trend. Thus, the reported
130 confirmed COVID-19 cases were fitted using the EVDs to project the number of future
131 confirmed cases.

132
133 In general, the reported cases on any given days were lower than the actual infected cases due to
134 various reasons including the delay in the onset of acute symptoms, inefficiency in the testing
135 methods, lack of sufficient testing facilities etc. There is also significant risk of Covid-19 patients

136 to be tested positive after initially being tested negative due to the inaccuracies in testing and
137 latent symptoms [22]. However, government authorities mainly health care professionals should
138 be aware and be informed about the discrepancy between the reported and actual infected cases
139 to effectively tackle the current COVID-19 situations. Thus, we also estimated the fold increase
140 in the number of confirmed cases due to the delay in reporting. To account the effect of the delay
141 in reporting on the confirmed cases, we considered a Reporting delay (R_d) of 1 to 7 days and
142 have proposed a simple statistical lagging approach to estimate the fold increase in the number of
143 confirmed cases. For this analysis, the ratio $\left[\frac{C_t}{C_{t-lag}} \right]$ of confirmed cases on the current date C_t to
144 the previously lagged date C_{t-lag} were computed using the complete data having n data points.
145 The mean value of the fold increase was estimated from the ratios (For example, lag of 5 days
146 will have $n-5$ number of ratios) for each R_d . Furthermore, the estimate of the fold increase (with
147 its uncertainty) was coupled with the future projection of the confirmed cases for quantifying
148 associated uncertainty in the projection. Note that other than R_d , no other sources of uncertainty
149 such as incubation period, communal spread, the effects of lockdown or other containment
150 strategies, healthcare capacity, etc. were included in our study.

151 **Results and discussion**

152 **Statistical tests and model performance**

153 The daily number of the confirmed cases from the selected 42 countries were computed from the
154 reported cumulative data. These data were further processed with the modified Mann-Kendall
155 test to check for the presence of non-parametric trend and we found that there is no trend in the
156 entire dataset. This proves that the data are statistically independent and identically distributed

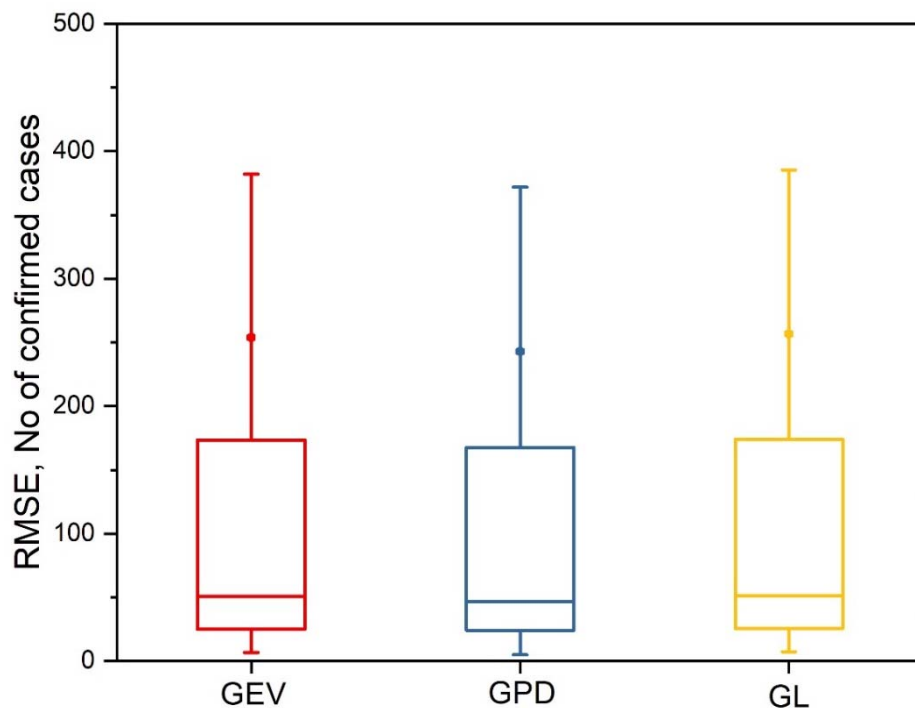
157 during the critical stage of pandemic situation. Hence, we applied extreme value distribution to
158 model the available data as well as to project the COVID-19 cases.

159

160 As mentioned earlier, three different EVDs were explored to fit the datasets for the numbers of
161 confirmed cases. The root mean squared error (RMSE) computed for each distribution against
162 observation for all datasets are plotted in the boxplot (Fig 1). It is evident from Fig 1 that all the
163 three distributions (GEV, GP and GL) performed equivalently. Furthermore, the fitting
164 performance was slightly improved when using the GP distribution compared to GEV and GL
165 distributions and in particular the GP performed consistently well across all the datasets. In
166 addition, large variations in the estimated distribution parameters (i.e., location, scale, and shape
167 parameters) were identified. The results of parameter variations of GP have been shown in S2
168 Fig. These variations would reflect the variations in the statistical characteristics of the datasets
169 of different countries. As these models are data specific with parameters not having direct
170 physical meaning, it is hard to link the behaviour of parameters with the modelled variables.
171 Since we observed better performance and lower RMSE using GP distribution, in this study, we
172 are projecting the estimates of confirmed cases for selected countries using the GP models.

173 **S2 Fig: Estimated parameters of the GP Models**

174



175

176 **Fig 1. Performance of three different Extreme value probability distributions for fitting the**
177 **confirmed cases of all 42 countries**

178

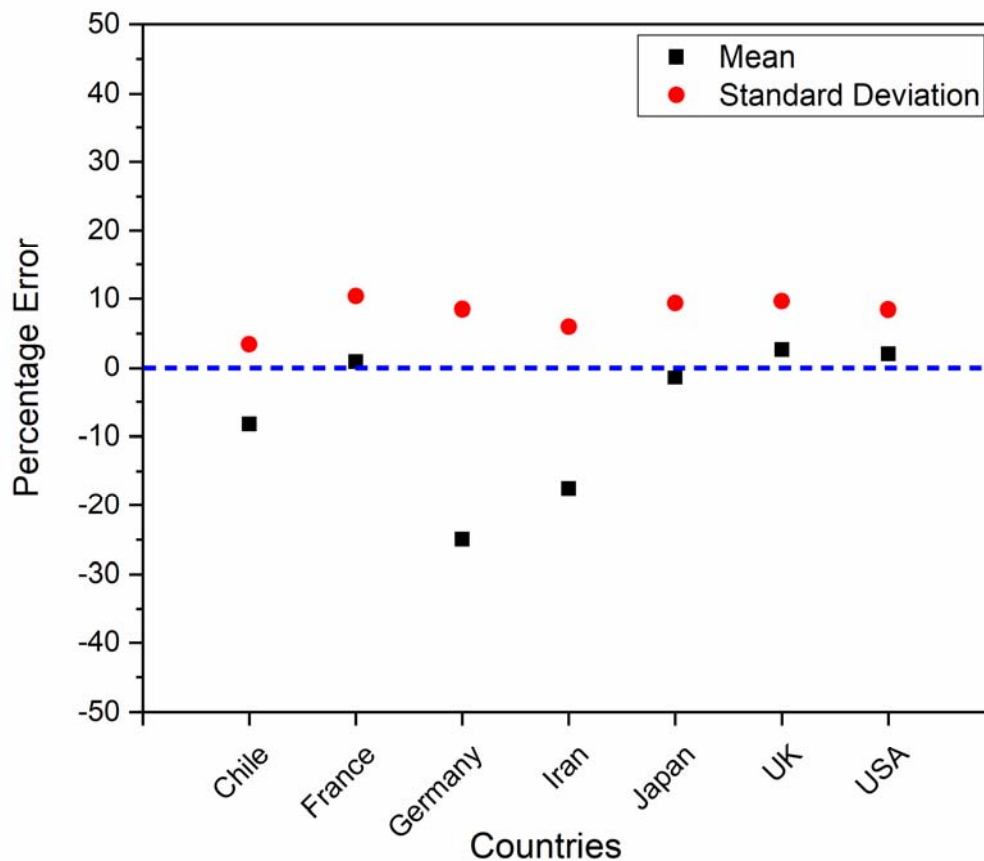
179 Although the outbreak had started in the beginning of January 2020 in Wuhan, China, majority
180 of the countries started experiencing new cases in the beginning of March 2020. All these
181 demonstrates the different timing of the onset of the spread in different countries. However, we
182 considered same period (January 22, 2020 to March 30, 2020) for the model calibration which in
183 turn resulted a uniform data length of all the selected countries for estimating the parameters of
184 distribution.

185

186 The model performance has been validated by comparing the model's projections and the
187 confirmed cases observed for the selected countries for the period of 10 days (March 31 – April
188 9, 2020). The mean and standard deviation of the resulting percentage error (i.e. ratio of
189 difference between observed minus projected to observed cases) has been shown in Fig 2. Please

190 note that the positive and negative value of mean of the percentage error indicates under and
191 overestimate of the projected value, respectively. From Fig 2, it is evident that the mean and
192 standard deviation of the percentage error are within the $\pm 5\%$ and $\pm 10\%$ respectively for most of
193 the countries, and the overall performance of model is quite satisfactory for majority of the cases
194 except few. The poor performance of model for few countries, for example Germany, Australia,
195 and Iran, might be due to high variation in the infected cases. As more data is available in future,
196 more critical validation of these models can be performed to bring additional insights on the
197 reliability of the model projection.

198



199

200 **Fig 2 Percentage error in the projected confirmed cases for the validation period from**
201 **March 31, 2020 to April 09, 2020**

202 Further, the model was also validated using the projection at global scale. As shown in Table 1,
203 the projections of the confirmed cases are very close to actual value with slight over estimation
204 and the increasing trend is captured well. As mentioned earlier, more data is required to validate
205 the global scale long term projection of model. However, the long-term projection of deaths
206 estimated using these model projections are very close to the projection reported by IHME for
207 the countries such as United States of America (USA) and United Kingdom (UK) [23].

208

209 **Table 1. Global scale estimates of confirmed cases for validation period**

Date	Projected	Actual
March 31, 2020	869,406	857,487
April 1, 2020	957,136	932,605
April 2, 2020	1,045,551	1,013,320
April 3, 2020	1,134,644	1,095,917
April 4, 2020	1,224,412	1,197,405
April 5, 2020	1,314,850	1,272,115
April 6, 2020	1,405,953	1,345,101
April 7, 2020	1,497,716	1,426,096
April 8, 2020	1,590,135	1,511,104
April 9, 2020	1,683,203	1,595,350

210

211 **Projection of COVID-19 confirmed cases**

212 We observed two main behavioural changes in the number of confirmed cases curves (Fig 3). (i)
213 plateau-peak shift: number of confirmed cases in the USA, Australia, and Italy has plateaued
214 (with approximately 10 cases) soon after the beginning until last week of February 2020 and
215 increased to peak thereafter. A possible reason could be an inflow of passengers soon after the
216 Chinese New Year [24]. (ii) cross-over during March 2020: though the number of confirmed
217 cases around end of February 2020 (approximately 10 cases) was almost similar for the USA,
218 Australia, Italy, and Iran, Australia's number of confirmed cases were less when compared with

219 the USA, Italy, and Iran during March 2020. The USA's number of confirmed cases, though
220 lesser than Italy and Iran during the first three weeks of March 2020, has surpassed the cases of
221 Italy and Iran during the end of third week of March 2020 and stayed at peak thereafter. The
222 possible explanation for this behaviour could be attributed to the difference in intervention
223 measures imposed in the respective countries. However, it is important to note that, testing
224 capacity might be one of the keys which dictate the success of non-medical measures such as
225 social-distancing and lockdown to contain the virus [25].

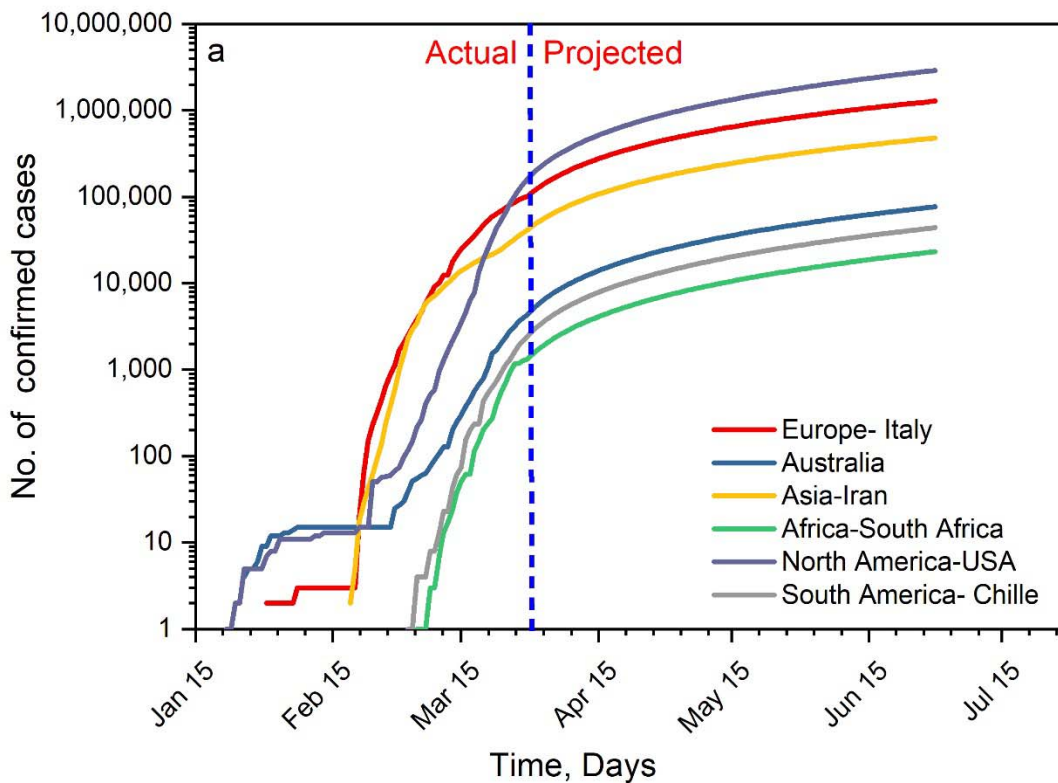
226
227 Note that due to the limited observed data for fitting the model, the error in the future projection
228 is expected to be high and increases with time. Thus, the projection was made till June 30, 2020
229 (three months from March 30, 2020). Along with the estimate of probability, the shape, scale and
230 location parameters of GP distribution were used to project the confirmed cases for the future
231 period. Fig 3 shows the estimate of the number of confirmed cases for the present and future (up
232 to June 30, 2020) for the most affected country of each continent. In Chile and South Africa,
233 during the first week of March 2020, fewer (i.e. around 10 cases) confirmed cases were reported;
234 then an exponential increase was clearly seen during the subsequent weeks. Although number of
235 daily confirmed cases show randomness, the varying pattern across different countries were
236 similar.

237
238 As of June 30, 2020, the numbers of the confirmed cases for Italy, Iran, Australia, South Africa
239 and Chile would reach around 1,281,708, 479,531, 76,795, 23,281, and 44,041 respectively. The
240 number of the confirmed cases for the USA would likely to be at least more than one million
241 (highest among all the countries) in the early May 2020, though the initial progression of the

242 number infected was much lower than other highly affected countries such as Italy and Iran.
243 Similar behaviour in the number of projected confirmed cases was observed for the other less
244 affected countries such as India (refer S3 Fig), however, with less magnitude mainly because of
245 delay in onset of disease spread (early March 2020) and reporting of cases. Majority of the
246 countries (mostly developing and under-developed) at the onset of surging increase in COVID-
247 19 spread would neither be equipped with testing facilities nor have the medical infrastructure to
248 tackle the crisis [26]. Therefore, the availability of more data in the forthcoming days would help
249 in producing a more reliable projection of the confirmed cases in the future.

250 **S3 Fig: Projection of confirmed cases for the countries with delayed onset of COVID-19**

251



252

253 **Fig 3. Actual and projected trend of cumulative increase of confirmed cases for the selected**
254 **countries**

255
256 Note that these projections did not explicitly include the effect of the actual stringent control
257 measures (eg. social distancing, travel bans, isolation/quarantine, lockdown) adopted in the
258 various countries at local/regional and national level. However these projections might vary and
259 the number of actual confirmed cases could be less if all the countries apply inter and intra
260 circuit breaking measures (control measures) to reduce the COVID-19 spread. From the
261 projection across the world on June 30, 2020 as illustrated in Fig. 4, 17 thousand (17k) to 3000
262 thousand (3000k) number of confirmed cases were observed across different countries since
263 January 22, 2020. The maximum number of confirmed cases between 1522k and 2906k were
264 observed in the USA and Spain. Following that Italy and Germany are likely to have more than 1
265 million confirmed cases.

266
267 Globally the total number of confirmed cases would reach 11.4 million by the end of June 2020.
268 Several countries will exceed one million COVID-19 infections within the next two months. For
269 example, USA with the highest number of confirmed cases globally will be the first to reach the
270 one million count in the first week of May, followed by Spain and Italy in the first week of June
271 2020. We also estimated that Germany and France will also exceed a million cases in the middle
272 and end of June respectively. Countries like Iran, UK, Chile and Portugal are also identified to be
273 at high risk since confirmed cases in these countries will exceed half a million by the end of June
274 2020.

275

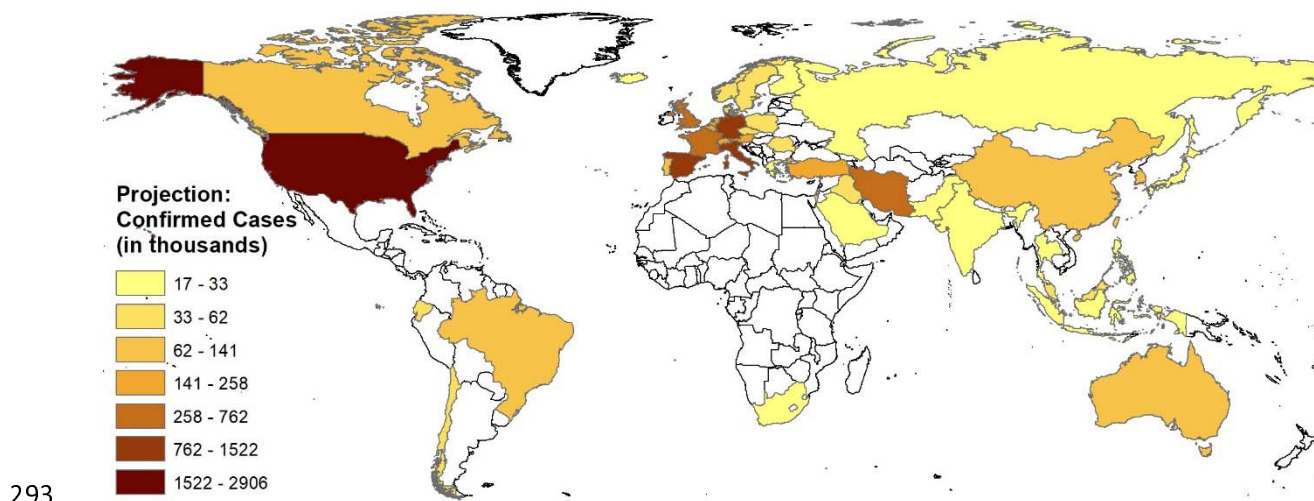
276 The current number of the confirmed cases in India is still in the range of few thousands and the
277 future projection is estimated to be around 28,028, which is considerably lesser than the USA
278 and the UK. However currently in India, stringent measures such as 42 days nation-wide
279 lockdown has been imposed to control and stabilize the communal spread in order to prevent
280 from becoming a global hotspot of COVID-19. It is difficult to project realistic estimates for
281 countries like India and Indonesia due to lack of sufficient data which is attributed to the delay in
282 onset of COVID-19 (first week of March 2020). However, upon getting more data with time, our
283 model could be used to project more realistic values.

284
285 In other countries including Japan and South Korea that were severely affected but imposed
286 many preventive measures, would likely to have 23k and 118k confirmed cases respectively.
287 However, the number of confirmed cases might still be reduced depending on the effectiveness
288 of the preventive measures. Please refer S4 Table for the country-wise estimates of projected
289 confirmed cases.

290 **S4 Table: Country wise estimates of projected confirmed cases**

291

292

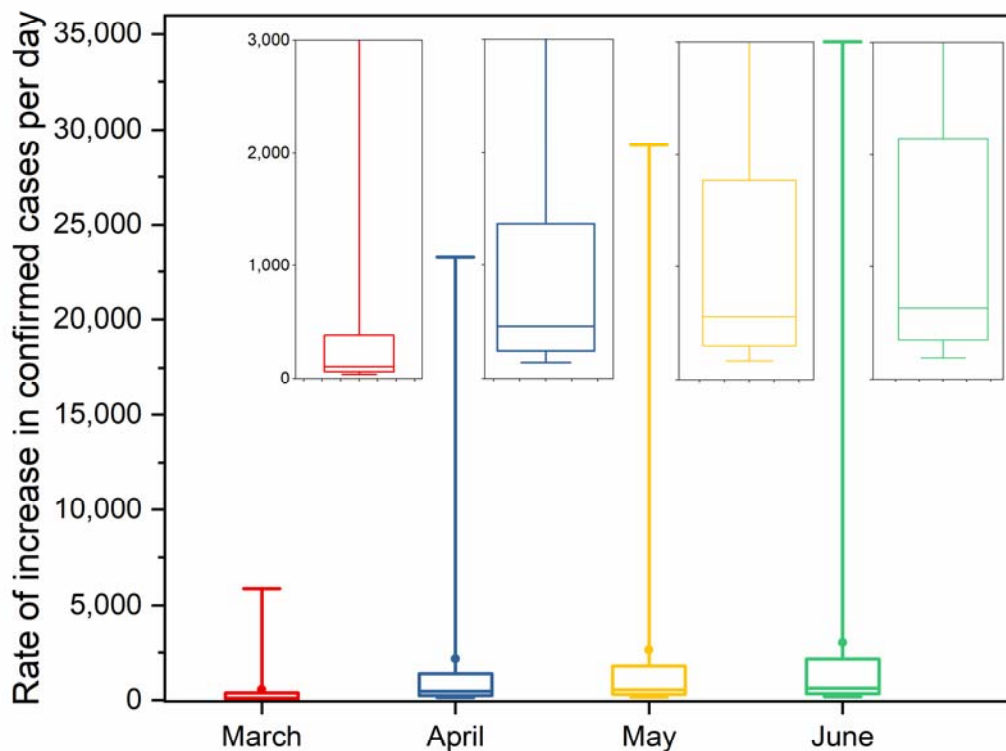


293

294 **Fig 4. Number of confirmed cases as on June 30, 2020**

295

296 The rate of increase in the projected confirmed cases were estimated at the end of each months
297 (i.e. April, May and June 2020) including the actual confirmed case (March). This daily rate of
298 confirmed cases was computed from estimating the difference between the cumulative confirmed
299 cases at each month interval and dividing by the total number of days (one month in this case)
300 using the projection of 42 countries. The rate of increase falling in the box (i.e. 25 to 75
301 percentiles) indicates that in several countries, the impact would be less as the number of
302 confirmed cases per day ranges between few thousand for the projection period varying from
303 April 2020 to June 2020. However, a very high rate of increase in confirmed cases were found
304 in USA, Spain, Italy and Germany followed by France, Iran and UK.



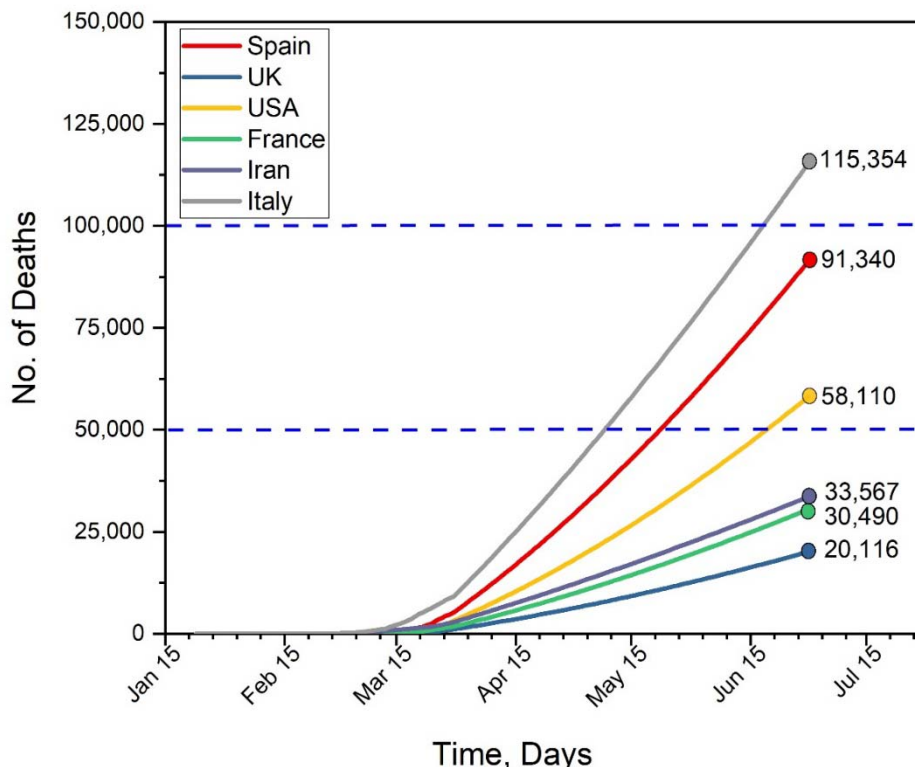
305
306 **Fig 5. Rate of increase in confirmed cases per day (box plot illustrates variation in the rate**
307 **from the data of 42 countries)**

308
309 Besides the projection of confirmed cases, we also estimated the likely number of deaths. CFR is
310 commonly used to estimate the risk of death due to any infectious disease. Please note that
311 though CFR (usually represented as percentage of the ratio of confirmed cases to confirmed
312 deaths) is not constant and changes with the context (e.g., it can vary with time, age and the
313 characteristics of infected population, etc.)[4], it can give an approximate estimate of the number
314 of deaths [4, 27, 28, 29].

316 We estimated the average value of the CFR for the selected 42 countries (from March 15, 2020
317 to March 30, 2020) and found the CFR values as 9%, 8%, 7%, 6%, 5% and 2% for the countries
318 Italy, Indonesia, Iran, Spain, UK, France and USA respectively. Based on these CFR estimates
319 along with the model projection for the confirmed cases (Fig 6), we estimated the likely number
320 of deaths on June 30, 2020 and found the number of deaths to be highest in Italy (115,354) and
321 Spain (91,340). The number of deaths in countries such as USA, Iran, France and UK are also
322 likely to be high with 58,110, 33,567, 30,490 and 20,116 deaths respectively by the end of June
323 2020. Refer S5 Table5 for the estimated number of deaths of other countries having CFR greater
324 than 1 percent. Our estimate especially for USA (82,638 deaths by the end of July 2020) is very
325 close with the number of deaths projected by the IHME health service utilization forecasting
326 team (81,114 deaths) [12].

327 **S5 Table: Estimated deaths for selected countries based on the projection of confirmed**
328 **cases and average CFR value higher than 1% on June 30, 2020**

329



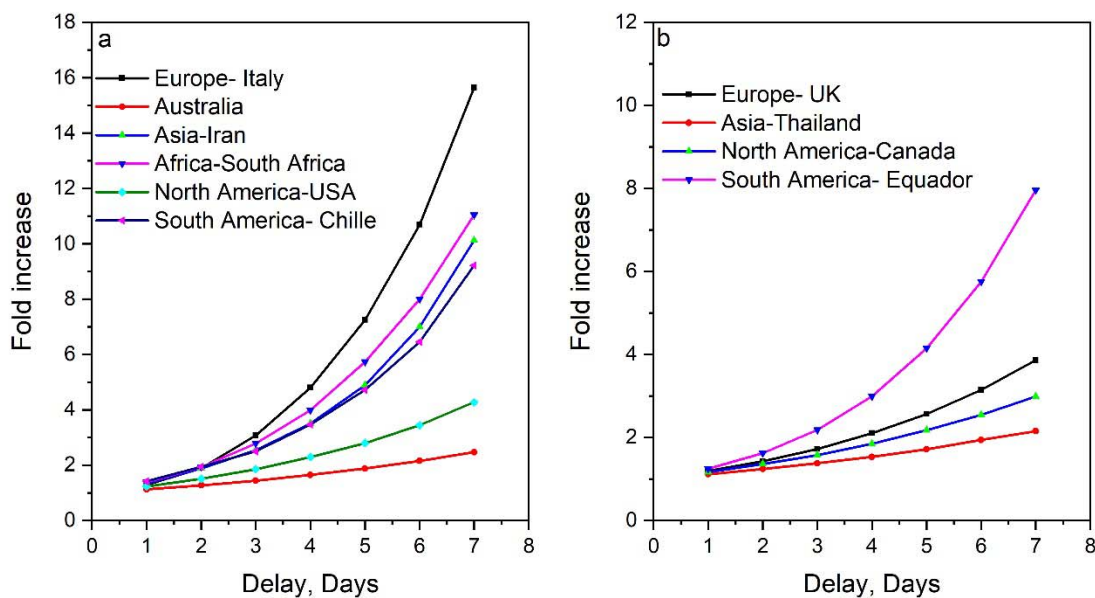
330

331 **Fig 6. The estimated deaths in selected countries based the results of CFR and projection of**
332 **confirmed cases**

333 **Estimating fold increase due to the delay in reporting confirmed** 334 **cases**

335 Effective lag length is a key variable to estimate the fold increase due to delay in reporting. As
336 different countries follow different testing procedure and also the capacity of health care systems
337 largely varies between the countries, the results of confirmed cases on any day is lower than the
338 actual number of people infected. We varied the minimum and maximum lag length of 1 and 7
339 days respectively to analyse the impact of delay on number of confirmed cases. Fig 7 is plotted
340 between the number of days delayed and the number of folds increase in the confirmed cases for
341 the selected countries across the world. The mean estimate of fold increase was calculated for
342 each day lag (Fig 7). It is well known as illustrated in Fig 7 that increasing the number of days
343 delay elevates the magnitude of fold increase. The fold increase of 16 and 10 would reach for the

344 delay of 7 days in the context of extreme scenario as currently Italy and USA respectively are
345 facing (Fig 7a). In specific, a steep increase was found when the lag length (i.e. delay in
346 reporting) is more than 5 days. Therefore, it is to be noted that sooner the case is identified and
347 reported, better the preventive measures could be ensured without much communal spread [30].
348 Further, we noticed that although USA is experiencing high surge in confirmed cases, the fold
349 increase for even 1-week delay was considerably less compared to other countries. Perhaps this
350 observation is likely to change as more data will be available. This is a positive point that USA
351 can manage the situation if adequate preventive measures are in place. Overall, it was observed
352 that many of the European countries exhibited similar results. As shown in Fig 7b, though the
353 magnitude of fold increase seems comparatively low in less affected countries, it might increase
354 when the number of confirmed cases increases. Therefore, this is a right time for them to enforce
355 preventive measures to safeguard people from COVID-19.



356

357

358 **Fig 7. Fold increase in number of confirmed cases for different lag length a) severely**
359 **affected b) less affected**

360

361 **Assumptions, limitations, and quantification of uncertainty**

362 The uncertainties associated with the epidemic modeling studies are due to various reasons such
363 as (i) availability, length, and correctness of the data [31], (ii) model parameter values: either
364 assumed or estimated or adopted from previous modeling studies (assumption of incubation
365 period and reproduction number) [32, 33] (iii) assumptions or limitations of the model being
366 used. For example, the SEIR model assumes all the population is susceptible to infection.
367 Dynamics transmission model's assumption that symptomatic individuals are more (50%)
368 susceptible to infection than asymptomatic individuals [32]. Assumptions while conceptualizing
369 the non-pharmaceutical interventions such as duration of stay at home during isolation, percent
370 contact reduction in workplaces, impact of non-pharmaceutical interventions are constant with
371 time and same across all countries etc. [32].

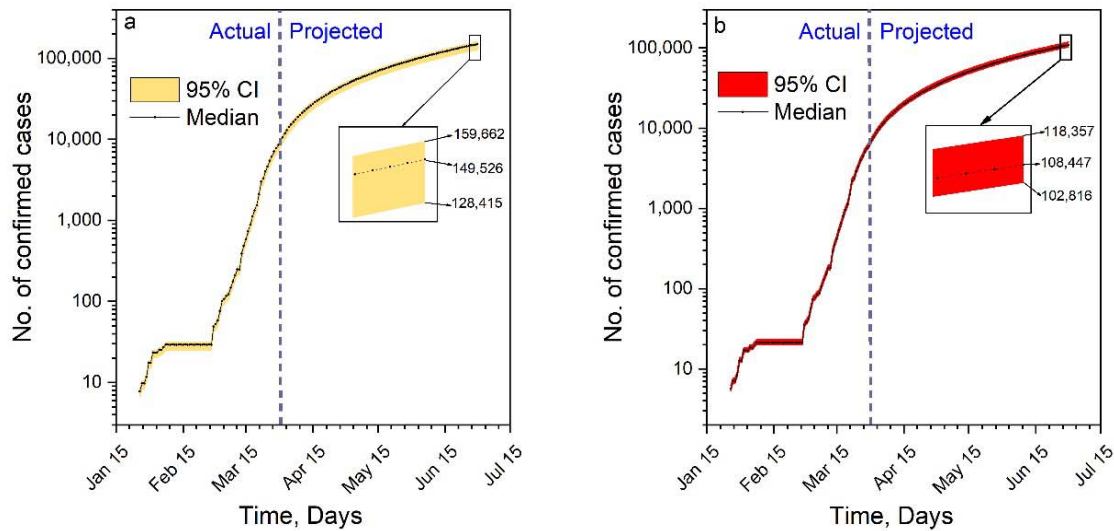
372

373 As mentioned earlier, the uncertainty in the projected confirmed cases were quantified only
374 based on the R_d , and the fold increase estimated from each data point for a fixed lag length will
375 have inherent variability (Fig. 8). This is mainly because the data is highly random and during
376 the initial phase of pandemic the effect of delay will be less and gradually increase with time.
377 Although the mean estimate as reported in the previous section is a good choice to quantify the
378 delay effect in the projection, ignoring the uncertainty might under predict the likely estimate of
379 future period. Therefore, we estimated 95% confidence interval from the estimate of fold
380 increase. From the lag length period of 1 to 7 days, we considered on an average of 3- and 5-days

381 lag based on the delay in reporting to estimate the range of variation in the projection especially
382 in the confirmed cases. We chose Australia, randomly, to illustrate the impact of uncertainty in
383 the projection (Fig 8). It is evident from Fig 8 that the median projected confirmed cases and
384 uncertainty increases with time. For example, the median projected confirmed cases increased
385 from 6,958 on March 30 2020 to 108,447 on 30 June 2020 for R_d of 3 days. Similarly, the width
386 of uncertainty band for projected confirmed cases on 31 March 2020 was 6,596-7,593 (upper
387 bound – lower bound) while it has increased to 102,816-118,357 on June 30 2020. Refer Table 2
388 for median and CI intervals for selected seven countries.

389
390 It is expected that the unbiased estimate of uncertainty band will have the median estimate closer
391 to the mid-portion of the band. In other words, any deviation from this mid-portion of the band
392 represents the bias in the uncertainty estimate. However, we found the median falling towards
393 the upper bound in most of the countries mainly due to the drastic increase in number of cases
394 with during the critical period. As shown earlier in Fig 7, the delay in reporting increases the
395 fold-increase in confirmed cases which in turn will significantly increase the projection range as
396 well as the uncertainty (Table 2).

397



398

399 **Fig 8. Actual and projected confirmed cases with uncertainty band for a) 5 days delay and**
 400 **b) 3 days delay for Australia**

Table 2. Projection of confirmed cases with uncertainty in Reporting delay of 3 and 5 days respectively for selected countries

401

S.No	Country	Continent	95% CI for $R_d=3$ days	95% CI for $R_d=5$ days
1	Australia	Australia	102,816-118,357	128,415-159,662
2	Chile	South America	95,322-124,776	167,684-247,728
3	Iran	Asia	836,231-1,602,385	1,371,580-3,314,668
4	Italy	Europe	1,496,111-6,387,634	2,866,670-15,697,496
5	South Africa	Africa	51,117-78,259	91,763-175,052
6	USA	North America	4,795,011-5,971,468	7,037,335-9,191,516

402

403 **Summary and limitations**

404 This study explored the (i) applicability of EVDs in predicting the COVID-19 confirmed cases,
 405 and (ii) possible relation between the delay in reporting the cases and the potential increase in the
 406 number of infection (number of confirmed cases). The results of the projection indicate that the

407 USA would have the highest number of confirmed cases of 2,905,522 (4,795,011-5,971,468 for
408 $R_d= 3$ days) and Iceland to have the minimum number of confirmed cases of 21,166 (27222-
409 58,008 for $R_d= 3$ days) by June 30, 2020. The number of deaths has also been estimated for the
410 42 countries and found that the deaths to be maximum in Italy (115,354 deaths) followed by
411 Spain (91,340) by June 30, 2020.

412
413 It may be noted that we have not considered any intervention measures (i.e. lockdown, social
414 distancing, school closures etc.,) in the model rather we focused on projecting the actual trend
415 exist in the data, thereby informing the likely increase in number of confirmed cases due to
416 COVID-19 outbreak. As inferred from this study, the reporting delay should be minimized to get
417 more accurate information on the confirmed cases. Therefore, the uncertainty due to delay in
418 reporting should not be ignored in the projection in order to estimate the reliable number of
419 confirmed cases. However, future studies will include other sources of uncertainty such as
420 model, parameter and input for the more realistic projection. The projected confirmed cases are
421 based on the data collected until March 30, 2020. However, we will be updating the model
422 projection once in every two weeks and our results will be posted in twitter handle
423 @Hydroviswa and @ IdhayaI.

424

425 **Reference**

- 426 1. Koo KR, Cook AR, Park M, Sun Y, Sun H, Lim JT. Et al. (2020) Interventions to
427 mitigate early spread of SARS-CoV-2 in Singapore: a modelling study. T. Lan. *Infec.*
428 *Dis.* doi: 10.1016/S1473-3099(20)30162-6

- 429 2. Bolddog P, Tekeli T, Vizi Z, Denes A, Bartha FA, Rost G. (2020) Risk Assessment of
430 Novel Coronavirus COVID-19 Outbreaks Outside China. *J. Clin. Med.* 9(2)-571. doi:
431 10.3390/jcm9020571
- 432 3. Data Repository by Johns Hopkins CSSE [Internet] -2019 Novel Coronavirus COVID-
433 19. [cited 2020 Apr 03] Available from:
434 [https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299](https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6)
435 [423467b48e9ecf6](https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6)
- 436 4. Roser M, Ritchie H, Ortiz-Ospine E. (2020) Coronavirus Disease (COVID-19) Statistics
437 and research. Oxford Martin School, Published online at OurWorldInData.org. [Cited
438 2020 April 03]. Available from: <https://ourworldindata.org/coronavirus>
- 439 5. Stehlé J, Voirin N, Barrat A, Cattuto C, Colizza V, Isella L. et al. (2011) Simulation of an
440 SEIR infectious disease model on the dynamic contact network of conference attendees.
441 *BMC Med.* 9-87. doi: 10.1186/1741-7015-9-87.
- 442 6. Tuite AR, Fisman DN. (2020) Reporting, Epidemic Growth, and Reproduction Numbers
443 for the 2019 Novel Coronavirus (2019-nCoV) Epidemic. *Annals of Internal Medicine.*
444 doi: 10.7326/M20-0358.
- 445 7. Hu Z, Ge Q, Li S, Jin L, Xiong M. (2020) Artificial Inteligence Forecasting of Covid-19
446 in China. *q-bio.OT.* Avaialble From: <https://arxiv.org/abs/2002.07112>
- 447 8. Wu JT, Leung K, Leung GM. (2020) Nowcasting and forecasting the potential domestic
448 and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a
449 modelling study. *The Lancet.* 395(10225):689-8. doi: [10.1016/S0140-6736\(20\)30260-9](https://doi.org/10.1016/S0140-6736(20)30260-9).

- 450 9. Kucharski A, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S. (2020) Early
451 dynamics of transmission and control of COVID-19: A mathematical modelling study. *T.*
452 *Lan. Infect. Dis.* doi: 10.1016/S1473-3099(20)30144-4
- 453 10. Sen M, Peng Z, Xiao Y, Zhang L. (2020) Modelling the epidemic trend of the 2019 novel
454 coronavirus outbreak in China. *bioRxiv* [Preprint]. 2020 [Cited 2020 April 5]. Available
455 from: <https://www.biorxiv.org/content/10.1101/2020.01.23.916726v1.article-info>. doi:
456 10.1101/2020.01.23.916726
- 457 11. Zhao S, Musa SS, Lin Q, Ran J, Yang G, Wang W. (2020) Estimating the unreported
458 number of Novel Coronavirus (2019-nCoV) cases in China in the first half of January
459 2020: A data-driven modelling analysis of the early outbreak. *J. Clin. Med.* 2020, 9(2),
460 388. doi: 10.3390/jcm9020388
- 461 12. Thomas M, Lemaitre M, Wilson ML, Viboud C. (2016) Applications of Extreme Value
462 Theory in Public Health. *PLoS ONE* 11(7): e0159312. doi:
463 10.1371/journal.pone.0159312
- 464 13. Tariq A, Lee Y, Roosa K, Blumberg S. (2020) Real-time monitoring the transmission
465 potential of COVID-19 in Singapore, February 2020. *MedRxiv* [Preprint]. 2020 [Cited
466 2020 April 5]. Available from:
467 <https://www.medrxiv.org/content/10.1101/2020.02.21.20026435v6.article-info>. doi:
468 10.1101/2020.02.21.20026435
- 469 14. Kraemer MUG, Yang CH, Gutierrez B, Wu CH. (2020) The effect of human mobility
470 and control measures on the COVID-19 epidemic in China. *Science*. doi:
471 10.1126/science.abb4218

- 472 15. Azman AS, Luquero F. (2020) From China: hope and lessons for COVID-19 control. T.
473 Lan. *Infec. Dis.* doi: 10.1016/S1473-3099(20)30264-4
- 474 16. Zhao S, Lin Q, Ran J, Musa SS. (2020) Preliminary estimation of the basic reproduction
475 number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven
476 analysis in the early phase of the outbreak. *Int. J. Infec. Dis.* doi:
477 10.1016/j.ijid.2020.01.050
- 478 17. Zhi Z. (2020) The epidemiological characteristics of an outbreak of 2019 novel
479 coronavirus diseases (COVID-19) in China. *Chinese Journal of Epidemiology [Preprint].*
480 2020 [Cited 2020 April 07]. doi: 10.3760/cma.j.issn.0254-6450.2020.02.003
- 481 18. Chiu Y, Chebana F, Abdous. (2018) Mortality and morbidity peaks modeling: An
482 extreme value theory approach. *Statistical Methods in Medical Research.* doi:
483 10.1177/0962280216662494
- 484 19. Northrop PJ. (2004) Likelihood-based approaches to flood frequency estimation. *Journal*
485 *of Hydrology.* doi: 10.1016/j.jhydrol.2003.12.031
- 486 20. Sen S, He J, Kasiviswanathan K.S. (2020) Uncertainty quantification using the particle
487 filter for non-stationary hydrological frequency analysis. *Journal of Hydrology.* doi:
488 10.1016/j.jhydrol.2020.124666
- 489 21. Hosking JRM. (1990) L-Moments: Analysis and Estimation of Distributions Using
490 Linear Combinations of Order Statistics. *Journal of Royal Statistical Society. Series B*
491 (Methodological). From: <https://www.jstor.org/stable/2345653>
- 492 22. Nicoletta Lanse. (2020) Even if you test negative for COVID-19, assume you have it,
493 experts say. *Live Science.* 2020 April 3 [Cited 2020 April 9] Available from:
494 <https://www.livescience.com/covid19-coronavirus-tests-false-negatives.html>

- 495 23. Murray JL. (2020) Forecasting COVID-19 impact on hospital bed-days, ICU-days,
496 ventilator days and deaths by US state in the next 4 months. medRxiv [Preprint]. 2020
497 [Cited 2020 April 7] Available from:
498 <https://www.medrxiv.org/content/10.1101/2020.03.27.20043752v1>. doi:
499 10.1101/2020.03.27.20043752
500
- 501 24. Eder S, Fountain H, Keller MH, Xiao M. 43,000 People Have Travelled From China to
502 U.S. Since Coronavirus Surfaced. The New York Times. 2020 April 4 [Cited 2020 April
503 9]. Available From: [https://www.nytimes.com/2020/04/04/us/coronavirus-china-travel-](https://www.nytimes.com/2020/04/04/us/coronavirus-china-travel-restrictions.html)
504 [restrictions.html](https://www.nytimes.com/2020/04/04/us/coronavirus-china-travel-restrictions.html)
- 505 25. Colbourn T. (2020) COVID-19: extending or relaxing distancing control measures. T.
506 LANCET Pub. Health. doi: 10.1016/ S2468-2667(20)30072-4
- 507 26. Cavallo JJ, Donoho DA, Forman HP. (2020) Hospital Capacity and Operations in the
508 Coronavirus Disease 2019 (COVID19) Pandemic—Planning for the Nth Patient. Insights.
509 2020 March 17 [Cited 2020 April 10]. Available from:
510 <https://jamanetwork.com/channels/health-forum/fullarticle/2763353>
- 511 27. Wong JY, Kelly H, Dennis KM, Wu JT. (2013). Case fatality risk of influenza A
512 (H1N1pdm09): a systematic review. Epidemiology.
513 doi: [10.1097/EDE.0b013e3182a67448](https://doi.org/10.1097/EDE.0b013e3182a67448)
- 514 28. Lipsitch M, Donnelly CA, Fraser C, Blake IM, Cori A, et al. (2015) Potential Biases in
515 Estimating Absolute and Relative Case-Fatality Risks during Outbreaks. PLOS Neglected
516 Tropical Diseases 9(7): e0003846. doi: [10.1371/journal.pntd.0003846](https://doi.org/10.1371/journal.pntd.0003846)

- 517 29. Kobayashi T, Jung SM, Linton NM, Kinoshita R. (2020). Communicating the Risk of
518 Death from Novel Coronavirus Disease (COVID-19). J. Clin. Med. doi:
519 10.3390/jcm9020580
- 520 30. Shear MD, Goodnough A, Kaplan S, Fink S. (2020) The lost month: How a failure to test
521 blinded the US to COVID-19. The Economic Times. 2020 March 30 [Cited 2020 April
522 5]. Available from: [https://economictimes.indiatimes.com/news/international/world-](https://economictimes.indiatimes.com/news/international/world-news/the-lost-month-how-a-failure-to-test-blinded-the-us-to-covid-19/articleshow/74876897.cms)
523 [news/the-lost-month-how-a-failure-to-test-blinded-the-us-to-covid-](https://economictimes.indiatimes.com/news/international/world-news/the-lost-month-how-a-failure-to-test-blinded-the-us-to-covid-19/articleshow/74876897.cms)
524 [19/articleshow/74876897.cms](https://economictimes.indiatimes.com/news/international/world-news/the-lost-month-how-a-failure-to-test-blinded-the-us-to-covid-19/articleshow/74876897.cms)
- 525 31. Jennifer AG, Lauren AM, Alison PG, Jeffrey PT. (2020) Probabilistic uncertainty
526 analysis of epidemiological modeling to guide public health intervention policy.
527 Epidemics. doi: 10.1016/j.epidem.2013.11.002
- 528 32. Flaxman S, Mishra S, Gandy A. (2020) Estimating the number of infections and the
529 impact of non-pharmaceutical interventions on COVID-19 in 11 European countries.
530 Imperial College London. doi: 10.25561/77731
- 531 33. Özmen Ö, Nutaro JJ, Pullum LL, Ramanathan A. (2016) Analyzing the impact of
532 modeling choices and assumptions in compartmental epidemiological models.
533 Simulation. doi: 10.1177/0037549716640877.

534 **Ethics approval and consent to participate**

535 The ethical approval or individual consent was not applicable

536

537 **Availability of data and materials**

538 The data were retrieved by the Center for Systems Science and Engineering(CSSE) at Johns
539 Hopkins University: <https://github.com/CSSEGISandData/COVID-19> [accessed
540 2020].

541

542 **Funding**

543 The authors KSK and II would like to thank Indian Institute of Technology Roorkee for
544 supporting this research financially.

545

546 **Competing Interests**

547 The authors have declared that no competing interests exist.

548

549 **Acknowledgement**

550 Authors would like to thank Editor and anonymous reviewers for reviewing the paper. All
551 authors would like to sincerely acknowledge their family members for supporting and
552 encouraging to carry out the research work during this critical lockdown period.

553

554 **Authors Contribution**

555 Conceived and designed the experiments: KSK. Performed the experiments: MA, KSK.
556 Analyzed the data: KSK, MA, II, BS. Contributed reagents/materials/analysis tools: MA, MB.
557 Wrote the paper: KSK, II, MA, BS. Interpretation of results: KSK, II, MA, JH, MB. Developed
558 the codes: JH, KSK, MA.

