Improving inference in environmental surveillance by modeling 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

Digital polymerase chain reaction (dPCR) is a powerful technique for quantifying gene targets in environmental samples, with various applications such as biodiversity monitoring and wastewater-based epidemiology. However, statistical analyses of environmental dPCR data often assume, explicitly or implicitly, that concentration measurements have a normal or log-normal error structure, which does not reflect the underlying partitioning statistics of dPCR. Using simulations and real-world environmental data, we show that (log-)normality assumptions are violated for dPCR measurements, leading to inaccurate estimates of gene concentrations and underlying biological processes. To enable reliable analyses of environmental dPCR data, we present a dPCR-specific likelihood model that accounts for concentration-dependent measurement noise and non-detects as characteristic of dPCR assays. We demonstrate that this approach overcomes biases in inference from environmental data, such as estimating free-eDNA decay in seawater or pathogen transmission from wastewater monitoring. Our method is implemented in the R packages "dPCRfit" for regression analyses and "EpiSewer" for wastewater surveillance.

Keywords: eDNA, wastewater-based epidemiology, dPCR, limit of detection, limit of quantification

1 Introduction

The quantification of gene concentrations from environmental samples has become an important technique in environmental science and engineering, with diverse applications in ecology, conservation, and public health. For example, the collection and analysis of environmental DNA (eDNA), i. e. genetic material shed into the environment by living organisms, enables the detailed monitoring of biodiversity and population dynamics [1,2]. By quantifying the abundance of a specific target gene in samples taken at different locations or times, insights into the occurrence, distribution, movements, and interactions of species can be gained [3,4]. Similarly, in wastewater-based epidemiology, genetic material from pathogens shed by infected individuals into wastewater is sampled at municipal treatment plants to track infection dynamics in human populations without reliance on clinical testing [5,6]. This has enabled the cost-effective monitoring of community-level pathogen spread [7–12] and the estimation of epidemiological parameters, such as the effective reproduction number [13].

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 before PCR amplification [14]. Using the ratio of positive partitions to the total number of partitions, the concentration in the sample can be measured [15]. 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 [16–18].

To analyze gene concentrations measured in environmental samples from different locations or points in time, various statistical approaches such as non-parametric smoothing [19], correlation analysis [20–22], linear regression [23–25], generalized linear models [26, 27] and latent variable models [28–30] have been used. Although appropriate statistical models exist for the number of positive partitions in a dPCR assay [15], many approaches directly analyze measured concentrations, treating them as normally [19–22, 25, 27, 28] or log-normally [9, 23, 24, 26, 29] distributed. However, such assumptions do not reflect the statistical properties of dPCR assays, which can affect data analysis in several ways. For example, the measurement error in dPCR data increases in a non-linear

fashion as sample concentrations become lower [31]. This can lead to unreliable inferences due to heteroskedasticity in ordinary least squares regression or Pearson correlation coefficients [32]. In addition, dPCR measurements may result in non-detects, corresponding to a measured concentration of zero. However, when modeling dPCR data as log-normally distributed, such zero measurements must be removed from the analysis, biasing concentration estimates upwards. These problems are particularly pronounced when analyzing dPCR data with limited sample size and low concentrations, such as estimating species prevalence from a few sampled sites, or monitoring pathogen concentrations at the start of an epidemic wave from weekly wastewater samples.

In this study, we develop an accurate statistical model for analyzing concentration measurements from dPCR. Using established theory of partitioning statistics in dPCR, we derive estimators for the coefficient of variation and the probability of non-detection given the true concentration in a sample. Unlike previous work, we account for noise through extraction, preprocessing, and other sources of variation, which can be substantial in the case of environmental samples. We then propose a likelihood function for estimating gene concentrations and underlying biological process characteristics from reported dPCR measurements. This offers a practical and interpretable alternative to explicitly modeling the number of positive partitions in an assay [15], which is often unavailable from publicly reported data. We demonstrate that, compared to normal and log-normal models, our method achieves more reliable and precise estimates of i) decay rates for marine species free-eDNA in seawater and ii) infection dynamics of respiratory viruses from wastewater.

2 Materials and Methods

2.1 Partitioning statistics in digital PCR

In digital PCR, the reaction mixture is split uniformly into m independent partitions, where m is often in the order of tens of thousands. The number of target molecules in each partition is assumed to follow a Poisson process [33], i. e. partition $j \in \{1, \ldots, m\}$ will contain $N_j \sim \text{Pois}(\lambda c)$ molecules. Here, λ is the expected concentration in the analyzed sample, and c is a "conversion factor" that expresses the expected number of gene copies per partition for one concentration unit of the original sample. This factor can be expressed as a product of the partition volume v and a scaling factor s, summarizing concentration differences due to preprocessing (e. g. extraction, adding of reagents,

dilution, and suboptimal recovery efficiency [34]). During PCR, a fluorescent signal is generated if one or several gene copies are present in a partition, leading to a binary result (positive or negative) for each partition. Since the reaction in each partition is regarded as independent, the total number of positive partitions Y is binomially distributed with

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

where $1 - \exp(-\lambda c)$ is the probability for at least one gene copy being present in a partition.

Under this model, Dorazio and Hunter [15] show that the maximum likelihood estimate for the concentration λ given $Y = y$ observed positive partitions is

$$\tilde{\lambda} = -\frac{1}{c} \log\left(1 - \frac{y}{m}\right). \tag{2}$$

It is furthermore common to run n technical replicates $i \in \{1, \ldots, n\}$ of the same sample. Replicate measurements can then be combined by applying the maximum likelihood estimator to the pooled partition counts or by computing the arithmetic mean of individual estimates (see Supporting Information A for a comparison of both approaches). In the following, we assume an equal number of partitions $m_i = m$ for all technical replicates, in which case our results are valid for both approaches. Naturally, by letting $n = 1$, our model also applies to single measurements.

In environmental surveillance, studies often report the maximum likelihood estimate $\hat{\lambda}$ for the target concentration λ rather than the number of positive partitions y. Here we develop a suitable model for $\hat{\lambda}$ by deriving the coefficient of variation and probability of non-detection of dPCR measurements and proposing an appropriate likelihood function.

2.2 Coefficient of variation

We now regard the dPCR measurement $\hat{\lambda}$ as a random variable $\hat{\Lambda}$. Assuming that λ is the true expected sample concentration, and using the Binomial model of positive partition counts Y, it can be shown that the asymptotic variance of $\hat{\Lambda}$ is

$$\text{Var}[\hat{\Lambda}] = \frac{1}{c^2 mn} (\exp(\lambda c) - 1), \tag{3}$$

see Supporting Information B.1 for details. Using the fact that $\hat{\Lambda}$ is an asymptotically unbiased estimator of λ [33], we can estimate the coefficient of variation (CV) of the measurements as

$$\nu_{\hat{\Lambda}}(\lambda) = \frac{1}{c\sqrt{mn}} \frac{\sqrt{\exp(\lambda c) - 1}}{\lambda}. \tag{4}$$

Eq. (4) implies that the CV of measurements is concentration-dependent, which can lead to heteroskedasticity in linear regression even if measurements are log-transformed [32]. However, this estimate 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 extend our model by defining the unscaled concentration in the PCR as a random variable Λ_{pre} that follows a parametric distribution with mean λ and coefficient of variation $\nu_{\text{pre}} > 0$. Here, the parameter ν_{pre} describes random variation before the PCR. As shown in Supporting Information B.2, the variance of $\hat{\Lambda}$ under this extended model is

$$\text{Var}[\hat{\Lambda}] = \lambda^2 \nu_{\text{pre}}^2 + \frac{1}{c^2 mn} \left(\text{E}[\exp(\Lambda_{\text{pre}} c)] - 1\right), \tag{5}$$

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

$$E[\exp(\Lambda_{\text{pre}} c)] = \left(1 - \lambda c \nu_{\text{pre}}^2\right)^{-\frac{1}{\nu_{\text{pre}}^2}}, \tag{6}$$

yielding a coefficient of variation

$$\nu_{\hat{\Lambda}}(\lambda) = \sqrt{\nu_{\text{pre}}^2 + \frac{1}{\lambda^2 c^2 mn} \left(\left(1 - \lambda c \nu_{\text{pre}}^2\right)^{-\frac{1}{\nu_{\text{pre}}^2}} - 1 \right)}. \tag{7}$$

In Supporting Information B.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

A dPCR run can have zero positive partitions even for non-zero concentrations. If this happens for all technical replicates, the target 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 in Eq. (1), i. e.

$$p_{\text{zero}}(\lambda) = \prod_{i=1}^{n} \Pr(\hat{\Lambda} = 0|\lambda) = \prod_{i=1}^{n} \exp(-\lambda cm) = \exp(-\lambda cmn). \tag{8}$$

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 [35]. Leaving the possibility of false positive partitions aside, we note that $\alpha = 1 - p_{\text{zero}}(\lambda_{\text{LOD}}) = 1 - \exp(-\lambda_{\text{LOD}} cmn)$, hence λ_{LOD} sufficiently determines the non-detection probability under Eq. (8) for arbitrary concentrations (see Supporting Information C.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

$$p_{\text{zero}}(\lambda) = \mathrm{E}_{\Lambda_{\text{pre}}}\left[\prod_{i=1}^{n} \Pr(\hat{\Lambda} = 0|\Lambda_{\text{pre}} = \lambda_{\text{pre}})\right] = \mathrm{E}[\exp(-\Lambda_{\text{pre}} cmn)]. \tag{9}$$

Similar to Eq. (5), 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 Supporting Information C.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 $\hat{\lambda}$:

$$L_{\hat{\Lambda}}(\hat{\lambda}|\lambda) = \begin{cases} p_{\text{zero}}(\lambda), & \text{if } \hat{\lambda} = 0 \\ f_{\text{PCR}}(\hat{\lambda})\,(1 - p_{\text{zero}}(\lambda)), & \text{otherwise} \end{cases}, \tag{10}$$

where $p_{\text{zero}}(\lambda)$ is the probability of non-detection as defined in Eq. (9) and f_{PCR} is the density function of a positive continuous probability distribution with conditional mean

$$\mathrm{E}[\hat{\Lambda}|\hat{\Lambda} > 0] = \frac{\lambda}{1 - p_{\text{zero}}(\lambda)} \tag{11}$$

and conditional coefficient of variation

$$\mathrm{CV}[\hat{\Lambda}|\hat{\Lambda} > 0] = \sqrt{\nu_{\hat{\Lambda}}(\lambda)^2(1 - p_{\mathrm{zero}}(\lambda)) - p_{\mathrm{zero}}(\lambda)}. \tag{12}$$

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. (7)) for the conditioning on a non-zero outcome (see Supporting Information F.1 for details). The likelihood defined in Eq. (10) 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 $f_{\mathrm{PCR}}(x)$ 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 Supporting Information 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 Supporting Information F.3).

2.5 Inference from dPCR measurements

Given a set of dPCR measurements $\hat{\lambda}_{i\in\{1,\ldots,k\}}$ from k samples, we can use the likelihood $P(\hat{\lambda}_{i\in\{1,\ldots,k\}}|\lambda_{i\in\{1,\ldots,k\}}) = \prod_{i=1}^{k} L_{\hat{\Lambda}}(\hat{\lambda}_i|\lambda_i)$ as proposed in Eq. (10) to estimate the underlying concentrations $\lambda_{i\in\{1,\ldots,k\}}$ via Bayesian inference. Specifically, given a concentration model $\lambda_i = f(i,\theta)$ with parameters θ and corresponding prior $P(\theta)$, we can sample from the posterior distribution $P(\lambda,\theta|\hat{\lambda})$ via Markov Chain Monte Carlo (MCMC). In this study, we implemented a generalized linear model (GLM) for $f(i,\theta)$, allowing us to estimate the association of covariates with the target concentration under different link functions. Moreover, as an example of a domain-specific model, we also implemented a model of pathogen concentrations in municipal wastewater for $f(i,\theta)$. Both models are implemented in the probabilistic programming language "stan" [36] and can be fitted through an interface in the R programming language (see Supporting Information G.1 for details on the estimation). We provide a dedicated R package for the GLM ("dPCRfit") [37] and for the wastewater model ("EpiSewer") [38].

To apply Eq. (10), 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 are known in a typical laboratory setting, they may be unknown in other settings, e. g. when modeling publicly reported concentration measurements without sufficient metadata. In this case, it is possible to provide broad priors for these parameters and estimate them jointly with the concentrations λ. We provide an example of such a joint estimation using only measurement data when applying our model to wastewater concentration measurements.

2.6 Model validation

We used simulations of positive partition counts according to Eq. (1) under different types of pre-PCR noise to validate our approximations for the CV and the probability of non-detection (Supporting Information D). We compared the theoretical predictions for the CV and the probability of non-detection with empirical data from ongoing wastewater surveillance by Eawag, the Swiss Federal Institute of Aquatic Science and Technology (Dübendorf, Switzerland). 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 14 wastewater treatment plants in Switzerland from July 10, 2023 to March 5, 2024 obtained on the Naica Crystal Digital PCR Platform with Sapphire Chips (Stilla Technologies, France) using the laboratory methods described in Nadeau et al. [39]. The empirical CV was estimated based on the bias-corrected empirical standard deviation of two technical replicates divided by their empirical mean. The empirical probability of non-detection was estimated as 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 (Supporting Information E). To validate our proposed likelihood, we estimated concentrations from simulated measurements under the positive partition count model described in Section 2.1 for concentrations of 1, 3, and 5 gc/mL, assuming a scaling factor of $s = 30$, a partition volume of $v = 0.519\,\text{nL}$, and $m = 25000$ partitions per measurement (Supporting Information G.2).

2.7 Application to eDNA-based biomonitoring

We conducted a reanalysis of data from an aquarium study by Scriver et al. [27], which tested a direct droplet dPCR assay for eDNA-based marine species monitoring. In the study, the abundance of free-eDNA of the brown bryozoan “Bugula neritina” in aquarium seawater was measured by concentrations of the Cytochrome c oxidase subunit 1 (COI) gene at different points in time up

to 72 hours after removal of the organism from the aquarium. The speed of decline in detectable gene concentrations was then estimated using an exponential decay model. We here applied a GLM with logarithmic link using the "dPCRfit" package to model the exponential decay in free-eDNA, measured in terms of the half-life period (see Supporting Information H for details). In the original analysis, the authors used the averages of measurements across three tanks with low, medium, and high initial species biomass. Here, we instead applied our GLM to the individual measurements, and additionally stratified our analysis across tanks. We then compared the consistency of eDNA half-life estimates between tanks and with the original result by Scriver et al. when using a normal, log-normal, or dPCR-specific likelihood.

2.8 Application to wastewater-based epidemiology

We combined our dPCR-specific likelihood function with a wastewater model implemented in the R package "EpiSewer" [38] to analyze real-world duplicate dPCR measurements of Influenza A virus from the wastewater treatment plants of Zurich and Geneva, Switzerland, during the 2022/23 Influenza season. We used discretized estimates from the literature to characterize the generation time distribution (shifted Gamma distribution, mean = 2.6, sd = 1.5 [40, 41]) and the shedding load distribution since time of infection (Gamma distribution, mean = 2.49, sd = 1 [42]) of Influenza A virus. Based on a comparison with confirmed case numbers, we assumed an average total shedding load per case of 1×10^{11} gc. We provided broad priors for the parameters of the dPCR assay (c and m) and estimated these jointly from the measurement data (Supporting Information I.1).

We applied the model to longitudinal wastewater measurements to estimate the viral load in wastewater and the effective reproduction number R_t over time. We compared estimates obtained under a normal, log-normal, and dPCR-specific likelihood. We also produced short-term forecasts of concentrations up to one week ahead (assuming continued transmission dynamics) and assessed forecast performance using the continuous ranked probability score (CRPS) [43]; see Supporting Information I.2 for details.

3 Results

3.1 Coefficient of variation

The coefficient of variation (CV) of measurements from dPCR depends on the target concentration in the sample. As shown in Figure 1A, the CV is approximately constant for high concentrations but increases exponentially as the concentration approaches zero. It also increases when concentrations become extremely high, such that the share of positive partitions approaches one (Figure S5) [31]. The presence of pre-PCR variation further increases the CV of measurements, as shown for gamma distributed pre-PCR noise in Figure 1. Apart from extremely high concentrations, we find that the CV is almost identical for log-normally and gamma-distributed pre-PCR noise and can be well approximated using a Taylor series expansion (Figures S4–S5). The averaging of n technical replicates scales the CV by a factor of $\frac{1}{\sqrt{n}}$. For example, absent pre-PCR noise, the pooling of two wells or, equivalently, the use of an average of two technical replicates reduces the CV by $\approx 30\%$. Our theoretical CV model also agreed well with empirical CV estimates, based on technical replicates of dPCR measurements in wastewater samples from Switzerland (Figures 1A and S10). For Influenza A virus, we found a median absolute deviation of 4.64 percentage points between the theoretical curve predicted by Eq. (4) and the locally estimated scatterplot smoothing (LOESS) curve for empirical estimates. Since the empirical CV is based on technical replicates of the same preprocessed sample, we could not assess the effect of pre-PCR noise as predicted by Eq. (7).

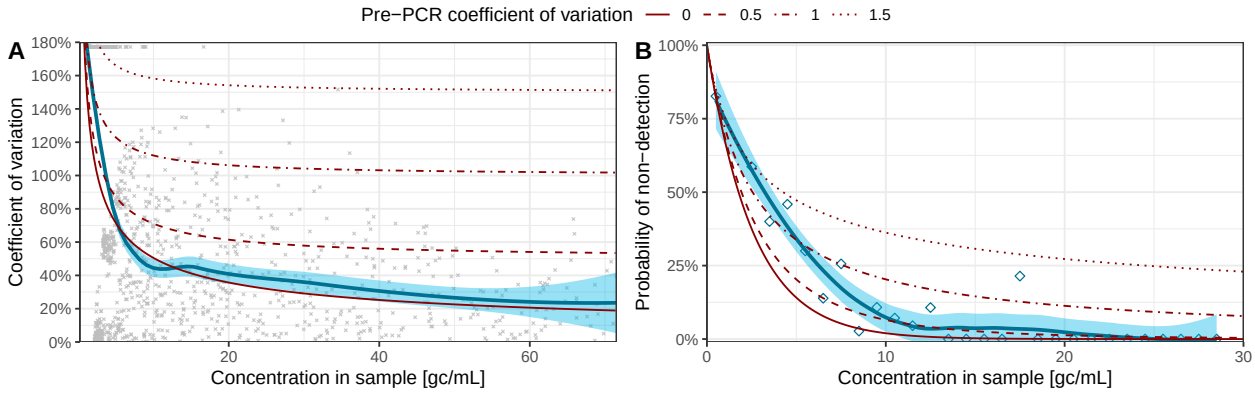

Fig 1. Relationship of the sample concentration with the coefficient of variation and probability of non-detection in digital PCR. Shown is (A) the coefficient of variation (CV) and (B) probability of non-detection (p_0) for a single digital PCR measurement, as a function of the sample concentration λ. Empirical estimates of the CV (dots) and p_0 (diamonds) were computed from duplicate dPCR measurements of Influenza A virus from 14 wastewater treatment plants in Switzerland (Naica Crystal Digital PCR Platform, scaling factor $s = 30$, partition volume $v = 0.519\,\text{nL}$, average partition number $m = 22\,665$). Blue lines with 95% uncertainty intervals show LOESS-based kernel regressions of the empirical estimates, indicating the relationship with the sample concentration λ, respectively. Red lines show the CV and p_0 as predicted by Eq. (7) and Eq. (9), for different strengths of the pre-PCR coefficient of variation ν_{pre} under gamma distributed pre-PCR noise (dashed and dotted lines). As the empirical estimates are based on technical replicates of the same sample, they do not include pre-PCR variation and therefore correspond to $\nu_{\text{pre}} = 0$.

3.2 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 modeling 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 the PCR parameters described in Section 2.6, Eq. (8) predicts that even without pre-PCR noise, there is an 11.5% probability that a concentration of $\lambda = 5\,\text{gc/mL}$ will be measured as zero. For gamma-distributed pre-PCR noise with a CV of $\nu_{\text{pre}} = 0.5$, the non-detection probability increases by another 5.5 percentage points (Figure 1B). Importantly, we find that the non-detection probability curve changes significantly when we model the pre-PCR noise as log-normal instead of gamma-distributed (Figure S9), with gamma-distributed noise leading to higher non-detection probabilities. Moreover, the probability of non-detection decreases with the number of partitions and the number of replicates (Eq. (8)). 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. We compared our model for the probability of non-detection with empirical estimates based on the same real-world measurements as

in Section 3.1 (Figures 1B and S10). Although the empirical estimates rely on a moving window approach and are therefore less accurate, we find a general agreement between the theoretical curves predicted by Eq. (8) and the LOESS curve of the empirical estimates (median absolute deviation: 3.40 percentage points for Influenza A).

3.3 Application to eDNA-based biomonitoring

We used a generalized linear model with our dPCR-specific likelihood function to estimate the half-life of free-eDNA of a marine species in an aquarium seawater experiment by Scriver et al. [27]. In the study, concentrations of the Cytochrome c oxidase subunit 1 (COI) gene of the brown bryozoan "Bugula neritina" were measured in different aquariums with low, medium, and high biomass of the species, using samples taken 0, 4, 8, 24, 48, and 72 hours after removal of the organism from the aquarium. Scriver et al. estimated the exponential decay rate of detectable free-eDNA from the averages of all measurements at each point in time. We reproduced this analysis by fitting a GLM with normal likelihood to the averaged measurements. However, when stratifying the analysis to individual aquariums, i. e. each with smaller sample size, this averaging approach yielded inconsistent results (Figure 2A). When fitting to individual samples using a normal or log-normal likelihood, results remained inconsistent. Specifically, the normal likelihood yielded half-life estimates much lower than the original estimate by Scriver et al. Under the log-normal likelihood, half-life estimates were higher than the original estimate and highly uncertain. In contrast, our dPCR-specific likelihood yielded precise and consistent results across aquariums. The median estimates of the free-eDNA half-life were 19.51 (95% CrI 12.07–45.24), 12.63 (95% CrI 8.95–20.13), and 15.50 (95% CrI 11.17–25.87) hours for low, medium, and high initial biomass, and 18.75 (95% CrI 14.36–26.85) hours for all aquariums combined, which agrees well with the original estimate by Scriver et al. When comparing median concentrations as predicted by the models, we found that predictions under the dPCR-specific likelihood were generally between those under the normal and log-normal likelihood (Figure 2B).

3.4 Application to wastewater-based epidemiology

The use of a dPCR-specific likelihood can improve wastewater-based surveillance of pathogen transmission. Using the normal, log-normal, and dPCR-specific likelihood, we computed estimates and forecasts of concentrations, viral loads, and effective reproduction number R_t values from dPCR

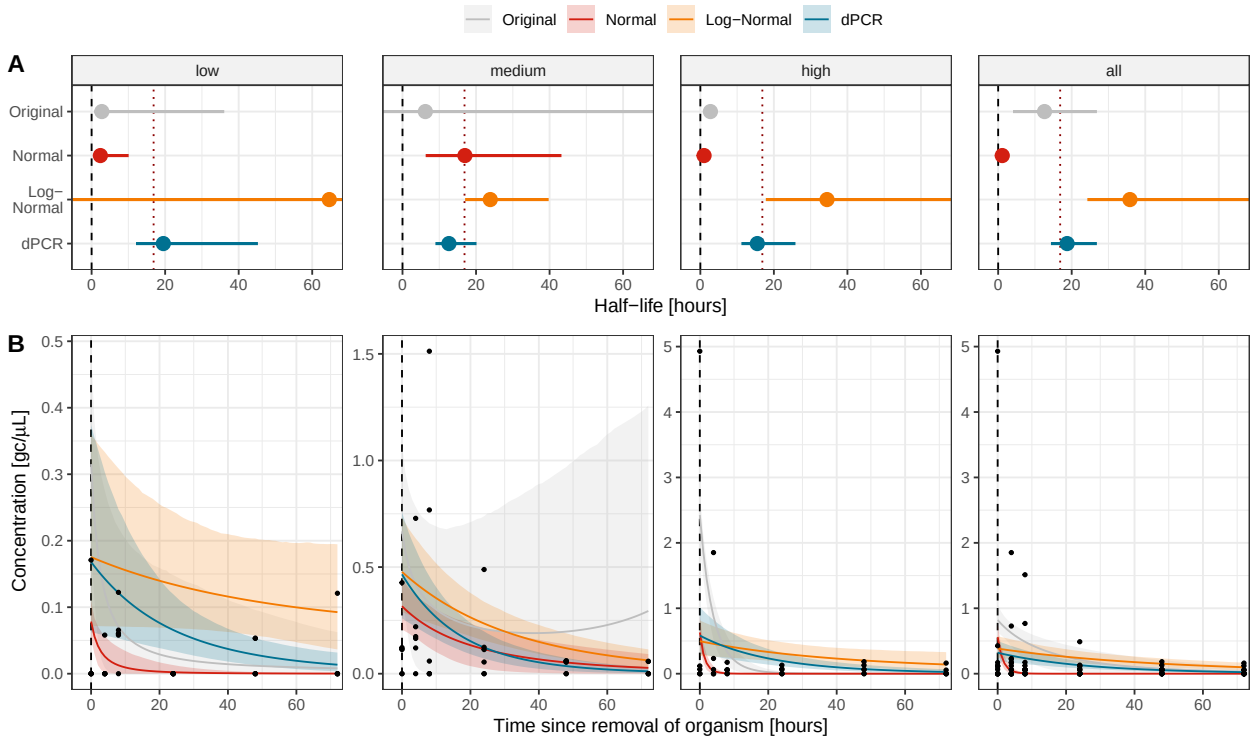

Fig 2. Modeling the exponential decay of eDNA of a marine species in aquarium seawater. (A) Estimated half-life period for free-eDNA of Bugula neritina in aquarium seawater based on longitudinal dPCR measurements after removal of the organism by Scriver et al. [27]. Shown are median estimates (dots) with 95% credible intervals (bars) from a generalized linear model (GLM) applied to data from an aquarium with low, medium, and high initial biomass, respectively, and from all aquariums combined. The GLM was applied to averages of measurements as in the original analysis by Scriver et al. (grey) and to individual measurements using a normal (red), log-normal (orange), and dPCR-specific (blue) likelihood function. Red vertical lines show the original estimate reported by Scriver et al. (B) Measured concentrations (dots) and model predictions using the different likelihoods (colored lines) with 95% credible intervals (shaded areas).

measurements of Influenza A virus at the municipal treatment plant of Zurich, Switzerland (Figure 3). Under the normal likelihood, measurements are modeled to have constant variance, leading to an overestimation of measurement noise at low concentrations. 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 (Figure 3A). The concentration curve predicted by the normal likelihood model was also highly volatile, with the median estimated concentration falling within the previous 95% CrIs on only 56.0% of days. This also resulted in volatile estimates of the viral load (Figure 3B) and the effective reproduction number (Figure 3C), likely due to overfitting of the model at the peak of the seasonal wave. R_t estimates at the start of the seasonal wave were highly uncertain (Figure 3C). Under the log-normal likelihood, measurements are assumed to have a constant coefficient of variation, which is approximately valid at high concentrations. This resulted

in more similar estimates by the log-normal and dPCR-specific likelihood models. Nevertheless, measurement noise tended to be underestimated for low concentrations (80% CrIs covered only 59.7% of observations before Dec 1, 2022) and overestimated for high concentrations (80% CrIs covered 100% of observations after Dec 1, 2022) by the log-normal model. As a result, the estimated and forecasted load curve and R_t estimates were more uncertain than under the dPCR-specific likelihood. By modeling the concentration-dependent coefficient of variation, the dPCR-specific likelihood model 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) and estimated viral load and R_t values with reduced uncertainty. It also achieved more accurate forecasts of measurements after Dec 26, 2022 ($\overline{\text{CRPS}} = 12.04$) than the normal ($\overline{\text{CRPS}} = 20.08$) and log-normal likelihood ($\overline{\text{CRPS}} = 18.19$). In Figures S16–S20, we show additional results from fitting the models on different date ranges of measurements at the treatment plants of Zurich and Geneva, Switzerland.

When fitting the dPCR-specific model to wastewater measurements, we assumed that the precise characteristics of the wastewater assay are unknown and only assigned broad priors for the number of total partitions m and the conversion factor c. Despite the additional parameter uncertainty, this yielded highly similar estimates as a fully informed model where m and c are fixed to their true values (Figure S21). Moreover, additional insights about the measurement process can be gained from the estimated laboratory parameters. For example, the posterior distribution for ν_{pre} had a median of 17.0% (95% credible interval: 5.3% – 28.1%) and differed clearly from its prior (Figure S15). As this parameter describes the variation in concentrations before the PCR, it allows us to distinguish the technical noise of dPCR from other 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 theoretical minimum CV that would be achieved without preprocessing noise or other unexplained variation (15.1%).

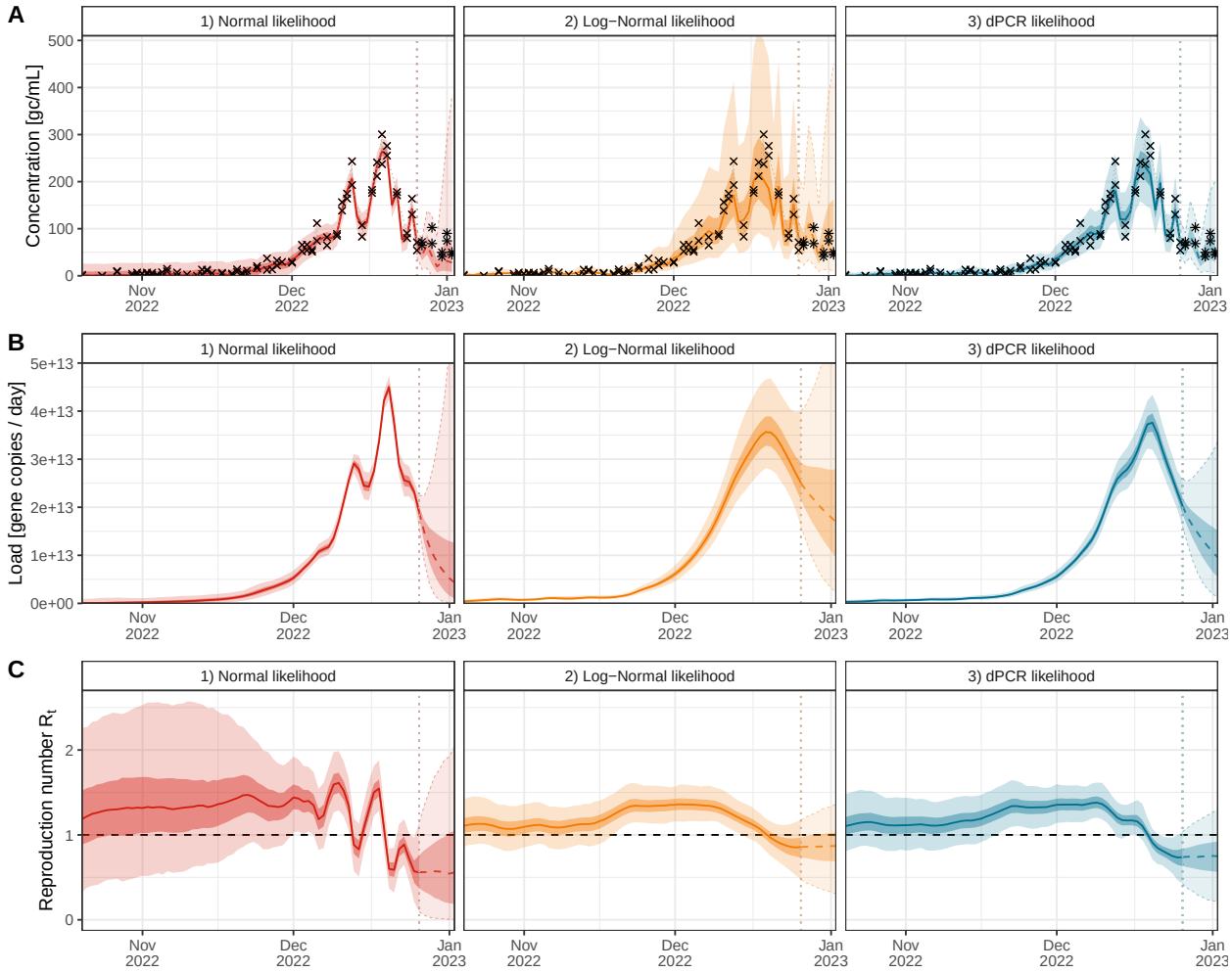

Fig 3. Comparison of inference and short-term forecasts from wastewater dPCR measurements using different likelihood functions: Shown are the results of fitting an epidemiological wastewater model (EpiSewer) to duplicate dPCR measurements of Influenza A virus concentrations at the municipal treatment 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.

4 Discussion

In this work, we have developed a model for the statistical analysis of dPCR concentration measurements. Our model reflects two important properties of dPCR data that violate standard assumptions in regression analysis. First, measurement noise in dPCR is concentration-dependent, with the coefficient of variation increasing exponentially as concentrations become small. When not accounted

for, this heteroskedasticity leads to unreliable estimates especially at small sample sizes. Second, even with the target gene present in a sample, there can be a significant probability of non-detection, resulting in zero measurements. These data must be removed or imputed when modeling measurements as log-normally distributed, leading to biased estimates. Our likelihood function avoids these problems by modeling both concentration-dependent measurement noise and the probability of zero measurements while accounting for pre-PCR variation, e. g. from sample extraction and preprocessing.

Using real-world examples of dPCR measurements, we have shown that our model achieves more precise and reliable estimates from dPCR data than simpler likelihoods currently used in practice, such as normal or log-normal distributions. This can impact inferences about underlying biological processes, such as eDNA decay in biomonitoring or pathogen transmission dynamics in wastewater-based epidemiology. We argue that a log-normal distribution with a constant coefficient of variation (or, equivalently, modeling log-transformed measurements as normally distributed) may provide a sufficient approximation, but only if all concentrations are sufficiently high, e. g. above the Limit of Quantification (LoQ), which is operationally defined as a concentration below which the CV surpasses a fixed threshold such as 30% [35]. When concentrations are lower, it is important to model a concentration-dependent coefficient of variation and the probability of zero measurements to avoid biased or highly uncertain estimates.

Our findings also have broader implications for data analysis in dPCR-based environmental surveillance. In particular, we argue that statistical models should be applied directly to measured concentrations rather than to derived indicators. For example, in wastewater surveillance, measured pathogen concentrations are often normalized by flow rates ("loads") or fecal biomarker concentrations, such as crAssphage or PMMoV [44, 45]. However, these normalized quantities are not ideal as direct inputs to wastewater models since their measurement distribution depends on the normalization factor. For example, depending on the flow rate, the same "load" may correspond to vastly different concentrations, with different measurement noise and probabilities of non-detection. The same applies to concentrations normalized by fecal markers, with the added complication that marker concentrations are themselves subject to measurement noise. Therefore, we recommend using non-normalized concentrations as observations and incorporating normalization factors into the model through a latent variable approach, as e. g. done in the R package "EpiSewer" [38].

Moreover, for modeling dPCR measurements at low concentrations, the likelihood proposed in this work should be more appropriate than censoring approaches in which zero measurements or measurements below the LoD are replaced with a predefined value [24, 46]. This is because censoring uses a fixed threshold, while the probability of non-detection gradually increases as the concentration approaches zero. Furthermore, censoring applies to the measurements, which are subject to noise, not to the true concentration in the sample. Similar arguments apply to the LoQ, considering the gradual increase of the CV as concentrations approach zero and the fact that the CV depends on the true sample concentration, not the measurement. We thus argue that it is better to model the increase in CV at low concentrations than to classify measurements by a predefined threshold. Indeed, removing measurements below the LoQ from statistical analyses may lead to an overestimation of concentrations.

Using Bayesian inference, our dPCR-specific likelihood can be fitted directly to concentration measurements, as are primarily reported by studies and laboratories. This makes our approach widely applicable, even in settings where the number of positive and total partition counts is not reported. We note that in our examples, concentrations were measured in gene copies per volume, but in the case of solids-based extraction, measurements can also have other units, such as gene copies per gram of dry weight [44]. Information about sample dilution and parameters of the PCR, such as partition number and volume, can be integrated by fixing the respective parameters to their values. However, in our application to wastewater-based epidemiology, our model achieved precise estimates even when the PCR parameters were unknown and broad priors were used.

We note a few limitations of our model. First, we have tested our estimators for the CV and probability of non-detection only for log-normally and gamma-distributed pre-PCR noise. In principle, any positive continuous distribution with a well-defined moment generating function (MGF) could be used, as long as a closed-form expression or accurate approximation of the MGF is available. We acknowledge that choosing a pre-PCR noise distribution can be difficult, as this parameter summarizes all unexplained variation in a model. For the CV, we found that the exact distribution of pre-PCR noise is irrelevant 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. Second, we assumed that in expectation, the dPCR measurements are equal to the true sample

concentration, i. e. we do not account for unknown systematic distortions of dPCR measurements such as from miscalibration. When estimating PCR parameters jointly from the data, we also assumed that these are stable over time, i. e. systematic changes in the laboratory protocol may not be detected without additional information. Third, we did not account for variation in the volume of partitions, which can be an additional but minor source of noise in droplet-based systems [31, 47]. Fourth, 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. Fifth, 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. 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. Finally, since this work focused on the quantification of targets already present in the environment, we assumed that the true concentration can be arbitrarily small but is larger than zero. If, instead, the objective is to determine whether a particular target is present or absent, it may also be important to model the probability of false positives [35].

We have implemented our dPCR-specific likelihood in the R package "dPCRfit" [37], providing a Bayesian generalized linear model for regression analyses of dPCR data, e. g. to improve eDNA abundance estimates in biomonitoring [25, 26]. We have also integrated the likelihood as a module in the "EpiSewer" package, e. g. to improve wastewater-based estimates of pathogen transmission rates [38]. Our theoretical results can also benefit assay development and optimization by informing the effect of laboratory parameters on the precision of measurements. For example, we show that pre-PCR noise strongly influences the probability of non-detection, which highlights the importance of including extraction and preprocessing steps when determining the Limit of Detection [34, 35]. While this work has focused on digital PCR, we expect that similar statistical relationships and models can be established for quantitative PCR methods [48]. Overall, we envision that accurate modeling of laboratory and measurement processes will improve environmental surveillance and provide new insights into environmental variability by disentangling it from measurement noise.

References

1. 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.
2. Beng KC, Corlett RT. Applications of Environmental DNA (eDNA) in Ecology and Conservation: Opportunities, Challenges and Prospects. Biodiversity and Conservation. 2020;29(7):2089–2121.
3. Barnes MA, Turner CR. The Ecology of Environmental DNA and Implications for Conservation Genetics. Conservation Genetics. 2016;17(1):1–17.
4. Rees HC, Maddison BC, Middleditch DJ, Patmore JRM, Gough KC. REVIEW: The Detection of Aquatic Animal Species Using Environmental DNA – a Review of eDNA as a Survey Tool in Ecology. Journal of Applied Ecology. 2014;51(5):1450–1459.
5. 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.
6. 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.
7. 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.
8. Bivins A, Bibby K. Wastewater Surveillance during Mass COVID-19 Vaccination on a College Campus. Environmental Science & Technology Letters. 2021;8(9):792–798.
9. 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.
10. 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.

11. 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.

12. 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.

13. 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.

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

15. 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.

16. Doi H, Takahara T, Minamoto T, Matsuhashi S, Uchii K, Yamanaka H. Droplet Digital Polymerase Chain Reaction (PCR) Outperforms Real-Time PCR in the Detection of Environmental DNA from an Invasive Fish Species. Environmental Science & Technology. 2015;49(9):5601–5608.

17. Nathan LM, Simmons M, Wegleitner BJ, Jerde CL, Mahon AR. Quantifying Environmental DNA Signals for Aquatic Invasive Species Across Multiple Detection Platforms. Environmental Science & Technology. 2014;48(21):12800–12806.

18. 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.

19. 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.

20. Capo E, Spong G, Königsson H, Byström P. Effects of Filtration Methods and Water Volume on the Quantification of Brown Trout (Salmo Trutta) and Arctic Char (Salvelinus Alpinus) eDNA

Concentrations via Droplet Digital PCR. Environmental DNA. 2020;2(2):152–160.

21. Wood SA, Pochon X, Laroche O, von Ammon U, Adamson J, Zaiko A. A Comparison of Droplet Digital Polymerase Chain Reaction (PCR), Quantitative PCR and Metabarcoding for Species-Specific Detection in Environmental DNA. Molecular Ecology Resources. 2019;19(6):1407–1419.

22. Thalinger B, Wolf E, Traugott M, Wanzenböck J. Monitoring Spawning Migrations of Potamodromous Fish Species via eDNA. Scientific Reports. 2019;9(1):15388.

23. Renninger N, Nastasi N, Bope A, Cochran SJ, Haines SR, Balasubrahmaniam N, et al. Indoor Dust as a Matrix for Surveillance of COVID-19. mSystems. 2021;6(2):10.1128/msystems.01350–20.

24. Uthicke S, Lamare M, Doyle JR. eDNA Detection of Corallivorous Seastar (Acanthaster Cf. Solaris) Outbreaks on the Great Barrier Reef Using Digital Droplet PCR. Coral Reefs. 2018;37(4):1229–1239.

25. Chucholl F, Fiolka F, Segelbacher G, Epp LS. eDNA Detection of Native and Invasive Crayfish Species Allows for Year-Round Monitoring and Large-Scale Screening of Lotic Systems. Frontiers in Environmental Science. 2021;9.

26. Capo E, Spong G, Norman S, Königsson H, Bartels P, Byström P. Droplet Digital PCR Assays for the Quantification of Brown Trout (Salmo Trutta) and Arctic Char (Salvelinus Alpinus) from Environmental DNA Collected in the Water of Mountain Lakes. PLOS ONE. 2019;14(12):e0226638.

27. Scriver M, von Ammon U, Youngbull C, Pochon X, Stanton JAL, Gemmell NJ, et al. Drop It All: Extraction-Free Detection of Targeted Marine Species through Optimized Direct Droplet Digital PCR. PeerJ. 2024;12:e16969.

28. 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.

29. 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.

30. 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.

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

32. White H. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. Econometrica. 1980;48(4):817–838.

33. 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.

34. 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.

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

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

37. Lison A. dPCRfit v0.0.1; 2025. Zenodo. Available from: https://doi.org/10.5281/zenodo.15062625.

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

39. 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.

40. 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.

41. 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.

42. 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.

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

44. 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.

45. 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.

46. 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.

47. 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.

48. 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

Conceptualization: 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

Visualization: AL

Writing – original draft: AL

Writing – review & editing: AL, TJ, TS

Acknowledgments

This work contains a secondary analysis of open wastewater data provided by EAWAG, Swiss Federal Institute of Aquatic Science and Technology. We kindly thank the members of the Wastewater Monitoring Laboratory, including 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.

Supporting Information: Improving inference in environmental surveillance by modeling 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 Averaging of dPCR measurements **2**

B Coefficient of variation of concentration estimates **5**

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

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

Gamma distribution ▪ Log-normal distribution ▪ Taylor series expansion

C Probability of non-detection **8**

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

C.2 Probability of non-detection with pre-PCR variation . . . 8

D Simulation analysis **9**

D.1 Coefficient of variation . . . 9

D.2 Probability of non-detection . . . 14

E Comparison with empirical data **15**

F Likelihoods for non-zero dPCR measurements **18**

F.1 Conditional distribution of non-zero measurements . . . 18

F.2 Continuous vs. discrete measurement distributions . . . 19

F.3 Comparison of continuous measurement distributions . . . 20

G Inference from dPCR measurements **21**

G.1 Estimation 21

G.2 Validation 22

H Application to eDNA-based biomonitoring **23**

I Application to wastewater-based epidemiology **24**

I.1 PCR parameters 25

Prior for number of partitions ▪ Prior for conversion factor ▪ Prior for pre-PCR coefficient of variation ▪ Estimated parameters

I.2 Forecasts 26

I.3 Additional results 28

References **35**

A Averaging of dPCR measurements

There exist different approaches to combine replicate measurements of a sample into a single concentration estimate. Dorazio and Hunter [1] describe a maximum likelihood estimator for λ

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

based on 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. This approach is typically used when pooling droplet counts from several wells.

A different approach also used in practice is to compute the arithmetic mean of individual ML estimates for each replicate

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

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 dPCR 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 in order to study the effect of differences in the total number of partitions. 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 S1 shows how the arithmetic mean of measurements deviates from the unweighted ML estimate. We find that the arithmetic mean has a small positive bias compared to the unweighted ML estimate, as 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 $1000\,\mathrm{gc/mL}$, the deviation is considerably less than $1\,\mathrm{gc/mL}$.

Next, we compared the unweighted ML estimate defined in Eq. (3) to the weighted ML estimate defined in Eq. (1). We found larger deviations than between the arithmetic mean and unweighted ML estimate, but no systematic bias (Figure S2). Deviations were particularly large in cases where one of the replicates only had few total partitions, such that the corresponding concentration estimate was noisy. In the arithmetic mean and the unweighted ML estimate, such replicates are weighted equally to other replicates. 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. 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 individual estimates, but in practice the differences will be small compared to the overall uncertainty of measurements, except in cases where the total number of partitions varies greatly.

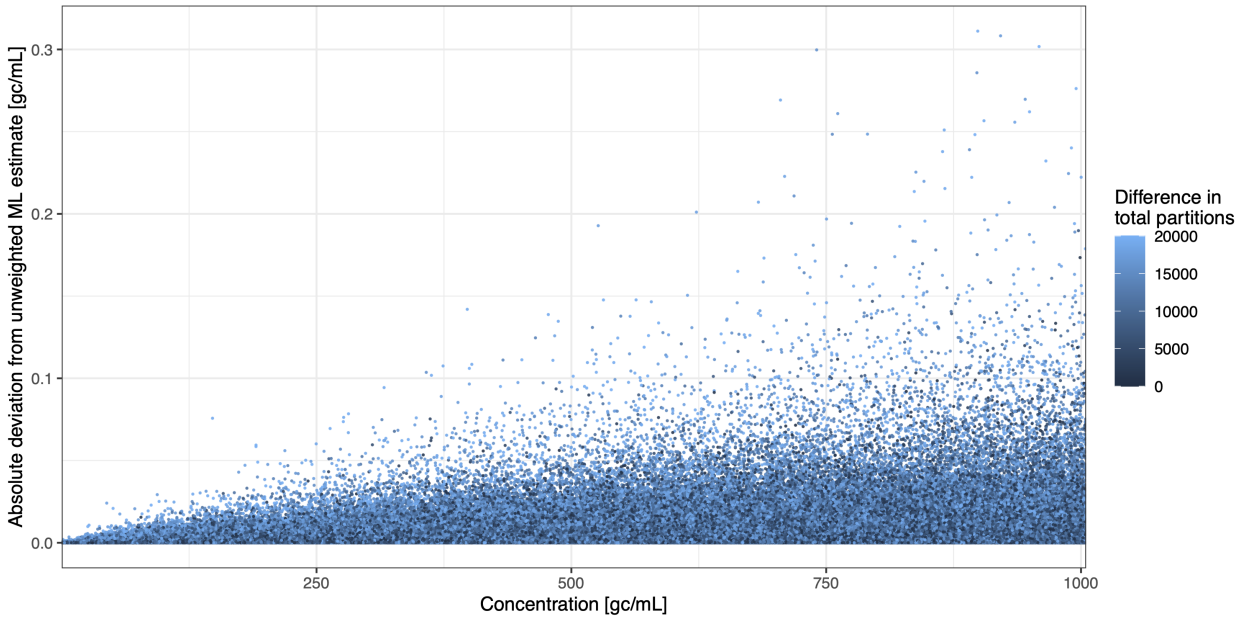

Fig S1. Comparison of arithmetic mean estimate with unweighted ML estimate of concentrations. Shown are deviations of the arithmetic mean of replicate measurements from an unweighted ML estimate for simulated measurements as a function of concentration, and for different degrees of partition number variation between replicates. Simulated measurements assume a scaling factor of $s = 30$ and partition volume of $v = 0.519\,\mathrm{nL}$.

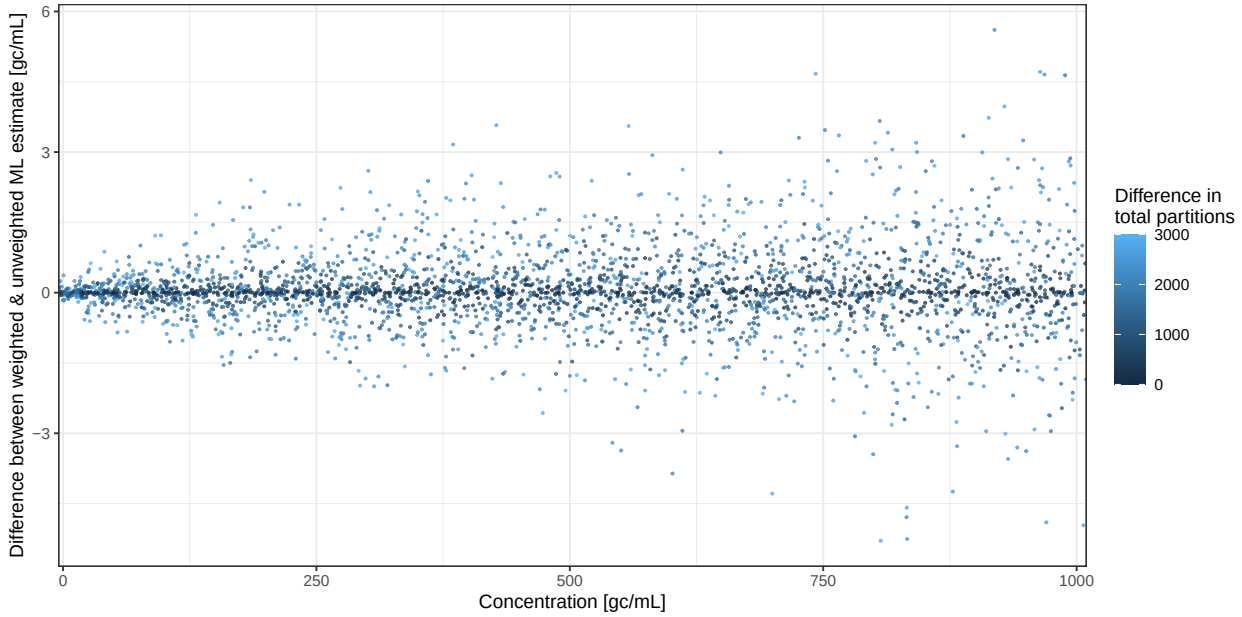

Fig S2. Comparison of unweighted and weighted ML estimates of concentrations. Shown are deviations of the unweighted ML estimate from the weighted ML estimate for simulated measurements as a function of concentration, and for different degrees of partition number variation between replicates. Simulated measurements assume a scaling factor of $s = 30$ and partition volume of $v = 0.519\,\mathrm{nL}$.

B Coefficient of variation of concentration estimates

B.1 CV without pre-PCR variation

To derive the variance of a single maximum likelihood concentration estimate $\tilde{\Lambda} = -\frac{1}{c}\log\left(1-\frac{Y}{m}\right)$ [1], 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 $p = 1-\exp(-\lambda c)$ [1], 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 $\tilde{\Lambda}$ [1].

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

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

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

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

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 [2], 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

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

B.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 $s\Lambda_{\text{pre}}$, where Λ_{pre} is a random variable drawn from a parametric distribution with mean λ and coefficient of variation $\nu_{\text{pre}} = \frac{\sigma_{\text{pre}}}{\lambda} > 0$. Using the Law of Total Variance, we can express the variance of $\hat{\Lambda}$ under pre-PCR noise as

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

with the "explained" component of the variance

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

and the "unexplained" component of the variance

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

B.2.1 Gamma distribution

If we model Λ_{pre} as gamma distributed with parameters $\alpha = \frac{1}{\nu_{\text{pre}}^2}$ and $\beta = \frac{1}{\lambda \nu_{\text{pre}}^2}$, 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

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

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

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

B.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 [3] to obtain

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

where w is an evaluation of the Lambert W function [4] with

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

B.2.3 Taylor series expansion

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

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

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

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

and thus

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

C Probability of non-detection

C.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 [5]. 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.

C.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

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

D Simulation analysis

In the following section, we provide results of simulation analysis in which we compared our analytical approximations for the coefficient of variation and the probability of non-detection with simulated dPCR measurements. For our simulations, we assumed exemplary PCR parameters based on the Naica Crystal Digital PCR Platform with Sapphire Chips (Stilla Technologies, France), a scaling factor of $s = 30$, a partition volume of $v = 0.519\,\text{nL}$, and $m = 25000$ partitions per measurement, but we also provide a sensitivity analysis of these parameters in Figures S6 and S7. Positive partition counts were simulated under the binomial model described in the main text and transformed into a concentration estimate using the maximum likelihood estimator by Dorazio and Hunter [1], see main text.

D.1 Coefficient of variation

Figure S3 shows the coefficient of variation (CV) as a function of the sample concentration under gamma distributed pre-PCR noise. The theoretical prediction for the CV based on Eq. (16) matches the simulated dPCR data well. 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.

Figure S4 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, two theoretical approximations are shown. While dashed lines show an approximation

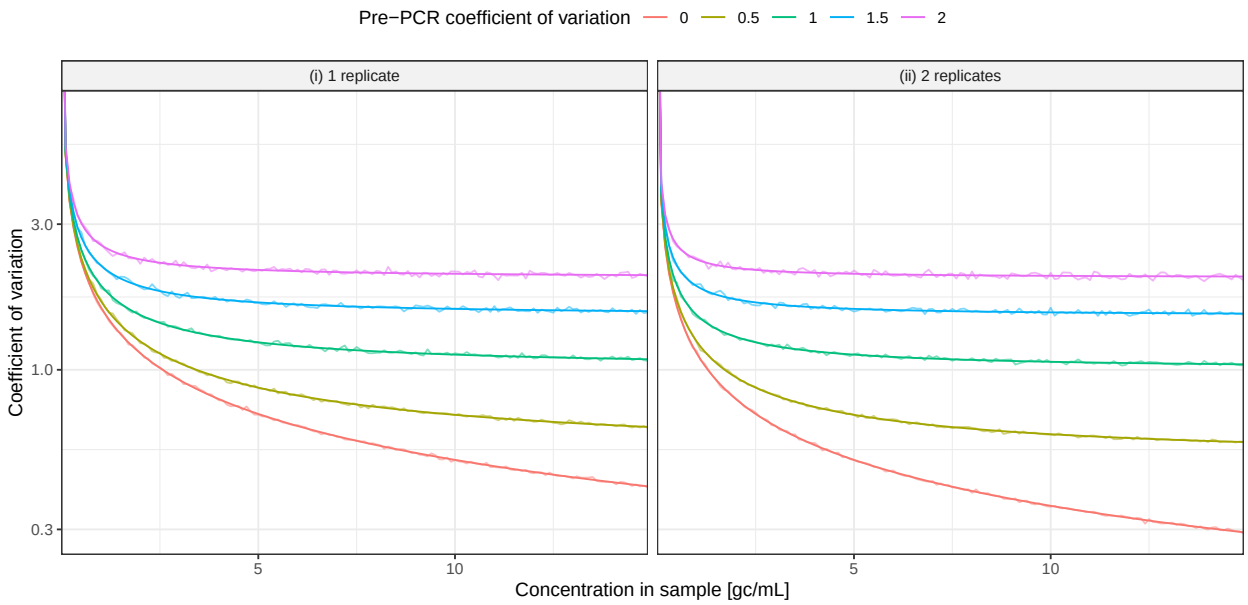

Fig S3. Relationship between concentration and coefficient of variation of dPCR measurements under gamma distributed pre-PCR noise. Shown is the coefficient of variation (CV) of measurements from digital PCR for (i) a single replicate and (ii) the arithmetic mean of two replicates, as a function of the sample concentration λ under gamma 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 predicted by Eq. (16). Colors correspond to different strengths of pre-PCR noise, as measured by the pre-PCR coefficient of variation ν_{pre}.

of the log-normal MGF by Asmussen et al. [3] (see Supplement B.2.2), straight lines show the CV as approximated using a Taylor series expansion (see Supplement B.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.

Figure S5 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

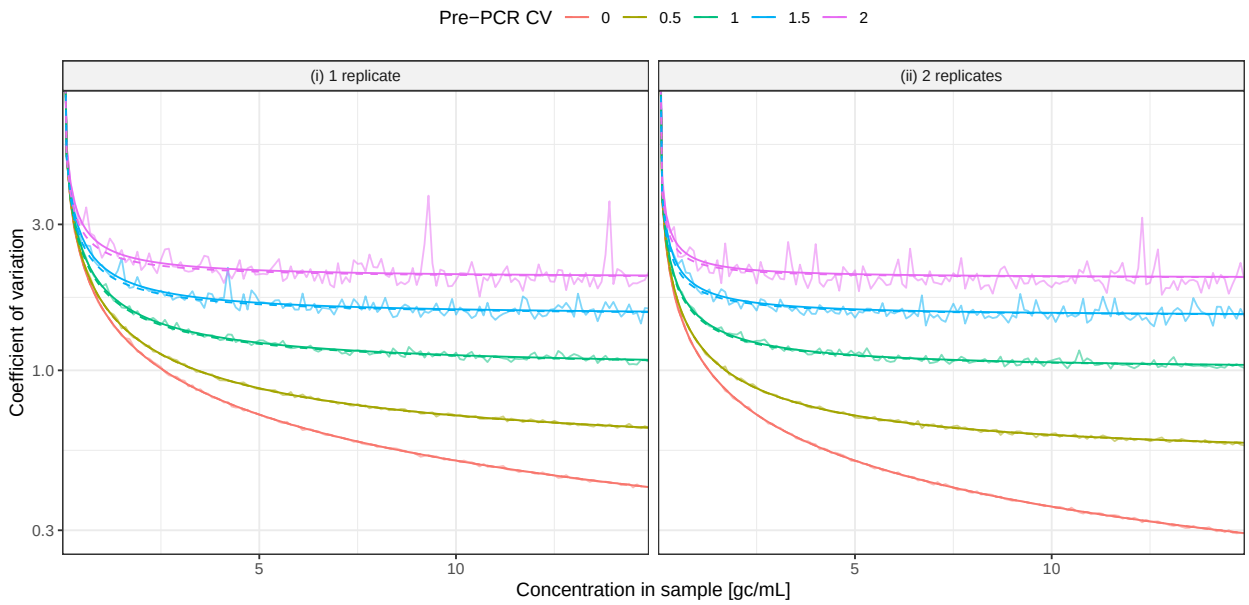

Fig S4. Relationship between concentration and coefficient of variation of dPCR measurements under log-normally distributed pre-PCR noise. Shown is the coefficient of variation (CV) of measurements from digital PCR for (i) a single replicate and (ii) the arithmetic mean of two replicates, as a function 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. [3]. Colors correspond to different strengths of pre-PCR noise, as measured by the pre-PCR coefficient of variation ν_{pre}.

gamma case and the log-normal approximation.

Figure S6 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 S7 shows a sensitivity analysis of the CV as a function of the sample concentration to 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.

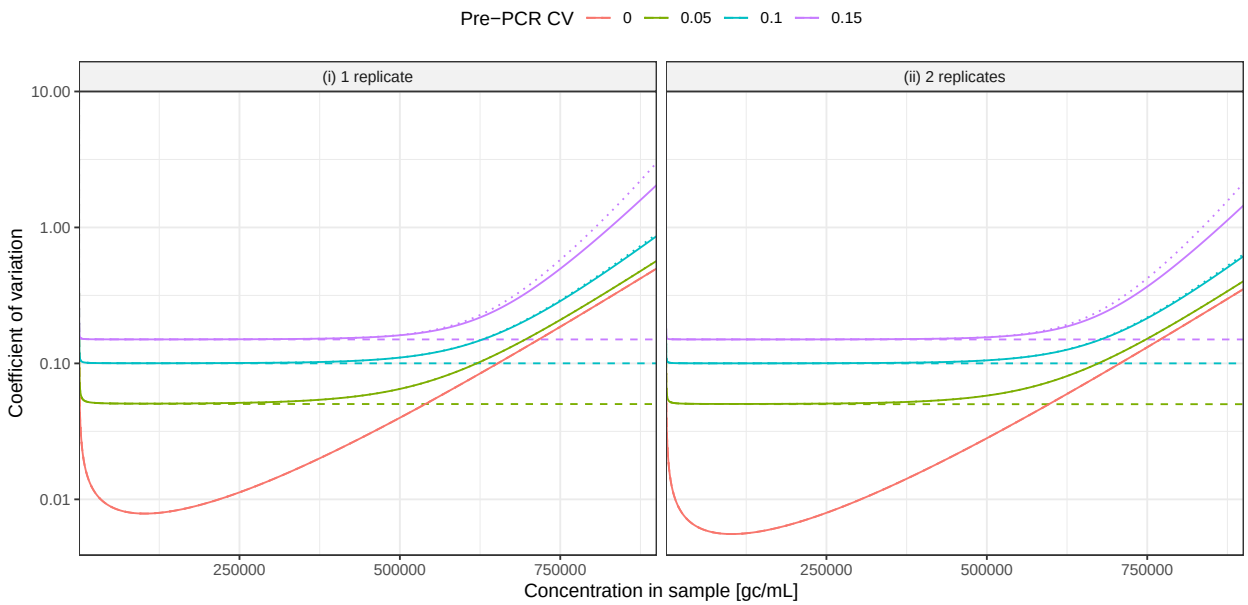

Fig S5. Relationship between concentration and coefficient of variation of dPCR measurements across wide range of concentrations. Shown is the coefficient of variation (CV) of measurements from digital PCR for (i) a single replicate and (ii) the arithmetic mean of two replicates, as a function of the 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}.

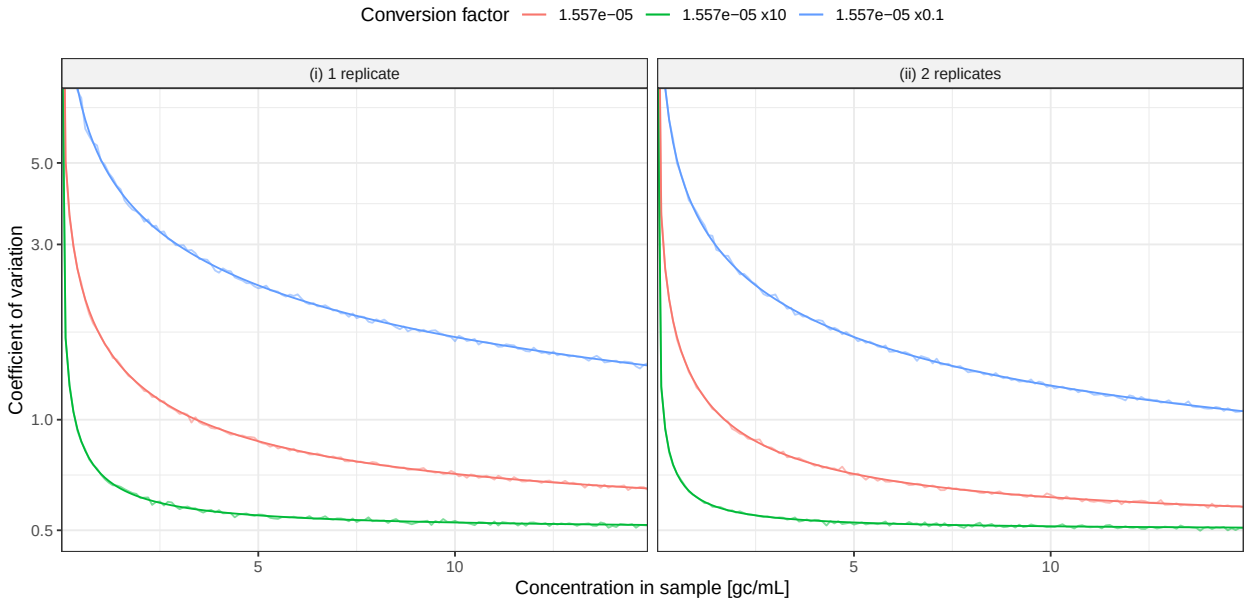

Fig S6. Sensitivity of the coefficient of variation of dPCR measurements to the conversion factor. Shown is the coefficient of variation (CV) of measurements from digital PCR for a conversion factor as in the main analysis ($c = 30 \cdot 0.519\,\text{nL}$, red), 10x as large (green), and 10x as small (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).

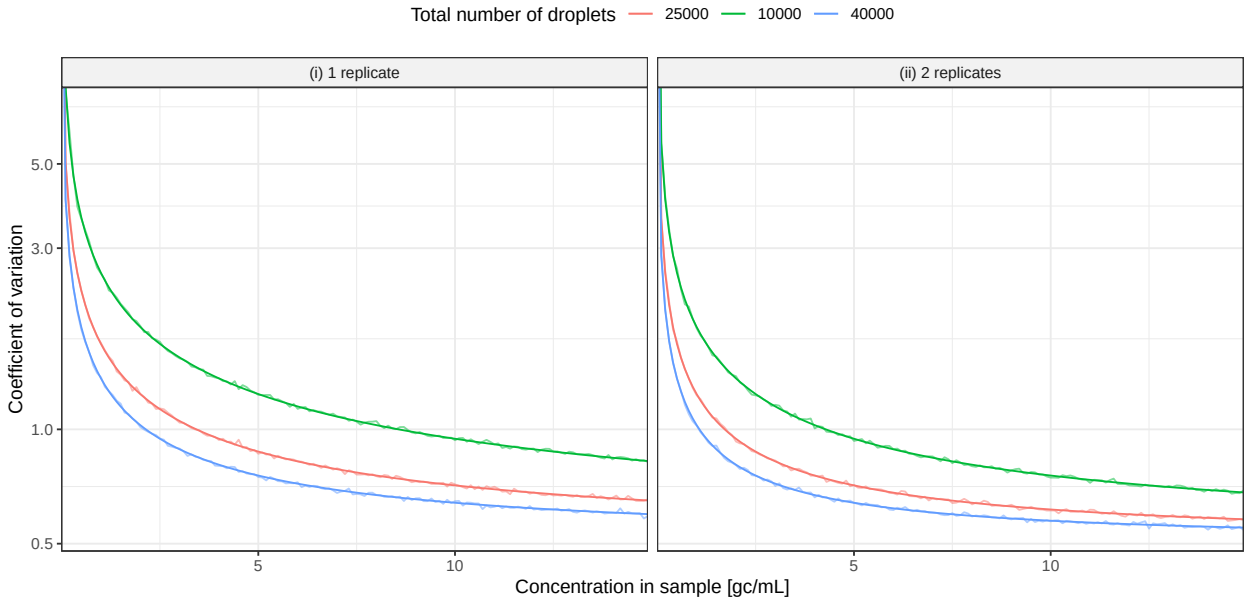

Fig S7. Sensitivity of the coefficient of variation of dPCR measurements to the total number of partitions. Shown is the coefficient of variation (CV) of measurements from digital PCR for 25000 total partitions, i. e. as in the main analysis (red), 10000 total partitions (green), and 40000 total partitions (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).

D.2 Probability of non-detection

Figure S9 shows the probability of non-detection as a function of the sample concentration under gamma distributed pre-PCR noise. The presence of pre-PCR noise considerably increases the probability of zero measurements. Moreover, from Eq. (24) we see that the probability of non-detection decreases with the number of partitions and the number of replicates.

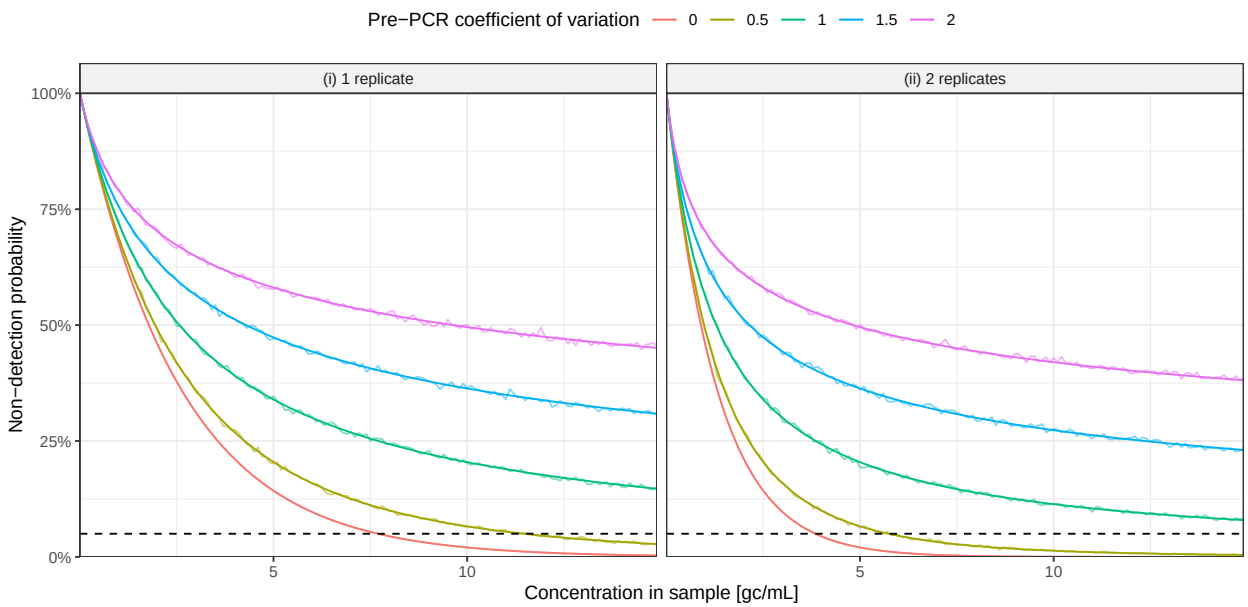

Fig S8. Relationship between concentration and probability of non-detection of dPCR measurements under gamma distributed pre-PCR noise. Shown is the probability of a zero measurement (i. e. non-detection) in digital PCR for (i) a single replicate and (ii) the arithmetic mean of two replicates, as a function of the sample concentration λ under gamma 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 by Eq. (24). 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}.

Figure S9 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. [3] (see Supplement C.2) is shown. As can be seen, the approximation matches the simulated CV values well, but slightly underestimates the CV at low concentrations when the pre-PCR variation is very high (Figure S9). Compared to the case of gamma distributed pre-PCR noise (Figure S8), the probability of non-detection falls considerably faster under log-normally distributed pre-PCR noise.

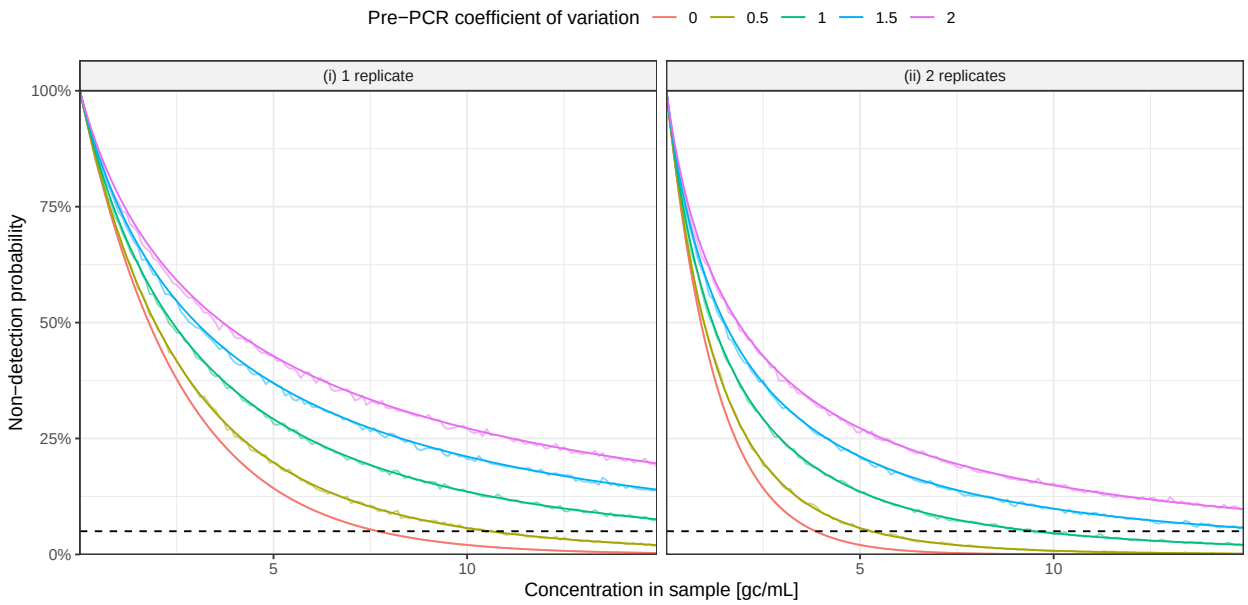

Fig S9. Relationship between concentration and probability of non-detection of dPCR measurements under log-normally distributed pre-PCR noise. Shown is the probability of a zero measurement (i. e. non-detection) in digital PCR for (i) a single replicate and (ii) the arithmetic mean of two replicates, 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. [3]. 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}.

E Comparison with empirical data

We used dPCR data of pathogen concentrations in Swiss wastewater to validate our statistical model of dPCR measurements. For this 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. [6] and Nadeau et al. [7], 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.

For each target and sample, the CV was empirically estimated from two technical replicates, using the formula

$$\hat{\nu}_{\hat{\lambda}} = \sqrt{\frac{\pi}{2}} \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\lambda}_i - \mu_{\hat{\lambda}})^2}}{\mu_{\hat{\lambda}}}, \tag{26}$$

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) [8] to estimate a smooth relationship of the target concentration with the CV and probability of non-detection, respectively.

Figure S10 shows the comparison between theoretical predictions and the real-world dPCR measurements for Influenza A and B virus, RSV, and SARS-CoV-2.

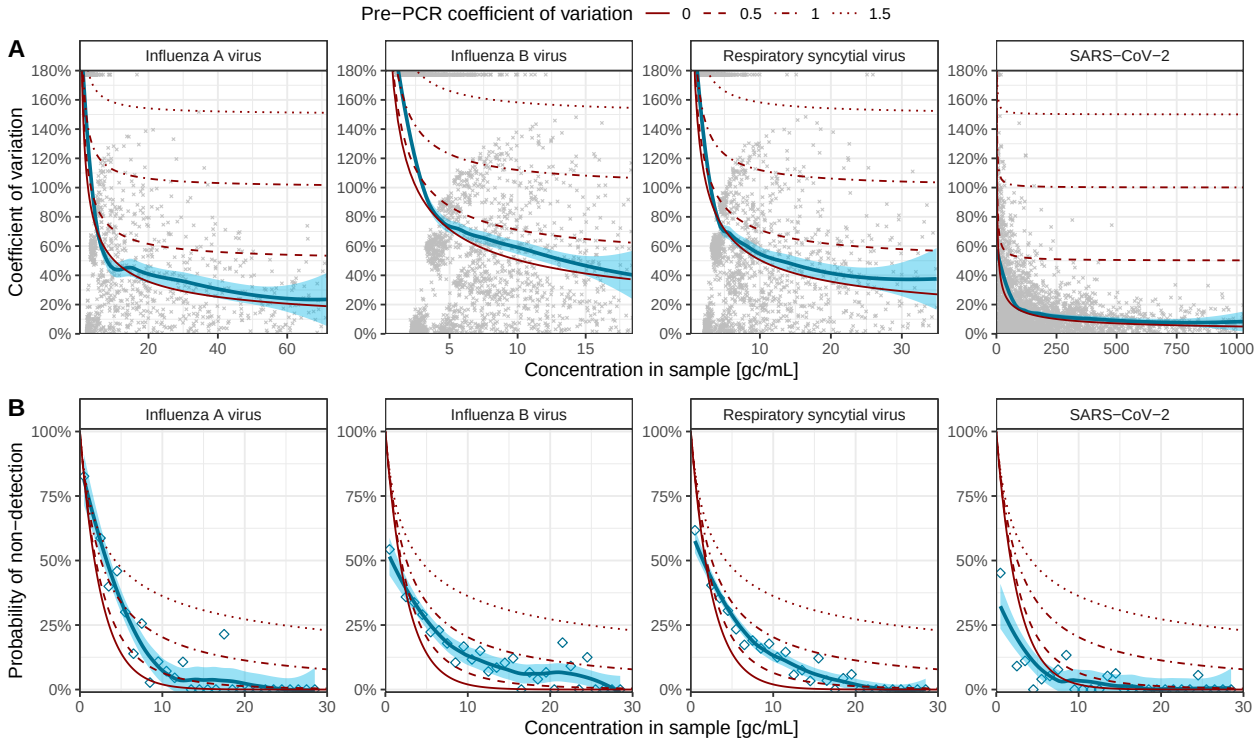

Fig S10. Comparison of theoretical predictions with real-world dPCR measurements. Shown are empirical estimates of (A) the coefficient of variation and (B) the probability of non-detection for dPCR measurements of Influenza A and B virus, RSV, and SARS-CoV-2 concentrations from 14 wastewater treatment 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, for different strengths of the pre-PCR coefficient of variation ν_{pre} under gamma distributed pre-PCR noise (dashed and dotted lines). 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.

F Likelihoods for non-zero dPCR measurements

F.1 Conditional distribution of non-zero measurements

In the proposed hurdle model, $f_{\mathrm{PCR}}(x)$ 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.

$$\mathbb{E}\left[\hat{\Lambda}\right] = \mathbb{E}\left[\hat{\Lambda}\middle|\hat{\Lambda}=0\right]\cdot P(\hat{\Lambda}=0) + \mathbb{E}\left[\hat{\Lambda}\middle|\hat{\Lambda}>0\right]\cdot P(\hat{\Lambda}>0), \tag{27}$$

we know that

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

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

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

Expanding the conditional variance and inserting the above, we get

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

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

$$\mathrm{Var}\left[\hat{\Lambda}\middle|\hat{\Lambda}>0\right] = \frac{\mathrm{Var}\left[\hat{\Lambda}\right] + \mathbb{E}\left[\hat{\Lambda}\right]^2}{1-p_{\mathrm{zero}}(\lambda)} - \frac{\mathbb{E}\left[\hat{\Lambda}\right]^2}{(1-p_{\mathrm{zero}}(\lambda))^2} = \frac{Var\left[\hat{\Lambda}\right]}{1-p_{\mathrm{zero}}(\lambda)} - \mathbb{E}\left[\hat{\Lambda}\right]^2 \frac{p_{\mathrm{zero}}(\lambda)}{(1-p_{\mathrm{zero}}(\lambda))^2}. \tag{31}$$

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

$$\mathrm{CV}[\hat{\Lambda}|\hat{\Lambda}>0] = \sqrt{\frac{\mathrm{Var}\left[\hat{\Lambda}\middle|\hat{\Lambda}>0\right]}{\mathbb{E}\left[\hat{\Lambda}\middle|\hat{\Lambda}>0\right]^2}} \tag{32}$$

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

$$= \sqrt{\frac{\mathrm{Var}\left[\hat{\Lambda}\right]}{\mathbb{E}\left[\hat{\Lambda}\right]^2}(1-p_{\mathrm{zero}}(\lambda)) - p_{\mathrm{zero}}(\lambda)} \tag{34}$$

$$= \sqrt{\nu_{\hat{\Lambda}}(\lambda)^2(1-p_{\mathrm{zero}}(\lambda)) - p_{\mathrm{zero}}(\lambda)}. \tag{35}$$

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 visualized in Figure S11.

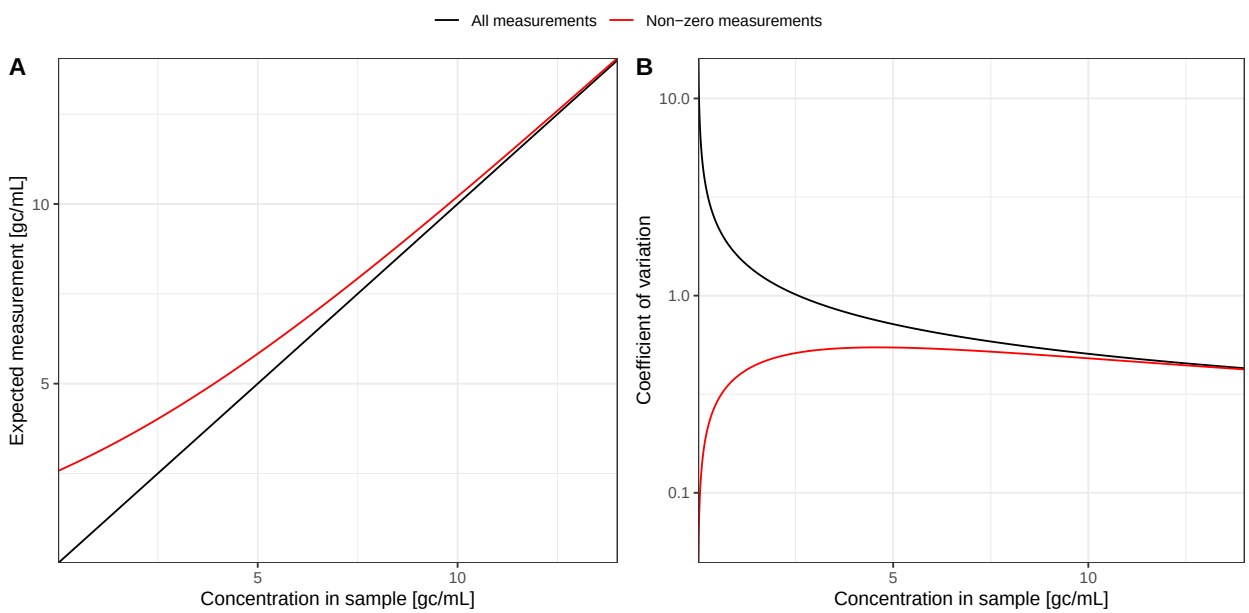

Fig S11. Effect of conditioning on non-zero measurements on the mean and coefficient of variation of measurements. Shown are the (A) mean and (B) coefficient of variation of the unconditional (black, all measurements) and conditional (red, non-zero measurements) distribution of measurements as a function 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

$$y = m(1 - \exp(-\tilde{\lambda}c)), \tag{36}$$

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 S12 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 S12A). 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 S12B), as indicated by their discretized Kullback-Leibler divergence

to true distribution ($\mathrm{KL}_{\mathrm{Gamma}} = 0.01$, $\mathrm{KL}_{\text{Log-normal}} = 0.05$, $\mathrm{KL}_{\text{Inverse Gaussian}} = 0.06$). 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.

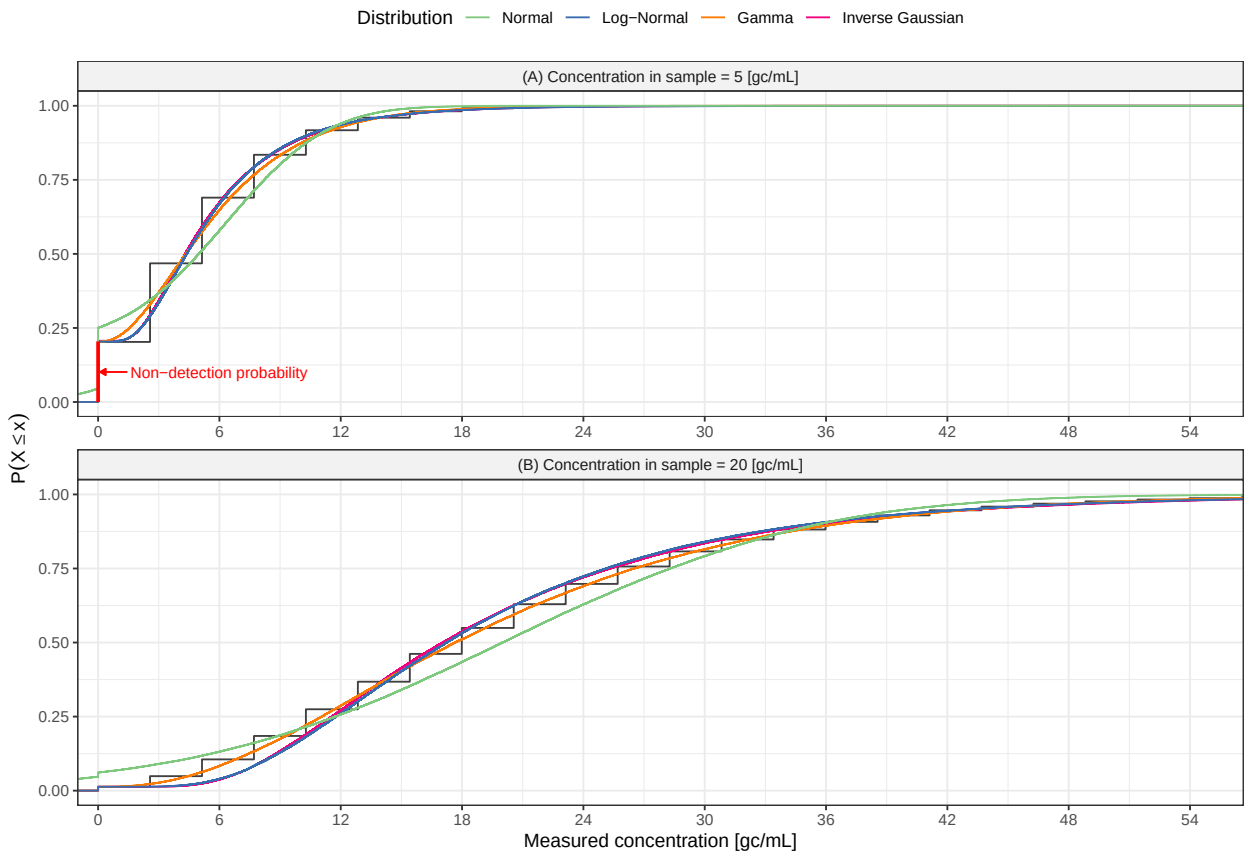

Fig S12. Continuous distributions approximating the distribution of dPCR measurements. Shown in black is the discrete cumulative distribution function of dPCR measurements obtained from the number of positive partitions in the assay for a concentration of (A) 5 gc/mL and (B) 20 gc/mL. This distribution 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 Estimation

The dPCR-specific likelihood was implemented as a function in the probabilistic programming language stan [9]. We integrated this likelihood function in a validation model for estimating a single

concentration from replicate measurements, ia generalized linear model (GLM) for estimating the association of covariates with the target concentration, and in the EpiSewer wastewater model for estimating reproduction numbers and other transmission parameters from wastewater measurements over time.

We fitted models via Markov chain Monte Carlo (MCMC) using cmdstan version 2.34.1 via cmdstanr version 0.8.1 [10]. 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) [11], and for convergence problems using the number of divergent transitions [12], the Bayesian fraction of missing information (E-BFMI < 0.2) [13] and the Gelman-Rubin convergence diagnostic ($\hat{R} > 1.01$) [14]. The checks indicated sufficient effective sampling sizes and good convergence and mixing of chains.

G.2 Validation

We validated inference via the dPCR-specific likelihood by fitting the model for a single concentration to simulated dPCR measurements with different true sample concentrations (1, 3, and 5 gc/mL). 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.

Figure S13 shows the true concentration and estimated concentrations using the dPCR-specific as well as the normal and log-normal likelihood function, together with the posterior predictive distributions for the dPCR measurements. When using a dPCR-specific likelihood, the probability of zero measurements is explicitly represented and the true concentration is estimated without bias. 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 but zero measurements must be dropped from the analysis. This leads to a significant bias in the estimated concentration when the true concentration is small.

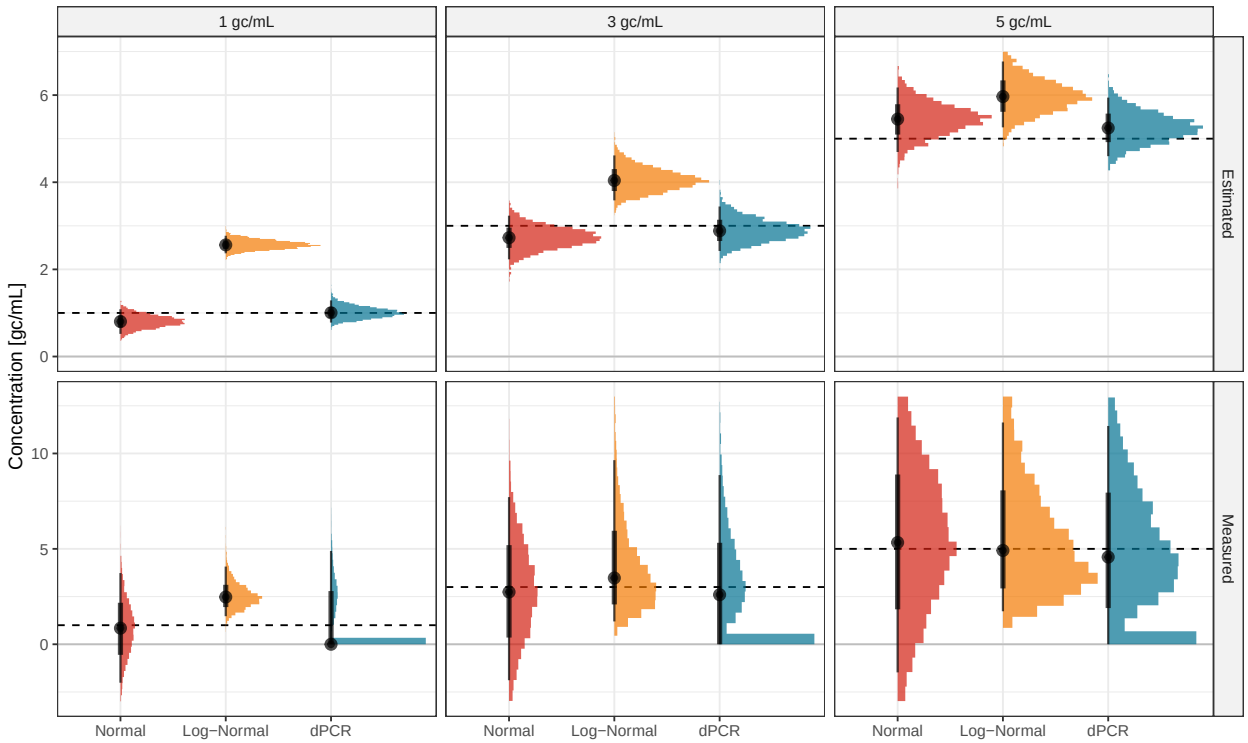

Fig S13. Comparison of estimated concentration and predicted dPCR measurements under different likelihood functions. Shown is the posterior distribution of the estimated true concentration (top) and the posterior predictive distribution of measurements (bottom) from a Bayesian model fitted with a normal (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.

H Application to eDNA-based biomonitoring

We modeled the expected concentration of free-eDNA of Bugula neritina t hours after removal of the organism as

$$\lambda_{t,i} = \exp(a + bt + \delta_i), \tag{37}$$

where $\exp(a)$ is the concentration at $t = 0$ (intercept), b the exponential decay factor, and $\exp(\delta_i)$ a multiplicative fixed effect for the sample location in the aquarium (back left, back right, front left, front right, and middle). The half-life period can be computed as $-\frac{\log(2)}{b}$. When analyzing data from all aquariums combined, we accounted for differences between aquariums using

$$\lambda_{t,i,j} = \exp(a + bt + \delta_i + \eta_j), \tag{38}$$

where $\exp(\eta_j)$ is a multiplicative effect for the aquarium of type j (low, medium, or high biomass).

We fitted the model with a normal, log-normal, and dPCR-specific likelihood respectively (see Supplement G.1 for estimation details). For the log-normal likelihood, we placed a half-normal prior with a standard deviation of 0.5 on the coefficient of variation. For the normal likelihood, we assumed a fixed variance, but parameterized it in terms of the average coefficient of variation to use the same noise prior as in the log-normal model. For the dPCR-specific likelihood, based on the lab protocol in Scriver et al. [15], we assumed a scaling factor of $s = \frac{1}{21}$ for the reaction mix and $m = 20\,000$ partitions with a volume of $v = 1\,\text{nL}$ for the BioRad200^{TM} droplet generator without droplet loss. Again, we used a half-normal prior with a standard deviation of 0.5 for the remaining pre-PCR variation (ν_{pre}).

To reproduce the original approach by Scriver et al., we simply fitted a normal likelihood to the averages of measurements at each point in time. As in Scriver et al., measurements above the maximum value of the negative controls in the experiment ($0.08\,\text{gc}/\mu\text{L}$) were removed before averaging.

I Application to wastewater-based epidemiology

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 described in Supplement E 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. For each treatment plant separately, we fitted the wastewater model implemented in the EpiSewer R package to the longitudinal measurements of Influenza A virus. In this model, the effective reproduction number was smoothed using a random walk prior, and infections were assumed to arise from a stochastic renewal process with Poisson distributed noise. The full model specification is provided as reproducible code at `https://github.com/adrian-lison/dPCR-observation-model-study`.

I.1 PCR parameters

I.1.1 Prior for number of partitions

To specify a prior for the number of partitions m in the assay, we took into account that m varies between PCR runs due to random loss of partitions. As can be seen in Figure S14, the number of lost partitions per PCR run can be approximately modeled as log-normally distributed. In principle, m could thus be modeled 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 wanted to parameterize m instead via hyperpriors for the mean number of partitions μ_m and coefficient of variation of partitions ν_m. We thus assumed that the maximum number of partitions corresponds to $\mu_m + 3\mu_m\nu_m$ and chose 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 used $\text{Normal}^+(\mu = 20000, \sigma = 5000)$ to accommodate dPCR systems with between 10000 and 30000 partitions. As a hyperprior for ν_m, we used $\text{Normal}^+(\mu = 0, \sigma = 0.05)$, reflecting our expectation of less than 10% variation in the number of partitions.

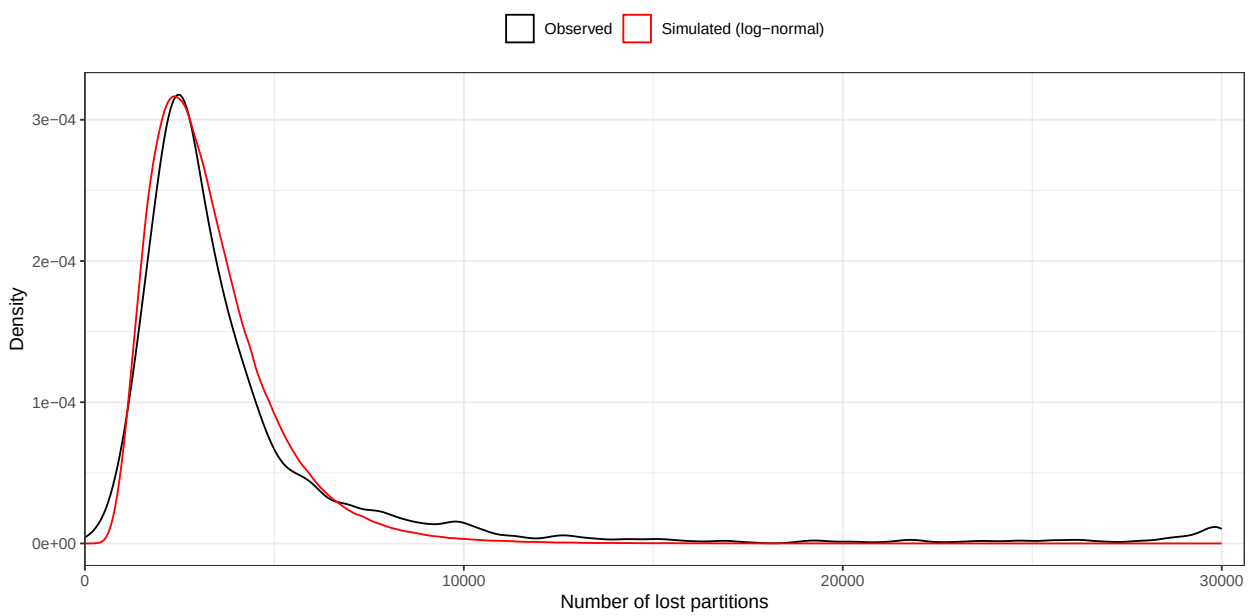

Fig S14. Distribution of lost partitions in a dPCR assay. Shown in black is a kernel density estimate for the distribution of the number of lost partitions in a dPCR assay based on 2733 runs on a Stilla® Sapphire chip with a maximum of 30000 partitions. Shown in red is a log-normal distribution matched to the observed density.

I.1.2 Prior for conversion factor

Since the scaling factor s and partition volume v are not individually unidentifiable, we specified a prior for the "conversion factor" factor $c = sv$. This factor expresses how many expected gene copies per partition a concentration of 1 gc/mL in the original sample corresponds to. Here, we used a broad prior for c for the case of digital droplet PCR based on a typical laboratory protocol for wastewater-based epidemiology. Specifically, we assumed that extraction volumes from wastewater samples are a few dozen mL (e. g. 40mL in the assay described in Supplement E). 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 E). 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 in the nanoliter range. 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 E, the conversion factor is 0.173×10^{-4}. Based on these assumptions, we used 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}.

I.1.3 Prior for pre-PCR coefficient of variation

As a prior for ν_{pre}, we used a standard regularizing prior, i. e. $\nu_{\text{pre}} \sim \text{Normal}^{+}(0,1)$.

I.1.4 Estimated parameters

Figure S15 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.

I.2 Forecasts

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

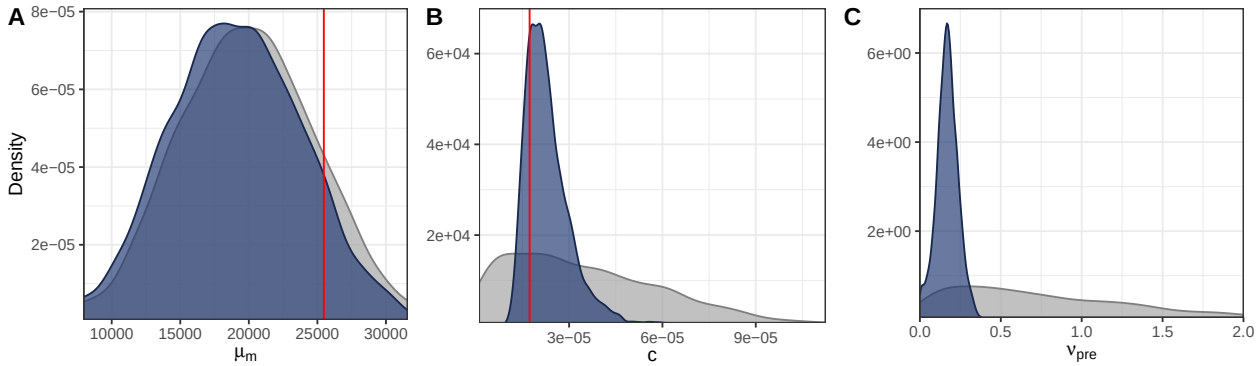

Fig S15. Inference of dPCR parameters and pre-PCR variation from a time series of dPCR measurements: Shown are prior and posterior distributions for (A) the average number of partitions μ_m, (B) the conversion factor c, and (C) the pre-PCR coefficient of variation ν_{pre} of a model applied to dPCR measurements 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.

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.

We assessed the forecasts of concentration measurements using the Continuous ranked probability score (CRPS) [16], i. e.

$$\text{CRPS}(F_{\hat{\Lambda}}, \hat{\lambda}) = \int_{-\infty}^{\infty} \left(F_{\hat{\Lambda}}(y) - \mathbb{1}\{y \geq \hat{\lambda}\} \right)^2 dy, \tag{39}$$

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. [17] implemented in the R package "scoringutils" [18]. We summarized the CRPS scores from different days using the arithmetic mean, denoted $\overline{\text{CRPS}}$.

I.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 S16–S20). We also show a comparison of estimates obtained using the dPCR-specific likelihood with and without exact knowledge of the dPCR parameters (Figure S21).

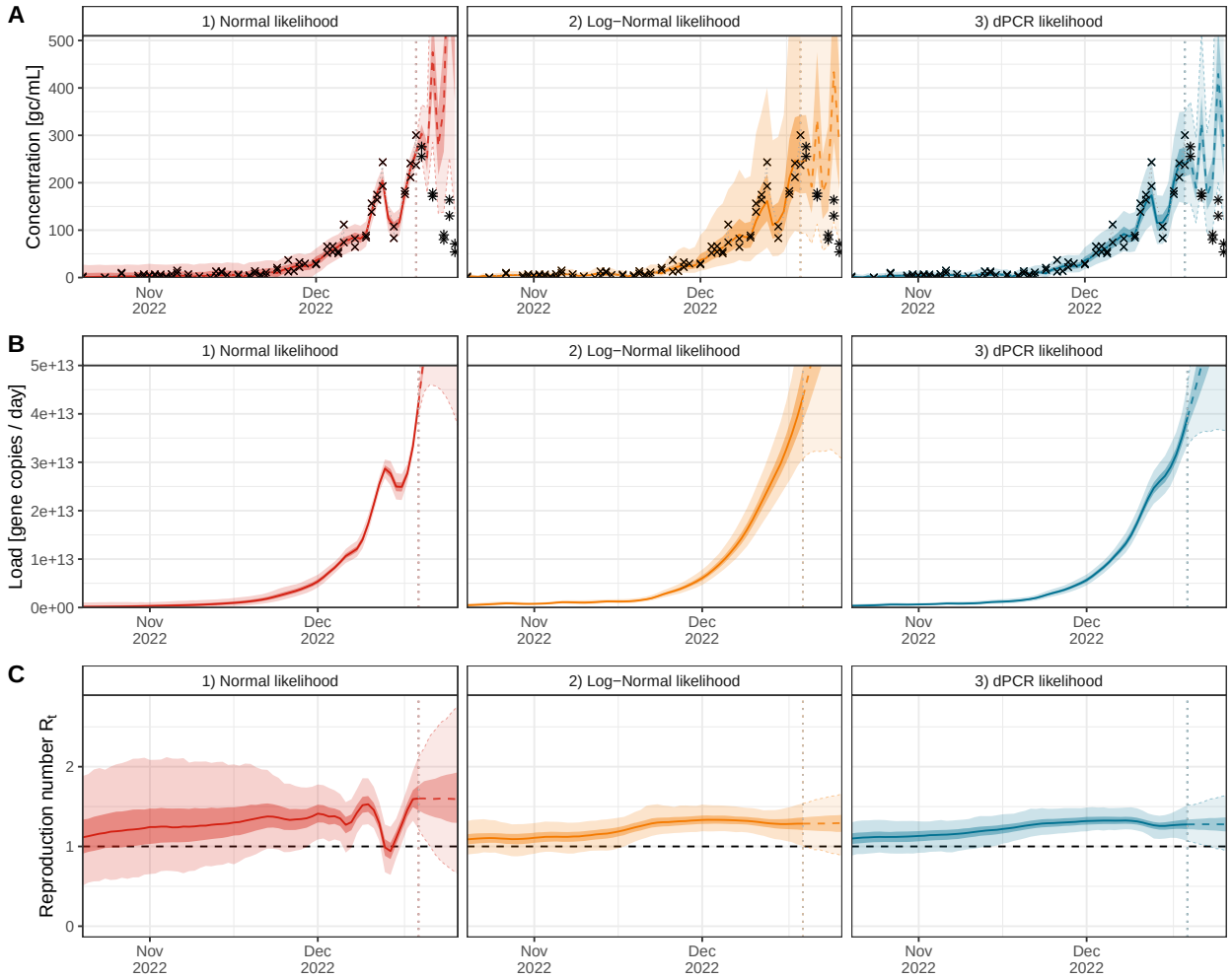

Fig S16. Comparison of likelihood functions on dPCR measurements of Influenza A virus concentrations at the municipal treatment plant of Zurich, Switzerland, Oct 20 – Dec 19, 2022: Shown are the results of fitting an epidemiological wastewater model (EpiSewer) to duplicate dPCR measurements of Influenza 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.

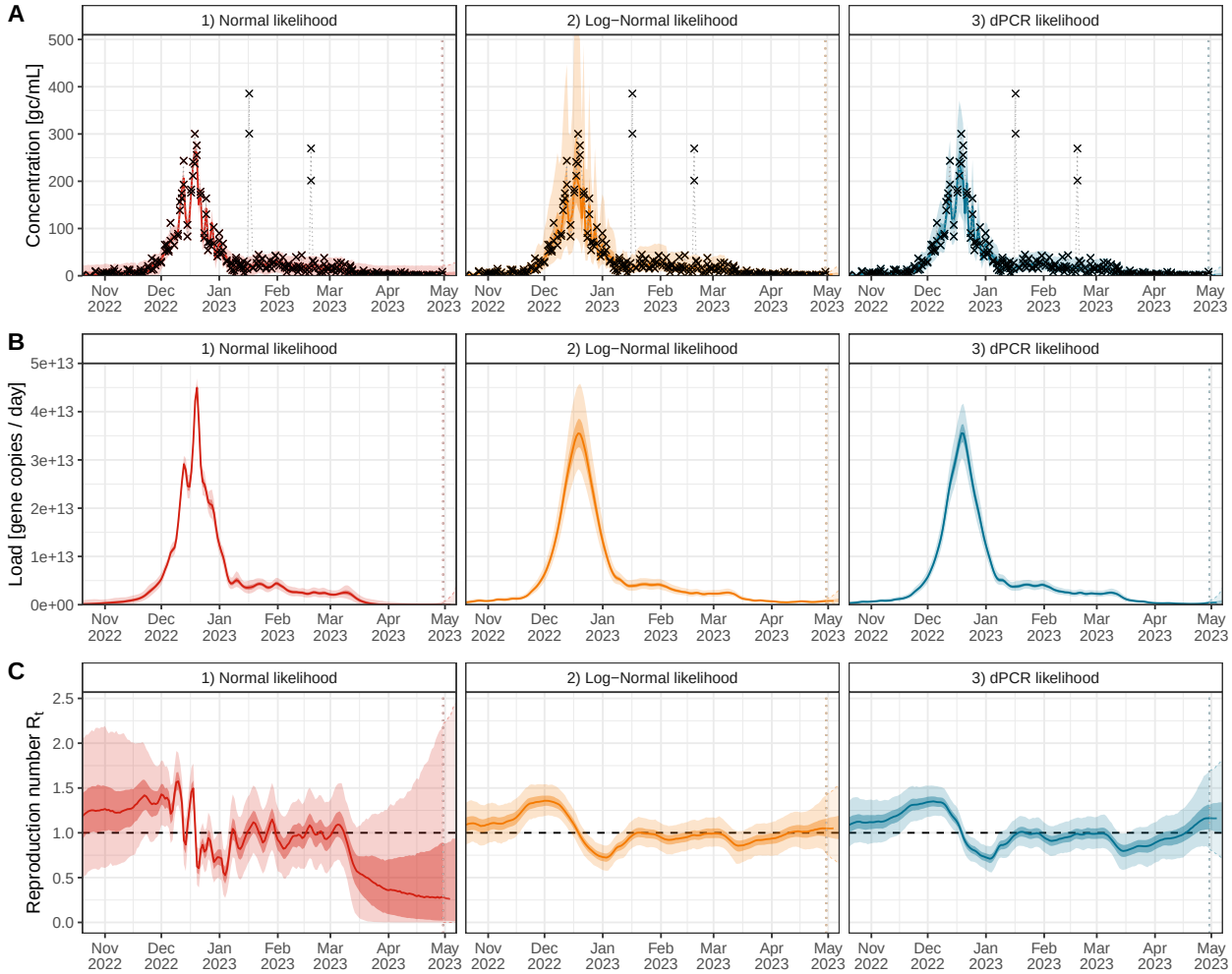

Fig S17. Comparison of likelihood functions on measurements of Influenza A virus concentrations at the municipal treatment plant of Zurich, Switzerland, Oct 20 – May 01, 2023: Shown are the results of fitting an epidemiological wastewater model (EpiSewer) to duplicate concentration measurements of Influenza 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.

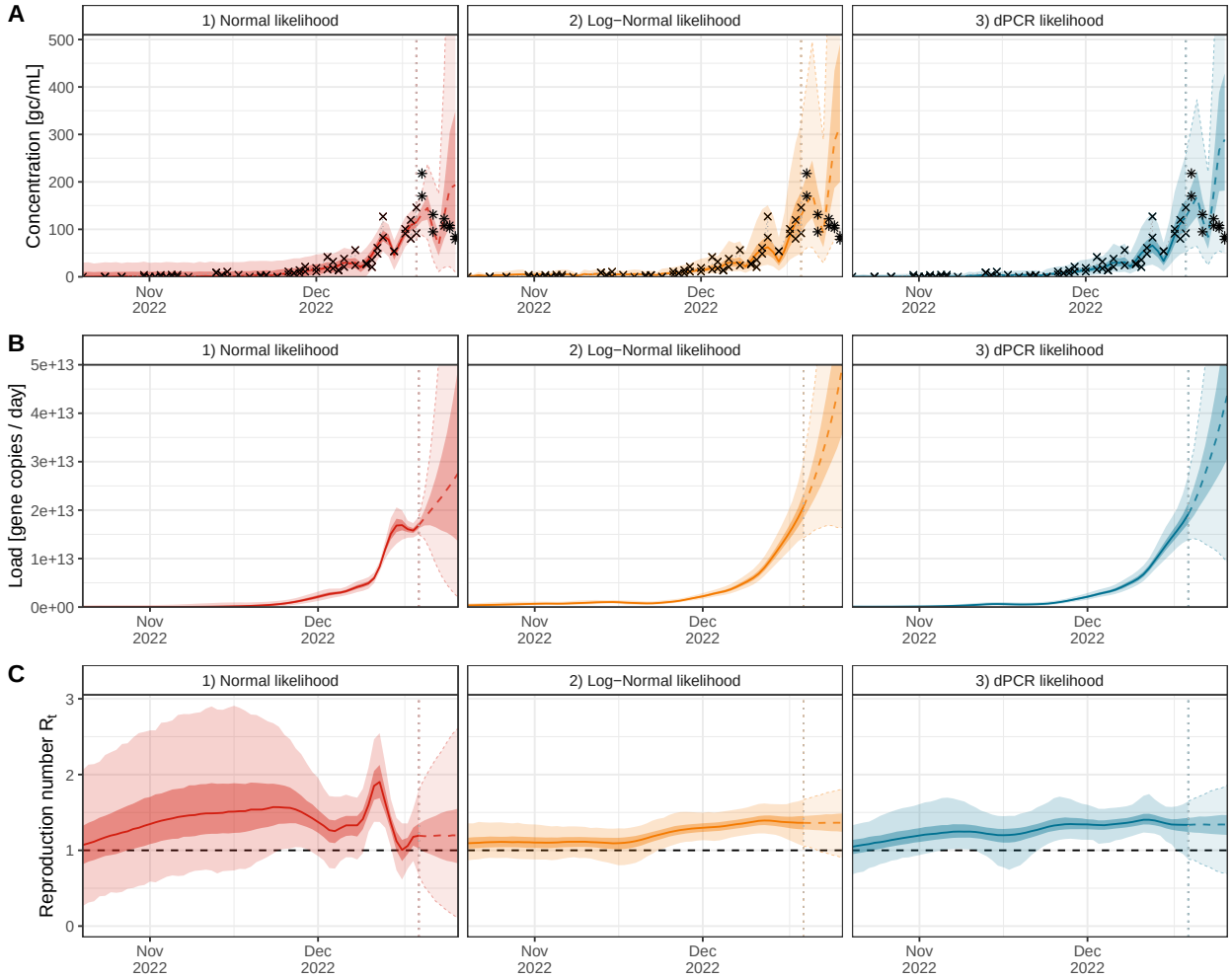

Fig S18. Comparison of likelihood functions on measurements of Influenza A virus concentrations at the municipal treatment plant of Geneva, Switzerland, Oct 20 – Dec 19, 2022: Shown are the results of fitting an epidemiological wastewater model (EpiSewer) to duplicate dPCR measurements of Influenza A 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.

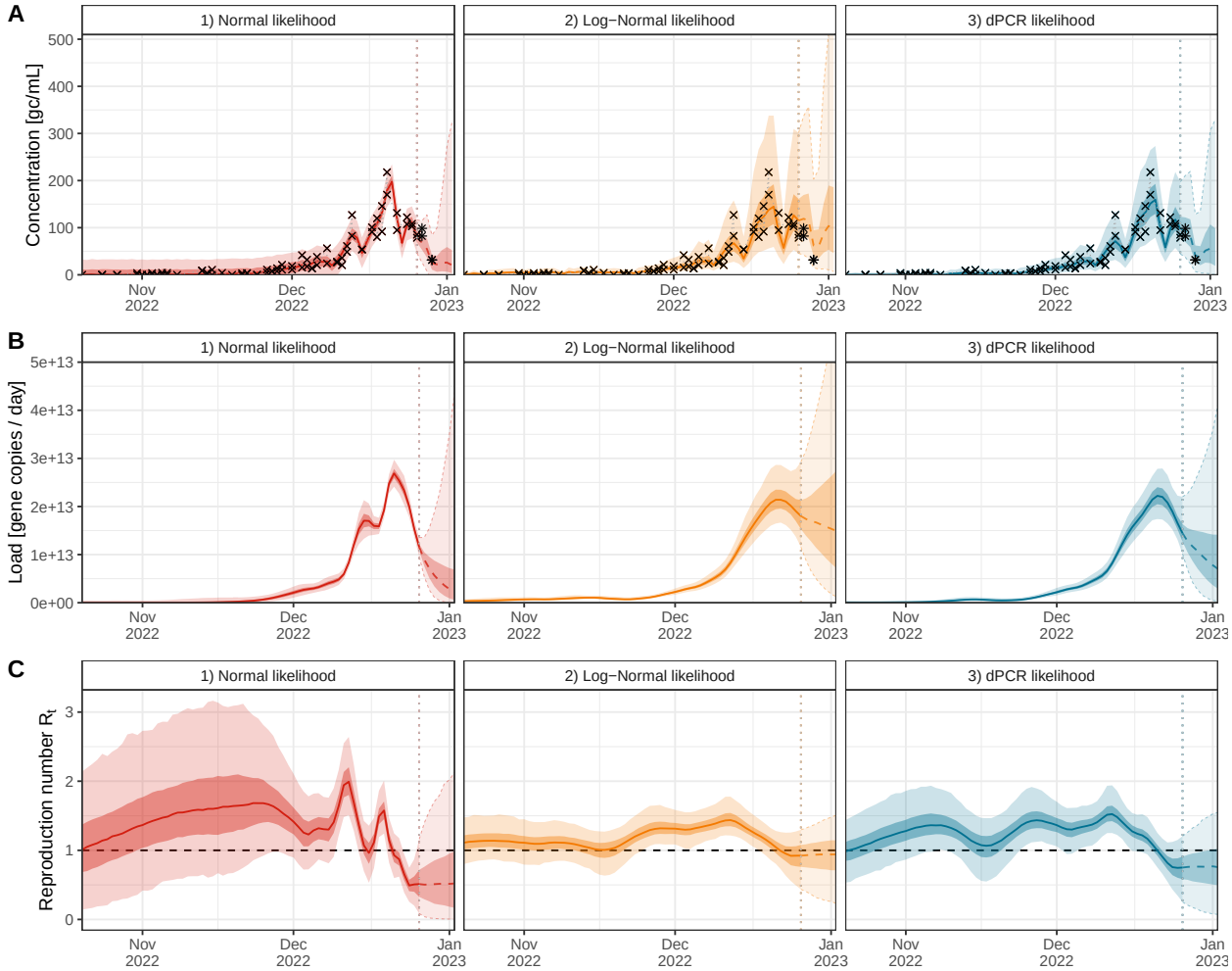

Fig S19. Comparison of likelihood functions on measurements of Influenza A virus concentrations at the municipal treatment plant of Geneva, Switzerland, Oct 20 – Dec 26, 2022: Shown are the results of fitting an epidemiological wastewater model (EpiSewer) to duplicate dPCR measurements of Influenza A 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.

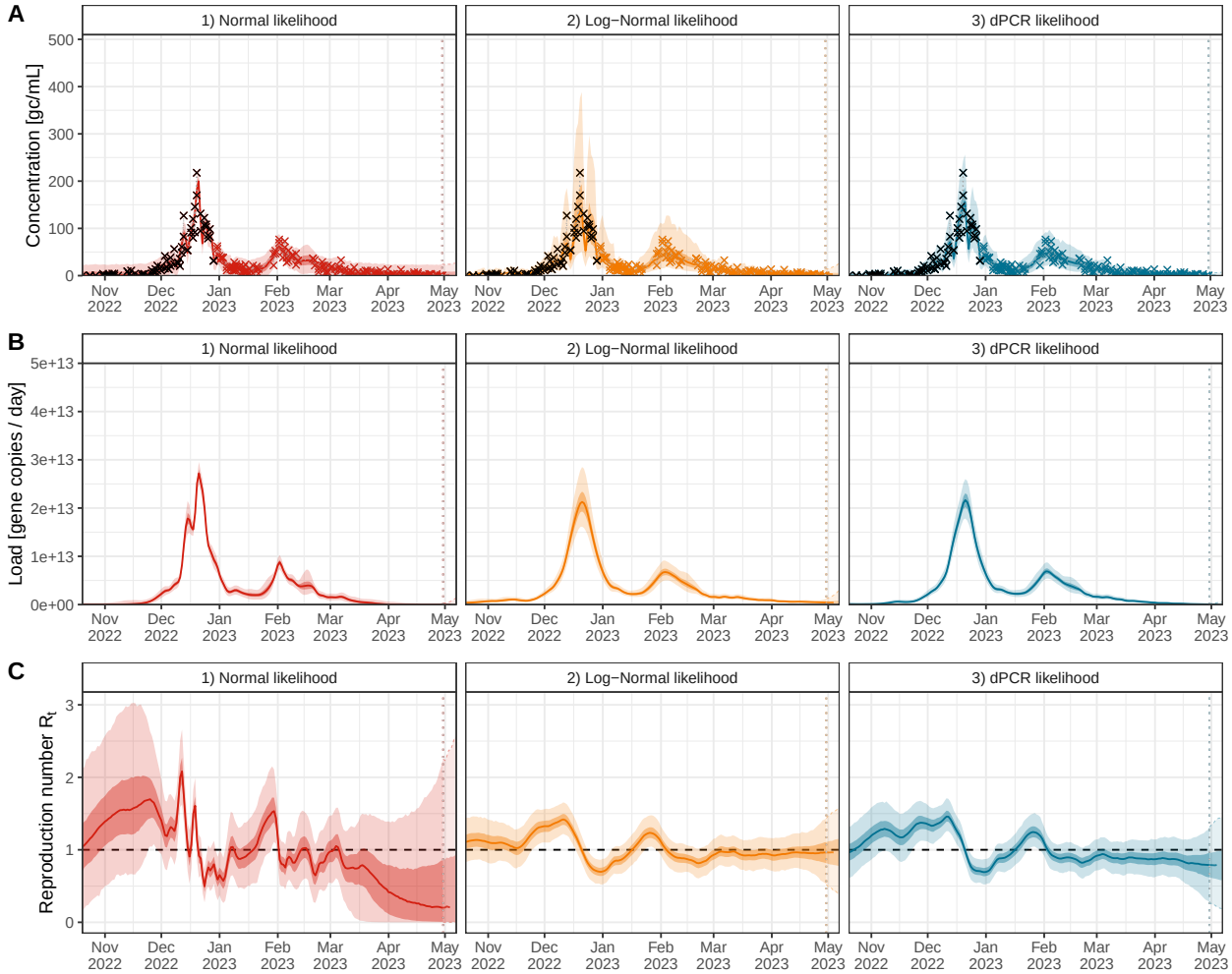

Fig S20. Comparison of likelihood functions on measurements of Influenza A virus concentrations at the municipal treatment plant of Geneva, Switzerland, Oct 20 – May 01, 2023: Shown are the results of fitting an epidemiological wastewater model (EpiSewer) to duplicate dPCR measurements of Influenza A 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.

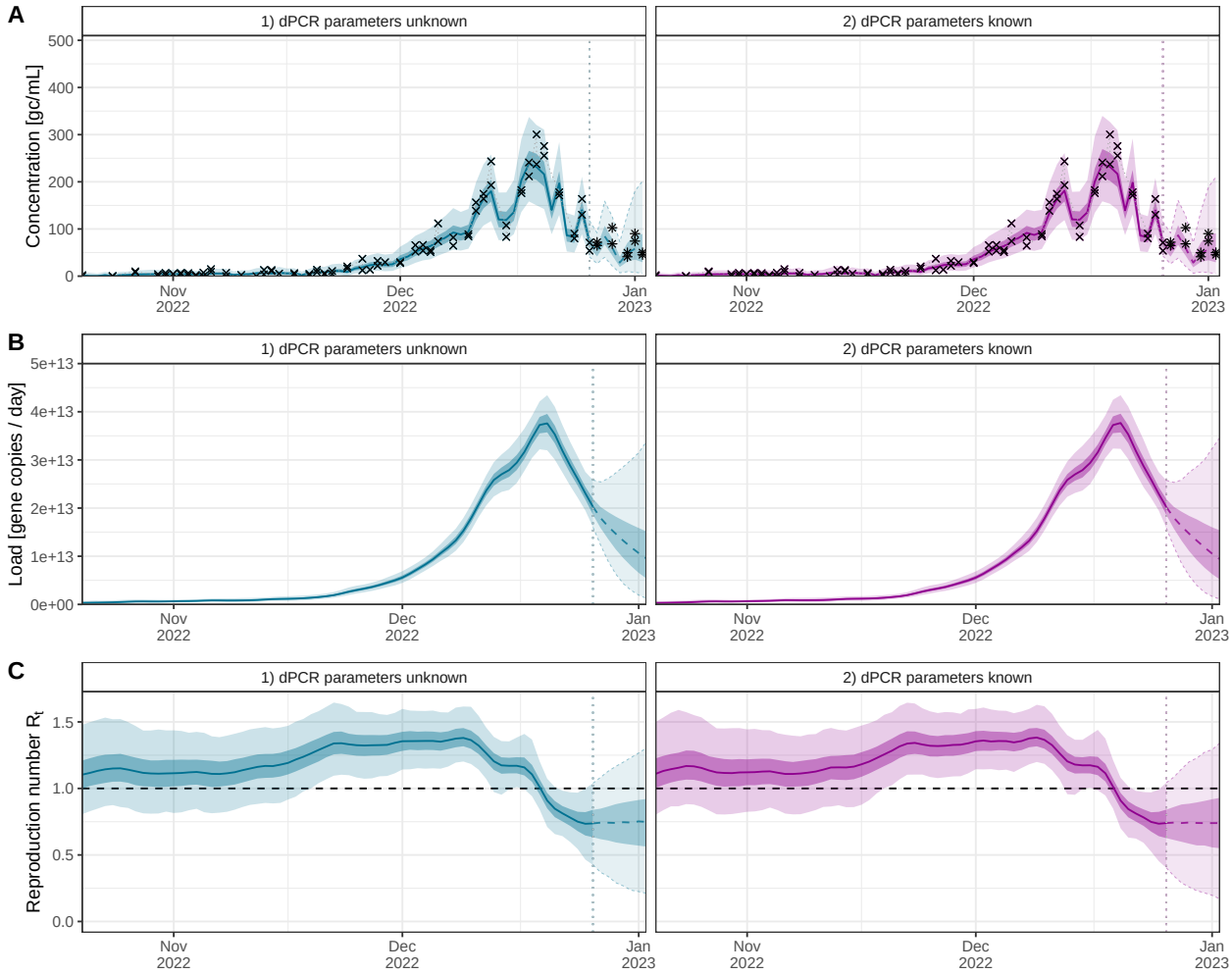

Fig S21. Comparison of inference and short-term forecasts from wastewater measurements using a dPCR-specific likelihood with and without knowledge of the dPCR parameters: Shown are the results of fitting an epidemiological wastewater model (EpiSewer) to duplicate dPCR measurements of Influenza A virus 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. 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.
2. 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.
3. 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.
4. Corless RM, Gonnet GH, Hare DEG, Jeffrey DJ, Knuth DE. On the LambertW Function. Advances in Computational Mathematics. 1996;5(1):329–359.
5. Armbruster DA, Pry T. Limit of Blank, Limit of Detection and Limit of Quantitation. The Clinical Biochemist Reviews. 2008;29(Suppl 1):S49–S52.
6. 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.
7. 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.
8. 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.
9. Stan development team. Stan Modeling Language Users Guide and Reference Manual, Version 2.35; 2024.
10. Gabry J, Češnovar R. Cmdstanr: R Interface to 'CmdStan' [Manual]; 2024.
11. 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.
12. Betancourt M. A Conceptual Introduction to Hamiltonian Monte Carlo. arXiv; 2018.

13. Betancourt M. Diagnosing Suboptimal Cotangent Disintegrations in Hamiltonian Monte Carlo. arXiv; 2016.

14. Gelman A, Rubin DB, et al. Inference from Iterative Simulation Using Multiple Sequences. Statistical Science. 1992;7(4):457–472.

15. Scriver M, von Ammon U, Youngbull C, Pochon X, Stanton JAL, Gemmell NJ, et al. Drop It All: Extraction-Free Detection of Targeted Marine Species through Optimized Direct Droplet Digital PCR. PeerJ. 2024;12:e16969.

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

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

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