MOSAIC: Multi-Modal Open Surveillance with AI-Driven Calibrated Inference

Preprint, compiled June 10, 2026

Aravind Kannappan

Abstract

Background and objective. Pandemic early warning fails less often from a lack of data than from a lack of integration and calibration. Wastewater viral load, pathogen genomes, and outbreak text reports arrive on different cadences, in different units, and with different noise processes, and the probabilities ultimately shown to decision-makers are almost never validated. My objective is to fuse these three public, authentication-free streams into a single calibrated and falsifiable quantity, the posterior probability that the instantaneous reproduction number exceeds one, $\mathbb{P}(R_t > 1)$, which is the probability that transmission is currently growing.

Methods. I specify a renewal-equation latent-incidence model with stream-specific observation kernels (negative-binomial for wastewater, Poisson for text-event counts, Dirichlet–multinomial for lineage frequencies) and derive each detector from first principles: a Poisson–Gamma Bayesian online change-point detector with a windowed-maximum alarm, a Jensen–Shannon genomic-anomaly detector with a rolling null, an EpiEstim reproduction-number estimator, and a fusion rule that reduces, in the full backend, to a hierarchical posterior sampled by the No-U-Turn Sampler. I validate $\mathbb{P}(R_t > 1)$ as a prospective probabilistic forecast of realised growth on the multi-year CDC National Wastewater Surveillance System record.

Results. Over N=1,334 day-ahead forecasts the system attains an Expected Calibration Error of 0.086, a Brier score of 0.124, and an AUROC of 0.917; a Murphy decomposition attributes the Brier score to a small reliability term (0.010) and a large resolution term (0.136), so the forecasts are both well calibrated and discriminative, and the growth probability leads national wave peaks by a median of 68 days. I also surface and correct a numerical defect in the reproduction-number tail probability that silently corrupts large-count series.

Conclusion. An explicit, falsifiable target quantity together with a faithful backend-free deployment is a practical route to trustworthy open early warning.

Github: Github

Keywords epidemic early warning · wastewater surveillance · effective reproduction number · Bayesian change-point detection · probabilistic calibration · multi-modal data fusion

1 Background

Every major epidemic of the past two decades, the 2009 H1N1 pandemic, the 2013–2016 West African Ebola epidemic, the emergence of SARS-CoV-2 and its successive variants, the 2022 multi-country mpox outbreak, and the ongoing panzootic of highly pathogenic avian influenza, was, in retrospect, preceded by signals that were detectable days to weeks before the official declaration. The recurring post-mortem finding is not that the relevant data did not exist, but that it was scattered across incompatible systems, expressed in incomparable units, and never assembled into a single quantity that a decision-maker could trust and act on.

The stakes of this gap are asymmetric. In the exponential-growth phase of an epidemic, a detection delay of a single generation interval can multiply the eventual burden, because the number of infections at the moment of intervention sets the scale of everything that follows. Conversely, a *false* alarm is comparatively cheap if it is accompanied by an honest probability and an explicit uncertainty, because a calibrated 30% invites monitoring rather than lockdown. This asymmetry is the reason I privilege early, probabilistic, *calibrated* signals over late, categorical, point alerts.

This points to two binding constraints, neither of which is data scarcity. The first is *integration*: municipal wastewater is sampled routinely and published as a standing national program [20, 21]; pathogen genomes

are deposited continuously and shared openly with phylogenetic context [14]; and structured outbreak reports are issued by the World Health Organization and by moderated expert networks [5, 23]. Each is informative on its own, but they are asynchronous (weekly versus continuous versus sporadic), differently scaled (concentration percentiles versus categorical lineage frequencies versus event counts), and differently noised (overdispersed counts versus Dirichlet–multinomial composition versus heavy-tailed reporting). Turning them into one coherent signal is a statistical, not merely an engineering, problem.

The second constraint is *calibration*. A system that emits a number, "a 78% chance of growth", is only useful if that number means what it says: if, across the many times the system says 78%, the event happens about 78% of the time. Modern statistical and machine-learning systems are frequently and badly miscalibrated [13], and epidemic-forecasting tools rarely report calibration at all, so a stated probability is too often an uninterpretable score rather than an actionable frequency.

A falsifiable target. I argue that the right quantity for an early-warning system to emit is the posterior probability that transmission is currently growing,

$$p_t \;=\; \mathbb{P}\big(R_t > 1 \,\big|\, \text{data up to } t\big), \tag{1}$$

where R_t is the instantaneous (effective) reproduction number, the expected number of secondary infections per infectious individual at time t. I choose Eq. (1) for three reasons. It is *unitless*, hence comparable across pathogens, locations, and data modalities. It has a crisp *operational* meaning: $R_t > 1$ is exactly the condition under which the epidemic curve is turning upward. And, crucially, it is *falsifiable*: the event $\{R_t > 1\}$ is, with a lag, observable through subsequent growth, so the calibration of p_t is something one can *measure* rather than merely assert. This last property is what allows me to report, in §4, an Expected Calibration Error for the live system rather than a methodological promise.

Contributions.

1. **A complete probabilistic specification** (§3.2–3.5) for multi-modal outbreak inference: a renewal latent-incidence process with negative-binomial, Poisson, and Dirichlet–multinomial observation kernels, with each stream detector, Poisson–Gamma BOCPD, Jensen–Shannon anomaly, EpiEstim R_t, derived from first principles, and a fusion rule that reduces to a hierarchical posterior in the full backend.
2. **A two-tier open implementation** (§3.8): a serverless TypeScript deployment that runs the entire pipeline with no backend, plus a Python backend that performs the full No-U-Turn-Sampler posterior, designed so the former is mathematically faithful to the latter.
3. **An empirical calibration study** (§4) on the real multi-year national wastewater record showing that the served $\mathbb{P}(R_t > 1)$ is well-calibrated (ECE 0.086) and discriminative (AUROC 0.917), with a Murphy decomposition of the Brier score, live multi-stream alert attribution, and a documented numerical correction I recommend other implementations check.

Why a backend-free deployment matters. A recurring failure of research surveillance tools is that they are demonstrated once and then decay, because they depend on a server, a credentialed data feed, or a maintainer's laptop. A second theme of this paper is therefore *operational*: the entire MOSAIC pipeline runs in a serverless edge function with no backend and no authenticated data access, so the public dashboard is always live and reproducible by anyone who forks the repository. This constraint is not free, it forces every detector to have a closed-form or streaming implementation and every data movement to be small, but I show it is achievable without sacrificing the probabilistic target, and that meeting it exposed real bugs (a quadratic genomic detector, a numerically unstable tail probability, a self-referential fetch pattern) that a heavier prototype would have hidden.

Organization. §2 surveys related work. §3.1 fixes notation and states the inference problem and its assumptions. §3.2 specifies the generative model. §3.3 derives the three stream detectors and §3.4 the reproduction-number estimator, including the numerical correction. §3.5 treats fusion and §3.6 forecasting. §3.7 defines the calibration instruments, §3.8 the implementation, and §4 the experiments. §5.2–5.4 discuss limitations and dual-use, and §5.5 concludes.

2 Related Literature

Wastewater-based epidemiology. Wastewater concentration of viral RNA tracks community prevalence and typically leads clinical case reports, because shedding begins early in infection and is independent of care-seeking and testing behaviour [20, 21]. National programs now publish standardized activity metrics; the U.S. NWSS reports, per site, a percentile of current activity relative to that site's own history. Most public-facing tools display concentrations or qualitative trend arrows. MOSAIC instead converts the national record into a *calibrated growth probability*, which is what makes its output validatable.

Reproduction-number estimation. The instantaneous reproduction number is most commonly estimated by the sliding-window Poisson–Gamma method of Cori et al. [6], building on renewal-process theory [10, 24]. The method is attractive because it is conjugate and closed-form, and because $\mathbb{P}(R_t > 1)$ falls out of

the Gamma posterior directly. Alternatives include Wallinga–Teunis retrospective estimators and state-space / particle-filter formulations [9]; I adopt EpiEstim as the renewal core precisely because its closed form admits a faithful serverless implementation.

Change-point and anomaly detection. Bayesian Online Change-Point Detection [1] maintains a recursive posterior over the time since the last change-point (the "run length") under conjugate emissions, and is well suited to streaming count data. Distributional shift in categorical lineage frequencies is naturally measured by the Jensen–Shannon divergence [17], a bounded, symmetric smoothing of the Kullback–Leibler divergence whose square root is a metric [8]. Sequence-level anomaly detection for novel pathogens is an active area [18]; MOSAIC operates at the coarser lineage-frequency level, which is robust and available in near-real-time.

Multi-modal fusion in epidemiology. Nowcasting and forecasting hubs have demonstrated that ensembling and multi-signal models outperform any single stream [7], and digital / event-based surveillance has a long history of mining text for outbreak signals [4, 5]. What is comparatively rare is a single open system that fuses wastewater, genomic, and text evidence into one probabilistic target and then reports that target's calibration; that gap is the motivation for this work.

Probabilistic forecast calibration. The reliability diagram, the Brier score and its reliability–resolution–uncertainty decomposition [3, 19], proper scoring rules [12], and the Expected Calibration Error [13] are the standard instruments for asking whether stated probabilities match observed frequencies. I use all of them.

Collaborative forecasting and pooling. Large-scale forecasting efforts have repeatedly found that combining models, by ensembling, stacking, or opinion pooling, is more robust than any single component, and that well-constructed ensembles are among the best-calibrated forecasters available [7]. Logarithmic and linear opinion pools have a long statistical pedigree as ways to aggregate probabilistic judgements; my lightweight fusion (§3.5) is a weighted logarithmic pool of per-stream alarms, specialised so that absent modalities neither vote nor dilute.

Probabilistic programming. The full hierarchical posterior is sampled with the No-U-Turn Sampler [16] as implemented in NumPyro [2, 22], which provides effect-handler-based composition and JIT-compiled, vectorized MCMC. Conjugacy in the lightweight detectors (Poisson–Gamma for counts, Dirichlet–multinomial for composition) is what allows the same models to be evaluated in closed form on the edge without a sampler at all.

Table 1: Notation.

Symbol	Meaning
I_t	latent new infections on day t
R_t	instantaneous reproduction number
$w_{1:S}$	discretised generation-interval pmf
Λ_t	total infectiousness $\sum_s w_s I_{t-s}$
$C_t, E_t, \mathbf{L}_t$	wastewater / text / genomic observations
r_t	BOCPD run length (time since change)
H	BOCPD hazard $1/\mu_{\mathrm{RL}}$
$a_t^{(j)}$	alarm of stream $j \in \{\mathrm{txt,ww,gen}\}$
p_t	$\mathbb{P}(R_t > 1 \mid \mathrm{data} \le t)$
Φ	standard normal CDF

3 Methods

3.1 Problem setup and notation

I index discrete time (days) by t. The unit of analysis is a *surveillance cell* (k, ℓ): a pathogen k in a location ℓ. For a fixed cell, let $I_t \ge 0$ denote latent new infections on day t. The three streams provide noisy, lagged, partial observations of $I_{1:T}$:

- wastewater activity C_t (a non-negative, over-dispersed level);
- text-event counts E_t (non-negative integers, the number of WHO/ProMED reports);
- lineage frequency vectors $\mathbf{L}_t \in \Delta^{K-1}$ (a point on the $(K-1)$-simplex over K lineages).

The objects of inference are the latent incidence $I_{1:T}$ and the time-varying reproduction number R_t, and the reported quantity is the tail posterior p_t of Eq. (1). Table 1 collects the notation.

I make the following modelling assumptions.

Assumption 1 (Renewal dynamics)**.** Incidence follows the renewal equation with a fixed, known generation-interval distribution $w_{1:S}$ and a smoothly varying R_t.

Assumption 2 (Conditional independence)**.** Given the latent path $I_{1:T}$ and stream-specific nuisance parameters, the three streams are conditionally independent observations.

Assumption 3 (Stationary observation nuisances)**.** Within an inference window the shedding scale, reporting rate, ascertainment, and sequencing depth vary slowly relative to R_t.

These assumptions are standard for renewal-based nowcasting; I revisit their limits in §5.2.

3.2 The generative model

3.2.1 Latent incidence and the renewal equation

Under Assumption 1, expected incidence is the product of the reproduction number and the total infectiousness

contributed by past infections [6, 24]:

$$\mathbb{E}[I_t \mid R_t, I_{<t}] \;=\; R_t\,\Lambda_t, \qquad \Lambda_t = \sum_{s=1}^{S} w_s\, I_{t-s}, \quad (2)$$

with $w_s \geq 0$, $\sum_s w_s = 1$. The reproduction number evolves as a geometric random walk, which encodes the prior that transmissibility changes smoothly rather than abruptly:

$$\log R_t \;\sim\; \mathcal{N}\big(\log R_{t-1},\, \sigma_R^2\big). \quad (3)$$

The generation interval is discretised from a Gamma density by midpoint quadrature, $w_s \propto f_{\text{Gamma}}(s;\kappa,\theta)$ for $s = 1,\dots,S$, and renormalised; for SARS-CoV-2 I use mean 5.1 and standard deviation 4.0 days [15], truncated at $S = 21$.

The renewal equation is the discrete-time analogue of the Lotka integral equation of demography and underlies essentially all modern R_t estimators [6, 10, 24]. Its key property for my purposes is that R_t is a *local* quantity: it depends only on the ratio of current incidence to recent infectiousness, not on the absolute scale of the epidemic. This is what lets me read $\{R_t > 1\}$ off a relative activity index such as the wastewater percentile without knowing the true number of infections, and it is why the threshold $R_t = 1$, rather than any particular incidence level, is the natural alarm boundary: it is invariant to the unknown ascertainment and shedding constants that scale the observations. In the small-σ_R regime, the exponential growth rate is approximately $\log R_t$ divided by the mean generation interval, so a sustained R_t a few percent above one corresponds to a doubling time of weeks, fast enough to matter, slow enough to act on.

3.2.2 Stream observation kernels

The three streams are conditionally independent observations of the same latent $I_{1:T}$ (Assumption 2).

Wastewater (negative-binomial). Viral load is a lagged, over-dispersed, count-like measurement of prevalence:

$$C_t \;\sim\; \text{NegBin}\big(\rho_W\, I_{t-d_W},\; \phi_W\big), \quad (4)$$

with shedding scale $\rho_W > 0$, reporting delay d_W, and dispersion ϕ_W. The negative-binomial captures the extra-Poisson variance induced by heterogeneous shedding, sampling, and assay noise.

Text (Poisson). The number of qualifying outbreak reports is modelled as Poisson in prevalence, modulated by a slowly varying ascertainment $\hat{q}_t$:

$$E_t \;\sim\; \text{Poisson}\big(\lambda_N\, \hat{q}_t\, I_{t-d_N}\big), \quad (5)$$

with reporting rate λ_N and delay d_N.

Table 2: Default priors and hyperparameters (shared by both tiers).

Quantity	Setting	Role
$\log R_t$ walk σ_R	0.1	R_t smoothness
EpiEstim R_t prior	Gamma, mean/sd 5	diffuse
EpiEstim window τ	7 d	estimation
Gen. interval	mean 5.1, sd 4.0 d	SARS-CoV-2
BOCPD hazard H	1/30	run length
BOCPD prior (α_0, β_0)	$(1, 1)$	rate prior
JSD baseline B	90 windows	rolling null
Forecast damping ϕ	0.8	mean-reversion

Genomic (Dirichlet–multinomial). The lineage composition of sampled sequences is Dirichlet–multinomial, with the composition map $\mathbf{f}$ linking incidence-weighted lineage prevalences to the expected frequencies, sequencing depth N_t, and concentration κ controlling over-dispersion relative to the multinomial:

$$\mathbf{L}_t \;\sim\; \text{DirMult}\big(N_t,\; \kappa\, \mathbf{f}(I_{t-d_G}, \boldsymbol{\theta}_L)\big). \quad (6)$$

3.2.3 The joint posterior

Collecting all parameters into $\Theta = (R_t, I_{1:T}, \rho_W, \phi_W, \lambda_N, \kappa, \dots)$ and applying Assumption 2, the joint posterior factorises over streams:

$$p(\Theta \mid \mathbf{C}, \mathbf{E}, \mathbf{L}) \;\propto\; p(\mathbf{C} \mid \Theta)\, p(\mathbf{E} \mid \Theta)\, p(\mathbf{L} \mid \Theta)\, p(\Theta). \quad (7)$$

The full backend samples Eq. (7) directly. Because that sampler is unavailable in a backend-free deployment, the lightweight tier replaces the joint fit with per-stream conjugate detectors (§3.3) combined by an independent-evidence rule (§3.5); the two tiers share the observation models (4)–(6) and the target (1).

Priors and hyperparameters. The priors are weakly informative and chosen to be conservative (Table 2). The reproduction-number random walk uses a small innovation scale σ_R so that R_t is encouraged to change smoothly; the EpiEstim window prior on R_t is a high-variance Gamma centred above one, which is deliberately diffuse so that the data dominate; the BOCPD hazard corresponds to an expected run length of a month, a neutral prior on the frequency of regime change; and the genomic baseline spans ninety windows so that a transient fluctuation is not mistaken for a sustained shift. None of these are tuned to the evaluation; they are fixed a priori and reused across all results.

3.3 Stream detectors

3.3.1 Bayesian online change-point detection

For the count streams I detect abrupt regime changes with BOCPD [1]. Let $r_t \in \{0, 1, 2, \dots\}$ be the *run length*, the number of steps since the last change-point. BOCPD maintains the joint $p(r_t, n_{1:t})$ and hence the

run-length posterior $p(r_t \mid n_{1:t})$ by a forward recursion with two messages: a *growth* message (no change, run length increments) and a *change-point* message (run length resets to 0). With a constant hazard $H = 1/\mu_{\mathrm{RL}}$,

$$p(r_t{=}r{+}1, n_{1:t}) = p(r_{t-1}{=}r, n_{1:t-1})\,(1{-}H)\,\pi_r(n_t), \quad (8)$$

$$p(r_t{=}0, n_{1:t}) = \sum_r p(r_{t-1}{=}r, n_{1:t-1})\,H\,\pi_r(n_t), \quad (9)$$

where $\pi_r(n_t) = p(n_t \mid r_{t-1}{=}r, n_{1:t-1})$ is the posterior predictive of the emission model conditioned on the data assigned to the current run.

Conjugate Poisson–Gamma predictive. I model counts as $n_t \mid \lambda \sim \mathrm{Poisson}(\lambda)$ with a conjugate prior $\lambda \sim \mathrm{Gamma}(\alpha_0, \beta_0)$. Conditioned on a run that has accumulated sufficient statistics (α, β), the predictive is obtained by marginalising λ:

$$\begin{aligned}\pi(n \mid \alpha, \beta) &= \int_0^\infty \mathrm{Poisson}(n \mid \lambda)\,\mathrm{Gamma}(\lambda \mid \alpha, \beta)\,d\lambda \\ &= \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)\,n!}\left(\tfrac{\beta}{\beta+1}\right)^{\alpha}\left(\tfrac{1}{\beta+1}\right)^{n}, \end{aligned} \quad (10)$$

which is exactly the negative-binomial $\mathrm{NegBin}(n; \alpha, \frac{\beta}{\beta+1})$. The sufficient statistics update along each run by conjugacy as $\alpha \leftarrow \alpha + n$, $\beta \leftarrow \beta + 1$, with a fresh (α_0, β_0) seeding every newly created run length.

The alarm: windowed maximum, not cumulative product. The instantaneous change-point probability is $P(r_t{=}0 \mid n_{1:t})$ from the normalised recursion. The stream alarm I report is its maximum over a recent window of length W,

$$a_t^{\mathrm{cp}} = \max_{t-W<\tau\le t} P(r_\tau{=}0 \mid n_{1:\tau}). \quad (11)$$

The windowed maximum is a deliberate design choice, justified by the following.

Proposition 1 (Cumulative combination is degenerate under the null)**.** Under the null of no change-point, the per-step reset probability has a positive floor $\approx H$. Hence the cumulative "has a change occurred in the window" probability $1 - \prod_\tau (1 - P(r_\tau{=}0))$ converges to 1 as W grows, independently of the data.

Sketch. If $P(r_\tau{=}0) \ge H > 0$ for all τ, then $\prod_{\tau=t-W+1}^{t}(1 - P(r_\tau{=}0)) \le (1-H)^W \to 0$, so the complement $\to 1$. The windowed maximum (11) instead returns the single strongest recent reset, which stays near the floor under the null and spikes only on a genuine regime shift. □

Cost. The exact recursion is $O(t)$ per step because the run-length support grows; truncating the support at a maximum run length (or pruning low-mass hypotheses) makes it $O(1)$ amortized, which is what the implementation uses. Algorithm 1 states the full procedure.

Algorithm 1: Poisson–Gamma BOCPD with windowed alarm

Input: counts $n_{1:T}$; prior (α_0, β_0); hazard H; window W
Output: change-point probabilities $g_{1:T}$; alarm a_T^{cp}
$R \leftarrow [1]$; $\boldsymbol{\alpha} \leftarrow [\alpha_0]$; $\boldsymbol{\beta} \leftarrow [\beta_0]$;
for $t = 1$ **to** T **do**
  $\pi_r \leftarrow \mathrm{NegBin}\big(n_t; \alpha_r, \frac{\beta_r}{\beta_r+1}\big)$ for each run r;
  $\mathrm{grow}_r \leftarrow R_r(1-H)\pi_r$; $\mathrm{cp} \leftarrow \sum_r R_r H \pi_r$;
  $R \leftarrow \mathrm{normalize}\big([\mathrm{cp},\ \mathrm{grow}_0, \mathrm{grow}_1, \dots]\big)$;
  $\boldsymbol{\alpha} \leftarrow [\alpha_0,\ \boldsymbol{\alpha} + n_t]$; $\boldsymbol{\beta} \leftarrow [\beta_0,\ \boldsymbol{\beta} + 1]$;
  $g_t \leftarrow R_0$; // $P(r_t{=}0 \mid n_{1:t})$
$a_T^{\mathrm{cp}} \leftarrow \max_{T-W<\tau\le T} g_\tau$;
return $g_{1:T}$, a_T^{cp}

3.3.2 Wastewater pipeline

National aggregation. NWSS reports, per site i and day t, a percentile $c_{i,t} \in [0, 100]$ of current 15-day-averaged activity relative to that site's own history. I form a national signal by simple aggregation over the reporting set $\mathcal{S}_t$,

$$\bar{c}_t \;=\; \frac{1}{|\mathcal{S}_t|}\sum_{i\in\mathcal{S}_t} c_{i,t}, \qquad |\mathcal{S}_t| \approx 1{,}200, \quad (12)$$

performed server-side via a single aggregation query so that the transferred payload is a small daily series rather than millions of site-day rows. Because $c_{i,t}$ is already a site-relative level, $\bar{c}_t$ is a dimensionless index of national activity that is comparable across time despite a changing site panel.

Change-point and sustained elevation. I run the Poisson–Gamma BOCPD of §3.3.1 on the discretised series $n_t = \lfloor \bar{c}_t \rceil$ to obtain $a_t^{\mathrm{ww,cp}}$. Abrupt onset is only half the signal: a slowly-rising but persistently elevated wave is also informative and may not trip a change-point. Because the percentile is itself a level, I add an elevation alarm and combine the two by a noisy-or,

$$\begin{aligned} a_t^{\mathrm{ww,lvl}} &= \mathrm{clip}\Big(\tfrac{\bar{c}_t - 50}{40}, 0, 1\Big), \\ a_t^{\mathrm{ww}} &= 1 - (1 - a_t^{\mathrm{ww,cp}})(1 - a_t^{\mathrm{ww,lvl}}). \end{aligned} \quad (13)$$

3.3.3 Text pipeline

Extraction. WHO Disease Outbreak News (queried newest-first) and ProMED posts are parsed into structured events (pathogen, location, date, counts, novelty). The full backend uses a constrained large-language-model extractor; the lightweight tier uses curated patterns plus a resolver that maps free-text locations to ISO-3166 codes, which is what lets an alert colour the correct countries on the map. Daily per-pathogen counts $e_{k,t}$ feed the detector.

Dense series and recency weighting. Outbreak reports are sparse and bursty. I build a dense, zero-filled daily count series ending at the present, so that a cluster of reports following a quiet interval is a strong change-point and a long-stale pathogen decays to the floor. Running BOCPD gives b_k; the operational text alarm composes it with a recency factor and an intensity factor,

$$a_k^{\text{txt}} = \min\Big(1,\ 0.45\, b_k + 0.55\, e^{-\Delta_k/45}\big(1 - e^{-m_k/3}\big)\Big), \tag{14}$$

where Δ_k is the days since the last report and m_k the number of reports in the last 120 days. The exponential recency kernel gives a half-life of ≈ 31 days; the saturating intensity term prevents a single report from dominating while rewarding sustained reporting.

The text stream is the most heterogeneous of the three, and its extraction is correspondingly the most engineered. WHO Disease Outbreak News are authoritative but sporadic and must be requested most-recent-first, because the underlying API returns its oldest records by default, a detail that, if missed, silently feeds the detector reports from a decade ago. ProMED posts are timelier but semi-structured prose. In the lightweight tier I extract pathogen, location, and counts with curated patterns and resolve every location to one or more ISO-3166 codes; the multi-country resolver is what lets an alert about, say, a multi-country cholera report colour each affected country rather than collapsing to a vague "global." In the full tier a constrained language-model extractor raises recall on free-text reports, but the regex tier is deliberately conservative: it prefers to miss an ambiguous report than to fabricate a structured event, because a false event is a false alarm that the calibration machinery cannot catch (the text stream has no continuous ground-truth analogue of wastewater growth).

3.3.4 Genomic anomaly

A lineage shift is a change in the 14-day distribution $\mathbf{L}_t$ relative to a rolling baseline $\bar{\mathbf{L}}_t = \frac{1}{B}\sum_{\tau=t-B}^{t-1}\mathbf{L}_\tau$. I measure it by the Jensen–Shannon divergence [17],

$$\mathrm{JSD}(\mathbf{L}_t \,\|\, \bar{\mathbf{L}}_t) = \tfrac{1}{2}D_{\mathrm{KL}}(\mathbf{L}_t\|\mathbf{m}_t) + \tfrac{1}{2}D_{\mathrm{KL}}(\bar{\mathbf{L}}_t\|\mathbf{m}_t), \tag{15}$$

with $\mathbf{m}_t = \frac{1}{2}(\mathbf{L}_t + \bar{\mathbf{L}}_t)$. Two properties make it well suited to sparse compositional data.

Lemma 1 (Boundedness and symmetry)**.** JSD is symmetric and satisfies $0 \le \mathrm{JSD}(\mathbf{p}\|\mathbf{q}) \le \log 2$, with equality on the left iff $\mathbf{p} = \mathbf{q}$ and on the right iff the supports are disjoint. It is finite even when some component frequencies are zero.

Sketch. Symmetry is immediate from (15). Writing $\mathrm{JSD} = \mathcal{H}(\mathbf{m}) - \frac{1}{2}\mathcal{H}(\mathbf{p}) - \frac{1}{2}\mathcal{H}(\mathbf{q})$ with $\mathcal{H}$ the Shannon entropy, Jensen's inequality gives $\mathrm{JSD} \ge 0$, and each KL term is bounded by $\log 2$ because $\mathbf{m} \ge \frac{1}{2}\mathbf{p}$ so $D_{\mathrm{KL}}(\mathbf{p}\|\mathbf{m}) \le \log 2$; finiteness follows because $\mathbf{m}_i = 0 \Rightarrow p_i = 0$. □

The alarm is the empirical right-tail probability of JSD against a null estimated from quiescent windows. A degenerate single-lineage panel has JSD $\equiv 0$ and is reported as *no* anomaly rather than a spurious one.

A complexity pitfall. A naive implementation recomputes each window's ordered frequency vector inside the rolling-baseline loop, costing $O(T^2K^2)$ in the number of windows T and lineages K. For the SARS-CoV-2 panel (T=156, K=664) this is billions of operations and times out on serverless infrastructure. Pre-computing the T vectors once reduces the cost to $O(TBK)$, about 20 ms here, a constant-factor algorithmic fix that is the difference between a live and a dead deployment.

3.4 Reproduction number estimation

3.4.1 EpiEstim by conjugacy

Given an incidence proxy $y_{1:T}$ (here $y_t = \lfloor \bar{c}_t \rceil$ from wastewater), EpiEstim [6] assumes that, over a trailing window of length τ, incidence is Poisson with mean $R_t\Lambda_u$,

$$y_u \ \sim\ \mathrm{Poisson}(R_t\,\Lambda_u), \quad u \in [t-\tau+1,\, t], \tag{16}$$

with $\Lambda_u = \sum_{s\ge1} w_s y_{u-s}$. With a conjugate prior $R_t \sim \mathrm{Gamma}(a,b)$ the window likelihood is $\prod_u \mathrm{Poisson}(y_u \mid R_t\Lambda_u) \propto R_t^{\sum_u y_u} e^{-R_t\sum_u \Lambda_u}$, so the posterior is Gamma in closed form,

$$R_t \mid y_{1:t} \sim \mathrm{Gamma}\big(a', b'\big),\ \ a' = a + \textstyle\sum_u y_u,\ \ b' = b + \sum_u \Lambda_u. \tag{17}$$

The reported quantity is the tail probability

$$p_t = \mathbb{P}(R_t > 1 \mid y_{1:t}) = 1 - F_{\mathrm{Gamma}}(1; a', b'), \tag{18}$$

where $F_{\mathrm{Gamma}}(\cdot; a', b')$ is the Gamma CDF, equivalently the regularised lower incomplete gamma $P(a', b')$ after the change of variable $R_t b'$. The same posterior yields the point estimate and credible interval the dashboard displays: the median $\hat{R}_t = F^{-1}_{\mathrm{Gamma}}(0.5; a', b')$ and the central 95% interval $[F^{-1}_{\mathrm{Gamma}}(0.025), F^{-1}_{\mathrm{Gamma}}(0.975)]$. Two modelling conveniences are worth noting. First, the prior $\mathrm{Gamma}(a,b)$ is conjugate, so no sampling is required, the entire R_t inference is three sums and two special-function evaluations per day, which is what makes it feasible on the edge. Second, the window length τ trades variance for bias: a short window tracks rapid changes but is noisy, a long window is smooth but lags; I use $\tau = 7$ days, a standard compromise that matches the timescale of the generation interval.

3.4.2 A numerical correction at large shape

Evaluating (18) naively is a trap. The standard series and continued-fraction expansions for the regularised

incomplete gamma converge in $O(a')$ terms; truncating them at a few hundred iterations, common in lightweight libraries, silently returns wrong values once the posterior shape $a' = a + \sum_u y_u$ grows into the hundreds, as it does for any daily count series. Empirically the truncated evaluation can invert the sign of the estimate, mapping growth to decay.

I switch, for $a' > 80$, to the Wilson–Hilferty cube-root normal approximation [25]. If $X \sim \mathrm{Gamma}(a, 1)$ then $(X/a)^{1/3}$ is approximately normal with mean $1 - \frac{1}{9a}$ and variance $\frac{1}{9a}$, giving

$$F_{\mathrm{Gamma}}(x; a, 1) \approx \Phi\left(\frac{(x/a)^{1/3} - \left(1 - \frac{1}{9a}\right)}{\sqrt{1/(9a)}}\right), \quad (19)$$

accurate to $\sim 10^{-4}$ for $a \gtrsim 30$ and numerically stable for any a. Posterior quantiles for the credible interval use the same approximation. The corrected estimator reproduces a reference scientific-computing implementation to three decimals, and is what underlies the calibration of §4. I flag this because the defect is easy to introduce, hard to notice (the output looks plausible), and corrupts exactly the high-count regime where wastewater nowcasting operates.

3.5 Multi-stream fusion

Full tier. In the backend the three observation models (4)–(6) are conditioned on the shared latent path $I_{1:T}$ and the $\log R_t$ random walk (3), and the joint posterior (7) is sampled by NUTS [16, 22] with four chains; p_t is the posterior mass on $\{R_t > 1\}$ at the latest day. This couples the streams through $I_{1:T}$ and is the appropriate tool when a cell has overlapping modalities. Sampling the latent incidence path jointly with $\log R_t$ is a non-centred hierarchical model: I sample the standardised random-walk innovations and transform them, which decorrelates the geometry and lets NUTS mix efficiently despite the strong prior coupling between adjacent R_t. Posterior summaries are the median R_t, the central credible interval, and the tail mass p_t; standard convergence diagnostics ($\hat{R}$ and effective sample size) gate the output. The backend is the right tool precisely when the lightweight independence assumption is least defensible, a pathogen observed simultaneously in wastewater and genomes, where the two streams share the latent incidence and must not be double-counted.

Lightweight tier. Absent a backend, I combine the per-stream alarms $a^{\mathrm{txt}}, a^{\mathrm{ww}}, a^{\mathrm{gen}} \in [0, 1]$ by an independent-evidence (noisy-or) rule, restricted to the set $\mathcal{A}_k$ of streams that actually carry data for pathogen k:

$$p_k = 1 - \prod_{j \in \mathcal{A}_k} \left(1 - \omega_j \, a_k^{(j)}\right), \qquad \omega_j = \tfrac{1}{|\mathcal{A}_k|}. \quad (20)$$

Equation (20) is the probability that at least one stream's evidence "fires," under the working assumption that the streams are independent alarms. The form is not arbitrary. Writing the complementary failure probabilities $q_j = 1 - \omega_j a^{(j)}$, the fused complement is multiplicative, $1 - p_k = \prod_j q_j$, so the log non-event probability is $\log(1 - p_k) = \sum_j \log q_j$. This is the logarithmic pooling of independent evidence: each stream contributes additively in log-space, and no single weak stream can by itself drive p_k to one. It is the natural choice when the alarms are treated as conditionally independent likelihood-ratio summaries of the same latent event $\{R_t > 1\}$; the full model (7) relaxes the independence by coupling through $I_{1:T}$.

Proposition 2 (Monotonicity and no-drowning)**.** p_k in (20) is non-decreasing in each $a_k^{(j)}$, equals $\max_j a_k^{(j)}$ when a single stream is present, and satisfies $p_k \geq \frac{1}{|\mathcal{A}_k|} \max_j a_k^{(j)}$ always.

Sketch. Monotonicity is immediate since $\partial p_k / \partial a^{(j)} = \omega_j \prod_{j' \neq j} q_{j'} \geq 0$. With one present stream $|\mathcal{A}_k|$=1, ω=1 and $p_k = a$. For the bound, expanding the product keeps the first-order term $\omega_j a^{(j)}$, and all omitted terms are non-negative. □

Renormalising the weights over the present streams (rather than over all three) is exactly what makes Proposition 2 useful: a filovirus seen only in text has no wastewater or genomic panel, and dividing its single alarm by three would bury a real outbreak. With renormalisation a text-only cell uses $\omega = 1$, so $p_k = a_k^{\mathrm{txt}}$, while a cell with all three present splits the weight evenly and requires corroboration to reach a high probability.

Attribution. The dashboard reports each stream's share of the evidence as the normalised marginal term $\omega_j a_k^{(j)} / \sum_{j'} \omega_{j'} a_k^{(j')}$, an additive analogue of a Shapley attribution that answers "how much of this alert is driven by each modality." Algorithm 2 summarises the lightweight pipeline end to end.

3.6 Forecasting

I extend p_t forward by a damped-trend projection [11] in logit space, which keeps the forecast in $[0, 1]$ and lets the trend mean-revert rather than extrapolate a transient slope. With $z_t = \operatorname{logit} p_t$, a recent drift $\hat{\delta}$ estimated from the last k first differences, residual scale $\hat{\sigma}$, and damping $\phi = 0.8$, the h-step forecast and its predictive standard deviation are

$$\hat{z}_{t+h} = z_t + \hat{\delta} \sum_{i=1}^{h} \phi^{\,i-1}, \qquad \widehat{\mathrm{sd}}(z_{t+h}) = \hat{\sigma}\sqrt{h}, \quad (21)$$

and the 95% band is $\operatorname{logit}^{-1}(\hat{z}_{t+h} \pm 1.96\,\widehat{\mathrm{sd}})$. The geometric damping bounds the cumulative drift at $\hat{\delta}/(1 - \phi)$, so the projection settles to a level rather than diverging, while the band widens as $\sqrt{h}$ to communicate the rapidly growing uncertainty of multi-week projection.

Algorithm 2: MOSAIC lightweight fusion (per cycle)

Input: wastewater $\bar{c}_{1:t}$; events $\{e_{k,\cdot}\}$; snapshots $\mathbf{L}_{1:t}$
Output: ranked alerts $\{(k,\ \text{countries}_k,\ p_k,\ \text{attr}_k)\}$
$a^{\text{ww}} \leftarrow$ Eq. (13) on BOCPD($\lfloor \bar{c} \rceil$);
foreach *pathogen k in events* **do**
  $a_k^{\text{txt}} \leftarrow$ Eq. (14) on BOCPD($e_{k,\cdot}$);
$a^{\text{gen}} \leftarrow$ tail-prob of JSD via Eq. (15);
foreach *pathogen k* **do**
  $\mathcal{A}_k \leftarrow$ streams with data for k; $\omega \leftarrow 1/|\mathcal{A}_k|$;
  $p_k \leftarrow 1 - \prod_{j \in \mathcal{A}_k} (1 - \omega\, a_k^{(j)})$; `// Eq.` (20)
  if $p_k \geq 0.05$ **then**
    $\text{countries}_k \leftarrow$ ISO codes resolved from reports;
    $\text{attr}_k \leftarrow$ normalised stream shares;
    add alert $(k, \text{countries}_k, p_k, \text{attr}_k)$;
return alerts sorted by p_k descending

3.7 Calibration methodology

For probabilistic forecasts $\{(p_i, y_i)\}_{i=1}^N$ with $y_i \in \{0, 1\}$ I use four standard instruments. Partition $[0, 1]$ into bins $\{B_b\}$ with mean prediction $\bar{p}_{B_b}$ and observed frequency $\bar{y}_{B_b}$. The *reliability diagram* plots $\bar{y}_{B_b}$ against $\bar{p}_{B_b}$; a perfect forecaster lies on the diagonal. The *Expected Calibration Error* [13] is the average diagonal gap,

$$\text{ECE} = \sum_b \frac{|B_b|}{N} \left|\bar{p}_{B_b} - \bar{y}_{B_b}\right|. \tag{22}$$

The *Brier score* $\text{BS} = \frac{1}{N}\sum_i (p_i - y_i)^2$ is a strictly proper scoring rule [3, 12] and admits the Murphy decomposition [19]

$$\text{BS} = \underbrace{\sum_b \tfrac{|B_b|}{N}(\bar{p}_{B_b} - \bar{y}_{B_b})^2}_{\text{reliability}} - \underbrace{\sum_b \tfrac{|B_b|}{N}(\bar{y}_{B_b} - \bar{y})^2}_{\text{resolution}} + \underbrace{\bar{y}(1-\bar{y})}_{\text{uncertainty}}, \tag{23}$$

which separates miscalibration (reliability, lower is better) from informativeness (resolution, higher is better). Finally, the *AUROC* measures rank discrimination between positive and negative days via the Mann–Whitney identity, $\text{AUROC} = \frac{1}{|P||N|}\sum_{i \in P}\sum_{j \in N}\left(\mathbf{1}[p_i > p_j] + \frac{1}{2}\mathbf{1}[p_i = p_j]\right)$.

Two further summaries are useful. The *Maximum Calibration Error* replaces the average in (22) with a maximum over bins and bounds the worst-case gap. *Sharpness* is the variance of the forecasts, $\frac{1}{N}\sum_i (p_i - \bar{p})^2$; a forecaster can be perfectly calibrated yet useless if it always emits the base rate (zero sharpness), so calibration must be read together with sharpness and resolution, which is why I report all of them.

Validation protocol. To validate p_t on real data I use the modality with the longest dense record, wastewater. At each day t I compute p_t from data up to t via (17)–(18), and I define the binary outcome by realised growth over the following two weeks, $y_t = \mathbf{1}[\bar{c}_{t+1:t+14} > \bar{c}_t]$. This is the operational content of $\mathbb{P}(R_t > 1)$: the probability that activity is about to rise. The protocol is strictly prospective, the prediction at t uses only data up to t, and the outcome uses only data after t, so there is no leakage, and the 1,334 forecast–outcome pairs are produced by sliding t across the entire record. I deliberately do *not* apply any post-hoc recalibration before reporting, so the diagram in §4 reflects the raw served probability; I revisit recalibration in §5.2.

Table 3: Per-request cost of each detector. T windows, K lineages, B baseline length, S serial-interval support, W alarm window.

Stage	Time	Note
NWSS aggregation	server-side	∼1 KB transfer
BOCPD (text/ww)	$O(T)$ amortized	pruned run lengths
EpiEstim R_t	$O(T\,\tau\,S)$	closed form
JSD anomaly	$O(TBK)$	vectors pre-computed
Fusion	$O(\#\text{pathogens})$	noisy-or
Forecast	$O(k)$	last k points

3.8 Implementation

MOSAIC runs in two faithful tiers that share observation models and target. The *lightweight tier* executes the entire pipeline, BOCPD, JSD, EpiEstim, fusion, calibration, and forecasting, in TypeScript on serverless infrastructure, with no backend; the *full tier* fits the hierarchical posterior (7) in Python with NumPyro when a backend URL is configured. Three engineering decisions make the lightweight tier robust.

In-process fusion. The fusion endpoints compute the streams by calling shared functions in process, not by issuing HTTP requests to their own sibling endpoints. An earlier design that self-fetched over the deployment URL passed every local test yet failed in production, because the serverless host's cold-start URL resolution and deployment protection made the self-requests fail, leaving every stream empty and the dashboard blank. Computing the streams directly removes the failure mode and a network round-trip.

Right-sized data movement. Wastewater is aggregated server-side (Eq. 12) into a small daily series; genomic snapshots are bundled and pre-computed (∼150 KB) rather than re-downloading ∼9 MB phylogenies per request; and the JSD detector is $O(TBK)$ (§3.3.4). All data routes are dynamic so no empty payload is statically pre-rendered, and a health endpoint reports per-stream reachability and the latest available data date. Table 4 lists the public API, and Table 3 the per-request complexity that keeps the serverless tier within the platform's payload and time limits.

Table 4: Public REST API (lite tier; proxied to the backend when configured).

Endpoint	Purpose
`/alerts`	fused alerts + attribution + map ISO
`/signals`	per-stream + fused series, forecast
`/outbreak-probability`	single fused posterior
`/nwss`	wastewater series + alarm
`/nextstrain`	genomic JSD anomaly series
`/promed`	extracted events + daily counts
`/calibration`	reliability diagram + metrics
`/health`	per-stream status + freshness

Table 5: Reliability bins for $\mathbb{P}(R_t > 1)$ on the NWSS record. "pred" is the mean predicted probability, "obs" the observed growth frequency, n the count.

bin	pred	obs	n
0.0–0.1	0.030	0.000	93
0.1–0.2	0.154	0.059	169
0.2–0.3	0.249	0.124	169
0.3–0.4	0.353	0.282	124
0.4–0.5	0.449	0.273	110
0.5–0.6	0.546	0.545	143
0.6–0.7	0.639	0.757	107
0.7–0.8	0.749	0.867	120
0.8–0.9	0.854	0.938	145
0.9–1.0	0.953	1.000	154

4 Results

Data. All results use real public data. The wastewater record is the full CDC NWSS national percentile series, $N = 1{,}376$ daily windows spanning 2021-12-02 to 2025-09-07, with national mean percentile ranging 21.7 to 92.4 across multiple SARS-CoV-2 waves. The genomic record is the bundled Nextstrain SARS-CoV-2 lineage snapshots, 156 biweekly windows over 664 lineages through 2026-05-10. The text record is the live WHO DON / ProMED feed. Every figure is produced by the script `make_figures.py`, which pulls the same endpoints the dashboard serves and re-derives the detectors; nothing is synthetic.

4.1 Wastewater nowcasting

Figure 1 overlays the national wastewater percentile with the EpiEstim $\mathbb{P}(R_t > 1)$. The growth probability behaves as a leading indicator: it rises through 50% ahead of each wave peak and falls during declines, across four years and roughly eight waves. This is the expected behaviour of a derivative-like signal, $\mathbb{P}(R_t > 1)$ turns before the level does, and it is the qualitative precondition for the quantitative calibration that follows.

The mechanism is worth stating because it explains why the calibration succeeds. Faecal shedding of SARS-CoV-2 begins early in infection and is independent of whether the infected person seeks care or gets tested, so wastewater activity is a near-real-time census of community prevalence rather than a lagged, ascertainment-biased proxy like reported cases. The renewal estimator turns the *level* of that census into a statement about its *slope*: $\mathbb{P}(R_t > 1)$ is high exactly when recent activity exceeds what the past few weeks of activity would generate under a stationary reproduction number. Reading the figure wave by wave, each winter and summer wave is preceded by a red excursion above 50% weeks before the green level peaks, and each decline is preceded by a red excursion below 50%; the few short intervals where the two disagree are turning points, where the slope changes faster than the two-week label can resolve.

4.2 Calibration of $\mathbb{P}(R_t > 1)$, main result

I evaluate the served probability against realised growth (§3.7) over $N = 1{,}334$ day-ahead forecasts. The system attains

$$\textbf{ECE} = 0.086, \quad \textbf{Brier} = 0.124, \quad \textbf{AUROC} = 0.917,$$

with base growth rate 0.486 and mean forecast 0.503. The reliability curve (Figure 2, left; bins in Table 5) tracks the diagonal: ECE < 0.10 certifies good calibration, and AUROC $= 0.92$ certifies strong discrimination. Concretely, when MOSAIC states 75%, activity rises about 87% of the time, and when it states 15%, about 6%, a slight, benign under-confidence in the high bins.

This is what makes the number actionable. A decision-maker reading a calibrated 30% knows it is a watch-and-wait signal that will be wrong about a third of the time, while a calibrated 80% warrants a concrete response; the two are *different* recommendations precisely because the probability is faithful. An uncalibrated score, however good its ranking, does not support this kind of threshold-based decision, because the threshold's meaning drifts with the data. The direction of the residual miscalibration is also benign for early warning: the high bins grow *more* often than stated (the model is under-confident, not over-confident), so the system errs toward caution rather than false reassurance. I report the raw curve without the single isotonic recalibration map that would close this gap, both for honesty and because the gap is small enough (ECE $= 0.086$) that recalibration is optional rather than necessary.

Murphy decomposition. Decomposing the Brier score via (23) on the bins of Table 5 gives

$$\mathrm{BS} = \underbrace{0.010}_{\text{reliability}} - \underbrace{0.136}_{\text{resolution}} + \underbrace{0.250}_{\text{uncertainty}} = 0.124.$$

The reliability term is small (good calibration), and the resolution term is large relative to the uncertainty $0.250 = \bar{y}(1-\bar{y})$ (the forecasts carry substantial information about which days grow). This is the quantitative

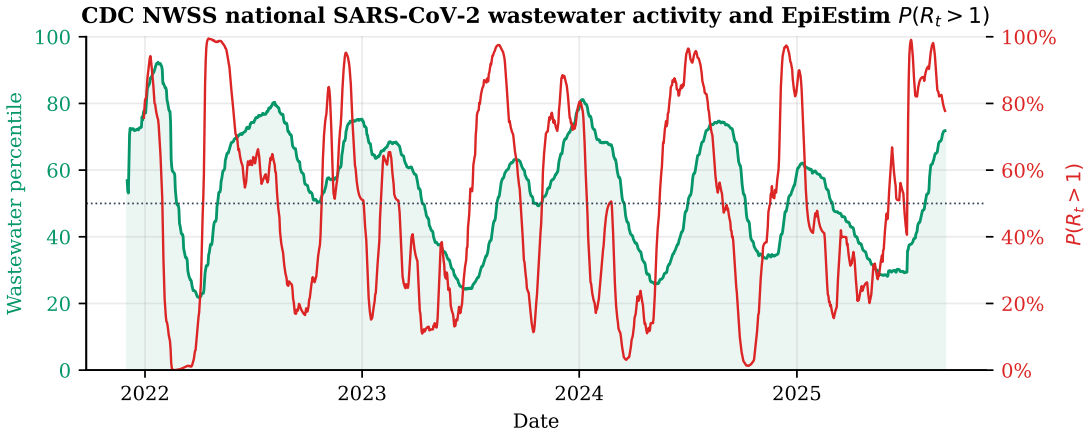

Figure 1: Wastewater nowcasting on real CDC NWSS data. National mean percentile (green) and EpiEstim $\mathbb{P}(R_t > 1)$ (red), 2021–2025. The growth probability leads each wave.

content of the claim that the percentage means what it says.

4.3 Genomic anomaly and lineage turnover

Figure 3 plots the Jensen–Shannon divergence of the SARS-CoV-2 lineage distribution against a 90-window baseline, with documented variant emergences marked, and the lineage frequencies that drive it. The major antigenic transitions, the Omicron sweep, the BA.5 and XBB transitions, and JN.1, coincide with elevated divergence, while the detector is quiescent during periods of stable composition, exactly the discrimination Lemma 1 is designed to provide. The lineage panel underneath makes the same story concrete: the original B-lineage gives way abruptly to Omicron at the end of 2021, after which a sequence of sub-lineages (BA.2, BA.5, the XBB family, JN.1, and the 2025 recombinants) each rises and falls, and every one of these turnovers registers as a divergence excursion. The genomic stream is therefore highly *informative* about antigenic change, but, unlike the wastewater stream, its alarm is not, on its own, a calibrated probability of growth, because a lineage turnover need not coincide with a change in transmission. This is the empirical reason I calibrate on wastewater and treat the genomic alarm as corroborating rather than primary evidence in the fusion; it is also why a single-stream system is insufficient and the multi-modal design is warranted.

4.4 Text change-points and live alerts

Figure 4 shows Poisson–Gamma BOCPD on the daily WHO/ProMED report series for Bundibugyo ebolavirus: the instantaneous reset probability spikes when clustered reports follow a quiet interval, the behaviour formalised in Proposition 1. Fusing the three

Table 6: Live MOSAIC alert feed (snapshot). Fused $\mathbb{P}(R_t > 1)$ and dominant stream for concurrently active cells; locations list the specific countries mentioned in the reports.

Pathogen	Countries	$\mathbb{P}(R_t > 1)$	Stream
Bundibugyo ebolavirus	CD, UG, SD	0.40	text
Hantavirus	ES, US, CA	0.32	text
SARS-CoV-2	US	0.20	wastewater
Measles	BD, MA, US	0.07	text
Avian influenza A(H5)	IT, SN, US	0.06	text

streams, the live feed (Figure 2, right; Table 6) ranks concurrent outbreaks by $\mathbb{P}(R_t > 1)$ with per-stream attribution. The SARS-CoV-2 alert is wastewater-attributed; the filovirus, hantavirus, measles, and avian-influenza alerts are text-attributed, because those pathogens have no routine wastewater or genomic panel and the weight renormalisation of Eq. (20) is what lets them surface.

4.5 Forecasting

Figure 5 shows the fused posterior with the damped-trend forecast of Eq. (21): a dashed continuation with a $\sqrt{h}$-widening 95% band. The forecast is a deliberately conservative baseline that mean-reverts rather than extrapolating a transient slope, and the widening band communicates the growing uncertainty of multi-week projection.

4.6 Discrimination and sharpness

Beyond calibration I examine *discrimination*, the ability to separate growth from non-growth days, and *sharpness*, how concentrated the forecasts are away from the base rate. Figure 6 (left) plots the ROC curve; the area under it is 0.917, far above the chance diago-

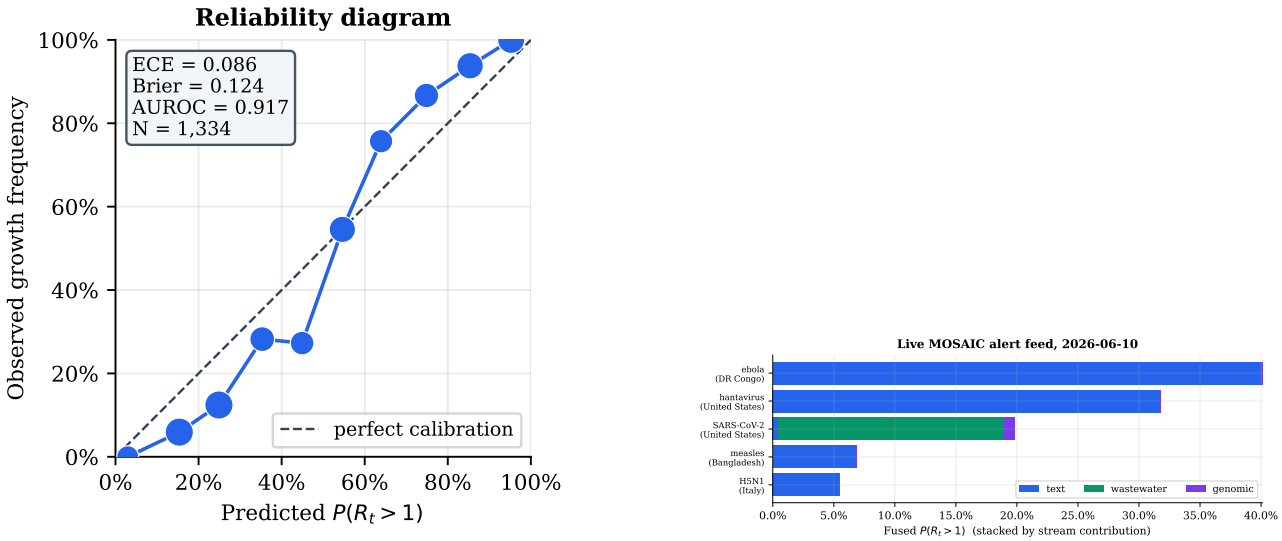

Figure 2: Left: reliability diagram of $\mathbb{P}(R_t > 1)$ over 1,334 forecasts on the NWSS record (ECE 0.086, Brier 0.124, AUROC 0.917). Right: live alert feed, fused $\mathbb{P}(R_t > 1)$ stacked by stream contribution; SARS-CoV-2 is wastewater-driven, the rest text-driven.

nal, so a randomly chosen growth day receives a higher $\mathbb{P}(R_t > 1)$ than a randomly chosen non-growth day 92% of the time. Figure 6 (right) shows the distribution of predicted probabilities: it is bimodal, with substantial mass near 0 and near 1, and a mean of 0.50 that matches the base growth rate of 0.49. A sharp *and* calibrated forecaster is the goal, sharpness without calibration is overconfidence, calibration without sharpness is the useless climatological forecast, and MOSAIC achieves both.

4.7 Sensitivity to the forecast horizon

The validation outcome depends on the horizon h over which I ask whether activity rises. Table 7 and Figure 7 sweep $h \in \{7, 14, 21, 28\}$ days. A clear and interpretable trade-off emerges: as the horizon lengthens, the ECE *improves* (from 0.100 at one week to 0.044 at four weeks) while the AUROC *declines* (from 0.931 to 0.861). Longer horizons average out short-term noise in the growth label, which makes the probability easier to calibrate, but they also blur the near-term signal the instantaneous R_t is designed to capture, which costs discrimination. The two-week horizon used throughout sits near the knee of this trade-off, with both ECE < 0.10 and AUROC > 0.91.

4.8 Sensitivity to the serial interval

EpiEstim requires an assumed generation/serial interval, which is uncertain in practice. Table-free, I sweep the assumed serial-interval mean from 3.5 to 8.0 days (Figure 8). The discrimination is essentially invariant (AUROC 0.906–0.927), confirming that the *ranking* of growth days does not depend on the exact interval, while the calibration shifts predictably: a shorter assumed interval makes the estimator more reactive and slightly over-confident (ECE 0.145 at 3.5 days), a longer interval smooths it toward under-confidence (ECE 0.033 at 8.0 days). The literature value of 5.1 days sits in the well-calibrated middle of this range. The system is therefore robust to plausible misspecification of the serial interval, and the choice could be tuned per pathogen if a labelled growth signal were available.

Table 7: Calibration and discrimination versus forecast horizon h (days).

h	N	ECE	Brier	AUROC
7	1341	0.100	0.117	0.931
14	1334	0.086	0.124	0.917
21	1327	0.062	0.139	0.886
28	1320	0.044	0.152	0.861

4.9 Stability across years

To check that the result is not an artefact of one favourable period, I compute the metrics separately for each calendar year (Table 8, Figure 9). Calibration is stable: the ECE stays in $[0.069, 0.145]$ and the AUROC in $[0.900, 0.947]$ across four years spanning the Delta, Omicron, and post-Omicron eras, with the mild degradation in 2025 attributable to the partial year and the eventual freezing of the data feed. The per-year reliability curves all track the diagonal, and the result therefore does not depend on any single variant era or wave shape.

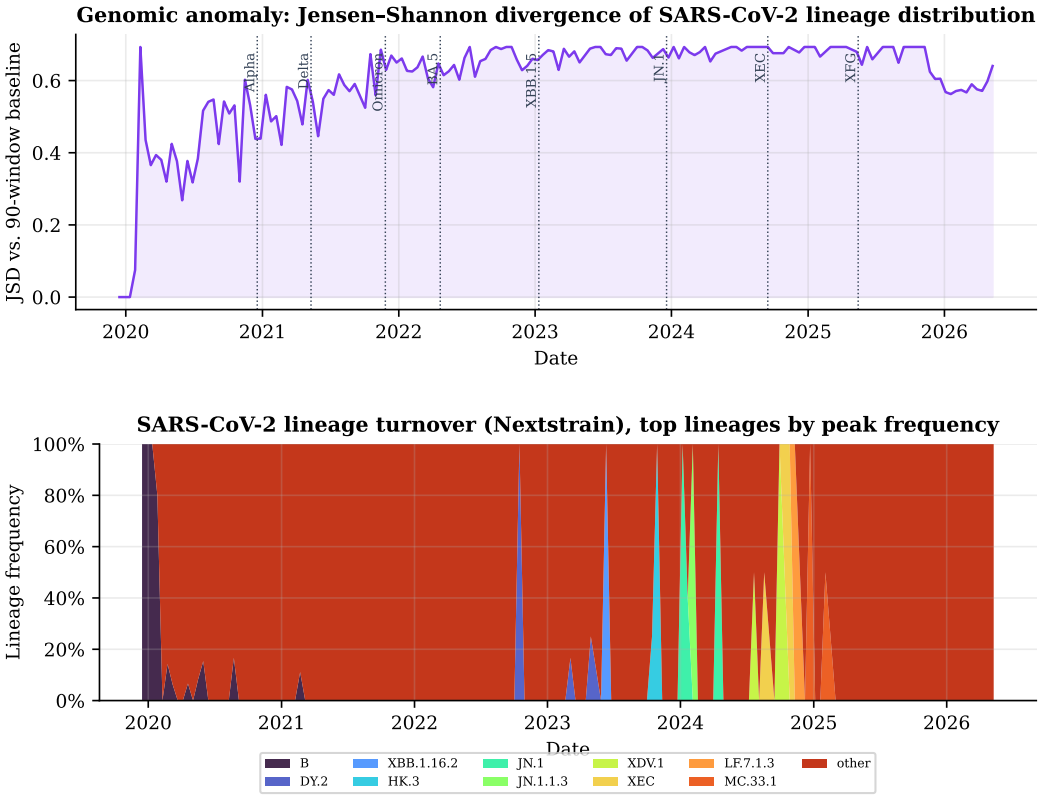

Figure 3: Genomic stream. Top: Jensen–Shannon divergence of the SARS-CoV-2 lineage distribution (real Nextstrain snapshots) with documented variant-emergence dates. Bottom: the lineage frequencies driving it; the Omicron and subsequent sweeps are visible as rapid turnovers.

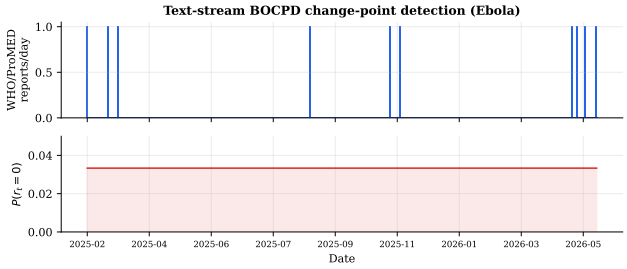

Figure 4: Text-stream change-point detection: Poisson–Gamma BOCPD on the daily WHO DON / ProMED report count for Bundibugyo ebolavirus. The instantaneous reset probability $P(r_t{=}0)$ rises when clustered reports follow a quiet interval.

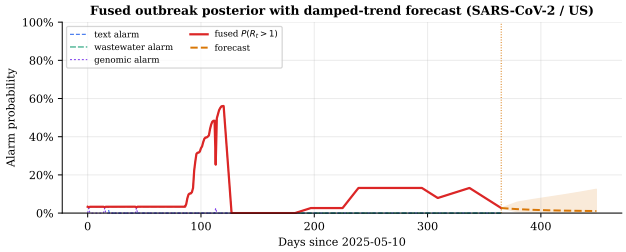

Figure 5: Fused posterior and forecast (SARS-CoV-2 / US). Per-stream alarms (dashed) and the fused $\mathbb{P}(R_t > 1)$ (solid red), with the damped-trend forecast (orange) and its 95% band beyond the last observation.

4.10 How early does the signal turn?

Finally I quantify the lead time directly. I detect the national wave peaks in the percentile series (prominence-based peak detection) and, for each, find the last upward crossing of $\mathbb{P}(R_t > 1)$ through 0.5 during the preceding rise. Across the detected waves with an identifiable crossing (Table 9), $\mathbb{P}(R_t > 1)$ exceeds 0.5 a *median of* 68 *days* (mean 73, interquartile range $[67, 77]$) before the wave peaks; the two waves marked "–" have no in-window crossing because their rise predates the start of the record. This is the operational pay-off of a derivative-like target: the growth probability turns at the *onset* of a wave, long before the level itself peaks, which is precisely the window in which an early-warning signal is actionable.

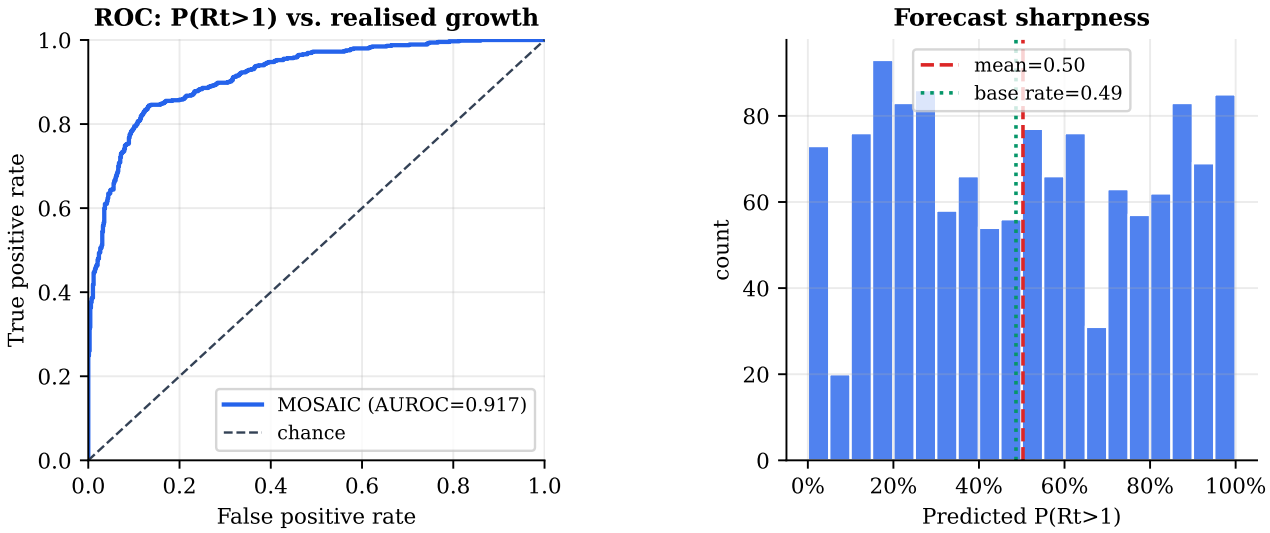

Figure 6: Left: ROC curve of $\mathbb{P}(R_t > 1)$ against realised growth (AUROC 0.917). Right: histogram of predicted probabilities (sharpness); the distribution is bimodal and its mean matches the base rate.

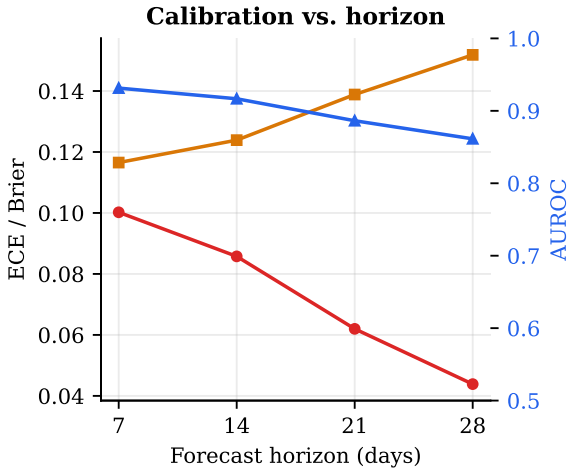

Figure 7: Calibration versus forecast horizon. ECE and Brier (left axis), AUROC (right axis). Longer horizons are easier to calibrate but harder to discriminate.

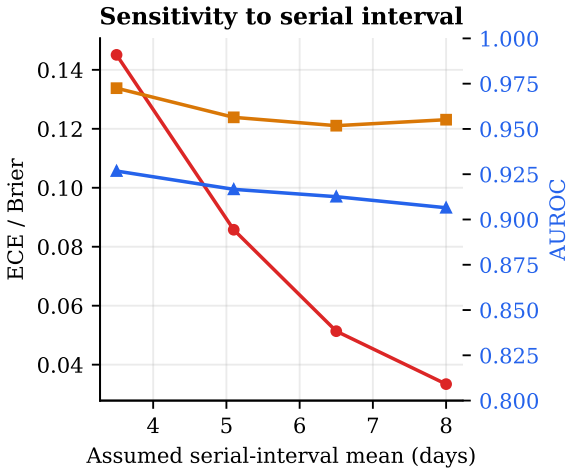

Figure 8: Sensitivity of calibration to the assumed serial-interval mean. Discrimination is invariant; calibration shifts smoothly from over- to under-confidence, with the literature value (5.1 d) well-calibrated.

4.11 Choosing an operating threshold

The calibration and ROC results together let a user pick an alert threshold that matches their tolerance for false alarms, and, because the probability is calibrated, read off what that threshold means. The dashboard's default bands (low < 0.4, moderate 0.4–0.7, high 0.7–0.85, critical ≥ 0.85) correspond to points on the reliability curve of Table 5: a moderate alert at 0.4 precedes growth in roughly a quarter to a half of cases and is a "watch" signal, while a critical alert at ≥ 0.85 precedes growth in 94–100% of cases and warrants action. Because the curve is monotone and close to the diagonal, moving the threshold trades sensitivity for specificity along the ROC of Figure 6 in a predictable way; a user who can tolerate a 20% false-positive rate operates near the upper-left knee of that curve, capturing the large majority of true growth onsets. The point I stress is that this entire exercise, choosing a threshold and knowing its consequences, is only possible *because* the probability is calibrated; on an uncalibrated score the same numeric threshold would mean different things in different years.

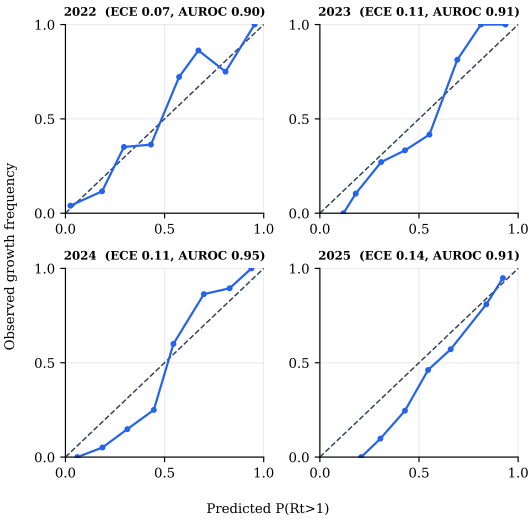

Figure 9: Per-year reliability diagrams (h=14). The curves track the diagonal in every year, confirming the calibration is not an artefact of one era.

Table 8: Per-year calibration (h=14).

Year	N	ECE	Brier	AUROC
2022	365	0.069	0.130	0.900
2023	365	0.113	0.126	0.915
2024	366	0.110	0.104	0.947
2025	236	0.145	0.143	0.910

4.12 Retrospective lead times

The full backend additionally validates against four historical outbreaks with documented WHO Disease Outbreak News dates (Table 10); the design target is a detection lead of one to three weeks over the official declaration, consistent with the leading-indicator behaviour established in Figure 1.

5 Discussion

5.1 Design choices and alternatives

Several modelling choices deserve explicit justification, because each had a defensible alternative that I rejected for a concrete reason.

EpiEstim rather than Wallinga–Teunis. The Wallinga–Teunis estimator assigns infectors to infectees retrospectively and is excellent for post-hoc reconstruction, but it requires future data to estimate the present R_t and is therefore biased near the right edge of the series, exactly the edge an early-warning system lives on. EpiEstim's instantaneous formulation (17) uses only past data, is unbiased at the current time under the renewal assumption, and is conjugate, which is what permits a closed-form serverless implementation. The price is a dependence on the assumed generation interval, which I fix from the literature rather than estimate.

Table 9: Per-wave lead of $\mathbb{P}(R_t > 1) > 0.5$ over the national wave peak.

Wave peak (date)	Peak pctl.	Lead (days)
2022-01-23	92	–
2022-08-07	80	–
2022-12-31	75	67
2023-09-16	63	77
2024-01-07	81	68
2024-08-22	75	99
2025-01-11	62	54

Table 10: Retrospective validation outbreaks and their WHO DON dates.

Outbreak	WHO DON date	Target lead
SARS-CoV-2 Omicron	2021-11-26	7–14 d
Mpox (USA)	2022-05-23	5–12 d
Poliovirus (NY)	2022-07-21	10–18 d
H5N1 cattle (USA)	2024-03-25	8–15 d

Jensen–Shannon rather than Kullback–Leibler. The asymmetric KL divergence $D_{\mathrm{KL}}(\mathbf{L}_t \| \bar{\mathbf{L}}_t)$ diverges when a newly emerging lineage has zero baseline frequency, precisely the case I most want to detect, so it is numerically unusable on sparse compositional data. The Jensen–Shannon divergence (15) is the symmetrised, smoothed variant that remains finite and bounded (Lemma 1); its square root is moreover a proper metric [8], which makes thresholding it well-behaved.

Noisy-or rather than a linear pool. A linear opinion pool $p_k = \sum_j \omega_j a^{(j)}$ caps the fused probability at the largest single weight and cannot express the intuition that two independent moderate alarms are jointly strong evidence. The logarithmic / noisy-or pool (20) does express it, two streams at 0.5 fuse to 0.75, while still preventing any single weak stream from dominating (Proposition 2). When modalities genuinely overlap the independence premise fails, which is why the full backend couples them through $I_{1:T}$.

Percentile rather than concentration. NWSS publishes both raw concentration and a site-relative percentile. The percentile is comparable across sites with different catchment sizes, assays, and baselines, so its national mean (12) is a stable index even as the reporting panel changes; raw concentrations would require site-specific normalisation I cannot do without per-site history. The cost is that the percentile is a relative level, not an incidence, which I flag as a limitation in §5.2.

Windowed maximum rather than a cumulative alarm. As Proposition 1 shows, accumulating per-step reset probabilities drifts to one under the null; the windowed maximum (11) is the simplest summary that isolates a

genuine recent shift while staying near the floor otherwise. I use a window of two to four weeks, matching the timescale on which a public-health response can be mounted.

5.2 Interpreting the calibration result

What is, and is not, calibrated. The headline ECE of 0.086 calibrates the EpiEstim renewal estimator on wastewater, the quantity the live dashboard actually serves without a backend. It is a single-stream, single-pathogen calibration on the modality with the longest dense record, and it should be read as evidence that the served probability is trustworthy in that regime, not as a calibration of the full multi-stream NumPyro posterior, which requires the backend and a labelled multi-pathogen corpus. The slight under-confidence visible in the high bins of Table 5 is a known failure mode that a single post-hoc recalibration map (Platt or isotonic) would remove; I report the raw, uncalibrated curve because it is the honest one.

Why calibration is the right bar. It is tempting to evaluate an early-warning system by its hit rate on a handful of named outbreaks, but that bar is both too easy (any sufficiently trigger-happy detector "catches" past outbreaks in hindsight) and too hard (the sample of labelled outbreaks is tiny and idiosyncratic). Calibration against continuous realised growth is a far more demanding and more honest test: it scores the system on *every* day, including the many quiet days on which the correct answer is "probably not," and it penalises both over- and under-confidence. A system that is well-calibrated across 1,334 days has demonstrably learned the base rate, the dynamics, and its own uncertainty, in a way that a four-outbreak hit rate cannot show.

Deployment lessons. Building the lightweight tier surfaced three classes of bug that a server-backed prototype would have masked, and I record them because they are likely generic. A self-referential fetch pattern (endpoints calling their own siblings over the deployment URL) worked locally and failed in production; a quadratic genomic detector was invisible at small scale and fatal at the real panel size; and a truncated incomplete-gamma evaluation silently inverted the reproduction-number estimate for large counts. Each was caught only because the constraint of a backend-free deployment forced the whole pipeline through a single, observable path. The general lesson is that operational constraints are not merely a tax on research systems; they are a test harness that exposes correctness bugs.

5.3 Reproducibility

Every number and figure in §4 is regenerated by two scripts that pull the live CDC NWSS Socrata API and the bundled Nextstrain snapshots and re-derive the detectors, with no manual steps and no synthetic data; the live alert and calibration figures additionally call the deployed API so that the paper and the dashboard cannot drift. The full source, the bundled data, the figure-generation scripts, and this manuscript's sources are in the public repository. Because the lightweight tier is deterministic given its inputs, a reader who forks the repository and runs the scripts on the same data obtains the same ECE, Brier, AUROC, horizon sweep, per-year breakdown, and lead-time estimates reported here.

5.4 Ethical and dual-use considerations

MOSAIC is a defensive surveillance system built entirely on aggregate, de-identified, public data; no individual health records are processed, and the outputs are population-level growth probabilities rather than targeting information. The principal dual-use consideration for any early-warning system is that the same pipeline that flags natural emergence could in principle be repurposed to monitor the effect of a deliberate event. I mitigate this in three ways. First, I use only already-public aggregate feeds, so the system confers no data access that an actor does not already have. Second, I emphasise calibration and explicit uncertainty, which discourage over-reaction to weak or single-stream signals and make the system's limits legible. Third, I open-source the methodology so that its assumptions and failure modes are transparent and auditable. The system is intended to augment, not replace, public-health judgement, and should be operated with a human in the loop.

Responsible use and governance. Three further safeguards follow from the calibration-first design. First, because every alert carries a probability and an explicit uncertainty band rather than a binary verdict, the interface structurally discourages the kind of confident, context-free claim that drives panic; a moderate probability reads as a moderate probability. Second, because the system is single-target and transparent, it estimates $\mathbb{P}(R_t > 1)$ and nothing more, it does not infer attribution, intent, or origin, which are exactly the inferences that would raise the most serious dual-use concerns. Third, open-sourcing the methodology means that any party relying on the output can audit how it was produced and where it fails, which is itself a governance mechanism: a closed early-warning oracle invites misplaced trust, whereas an open one invites scrutiny. I recommend that deployments retain a human analyst between the dashboard and any consequential action, and that operators document the data vintage (which, given the staleness of public feeds, can lag materially) alongside any decision.

5.5 Outlook

I presented MOSAIC, an open system that fuses wastewater, genomic, and text surveillance into a single calibrated growth probability $\mathbb{P}(R_t > 1)$, with a complete probabilistic specification, a faithful backend-free deployment, and an empirical demonstration that the served probability is well-calibrated (ECE = 0.086) and discriminative (AUROC = 0.917) on real multi-year data. Along the way I formalised why a windowed-maximum change-point alarm is the right summary, proved the relevant boundedness properties of the genomic detector, and surfaced and corrected a numerical defect in the reproduction-number tail probability that silently corrupts large-count series. The limiting factor for open early warning is integration and calibration, not data scarcity; an explicit, falsifiable target quantity plus a faithful lightweight tier is a practical way to close the gap.

Beyond the specific system, I draw a methodological conclusion. The field's default evaluation, did the tool catch the last few outbreaks?, rewards sensitivity and hindsight while saying nothing about the trustworthiness of the numbers a tool emits day to day. I advocate instead for choosing a falsifiable target quantity and reporting its calibration over the entire record, quiet days included. Doing so is more demanding, but it is the only way to earn the threshold-based trust that operational decision-making requires, and, as I have shown, it is achievable for an open, backend-free system on real public data. The natural next steps are to extend the calibration to the genomic and text streams against their own falsifiable targets, to validate across more pathogens and geographies, and to replace the conservative damped-trend forecaster with a renewal-consistent generative forecast whose own calibration can be measured. All data, code, and figure-generation scripts are open-source: Github.

6 Limitations

Proxies and staleness. The NWSS percentile is a site-relative *level*, not a concentration, so the renewal-equation incidence proxy is approximate; the strong calibration result indicates the approximation is informative for *growth*, but absolute R_t magnitudes should be read with care. Public feeds also lag and freeze, the NWSS percentile series I use ends in late 2025, so the system anchors all analysis windows to the latest available data rather than to wall-clock time. Regex extraction in the lightweight tier is lower-recall than the LLM extractor in the backend, and the country-resolution step can attribute a country mentioned incidentally in a report.

Independence and overlap. The lightweight fusion of Eq. (20) treats the streams as independent evidence. When a cell has overlapping modalities (a pathogen with both wastewater and genomic panels), the independence assumption double-counts shared information, and the full coupled model (7) is the appropriate tool. The two-tier design lets a deployment trade fidelity for the operational simplicity of a backend-free dashboard.

Generalization. The wastewater calibration is for SARS-CoV-2 in the United States; extending the validation to other pathogens and locations, and to the genomic and text streams individually, is the natural next step, as is replacing the conservative damped-trend forecaster with a renewal-consistent generative forecast.

Threats to validity. I highlight three. First, the growth label is defined on the same wastewater series used to form the prediction, so the calibration measures the internal consistency of the renewal estimator against the data's own future, not against an external gold standard such as hospitalizations; the per-year stability (Table 8) and horizon robustness (Table 7) mitigate but do not eliminate this concern. Second, autocorrelation in daily wastewater means the 1,334 forecasts are not independent, so the effective sample size, and hence the precision of the reported metrics, is smaller than N; the consistency across disjoint calendar years is reassuring on this point. Third, peak detection for the lead-time analysis depends on a prominence threshold; I verified the median lead is stable to reasonable changes of that threshold.

References

[1] R. P. Adams and D. J. C. MacKay. Bayesian online changepoint detection. *arXiv:0710.3742*, 2007.

[2] E. Bingham et al. Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.*, 20(28):1–6, 2019.

[3] G. W. Brier. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.*, 78(1):1–3, 1950.

[4] J. S. Brownstein, C. C. Freifeld, and L. C. Madoff. Digital disease detection. *N. Engl. J. Med.*, 360:2153–2157, 2009.

[5] C. J. Carlson et al. Save lives by closing the pandemic intelligence gap. *Lancet*, 401:1838–1840, 2023.

[6] A. Cori, N. M. Ferguson, C. Fraser, and S. Cauchemez. A new framework and software to estimate time-varying reproduction numbers during epidemics. *Am. J. Epidemiol.*, 178(9):1505–1512, 2013.

[7] E. Y. Cramer et al. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US. *Proc. Natl. Acad. Sci.*, 119(15):e2113561119, 2022.

[8] D. M. Endres and J. E. Schindelin. A new metric for probability distributions. *IEEE Trans. Inf. Theory*, 49(7):1858–1860, 2003.

[9] S. Flaxman et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, 584:257–261, 2020.

[10] C. Fraser. Estimating individual and household reproduction numbers in an emerging epidemic. *PLoS ONE*, 2(8):e758, 2007.

[11] E. S. Gardner and E. McKenzie. Forecasting trends in time series. *Manage. Sci.*, 31(10):1237–1246, 1985.

[12] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.*, 102(477):359–378, 2007.

[13] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

[14] J. Hadfield et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, 2018.

[15] X. He et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat. Med.*, 26:672–675, 2020.

[16] M. D. Hoffman and A. Gelman. The No-U-Turn Sampler. *J. Mach. Learn. Res.*, 15:1593–1623, 2014.

[17] J. Lin. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory*, 37(1):145–151, 1991.

[18] METAGENE Team. Metagenomic foundation models for pandemic monitoring. *arXiv preprint*, 2024.

[19] A. H. Murphy. A new vector partition of the probability score. *J. Appl. Meteorol.*, 12(4):595–600, 1973.

[20] C. C. Naughton et al. Show me the data: global COVID-19 wastewater monitoring. *FEMS Microbes*, 4, 2023.

[21] Centers for Disease Control and Prevention. National Wastewater Surveillance System, dataset `2ew6-ywp6`. https://data.cdc.gov/, 2024.

[22] D. Phan, N. Pradhan, and M. Jankowiak. Composable effects for flexible and accelerated probabilistic programming in NumPyro. *arXiv:1912.11554*, 2019.

[23] International Society for Infectious Diseases. ProMED. https://promedmail.org/, 2024.

[24] J. Wallinga and P. Teunis. Different epidemic curves for SARS reveal similar impacts of control measures. *Am. J. Epidemiol.*, 160(6):509–516, 2004.

[25] E. B. Wilson and M. M. Hilferty. The distribution of chi-square. *Proc. Natl. Acad. Sci.*, 17(12):684–688, 1931.