Improving inference in wastewater-based epidemiology by modelling the statistical features of digital PCR

Adrian Lison[1,2*], Timothy R. Julian[3,4,5], and Tanja Stadler[1,2]

[1]ETH Zurich, Department of Biosystems Science and Engineering, Zurich, Switzerland

[2]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland

[3]Eawag, Swiss Federal Institute of Aquatic Science and Technology, 8600 Dubendorf, Switzerland

[4]Swiss Tropical and Public Health Institute, 4123 Allschwil, Switzerland

[5]University of Basel, 4055 Basel, Switzerland

[*]Corresponding author: adrian.lison@bsse.ethz.ch

Abstract

The growing field of wastewater-based infectious disease surveillance relies on the quantification of pathogen concentrations in wastewater using polymerase chain reaction (PCR) techniques. However, existing models for monitoring pathogen spread using wastewater have often been adapted from methods for case count data and neglect the statistical features of PCR techniques. In this paper, we seek to overcome the widespread simplistic modelling of wastewater PCR measurements as normally or log-normally distributed by proposing an appropriate model for digital PCR (dPCR). Building on established statistical theory of dPCR, we derive approximations for the coefficient of variation of measurements and the probability of non-detection and propose a hurdle model-based likelihood for estimating concentrations from dPCR measurements. Using simulations and real-world data, we show that simple likelihoods based on normal or log-normal distributions are misspecified, affecting the estimation of pathogen concentrations and infection trends over time. In contrast, the proposed dPCR-specific likelihood accurately models the distribution of dPCR measurements, improving epidemiological estimates and forecasts even if details of the laboratory protocol are unknown. The method has been implemented in the open-source R package "EpiSewer" to improve wastewater-based monitoring of pathogens.

Keywords: Wastewater-based epidemiology, digital PCR, Limit of Detection, Limit of Quantification, infectious disease surveillance, environmental surveillance

1 Introduction

Wastewater-based epidemiology has become a popular approach to studying the community-level spread of pathogens, including SARS-CoV-2, Influenza A and B virus, respiratory syncytial virus, monkeypox virus, and antimicrobial-resistant bacteria [1–9]. A major strength of wastewater-based pathogen surveillance is that it allows the cost-effective monitoring of infection dynamics in a large population, independent of clinical testing and from only a small number of wastewater samples per week [10]. If genetic material of a pathogen is shed into the sewerage system by infected individuals in sufficiently large quantities, its concentration in wastewater samples, often reported in "gene copies per mL", can be quantified using polymerase chain reaction (PCR) techniques [11]. Moreover, the influent wastewater received by municipal treatment plants represents the combined effluent of many households, acting as a pooled sample. Thus, gene concentrations measured at a treatment plant, when normalised by wastewater flow or shedding markers, are somewhat representative of the pathogen load shed by the population in the associated catchment area. This allows the monitoring of transmission dynamics and estimation of epidemiological parameters such as the effective reproduction number from wastewater measurements over time [1].

To analyse a time series of wastewater concentration measurements, various statistical approaches such as non-parametric smoothing [12,13], deconvolution [1,14], and latent variable modelling [15–18] have been proposed. These approaches are often based on existing models originally developed for epidemiological count data (e. g. confirmed cases or hospitalisations) and have been adapted to account for the kinetics of pathogen shedding into wastewater. This analogy is justified for many infectious diseases because clinical cases and pathogen loads in wastewater arise from the same underlying infection dynamics, and have been shown to correlate [4, 19]. However, current methods often analyse PCR measurements from wastewater samples like case data (e. g. by explicitly or implicitly treating them as Poisson distributed case counts [1, 14]) or by modelling them as normally [12, 15], Student's t [16] or log-normally [5, 13, 17] distributed. These simplifications ignore that PCR measurements from environmental samples have distinct statistical properties that influence the bias and error structure of observations. However, such statistical properties are particularly important in wastewater surveillance, where the spread of pathogens in a population is monitored from only a few measurements per week.

The most widespread techniques for quantifying genetic targets in a sample are quantitative PCR (qPCR) and digital PCR (dPCR). While qPCR quantification relies on a comparison of replication cycle thresholds with a standard curve established from reference samples, dPCR works by partitioning a sample into thousands of independent nanoscale chambers or droplets prior to PCR amplification [20]. Using the ratio of positive partitions to the total number of partitions, the concentration in the sample can be measured [21]. Due to its independence from standard curves, greater robustness and repeatability, improved handling of inhibition, and increased potential for multiplexing, dPCR is becoming an increasingly popular method for quantifying molecular targets [22, 23].

In this paper, we develop an observation model for dPCR measurements from wastewater samples that accounts for the statistical properties of digital PCR. Using an established model for the number of positive partitions in a dPCR experiment, we derive estimates for the coefficient of variation of the measured concentrations and the probability of non-detection (i. e. a zero measurement). Unlike previous work, we here account for noise through extraction, preprocessing and other sources of variation, which can be substantial in the case of wastewater samples. Our derived estimates allow us to propose a likelihood function for inference of pathogen concentrations from a time series of reported dPCR measurements. Using data from two wastewater treatment plants in Switzerland, we demonstrate how our approach improves model fit, inferred pathogen loads, and estimates of epidemiological parameters such as the effective reproduction number.

2 Model

2.1 Partitioning statistics in digital PCR

Statistics in digital PCR are based on a mathematical model of the number of gene copies found in each partition of the PCR. This number depends on the expected concentration λ in the analysed sample and the volume of each partition v. We account for the higher concentration of the target in the PCR compared to the original sample due to an extraction step (which may, however, be partly counteracted by dilution and suboptimal recovery efficiency [24]) via a scaling factor s, i. e. the concentration in the PCR is $s\lambda$. We note that s is a constant factor in our model and that variation in dilution and recovery efficiency will later be modelled through a pre-PCR coefficient of

variation. The gene copies are typically assumed to be uniformly distributed in the sample and the number of copies in each of the m partitions is modelled via a Poisson process [25], where partition $j \in \{1, \ldots, m\}$ has N_j copies with

$$N_j \sim \text{Pois}(\lambda c). \tag{1}$$

Here, $c = sv$ is a "conversion factor" that expresses how many expected gene copies per partition one concentration unit in the original sample corresponds to. The PCR in each partition generates a fluorescent signal if one or several gene copies are present, which is used to obtain a binary result (positive or negative) for each partition. Under the Poisson model, the probability of a positive result in partition j, indicated by $Z_j = 1$, is thus

$$Pr(Z_j = 1) = Pr(N_j > 0) = 1 - Pr(N_j = 0) = 1 - \exp(-\lambda c). \tag{2}$$

Moreover, since the PCR in each partition can be regarded as independent from the PCRs in the other partitions, the total number of partitions with a positive result Y is binomially distributed with

$$Y \sim \text{Binom}(m, 1 - \exp(-\lambda c)). \tag{3}$$

Under this model, Dorazio and Hunter [21] show that the maximum likelihood estimate for the concentration λ given an observed number of positive partitions $Y = y$ is

$$\tilde{\lambda} = -\frac{1}{c} \log\left(1 - \frac{y}{m}\right). \tag{4}$$

It is furthermore common to run n technical replicates $i \in \{1, \ldots, n\}$ of the same sample to assess technical variation and reduce overall uncertainty. For reporting, however, replicate measurements are typically combined into a single estimate. In this case, Dorazio and Hunter show that the combined maximum likelihood estimate is $\hat{\lambda} = -\frac{1}{c} \log\left(1 - \frac{\bar{y}}{\bar{m}}\right)$, i. e. y and m are replaced with the mean number of positive and total partitions over all replicates, respectively [21]. Importantly, this estimate differs from the arithmetic mean of individual concentration estimates $\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} \tilde{\lambda}_i$, which

is however also used in practice (see electronic supplementary material B for a comparison of both approaches). In the following, we assume an equal number of partitions $m_i = m$ for all replicates, in which case our results are valid for both approaches. Naturally, by letting $n = 1$, our model also applies to single measurements.

In the context of wastewater surveillance, $\hat{\lambda}$ is reported by laboratories to indicate the measured target concentration in wastewater on a given day. Using a time series of such dPCR measurements as observations, models for wastewater-based epidemiology try to infer trends in pathogen shedding and infections in the respective population. Since the individual dPCR measurements are subject to measurement noise, however, this requires an adequate error model. In the following, we develop such a model by deriving the coefficient of variation and probability of non-detection of dPCR measurements and proposing a suitable likelihood function for $\hat{\lambda}$.

2.2 Coefficient of variation

We regard the dPCR measurement $\hat{\lambda}$ as a random variable itself and therefore denote it by $\hat{\Lambda}$. Assuming a true concentration λ, and using the previously introduced model of positive partition counts Y, it can be shown that the asymptotic variance of $\hat{\Lambda}$ is

$$\mathrm{Var}[\hat{\Lambda}] = \frac{1}{c^2 mn}\left(\exp(\lambda c) - 1\right), \tag{5}$$

see electronic supplementary material C.1 for details. Moreover, using the fact that $\hat{\Lambda}$ is an asymptotically unbiased estimator of λ [25], we can estimate the coefficient of variation (CV) of the measurements as

$$\nu_{\hat{\Lambda}}(\lambda) = \frac{1}{c\sqrt{mn}} \frac{\sqrt{\exp(\lambda c) - 1}}{\lambda}. \tag{6}$$

An important implication of Eq. (6) is that the CV of measurements is concentration-dependent (see Figure 1). However, this estimate of the CV assumes that the concentration in the PCR is exactly proportional to the original concentration of interest λ. To account for extraction and preprocessing noise and other sources of variation, we here model the unscaled concentration in the PCR as Λ_{pre}, which is a random variable drawn from a parametric distribution with mean λ and coefficient of variation $\nu_{\mathrm{pre}} > 0$. That is, the parameter ν_{pre} describes random variation before the PCR. As shown

in electronic supplementary material C.2, the variance of $\hat{\Lambda}$ under this extended model is

$$\mathrm{Var}[\hat{\Lambda}] = \lambda^2 \nu_{\mathrm{pre}}^2 + \frac{1}{c^2 mn} \left(\mathrm{E}[\exp(\Lambda_{\mathrm{pre}} c)] - 1\right), \tag{7}$$

where $E[\exp(\Lambda_{\mathrm{pre}} c)]$ can be obtained from the moment-generating function (MGF) of Λ_{pre}. For example, if Λ_{pre} is gamma distributed, then

$$E[\exp(\Lambda_{\mathrm{pre}} c)] = \left(1 - \lambda c \nu_{\mathrm{pre}}^2\right)^{-\frac{1}{\nu_{\mathrm{pre}}^2}}, \tag{8}$$

yielding a coefficient of variation

$$\nu_{\hat{\Lambda}}(\lambda) = \sqrt{\nu_{\mathrm{pre}}^2 + \frac{1}{\lambda^2 c^2 mn} \left(\left(1 - \lambda c \nu_{\mathrm{pre}}^2\right)^{-\frac{1}{\nu_{\mathrm{pre}}^2}} - 1 \right)}. \tag{9}$$

In electronic supplementary material C.2 we also provide an approximation for the case where Λ_{pre} is log-normally distributed, and use a Taylor series expansion to approximate $\nu_{\hat{\Lambda}}(\lambda)$ if only its mean and variance are known.

2.3 Probability of non-detection

From the binomial model of positive partitions counts described in Eq. (3), it follows that even for non-zero concentrations, the PCR can result in zero positive partitions. If this happens in all technical replicates, the target of interest is not detected and the maximum likelihood estimate for the concentration is $\hat{\lambda} = 0$. Without accounting for pre-PCR variation, the probability for such a non-detection follows directly from the binomial model, i. e.

$$p_{\mathrm{zero}}(\lambda) = \prod_{i=1}^{n} \Pr(\hat{\Lambda} = 0 | \lambda) = \prod_{i=1}^{n} \exp(-\lambda c m) = \exp(-\lambda c m n). \tag{10}$$

This probability is closely related to the Limit of Detection (LoD) of an assay, which is typically defined as the lowest concentration λ_{LOD} at which detection is feasible with sufficient confidence α [26]. Leaving the possibility of false positive partitions aside, we note that $\alpha = 1 - p_{\mathrm{zero}}(\lambda_{\mathrm{LOD}}) = 1 - \exp(-\lambda_{\mathrm{LOD}} c m n)$, hence λ_{LOD} sufficiently determines the non-detection probability under Eq. (10) for arbitrary concentrations (see electronic supplementary material D.1 for further discussion).

To account for pre-PCR variation as in Section 2.2, we compute the marginal probability of

non-detection, which can be expressed as

$$p_{\text{zero}}(\lambda) = \mathrm{E}_{\Lambda_{\text{pre}}}\left[\prod_{i=1}^{n} \Pr(\hat{\Lambda} = 0 | \Lambda_{\text{pre}} = \lambda_{\text{pre}})\right] = \mathrm{E}[\exp(-\Lambda_{\text{pre}} cmn)]. \tag{11}$$

Similar to Eq. (7), the expectation $\mathrm{E}[\exp(-\Lambda_{\text{pre}} cmn)]$ can be obtained using the MGF of Λ_{pre}, and we provide solutions for the case where Λ_{pre} is gamma or log-normally distributed in electronic supplementary material D.2.

2.4 Likelihood function

Based on the concentration-dependent noise and probability of non-detection in digital PCR as derived above, we propose the following likelihood for a dPCR measurement $\hat{\lambda}$:

$$L_{\hat{\Lambda}}(\hat{\lambda}|\lambda) = \begin{cases} p_{\text{zero}}(\lambda), & \text{if } \hat{\lambda} = 0 \\ f_{\text{PCR}}(\hat{\lambda})\,(1 - p_{\text{zero}}(\lambda)), & \text{otherwise} \end{cases}, \tag{12}$$

where $p_{\text{zero}}(\lambda)$ is the probability of non-detection as defined in Eq. (11) and f_{PCR} is the density function of a positive continuous probability distribution with conditional mean

$$\mathrm{E}[\hat{\Lambda}|\hat{\Lambda} > 0] = \frac{\lambda}{1 - p_{\text{zero}}(\lambda)} \tag{13}$$

and conditional coefficient of variation

$$\mathrm{CV}[\hat{\Lambda}|\hat{\Lambda} > 0] = \sqrt{\nu_{\hat{\Lambda}}(\lambda)^2 (1 - p_{\text{zero}}(\lambda)) - p_{\text{zero}}(\lambda)}. \tag{14}$$

That is, f_{PCR} models the distribution of non-zero measurements by adjusting the expected concentration λ and the coefficient of variation $\nu_{\hat{\Lambda}}(\lambda)$ (see Eq. (9)) for the conditioning on a non-zero outcome (see electronic supplementary material F.1 for details). The likelihood defined in Eq. (12) thus corresponds to a hurdle model in which concentrations can be measured as zero with a certain probability, and otherwise follow the specified continuous probability distribution f_{PCR}. We note that while $\hat{\lambda}$ is a real number, its distribution is not continuous due to the underlying count nature of the digital PCR. Therefore, the use of a continuous distribution for $f_{\text{PCR}}(x)$ is an approximation of the

true distribution, with the advantage of being computationally more efficient than a back-calculated binomial likelihood and having an unbiased mean even if the assumed size or number of partitions is unknown or misspecified (see electronic supplementary material F.2). We here choose a gamma distribution for $f_{\mathrm{PCR}}(x)$, as suggested by a simulation-based comparison of different continuous distributions (see electronic supplementary material F.3).

3 Inference from dPCR measurements

Given a set of k dPCR measurements $\hat{\lambda}_{i\in\{1,\dots,k\}}$, we can use the likelihood proposed in Eq. (12) to estimate the underlying concentration λ via Bayesian inference. Specifically, given a prior $P(\lambda)$ and our likelihood $P(\hat{\lambda}_{i\in\{1,\dots,k\}}|\lambda) = \prod_{i=1}^{k} L_{\hat{\Lambda}}(\hat{\lambda}_i|\lambda)$, we can sample from the posterior distribution $P(\lambda|\hat{\lambda}_{i\in\{1,\dots,k\}})$ via Markov Chain Monte Carlo (MCMC). We implemented a corresponding Bayesian model using the probabilistic programming language "stan" [27] that can estimate the concentration in a single sample from a set of dPCR measurements (see electronic supplementary material G.2 for details on the estimation). In wastewater-based epidemiology, however, the goal is often to infer a time series of concentrations $\lambda_{\{t|t\leq T\}}$ and other epidemiological variables from longitudinal measurements $\hat{\lambda}_{\{t|t\leq T,\, i\in\{1,\dots,k\}\}}$, where t is a time index and i is a replicate index. For this purpose, the prior $P(\lambda)$ of our previous model can be extended to a time series prior $P(\lambda_{\{t|t\leq T\}})$ that reflects the autocorrelation of concentrations in wastewater and infection trends. To obtain estimates of the concentration over time, we can then sample from the posterior distribution of concentrations $P(\lambda_{\{t|t\leq T\}}|\hat{\lambda}_{\{t|t\leq T,\, i\in\{1,\dots,k\}\}})$ via MCMC as described above. Depending on the application, the prior $P(\lambda_{\{t|t\leq T\}})$ can be a simple non-parametric smoothing prior, e. g. a random walk model, or a more complex hierarchical prior that models further quantities of interest. For the example application in this paper, we use an epidemiological model implemented in the R package "EpiSewer" [28] that accounts for the shedding load kinetics of infected individuals in the catchment and can estimate the effective reproduction number over time.

To apply Eq. (12), we in principle require knowledge of the PCR parameters (number of partitions m, scaling factor s, partition volume v) and the pre-PCR coefficient of variation ν_{pre}. While these quantities could be derived from protocols and lab experiments, they may be unknown in many realistic settings, e. g. when modelling publicly reported concentration measurements without sufficient metadata. In practice, we thus provide broad priors for these parameters and

estimate them jointly with $\lambda_{\{t|t\leq T\}}$ using only the measurement data. To specify a prior for the number of partitions m in the assay, we take into account that m varies between PCR runs due to random loss of partitions. Modelling the number of lost partitions as log-normally distributed, we parameterize m via a hierarchical prior with hyperpriors for the mean number of partitions μ_m and coefficient of variation of partitions ν_m (see electronic supplementary material G.1). For μ_m, we use a $\text{Normal}^+(\mu = 20000, \sigma = 5000)$ prior to accommodate dPCR systems with 10000 to 30000 partitions. For ν_m, we use a $\text{Normal}^+(\mu = 0, \sigma = 0.05)$ prior, reflecting our expectation of less than 10% variation in the number of partitions. Since the scaling factor s and partition volume v are not individually unidentifiable, we specify a prior for the "conversion factor" factor $c = sv$. In our application, we place a $\text{Normal}^+(\mu = 1 \times 10^{-5}, \sigma = 4 \times 10^{-5})$ prior on c to reflect that extraction and preprocessing typically increases target concentrations in the assay by less than a factor of 100 and that partition volumes in digital droplet PCR are typically in the nanoliter range (see electronic supplementary material G.1 for explanation). The prior for ν_{pre} can be a standard regularizing prior, such as $\nu_{\text{pre}} \sim \text{Normal}^+(0, 1)$.

4 Evaluation

4.1 Model validation

We used simulations of the positive partition counts according to Eq. (1)–(4) under different types of pre-PCR noise to assess the accuracy of our approximations for the coefficient of variation (CV) of dPCR measurements and the probability of non-detection. For our simulations, we assumed exemplary PCR parameters based on the Naica Crystal Digital PCR Platform with Sapphire Chips (Stilla Technologies, France) and a laboratory protocol as used by EAWAG, Swiss Federal Institute of Aquatic Science and Technology, i. e. an estimated scaling factor of $s = 30$, a partition volume of $v = 0.519\,\text{nL}$, and $m = 25000$ partitions per measurement. We provide a sensitivity analysis of these parameters in electronic supplementary material E.

We also compared the theoretical predictions for the CV and the probability of non-detection with empirical data from ongoing wastewater surveillance by EAWAG. For this, we used dPCR measurements for Influenza A virus (IAV), Influenza B virus (IBV), Respiratory syncytial virus (RSV), and SARS-CoV-2 of samples from six wastewater treatment plants in Switzerland from July

10, 2023 to March 5, 2024 obtained using the laboratory methods described in Nadeau et al. [4] (see electronic supplementary material A for details) to compute empirical estimates of the CV and probability of non-detection. For each target and sample, the CV was empirically estimated from two technical replicates, using the formula

$$\hat{\nu}_{\hat{\lambda}} = \sqrt{\frac{\pi}{2}} \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\lambda}_i - \mu_{\hat{\lambda}})^2}}{\mu_{\hat{\lambda}}}, \tag{15}$$

where $\mu_{\hat{\lambda}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\lambda}_i$ is the empirical mean of the replicates and $\sqrt{\frac{\pi}{2}}$ is a bias correction factor for the empirical standard deviation. We note that $\hat{\nu}_{\hat{\lambda}} = \sqrt{\pi}$ in the special case where one of the replicate measurements is zero. To estimate the probability of non-detection, we used the technical replicates by treating the first replicate as an estimate of the true target concentration, and the second replicate as an indicator of the probability of zero. Specifically, for a given window $w \in \{0, \ldots, 30\}$, we selected all replicate pairs where the result of the first replicate was between w and $w+1$ gc/mL. Of these, we then computed the percentage of second replicates with a result of zero to estimate $p_{\text{zero}}(w+0.5)$. The moving window was here used to group samples with a similar concentration together while accounting for the noise of the measurements. Given the empirical estimates for different target concentrations, we applied locally estimated scatterplot smoothing (LOESS) [29] to estimate a smooth relationship of the target concentration with the CV and probability of non-detection, respectively. We compared the resulting curves with the $\nu_{\hat{\Lambda}}(\lambda)$ and $p_{\text{zero}}(\lambda)$ functions as predicted by Eq. (6) and Eq. (10). As technical replicates of the same preprocessed sample were used in this comparison, the estimates did not include potential preprocessing noise and we assumed $\nu_{\text{pre}} = 0$ accordingly.

To validate the likelihood function, we simulated measurements using the positive partition count model described in Section 2.1 for concentrations of 1, 3, and 5 gc/mL (assuming the same PCR parameters as above, i. e. $s = 30$, $v = 0.519$ nL, and $m = 25000$) and estimated the respective concentration from the simulated dPCR measurements using the Bayesian model for single samples described in Section 3. For comparison purposes, we also tested the model with two simpler likelihoods, i. e. i) using a normal distribution with constant variance and ii) using a log-normal distribution with constant CV. When using the log-normal likelihood, zero measurements had to be dropped from the analysis, otherwise, the models and data used were identical for all likelihoods.

We deliberately used a large number of 100 simulated replicates per concentration for estimation to study asymptotic differences between the different likelihoods.

4.2 Application to real-world data

We applied an epidemiological model from the R package "EpiSewer" [28] with our likelihood function to real-world duplicate dPCR measurements of Influenza A virus from the wastewater treatment plants of Zurich and Geneva, Switzerland during the Influenza season 2022/2023. For the generation time distribution of Influenza A virus, we used a discretised shifted Gamma distribution [30] with a mean of 2.6 and a standard deviation of 1.5 [31]. For the shedding load distribution since the time of infection, we used a discretised Gamma distribution with shape parameter 6.15 and scale parameter 0.40 [32], corresponding to a mean of 2.49 and a standard deviation of 1. Based on a comparison with confirmed case numbers, we assumed an average total shedding load per case of 1×10^{11} gc. No variation in the total shedding load was modelled. The sewer residence time was assumed to be less than 1 day. We conducted retrospective real-time inference using our dPCR-specific likelihood and, for comparison, using a normal distribution with constant variance or a log-normal distribution with constant coefficient of variation. The likelihoods were compared in terms of i) the fit of the model to the observations, ii) estimates of the time-varying load, and iii) estimates of the effective reproduction number R_t. We also produced short-term forecasts of these quantities up to one week ahead using EpiSewer based on the assumption of continued transmission dynamics (see electronic supplementary material G.2 for details). We assessed the forecasts of concentration measurements using the Continuous ranked probability score (CRPS) [33], i. e.

$$\mathrm{CRPS}(F_{\hat{\Lambda}}, \hat{\lambda}) = \int_{-\infty}^{\infty} \left(F_{\hat{\Lambda}}(y) - \mathbb{1}\{y \geq \hat{\lambda}\} \right)^2 dy, \tag{16}$$

where $\hat{\lambda}$ is the observed dPCR measurement and $F_{\hat{\Lambda}}$ is the cumulative distribution function of the corresponding probabilistic forecast by EpiSewer. To compute the CRPS from posterior samples of our model fitted with MCMC, we used a method by Jordan et al. [34] implemented in the R package "scoringutils" [35]. We summarised the CRPS scores from different days using the arithmetic mean, denoted $\overline{\mathrm{CRPS}}$. Using our dPCR-specific model, we also compared prior and posterior distributions of the jointly estimated PCR parameters to assess the extent to which they were informed by the data.

All analyses were performed in R 4.1.0. We used the R package "targets" [36] to build a pipeline for inference from dPCR measurements.

5 Results

5.1 Coefficient of variation

Figure 1A shows the coefficient of variation (CV) of dPCR measurements as a function of the concentration in the sample. We see that the CV increases exponentially as the concentration approaches zero but is approximately constant for higher concentrations. However, this stabilisation of the CV does not apply for arbitrarily high concentrations (electronic supplementary material E, Figure 4), since the variance of the digital PCR increases drastically as the share of positive partitions approaches one [37]. Figure 1A also compares the CV functions for different levels of pre-PCR variation under the assumption of gamma distributed noise. As can be seen, pre-PCR variation acts as an intercept for the CV function only at higher concentrations, and the CV functions converge as the concentration approaches zero. Our expression for the CV furthermore shows that the averaging of technical replicates scales the CV by a factor of $\frac{1}{\sqrt{n}}$, where n is the number of replicates. For example, absent pre-PCR noise, using an average of two technical replicates reduces the CV by $\approx 30\%$. Under pre-PCR noise, however, this effect is smaller. For log-normally distributed noise, the CV cannot be exactly computed but can be accurately approximated using a Taylor series expansion or an approximation of the log-normal MGF by Asmussen et al. [38] (see electronic supplementary material E, Figure 3). Apart from extremely high concentrations, we find that the CV is almost identical for log-normally and gamma distributed pre-PCR noise (electronic supplementary material E, Figure 4).

To verify our theoretical CV model on real-world data, we computed empirical CV estimates based on technical replicates of dPCR measurements for different respiratory viruses in wastewater (see Methods, Section 4.1). Figure 2A shows such CV estimates for Influenza A and B virus, RSV, and SARS-CoV-2 from 14 wastewater treatment plants in Switzerland, together with a locally estimated scatterplot smoothing (LOESS) curve to model the relationship between the empirical concentration and CV. For all targets, we find that the theoretical curve predicted by Eq. (6) agrees well with the empirical CV curve (median absolute deviation: 1.89 percentage points). We note that

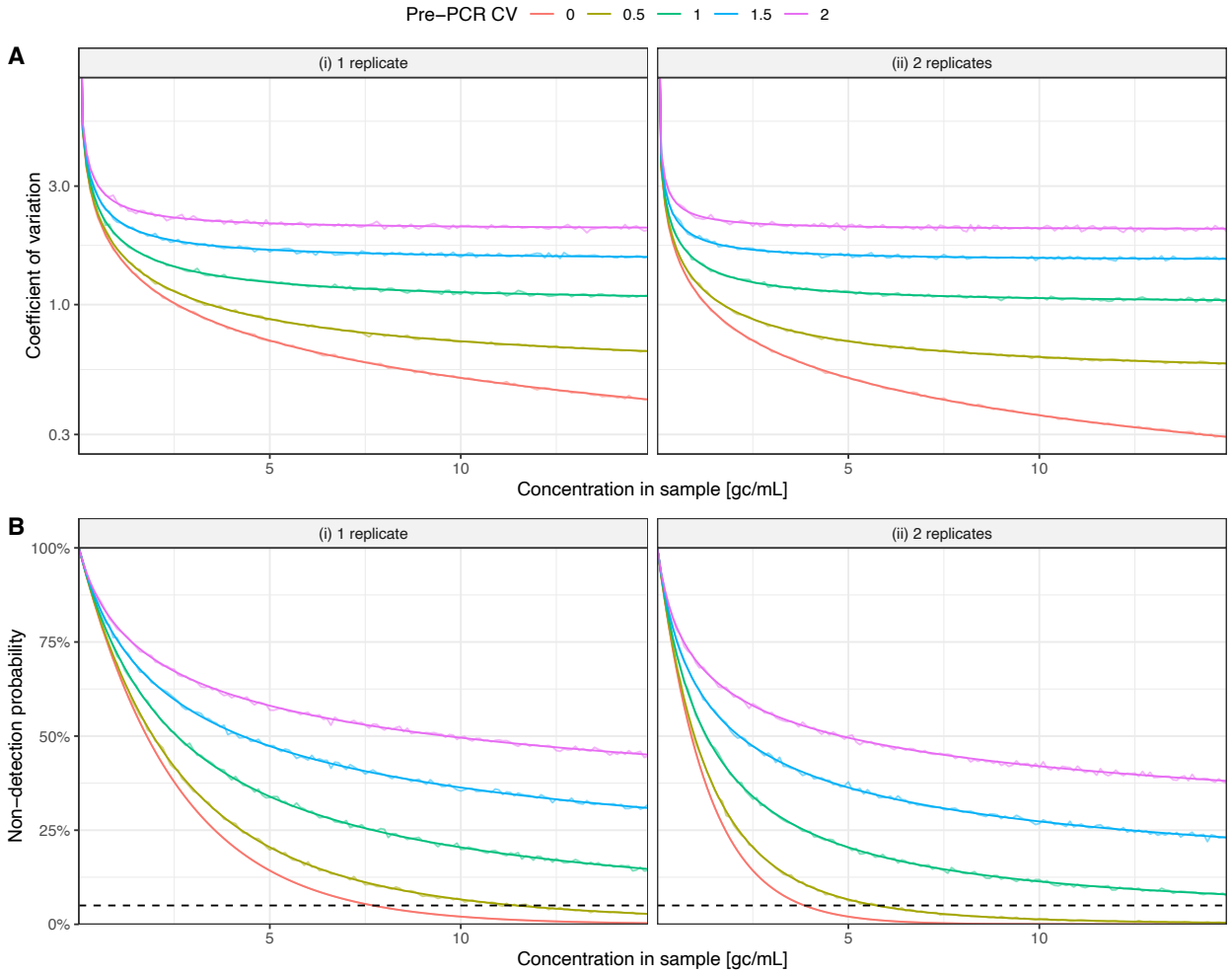

Fig 1. Theoretical relationship of the sample concentration with the coefficient of variation and probability of non-detection in digital PCR. Shown is (A) the coefficient of variation of measurements and (B) probability of non-detection from digital PCR for (i) a single replicate and (ii) the arithmetic mean of two replicates, as a function of the sample concentration λ under gamma distributed pre-PCR noise. Jagged lines show the CV and probability of non-detection from simulated measurements, respectively, for different concentrations in steps of 0.1 gc/mL. Straight lines show the corresponding CV and probability of non-detection as predicted by Eq. (9) and Eq. (11), respectively. Colors correspond to different strengths of pre-PCR noise, as measured by the pre-PCR coefficient of variation ν_{pre}. Simulated and predicted measurements assume a scaling factor of $s = 30$, a partition volume of $v = 0.519\,\text{nL}$, and $m = 25000$ partitions per measurement. In panel (B), dashed horizontal lines indicate the limit of detection (LoD), defined as a 5% probability of non-detection.

292 since the empirical CV is based on technical replicates of the same preprocessed sample, we could
293 not assess the effect of pre-PCR noise as predicted by Eq. (9).

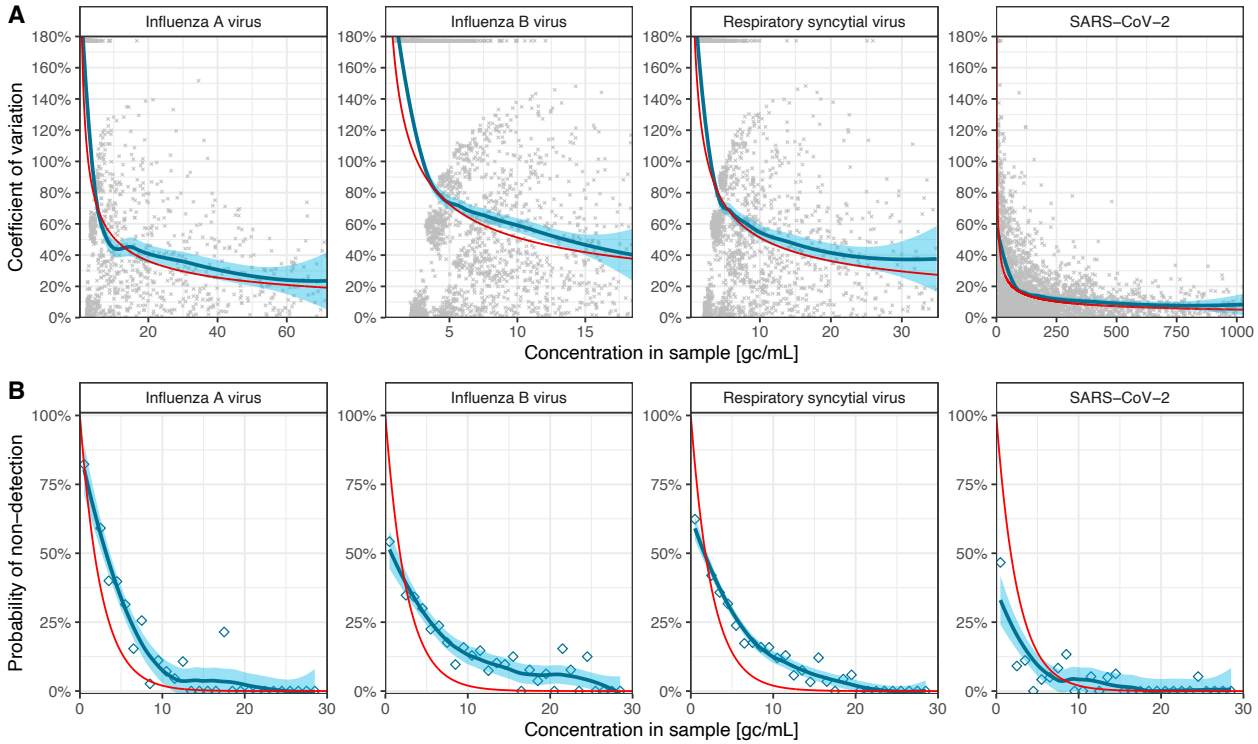

Fig 2. Comparison of theoretical predictions with real-world dPCR measurements. Shown are empirical estimates of (A) the coefficient of variation and (B) the probability of non-detection for dPCR measurements of Influenza A and B virus, RSV, and SARS-CoV-2 concentrations from 14 wastewater treatment plants in Switzerland. Coefficient of variation (CV) estimates are based on the bias-corrected empirical standard deviation of two technical replicates, respectively, divided by their empirical mean (dots). Probability of non-detection estimates are based on the percentage of second replicates with a result of zero among all replicate pairs where the result of the first replicate falls within a moving window of 1 gc/mL width (diamonds). Shown in blue with 95% uncertainty intervals are LOESS-based kernel regressions of the empirical estimates, indicating the relationship between the concentration in the sample and empirical CV and probability of non-detection, respectively. Shown in red are the CV and probability of non-detection as predicted by the dPCR-specific measurement model. Note that as the empirical CV is based on technical replicates of the same sample, respectively, it does not include variation from pre-PCR noise.

5.2 Probability of non-detection

The probability of non-detection describes the chance of zero measurements despite the presence of the target in the sample. Such "false negatives" are problematic when modelling dPCR measurements via strictly positive likelihoods such as a log-normal distribution, as they assign zero probability to such measurements. If concentrations are sufficiently low, non-detects will however occur. For example, using the exemplary PCR parameters described in Section 4.1, Eq. (10) predicts that even without pre-PCR noise, there is an 11.5% probability that a concentration of $\lambda = 5\,\mathrm{gc/mL}$ will be measured as zero. This means that the corresponding observations cannot be modelled via a strictly positive distribution, motivating the use of a hurdle model as introduced in Section 2.4.

Figure 1B shows simulated and theoretically predicted non-detection probabilities under pre-PCR noise of different strengths. As can be seen, the presence of pre-PCR noise further increases the probability of zero measurements. For example, under gamma distributed pre-PCR noise with a CV of $\nu_{\text{pre}} = 0.5$ and using the exemplary PCR parameters described in Section 4.1, the non-detection probability for a concentration of $\lambda = 5\,\text{gc/mL}$ is 5.5 percentage points higher than without. From our expression derived in Eq. (10), we also see that the probability of non-detection for a given concentration decreases with the number of partitions and the number of replicates. For example, a doubling of the number of replicates has the same effect on the probability of non-detection as a doubling of the concentration. As with the CV, computing the probability of non-detection under log-normally distributed pre-PCR noise requires an approximation of the log-normal MGF. We find that the approximation by Asmussen et al. [38] is accurate in general, but slightly underestimates the CV at low concentrations when the pre-PCR variation is very high (electronic supplementary material E, Figure 7). Moreover, the non-detection probability curve changes significantly if we model the pre-PCR noise as log-normally instead of gamma distributed, with gamma distributed noise leading to higher non-detection probabilities.

To verify our model for the probability of non-detection on real-world data, we used the same real-world measurements as in Section 5.1 to compute empirical estimates of the probability of non-detection p_{zero} (see Methods, Section 4.1). Figure 2B shows the p_{zero} estimates together with a locally estimated scatterplot smoothing (LOESS) curve to model their relationship with the underlying concentration. Here we note that the empirical estimates of p_{zero} rely on a moving window approach and are therefore less accurate than the empirical CV estimates in Figure 2A. Nevertheless, we find a general agreement between the theoretical curve predicted by Eq. (10) and the empirical $p_{\text{zero}}(\lambda)$ curve (median absolute deviation: 3.79 percentage points).

5.3 Inference from concentration measurements

To compare our dPCR-specific likelihood with simpler likelihoods, we fitted a Bayesian model to simulated dPCR measurements of different concentrations (see Methods, Section 4.2). Figure 3 shows the true and estimated concentrations using a normal, log-normal, and dPCR-specific likelihood function, together with the posterior predictive distributions for the dPCR measurements. When using a normal likelihood, estimates of the concentration are unbiased, however, the model can

predict negative measurements. When using a log-normal likelihood, no negative measurements are predicted, however, zero measurements must be dropped from the analysis. This leads to a significant bias in the estimated concentration when the true concentration is small. When using a dPCR-specific likelihood, the inflated probability of zero measurements is explicitly represented and the true concentration is estimated without bias.

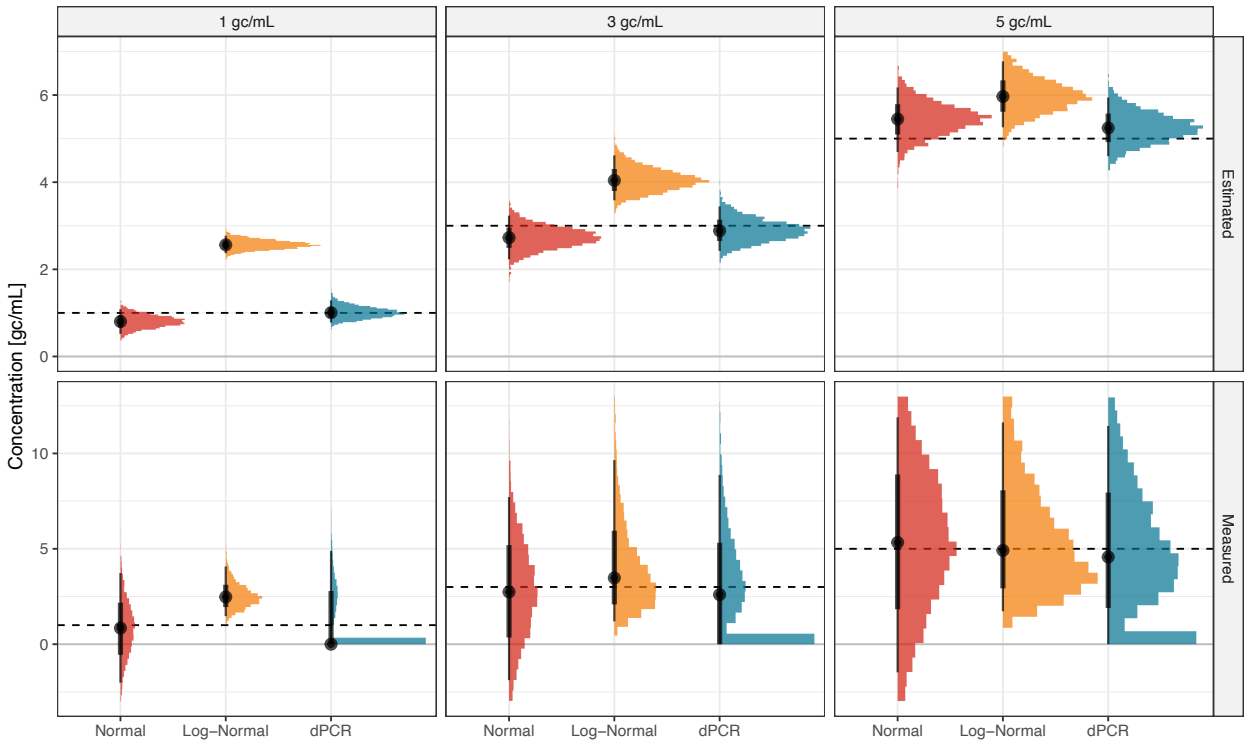

Fig 3. Comparison of estimated concentration and predicted dPCR measurements under different likelihood functions. Shown is the posterior distribution of the estimated true concentration (top) and the posterior predictive distribution of measurements (bottom) from a Bayesian model fitted with a normal (red), log-normal (orange), and dPCR-specific (blue) likelihood to simulated dPCR measurements of different concentrations (1, 3, and 5 gc/mL). Dashed lines show the true concentration, respectively.

The dPCR-specific likelihood also improved inference from real-world dPCR measurements over time. Using the normal, log-normal, and dPCR-specific likelihood, respectively, we fitted a model for estimation of the effective reproduction number R_t to dPCR measurements of Influenza A virus at the municipal treatment plant of Zurich, Switzerland. Figure 4 compares estimated and forecasted dPCR measurements, time-varying viral loads, and R_t values obtained with the different likelihoods. Under the normal likelihood, measurements are erroneously modelled to have constant variance. At low concentrations, we find that this can lead to an overestimation of measurement noise. For example, before Dec 1, 2022, the 80% credible intervals (CrIs) of measurements predicted by the normal

likelihood covered already 98.5% of observations, suggesting a miscalibration of the model. Under the log-normal likelihood, in contrast, measurements are modelled to have a constant coefficient of variation. However, our results in Section 5.1 show that this assumptions is equally violated. As a consequence, measurement noise tends to be underestimated at low concentrations (80% CrIs covered only 59.7% of observations before Dec 1, 2022) and overestimated at high concentrations (80% CrIs covered 100% of observations after Dec 1, 2022). By modelling a concentration-dependent coefficient of variation, the dPCR-specific likelihood achieved similar coverage at low and high concentrations (80% CrIs covered 88.0% of observations before Dec 1, 2022 and 84.2% after Dec 1, 2022). We furthermore find that concentration curve predicted by the normal likelihood model was more volatile than that of the dPCR-specific likelihood model. For example, the median estimated concentration falls within the previous 95% CrIs on only 56.0% of days for the normal likelihood, and on 84.0% of days for the dPCR-specific likelihood. This can indicate an overfitting on measurement noise by the normal likelihood model. Measurements after Dec 26, 2022 were forecasted, here the dPCR-specific likelihood achieved better forecast scores ($\overline{\text{CRPS}} = 12.04$) than both the normal ($\overline{\text{CRPS}} = 20.08$) and log-normal likelihood ($\overline{\text{CRPS}} = 18.19$).

The dPCR-specific likelihood model also reduced uncertainty in estimating and forecasting unobserved quantities such as the viral load (Figure 4B) and the effective reproduction number (Figure 4C). Specifically, we find that when using the normal likelihood, the time-varying load and R_t estimates were considerably more volatile than when using the other likelihoods. While no exact ground truth for R_t is known, the strong fluctuations around $R_t = 1$ (Figure 4C1) under the normal likelihood further suggest an overfitting of the model at the peak of the seasonal wave. At the start of the seasonal wave, R_t estimates under the normal likelihood were highly uncertain, which can be explained by the overestimation of measurement noise at low concentrations. On the other hand, estimates from the log-normal and dPCR-specific likelihood models were more similar. Nevertheless, as the log-normal likelihood model overpredicts the measurement noise at higher concentrations, the load curve (Figure 4B2) and R_t estimates (Figure 4C2) were more uncertain than under the dPCR-specific likelihood. Generally, both the normal and log-normal likelihood showed more uncertain forecasts of the load trend. In electronic supplementary material G.3, Figures 12–16, we show additional results from fitting the models on different date ranges of measurements at the treatment plants of Zurich and Geneva, Switzerland.

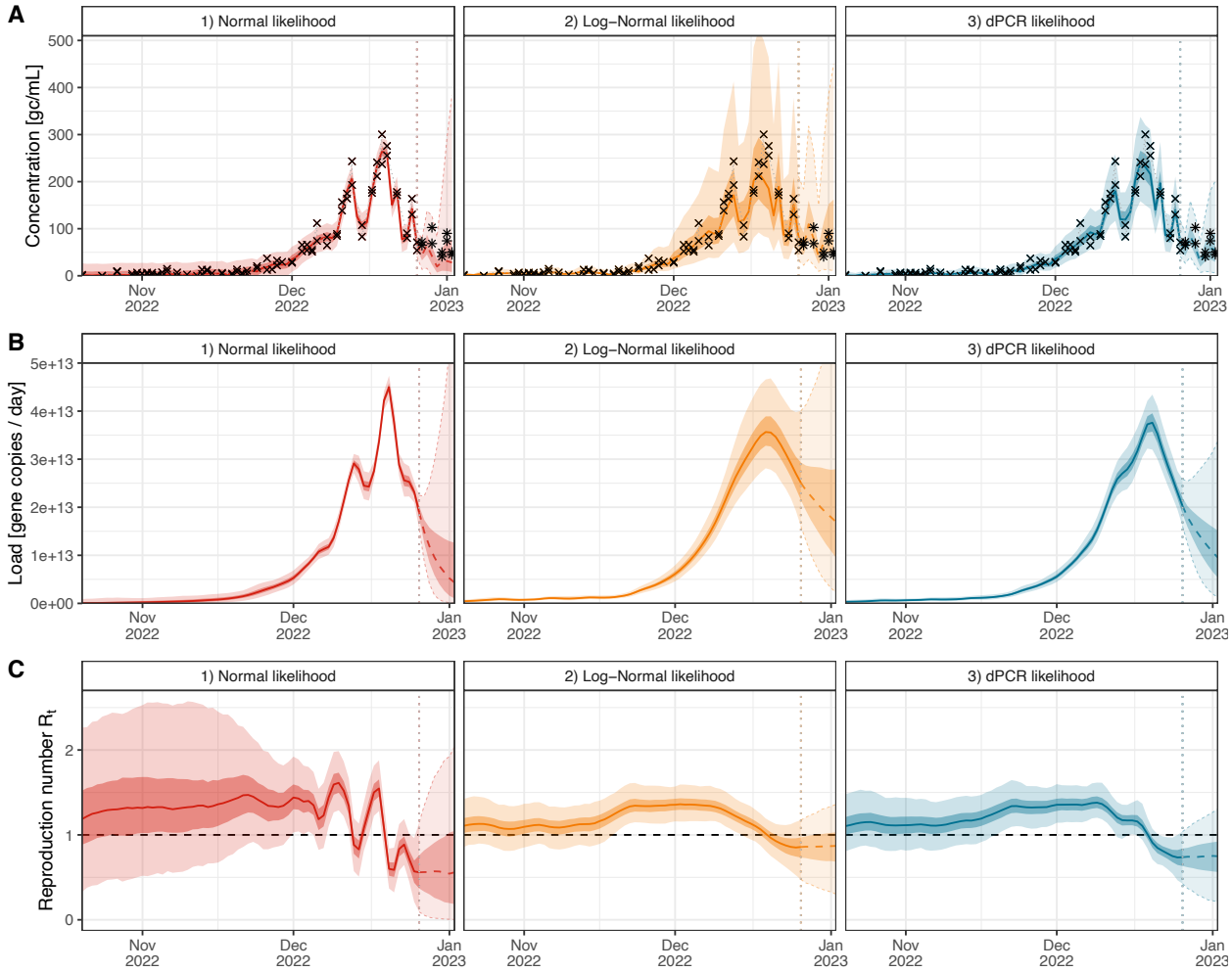

Fig 4. Comparison of inference and short-term forecasts from wastewater dPCR measurements using different likelihood functions: Shown are the results of fitting an epidemiological wastewater model (EpiSewer) to duplicate dPCR measurements of Influenza A virus concentrations at the municipal treatment plant of Zurich, Switzerland (Oct 20 – Dec 26, 2022) using either 1) a normal likelihood (red), 2) a log-normal likelihood (orange), and 3) a dPCR-specific likelihood (blue) for the observations. Each panel shows the median (lines), and 50% and 95% credible intervals (strong and weakly shaded areas) of A) posterior predictive distributions for dPCR measurements, B) estimated viral loads in wastewater over time, and C) the estimated effective reproduction number over time. Estimates shown beyond Dec 26, 2022 (vertical dotted line) are forecasted based on a continuation of the latest transmission dynamics. Panel A also shows the true observed (crosses) and future (stars) dPCR measurements.

Our proposed likelihood also improves model fit even if the precise characteristics of the assay are unknown. For example, when fitting the dPCR-specific model to real-world measurements as shown in Figure 4, we did not supply the specific parameters of the dPCR assay. Instead, we jointly estimated the average number of partitions μ_m and the concentration conversion factor c as additional parameters (see Methods, Section 3). When comparing to an informed model where the number of total partitions and the conversion factor are fixed to their true values, we find that

the uninformed model yields highly similar estimates despite the parameter uncertainty (electronic supplementary material G.3, Figure 17). When inspecting the posterior distribution of the PCR parameters, we found that μ_m remained uncertain and did not differ notably from its prior, but c was informed by the data, with its posterior mode closely matching the true value of c (electronic supplementary material G.1, Figure 11). The posterior distribution for ν_{pre} also differed clearly from its prior, with a median estimate of 17.0% (95% credible interval: 5.3% – 28.1%). As this parameter describes the variation in concentrations before the PCR, it allows us to distinguish the technical noise of digital PCR from the extraction and preprocessing noise and other unexplained sources of variation. For instance, using the median estimate for ν_{pre} and for the given values of m and c, the overall CV at a concentration of 100 gc/mL is 22.9%, which is 7.8 percentage points higher than the optimal CV that would be achieved without preprocessing noise or other unexplained sources of variation (15.1%).

6 Discussion

In this work, we have developed a model for statistical inference from dPCR concentration measurements with an application focus on wastewater-based epidemiology. Our model reflects two important statistical properties of dPCR, namely i) concentration-dependent measurement noise and ii) the probability of zero measurements (i. e. non-detection). We derived and verified approximations of these quantities while accounting for additional noise such as from extraction and preprocessing of samples or other sources of variation, and for the averaging over multiple technical replicates. Based on these results, we proposed i) the use of a hurdle model to account for the probability of zero measurements, ii) the use of a gamma distribution parameterised by its mean and concentration-dependent coefficient of variation for non-zero measurements.

Using real-world wastewater measurements over time, we have shown that our proposed likelihood provides a better fit to dPCR data than simpler likelihoods currently used in practice [1, 5, 12–17]. Specifically, we find that a normal distribution with constant variance does not adequately reflect the data generating process of dPCR measurements and can lead to overfitting on measurement noise that severely impacts epidemiological models. In contrast, the use of a log-normal distribution with a constant coefficient of variation (or, equivalently, the modelling of log-transformed measurements as normally distributed) may provide a sufficient approximation to the measurement distribution,

but only if all concentrations are well above the Limit of Detection. When concentrations are lower, it becomes important to model a concentration-dependent coefficient of variation and the probability of zero measurements. Otherwise, our results show that the resulting model misspecification leads to biased concentration estimates at low concentrations and overestimated measurement noise at high concentrations.

Our findings have several important implications for real-time wastewater surveillance. First, the use of a misspecified observation model for wastewater measurements may not only affect the reliability of concentration estimates but also increase the uncertainty of epidemiological indicators and forecasts, potentially hindering the detection of changes in transmission dynamics. Thus, for robust tracking of ongoing epidemics and detection of new outbreaks, an accurate model of PCR measurements is important. Second, our results also imply that wastewater loads (concentrations multiplied by flow volume) or shedding-normalised concentrations (concentrations divided by a faecal biomarker such as crAssphage or PMMoV concentration) [39, 40] are not recommended as direct inputs to wastewater models, because their measurement distribution depends on the normalisation factor. For example, depending on the daily flow at the treatment plant, the same wastewater load may correspond to vastly different concentrations, with different measurement noise and probabilities of non-detection. The same applies to concentrations normalised by faecal markers, with the added complication that marker concentrations are themselves subject to measurement noise. Therefore, we recommend using non-normalised concentrations as observations instead and incorporating normalisation factors into the model through a latent variable approach. In our real-world example, we have demonstrated this for flow normalisation using the R package "EpiSewer", but the approach can also be adapted for normalisation with faecal markers. Third, we suspect that epidemiological methods developed for case count data may not be directly applicable to wastewater data. For example, "EpiEstim", a popular method for R_t estimation by Cori et al. [30], assumes Poisson distributed case counts as observations. If directly fitted to wastewater concentrations or loads, the resulting misspecification could distort the uncertainty of reproduction number estimates. Finally, for modelling dPCR measurements at low concentrations, we argue that the hurdle model proposed in this work is more appropriate than previously discussed censoring approaches in which measurements below the LoD are censored by replacing them with zero or the LoD value [13, 16, 41]. This is because censoring assumes a fixed threshold at which values are censored, but it follows from Eq. (10) that the

probability of non-detection gradually increases as the concentration approaches zero. Furthermore, censoring applies to the dPCR measurements, which are subject to noise, but the probability of non-detection depends on the true concentration in the sample. For example, concentrations above the LoD can lead to measurements below the LoD. In a censoring model, these observations would be incorrectly treated as below the LoD, and vice versa. Similar arguments apply to the Limit of Quantification (LoQ) [26], a threshold based on a predefined CV below which measurements should not be interpreted quantitatively. Again, considering the gradual increase of the CV as concentrations approach zero and the fact that the CV depends on the sample concentration, not the measurement, such a threshold can be regarded as arbitrary and potentially inaccurate. The likelihood proposed here avoids using such thresholds by accurately modelling the distribution of dPCR measurements at low concentrations.

When analysing wastewater measurements, it is in principle possible to model the number of positive partitions in dPCR [42]. In practice, however, laboratories typically report concentrations (e. g. in gene copies per mL wastewater), which are more interpretable and independent of laboratory parameters. To be widely applicable, especially for publicly reported data, models must therefore use concentration measurements as observations. Using Bayesian inference, our dPCR-specific model can be fitted directly to such concentration measurements, with the option to include multiple replicates per sample or averages thereof. Information from the laboratory protocol, such as the dilution factor used or the parameters of the PCR, can be integrated by fixing the respective parameters to their values. Importantly, however, the model provides precise estimates of the load trajectory even when broad priors for the laboratory parameters are used. This highlights the applicability of our model even in settings where the laboratory details are unknown. We thus argue that it is preferable to express a lack of knowledge about laboratory parameters through broad priors rather than using simple but misspecified likelihoods for inference.

We note a few limitations of our model. First, we have tested our estimators for the coefficient of variation (CV) of measurements and the probability of non-detection only for log-normally and gamma distributed pre-PCR noise. In principle, any positive continuous parametric distribution with a well-defined moment generating function (MGF) could be used, but since closed-form expressions or accurate approximations of the MGF are not known for all distributions, some cases may require the use of numerical integration. Moreover, a difficulty in choosing a pre-PCR noise distribution

in practice is that this parameter summarizes all sources of unexplained variation before the PCR. In the context of wastewater surveillance, this may not only include extraction and preprocessing noise, which may be analysed through wet-lab experiments, but also variation in sampling at the treatment plant, sample storage, and other factors. We found that for the CV of measurements, the exact distribution of the pre-PCR noise is not highly relevant for most concentration values and can be well approximated using a Taylor series expansion. For the probability of non-detection however, we found notable differences between log-normally and gamma distributed pre-PCR noise and no suitable Taylor series approximation. Although general advice is difficult, we remark that if the pre-PCR noise is assumed to be the product of many independent noise components, the approximate use of a log-normal distribution may be justified by the multiplicative central limit theorem [43].

In our model, we assume that in expectation the dPCR measurements are equal to the true sample concentration. In other words, our modelling only accounts for random measurement error, not for unknown systematic distortions of dPCR measurements, e. g. due to false calibration. Similarly, while our model works well without specific knowledge of the laboratory parameters, we assume that these are stable over time, i. e. systematic changes in the laboratory protocol may not be detected without additional information. We also did not account for variation in the volume of partitions, which can be an additional but minor source of noise in droplet-based systems [37, 44]. Moreover, our hurdle model assumes that, given a true concentration and the parameters of the PCR assay, the probability of non-detection is stochastically independent of the CV of the measurements. In practice, these quantities are correlated due to the same underlying pre-PCR noise, however, this correlation has limited impact as it only applies to measurements that are non-zero but small. When computing the CV and the probability of non-detection for measurements that have been averaged over several replicates, we assume for simplicity that the number of partitions does not vary between the replicates and that all replicates are based on the same preprocessed sample. However, if the exact total and positive partition numbers are known for each replicate, this information can be used by fitting our model to individual replicate measurements, as demonstrated in the real-world example in this paper.

A further limitation of our model is that we do not distinguish between noise at different stages of the data-generating process, e. g. sampling vs. extraction [24]. This can however be achieved

through hierarchical modeling of noise, as for example implemented in the EpiSewer model [28]. We also note that our notation and examples assumed that concentrations are measured in gene copies per volume but, in the case of solids-based extraction, measurements may have other units such as gene copies per g of dry weight [39]. Still, our model can be equally applied to such units by adapting the scaling factor s accordingly. Finally, since this work focused on the surveillance of pathogens already present in the population, we assumed that the true concentration can be arbitrarily small but will be larger than zero. If, instead, the objective is to determine whether a particular pathogen is present or absent, it may also be important to model the probability of false positives and the Limit of Blank [26].

The model developed in this work has been integrated as a user-friendly module in the open-source R package EpiSewer [28]. We envision our proposed likelihood function to improve pathogen monitoring by overcoming the limitations of simplistic likelihoods currently used by many methods for wastewater-based epidemiology. Beyond this application, our results may also benefit other domains involving dPCR measurements, such as dPCR-based eDNA monitoring [45]. From an experimental perspective, the approximations for the coefficient of variation and probability of non-detection under pre-PCR noise derived in this work could inform the effect of laboratory protocols and parameters on the precision of measurements. For example, our finding that pre-PCR noise strongly influences the probability of non-detection highlights the importance of including extraction and preprocessing steps in wet-lab recommendations for determining the Limit of Detection [24, 26]. While this work has focused on digital PCR, we expect that similar statistical relationships and models could be established for quantitative PCR methods [46]. We therefore hope that accurate modelling of the measurement process will not only improve real-time monitoring and forecasting in wastewater-based epidemiology, but also help to disentangle measurement noise from other relevant sources of variation, such as in-sewer environmental factors and individual-level differences in shedding profiles.

References

1. Huisman JS, Scire J, Caduff L, Fernandez CX, Ganesanandamoorthy P, Kull A, et al. Wastewater-Based Estimation of the Effective Reproductive Number of SARS-CoV-2. Environmental Health Perspectives. 2022;130(5):057011.

2. Mercier E, D'Aoust PM, Thakali O, Hegazy N, Jia JJ, Zhang Z, et al. Municipal and Neighbourhood Level Wastewater Surveillance and Subtyping of an Influenza Virus Outbreak. Scientific Reports. 2022;12(1):15777.

3. Ho J, Stange C, Suhrborg R, Wurzbacher C, Drewes JE, Tiehm A. SARS-CoV-2 Wastewater Surveillance in Germany: Long-term RT-digital Droplet PCR Monitoring, Suitability of Primer/Probe Combinations and Biomarker Stability. Water Research. 2022;210:117977.

4. Nadeau S, Devaux AJ, Bagutti C, Alt M, Hampe EI, Kraus M, et al. Influenza transmission dynamics quantified from RNA in wastewater in Switzerland. Swiss Medical Weekly. 2024;154(1):3503–3503.

5. Toribio-Avedillo D, Gómez-Gómez C, Sala-Comorera L, Rodríguez-Rubio L, Carcereny A, García-Pedemonte D, et al. Monitoring Influenza and Respiratory Syncytial Virus in Wastewater. Beyond COVID-19. Science of The Total Environment. 2023;892:164495.

6. Julian TR, Devaux AJ, Brülisauer L, Conforti S, Rusch JC, Gan C, et al. Monitoring an Emergent Pathogen at Low Incidence in Wastewater Using qPCR: Mpox in Switzerland. Food and Environmental Virology. 2024;p. 1–11.

7. Wolfe MK, Yu AT, Duong D, Rane MS, Hughes B, Chan-Herur V, et al. Use of Wastewater for Mpox Outbreak Surveillance in California. New England Journal of Medicine. 2023;388(6):570–572.

8. Boehm AB, Hughes B, Duong D, Chan-Herur V, Buchman A, Wolfe MK, et al. Wastewater Concentrations of Human Influenza, Metapneumovirus, Parainfluenza, Respiratory Syncytial Virus, Rhinovirus, and Seasonal Coronavirus Nucleic-Acids during the COVID-19 Pandemic: A Surveillance Study. The Lancet Microbe. 2023;4(5):e340–e348.

9. Conforti S, Holschneider A, Sylvestre É, Julian TR. Monitoring ESBL-Escherichia Coli in Swiss Wastewater between November 2021 and November 2022: Insights into Population Carriage. mSphere. 2024;9(5):e00760–23.

10. Keshaviah A, Diamond MB, Wade MJ, Scarpino SV, Ahmed W, Amman F, et al. Wastewater Monitoring Can Anchor Global Disease Surveillance Systems. The Lancet Global Health. 2023;11(6):e976–e981.

11. Fernandez-Cassi X, Scheidegger A, Bänziger C, Cariti F, Tuñas Corzon A, Ganesanandamoorthy P, et al. Wastewater Monitoring Outperforms Case Numbers as a Tool to Track COVID-19 Incidence Dynamics When Test Positivity Rates Are High. Water Research. 2021;200:117252.

12. Vallejo JA, Trigo-Tasende N, Rumbo-Feal S, Conde-Pérez K, López-Oriona Á, Barbeito I, et al. Modeling the Number of People Infected with SARS-COV-2 from Wastewater Viral Load in Northwest Spain. Science of The Total Environment. 2022;811:152334.

13. Courbariaux M, Cluzel N, Wang S, Maréchal V, Moulin L, Wurtzer S, et al. A Flexible Smoother Adapted to Censored Data With Outliers and Its Application to SARS-CoV-2 Monitoring in Wastewater. Frontiers in Applied Mathematics and Statistics. 2022;8:836349.

14. Champredon D, Papst I, Yusuf W. Ern: An R Package to Estimate the Effective Reproduction Number Using Clinical and Wastewater Surveillance Data. PLOS ONE. 2024;19(6):e0305550.

15. Kaplan EH, Wang D, Wang M, Malik AA, Zulli A, Peccia J. Aligning SARS-CoV-2 Indicators via an Epidemic Model: Application to Hospital Admissions and RNA Detection in Sewage Sludge. Health Care Management Science. 2020;p. 1–10.

16. Lewis-Borrell L, Irving J, Lilley CJ, Courbariaux M, Nuel G, Danon L, et al. Robust Smoothing of Left-Censored Time Series Data with a Dynamic Linear Model to Infer SARS-CoV-2 RNA Concentrations in Wastewater. AIMS Mathematics. 2023;8(7):16790–16824.

17. Li G, Denise H, Diggle P, Grimsley J, Holmes C, James D, et al. A Spatio-Temporal Framework for Modelling Wastewater Concentration during the COVID-19 Pandemic. Environment International. 2023;172:107765.

18. Watson LM, Plank MJ, Armstrong BA, Chapman JR, Hewitt J, Morris H, et al. Jointly Estimating Epidemiological Dynamics of Covid-19 from Case and Wastewater Data in Aotearoa

New Zealand. Communications Medicine. 2024;4(1):1–9.

19. Li X, Zhang S, Sherchan S, Orive G, Lertxundi U, Haramoto E, et al. Correlation between SARS-CoV-2 RNA Concentration in Wastewater and COVID-19 Cases in Community: A Systematic Review and Meta-Analysis. Journal of Hazardous Materials. 2023;441:129848.

20. Basu AS. Digital Assays Part I: Partitioning Statistics and Digital PCR. SLAS TECHNOLOGY: Translating Life Sciences Innovation. 2017;22(4):369–386.

21. Dorazio RM, Hunter ME. Statistical Models for the Analysis and Design of Digital Polymerase Chain Reaction (dPCR) Experiments. Analytical Chemistry. 2015;87(21):10886–10893.

22. Salipante SJ, Jerome KR. Digital PCR—An Emerging Technology with Broad Applications in Microbiology. Clinical Chemistry. 2020;66(1):117–123.

23. Ciesielski M, Blackwood D, Clerkin T, Gonzalez R, Thompson H, Larson A, et al. Assessing Sensitivity and Reproducibility of RT-ddPCR and RT-qPCR for the Quantification of SARS-CoV-2 in Wastewater. Journal of Virological Methods. 2021;297:114230.

24. Borchardt MA, Boehm AB, Salit M, Spencer SK, Wigginton KR, Noble RT. The Environmental Microbiology Minimum Information (EMMI) Guidelines: qPCR and dPCR Quality and Reporting for Environmental Microbiology. Environmental Science & Technology. 2021;55(15):10210–10223.

25. Dube S, Qin J, Ramakrishnan R. Mathematical Analysis of Copy Number Variation in a DNA Sample Using Digital PCR on a Nanofluidic Device. PLOS ONE. 2008;3(8):e2876.

26. Armbruster DA, Pry T. Limit of Blank, Limit of Detection and Limit of Quantitation. The Clinical Biochemist Reviews. 2008;29(Suppl 1):S49–S52.

27. Stan development team. Stan Modeling Language Users Guide and Reference Manual, Version 2.31; 2022.

28. Lison A. EpiSewer v0.0.3; 2024. Zenodo. Available from: https://zenodo.org/doi/10.5281/zenodo.13899759.

29. Cleveland WS, Loader C. Smoothing by Local Regression: Principles and Methods. In: Härdle W, Schimek MG, editors. Statistical Theory and Computational Aspects of Smoothing. Heidelberg: Physica-Verlag HD; 1996. p. 10–49.

30. Cori A, Ferguson NM, Fraser C, Cauchemez S. A New Framework and Software to Estimate Time-Varying Reproduction Numbers during Epidemics. American Journal of Epidemiology. 2013;178(9):1505–1512.

31. Ferguson NM, Cummings DAT, Cauchemez S, Fraser C, Riley S, Meeyai A, et al. Strategies for Containing an Emerging Influenza Pandemic in Southeast Asia. Nature. 2005;437(7056):209–214.

32. Carrat F, Vergu E, Ferguson NM, Lemaitre M, Cauchemez S, Leach S, et al. Time Lines of Infection and Disease in Human Influenza: A Review of Volunteer Challenge Studies. American Journal of Epidemiology. 2008;167(7):775–785.

33. Matheson JE, Winkler RL. Scoring Rules for Continuous Probability Distributions. Management Science. 1976;22(10):1087–1096.

34. Jordan A, Krüger F, Lerch S. Evaluating Probabilistic Forecasts with scoringRules. Journal of Statistical Software. 2019;90:1–37.

35. Bosse NI, Gruson H, Cori A, van Leeuwen E, Funk S, Abbott S. Evaluating Forecasts with Scoringutils in R. arXiv; 2022.

36. Landau WM. The Targets R Package: A Dynamic Make-like Function-Oriented Pipeline Toolkit for Reproducibility and High-Performance Computing. Journal of Open Source Software. 2021;6(57):2959.

37. Jacobs BK, Goetghebeur E, Clement L. Impact of Variance Components on Reliability of Absolute Quantification Using Digital PCR. BMC Bioinformatics. 2014;15(1):283.

38. Asmussen S, Jensen JL, Rojas-Nandayapa L. On the Laplace Transform of the Lognormal Distribution. Methodology and Computing in Applied Probability. 2016;18(2):441–458.

39. Graham KE, Loeb SK, Wolfe MK, Catoe D, Sinnott-Armstrong N, Kim S, et al. SARS-CoV-2 RNA in Wastewater Settled Solids Is Associated with COVID-19 Cases in a Large Urban Sewershed. Environmental Science & Technology. 2021;55(1):488–498.

40. Greenwald HD, Kennedy LC, Hinkle A, Whitney ON, Fan VB, Crits-Christoph A, et al. Tools for Interpretation of Wastewater SARS-CoV-2 Temporal and Spatial Trends Demonstrated with Data Collected in the San Francisco Bay Area. Water Research X. 2021;12:100111.

41. Safford H, Zuniga-Montanez RE, Kim M, Wu X, Wei L, Sharpnack J, et al. Wastewater-Based Epidemiology for COVID-19: Handling qPCR Nondetects and Comparing Spatially Granular Wastewater and Clinical Data Trends. ACS ES&T Water. 2022;2(11):2114–2124.

42. Jahne MA, Brinkman NE, Keely SP, Zimmerman BD, Wheaton EA, Garland JL. Droplet Digital PCR Quantification of Norovirus and Adenovirus in Decentralized Wastewater and Graywater Collections: Implications for Onsite Reuse. Water Research. 2020;169:115213.

43. Limpert E, Stahel WA, Abbt M. Log-Normal Distributions across the Sciences: Keys and Clues: On the Charms of Statistics, and How Mechanical Models Resembling Gambling Machines Offer a Link to a Handy Way to Characterize Log-Normal Distributions, Which Can Provide Deeper Insight into Variability and Probability—Normal or Log-Normal: That Is the Question. BioScience. 2001;51(5):341–352.

44. Emslie KR, H McLaughlin JL, Griffiths K, Forbes-Smith M, Pinheiro LB, Burke DG. Droplet Volume Variability and Impact on Digital PCR Copy Number Concentration Measurements. Analytical Chemistry. 2019;91(6):4124–4131.

45. Kelly RP, Lodge DM, Lee KN, Theroux S, Sepulveda AJ, Scholin CA, et al. Toward a National eDNA Strategy for the United States. Environmental DNA. 2024;6(1):e432.

46. Forootan A, Sjöback R, Björkman J, Sjögreen B, Linz L, Kubista M. Methods to Determine Limit of Detection and Limit of Quantification in Quantitative Real-Time PCR (qPCR). Biomolecular Detection and Quantification. 2017;12:1–6.

Declarations

Ethics approval

Ethics approval was not required for this study.

Competing interests

The authors declare no competing interests.

Funding

The project was supported by the Swiss National Science foundation through a Sinergia grant (CRSII5_205933 / 1; WISE: Wastewater-based Infectious Disease Surveillance).

Author Contributions

Conceptualisation: AL

Data curation: AL

Formal Analysis: AL

Funding acquisition: TJ, TS

Investigation: AL

Methodology: AL

Project administration: AL

Resources: TS

Software: AL

Supervision: TJ, TS

Validation: AL

Visualisation: AL

Writing – original draft: AL

Writing – review & editing: AL, TJ, TS

Acknowledgements

This work contains a secondary analysis of open wastewater data provided by EAWAG, Swiss Federal Institute of Aquatic Science and Technology. We kindly thank Lea Caduff, Sheena Conforti, Seju Kang, Jolinda de Korne-Elenbaas, Charles Gan, Rachel McLeod, and Melissa Pitton for their contributions to the development of the underlying digital PCR assays and their application in ongoing wastewater surveillance to produce these data. We thank the Swiss Federal Office of Public Health FOPH for supporting the wastewater surveillance by EAWAG.

Data availability

All data, code for simulation, and analysis scripts are publicly available at https://github.com/adrian-lison/dPCR-observation-model-study.

Electronic supplementary material: Improving inference in wastewater-based epidemiology by modelling the statistical features of digital PCR

Adrian Lison[1,2*], Timothy R. Julian[3,4,5], and Tanja Stadler[1,2]

[1]ETH Zurich, Department of Biosystems Science and Engineering, Zurich, Switzerland

[2]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland

[3]Eawag, Swiss Federal Institute of Aquatic Science and Technology, 8600 Dubendorf, Switzerland

[4]Swiss Tropical and Public Health Institute, 4123 Allschwil, Switzerland

[5]University of Basel, 4055 Basel, Switzerland

[*]Corresponding author: adrian.lison@bsse.ethz.ch

Contents

A Digital PCR data **2**

B Averaging of dPCR measurements **2**

C Coefficient of variation of concentration estimates **5**

C.1 CV without pre-PCR variation . . . 5

C.2 CV with pre-PCR variation . . . 6

Gamma distribution ▪ Log-normal distribution ▪ Taylor series expansion

D Probability of non-detection **8**

D.1 Relationship between probability of non-detection and limit of detection . . . 8

D.2 Probability of non-detection with pre-PCR variation . . . 9

E Sensitivity analyses **9**

E.1 Coefficient of variation . . . 10

E.2 Probability of non-detection . . . 13

F Likelihoods for non-zero dPCR measurements **14**

F.1 Conditional distribution of non-zero measurements . . . 14

F.2 Continuous vs. discrete measurement distributions . . . 16

F.3 Comparison of continuous measurement distributions . . . 17

G Inference from dPCR measurements **18**

G.1 PCR parameters 18

Prior for number of partitions ▪ Prior for conversion factor ▪ Estimated parameters

G.2 Estimation 20

G.3 Additional results 21

References **28**

A Digital PCR data

We used dPCR data of pathogen concentrations in Swiss wastewater to validate our statistical model of dPCR measurements and to demonstrate inference from a time series of measurements. For the validation of our model, we used data from 3031 autosampler-based, 24-hour-composite samples taken at 14 different municipal treatment plants throughout Switzerland in the time period July 2023 – June 2024. For each sample, viral RNA was extracted from a 40 mL aliquot by vacuum filtration, yielding an elution of 80 µL, and subsequently diluted by a factor of 1:3. As described in Huisman et al. [1] and Nadeau et al. [2], for quantification via dPCR, a 25 µL reaction mix with 5 µL of the template and 20 µL of reagents was run in a droplet-based, sixplex dPCR assay with Influenza A (M gene), Influenza B (M gene), Respiratory Syncytial Virus (N gene), SARS-CoV-2 (N1 and N2 genes), and an internal extraction efficiency control as targets. Quantification was performed using the Naica Crystal Digital PCR System (Stilla® Technologies) with a Sapphire Chip with a maximum number of 30000 droplets and an average droplet volume of 0.519 nL. To demonstrate inference from a time series of measurements, we used samples taken on average 5 times per week at the treatment plants of Zurich and Geneva, Switzerland in the time period Oct 19, 2022 – April 30, 2023, extracted under the same protocol as above and quantified using a similar, fourplex dPCR assay with Influenza A (M gene), Influenza B (M gene), Respiratory Syncytial Virus (N gene), and SARS-CoV-2 (N2 gene) as targets.

B Averaging of dPCR measurements

Given replicate measurements $i \in \{1, \ldots, n\}$ of a sample with concentration λ, where the replicate runs had $m_1, \ldots, m_n$ total partitions and $y_1, \ldots, y_n$ positive partitions, respectively, the maximum

likelihood (ML) estimate for λ [3] is

$$\hat{\lambda}^{\mathrm{ML}} = -\frac{1}{c}\log\left(1-\frac{\overline{y}}{\overline{m}}\right) = -\frac{1}{c}\log\left(1-\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}\right). \tag{1}$$

Importantly, this is not equal to the arithmetic mean of individual ML estimates for each replicate

$$\hat{\lambda}^{\mathrm{AVG}} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\lambda}_i = \frac{1}{n}\sum_{i=1}^{n}\left(-\frac{1}{c}\log\left(1-\frac{y_i}{m_i}\right)\right), \tag{2}$$

which is also used in practice. In essence, the ML estimate in Eq. (1) differs from the arithmetic mean of estimates in Eq. (2) in two regards. First, the ML estimate computes an average of the estimated probabilities $p_1, \ldots, p_n$ for positive partitions, not an average of the concentrations. Second, this average is weighted by the total number of partitions of each replicate, while the arithmetic mean gives equal weight to all replicates.

To assess the magnitude of these differences, we simulated duplicate dOCR measurements for various concentrations. While the first replicate was assumed to have 25000 total partitions, we simulated the second replicate to have between 5000 and 25000 partitions. This way, we could study the effect of differences in the total number of partitions. As in the other simulations, we assumed a scaling factor of $s = 30$ and a partition volume of $v = 0.519\,\mathrm{nL}$. As a first step, we compared the arithmetic mean estimate $\hat{\lambda}^{\mathrm{AVG}}$ with an unweighted ML estimate

$$\hat{\lambda}^{\mathrm{uML}} = -\frac{1}{c}\log\left(1-\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{m_i}\right), \tag{3}$$

in which each replicate receives the same weight. Thereby, we could isolate the effect of averaging concentrations instead of averaging probabilities. Figure 1 shows how the arithmetic mean of measurements deviates from the unweighted ML estimate. As can be seen, the arithmetic mean has a small positive bias compared to the unweighted ML estimate, because the concentration estimate is a non-linear function of the probability of positive droplets, such that the average of concentrations is not equal to the concentration estimate based on the average probability. The potential deviation increases linearly with the concentration in the PCR, however even for high concentrations of 1000 gc/mL, the deviation is considerably less than 1 gc/mL.

Next, we compared the unweighted ML estimate defined in Eq. (3) to the weighted ML estimate

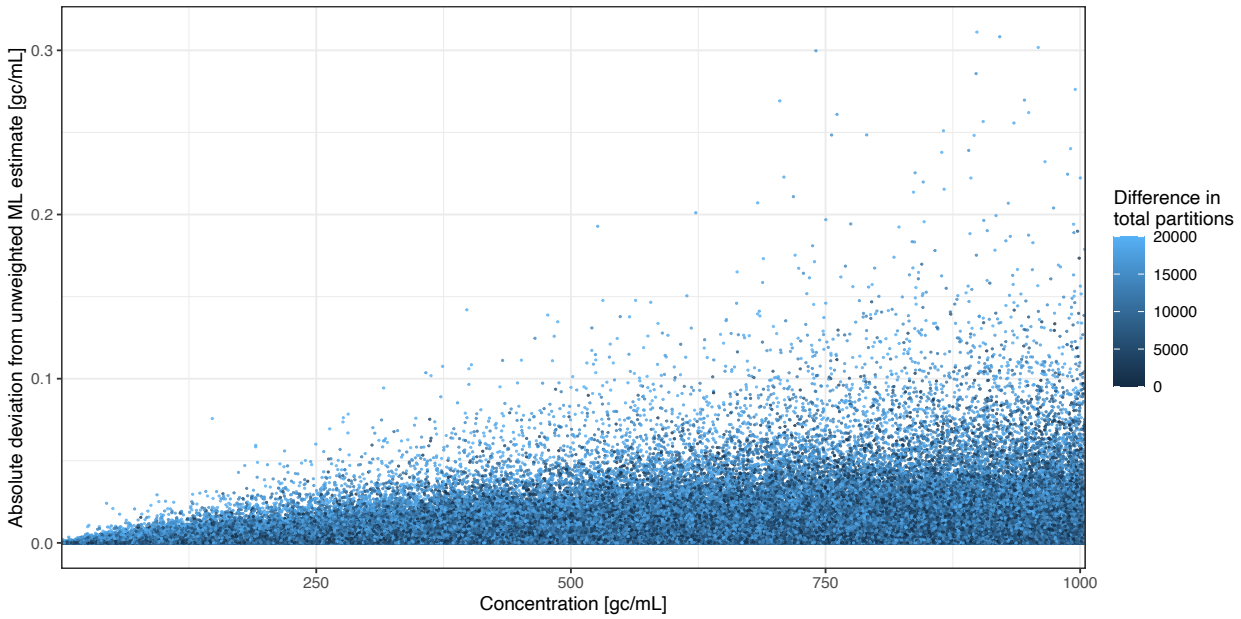

Fig 1. Comparison of arithmetic mean estimate with unweighted ML estimate of concentrations. Shown are deviations of the arithmetic mean of replicate measurements from an unweighted ML estimate for simulated measurements as a function of concentration, and for different degrees of partition number variation between replicates. Simulated measurements assume a scaling factor of $s = 30$ and partition volume of $v = 0.519\,\mathrm{nL}$.

defined in Eq. (1). As can be seen in Figure 2, larger deviations than between the arithmetic mean and unweighted ML estimate occur, but no systematic bias. Deviations are particularly large in cases where one of the replicates only has a small number of total partitions, such that the corresponding concentration estimate is noisy. In the arithmetic mean and the unweighted ML estimate, such replicates are weighted equally to other replicates. In contrast, the ML estimate defined in Eq. (1) gives more weight to replicates with more partitions, making it less sensitive to noisy measurements based on few partitions. We note that in practice, however, differences in the total number of partitions larger than 5000 do not occur frequently in dPCR assays. For smaller differences in the total number of partitions, the differences between unweighted and weighted estimates are considerably smaller. Overall, we conclude that the ML estimate of multiple replicates has higher accuracy and precision than the arithmetic mean of their individual concentration estimates, but that in practice the differences will be rather small compared to the overall uncertainty of the measurements, except in cases where the total number of partitions varies greatly between replicates.

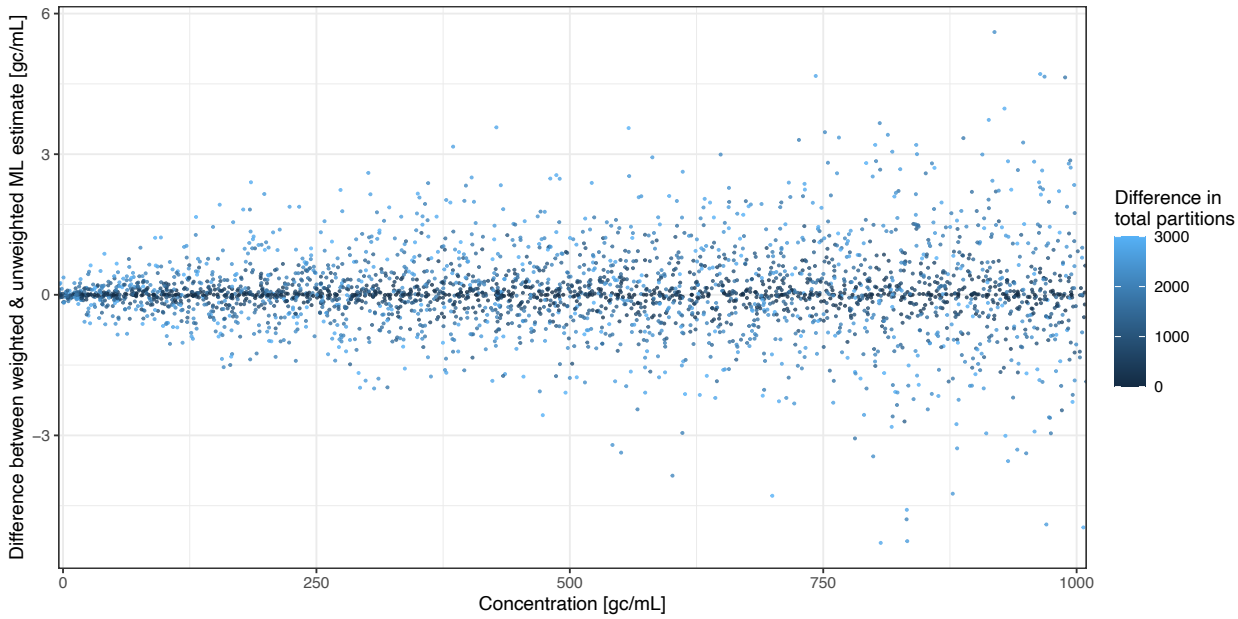

Fig 2. Comparison of unweighted and weighted ML estimates of concentrations. Shown are deviations of the unweighted ML estimate from the weighted ML estimate for simulated measurements as a function of concentration, and for different degrees of partition number variation between replicates. Simulated measurements assume a scaling factor of $s = 30$ and partition volume of $v = 0.519\,\text{nL}$.

C Coefficient of variation of concentration estimates

C.1 CV without pre-PCR variation

To derive the variance of a single maximum likelihood concentration estimate $\tilde{\Lambda} = -\frac{1}{c}\log\left(1-\frac{Y}{m}\right)$ [3], we use the approximation $\mathrm{Var}[\log(X)] \approx \frac{\mathrm{Var}[X]}{\mathbb{E}[X]^2}$ to obtain

$$\mathrm{Var}[\tilde{\Lambda}] = \mathrm{Var}\left[-\frac{1}{c}\log\left(1-\frac{Y}{m}\right)\right] = \frac{1}{c^2}\mathrm{Var}\left[\log\left(1-\frac{Y}{m}\right)\right] \approx \frac{1}{c^2 m^2}\frac{\mathrm{Var}[Y]}{\left(1-\mathbb{E}\left[\frac{Y}{m}\right]\right)^2}. \tag{4}$$

Since Y is Binomially distributed with $\mathrm{Var}[Y] = m\,p\,(1-p)$ and $\mathbb{E}\left[\frac{Y}{m}\right] = p$, we further get

$$\mathrm{Var}[\tilde{\Lambda}] \approx \frac{1}{c^2 m^2}\frac{\mathrm{Var}[Y]}{(1-\mathbb{E}\left[\frac{Y}{m}\right])^2} = \frac{1}{c^2 m^2}\frac{m\,p\,(1-p)}{(1-p)^2} = \frac{1}{c^2 m}\frac{p}{(1-p)}. \tag{5}$$

Finally, using the probability for a positive partition $p = 1-\exp(-\lambda c)$ [3], we obtain

$$\mathrm{Var}[\tilde{\Lambda}] \approx \frac{1}{c^2 m}\frac{1-\exp(-\lambda c)}{(\exp(-\lambda c))} = \frac{1}{c^2 m}(\exp(\lambda c)-1). \tag{6}$$

That is, the variance of $\tilde{\Lambda}$ increases exponentially with the concentration in the PCR. This relationship between the variance of $\tilde{\Lambda}$ and λ can also be obtained through the Fisher information of $\tilde{\Lambda}$ [3].

If we have multiple technical replicates $i \in \{1, \ldots, n\}$, the combined maximum likelihood estimate by Dorazio and Hunter as described in Eq. (1) corresponds to a pooled analysis where the m_i partitions from each well are combined into one large assay. Assuming an equal number of partitions $m_i = m$ for all replicates, the variance of this estimate is simply

$$\mathrm{Var}[\Lambda^{\hat{\mathrm{ML}}}] = \frac{1}{c^2 \sum_{i=1}^{n} m_i} (\exp(\lambda c) - 1) = \frac{1}{c^2 mn} (\exp(\lambda c) - 1). \tag{7}$$

Moreover, if all $m_i = m$, the concentration estimates from individual replicates have identical variances $\mathrm{Var}[\tilde{\Lambda}_i]$, such that the their arithmetic mean as described in Eq. (2) has variance

$$\mathrm{Var}[\hat{\Lambda}^{\mathrm{AVG}}] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} \hat{\Lambda}_i\right] = \frac{1}{n^2}\left(n \frac{1}{c^2 m} (\exp(\lambda c) - 1)\right) = \frac{1}{c^2 mn} (\exp(\lambda c) - 1), \tag{8}$$

which is identical to the variance of the maximum likelihood estimate in Eq. (7).

Since $\hat{\Lambda}$ is an asymptotically unbiased estimator of the true concentration [4], i. e.

$$\lim_{m \to \infty} \mathbb{E}[\hat{\Lambda}_m] = \lambda, \tag{9}$$

the coefficient of variation of $\hat{\Lambda}$ can be estimated as

$$\nu_{\hat{\Lambda}}(\lambda) = \frac{1}{c\sqrt{mn}} \frac{\sqrt{\exp(\lambda c) - 1}}{\lambda}. \tag{10}$$

C.2 CV with pre-PCR variation

As described in the main text, we account for pre-PCR noise by extending the above model such that the concentration in the PCR is $s\Lambda_{\mathrm{pre}}$, where Λ_{pre} is a random variable drawn from a parametric distribution with mean λ and coefficient of variation $\nu_{\mathrm{pre}} = \frac{\sigma_{\mathrm{pre}}}{\lambda} > 0$. Using the Law of Total Variance, we can express the variance of $\hat{\Lambda}$ under pre-PCR noise as

$$\mathrm{Var}[\hat{\Lambda}] = \mathrm{Var}[\mathbb{E}[\hat{\Lambda}|\Lambda_{\mathrm{pre}}]] + \mathbb{E}[\mathrm{Var}[\hat{\Lambda}|\Lambda_{\mathrm{pre}}]], \tag{11}$$

with the "explained" component of the variance

$$\mathrm{Var}[\mathbb{E}[\hat{\Lambda}|\Lambda_{\mathrm{pre}}]] \overset{\text{Eq. (9)}}{=} \mathrm{Var}[\Lambda_{\mathrm{pre}}] = \lambda^2\,\nu_{\mathrm{pre}}^2, \tag{12}$$

and the "unexplained" component of the variance

$$\mathbb{E}[\mathrm{Var}[\hat{\Lambda}|\Lambda_{\mathrm{pre}}]] = \mathbb{E}\left[\frac{1}{c^2 mn}\left(\exp(\Lambda_{\mathrm{pre}}c)-1\right)\right] = \frac{1}{c^2 mn}(\mathbb{E}[\exp(\Lambda_{\mathrm{pre}}c)]-1). \tag{13}$$

C.2.1 Gamma distribution

If we model Λ_{pre} as gamma distributed with parameters $\alpha = \frac{1}{\nu_{\mathrm{pre}}^2}$ and $\beta = \frac{1}{\lambda\nu_{\mathrm{pre}}^2}$, an exact solution for the "unexplained" component of the variance can be obtained. That is, by the MGF of the gamma distribution, we know that

$$\mathbb{E}[\exp(\Lambda_{\mathrm{pre}}c)] = M_{\Lambda_{\mathrm{pre}}}(c) = \left(1-\frac{c}{\beta}\right)^{-\alpha} = (1-\lambda\,c\,\nu_{\mathrm{pre}}^2)^{-\frac{1}{\nu_{\mathrm{pre}}^2}}. \tag{14}$$

Substituting back, we get the variance

$$\mathrm{Var}[\hat{\Lambda}] = \lambda^2\,\nu_{\mathrm{pre}}^2 + \frac{1}{c^2 mn}\left((1-\lambda\,c\,\nu_{\mathrm{pre}}^2)^{-\frac{1}{\nu_{\mathrm{pre}}^2}}-1\right) \tag{15}$$

and thus

$$\nu_{\hat{\Lambda}}(\lambda) = \sqrt{\nu_{\mathrm{pre}}^2 + \frac{1}{\lambda^2 c^2 mn}\left((1-\lambda\,c\,\nu_{\mathrm{pre}}^2)^{-\frac{1}{\nu_{\mathrm{pre}}^2}}-1\right)}. \tag{16}$$

C.2.2 Log-normal distribution

We can also model Λ_{pre} as log-normally distributed with parameters $\tilde{\mu} = \log(\lambda) - \frac{\log(1+\nu_{\mathrm{pre}}^2)}{2}$ and $\tilde{\sigma}^2 = \log(1+\nu_{\mathrm{pre}}^2)$. While no closed-form solution for the MGF of the log-normal distribution exists, we can use an approximation [5] to obtain

$$\mathbb{E}[\exp(\Lambda_{\mathrm{pre}}c)] \approx \frac{1}{\sqrt{1+w}}\exp\left(-\frac{(w^2+2\,w)}{2\,\tilde{\sigma}^2}\right) = \frac{1}{\sqrt{1+w}}\exp\left(-\frac{(w^2+2\,w)}{2\log(1+\nu_{\mathrm{pre}}^2)}\right), \tag{17}$$

where w is an evaluation of the Lambert W function [6] with

$$w = W\left(c\tilde{\sigma}^2 \exp(\tilde{\mu})\right) = W\left(\lambda c \frac{\log(1+\nu_{\text{pre}}^2)}{\sqrt{1+\nu_{\text{pre}}^2}}\right). \tag{18}$$

C.2.3 Taylor series expansion

If only the mean and coefficient of variation of Λ_{pre} are known, we can approximate $E[\exp(\Lambda_{\text{pre}}c)]$ using a Taylor polynomial, i. e.

$$\mathbb{E}[\exp(\Lambda_{\text{pre}}c)] \approx 1 + \mathbb{E}[\Lambda_{\text{pre}}c] + \frac{\mathbb{E}[\Lambda_{\text{pre}}c]^2 + \text{Var}[\Lambda_{\text{pre}}c]}{2} = 1 + \lambda c + \frac{1}{2}\lambda^2 c^2 (1+\nu_{\text{pre}}^2). \tag{19}$$

Substituting this back into Eq. (11), we obtain

$$\text{Var}[\hat{\Lambda}] = \lambda^2 \nu_{\text{pre}}^2 + \frac{1}{c^2 mn}\left(1 + \lambda c + \frac{1}{2}\lambda^2 c^2 (1+\nu_{\text{pre}}^2) - 1\right) = \lambda^2 \left(\nu_{\text{pre}}^2 + \frac{1}{\lambda cmn} + \frac{1+\nu_{\text{pre}}^2}{2nm}\right) \tag{20}$$

and thus

$$\nu_{\hat{\Lambda}}(\lambda) = \sqrt{\nu_{\text{pre}}^2 + \frac{1}{\lambda cmn} + \frac{1+\nu_{\text{pre}}^2}{2nm}}. \tag{21}$$

D Probability of non-detection

D.1 Relationship between probability of non-detection and limit of detection

The Limit of Detection (LoD) of an assay is typically defined as the lowest concentration λ_{LOD} at which detection is feasible with sufficient confidence α [7]. If we leave aside the probability of false positive partitions, this corresponds to the converse of the probability of a zero measurement, i. e. $\alpha = 1 - p_{\text{zero}}(\lambda_{\text{LOD}}) = 1 - \exp(-\lambda_{\text{LOD}} cmn)$.

If the specific parameters of an assay are unknown, but the LoD is reported, it may thus be possible to obtain a rough estimate of cmn from the reported LoD via

$$cmn = -log(1-\alpha)/\lambda_{\text{LOD}}. \tag{22}$$

Once an estimate of cmn is available, the probability of non-detection can be computed for arbitrary

concentrations in the assay. This may be more accurate than censoring measurements below the LoD, as discussed in the main text. On the other hand, as the details of the experiment used to establish the LoD may also be unknown and could violate the above simple relationship, we believe it is generally better to jointly estimate the unknown PCR parameters from the provided non-zero measurements, as described in the Methods.

D.2 Probability of non-detection with pre-PCR variation

As described in the main text, the probability of non-detection under pre-PCR variation is

$$p_{\text{zero}}(\lambda) = \mathbb{E}_{\Lambda_{\text{pre}}}\left[\prod_{i=1}^{n} \Pr(\hat{\Lambda} = 0 | \Lambda_{\text{pre}} = \lambda_{\text{pre}})\right] = \mathbb{E}[\exp(-\Lambda_{\text{pre}} cmn)], \tag{23}$$

where, as before, Λ_{pre} is drawn from a parametric distribution with mean λ and coefficient of variation $\nu_{\text{pre}} = \frac{\sigma_{\text{pre}}}{\lambda} > 0$. Moreover, $\mathbb{E}[\exp(-\Lambda_{\text{pre}} cmn)]$ is again an evaluation of the MGF of Λ_{pre}. As the Taylor series expansion of Eq. (23) is alternating and its partial sum up to the 2nd degree has a large error, we cannot provide a suitable approximation without further assumptions about Λ_{pre}. However, if we assume that Λ_{pre} is gamma distributed, we can again obtain an exact solution using its MGF, i. e.

$$\mathbb{E}[\exp(-\Lambda_{\text{pre}} cmn)] = M_{\Lambda_{\text{pre}}}(-cmn) = \left(1 - \frac{-cmn}{\beta}\right)^{-\alpha} = (1 + \lambda cmn\, \nu_{\text{pre}}^2)^{-\frac{1}{\nu_{\text{pre}}^2}}. \tag{24}$$

If we instead assume that Λ_{pre} is log-normally distributed, we can use the approximation of its MGF as in Eq. (17), but now with

$$w = W\left(-\lambda cmn \frac{\log(1 + \nu_{\text{pre}}^2)}{\sqrt{1 + \nu_{\text{pre}}^2}}\right). \tag{25}$$

E Sensitivity analyses

In the following section, we provide results of additional comparisons and sensitivity analyses for the coefficient of variation and probability of non-detection with respect to Figure 1 in the main text.

E.1 Coefficient of variation

Figure 3 shows the coefficient of variation (CV) as a function of the sample concentration under log-normally instead of gamma distributed pre-PCR noise. In addition to the CV of simulated measurements, to theoretical approximations are shown. While dashed lines show an approximation of the log-normal MGF by Asmussen et al. [5] (see Supplement C.2.2), straight lines show the CV as approximated using a Taylor series expansion (see Supplement C.2.3). As can be seen, the approximations yield fairly similar results, however, we find that for high levels of pre-PCR variation, the MGF-based approximation gives slightly lower CV values than the Taylor series-based approximation.

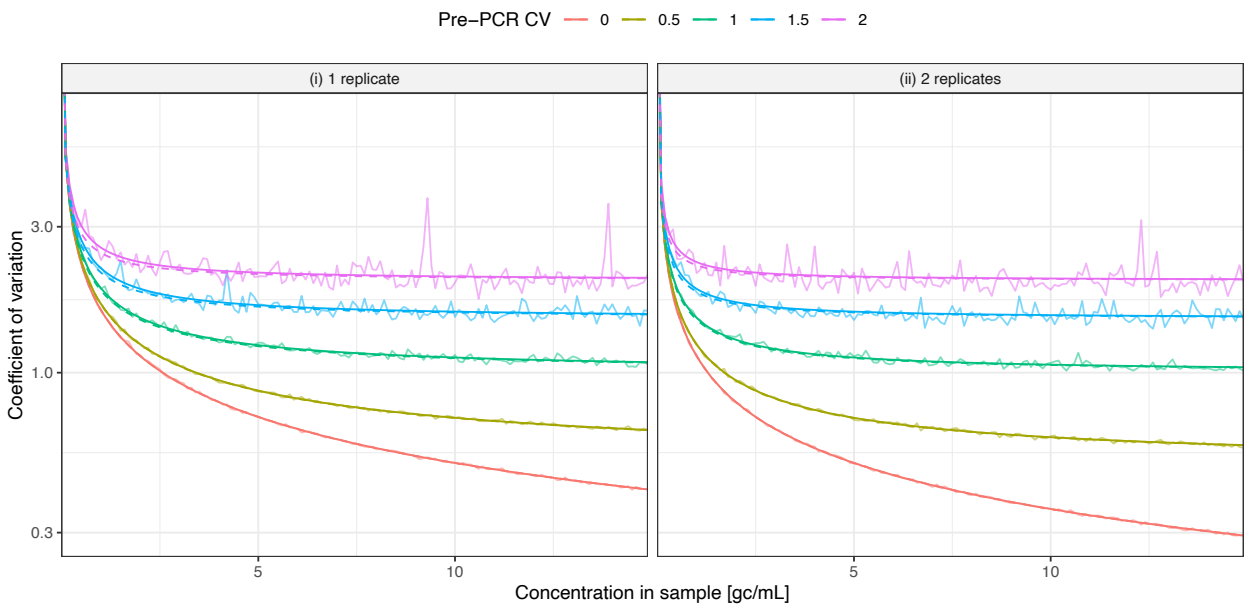

Fig 3. Theoretical relationship betweeen concentration and coefficient of variation of dPCR measurements under log-normally distributed pre-PCR noise. Shown is the coefficient of variation (CV) of measurements from digital PCR for (i) a single replicate and (ii) the arithmetic mean of two replicates, as a function of the sample concentration λ under log-normally distributed pre-PCR noise. Jagged lines show the CV of simulated measurements for different concentrations in steps of 0.1 gc/mL. Straight lines show the corresponding CV as approximated using a Taylor series expansion, and dashed lines show an alternative approximation of the log-normal MGF by Asmussen et al. [5]. Colors correspond to different strengths of pre-PCR noise, as measured by the pre-PCR coefficient of variation ν_{pre}. Simulated and predicted measurements assume a scaling factor of $s = 30$, a partition volume of $v = 0.519\,\text{nL}$, and $m = 25000$ partitions per measurement.

Figure 4 compares the exact solution for gamma distributed pre-PCR noise with the log-normal MGF approximation and the Taylor series-based approximation across a wide range of concentrations. We note that the upper limit of concentrations shown (900 000 gc/mL) corresponds to an extreme

case where the expected number of gene copies per partition is 15 or higher (in practice, dPCR assays aim for concentrations that correspond to an expected number of gene copies well below 1, which can be achieved via dilution). As the share of positive partitions in the PCR approaches 1, the CV increases considerably. This increase in CV at extreme concentrations is not captured by the Taylor series expansion approximation but by the exact solution under gamma noise and the MGF approximation under log-normal noise. There are small differences in the CV between the gamma case and the log-normal approximation.

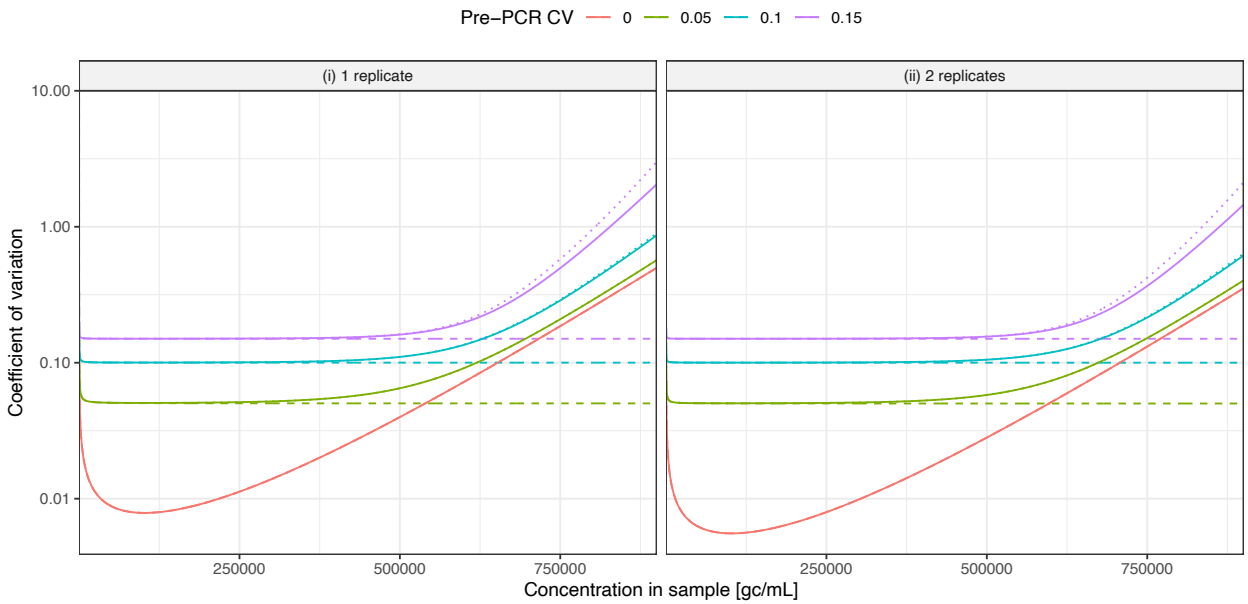

Fig 4. Theoretical relationship betweeen concentration and coefficient of variation of dPCR measurements across wide range of concentrations. Shown is the coefficient of variation (CV) of measurements from digital PCR for (i) a single replicate and (ii) the arithmetic mean of two replicates, as a function of the sample concentration λ under gamma distributed pre-PCR noise for different concentrations. Solid lines show the CV as predicted under gamma distributed pre-PCR noise, dashed lines show an approximation using a Taylor series expansion, and dotted lines show an approximation under log-normally distributed pre-PCR noise. Colors correspond to different strengths of pre-PCR noise, as measured by the pre-PCR coefficient of variation ν_{pre}. Predicted measurements assume a scaling factor of $s = 30$, a partition volume of $v = 0.519\,\text{nL}$, and $m = 25000$ partitions per measurement.

Figure 5 shows a sensitivity analysis of the CV as a function of the sample concentration to the conversion factor c for the dPCR assay. Compared are results for a conversion factor 10x as large (green), and 10x as small (blue) as assumed in the main analysis (red). The simulated and predicted CV values are based on an assumed pre-PCR coefficient of variation of 0.5. As can be seen, a larger conversion factor reduces the CV, while a smaller conversion factor increases the CV.

Figure 6 shows a sensitivity analysis of the CV as a function of the sample concentration to

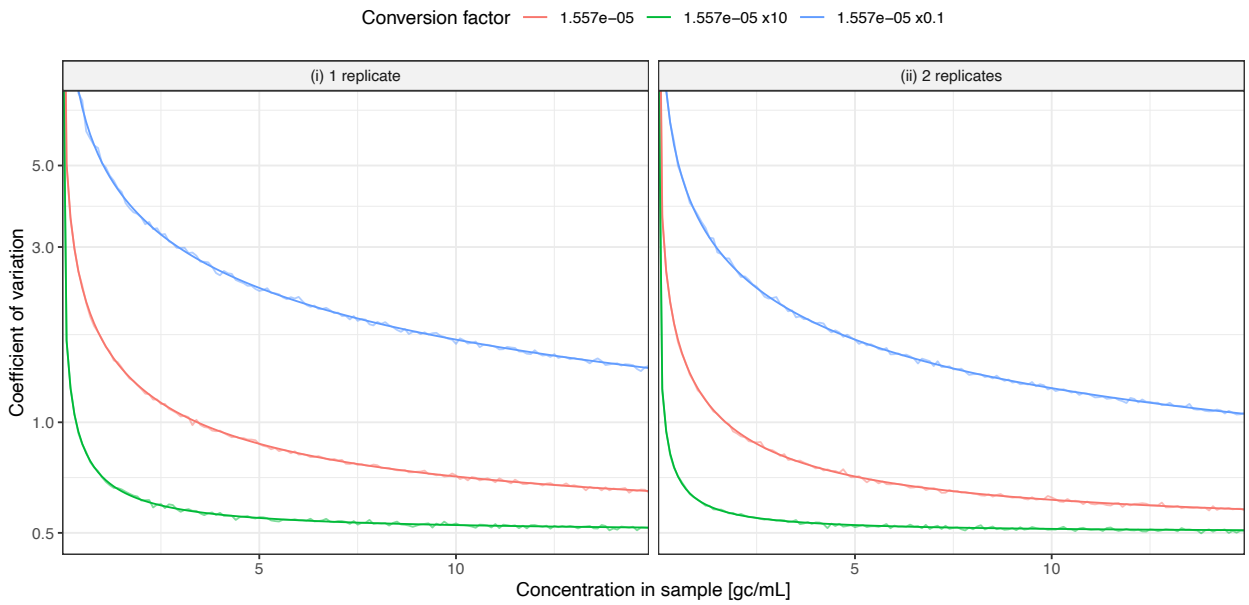

Fig 5. Sensitivity of the coefficient of variation of dPCR measurements to the conversion factor. Shown is the coefficient of variation (CV) of measurements from digital PCR for a conversion factor as in the main analysis ($c = 30 \cdot 0.519\,\text{nL}$, red), 10x as large (green), and 10x as small (blue). The CV is shown for (i) a single replicate and (ii) the arithmetic mean of two replicates, as a function of the sample concentration λ under gamma distributed pre-PCR noise with a pre-PCR coefficient of variation of 0.5. Jagged lines show the CV of simulated measurements for different concentrations in steps of 0.1 gc/mL, straight lines show the corresponding CV as predicted by Eq. (16). Simulated and predicted measurements assume $m = 25000$ partitions per measurement.

the total number of droplets in the dPCR assay. Compared are results for 15000 partitions less (green), and 15000 partitions more (blue) as assumed in the main analysis (red). The simulated and predicted CV values are based on an assumed pre-PCR coefficient of variation of 0.5. As can be seen, a larger number of partitions reduces the CV, while a smaller number of partitions increases the CV.

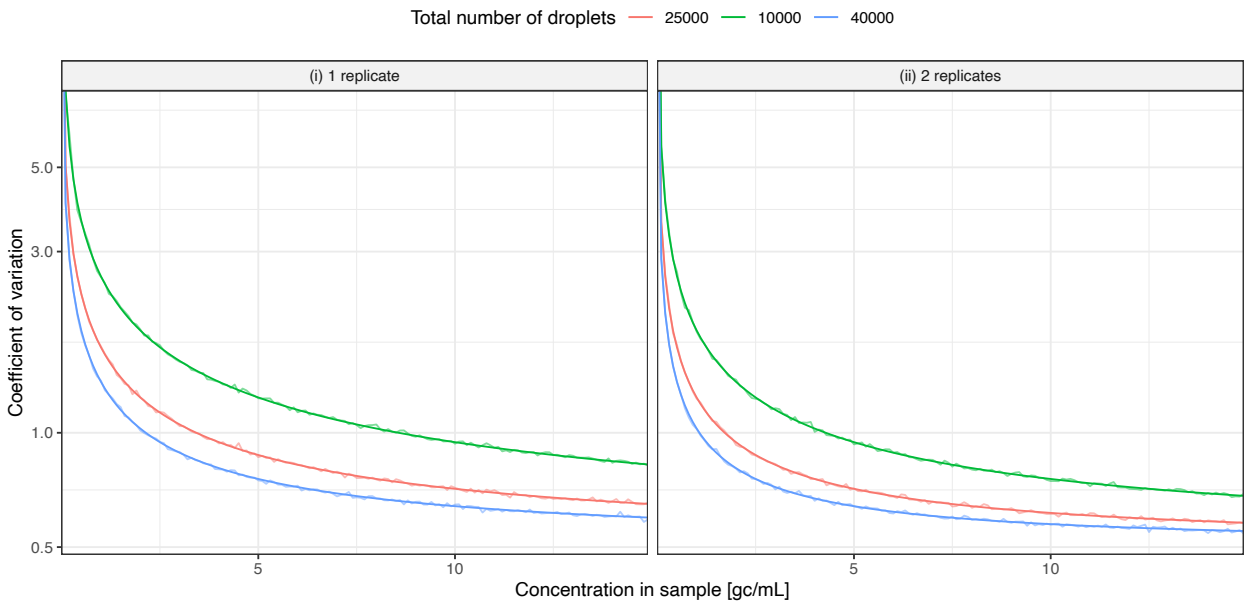

Fig 6. Sensitivity of the coefficient of variation of dPCR measurements to the total number of partitions. Shown is the coefficient of variation (CV) of measurements from digital PCR for 25000 total partitions, i. e. as in the main analysis (red), 10000 total partitions (green), and 40000 total partitions (blue). The CV is shown for i) a single replicate and (ii) the arithmetic mean of two replicates, as a function of the sample concentration λ under gamma distributed pre-PCR noise with a pre-PCR coefficient of variation of 0.5. Jagged lines show the CV of simulated measurements for different concentrations in steps of 0.1 gc/mL, straight lines show the corresponding CV as predicted by Eq. (16). Simulated and predicted measurements assume a scaling factor of $s = 30$ and a partition volume of $v = 0.519\,\text{nL}$.

E.2 Probability of non-detection

Figure 7 shows the probability of non-detection as a function of the sample concentration under log-normally instead of gamma distributed pre-PCR noise. In addition to the CV of simulated measurements, a theoretical prediction using an approximation of the log-normal MGF by Asmussen et al. [5] (see Supplement D.2) is shown. As can be seen, the approximation matches the simulated CV values well. Compared to the case of gamma distributed pre-PCR noise (see Figure 1B in the main text), the probability of non-detection falls considerably faster under log-normally distributed pre-PCR noise.

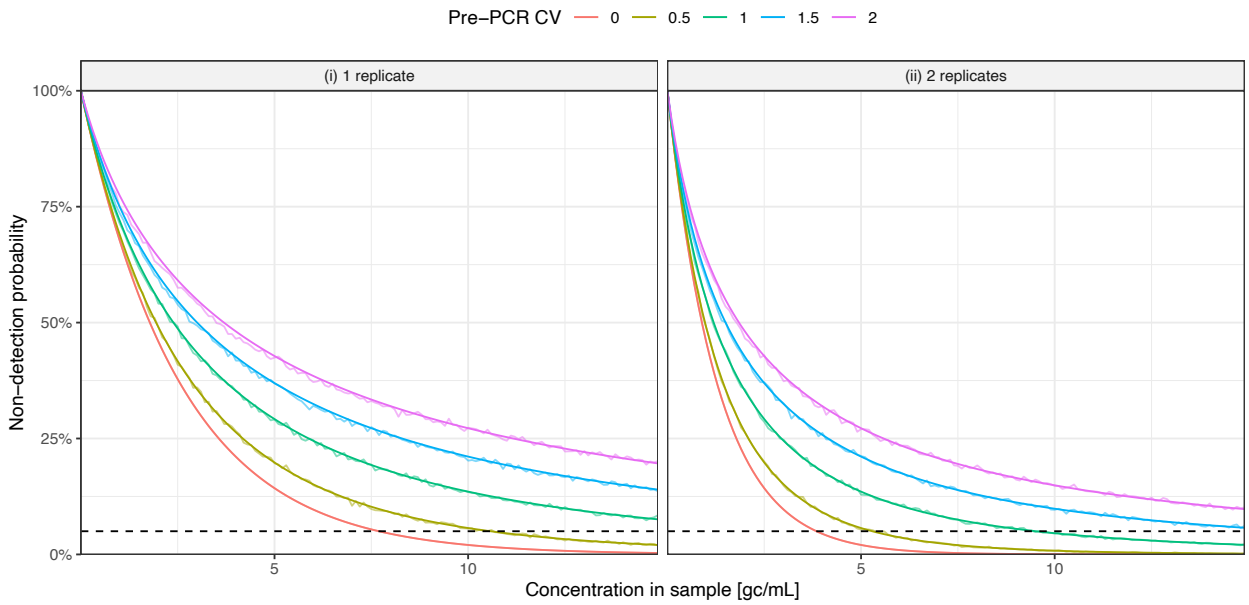

Fig 7. Theoretical relationship betweeen concentration and probability of non-detection of dPCR measurements under log-normally distributed pre-PCR noise. Shown is the probability of a zero measurement (i. e. non-detection) in digital PCR for (i) a single replicate and (ii) the arithmetic mean of two replicates, as a function of the sample concentration λ under log-normally distributed pre-PCR noise. Jagged lines show the probability of non-detection of simulated measurements for different concentrations in steps of 0.1 gc/µL, while straight lines show the corresponding probability of non-detection as predicted using an approximation of the log-normal MGF by Asmussen et al. [5]. Dashed horizontal lines indicate the limit of detection (LoD), defined as a 5% probability of non-detection. Colors correspond to different strengths of pre-PCR noise, as measured by the pre-PCR coefficient of variation ν_{pre}. Simulated and predicted measurements assume a scaling factor of $s = 30$, a partition volume of $v = 0.519\,\text{nL}$, and $m = 25000$ partitions per measurement.

F Likelihoods for non-zero dPCR measurements

F.1 Conditional distribution of non-zero measurements

In the proposed hurdle model, $f_{\text{PCR}}(x)$ is the distribution of measurements conditioned on a non-zero result. While the overall distribution of measurements has mean λ and coefficient of variation $\nu_{\hat{\Lambda}}(\lambda)$, this is not true for the conditional distribution. Instead, using the law of total expectation, i. e.

$$\mathbb{E}\left[\hat{\Lambda}\right] = \mathbb{E}\left[\hat{\Lambda}\middle|\hat{\Lambda} = 0\right] \cdot P(\hat{\Lambda} = 0) + \mathbb{E}\left[\hat{\Lambda}\middle|\hat{\Lambda} > 0\right] \cdot P(\hat{\Lambda} > 0), \tag{26}$$

we know that

$$\mathbb{E}\left[\hat{\Lambda}\middle|\hat{\Lambda} > 0\right] = \frac{\mathbb{E}\left[\hat{\Lambda}\right]}{P(\hat{\Lambda} > 0)} = \frac{\mathbb{E}\left[\hat{\Lambda}\right]}{1 - p_{\text{zero}}(\lambda)} = \frac{\lambda}{1 - p_{\text{zero}}(\lambda)}. \tag{27}$$

That is, the conditional mean increases relative to the unconditional mean as the probability of non-detection becomes large. Moreover, by the same argument, we see that

$$\mathbb{E}\left[\hat{\Lambda}^2 \middle| \hat{\Lambda} > 0\right] = \frac{\mathbb{E}\left[\hat{\Lambda}^2\right]}{P(\hat{\Lambda} > 0)} = \frac{\mathbb{E}\left[\hat{\Lambda}^2\right]}{1 - p_{\text{zero}}(\lambda)}. \tag{28}$$

Expanding the conditional variance and inserting the above, we get

$$\text{Var}\left[\hat{\Lambda} \middle| \hat{\Lambda} > 0\right] = \mathbb{E}\left[\hat{\Lambda}^2 \middle| \hat{\Lambda} > 0\right] - \mathbb{E}\left[\hat{\Lambda} \middle| \hat{\Lambda} > 0\right]^2 = \frac{\mathbb{E}\left[\hat{\Lambda}^2\right]}{1 - p_{\text{zero}}(\lambda)} - \frac{\mathbb{E}\left[\hat{\Lambda}\right]^2}{(1 - p_{\text{zero}}(\lambda))^2}. \tag{29}$$

Then, using $\mathbb{E}\left[\hat{\Lambda}^2\right] = \text{Var}\left[\hat{\Lambda}\right] + \mathbb{E}\left[\hat{\Lambda}\right]^2$, we obtain

$$\text{Var}\left[\hat{\Lambda} \middle| \hat{\Lambda} > 0\right] = \frac{\text{Var}\left[\hat{\Lambda}\right] + \mathbb{E}\left[\hat{\Lambda}\right]^2}{1 - p_{\text{zero}}(\lambda)} - \frac{\mathbb{E}\left[\hat{\Lambda}\right]^2}{(1 - p_{\text{zero}}(\lambda))^2} = \frac{Var\left[\hat{\Lambda}\right]}{1 - p_{\text{zero}}(\lambda)} - \mathbb{E}\left[\hat{\Lambda}\right]^2 \frac{p_{\text{zero}}(\lambda)}{(1 - p_{\text{zero}}(\lambda))^2}. \tag{30}$$

From the conditional variance, we can derive the conditional coefficient of variation as

$$\text{CV}[\hat{\Lambda} | \hat{\Lambda} > 0] = \sqrt{\frac{\text{Var}\left[\hat{\Lambda} \middle| \hat{\Lambda} > 0\right]}{\mathbb{E}\left[\hat{\Lambda} \middle| \hat{\Lambda} > 0\right]^2}} \tag{31}$$

$$= \sqrt{\left(\frac{\text{Var}\left[\hat{\Lambda}\right]}{1 - p_{\text{zero}}(\lambda)} - \mathbb{E}\left[\hat{\Lambda}\right]^2 \frac{p_{\text{zero}}(\lambda)}{(1 - p_{\text{zero}}(\lambda))^2}\right) \frac{(1 - p_{\text{zero}}(\lambda))^2}{\mathbb{E}\left[\hat{\Lambda}\right]^2}} \tag{32}$$

$$= \sqrt{\frac{\text{Var}\left[\hat{\Lambda}\right]}{\mathbb{E}\left[\hat{\Lambda}\right]^2}(1 - p_{\text{zero}}(\lambda)) - p_{\text{zero}}(\lambda)} \tag{33}$$

$$= \sqrt{\nu_{\hat{\Lambda}}(\lambda)^2 (1 - p_{\text{zero}}(\lambda)) - p_{\text{zero}}(\lambda)}. \tag{34}$$

Thus, the conditional CV decreases relative to the unconditional CV as the probability of non-detection becomes large. The resulting relationship between the unconditional and conditional mean and CV is also visualised in Figure 8.

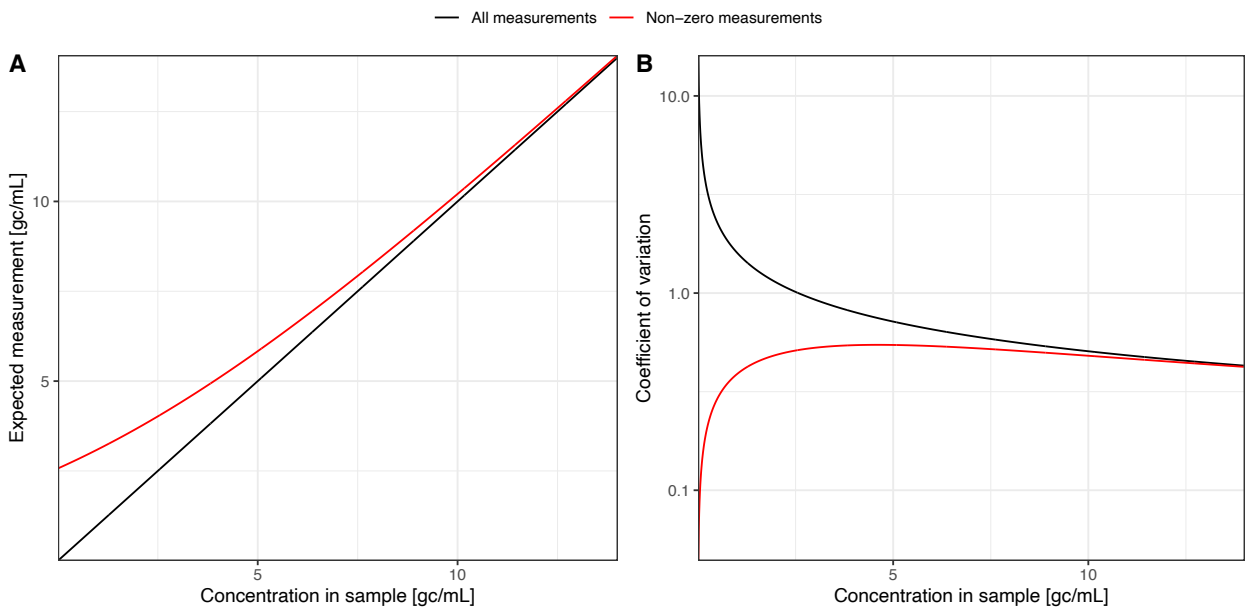

Fig 8. Effect of conditioning on non-zero measurements on the mean and coefficient of variation of measurements. Shown are the (A) mean and (B) coefficient of variation of the unconditional (black, all measurements) and conditional (red, non-zero measurements) distribution of measurements as a function of the true concentration in the sample. The differences between the unconditional and conditional distribution increase as the concentration becomes small.

F.2 Continuous vs. discrete measurement distributions

In theory, if the total number of partitions, the partition volume, and the scaling factor of each measurement are known exactly, the number of positive partitions in the PCR could be back-calculated from the concentration estimate $\tilde{\lambda}$ as

$$y = m(1 - \exp(-\tilde{\lambda}c)), \tag{35}$$

although we note that in such a situation y is likely already known. Given these quantities, it would also be possible to fit a Binomial distribution to the observed y according to Eq (3) from the main text. The goal of this work was however to propose a likelihood model that also works well if the specific parameters of the PCR are not exactly known. In such a situation, the use of a Binomial likelihood has two important drawbacks. First, and most importantly, when using a Binomial distribution, any misspecification or uncertainty about the PCR parameters directly influences the expected concentration estimate, not only its variance. Thus, a lack of knowledge about the PCR parameters can bias concentration estimates even if the measurements $\tilde{\Lambda}$ are unbiased.

Second, in contrast to our continuous distribution approach, if only $\tilde{\lambda}$ is known exactly, then the back-computation of y depends on the assumed PCR parameters m, s, and v, which makes a joint estimation of these parameters from dPCR data difficult.

F.3 Comparison of continuous measurement distributions

When using a continuous distribution to model dPCR measurements, we only approximate the true distribution of measurements, which are derived from positive partition counts and therefore discretely distributed. Figure 9 compares the cumulative distribution functions (CDFs) of different continuous probability distributions to the discrete CDF of the simulated true distribution of measurements. As can be seen, the normal distribution is an inappropriate approximation as it assigns non-zero probability to negative measurements. Since zero measurements are represented by our hurdle model, we can use strictly positive distribution to represent non-zero measurements (see Figure 9A). As strictly positive distributions, we compared the log-normal, inverse Gaussian, and gamma distributions. While the log-normal distribution is often used as a default to model positive measurements, we note that the concentration estimates $\hat{\lambda}$ are a log-linear transform of the Binomially (and therefore approximately normally) distributed partition counts. The log-normal distribution, however, would fit well for an exponential transform. Indeed, we find that a gamma distribution offers a better fit than the log-normal or inverse Gaussian distribution, especially at larger concentrations (see Figure 9B), as indicated by their discretised Kullback-Leibler divergence to true distribution ($\mathrm{KL}_{\text{Gamma}} = 0.01$, $\mathrm{KL}_{\text{Log-normal}} = 0.05$, $\mathrm{KL}_{\text{Inverse Gaussian}} = 0.06$). Other strictly positive distributions such as the Truncated normal distribution could also provide a suitable approximation, but are difficult to parameterize by their mean and coefficient of variation. We therefore use a gamma distribution to model non-zero measurements in this work.

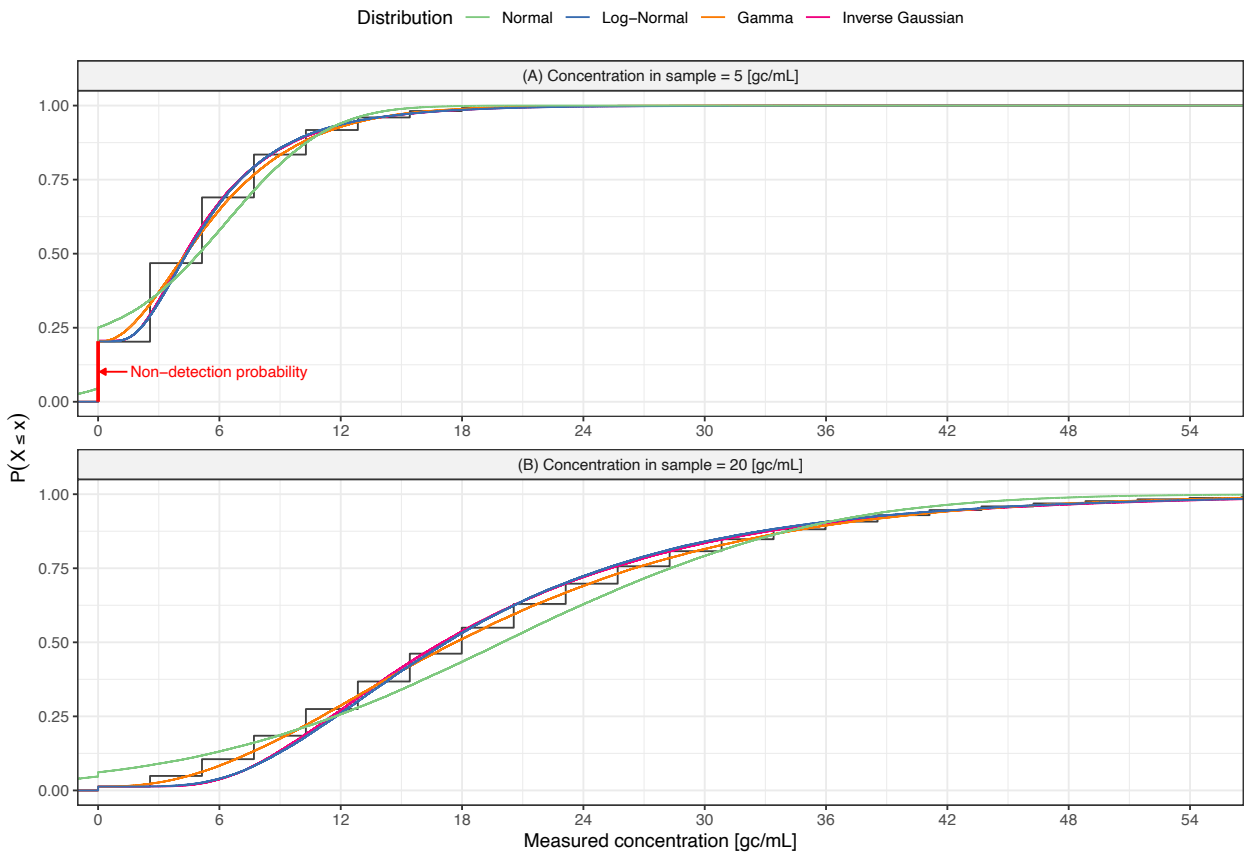

Fig 9. Continuous distributions approximating the distribution of dPCR measurements. Shown in black is the discrete cumulative distribution function of dPCR measurements obtained from the number of positive partitions in the assay for a concentration of (A) 5 gc/mL and (B) 20 gc/mL. This distribution can be approximated using a hurdle model with a probability of non-detection as defined in Eq. (23) and continuously distributed non-negative measurements with conditional mean $\mathbb{E}[\hat{\Lambda}|\hat{\Lambda} > 0]$ and conditional coefficient of variation $\mathrm{CV}[\hat{\Lambda}|\hat{\Lambda} > 0]$. Shown in colors are the continuous cumulative distribution functions of a hurdle model using a normal (green), log-normal (blue), gamma (orange), and inverse Gaussian (pink) distribution for the non-negative measurements, respectively.

G Inference from dPCR measurements

G.1 PCR parameters

G.1.1 Prior for number of partitions

The number of partitions m varies between PCR runs due to random loss of partitions. As can be seen in Figure 10, the number of lost partitions per PCR run can be approximately modelled as log-normally distributed. In principle, m could thus be modelled using a hyperprior for the maximum number of partitions, and hyperpriors for the mean and standard deviation of the number of lost partitions. To simplify estimation, however, we want to parameterize m instead via hyperpriors for

the mean number of partitions μ_m and coefficient of variation of partitions ν_m. We thus assume that the maximum number of partitions corresponds to $\mu_m + 3\mu_m\nu_m$ and choose the parameters of the log-normal distribution for lost partitions accordingly, i. e. such that the mean number of lost partitions is $3\mu_m\nu_m$ and the standard deviation is $\mu_m\nu_m$. As a hyperprior for μ_m, we use $\text{Normal}^+(\mu = 20000, \sigma = 5000)$ to accommodate dPCR systems with between 10000 and 30000 partitions. As a hyperprior for ν_m, we use $\text{Normal}^+(\mu = 0, \sigma = 0.05)$, reflecting our expectation of less than 10% variation in the number of partitions.

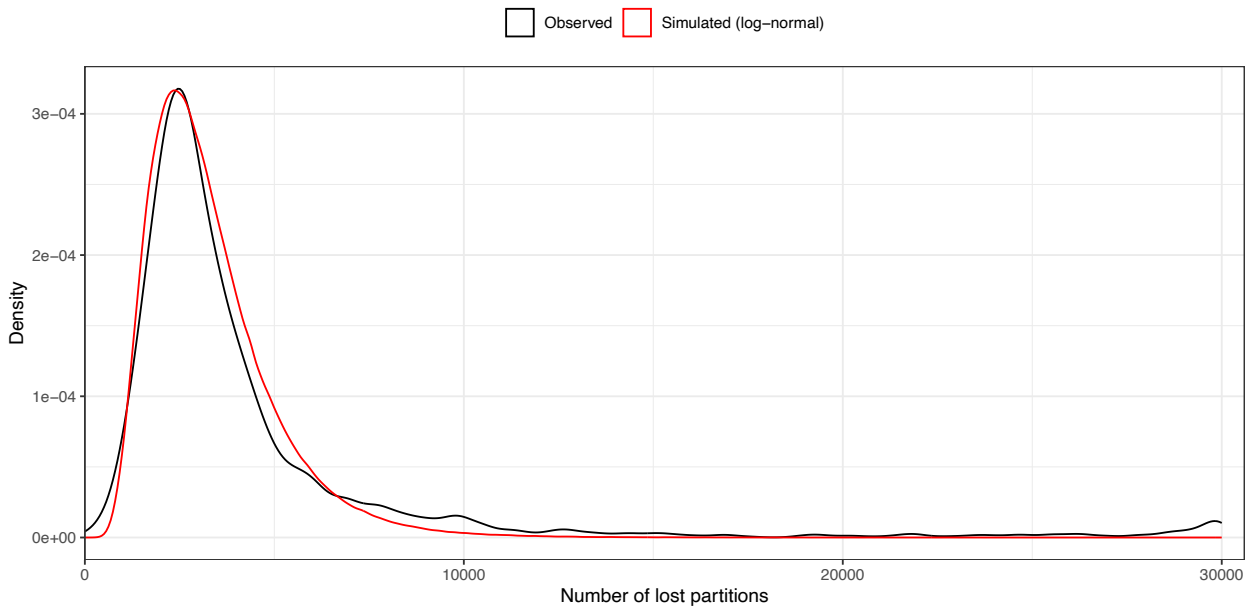

Fig 10. Distribution of lost partitions in a dPCR assay. Shown in black is a kernel density estimate for the distribution of the number of lost partitions in a dPCR assay based on 2733 runs on a Stilla® Sapphire chip with a maximum of 30000 partitions. Shown in red is a log-normal distribution matched to the observed density.

G.1.2 Prior for conversion factor

The conversion factor $c = sv$ is the product of the scaling of the concentration in the assay and the partition volume. It expresses how many expected gene copies per partition a concentration of 1 gc/mL in the original sample corresponds to. Here, we propose a broad prior for c for the case of digital droplet PCR based on a typical laboratory protocol for wastewater-based epidemiology. Specifically, we assume that extraction volumes from wastewater samples are a few dozen mL (e. g. 40mL in the assay described in Supplement A). From these, an elution of several dozen uL is produced, which corresponds to an increase in concentration by up to a factor of 1000. Due

to dilution and adding of reagents for the PCR reaction, this concentration is however decreased, typically by at least a factor of 10 (e. g. by a factor of 15 in the assay described in Supplement A). This means that overall the scaling s in concentration through preprocessing is likely less than by a factor of 100. Importantly, while there may be further factors influencing the relationship between the original concentration in the sample and the concentration in the PCR, e. g. recovery efficiency or inhibition, s should only include factors that were accounted for by the lab when reporting the sample concentration. Moreover, the droplet volume v in digital droplet PCR is typically below 1 nL. Overall, we therefore expect c to be below $100 \times 10^{-6} = 1 \times 10^{-4}$, but typically much lower than that. For example, in the assay described in Supplement A, the conversion factor is 0.173×10^{-4}. Based on these assumptions, we use a $\text{Normal}^{+}\left(\mu = 1 \times 10^{-5}, \sigma = 4 \times 10^{-5}\right)$ prior for c, which gives a rather flat prior that has most of its probability mass below 1×10^{-4}.

G.1.3 Estimated parameters

Figure 11 shows prior and posterior distributions for the average number of partitions, the conversion factor, and the pre-PCR coefficient of variation based on a model fitted to real-world measurements of Influenza A virus as presented in the main text.

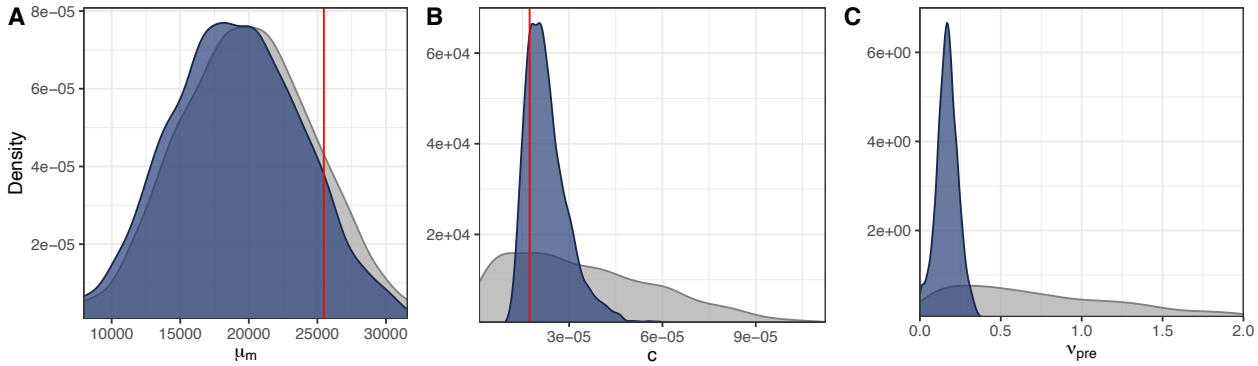

Fig 11. Inference of dPCR parameters and pre-PCR variation from a time series of dPCR measurements: Shown are prior and posterior distributions for (A) the average number of partitions μ_m, (B) the conversion factor c, and (C) the pre-PCR coefficient of variation ν_{pre} of a model applied to dPCR measurements of Influenza A virus concentrations at the municipal treatment plant of Zurich, Switzerland (Oct 19 – Dec 12, 2022). Vertical dashed lines show the true μ_m and c values based on the laboratory setup.

G.2 Estimation

The different likelihoods were implemented using the probabilistic programming language stan [8], both in a standalone model for estimation of a single concentration from replicate measurements

and in the R_t estimation model of the EpiSewer R package for inference from measurements over time. In our examples, we used a standard epidemiological model of EpiSewer, where the effective reproduction number was modelled as a random walk, with infections arising from a stochastic infection process with Poisson distributed noise.

For all models, parameters were estimated via Markov chain Monte Carlo (MCMC) using cmdstan version 2.34.1 via cmdstanr version 0.8.1 [9]. We ran four No-U-Turn sampler chains with 1,000 warm-up and 1,000 sampling iterations each, with a maximum tree depth of 15, an initial step size of 0.01, and an adaptation target acceptance statistic of 0.99. We checked the fitted models for low effective sample sizes (ESS < 400) [10], and for convergence problems using the number of divergent transitions [11], the Bayesian fraction of missing information (E-BFMI < 0.2) [12] and the Gelman-Rubin convergence diagnostic ($\hat{R} > 1.01$) [13]. The checks indicated sufficient effective sampling sizes and good convergence and mixing of chains.

One week-ahead forecasts in EpiSewer were produced by assuming a continuation of the recent transmission dynamics. For this, the random walk model estimated for R_t was projected forward, which assumes that R_t will remain at its current value in expectation, but could vary from this depending on the estimated random walk variance. Thereby, the uncertainty of the projected R_t is higher for forecasts further into the future. Given the projected R_t estimates, all other latent variables such as the number of infections and the expected load in wastewater could be sampled from the generative model of EpiSewer. Since we conducted retrospective forecasts, we could use the observed flow volumes at the respective treatment plants also for the forecast dates to compute the expected concentration from on the predicted load. Finally, the sampling model corresponding to the respective likelihood function was used to predict dPCR measurements based on the forecast. This procedure was repeated for each posterior sample of the fitted model, yielding a posterior predictive distribution of R_t, loads, and measurements.

G.3 Additional results

In the following, we provide additional results from fitting models using 1) a normal likelihood with constant variance, 2) a log-normal likelihood with constant coefficient of variation, and 3) the dPCR-specific likelihood on different date ranges of measurements at the treatment plants of Zurich and Geneva, Switzerland (Figures 12–16). We also show a comparison of estimates obtained using

the dPCR-specific likelihood with and without exact knowledge of the dPCR parameters (Figure 17).

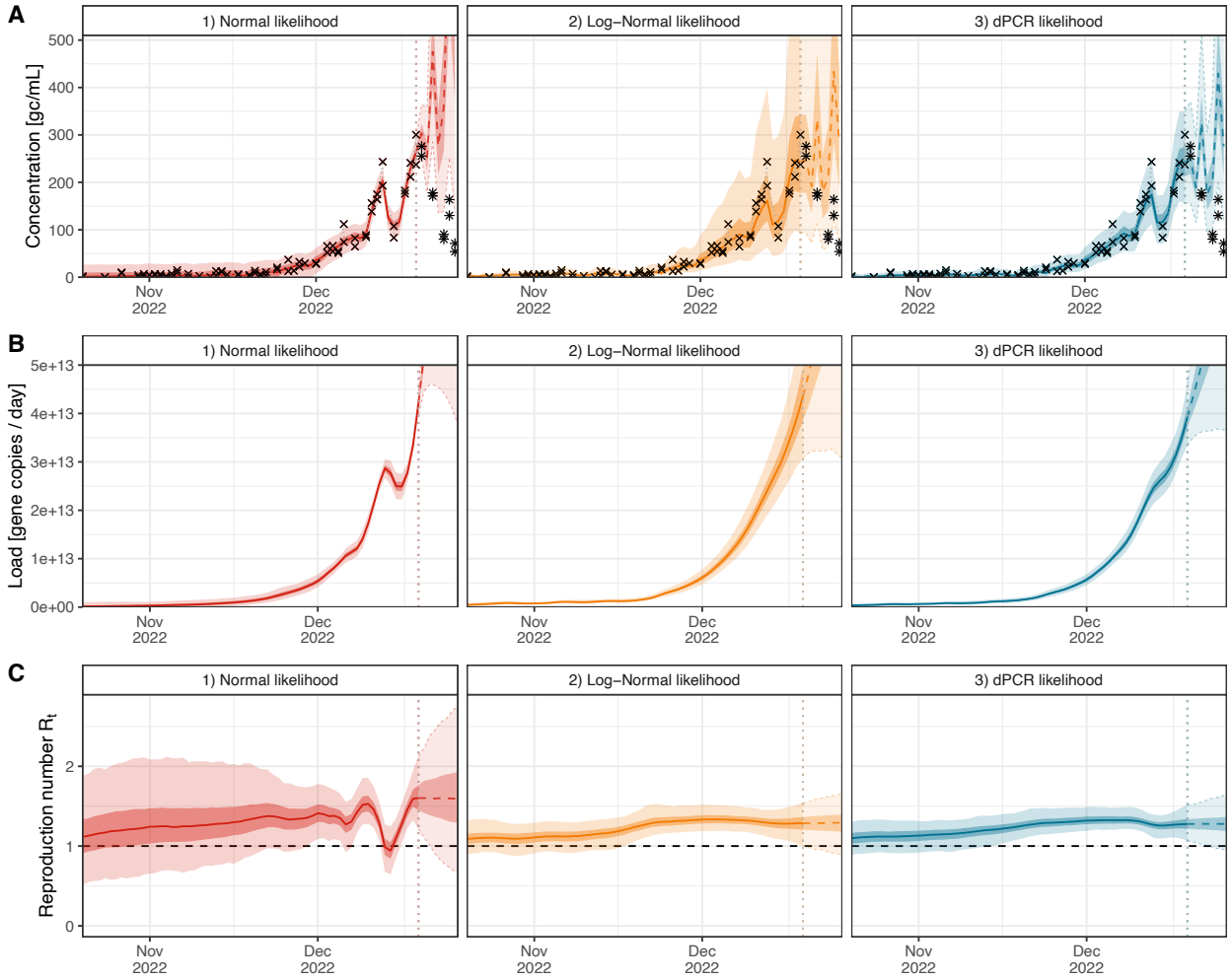

Fig 12. Comparison of likelihood functions on dPCR measurements of Influenza A virus concentrations at the municipal treatment plant of Zurich, Switzerland, Oct 20 – Dec 19, 2022: Shown are the results of fitting an epidemiological wastewater model (EpiSewer) to duplicate dPCR measurements of Influenza A virus at the municipal treatment plant of Zurich, Switzerland (Oct 20 – Dec 19, 2022) using either 1) a normal likelihood (red), 2) a log-normal likelihood (orange), and 3) a dPCR-specific likelihood (blue) for the observations. Each panel shows the median (lines), and 50% and 95% credible intervals (strong and weakly shaded areas) of (A) posterior predictive distributions for dPCR measurements, (B) estimated viral loads in wastewater over time, and (C) the estimated effective reproduction number over time. Estimates shown beyond Dec 26, 2022 (vertical dotted line) are forecasted based on a continuation of the latest transmission dynamics. Panel A also shows the true observed (crosses) and future (stars) dPCR measurements.

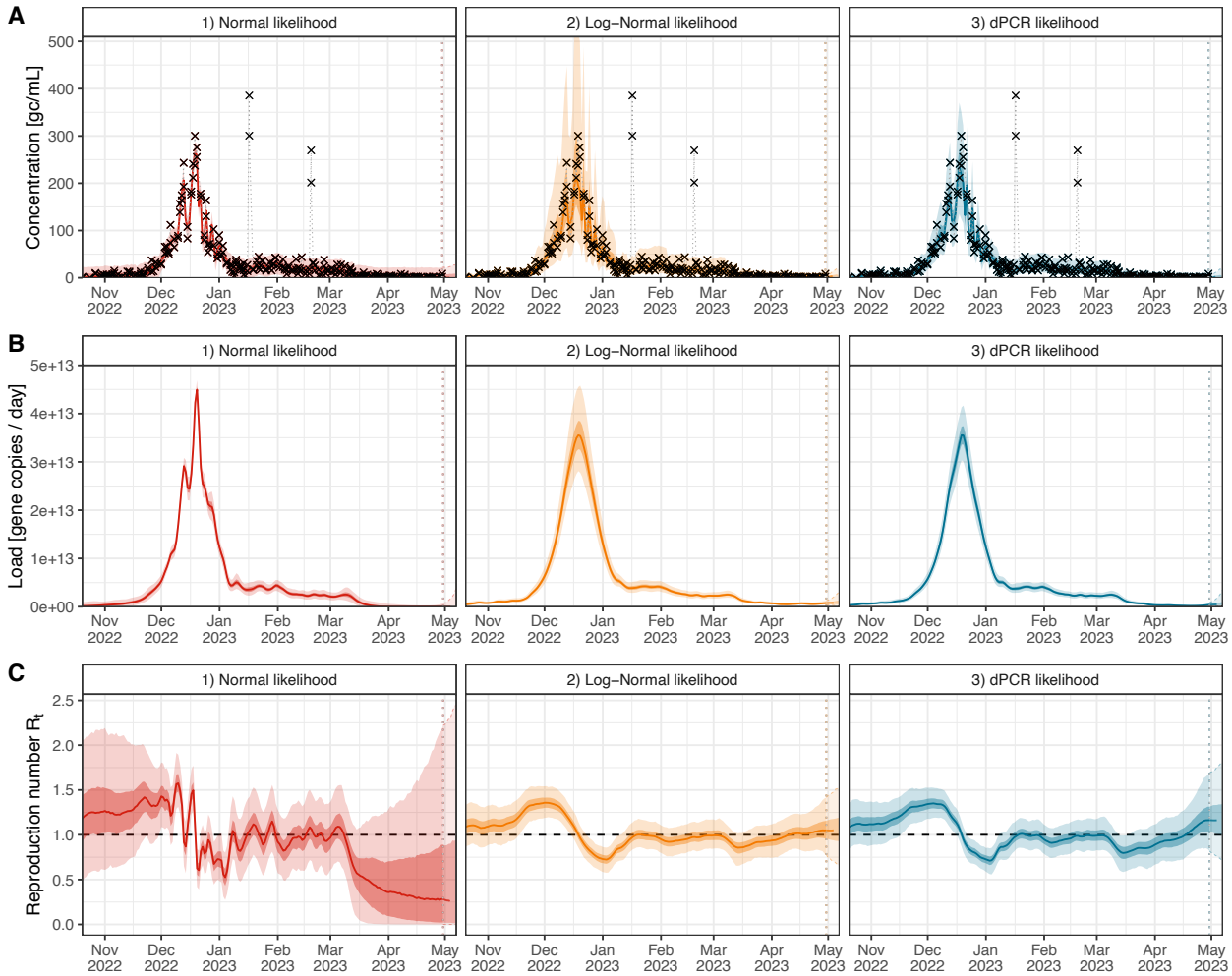

Fig 13. Comparison of likelihood functions on measurements of Influenza A virus concentrations at the municipal treatment plant of Zurich, Switzerland, Oct 20 – May 01, 2023: Shown are the results of fitting an epidemiological wastewater model (EpiSewer) to duplicate concentration measurements of Influenza A virus at the municipal treatment plant of Zurich, Switzerland (Oct 20 – May 01, 2023) using either 1) a normal likelihood (red), 2) a log-normal likelihood (orange), and 3) a dPCR-specific likelihood (blue) for observed measurements. Each panel shows the median (lines), and 50% and 95% credible intervals (strong and weakly shaded areas) of (A) posterior predictive distributions for concentration measurements, (B) estimated viral loads in wastewater over time, and (C) the estimated effective reproduction number over time. Estimates shown beyond Dec 26, 2022 (vertical dotted line) are forecasted based on a continuation of the latest transmission dynamics. Panel A also shows the true observed (crosses) and future (stars) measurements, respectively. Measurements on January 17, 2023, and February 2, 2023 were removed as outliers before model fitting.

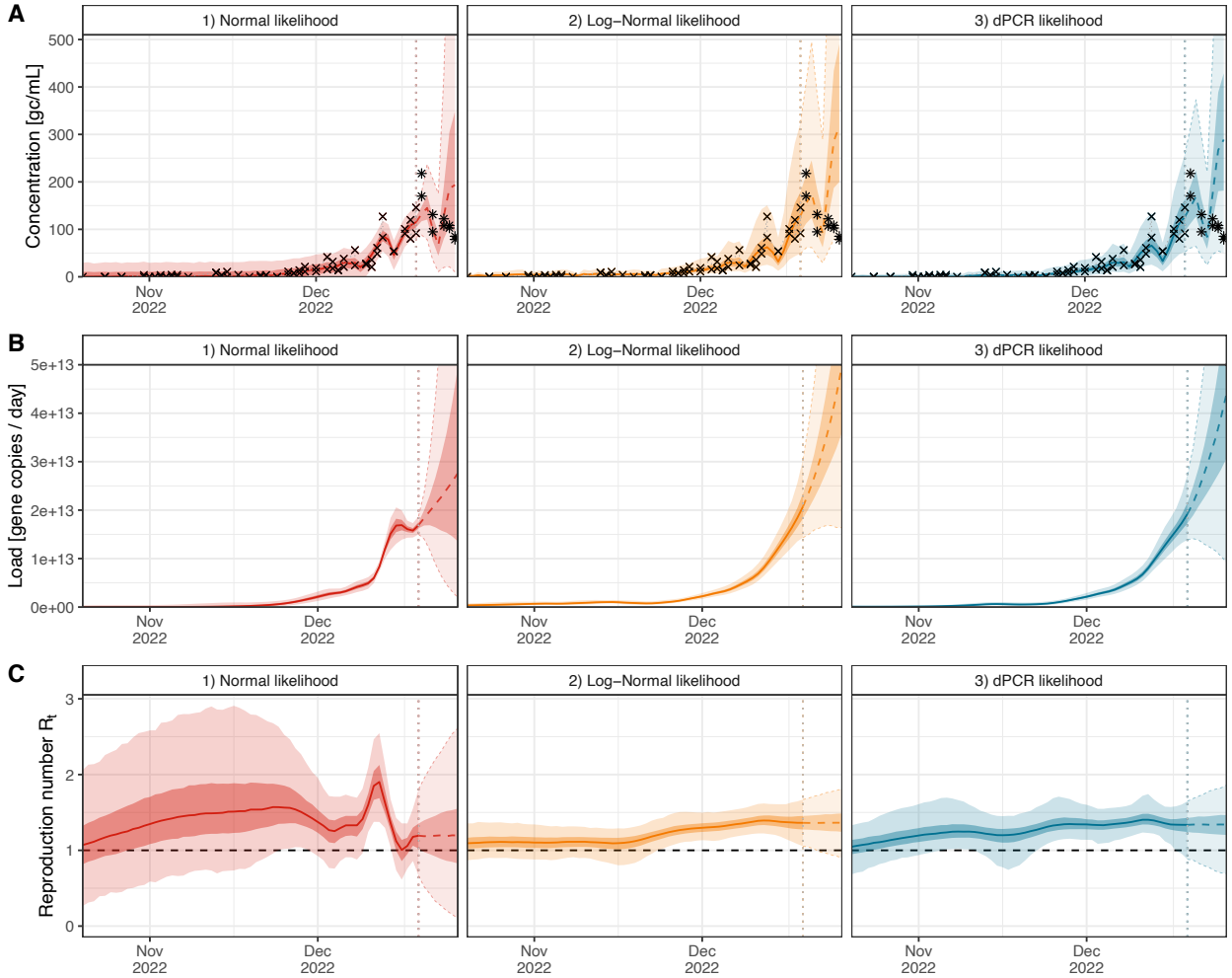

Fig 14. Comparison of likelihood functions on measurements of Influenza A virus concentrations at the municipal treatment plant of Geneva, Switzerland, Oct 20 – Dec 19, 2022: Shown are the results of fitting an epidemiological wastewater model (EpiSewer) to duplicate dPCR measurements of Influenza A virus at the municipal treatment plant of Geneva, Switzerland (Oct 20 – Dec 19, 2022) using either 1) a normal likelihood (red), 2) a log-normal likelihood (orange), and 3) a dPCR-specific likelihood (blue) for the observations. Each panel shows the median (lines), and 50% and 95% credible intervals (strong and weakly shaded areas) of (A) posterior predictive distributions for dPCR measurements, (B) estimated viral loads in wastewater over time, and (C) the estimated effective reproduction number over time. Estimates shown beyond Dec 26, 2022 (vertical dotted line) are forecasted based on a continuation of the latest transmission dynamics. Panel A also shows the true observed (crosses) and future (stars) dPCR measurements.

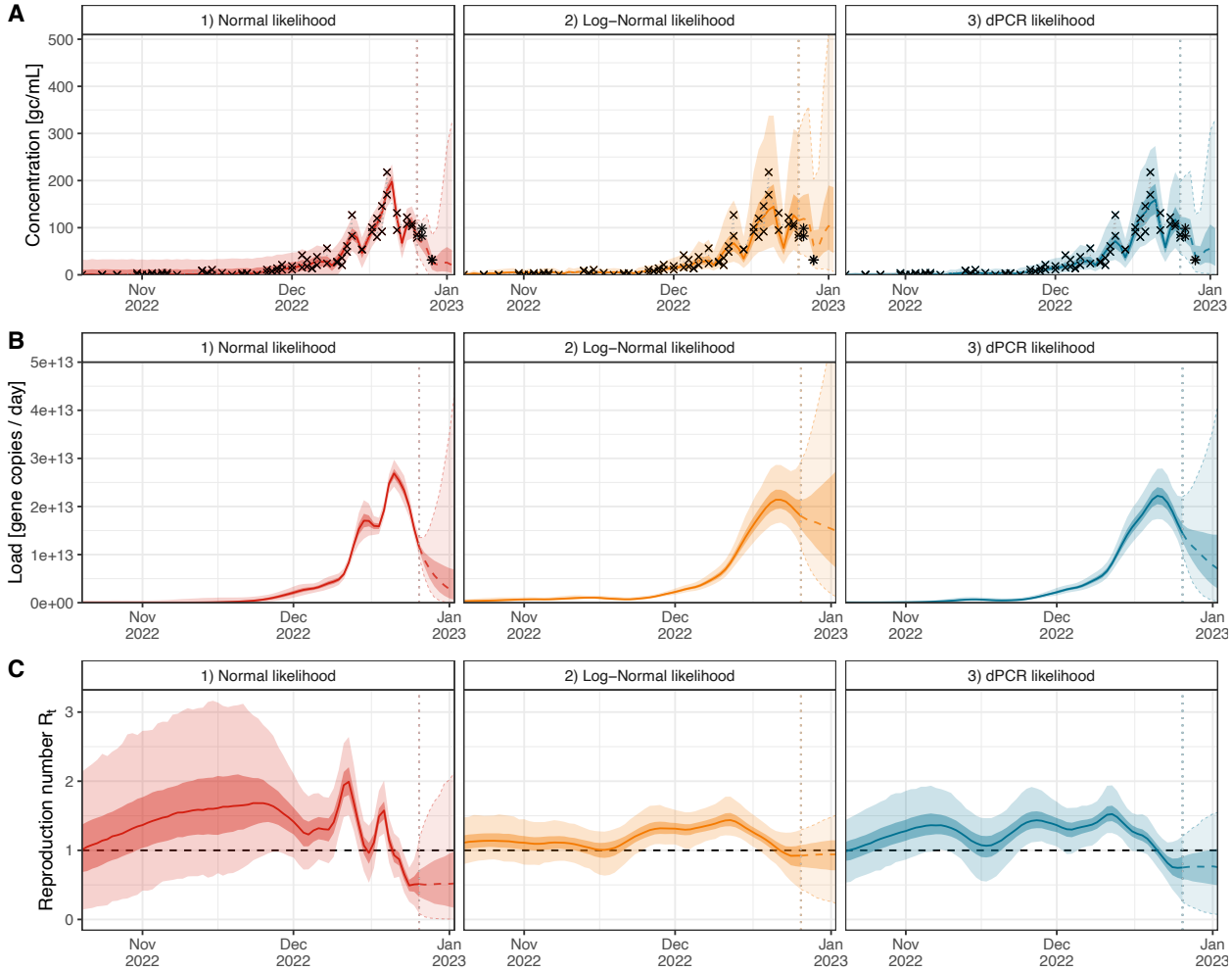

Fig 15. Comparison of likelihood functions on measurements of Influenza A virus concentrations at the municipal treatment plant of Geneva, Switzerland, Oct 20 – Dec 26, 2022: Shown are the results of fitting an epidemiological wastewater model (EpiSewer) to duplicate dPCR measurements of Influenza A virus at the municipal treatment plant of Geneva, Switzerland (Oct 20 – Dec 26, 2022) using either 1) a normal likelihood (red), 2) a log-normal likelihood (orange), and 3) a dPCR-specific likelihood (blue) for the observations. Each panel shows the median (lines), and 50% and 95% credible intervals (strong and weakly shaded areas) of (A) posterior predictive distributions for dPCR measurements, (B) estimated viral loads in wastewater over time, and (C) the estimated effective reproduction number over time. Estimates shown beyond Dec 26, 2022 (vertical dotted line) are forecasted based on a continuation of the latest transmission dynamics. Panel A also shows the true observed (crosses) and future (stars) dPCR measurements.

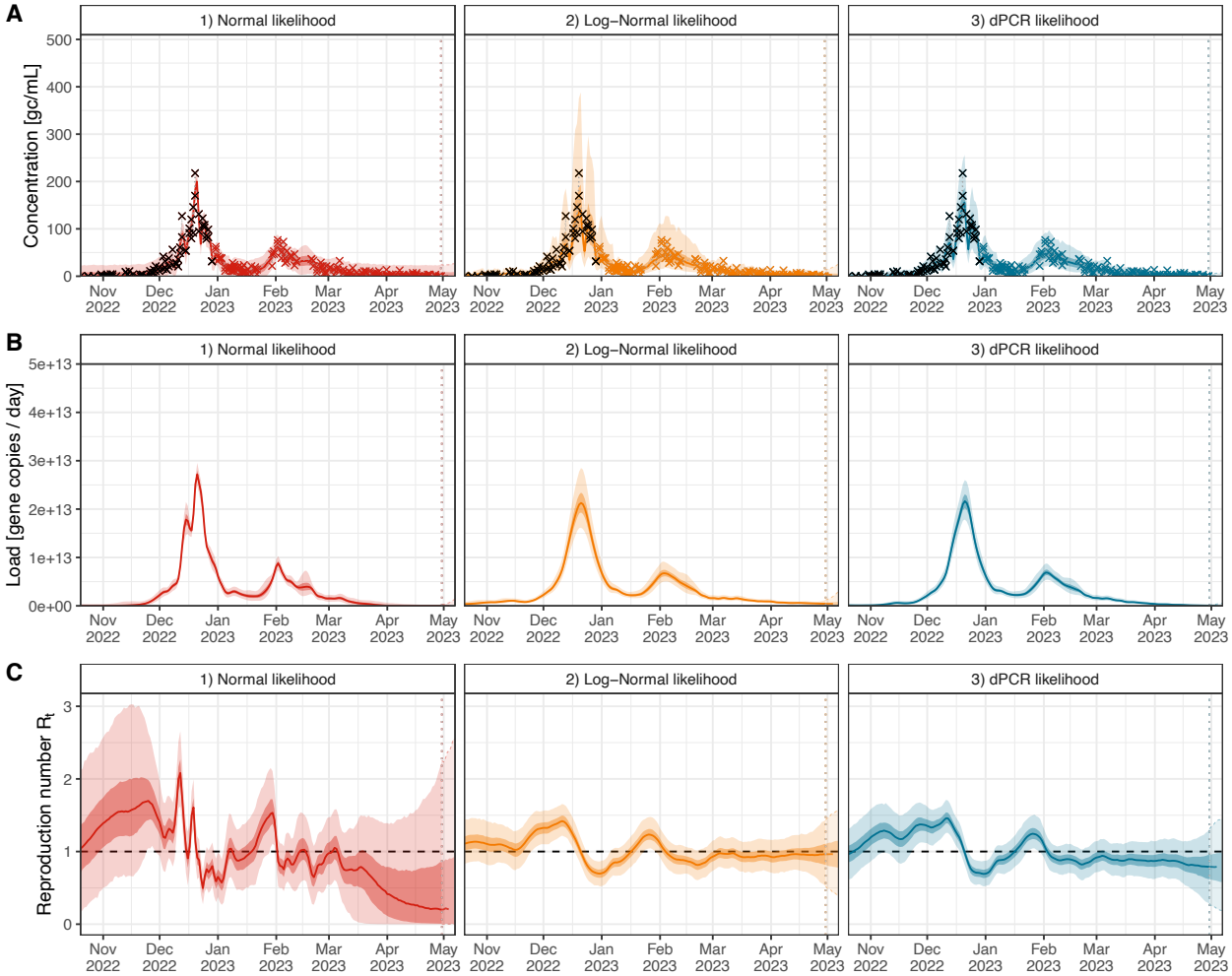

Fig 16. Comparison of likelihood functions on measurements of Influenza A virus concentrations at the municipal treatment plant of Geneva, Switzerland, Oct 20 – May 01, 2023: Shown are the results of fitting an epidemiological wastewater model (EpiSewer) to duplicate dPCR measurements of Influenza A virus at the municipal treatment plant of Geneva, Switzerland (Oct 20 – May 01, 2023) using either 1) a normal likelihood (red), 2) a log-normal likelihood (orange), and 3) a dPCR-specific likelihood (blue) for the observations. Each panel shows the median (lines), and 50% and 95% credible intervals (strong and weakly shaded areas) of (A) posterior predictive distributions for dPCR measurements, (B) estimated viral loads in wastewater over time, and (C) the estimated effective reproduction number over time. Estimates shown beyond Dec 26, 2022 (vertical dotted line) are forecasted based on a continuation of the latest transmission dynamics. Panel A also shows the true observed (crosses) and future (stars) dPCR measurements.

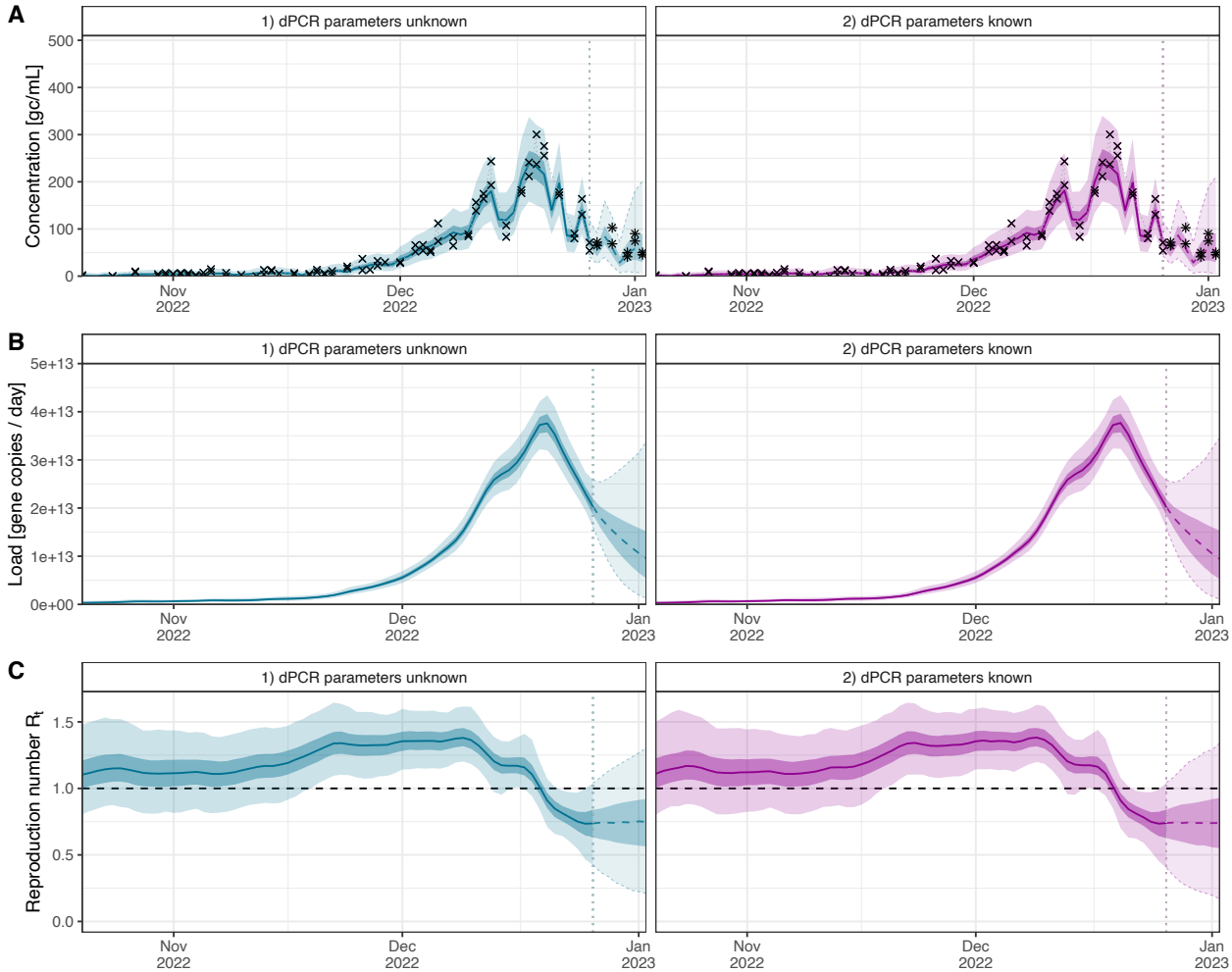

Fig 17. Comparison of inference and short-term forecasts from wastewater measurements using a dPCR-specific likelihood with and without knowledge of the dPCR parameters: Shown are the results of fitting an epidemiological wastewater model (EpiSewer) to duplicate dPCR measurements of Influenza A virus at the municipal treatment plant of Zurich, Switzerland (Oct 20 – Dec 26, 2022) using a dPCR-specific likelihood where 1) broad priors are used for the number of total partitions and the conversion factor (blue) and 2) the number of total partitions and the conversion factor are fixed to their true values (purple). Each panel shows the median (lines), and 50% and 95% credible intervals (strong and weakly shaded areas) of (A) posterior predictive distributions for dPCR measurements, (B) estimated viral loads in wastewater over time, and (C) the estimated effective reproduction number over time. Estimates shown beyond Dec 26, 2022 (vertical dotted line) are forecasted based on a continuation of the latest transmission dynamics. Panel A also shows the true observed (crosses) and future (stars) dPCR measurements.

References

1. Huisman JS, Scire J, Caduff L, Fernandez CX, Ganesanandamoorthy P, Kull A, et al. Wastewater-Based Estimation of the Effective Reproductive Number of SARS-CoV-2. Environmental Health Perspectives. 2022;130(5):057011.
2. Nadeau S, Devaux AJ, Bagutti C, Alt M, Hampe EI, Kraus M, et al. Influenza transmission dynamics quantified from RNA in wastewater in Switzerland. Swiss Medical Weekly. 2024;154(1):3503–3503.
3. Dorazio RM, Hunter ME. Statistical Models for the Analysis and Design of Digital Polymerase Chain Reaction (dPCR) Experiments. Analytical Chemistry. 2015;87(21):10886–10893.
4. Dube S, Qin J, Ramakrishnan R. Mathematical Analysis of Copy Number Variation in a DNA Sample Using Digital PCR on a Nanofluidic Device. PLOS ONE. 2008;3(8):e2876.
5. Asmussen S, Jensen JL, Rojas-Nandayapa L. On the Laplace Transform of the Lognormal Distribution. Methodology and Computing in Applied Probability. 2016;18(2):441–458.
6. Corless RM, Gonnet GH, Hare DEG, Jeffrey DJ, Knuth DE. On the LambertW Function. Advances in Computational Mathematics. 1996;5(1):329–359.
7. Armbruster DA, Pry T. Limit of Blank, Limit of Detection and Limit of Quantitation. The Clinical Biochemist Reviews. 2008;29(Suppl 1):S49–S52.
8. Stan development team. Stan Modeling Language Users Guide and Reference Manual, Version 2.35; 2024.
9. Gabry J, Češnovar R. Cmdstanr: R Interface to 'CmdStan' [Manual]; 2024.
10. Geyer C. Introduction to Markov Chain Monte Carlo. In: Brooks S, Gelman A, Jones G, Meng XL, editors. Handbook of Markov Chain Monte Carlo. Chapman and Hall/CRC; 2011. p. 3–48.
11. Betancourt M. A Conceptual Introduction to Hamiltonian Monte Carlo. arXiv; 2018.
12. Betancourt M. Diagnosing Suboptimal Cotangent Disintegrations in Hamiltonian Monte Carlo. arXiv; 2016.
13. Gelman A, Rubin DB, et al. Inference from Iterative Simulation Using Multiple Sequences. Statistical Science. 1992;7(4):457–472.