Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: http://ees.elsevier.com

Comment on "A global-scale ecological niche model to predict SARS-CoV-2 coronavirus infection rate", author Coro

Andrea Contina [a,*], Scott W. Yanco [a], Allison K. Pierce [a], Michelle DePrenger-Levin [a,b], Michael B. Wunder [a], Andreas M. Neophytou [c], C. Phoebe Lostroh [d], Richard J. Telford [e], Blas M. Benito [f], Joseph Chipperfield [g], Robert B. O'Hara [h], Colin J. Carlson [i,*]

[a] University of Colorado Denver, Department of Integrative Biology, Science Building 2074, Denver, CO 80217, USA
[b] Denver Botanic Gardens, Research and Conservation, 909 York Street, Denver, CO 80206, USA
[c] Colorado State University, Department of Environmental and Radiological Health Sciences, Fort Collins, CO 80523, USA
[d] Colorado College Department of Molecular Biology, 14 E Cache La Poudre Street, Colorado Springs, CO 80903, USA
[e] Department of Biological Sciences, University of Bergen and Bjerknes Centre for Climate Research, Bergen, Norway
[f] Department of Ecology & Multidisciplinary Institute for Environmental Studies "Ramon Margalef", University of Alicante, Alicante, Spain
[g] Norwegian Institute for Nature Research, Thormøhlensgate 55, 5006 Bergen, Norway
[h] Dept of Mathematical Sciences and Centre for Biodiversity Dynamics, Norwegian Univ. of Science and Technology (NTNU), Trondheim, Norway
[i] Center for Global Health Science and Security, Georgetown University Medical Center, 6 Georgetown University, DC 20007, USA

ARTICLE INFO

Keywords
COVID-19
Pandemic
Data validation
Open science
Interdisciplinary research

ABSTRACT

In this letter we present comments on the article "A global-scale ecological niche model to predict SARS-CoV-2 coronavirus" by Coro published in 2020.

1. Introduction

The recent outbreak of SARS-CoV-2 infections, which causes COVID-19 in humans, has accelerated the development of global health policies to manage and mitigate the risks of infectious diseases (Wilder-Smith et al., 2020). Concurrently, the medical scientific community has been mobilized to provide clinical support during the pandemic (Kupferschmidt and Cohen, 2020). This pandemic has also attracted the interest of non-medical researchers hoping to understand potential environmental drivers of SARS-CoV-2 prevalence and to predict future outbreaks via correlative methods such as species distribution models (SDMs).

Coro (2020) presented an implementation of a SDM to predict global areas of high SARS-CoV-2 infection rate from geophysical and social-related spatial covariates that he found to be correlated with high infection rates. While contributing to a broader understanding of the potential geographic scope of the pandemic is a commendable effort, the application of a correlative model to estimate SARS-CoV-2's ecological niche given the epidemiology of transmission, as well as technical aspects of model implementation, highlights the drawbacks of using this approach to predict future outbreaks of SARS-CoV-2. We contend that such SDM implementations are prone to spurious results, meaning that they have a high potential for finding correlations between mechanistically unrelated variables, especially when model assumptions are violated and spatial scaling between covariates and mechanistic processes are mismatched. Therefore, extreme care should be taken when disseminating and overinterpreting correlative research without strong *a priori* mechanistic hypotheses, particularly in the midst of a pandemic, as incorrect conclusions about drivers of viral spread may have negative consequences if used as guidance by policy makers. Indeed, Coro's model did identify correlations between putative explanatory variables and high infection rates. However, below we argue that such correlations are unlikely to be biologically relevant both because the variables considered lack mechanistic plausibility in this setting as well as other flaws in model implementation.

Herein, we present our concerns with the foundational premise of Coro 2020 to model the ecological niche of SARS-CoV2 as well as the implementation and interpretation of modeling results. We suggest that

* Corresponding authors.
E-mail addresses: andrea.contina@ucdenver.edu (A. Contina); Colin.Carlson@georgetown.edu (C.J. Carlson)

https://doi.org/10.1016/j.ecolmodel.2020.109288
Received 18 July 2020; Received in revised form 11 September 2020; Accepted 12 September 2020
Available online xxx
0304-3800/

Coro's (2020) claim that "Generally, the model indicates a high infection rate in areas characterized by an annual moderate-high level of CO_2, moderate-low temperatures, and moderate precipitation" is problematic for two primary reasons: 1) SDMs do not accurately predict the dynamic geography of SARS-CoV-2 transmission because the underlying drivers of viral spread are dominated by human behavior (Carlson et al., 2020); and 2) the paper overlooks substantial predictive misclassifications, model assumptions and validations.

2. SDMs are not suitable for modeling emerging SARS-CoV-2 spread

SDMs typically use the locations where a species is recorded as being present and, if possible, where the species is recorded as being absent, to build a statistical model of the occurrence of the species with environmental covariates in order to predict its distribution. The correct application of an SDM requires that the species has a clear environmental niche, even if that niche is unknown to the researcher, and that its distributional data are reflective of that niche (Václavík and Meentemeyer, 2012; Gallien et al., 2012). The use of an SDM to predict viral spread does not meet these requirements because SARS-CoV-2 presumably originated from a single spillover event (Andersen et al., 2020) and further transmission has resulted from human-to-human contact and not through a vector with an identifiable ecological niche (Carlson et al., 2020). Moreover, the virus is continuing to spread across the globe into new environments, driven primarily by human social contact patterns (Liu et al., 2020), such that the distribution of occurrence at any one point in time likely does not accurately reflect any sort of theoretical ecological niche (Chipperfield et al. 2020). Although Coro's (2020) model includes human-related parameters of population density and CO_2 levels in addition to climatic variables, they inadequately describe other social-related factors that drive transmission beyond human abundance such as, but not limited to, social contact dynamics, behavior changes from perceived threat of infection, governmental policy and its timing of implementation, socioeconomic factors across different regions of a country and across countries, and the timing of infection peaks.

2.1. The pandemic is a dynamic process

Coro (2020) derived an "occurrence of high infection rate" response variable that is constructed by selecting the provinces in Italy that have a higher number of confirmed cases per capita than the national average (up until March 2020 representing the peak of the pandemic in Italy). However, this process does not improve the suitability of the application of an SDM. Firstly, the pandemic is a *dynamic* process; Markov processes operating within human contact networks alone are sufficient to expect that areas of Italy which don't currently have high infection rates, may later become the high infection areas as epidemics potentially take hold in those places whilst declining in the areas affected earlier. Given the north-south gradient of environmental conditions in Italy, then we might expect the entire correlation between occurrence and climate to change depending on when during the outbreak we perform the analysis. The proposed climate-occurrence signal may in fact be noise because the provinces identified as high-infection rate can and do change if case data from one day preceding or following the period analyzed by Coro are omitted or included. Therefore, the model is not robust to making accurate predictions outside of that time period. It is widely appreciated that SDMs and similar correlative spatial models are susceptible to spurious findings through random covariance of spatially structured variables alone with no causal link (Currie et al., 2019; Bahn and McGill, 2007; Fourcade et al., 2018). Therefore, *a priori* care in selecting covariates and accounting for spatial autocorrelation is paramount to avoiding spurious inferences in SDMs (Currie et al., 2019; Austin, 2007; Merow et al., 2013). In this case, transmission of SARS-CoV-2 is largely by droplets generated from human-to-human interaction, so there is unlikely to be a strong effect of climate, especially as most transmission occurs indoors (Allen and Marr, 2020).

3. Flaws in model implementation and validation

Coro (2020) dichotomized continuous values of infection rate using as threshold the average number of confirmed cases per capita at the peak of the pandemic in Italy to derive a response variable with the categories "occurrence" and "absence" of "high infection rate". The dichotomization of continuous data has well-known negative consequences, such as overestimation of effect sizes and loss of measurement reliability (MacCallum et al., 2002), overestimation of the differences between groups (Altman and Royston, 2006), and distortion of the relationship between predictive and response variables (Selving, 1987).

Furthermore, Howard et al. (2014) demonstrate that using a proxy of abundance (such as per-capita infection rate) as input for a SDM lead to distribution estimates significantly better than those derived from presence-absence data because the signal abundance-suitability is lost when all presences are treated as equals. For example, treating above-average per-capita infection rates as equals in the presence category may lead to a high bias if "very high" infection rates are scarce within the dataset and "just above average" infection rates are environmentally clustered by chance due to the North-to-South climatic and infection gradients in Italy. In such a case, high suitability values yielded by MaxEnt will be centered around the presences with lower infection rates within the "presence" category. Such a problem may only worsen if there is not much difference between infection rates at both sides of the threshold.

Using such knowledge, Coro (2020) could have made a sound choice by using per-capita infection rates as a continuous response variable in a regression model, rather than obscuring potential relationships between the response variable and the predictors through a convenient (MaxEnt can only use a dichotomous variable as response) albeit arbitrary dichotomization. As a result, biases in Coro's analysis cannot be assessed, and remain masked behind the apparent reliability of a colorful map.

Moreover, Coro's paper does not acknowledge or address violations of model assumptions beyond the existence of a viral niche that significantly affect results and their interpretation such as scaling mismatch between species' data (modeled here as locations with high viral incidence rate) and predictor variables, and equilibrium of the species with its environment (Guisan and Zimmerman, 2000). Additionally, the poor model validation described in the manuscript undermines the paper's primary conclusions. To wit, the conclusions 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" is unsupported based on the model diagnostics presented in the paper. We elaborate these details in the next three sections.

3.1. Scale mismatch and prediction outside model domain

Coro (2020) used a MaxEnt model which is known to be highly prone to bias resulting from non-random training samples (Elith et al., 2011; Merow et al., 2013; Gurutzeta et al., 2015). Coro's model was initially trained on data restricted to point locations of capital cities of Italian provinces with a high-incidence of infections which is a non-representative sample for predictive extrapolation to the whole planet (Jarnevich et al., 2015). To test whether the geographic scope of the training data influenced model performance, the author retrained the model first by adding areas that were initially well-predicted by the base model and re-assessing model fit. Unsurprisingly, the inclusion of these areas did not substantially affect model performance.

However, when the training area was expanded to include areas with poor predictive validation, model performance dropped. The author states that this result "indicates that the used input parameters are insufficient to understand the infection rate increase in these areas". Thus, the fitted model is not suitable for predicting areas of high infection rate on a global scale. We find the poor predictive performance of the model unsurprising because the restricted geographic extent of the training data limits the explanatory power of relatively coarse resolution (0.5˚) covariate data used by Coro (Mertes and Jetz, 2018), particularly for the CO_2 flux data that was reprojected to a 0.5˚ resolution from the original Copernicus CO_2 data product (Coro and Trumpy, 2020) which has an even coarser resolution of 3.75˚ in longitude by 1.87˚ in latitude (CAMS, 2019). Model results suggest CO_2 flux as the variable with the highest predictive power, however, even after resampling from the original coarser resolution, correlative associations between CO_2 and high incident rates of COVID-19 are drawn from only 26 unique values that are unrepresentative of variation of values globally. Thus, global extrapolation is outside of the original model domain and based on a limited amount of information. Furthermore, the 0.5˚ resolution of covariate data is mismatched in scale with high-infection rates because it is derived from case data summarized by province and geo-referenced by the location of the province capital city. This leads to substantial model bias because any grid-cells within the province that do not overlap the point location of the capital city are falsely ignored by the model as true presence locations leading to missed covariate associations, particularly for CO_2 flux and population density which may have substantially different values in grid-cells not overlapping with the capital city.

Given the bias introduced by an unrepresentative sample and scaling mismatch, rather than implying the modeled viral "niche" does validate across the "range" of SARS-CoV-2 due to an undiscovered separate set of environmental correlates for viral spread in these poorly-predicted areas, we suggest that incomplete predictive accuracy implies estimated correlations that might be spurious (i.e., not causally related).

3.2. SARS-CoV-2 is spreading through human social interactions; a system not in equilibrium

Another key assumption of SDMs, including MaxEnt models, is that the modeled species distribution is a stationary function of some environmental variable(s) (Austin, 2002; Guisan and Zimmerman, 2000; Elith et al., 2011). However, this assumption does not hold true for invasive species at the onset of immigration into a new area since the geography of occurrence is still shifting and likely expanding (Theoharides and Dukes, 2007; Phillips et al., 2008). Thus, the assumption of an equilibrium distribution is unmet by the emerging pandemic whose emerging spatial distribution is dynamic and primarily influenced by human social contact networks rather than abiotic environmental factors (Liu et al., 2020; see also Section 2.1).

3.3. Predictive misclassification

Coro's (2020) model misclassifies 22.75% of known high infection rate areas (where areas are defined by a mix of geopolitical units). Based on World Health Organization data (WHO, 2020), at the time the revised paper was submitted to the journal (June 11, 2020) the missed areas, from a visual inspection of Fig. 3 in Coro 2020, cover over a quarter of the cumulative global case load; today those areas cover nearly 30% of the cumulative global case load. We suggest that such substantial underprediction of extant outbreaks indicates poor model performance rendering inferences drawn from the model not informative.

4. Conclusion

We found little convincing evidence that this global-scale ecological niche model predicts SARS-CoV2-infection rate due to the drawbacks that we identified in the model implementation, validation, and interpretation as well as in the premise of an identifiable ecological niche for the spread of this particular virus. All modelling exercises require careful consideration of modeling assumptions in relation to study design and data collection for making interpretations that are meaningful and repeatable. Given the limitations of SDMs in particular, we recommend that future efforts to forecast SARS-CoV-2 outbreaks and/or to predict the spatial occurrence of other viruses with a similar ecology to SARS-CoV-2 be avoided (Carlson et al., 2020). Fundamentally, these outbreaks represent a non-equilibrium process with an environmental niche constrained almost entirely by that of its host organism, in this case humans, and with range dynamics that SDMs have a poor track record of predicting well. There may be cases for limited application of well-considered SDMs for directly transmitted diseases in appropriate scenarios such as the modelling of the occurrence of reservoir hosts of infectious disease transmissible from animals to humans (Zhu and Peterson, 2014; Carlson et al., 2016) and/or to make assessments of wildlife ecology and conservation strategies (Higgins et al., 2012). However, the direct application of an SDM for the occurrence of a directly-transmitted disease with a host species that has affected an estimated 95% of the earth's land surface is never likely to be informative.

Declaration of Competing Interest

The authors whose names are listed immediately below certify that they have NO conflict of interests in the subject matter or materials discussed in this manuscript.

Acknowledgements

AC is supported by NSF grant DBI-1565128.

References

Allen, J.G., Marr, L.C., 2020. Recognizing and controlling airborne transmission of SARS-CoV-2 in indoor environments. Indoor Air 30 (4), 557.
Altman, D.G., Royston, P., 2006. The cost of dichotomising continuous variables. BMJ 332 (7549), 1080.
Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C., Garry, R.F., 2020. The proximal origin of SARS-CoV-2. Nat. Med.. https://doi.org/10.1038/s41591-020-0820-9.
Austin, M.P., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecol. Model. 157, 101–118. doi:10.1016/S0304-3800(02)00205-3.
Austin, M., 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. Ecol. Model. 200, 1–19. doi:10.1016/j.ecolmodel.2006.07.005.
Bahn, V., McGill, B.J., 2007. Can niche-based distribution models outperform spatial interpolation? Glob. Ecol. Biogeogr. 16 (6), 733–742.
CAMS (Copernicus Atmosphere Monitoring Service)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.
Carlson, C.J., Dougherty, E.R., Getz, W., 2016. An ecological assessment of the pandemic threat of Zika virus. PLoS Negl. Trop. Dis. 10 (8), e0004968.
Carlson, C.J., Chipperfield, J.D., Benito, B.M., Telford, R.J., O'Hara, R.B., 2020. Species distribution models are inappropriate for COVID-19. Nat. Ecol. Evol. 4 (6), 770–771.
Chipperfield, J.D., Benito, B.M., O'Hara, R.B., Telford, R.J., Carlson, C.J., 2020. On the inadequacy of species distribution models for modelling the spread of SARS-CoV-2: response to Araújo and Naimi. In: EcoEvoRxiv. doi:10.32942/osf.io/mr6pn.
Coro, G., 2020. A global-scale ecological niche model to predict SARS-CoV-2 coronavirus infection rate. Ecol. Model. 109187.
Coro, G., Trumpy, E., 2020. Predicting geographical suitability of geothermal power plants. J. Clean. Prod. 121874. https://doi.org/10.1016/j.jclepro.2020.121874.
Currie, D.J., Pétrin, C., Boucher-Lalonde, V., 2019. How Perilous are Broad-Scale Correlations with Environmental Variables? Frontiers of Biogeography.
Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E., Yates, C.J., 2011. A statistical explanation of MaxEnt for ecologists. Divers. Distrib. 17, 43–57. doi:10.1111/j.1472-4642.2010.00725.x.

Fourcade, Y., Besnard, A.G., Secondi, J., 2018. Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. Glob. Ecol. Biogeogr. 27 (2), 245–256.

Gallien, L., Douzet, R., Pratte, S., Zimmermann, N.E., Thuiller, W., 2012. Invasive species distribution models–how violating the equilibrium assumption can create new insights. Glob. Ecol. Biogeogr. 21 (11), 1126–1136.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135, 147–186. doi:10.1016/S0304-3800(00)00354-9.

Gurutzeta, G., José, L., Elith, J., Gordon, A., Kujala, H., Lentini, P., Michael, M., Tingley, R., Wintle, B., 2015. Is my species distribution model fit for purpose? Matching data and models to applications. Glob. Ecol. Biogeogr. 24, 276–292. doi:10.1111/geb.12268.

Higgins, S.I., O'Hara, R.B., Römermann, C., 2012. A niche for biology in species distribution models. J. Biogeogr. 39 (12), 2091–2095.

Howard, C., Stephens, P.A., Pearce Higgins, J.W., Gregory, R.D., Willis, S.G., 2014. Improving species distribution models: the value of data on abundance. Methods Ecol. Evol. 5 (6), 506–513.

Liu, Y., Gu, Z., Xia, S., Shi, B., Zhou, X.N., Shi, Y., Liu, J., 2020. What are the underlying transmission patterns of covid-19 outbreak?–an age-specific social contact characterization. EClinicalMedicine 100354.

Jarnevich, C.S., Stohlgren, T.J., Kumar, S., Morisette, J.T., Holcombe, T.R., 2015. Caveats for correlative species distribution modeling. Ecol. Inform. 29, 6–15.

Kupferschmidt, K. and Cohen, J., 2020. Race to find COVID-19 treatments accelerates.

MacCallum, R.C., Zhang, S., Preacher, K.J., Rucker, D.D., 2002. On the practice of dichotomization of quantitative variables. Psychol. Methods 7 (1), 19.

Mertes, K., Jetz, W., 2018. Disentangling scale dependencies in species environmental niches and distributions. Ecography 41 (10), 1604–1615.

Merow, C., Smith, M.J., Silander, J.A., Jr., 2013. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. Ecography 36, 1058–1069. doi:10.1111/j.1600-0587.2013.07872.x.

Phillips, B.L., Chipperfield, J.D., Kearney, M.R., 2008. The toad ahead: challenges of modelling the range and spread of an invasive species. Wildl. Res. 35, 222–234. doi:10.1071/WR07101.

Selvin, S., 1987. Two issues concerning the analysis of grouped data. Eur. J. Epidemiol. 3 (3), 284–287.

Theoharides, K.A., Dukes, J.S., 2007. Plant invasion across space and time: factors affecting nonindigenous species success during four stages of invasion. New Phytol. 176, 256–273. doi:10.1111/j.1469-8137.2007.02207.x.

Václavík, T., Meentemeyer, R.K., 2012. Equilibrium or not? Modelling potential distribution of invasive species in different stages of invasion. Divers. Distrib. 18, 73–83.

World Health Organization (WHO), 2020. In: WHO Coronavirus Disease (COVID-19) Dashboard. Geneva, 2020. World Health Organization Available online. https://covid19.who.int accessed: 07/08/2020.

Wilder-Smith, A., Chiew, C.J., Lee, V.J., 2020. Can We Contain the COVID-19 Outbreak with the Same Measures as for SARS? The Lancet Infectious Diseases.

Zhu, G., Peterson, A.T., 2014. Potential geographic distribution of the novel avian-origin influenza A (H7N9) virus. PLoS ONE 9 (4), e93390.