Leveraging pathogen community distributions to understand outbreak and emergence potential

Tad A. Dallas^{a,†}, Colin J. Carlson^{b,c,†} and Timothée Poisot^d

^a*Centre for Ecological Change, University of Helsinki, 00840 Helsinki, Finland*

^b*National Socio-Environmental Synthesis Center, University of Maryland, Annapolis, Maryland 21401, USA.*

^c*Department of Biology, Georgetown University, Washington, D.C. 20057, USA.*

^d*Dépt de Sciences Biologiques, Univ. de Montréal, Montréal, Canada*

[†]*These authors contributed equally to this study.*

*Corresponding author: tad.a.dallas@gmail.com

1 **Running title:** Community ecology and pathogen emergence

2

3 **Author contributions:** TD, CJC, and TP conceived of the idea for the study.
4 TD and TP designed the model. All authors contributed to the writing of the
5 manuscript.

6

7 **Data accessibility:** R code is available on figshare at
8 <https://doi.org/10.6084/m9.figshare.6364955>.

9

10 **Ethics:** This study used existing data on pathogen outbreak and emergence events
11 in human populations.

12

13 **Funding:** This work was supported by the National Socio-Environmental Synthe-
14 sis Center (SESYNC) under funding received from the National Science Foundation
15 DBI-1639145.

16

17 **Competing interests:** The authors declare no competing interests.

18

19 **Keywords:** Emerging infectious disease, community ecology, community dissim-
20 ilarity, disease forecasting, pathogen biogeography

21

22 Abstract

23 Understanding pathogen outbreak and emergence events has important implica-
 24 tions to the management of infectious disease. Apart from preempting infectious
 25 disease events, there is considerable interest in determining why certain pathogens
 26 are consistently found in some regions, and why others spontaneously emerge or re-
 27 emerge over time. Here, we use a trait-free approach which leverages information
 28 on the global community of human infectious diseases to estimate the potential for
 29 pathogen outbreak, emergence, and re-emergence events over time. Our approach
 30 uses pairwise dissimilarities among pathogen distributions between countries and
 31 country-level pathogen composition to quantify pathogen outbreak, emergence,
 32 and re-emergence potential as a function of time (e.g., number of years between
 33 training and prediction), pathogen type (e.g., virus), and transmission mode (e.g.,
 34 vector-borne). We find that while outbreak and re-emergence potential are well
 35 captured by our simple model, prediction of emergence events remains elusive, and
 36 sudden global emergences like an influenza pandemic seem beyond the predictive
 37 capacity of the model. While our approach allows for dynamic predictability of
 38 outbreak and re-emergence events, data deficiencies and the stochastic nature of
 39 emergence events may preclude accurate prediction. Together, our results make a
 40 compelling case for incorporating a community ecological perspective into existing
 41 disease forecasting efforts.

Introduction

The emergence of infectious diseases in humans and wildlife is a continuous and natural process that is nevertheless rapidly intensifying with global change (1). Around the world, the diversity, and frequency, of infectious outbreaks is rising over time (2; 1), and the vast majority of pathogens with zoonotic potential still have yet to emerge in human populations, with an estimated 600,000 minimum viruses with zoonotic potential (3). Intensifying pathways of contact between wildlife reservoirs and humans, and rapid spread of new pathogens among human populations around the globe, are considered major drivers in this accelerating process (4; 5). Changes in climate and land-use, as well as food insecurity and geopolitical conflict, are expected to exacerbate feedbacks between socio-ecological change and emerging infectious diseases (EIDs). In the face of these threats, the anticipation of disease emergence events is a seminal but elusive challenge for public health research (6).

One forecasting approach recognizes that the drivers of emergence events are distributed non-randomly in space and time, and follow predictable regional patterns that inherently predispose some areas to a higher burden of EIDs (7). Different classes of emerging pathogens (e.g., new pathogens versus drug-resistant strains of familiar ones; vector-borne and/or zoonotically transmitted diseases) follow different spatial risk patterns at a global scale (1). In part, this can be explained by the non-random distribution of host groups that disproportionately contribute to zoonotic emergence events, like bats and rodents (8; 9), and are likely to continue to do so (10; 11; 12). However, additional factors are strongly associated with the distribution of emerging infection risk; notably human population density, land cover, and land use change (7). In addition to these factors, deterministic emergence of disease is influenced by social, cultural, and economic factors (13; 14; 15; 16; 17).

As a consequence of this heterogeneity in host distributions and other contributing factors, emerging pathogens may follow Tobler’s First Law (“near things are more related than distant things”; (18)), and fall into a handful of global biogeographic regions with similar pathogen communities (19). However, with increasing global connectivity, both pathogens and the free-living organisms that host them are spreading around the world at an accelerating rate, and consequently the spatial structure of pathogen diversity is becoming less pronounced. One study examining a global pathogen-country network showed that modularity is decreasing while connectance is increasing over time: pathogen ranges are on average expanding, and over time, geographically-separate regions are facing more threats (20; 21). This process of biotic homogenization has critical implications for public health, as known diseases can become unfamiliar problems in novel locations, or can re-emerge in landscapes from which they were previously eradicated.

Leveraging disease ecology in global health settings requires models that consider disease emergence as a long-term process over space and time, extending beyond initial spillover events. Work that models the impact of human mobility networks has arisen out of the pandemic influenza literature (22; 23; 24), and has recently been successful in developing a multi-scale approach to anticipating emergence risk for hemorrhagic viruses in Africa (25). However, conceptually-similar work capable of modeling numerous pathogen species at large spatial scales is presently undeveloped. It has been suggested that countries who share pathogens might be more likely targets during a given pathogen outbreak (19), but this approach does not leverage information on the identity of the shared pathogens. Given the inherent need in estimating outbreak potential, and the current availability of data on outbreak events, there is a current pressing need to leverage existing data on numerous pathogen species to allow for dynamic prediction of potential pathogen outbreak or emergence events.

Here, we examine the predictability of pathogen biogeography over time using a similarity-based approach that utilizes data on all pathogen outbreaks in all countries, but does not require information on pathogen traits or spatial structure. In the process of modeling outbreak predictability, we test a basic but important hypothesis: do recurring outbreaks have a more predictable signal than emergence events (and, implicitly, are emergence events predictable)? Within emergences, we further note the subtle difference between emergence and re-emergence, and hypothesize the factors driving these might be subtly different. While both may be driven by genetic shifts in pathogens or changing land use patterns enhancing transmission risk, re-emergence events are more likely to be related to weakened healthcare infrastructure, prematurely-terminated eradication campaigns (26; 27), or low detection long-term persistence of environmental pathogen reservoirs (e.g., anthrax spores in the soil; (28)).

Finally, we examine whether pathogens show any differences in predictability based on agent, class, or transmission mode. Diseases of zoonotic origin (i.e. with animal hosts) and with vector-borne transmission might be harder to predict due to hidden constraints on their distribution and more complicated outbreak dynamics than directly-transmitted pathogens have. On the other hand, commonalities between species that share vectors or reservoir hosts might lead to similarities in distributions (a common notion in pathogen biogeography, as in how dengue models were frequently used in the early days of the Zika pandemic, given the shared vector *Aedes aegypti*; (29; 30)). In this case, community-based prediction could be more powerful for zoonotic and vector-borne diseases. Differential frequency of zoonotic and vector-borne transmission might also make different pathogen classes (viruses, bacteria, fungi, and macroparasites) more or less predictable, as might different dispersal ability on a global scale, with respiratory viruses usually presumed to spread the fastest, and macroparasites generally treated as the most

123 dispersal-limited. Understanding how the role of community structure changes for
124 these different pathogens can help contextualize the method we use, and under-
125 stand how it might be built upon to account for these differences.

126 **Methods**

127 **Pathogen emergence data**

128 Data from the Global Infectious Diseases and Epidemiology Network (GIDEON)
129 contains pathogen outbreak information at the country level obtained from case
130 reports, governmental agencies, and published literature records (31; 32). Records
131 with multiple etiological agents (e.g., “*Aeromonas* and marine *Vibrio* infx.”) and
132 unresolved to agent level (e.g., “Respiratory viruses - miscellaneous”) were ex-
133 cluded from the model. In a handful of cases, we kept divisions between clinical
134 presentations from the same pathogens, like cutaneous versus visceral leishmania-
135 sis. The data obtained were yearly records between 1990 and 2016, and consisted
136 of pathogen outbreak and emergence events for 234 pathogens across 224 coun-
137 tries. While there are some data for pathogen events between 1980 and 1990, the
138 number of pathogen events reported was fewer than from 1990 onward, suggesting
139 some potential reporting or sampling bias in these earlier years. Therefore, we
140 restrict our analyses to pathogen occurrences after 1990. Based on supplemental
141 data from (20) and updated with recent literature given several misclassifications,
142 each was manually classified as a bacterial, viral, fungal, protozoan, or macropar-
143 asitic disease, and as vector-borne and/or zoonotic or neither. In some rare cases,
144 these were left as unknown; for example, Oropouche virus is vector-borne but
145 its sylvatic cycle remains uncertain, while the environmental origin of Bas-Congo
146 virus is altogether unknown.

147 While much can be gained by leveraging data on multiple pathogens to predict

148 outbreak or emergence potential, there are some drawbacks. The most pronounced
149 is that pandemic events may strongly influence model predictions, such that a pan-
150 demic of one pathogen will decrease model performance when attempting to predict
151 outbreak or emergence potential of other pathogens. We explore this further in the
152 supplement, where we see the inclusion of influenza and the corresponding 2009 flu
153 pandemic noticeably affects our model performance. As such, we remove influenza
154 from the main text analyses, and place analyses containing flu in the supplement
155 for comparison.

156 We distinguish between three different types of pathogen events; outbreak, re-
157 emergence, and emergence. Outbreaks are pathogen events are recurrent pathogen
158 events, quantified as having occurred in a given country within three years of a
159 given year. Re-emergence events are those that did not occur within three years,
160 but have occurred at some time in a given country in the past (a cutoff we chose
161 inspired by World Health Organization guidelines for certifying regional eradica-
162 tion of poliovirus or dracunculiasis). Lastly, emergence events were considered as
163 the first record of a pathogen within a country.

164 **Model structure**

165 We developed a dissimilarity-based approach to forecast pathogen outbreak and
166 emergence events that does not require country-level or pathogen traits data. Ap-
167 plying tools from community ecology, we calculated mean pairwise dissimilarity
168 (Bray-Curtis index, $\overline{BC}$) values for countries (how dissimilar are the pathogen
169 communities between countries) and pathogens (how dissimilar are the geographic
170 distributions of pathogens). For a given pair of countries a, b with P_a and P_b
171 pathogens each, and S shared pathogens among those, the Bray-Curtis index is
172 given as:

$$BC_{a,b} = 1 - \frac{2S}{P_a P_b} \quad (1)$$

173 This can be treated as a measure of dissimilarity between different countries’
 174 pathogen communities. We then considered the potential for a pathogen to be
 175 found in a country proportional to the product of these dissimilarity values. We
 176 also included year as a covariate, resulting in a set of four variables for model
 177 training.

178 Using these data, we applied a statistical approach previously used for species
 179 distribution modeling (33) and link prediction in ecological networks (34) called
 180 **plug-and-play** (PNP). This approach utilizes information on pathogen occur-
 181 rence events, and also on background interactions — country-pathogen pairs which
 182 did not have a recorded outbreak — to estimate the suitability of a country for
 183 pathogen emergence from a particular pathogen (Figure 1). These suitability val-
 184 ues can then be used to quantify model performance on data not used to train the
 185 model.

186 If pathogen outbreak events occur in the same countries probabilistically based
 187 on some propensity for the pathogen to occur at that location, we might expect that
 188 using past data on pathogen outbreaks could be used to forecast pathogen events.
 189 If a pathogen were to occur in a given country in one year out of four, a naive
 190 assumption would be that it has a 25% chance of occurrence in the subsequent
 191 sampling event. We examined how this null expectation compares against our
 192 approach, which uses information on country and pathogen similarity values (light
 193 grey lines in figures).

194 **Assessing model performance**

195 We used the PNP modeling approach to address the possibility of predicting
 196 pathogen outbreak and emergence events compared to a null model. Model per-

197 formance was quantified using Area Under the Curve (AUC), which captures the
198 ability of the classifier to rank positive instances higher than negative instances.
199 To assess model performance, we examined three different potential scenarios.

200 First, we examined how the inclusion of pathogen events from previous years
201 influenced model accuracy. That is, we predicted pathogen events of 2016 using
202 data starting at 2015 and then including additional years until 1995. This was
203 performed to determine the amount of data necessary to make accurate forecasts.
204 Second, we examined how predictive accuracy was maintained as we attempted to
205 predict both past (hindcast) and future (forecast) pathogen events. To do this, we
206 trained models on a ten year period (either 2005-2015 for hindcasting, or 1990-2000
207 for forecasting), and used these models to predict pathogen events between 1990
208 and 2004 for hindcasting, and between 2001 and 2015 for forecasting. Lastly, we
209 examined how the accuracy of predictions might have changed over time. Given
210 increased surveillance in more recent years, predictive accuracy might be dependent
211 on the time period at which models are trained and predictions made. To test this,
212 we trained models along a rolling window of 4 years from 1990-2015, using these
213 models to predict pathogen events in the year following the final year of model
214 training (e.g., a model trained on 1990-1994 would be used to predict pathogen
215 events in 1995).

216 Results

217 We find that our dissimilarity-based model can predict outbreak events accurately,
218 re-emergence events slightly less accurately, and emergence events only slightly
219 better than random (i.e., $AUC = 0.5$). This makes intuitive sense, as outbreak
220 events occur repeatedly, providing not only ample data for model training, but
221 also a clear tendency of a pathogen to occur in a country. That is, if the model

is allowed to see 5 years of data, and the country has an outbreak of a particular pathogen in 4 of the 5 years, a naive model would predict that an outbreak will likely occur with an 80% probability. This situation corresponds to the null model (Figure 2), which performed poorly until enough temporal data was available, at which time the naive null model still underperformed our approach. Meanwhile, emergence events are determined by many unique drivers (7), which may not be consistent across any two given emergence events, and which we evidently lack sufficient data to predict using our method. While our model allows for dynamic predictability of outbreak and re-emergence events, data deficiencies and the stochastic nature of emergence events may thus preclude accurate prediction. However, our approach outperforms the naive null model in all modeled scenarios, especially when temporal data were limited (Figure 2 - 4).

Our predictive model was sensitive to the number of training years (Figure 2), with accuracy plateauing around 5-10 years of training data; however, models also just trained on a single year (the temporally closest community matrix) seemed to perform disproportionately well, which would make sense if the community changes in a Markov-like process. We further examined the limits of predictability in terms of both hindcasting and forecasting pathogen outbreak and emergence events by training the model on a known period of 10 years, and then either forecasting or hindcasting t years into the past or future (Figure 3). Interestingly, our accuracy – measured as area under the receiver operating characteristic – did not decline at the same rate when hindcasting and forecasting. That is, model accuracy was higher when hindcasting relative to the accuracy of forecasts of the same duration of time away from the training data (Figure 3). This perhaps indicates that as the country-pathogen network becomes asymptotically more connected and stable (21), the network accumulates information content, reducing the time sensitivity of hindcasting performance.

Examining a rolling window of t years ($t = 4$ years) over the last two decades, we failed to detect evidence that the enhanced reporting and surveillance in more recent years influenced our model’s predictive ability (Figure 4). This also suggests that even though there were annual variations in the sample size of both pathogens and countries, there was still consistency in the structure of the country-pathogen interaction matrix over time. We explore the sensitivity of this finding to the size of the rolling window in the Supplemental Materials.

Differences in PNP model accuracy among pathogen types existed when examining the effect of the amount of data used for model training (Figure 2), with viruses having lower accuracy relative to bacteria, fungi, or other parasites. The simplest explanation for this is that accuracy is sensitive to the number of events. However, the average number of viral occurrences over time ($\bar{x} = 179$) was only slightly less than the average number of bacterial ($\bar{x} = 185$) occurrences, and far greater than the average number of fungal ($\bar{x} = 10$) or macroparasite ($\bar{x} = 17.7$) or protozoan parasite ($\bar{x} = 22.5$) occurrence events. The average number of pathogen occurrences over time is qualitatively proportional to the number of unique viruses ($n = 83$), bacteria ($n = 81$), fungi ($n = 14$), macroparasites ($n = 38$), and protozoans ($n = 15$) we examined. Interestingly, differences among pathogen types were not found when examining the ability of the modeling approach to hind-cast/forecast (Figure 3) or when examining predictive accuracy along a rolling window (Figure 4).

For our 2016 explanatory PNP model, differentiating pathogens based on zoonotic and vector-borne transmission modes suggested that both classes of pathogens were more difficult to forecast (Figure 2). While it is possible that class imbalances between groups might drive this pattern (i.e., more event occurrences may increase model predictive accuracy), this seems unlikely: the majority of pathogens (144 of 228) were zoonotic, and many (59 of 233) were vector-borne. A more compelling

276 explanation is that this year was an anomalous result; transmission mode did not
 277 influence accuracy when hindcasting/forecasting (Figure S5) or when models were
 278 trained along a rolling window (Figure 4), though there was notable year-to-year
 279 variation in the latter.

280 Discussion

281 Community ecology and biogeography have a history as deeply linked fields, and
 282 both play an increasingly significant role in emerging infectious disease research.
 283 (35; 19; 36) However, research connecting the two for global pathogen diversity
 284 is fairly limited so far. Our goal was to examine whether the intrinsic structure
 285 of pathogen biogeography, approached as a bipartite network, was predictable
 286 enough to enable forecasting of different outbreak types—even in the absence of
 287 any other mechanistic predictors, like transmission mode, phylogenetic data, or
 288 environmental covariates.

289 Despite obvious stochasticity and data limitations, the modeling approach per-
 290 formed well with as little as 7 to 10 years of training data, and when predicting
 291 country-pathogen network structure across large time windows. The model was
 292 able to capture pathogen outbreak and re-emergence potential well, suggesting
 293 that, at least at administrative levels, pathogen outbreak and re-emergence events
 294 are both recurrent and predictable (and that community assembly patterns are
 295 structured and predictive of outbreak potential). However, our model generally
 296 failed to forecast pathogen emergence events. This is maybe unsurprising, as pre-
 297 dicting when and where the next major public health threat will emerge is an
 298 incredibly difficult task which remains unsolved despite having received decades
 299 of attention (7; 1; 6). However, the failure of community information to help an-
 300 ticipate local emergences is still disappointing, especially given the proposal that

301 biogeographic “co-zones” could be useful strategic tools for pandemic forecasting.
302 (16)

303 We found some indications of differences in the predictability of pathogen
304 events as a function of pathogen type and transmission modes. In the 2016
305 model breakdown, bacteria were the most predictable while viruses were dispro-
306 portionately unpredictable, as were zoonotic and vector-borne pathogens. Given
307 how clearly unpredictable emergence events were, this might make intuitive sense:
308 zoonotic pathogens make up the majority of emerging diseases (1), and single-
309 stranded RNA viruses (many vector-borne) have been responsible for many of the
310 biggest recent emergence events (8). However, this pattern did not appear to hold
311 up across all or even most years, and the factors that reduce model performance
312 on a year-by-year basis are mostly unclear at the community level.

313 One contributor to interannual variation is large-scale events such as pan-
314 demics, which appeared to strongly influence prediction of the entire country-
315 pathogen network. While pandemic spread may be predictable using detailed
316 information on climate, human movement, and local environmental suitability
317 (6; 37; 38), our approach lacks these mechanistic predictors and is sensitive to
318 these black swan events. This can be seen in reduced model performance during
319 the 2009 flu pandemic, including for pathogens with no relationship to flu, although
320 viruses and vector-borne pathogens are more severely affected (see Supplemental
321 Materials). So while the model benefits from pathogen community data, rare and
322 widespread events can strongly reduce model accuracy. Future work to differen-
323 tially weight these stochastic events would probably improve model performance.

324 While this approach enhances estimation of outbreak and emergence potential
325 for rare pathogens or poorly sampled countries, it is also worth noting that our
326 approach is *not* a valid standalone forecasting tool. This is in large part due to
327 how time is used in the model: though year is a covariate, the model itself is not

temporally explicit, meaning that the model can predict a certain link following on previous years, but it would be erroneous to interpret that as a forecast for a given point in time. However, the tool can be used to investigate pathogen outbreak and emergence potential under different pathogen range expansion scenarios. That is, researchers could construct artificial data which differs from empirical data slightly, and quantify the ability of the model to predict those novel events. Since the method is based on dissimilarity of countries and pathogen distributions at its core, it is possible to examine the expected outcome as pathogen distributions become more (or less) homogeneous, or countries become more (or less) dissimilar in their pathogen communities.

Within infectious disease ecology, a disproportionate focus has emerged on the drivers and predictability of emergence events. (7) Recent work offers a compelling case that community ecology might bring predictive tools to bear on that problem (35), and modeling work suggests that community assembly data can be leveraged to better predict how pathogens spread (19), the host range of emerging diseases (39; 8), and the dynamics of diseases within an ecosystem (40; 41). Our results show how a simple model considering the entire pathogen community captures important global geographic variation in outbreak potential, but as a standalone tool, still struggles to predict when a pathogen will first arrive in a new region. Though this casts doubt on biogeographic tools like “co-zones” as standalone tools for surveillance or outbreak response, our study is a compelling indicator that community data could be very easily leveraged alongside other socioecological predictors to forecast disease emergence as an ecosystem process rather than a single-species one. With a Nipah virus outbreak in India and an Ebola virus outbreak in the Democratic Republic of the Congo alone both concurrent to the completion of this manuscript, the priority of prediction in emerging disease research only continues to grow.

355 **Acknowledgements**

356 We thank GIDEON (<https://www.gideononline.com/>) for their collection and
357 curation of the data.

References

- [1] Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L. & Daszak, P., 2008 Global trends in emerging infectious diseases. *Nature* **451**, 990–993.
- [2] Smith, K. F., Goldberg, M., Rosenthal, S., Carlson, L., Chen, J., Chen, C. & Ramachandran, S., 2014 Global rise in human infectious disease outbreaks. *Journal of the Royal Society Interface* **11**, 20140950.
- [3] Carroll, D., Daszak, P., Wolfe, N. D., Gao, G. F., Morel, C. M., Morzaria, S., Pablos-Méndez, A., Tomori, O. & Mazet, J. A., 2018 The global virome project. *Science* **359**, 872–874.
- [4] Cleaveland, S., Haydon, D. & Taylor, L., 2007 Overviews of pathogen emergence: which pathogens emerge, when and why? *Current Topics in Microbiology and Immunology* pp. 85–111.
- [5] Tatem, A. J., Rogers, D. J. & Hay, S., 2006 Global transport networks and infectious disease spread. *Advances in Parasitology* **62**, 293–343.
- [6] Morse, S. S., Mazet, J. A., Woolhouse, M., Parrish, C. R., Carroll, D., Karesh, W. B., Zambrana-Torrel, C., Lipkin, W. I. & Daszak, P., 2012 Prediction and prevention of the next pandemic zoonosis. *The Lancet* **380**, 1956–1965.
- [7] Allen, T., Murray, K. A., Zambrana-Torrel, C., Morse, S. S., Rondinini, C., Di Marco, M., Breit, N., Olival, K. J. & Daszak, P., 2017 Global hotspots and correlates of emerging zoonotic diseases. *Nature Communications* **8**, 1124.
- [8] Johnson, C. K., Hitchens, P. L., Evans, T. S., Goldstein, T., Thomas, K., Clements, A., Joly, D. O., Wolfe, N. D., Daszak, P., Karesh, W. B. *et al.*,

- 381 2015 Spillover and pandemic properties of zoonotic viruses with high host
382 plasticity. *Scientific Reports* **5**.
- 383 [9] Olival, K. J., Hosseini, P. R., Zambrana-Torrel, C., Ross, N., Bogich, T. L.
384 & Daszak, P., 2017 Host and viral traits predict zoonotic spillover from mam-
385 mals. *Nature* **546**, 646–+.
- 386 [10] Han, B. A., Kramer, A. M. & Drake, J. M., 2016 Global patterns of zoonotic
387 disease in mammals. *Trends in Parasitology* **32**, 565–577.
- 388 [11] Han, B. A., Schmidt, J. P., Alexander, L. W., Bowden, S. E., Hayman, D. T.
389 & Drake, J. M., 2016 Undiscovered bat hosts of filoviruses. *PLoS Neglected*
390 *Tropical Diseases* **10**, e0004815.
- 391 [12] Han, B. A., Schmidt, J. P., Bowden, S. E. & Drake, J. M., 2015 Rodent
392 reservoirs of future zoonotic diseases. *Proceedings of the National Academy of*
393 *Sciences* **112**, 7039–7044.
- 394 [13] Bonds, M. H., Keenan, D. C., Rohani, P. & Sachs, J. D., 2010 Poverty trap
395 formed by the ecology of infectious diseases. *Proceedings of the Royal Society*
396 *of London B: Biological Sciences* **277**, 1185–1192.
- 397 [14] Farmer, P., 1996 Social inequalities and emerging infectious diseases. *Emerg-*
398 *ing Infectious Diseases* **2**, 259.
- 399 [15] McMichael, A. J., 2004 Environmental and social influences on emerging in-
400 fectious diseases: past, present and future. *Philosophical Transactions of the*
401 *Royal Society of London B: Biological Sciences* **359**, 1049–1058.
- 402 [16] Murray, D. R. & Schaller, M., 2010 Historical prevalence of infectious diseases
403 within 230 geopolitical regions: A tool for investigating origins of culture.
404 *Journal of Cross-Cultural Psychology* **41**, 99–108.

- 405 [17] Parkes, M. W., Bienen, L., Breilh, J., Hsu, L.-N., McDonald, M., Patz, J. A.,
406 Rosenthal, J. P., Sahani, M., Sleight, A., Waltner-Toews, D. *et al.*, 2005 All
407 hands on deck: transdisciplinary approaches to emerging infectious disease.
408 *EcoHealth* **2**, 258–272.
- 409 [18] Tobler, W. R., 1970 A computer movie simulating urban growth in the detroit
410 region. *Economic Geography* **46**, 234–240.
- 411 [19] Murray, K. A., Preston, N., Allen, T., Zambrana-Torrel, C., Hosseini, P. R.
412 & Daszak, P., 2015 Global biogeography of human infectious diseases. *Pro-*
413 *ceedings of the National Academy of Sciences* **112**, 12746–12751.
- 414 [20] Smith, K. F., Sax, D. F., Gaines, S. D., Guernier, V. & Guégan, J.-F., 2007
415 Globalization of human infectious disease. *Ecology* **88**, 1903–1910.
- 416 [21] Poisot, T., Nunn, C. & Morand, S., 2014 Ongoing worldwide homogenization
417 of human pathogens. *bioRxiv* p. 009977.
- 418 [22] Balcan, D., Hu, H., Goncalves, B., Bajardi, P., Poletto, C., Ramasco, J. J.,
419 Paolotti, D., Perra, N., Tizzoni, M., Van den Broeck, W. *et al.*, 2009 Seasonal
420 transmission potential and activity peaks of the new influenza a (h1n1): a
421 monte carlo likelihood analysis based on human mobility. *BMC Medicine* **7**,
422 45.
- 423 [23] Russell, C. A., Jones, T. C., Barr, I. G., Cox, N. J., Garten, R. J., Gregory, V.,
424 Gust, I. D., Hampson, A. W., Hay, A. J., Hurt, A. C. *et al.*, 2008 The global
425 circulation of seasonal influenza a (h3n2) viruses. *Science* **320**, 340–346.
- 426 [24] Khan, K., Arino, J., Hu, W., Raposo, P., Sears, J., Calderon, F., Heidebrecht,
427 C., Macdonald, M., Liauw, J., Chan, A. *et al.*, 2009 Spread of a novel influenza
428 a (h1n1) virus via global airline transportation. *New England Journal of*
429 *Medicine* **2009**, 212–214.

- 430 [25] Pigott, D. M., Deshpande, A., Letourneau, I., Morozoff, C., Reiner, R. C.,
431 Kraemer, M. U., Brent, S. E., Bogoch, I. I., Khan, K., Biehl, M. H. *et al.*, 2017
432 Local, national, and regional viral haemorrhagic fever pandemic potential in
433 africa: a multistage analysis. *The Lancet* .
- 434 [26] Chiappini, E., Stival, A., Galli, L. & De Martino, M., 2013 Pertussis re-
435 emergence in the post-vaccination era. *BMC Infectious Diseases* **13**, 151.
- 436 [27] Minor, P. D., 2004 Polio eradication, cessation of vaccination and re-
437 emergence of disease. *Nature Reviews Microbiology* **2**, 473.
- 438 [28] Carlson, C. J., Getz, W. M., Kausrud, K. L., Cizauskas, C. A., Blackburn,
439 J. K., Bustos Carrillo, F. A., Colwell, R., Easterday, W. R., Ganz, H. H., Ka-
440 math, P. L. *et al.*, 2018 Spores and soil from six sides: interdisciplinarity and
441 the environmental biology of anthrax (*Bacillus anthracis*). *Biological Reviews*
442 **0**, in press.
- 443 [29] Bogoch, I. I., Brady, O. J., Kraemer, M. U., German, M., Creatore, M. I.,
444 Kulkarni, M. A., Brownstein, J. S., Mekaru, S. R., Hay, S. I., Groot, E. *et al.*,
445 2016 Anticipating the international spread of zika virus from brazil. *The*
446 *Lancet* **387**, 335–336.
- 447 [30] Carlson, C. J., Dougherty, E. R. & Getz, W., 2016 An ecological assessment
448 of the pandemic threat of zika virus. *PLoS Neglected Tropical Diseases* **10**,
449 e0004968.
- 450 [31] Berger, S. A., 2005 GIDEON: a comprehensive web-based resource for geo-
451 graphic medicine. *International journal of health geographics* **4**, 10.
- 452 [32] Yu, V. L. & Edberg, S. C., 2005 Global infectious diseases and epidemiol-
453 ogy network (gideon): a world wide web-based program for diagnosis and
454 informatics in infectious diseases. *Clinical Infectious Diseases* **40**, 123–126.

- 455 [33] Drake, J. & Richards, R., 2017 Estimating environmental suitability. *bioRxiv*
456 p. 109041.
- 457 [34] Dallas, T., Huang, S., Nunn, C., Park, A. W. & Drake, J. M., 2017 Estimating
458 parasite host range. *Proceedings of the Royal Society B: Biological Sciences*
459 **284**, 20171250.
- 460 [35] Johnson, P. T., De Roode, J. C. & Fenton, A., 2015 Why infectious disease
461 research needs community ecology. *Science* **349**, 1259504.
- 462 [36] Stephens, P. R., Altizer, S., Smith, K. F., Alonso Aguirre, A., Brown, J. H.,
463 Budischak, S. A., Byers, J. E., Dallas, T. A., Jonathan Davies, T., Drake,
464 J. M. *et al.*, 2016 The macroecology of infectious diseases: a new perspective
465 on global-scale drivers of pathogen distributions and impacts. *Ecology Letters*
466 **19**, 1159–1171.
- 467 [37] Tizzoni, M., Bajardi, P., Poletto, C., Ramasco, J. J., Balcan, D., Gonçalves,
468 B., Perra, N., Colizza, V. & Vespignani, A., 2012 Real-time numerical forecast
469 of global epidemic spreading: case study of 2009 a/h1n1pdm. *BMC Medicine*
470 **10**, 165.
- 471 [38] Zhang, Q., Sun, K., Chinazzi, M., y Piontti, A. P., Dean, N. E., Rojas, D. P.,
472 Merler, S., Mistry, D., Poletti, P., Rossi, L. *et al.*, 2017 Spread of zika virus in
473 the americas. *Proceedings of the National Academy of Sciences* **114**, E4334–
474 E4343.
- 475 [39] Dallas, T., Park, A. W. & Drake, J. M., 2017 Predictability of helminth
476 parasite host range using information on geography, host traits and parasite
477 community structure. *Parasitology* **144**, 200–205.
- 478 [40] Parker, I. M., Saunders, M., Bontrager, M., Weitz, A. P., Hendricks, R.,

- 479 Magarey, R., Suiter, K. & Gilbert, G. S., 2015 Phylogenetic structure and
480 host abundance drive disease pressure in communities. *Nature* **520**, 542.
- 481 [41] Johnson, P. T., Preston, D. L., Hoverman, J. T. & LaFonte, B. E., 2013 Host
482 and parasite diversity jointly control disease risk in complex communities.
483 *Proceedings of the National Academy of Sciences* **110**, 16916–16921.

484 **Figure captions**

Figure 1: The dissimilarity-based model used takes mean dissimilarity values of pathogen distributions and between countries in a given year, and uses this information in addition to the product of these two values to train the PNP model. Pathogen occurrences among countries are present or absent (black dots in panel a indicate pathogen occurrences), and the density of dissimilarities where the pathogen occurred relative to the overall density of dissimilarities provides information on the suitability of pathogen occurrence in a given country (b), and forms the basis of the PNP model approach.

Figure 2: Pathogen events from previous years increased model predictive accuracy after an initial small decrease, suggesting that five years or more of data improves predictions, but accuracy could actually decrease in some data sparse situations where only two or three years of data were available. Performance of the null expectation (grey line) was less than our approach, except when the null was given more than 15 years of previous data.

Figure 3: Predictive accuracy decreased when attempting to forecast far into the past or future. Models were trained on either the period between 2005-2015 (for prediction into the past) or 1990-2000 (for prediction into the future). The null expectation (grey line) performed consistently worse than our approach.

Figure 4: Using a rolling window ($t = 4$ years), we found that predictive accuracy did not increase as a result of enhanced surveillance and data collection of more recent years. The null expectation (grey line) performed consistently worse than our approach.

485 **Figures**

486 **Figure 1**

487

Figure 2

Figure 3

Figure 4