Establishing a Global Wastewater Surveillance Network at Airports for Early Detection of Emerging Pathogens: a Modeling Study

Guillaume St-Onge[1], Jessica T. Davis[1], Laurent Hébert-Dufresne[2,3], Antoine Allard[2,3], Alessandra Urbinati[1], Samuel V. Scarpino[4,5,6], Matteo Chinazzi[1,7], and Alessandro Vespignani[1]

[1] *Laboratory for the Modeling of Biological and Socio-technical Systems, Northeastern University, Boston, MA 02115, USA*
[2] *Vermont Complex Systems Center, University of Vermont, Burlington, VT 05401, USA*
[3] *Département de physique, de génie physique et d'optique, Université Laval, Québec City, QC G1V 0A6, Canada*
[4] *Institute for Experiential AI, Northeastern University, Boston, MA 02115, USA*
[5] *Network Science Institute, Northeastern University, Boston, MA 02115, USA*
[6] *Santa Fe Institute, Santa Fe, NM 87501, USA*
[7] *The Roux Institute, Northeastern University, Portland, ME 04101, USA*

June 5, 2024

Abstract

Aircraft wastewater surveillance has been proposed as a novel approach to monitor the global spread of pathogens. Here we use an advanced epidemic and mobility model covering the entire air-travel network, combined with probability generating function analytics, to estimate the performance and optimization of global wastewater surveillance networks (WWSNs). We study detection times for respiratory diseases of varying transmission potentials and characterize how the network performance is affected by the location and number of monitoring sites. We show that 10 to 20 strategically placed wastewater sentinel sites offer timely situational awareness and function as an effective early warning system, although geographic blind spots exist. We propose optimization strategies to improve WWSN effectiveness and minimize resource use. The analysis of the SARS-CoV-2 dissemination scenarios illustrates the WWSN's ability to quickly assess pandemic potential. The presented approach offers a realistic analytic framework for the analysis WWSN at entry points.

Recent health crises have highlighted the dual role of airports both in spreading infectious diseases globally and simultaneously acting as convenient frontlines for detecting and monitoring emerging health threats [1–5]. Traditionally, wastewater surveillance has been used to monitor community prevalence of pathogens such as SARS-CoV-2 variants [6, 7], polio [8, 9], and influenza [10, 11]. Expanding wastewater surveillance to airports to create a global wastewater surveillance network (WWSN)has been recently proposed as a novel, early warning system against emerging pathogens [12–17].Although direct sampling from an aircraft provides a non-invasive method to quickly detect and monitor the spread of pathogens and eliminates the need for traveler participation, the creation of a global WWSN poses strategic and

operational challenges. These challenges include efficient sample collection, genomic surveillance set up at airports, and making strategic decisions such as selecting optimal airports for surveillance, scaling the network, and addressing surveillance blind spots to balance effectiveness and cost [18]. For this reason, while there have been studies on the feasibility of aircraft wastewater surveillance at several major airports [19–22], fully understanding the performance of a WWSN—in terms of its size, location, and operations—poses significant challenges.

Here, we use the Global Epidemic and Mobility Model (GLEAM) to effectively model the performance of a global WWSN with synthetic simulations, providing valuable insights into how pathogens spread and are detected within these networks. GLEAM is a stochastic, spatial, age-structured metapopulation model. It divides the global population into over 3,200 subpopulations across approximately 200 countries and territories, all interconnected by a network of air travel and commuting. The airline network component of the model includes data on flight segments and origin-destination information for more than 4,600 airports, obtained from the Official Aviation Guide (OAG) database (see Methods and Sec. 1 in the Supplementary Information, SI). The model incorporates country-specific age-stratified contact patterns and local mobility data [23]. The mobility framework is coupled with an epidemic compartmental modeling scheme that describes the disease statuses of individuals (ex. Susceptible, Latent, Infectious, etc.) and tracks how they transition through these stages. This allows us to the track the movement of individuals within specific compartments globally. The GLEAM model has been successfully applied to global modeling health threats including pandemic influenza, Ebola and SARS-CoV-2 [24–27]. To simulate a surveillance system within GLEAM, we create a global WWSN that consists of N surveillance sites—called *sentinels*. We assume each sentinel airport will test the wastewater from a certain fraction of arriving international flights per day.

Given any initial conditions for an outbreak, the model generates stochastic realizations of the global epidemic spread. Simulated data includes international and domestic infection importations, incidence of infections, and individual level detection at sentinel sites with a daily resolution. The early growth phase can be mapped onto a multitype branching process allowing for the computationally efficient derivation in terms of probability generating functions (PGFs) of several key analytics that measure the efficiency of the WWSN. These include the time to first detection of an emerging pathogen, the uncertainty in source location and estimates of the initial size of the outbreak; and other quantities such as the reproduction number and timing of the outbreak's onset. These metrics provide a general framework for assessing the WWSN's effectiveness in real-time surveillance and public health response.

Results

We established a baseline Wastewater Surveillance Network (WWSN) with 20 sentinel sites. To achieve comprehensive regional coverage, we selected the three busiest international airports from each of the six World Health Organization regions and added two additional sites in South America and Oceania. While not rigorously optimized, these sites address gaps in regional monitoring. The locations are shown by airport markers in Fig. 1 and reported in Table S4 in the SI.

The efficiency of the WWSN is related to the intrinsic characteristics of a pathogen as well as its detectability. To test the effectiveness of the WWSN we assume a SARS-CoV-2-like respiratory infection is spreading with a detectability period in wastewater similar to what is reported in Refs. [6, 28, 29]. We map an individual's disease history to a Susceptible–Latent–Detectable–Recovered (SLDR) compartmental structure, as shown Fig. 1. Susceptible (S) individuals can get infected through exposure to Infectious individuals. Latent (L) individuals are infected with the pathogen, but are not actively shedding infectious material which makes them unable to infect others and undetectable. Detectable (D) individuals are both infectious (I) individuals who can transmit the pathogen and post-infectious (P) individuals who no longer infect others but are still detectable through wastewater. Finally, removed (R) individuals are no longer detectable and we assume they cannot be reinfected (See Methods for details and key time-to-event intervals).

We apply a detection probability p_{det} that combines the fraction of sampled aircrafts, the probability an individual uses the lavatory during a flight, and the probability a detectable individual is shedding enough virus to lead to a detection to each traveling and detectible individual arriving at a sentinel. The current detectability of SARS-CoV-2 in aircraft wastewater varies [14, 16, 21], therefore, our analysis varies the detection probabilities, p_{det}, from 4 to 32% (see Methods). While sampling individual aircrafts independently is optimal for detection accuracy, it may be more cost-effective to test combined wastewater from multiple aircrafts at a consolidation point such as the airport triturator. Consequently, we assume that through pooled sampling, multiple detectable individuals traveling to the same sentinel on the same day cannot lead to more than *one detection*. It is worth remarking that most of these assumption can be relaxed in order to use different detection schemes, cadence and sentinel sites location.

Assessment of the WWSN efficiency.

A key metric for evaluating the effectiveness of a WWSN is the *time to first detection* t_{fd} of an emerging pathogen. This metric is defined as the number of days from the onset of an outbreak until the first detection at any sentinel. In our analysis, we seed an epidemic in a single basin with a small, initial cluster of 10 latent and 10 infectious individuals. The time to first detection is inherently dependent on

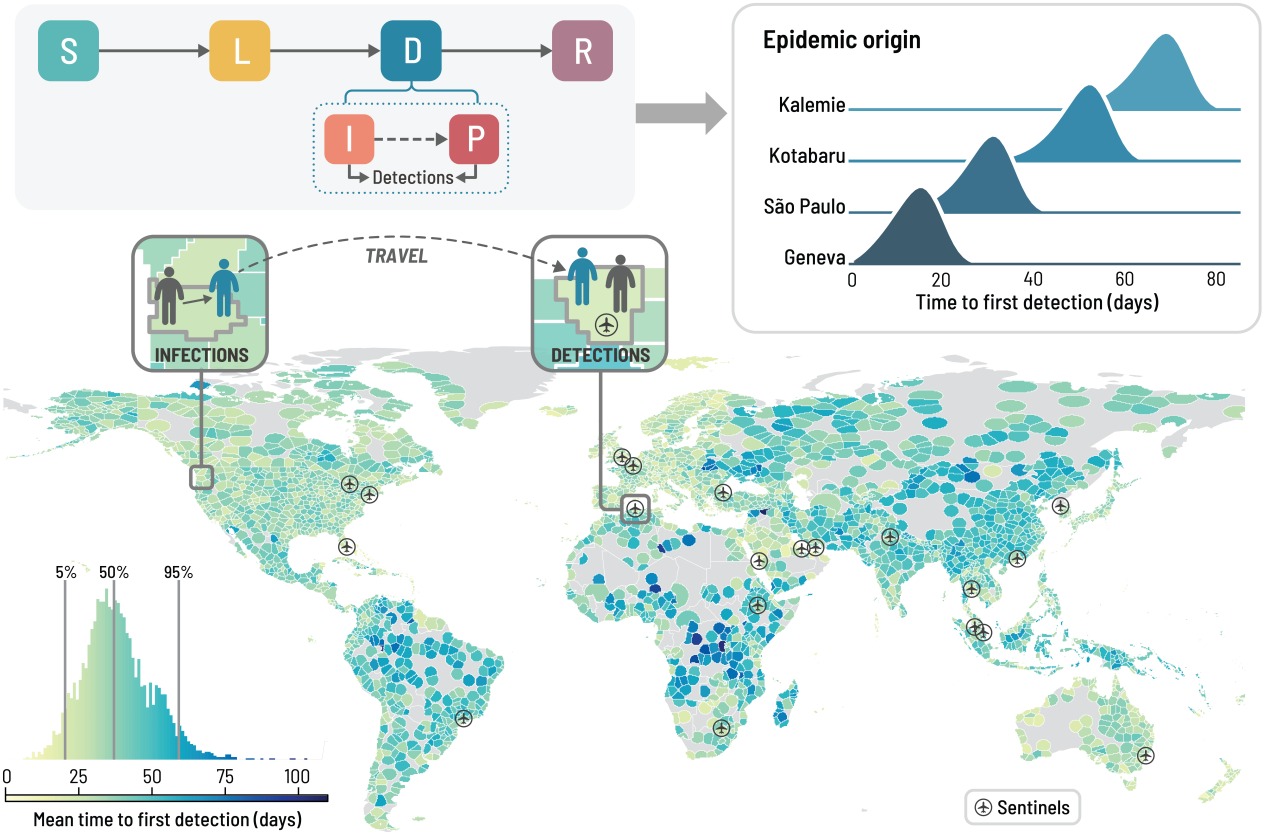

Figure 1: **Time to detect a novel pathogen with a global surveillance network at airports.** The surveillance system corresponds to a baseline network of 20 sentinel airports, chosen based on their high volume of international passengers and favoring geographical diversity. We use an average basic reproduction number $\langle \mathcal{R}_0 \rangle = 2$ at the source, a generation time of 4 days, and a post-infectious period of 10 days, resulting in a detectable period of ∼12.7 days. For detectable individuals, the probability of detection on an international flight incoming to a sentinel is 16%. We consider each of subpopulations as the potential origin for an epidemic and evaluate the mean time to obtain a single detection within the surveillance system. The histogram in the lower left corner compiles the mean time to first detection from 3244 subpopulations.

the WWSN configuration, the origin of the outbreak, the pathogen characteristics, and the operational detection rate. At the same time, there are also fluctuations stemming from the stochastic nature of each simulated outbreak and the individuals' traveling and detection events. More rigorously, we characterize this time to first detection using the probability distribution $P(t_{\rm fd} = t) = P(d_{t-1} = 0, d_t \geq 1)$, where d_t represents the total number of detections t days after the outbreak starts. In Fig. 1, we show the full probability distribution $P(t_{\rm fd})$ for the time to first detection for four different origins: Geneva (Switzerland, CH), São Paulo (Brazil, BR), Kotabaru (Indonesia, IN), and Kalemie (Democratic Republic of the Congo, DR).

We see that the time to first detection can vary significantly, with a mean of 14.2 (90% PI, 4–22) days for

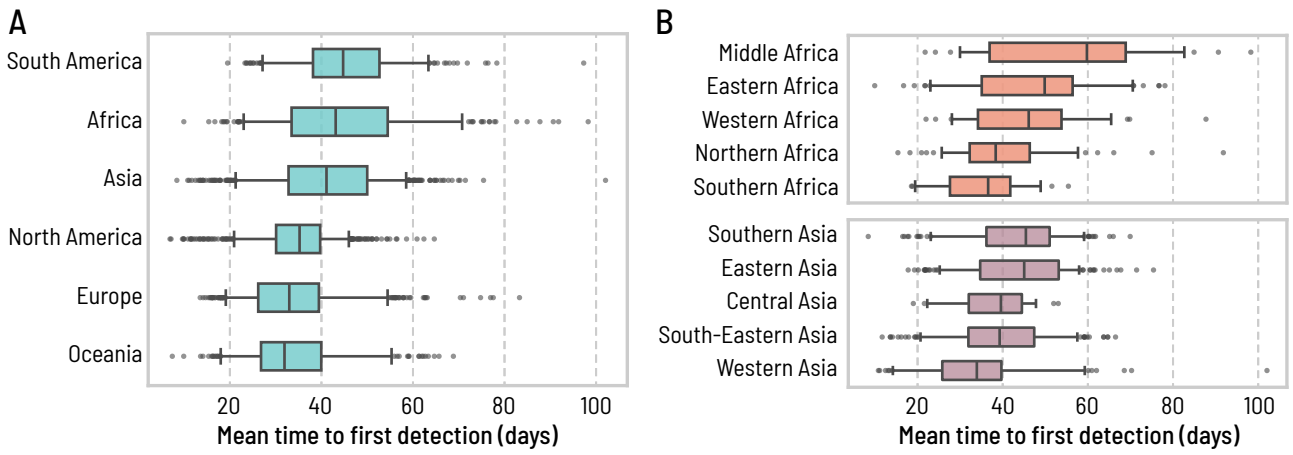

Figure 2: **Heterogeneity of the time to first detection at different geographical scales.** We aggregate the mean time to first detection T_{fd} obtained in Fig. 1 over (A) continents (S. America, $n = 297$; Africa $n = 338$; Asia $n = 867$; N. America $n = 854$; Europe $n = 596$; Oceania $n = 282$) and (B) statistical subregions as defined by the United Nations geoscheme in Africa (Middle, $n = 45$; Eastern $n = 121$; Western $n = 52$; Northern $n = 89$; Southern $n = 31$) and Asia (Southern $n = 205$; Eastern $n = 269$; S.-Eastern $n = 236$; Central $n = 37$; Western $n = 120$). The whiskers of the boxplot cover the 90% central prediction interval and black dots correspond to outliers outside this interval.

Geneva (CH), to 66.5 (90% PI, 53–76) days for Kalemie (CD), where PI refers to the central prediction interval. In this figure, we assume that the detection rate in the WWSN is $p_{\mathrm{det}} = 16\%$ and is uniform across all 20 sentinels. To obtain a global picture of the WWSN performance given different epidemic origin locations, we calculate the average time to first detection, T_{fd}, for a given source across all of the 3,200+ subpopulations in our model (see Fig. 1). There is notable spatial heterogeneity of T_{fd} that depends on the origin of the epidemic. For certain locations in Central Africa, T_{fd} is of the order of 100 days, while for many places in Europe, 15-25 days is more typical.

This heterogeneity is further highlighted in Fig. 2 at different geographical scales. While Fig. 2A shows that epidemics emerging from some continents take more time than others to be detected by a global WWSN, we note an important heterogeneity within continents as well. For instance, in Africa, the 90% PI of the T_{fd} ranges from 23 to 71 days. Even looking down at the level of statistical subregions as defined by the United Nations geoscheme in Fig. 2B, we still find very broad distributions of T_{fd} for all subregions. Middle Africa for instance is very dispersed, with a 90% PI ranging from 28.2 to 84.5 days. This means that within all regions, and at various scales, there are potential *blind spots* for a global WWSN: locations that if they are the source of an epidemic, will take a very long time to lead to a detection.

These blind spots within the WWSN can be attributed to the per capita volume of travel in each location. In the SI we show a strong indirect correlation between the per capita volume of travel and the mean time to first detection (Fig. S6). Often, detections at sentinel sites are not solely due to direct importations from

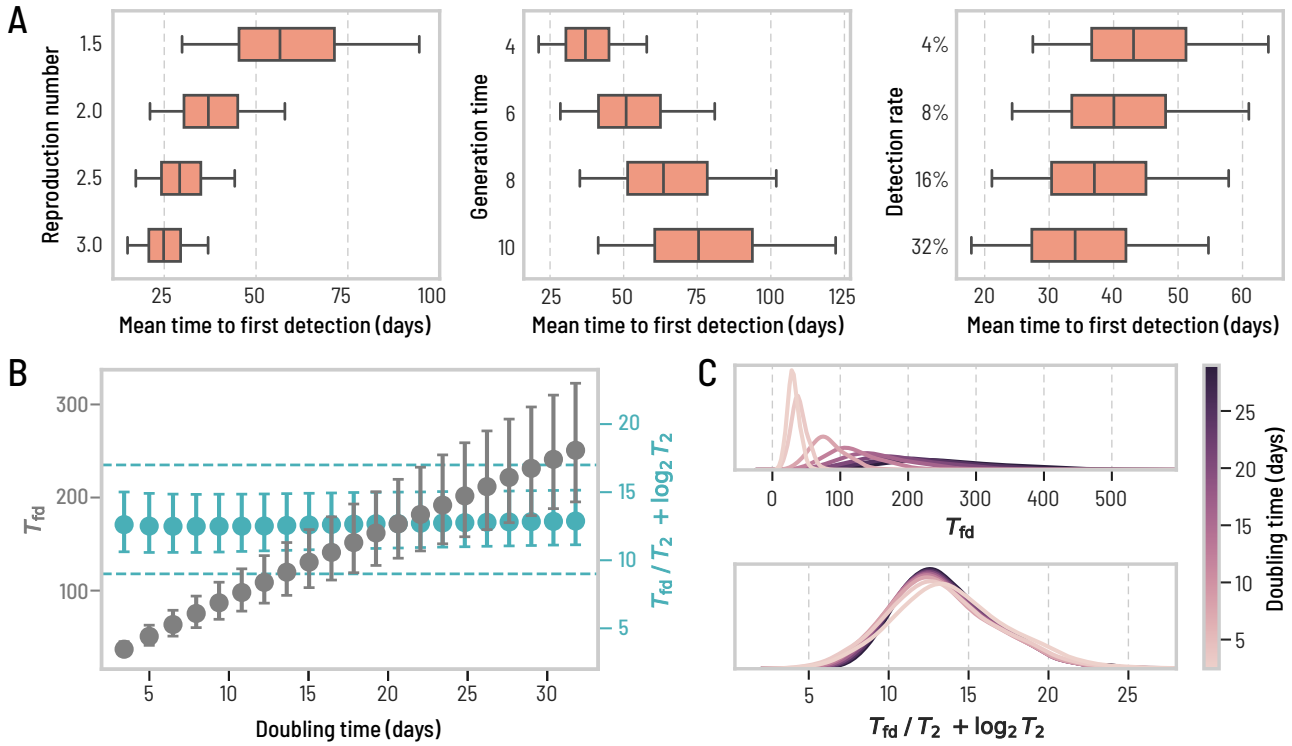

Figure 3: **Changing the transmission dynamics predictably affects the time to first detection**. We use the same baseline WWSN as in Fig. 1 and the same detectable period. Unless specified, we keep an average reproduction number of 2, a generation time of 4 days, and a 16% detection rate at sentinels. All prediction intervals are obtained with $n = 3244$ subpopulations. (A) T_{fd} from all origins, with varying reproduction number, generation time, and detection rate. The whiskers of the boxplot cover the 90% central prediction interval, the outliers outside the interval are not shown. (B-C) We vary the generation time between 2 and 42 days, resulting in doubling times between 3.4 and 31.8 days. (B) T_{fd} and $T_{\mathrm{fd}}/T_2 + \log_2 T_2$ as a function of the doubling time. Circles indicate the median and the whiskers the interquartile range. The dashed lines are there to guide the eyes. (C) Distributions of T_{fd} and $T_{\mathrm{fd}}/T_2 + \log_2 T_2$ over all origins for different doubling times. We use kernel density estimates (KDE) for the distributions to improve the visualization.

the outbreak's origin, but rather due to importations from other secondary outbreak locations experiencing community transmission. This complexity highlights the need for a detailed modeling framework and accurate flight data to evaluate the effectiveness of a global WWSN.

Additionally, the natural history of a disease affects the T_{fd}. In Fig. 3A, we show how the global distribution of T_{fd}, aggregated over all locations, changes as the reproduction number $\mathcal{R}_0$, the generation time T_{gen}, and the surveillance detection rates p_{det} change. A larger reproduction number and a smaller generation time lead to faster T_{fd}, and vice-versa. However, the smaller the probability of detection the larger the T_{fd}, although with limited impact. This can be explained by the exponential growth of epidemics in the early stages. The WWSN will typically start detecting cases when there is a sufficient number of detectable individuals, D, which is approximately $D \propto 2^{T_{\mathrm{fd}}/T_2}$, where T_2 is the *doubling time* of the epidemic

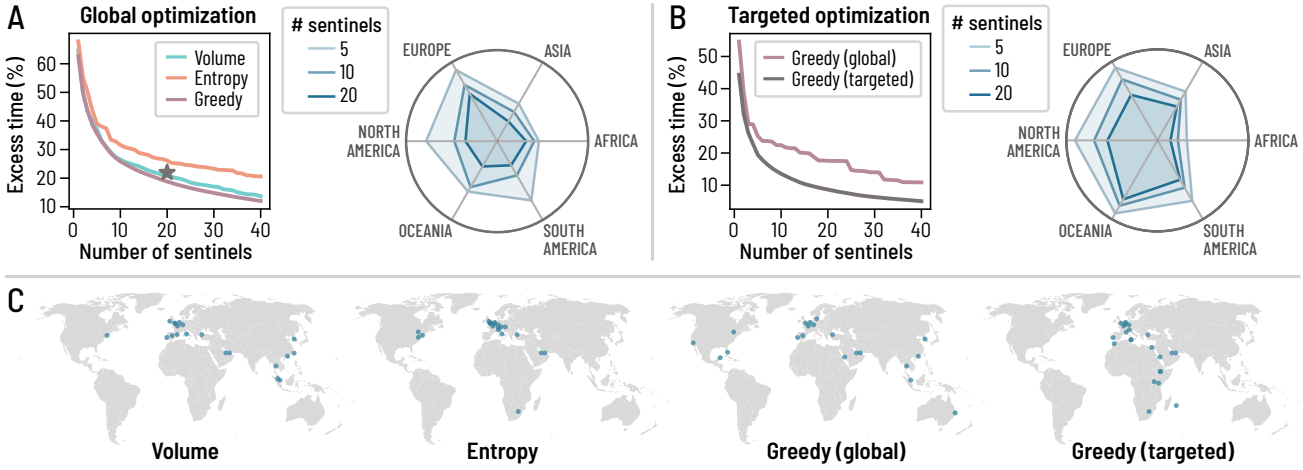

Figure 4: **Scaling and optimization of a global surveillance network at airports**. We use the same disease parametrization as in Fig. 1, but we use different strategies to choose the sentinel airports and we vary their number. We evaluate T_{fd} and compute the excess time to first detection compared to the idealized WWSN. (A) Global optimization: we assume all subpopulations are equiprobable for the origin of an epidemic. The star corresponds to the excess time for the baseline network. (B) Targeted optimization: the greedy strategy is tuned to minimize the excess time for epidemics originating from Africa. (A-B) The radar charts illustrate the relative excess time in each continent for the greedy strategy. (C) Spatial distribution of the 20 first sentinels for each strategy.

(days). Altering the basic reproduction number $\mathcal{R}_0$ or the generation time T_{g} significantly impacts the time to first detection because of the change of T_2. However, this is not the case for the detection probability p_{det}. Notably, a two-fold decrease in detection probability results in a corresponding two-fold increase in D, the number of individuals infected before detections begin. But it's important to recognize that this increase in D is rapidly offset by the exponential growth nature of the infectious disease in a single doubling time T_2.

The exponential growth also suggests the ratio T_{fd}/T_2 should be approximately constant as we vary the doubling time of the epidemic. However, as shown in Fig. 3B, a more appropriate invariant quantity is

$$T_{\mathrm{fd}}/T_2 + \log_2 T_2 = \text{const.} \tag{1}$$

The correction term $\log_2 T_2$ is necessary to account for the stochastic nature of the detection process [30] (see SI, Sec. 2). In Fig. 3C, we also show how the distributions of mean time to first detection collapse onto one another when instead considering the invariant quantity in Eq. (1). Altogether, this shows changing the disease characteristics amounts to a simple linear transformation of T_{fd} for all locations (see also Fig. S8 in the SI), and consequently, we can focus on a specific parametrization without loss of generality. Other aspects of disease transmission affecting T_{fd}—overdispersion of the secondary-infection distribution, length of the detectable period, and seasonal change in the air-travel network—have a very limited impact (see Table S3 in the SI).

Scaling and optimization of WWSNs. To determine the most effective configurations for a global WWSN, two critical decisions must be made: the optimal number of sentinel airports and their geographical distribution. This scenario presents a classic resource-constrained optimization problem. Using the mean time to first detection as our efficiency measure, we more precisely define $T_{\mathrm{fd}}(\mathcal{S}, l)$ as the mean time to first detection for a WWSN configuration consisting of the set of sentinels $\mathcal{S}$ and an epidemic origin in subpopulation l. We average this measure over multiple origins l by weighing each location according to a prior distribution $P(l)$ for the occurrence of an outbreak, resulting in the (average) mean time to first detection

$$T_{\mathrm{fd}}(\mathcal{S}) = \sum_{l} P(l)\, T_{\mathrm{fd}}(\mathcal{S}, l) \ . \tag{2}$$

While $T_{\mathrm{fd}}(\mathcal{S})$ is a well-defined indicator of performance, its value is sensitive to variation of the disease transmission characteristics, as highlighted by Eq. (1). To provide a more informative measure, we compare $T_{\mathrm{fd}}(\mathcal{S})$ with $T_{\mathrm{fd}}(\mathcal{I})$, the average mean time to first detection we would obtain with the *idealized* WWSN, whose set of sentinels $\mathcal{I}$ contains every airport in the world. This allows us to define the *excess time* to first detection for a particular sentinel system $\mathcal{S}$ as

$$E(\mathcal{S}) = 100 \times \frac{T_{\mathrm{fd}}(\mathcal{S}) - T_{\mathrm{fd}}(\mathcal{I})}{T_{\mathrm{fd}}(\mathcal{I})} \ , \tag{3}$$

which is interpreted as the additional time it takes for a sentinel system $\mathcal{S}$ to get a first detection, in percentage of $T_{\mathrm{fd}}(\mathcal{I})$.

We use three different strategies to define the geographic distribution of the sentinel network: (1) ranks airports based on their international inbound passenger *volume* [15], (2) ranks airports by their *entropy*—a measure of diversity that favors airports with a broad and homogeneous geographical coverage, (3) uses a *greedy* optimization strategy that minimizes the mean time to first detection (see Methods) In Fig. 4A, we show the excess time to first detection for the three different strategies considered, assuming an equiprobable source of an epidemic, irrespective of the area or population size ($P(l) = \text{const. fro all } l$). While the greedy approach systematically provides the smallest excess time, all three strategies have a similar performance, despite different network configurations (see Fig. 4C). The radar charts suggest that the greedy strategy provides a relatively balanced geographical surveillance when compared to the idealized WWSN. This means different WWSNs can provide a comparable global coverage on average, as evaluated by the excess time to first detection, and close to being optimal for a given number of sentinels.

There is the clear diminishing return on the reduction of the excess time with the addition of sentinels. This indicates there is an optimal balance between the desired efficiency of the WWSN and the resources available for such a system. While the deployment and maintenance of a global WWSN would necessitate

rigorous cost evaluation, our analysis suggests that a WWSN with 20 sentinels would only take 20% longer to provide a first detection than a surveillance system composed of thousands of airports. Doubling the number of sentinels would introduce marginal gains of less than 10%.

While minimizing the global T_{fd} is desirable, we might want to take a more spatially targeted approach. Some diseases are only endemic in certain parts of the world or have clear seasonal patterns. Similarly, from Fig. 2, the T_{fd} is higher in some regions than others. Therefore, it might be preferable to bias the optimization procedure to improve the detection capabilities for regions with high T_{fd}s. The greedy optimization approach can be adapted to do so by tuning the prior function $P(l)$. If we aim to minimize the excess time to first detection for epidemics originating in Africa, we can set $P(l) = \text{const.}$ if l is in Africa, otherwise $P(l) = 0$. In Fig. 4B, we show the excess time to first detection and compare the previous (global) greedy method with this *targeted* greedy optimization strategy, which displays a much better performance for the task. The radar chart in Fig. 4B illustrates the bias introduced by the targeted optimization procedure, providing a biased coverage of the African continent at the expense of performance elsewhere.

An aspect of the greedy optimization approach is its potential to optimize the surveillance system dynamically, or in an *adaptive* manner, based on real-time data. For instance, if initial detections—whether from the WWSN or local case reports—indicate certain regions as the likely sources of an epidemic, we can apply this method adaptively. This involves recommending additional sentinel sites and identifying specific routes for targeted screening to maximize detections. Such an adaptive strategy enhances situational awareness by responding promptly and effectively to emerging epidemic trends.

Situational awareness capabilities of WWSNs. An efficiently functioning WWSN is crucial for the early detection of pathogens, providing timely situational awareness insights. To demonstrate the potential impact of such a system, we analyze a hypothetical scenario that assesses the utility of a WWSN prior to the onset of the SARS-CoV-2 Alpha variant (B.1.1.7) during Fall 2020 [31–33]. This counterfactual study uses air-travel data from September to November 2020 and the baseline WWSN illustrated in Fig. 1. We considered an effective reproduction number of $\mathcal{R}_{\mathrm{eff}}^{\mathrm{alpha}} = 1.7$ for the alpha variant, corresponding to an increased transmissibility of 55% compared to the SARS-CoV-2 wild strain (see SI Sec. 3).

In Fig. 5A, we show the plausible distributions for the time to first detection. We find that even with a low detection rate (4%), we would likely have started to detect the variant in November, with a median time to first detection on November 13 (90% PI, October 15–December 1). With a 16% detection rate—more in line with estimates for a triturator sampling scheme capturing all international inbound flights—the first detections would have likely occurred at the end of October (median value, October 29, 2020, (90% PI, October 2–November 16). Considering the Alpha variant was first reported by the United Kingdom government on December 14, 2020, our retrospective analyses shows the potential of a global WWSN as an early warning system.

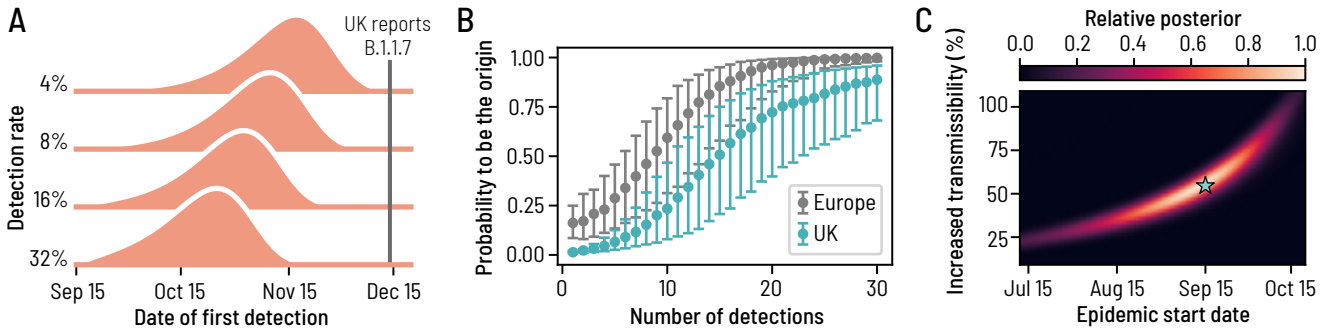

Figure 5: A global WWSN would provide an early warning system for international spreading and timely inferential capabilities. We consider a counterfactual scenario of the emergence of the SARS-CoV-2 Alpha variant where a global WWSN would have been available. We use the baseline surveillance system illustrated in Fig. 1. We considered an effective reproduction number of $\mathcal{R}_{\text{eff}}^{\text{ws}} = 1.1$ for the wild strain SARS-CoV-2 virus, and an increased transmissibility of 55% for the Alpha variant, resulting in an effective reproduction number of $\mathcal{R}_{\text{eff}}^{\text{alpha}} = 1.7$. We also assumed a generation time of 6.5 days, and a post-infectious period of 10 days. We assume an initial cluster of 20 infectious and 20 latent individual in the London and South East England region on September 15, 2020. (A) Distributions for the time to first detection with varying detection rates. (B)-(C) Inference experiment using data generated by the mechanistic GLEAM model, assuming a 16% detection rate. (B) Geolocalisation of the source as more detections cumulate. We compute the posterior distribution for the origin of the epidemic based on the detection counts at each sentinel. The markers indicate the median posterior value and the whiskers the interquartile range obtained from 1250 detection time series. (C) Joint posterior distribution on the increased transmissibility of the Alpha variant and the epidemic start date, average over 125 detection time series. Gouraud interpolation is used to improve the visualization. The blue star indicates the ground truth values for the simulation experiment, corresponding to September 15 and an increased transmissibility of 55%.

Beyond detecting the initial international spread, the WWSN can provide a critical and timely information that can be instrumental in pinpointing the origin of an outbreak and understanding its growth dynamics. In Fig. 5(B), we show reliable estimations of the continent and country of origin as detections accumulate, and this without even matching the detection with a specific inbound flight. This is achieved by calculating the posterior distribution $P(l|\boldsymbol{d})$ for each subpopulation l to be the origin of an epidemic based on the the cumulative number of detections at each sentinel $\boldsymbol{d} = (d_\nu)_{\nu\in\mathcal{S}}$ (see SI Sec. 3). We see that the source country would have been correctly geolocalized after around 20 detections, which would have been observed by December 5th in more than 50% of realizations when considering a detection rate of 16%. In a practical scenario, an even more efficient adaptive localization strategy could be implemented, involving targeted sampling from individual aircraft in regions suspected to be the outbreak's origin, thereby facilitating more precise identification of the source.

Data stream of cumulative detections can also be utilized to estimate key epidemic parameters, such as the growth rate, the onset time, and, given some knowledge of the generation time, the reproduction number. Here, we more precisely consider the inference of the epidemic start date and the increased transmissibility of the Alpha variant compared to SARS-CoV-2 wild strain. As of December 14, 2020, we

would have had access to reliable intelligence on the growth dynamics, as illustrated in Fig. 5C by the joint posterior distribution averaged over multiple detection time series. This average posterior distribution summarizes the expected confidence we would have about the inferred epidemic start date (90% CI, July 29–Oct. 12) and increased transmissibility (90% CI, 25–91%), solely based on aircraft wastewater detections that would have been available at the time the United Kingdom first reported the Alpha variant. The high posterior density region also matches closely the value of the simulation experiment, i.e., a start date on September 15 and a 55% increased transmissibility. The detailed inference procedure and individual posterior distributions for different time series are reported in Sec. 3 of the Supplementary Information.

For additional evidence of the timely situational awareness capacities provided by a global WWSN, we present in the Supplementary Information Sec. 3 another counterfactual scenario where a global WWSN would have been operational at the time of emergence of SARS-CoV-2 in Wuhan.

Discussion

The systematic detection and reporting of international cases during an emerging epidemic are pivotal for assessing pandemic risk, calibrating models, and providing global situational awareness, ultimately enhancing pandemic preparedness. However, gathering and processing such data in the early outbreak phase is challenging due to varied reporting and surveillance capabilities across regions [34].

The modeling study and analytical results presented here show clearly how a global WWSN at airports can systematically characterize international spread, offering timely insights into epidemic outbreaks. To leverage the WWSN as an early warning system, our focus has been on minimizing the time to first detection. Our findings indicate an inherent heterogeneity in T_{fd} depending on the pathogen's origin, persisting even in an ideal WWSN with all airports as sentinels. This suggests the implementation of WWSN strategies where sentinel activation is dynamically based on the geolocalization of ongoing community transmission. Such strategies would not only aid in screening diverse pathogens but also facilitate a resource-efficient adaptive WWSN. The presence of blind spots also underscores the need for complementary surveillance methods. For instance, environmental wastewater monitoring in specific areas of interest could be leveraged to achieve a more performing global surveillance system [35]. While our study focuses on airport wastewater surveillance, our models can be adapted to environmental monitoring and other travel-based surveillance methods, such as nasal-swab testing, thereby providing a robust and comprehensive computational platform for disease surveillance.

The optimization experiments and the strategies were mainly designed to evaluate and showcase the performance of a global WWSN, and not tailored to any specific disease and its history. In future studies, these should be compared with other surveillance strategies that have been developed in the context of

arbovirus and influenza surveillance for instance [36–38]. Also, moving forward with the detection of novel pathogens, we should start to incorporate knowledge about the propensity of zoonotic spillovers based on socioeconomic, environmental, and ecological factors [39–43].

As with every modeling study, our analysis contains assumptions and limitations that must be clearly identified. Our approach considers air travel as an independent process for each individual and thus neglects clusters and household travel. While our study presents a comprehensive framework for a global Wastewater Surveillance Network (WWSN), it does not delve into the logistics of wastewater sampling at airports. For example, the feasibility of daily testing may not always be practical. Future studies should explore alternative strategies, such as rotating testing schedules across different sentinel sites, to optimize network efficiency while navigating logistical constraints. A full-scale implementation of the model should be developed in conjunction with the actual deployment of the surveillance system, considering the practicalities and challenges of on-site operations.

Another important aspect to investigate in future modeling studies is the eventuality of false positive test results and the occurrence of positive tests caused by wastewater tanks not being cleaned in between flights [21]. While this should not impact substantially our analysis of the time to first detection, it could affect the situational awareness capabilities of the WWSN. Therefore, future analysis should incorporate the prior probability of a pathogen circulating in the statistical model, along with the test specificity and the possibility of wastewater tank contamination between flights. This will be important for decision-making, especially if a rare but very high-consequence pathogen is detected.

On a more technical side, the analytics developed in this study rely on a multitype branching process, which neglects saturation effects from finite population sizes. Although these effects are minor and do not alter our conclusions at the early stage of an outbreak, they should be carefully considered when analyzing the performance of the WWSN for the inference of incidence/prevalence of large epidemics or endemic situations.

Despite the assumptions and limitations inherent to any modeling study, the computational platform we have developed provides essential analytics for optimizing the functionality of wastewater surveillance at airports, and a more informed and strategic deployment of WWSNs. The quantitative insight provided by our approach holds significant implications for a range of stakeholders in public health, policy, and global health security. By demonstrating the effectiveness of a Wastewater Surveillance Network in early pathogen detection and spread assessment, our computational platform offers a powerful tool for informed decision-making.

Methods

The Global Epidemic and Mobility model. The Global Epidemic and Mobility model (GLEAM) is a computational platform used for modeling epidemic spread, combining stochastic elements and spatial data in an age-structured, metapopulation framework [26, 27]. GLEAM divides the world into distinct geographic subpopulations using a Voronoi tessellation of the Earth's surface, with each subpopulation centered around major transportation hubs such as airports. These subpopulations are detailed with high-resolution data about population demographics, age-specific contact patterns, health infrastructure, and other relevant attributes based on available data.

GLEAM incorporates a human mobility layer into its modeling, using data from various sources, including the Official Aviation Guide (OAG) and IATA databases. This layer includes both short-range (e.g., commuting) and long-range (e.g., flights) mobility data, and creates a network of daily passenger flows between airports worldwide. The model uses a worldwide homogeneous standard for commuting data and compensates for missing information with synthetic data based on the "gravity law" calibrated with real data [44–46].

The model sets initial conditions by specifying the number and location of individuals capable of transmitting the infection. GLEAM can track the number of individuals in each disease state for all subpopulations over time. It simulates travelers' movements through the flight network, with air travel probabilities varying by age group. Each day in the model is simulated in 12 distinct time steps, and this process repeats for every simulated day, with individuals and their travel patterns monitored.

Finally, the disease dynamics and the detection process at airports within GLEAM are simulated at the individual level using stochastic binomial chain processes. These processes rely on parameter values sourced from existing literature, defining the natural history of the infection being modeled. This approach allows GLEAM to provide detailed and realistic simulations of how infectious diseases might spread across different populations and regions, taking into account various factors like human mobility and local health infrastructures. We refer to this implementation as the *mechanistic* GLEAM approach. See Sec. 1 of the Supplementary Information for a more technical description of the model. All our analyses make use of a global air-travel network capturing the period of September 2022 to August 2023, except the case studies on the emergence of COVID-19 and the SARS-CoV-2 Alpha variant where we use data from December 2018 to February 2019[1] and September to November 2020 respectively.

Disease progression and transmission dynamics. To model the disease transmission within the subpopulations and the detections at airports following air travel, we make use of a standard compartmental-

[1]These were the available air-travel networks at the beginning of the COVID-19 pandemic, and the one that was used to generate stochastic simulations from GLEAM at the time.

Table 1: Range of disease parameters explored in the main text and Supplementary Information compared with estimates for the early transmission of SARS-CoV-2 (wild strain). The white cells correspond to free parameters, while the shaded cells correspond to quantities depending on the other parameters. *Estimates are based on the incubation period; the latent period can be shorter in the presence of pre-symptomatic transmission. †Estimates combine mean generation time and serial interval.

Parameter	Range considered	SARS-CoV-2 (wild strain)	References
Reproduction number	[1.5, 3]	[2, 3.5]	[47–50]
Mean latent period	[2, 42] days	[5, 7] days*	[48, 49, 51]
Mean infectious period	[2, 14] days	—	—
Mean post-infectious period	[5, 20] days	—	—
Mean generation time	[4, 44] days	[4, 7.5] days†	[48]
Mean detectable period	[7.7, 22.7] days	[7, 22] days	[6, 28, 29]

ization scheme for the disease progression. Each individual, at any time point, is assigned to a compartment corresponding to their particular disease-related state. An individual who gets infected will go through the following sequence of states: susceptible (S; pre-exposure), latent (L; exposed, but do not yet transmit the infectious pathogen), infectious (I; can transmit the disease), post-infectious (P; no longer infectious), and recovered (R). In our model, we assume that only infectious and post-infectious can be detected through wastewater, which we regroup under the *detectable* (D) state. The addition of the post-infectious state is necessary because diseases like COVID-19 can be detected through wastewater for a long period of time beyond the infectious period [6, 28, 29].

The contagion dynamics in each location is strongly influenced by the basic reproduction number $\mathcal{R}_0$—the average number of secondary infections caused by an infectious individual in a fully healthy population. In our model, the reproduction number $\mathcal{R}_0$ varies slightly from one subpopulation to another, since it is proportional to the largest eigenvalue of the age-structured contact matrices [23]. Therefore, we either fix the mean reproduction number $\langle \mathcal{R}_0 \rangle$, averaged over all subpopulations, or the reproduction number at the source for the Alpha variant case study, thereby fixing the transmissibility of the pathogen.

Another important driver of the epidemic is the mean generation time T_{gen}, i.e., the average time between the exposures of an infector-infectee pair. It is a function of the mean latent period T_{lat} and the mean infectious period T_{inf}. Since the period an individual spends in a certain compartment is typically not exponentially distributed [52, 53], we add realism to our model by further decomposing the infectious states in two substates, namely I_1 and I_2. The resulting infectious period is then gamma-distributed. Because of these two substates, the mean generation time T_{gen} in our model is [53]

$$T_{\mathrm{gen}} = T_{\mathrm{lat}} + \frac{3}{4} T_{\mathrm{inf}} \, . \tag{4}$$

The detection process on the other hand will depend on the detectable period of the disease, T_{det}, which

corresponds to

$$T_{\text{det}} = T_{\text{inf}} + T_{\text{post}} \ , \tag{5}$$

where T_{post} is the mean post-infectious period. Similarly to the infectious state, we decompose the post-infectious state in two substates, namely P_1 and P_2, but this does not affect the expression for T_{det}. In Table 1, we compile the ranges considered in this study for these important parameters of our model and compare them with estimated ranges for relevant diseases.

Lastly, an important quantity introduced in the main text to characterize the variation of the time to first detection with the contagion parameter is the doubling time T_2. For the specific compartmental model considered here, it can obtained from the following implicit nonlinear equation [52] relating $\mathcal{R}_0$ to the *growth rate* λ:

$$\mathcal{R}_0 = \frac{\lambda T_{\text{inf}} \left(1 + \lambda T_{\text{lat}}\right)}{\left[1 - \left(1 + \frac{\lambda T_{\text{inf}}}{2}\right)^{-2}\right]} \ . \tag{6}$$

We solve this equation for λ, then we have $T_2 = \ln 2/\lambda$.

Aircraft wastewater detection. In our model, a detectable individual passing through a sentinel site is detected with probablity p_{det}, a multifaceted measure anticipated to fluctuate based on varying travel pathways. This probability hinges on several factors, including the capabilities of airport wastewater surveillance, the duration of the flight, the diverse sociodemographic profiles of the passengers [16], etc. In our analysis, we neglect this heterogeneity and assume that, on average, the detection probability p_{det} is uniform across all inbound international flights arriving at any given sentinel site. To provide a rationale for the spectrum of detection rates examined in this study, we break down the probability into the following components

$$p_{\text{det}} = p_{\text{lav}} \times p_{\text{shed}} \times p_{\text{sample}} \ , \tag{7}$$

where p_{lav} represents the likelihood that an individual will utilize the lavatory and consequently deposit detectable genetic traces of the pathogen in the wastewater, p_{shed} denotes the probability that a detectable individual is actively shedding the pathogen at levels sufficient for detection in the wastewater, and p_{sample} refers to the proportion of flights that are subjected to sampling."

The probability of adult passengers using the lavatory on flights, crucial for estimating p_{lav}, is surveyed to be less than 13% on short-haul and less than 36% on long-haul flights [16]. This estimation assists in determining the probability of pathogen traces being left in wastewater. Further, p_{shed}, the probability of detectable pathogen shedding in fecal matter, ranges between 30% and 60% [16]. Assuming that

all international flights undergo sampling ($p_{\text{sample}} = 100\%$), the resulting detection probability (p_{det}) is calculated to be between 12% and 22%, corresponding to a detection rate of about 16% on long-haul flights. This estimate might be low for viruses like SARS-CoV-2, as individuals can leave genetic material in the wastewater without defecating [54], such as by disposing of a used tissue in the toilet. Furthermore, previous studies [21] have evaluated the effectiveness of aircraft wastewater surveillance by analyzing long-haul repatriation flights to Australia and found a prediction accuracy of 83.7% to identify flights with passengers who would test positive for COVID-19 during the subsequent two-week isolation period. Translating this value to an individual's marginal detection probability (p_{det}) is complex, as the number of COVID-19 cases per flight varied significantly, averaging 4.62 cases. Accounting for false-positive wastewater results, assuming each case had an equal probability to be detected, and a flat prior on p_{det}, we find a median marginal detection rate of 51% (90% CI, 28–72). However, this value might be inflated, notably due to the persistent nature of fecal RNA shedding compared to respiratory shedding [55].

Given the varying estimates discussed earlier, we explore detection rates up to 32%. However, acknowledging that not all international flights are long-haul and that only a fraction (e.g., 25%) of flights might be sampled, p_{det} could be as low as 4%. Our study comprehensively investigates and presents results across this entire range of p_{det}

Probability generating function methodology. The mechanistic GLEAM model relies on large-scale stochastic simulations, which are computationally expensive. For most results in this paper, we instead make use of an alternative description in terms of probability generating functions (PGFs). This mathematical tool is standard in computational epidemiology [56–58], and has found many applications, notably to quantify the risk of introduction and predict the arrival time of diseases [59–63].

PGFs are useful to *count* various things. Here we are counting individuals based on certain properties: their age, their location, and their epidemiological state. We define s_σ as the number of individuals of *type* σ. For instance, s_σ could represent the current number of latent individuals in a given location and a certain age, or the cumulative number of individuals of a certain age that have been detected on a particular travel route. We use the vector $\boldsymbol{s}$ to encapsulate all these numbers.

To capture the full stochastic evolution of the system, we need to characterize the probability distribution $P(\boldsymbol{s}, t)$. We encode this distribution with a multivariate PGF

$$\Psi^t(\boldsymbol{x}) = \sum_{\boldsymbol{s}} P(\boldsymbol{s}, t) \prod_{\sigma} x_\sigma^{s_\sigma} \ , \tag{8}$$

where the sum (product) runs over all possible values of $\boldsymbol{s}$ (σ) and each x_σ is a variable that act as a placeholder—it does not mean anything and only serve to encode the probability distribution. The vector $\boldsymbol{x}$ encapsulates all these variables.

In the early stage of an epidemic, the GLEAM model is equivalent to a *multitype branching process*, in which case we can solve the PGF through a recursive equation of the form

$$\Psi^{t+1}(\boldsymbol{x}) = \Psi^{t}[\boldsymbol{F}(\boldsymbol{x})] \,, \tag{9}$$

where $\boldsymbol{F}(\boldsymbol{x})$ is a vector of PGFs and each element $F_\sigma(\boldsymbol{x})$ is itself a PGF that characterizes the offspring distribution of an individual of type σ. For instance, the offspring distribution of an individual in the I_1 state would give the probability that this individual generates a certain number of new latent individuals of each type through infections at the next time steps and the probability that this individual transitions to the I_2 state. Computing the full distribution $P(\boldsymbol{s}, t)$ is generally out of reach—the number of terms explodes combinatorially. However, computing marginal or joint distributions for a few *observables*, like the total number of individuals in a particular state or the cumulative number of detections, is possible (see Supplementary Information Sec. 1).

Altogether, the recursive evaluation of PGFs and their numerical inversion to recover probability distributions represents a very efficient computational alternative to Monte-Carlo simulations of GLEAM. Most notably, scanning different initial conditions is computationally cheap, since in Eq. (9), the PGF $\Psi^0(\boldsymbol{x})$ specifying the initial conditions is evaluated at the *end* of the recursion. This crucially allows us to extract distributions of observables, like the time to first detection, assuming the epidemic could have started from any of the 3200+ subpopulations of our model, a task that would be prohibitive with a purely simulation-based framework. See Sec. 1 of the Supplementary Information for an in-depth description and characterization of the PGF methodology.

Optimization of the global WWSN. As a first approach to maximize the detections of a global WWSN, we compute centrality scores for the airports and select the ones with the highest scores. For each subpopulation l, we have a number $N_{l\to\nu}$ of individuals per day who will travel and arrive at airport ν on an international flight, either as a final destination or for a connection. This flow of international passengers forms a weighted bipartite network connecting international airports ν to subpopulations l. We rank airports based on their *volume* of international travel,

$$c_\nu^{\mathrm{vol}} = \sum_l N_{l\to\nu} \,. \tag{10}$$

We additionally rank airports based on their *entropy*,

$$c_\nu^{\mathrm{ent}} = -\sum_l \left(\frac{N_{l\to\nu}}{\sum_{l'} N_{l'\to\nu}}\right) \log\left(\frac{N_{l\to\nu}}{\sum_{l'} N_{l'\to\nu}}\right) \,. \tag{11}$$

Also known as Shannon's diversity index, this measure favors airports with a broad and homogeneous coverage of the different subpopulations.

As a second approach, we aim to directly minimize the mean time to first detection of an epidemic, averaged over all potential origins. We can assign an arbitrary prior probability $P(l)$ for location l to be the origin of an epidemic, resulting in the following objective function

$$\Phi(\mathcal{S}) \equiv T_{\mathrm{fd}}(\mathcal{S}) = \sum_{l} P(l)\, T_{\mathrm{fd}}(\mathcal{S}, l) \;, \tag{12}$$

where $T_{\mathrm{fd}}(\mathcal{S}, l)$ is the mean time to first detection, assuming the epidemic started in subpopulation l and that the WWSN consists of the set of sentinel airports $\mathcal{S}$. This corresponds to the posterior mean of T_{fd} over all locations, which is proportional to $E(\mathcal{S})$, the excess time to first detection. For the global optimization, we use $P(l) = \text{const.}\ \forall l$, i.e., all subpopulations are equiprobable source. For the targeted optimization, we use $P(l) = \text{const.}$ for the locations in the targeted region and $P(l) = 0$ otherwise.

We conjecture that $-\Phi(\mathcal{S})$ is a *monotone submodular set function* [64]. We proved this statement in Sec. 2 of the Supplementary Information for a very accurate approximation of $-\Phi(\mathcal{S})$, but the exact case remains to be proven. Monotone submodular functions have desirable properties when it comes to discrete optimization problems. While minimizing $\Phi(\mathcal{S})$ (equivalently maximizing $-\Phi(\mathcal{S})$) is an NP-hard problem, we have a guarantee on the performance of a *greedy* optimization algorithm—there exists an upper bound on the value of $\Phi(\mathcal{S})$ obtained through this approach [64]. But most importantly, it is known that in practice, a greedy algorithm will find a solution that is very close to the optimal one. Consequently, to minimize the objection function Eq. (12), we use the following greedy optimization scheme:

1. define an initial set $\mathcal{S}$ (can be empty);
2. for each airport $\nu \notin \mathcal{S}$, compute $\Phi(\mathcal{S} \cup \{\nu\})$;
3. update the set $\mathcal{S} \leftarrow \mathcal{S} \cup \{\nu^\star\}$, where $\nu^\star$ is the sentinel airport that minimizes the objective function;
4. repeat steps 2-3 until a desired number of sentinels is reached.

The 'greedy' name comes from the fact that we are successively choosing a locally optimal choice at each stage of the algorithm (here in step 3).

Acknowledgements

This work was in part supported by the Bill & Melinda Gates Foundation INV-058220. Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 Generic License has already been assigned to the Author Accepted Manuscript version that might arise from this submission. We acknowledge support from the CDC-75D301Cl4810 contract and the cooperative agreement CDC-RFA-FT-23-0069 from the CDC's Center for Forecasting and Outbreak Analytics. GS acknowledges financial support from

the Fonds de recherche du Québec – Nature et technologies (project 313475). LHD is supported by the National Institutes of Health 2P20GM125498-06 Centers of Biomedical Research Excellence Award. AA acknowledges support from the Natural Sciences and Engineering Research Council of Canada (projects 2019-05183 and 2024-05626). The findings and conclusions in this study are those of the authors and do not necessarily represent the official position of the funding agencies, the CDC, or the U.S. Department of Health and Human Services of the United States.

References

1. Williams, G. H. *et al.* SARS-CoV-2 testing and sequencing for international arrivals reveals significant cross border transmission of high risk variants into the United Kingdom. *eClinicalMedicine* **38,** 101021 (Aug. 2021).
2. Bart, S. M. *et al.* Effect of Predeparture Testing on Postarrival SARS-CoV-2–Positive Test Results Among International Travelers — CDC Traveler-Based Genomic Surveillance Program, Four U.S. Airports, March–September 2022. *Morb. Mortal. Wkly. Rep.* **72,** 206–209 (8 Feb. 2023).
3. Wegrzyn, R. D. *et al.* Early Detection of Severe Acute Respiratory Syndrome Coronavirus 2 Variants Using Traveler-based Genomic Surveillance at 4 US Airports, September 2021–January 2022. *Clin. Infect. Dis.* **76,** e540–e543 (2022).
4. National Academies of Sciences, Engineering, and Medicine. Improving the CDC Quarantine Station Network's Response to Emerging Threats (The National Academies Press, Washington, DC, 2022).
5. Kucharski, A. J. *et al.* Real-time surveillance of international SARS-CoV-2 prevalence using systematic traveller arrival screening: An observational study. *PLOS Med.* **20,** 1–15 (Sept. 2023).
6. Foladori, P. *et al.* SARS-CoV-2 from faeces to wastewater treatment: What do we know? A review. *Sci. Total Environ.* **743,** 140444 (2020).
7. Keshaviah, A. *et al.* Separating signal from noise in wastewater data: An algorithm to identify community-level COVID-19 surges in real time. *Proc. Natl. Acad. Sci. U.S.A.* **120,** e2216021120 (2023).
8. Hovi, T. *et al.* Role of environmental poliovirus surveillance in global polio eradication and beyond. *Epidemiol. Infect.* **140,** 1–13 (2012).
9. Brouwer, A. F. *et al.* Epidemiology of the silent polio outbreak in Rahat, Israel, based on modeling of environmental surveillance data. *Proc. Natl. Acad. Sci. U.S.A.* **115,** E10625–E10633 (2018).
10. Wolfe, M. K. *et al.* Wastewater-Based Detection of Two Influenza Outbreaks. *Environ. Sci. Technol. Lett.* **9,** 687–692 (2022).
11. Mercier, E. *et al.* Municipal and neighbourhood level wastewater surveillance and subtyping of an influenza virus outbreak. *Sci. Rep.* **12,** 15777 (2022).
12. Krzysztoszek, A. *et al.* Investigation of airport sewage to detect importation of poliovirus, Poland, 2017 to 2020. *Eurosurveillance* **27** (2022).

13. La Rosa, G. *et al.* Detection of Monkeypox Virus DNA in Airport Wastewater, Rome, Italy. *Emerg. Infect. Dis.* **29,** 193 (2023).

14. Farkas, K. *et al.* Wastewater-based monitoring of SARS-CoV-2 at UK airports and its potential role in international public health surveillance. *PLOS Glob. Public Health* **3,** 1–16 (2023).

15. Li, J. *et al.* A global aircraft-based wastewater genomic surveillance network for early warning of future pandemics. *Lancet Glob. Health.* **11,** e791–e795 (2023).

16. Jones, D. L. *et al.* Suitability of aircraft wastewater for pathogen detection and public health surveillance. *Sci. Total Environ.* **856,** 159162 (2023).

17. Shingleton, J. W., Lilley, C. J. & Wade, M. J. Evaluating the theoretical performance of aircraft wastewater monitoring as a tool for SARS-CoV-2 surveillance. *PLOS Glob. Public Health* **3,** 1–11 (June 2023).

18. Bivins, A. *et al.* The lavatory lens: Tracking the global movement of pathogens via aircraft wastewater. *Crit. Rev. Environ. Sci. Technol.,* 1–21 (2023).

19. Albastaki, A. *et al.* First confirmed detection of SARS-COV-2 in untreated municipal and aircraft wastewater in Dubai, UAE: The use of wastewater based epidemiology as an early warning tool to monitor the prevalence of COVID-19. *Sci. Total Environ.* **760,** 143350 (2021).

20. Ahmed, W. *et al.* Detection of the Omicron (B.1.1.529) variant of SARS-CoV-2 in aircraft wastewater. *Sci. Total Environ.* **820,** 153171 (2022).

21. Ahmed, W. *et al.* Wastewater surveillance demonstrates high predictive value for COVID-19 infection on board repatriation flights to Australia. *Environ. Int.* **158,** 106938 (2022).

22. Morfino, R. C. *et al.* Notes from the Field: Aircraft Wastewater Surveillance for Early Detection of SARS-CoV-2 Variants — John F. Kennedy International Airport, New York City, August–September 2022. *Morb. Mortal. Wkly. Rep.* **72,** 210–211 (8 2023).

23. Mistry, D. *et al.* Inferring high-resolution human mixing patterns for disease modeling. *Nat. Commun.* **12,** 323 (2021).

24. Gomes, M. F. C. *et al.* Assessing the International Spreading Risk Associated with the 2014 West African Ebola Outbreak. *PLoS Currents* **6** (2014).

25. Tizzoni, M. *et al.* Real-time numerical forecast of global epidemic spreading: Case study of 2009 A/H1N1pdm. *BMC Med.* **10,** 1–31 (1 Dec. 2012).

26. Chinazzi, M. *et al.* The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* **368,** 395–400 (2020).

27. Davis, J. T. *et al.* Cryptic transmission of SARS-CoV-2 and the first COVID-19 wave. *Nature* **600,** 127–132 (2021).

28. Weiss, A., Jellingsø, M. & Sommer, M. O. A. Spatial and temporal dynamics of SARS-CoV-2 in COVID-19 patients: A systematic review and meta-analysis. *EBioMedicine* **58,** 102916 (2020).

29. Li, Q. *et al.* Number of COVID-19 cases required in a population to detect SARS-CoV-2 RNA in wastewater in the province of Alberta, Canada: Sensitivity assessment. *J. Environ. Sci.* **125,** 843–850 (2023).

30. Gautreau, A., Barrat, A. & Barthélemy, M. Global disease spread: Statistics and estimation of arrival times. *J. Theor. Biol.* **251,** 509–522 (2008).

31. Kraemer, M. U. G. *et al.* Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence. *Science* **373,** 889–895 (2021).

32. Davies, N. G. *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* **372,** eabg3055 (2021).

33. Volz, E. *et al.* Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature* **593,** 266–269 (2021).

34. Niehus, R. *et al.* Using observational data to quantify bias of traveller-derived COVID-19 prevalence estimates in Wuhan, China. *Lancet Infect. Dis.* **20,** 803–808 (2020).

35. Keshaviah, A. *et al.* Wastewater monitoring can anchor global disease surveillance systems. *Lancet Glob. Health* **11,** e976–e981 (2023).

36. Polgreen, P. M. *et al.* Optimizing Influenza Sentinel Surveillance at the State Level. *Am. J. Epidemiol.* **170,** 1300–1306 (Oct. 2009).

37. Scarpino, S. V., Dimitrov, N. B. & Meyers, L. A. Optimizing Provider Recruitment for Influenza Surveillance Networks. *PLOS Comput. Biol.* **8,** 1–12 (Apr. 2012).

38. Scarpino, S., Meyers, L. A. & Johansson, M. Design Strategies for Efficient Arbovirus Surveillance. *Emerg. Infect. Dis.* **23,** 642 (2017).

39. Jones, K. E. *et al.* Global trends in emerging infectious diseases. *Nature* **451,** 990–993 (2008).

40. Grace, D. *et al.* Mapping of poverty and likely zoonoses hotspots. Zoonoses Project 4. Report to the UK Department for International Development. (2012).

41. Watts, N. *et al.* Health and climate change: policy responses to protect public health. *Lancet* **386,** 1861–1914 (2015).

42. Carlson, C. J. *et al.* Climate change increases cross-species viral transmission risk. *Nature* **607,** 555–562 (2022).

43. Becker, D. J. *et al.* Optimising predictive models to prioritise viral discovery in zoonotic reservoirs. *Lancet Microbe* **3,** e625–e637 (2022).

44. Zipf, G. K. The P1 P2/D Hypothesis: On the Intercity Movement of Persons. *Am. Sociol. Rev.* **11,** 677–686 (1946).

45. Balcan, D. *et al.* Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 21484–21489 (2009).

46. Simini, F. *et al.* A universal model for mobility and migration patterns. *Nature* **484,** 96–100 (Apr. 2012).

47. Riou, J. & Althaus, C. L. Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. *Eurosurveillance* **25** (2020).

548 48. Li, Q. *et al.* Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *N.*
549 *Engl. J. Med.* **382,** 1199–1207 (2020).

550 49. Adhikari, S. P. *et al.* Epidemiology, causes, clinical manifestation and diagnosis, prevention and control of
551 coronavirus disease (COVID-19) during the early outbreak period: a scoping review. *Infect. Dis. Poverty* **9,** 29
552 (2020).

553 50. Read, J. M. *et al.* Novel coronavirus 2019-nCoV (COVID-19): early estimation of epidemiological parameters
554 and epidemic size estimates. *Philos. Trans. R. Soc. B* **376,** 20200265 (2021).

555 51. Backer, J. A., Klinkenberg, D. & Wallinga, J. Incubation period of 2019 novel coronavirus (2019-nCoV) infections
556 among travellers from Wuhan, China, 20–28 January 2020. *Eurosurveillance* **25** (2020).

557 52. Wearing, H. J., Rohani, P. & Keeling, M. J. Appropriate Models for the Management of Infectious Diseases.
558 *PLOS Med.* **2,** 621–627 (July 2005).

559 53. Krylova, O. & Earn, D. J. D. Effects of the infectious period distribution on predicted transitions in childhood
560 disease dynamics. *J. R. Soc. Interface* **10,** 20130098 (2013).

561 54. Crank, K. *et al.* Contribution of SARS-CoV-2 RNA shedding routes to RNA loads in wastewater. *Sci. Total*
562 *Environ.* **806,** 150376 (2022).

563 55. Zhang, Y. *et al.* Prevalence and Persistent Shedding of Fecal SARS-CoV-2 RNA in Patients With COVID-19
564 Infection: A Systematic Review and Meta-analysis. *Clinical and Translational Gastroenterology* **12** (2021).

565 56. Brauer, F. *et al.* Mathematical epidemiology (Springer, 2008).

566 57. Allen, L. J. S. An introduction to stochastic processes with applications to biology (CRC press, 2010).

567 58. Miller, J. C. A primer on the use of probability generating functions in infectious disease modeling. *Infect. Dis.*
568 *Model.* **3,** 192–248 (2018).

569 59. Johansson, M. A. *et al.* Assessing the Risk of International Spread of Yellow Fever Virus: A Mathematical
570 Analysis of an Urban Outbreak in Asunción, 2008. *Am. J. Trop. Med. Hyg.* **86,** 349–358 (2012).

571 60. Johansson, M. A. *et al.* Nowcasting the Spread of Chikungunya Virus in the Americas. *PLOS ONE* **9,** e104915
572 (2014).

573 61. Mier-y-Teran-Romero, L., Tatem, A. J. & Johansson, M. A. Mosquitoes on a plane: Disinsection will not stop
574 the spread of vector-borne pathogens, a simulation study. *PLOS Negl. Trop. Dis.* **11,** 1–13 (July 2017).

575 62. Lai, S. *et al.* Seasonal and interannual risks of dengue introduction from South-East Asia into China, 2005-2015.
576 *PLOS Negl. Trop. Dis.* **12,** 1–16 (Nov. 2018).

577 63. Truelove, S. *et al.* Epidemics, Air Travel, and Elimination in a Globalized World: The Case of Measles. *medRxiv*
578 (2020).

579 64. Nemhauser, G. L., Wolsey, L. A. & Fisher, M. L. An analysis of approximations for maximizing submodular set
580 functions—I. *Math. Program.* **14,** 265–294 (1978).

Establishing a Global Wastewater Surveillance Network at Airports for Early Detection of Emerging Pathogens: a Modeling Study

Supplementary Information

Guillaume St-Onge[1], Jessica T. Davis[1], Laurent Hébert-Dufresne[2,3], Antoine Allard[2,3], Alessandra Urbinati[1], Samuel V. Scarpino[4,5,6], Matteo Chinazzi[1,7], and Alessandro Vespignani[1]

[1] *Laboratory for the Modeling of Biological and Socio-technical Systems, Northeastern University, Boston, MA 02115, USA*
[2] *Vermont Complex Systems Center, University of Vermont, Burlington, VT 05401, USA*
[3] *Département de physique, de génie physique et d'optique, Université Laval, Québec City, QC G1V 0A6, Canada*
[4] *Institute for Experiential AI, Northeastern University, Boston, MA 02115, USA*
[5] *Network Science Institute, Northeastern University, Boston, MA 02115, USA*
[6] *Santa Fe Institute, Santa Fe, NM 87501, USA*
[7] *The Roux Institute, Northeastern University, Portland, ME 04101, USA*

June 5, 2024

Contents

1 Model description

1.1 Global Epidemic and Mobility Model

The GLEAM (Global Epidemic and Mobility) model is a stochastic epidemic metapopulation framework that incorporates age-based contact matrices and data on human mobility. The approach has been documented in previous publications [1, 2] and has been used to study the spread of diseases such as Ebola [3], Zika [4], and COVID-19 [5, 6]. The model uses a Voronoi tessellation to create a metapopulation network with over 3,200 subpopulations, covering areas of the globe inhabited by humans. These subpopulations are anchored around major transport hubs like airports and are themselves divided into cells measuring around 25 x 25 kilometers, equivalent to 15 x 15 arc minutes. In

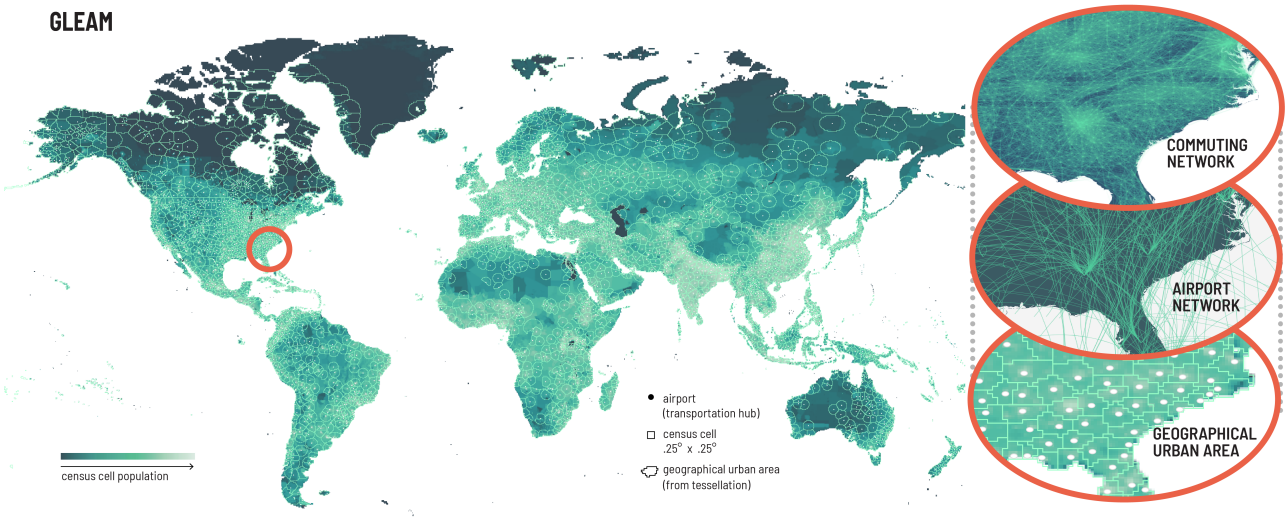

Figure S1: Schematic representation of GLEAM. (left) Voronoi tessellation of the globe centered around transportation hubs (airports) creating each of the 3200+ geographical units that we refer as subpopulations. Each subpopulation is constructed from census cells of approximately 25km × 25km. (right) For a particular region, we illustrate the subpopulations and the two mobility layers— air travel (long-range) and commuting (short-range).

addition, the model integrates cell-level population data [7] and subpopulation-level age-specific contact patterns using the contact matrices developed in [8]. Here, we consider individuals divided into 5 age groups: [0-4, 5-17, 18-49, 50-64, 65+].

The individual subpopulations are connected through a human mobility layer that combines both short-range (i.e., commuting) and long-range (i.e., flights) mobility data. Commuting data is sourced from the Offices of Statistics for 30 countries on 5 continents. To harmonize the varying spatial resolutions of commuting data across different countries and to address data availability gaps, the short-range mobility layer is synthetically generated where necessary. This is achieved by relying on the "gravity law" [1, 9, 10], which is calibrated on the available data. Air-travel data from the Official Aviation Guide (OAG) and IATA databases is used to build an origin-destination network, incorporating connecting flight information. The network provides daily passenger flows between airports globally, which we systematically map and aggregate at the subpopulation level. Figure S1 displays the geographical resolution of the model for selected regions, illustrating both the short-range and long-range mobility networks and the global population structure.

Compartmental model

The synthetic world created by the human mobility layer couples the epidemic dynamics unfolding within each subpopulation. To model the infection process, we adopt an extended SLIR-like model in which individuals are either susceptible (S), latent (L), infectious (I), post-infectious (P), or removed (R). We further subdivide the infectious and post-infectious compartments into two stages, namely I_1, I_2, P_1, and P_2 to more realistically model the timing of the disease progression [11, 12]. Susceptible individuals become latent through interactions with infectious individuals, at a rate Λ, the *effective force of infection*, which depends on the age of the susceptible individual but also on the whole state of the system, and thus varies in time. Latent individuals progress to the first infectious stage at a constant rate inversely proportional to the mean latent period $\eta \equiv T_{\rm lat}$. Infectious individuals in the first stage I_1 progress to the second stage I_2 at a rate $\mu \equiv 2 \times T_{\rm inf}^{-1}$, where $T_{\rm inf}$ is the mean infectious period; the process is identical for

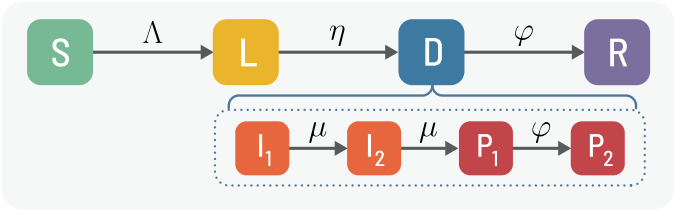

Figure S2: **Compartmental model for wastewater surveillance at airports.**

the transition of I_2 individuals to P_1. Similarly, post-infectious individuals in the first stage (P_1) progress to the second stage P_2 at a rate $\varphi \equiv 2 \times T_{\rm post}^{-1}$, where $T_{\rm post}$ is the mean post-infectious period, then progress to the recovered stage at the same rate φ. The post-infectious period is the length of time that an individual can still shed the virus and remain detectable through wastewater, but not generate any more new infections. Once an individual is in the removed compartment, it can no longer be detected. Here we assume only infectious and post-infectious individuals can be detected through wastewater, which is why we regroup them in a *detectable* (D) meta-compartment. The various transitions and their rates are portrayed in Fig. S2.

Because of the subdivision of the infectious and post-infectious states, the infectious and post-infectious periods are gamma-distributed, while the latent period is exponentially distributed. The *generation time*, the time between the exposures of an infector-infectee pair, will also be gamma-distributed with mean $T_{\rm gen}$ expressed as [11, 12]

$$T_{\rm gen} = T_{\rm lat} + \left(\frac{n_{\rm inf}+1}{2n_{\rm inf}}\right) T_{\rm inf} \,, \tag{1}$$

where $n_{\rm inf}$ is the number of infectious states—in our model $n_{\rm inf} = 2$. Similarly, the *detectable period*—the length of time an individual can be detected through wastewater—is gamma-distributed, with mean $T_{\rm det} = T_{\rm inf} + T_{\rm post}$.

Stochastic simulation of the transmission and mobility dynamics

With the mobility data layers and the disease dynamics defined, the number of individuals in each compartment c, age bracket a, and subpopulation l follows a discrete and stochastic dynamical equation that reads as

$$X_l^{[c,a]}(t+\Delta t) - X_l^{[c,a]}(t) = \Delta X_l^{[c,a]} + \Omega_l([c,a]) \tag{2}$$

where the term $\Delta X_l^{[c,a]}$ represents the changes induced by the disease dynamics and $\Omega_l([c,a])$ represents the variations due to air travel. In this study, each day is subdivided in $m = 12$ time steps, i.e., $\Delta t = 1/m$ days. While the disease dynamics part $\Delta X_l^{[c,a]}$ is applied at every of these time steps, the variations due to traveling $\Omega_l([c,a])$ are introduced daily. The latter $\Omega_l([c,a])$ is directly extracted from a multinomial distribution associated with the age-specific probability of travel per day, as defined by the global air-travel network.

The variation $\Delta X_l^{[c,a]}$ is determined by summing over all transitions in and out of the disease compartment c for the age group a,

$$\Delta X_l^{[c,a]} = \sum_{[c',a]} \left\{ -\mathcal{D}_l([c,a],[c',a]) + \mathcal{D}_l([c',a],[c,a]) \right\} , \tag{3}$$

where $\mathcal{D}_l([c,a],[c',a])$ represents the number of transitions from $[c,a]$ to $[c',a]$ during the time interval Δt. For all spontaneous transitions, like from latent to infectious, $\mathcal{D}_l([c,a],[c',a])$ is simply extracted from a binomial distribution.

The generation of new infections is a more complex procedure. First, it hinges on the age-structured contacts matrix $\mathbf{C}$, which gives the expected number of contacts per day between each age pair (a,a') in a given location l. We consider interactions in four social settings: contacts at school ($\mathbf{C}_{school}$), workplace ($\mathbf{C}_{work}$), home ($\mathbf{C}_{home}$), and in the general community ($\mathbf{C}_{community}$), which are linearly combined to create $\mathbf{C}$, as defined in Ref. [8]. Second, the mobility due to the commuting flows is also taken into account using a time scale separation approximation, as

detailed in Ref. [1]. Altogether, these two factors contribute to the effective force of infections $\Lambda([l,a])$ acting on susceptible individuals in subpopulation l and age a, resulting in new transitions of the form $\mathcal{D}_l([c,a],[c',a])$.

Initial conditions are established by defining the quantity and location of individuals who can spread the infection. Subsequently, GLEAM tracks the number of individuals in each disease compartment for every subpopulation over time. Please see Ref. [2] for a more in-depth discussion of the simulation framework.

Wastewater detection at airports

We assume that a subset of all airports $\mathcal{S} = \{\nu_1, \nu_2, \dots\}$—*sentinels*—monitor the wastewater of incoming *international* aircrafts. Consequently, each detectable international traveler passing through a sentinel has a probability $p_{\rm det}$ of leading to a detection. As detailed in the Methods section of the main text, the probability $p_{\rm det}$ is a complex quantity combining multiple factors, such as sampling frequency of aircrafts, length of the flight, demography, etc. We settled on performing an extensive sensitivity analysis spanning the range 4–32%.

In GLEAM's stochastic simulation process, detections at sentinels are aggregated during the post-processing of the simulation data, which keeps track of all mobility-induced changes $\Omega_l([c,a])$. Since GLEAM produces origin-destination travel patterns at the level of subpopulations l, we extract from the global air-travel network the probability of detection $p_{l,l'}$ for each travel $l \to l'$. Each travel $l \to l'$ is in fact associated with a set of potential airport *paths* of the form $\mathcal{P} = \nu_1 \to \nu_2 \to \cdots \to \nu_k$. If the path $\mathcal{P}$ contains an international flight $\nu \to \nu'$, where $\nu' \in \mathcal{S}$, then detectable individuals taking this path will be detected with probability $p_{\rm det}$. If a path $\mathcal{P}$ contains $\mathcal{N}(\mathcal{P})$ such international flight to any sentinel, the probability of detection is $1-(1-p_{\rm det})^{\mathcal{N}(\mathcal{P})}$. We therefore calculate

$$p_{l,l'} = \sum_{\mathcal{P}} P(\mathcal{P}|l \to l') \left[1-(1-p_{\rm det})^{\mathcal{N}(\mathcal{P})}\right] , \tag{4}$$

where $P(\mathcal{P}|l \to l')$ is the probability an individual will take the path $\mathcal{P}$, given the travel $l \to l'$.

1.2 Probability generating function framework

To characterize the early phase of an epidemic, we can map the transmission tree to a branching process, from which an efficient *probability generating function* (PGF) methodology can be leveraged [13–17]. This is a standard approach in computational epidemiology and has been used, among others, to characterize the time evolution of contagion on heterogeneous networks [18–20], and to quantify the risk of introduction in metapopulation models [21–25]. Here we generalize and adapt this methodology by mapping to a multitype branching process the age-structured, stochastic, metapopulation dynamics of GLEAM, the disease progression, and the wastewater surveillance at airports (see Fig. S3).

PGFs are used to encode discrete probability distributions with functions. In the present case, we want to encode the full distribution for the state of the epidemic at all times t—the number of days since the start of the epidemic. The state of the epidemic includes the number of individuals in each stage of the disease, of each age, and in each subpopulation, but also other quantities we want to "measure" (evaluate an associated probability distribution), like the cumulative number of exported cases and the cumulative number of detections on each origin-destination travel ending at a sentinel.

To introduce our formalism, let us define the multi-index $\alpha = (l,a)$ characterizing the location l and the age a of an individual—we refer to α as the *category* of an individual or agent. We define ℓ_α as the number of latent individuals of category α. Since we divide the number of infectious and post-infectious states into two stages ($\rm I_1$,$\rm I_2$,$\rm P_1$, and $\rm P_2$) we identify them by the numbers $i_{\alpha,1}$, $i_{\alpha,2}$, and $j_{\alpha,1}$,$j_{\alpha,2}$ respectively. To track the number of exported cases, we introduce e_α as the cumulative number of α-agents, either latent or infectious, who were infected previously in another location—they were previously of another category—then traveled, thereby becoming of category α. Finally, to track wastewater detections at sentinel airports, we define $d_{\alpha,\alpha'}$ the cumulative number of detections for international travel $l \to l'$ resulting in a change of category $\alpha \to \alpha'$ for a given age group a. Note that we do not track susceptible or

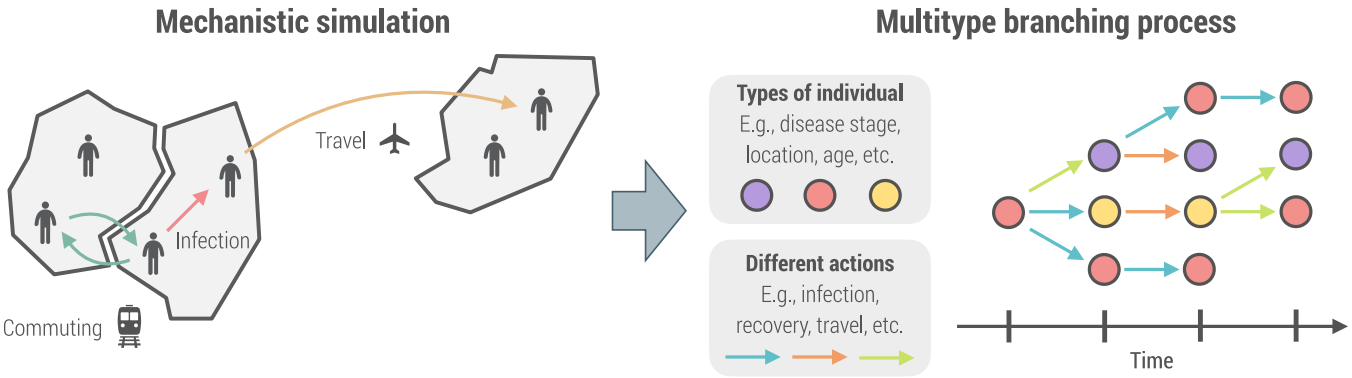

Figure S3: Mapping GLEAM to a multitype branching process. Infections happen within subpopulations—and across adjacent subpopulations when accounting for commuting—and agents can travel between the subpopulations. Individuals are distinguished by their *type*, which encapsulates everything that distinguishes them (like disease stage, age, and location). The early phase of an epidemic naturally takes the form of an event tree encoding all actions happening at each time step, including travel and transmission. This event tree is mathematically described via a multitype branching process.

recovered individuals in this framework, since they do not play a significant role in the early phase of an outbreak. The whole state of the system is then described by the tuple of vectors $(\boldsymbol{\ell}, \boldsymbol{i}_1, \boldsymbol{i}_2, \boldsymbol{j}_1, \boldsymbol{j}_2, \boldsymbol{e}, \boldsymbol{d})$, where for instance $\boldsymbol{\ell} = [l_{\alpha_1}, l_{\alpha_2}, \dots]$.

To simplify the notation, we define the *state vector*

$$\boldsymbol{s} = [l_{\alpha_1}, l_{\alpha_2}, \dots, i_{\alpha_1,1}, i_{\alpha_2,2}, \dots, d_{\alpha_1,\alpha_2}, d_{\alpha_1,\alpha_3}, \dots] \equiv [s_1, s_2, \dots] \,, \tag{5}$$

as the concatenation of previous vectors, encoding all information about the system. Each component s_σ of the vector identifies a number we keep track of (e.g., number of infectious of a specific category α) and is associated with nodes of a specific color in Fig. S3; we refer to σ as the *type* of each node in the multitype branching process representation of the epidemic and mobility dynamics.

Since the state vector is a random variable $\boldsymbol{s}_t$ at each time t, we write the probability of observing a particular state at time t as $P(\boldsymbol{s}_t = \boldsymbol{s}) \equiv P(\boldsymbol{s}, t)$. Finally, we encode this distribution with a PGF (in the compact format) as

$$\Psi^t(\boldsymbol{x}) \equiv \sum_{\boldsymbol{s}} P(\boldsymbol{s}, t) \prod_\sigma x_\sigma^{s_\sigma} \,, \tag{6}$$

where the sum runs over all potential values of the vector $\boldsymbol{s}$ and each x_σ is a dummy variable used to track each quantity s_σ.

If we unravel the vector $\boldsymbol{s}$, the PGF is written as

$$\Psi^t(\boldsymbol{z}, \boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{v}, \boldsymbol{u}) = \sum_{\boldsymbol{\ell}} \sum_{\boldsymbol{i}_1} \sum_{\boldsymbol{i}_2} \sum_{\boldsymbol{j}_1} \sum_{\boldsymbol{j}_2} \sum_{\boldsymbol{e}} \sum_{\boldsymbol{d}} P(\boldsymbol{\ell}, \boldsymbol{i}_1, \boldsymbol{i}_2, \boldsymbol{j}_1, \boldsymbol{j}_2, \boldsymbol{e}, \boldsymbol{d}, t) \prod_{\alpha,\alpha'} z_\alpha^{\ell_\alpha} y_{\alpha,1}^{i_{\alpha,1}} y_{\alpha,2}^{i_{\alpha,2}} w_{\alpha,1}^{j_{\alpha,1}} w_{\alpha,2}^{j_{\alpha,2}} v_\alpha^{e_\alpha} u_{\alpha,\alpha'}^{d_{\alpha,\alpha'}} \,, \tag{7}$$

where the vector of variables $(\boldsymbol{z}, \boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{v}, \boldsymbol{u})$ tracks the quantities $(\boldsymbol{\ell}, \boldsymbol{i}_1, \boldsymbol{i}_2, \boldsymbol{j}_1, \boldsymbol{j}_2, \boldsymbol{e}, \boldsymbol{d})$. For obvious reasons, we will favor the compact representation as much as possible, and use the unraveled vectors only when necessary to specify operations on a subset of the variables.

For a general multitype branching process, the solution is obtained by recursion [18]

$$\Psi^{t+1}(\boldsymbol{x}) = \Psi^t \left[\boldsymbol{F}(\boldsymbol{x}) \right] \,, \tag{8}$$

Table S1: Offspring PGFs associated to a single step of the reaction phase for each element of the state vector.

Element identified by the type σ	Reaction phase PGF $r_\sigma(\boldsymbol{x})$
Latent individuals of category α	$y_{\alpha,1}\eta\Delta t + z_\alpha(1-\eta\Delta t)$
Infectious individuals of category α in the first stage	$\exp\left[\sum_{\alpha'}\beta_{\alpha,\alpha'}(z_{\alpha'}-1)\Delta t\right]\left[y_{\alpha,2}\mu\Delta t + y_{\alpha,1}(1-\mu\Delta t)\right]$
Infectious individuals of category α in the second stage	$\exp\left[\sum_{\alpha'}\beta_{\alpha,\alpha'}(z_{\alpha'}-1)\Delta t\right]\left[w_{\alpha,1}\mu\Delta t + y_{\alpha,2}(1-\mu\Delta t)\right]$
Post-infectious individuals of category α in the first stage	$w_{\alpha,2}\varphi\Delta t + w_{\alpha,1}(1-\varphi\Delta t)$
Post-infectious individuals of category α in the second stage	$\varphi\Delta t + w_{\alpha,2}(1-\varphi\Delta t)$
Exported individuals of category α	v_α
Wastewater detections associated with a change of category $\alpha \to \alpha'$	$u_{\alpha,\alpha'}$

where $F(\boldsymbol{x}) = [F_1(\boldsymbol{x}), F_2(\boldsymbol{x}), \dots]$ is a vector of PGFs, and each PGF $F_\sigma(\boldsymbol{x})$ characterizes the *offspring* distribution at the next time step for each node of type σ. For instance, assuming σ identifies infectious α-agents in the first stage (I_1), each individual could lead to a certain number of new latent individuals (through transmission), an α-agent in the second stage (I_2) through disease progression, detection at airports through travel, and so forth. Each possible combination of offspring is encoded in a multivariate PGF, $F_\sigma(\boldsymbol{x})$, similar to Eq. (6).

We now decompose the vector of offspring PGFs $\boldsymbol{F}(\boldsymbol{x})$ by describing separately the reaction phase—modeling disease transmission and progression—, and the mobility phase of GLEAM in terms of vectors of PGFs $\boldsymbol{R}(\boldsymbol{x})$ and $\boldsymbol{M}(\boldsymbol{x})$ respectively. From a similar argument justifying Eq. (8), the vector of offspring PGFs is the composition of the PGFs for each phase [18], i.e., $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{M}(\boldsymbol{R}(\boldsymbol{x}))$. The order of the composition indicates that air travel is happening *before* the reaction phase, in line with the convention in GLEAM.

Reaction phase: disease transmission and progression

While the PGF framework Eq. (8) is defined at a daily resolution, we model the reaction phase at a finer temporal resolution by dividing each day into $m = 12$ time periods of duration $\Delta t = 1/m$, as in the GLEAM simulation procedure. This provides a more realistic description of the intraday dynamics, especially for rapidly evolving epidemics, but one could work at any temporal resolution without significantly affecting the results. Consequently, the vector of PGFs $\boldsymbol{R}(\boldsymbol{x})$ for the reaction phase itself corresponds to the m-th composition of the vector of PGFs $\boldsymbol{r}(\boldsymbol{x})$, i.e.,

$$\boldsymbol{R}(\boldsymbol{x}) = \underbrace{\boldsymbol{r}(\boldsymbol{r}(\dots\boldsymbol{r}(}_{m \text{ times}}\boldsymbol{x})\dots)) \,. \tag{9}$$

Each transition in the disease progression has a rate: latent individuals become infectious at rate η, infectious individuals in the first stage transition to the second stage (and then in the post-infectious stage) at rate μ, post-infectious individuals in the first stage transition to the second stage (and then in the removed stage) at rate φ. Also, infectious α-agents (at any stage) interact and transmit the disease to α'-susceptible individuals at rate $\beta_{\alpha,\alpha'}$, which is constructed from the age-structured contact matrix $\mathbf{C}$ and take into account commuting.

Considering all these possible transitions, we report in Table S1 the offspring PGF associated to a single step of the reaction phase, $r_\sigma(\boldsymbol{x})$, for each type of node. Disease progression transitions are represented by multinomial PGFs, while disease transmission (for infectious individuals only) is represented by a multivariate Poisson PGF. The multivariate Poisson PGF corresponds to an infinite-size subpopulation approximation for binomial draws used in GLEAM simulations for the infection process. Note that the offspring PGFs for elements of type σ identifying a cumulative number of exported individuals or a cumulative number of wastewater detections at airports are simply the identity PGF, i.e., $f(x) = x$. Indeed, these quantities do not generate new infections or transition to other states, they only serve to keep track of a sum [18].

Table S2: Offspring PGFs associated with the mobility phase for each element of the state vector.

Element identified by the type σ	Mobility phase PGF $M_\sigma(\boldsymbol{x})$
Latent individuals of category α	$\sum_{\alpha' \neq \alpha} m_{\alpha,\alpha'} z_{\alpha'} v_{\alpha'} + m_{\alpha,\alpha} z_\alpha$
Infectious individuals of category α in the first stage	$\sum_{\alpha' \neq \alpha} m_{\alpha,\alpha'} y_{\alpha',1} v_{\alpha'} (1 - p_{\alpha,\alpha'} + p_{\alpha,\alpha'} u_{\alpha,\alpha'}) + m_{\alpha,\alpha} y_{\alpha,1}$
Infectious individuals of category α in the second stage	$\sum_{\alpha' \neq \alpha} m_{\alpha,\alpha'} y_{\alpha',2} v_{\alpha'} (1 - p_{\alpha,\alpha'} + p_{\alpha,\alpha'} u_{\alpha,\alpha'}) + m_{\alpha,\alpha} y_{\alpha,2}$
Post-infectious individuals of category α in the first stage	$\sum_{\alpha' \neq \alpha} m_{\alpha,\alpha'} w_{\alpha',1} (1 - p_{\alpha,\alpha'} + p_{\alpha,\alpha'} u_{\alpha,\alpha'}) + m_{\alpha,\alpha} w_{\alpha,1}$
Post-infectious individuals of category α in the second stage	$\sum_{\alpha' \neq \alpha} m_{\alpha,\alpha'} w_{\alpha',2} (1 - p_{\alpha,\alpha'} + p_{\alpha,\alpha'} u_{\alpha,\alpha'}) + m_{\alpha,\alpha} w_{\alpha,2}$
Exported individuals of category α	v_α
Wastewater detections associated with a change of category $\alpha \to \alpha'$	$u_{\alpha,\alpha'}$

Mobility phase: air-travel and detection at airports

Each day, latent, infectious, and post-infectious α-agents in location l move to a new subpopulation l' of multi-index α' with probability $m_{\alpha,\alpha'}$ and stay with probability $m_{\alpha,\alpha}$; agents that move and are detectable (infectious or post-infectious) are detected with probability $p_{\alpha,\alpha'} = p_{l,l'}$, calculated in Eq. (4). Similarly to the reaction phase, we report in Table S2 the offspring PGF associated with the mobility phase, $M_\sigma(\boldsymbol{x})$, for each type of node. Except for the types identifying exported individuals and wastewater detections, $M_\sigma(\boldsymbol{x})$ is a multinomial PGF.

Initial conditions

The solution of Eq. (8) takes the form

$$\Psi^t(\boldsymbol{x}) = \Psi^0(\underbrace{\boldsymbol{F}(\boldsymbol{F}(\cdots \boldsymbol{F}(}_{t \text{ times}} \boldsymbol{x}) \cdots))) \, . \tag{10}$$

The PGF $\Psi^0(\boldsymbol{x})$ specifies the initial condition. In all cases in this study, we specify a number ℓ and i_1 of initial latent and infectious (I_1) individuals respectively. Their age is randomly selected according to the age distribution of the origin subpopulation l, which we represent by the conditional category distribution $P(\alpha|l)$. The PGF for the initial conditions then takes the form

$$\Psi^0(\boldsymbol{x}) = \left(\sum_\alpha P(\alpha|l) z_\alpha \right)^\ell \left(\sum_\alpha P(\alpha|l) y_{\alpha,1} \right)^{i_1} , \tag{11}$$

It is worth stressing that since Ψ^0 is applied *last* when evaluating Eq. (10), it is computationally inexpensive to test various initial conditions, which we use in this study to vary the origin of the outbreak.

Evaluation of probability distributions and cumulants

For a large number of multi-indices σ, evaluating the full joint distribution $P(\boldsymbol{s}, t)$ is computationally prohibitive. However, it is not our goal: In general, we want to evaluate a marginal or a joint distribution of some *observables*, like d_t, the total number of detections at any sentinels by time t, which can be estimated from the PGF $\Psi^t(\boldsymbol{x})$. For instance, the PGF for the distribution $P(d_t = d)$, is

$$\zeta^t(x) \equiv \sum_d P(d_t = d) x^d = \Psi^t \left[\boldsymbol{A}(x) \right] \, , \tag{12}$$

where $\boldsymbol{A}(x)$ is a vector where $A_\sigma(x) = x$ if σ identifies any of the cumulative detection at a sentinel for a specific pair of category, $d_{\alpha,\alpha'}$, and $A_\sigma(x) = 1$ otherwise. This allows us to implicitly sum over all possible combinations of $d_{\alpha,\alpha'}$ resulting in a total number of detection d_t. PGFs $\zeta^t(x, y)$ for joint distributions of observables are constructed in similar fashion.

For any observable n_t, we can extract the distribution $P(n_t)$ from its PGF $\zeta^t(x) = \sum_n P(n_t = n)x^n$ using the following identity

$$P(n_t = n) = \frac{1}{n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\zeta^t(x)\Big|_{x=0} , \\ = \frac{1}{2\pi\rho^n}\int_0^{2\pi} \zeta^t\left(\rho e^{i\omega}\right) e^{-i\omega n}\mathrm{d}\omega , \tag{13}$$

where $0 < \rho < 1$ is a free control parameter. In practice, we use the following discrete Fourier transform approximation (efficiently calculated using Fast Fourier Transform algorithms)

$$P(n_t = n) \approx \frac{1}{k\rho^n}\sum_{j=0}^{k-1} \zeta^t\left(e^{2\pi ij/k}\right) e^{-2\pi ijn/k} . \tag{14}$$

Note that it is possible to bound and control the error committed by choosing k and a suitable ρ value [26], making this approximation exact for all practical purposes. An analogous multidimensional discrete Fourier transform numerical solution is used for distributions of joint observables.

Finally, in some cases, it is more convenient and computationally efficient to extract a few cumulants—mean, variance, etc.—instead of the full distribution. Fortunately, the *cumulant generating function* (CGF) $K^t(x)$ for an observable is directly related to its PGF $\zeta^t(x)$ by the relation $K^t(x) = \ln \zeta^t(e^x)$ [27]. Cumulants are extracted from the CGF using a discrete Fourier transform procedure similar to Eq. (14), which is also directly generalizable to higher dimensions (joint cumulants).

2 Time to first detection for a global wastewater surveillance system at airports

An important quantity introduced in the main text is the time to first detection, t_{fd}, with distribution

$$P(t_{\mathrm{fd}} = t) = P(d_{t-1} < 1, d_t \geq 1) . \tag{15}$$

We can simplify the joint distribution on the right-hand side. Note that the statements $A \equiv d_{t-1} < 1$ and $B \equiv d_t \geq 1$ are Boolean random variables. We can write the probability $P(A, B) = P(A) - P(A, \neg B)$. Since $P(\neg A, \neg B) = 0$ because it is impossible, then $P(A, \neg B) = P(\neg B)$. Therefore,

$$P(d_{t-1} < 1, d_t \geq 1) \equiv P(A) - P(\neg B) = P(d_{t-1} < 1) - P(d_t < 1) = P(d_{t-1} = 0) - P(d_t = 0) , \tag{16}$$

which is straightforward to evaluate using the PGFs. Let us emphasize that the time to first detection is independent of the wastewater methodology being used (pooled sampling or individual aircraft sampling), in the sense that it does not matter if more than one detectable individual contributed to a positive test at a sentinel. It is also worth highlighting the close connection with the extensive line of work on the *arrival time* of a disease in metapopulation networks [28–32].

First, let us validate the PGF solutions with GLEAM simulations. We use the same baseline sentinel system as in the main text. In Fig. S4, we compare the distributions we obtain for t_{fd} using different epidemic origins. In Fig. S4(A), the PGF prediction accurately reproduces the distribution, with T_{fd}, the mean time to first detection, being within 1 standard error of the simulation estimate. In Fig. S4(B), there are approximately 5 days of difference between the two estimates of T_{fd} (i.e., 6% difference), which is due to finite subpopulation effects not taken into account by the PGFs. Note that Boende is one of the subpopulations taking the most time to detect with this surveillance system and thus represents an example of worst-case scenario.

An important surveillance network introduced in the main text is the idealized WWSN, where all airports act as surveillance sites. In Fig S5, we show T_{fd} from every potential origin in the world for the idealized WWSN. We note

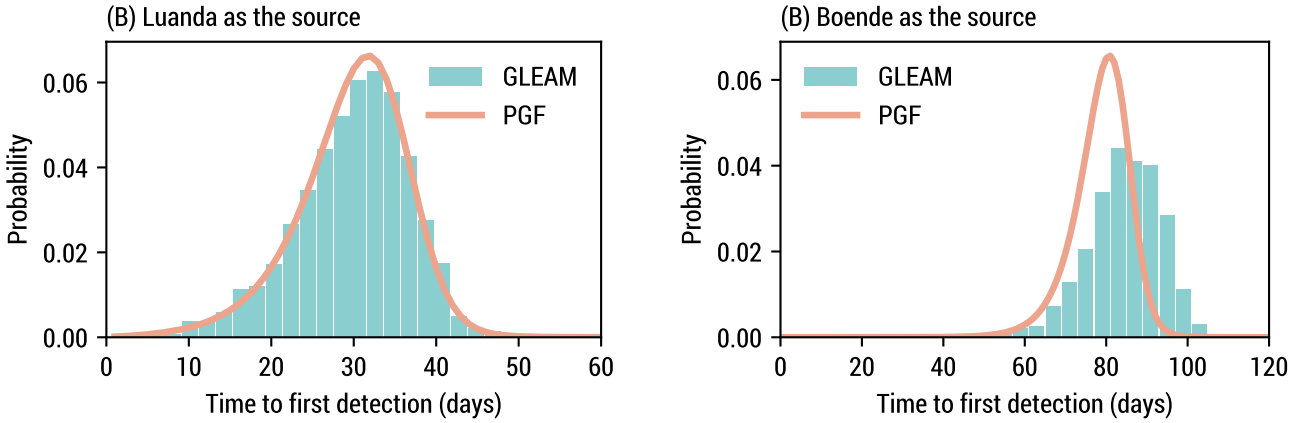

Figure S4: Comparison of the time to first detection. We use the same model, parametrization, and basline WWSN as in Fig. 1, but we use the 2019 summer air-travel network. (A) The epidemic starts in Luanda. GLEAM estimates are based on 2360 simulation runs. The mean time to first detection is 29.84 (SE, 0.14) days for GLEAM and 29.79 days using the PGFs. (B) The epidemic starts in Boende. GLEAM estimates are based on 1050 simulation runs. The mean time to first detection is 83.85 (SE, 0.29) days for GLEAM and 78.72 days using the PGFs.

the similarity with the global map obtained in Figs. 1. The mean time to first detection is still very heterogeneous, with some locations taking a week and others as much as 100 days before a first detection by the WWSN. To better explain the source of this heterogeneity, in Fig. S6, we show the relation between the probability per day of traveling to an international destination for an individual in each location and T_{fd}. There is a clear negative correlation between the two—a higher probability lead to a faster detection—, but other factors are important. For instance, for some subpopulations, international travel is not directly possible—an individual must first travel to another subpopulation within the same country before moving abroad. In other words, one must take into account additional seeding events through importations or commuting to fully characterize the time to first detection.

2.1 Variation with the disease natural history

Changing the disease's natural history affects significantly T_{fd}, as can be seen in Fig. 3 in the main text. For the sake of completeness, we also show in Fig. S7 a similar analysis, where instead of varying generation time via the latent period, we do it via the infectious period, keeping the detectable period fixed. We obtain similar results for the mean time to first detection as in Fig. 3.

By changing the generation time or the reproduction number, we ultimately change the growth rate λ of the epidemic—or equivalently the doubling time $T_2 \equiv \ln 2/\lambda$. For the model we consider, the growth rate can be obtained by solving the following implicit nonlinear equation [11]:

$$\mathcal{R}_0 = \lambda T_{\mathrm{inf}} \frac{\left(1+\frac{\lambda T_{\mathrm{lat}}}{n_{\mathrm{lat}}}\right)^{n_{\mathrm{lat}}}}{\left[1-\left(1+\frac{\lambda T_{\mathrm{inf}}}{n_{\mathrm{inf}}}\right)^{-n_{\mathrm{inf}}}\right]}\,, \tag{17}$$

where n_{lat} and n_{inf} are the numbers of latent and infectious states respectively (here we use $n_{\mathrm{lat}} = 1$ and $n_{\mathrm{inf}} = 2$).

In Fig. 3 of the main text, we show that for all practical purposes, the following relationship holds:

$$\frac{T_{\mathrm{fd}}}{T_2} + \log_2 T_2 \approx \text{const.} \tag{18}$$

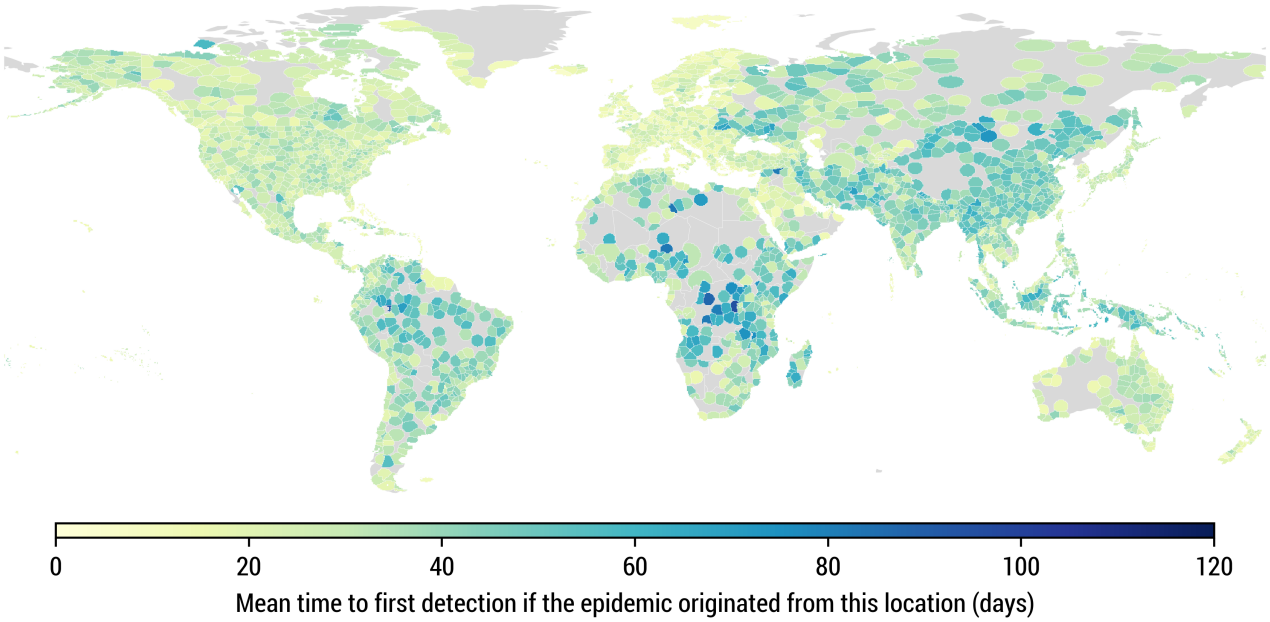

Figure S5: Mean time to first detection with the idealized global surveillance network, with a sentinel at every airport. We use the same model parametrization as in Fig. 1.

Another way to interpret this equation is that changing the doubling time from T_2 to $\hat{T}_2$ ultimately amounts to a linear transformation of the form

$$\hat{T}_{\rm fd} = aT_{\rm fd} + b \quad ; \quad a = \left(\frac{\hat{T}_2}{T_2}\right) \quad ; \quad b = \hat{T}_2 \log_2\left(\frac{T_2}{\hat{T}_2}\right) , \tag{19}$$

where $\hat{T}_{\rm fd}$ is the mean time to first detection with doubling time $\hat{T}_2$. This is illustrated in Fig. S8 for all potential origins, when changing the mean generation time or the reproduction number. The Pearson correlation coefficients are very high (¿0.99).

To justify Eq. (18), let us rephrase more formally the argument introduced in the main text. If we neglect the stochastic fluctuations of the epidemic, the number of detectable individuals is approximately $D(t) \simeq D_0 e^{\lambda t}$, t days after the beginning of the outbreak. Let us also assume that each day, a detectable individual has some constant probability ω to travel and be detected by the WWSN. In that case, we can approximate the probability of having a first detection on day t as [28]

$$P(t_{\rm fd} = t) \approx [1 - (1-\omega)^{D(t)}] \prod_{t'=1}^{t-1} (1-\omega)^{D(t')} \tag{20}$$

$$\approx \xi e^{\lambda t} \exp\left(-\frac{\xi}{\lambda} e^{\lambda t}\right) , \tag{21}$$

where $\xi = \omega D_0$, and the approximations hold when $\xi/\lambda \ll 1$. Since the probability per day of air-travel is very small in general (see Fig. S6), this approximation is almost always valid. We recognize a Gumbel distribution with mean

$$T_{\rm fd} = \frac{1}{\lambda}\left[\ln\left(\frac{\lambda}{\xi}\right) - \gamma\right] , \tag{22}$$

where γ is the Euler–Mascheroni constant. Rearranging the terms and using $\lambda = \ln 2/T_2$, we get

$$\frac{T_{\rm fd}}{T_2} + \log_2 T_2 = \frac{1}{\ln 2}\left(\ln(\ln 2) - \ln \xi - \gamma\right) = \text{const.} , \tag{23}$$

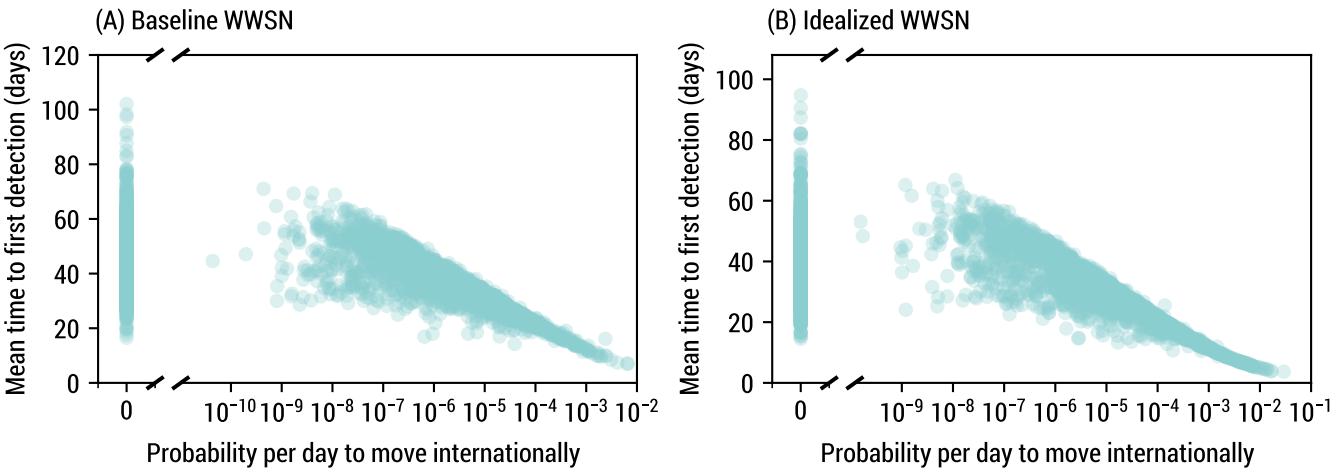

Figure S6: Mean time to first detection considering each of the 3200+ subpopulations as potential origin, against the probability per day to move internationally from each subpopulation. We use the same model parametrization as in Fig. 1. (A) We use the same baseline WWSN as in Fig. 1. The Pearson correlation coefficient is -0.862 (two-sided P value ¡0.001) between the mean time to first detection and the logarithm of the probability per day to move internationally (for subpopulations where this probability is larger than zero). (B) We use the idealized WWSN, with a sentinel at every airport. The Pearson correlation coefficient is -0.907 (two-sided P value ¡0.001) between the mean time to first detection and the logarithm of the probability per day to move internationally (for subpopulations where this probability is larger than zero).

which is identical to Eq. (18).

In Fig. S9, we show that the Gumbel distribution approximates well the distribution for the time to first detection, especially when the initial number of latent and infectious is high in Fig. S9(B). This is due to the fact that the Gumbel approximation neglects the stochastic fluctuations of the epidemic, which are more important when starting with fewer exposed individuals.

To obtain the Gumbel approximations in Fig. S9, we calculate λ and T_{fd} from our PGF framework, which allows us to fix ξ in Eq. (22). While ξ should not depend on λ, we do find small variations, which also explain the small variations of the approximate invariant quantity in Eq. (18) and in Fig. 3. This is because we treated the detectable individuals $D(t)$ as one homogeneous population with a fixed probability of moving and being detected ω, while in fact it is a heterogeneous group of individuals distributed across a complex metapopulation network.

2.2 Sensitivity analyses

Other than the reproduction number and the generation time that directly affect the growth rate, other aspects of the disease's natural history impact the time to first detection, but less significantly so. Here we vary the length of the post-infectious period—thereby changing the detectable period—and the shape of the secondary-infection distribution. In our framework, we do not fix directly the secondary-infection distribution, but rather the secondary-infection distribution per time step. To tune the variance, we replace the multivariate Poisson PGF term in the reaction phase (see Table S1) by the composition of a multinomial and a negative binomial PGF, namely

$$\exp\left[\sum_{\alpha'}\beta_{\alpha,\alpha'}(z_{\alpha'}-1)\Delta t\right] \mapsto \left[1+\frac{\Delta t}{\kappa}\sum_{\alpha'}\beta_{\alpha,\alpha'}(1-z_{\alpha'})\right]^{-\kappa}. \tag{24}$$

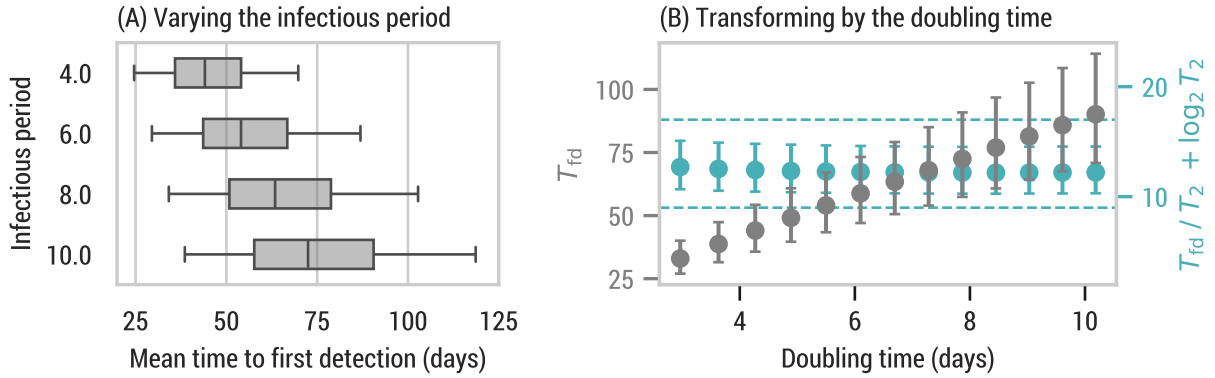

Figure S7: Varying the generation time via the infectious period instead of the latent period. We keep the detectable period fixed to 15 days by also changing the post-infectious detectable period as $T_{\text{post}} = T_{\text{det}} - T_{\text{inf}}$. Otherwise, we use the same WWSN and parametrization as in Fig. 1.

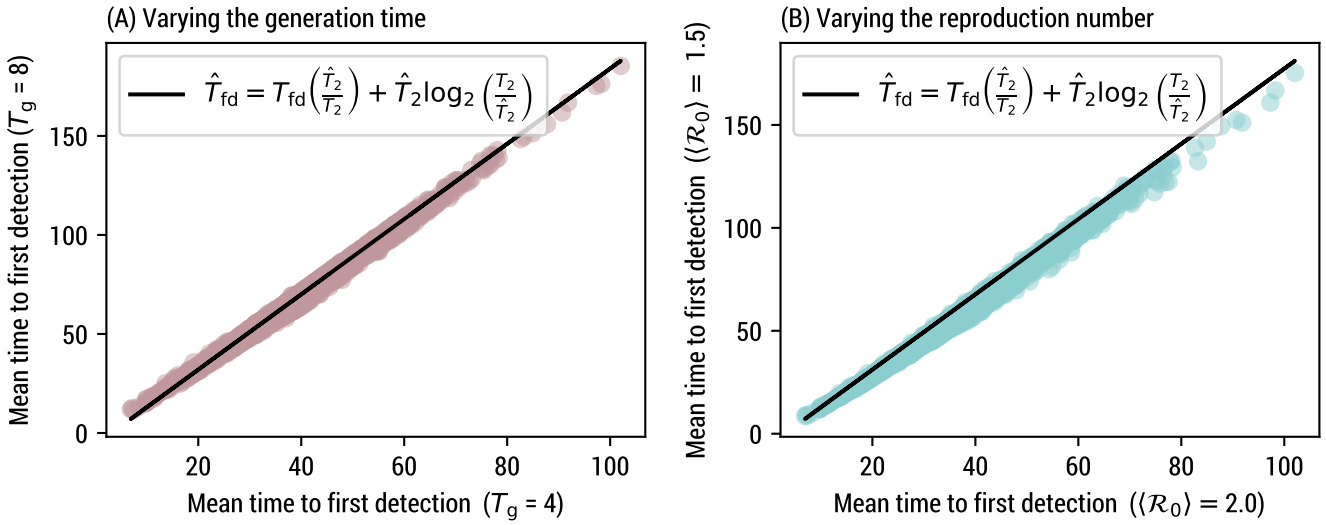

Figure S8: Mean time to first detection considering each of the 3200+ subpopulations as potential origin, for varying natural disease history. We use the same WWSN as in Fig. 1. The detection probability is 16% and the detectable period is 12.7 days. The epidemic starts with 5 infectious and 5 latent individuals. (A) The mean reproduction number is fixed $\langle \mathcal{R}_0 \rangle = 2$, and the mean generation time varies. The Pearson correlation coefficient is 0.997 with a two-sided P value < 0.001. (B) The mean generation time is fixed $T_{\text{gen}} = 4$, and the reproduction number varies. The Pearson correlation coefficient is 0.996 with a two-sided P value < 0.001.

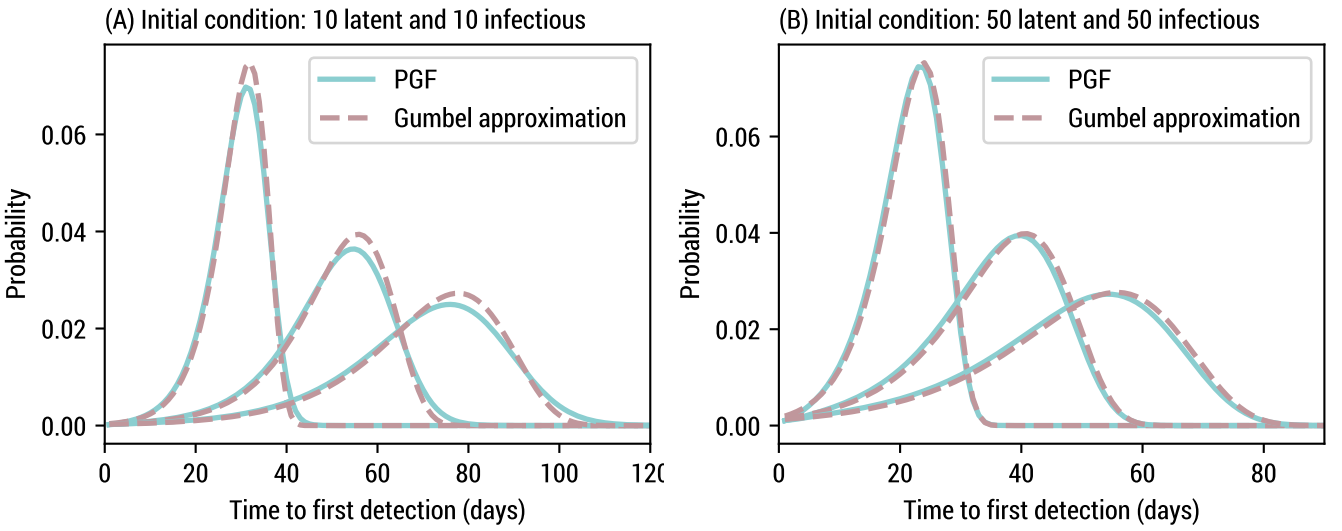

Figure S9: Approximation of the time to first detection using a Gumbel distribution. We use the same WWSN as in Fig. 1 and focus on an epidemic originating in São Paulo. The detection probability is 16%, the post-infectious detectable period is 12.7 days and the average reproduduction number is $\langle \mathcal{R}_0 \rangle = 2$. We show the results for mean generation times of 4, 8, and 12 days. We consider two initial conditions: (A) 10 latent and 10 infectious individuals and (B) 50 latent and 50 infectious individuals.

Table S3: Statistics of the mean time to first detection for various post-infectious periods and overdispersion parameters for the secondary-infection distribution (per time step). The rest of the model parameters and the WWSN are the same as in Fig. 1. We report the median and the 5th and 95th percentile in parentheses.

Overdispersion κ		0.01	0.03	0.1	∞
Infections caused by the top 20%		81.9%	66.2%	56.8%	51.3%
Post-Infec. period	5	39.6 (22.4,62.1)	38.3 (21.4,60.7)	37.9 (21.1,60.1)	37.7 (21.0,59.9)
	10	38.9 (21.7,61.4)	37.7 (20.8,60.0)	37.2 (20.4,59.4)	37.0 (20.3,59.1)
	20	38.5 (21.3,60.9)	37.2 (20.4,59.5)	36.8 (20.1,58.9)	36.6 (20.0,58.7)

Increasing the overdispersion parameter κ reduces the variance while reducing κ increases the variance[1]; in the limit $\kappa \to \infty$, we recover the multivariate Poisson PGF.

Another way to communicate the overdispersion in epidemiology is to assess what portion of infections are caused by the top 20% of infectors [33]. We compile our results in Table S3 for various combinations of post-infectious period and overdispersion parameters. Overall, both have a limited impact on the distribution of T_{fd} for all subpopulations. A broader secondary-infection distribution will mainly affect the beginning of an epidemic—once the outbreak is large enough, the stochastic fluctuations are averaged out. Similar results were obtained in Ref. [6]. The small impact of changing the detectable period (through the post-infectious period) is due to the fact that the number of infectious grows exponentially at the beginning of an epidemic. Consequently, unless the growth rate of the disease is very small, detection at airports should be predominantly caused by newly infectious individuals traveling.

Variation of the air-travel patterns also do not lead to large variations of the global statistics of T_{fd} for all locations, as illustrated in Fig. S10. The median of the distribution for T_{fd} varies from 36 days in the summer to 38 days in the

[1] It is important to distinguish the overdispersion parameter κ given here, associated to the secondary-infection distribution per time step of duration Δt, versus what one would expect using a negative binomial for the secondary-infection distribution over the whole infectious period.

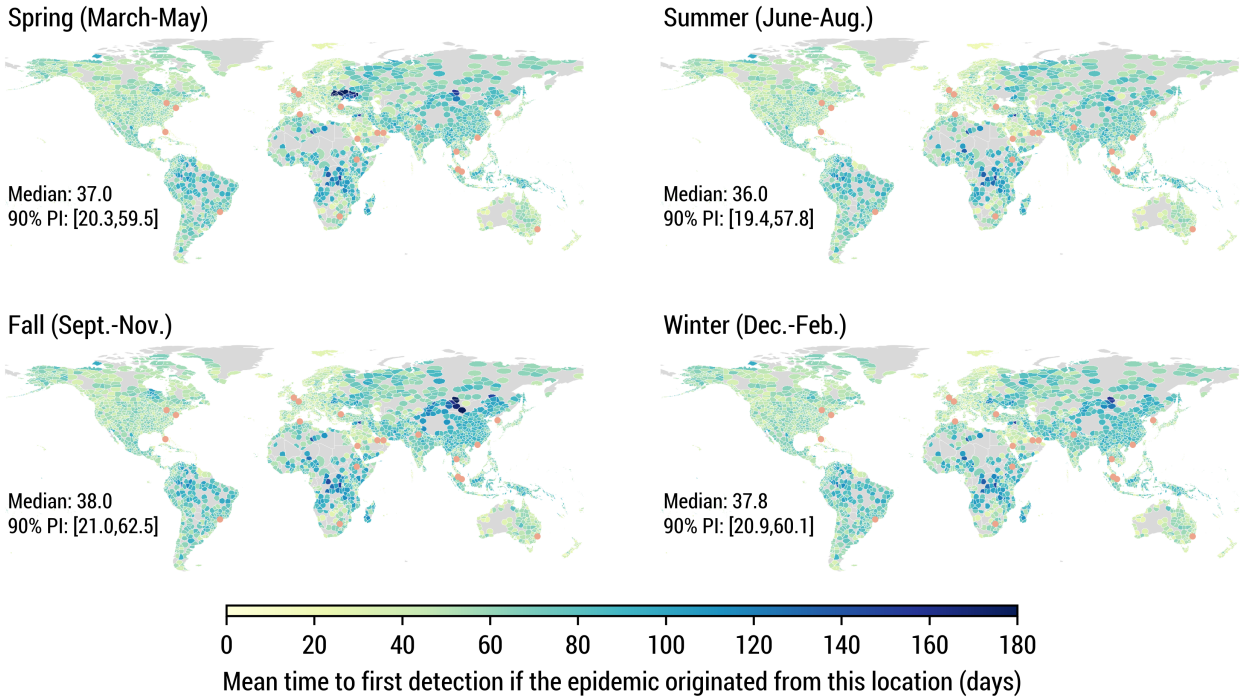

Figure S10: Mean time to first detection using different air-travel mobility networks associated with different seasons. We use the same WWSN and model parametrization as in Fig. 1.

Fall, and the 90% prediction intervals remain relatively stable.

2.3 Optimization of the time to first detection

In the Methods section of the main text, we introduce a greedy optimization scheme to minimize the mean time to first detection, averaged over all subpopulations as a potential source

$$\Phi(\mathcal{S}) = \sum_{l} P(l) T_{\mathrm{fd}}(\mathcal{S}, l) \,, \tag{25}$$

with probability $P(l)$ of being the source. In the second step of the optimization procedure, it requires the evaluation of $\Phi(\mathcal{S} \cup \{\nu\})$ for each airport $\nu \notin \mathcal{S}$ to assess the reduction in time to first detection associated with adding ν as a sentinel. However, evaluating $\Phi(\mathcal{S} \cup \{\nu\})$ for each new potential sentinel is computationally expensive.

Instead, we use an approximation for each $T_{\mathrm{fd}}(\mathcal{S}, l)$ hinging on two assumptions. First, we assume the full distribution of the time to first detection from each source subpopulation l is well-described by a Gumbel distribution—which is validated in Fig. S9. Second, we assume that the detection at all sentinels $\nu \in \mathcal{S}$ are independent processes. From these assumptions, we can leverage the following identity [28]

$$\exp\left(-\lambda T_{\mathrm{fd}}(\mathcal{S}, l)\right) = \sum_{\nu \in \mathcal{S}} \exp\left(-\lambda T_{\mathrm{fd}}(\{\nu\}, l)\right) \,. \tag{26}$$

This allows us to efficiently estimate $T_{\mathrm{fd}}(\mathcal{S}, l)$ from the individual $T_{\mathrm{fd}}(\{\nu\}, l)$ for all $\nu \in \mathcal{S}$. Figure S11 validates the accuracy of the approximation.

Submodularity proof

In the context of the approximation provided by Eq. (26), we want to show that $-\Phi(\mathcal{S})$ is a monotone submodular set function. Since $-\Phi(\mathcal{S})$ is a positive linear combination of $-T_{\mathrm{fd}}(\mathcal{S}, l)$ for each potential source l, it suffices to show

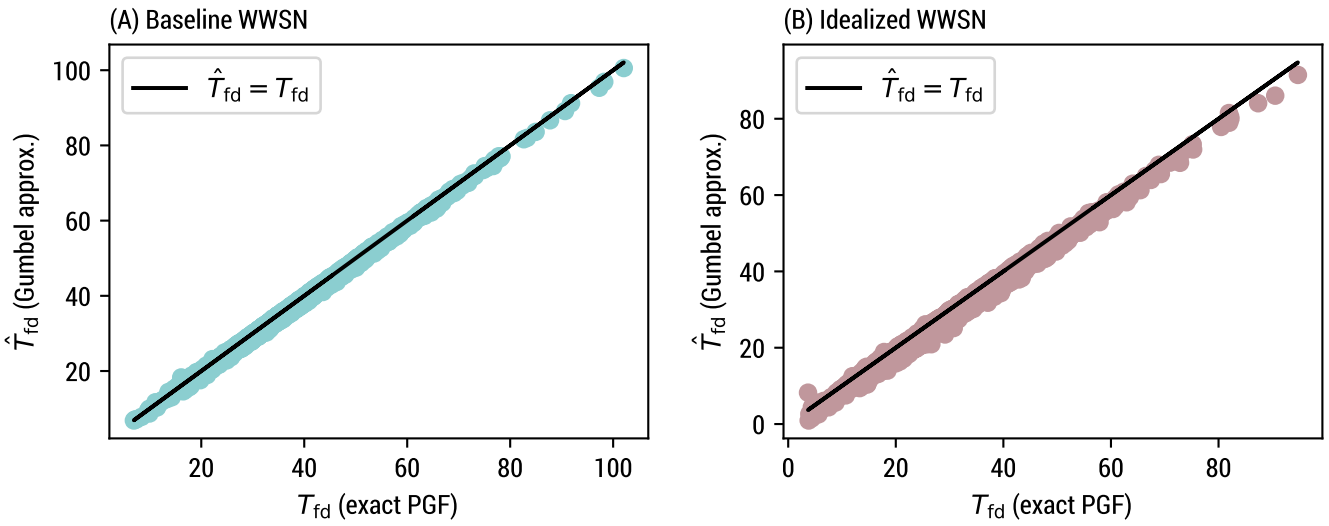

Figure S11: Mean time to first detection considering each of the 3200+ subpopulations as potential origin, using exact PGF calculation and the Gumbel approximation provided by Eq. (26). We use the same model parametrization as in Fig. 1. (A) We use the baseline WWSN. The Pearson correlation coefficient is > 0.999 with a two-sided P value < 0.001. (B) We use the idealized WWSN. The Pearson correlation coefficient is 0.997 with a two-sided P value < 0.001.

that $-T_{\mathrm{fd}}(\mathcal{S}, l)$ is a monotone submodular set function for all l [34].

We define the set of all potential sentinels as Ω. Let us also simplify the notation by writing $\tau(B) \equiv T_{\mathrm{fd}}(B, l)$ and $\tau(b) \equiv T_{\mathrm{fd}}(\{b\}, l)$ for the mean time to first detection for every set of sentinels $B \subseteq \Omega$ and every sentinel b respectively[2]. Note that $\tau : 2^{\Omega} \to \mathbb{R}^{+}$ is a set function, where 2^{Ω} is the set of all subsets of Ω

Lemma 1. *τ and $-\tau$ are monotone functions, i.e., $\tau(B) \leq \tau(A) \iff -\tau(B) \geq -\tau(A)$ for $A \subseteq B$.*

Proof. Let us define $C \equiv A \cap B$. From Eq. (26), we have

$$e^{-\lambda\tau(B)} = e^{-\lambda\tau(A)} + e^{-\lambda\tau(C)} \geq e^{-\lambda\tau(A)} , \tag{27}$$

where $\lambda > 0$. Taking the logarithm on both sides and dividing by λ, this implies $-\tau(B) \geq -\tau(A)$. □

Definition 1. *Let Ω be a finite set and $f : 2^{\Omega} \to \mathbb{R}$ a set function. The function f is submodular if for two subsets $A, B \subseteq \Omega$ with $A \subseteq B$ and $e \in \Omega \setminus B$, we have $f(A \cup \{e\}) - f(A) \geq f(B \cup \{e\}) - f(B)$ [34].*

Theorem 1. *The function $-\tau$ is monotone and submodular.*

Proof. The function $-\tau$ is monotone from lemma 1. Let us consider two sets A and B such that $A \subseteq B \subseteq \Omega$, $C \equiv B \setminus A$, and an element $e \in \Omega \setminus B$. Using Eq. (26), we have

$$\begin{aligned} e^{-\lambda\tau(B\cup\{e\})} &= e^{-\lambda\tau(B)} + e^{-\lambda\tau(e)} , \\ &= e^{-\lambda\tau(A)} + e^{-\lambda\tau(C)} + e^{-\lambda\tau(e)} , \end{aligned}$$

where $\lambda > 0$. Dividing both sides by $e^{-\lambda\tau(B)}$ and applying again Eq. (26), we find

$$e^{-\lambda[\tau(B\cup\{e\})-\tau(B)]} = \frac{e^{-\lambda\tau(A\cup\{e\})} + e^{-\lambda\tau(C)}}{e^{-\lambda\tau(A)} + e^{-\lambda\tau(C)}} \leq e^{-\lambda[\tau(A\cup\{e\})-\tau(A)]} , \tag{28}$$

[2]The ensuing results are general for all epidemic source l and therefore we drop the index to simplify the notation.

where the inequality holds if and only if $\tau(A \cup \{e\}) \leq \tau(A)$, which is true from lemma 1, but also more intuitively because adding sentinels can only reduce the time to first detection. Taking the logarithm on both sides of Eq. (28) and dividing by λ, we arrive at

$$-\tau(B \cup \{e\}) + \tau(B) \leq -\tau(A \cup \{e\}) + \tau(A) \, . \tag{29}$$

From definition 1, this implies that $-\tau$ is a submodular set function. □

3 Retrospective counterfactual scenarios

3.1 SARS-CoV-2 Alpha variant emergence

In the main text, we present the results of a counterfactual scenario where we would have had a global WWSN to track the international dissemination of the SARS-CoV-2 Alpha (B.1.1.7) variant. We utilize air-travel data from September 2020 to November 2020, along with the baseline WWSN of 20 sentinels (see Table S4).

This variant was first identified by health authorities in the United Kingdom and retrospective analyses trace the first identified case in Kent, South East England, on September 20, 2020 [35]. According to data published by the United Kingdom government [36], the effective reproduction number for SARS-CoV-2 between September 11 and October 30, 2020, was between 1.1 and 1.4 (90% CI) in both London and South East England. The Alpha variant was found to be more transmissible, with an estimated increased reproduction number ranging from 40 to 100% [35, 37, 38]. In this counterfactual study, we considered a value of $\mathcal{R}_{\text{eff}}^{\text{alpha}} = 1.7$ for the Alpha variant, on the lower side of available estimates, equivalent to having an effective reproduction number of $\mathcal{R}_{\text{eff}}^{\text{ws}} = 1.1$ for the wild strain and an increased transmissibility of 55%. Since approximately 5% of positive cases were sequenced at the source [35], we considered an initial cluster of 20 infectious and 20 latent individuals on September 15. The generation time is kept fixed at 6.5 days, with a latency period of 4.5 days.

Distributions for the time to first detection in Fig. 5A are calculated using the PGF methodology and are not influenced by the wastewater sampling scheme. The geolocalization of the source (Fig. 5B) and the parameter inference (Fig. 5C), however, require us to transform distributions for the number detections to acccount for wastewater pool sampling at sentinel sites.

Distribution for the cumulative number of detections by a WWSN

Our PGF methodology allows us to estimate $P(d_t = d)$, the distribution for the cumulative number of detections at time t by the WWSN. However, as defined, all detectable agents traveling through a screened route could contribute to d_t *independently*. While this can be a reasonable assumption for individual testing, this is not the case for wastewater sampling, where it is not usually feasible to identify precisely the number of detectable individuals in an aircraft—we only get a binary *yes* or *no* answer to whether or not the pathogen was detected. Depending on how the testing is performed—on individual aircraft, or pooled sampling of multiple aircraft at a triturator—we would obtain different results for the cumulative number of detections.

We assume the pooled testing of all aircraft within a day at each sentinel airport. To take into account the potential copresence on the same day of detectable individuals traveling to the same destination, we need to define the probability $P(\tilde{d}|d)$, where $\tilde{d}$ represents the cumulative "wastewater" detections and d the cumulative "individual" detections we would obtain if we were able to detect each individual independently. Then we get the adjusted distribution

$$P(\tilde{d}) = \sum_{d} P(\tilde{d}|d)P(d) \, . \tag{30}$$

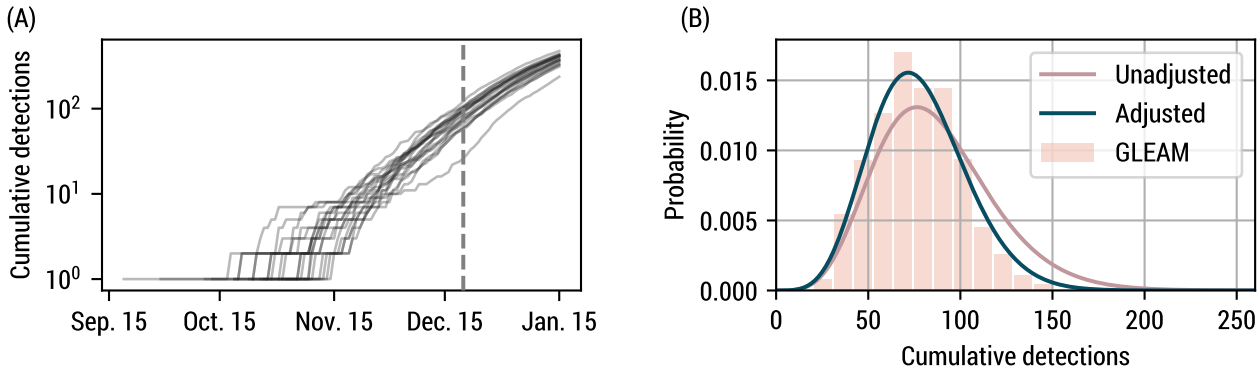

Figure S12: Cumulative number of wastewater detections for the Alpha variant counterfactual study. (A) Example of 20 detection time series from GLEAM simulations. The vertical dashed line indicates December 20, 2020. (B) Distribution for the cumulative number of detections on December 20, 2020, from 1250 GLEAM simulations and the PGF methodology, with and without adjustment for wastewater pool sampling.

We evaluate $P(\tilde{d}|d)$ by first characterizing $q_{\nu,t}$, the relative probability an individual detection happens at sentinel ν and at time t. In general, $q_{\nu,t} \propto \langle d^*_{\nu,t}\rangle$, where $\langle d^*_{\nu,t}\rangle$ is the mean number of individual detections at sentinel ν and *incident* on time t—not the cumulative detections. The ratio $\langle d^*_{\nu,t}\rangle/\langle d^*_{\nu',t}\rangle$ is very stable in the early phase of an outbreak for any pair of airports (ν,ν'). Therefore, $\langle d^*_{\nu,t}\rangle \sim A(\nu)e^{\lambda t}$, where λ is the growth rate at the source and $A(\nu)$ is the relative propensity of detection at each sentinel airports in the early phase. After normalization, we encapsulate all probabilities in a vector $\boldsymbol{q}$. Secondly, we assume that the number of *incident* detections for each sentinel and day $d^*_{\nu,t}$ ($\boldsymbol{d}^*$ in vector format) is distributed according to a multinomial $P(\boldsymbol{d}^*)$ with parameters d (the sum of all $\boldsymbol{d}^*$) and $\boldsymbol{q}$. The total number of *wasterwater* detections corresponds to the number of nonzero counts in $\boldsymbol{d}^*$, and thus $P(\tilde{d}|d)$ is obtained by summing $P(\boldsymbol{d}^*)$ over all configurations such that there are a total of $\tilde{d}$ nonzero counts. While this is technically a difficult combinatorial task, this can be carried out efficiently numerically (see Ref. [39]).

In Fig. S12(A), we show examples of time series for the cumulative number of detections at all sentinels of the WWSN generated by GLEAM. In Fig. S12(B), we illustrate the importance of accounting for wastewater pooled sampling when the number of detections is sufficiently large; the number of wastewater detections is effectively reduced due to the copresence of detectable individuals transiting through as sentinel during the same day.

Geolocalization of the source

Even without information about which flight paths led to detections, it is possible to recover information about the location of the source of an epidemic. Indeed, we can construct a likelihood $P(\tilde{\boldsymbol{d}}|\text{source} = l) \equiv P(\tilde{\boldsymbol{d}}|l)$ for the probability to have observed the wastewater detections $\tilde{\boldsymbol{d}} = (\tilde{d}_\nu)_{\nu\in\mathcal{S}}$ at the sentinel airports.

First, let us ignore wastewater pool sampling and consider individual detections $\boldsymbol{d} = (d_\nu)_{\nu\in\mathcal{S}}$ and the associated likelihood function $P(\boldsymbol{d}|l)$. Then $P(\boldsymbol{d}|l)$ can be modeled by a multinomial distribution with parameter $d = \sum_\nu d_\nu$ and probability vector $\boldsymbol{q} = (q_\nu)_{\nu\in\mathcal{S}}$, where $q_\nu \propto \langle d_\nu\rangle$, the expected number of detections at sentinel ν when we have d detections. In practice, we approximate q_ν by the expected number of detections at each sentinel when we have a first detection.

To account for wastewater pool sampling, we transform the likelihood as

$$P(\tilde{\boldsymbol{d}}|l) = \sum_{\boldsymbol{d}} P(\tilde{\boldsymbol{d}}, \boldsymbol{d}|l) = \sum_{\boldsymbol{d}} P(\tilde{\boldsymbol{d}}|\boldsymbol{d}, l)P(\boldsymbol{d}|l)\ . \tag{31}$$

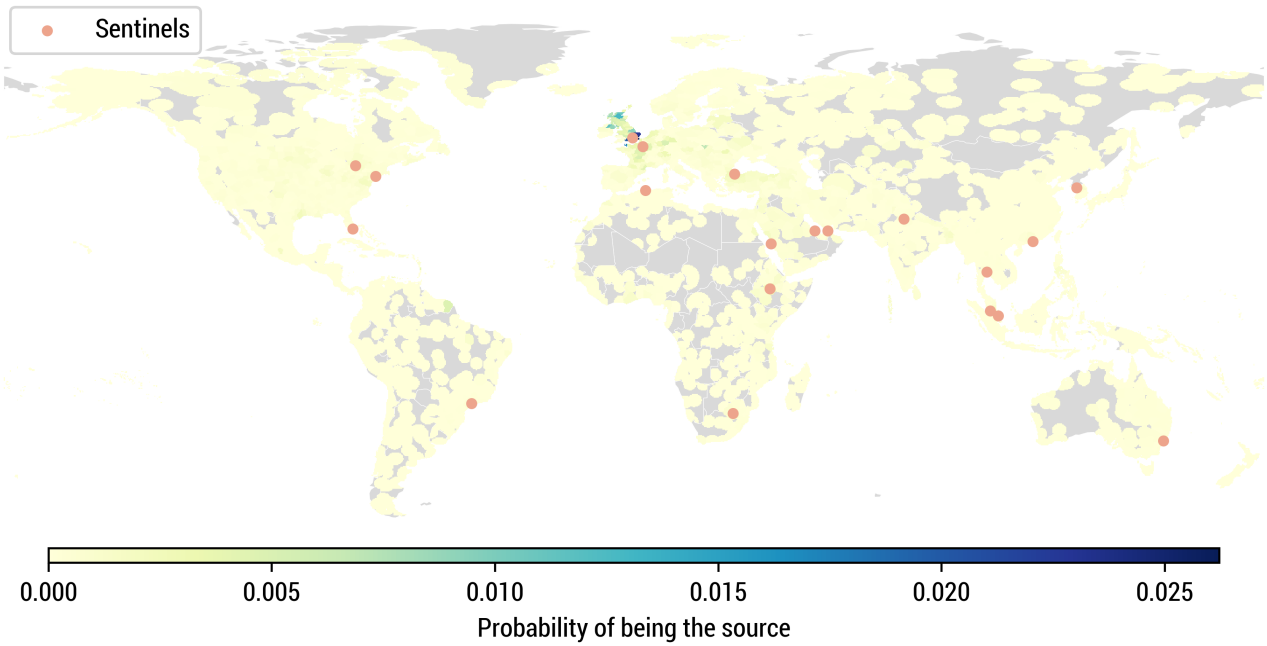

Figure S13: Geolocalization of the source for the SARS-CoV-2 Alpha variant counterfactual scenario. We show the posterior distribution over subpopulations averaged over 1250 GLEAM simulations, when ≥ 10 cumulative wastewater detections have occurred. We use the same model parametrization as in Fig. 5 for the geolocalization.

Assuming that $P(\tilde{\boldsymbol{d}}|\boldsymbol{d}, l)$ is concentrated around its mean, we can use the following approximation

$$P(\tilde{\boldsymbol{d}}|l) \approx \sum_{\boldsymbol{d}} \delta_{\boldsymbol{d},\boldsymbol{b}} P(\boldsymbol{d}|l) = P(\boldsymbol{b}|l) \ , \tag{32}$$

where $\boldsymbol{b}$ is the expected number of "individual detections" given the observed pooled wastewater detections $\tilde{\boldsymbol{d}}$, rounded to the nearest integer.

Using the likelihood in Eq. (32), we compute the following posterior distribution for the source

$$P(l|\tilde{\boldsymbol{d}}) \propto P(\tilde{\boldsymbol{d}}|l)P(l) \ , \tag{33}$$

where $P(l)$ is a prior distribution on the source location. In this work, we consider a uniform prior $P(l) = \text{const}$. In Fig. S13, we show this posterior distribution for at least 10 wastewater detections, averaged over 1250 simulations. We see that most of the posterior density is concentrated in Europe, and especially in the United Kingdom. The same simulations and posterior distributions are used to assess the geolocalization capacities of a WWSN in Fig. 5, as detections accumulates at the sentinels.

Characterization of the growth dynamics

The time series of cumulative detections can also be utilized to estimate key epidemic parameters. For generic parameter inference, we use the following approach. Given the observed cumulative number of wastewater detections $\tilde{d}$ and $\tilde{d}'$ at different time t and t', we calculate the posterior distribution over $\boldsymbol{\theta}$ (the parameters) using

$$P(\boldsymbol{\theta}|\tilde{d}', \tilde{d}) \propto P(\tilde{d}', \tilde{d}|\boldsymbol{\theta})P(\boldsymbol{\theta}) \ . \tag{34}$$

To compute more efficiently the joint likelihood $P(\tilde{d}', \tilde{d}|\boldsymbol{\theta})$, we evaluate the joint cumulants from the CGF and employ the method of moments with a negative multinomial distribution.

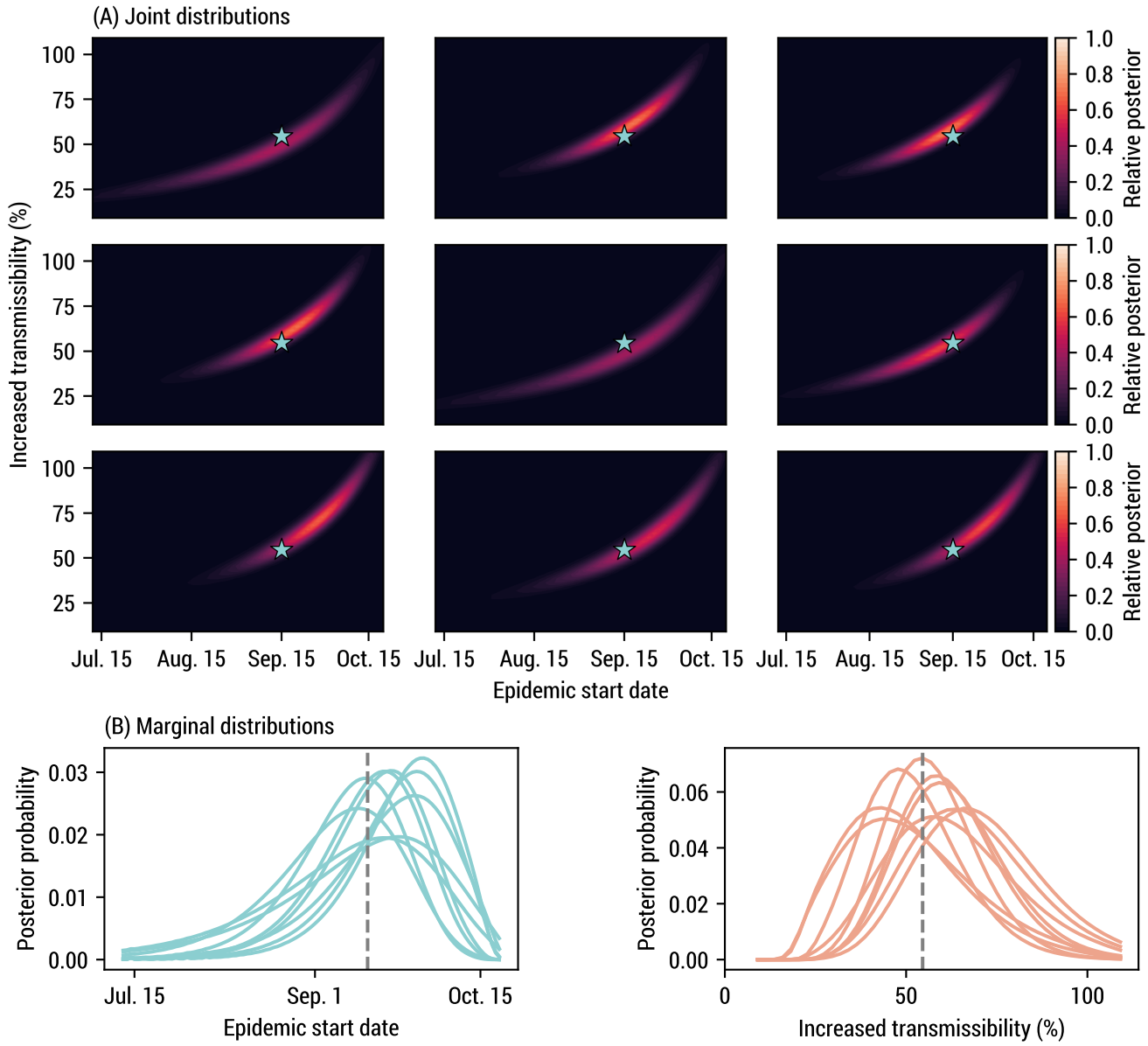

Figure S14: Examples of joint and marginal posterior distributions obtained for a subset of time series. The blue star (A) and vertical dashed line (B) represent the ground truth (55% increased transmissibility and September 15, 2020 as the start date).

In Fig. 5C of the main text, we jointly infer the epidemic start date and the increased transmissibility of the Alpha variant with respect to the SARS-CoV-2 wild strain. We use a flat prior $P(\boldsymbol{\theta}) = \text{const.}$, between 9% and 109% for the increased transmissibility ($\mathcal{R}_{\text{eff}}^{\text{alpha}} \in [1.2, 2.3]$) and between July 12 and October 20, 2020 for the starting date (cluster of 20 latent and 20 infectious). In Fig. S14, we also illustrate a subset of the posterior distributions from individual time series.

3.2 SARS-CoV-2 (wild strain) emergence

To assess the potential effectiveness of a WWSN in early detection and response to a pandemic situation, we analyze a second hypothetical scenario: the operation of a WWSN prior to the onset of the COVID-19 pandemic. This

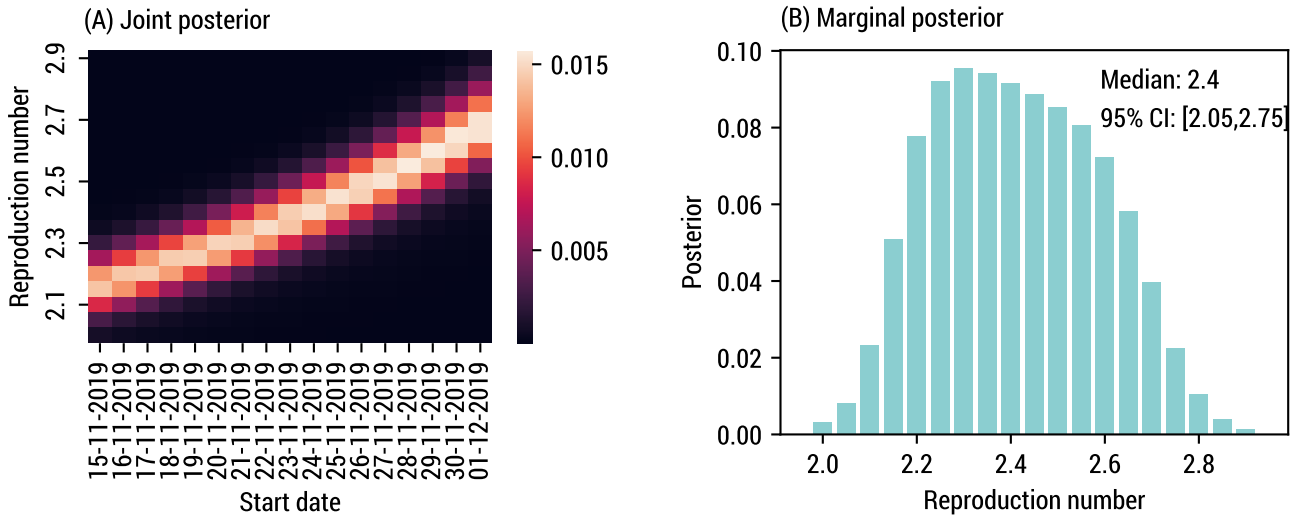

Figure S15: Posterior distributions given 69 international COVID-19 cases with a travel history and an arrival date before or on Jan. 23, 2020.

counterfactual study utilizes air-travel data from December 2018 to February 2019, along with the baseline WWSN of 20 sentinels (see Table S4).

Calibration on importations

Considering the high uncertainty at the beginning of the pandemic, we first calibrate our model to available data. To do so, we use the same approach as in Refs. [5, 6], i.e., we calibrate our model on the number of international importations until January 23rd, 2020, when travel restrictions were imposed.

Only a fraction of importations are identified at the destination, notably because of asymptomatic individuals—we considered a 40% rate of asymptomatic individuals [40]—, but also because of different levels of capacity for detection. To account for the heterogeneity in case detection, we stratify countries into three groups based on the second component of the Global Health Security Index: high ($\geq$80th percentile), low ($\leq$20th percentile), and medium (the rest) surveillance capacity. We then use the estimates provided in Ref. [41] to assign a probability of detection relative to Singapore, which has had strong epidemiological surveillance in past infectious disease outbreaks including the COVID-19 pandemic. We used a 60% probability of detection for symptomatic individuals in Singapore, and then countries in the high, medium, and low categories were assigned relative capacities corresponding to 40%, 37%, and 11% compared to Singapore.

To calibrate our model, we compute the distribution $P(E|\boldsymbol{\theta})$ for the number of importations E given the parameters $\boldsymbol{\theta}$ of the model—in the present case, $\boldsymbol{\theta}$ corresponds to the start date and the basic reproduction number. This likelihood is used to evaluate the posterior distribution $P(\boldsymbol{\theta}|E) \propto P(E|\boldsymbol{\theta})$. We use a uniform prior between November 15th and December 1st, 2019 for the starting date with an initial cluster of 10 infectious and 10 latent individuals in Wuhan, consistent with estimates placing the index case mid-October to mid-November [42]. We also consider a uniform prior distribution on $\mathcal{R}_0$ between 2 and 2.9. The generation time is kept fixed at 6.5 days, with a latency period of 4.5 days.

In Fig. S15(A), we show the posterior distribution obtained for $\mathcal{R}_0$ and the starting date using as evidence 69 identified importations by Jan. 23, 2020 (see Table S1 in the Supplementary Material of Ref. [5]). Summing the joint posterior over the plausible starting date, we obtain the marginal posterior distribution on $\mathcal{R}_0$ in Fig. S15(B), with a

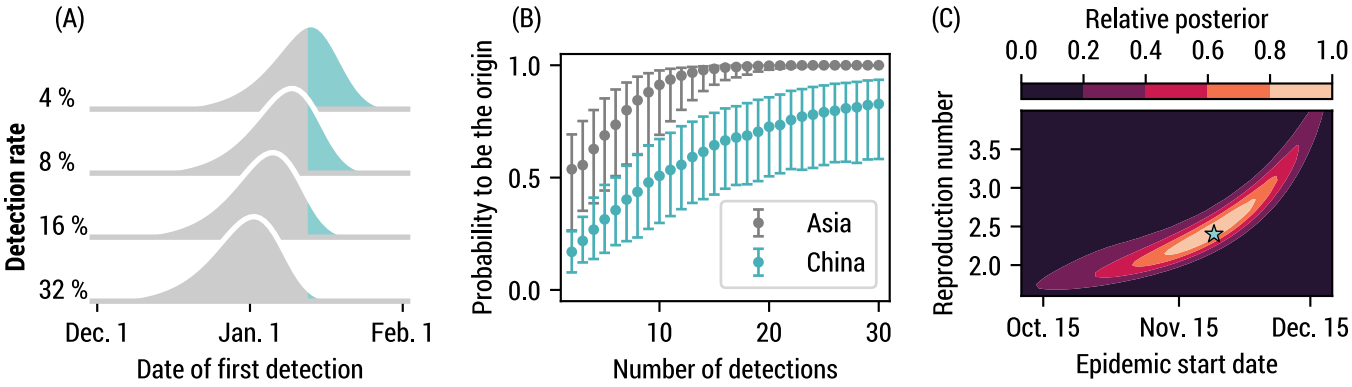

Figure S16: A global WWSN would provide an early warning system for international spreading and timely inferential capabilities. We consider a counterfactual scenario of the emergence of COVID-19 where a global WWSN would have been available. We use the baseline surveillance system consisting of 20 sentinels (see Table S4). We calibrate our model on international importations with an arrival date on or before Jan. 23 2020. (A) Distributions for the time to first detection with varying detection rates. The blue portion corresponds to dates after the first international case was reported in Thailand on January 13 [46]. (B)-(C) Inference experiment using data generated by the mechanistic GLEAM model with $\mathcal{R}_0 = 2.4$ and a start date of November 23 for the initial cluster. (B) Geolocalization of the source as more detections cumulate. We compute the posterior distribution for the origin of the epidemic based on the detection counts at each sentinel. Geolocalization of the source as more detections cumulate. We compute the posterior distribution for the origin of the epidemic based on the detection counts at each sentinel. The markers indicate the median posterior value and the whiskers the interquartile range obtained from 1200 detection time series. (C) Joint posterior distribution on $\mathcal{R}_0$ and the start date, averaged over 59 detection time series. The blue star indicates the ground truth for the simulation experiment.

median of 2.4 (95% CI, 2.05–2.75). The resulting median doubling time is 4.05 days (95% CI, 3.35—5.15), broadly consistent with other estimates [5, 6, 43–45].

Results

In Fig. S16, we present similar results as in Fig. 5, but for the emergence of SARS-CoV-2 in Wuhan, demonstrating the robust early warning and situational awareness capacities of a global WWSN.

We use the joint posterior of Fig. S15(A) to compute the posterior predictive distribution for the time to first detection in Fig. S16(A). We find that even with a low detection rate (4%), there is a 54% probability that an international case outside China would have been detected earlier than the first reported case in Thailand (January 13th), with a median time to first detection on January 12th. With a 16% detection rate—more in line with estimates for a triturator sampling scheme capturing all international inbound flights—the probability to have detection prior to January 13th is 91% and the median time to first detection is on January 4th.

This hypothetical scenario presupposes the capability of detecting SARS-CoV-2 in aircraft wastewater. The key point, however, is that an operational WWSN has the potential to significantly expedite the detection of international pathogen introductions. Any reduction in the time taken to sequence the genome of emerging pathogens would directly correlate to earlier detections. Furthermore, frozen samples can be stored and analyzed in the future.

Similarly to the Alpha variant retrospective study, we use synthetic time series data from GLEAM to illustrate how we can evaluate the source and key epidemic parameters. We fix the basic reproduction number to $\mathcal{R}_0 = 2.4$ and an initial cluster of 10 infectious and 10 latent individuals on November 23, 2019, based on the joint posterior in Fig. S15(A). We obtain the posterior distribution over subpopulations as a function of the number of wastewater

detections and aggregate the posterior distribution at the continent and country level in Fig. S16(C). We find reliable geolocalization capacities at the country level after about 10 detections, with a median posterior probablity higher than 50%.

We also infer the basic reproduction number and the epidemic start date using the cumulative number of detections up until January 16 as the first observation $\tilde{d}$, and the cumulative number of detections between January 16 and January 23 as the second observation $\tilde{d}'$. We use a flat prior $P(\boldsymbol{\theta}) = \text{const.}$, between 1.5 and 4 for $\mathcal{R}_0$ and between October 5th and December 20th for the starting date (cluster of 10 latent and 10 infectious). The posterior distribution averaged over 59 time series is illustrated in Fig. S16(C). This average posterior indicates a median basic reproduction number of 2.4 (90% CI, 1.75–3.55), and a median starting date on November 20th (90% CI, Oct. 17–Dec. 14), illustrating the timely analytics a global WWSN could provide about the growh dynamics of emerging outbreaks.

4 Airport table for the sentinel surveillance system

We report in Table S4 the list of sentinel airports for the baseline WWSN and the different optimization schemes (volume, entropy, greedy), when considering all subpopulations as equiprobable sources of an epidemic. Each airport is identified by its 3 letter IATA code. We additionally list the sentinels obtained by the greedy approach in the context of targeted optimization, assuming all subpopulations within a specific continent as equiprobable sources of an epidemic, but all subpopulations outside the continent are ignored.

References

1. Balcan, D. *et al.* Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 21484–21489 (2009).
2. Balcan, D. *et al.* Modeling the spatial spread of infectious diseases: The GLobal Epidemic and Mobility computational model. *J. Comput. Sci.* **1,** 132–145 (2010).
3. Pastore-Piontti, A. *et al.* Real-time assessment of the international spreading risk associated with the 2014 West African Ebola outbreak. *Mathematical and Statistical Modeling for Emerging and Re-emerging Infectious Diseases,* 39–56 (Jan. 2016).
4. Zhang, Q. *et al.* Spread of Zika virus in the Americas. *Proc. Natl. Acad. Sci. U.S.A.* **114** (2017).
5. Chinazzi, M. *et al.* The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* **368,** 395–400 (2020).
6. Davis, J. T. *et al.* Cryptic transmission of SARS-CoV-2 and the first COVID-19 wave. *Nature* **600,** 127–132 (2021).
7. Gridded Population of the World (GPW). Socioeconomic Data and Applications Center (SEDAC).
8. Mistry, D. *et al.* Inferring high-resolution human mixing patterns for disease modeling. *Nat. Commun.* **12,** 323 (2021).
9. Zipf, G. K. The P1 P2/D Hypothesis: On the Intercity Movement of Persons. *Am. Sociol. Rev.* **11,** 677–686 (1946).
10. Simini, F. *et al.* A universal model for mobility and migration patterns. *Nature* **484,** 96–100 (Apr. 2012).
11. Wearing, H. J., Rohani, P. & Keeling, M. J. Appropriate Models for the Management of Infectious Diseases. *PLOS Med.* **2,** 621–627 (July 2005).
12. Krylova, O. & Earn, D. J. D. Effects of the infectious period distribution on predicted transitions in childhood disease dynamics. *J. R. Soc. Interface* **10,** 20130098 (2013).

13. Brauer, F. *et al.* Mathematical epidemiology (Springer, 2008).
14. Allen, L. J. S. An introduction to stochastic processes with applications to biology (CRC press, 2010).
15. Miller, J. C. A primer on the use of probability generating functions in infectious disease modeling. *Infect. Dis. Model.* **3,** 192–248 (2018).
16. Roberts, D. *et al.* Quantifying the impact of individual and collective compliance with infection control measures for ethical public health policy. *Sci. Adv.* **9,** eabn7153 (2023).
17. Allard, A. *et al.* Complex Systems Modeling notes. COSMO-notes. (version: 2024-03-21).
18. Noël, P.-A. *et al.* Time evolution of epidemic disease on finite and infinite networks. *Phys. Rev. E* **79,** 026101 (2 Feb. 2009).
19. Allen, A. J. *et al.* Predicting the diversity of early epidemic spread on networks. *Phys. Rev. Res.* **4,** 013123 (1 2022).
20. Boudreau, M. C. *et al.* Temporal and Probabilistic Comparisons of Epidemic Interventions. *Bull. Math. Biol.* **85,** 118 (2023).
21. Johansson, M. A. *et al.* Assessing the Risk of International Spread of Yellow Fever Virus: A Mathematical Analysis of an Urban Outbreak in Asunción, 2008. *Am. J. Trop. Med. Hyg.* **86,** 349–358 (2012).
22. Johansson, M. A. *et al.* Nowcasting the Spread of Chikungunya Virus in the Americas. *PLOS ONE* **9,** e104915 (2014).
23. Mier-y-Teran-Romero, L., Tatem, A. J. & Johansson, M. A. Mosquitoes on a plane: Disinsection will not stop the spread of vector-borne pathogens, a simulation study. *PLOS Negl. Trop. Dis.* **11,** 1–13 (July 2017).
24. Lai, S. *et al.* Seasonal and interannual risks of dengue introduction from South-East Asia into China, 2005-2015. *PLOS Negl. Trop. Dis.* **12,** 1–16 (Nov. 2018).
25. Truelove, S. *et al.* Epidemics, Air Travel, and Elimination in a Globalized World: The Case of Measles. *medRxiv* (2020).
26. Abate, J. & Whitt, W. Numerical inversion of probability generating functions. *Oper. Res. Lett.* **12,** 245–251 (1992).
27. Hébert-Dufresne, L. *et al.* Beyond ¡i¿R¡/i¿¡sub¿0¡/sub¿: heterogeneity in secondary infections and probabilistic epidemic forecasting. *J. R. Soc. Interface* **17,** 20200393 (2020).
28. Gautreau, A., Barrat, A. & Barthélemy, M. Global disease spread: Statistics and estimation of arrival times. *J. Theor. Biol.* **251,** 509–522 (2008).
29. Brockmann, D. & Helbing, D. The Hidden Geometry of Complex, Network-Driven Contagion Phenomena. *Science* **342,** 1337–1342 (2013).
30. Iannelli, F. *et al.* Effective distances for epidemics spreading on complex networks. *Phys. Rev. E* **95,** 012313 (1 Jan. 2017).
31. Chen, L. M., Holzer, M. & Shapiro, A. Estimating epidemic arrival times using linear spreading theory. *Chaos* **28,** 013105 (Jan. 2018).
32. Jamieson-Lane, A. & Blasius, B. Calculation of epidemic arrival time distributions using branching processes. *Phys. Rev. E* **102,** 042301 (4 Oct. 2020).
33. Lloyd-Smith, J. O. *et al.* Superspreading and the effect of individual variation on disease emergence. *Nature* **438,** 355–359 (2005).
34. Nemhauser, G. L., Wolsey, L. A. & Fisher, M. L. An analysis of approximations for maximizing submodular set functions—I. *Math. Program.* **14,** 265–294 (1978).

35. COVID-19 - United Kingdom of Great Britain and Northern Ireland. https://www.who.int/emergencies/disease-outbreak-news/item/2020-DON304, Accessed on March 19, 2024.

36. The R value and growth rate. www.gov.uk/guidance/the-r-value-and-growth-rate, Accessed on September 15, 2023.

37. Davies, N. G. *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* **372,** eabg3055 (2021).

38. Volz, E. *et al.* Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature* **593,** 266–269 (2021).

39. St-Onge, G. Distribution of number of non-zero counts of a multinomial distributed set. Mathematics Stack Exchange. (version: 2023-09-12).

40. Oran, D. P. & Topol, E. J. Prevalence of Asymptomatic SARS-CoV-2 Infection. *Annals of Internal Medicine* **173,** 362–367 (2020).

41. Niehus, R. *et al.* Using observational data to quantify bias of traveller-derived COVID-19 prevalence estimates in Wuhan, China. *Lancet Infect. Dis.* **20,** 803–808 (2020).

42. Pekar, J. *et al.* Timing the SARS-CoV-2 index case in Hubei province. *Science* **372,** 412–417 (2021).

43. Riou, J. & Althaus, C. L. Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. *Eurosurveillance* **25** (2020).

44. Adhikari, S. P. *et al.* Epidemiology, causes, clinical manifestation and diagnosis, prevention and control of coronavirus disease (COVID-19) during the early outbreak period: a scoping review. *Infect. Dis. Poverty* **9,** 29 (2020).

45. Read, J. M. *et al.* Novel coronavirus 2019-nCoV (COVID-19): early estimation of epidemiological parameters and epidemic size estimates. *Philos. Trans. R. Soc. B* **376,** 20200265 (2021).

46. Novel Coronavirus – Thailand (ex-China). https://www.who.int/emergencies/disease-outbreak-news/item/2020-DON234, Accessed on September 15, 2023.

Table S4: Ordered list of sentinel airports (most important first) identified by the different optimization schemes. We use the aggregated global air-travel network from September 2022 to August 2023. We also list the (unordered) baseline sentinels as determined in the main text.

Global				Targeted (greedy)					
baseline	volume	entropy	greedy	Africa	Asia	Europe	N. America	Oceania	S. America
ADD	DXB	FRA	LHR	DXB	DXB	IST	LHR	AKL	MIA
JNB	LHR	CDG	DXB	CDG	SIN	AMS	CUN	SIN	PTY
ALG	CDG	AMS	CDG	ADD	ICN	STN	YYZ	BNE	SCL
DXB	IST	LHR	SIN	IST	BKK	AYT	CDG	LAX	MAD
DOH	AMS	IST	IST	JNB	IST	FRA	MIA	HNL	LIS
JED	SIN	MUC	FRA	LIS	HKG	LGW	YVR	SYD	LIM
LHR	FRA	BRU	MIA	JED	JED	CPH	FRA	DPS	GRU
CDG	ICN	DXB	ICN	NBO	KUL	DXB	LAX	NAN	BOG
IST	MAD	VIE	AMS	LHR	TPE	DUB	MEX	MNL	AEP
JFK	DOH	DOH	CUN	MRS	DOH	VIE	AMS	SFO	EZE
YYZ	BKK	FCO	MAD	DOH	LHR	LHR	CPH	NRT	FLL
MIA	LGW	MXP	BKK	RUN	CAI	TLV	ICN	DXB	JFK
ICN	BCN	MAN	LAX	CAI	FRA	CDG	IAH	CDG	MEX
BKK	DUB	YYZ	DOH	BRU	DMK	EVN	PVR	KUL	CDG
DEL	HKG	STN	HKG	FRA	SHJ	MUC	PUJ	HKG	CUN
SIN	JFK	ZRH	YYZ	AMS	NRT	BCN	SJD	MEL	LHR
HKG	KUL	JFK	JED	EBB	SVO	TAS	FLL	ICN	MCO
KUL	FCO	LGW	LIS	CMN	KWI	ARN	YUL	LHR	SDQ
SYD	LIS	IAD	BNE	MXP	MNL	LTN	DFW	NOU	MVD
GRU	TPE	JNB	CPH	TUN	SGN	MAN	NRT	DOH	AMS
–	VIE	MAD	PTY	LYS	AUH	FCO	JFK	GUM	FRA
–	MUC	PRG	AYT	ABJ	MED	OSS	MNL	POM	HAV
–	ORY	DUS	TPE	MAD	CGK	PMI	YYC	BKK	PUJ
–	JED	ADD	KUL	BOM	CDG	DUS	PTY	CHC	FCO
–	YYZ	WAW	ADD	DZA	MCT	MAD	EWR	CNS	IAH
–	CPH	ORD	LGW	LPA	DEL	GYD	MBJ	YVR	YYZ
–	STN	YUL	NRT	KGL	KIX	ZRH	GDL	DEL	ATL
–	ZRH	EDI	JFK	ACC	DUS	BER	ATL	HND	BCN
–	TLV	DUB	FCO	RUH	SYD	FRU	BOG	SGN	ASU
–	CAI	BCN	AKL	BCN	SAW	AGP	MCO	ZQN	SJO
–	MAN	CPH	MNL	FCO	RUH	LIS	MAD	VLI	VVI
–	MIA	ZAG	JNB	DAR	MFM	BRU	HND	KIX	MXP
–	MXP	BER	MUC	LFW	AMS	SVO	DUB	TPE	UIO
–	BRU	CPT	GRU	LGW	HAN	ALC	MUC	WLG	LAX
–	CUN	LAX	DUB	DSS	LED	DYU	TPE	SEA	ZRH
–	NRT	BUD	CAI	KWI	PUS	MXP	ORD	CGK	EWR
–	MNL	MIA	MEX	DEL	BOM	CRL	FCO	SUV	IST
–	PMI	CRL	YVR	CPT	MEL	OSL	GRU	PVG	DXB
–	AUH	TLV	STN	KRT	AMM	OPO	LAS	OOL	GYE
–	LAX	BHX	SYD	LOS	DAC	LBD	DOH	FRA	DOH