Science of the Total Environment 806 (2022) 150639

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Model-based ensembles: Lessons learned from retrospective analysis of COVID-19 infection forecasts across 10 countries

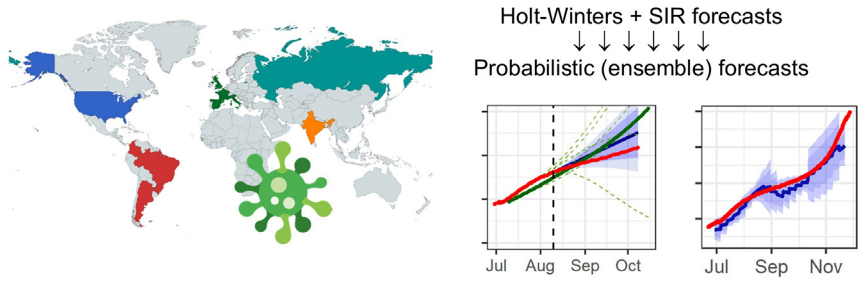
Martin Drews [a,*], Pavan Kumar [b], Ram Kumar Singh [c], Manuel De La Sen [d], Sati Shankar Singh [b], Ajai Kumar Pandey [b], Manoj Kumar [e], Meenu Rani [f], Prashant Kumar Srivastava [g]

[a] *Department of Technology, Management and Economics, Technical University of Denmark, Kgs. Lyngby 2800, Denmark*
[b] *Rani Lakshmi Bai Central Agricultural University, Jhansi 284003, India*
[c] *Department of Natural Resources, TERI School of Advanced Studies, New Delhi 110070, India*
[d] *Institute of Research and Development of Processes IIDP, Department of Electricity and Electronics, University of the Basque Country, PO Box 48940, Leioa, Spain*
[e] *Forest Research Institute, Dehradun, Uttarakhand 248006, India*
[f] *Department of Geography, Kumaun University, Nainital, Uttarakhand 263001, India*
[g] *Institute of Environment and Sustainable Development, Banaras Hindu University, Varanasi 221005, India*

HIGHLIGHTS

- COVID-19 forecast models are critically needed but highly uncertain.
- Retrospective analyses across different countries/environments portray model biases.
- *Re*-forecasts for different seasons and pandemic states are explored in 10 countries.
- Probabilistic (ensemble) forecasts provide added value but must be further explored.
- Ensemble forecasts show reasonable skill 20 days ahead (20% relative errors).

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 19 March 2021
Received in revised form 22 September 2021
Accepted 23 September 2021
Available online 27 September 2021

Editor: Jay Gan

Keywords:
COVID-19 ensemble forecasts
Environmental factors
SIR model
Holt-Winters model
Retrospective analysis

ABSTRACT

Mathematical models of different types and data intensities are highly used by researchers, epidemiologists, and national authorities to explore the inherently unpredictable progression of COVID-19, including the effects of different non-pharmaceutical interventions. Regardless of model complexity, forecasts of future COVID-19 infections, deaths and hospitalization are associated with large uncertainties, and critically depend on the quality of the training data, and in particular how well the recorded national or regional numbers of infections, deaths and recoveries reflect the the actual situation. In turn, this depends on, e.g., local test and abatement strategies, treatment capacities and available technologies. Other influencing factors including temperature and humidity, which are suggested by several authors to affect the spread of COVID-19 in some countries, are generally only considered by the most complex models and further serve to inflate the uncertainty. Here we use comparative and retrospective analyses to illuminate the aggregated effect of these systematic biases on ensemble-based model forecasts. We compare the actual progression of active infections across ten of the most affected countries in the world until late November 2020 with "re-forecasts" produced by two of the most commonly used model types: (i) a compartment-type, susceptible–infected–removed (SIR) model; and (ii) a statistical (Holt-Winters) time series model. We specifically examine the sensitivity of the model parameters, estimated systematically

[*] Corresponding author.
E-mail addresses: mard@dtu.dk (M. Drews), pawan2607@gmail.com (P. Kumar), singhramkumar@gmail.com (R.K. Singh), manuel.delasen@ehu.eus (M. De La Sen), directorextension.rlbcau@gmail.com (S.S. Singh), pandey.ajai1@gmail.com (A.K. Pandey), manojfri@gmail.com (M. Kumar), meenurani06@gmail.com (M. Rani), prashant.just@gmail.com (P.K. Srivastava).

https://doi.org/10.1016/j.scitotenv.2021.150639

from different subsets of the data and thereby different time windows, to illustrate the associated implications for short- to medium-term forecasting and for probabilistic projections based on (single) model ensembles as inspired by, e.g., weather forecasting and climate research. Our findings portray considerable variations in forecasting skill in between the ten countries and demonstrate that individual model predictions are highly sensitive to parameter assumptions. Significant skill is generally only confirmed for short-term forecasts (up to a few weeks) with some variation across locations and periods.

1. Introduction

Since the earliest days of the global COVID-19 pandemic, a wide range of mathematical and epidemiological models have been proposed as means of exploring the transmission properties of the disease or as instruments for delivering indicative forecasts of, e.g., total infections, hospitalizations and mortalities, including forecast scenarios assuming combinations of different non-pharmaceutical countermeasures (Jewell et al., 2020; Li et al., 2020; Diaz-Quijano et al., 2020). Worldwide, the latter is extensively used by local and national (health) authorities to inform not only policies aimed at limiting the spread of COVID-19, but also to help manage the implications for the rest of society, including the economy. Variations of the "classical" compartmental-type models, where the susceptible population is divided into different "compartments" so far rank amongst the most used. This includes variants of the susceptible-infectious-removed (SIR) model (Porter and Oleson, 2013, Biswas et al., 2020, Wangping et al., 2020) and the extended susceptible-exposed-infectious-removed (SEIR) model (Sun et al., 2020; Yang, 2020). Other authors have explored data-driven prediction models of different complexities based on classical (e.g. curve fitting) as well as advanced statistical and Bayesian approaches (Cássaro and Pires, 2020; Ceylan, 2020; Chatterjee et al., 2020; Petropoulos and Makridakis, 2020; Remuzzi and Remuzzi, 2020; Singh et al., 2020a; Singh et al., 2020b; Tomar and Gupta, 2020; Verity et al., 2020), geographically-based transmission models (Wu et al., 2019), stochastic transmission models (Kucharski et al., 2020; Bi et al., 2020), agent-based models (Koo et al., 2020) and hybrid model types utilizing artificial intelligence techniques (Tiwari et al., 2021; Yang, 2020; Zheng et al., 2020) as means of providing forecasts of COVID-19 progression. In most of these referred cases, the domain focus is regional or national. Some of the exceptions include Wu et al., 2019, who in the early phase of the pandemic (based on flight bookings and human mobility) predicted the geographic pattern of COVID-19 spread originating from Wuhan and surrounding cities, both inside China and internationally, using a compartment model trained using a Markov Chain Monte Carlo technique; Singh et al. (2020b), who applied an autoregressive integrated moving average (ARIMA) model to time series data drawn from the "top 15" countries in terms of cumulative infections; and Ceylan (2020), who estimated the COVID-19 prevalence in Italy, Spain, and France also using ARIMA models.

As recently proposed by several authors, including Castro et al. (2020), and Wilke and Bergstrom (2020), COVID-19 forecasts are inherently associated with large uncertainties that particularly prevent reliable prediction of intermediate and long-term COVID-19 trajectories (IHME, 2020; Scudellari, 2020). For example, a common denominator in all of these varied modelling efforts (and the countless more not referred here) is that the large uncertainties associated with model structure and estimated model parameters effectively propagate to the predictions. This is regardless of the modelling philosophy and complexity of the models used, and includes amongst other factors hypothesized (and typically non-modelled) correlations with atmospheric conditions such as temperature, precipitation and humidity and their potential influence on the local incidence of the disease (Bashir, 2020; Briz-Redón and Serrano-Aroca, 2020; Gupta et al., 2020; Menebo, 2020; Runkle et al., 2020; Şahin, 2020). As an alternative, Castro et al. (2020) promotes probabilistic forecasts similar to the ones used by weather forecasters. The COVID-19 ensemble forecasts for the United States (US) produced by the Centers for Disease Control and Prevention (CDC - Centers for Disease Control and Prevention, 2020), which are based on contributions from more than 30 expert modelling groups, is arguably the premier example of such an approach. Wilke and Bergstrom (2020) identifies the way that the many different sources of uncertainties compound as the fundamental problem for COVID-19 prediction and adds that (quoting) "predicting the trajectory of a novel emerging pathogen is like waking in the middle of the night and finding yourself in motion—but not knowing where you are headed, how fast you are traveling, how far you have come, or even what manner of vehicle conveys you into the darkness". Or, expressed differently, that there are still many questions about the new coronavirus that remain largely unresolved, and that mathematical-epidemiological modellers are not different from anyone else - we are presently all learning as we go.

This study does not claim to have solved these challenges. Rather, we hypothesize that there are important lessons to be learned from systematic analysis of the retrospective performance of COVID-19 forecast models in different (e.g., natural) environments. This is in line with the multi-model comparison collaboration suggested by the Center for Global Development already in late May 2020 (Chalkidou et al., 2020). It is also yet another reference to weather forecasters (Castro et al., 2020). Or even more to regional climate modellers, who until recently (when this practice was replaced by large collaborative multi-model intercomparison experiments (Gutowski et al., 2016)) would often evaluate (and subsequently improve) in-house regional models on the basis of lessons learned from dedicated experiments within climatic domains other than the "native" ones (Refsgaard et al., 2014). Transferring this analogy to COVID-19 modelling, this paper analyses the results of a systematic and intercomparable modelling effort, where we explore the properties of an ensemble of retrospective forecasts ("re-forecasts") of the COVID-19 development until late November 2020 across ten of the most highly infected countries in the world and compare them to real-life records reported by, e.g., the Johns Hopkins Corona Virus Resource Centre (Johns Hopkins 2020). The ten countries are the US, India, Brazil, Russia, France, United Kingdom (UK), Italy, Spain, Argentina and Colombia. COVID-19 took hold in these countries at different times, and so they conceptually represent slightly different phases of the pandemic. Inherently, each of these records represent the "sum" of the local circumstances, with the sampled countries spanning multiple continents, environmental and climatic conditions, developed as well as developing countries, economic and technological capacities, cultures, different interventional strategies, etc. Accordingly, the observed numbers could be expected to show distinct features that - ideally - might be attributed to different key factors, including, if this is applicable, environmental influences. In this paper, we assert that such features could be used to uniquely sample the properties and biases of general forecast model types from a more principal perspective. Here, we exclude data from 2021, where the widespread but unevenly distributed emergence of pharmaceutical interventions have significantly affected COVID-19 trajectories in some countries.

In the following, we consider two very simple forecast models – one epidemiological and one statistical: the most basic form of the SIR model (Rodrigues, 2016; Liu et al., 1987) and a Holt-Winters triple exponential smoothing model (Holt, 1957; Winters, 1960). Both test models are first trained on samples of observed numbers of active COVID-19 infections, and forecasts are subsequently tested against independent samples of the data. The Holt-Winters model is drawn from classical time series analysis and as implied by its name, a triple exponential smoothing procedure is applied to derive parameters corresponding to the level, trend and seasonality of the time series, which subsequently are used to make statistical forecasts (Singh et al., 2020a). The reason for working with such minimal models is threefold. Firstly, as mentioned above, epidemic forecast models rely on uncertain parameters and cannot in general provide exact COVID-19 predictions. As noted above they could however be used to provide ranges of trajectories. With our choice of test models, we implicitly assume that predictive skill is principally related to the estimated parameters and not to the form of the model. Secondly, and within this framework, we explore the potential of probabilistic forecasting as proposed by several authors. Compared to the large number of studies focusing on specific forecast models, probabilistic approaches remain relatively unexplored. In this paper, we pursue the concept of an “initial condition, single-model ensemble” from the geosciences, where it is used by, e.g., weather forecasters and climate scientists (Haughton et al., 2014). Essentially, we create a small ensemble of COVID-19 forecast models by sampling different training periods, yielding a set of forecasts that are intrinsically linked to local phases of the pandemic and to local country conditions (i.e., fully exploiting our multi-country approach). To our knowledge, this is the first attempt at using specifically this kind of methodology for COVID-19 analysis. Using our ensemble-based approach, we quantify the predictive skill of COVID-19 forecasts across the range of underlying conditions implicitly represented by the ten countries. In particular, we put numbers on what is defined in the scientific literature as the “limited short-term forecasting skill” of COVID-19 models (Castro et al., 2020; Wilke and Bergstrom, 2020), quantitatively assessing the forecasting skill as a function of the number of forecasted days ahead. Thirdly, we suggest that simplicity goes a long way towards ensuring that results for the different countries are replicable and intercomparable. This comes with the significant caveat that at best our simple models are only likely to be applicable at shorter time scales as real-life trajectories of COVID-19 progression do not follow simple paths. Applying them even at intermediate time scales however allows us to examine how poorly performing ensemble members could influence, e.g., central estimates based on the full model ensemble. The latter is an entirely realistic situation often found, e.g., within weather forecasting and climate modelling.

In practice, the capacity to model COVID-19 varies between countries - including the ten countries, we are studying – and existing models are for obvious reasons optimized for local conditions. Given the assumed differences in COVID-19 trajectories across our suite of countries, our choice of models denote general approaches that would be equally applicable in all environments, model parameters are readily interpretable, and the two models represent typical modelling philosophies (epidemiological and statistical) currently used for COVID-19 predictions all over the world.

2. Materials and methods

2.1. Data

The COVID-19 data used in this study was retrieved from the Johns Hopkins Coronavirus Resource Centre (Johns Hopkins 2020) and from the Worldometer (Worldometer, 2020) and describes the number of active infections reported by ten different countries (US, India, Brazil, Russia, France, United Kingdom (UK), Italy, Spain, Argentina and Colombia). The data extracted span the period from 22 January 2020 to 28 November 2020.

Population totals for each of the ten countries was extracted from the World Bank database (https://data.worldbank.org).

2.2. Fitting test models to COVID-19 data

To illustrate the performance of our two test models, forecasts were initially compared with COVID-19 observations within five overlapping time windows as indicated in Table 1.

The outline of our analysis scheme is shown in Fig. 1. For each time window, a corresponding Holt-Winters model was first trained on a 120-day subset of the data (reaching beyond the analysed time windows) and subsequently used to provide a 60-day forecast, which was compared to an independent sample of the data. Analogously, we estimated the parameters for an ensemble of 25 SIR models based on sliding and relatively tight monthly subsets of the data (see Table S1 in the Supplementary Material). The first SIR model was trained on data from 1 May – 31 May, which were consecutively shifted by one-week (i.e., 8 May – 7 June) when used for training the second SIR model, etc. The last model was trained on data from 16 October – 15 November. For each of the five time windows, six of the associated SIR model forecasts as well as an ensemble mean were compared with independent observations and with the Holt-Winters forecasts as indicated in Fig. 1.

As shown in Fig. 1, the SIR models considered within each overlapping 60-day analysis period differ with respect to “age”, i.e., the “newest” model is trained on observed COVID-19 records immediately preceding the cross-validation period, whereas the “oldest” SIR model is trained on data that is five weeks behind (and thereby not expected to capture the most current developments unless the situation would be stationary). The consideration of gradually “older models” here serves as means of sampling uncertainty and demonstrating the potential robustness of our ensemble approach. Evidently, the individual model performance is dependent on what strategy we use to train and validate the models. In our study, the use of independent samples ensures that our test results are comparable, however, one could easily make the case for a variant approach for training and validation.

2.3. Holt-Winters modelling

The Holt-Winters method (Holt, 1957; Winters, 1960) comprises a forecast equation and three smoothing equations — one for the level l, one for the trend t and one for the seasonal time series component s with corresponding smoothing parameters α, β and γ. The model parameters are estimated from observed time series data using, e.g., least-squares fitting. The Holt-Winters method distinguishes between two kinds of seasonality: an additive formulation is used when the seasonal variations are roughly constant through the time series, whilst a multiplicative formulation is used when seasonal variations are changing proportionally with the level of the time series. We used the implementation of the Holt-Winters method provided by the *forecast* software package (Hyndman and Khandakar, 2008) made for the R statistical programming environment (R Core Team, 2013). In all the five cases illustrated on Figs. 2–4, S1–S2 we used a 120-day training period as the basis for fitting a Holt-Winters model in R, which was then subsequently used to provide a 60-day forecast (Table 1). We opted for a 120-day training period for two reasons: *(i)* to achieve a reasonable ratio between the training and evaluation periods of 2/3 to 1/3, and *(ii)* to have it long enough that the Holt-Winters model would pick up any

Table 1
The five time windows analysed in Figs. 2–4 and S1–S2 (Supplementary Material) using the scheme outlined in Fig. 1. The date in the third column indicates the first day of the Holt-Winters forecast, which is also the first day of the analysis (cross-validation) period (Fig. 1) considering the entire “mini-ensemble”.

No.	Time window	Holt-Winters 60 d forecast	Analysis
1	1 May–9 August	11 June	Fig. 2
2	1 June–8 September	11 July	Fig. 3
3	30 June–8 October	10 August	Fig. S1
4	1 August–11 November	9 September	Fig. S2
5	1 September–28 November	30 September	Fig. 4

"seasonal" (periodic) components. Periodic components were however not found, leaving our Holt-Winters models a function only of the level l and trend t. The confidence intervals (80%, 95%) on the figures below and in the Supplementary Material reproduce the diagnostics produced by the R implementation.

2.4. SIR modelling

The SIR modelling discussed above was implemented in R. In this study we use the basic formulation of the susceptible-infectious-removed (SIR) compartment model (Rodrigues, 2016; Liu et al., 1987). In this form, the development of an infectious disease (here: COVID-19) is expressed by three differential equations:

$$\frac{dS}{dt} = -\frac{\beta IS}{N}$$

$$\frac{dI}{dt} = \frac{\beta IS}{N} - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

where N indicates the total population, S is the number of individuals susceptible to the disease (or alternatively the proportion of the total population), I is the number of infected individuals and R is the number of individuals removed from the model. That is, they are no longer susceptible to infection due to recovery/ immunization or death. In the basic form used here, the SIR model considers the population to be homogenous, and that infection of susceptible individuals occurs simultaneously (Wu and McGoogan, 2020). We further assumed that the number of deaths (and births) is negligible with respect to the total population, and that N was equal to the total country population. The transmission parameter β represents the transition rate from the infected to susceptible individuals. Theoretically, β is a function of the average number of contacts per person per time, multiplied by the probability of disease transmission in a contact between a susceptible and an infectious subject. Accordingly SI/N^2 is the fraction of contacts between infectious and susceptible individuals that result in the susceptible becoming infected. The "recovery" parameter γ represents the transition rate between I and R and can be assumed proportional to the number of infectious individuals (equivalently, the probability of an infectious individual recovering during any time interval dt is γ times dt). If an individual is infectious for an average time period D then $\gamma = 1/D$.

For each of the ten countries, we derived 25 SIR models (β, γ) using a least-squares optimization, based on monthly-long training periods representing a sliding one-week time window starting on 1 May 2020 (see Table S1 in the Supplementary Material, which includes the estimated β, γ for all countries and all models). The relatively short monthly time windows ensure that the SIR models (β, γ) developed directly relates to the current epidemic situation at specific times including the effects of time-varying interventions. Trial attempts at using shorter training periods (2–3 weeks) were also carried out (results not shown). In most cases, observed improvements in short-term forecast skill were found to be marginal, and it often proved more difficult to achieve a stable convergence of the least-squares optimization. We also investigated the use of sliding two-week time windows as means of generating ensemble members. This was found to produce significantly worse results for the ensemble as a whole (not shown). Using the model parameters derived from the least-squares fits, 200 future days were simulated for each SIR model.

2.5. Ensemble forecasts

To study the feasibility of a probabilistic, ensemble-based approach, we calculated an ensemble mean based on forecasted values from the SIR model ensemble (Table S1). Iteratively, for each week, starting from July 2020, we average five SIR models as indicated in Fig. 1. This results in a rolling ensemble mean, where for each week we add forecasts from the SIR model that was trained on data up to but excluding that week, while the "oldest" model drops out. Since the five averaged SIR models differ only in the (β, γ) parameters, which are estimated from different training periods, our mini-ensembles are similar to "initial condition" ensembles (Haughton et al., 2014).

To assess the skill of our (short-term) probabilistic forecast, for each weekly "mini-ensemble" (five SIR models), we calculated the relative prediction errors ($\frac{|prediction-observed|}{observed}$) of the ensemble mean forecasts up to 35 days ahead (Fig. 6).

3. Results

Figs. 1 through 3 compare the performance of our test models within three of the time windows mentioned above (see also S1 and S2 in the

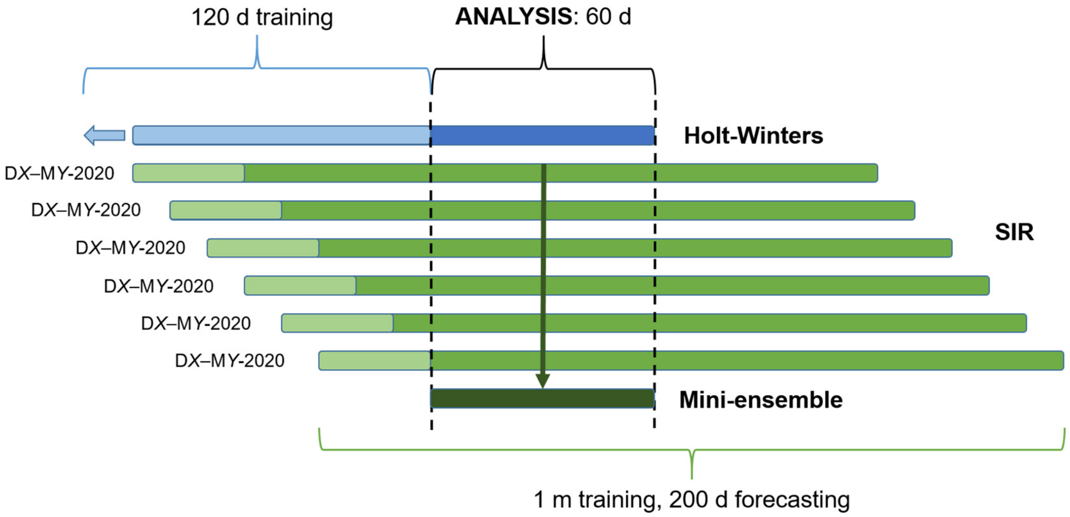


Fig. 1. Analysis scheme. The blue colors indicate the Holt-Winters model and the green colors the (6) SIR models. The Holt-Winters model is trained on the 120-days of observations (light blue) preceding the beginning of the 60-day forecast period (darker blue). The SIR models are trained on one month of COVID-19 observations based on one-week "sliding windows" (light green) and then used to forecast the situation 200-days ahead (darker green). An ensemble mean is calculated based on the average of the SIR models for the 60-day part of the forecast periods overlapping the Holt-Winters forecast (dark green). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

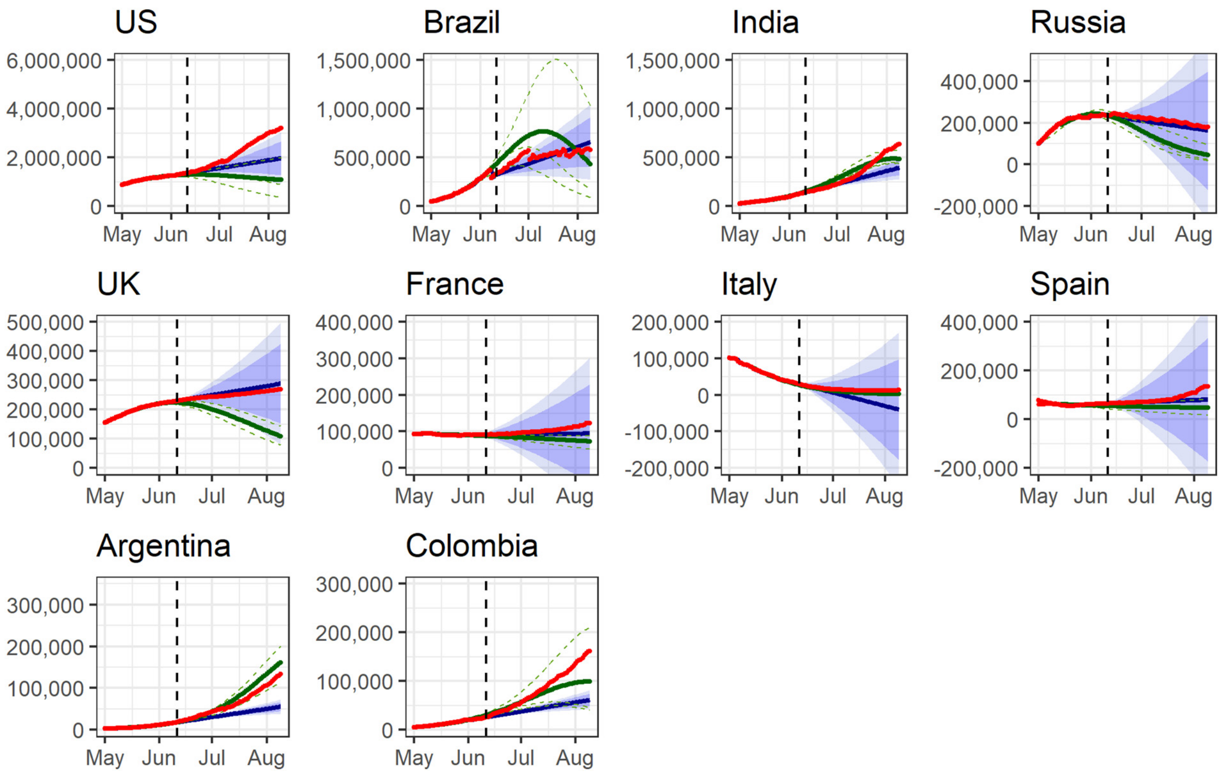


Fig. 2. Retrospective analyses of 60-day forecasts of COVID-19 progression across ten countries (US = United States of America) from June to August 2020. The red curves indicate the recorded number of active cases in each country ("observations"). The dark blue curves display the Holt-Winters forecasts with associated 80% (blue) and 95% (light blue) confidence intervals. The dark green curves show the ensemble mean of six SIR models (Fig. 1), whereas the dashed green curves depict the individual SIR model forecasts. The vertical dashed line indicates the first day of the Holt-Winters forecast and defines the analysis period. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Supplementary Material). The red curves indicate the observed number of active daily COVID-19 records in the ten different countries (Section 2.1); the blue curves are the associated predictions from the Holt-Winters model (with the shaded areas indicating the estimated 80% and 95% confidence intervals respectively; Section 2.3); while the remaining curves show predictions obtained using the basic SIR model with parameters estimated from sliding monthly data sets (Section 2.4).

Considering the first time window (Fig. 2), in six out of ten cases (Brazil, Russia, UK, France, Italy and Spain) and also partially in the case of India, the re-forecasts using the time series model fall within the estimated 80% confidence intervals and in a few instances almost overlap the observations. The equivalent SIR predictions including the ensemble mean (dark green curves) in most of these cases compare less convincingly to the observed behavior or deviate all together (Brazil, UK). For Brazil, this is partially explained by the observed data (and hence the training data used for the SIR models) being noisier than in the other cases. Conversely, Argentina and Colombia the predicted trends and levels using the Holt-Winters method clearly fail to capture the observed (near-term) changes in COVID-19 transmission rates in these countries, as these abrupt changes are not reflected in the training data, while the ensemble mean of the fitted SIR models rightly captures the trend. For the US, the Holt-Winters and the "newest" of the individual SIR forecasts agree reasonably well with the observations but only for the initial two week period, after which neither of the individual SIR models, the SIR model ensemble mean or the time seriest models demonstrate any skill. Again, this can be attributed to the fact that knowing the "past" trend alone is not enough to forecast future trends. In general, the ensemble mean seems to perform better than individual SIR models in reproducing the observed COVID-19 trajectories.

Fig. 3 shows the equivalent results for the second time window. Except for Columbia (where this only holds for the first ~30 days), essentially all of the observations now largely lie within the 80% (US, Russia, UK, Italy) or 95% (Argentina, Brazil, Italy and Spain) confidence intervals associated the Holt-Winters forecasts - or forecasts are at least trending the same way (India). For all countries the situation thereby seems to have been relatively stable in spite of national trends being very different. The SIR ensemble forecasts generally agrees with the time series forecasts though with a tendency of underestimating slightly. Notable exceptions are Brazil (where the noisiness of the training data leads to highly varying SIR model fits, but where the ensemble mean actually performs better than the Holt-Winters model); Argentina (where both the ensemble mean and ensemble members follow the observations well until the number of actives suddenly drop dramatically to a lower level in early August and then start to rise again; this is probably linked to changes in the way data was collected in Argentina); and Colombia where two out of six SIR models correctly predict the right shape of the curve despite a sharp drop in August (the ensemble mean does not).

The onset of the second wave of COVID-19 infections in Europe is clearly depicted in the third time window (Fig. 4). For the UK, Italy, and Russia both the Holt-Winters forecasts and the SIR-based forecasts clearly underestimate the sudden and highly rising numbers of infections in these countries. For Spain and France, the observed trend is clearly captured by both individual SIR models and reflected in the SIR ensemble mean, albeit in the latter case the observed forecasts eventually underestimate the observed behavior same as the Holt-Winters models. For India and Argentina, the agreement with observations for the Holt-Winters model is fair (close to observations or within the 80% confidence interval) for most of the evaluation period, whereas SIR (medium- as well as short-term) forecasts are generally completely off. The same is generally found also for the US, except that here the Holt-Winters forecasts only keep pace with observed COVID-19 instances for the first few weeks. For Brazil and Columbia it is evident that variations and potential deficiencies in the training data impact

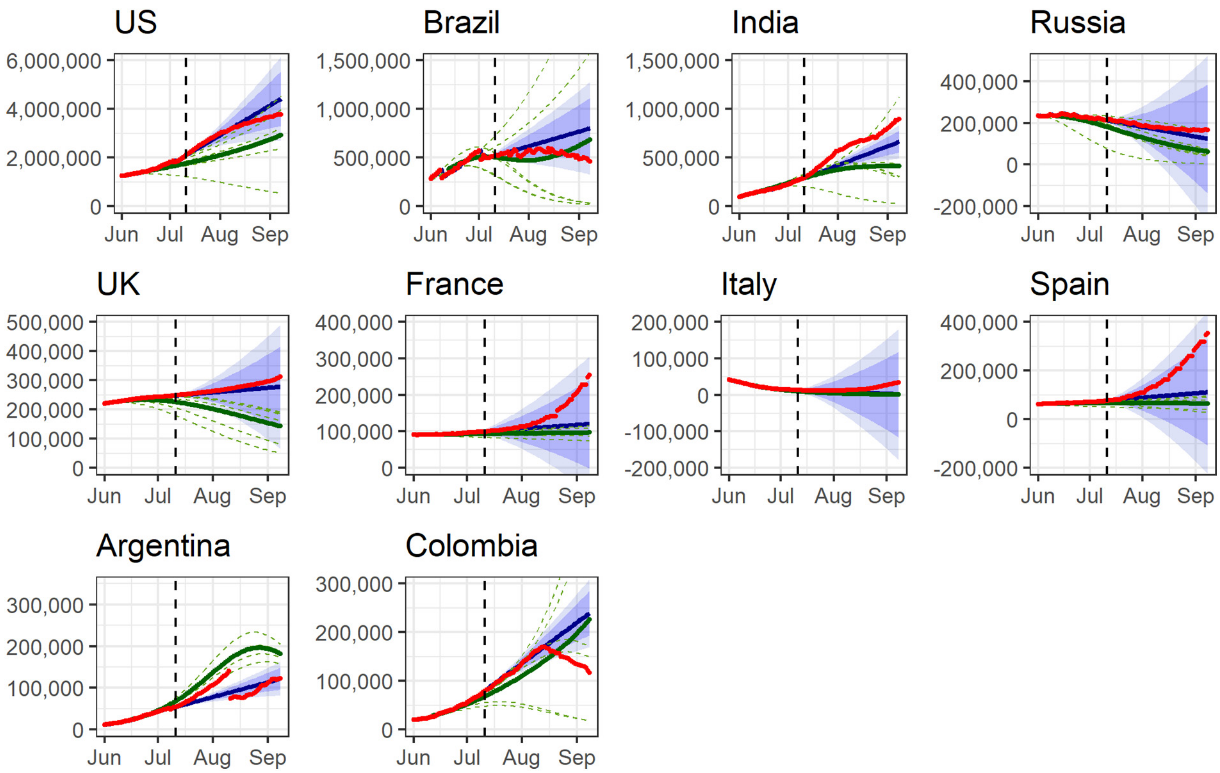


Fig. 3. Same as Fig. 2, except for July to September.

the performance of both types of test models. That said, for Columbia the SIR ensemble mean (and most of the individual models) still adequately captures the observed trend, whereas for Brazil it mimicks the Holt-Winters forecast.

As expected, we see that larger prediction errors are typically associated with observed instances of abrupt change (e.g., autumn 2020), where a curve-fitting approach naturally fails. Overall, the test results found for France, Italy and Spain exhibit the same general characteristics

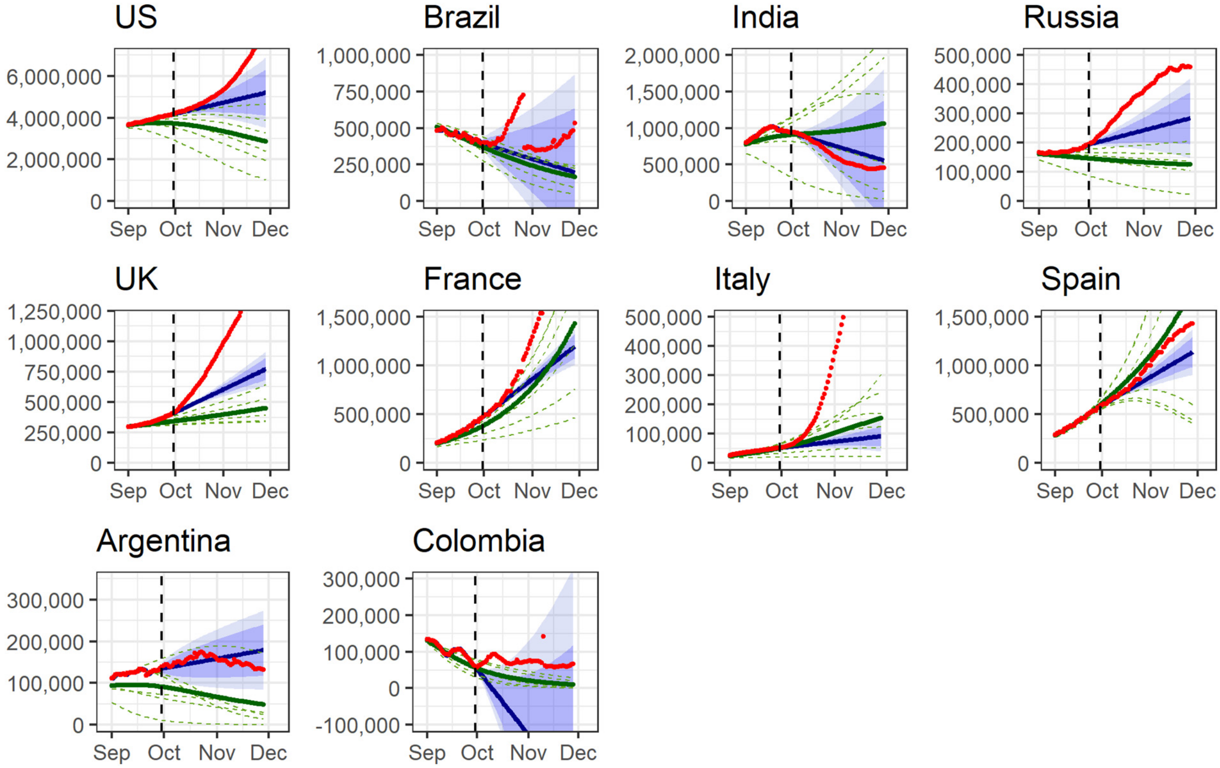


Fig. 4. Same as Fig. 2, except for October to late November. Note the shift of the y-axes for several countries compared to Figs. 2–3.

(Figs. 2–4, S1–S2). However, since all three neighbouring countries stem from the same (South-Eastern) European region and resemble each other in many different ways, this is perhaps not surprising. Analogous similarities are found in the (rather negative) test results from the UK and US, where individual SIR models as well as the ensemble mean predictors do not in general account well for the observed rates of change, and the Holt-Winters forecasts perform only slightly better. Conversely, for the South American countries Brazil, Argentina and Columbia our simple test models are found to describe the observed trends reasonably well between them despite challenges with the observed COVID-19 data. Roughly the same conclusion applies to India, whereas in the case of Russia a fairly good agreement in the first and second time window (Figs. 2 and 3) is replaced by very poor agreement in the third time window (Fig. 4). It is tempting to suggest – specultatively - that perhaps the abovementioned differences could be attributed to the difference between developed and developing countries (e.g. Argentina, Brazil, Colombia and India) and/or to cultural and regional differences (South America/Asia vs. Western). To validate such a hypothesis, however, is beyond the scope of this paper and would require a much more comprehensive analysis.

Interestingly, for nearly all countries and all time windows there are always at least a few of the “re-forecasts” using either of the two types of test models, which demonstrate some level of skill even within a full 60-day time horizon, whereas other models completely fail to capture the right trend. This includes the ensemble average of six SIR models (Fig. 1), which generally seems to be a decent predictor. On this background, Fig. 5 compares the results from a *generalized rolling ensemble forecast* (dark blue curves with 80% and 95% confidence intervals indicated by blue and light blue colors, respectively) based on five SIR models (Section 2.5) with observations (red curves). In all ten cases (Fig. 5) the ensemble mean largely reproduces the observations, whereas the varying widths of the confidence intervals are easily attributed to, for example the observed inconsistencies in the COVID-19 time series used for training the forecast models (e.g. Brazil, Argentina, Columbia). That said, using a rolling ensemble mean in our setup clearly optimizes short-term forecasting. Hence, the collective results shown in Fig. 5 seems to indicate that our ensemble-based approach yields decent and robust short-term forecasting skill.

Fig. 6 shows the relative prediction errors (including the associated 80% confidence intervals) when comparing the rolling ensemble mean forecasts (Fig. 5) with the observed COVID-19 incidences (Section 2.5). If one considers a forecast 20-days ahead, the relative error of the ensemble mean is found to be *less than about 20% (plus or minus about 10%)* for six of the ten countries: US, India, Russia, UK, France, and Spain. For the US, UK and France the relative error of the ensemble mean stays under 20% even up to 30-days ahead while the associated uncertainty increases. In India, Russia and Spain on the other hand both the relative error of the ensemble mean (30–40%) and the associated uncertainty increase, indicating that forecasting 30-days ahead has virtually no skill. For Brazil and Colombia the relative error for a forecast 20-days ahead is about 25–30% with an associated uncertainty that is slightly higher than for the six countries just mentioned. For Italy and Argentina, it is higher still and approaches 40–50%. The poor performance for these two countries can, however, be explained by Figs. 3–4. In both analyses, the SIR ensemble mean clearly fails to capture the COVID-19 trajectories in Argentina due in no small part to inconsistencies in the training data, whereas none of the test models come close to representing the steeply increasing incidence in Italy in the autumn of 2020 (Fig. 4).

4. Discussion and conclusions

While neither of the two test models explored in our comparative analyses (Figs. 2–4, S1–S2) demonstrate superior skill in re-forecasting observed trends and variability across all ten countries, our numerical experiments, and in particular our trial ensemble forecasts, clearly suggest that COVID-19 predictions can be consistently skillful (Figs. 5–6). Unsurprisingly, the statistical extrapolation is found to be best suited

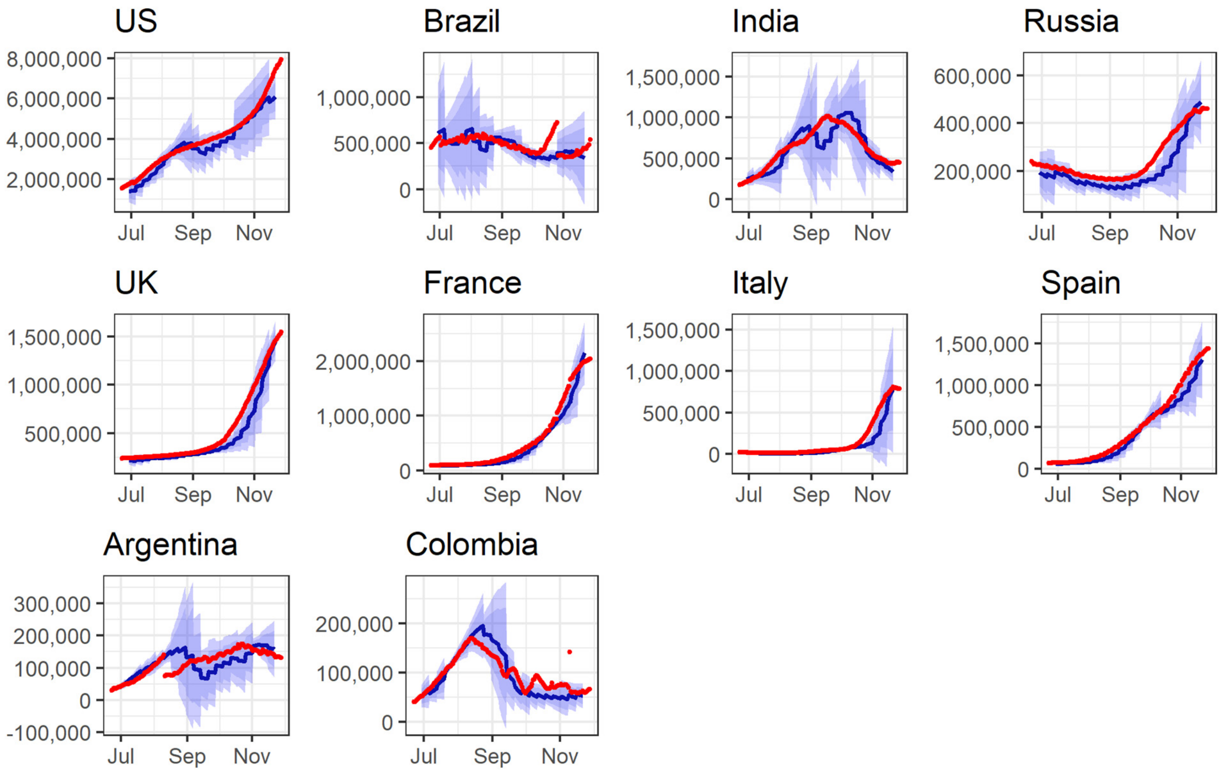


Fig. 5. Ensemble forecasts of COVID-19 progression from mid-July 2020 to late-November 2020 based on our SIR model ensemble (Section 2.5). The dark blue lines indicate the rolling ensemble averages over five models, whereas the colored ribbons are the associated 80% (blue) and 95% (dark blue) confidence intervals. The red curves indicate the recorded number of active cases in each country (“observations”). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

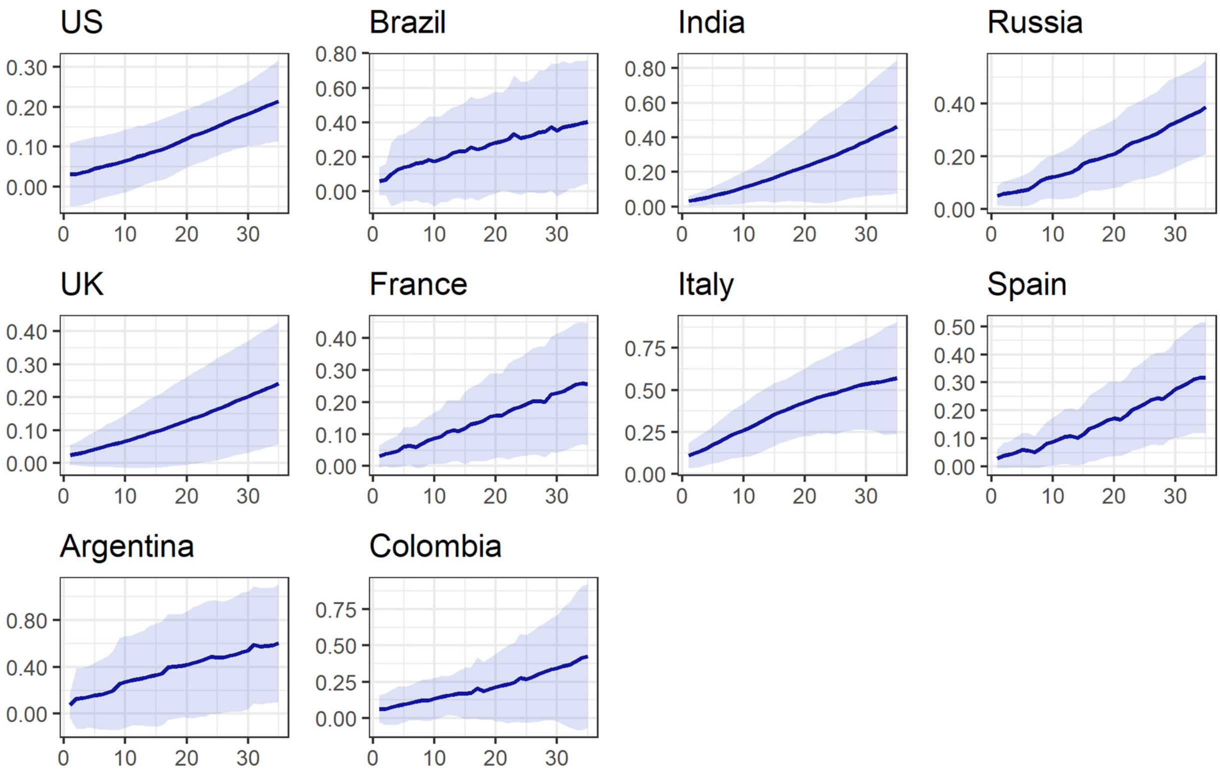


Fig. 6. Relative forecast errors as a function of the number of days predicted ahead. The dark blue curves indicate the relative prediction errors ($\frac{|prediction-observed|}{observed}$) of the ensemble means shown in Fig. 5 as a function of the number of days predicted ahead. The light blue shading is the associated 80% confidence interval inferred from the relative prediction errors of individual ensemble members. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

for situations, where the transmission rates of the coronavirus do not change abruptly as was the case in Europe in the last months of 2020. Thus, in periods where there is abrupt and non-stationary growth in the number of active COVID-19 cases that cannot be attributed to "seasonal" components, Holt-Winters, ARIMA, and similar statistical curve fitting approaches should generally exhibit very poor performance until the new trend and a more stationary situation is picked up.

That said, in spite of the simplicity and evident limitations of the underlying model formulation, our initial condition single-model ensemble forecasts are found systematically to reproduce approx. the right levels and trends, and they generally outperform any of the individual model forecasts (except for a few cases where the ensemble mean is actually not strictly the best overall predictor). This confirms the potential of probabilistic COVID-19 forecasting schemes in a quantitative sense - whether based on single or multiple models – as means of coping with the inherently large and compounding uncertainties (Wilke and Bergstrom, 2020). It is quite possible that this conclusion is mainly a result of the way, we construct our test ensemble. However, emphasis on the ensemble mean is known from other key research areas, in particular climate science, where it is commonly noted that the ensemble mean (i.e., global mean surface temperature) is found to perform better than any ensemble member in comparison to observations (i.e., observed global mean surface temperature) (Annan and Hargreaves, 2011). In this view, the current study could serve to remind us that in terms of providing the most robust model-based information on COVID-19 progression, there is arguably a lot of potential but also a lot of research and learning to be done on how to construct optimal probabilistic forecasts; and that knowledge from other disciplines could help.

Due to the inherently stochastic and constantly changing dynamics of COVID-19 transmissions at local levels, which is compounded by a lot of different factors including the nature of non-pharmaceutical interventions, several authors including Castro et al. (2020) propose that only short-term predictions can be reasonable accurate - generally without providing quantitive numbers on what that means, since this is likely to be model-dependent and situational. Short-term forecasting is arguably a matter of extrapolation and trend detection, which is what simple, parametric models excel at. In this view, our study suggest (Fig. 6) that about 20 days may not be an unreasonable definition of "short term", although we do see individual examples of comparable performance up to 30–35 days, and also instances where the predictive skill is less than 2 weeks amongst our cohort of numerical experiments. This is asserting a ~20% relative prediction error, which seems to be in line with what is found from the recent modelling studies cited above using both simple and highly complex models. This is not meant to imply that simple models are superior to more complex, epidemiological models on these time scales! On the contrary, it is important to point out that our simple test models and equivalent parametric forecast models come with significant weaknesses; including that they don't explicitly take into account the effect of known developments or planned interventions that might affect future transmission dynamics and hence the outcome in terms of estimated numbers of new COVID infection. This is obviously a severe limitation of their use as decision-support tools. Rather, it is probably more reasonable to consider simpler models as means of exploring the overarching issue of predictability – as in the current paper - and for benchmarking (more realistic) COVID-19 forecast models used operationally or academically.

The fact that we have replicated the same retrospective model analyses across ten different countries representing different climates, seasonality and in slightly different phases of the COVID-19 pandemic grant us a unique possibility to study the skill of COVID-19 forecasting in a multi-variate perspective, including aspects of data availability, data quality, the different temporal evolution of the spread of COVID-19 under the compound influence of a variety of environmental and local factors, etc. While the test modelling performed in our study obviously cannot stand alone, further studies along this direction – including detailed attribution of the different drivers, which has so far been primarily done from a univariate perspective, e.g., for select environmental parameters like temperature, could provide important insights to improve – if not prediction models directly – then our understanding of the uncertainties involved and contribute to the development of new robust and probabilistic methodologies and tools for projecting the continued development of COVID-19 infections.

5. Perspectives

As of March 2021, the expected "second" (Xu and Li, 2020) and "third waves" of the COVID-19 pandemic were still lingering on or on the rise, in particular in some European countries, driven by new and more infectious COVID-19 variants, including B.1.1.7 (discovered in United Kingdom in the autumn of 2020) and B.1.351 (discovered in South Africa in October 2020). Meanwhile, pharmaceutical interventions became increasingly available. Looking at recorded numbers across the ten countries addressed in our study tell an intriguing story of how the pandemic continues to manifest in quite different ways whether one is looking at, say, Europe, South America, North America or India, e.g., as affected by seasonal variations, policies and different levels of pharmaceutical and non-pharmaceutical interventions. In the end, even considering the widespread application of vaccines, implications are that the world (and the world economy) will still for considerable time have to account for and cope with a coronavirus – and the new variants thereof - that may be strongly present in some countries and/or parts of the world and under control in others. Hence, increased future international collaboration on developing and evaluating improved COVID-19 modelling and forecasting techniques, e.g., collaborative model intercomparison studies, could be of the outmost importance to help us cope with the current pandemic - and those to come.

CRediT authorship contribution statement

Martin Drews: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft, Visualization. **Pavan Kumar:** Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft. **Ram Kumar Singh:** Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft. **Manuel De La Sen:** Formal analysis, Writing – original draft, Funding acquisition. **Sati Shankar Singh:** Formal analysis, Writing – review & editing. **Ajai Kumar Pandey:** Formal analysis, Writing – review & editing. **Manoj Kumar:** Formal analysis, Writing – review & editing. **Meenu Rani:** Formal analysis, Writing – review & editing. **Prashant Kumar Srivastava:** Formal analysis, Writing – review & editing.

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.

Acknowledgments

The authors gratefully acknowledge Johns Hopkins Corona Virus Resource Center and the Worldometer for collecting COVID-19 data from official sources, including national and international authorities and organisations, for making them available for analyses, and for promoting best practices in collecting and documenting global health.

Funding

The authors acknowledge financial support from the Spanish Government, Grant RTI2018-094336-B-I00 (MCIU/AEI/FEDER, UE), and the Spanish Carlos III Health Institute, COV 20/01213.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2021.150639.

References

Annan, J.D., Hargreaves, J.C., 2011. Understanding the CMIP3 multimodel ensemble. J. Clim. 24, 4529–4538.

Bashir, M.F., et al., 2020. Correlation between climate indicators and COVID-19 pandemic in New York, USA. Sci. Total Environ. 728, 138835.

Bi, Q., et al., 2020. Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. Lancet Infect. Dis. 20, 911–919.

Biswas, K., Khaleque, A., Sen, P., 2020. Covid-19 Spread: Reproduction of Data and Prediction Using a SIR Model on Euclidean Network. arXiv, 2003.07063.

Briz-Redón, A., Serrano-Aroca, Á., 2020. A spatio-temporal analysis for exploring the effect of temperature on COVID-19 early evolution in Spain. Sci. Total Environ. 728, 138811.

Cássaro, F.A.M., Pires, L.F., 2020. Can we predict the occurrence of COVID-19 cases? Considerations using a simple model of growth. Sci. Total Environ. 728, 138834.

Castro, M., Ares, S., Cuesta, J.A., Manrubia, S., 2020. The turning point and end of an expanding epidemic cannot be precisely forecast. Proc. Natl. Acad. Sci. U. S. A. 117, 26190–26196.

CDC - Centers for Disease Control and Prevention, 2020. Covid-19 forecasts. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html. (Accessed 20 November 2020).

Ceylan, Z., 2020. Estimation of COVID-19 prevalence in Italy, Spain, and France. Sci. Total Environ. 729, 138817.

Chalkidou, K., Gorgens, M., Hutubessy, R., Teerawattananon, Y., Wilson, D., 2020. Introducing the COVID-19 Multi-model Comparison Collaboration.

Chatterjee, K., Chatterjee, K., Kumar, A., Shankar, S., 2020. Healthcare impact of COVID-19 epidemic in India: a stochastic mathematical model. Med. J. Armed Forces India. https://doi.org/10.1016/j.mjafi.2020.03.022.

Diaz-Quijano, F.A., Rodriguez-Morales, A.J., Waldman, E.A., 2020. Translating transmissibility measures into recommendations for coronavirus prevention. Rev. Saude Publica 54, 43. https://doi.org/10.11606/s1518-8787.2020054002471.

Gupta, S., Raghuwanshi, G.S., Chanda, A., 2020. Effect of weather on COVID-19 spread in the US: a prediction model for India in 2020. Sci. Total Environ. 728, 138860.

Gutowski, W.J., et al., 2016. WCRP COordinated Regional Downscaling EXperiment (CORDEX): a diagnostic MIP for CMIP6. Geosci. Model Dev. 9, 4087–4095.

Haughton, N., Abramowitz, G., Pitman, A., et al., 2014. On the generation of climate model ensembles. Clim. Dyn. 43, 2297–2308.

Holt, C.C., 1957. Forecasting seasonals and trends by exponentially weighted moving averages. ONR Research Memorandum. 52. Carnegie Institute of Technology.

Hyndman, R.J., Khandakar, Y., 2008. Automatic time series forecasting: the forecast package for R. J. Stat. Softw. 27 (3). https://doi.org/10.18637/jss.v027.i03.

IHME - Institute for Health Metrics and Evaluation. First COVID-19 Global Forecast: IHME Projects Three-Quarters of a Million Lives Could be Saved by January 1. http://www.healthdata.org/news-release/first-covid-19-global-forecast-ihme-projects-three-quarters-million-lives-could-be. Accessed on November 20, 2020.

Jewell, N.P., Lewnard, J.A., Jewell, B.L., 2020. Predictive mathematical models of the COVID-19 pandemic: underlying principles and value of projections. JAMA 323 (19). https://doi.org/10.1001/jama.2020.6585.

Johns Hopkins Corona Virus Resource Centre. https://coronavirus.jhu.edu. (Accessed 17 October 2020).

Koo, J.R., et al., 2020. Interventions to mitigate early spread of SARS-CoV-2 in Singapore: a modelling study. Lancet Infect. Dis. 20, 678–688.

Kucharski, A.J., et al., 2020. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. Lancet Infect. Dis. 20, 553–558.

Li, Q., et al., 2020. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. N. Engl. J. Med. https://doi.org/10.1056/nejmoa2001316.

Liu, W., HW, H.W.Hethcote, Levin, S.A., 1987. Dynamical behavior of epidemiological models with nonlinear incidence rates. J. Math. Biol. 25, 359–380.

Menebo, M.M., 2020. Temperature and precipitation associate with Covid-19 new daily cases: a correlation study between weather and Covid-19 pandemic in Oslo, Norway. Sci. Total Environ. 737, 139659.

Petropoulos, F., Makridakis, S., 2020. Forecasting the novel coronavirus COVID-19. PLOS ONE 15, e0231236. https://doi.org/10.1371/journal.pone.0231236.

Porter, A.T., Oleson, J.J., 2013. A path-specific SIR model for use with general latent and infectious time distributions. Biometrics. https://doi.org/10.1111/j.1541-0420.2012.01809.x.

R Core Team, 2013. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria 3-900051-07-0. http://www.R-project.org/.

Refsgaard, J.C., et al., 2014. A framework for testing the ability of models to project climate change and its impacts. Clim. Chang. 122, 271–282.

Remuzzi, A., Remuzzi, G., 2020. COVID-19 and Italy: what next? Lancet 395, 1225–1228.

Rodrigues, H.S., 2016. Application of SIR epidemiological model: new trends. Int. J. Appl. Math. Inf. 10, 92–97.

Runkle, J.D., et al., 2020. Short-term effects of specific humidity and temperature on COVID-19 morbidity in select US cities. Sci. Total Environ. 740, 140093.

Şahin, M., 2020. Impact of weather on COVID-19 pandemic in Turkey. Sci. Total Environ. 728, 138810.

Scudellari, M., 2020. How the pandemic might play out in 2021 and beyond. Nature 584, 22–25.

Singh, R.K., et al., 2020. Short-term statistical forecasts of COVID-19 infections in India. IEEE Access 8, 186932–186938.

Singh, R.K., et al., 2020. Prediction of the COVID-19 pandemic for the top 15 affected countries: advanced autoregressive integrated moving average (ARIMA) model. JMIR Public Health Surveill. https://doi.org/10.2196/19115.

Sun, J., Chen, X., Zhang, Z., et al., 2020. Forecasting the long-term trend of COVID-19 epidemic using a dynamic model. Sci. Rep. 10, 21122.

Tiwari, A., Dadhania, A.V., Ragunathrao, V.A.B., Oliveira, E.R.A., 2021. Using machine learning to develop a novel COVID-19 vulnerability index (C19VI). Sci. Total Environ. 773, 145650.

Tomar, A., Gupta, N., 2020. Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. Sci. Total Environ. 728, 138762.

Verity, R., et al., 2020. Estimates of the severity of coronavirus disease 2019: a model-based analysis. Lancet Infect. Dis. 20, 669–677.

Wangping, J., et al., 2020. Extended SIR prediction of the epidemics trend of COVID-19 in Italy and compared with Hunan, China. Front. Med. 7, 169.

Wilke, C.O., Bergstrom, C.T., 2020. Predicting an epidemic trajectory is difficult. Proc. Natl. Acad. Sci. U. S. A. 117, 28549–28551.

Winters, P.R., 1960. Forecasting sales by exponentially weighted moving averages. Manag. Sci. 6, 324–342.

Worldometer, 2020. COVID-19 Coronavirus Pandemic. https://www.worldometers.info/coronavirus/. (Accessed 17 October 2020).

Wu, J.T., Leung, K., Leung, G.M., 2019. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet 395, 689–697.

Wu, Z., McGoogan, J.M., 2020. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China. JAMA. https://doi.org/10.1001/jama.2020.2648.

Xu, S., Li, Y., 2020. Beware of the second wave of COVID-19. Lancet 395, 1321.

Yang, Z., 2020. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. J. Thorac. Dis. 12, 165.

Zheng, N., et al., 2020. Predicting COVID-19 in China using hybrid AI model. IEEE Trans. Cybern. 50, 2891–2904.