Quantifying impacts of an environmental intervention using environmental DNA: Appendix S3

Elizabeth Andruszkiewicz Allan, Ryan P. Kelly, Erin D'Agnese,
Maya Garber-Yonts, Megan Shaffer, Zachary Gold, Andrew O. Shelton

2023

Supplemental material for "Quantifying impacts of an environmental intervention using environmental DNA" in *Ecological Applications*

Our analysis depends upon a set of quantitative models, each linking our observations of metabarcoding reads or qPCR cycle-threshold values to an underlying concentration of target-species DNA in water samples.

In summary, we (1) use a mock community with a known composition to calibrate our environmental metabarcoding data as described in (Shelton et al. 2022). The result is a set of estimated proportions of DNA from each species in each sample. We then (2) relate qPCR cycle-threshold values for a reference species (here, cutthroat trout (*O. clarkii*)) from the same set of samples to a standard curve to yield quantitative estimates of the concentration of our reference species in each sample. We (3) use these absolute estimates of DNA concentration to expand the metabarcoding-derived proportion data into a complete set of quantitative estimates of DNA concentrations for each species in each sample. We account for the variable water-flow-rates of the sampled creeks by converting these concentrations from units of copies/L into units of copies/s, given an flow rate in L/s. Finally, we (4) construct a model describing changes in these species-specific concentrations over time. We give the statistical details of these steps below.

Calibration with a Mock Community

See Shelton et al. (2022); McLaren et al. (2019); Silverman et al. (2021) for similar analyses.

For ease of computation, we ran the metabarcoding-calibration model on data for each of our five creeks separately, using the same mock communities to calibrate each.

Model Diagnostics: 3 chains, 1500 iterations, for all parameters, $\hat{R} \leq 1.02$

qPCR Calibration

See Shelton et al. (2019); McCall et al. (2014) for similar analyses.

For all samples i, on qPCR plates j, we either observe ($z_{i,j} = 0$ or do not observe $z_{i,j} = 1$) amplification; we omit the subscripts i and j from the following description except where necessary for clarity. We assume an intercept of zero.

We model the probability of detection $P(z = 1))$ as a linear function of concentration and slope parameter ϕ, ($P(z = 1) = \theta = c\phi$), with a logit transform to constrain the inferred probability to between 0 and 1.

For those samples that amplify ($z = 1$), we model the observed Ct value (y) as a linear function of our parameter of interest, the log-concentration of target-species DNA under analysis (c). We treat y as drawn from a normal distribution $y \sim N(\mu_{i,j}, \sigma_{i,j})$), where each triplicate sample on each qPCR plate has its own estimated mean and standard deviation. The means are estimated as a straightforward linear model, $\mu = \beta_{0,j} + \beta_{1,j}c$, but we allow the standard deviation to vary as a linear function of log-concentration so as to accurately capture decreasing precision with decreasing concentration: $\sigma = e^{\gamma_0 + \gamma_{1,j}c}$; we estimate these parameters as an exponent to constrain $\sigma > 0$.

Samples with known concentrations (i.e., standards) were fit jointly with unknown samples (i.e., environmental samples); because qPCR plate identity was shared among all environmental samples and standards within a plate, this has the effect of applying plate-specific slope and intercept values for the standard curve to each of the environmental samples on the plate (Figure S1).

We apply moderately informative priors that make use of background information in hand. For example, because qPCR standard curves of all kinds have slopes near -3, this slope becomes our background expectation as embodied in the prior on β_1, but the standard deviation of that prior leaves plenty of room for this background to be overwhelmed by the observed data. The same logic applies to the intercept of the standard curve, which in qPCR (for any given species) generally falls near 39 cycles, an expectation that we formalize by having β_0 drawn from a normal distribution with $\mu = 39$ and $\sigma = 3$.

Taken together with priors, the model is:

$$z_{i,j} \sim Bernoulli(\theta_{i,j})$$

$$\theta_{i,j} = logit^{-1}(\phi c_{i,j})$$

$$y_{i,j} \sim Normal(\mu_{i,j}, \sigma_{i,j}) \text{ if } z_{i,j} = 1$$

$$\mu_{i,j} = \beta_{0,j} + \beta_{1,j} c_{i,j}$$

$$\sigma_{i,j} = e^{\gamma_0 + \gamma_{1,j} c_{i,j}}$$

$$\beta_0 \sim normal(39, 3)$$

$$\beta_1 \sim normal(-3, 1)$$

$$\gamma_1 \sim normal(0, 5)$$

$$\gamma_0 \sim normal(-2, 1)$$

Model Diagnostics: 3 chains, 2500 iterations, for all parameters, $\hat{R} \leq 1.002$.

Expanding Proportions into Absolute Abundances

See Pont et al. (2022) and McLaren et al. (n.d.) (preprint) for examples of similar expansions.

As described in the main text, calibrated metabarcoding analysis yielded quantitative estimates of the proportions of species' DNA in environmental samples prior to PCR.

We then converted these proportions into absolute abundances by expansion, in light of the qPCR results for our reference species *O. clarkii*. We estimated the total amplifiable salmonid DNA in environmental sample i as $DNA_{salmonid_i} = \frac{[qPCR_{reference_i}]}{Proportion_{reference_i}}$, and then expanded species' proportions into absolute concentrations by multiplying these sample-specific total concentrations by individual species' proportions, such that for species j in sample i, $DNA_{i,j} = DNA_{salmonid_i} * Proportion_{i,j}$.

We transformed the resulting abundances to account for the creeks' flow-rates as described in the main text.

Ideally, we would have fit a joint model that simultaneously estimated species proportions (metabarcoding), absolute concentrations (qPCR), and developed the time-series trends for all species. As a practical computational matter, we had to create these models individually, which entailed some loss of information about parameter variability and cross correlation. For the mixed-effects model describing trends over time (described below), we used the product of posterior means from the metabarcoding and the concentrations

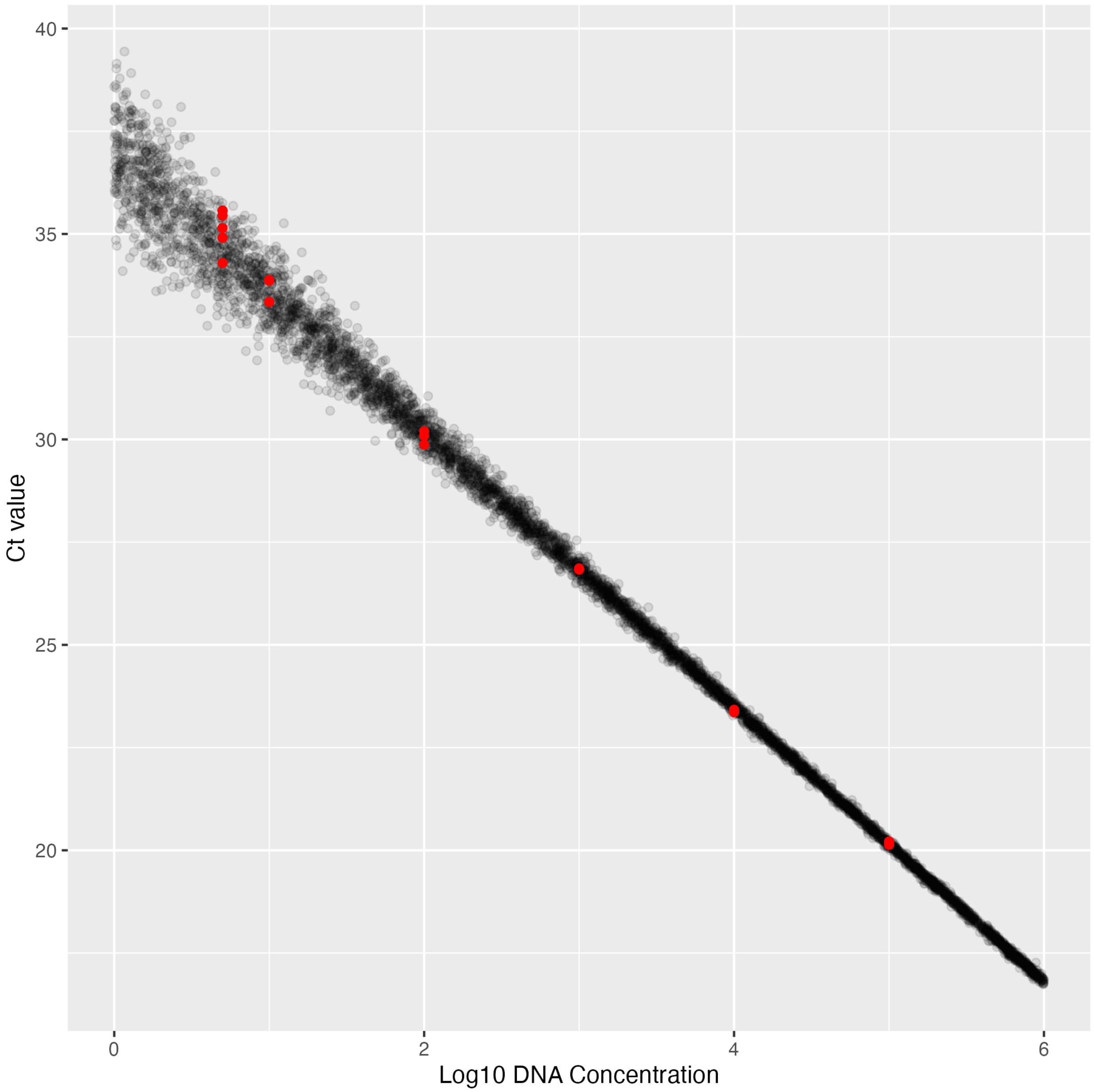

Figure S1. Example of 2500 samples from the joint posterior distribution of the model fit for a single representative qPCR plate. Red dots are standard-curve observations with known starting concentrations. The spread of black dots (posterior samples) indicates the shape of the calibration curve, with standard deviation increasing as concentration decreases.

of the qPCR model as observations, rather than being able to use the full posteriors for each input to the model. We deemed this acceptable because our metabarcoding proportions were quite precisely estimated: for example, in our focal Padden Creek, the coefficient of variation for estimated proportions of our reference species (*O. clarkii*) ranged from 0.008 to 0.25.

Modeling Changes in Concentration over Time

At a given station in a given creek, some DNA concentration exists for each species. For simplicity, we focus on a single species and a single station (downstream or upstream) for the moment.

Our observations of the (log) DNA concentration in creek i at time t are distributed as $Y_{i,t} \sim \mathcal{N}(\mu_{i,t}, \sigma^2)$. More complex versions of the model may let σ vary across creeks, time points, species, or with environmental covariates of interest.

We are interested in how the DNA concentration changes over time, so we model the expected value of DNA in a creek at time t, $\mu_{i,t}$.

We considered three ways of modeling the salmonid eDNA data, each in a Bayesian framework, but each treating non-independence among time points somewhat differently:

- A linear auto-regressive (AR(1)) model, written in `stan`. For each species in each creek, the expected concentration of eDNA of each month is a linear function of the expected value from the previous month. Within a species, the monthly autoregressive parameters are shared across creeks. For each species j – the subscript for which we omit here for clarity – we have a single overall model of the change in eDNA concentration among species, creeks (i), timepoints (t), and stations (d).

$$Y_{i,t,d} \sim \mathcal{N}(\mu_{i,t,d},\, \sigma^2_{\text{obs}})$$

$$\mu_{i,t,d} = \alpha_{i,t} + \epsilon_{i,t,d} + \eta_{i,t,d}$$

$$\epsilon_{i,t,d} \sim \mathcal{N}(\beta\mu_{i,t-1,d},\, \phi^2)$$

$$\alpha_{i,t} \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha)$$

$$\beta \sim \mathcal{U}(-1, 1)$$

$$\sigma_{\text{obs}} \sim gamma(1, 2)$$

$$\sigma_\alpha \sim gamma(1, 2)$$

$$\eta \sim \mathcal{N}(\mu_\eta, 1)$$

$$\phi \sim gamma(1, 2)$$

$$\mu_\alpha \sim \mathcal{N}(\mathbf{0}, \mathbf{10})$$

$$\mu_\alpha \sim \mathcal{N}(\mathbf{0}, \mathbf{5})$$

- A generalized additive model (GAM), written in `brms` (which itself writes a `stan` model). For each species in each creek, an independent set of spline (weighting) parameters describes the temporal trends in expected eDNA concentration; the number of spline knots is shared across species and creeks. We follow (Pedersen et al. 2019) to create a hierarchical GAM in which the expected value for each species in each creek at each time point is a spline function of time, time-by-creek, and time-by-station, with random effects for creek and station. Here, time-by-creek and time-by-station splines are centered, requiring additional fixed-effect terms for station and creek. Because no information is shared across species in this model, we fit the model each species independently.

$$\mu_{idt} = \beta_0 + s(t) + s_d(t) + s_i(t) + s(d) + s(i)$$

In `R` code using `brms`, this model is coded as

```
brm(
  bf(
    log(observed) ~
               s(time_idx, bs="cc") +
               s(time_idx, by=station, m=1, bs="cc")+  #main effect, station
```

```
                s(station, bs="re") +  #random effect, station
                s(time_idx, by=creek, m=1, bs="cc")+  #main effect, creek
                s(creek, bs="re") + #random effect, creek
  )
)
```

- A linear mixed-effects (LME) model, written in `rstanarm`. For each species in each creek, time (i.e., sampling month) is treated as a random effect. Each species-creek-month effect is treated as an independent draw from a common distribution.

$$\mu_{ijdt} = \beta_0 + \beta_{1_{ij}} + Month * \beta_{2_{ij}} + \beta_{3_{ijdt}}$$

In `R` code using `rstanarm`, this model is coded as

```
stan_glmer(log(observed) ~ (1 + time_idx|creek:species) + (1|station:creek:species:time_idx)
```

Ultimately, the three models yielded very similar results (Figure S2). The LME model proved simplest and most flexible insofar as it could easily handle datasets with uneven sets of observations – for example, cases in which a species was detected downstream of a barrier, but not upstream. We accordingly used the LME as the model for the analysis given in the main manuscript.

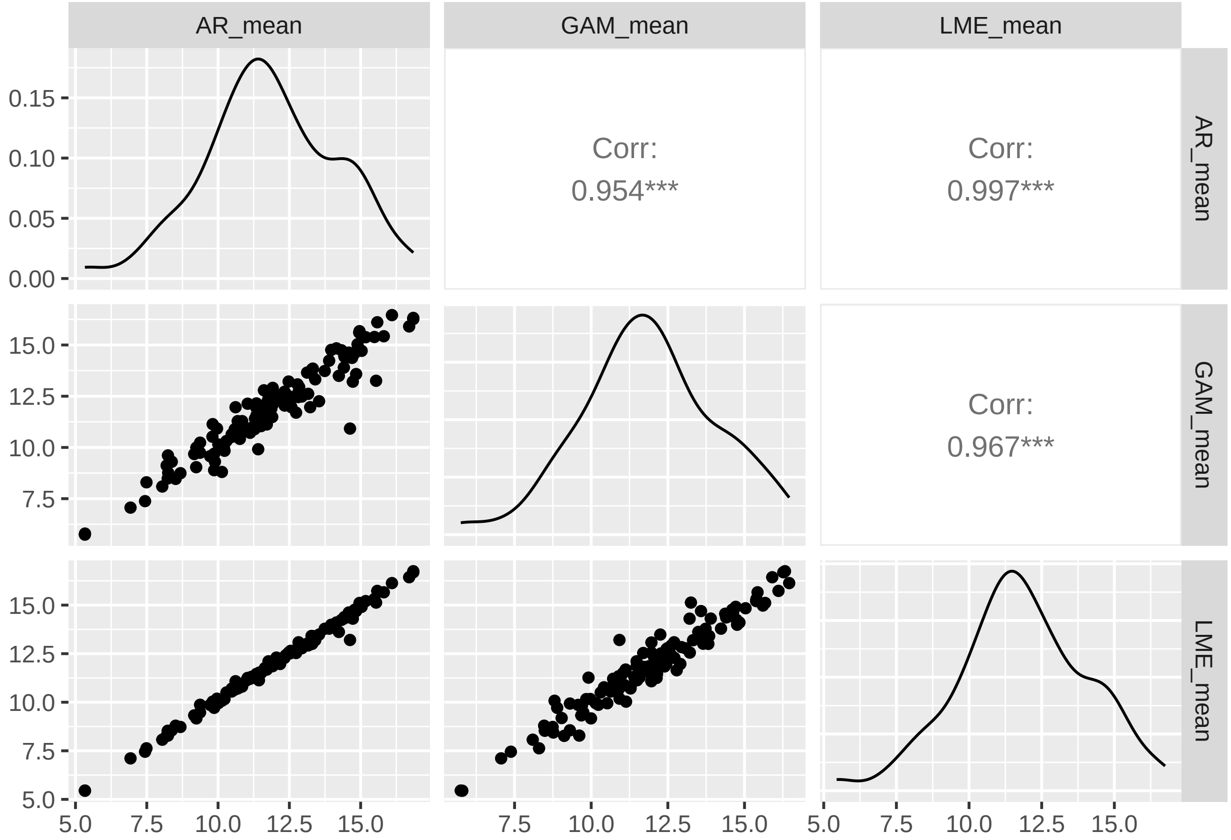

Figure S2. Comparison of three models (linear autoregressive, generalized additive model, and linear mixed effects model) shown for a subset of the data used in the main manuscript.

References

McCall, M. N., H. R. McMurray, H. Land, and A. Almudevar. 2014. On non-detects in qPCR data. Bioinformatics 30:2310–2316.

McLaren, M. R., J. T. Nearing, A. D. Willis, K. G. Lloyd, and B. J. Callahan. (n.d.). Implications of taxonomic bias for microbial differential-abundance analysis.

McLaren, M. R., A. D. Willis, and B. J. Callahan. 2019. Consistent and correctable bias in metagenomic sequencing experiments. eLife 8:e46923.

Pedersen, E. J., D. L. Miller, G. L. Simpson, and N. Ross. 2019. Hierarchical generalized additive models in ecology: an introduction with mgcv. PeerJ 7:e6876.

Pont, D., P. Meulenbroek, V. Bammer, T. Dejean, T. Erős, P. Jean, M. Lenhardt, C. Nagel, L. Pekarik, M. Schabuss, B. C. Stoeckle, E. Stoica, H. Zornig, A. Weigand, and A. Valentini. 2022. Quantitative monitoring of diverse fish communities on a large scale combining eDNA metabarcoding and qPCR. Molecular Ecology Resources n/a.

Shelton, A. O., Z. J. Gold, A. J. Jensen, E. D'Agnese, E. Andruszkiewicz Allan, A. Van Cise, R. Gallego, A. Ramón-Laca, M. Garber-Yonts, K. Parsons, and R. P. Kelly. 2022. Toward quantitative metabarcoding. Ecology n/a:e3906.

Shelton, A. O., R. P. Kelly, J. L. O'Donnell, L. Park, P. Schwenke, C. Greene, R. A. Henderson, and E. M. Beamer. 2019. Environmental DNA provides quantitative estimates of a threatened salmon species. Biological Conservation 237:383–391.

Silverman, J. D., R. J. Bloom, S. Jiang, H. K. Durand, E. Dallow, S. Mukherjee, and L. A. David. 2021. Measuring and mitigating PCR bias in microbiota datasets. PLoS Computational Biology 17:e1009113.