**Supplementary Material**

**Davies-Colley et al. (NZJM-2018-0108)**

**Further notes on analysis of paired water quality data**

(duplicate measurements of river water quality in the Wellington Region)

Measurements in the field by Greater Wellington Regional Council (GWRC), and analytical results by Hill Laboratories on water samples collected by GWRC, were paired with field or laboratory measurements by NIWA. Refer to the accompanying paper for details on sites and on statistics of agreement and their interpretation.

Logarithmic transformations were used to improve symmetry of water quality variables before calculating statistics of agreement between NIWA and GWRC measurements (Table 3). Logarithmic scales were also used where appropriate in the X-Y plots (displayed below) comparing the paired measurements. Plotting of typically strongly right-skewed water quality data on linear scales often reduces low value points to an unresolved cluster. Only water temperature, DO, and pH (which is already log-transformed) were sufficiently near-symmetrically distributed to be plotted on linear scales. Logarithmic statistics of agreement have the very desirable feature that they permit comparison between variables of very different untransformed scales. This permits us to compare numerical agreement for a diverse range of variables. We show, for example, that total nitrogen measurements are in closer numerical agreement than *E. coli*, despite completely different (incomparable) untransformed measurement scales for these quantities.

Correlation analysis is typically used to assess linear agreement of variables in X-Y plots. For comparison of different methods for measuring the same attribute, Lin (1989) introduced concordance correlation which measures closeness to the 1:1 line of perfect agreement rather than the regression line (Lin’s coefficient is smaller than Pearson’s except in the special case that the data are perfectly symmetric about the 1:1 line). However, Bland and Altman (1986) argue that numerical similarity is of greater interest than linear agreement in clinical applications, which we believe also applies in environmental applications. One problem with correlations for assessing agreement is that correlation coefficients such as Pearson’s and Lin’s are *range-dependent*. For example, it is possible to have a high correlation for measurements that range widely despite weak numerical agreement (e.g., for *E. coli*); conversely the correlation may be quite poor for variables that range only slightly even when measurements are numerically similar in magnitude (as is typical for DO in well-aerated waters).

We tabulate correlation coefficients (Pearson’s and Lin’s) in Table 3, but also give mean differences as an index of systematic numerical difference as advocated by Bland and Altman (1986). As discussed in the article, we used the standard deviation measured *perpendicular* to the 1:1 line (line of equality) as a measure of numerical agreement. *S*R values are given on the X-Y plots displayed below comparing NIWA and GWRC measurements.

**Graphical comparisons of duplicate measurements of river water quality at state-of-environment sites in the Wellington Region**



Figure S1. Comparison of water temperature measurements by GWRC and NIWA (*N* = 29). On seven occasions on two of the Wainuiomata runs, two GWRC field staff (denoted GWRC1 and GWRC2) made independent temperature measurements which are compared in the plot (triangles versus Xs). The regression fit (geometric average of Y-on-X and X-on-Y lines) is *not* shown, being almost coincident with the 1:1 line.



Figure S2. Comparison of dissolved oxygen (DO; % saturation) measurements by GWRC and NIWA (*N* = 29). On seven occasions on two of the Wainuiomata runs, two GWRC field staff (denoted GWRC1 and GWRC2) made independent DO measurements which are compared in the plot (triangles versus Xs). The regression fit (geometric average of Y-on-X and X-on-Y lines) is *not* shown, being almost coincident with the 1:1 line.



Figure S3. Comparison of electrical conductivity (EC) measurements in the field by GWRC versus NIWA laboratory measurements (*N* = 28; a transcription error on one occasion gave rise to an outlier point indicated by an arrow). On seven occasions on the Wainuiomata runs, two GWRC field staff (denoted GWRC1 and GWRC2) made independent conductivity measurements which are compared in Panel A (triangles versus Xs). The regression fit (geometric average of Y-on-X and X-on-Y lines) is *not* shown, being almost coincident with the 1:1 line.



Figure S4. Comparison of laboratory pH measurements by by Hill Laboratories (for GWRC) versus NIWA laboratory measurements (*N* = 25; laboratory pH was not measured on samples from the four sites on the initial Wainuiomata run). The regression fit (dashed line; the geometric average of Y-on-X and X-on-Y lines) is shown as well as the 1:1 line.



Figure S5. Comparison of visual clarity measurements in the field (black disc method) by GWRC versus NIWA (*N* = 28; a low-biased observation using a worn plastic viewer on one occasion is not shown). On seven occasions on the Wainuiomata runs, two GWRC field staff (denoted GWRC1 and GWRC2) took independent visual clarity observations which are compared in Panel A (triangles versus Xs). The regression fit (geometric average of Y-on-X and X-on-Y lines, dashed line) is shown as well as the 1:1 line.



Figure S6. Comparison of turbidity measurements by Hill Laboratories (for GWRC) versus laboratory measurements by NIWA (*N* = 17; only one data point, from 13 visits, was available for Kaitoke). The regression fit (geometric average of Y-on-X and X-on-Y lines, dashed line) is shown as well as the 1:1 line.



Figure S7. Comparison of ammoniacal nitrogen measurements by NIWA and Hill Laboratories (for GWRC). (*N* = 17: 11 measurements by Hill Laboratories, all but one at Kaitoke, were less than the method detection limit of 5 mg/m3). The regression fit (geometric average of Y-on-X and X-on-Y lines, dashed line) is shown as well as the 1:1 line.



Figure S8. Comparison of nitrate plus nitrite nitrogen measurements by NIWA and Hill Laboratories (for GWRC). (*N* = 28). The regression fit (geometric average of Y-on-X and X-on-Y lines, dashed line) is shown despite being almost coincident with the 1:1 line.



Figure S9. Comparison of total nitrogen measurements by NIWA and Hill Laboratories (for GWRC). (*N* = 29). The regression fit (geometric average of Y-on-X and X-on-Y lines, dashed line) is shown as well as the 1:1 line.



Figure S10. Comparison of dissolved reactive phosphorus measurements by NIWA and Hill Laboratories (for GWRC). (*N* = 27). The regression fit (geometric average of Y-on-X and X-on-Y lines, dashed line) is shown as well as the 1:1 line.



Figure S11. Comparison of total phosphorus measurements by NIWA and Hill Laboratories (for GWRC). (*N* = 29). The regression fit (geometric average of Y-on-X and X-on-Y lines, dashed line) is shown as well as the 1:1 line.



Figure S12. Comparison of faecal pollution as indicated by *E. coli* measurements by Hill Laboratories (for GWRC) and NIWA. (*N* = 25; three measurements were reported as < 1 cfu/100 mL and one result was missing). Hill Laboratories used membrane filtration and NIWA used the Colilert multi-well method. The regression fit (geometric average of Y-on-X and X-on-Y lines, dashed line) is shown as well as the 1:1 line. An apparent outlier point (discussed in the text) is indicated by an arrow.