Weather-inspired ensemble-based probabilistic prediction of COVID-19

Roberto Buizza

*Correspondence: roberto.buizza@santannapisa.it

CrossMark

← Click for updates

Scuola Superiore Sant'Anna, Pisa (Italy), and Center for Climate Change studies and Sustainable Actions (3CSA; Pisa Italy).

Abstract

The objective of this work is to generate probabilistic predictions of the spread of COVID-19 starting from observed data. The probabilistic forecasts are generated using an ensemble method inspired by probabilistic weather prediction systems operational today. Each ensemble member is defined by a logistic model: more specifically, each forecast is generated using logistic curves, determined by stochastically perturbing parameters.

Two are the main conclusions of this work. The first conclusion regards the ensemble method: results show that this method could provide valuable information on the probability of future scenario. The second conclusion regards the logistic model used to generate each single forecast: results indicate that the logistic model is too simple to be able to simulate the complex dynamic of the spread of COVID-19.

These conclusions indicate that, to generate more accurate and reliable probabilistic forecasts of the spread of diseases such as COVID-19, ensemble methods could be used, but each member of the ensemble of forecasts should be generated using a realistic model.

Keywords: COVID-19, probabilistic prediction, ensemble methods, uncertainty estimation

Summary: strengths and limitations of this study

- One of the strengths of this work is that **it demonstrates how ensemble methods could be used to develop a probabilistic prediction of the spread of a disease**, and thus provide decision makers not only with the most likely outcome but also with a range of possible scenarios, each with its associated probability of occurrence.
- A second strength of this work is that for the first time, to our knowledge, **an ensemble-based methodology**, used successfully in operational weather prediction by the main meteorological services, has been used to predict possible future outcomes of the spread of diseases such as COVID-19. An ensemble of 30 members, with parameters estimated by randomly perturbing the observed data in a way to simulate observation errors, is shown to provide valuable forecasts of the most likely outcome, possible future ranges and probabilities.
- **Results are promising:** the ensemble-method has been tested using a simple, data-driven model. For China, ensemble forecasts generated on day 25 have predicted correctly the outcome for the next 35 days. For Italy and Korea, forecasts based on the median of the ensemble have relative mean absolute errors of between 10-20%.
- **A limitation of this study** is that the logistic model used to generate each single forecast is too simple, not capable to capture detailed dynamics; despite this weakness, we think that the results are good enough to illustrate the value of applying an ensemble-based probabilistic method to build realistic, accurate and reliable operational ensemble diseases prediction systems (EDPS).
- We hope that this work will **inspire research in this area and lead to the development of accurate and reliable ensemble disease prediction systems**; it is suggested that ensemble methods are used with more realistic models, to predict future scenarios and their probability of occurrence, so that decision makers can take better-informed decisions than if they had access to a single forecast.

COVID-19: infection data analysis and prediction

Since the initial spread of the COVID-19 infection, different methods have been proposed and applied to predict future number of infected cases, either based only on data-driven models, or

on health models. This work falls into the first category, and it proposes to apply ensembles of forecasts generated using a data-driven logistic model to generate probabilistic forecasts of future scenarios. The idea of applying ensembles has been inspired by the ensemble prediction systems developed in the past 25 years to predict the weather [1,2].

If we look at Italy, which at the time of writing (31st of March 2020) is facing a very critical situation, some authors (see e.g. [3]) have spoken of an 'exponential growth rate. De Nicolao, for example, on 4 March talked about the possibility that the number of infected cases would reach 700+. Unfortunately, he severely under estimated the numbers that were going to be reached in the following days: 17,660 assessed cases (according to the Civil Protection Agency) by the end of the 31st of March.

Other authors (see, e.g., [4]) pointed out that the growth rate of diseases like COVID-19 is not exponential, but follows the 'S-shape' curve described by a logistic equation. It is interesting to quote the results of [5], who has also analysed the Italian data and tried to fit an exponential or a logistic curve to the existing data up to 12th March. [5] concluded that on that date it was still very difficult to predict the future number of confirmed infected cases.

Compared to the exponential one, a logistic curve has an asymptotic value. This growth limitation is to be expected, partly due to the fact that there is a physical limit to the number of people who can be infected, and partly because of measures that can be implemented to contain the spread of the disease. As soon as containment measures start working, one should expect that the growth rate slows down, the curve changes concavity, and starts evolving from an exponential one toward a curve that tends to an asymptotic value. The solution of a logistic equation behaves precisely in this way. In fact, a logistic model has been used, for example, to predict the risk of developing a given disease (e.g. diabetes; coronary heart disease), based on observed characteristics of the patient [6,7]. It is worth stressing the distinction between the 'ensemble method' and the 'data-driven logistic model':

- With the terms 'ensemble method', we mean that forecasts are generated using an ensemble of forecasts, instead of a single one; each forecast is generated using the data-driven model in such a way as to simulate initial condition and model uncertainties;
- With the term 'data-driven logistic model', we refer to the model based on the solution of the logistic equation, used to generate each single forecast.

This work aims to discuss the value of following an ensemble method, and to test the quality of the data-driven model based on the logistic equation, and address the following three questions:

- a) Is the logistic equation capable to fit well observed data of COVID-19?
- b) Is the ensemble method capable to generate accurate and reliable probabilistic diseases' predictions?
- c) Can we use the combination of the two, an ensemble of stochastically-perturbed logistic curves, to generate

COVID-19 probabilistic predictions of future infected people? The paper is structured as follows. In section 3, the logistic equation is introduced. In Section 4, an ensemble-based Monte Carlo method designed to estimate future forecast probabilities is described. In section 5, COVID-19 data from China are analysed using the logistic equation, and in Section 6, tests are performed to assess the accuracy of ensemble-based forecast for China. In section 7, the same analysis is applied to Italy and South Korea. Conclusions are then drawn in Section 8.

The 'logistic' growth model

The growth model equation used to analyse COVID-19 infection data and predict future numbers is based on the logistic equation, an equation used also to investigate and understand the predictability of weather systems. The equation, first proposed by [8] Lorenz (1982), then modified by [9,10], was used by [11] to investigate how resolution increases could improve weather forecasts.

The fact that a model characterized by sub-exponential growth rates, rather than an exponential one, could fit better the data was pointed out also by authors studying the propagation of COVID-19. For example, [12] stated that 'The recent outbreak of COVID-19 in Mainland China is characterized by a distinctive algebraic, sub-exponential increase of confirmed cases with time during the early phase of the epidemic, contrasting an initial exponential growth expected for an unconstrained outbreak with sufficiently large reproduction rate.'

Our choice of using a logistic curve was inspired by its use in weather prediction, since both problems face a maximum asymptotic value, and growth rates are dominated by linear and quadratic terms [8,11]:

$$\frac{dE}{dt} = \alpha E - \beta E^2 + \gamma \quad (1a)$$

In Eq. (1a), 'E' is a measure of a quantity that is under investigation: the forecast error in weather prediction, or, in this work, the cumulative number of infected cases. In this equation, 't' represents time, and the symbol 'dE/dt' denotes the ratio between the variation of the quantity E, dE, and the time variation 'dt'. The parameters α and β are the coefficients that determines the relative weight of the linear and the quadratic terms, and γ is a constant.

Eq. (1a) can also be written as:

$$\frac{dE}{dt} = (aE + S)\left(1 - \frac{E}{E_\infty}\right) \quad (1b)$$

Where the parameters have been defined as follows:

$$\begin{cases} S \equiv \gamma \\ E_\infty \equiv \frac{\alpha}{2\beta} + \sqrt{\frac{\alpha^2}{4\beta^2} + \frac{\gamma}{\beta}} \\ a \equiv \beta E_\infty \end{cases} \quad 2$$

Note that in the computation of the asymptotic value in Eq. (2), the solution with the '+' rather than the '-' sign has been chosen, since we are looking for solutions that remain positive for all the forecast times 't'.

The solution of Eq. (1b) is given by:

$$E(t) = E_{\infty} \cdot \left[1 - \frac{1}{1 + C_1 e^{C_2 t}} \left(1 + \frac{S}{a E_{\infty}} \right) \right] \quad (3a)$$

where E_{∞} is the asymptotic value. Eq. (3a) can be written also in normalized form as:

$$\eta(t) \equiv \frac{E(t)}{E_{\infty}} = 1 - \frac{1}{1 + C_1 e^{C_2 t}} \left(1 + \frac{S}{a E_{\infty}} \right) \quad (3b)$$

The coefficients C_1 and C_2 are given by:

$$\begin{cases} C_1 \equiv \frac{1}{E_{\infty} - E_0} \left(E_0 + \frac{S}{a} \right) \\ C_2 \equiv a + \frac{S}{E_{\infty}} \end{cases} \quad (4)$$

where E_0 is the initial-time error.

As [9] indicated, when used in weather prediction we can interpret a as the initial rate of growth, S as the effect of model uncertainties on the error growth, and E_{∞} as the error asymptotic value.

Given a set of data (number of infected people on a number of consecutive days), the (α, β, γ) parameters of Eq. (1a) have been determined by fitting the curve to the data. Once these parameters have been computed, the coefficients a , S and E_{∞} have been determined by applying the definitions (2), and the analytical solutions (3a, 3b) have been defined (see, e.g., [11]). The solutions (3a,b) can be used to predict future values. Eq. (3b) can also be used to compute the doubling time at each forecast time t :

$$\tau(t) = \frac{\ln(2)}{\alpha + \frac{\gamma}{E(t)}} \quad (5)$$

An ensemble-based method to estimate confidence and future scenarios

The main weakness of generating forecasts using only a single, deterministic forecast is that it does not provide any confidence information, for example expressed in terms of the possible range of future values (see, e.g., discussions in the articles collected in [1]). Furthermore, it does not take into account the fact that the training data used to estimate the model's parameters are affected by observation errors, and that the model provides only an approximate description of the system's dynamics. In other words, it does not take into account initial condition and model uncertainties.

We could address these two weaknesses by using an ensemble of forecasts, based on logistic curves estimated by perturbing the training data by an amount that reflects the observations' uncertainty. We can then use the ensemble of forecasts to estimate the future range of scenario, which could be expressed, for example, in terms of the minimum and maximum value, and quantiles computed from the predicted ensemble members. The ensemble members have been generated by stochastically perturbed logistic curves, defined by perturbing the observed data covering the training period (e.g. the data up to the time the forecasts were initialized).

More precisely, each forecast has been defined by a logistic curve, with its governing parameters $(\alpha, \beta, \gamma_j)$ estimated by applying best-linear fit methods to the perturbed observations. In our case, each ensemble of forecasts included 30 members, i.e. $j=1, \dots, 30$. Each observation has been perturbed by a random number, sampled from a Gaussian distribution with a standard deviation defined to be 10% of the daily variations in the observed value. By stochastically perturbing the observations used to compute the parameters of each logistic curve, we generate an ensemble of curves, each defined by a slightly different parameters. In this way, we simulate not only initial but also model uncertainties.

Suppose, for example, that today is the evening of day 31 since the start of the infection (23rd of March 2020), and we have access to observations of the past 31 days (with day 1 defined as the first day when the number of infected people has jumped from zero to a positive value). This is how the forecasts for the future days are generated:

- The daily percentage increment in the number of infected people $\delta n_{obs}(d) = [n_{obs}(d) - n_{obs}(d-1)] / n_{obs}(d-1)$ is computed for each of the past 31 days;
- The standard deviation σ_{obs} of the increments daily percentage increments $\delta n_{obs}(d)$ is computed;
- 30 sets of observations are defined by randomly perturbing each observation by a random number $r(d)$ sampled from a Gaussian distribution with zero mean, and standard deviation given by σ_{obs} ; in other words, each perturbed observation is defined by multiplying the observed value by $[1+r(d)]$;
- Once the perturbed observations have been generated, 30 logistic curves are fitted to the data; these are the 30 ensemble members defined by the logistic curve of Eq. 3a with slightly different parameters;
- Forecast statistics (e.g. the ensemble-mean and the ensemble standard deviation) are computed using the 30 curves to estimate the probability distribution of the future number of infected people.

The perturbation strategy assumes that the statistics of the daily changes in the observed numbers is a good approximation of the statistic of possible observation errors: given no other indications of the possible observation errors, we believe that this is a reasonable choice. We recognize that

the choice of sampling from a Gaussian distribution with a standard deviation equal to 0.1 times σ_{obs} is an arbitrary choice that should be further tested in the future. Sensitivity tests for China suggest that this is a reasonable choice, and for the purpose of this work is reasonable and can help us illustrating the value of an ensemble-based, probabilistic method compared to one based on a single prediction.

The idea of perturbing the input observations to generate the ensemble has been inspired by the successful use of similar approaches in weather prediction (see, e.g., [13-15]). In weather prediction, a similar ensemble method is used to generate of an ensemble of initial states for the atmosphere or the ocean, by assimilating perturbed observations.

COVID-19: the case of China

Figure 1 shows the cumulative number of confirmed infected cases from China, reported from the 22nd of January to the 7th of March, and a logistic curves with parameters estimated using the observed numbers of the first 25 days. It shows that the logistic curve is capable to describe very well the data even past day 25: the asymptotic value of the logistic curve is 68,790, and the cumulative number of cases reported on the 7th of March is 68,482 (see **Table A.1** in the Appendix for the data used in this study). It is worth pointing out that on 13 February, the World Health Organization (WHO) reported a daily spike of 15,200, while the day before and the day after it reported 2,000 and 4,000: because of the very large difference between 15,200 and these two values, we decided to replace 15,200 with the average between the values reported on the 12th and the 14th of February. A similar filtering approach is used in weather prediction, when quality-control methods are applied to ensure that the data used to initialize the models are not affected by unexpected, large errors.

Let us consider the case of a logistic curve fitted between days d_0 and d_1 , and then used to predict the future numbers between d_1 and d_2 . To measure the fit of the curve to the data during the data assimilation period (d_0 to d_1) we use the correlation coefficient, computed considering the data covering the period (d_0 , d_1). To measure the error of the fitted curve compared to the data between forecast days (d_1+1) and d_2 , we compute the average relative mean-absolute-error ($\langle R\text{-MAE} \rangle$), defined by computing the average (between the forecast days) of the ratio between the mean-absolute-error and the observed number:

$$\langle R\text{-MAE} \rangle(d_1+1, d_2) = \frac{1}{(d_2 - d_1 - 1)} \sum_{d=d_1+1}^{d_2} \frac{|fc(d) - obs(d)|}{obs(d)} \quad (6)$$

The $\langle R\text{-MAE} \rangle$ gives an average measure of the daily percent-age error of the forecast.

If we consider the logistic curve shown in **Figure 1**, the correlation coefficient between day 0 and 24 is 99.8%, and the average relative mean-absolute error ($\langle R\text{-MAE} \rangle$) between days 25 and 45 of 1.9%, which is a very small number. The $\langle R\text{-MAE} \rangle$ between days 25-45 when all 45 days for a logistic

curve generated assimilating observations covering all 45 days(not shown) is 1.2%.

Two very interesting questions to address are the following:

- Since when could we have predicted correctly the asymptotic value?
- Could we identify when this would have been possible by looking at the observed data?

We will answer top these questions by analysing the behaviour of the observed data. **Figure 2** shows the number of infected cases in China during the first 30 days, **Figure 3** shows the first order derivative of this function (i.e. the daily trend), and **Figure 4** shows the second order derivative. The figures show both the observed values, and a fitted curve obtained by considering the observations covering the whole 45 day period.

The first order derivative gives us an information on the growth rate, and the second order derivative tells us whether the data follow a convex or a concave trend. Note that at around day 15 the first order derivative starts decreasing in amplitude (**Figure 3**), and the second derivative becomes negative (**Figure 4**),

Figure 3. COVID-19 China. Observed daily increments (i.e. the derivative of the observed curve shown in **Figure 2**; solid line) and increments computed from the fitted curve (dotted line).

Figure 4. COVID-19 China. Second derivative of the observed daily increment curve (solid line) and second derivative of the fitted curve (dotted line).

indicating a change in the curve concavity. Note that, compared to the observed data, the fitted curves (the dotted lines) show smoother changes. Day 18, when the second derivative changes sign, is when the fitted curve (**Figure 1**) changes concavity and starts bending towards the asymptotic value. Note also that the number of infected people started climbing again between day 21 and 23: during these days, the first derivatives increases and the concavity of the curve changes back to positive. It is only from day 25 that the concavity becomes, and stays negative.

Figures 3 and **4** indicate that by looking at the two derivatives (the growth rate and the concavity of the observation curve), it is possible to identify whether the observed curve starts behaving like a logistic curve. Only from this time, we could apply the logistic model and predict the range of possible values using an ensemble of stochastically-perturbed logistic curves. Trying to fit a logistic curve to data that do not show yet a reduced growth rate and a change in concavity can lead to errors: indeed, our tests indicate that in some cases it is even impossible to fit a logistic curve to the data. In other words, the simple, error-growth model based on the

logistic equation that we are using has some limitations in its predictions' capabilities.

Going back to the two questions that we posed above, we can then say that:

- We cannot aim to predict the asymptotic value until we do not detect that the observed data show a clear transition from an exponential growth to one similar to a logistic one.
- We could identify when this would be possible by analysis the observed data, and in particular at its first and second order derivatives.

COVID-19: ensemble-based predictions for China

Let us now try to predict the number of infected cases for China, both on days before and after day 25 from the start of the infection. The data shown in **Figures 3-4** indicate that only from day 25 the number of infected people was growing at a lower rate, and the curve's concavity turned negative. Thus, we should expect difficulties in predicting the future numbers before day 25, and more accurate forecasts afterwards.

Figures 5 and **6** show two probabilistic forecasts generated on day 22 and day 25. Each probabilistic forecast is expressed in terms of five curves: the minimum and maximum values, the median, and the 25th and 75th quantiles (no assumption has been made on the shape of the forecast probability density function). These figures show that the observed values are outside the range spanned by the forecast distribution generated on day 22 (**Figure 5**), but were included in the range predicted on day 25 (**Figure 6**).

In terms of curve fitting, the correlation coefficient (standard error) between the fitted logistic curve generated by the unperturbed member of the ensemble, is 95% (262) for the curve fitted between days 1-21, and 92% (349) for the curve

Figure 5. COVID-19 China. Ensemble-based probabilistic forecast issued on day 22. The blue symbols show the observations used to estimate the ensemble of 30 logistic curves. The red symbols show the reported number of confirmed cases of days 23-to-45. Five black curves are shown: the minimum and maximum values (dotted lines), the median (solid line), and the 25th and 75th quantiles (dashed lines).

fitted between days 1-24. Thus, in this case, despite the fact that at day 22 the data did not show yet a change in concavity, we could fit logistic curves to the perturbed data, and generate 30 forecasts. It is worth reporting that we tried to do the same for earlier days, and for them we could not fit logistic curves to the data.

To verify that all the ensemble forecasts issued on day 25 were better than the ensemble forecasts issued on day 22, we computed the $\langle R\text{-MAE} \rangle$ for all forecasts. As can be seen from **Table 1**, the forecasts issued on day 25 had smaller errors.

Table 1. $\langle R\text{-MAE} \rangle$ statistics of the forecasts issued on day 22 and day 25, computed between forecast day 25-45 (see text for more details).

	Fc issued on day 22	Fc issued on day 25
Mean	18.79%	9.30%
Std	11.37%	9.10%
minimum	12.72%	1.68%
median	18.31%	9.01%
maximum	23.28%	17.48%

If we consider the forecast issued at day-25, results indicate that:

- There is a ~25% probability that the number would be lower than the 60,000 (i.e. below the 25th quantile);
- There is a ~25% probability that numbers would be higher than 75,000 (i.e. above the 75th quantile);
- The most-likely outcome, given by the median forecast, gives for day 45 a number not far from the observed value;
- The observed values between day 25 and 45 were inside the range spanned by the forecast probability distribution.

These results suggest that a logistic curve model could be used to predict the number of infected people, only from the

time when the curve of the observed number of infected cases shows a change in concavity. Predictions issued before this time could have large errors. After day 25, the model could be used, and our results indicate that if the model is used to generate an ensemble of forecasts, these forecasts could provide valuable information of the probability of occurrence of different scenarios.

Figure 6 illustrates the value of an ensemble forecast compared to having a single forecast only: the spread among the curves, represented by the spread among the different quantiles of the forecast distribution, gives an indication of the uncertainty of the forecasts. Note that the spread is small up to about day 30, but then is large, especially in the upper part of the diagram. This suggests that it is more probable that the observed numbers at day 60 will be above than below the median forecast.

COVID-19: ensemble-based predictions for Italy, South Korea and the UK

Let's now consider Italy, South Korea and the United Kingdom: first, let's analyses whether the data suggest that we could generate ensemble forecasts, and if so let's generate them.

At the time of writing this document (1st of April 2020), the situation in Italy is very serious. Today is the 39th day since the first case was reported, and from the site of the Italian Protection Agency we had access to 39 days of observed numbers of confirmed infected cases above 1 (see **Table A.2**, in the Appendix). Authorities are struggling to estimate how fast the disease could spread, and how many cases they could be facing in the next few weeks. South Korea and the UK have also been facing an increasing number of cases.

Can we say something about the future numbers of total number of infected cases, starting from the observed data?

Figure 7 shows the total number of infected cases for Italy,

Figure 7. COVID-19 China (red lines), Italy (blue lines), the UK (green lines) and South Korea (black lines). Total number of observed cases (day 1 is the first day when a number different from zero was reported; data from WHO for China, the UK and South Korea, and the Civil Protection Agency for Italy).

South Korea and the UK (for all countries, day 1 is the first day when infected cases were reported), and **Figures 8 and 9** show the first and second order derivatives. If we look at the first and second order derivatives, we can see that:

- **Italy:** the curves indicate that the numbers are still rising fast; the first derivative indicates that the last two days had a slower growth rate, and the second derivative indicates that the concavity has become negative;
- **UK:** the curves indicate that the first derivative is still increasing, and the concavity is still positive;
- **South Korea:** the first derivative have been showing a slower increase than the other three countries, and the concavity of the curve has been very close to zero, with days with a slightly positive value followed by days with slightly negative values.

Figures 8-9 indicates that for Italy, and possibly South Korea, we should be able to generate probabilistic forecasts. To the contrary, since for the UK data there is still no indication of a flexion point, it will not be possible to fit a logistic curve to

the data. Indeed, we tried to fit the logistic curve to the data for the UK, and we could not.

Figures 10 and 11 shows the forecasts for Italy, issued on day 35 (27th of March) and day 39 (31st of March, at the time of writing this paper). The two forecasts are rather consistent: they all predict that in the next 25 days from now the total number of cases would be between 120k-180k, and the most recent forecast gives a 50% chance of being between 130k-170k. Note that there is still a rather large uncertainty in both forecasts.

Ensemble forecasts have been generated for all days, from day 30 to day 40. For all these forecasts, the fit of the logistic curves to the observations have been very good. For example, we have looked at the correlation coefficients for the ensemble member number 0 (the one generated without adding random perturbations to the observations): for this curve, the correlation coefficient between the observed data and the fitted curve has been always above 93.5% for forecasts initialized on day 30 and 31, and above 95% for the other days.

To assess the quality of the ensemble forecasts, we have computed the $\langle R\text{-MAE} \rangle$ of all the ensemble members, for the ensembles generated from day 30 to day 40. **Figure 12** shows the $\langle R\text{-MAE} \rangle$ for the ensemble-median: it indicates that the forecast quality improves as we get closer to day 40, with the $\langle R\text{-MAE} \rangle$ decreasing from ~25% on day 30-31, to less than 10%. **Table 2**, which lists the mean and standard deviation of the $\langle R\text{-MAE} \rangle$ of the ensemble members, confirms the conclusions drawn from **Figure 12**, that as the forecast lead time decreases, the forecast accuracy improves. If we consider, for example Italy, the ensembles issued on days 30 had mean $\langle R\text{-MAE} \rangle$ of 28%, with a ~25% standard deviation, while the ensemble issued on day 40 had mean $\langle R\text{-MAE} \rangle$ of 14%, with

Table 2. Error statistics for the ensemble forecasts generated on day 30-to-40, computed between verification days 40-60, for Italy and South Korea. For each country, the mean and the standard deviation of the <R-MAE>(40-60) of the 30 ensemble members are reported.

Initial day	Italy <R-MAE>(40-60)		South Korea <R-MAE>(40-60)	
	Mean	std	Mean	std
30	27.8%	25.1%	20.3%	7.9%
31	28.7%	22.9%	18.9%	7.7%
32	30.9%	25.0%	17.3%	8.5%
33	28.1%	22.3%	16.4%	8.0%
34	21.0%	15.1%	15.5%	6.7%
35	13.4%	12.2%	14.6%	5.4%
36	18.1%	13.2%	14.1%	6.9%
37	12.8%	10.1%	14.5%	6.0%
38	13.9%	11.6%	12.7%	4.5%
39	17.2%	11.4%	11.7%	6.4%
40	14.0%	10.4%	11.7%	6.9%

a ~10% standard deviation.

Figures 13 and 14 show the equivalent forecasts for South Korea, issued on day 35 and 39 (23rd and 28th of March). They show that, compared to the Italian case, the observed data do not follow a pure logistic curve, but keep increasing with a linear trend between day 20 and 40. The poorer fit of the logistic curve to the data is reflected in the correlation coefficient for all the forecasts, which oscillates between 70% and 90% and never goes above that level. The poorer fit of the data could be due to the fact that the South Korean national numbers are compounded by the sum of cases occurring in different regions, and possibly each region is at a different

stage of the epidemic. The logistic curve model is possibly too simple to be able to capture this small scale dynamics, which make the national average number growing linearly for a long time. One possible way to address this weakness could be to model the spread of the diseases at a finer scale, fitting logistic curves to each single region, and then computing the national average by summing the logistic curves computed for each region. This option was not tested due to the lack of detailed, regional data.

Also for South Korea ensemble forecasts have been generated for all days, from day 30 to 40. **Figure 12** shows the average relative mean-absolute-error (MAE) also of the ensemble

Figure 14. Forecasts of the 'reported number of cases' for South Korea, issued on day 39 (28 March). The blue symbols show the observations (from WHO) used to estimate the ensemble of 30 logistic curves. The red symbols show the observations not used to compute the curves, covering days 40-60. Five black curves are shown: the minimum and maximum values (dotted lines), the median (solid line), and the 25th and 75th quantiles (dashed lines).

median forecasts for South Korea. Results indicate that the average relative MAE decreases as we get closer to day 40, with values decreasing from around 20% to about 10%. Similar conclusions can be drawn by looking at the ensemble error statistics reported in [Table 2](#). Thus, despite the fact that the ensembles' spread was too narrow, and did not include the observed values in the range of predicted values, the ensemble median forecast had a similar quality as for Italy.

It is worth mentioning that simple models show similar deficiencies in weather prediction. If one tries to predict detailed weather patterns with a model with too coarse a resolution, either in physical space or in probability space, the forecasts fail. Here, we suffer from the same problem: more complex models capable to simulate different regions, or clusters, separately, should be used if one wants to be able to resolve the small scale features that makes the daily changes differ from the changes induced by linear and quadratic dynamics.

Conclusions

In this work, we discussed issues linked to the prediction of the evolution of the COVID-19 disease, and proposed that ensemble methods should be used to generate probabilistic disease predictions. The key advantage of following an ensemble-based probabilistic approach is that it allows to take into consideration initial condition (observation) and model uncertainties, and it predicts the future range of possible scenarios. The ensemble median can be used to predict the most likely scenario, and the spread among the ensemble members, represented for example by the minimum and maximum values, and the 25th and the 75th percentiles, can be provide information of the range of possible future outcomes.

In the ensemble method applied in this work, each member of the ensemble has been defined by a logistic model, with the three parameters that define a logistic curve computed by fitting the curve to stochastically perturbed observations.

This ensemble method has been inspired by work in numerical weather prediction, where observations and models are stochastically perturbed to generate the ensemble of single forecasts, have been used in operational weather prediction very successfully since 1992 [[1,2,13-15](#)].

Three questions have been addressed in this study:

- Is the logistic model capable to fit well the observed data?
- Can we develop diseases' probabilistic predictions by applying methods used in weather prediction?
- Can we use an ensemble of stochastically-perturbed logistic curves to generate COVID-19 probabilistic predictions of future infection numbers?

Concerning the first question, results indicate that the logistic curve can fit reasonably well the data of COVID-19 spread only from the time when the observed data show a clear transition from an exponential growth to one similar to a logistic one. This time could be identified by analysing the observed data, in particular the first and second order derivatives of the observed data trend. This is the limitation of the simple logistic-curve model. Concerning the second question, results indicate that valuable probabilistic forecasts can be generated using an ensemble method. Concerning the third question, results indicated that probabilistic forecasts based on stochastically-perturbed logistic curves can provide very valuable information on the probability of possible future scenarios. Clearly, this is true only for forecasts issued after the time when the observed data show a clear transition from an exponential growth to one similar to a logistic one.

Thus, results indicate that the use of ensemble methods should be promoted, and that ensemble generation methods based on perturbed observations could be used to generate the ensemble of forecast trajectories. They also indicate that the logistic model is too simple a model, that cannot simulate the detailed dynamics that characterized the spread of the COVID-19. More sophisticated models that can simulate, for example, the impact of the disease control policies (e.g. of the imposition of travel restrictions and of lockdowns) should be used to generate each single trajectory. In other words, to be able to make predictions that would take into account disease control policies, one should use an ensemble of forecasts generated using a Susceptible-Exposed-Infected-Removed model ([\[17,18\]](#)) that could include the impact of these measures on the spread of the disease. Such an approach would allow to estimate the probability of different outcomes that would follow the adoption of different measures.

We hope that this work can lead to the development of more realistic, accurate and reliable ensemble methods capable to predict the spread of diseases such as COVID-19, and that in the future we will see, as it is the case for weather, ensemble-

based probabilistic predictions being at the core of disease monitoring and risk management.

Ensemble-based probabilistic methods have been used for almost three decades to predict possible future weather scenarios. Different ensemble methods are used to generate these ensembles: perturbing the observations collected up to the time when the forecast is issued (as used in this work) is one of the method applied to generate them. Ensembles of forecasts are used routinely to predict day-to-day weather, including extreme events such as hurricanes and wind storms. Hurricane strike probabilities are generated by considering the tracks predicted by the ensemble individual members: the ensemble-mean or the ensemble-median are used to predict the most likely track, and the range spanned by the tracks is used to estimate the forecast uncertainty. Ensembles are used to generate monthly and seasonal forecasts of large-scale weather patterns, and to compute, for example, the probability that extreme heat or cold waves, or extreme dry or wet periods could affect regions of interest.

We invite the scientific community to develop similar methods for disease prediction, so that better-informed decisions could be taken in the future.

Additional files

Appendix A

Competing interests

The author declares that she has no competing interests.

Publication history

EIC: Jimmy Efird, East Carolina University, USA.

Received: 01-May-2020 Final Revised: 14-Sept-2020

Accepted: 24-Sept-2020 Published: 02-Oct-2020

References

1. Palmer, T. N., and Hagedorn, R., 2006: Predictability of weather and climate. Cambridge University Press, pp702 (ISBN-13 978-0-521-84882-4).
2. Buizza, R., 2018: Introduction to the Special Issue on '25 years of ensemble forecasting'. *Q. J. Roy. Meteorol. Soc.*, 1-11 (doi: 10.1002/qj.3370).
3. De Nicolao, G., 2020: COVID-19 in crescita esponenziale? Analizziamo morti, terapie intensive e ricoveri (<https://statisticallearningtheories.worldpress.com>).
4. Trafficante, A., 2020: La curva dei contagiati non è esponenziale, ma occorre tenere duro per evitare che lo diventi. (www.haffingtonpost.it).
5. Remuzzi, A., and Remuzzi, G., 2020: COVID-19 and Italy: what next? *The Lancet* (doi: 10.1016/S0140-6736(20)30627-9).
6. Truett, J, Cornfield, J, Kannel, W, 1967: A multivariate analysis of the risk of coronary heart disease in Framingham. *Journal of Chronic Diseases*. 20 (7): 511–24 (doi: 10.1016/0021-9681(67)90082-3).
7. Freedman, A. A., 2009: Statistical models: theory and practice. Cambridge University Press, pp. 128.
8. Lorenz, E. N., 1982: Atmospheric predictability experiments with a large numerical model. *Tellus*, 34, 505–513.
9. Dalcher, A., and Kalnay, E., 1987: Error growth and predictability in operational ECMWF forecasts. *Tellus*, **39A**, 474-491.
10. Simmons, A., J., and Hollingsworth, A., 2002: Some aspects of the improvement in skill of numerical weather prediction. *Q. J. Roy. Meteorol. Soc.*, 128, 647-677.
11. Buizza, R., 2010: Horizontal resolution impact on short- and long-range

- forecast error. *Q. J. Roy. Meteorol. Soc.*, 136, 1020-1035 (DOI 10.1002/qj.613; SCOPUS 2-s2.0-77954449861; WOS:000279407000014).
12. Maier, B. F., and Broskmann, D., 2020: Effective containment explains sub-exponential growth in confirmed cases of recent COVID-19 outbreak in Mainland China (available from www.arxiv.org: 2002.07572v1).
13. Houtekamer, P. L., Derome, J., Ritchie, H., and Mitchell, H. L., 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.*, 124, 1225-1242.
14. Buizza, R., Leutbecher, M., & Isaksen, L., 2008: Potential use of an ensemble of analyses in the ECMWF Ensemble Prediction System. *Q. J. R. Meteorol. Soc.*, 134, 2051-2066.
15. Zuo, H., Alonso-Balmaseda, M., Mogensen, K., and Tiedsche, M., 2018: OCEAN5: the ECMWF Ocean Reanalysis System and its real-time analysis component. ECMWF Research Department Technical Memorandum 823, pp 44 (available from ECMWF, Shinfield Park, Reading RG2-9AX, UK).
16. Linton, N. M., Kobayashi, T., Yang, Y, Hayashi, K., Akhmetzhanov, A. R., Jung, S.-M., Yuan, B., Kinoshita, R, Nishiura, H., 2020: Epidemiological characteristics of novel coronavirus infection: A statistical analysis of publicly available case data (available from www.medrxiv.org: doi 10.1101/2020.01.26.20018754).
17. J.C. Blackwood and L.M. Childs, 2018: An introduction to compartmental modelling for the budding infectious disease modeler. *Letters in Biomathematics*, 5(1), 195-221.
18. Wangping Jia, Ke Han, Yang Song, Wenzhe Cao, Shengshu Wang, Shanshan Yang, Jianwei Wang, Fuyin Kou, Penggang Tai, Jing Li, Miao Liu, Yao He, 2020: Extended SIR Prediction of the Epidemics Trend of COVID-19 in Italy and Compared With Hunan, China. *Frontiers in Medicine*, 7, pp. 169.

Citation:

Buizza R. **Weather-inspired ensemble-based probabilistic prediction of COVID-19.** *J Med Stat Inform.* 2020; **8**:4.
<http://dx.doi.org/10.7243/2053-7662-8-4>