A Global-Scale Ecological Niche Model to Predict SARS-CoV-2 Coronavirus Infection Rate

Gianpaolo Coro[a,1,2,*]

[a] *Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo" – CNR, Pisa, Italy*

Abstract

COVID-19 pandemic is a global threat to human health and economy that requires urgent prevention and monitoring strategies. Several models are under study to control the disease spread and infection rate and to detect possible factors that might favour them, with a focus on understanding the correlation between the disease and specific geophysical parameters. However, the pandemic does not present evident environmental hindrances in the infected countries. Nevertheless, a lower rate of infections has been observed in some countries, which might be related to particular population and climatic conditions.

In this paper, *infection rate* of COVID-19 is modelled globally at a 0.5° resolution, using a Maximum Entropy-based Ecological Niche Model that identifies geographical areas potentially subject to a high infection rate. The model identifies locations that could favour infection rate due to their particular geophysical (surface air temperature, precipitation, and elevation) and human-related characteristics (CO_2 and population density). It was trained by facilitating data from Italian provinces that have reported a high infection rate and subsequently tested using datasets from World countries' reports. Based on this model, a *risk index* was calculated to identify the potential World countries and regions that have a high risk of disease increment.

[*] Corresponding author

Email address: coro@isti.cnr.it (Gianpaolo Coro)

[1] Telephone Number: +39 050 315 2978

[2] Fax Number: +39 050 621 3464

The distribution outputs foresee a high infection rate in many locations where real-world disease outbreaks have occurred, e.g. the Hubei province in China, and reports a high risk of disease increment in most World countries which have reported significant outbreaks (e.g. Western U.S.A.). Overall, the results suggest that a complex combination of the selected parameters might be of integral importance to understand the propagation of COVID-19 among human populations, particularly in Europe. The model and the data were distributed through Open-science Web services to maximise opportunities for re-usability regarding new data and new diseases, and also to enhance the transparency of the approach and results.

Keywords: Ecological Niche Modelling, Coronavirus, SARS-CoV-2, COVID-19, Maximum Entropy

1. Introduction

The spread of the COVID-19 pandemic, caused by the SARS-CoV-2 virus, is significantly afflicting both society and the global economy, and urgently calls for the development of systems capable of monitoring and predicting the risk of infection. The modelling of SARS-CoV-2 spread is being approached with heterogeneous methodologies, ranging from pure time series analysis to ecological models using climatic parameters, especially temperature and humidity (Giuliani et al., 2020; Nickbakhsh et al., 2020; Sajadi et al., 2020; Wang et al., 2020). However, the pandemic seems to be spreading in all World cities without evident environmental hindrances. Nevertheless, some countries are experiencing a lower rate of disease cases that might be related to their particular population and climatic conditions, but the exact effect of these conditions on infection rate is still unclear (Roser et al., 2020). Several approaches have been used to estimate the potential spatial outreach of the virus and the geophysical and climatic data that may foster disease transmission. Ecological Niche Models (ENMs) have been extensively and effectively used in this context (Davison, 2007; Misra and Kalita, 2010; Wahlgren, 2011; Costa and Peterson, 2012; Zhang et al., 2019). ENMs' aim is to predict the presence of a particular species in a geographical area by correlating species-specific occurrence records in its native habitat (presence records) with specific environmental parameters (Elith and Leathwick, 2009). The species' niche can be defined as the space within a hypervolume of numerical vectors - corresponding to environmental parameter ranges - which is correlated with the species' presence, and that fosters population persistence (*Hutchinsonian* ecological niche). Accuracy in the identification of this hypervolume can also be enhanced if the species' absence information is included in the model, as either expert-estimated or mathematically simulated information (Pearson, 2012; Chuine and Beaubien, 2008; Peterson et al., 2011; Coro et al., 2015b, 2016). ENMs have heterogeneous approaches and implementations, for example they can explicitly model a species' environmental preferences and physiological limits (*mechanistic models*), or they can automatically estimate the correlation between the parameter vectors and the species' presence (*correlative models*). Once the model has estimated the species' ecological niche, it can then project the niche characteristics across the native geographical area to reproduce the actual species' distribution, and subsequently extrapolate across another area (even at the global scale) to discover new potential suitable places for the species' persistence. Most ENMs that predict virus' spread use *correlative* approaches implemented as machine-learning or statistical models. These models can reach a high prediction accuracy on disease outreach because viruses and pandemics are known to be supported by particular geophysical characteristics and, potentially, by ecological and socioeconomic changes (Earn et al., 2000; Scheffer, 2009; Morse et al., 2012; Carlson et al., 2016; Scheffer and Van Nes, 2018). ENMs have been extensively used to discover these characteristics directly, or indirectly by tracing viruses' principal vectors (Linden, 2006; Peterson et al., 2006; Tachiiri et al., 2006; Medley, 2010; Walton et al., 2010; Fuller et al., 2013; Valiakos et al., 2014; Zhu and Peterson, 2014; Signorini et al., 2014; Samy et al., 2016). In particular, the Maximum Entropy model (MaxEnt) has been often used as an ENM due to its flexibility to work with both presence and presence/absence data scenarios (Phillips et al., 2004; Elith et al., 2011; Coro et al., 2013, 2015b). Also, MaxEnt can estimate the influence of each parameter on the identification of the niche, i.e. the most important parameters to understand a virus' preferred conditions. For these reasons, MaxEnt has often been used to trace the ecological niche of a virus based on pure geophysical parameters or human-related parameters (e.g. population density and urbanised area), and also to understand how climate change might foster the virus' spread (Peristeraki et al., 2006; Miller et al., 2012; Koch et al., 2016; Samy and Peterson, 2016).

In this paper, MaxEnt is used to estimate a *global-scale distribution of SARS-CoV-2 high infection rate*, and consequently of potential COVID-19 high spread rate. Differing from the other cited works, this model concentrates on infection *rate* rather than on absolute *spread* numbers. Further, the proposed model uses a complex combination of parameters to identify locations that could favour infection due to their particular geophysical- and human-related characteristics. As a result, it predicts a high probability of infection increase in many actual known infection areas, e.g. the Hubei province in China. The presented ENM is trained based on locations in Italy that have reported a high rate of new infections. Also, it facilitates geophysical (surface air temperature, precipitation, and elevation) and human-related (carbon dioxide and population density) data-vectors associated with these locations. The implemented model produces a probability map where higher values indicate a correlation with high infection rate; lower non-zero values indicate a lower correlation, and zero indicates unsuitable conditions for infection increase. A *risk index* is also calculated out of the produced probability distribution and identifies most World countries, with known high COVID-19 spread rate, as high-risk zones. Overall, the present work suggests that the involved parameters may play a key role in monitoring COVID-19 spread rate. The research question answered by the present work is: *Given the climatic, geophysical, and human-related parameters that other studies have individually correlated with a high COVID-19 infection rate, and that are publicly accessible, can we infer their overall weights and predict infection rate with high accuracy?*

This paper is organised in the following way: Section 2 describes the used data and the modelling approach and subsequently Section 3 reports performance evaluation metrics, model's parametrisation, and performance at predicting global high-infection-rate zones. Section 4 discusses results and conclusions, reporting the possible applications and future extensions of the presented model.

2. Material and Methods

2.1. Data

2.1.1. Data Selection Methodology and Data Availability

The methodology presented in this paper aims to be repeatable, reproducible, and re-usable for experiments on COVID-19 and other diseases. For this reason, only data which met the principles of findability, accessibility, interoperability, and re-usability were used

(FAIR data). Geospatial data accessible through representational standards, published on public geospatial services, were preferred in order to maximise their usage in the implemented model and further experiments. All used data (Table 1) were post-processed and transformed into gridded raster files, and were made available through the Zenodo open-access repository (Coro, 2020a) and the Unidata Thredds service of the D4Science e-Infrastructure (Coro, 2020b) while respecting their primary sources' citation requirements. The model used an annual data set so as not to be limited to the last winter/spring season.

2.1.2. Training and Test Data

The Italian Civil Protection Department - the national body that deals with emergency events - publishes daily updates on the number of people infected, recovered, and mortalities from COVID-19 per region and province (Italian Civil Protection Department, 2020). Data up to the end of March 2020 (Figure 1-a), i.e. the period of maximum infection rate in Italy, were used as a reference to identify locations with high infection rates on the basis of the derivative of the values. Among all available COVID-19 global reports, Italian data are particularly applicable to train an ENM because (i) Italy has been the first European country to be both heavily impacted by the virus and to study the virus, and (ii) infections in Italy have been reported on the basis of tens of thousands blanket tests. In Italy, a correlation between temperature and humidity increase and COVID-19 spread has been assessed (Italian Ministry of Health, 2020; Tuscany Regional Health Agency, 2020; Scafetta, 2020), in agreement with studies on other areas (Section 2.1.3). Indeed, despite the easing of the lockdown to lower levels and the consequential increase of human interactions, the disease spread has been decreasing from May 2020 (GEDI, 2020). At the end of April 2020, the Italian Prime Minister presented a plan of progressive lockdown level reduction, which also included possible regional restrictions in the case of a localised disease rate increase (Italian Government, 2020). However, significant increments were not observed and thus special regional restrictions were not applied. To better understand this phenomenon, Italy has started national projects to investigate the cause and effect relationships between the lockdown, environmental factors, and tourism, and to publish data and results under FAIR principles (CNR, 2020). Due to this range of considerations, Italy presents an optimum scenario to apply the proposed analysis. However, other countries are experiencing a high infection rate but have climatic conditions that are very different from the European ones. The identification of all these conditions would require more significant research and data collection initiatives.

For the scopes of the presented experiment, Italian locations with a high virus infection rate were selected, by first calculating average rates of infected people per province and then by studying the distribution of these quantities. A total of 54 provinces was selected by applying this approach (the detailed table is available in Coro (2020a)). A Chi-squared test confirmed that the distribution of infection rates could be approximated by a log-normal distribution. Consequently, provinces with a high infection rate were identified and selected as those with infection rates over the geometric mean of the rates. These data were used as reference observations of the modelled phenomenon to train an ecological niche model. It is worth noting that using average infection *rate* instead of absolute infection *counts* helps reducing a data bias due to the number of undetected cases of infection in Italy.

John Hopkins University publishes daily updates regarding COVID-19 infections and mortality statistics by collecting reports from the World countries (Dong et al., 2020). Data are given at a national scale for most countries, and at a regional scale for other countries (e.g. China, U.S.A., and Canada) (Figure 1-b). Unfortunately, reports from different countries are poorly comparable between them, given the different countries approaches to disease identification and monitoring (Reuters, 2020). Thus, mixing these data with Italian province data was not optimal for modelling. Nevertheless, global data were used as a reference to test the prediction performance of an aggregated *risk index* built upon the model's output (Section 2.3). To this aim, the countries/regions with the highest infection rates were selected using the same statistical analysis applied to Italian data, which resulted in 72 locations (the detailed table is available in Coro (2020a)).

2.1.3. Input Parameters

Surface air Temperature and Precipitation

The NASA Earth Exchange platform hosts long-term daily forecasts between 1950 and 2100 at a 0.25° resolution for minimum and maximum surface air temperature and precipitation at the surface (NASA-NEX, 2020). Forecasts come from 20 weather models developed by the Coupled Model Intercomparison Project Phase 5 (CMIP5, 2019). The D4Science e-Infrastructure hosts these data sets averaged in time and space, for 2018 and at a 0.5° resolution as gridded NetCDF-CF files (Coro and Trumpy, 2020a). In particular, data of average surface air temperature and precipitation (Figures 1-c and -d) were used due to their correlation with COVID-19 and similar viruses (Casanova et al., 2010; Chan et al., 2011; Chaudhuri et al., 2020; Ficetola and Rubolini, 2020; Ma et al., 2020; Oliveiros et al., 2020; Qi et al., 2020; Wang et al., 2020; Wu et al., 2020), and their general coupled involvement in virus ecological niche models (Patz, 1998; Fuller et al., 2013; Valiakos et al., 2014; Carlson et al., 2016). Additionally, precipitation was also used as a surrogate of humidity (Chen et al., 2012; Masunaga, 2012; Baskerville and Cobey, 2017). Italian provinces present a high variability of surface air temperature and precipitation. At the same elevation, there are temperature differences as high as 7° and precipitation differing of more than one order of magnitude. This variability increases the representativeness of Italian provinces as a training set.

Elevation

The United States National Geophysical Data Center (NGDC) hosts a global dataset of elevation and depth at a 0.33° resolution (ETOPO2, NOAA (2001)), which includes localised correction and integration of satellite, ocean sounding, and land data. Elevation has been used in several ecological niche models for viruses (Peterson et al., 2006; Miller et al., 2012; Valiakos et al., 2014) and thus was included in this experiment. The D4Science e-Infrastructure hosts a FAIR ETOPO2 dataset as a gridded NetCDF-CF file (Coro and Trumpy, 2020a,b) down-sampled at a 0.5° resolution (Figure 1-e).

2.1.4. Human-related Parameters

Carbon Dioxide

The Copernicus Atmosphere Monitoring Service hosts a global-scale uniform distribution of carbon dioxide (CO_2) flux with monthly estimates (CAMS, 2019) deriving from both human and natural activity. A FAIR dataset of averaged data from January 1979 to December 2013 with a 0.5° spatial resolution is hosted by D4Science (Coro and Trumpy, 2020a) as a gridded NetCDF-CF file (Figure 1-f). This dataset aims at combining CO_2 values preceding the higher industrialisation rate of the last decades with the natural presence of CO_2 in the soil. It summarises both natural emission and the evolution of human emission (Coro and Trumpy, 2020b). For the scope of this paper, this dataset was used as a surrogate of air pollution and human-related activity, which are generally correlated with virus spread and may foster COVID-19 spread (Lam et al., 2016; Ye et al., 2016; Clay et al., 2018; Tasci et al., 2018; Godzinski and Suarez Castillo, 2019; Liu et al., 2019; Han et al., 2020; ISPRA, 2020; BBC, 2020). Alternative parameters of CO_2, correlated with air pollution, were also tested but produced more adverse results (Section 3.2).

Population Density

Studies on complex systems' dynamics have highlighted that epidemics happen only beyond a critical threshold of population density that depends on infectivity, recovery, and mortality rates (Earn et al., 2000; Scheffer, 2009). The Center for International Earth Science Information Network openly publishes up-to-date population density data as gridded datasets with resolutions ranging from 30" to 1° (Warszawski et al., 2017). For the scopes of this paper, the Gridded Population of the World dataset - Version 4, was used at a 0.5° resolution (Figure 1-g) to include population density factors that could be correlated with infection rate.

2.2. Modelling

The experiment presented required training of MaxEnt models with several alternative parametrisations in order to identify the model with the highest performance and the best combination of parameters (Section 2.3). To this aim, the *gCube DataMiner* cloud computing platform was used. This is an open-source system that is able to process big data and offers over 400 free-to-use processes as-a-service from multiple domains (Coro et al., 2015a; Assante et al., 2019). This platform maximises the re-usability of processes through a standard Web Processing Service (WPS) interface (Coro et al., 2017). Further, DataMiner parallelises the training of models on a network of 100 machines while choosing the best computational configuration among a range of powerful multi-core virtual machines (Ubuntu 14.04.5 LTS x86 64 with 16 virtual CPUs, 16 GB of random access memory and 100 GB of storage capacity). Additionally, the system stores all trained models and their respective parametrisations under the standard and exportable Prov-O ontological format (Lebo et al., 2013). This representation allows to recover the complete set of input/output data and metadata which enable any other authorised user to reproduce and repeat an experiment (*provenance* of the computation). The Open Science concepts of re-usability of processes, and of reproducibility and repeatability of the experiments, allow the implementation of a methodology that can, in principle, be extended to analyse other diseases (Section 4). To this aim, DataMiner hosts a MaxEnt model as-a-service (CNR, 2019; Phillips et al., 2019), which can work on textual input files (CSVs) - that include pairs of coordinates related to a certain phenomenon - and FAIR input geospatial data. The WPS interface allows (i) inclusion of this service in complex workflows through a wide range of workflow management systems which support this standard (Berthold et al., 2009; QGis, 2011; Wolstencroft et al., 2013), and (ii) re-use of the service across multiple domains (Coro et al., 2013, 2015b, 2018; Coro and Trumpy, 2020b).

2.2.1. Model Description

MaxEnt is a machine learning model commonly used in ecological niche modelling (Phillips et al., 2004, 2006; Phillips and Dudik, 2008; Baldwin, 2009; Coro et al., 2015b, 2018). It simulates a probability density function $\pi(\bar{x})$ defined on real-valued vectors of parameters $\bar{x}$ taken at locations where a species occurs in its native habitat (Pearson, 2012; Coro et al., 2018). The advantage of MaxEnt with respect to other models is that it can learn from positive examples only. Thus, it does not necessarily need absence data, which are instead automatically estimated. Considering the high-infection-rate of Italian provinces as species occurrences, the parameters associated with these areas were treated as a positive example of input vectors to train the model. One drawback of MaxEnt, is that its prediction performance is very sensitive to data quality (Elith and Leathwick, 2009), an additional consideration for using only Italian data and not combining data from other countries (Reuters, 2020).

The MaxEnt training algorithm adjusts the model's internal variables so that (i) the simulated density function $\pi(\bar{x})$ is compliant with pre-calculated mean values at training-set locations and (ii) the entropy of the density function $H = -\sum \pi(\bar{x})\ ln(\pi(\bar{x}))$ is maximum for these locations (Elith et al., 2011). MaxEnt maximises the entropy function for training locations divided by the entropy values of the parameters of random points taken in the training-set area (*background points*, Phillips et al. (2006)). The model involves a linear combination of the input parameters, whose coefficients reproduce the influence of each variable on the prediction of the training set locations (*percent contribution*). Further, the model estimates the dependency of the performance on the permutation of each parameter in the training vectors (*permutation importance*).

In this experiment, MaxEnt uses the data vectors $\bar{x}$ of Italian high-infection-rate provinces (and of *background points* in Italy) to estimate the probability density $\pi(\bar{x}) = P(high - infection - rate|\bar{x})$ that a location would foster a high infection rate. To this aim, the model estimates the ratio between the probability density $f(\bar{x})$ of the vectors across Italy and the probability density in the high-infection-rate locations $f_1(\bar{x})$. The Bayes' rule defines the relation between $P(high-infection-rate|\bar{x})$, $f(\bar{x})$, and $f_1(\bar{x})$:

$$P(high-infection-rate|\bar{x}) = \frac{f_1(\bar{x})P(high-infection-rate)}{f(\bar{x})}$$

with $P(high-infection-rate)$ being the prior distribution of high-infection-rate zones in Italy (*prevalence*), fixed to 0.5 by default (i.e. no prior assumption is given). MaxEnt hypothesises that the optimal $f_1(\bar{x})$ distribution is the closest distribution to $f(\bar{x})$, because without any training-set location there would be no expectation about certain conditions over the others (i.e. $f(\bar{x})$ is a null model for $f_1(\bar{x})$). Also, the model constraints $f_1(\bar{x})$ to reflect the observations on the training set, i.e. $f_1(\bar{x})$ should estimate high probability on parameters' values close to the parameters' means over the training set. The model uses Kullback-Leibler divergence (relative entropy) to measure the distance between the two functions:

$$d(f_1(\bar{x}), f(\bar{x})) = \sum_{\bar{x}} f_1(\bar{x}) log_2 \left(\frac{f_1(\bar{x})}{f(\bar{x})} \right)$$

The aim of the training algorithm is to minimise this distance under the above constraints, which in turn maximises the entropy of the target probability density. It can be demonstrated that this characterization uniquely determines $f_1(\bar{x})$ as belonging to the following family of Gibbs distributions (Phillips et al., 2006):

$$f_1(\bar{x}) = f(\bar{x})e^{\eta(\bar{x})}$$

with $\eta(\bar{x}) = \alpha + \beta\ h(\bar{x})$; α being a normalization constant that makes $f_1(\bar{x})$ sum to 1; h being an optional transformation of the vectors $\bar{x}$ that possibly models complex relationships between parameters; β being the vector of coefficients that reports the *percent contribution* of each parameter. Thus, the ratio $f_1(\bar{x})/f(\bar{x})$ is equal to $e^{\eta(\bar{x})}$, i.e. MaxEnt needs to solve a log-linear model based on the background and training vectors to estimate the α and β parameters, which can be implemented through a penalised maximum likelihood algorithm (Phillips and Dudík, 2008).

After the training phase, the parameters' *percent contribution* can be used to select the most influential parameters for the model. This potentially allows to use MaxEnt as a filter to select those parameters carrying the highest quantity of information (Coro et al., 2015b, 2013, 2018). A MaxEnt model trained on 0.5° resolution parameters can be reasonably used to produce probability distributions at the same resolution. Given the semantics of the selected training locations, the model produced a distribution function that could be interpreted as a global-scale probability distribution for SARS-CoV-2 high infection rate.

2.3. Evaluation Metrics

The model training phase estimates the average Area Under the Curve (AUC), i.e. the integral of the Receiver Operating Characteristic (ROC) curve that plots *sensitivity* ($\frac{True\,Positives}{True\,Positives+False\,Negatives}$) against 1-*specificity* ($1-\frac{True\,Negatives}{True\,Negative+False\,Positives}$). AUC values closer to 1 indicate high classification performance of training sites. Reference cut-off thresholds on π were also calculated during the training phase (Phillips et al., 2019) and represent (i) the value balancing *omission rate* ($\frac{False\,Negatives}{True\,Positives+False\,Negatives}$) and *sensitivity* (*balanced threshold*), (ii) the value at which *sensitivity* and *specificity* are equal, and (iii) the minimum threshold at which all training locations are correctly classified as high-infection-rate areas.

In order to numerically estimate the prediction performance of the trained model, a *risk* *index* was also calculated, defined as the normalised density of non-zero MaxEnt probability locations (McGeoch et al., 2006; Coro et al., 2018) for all countries/regions reported in the global dataset of infection rates (Section 2.1). High-risk zones were identified as those with a *risk index* higher than the geometric mean of the *risk* values. Accuracy on the correct identification of *high-infection-rate countries/regions* as *high-risk zones* was calculated as $\frac{n.\,of\;high{-}infection{-}rate\;areas\;identified}{overall\;n.\,of\;high{-}infection{-}rate\;areas}$. Moreover, agreement between high-risk zones' classification and high-infection-rate country/region reports was calculated using Cohen's Kappa (Cohen et al., 1960). This statistical coefficient estimates the agreement between the two classifications with respect to purely random classifications (agreement by chance). An overall interpretation of this value was assigned using Fleiss' tables (Fleiss, 1971).

3. Results

3.1. Global-scale distribution and Performance

The MaxEnt model was trained using different combinations of parameters associated with Italian locations reporting a high rate of infections up to the end of March 2020 (Section 2.1). Training the model on all parameters produced the highest AUC and optimal estimates for the three model's thresholds (Table 2-a). When the model was trained with any other parameter subset, AUC resulted lower. This property indicates that all parameters bring useful information to estimate training set locations correctly. Nevertheless, the *percent contribution* and *permutation importance* of carbon dioxide, surface air temperature, and precipitation are much higher than the ones of elevation and population density (Table 3). The model using all parameters also indicates a correlation with high infection rate for particular parameter ranges (i.e. the boundaries of the niche hypervolume): CO_2 has the highest correlation around 0.03 (0.01;0.08) $g\ C\ m^{-2}\ day^{-1}$ (*moderate-high*), air temperature around 11.8 (8.0;16.0) °C (*moderate-low*), and precipitation around 0.3 (0.2;0.45) $10^{-4}\ kg\ m^{-2}\ s^{-1}$ (*moderate*).

The model was projected at the global scale to produce a global infection-rate probability distribution at a 0.5° resolution (Figure 2). For each cell, this map reports the probability that the cell has suitable conditions for infection increase. Locations with a value higher than the balanced threshold ($\pi(\bar{x}) \geq 0.4$) can be classified as high-infection-rate locations, whereas the other two thresholds indicate medium infection-rate ($0.1 \leq \pi(\bar{x}) < 0.4$) and low infection-rate ($0.008 \leq \pi(\bar{x}) < 0.1$) locations. Zero probability locations indicate unsuitable areas for an infection rate increase.

As a qualitative evaluation, it can be observed that the model correctly and precisely identifies the locations of real World high infection rates, e.g. the Hubei Chinese region, Western United States, and most of Europe. Instead, wrongly classified places are, for example, Peru and Brazil, that have parameter ranges out of the niche hypervolume. The identification of the climatic/geophysical parameters fostering infection rate increase in these countries would require further research, based on a more extensive and globally shared data collection (Section 3.3).

In order to quantify the prediction accuracy of the map, the *risk index* was used to select high-risk zones and compare them with global reports of high infection rates (Figure 3 and Table 2-b). Accuracy at predicting high-infection-rate countries/region reached 77.25%, and the overall agreement (0.46) was *good* according to Fleiss' classification. This result indicates that most countries/regions are correctly and non-randomly classified, and thus the model has extracted a correct characterisation of the actual risk of infection increase based on the considered parameters.

3.2. The weight of the CO_2 parameter

The high correlation of CO_2 with high infection rate requires a further investigation, starting from the correlation between air pollution and COVID-19 spread (Section 2.1.4). The Copernicus Atmosphere Monitoring Service provides FAIR data correlated with greenhouse gas concentration and fluxes, i.e. methane (CH_4), nitrous oxide (N_2O), and CO_2 (CAMS, 2020). The CH_4 and N_2O influence on prediction performance was evaluated by substituting these parameters to CO_2 in the all-parameter model (*individual* models), and then by using them together with CO_2 (*mixed* model). The aggregated data used for this analysis were published as FAIR data on Zenodo (Coro, 2020a). Executing the MaxEnt individual models revealed that CH_4 and N_2O have a much lower *percent contribution* (~52% for both models) to infection rate prediction than CO_2 (87.2%). Furthermore, their individual models reported a lower AUC (0.90 v.s. 0.994 of the CO_2 model). However, in these models, CH_4 and N_2O were always the parameters having the highest *percent contribution* to infection rate prediction. This property indicates that the parameters correlated with greenhouse gases concentration are of high importance for prediction accuracy, which confirms the correlation between air pollution and infection rate highlighted by other studies (Section 2.1.4). The mixed model further confirmed this result because it gained the same performance as the CO_2 individual model to predict high risk zones (77.25%). However, the mixed model reported a much higher *percent contribution* of CO_2 (85.9%) than of CH_4 (0.4%) and N_2O (0.4%). This result indicates that CH_4 and N_2O are not adding a substantially more predictive information than CO_2. Overall, this analysis indicates that CO_2 is the correct choice to represent air pollution in the experiment.

3.3. Training and input data completeness

In order to evaluate if Italian provinces were a sufficient representative training set for the reported experiment, the all-parameter MaxEnt model was executed by incrementally adding more World areas to the training set. First, the geographical areas of large cities correctly predicted by the original model were added, i.e. Madrid, London, Istanbul, Buenos Aires. This operation did not change the model's risk prediction performance (77.25%), which indicates that Italian provinces are strong representation of the correctly detected World cities. As an additional step, World city areas that were wrongly predicted by the original model were incrementally introduced, i.e. São Paulo, Lima, Santiago de Chile, Guayaquil. This process produced a continuously decreasing AUC, also if CH_4 and N_2O were used instead of CO_2. When involving these World cities, one major effect on the parameter ranges was a change in the upper confidence limit, which increased for temperature (from 16.0 to 18.8 °C) and precipitation (from 0.45 to 0.6 $10^{-4}\ kg\ m^{-2}\ s^{-1}$) and the decreased for CO_2 (from 0.08 to 0.05 $g\ C\ m^{-2}\ day^{-1}$). The decreasing AUC, indicates that these ranges are not able to make the model cover all the areas of the training set. This result indicates that the used input parameters are insufficient to understand the infection rate increase in these areas, independent of the use of Italian provinces as the training set.

4. Discussion and Conclusions

This paper has presented a methodology to estimate a geographical probability distribution of *high infection rate* for SARS-CoV-2, based on geophysical and human-related parameters. A *risk index* has been proposed based on this probability distribution, to identify global countries and regions that would mostly favour a high infection rate. A *good* concurrence with country-reported data and a moderate-high accuracy at predicting high-infection-rate countries/regions indicates that the model was able to identify real conditions of increased infection rate in many World areas. Generally, the model indicates a high infection rate in areas characterised by an annual moderate-high level of CO_2, moderate-low temperatures, and moderate precipitation. The most notable result is that, although the model was trained only with Italian cities, it assigns a high-infection-rate probability and a high-risk classification to most real World scenarios where a high infection rate has been actually reported. Also, the results indicate that climatic parameters such as air temperature and precipitation (or air humidity) play a critical role at defining locations that may be subject to a high infection rate. The model also indicates a temperature range which other studies have also correlated with the spread of COVID-19 (Sajadi et al., 2020). Additionally, estimated high-rates in moderate-precipitation regions might be related to reduced transmission in high-humidity zones (Wang et al., 2020). Carbon dioxide is the most influential parameter, which is correlated directly with pollution (which concurs with COVID-19 spread, Han et al. (2020)) and indirectly with population density. Correlation with population density could be one reason for the lower influence of this parameter on prediction performance. However, the fact that all parameters are necessary to achieve the optimal model performance indicates that they all contain complementary information. Thus, population density is not entirely covered by CO_2. Indeed, it affirms the complex system dynamics theory that if a population is vulnerable to a virus and its density exceeds a threshold, an epidemic will occur (Scheffer, 2009). In the case of

SARS-CoV-2, the presented results indicate a likely scenario where, after this threshold, population density does not influence infection rate anymore. This observation is valid in Italy, where provinces with population densities distant of almost two orders of magnitude have reported similar infection rates for a long period (e.g. Lucca and Naples). As for elevation, the model indicates that this is not a discriminant feature, as also demonstrated by the variability in the altitudes of high-infection-rate Italian provinces. However, elevation brings some information to the model - probably related to drier weather conditions - because without this parameter the model's AUC decreases.

Currently, the complete set of parameters correlated with COVID-19 infection rate increase remains unknown. The reported results indicate that the used parameters are sufficient to predict the situation in Europe and in many World countries, however there are additional unknown factors to be investigated in the misidentified countries (e.g. Brazil, Ecuador, and Peru). The identification of all these factors is a broader question that goes beyond this paper and would require on-the-field data collection and a global-scale effort, also to make data available under FAIR principles.

The proposed Open Science-oriented methodology is quickly reusable on new infections and epidemics, for example, to predict the risk that a particular country will be subject to a high rate of cases of a new infection. Also, the results may be the basis of other models that may refine the resolution of the presented model and revise the parameters used. One fundamental step is to collect and prepare FAIR data correlated to infection rate as open-access standardised geospatial datasets. The D4Science e-Infrastructure can be used freely and openly to this aim. Moreover, the Maximum Entropy process was published as a free-to-use service (CNR, 2019) intended for global health-care systems and epidemic prevention organizations, and for possibly contributing to COVID-19 spread control.

Overall, the presented results clearly indicate and identify that the influence of geophysical, climatic, and human-related parameters on COVID-19 infection rate should be further investigated. As a future extension, the model will be enhanced by increasing the projection resolution to 0.1° on specific areas to produce regional-scale distributions. The corresponding cloud computing service will be used to (i) explore a more extensive set of parameters taken from open-access repositories, (ii) understand the importance of climatic factors with respect to human-related factors in COVID-19 infection rate, and (iii) detect seasonal trends.

Acknowledgments

The author wants to thank Dr. Mattew Walsh, Dr. Chiara Magliozzi, and the Journal's Reviewers for revising the paper and giving valuable suggestions. The author acknowledges the European Open Science Cloud COVID-19 Fast Track Funding for supporting this research.

References

Assante, M., Candela, L., Castelli, D., Cirillo, R., Coro, G., Frosini, L., Lelii, L., Mangiacrapa, F., Pagano, P., Panichi, G., et al., 2019. Enacting open science by d4science. Future Generation Computer Systems 101, 555–563.

Baldwin, R. A., 2009. Use of maximum entropy modeling in wildlife research. Entropy 11 (4), 854–866.

Baskerville, E. B., Cobey, S., 2017. Does influenza drive absolute humidity? Proceedings of the National Academy of Sciences 114 (12), E2270–E2271.

BBC, 2020. How air pollution exacerbates covid-19. Online publication available at https://www.bbc.com/future/article/20200427-how-air-pollution-exacerbates-covid-19.

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Thiel, K., Wiswedel, B., 2009. Knime-the konstanz information miner: version 2.0 and beyond. AcM SIGKDD explorations Newsletter 11 (1), 26–31.

CAMS, 2019. Flux inversion reanalysis of global carbon dioxide - fluxes and atmospheric concentrations. https://atmosphere.copernicus.eu/catalogue#/product/urn:x-wmo:md:int.ecmwf::copernicus:cams:prod:rean:co2:pid286.

CAMS, 2020. greenhouse gas fluxes. https://atmosphere.copernicus.eu/greenhouse-gases-supplementary-products.

Carlson, C. J., Dougherty, E. R., Getz, W., 2016. An ecological assessment of the pandemic threat of zika virus. PLoS neglected tropical diseases 10 (8).

Casanova, L. M., Jeon, S., Rutala, W. A., Weber, D. J., Sobsey, M. D., 2010. Effects of air temperature and relative humidity on coronavirus survival on surfaces. Appl. Environ. Microbiol. 76 (9), 2712–2717.

Chan, K., Peiris, J., Lam, S., Poon, L., Yuen, K., Seto, W., 2011. The effects of temperature and relative humidity on the viability of the sars coronavirus. Advances in virology 2011.

Chaudhuri, S., Basu, S., Kabi, P., Unni, V. R., Saha, A., 2020. Modeling ambient temperature and relative humidity sensitivity of respiratory droplets and their role in determining growth rate of covid-19 outbreaks. arXiv preprint arXiv:2004.10929.

Chen, M.-J., Lin, C.-Y., Wu, Y.-T., Wu, P.-C., Lung, S.-C., Su, H.-J., 2012. Effects of extreme precipitation to the distribution of infectious diseases in taiwan, 1994–2008. PloS one 7 (6).

Chuine, I., Beaubien, E., 2008. Phenology is a major determinant of tree species range. Ecology Letters 4 (5), 500–510.

Clay, K., Lewis, J., Severnini, E., 2018. Pollution, infectious disease, and mortality: evidence from the 1918 spanish influenza pandemic. The Journal of Economic History 78 (4), 1179–1209.

CMIP5, 2019. Coupled Model Intercomparison Project Phase 5. pcmdi.llnl.gov/mips/cmip5/.

CNR, 2019. Maximum Entropy Model Web Processing Service. https://services.d4science.org/group/biodiversitylab/data-miner?OperatorId=org.gcube.dataanalysis.wps.statisticalmanager.synchserver.mappedclasses.transducerers.MAX_ENT_NICHE_MODELLING.

CNR, 2020. {The Snapshot CNR Inter-Departmental Project}. https://www.cnr.it/it/news/9418/snapshot-uno-sguardo-all-ambiente-marino-durante-e-dopo-la-pandemia.

Cohen, J., et al., 1960. A coefficient of agreement for nominal scales. Educational and psychological measurement 20 (1), 37–46.

Coro, G., 2020a. Suitability Map of COVID-19 Virus Spread. Data published on Zenodo Repository https://zenodo.org/record/3833230.

Coro, G., 2020b. Thredds Repository of COVID-19 data on the D4Science e-Infrastructure. Accessible at https://thredds.d4science.org/thredds/catalog/public/netcdf/covid-19/catalog.html.

Coro, G., Candela, L., Pagano, P., Italiano, A., Liccardo, L., 2015a. Parallelizing the execution of native data mining algorithms for computational biology. Concurrency and Computation: Practice and Experience 27 (17), 4630–4644.

Coro, G., Magliozzi, C., Berghe, E. V., Bailly, N., Ellenbroek, A., Pagano, P., 2016. Estimating absence locations of marine species from data of scientific surveys in obis. Ecological Modelling 323, 61–76.

Coro, G., Magliozzi, C., Ellenbroek, A., Pagano, P., 2015b. Improving data quality to build a robust distribution model for architeuthis dux. Ecological Modelling 305, 29–39.

Coro, G., Pagano, P., Ellenbroek, A., 2013. Combining simulated expert knowledge with neural networks to produce ecological niche models for latimeria chalumnae. Ecological modelling 268, 55–63.

Coro, G., Panichi, G., Scarponi, P., Pagano, P., 2017. Cloud computing in a distributed e-infrastructure using the web processing service standard. Concurrency and Computation: Practice and Experience 29 (18), e4219.

Coro, G., Trumpy, E., 2020a. Predicting geographical suitability of geothermal power plants. Journal of Cleaner Production, 121874.
URL http://www.sciencedirect.com/science/article/pii/S0959652620319211

Coro, G., Trumpy, E., 2020b. Predicting geographical suitability of geothermal power plants. Journal of Cleaner Production (under publication).

Coro, G., Vilas, L. G., Magliozzi, C., Ellenbroek, A., Scarponi, P., Pagano, P., 2018. Forecasting the ongoing invasion of lagocephalus sceleratus in the mediterranean sea. Ecological Modelling 371, 37–49.

Costa, J., Peterson, A. T., 2012. Ecological niche modeling as a tool for understanding distributions and interactions of vectors, hosts, and etiologic agents of chagas disease. In: Recent advances on model hosts. Springer, pp. 59–70.

Davison, A. J., 2007. Overview of classification. Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis, 3–9.

Dong, E., Du, H., Gardner, L., 2020. An interactive web-based dashboard to track covid-19 in real time. The Lancet Infectious Diseases. https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30120-1/fulltext.

Earn, D. J., Rohani, P., Bolker, B. M., Grenfell, B. T., 2000. A simple model for complex dynamical transitions in epidemics. Science 287 (5453), 667–670.

Elith, J., Leathwick, J. R., 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology, Evolution, and Systematics 40 (1), 677–697.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., Yates, C. J., Jan. 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions 17 (1), 43–57.

Ficetola, G. F., Rubolini, D., 2020. Climate affects global patterns of covid-19 early outbreak dynamics. medRxiv.

Fleiss, J. L., 1971. Measuring nominal scale agreement among many raters. Psychological bulletin 76 (5), 378.

Fuller, T. L., Gilbert, M., Martin, V., Cappelle, J., Hosseini, P., Njabo, K. Y., Aziz, S. A., Xiao, X., Daszak, P., Smith, T. B., 2013. Predicting hotspots for influenza virus reassortment. Emerging infectious diseases 19 (4), 581.

GEDI, 2020. Gedi group visual lab - coronavirus data and analysis. `https://lab.gedidigital.it/gedi-visual/2020/coronavirus-i-contagi-in-italia/`.

Giuliani, D., Dickson, M. M., Espa, G., Santi, F., 2020. Modelling and predicting the spread of coronavirus (covid-19) infection in nuts-3 italian regions. arXiv preprint arXiv:2003.06664.

Godzinski, A., Suarez Castillo, M., 2019. Short-term health effects of public transport disruptions: air pollution and viral spread channels. Ideas online publication. https://ideas.repec.org/p/nse/doctra/g2019-03.html.

Han, Y., Lam, J. C., Li, V. O., Guo, P., Zhang, Q., Wang, A., Crowcroft, J., Wang, S., Fu, J., Gilani, Z., et al., 2020. The effects of outdoor air pollution concentrations and lockdowns on covid-19 infections in wuhan and other provincial capitals in china. Online publication available at https://www.preprints.org/manuscript/202003.0364/v1.

ISPRA, 2020. Information on the relationship between air pollution and the spread of covid-19. Online publication available at https://www.isprambiente.gov.it/en/news/information-on-the-relationship-between-air-pollution-and-the-spread-o set_language=en.

Italian Civil Protection Department, 2020. Interface for browsing and downloading COVID-19 data. Accessible at http://opendatadpc.maps.arcgis.com/apps/opsdashboard/index.html#/b0c68bce2cce478eaac82fe38d4138b1.

Italian Government, 2020. Decreto del Presidente del Consiglio dei ministri della Repubblica Italiana - 26 Apr. 2020. http://www.governo.it/sites/new.governo.it/files/Dpcm_img_20200426.pdf.

Italian Ministry of Health, 2020. Faq on covid-19. http://www.salute.gov.it/

portale/malattieInfettive/dettaglioFaqMalattieInfettive.jsp?lingua=italiano&id=228.

Koch, L. K., Cunze, S., Werblow, A., Kochmann, J., Dörge, D. D., Mehlhorn, H., Klimpel, S., 2016. Modeling the habitat suitability for the arbovirus vector aedes albopictus (diptera: Culicidae) in germany. Parasitology research 115 (3), 957–964.

Lam, H. C.-y., Li, A. M., Chan, E. Y.-y., Goggins, W. B., 2016. The short-term association between asthma hospitalisations, ambient temperature, other meteorological factors and air pollutants in hong kong: a time-series study. Thorax 71 (12), 1097–1109.

Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., Zhao, J., 2013. Prov-o: The prov ontology. W3C Recommendation 30.

Linden, A., 2006. Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (roc) analysis. Journal of evaluation in clinical practice 12 (2), 132–139.

Liu, X.-X., Li, Y., Qin, G., Zhu, Y., Li, X., Zhang, J., Zhao, K., Hu, M., Wang, X.-L., Zheng, X., 2019. Effects of air pollutants on occurrences of influenza-like illness and laboratory-confirmed influenza in hefei, china. International journal of biometeorology 63 (1), 51–60.

Ma, Y., Zhao, Y., Liu, J., He, X., Wang, B., Fu, S., Yan, J., Niu, J., Zhou, J., Luo, B., 2020. Effects of temperature variation and humidity on the death of covid-19 in wuhan, china. Science of The Total Environment, 138226.

Masunaga, H., 2012. Short-term versus climatological relationship between precipitation and tropospheric humidity. Journal of climate 25 (22), 7983–7990.

McGeoch, M. A., Chown, S. L., Kalwij, J. M., 2006. A global indicator for biological invasion. Conservation Biology 20 (6), 1635–1646.

Medley, K. A., 2010. Niche shifts during the global invasion of the asian tiger mosquito, aedes albopictus skuse (culicidae), revealed by reciprocal distribution models. Global ecology and biogeography 19 (1), 122–133.

Miller, R. H., Masuoka, P., Klein, T. A., Kim, H.-C., Somer, T., Grieco, J., 2012. Ecological niche modeling to estimate the distribution of japanese encephalitis virus in asia. PLoS neglected tropical diseases 6 (6).

Misra, U. K., Kalita, J., 2010. Overview: japanese encephalitis. Progress in neurobiology 91 (2), 108–120.

Morse, S. S., Mazet, J. A., Woolhouse, M., Parrish, C. R., Carroll, D., Karesh, W. B., Zambrana-Torrelio, C., Lipkin, W. I., Daszak, P., 2012. Prediction and prevention of the next pandemic zoonosis. The Lancet 380 (9857), 1956–1965.

NASA-NEX, 2020. The NASA Earth Exchange Platform. `nex.nasa.gov`.

Nickbakhsh, S., Ho, A., Marques, D. F., McMenamin, J., Gunson, R. R., Murcia, P., 2020. Epidemiology of seasonal coronaviruses: Establishing the context for covid-19 emergence. medRxiv.

NOAA, 2001. ETOPO2 Global 2 Arc-minute Ocean Depth and Land Elevation from the US National Geophysical Data Center (NGDC). Available at https://doi.org/10.5065/D6668B75.

Oliveiros, B., Caramelo, L., Ferreira, N. C., Caramelo, F., 2020. Role of temperature and humidity in the modulation of the doubling time of covid-19 cases. medRxiv.

Patz, J. A., 1998. Predicting key malaria transmission factors, biting and entomological inoculation rates, using modelled soil moisture in kenya. Tropical Medicine & International Health 3 (10), 818–827.

Pearson, R. G., 2012. Species distribution modeling for conservation educators and practitioners. Synthesis. American Museum of Natural History. Available at http://ncep.amnh.org.

Peristeraki, P., Lazarakis, G., Skarvelis, C., Georgiadis, M., Tserpes, G., 2006. Additional records on the occurrence of alien fish species in the eastern mediterranean sea. Mediterranean Marine Science 7 (2), 61–66.

Peterson, A., Soberon, J., Pearson, R., Anderson, R., Martinez-Meyer, E., Nakamura, M., Araujo, M., 2011. Ecological Niches and Geographic Distributions (MPB-49). Vol. 49. Princeton University Press.

Peterson, A. T., Lash, R. R., Carroll, D. S., Johnson, K. M., 2006. Geographic potential for outbreaks of marburg hemorrhagic fever. The American journal of tropical medicine and hygiene 75 (1), 9–15.

Phillips, S. J., Anderson, R. P., Schapire, R. E., 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190 (3-4), 231–259.

Phillips, S. J., Dudik, M., 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31, 161–175.

Phillips, S. J., Dudík, M., Schapire, R. E., 2004. A maximum entropy approach to species distribution modeling. In: Proceedings of the twenty-first international conference on Machine learning. ACM, p. 83.

Phillips, S. J., Dudík, M., 2008. Modeling of species distributions with maxent: new extensions and a comprehensive evaluation. Ecography 31 (2), 161–175.

Phillips, S. J., Miroslav, D., E., S. R., 2019. Maxent software for modeling species niches and distributions (version 3.4.1). http://biodiversityinformatics.amnh.org/open_source/maxent/.

QGis, D., 2011. Quantum gis geographic information system. Open Source Geospatial Foundation Project 45.

Qi, H., Xiao, S., Shi, R., Ward, M. P., Chen, Y., Tu, W., Su, Q., Wang, W., Wang, X., Zhang, Z., 2020. Covid-19 transmission in mainland china is associated with temperature and humidity: A time-series analysis. Science of the Total Environment, 138778.

Reuters, 2020. Special Report: Italy and South Korea virus outbreaks reveal disparity in deaths and tactics. Accessible at https://www.reuters.com/article/us-health-coronavirus-response-specialre/special-report-italy-and-south-korea-virus-outbreaks-reveal-disparity-in-deaths-and-tactics-idUSKBN20Z27P.

Roser, M., Ritchie, H., Ortiz-Ospina, E., 2020. Coronavirus Disease (COVID-19) Statistics and Research. Online publication https://ourworldindata.org/coronavirus.

Sajadi, M. M., Habibzadeh, P., Vintzileos, A., Shokouhi, S., Miralles-Wilhelm, F., Amoroso, A., 2020. Temperature and latitude analysis to predict potential spread and seasonality for covid-19. Available at SSRN 3550308.

Samy, A. M., Peterson, A. T., 2016. Climate change influences on the global potential distribution of bluetongue virus. PloS one 11 (3).

Samy, A. M., Thomas, S. M., Wahed, A. A. E., Cohoon, K. P., Peterson, A. T., 2016. Mapping the global geographic potential of zika virus spread. Memorias do Instituto Oswaldo Cruz 111 (9), 559–560.

Scafetta, N., 2020. Distribution of the sars-cov-2 pandemic and its monthly forecast based on seasonal climate patterns. International Journal of Environmental Research and Public Health 17 (10), 3493.

Scheffer, M., 2009. Critical transitions in nature and society. Vol. 16. Princeton University Press.

Scheffer, M., Van Nes, E. H., 2018. Seeing a global web of connected systems. Science 362 (6421), 1357–1357.

Signorini, M., Cassini, R., Drigo, M., di Regalbono, A. F., Pietrobelli, M., Montarsi, F., Stensgaard, A.-S., 2014. Ecological niche model of phlebotomus perniciosus, the main vector of canine leishmaniasis in north-eastern italy. Geospatial health, 193–201.

Tachiiri, K., Klinkenberg, B., Mak, S., Kazmi, J., 2006. Predicting outbreaks: a spatial risk assessment of west nile virus in british columbia. International Journal of Health Geographics 5 (1), 21.

Tasci, S. S., Kavalci, C., Kayipmaz, A. E., 2018. Relationship of meteorological and air pollution parameters with pneumonia in elderly patients. Emergency medicine international 2018.

Tuscany Regional Health Agency, 2020. Recommendations for health operators. `https://www.ars.toscana.it/2-articoli/4276-nuovo-coronavirus-covid-19-informazioni-buone-pratiche-raccomanda` `html`.

Valiakos, G., Papaspyropoulos, K., Giannakopoulos, A., Birtsas, P., Tsiodras, S., Hutchings, M. R., Spyrou, V., Pervanidou, D., Athanasiou, L. V., Papadopoulos, N., et al., 2014. Use of wild bird surveillance, human case data and gis spatial analysis for predicting spatial distributions of west nile virus in greece. PLoS One 9 (5).

Wahlgren, J., 2011. Influenza a viruses: an ecology review. Infection ecology & epidemiology 1 (1), 6004.

Walton, N. A., Poynton, M. R., Gesteland, P. H., Maloney, C., Staes, C., Facelli, J. C., 2010. Predicting the start week of respiratory syncytial virus outbreaks using real time weather variables. BMC medical informatics and decision making 10 (1), 68.

Wang, J., Tang, K., Feng, K., Lv, W., 2020. High temperature and high humidity reduce the transmission of covid-19. Available at SSRN 3551767.

Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O., Zhang, X., Tang, Q., Pan, M., Tang, Y., Tang, Q., et al., 2017. Center for international earth science information networkciesincolumbia university.(2016). gridded population of the world, version 4 (gpwv4): Population density. palisades. ny: Nasa socioeconomic data and applications center (sedac). Atlas of Environmental Risks Facing China Under Climate Change, 228.

Wolstencroft, K., Haines, R., Fellows, D., Williams, A., Withers, D., Owen, S., Soiland-Reyes, S., Dunlop, I., Nenadic, A., Fisher, P., et al., 2013. The taverna workflow suite: designing and executing workflows of web services on the desktop, web or in the cloud. Nucleic acids research 41 (W1), W557–W561.

Wu, Y., Jing, W., Liu, J., Ma, Q., Yuan, J., Wang, Y., Du, M., Liu, M., 2020. Effects of temperature and humidity on the daily new cases and new deaths of covid-19 in 166 countries. Science of the Total Environment, 139051.

Ye, Q., Fu, J.-f., Mao, J.-h., Shang, S.-q., 2016. Haze is a risk factor contributing to the rapid spread of respiratory syncytial virus in children. Environmental Science and Pollution Research 23 (20), 20178–20185.

Zhang, J., Yoon, K.-J., Zimmerman, J. J., 2019. Overview of viruses. Diseases of Swine, 425–437.

Zhu, G., Peterson, A. T., 2014. Potential geographic distribution of the novel avian-origin influenza a (h7n9) virus. PLoS One 9 (4).

Data	**Primary Source**
Infection per Italian Province	Italian Civil Protection Department
World Infections	John Hopkins University
Surface Air Temperature	NASA Earth Exchange Platform
Precipitation	NASA Earth Exchange Platform
Elevation	United Stated National Geophysical Data Center
Carbon Dioxide	Copernicus Atmosphere Monitoring Service
World Population Density	Center for International Earth Science Information Network

Table 1: Summary of all used data along with their primary sources. Details about how these data were accessed and post-processed are given in the article.

Model Performance - a	
AUC	0.994
Balanced omission-sensitivity threshold	0.4
Equal training sensitivity and specificity threshold	0.1
Minimum training presence threshold	0.008
Risk Index Performance - b	
Accuracy	77.25%
Kappa	0.46
Kappa Interpretation	Good

Table 2: Report of (a) the performance and optimal thresholds of the trained MaxEnt model, and (b) the performance of the *risk index* on the identification of global high-infection-rate countries/regions.

Parameter name	Percent contribution	Permutation importance (%)
Carbon Dioxide	87.2	52.8
Surface Air Temperature	7.6	40
Precipitation	5.3	6.9
Elevation	0.01	0.01
Population Density	0.01	0.2

Table 3: Percent contribution and permutation importance of the parameters involved in the presented experiment, as estimated by the optimal Maximum Entropy model.

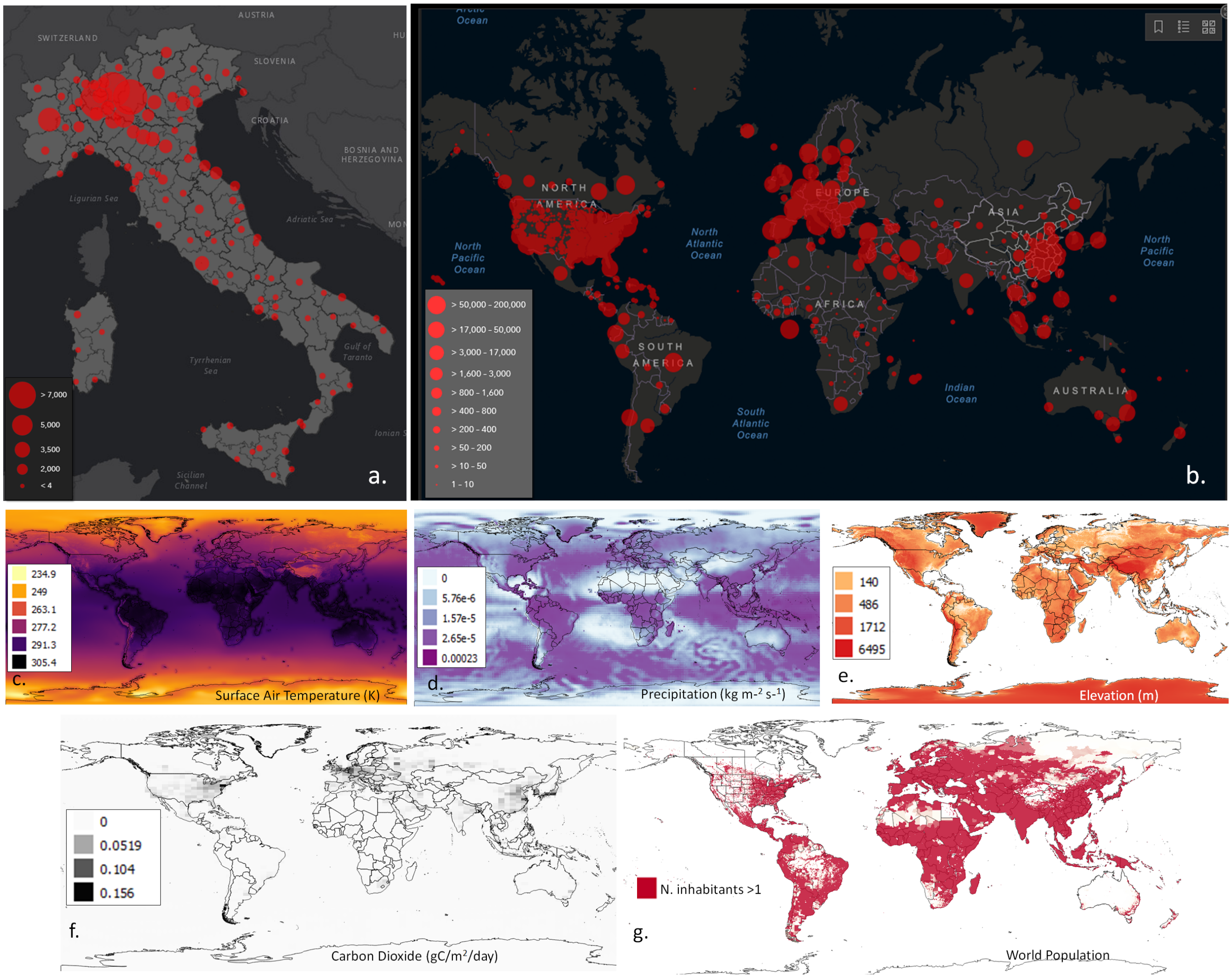

Figure 1: Visual comparison of the global-scale data used in the presented model: (a) number of infections in Italian provinces (31 March 2020), (b) global infections (31 March 2020), (c) surface air temperature, (d) precipitation, (e) elevation, (f) carbon dioxide, (g) World population.

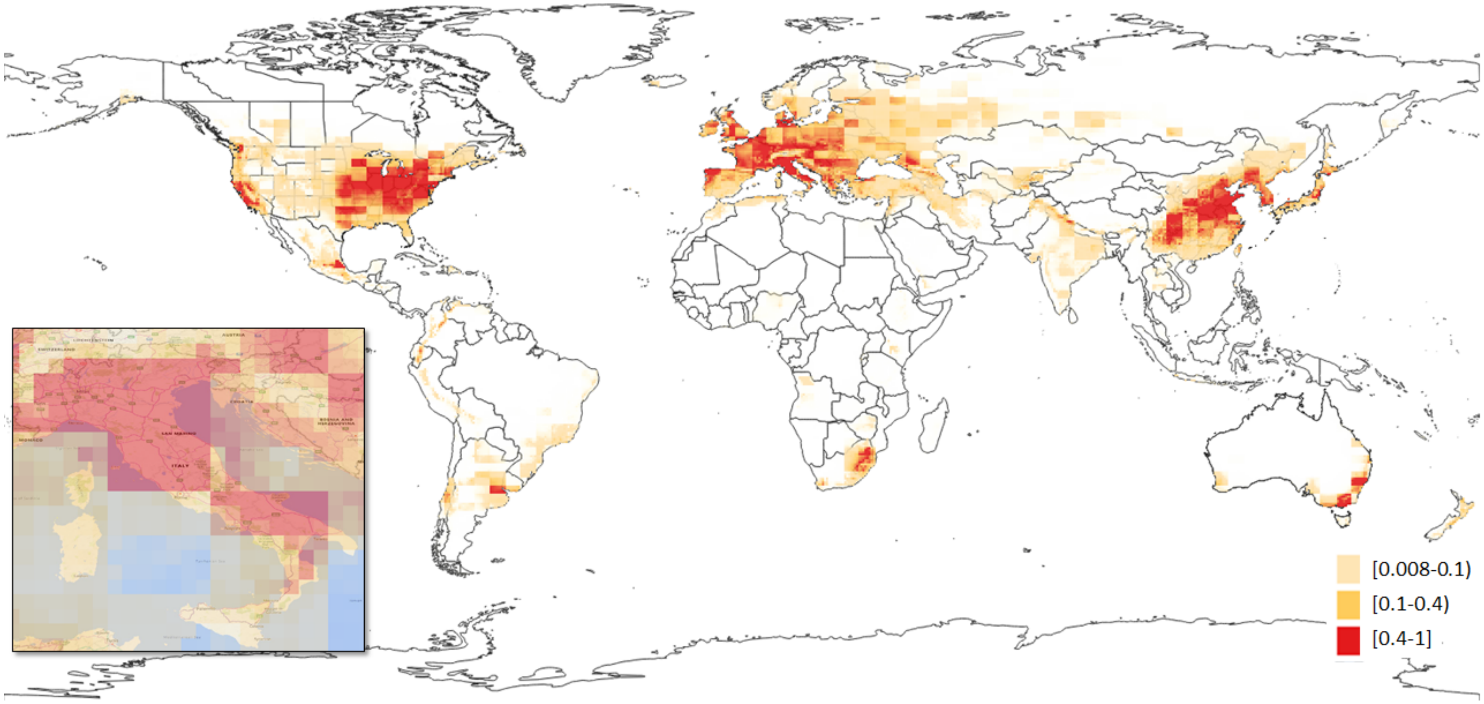

Figure 2: Global-scale probability distribution of SARS-CoV-2 infection rate produced by the presented model, with Italy magnified at the lower-left hand side.

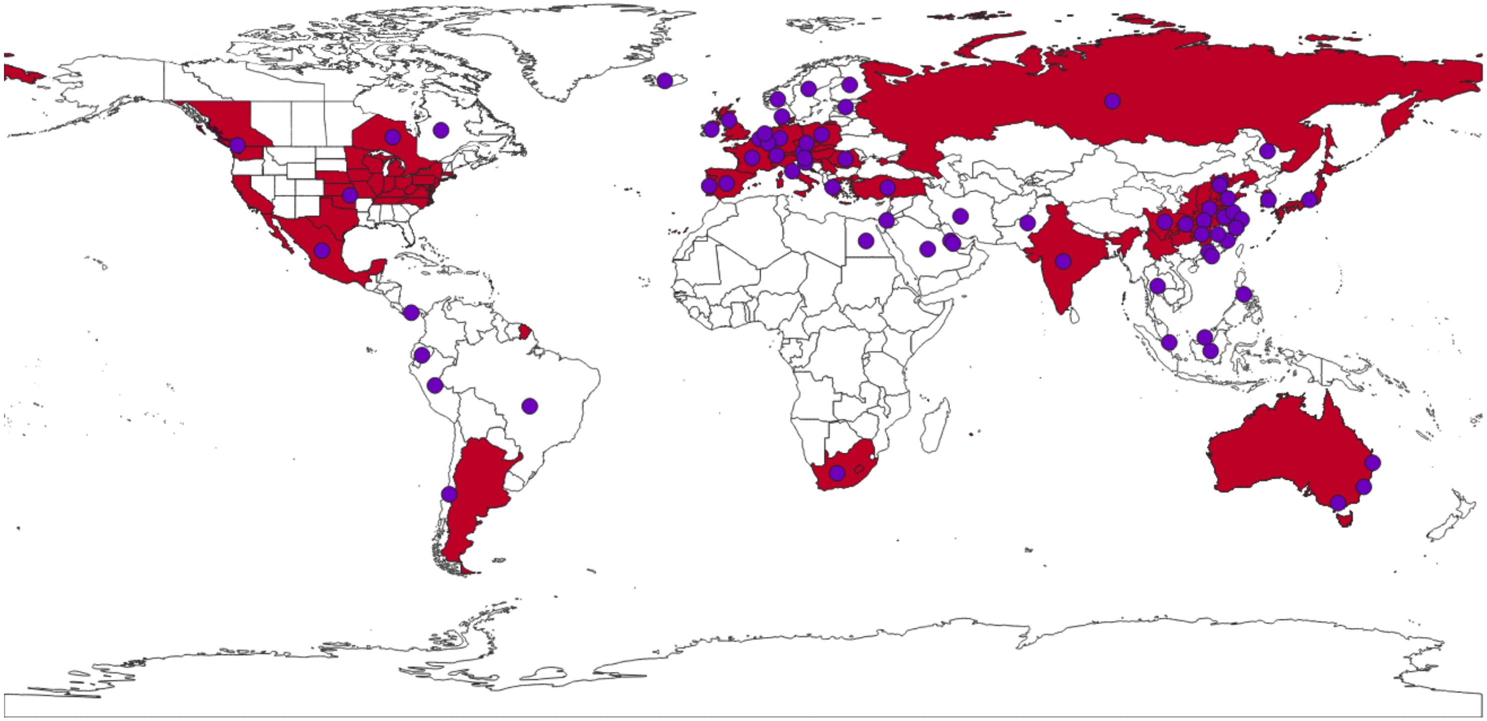

Figure 3: Overlap between estimated high-infection-rate risk zones (coloured countries/regions) and actual reported high-infection-rate countries/regions (circles).