A Method for Tryptophan-Like Fluorescence as an Alternative Fecal-Indicator in Natural Waters

Evan Thomas[1,2], Matthew R. V. Ross[2,3], Michael J. Vlah[2], and Whitney Knopp[1,2]

[1]Mortenson Center in Global Engineering and Resilience, University of Colorado Boulder, Boulder, CO 80309, United States
[2]Virridy, Inc., Boulder, CO 80301, United States
[3]Department of Ecosystem Science and Sustainability, Colorado State University, Fort Collins, CO 80523, United States

Draft v0.5
target: *Environmental Science & Technology*, Research Article

Abstract

Recreational water criteria are written on culturable fecal indicator bacteria (FIB), which return a result in 18–24 h from samples collected on fixed schedules. Tryptophan-like fluorescence (TLF) measures a fecal-associated substance in seconds. We specify TLF as a fecal-indicator method and validate it against the acceptance criteria of EPA's pathway for site-specific alternative recreational indicators and methods, using culture references that are themselves imprecise and non-unique. Such a method comprises an instrument and the analytics converting its readings to a reported concentration, evaluated together. On paired grabs the method reached an index of agreement (IA) of 0.96 against IDEXX Colilert ($n = 209$) and 0.91 against membrane filtration ($n = 206$); the two approved methods reached 0.79 with each other. Deployed as a continuous record and scored against grabs, IA was 0.79 at six stations on one stream ($n = 145$) and 0.74 at four sites on a second river system ($n = 98$). Propagating the reference confidence interval moved balanced accuracy by -0.012 and placed a reference-tolerant ceiling at 0.963. TLF can replace a culture reference where a state adopts it site-specifically, and augments one wherever a program needs temporal coverage rather than regulatory standing.

Keywords: tryptophan-like fluorescence; alternative fecal indicator; method validation; recreational water quality criteria; index of agreement; reference-method uncertainty; continuous monitoring

Synopsis. A continuous optical measurement of a fecal-associated substance can carry a recreational water quality criterion where a state adopts it, giving coverage across storms and seasons that a fixed grab schedule cannot reach.

Figure 1: *

TOC / abstract graphic. Photograph by the authors, Bow River at Banff, Alberta, Canada, June 2026. Supplied at 1950×1050 px, which is 3.25×1.75 in at 600 dpi, the ACS proportions. Submitted as a separate file rather than embedded.

1 Introduction

Fecal contamination is a leading cause of waterborne gastrointestinal illness, and recreational water criteria are set on culturable FIB. The 2012 U.S. EPA Recreational Water Quality Criteria (RWQC) recommend a culturable *Escherichia coli* geometric-mean criterion of 126 CFU/100 mL for fresh waters, and jurisdictions apply a range of single-sample, statistical-threshold and beach-action values around it[1]. Culture takes 18–24 h. A program built on it also samples on dates fixed weeks ahead, which may or may not fall on the high flows and storm recessions that carry contamination, and reports what the water was doing several days earlier.

TLF is the intrinsic fluorescence of tryptophan and tryptophan-like moieties in microbial cells and their metabolic byproducts, excited near 280 nm and emitting near 350 nm. It measures a different quantity from a culture assay: a fecal-associated substance rather than the organisms themselves.

A relationship between that substance and fecal indicator organisms has been reported repeatedly and independently. Sorensen et al. established in-situ TLF as a real-time indicator of faecal contamination in Zambian drinking-water supplies[2]; Nowicki et al. found it a useful measure of microbial contamination risk in Kenyan groundwater[3]; and Sorensen et al. later reported that in-situ fluorescence tracked faecal contamination risk in Ugandan drinking water more rapidly and more resiliently than the indicator organisms themselves[4]. Khamis et al. characterized the temperature and turbidity dependence that any deployed tryptophan-like fluorometer must correct for[5], and Bedell et al. coupled a continuous in-situ fluorescence sensor to a statistical model for fecal contamination risk in drinking water[6]. Those studies span groundwater, piped supply and surface water, several instruments, and several reference assays. What they establish is that TLF carries fecal-contamination information across matrices and hardware. What they leave open, and what this paper addresses, is the specification and acceptance criteria under which such a measurement can be adopted in place of a culture reference for a regulatory purpose.

That difference has a regulatory category. A site-specific alternative recreational criterion may rest on a different indicator organism or on a different indicator *substance*, the examples given being caffeine and detergent brighteners, both of them measured by fluorescence[7]. An alternative of that kind is judged on the consistency and predictability of its relationship to an approved indicator method rather than on organism confirmation.

Methods of this kind are usually validated against a single culture result treated as truth, and that convention understates them. Most-probable-number and colony counts carry large, well-characterized analytical confidence intervals, so near a regulatory threshold the reference cannot always assign a sample to a category. EPA also lists several approved culture methods for the same analyte, and they disagree with one another, which makes the choice of reference partly arbitrary. The EPA microbiological method-comparison framework already treats each method as carrying its own false-positive and false-negative rate[8], and the RWQC alternative-methods pathway uses a continuous, threshold-independent index of agreement in place of a threshold-specific confusion matrix[7]. Published sensor validations rarely follow either convention.

This paper specifies what a TLF method must fix to be reproducible, evaluates it against references whose own uncertainty is propagated into the estimate, and separates the regulatory uses a TLF record can carry from those it cannot. The acceptance criteria it is judged against are not devised here. They are those of EPA's *Site-Specific Alternative Recreational Criteria Technical Support Materials for Alternative Indicators and Methods*[7], the federal pathway by which a state adopts a different indicator or a different measurement method for a named waterbody, and the validation performed here is the validation that document asks for: its paired-sample minimum, its choice of reference, its agreement statistic and its threshold. Where a criterion in this paper carries a number, that number is the pathway's. A commercial sensor and its paired record serve as a worked example implementation. That instrument's design, calibration and validation, in natural waters against Colilert and in drinking waters against the compartment bag test, are reported separately[9,10]. Those studies characterize the device: its optical sensitivity, its response to tryptophan, turbidity and temperature, and the accuracy of its concentration estimates, reported as error and classification statistics. Neither asks what a TLF method must specify to be reproducible by another party, what criteria it should be accepted against, or which regulatory uses its record can carry. Those are the questions here, and they are answered with a different statistic: the threshold-independent index of agreement that the alternative-indicator pathway uses for adoption, in place of the error and accuracy measures appropriate to characterizing an instrument. A companion method document in EPA method-report format carries the reagents, quality-control acceptance criteria and sampling detail that a journal article cannot.

2 Materials and Methods

2.1 What a TLF method fixes

A TLF method is an instrument together with the analytics that turn its readings into a reported value. The combination is what is specified, evaluated and adopted, and it is evaluated as one object.

Five things are fixed. *The measurand*: emission near 350 nm under excitation near 280 nm, referenced to a clean-water baseline, reported as a continuous fecal-contamination index. *The corrections*: fluorescence is quenched by temperature and attenuated by turbidity[5], so readings become comparable across conditions only after normalization to a common reference temperature and a turbidity correction derived independently of the fluorescence signal. *The per-unit calibration*: sensitivity differs between physical instruments, so gain, temperature coefficient and clean-water baseline are characterized per unit against an L-tryptophan standard series and are not shared. *The quality control*: initial and continuing verification against a stable secondary fluorophore, a calibration blank and baseline-drift check, a fouling check on the independent optical channel, and a spike-recovery check for matrix interference. *The analytics*: the model form, every input it carries, how its coefficients are obtained, and when they are refitted.

2.2 Modes of operation

The method is operated two ways, and each figure below is labeled by the mode that produced it. In *grab-and-read* mode the instrument reads a captured sample under its own per-unit calibration, and the estimate comes from a regression on the instrument's own channels. In *continuous in-situ* mode the instrument is deployed in the stream and reports on a fixed interval; a fitted model combines the fluorescence signal with the independent fouling channel, water temperature and a temperature interaction, a site-level term fitted to that water's prior FIB results, and any declared exogenous covariates. Coefficients are fitted on the paired record and refitted as paired data accumulate. The in-situ arms scored below carry no exogenous covariate. The two records come from different operations and different models, and are not pooled.

2.3 Two outputs: an estimate, and a sampling trigger

A continuous method produces a concentration estimate at every reporting interval, and that is the output the acceptance criteria are written against. It also produces a second output that a grab program cannot: a statement of when the water is worth sampling. The two are specified together, because the trigger is what connects the method to the laboratory it is meant to work alongside rather than replace.

A triggered design collects a laboratory sample when the record gives a reason for one. The reasons are stated in advance and drawn from the record itself: an estimate crossing the action

limit, a storm and its falling limb, an unexplained step in the signal, a divergence between stations on the same reach, and the paired samples the method needs in order to hold its own calibration. That last category is not optional. The refit obligation of Section 3.3 means a deployment has to keep generating paired samples whether or not anything interesting happens, so a method of this kind never removes the laboratory from the program; it changes which analyses get run.

2.4 Model inputs

Two kinds of input are distinguished so that a submission discloses what is in a reported value. The acceptance test applies to the method as a whole.

Corrections act on the instrument's own reading to make it comparable across conditions: the temperature normalization and turbidity correction above. They belong to the measurement.

Exogenous covariates are independent indicators tiered alongside the measurement. Antecedent rainfall is the main example, driving wet weather to combined-sewer discharge to fecal loading, and its value is strongly site-dependent. Across the deployments behind this work, the Euclid Creek deployment in Cleveland carries a rainfall-only model over a 24 h window capped at 5 mm; in the Chicago waterways a 48 h rain term sits alongside the optical channels and carries more of the prediction than any other single input (pooled $r \approx 0.54$, against ≈ 0.1 for the fouling channel); and four further deployments carry no rainfall term at all, on Boulder Creek, the Anacostia in Washington DC, in Denver, and on the Seine and Marne in Paris. On Boulder Creek, basin streamflow does most of what any hydrological term can do and roughly halves false positives, and a directed upstream-neighbor term at a travel-time lag adds a marginal gain, moving day-blocked cross-validated balanced accuracy from 0.64 to about 0.73 at 126 CFU/100 mL. Antecedent rainfall adds nothing there, the gauge having already integrated the rain upstream. Those figures rest on 56 paired grabs carrying 11 exceedances, with about ten variants scored against one cross-validation, and the same terms overfit the continuous magnitude target. They establish that hydrology carries information at that site without fixing a coefficient.

A party seeking a site adoption states which inputs its model carries before running the agreement analysis.

2.5 Eligibility and exclusion

A limit of quantification (LOQ) of 10 CFU/100 mL is applied as the bottom of a validated working range, and non-detects are excluded, a zero having no logarithm and so no place in a $\log_{10}$ comparison.

Beyond that the regime has three tiers, and only the first is mechanical. *Attribution* is applied automatically: a reading counts only if it is a measurement of the water being characterized, so readings taken before deployment, after removal, or belonging to another location are removed without judgement. Establishing attribution is an operational matter and not a property of the

method; the worked example keeps a deployment record per instrument per site. *Recoverable effects are corrected, not excluded*: temperature quench, optical fouling, the step following a physical cleaning, and any instrument state change for which a correction has been characterized are corrected, and the reading is kept. *Everything else is a per-case judgement* against stated criteria, with the reasoning recorded and the count reported either way. The default is to keep.

The criteria are these. Is the effect already handled by an existing correction? Was the instrument measuring the intended water at all, a reading failing that test falling to the install window rather than to a judgement? Is the channel physically invalid, meaning outside the range over which it was characterized or carrying a diagnostic fault, rather than merely reading high? Did the flag come from an operator, or from an advisory auto-detect heuristic? Is the flagged window bounded, or open-ended and therefore a live status rather than a considered boundary? Is there corroboration from the control group, meaning other sensors on the same water over the same period, a step hitting only the units someone touched being instrumental where one hitting all of them is hydrological? And how much does the exclusion remove, in which direction? An exclusion removing more than about 10% of a calibration set, or one that improves the headline metric, is raised explicitly rather than taken as a default, and the count is reported either way.

No reading is dropped for disagreeing with the reference. That rule has teeth: the largest single disagreement in the second in-situ record, a 1935 CFU/100 mL grab against a prediction two orders of magnitude below it, is retained even though removing it would raise balanced accuracy on that arm from 0.81 to 0.91.

The requirement each tier places on an implementation is that the condition be decidable from a diagnostic independent of the fluorescence signal, so that no reading is removed on the strength of the value in question. Which diagnostic serves is a matter of instrument design: the worked example reads submersion and fouling from an optical backscatter channel, ambient light from a dark channel subtracted from every reading, and the validity of the fluorescence estimate from the completeness of its excitation sweep. Another instrument might establish the same conditions by other means, and the criteria above would be unchanged.

The distinction between an advisory flag and a considered exclusion is worth the emphasis it gets above, because the worked example got it wrong once. Enforcing that instrument's own suspect-window flags as exclusions removed 77 grabs, 30% of the calibration set, since 51 of 55 maintenance records carry such a flag and fouling is already corrected. The flags are now surfaced for inspection and not enforced.

2.6 Paired records

The evidence is the paired candidate/reference record from one implementation: environmental grabs paired with IDEXX Colilert Quanti-Tray/2000 and, where available, membrane filtration (MF), spanning laboratory dilution series and field deployments. The wider record behind the

companion studies covers 661 samples across 12 water bodies. The lead freshwater matrix, and the source of the first continuous in-situ arm, is Boulder Creek, Colorado, a reach listed under Clean Water Act section 303(d) for *E. coli*. A second in-situ record comes from an independent deployment on the Seine and Marne, Paris.

2.7 Thresholds

The *action limit* belongs to the jurisdiction and varies: 126 CFU/100 mL under the 2012 RWQC for fresh recreational water, 200, 235 and 410 CFU/100 mL at three other deployments behind this work, 900 CFU/100 mL for European bathing water. The *decision cut* is derived per deployment for whichever limit applies, usually as a protective buffer below it. Comparison is made at the decision cut and labeled against the action limit. In the worked example the action limit is 126 CFU/100 mL and the derived cut is 82 CFU/100 mL. The acceptance test is threshold-independent, so one validated method screens against any of these limits without revalidation.

2.8 Evaluation

(i) Reference confidence intervals. Each Colilert result carries the manufacturer's tabulated Quanti-Tray/2000 95% confidence limits[11]; MF counts carry Poisson limits. The tabulated limits are read as a $\log_{10}$-normal with

$$\sigma = \frac{\log_{10}(\text{upper}) - \log_{10}(\text{lower})}{2 \times 1.959964},$$

so that for a sample reading x against an applicable limit L the probability that the reference's true value exceeds the limit is $P(\text{true} \geq L) = 1 - \Phi\big((\log_{10} L - \log_{10} x)/\sigma\big)$, and hence the probability that the sample's true category differs from its point label.

(ii) Expected performance. Classification performance is reported against the reference point label and as an expectation over the reference uncertainty, each sample contributing its probability p_i in place of a hard 0/1, so that expected sensitivity is $\sum p_i \hat{y}_i / \sum p_i$ and expected specificity is $\sum (1-p_i)(1-\hat{y}_i) / \sum (1-p_i)$. A reference-tolerant ceiling credits a prediction whenever the reference interval cannot distinguish it from correct, and is reported as an upper bound.

(iii) Method non-uniqueness. Agreement between Colilert and MF is quantified by regression and by ISO 17994 mean relative difference[12]. The EPA comparability structure would also allow error rates to be scored against an independent standard of known concentration; no such standard exists for an environmental record, and the alternative-indicator pathway analyzes unspiked samples.

(iv) Index of agreement. The agreement demonstration follows EPA's *Site-Specific Alternative Recreational Criteria Technical Support Materials for Alternative Indicators and Methods* (EPA-820-R-14-011) and its companion Alternative Methods Calculator[7], which together define the

statistic, the threshold and the eligible comparison. Willmott's index of agreement (IA)[13] and R^2 are computed on $\log_{10}$-transformed pairs by the Calculator's convention. A candidate is compared against a designated EPA method (Method 1603, 1600, 1611, or an approved equivalent) on at least 30 paired environmental samples within the quantification limits of both assays; IA ≥ 0.70 permits the alternative to carry the unchanged numerical criterion, with $R^2 > 0.60$ as a secondary route to regression-adjusted criteria. Adoption is site-specific, with EPA Regional approval. The values here reproduce the Calculator's arithmetic, Willmott's d on $\log_{10}$ pairs; running the workbook itself belongs to an adopting party.

2.9 Acceptance criteria

Table 1 states the criteria. The rows marked RWQC are requirements of EPA-820-R-14-011 and carry its numbers: the paired-sample minimum, the eligible reference, the agreement threshold, the validation tier and the scope of adoption. The remainder are properties an implementation fixes and reports rather than meets, which follows that document's own structure: it sets no numeric thresholds on assay performance and records instead that performance is understood and deemed acceptable with a rationale given. A method meeting the left column and demonstrating the right has been validated in the sense the pathway means.

3 Results and Discussion

3.1 The achievable bar

Colilert and MF agreed with each other at $R^2 \approx 0.52$, with MF reading $\approx$0.34 $\log_{10}$ ($\sim$2.2$\times$) higher. Their ISO 17994 mean relative difference of +79% exceeds the $\pm$20% equivalence limit roughly fourfold, so the two approved methods are formally non-equivalent. Near any applied limit the reference also fails to resolve category: at 126 CFU/100 mL, 8.1% of the paired grab record (17/209) carries a tabulated Colilert 95% interval straddling the limit, and 11.0% of the Boulder in-situ record (16/145). That band moves with the limit and with where a site's concentrations sit relative to it, and sampling more grabs does not remove it.

3.2 Reference uncertainty propagated

At the worked example's operating point (decision cut 82 CFU/100 mL, action limit 126 CFU/100 mL, $n = 209$) point-label balanced accuracy was 0.900 (sensitivity 0.941, specificity 0.859). Weighting each sample by the probability that its true value exceeds the limit gave an expected balanced accuracy of 0.889 (sensitivity 0.910, specificity 0.867), a shift of -0.012. Crediting only predictions the reference cannot distinguish from correct raised balanced accuracy to 0.963 (sensitivity 1.000, specificity 0.927). Of the point predictions, 75.6% fell inside the reference's own 95% interval, re-

producing on this record the same result reported for the instrument in the companion validation[9]. A 126 CFU/100 mL cut on the prediction rather than the derived cut gives 0.864, 0.840 and 0.971.

3.3 Agreement against the adoption threshold

The grab-and-read record holds 209 paired observations, 176 from bench dilution series and 33 from field sampling; requiring both assays at or above the 10 CFU/100 mL LOQ leaves 190 (157 bench, 33 field). The Boulder in-situ record holds a further 145. Both clear the minimum of 30.

In grab-and-read mode (Table 2, Fig. 3) the method met the adoption threshold against both references: IA = 0.962 against Colilert ($n = 209$; 0.905 within the LOQ, $n = 190$) and 0.906 against MF ($n = 206$; 0.881, $n = 184$). Colilert and MF reached 0.794 with each other (0.705 within the LOQ, $n = 134$). The method therefore agrees with either reference more closely than the two approved methods agree with each other. A candidate cannot reasonably be required to match a reference more closely than a second approved method does. On a field-only subset ($n = 33$) agreement with Colilert held at IA 0.91 while agreement with MF fell to 0.52, so the pooled MF result is buoyed by the wide bench range and Colilert is the robust field reference.

Scored as the method operates in the water, on 145 pairs at six Boulder Creek stations with no per-sample handling and no site refitting, the deployed model reached IA = 0.79 on $R^2 = 0.47$ (Fig. 2A). The R^2 falls below the secondary route of 0.60, so this arm qualifies on the index of agreement alone.

Those figures are in-sample, so the arm was cross-validated by leaving out one calendar day at a time, every grab sharing a date being held out together. On the current record for that reach, 155 pairs across six sensors and 34 days carrying 33 exceedances, the deployed parameterisation gives a held-out IA of 0.711 against 0.799 in sample, and R^2 of 0.261 against 0.425. Refitting with slopes shared across sensors rather than fitted per sensor gives 0.741 held out against 0.770 in sample. The arm therefore clears the 0.70 acceptance threshold out of sample as well as in it, on either parameterisation, while the R^2 falls further below the secondary route. Classification is the more robust of the two: balanced accuracy at the deployment's decision cut moves only from 0.694 to 0.663. The Seine and Marne arm behaves the same way under the same protocol, falling from 0.742 in sample to 0.708 held out, and also clearing the threshold.

Holding out forward in time rather than by day is a harder question, and the arm does not pass it. Fitting on the earliest 80% of Boulder Creek sampling days and scoring the latest 20%, 36 grabs over the following six weeks, gives IA = 0.595 and $R^2 = 0.151$. The reason is specific, and it points at an operational requirement rather than at the measurement. Allowing a single fleet-wide level offset fitted on the test window recovers almost nothing, from 0.595 to 0.620; allowing a level offset per sensor recovers the arm to 0.779. Both offsets are diagnostics rather than predictors, since each is fitted on the window it is scored on. What they separate is the shape of the relationship, which survives six weeks, from the individual sensor levels, which do not. Balanced accuracy over

the same window rises to 0.792, so the class decision holds while the level drifts. A method of this kind therefore carries a refit obligation: the levels have to be maintained against continuing paired samples, and a fixed fit extrapolated forward is not supported by this record. Published out-of-sample evidence exists from a different water body: over the 2025 bathing season an independent Seine deployment produced 310 daily paired samples across three swimming docks, and a model trained on the first 80% and evaluated on the subsequent 20% classified the 900 CFU/100 mL bathing threshold at 96.8% overall accuracy and 94% balanced accuracy[9].

A second in-situ record comes from an independent deployment on the Seine and Marne, four sites scored against that jurisdiction's 900 CFU/100 mL bathing limit on 98 paired Colilert grabs. Holding the deployed functional form and refitting its per-site levels and amplitude coefficient on the full record gives IA = 0.742 and $R^2 = 0.391$ (Fig. 2B); the coefficients in service, fitted on 91 of the same rows, give 0.732 on all 98. No readings are excluded beyond the rules of Section 2.4. Sensitivity at that deployment's decision cut is 5 of 8 exceedances. Predictions there reach about 690 CFU/100 mL against a 900 limit, because levelling the site intercepts onto 900 requires roughly $+0.5$ $\log_{10}$ and degrades the continuous fit, so this arm clears the agreement threshold while undercalling the largest events. A second water body under a different jurisdiction's limit is what the arm demonstrates; detection at the limit itself is where it is weakest.

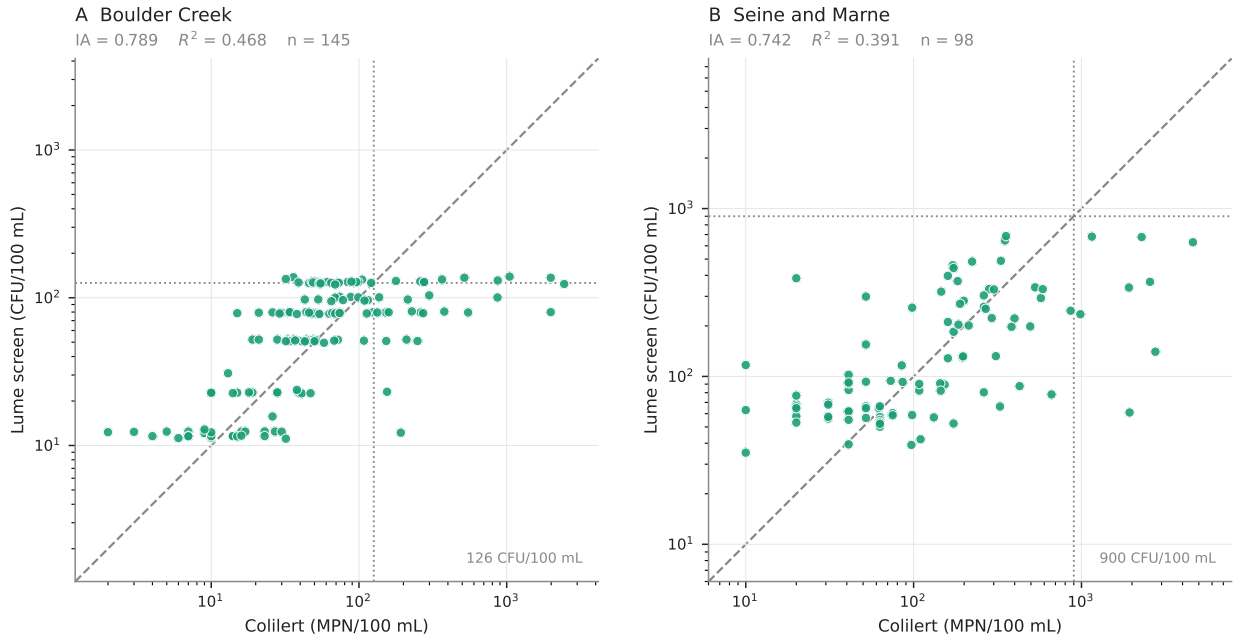

Figure 2: Continuous in-situ agreement, each record against its own jurisdiction's action limit (dotted crosshairs; dashed 1:1). **A**, Boulder Creek, six stations, 126 CFU/100 mL. **B**, Seine and Marne, four sites, 900 CFU/100 mL, on the refitted deployed form. Panels are square on a common log scale. Panel A pairs against the published Alternative Methods Calculator record; panel B uses exact-minute, install-window-enforced pairing. The compression of predicted against reference range in both panels is the magnitude limitation the R^2 values report.

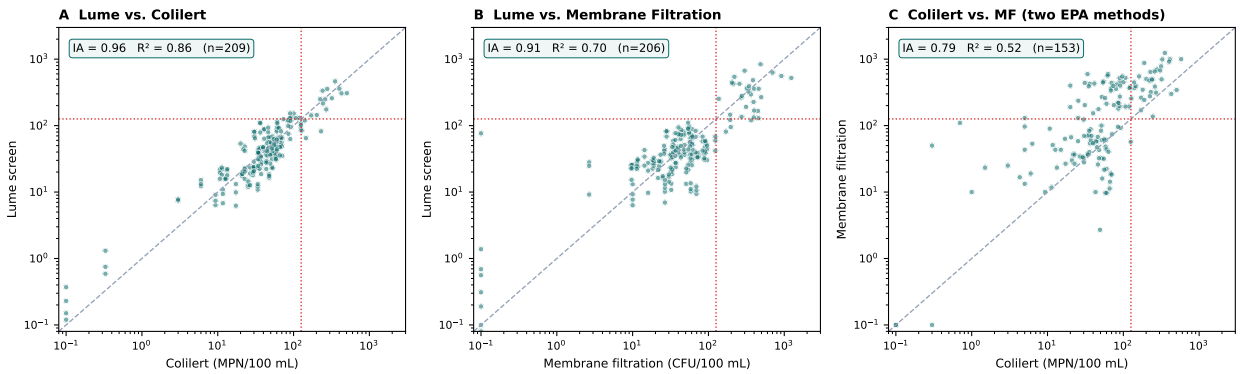

Figure 3: Grab-and-read agreement (index of agreement on $\log_{10}$ pairs; dashed 1:1, dotted 126 CFU/100 mL crosshairs). The method agrees with both EPA references above IA = 0.70 (A, B) and more closely than the two EPA methods agree with each other (C).

3.4 Agreement at the quantity the criterion is written on

Every figure above compares one reading against one grab, and no criterion is written that way. The RWQC recommend that a waterbody geometric mean not exceed the magnitude in any 30-day interval, and that excursions of the statistical threshold value not exceed 10% of that same interval; the alternative-indicator pathway directs a site-specific criterion to keep the same duration and frequency[1,7]. A regulator therefore computes a windowed geometric mean and an excursion frequency.

Scored that way on the Boulder Creek record (Table 3, Fig. 4A), agreement improves on the error measures and not on the acceptance statistic. Root-mean-square error falls monotonically with the averaging period, from 0.432 $\log_{10}$ sample-to-sample to 0.257 at 60 days, a reduction of 40%, and R^2 rises from 0.425 to 0.500. The index of agreement barely moves. It is normalized by the spread of the reference, and averaging compresses that spread along with the error, so the ratio is close to unchanged. Averaging makes the estimate better without making the agreement statistic look better, which is worth knowing before a submission is built on the expectation that it will.

The frequency half of the criterion behaves differently, and it is where a continuous record has no substitute. With n grabs in a window the only measurable excursion frequencies are multiples of $1/n$. The windows here hold 5 to 15 grabs, so the 10% threshold falls between 0 and 20% with nothing in between: in one five-grab window a single exceedance reads as 20%, twice the threshold, while the continuous record over the same window measures 0.2%. A grab schedule estimates the frequency component at a resolution coarser than the threshold it is tested against (Fig. 4B). A continuous record measures it directly.

Two limits bound this section. Six non-overlapping 30-day windows is a small basis, and rolling windows are not independent. And no window in this record carries a geometric mean above

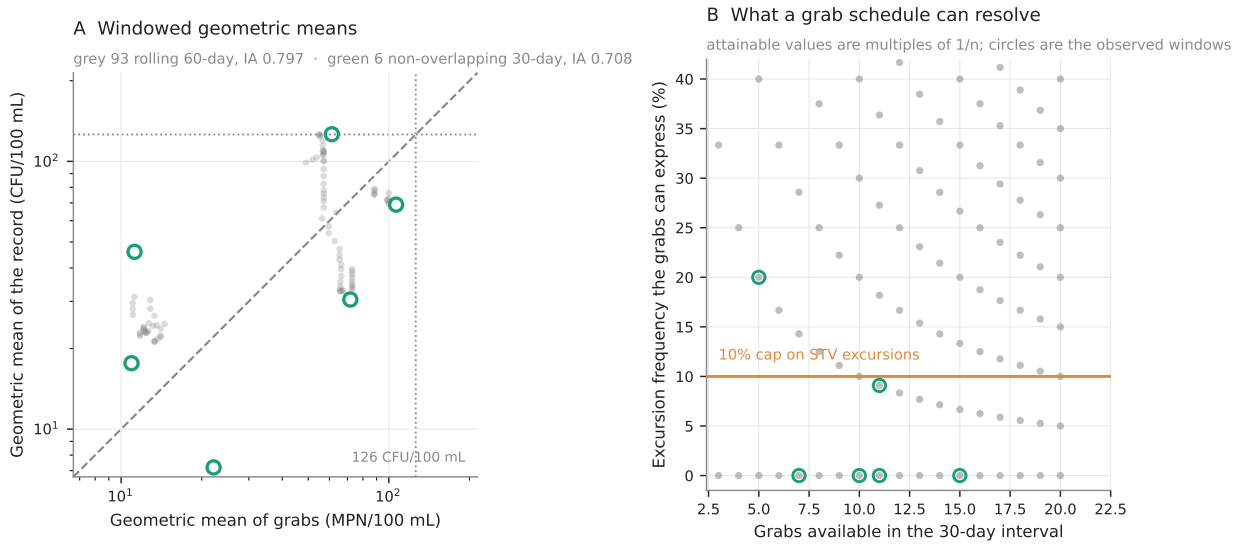

Figure 4: The method scored at the quantity a criterion is written on. **A**, windowed geometric means, record against grabs, on a square log scale with a dashed 1:1 line and the 126 CFU/100 mL magnitude as dotted crosshairs; grey points are 60-day windows rolled daily, which share data and form one streak per station, and green circles are the six independent non-overlapping 30-day windows. **B**, why the frequency half of the criterion cannot be estimated from grabs: with n grabs the attainable estimates are multiples of $1/n$, drawn as the grey ladder, and the 10% cap passes through the gaps for every n below ten. Circles are the observed windows, including a five-grab window in which one exceedance reads as 20%.

126 CFU/100 mL, the range being 7 to 126, so these data cannot test classification of a window as exceeding, only agreement on the value.

3.5 Which regulatory uses a TLF record carries

Where a state adopts TLF as a site-specific alternative indicator into its water quality standards for a named waterbody, with EPA Regional approval, the alternative carries the unchanged numerical criterion and its record becomes the record the recreation-use assessment is made from. Because the TLF–FIB relationship depends on the character and age of the fecal sources at a site, the demonstration is repeated per waterbody and re-confirmed on the state's triennial standards review.

Adoption brings the assessment rules with it, and those rules were written for sparse data. Under a geometric-mean criterion, whatever averaging period a jurisdiction sets, a record reporting in every period is assessed in every period, winter included, where a seasonal grab program produces no data and is not assessed. The small-sample allowances that protect a sparse record fall away. The averaging period and any excursion-frequency provision remain the jurisdiction's to set. A program should settle these before adoption.

Several uses need no regulatory status. Indicating when contact recreation carries elevated risk while the contamination is occurring is one. Directing where confirmatory culture samples are taken

is another, and it is worth quantifying, because it is the use that changes what a laboratory budget buys. Over the Boulder Creek record the deployed model resolves 78 distinct excursion events across the six stations, counting a maximal run of hours at or above the limit as one event. The 132 grabs collected over the same window on the program's routine schedule coincided with 24 of them, 31%. Two-thirds of what the continuous record identifies went unsampled, while the analyses that were run fell mostly on unremarkable water. Read this as a statement about sampling design and not about accuracy: the events are defined from the sensor's own estimate, so the comparison is between two sampling designs rather than against truth, and the grabs in question were collected on a calibration schedule and not to intercept excursions. What it shows is that a fixed schedule and the conditions that carry information are close to independent of one another, and that a record which reports continuously can tell a program which days are worth an analysis. Locating sources and testing whether an intervention changed anything depends on a dense record between stations. In each of these TLF augments a culture reference that remains the confirmatory measurement.

Approval of a method against a water quality standard is a different determination from approval of an analytical method for permit compliance. Analytical methods for NPDES purposes are governed by 40 CFR 136, whose Alternate Test Procedure route validates against spiked matrices where the alternative-indicator pathway analyzes unspiked environmental samples, and a correlative indicator measuring a different analyte is not straightforwardly eligible in any case. A site-specific standards revision therefore supports assessment, listing and progress-tracking against a criterion without authorizing the record for permit monitoring.

3.6 Scope and limitations

Separating the two modes matters for reading any of these numbers. Both are hardware and analytics together and differ in which analytics: grab-and-read carries the per-unit calibration alone, on a captured sample under controlled handling; continuous in-situ carries a fitted model with a site term and any declared covariates, under field conditions the bench does not reproduce. Reporting only the first overstates what a deployment delivers.

Inter-unit reproducibility currently exceeds the culture reference's measurement precision, so multi-unit deployments need the precision improvement reported as ongoing in the companion work. The IA values here replicate the Calculator's arithmetic rather than its output. The Boulder in-situ arm comes from six stations on one stream, which is what a site-specific adoption requires and is not evidence of transfer, and its R^2 sits below the secondary regression route in sample and further below it held out. Cross-validation also shows two things about the fit. The deployed parameterisation is overfitted in one specific way, in that fitting a separate response slope per sensor wins in sample and loses held out against slopes shared across sensors. And the per-sensor levels drift on a timescale of weeks, so the arm that clears the threshold under day-blocked validation does not clear it when a fixed fit is extrapolated six weeks forward. Per-unit calibration of the measurement is not in question, since that is fixed against standards rather than against the reference; what

does not survive is fitting a separate *model* slope per sensor on a paired record this size. The evidence throughout comes from one worked example implementation, which demonstrates that a method specified and evaluated as above meets the acceptance criteria on real environmental samples. The specification and the criteria are what transfer, and any implementation meeting the first is evaluated against the second in the same way.

The published TLF literature [10–14] is cited in the Introduction to establish that the underlying relationship is reported independently across matrices and instruments. Those studies report different metrics against different reference assays and are not pooled with the figures here.

Author Information

Corresponding Author. Evan Thomas, Mortenson Center in Global Engineering and Resilience, University of Colorado Boulder, Boulder, CO 80309, United States. Email: evan.thomas@colorado.edu.

ORCID. Evan Thomas, 0000-0003-3095-8407; Matthew R. V. Ross, 0000-0001-9105-4255; Michael J. Vlah, 0000-0002-6260-2416; Whitney Knopp, 0009-0008-3245-2059.

Notes. The authors declare the following competing financial interest. Several authors are affiliated with Virridy, Inc., the developer of the instrument used here as the worked example implementation, and are co-authors of the companion validation studies[9,10]. The paired records, the analysis code and the acceptance criteria are published in full so that every number reported here can be recomputed independently, and results are reported irrespective of performance outcomes.

Data and Code Availability

The paired records behind every figure and table are published with the validation record at `validation.thelume.ai/TLF-Method`, which serves the Boulder Creek in-situ pairs as CSV, the populated EPA Alternative Methods Calculator workbook, and the grab data behind each panel. Analysis code is in the project repository: `recompute_ia_table.py` regenerates Table 2; `reference_uncertainty_` regenerates the propagation of Section 3.2, reusing the per-barcode grab-and-read calibration of `generate_three_way_comparison.R` and reproducing $n = 209$ exactly; `paris_2026_in_situ_ia.py` and `paris_2026_trim_sensitivity.py` regenerate the Seine and Marne arm; `make_fig_insitu.py` draws Figure 2.

Acknowledgment

This work was supported by the National Science Foundation Convergence Accelerator (Award 24C0011), the National Science Foundation ASCEND Engine (Award NSF-2315760), and the National Aeronautics and Space Administration (Award 80NSSC24K0998). The authors thank

Michael Lawlor of the City of Boulder Utilities Department; colleagues at Colorado State University for the reference culture analysis; Marion Delarbre of the City of Paris, Françoise Lucas of the Université Paris-Est Créteil, Jean-Marie Mouchel of Sorbonne Université, and the Syndicat Marne Vive; Melissa Pierce and Tyler Stephen of Current Water; Jillian Knittle and Christina Miller of the Northeast Ohio Regional Sewer District; Natalie Van Scyoc and Jeff Pu of the Cleveland Water Alliance; and Ray Ozzie of Blues Wireless.

During the preparation of this work the authors used Anthropic's Claude to edit and revise manuscript text, to generate the tables and figures, and to check reported values against the underlying data. All analyses were re-derived from the primary records rather than transcribed, and the authors reviewed and edited the output and take full responsibility for the content of this article.

References

1. U.S. Environmental Protection Agency. *Recreational Water Quality Criteria*; EPA 820-F-12-058; Office of Water: Washington, DC, 2012.
2. Sorensen, J. P. R.; Lapworth, D. J.; Marchant, B. P.; Nkhuwa, D. C. W.; Pedley, S.; Stuart, M. E. In-Situ Tryptophan-like Fluorescence: A Real-Time Indicator of Faecal Contamination in Drinking Water Supplies. *Water Res.* **2015**, *81*, 38–46. DOI: 10.1016/j.watres.2015.05.035.
3. Nowicki, S.; Lapworth, D. J.; Ward, J. S. T.; Thomson, P.; Charles, K. Tryptophan-like Fluorescence as a Measure of Microbial Contamination Risk in Groundwater. *Sci. Total Environ.* **2019**, *646*, 782–791. DOI: 10.1016/j.scitotenv.2018.07.274.
4. Sorensen, J. P. R.; Nayebare, J.; Carr, A. F.; Lyness, R.; Campos, L. C.; Ciric, L. In-Situ Fluorescence Spectroscopy Is a More Rapid and Resilient Indicator of Faecal Contamination Risk in Drinking Water than Faecal Indicator Organisms. *Water Res.* **2021**, *206*, 117734. DOI: 10.1016/j.watres.2021.117734.
5. Khamis, K.; Sorensen, J. P. R.; Bradley, C.; Hannah, D. M.; Lapworth, D. J.; Stevens, R. In Situ Tryptophan-like Fluorometers: Assessing Turbidity and Temperature Effects for Freshwater Applications. *Environ. Sci.: Processes Impacts* **2015**, *17*, 740–752. DOI: 10.1039/C5EM00030K.
6. Bedell, E.; Harmon, O.; Fankhauser, K.; Shivers, Z.; Thomas, E. A Continuous, In-Situ, Near-Time Fluorescence Sensor Coupled with a Machine Learning Model for Detection of Fecal Contamination Risk in Drinking Water: Design, Characterization and Field Validation. *Water Res.* **2022**, *220*, 118644. DOI: 10.1016/j.watres.2022.118644.
7. U.S. Environmental Protection Agency. *Site-Specific Alternative Recreational Criteria Technical Support Materials for Alternative Indicators and Methods*; EPA-820-R-14-011; Office of Water: Washington, DC, 2014. Companion tool: *Alternative Methods Calculator*; EPA 821-B-21-002, user guide EPA 821-B-21-001, 2021.
8. U.S. Environmental Protection Agency. *Microbiological Alternate Test Procedure (ATP) Protocol*; EPA-821-B-10-001; Office of Water: Washington, DC, 2010.

9. Knopp, W.; Klaus, J.; Wilson, D.; Vlah, M. J.; Ross, M. R. V.; Thomas, E. Advancing Continuous In-Situ Quantification of Microbial Contamination in Environmental Waters Using Tryptophan-like Fluorescence: Sensor Design and Validation. *Water Res.* **2026**, *302*, 126099. DOI: 10.1016/j.watres.2026.126099.
10. Knopp, W.; Ecklu, J.; Ross, M. R. V.; Thomas, E. Validating Continuous In-Situ Quantification of Microbial Contamination in Drinking Water Using Tryptophan-like Fluorescence. *Water Res. X*, submitted for publication, 2026.
11. IDEXX Laboratories. *Quanti-Tray/2000 Most Probable Number Table with 95% Confidence Limits*; IDEXX Laboratories: Westbrook, ME.
12. *Water Quality — Requirements for the Comparison of the Relative Recovery of Microorganisms by Two Quantitative Methods*; ISO 17994; International Organization for Standardization: Geneva, 2014.
13. Willmott, C. J. On the Validation of Models. *Phys. Geogr.* **1981**, *2*, 184–194. DOI: 10.1080/02723646.1981.10642213.
14. U.S. Environmental Protection Agency. *Method 1603: Escherichia coli (E. coli) in Water by Membrane Filtration Using Modified Membrane-Thermotolerant Escherichia coli Agar*; EPA-821-R-14-010; Office of Water: Washington, DC, 2014.

Table 1: Acceptance criteria for a TLF fecal-indicator method. Rows marked *(RWQC)* are requirements of EPA-820-R-14-011 and carry that document's numbers; the rest are stated and reported by the implementation.

Specified	Demonstrated
Measurement	
Measurand	Emission and excitation bands, baseline reference, and the reported quantity as a continuous index.
Corrections	Temperature normalization to a stated reference temperature; turbidity correction derived independently of the fluorescence signal. Response characterized over the turbidity and temperature range claimed.
Per-unit calibration	Linearity against a tryptophan standard series over the working range; gain, temperature coefficient and low-fluorescence reference characterized per instrument and not shared.
Limit of quantification	A validated working range with the lower bound evidenced.
Analytics	
Model form and inputs	Every input carried, with corrections distinguished from exogenous covariates.
Fitting and refitting	How coefficients are obtained, and the cadence at which they are updated, with the cadence evidenced rather than asserted; inputs stated before the agreement analysis.
Exclusion rules	Criteria stated in advance, each decidable from a diagnostic independent of the fluorescence signal and applied without reference to the paired culture value; recoverable effects corrected rather than excluded; each judged exclusion recorded with its reasoning and its count.
Evidence	
Paired record	$\geq$30 paired environmental samples within the quantification limits of both assays, over the range of conditions at the site *(RWQC)*.
Reference	An EPA *E. coli* method or approved equivalent, analyzed unspiked *(RWQC)*.
Agreement	IA ≥ 0.70 permits the unchanged criterion; $R^2 > 0.60$ is a secondary route *(RWQC)*.
Reference uncertainty	Performance against the point label and as an expectation over the reference's interval, with reference-vs-reference agreement as the achievable bar.
Precision	Replicate and inter-unit reproducibility, each against the reference method's own precision.
Validation tier	Single-laboratory (Tier 1) where one organization analyzes the standards samples *(RWQC)*.
Deployment	
Action limit and decision cut	The action limit is the jurisdiction's; the decision cut is derived per deployment and validated against it.
Scope of adoption	Per waterbody, re-confirmed on the state's triennial standards review *(RWQC)*.
Triggered sampling, where used	The conditions that call for a laboratory sample, stated in advance, together with the paired samples the calibration requires regardless of conditions.

Table 2: EPA Alternative Methods Calculator agreement (Willmott IA and R^2 on $\log_{10}$ pairs). IA ≥ 0.70 permits the same numerical criterion. Grab-and-read is a per-unit calibration on the instrument's own channels; continuous in-situ is the deployed model. Modes carry different models and are not pooled. Restricted rows require both assays at or above the 10 CFU/100 mL LOQ.

Comparison (reference vs. candidate)	n	IA	R^2
Grab-and-read			
Colilert vs. candidate	209	0.962	0.861
both assays $\geq$ LOQ	190	0.905	0.680
Membrane filtration vs. candidate	206	0.906	0.701
both assays $\geq$ LOQ	184	0.881	0.618
Colilert vs. MF (two EPA methods)	153	0.794	0.522
both assays $\geq$ LOQ	134	0.705	0.417
Continuous in-situ			
Boulder Creek, six stations	145	0.790	0.468
Seine and Marne, four sites	98	0.742	0.391

Table 3: Agreement between the grab-derived and record-derived geometric means over the same window, Boulder Creek. Rolling windows share data across comparisons and are not independent; the non-overlapping row is.

Comparison	n	IA	R^2	RMSE ($\log_{10}$)
Sample to sample (from Table 2)	155	0.799	0.425	0.432
30-day windows, non-overlapping	6	0.708	0.262	0.394
30-day windows, rolling daily	152	0.745	0.326	0.331
60-day windows, rolling daily	93	0.797	0.500	0.257