The human factor: Weather bias in manual lake water quality monitoring

James M. Rand,1* Mirjam O. Nanko,2 Mikkel B. Lykkegaard,3 Danielle Wain,4,5 Whitney King,5 Lee D. Bryant,6 Alan Hunter1

1Department of Mechanical Engineering, University of Bath, Bath, UK
2Exeter Q-Step Centre, University of Exeter, Exeter, UK
3Institute for Data Science and AI, University of Exeter, Exeter, UK
47 Lakes Alliance, Belgrade Lakes, Belgrade, Maine
5Department of Chemistry, Colby College, Waterville, Maine
6Department of Architecture & Civil Engineering, University of Bath, Bath, UK

Abstract

Sampling bias due to weather conditions has been anecdotally reported; however, in this analysis we demonstrate that manual lake sampling is significantly more likely to take place in “fair weather” conditions. We show and quantify how a manual lake monitoring program in Maine, USA, is biased due to wind speed, rainfall intensity, and air temperature. Emulating a manually sampled water quality (WQ) data set, we show that, on average, manual sampling recorded, depending upon depth, higher water temperature (between 0.4°C and 1.2°C), lower dissolved oxygen (DO) (between −0.8 and −0.4 mgL⁻¹), and higher chlorophyll values (2.0 μgL⁻¹) than average automated monitoring. By analyzing the actual manual monitoring data, we show that manually collected lake water temperatures are on average 1.0°C higher in the epilimnion and 0.5°C (corrected for sensor lag) higher in the hypolimnion compared to those collected using automated methods. We attribute these differences in WQ measurement values to the weather-induced manual sampling bias. We believe that the nature of weather bias on manual monitoring will always record higher water temperatures, higher chlorophyll, and lower DO than automatic monitoring. The methodologies presented in this study will apply to similar manually sampled lake monitoring programs and the manual sampling bias will likely affect other WQ parameters. The weather-induced water temperature bias reported is of the same order of magnitude as the root mean square errors reported in many lake models and is therefore considered substantial. If generally applicable and not corrected for, these results will have important implications for climate models, and similar applications, where manually collected WQ data are employed.

The monitoring of lake water quality (WQ) is conducted by ecologists and limnologists for a myriad of reasons, including the assessment of biological and chemical health, meeting environmental or drinking water standards, and obtaining data for long-term time series and lake models. Formalized and regular WQ monitoring programs are an essential part of this measurement and characterization of lakes (Ban et al. 2014). WQ data can be collected by manual means, for example, sensors deployed from a boat, or automated methods, for example, a floating data buoy. Local, regional, or national guidance is often used in the design of such monitoring programs. These WQ monitoring programs, however, can be significantly affected by various sampling biases, both temporal and spatial.

Anecdotally, weather biases in WQ monitoring are intuitively recognized by lake scientists but not necessarily overtly addressed. For example, although regular monitoring is a cornerstone of most WQ sampling programs (British Standards 2006), lake scientists knowingly avoid working from boats in rain and high winds for practical and safety reasons but do not account for this behavior in their sampling programs. Weather bias in data collection is mentioned in the literature but its treatment is rarely addressed. Fair weather bias, in oceanographic terms, is the tendency for ships to avoid or be unable to work in bad weather and thus prevent data collection in such conditions (Dhanak and Xiros 2016). Data gathered by

*Correspondence: mjr46@bath.ac.uk

Additional Supporting Information may be found in the online version of this article.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
ships is invariably subject to such bias (Austin 2012). Guidance on the design of monitoring programs, such as the ISO 5667 series (British Standards 2006), is typically concerned with weather from a boat safety and station-keeping point of view but not from a data bias standpoint. It should be noted that such guidance, when looking for systematic changes over time, generally advises that lake WQ monitoring is conducted regularly on the same day and at the same time to minimize sample variations (British Standards 2006). If monitoring programs could adhere to such rigid timetables, then any weather-derived biases would potentially be eliminated.

In this paper, we examine weather influence on a lake WQ monitoring program conducted on the Belgrade Lakes, Maine, USA, and the influence of this on the WQ data collected. Although the methodology presented is for a single lake monitoring program, it could be applied to similar lake monitoring programs globally to quantify the effects of weather in lake WQ data.

The hypothesis examined is that manual sampling suffers (1) a statistically significant bias due to different types of weather conditions and that (2) these biases have a statistically significant impact on the WQ data collected. To test part one of our hypotheses, we use data from the WQ sampling program from all seven Belgrade Lakes to examine if weather conditions produce a data bias when manual sampling is conducted. Specifically, we examine if average daily wind speed, temperatures, and rainfall affect the probability of manual sampling. Based on the anecdotal evidence, we hypothesize that high winds, cold temperatures, and high rainfall decrease the likelihood of manual sampling or, conversely, that low wind speed, low rainfall, and high air temperatures increase the likelihood of manual sampling to take place, all other things being equal. To test the second part of the hypothesis, manual and automated lake sampling data from Great Pond (one of the Belgrade Lakes) is analyzed. Using the methodologies presented, we show that, for our Great Pond case study, manual lake sampling yields significantly different WQ values (higher temperature and chlorophyll and lower dissolved oxygen [DO] on average) as compared to automated WQ monitoring. This finding may have a more general relevance to all manual WQ sampling parameters and programs but further research would be required to investigate this.

**Monitoring program**

Most lake monitoring programs measure similar WQ parameters on a regular basis (Stanley et al. 2019), including Secchi disk transparency, temperature and DO profiles, and discrete or integrated water samples for parameters such as chlorophyll *a* and total phosphorus. Our analysis comprises two main elements: (1) to assess the presence of a temporal bias due to weather conditions on manual sampling, we have estimated the likelihood of manual sampling taking place on any given day within the sampling period depending on weather parameters obtained from the Belgrade Lakes weather station (see the Influence of weather on manual sampling section) and (2) to see the effect of this weather bias in the WQ data, we have analyzed WQ data from both manually obtained profiles and an automated data buoy on Great Pond (see the Effect of weather sampling bias on WQ parameters using only automated buoy measurements and Effect of weather sampling bias on water temperature using automated buoy and manually collected measurements sections).

The Belgrade Lakes are a chain of temperate lakes in central Maine. Temperature and DO profiles are measured throughout the open water season by the 7 Lakes Alliance (a regional watershed organization) in conjunction with Colby College. The sampling program is based upon the State of Maine WQ monitoring guidance (Webster 1997). Standard procedure is for manual profiles to be taken from a boat weekly in June–August, biweekly in May and September, and once in April and October at various buoyed sample stations in the lakes. Great Pond is a mesotrophic, dimictic lake with a mean depth of 6.4 m and a maximum depth of 21 m. On Great Pond (Fig. 1), there is an automated WQ data buoy at Sta. 2 (also used for mooring boats when sampling manually), which measures temperature, DO, chlorophyll, and photosynthetically active radiation (PAR) at fixed water-column depths every 15 min, and an automated weather station located on the lake shore that samples weather parameters every 15 min. These spatially adjacent and temporally aligned monitoring methods allow for weather and WQ data to be compared (Fig. 2).

Equipment used to monitor the weather and WQ parameters comprises:

- **Manual sampling:** Aqua TROLL 400 Multiparameter Probe measuring temperature (sensor accuracy: ±0.1°C, resolution: 0.01°C with a sensor lag of <30 s, factory calibrated with zero drift), DO (sensor accuracy: ±0.1 mg/L, resolution: 0.01 mg/L with a sensor lag of <60 s) and depth (sensor accuracy: ±0.08 m, resolution: ±0.008 m) deployed from a boat moored (in order to minimize wind induced drift and therefore minimize vertical offset of the manual sensor in the water) to the automated data buoy.
- **Automated data buoy:** fitted with NexSens T Node FR Water Temperature String (sensor accuracy: ±0.075°C, resolution: 0.01°C, sensor lag: 60 s, factory calibrated with zero drift. Sensors at 2-m intervals from 1 m down to 19 m), PONSEL Optical Dissolved Oxygen Sensor (sensor accuracy: ±0.1 mg/L, resolution: 0.01 mg/L. Sensors at 2, 6, 9.6, and 16 m), Turner Designs Cyclops-7 Chlorophyll Sensors (2 and 6 m), and LI-COR LI-190 and LI-192 series PAR sensors (surface and 2 m).
- **Weather station:** WS700-UMB Smart Weather Sensor.

**Influence of weather on manual sampling**

**Data**

For the analysis of weather bias in the sampling program, manual sampling data (see the Monitoring program section)
from all seven Belgrade Lakes and the data from the Great Pond weather station were combined. It is assumed that the weather parameters recorded at the weather station represent the Belgrade Lakes which are all within a 12 km radius of the weather station. For this study, we had access to valid observations from 26 July 2016 to 08 November 2019, a total of 1201 ds. Manual sampling is scheduled to take place each year between April and October, resulting in a total sample of 740 ds. The weather data were aggregated for each day, providing daily average measurements of air temperature, wind speed, and rainfall intensity (Table 1). Manual sampling took place on 221 ds (in at least one of the seven lakes), that is, on around 30% of the 740 ds in the sample. Several control variables were introduced to account for the temporal nature of the data, namely a binary variable measuring if the day was a working day (Monday–Friday) or not and temporal lag variables, accounting for the planned regularity of the sampling program.

Methods

To establish if the weather has an influence on the manual sampling program, we model the probability that manual sampling takes place on any given day, depending on the weather that same day. Specifically, we consider a linear time series model where the dependent variable $y_t$, $t = 1, 2, \ldots, T$, is a realization of a Bernoulli process (Nyberg 2010), representative of the manual sampling outcome at time $t$. As such, the dependent variable is operationalized in the following way:

$$y_t = \begin{cases} 1, & \text{if manual sampling took place on day } t \\ 0, & \text{if no manual sampling took place on day } t \end{cases}$$

We estimate this probability of sampling taking place using a logistic regression (LR) model with the independent variables of daily mean air temperature ($T_{air}$), mean wind speed ($S_{wind}$), and mean rainfall intensity ($I_{rain}$). Additionally, we control for whether the day was a working day or not (WD). The independent variables for each day $t$ can then be collected in a vector $x_t = \{T_{air}, S_{wind}, I_{rain}, WD\}^T$. We also initially included the prospective sampling schedule, that is, how many samples were supposed to be taken according to the seasonal schedule (see the Monitoring program section), as a control variable but did not include it in the final model since it had no significant effect and did not improve the model fit.
Fig. 2. Example of annual and daily time series of the Belgrade Lakes regional weather data and Great Pond water temperatures (sampled at Sta. 2 of Great Pond by both manual and automated means). (a) Peak daily air temperature and total daily rainfall for 2018. (b) Air temperature and rainfall (measured every 15 min) for 04 August 2018. (c) Great Pond water temperature (average daily) for 2016. (d) Great Pond water temperature (measured every 15 min) for 14 July 2016.
Logistic regression
Following the notation of Hosmer et al. (2013), the standard LR model can be written as

\[ \pi(x) = \frac{e^{\beta^T x}}{1 + e^{\beta^T x}} \]

where \( \pi = E(Y|x) \) is the conditional mean of the dependent variable \( Y \) given the independent variables \( x \). Here, \( x = (1, x_1, \ldots, x_n)^T \) is a vector of \( n \) independent variables, augmented with unity to allow for an intercept term and \( \beta = (\beta_0, \beta_1, \ldots, \beta_n)^T \) is a vector of the corresponding regression coefficients. These regression coefficients are determined by maximum likelihood estimation using observations of the pairs \((x_t, y_t)\), \(t = 1, 2, \ldots, T\). Note that, for a LR, the regression coefficients \( \beta \) represent the slope of each independent variable for the log-odds of the dependent variable. Hence, care must be taken when interpreting the numerical results. For ease of interpretation, we provide marginal effects plots (Fig. 3) to visualize the effect of each independent variable on the dependent variable.

Due to the temporal nature of these data and the fact that the manual sampling program was intended to follow a specified schedule, the observations were not strictly independent and temporal autocorrelation was expected to be an issue. Serial dependence was investigated using (1) the autocorrelation function (ACF), which calculates the correlation between the dependent variable with itself at previous times (lags), (2) the partial autocorrelation function (PACF), which estimates the partial correlation between different lagged observations while controlling for the effect of shorter lags (Metcalfe and Cowpertwait 2009), as well as (3) the Durbin–Watson test for autocorrelation. To control for temporal autocorrelation, which the standard LR model does not account for, two additional LR models were investigated: a lagged dependent variable model (LDV) and a generalized linear autoregressive moving-average model (GLARMA). For each of the models, the relevant assumption tests were completed following standard procedures (Menard 2010; Hosmer et al. 2013; Garson 2016).

Dependent variable model
The LDV model controls for the temporal autocorrelation by including the dependent variable \( y \) at some time lag \( k \) as an independent variable (Menard 2010). This is done by adding to the vector of independent variables \( x_t = (1, x_{1,t}, \ldots, x_{n,t}, y_{t-k})^T \) the dependent variable \( y_{t-k} \) where \( k \) is the lag. The vector of coefficients then becomes \( \beta = (\beta_0, \beta_1, \ldots, \beta_n, \alpha_k)^T \), where \( \alpha_k \) is an additional regression coefficient which, broadly speaking, measures the influence of the dependent variable at lag \( k \). The advantage of this approach is the ease of interpretation of the resulting model, as the lags can simply be interpreted in the same way as any other binary independent variable. However, the inclusion of lagged dependent variables can potentially bias the regression coefficients (Achen 2000; Keele and Kelly 2006).

Generalized linear autoregressive moving-average model
To validate the conclusions, we fitted a GLARMA model—an extension of generalized linear models which robustly controls for serial dependence by including autoregressive (AR - \( \phi_\theta \)) and moving-average (MA - \( \theta \)) terms (Dunsmuir 2015; Dunsmuir and Scott 2015). A more detailed description of the GLARMA model is given in Dunsmuir and Scott (2015) and Dunsmuir (2015).

Results
All three LR models were applied to the data and Table 2 shows the regression coefficients, standard errors, and significance levels of each. Figure 3 shows the marginal effects plots for each weather-related independent variable, for ease of model interpretation. Regardless of the model, it is evident that the air temperature and whether it is a working day (weekday) have a positive effect on sampling events and that wind speed and rainfall intensity have a negative effect on sampling events. According to the ACF and PACF, the standard LR model exhibited significant autocorrelation for lags \( k = \{1, 5, 7, 14\} \), which was also reflected in the Durbin–Watson test for autocorrelation (\( DW = 1.477, p < 0.001 \)). Accordingly, the regression coefficients of the standard LR model should not be interpreted and are only reported for completeness. We included the lags exhibiting significant autocorrelation in the LDV and GLARMA models, which eliminated all significant autocorrelation. We hypothesize that these 1, 5, 7, and 14 d lags broadly represent prevailing sampling patterns.

### Table 1. Descriptive statistics for the weather bias model (N = 740 d).

| Statistic                                    | Min | Median | Max | Mean |
|----------------------------------------------|-----|--------|-----|------|
| Manual sampling conducted (0 = no, 1 = yes)  | 0   | 0      | 1   | 0.30 |
| Mean air temperature (°C)                    | -1.64 | 17.55 | 30.01 | 16.49 |
| Mean wind speed (m s\(^{-1}\))               | 0.32 | 1.09   | 6.17 | 1.22 |
| Mean rainfall intensity (mm/15 min)          | 0   | 0      | 0.45 | 0.02 |
| Working day (0 = no, 1 = yes)                | 0   | 1      | 1   | 0.71 |
Introducing lags in the LDV model reduces the effect of air temperature, wind speed, and weekday, while the effect of rainfall intensity is increased. These differences are due partly to causal relationships exposed by controlling for past observations and partly to bias introduced through that same process (Achen 2000; Keele and Kelly 2006), as mentioned earlier. However, it is not particularly concerning in this case, considering that the differences are relatively minor and the coefficients remain significant in the LDV model. Considering the GLARMA model, we observe that all regression coefficients converge to values between those of the two previous models, and that the standard errors are mostly of a lower order than any of the other models. Since GLARMA is generally considered more robust than the LDV approach for time-series regression (Dunsmuir 2015; Dunsmuir and Scott 2015), we believe that the GLARMA model best represents the underlying “true” relationship between weather conditions and manual sampling events, even though the LDV model achieves a better Akaike Information Criterion (AIC, see Quinn and Keough 2002). It is likely that the relatively simplistic approach of including lagged dependent variables has introduced some bias in the LDV model, which would be reflected in a higher likelihood and consequently lower AIC, given that the models have the same number of estimated parameters. While the McFadden pseudo $R^2$ cannot be interpreted as analogous to the coefficient of determination in a linear least squares model, values above 0.3 indicate an excellent fit for a logistic model (McFadden 1979). This strongly supports our hypothesis that weather conditions influence the probability of undertaking manual WQ samples on a particular day. It should be noted that, while the exact meanings of “good” and “bad” weather are subjective, and especially depending upon

Fig. 3. Marginal effects plots for all the models. $E(Y|x)$ is the conditional mean of the dependent variable, and broadly represents the probability of a sample being taken. For each effects plot, the remaining variables were held constant at their means. Shaded areas show 95% confidence intervals. These results align with our hypothesis and the anecdotal behavior of lake scientists where “good weather” promotes manual sampling taking place and “bad weather” discourages or prevents manual sampling from taking place.
the time of year, the GLARMA model uses moving averages for the weather data so in effect accounts for such seasonal variations.

### Effect of weather sampling bias on WQ parameters using only automated buoy measurements

Having shown that weather conditions bias when manual sampling takes place, this and the next section examine the effect of the weather bias on the manually collected WQ data.

In the Effect of weather sampling bias on WQ parameters using only automated buoy measurements section, we use only data from the automated data buoy on Great Pond. By using this single data source, and using it to emulate the manual sampling data, any potential issues concerning sensor calibration, sensor drift, positional accuracy, environmental, and lake effects are eliminated. The analysis in this section demonstrates a consistent effect in the data, which we ascribe to the manual sampling weather bias.

The Effect of weather sampling bias on water temperature using automated buoy and manually collected measurements section analyses water temperature data to directly compare the automated and manually collected data. This relates our findings to actual manual measurements and additionally reveals sensor lag as another critical issue with manual sampling. While there are some caveats with respect to comparing measurements from two different sensors, the findings of both approaches align closely, hence mutually validating both analyses.

### Data and methods

For the sampling period April to October, obtained annually for 4 yr 2016–2019, we have approximately 35,000 automated measurements for each WQ sensor over a total of 422 d from the data buoy on Great Pond. The details and depths of these water temperature, DO, chlorophyll, and PAR sensors are given in the Monitoring program section. In this same period, we identified 51 ds when manual sampling took place on Great Pond.

By taking a subset of the automated data buoy measurements, filtered by the exact time window when actual manual sampling measurements took place, we obtained an emulated manual sampling data set. By comparison of the full automated data set and the emulated manual data set, we investigated the measurement bias introduced by the weather-dependent manual sampling. Under the assumption that the mean values of all automatic measurements of each WQ parameter represent the true population means ($\mu_{true}$), we estimated the difference in the subset means ($\mu_{sub}$) compared to the true means using one-sample t-tests (Ross 2021). For each t-test, we report the t-statistic (t), degrees of freedom (df) and

| Predictor       | (1) standard LR model | (2) LDV model | (3) GLARMA model |
|-----------------|-----------------------|---------------|-----------------|
|                 | $\beta$ | SE $\beta$ | $\beta$ | SE $\beta$ | $\beta$ | SE $\beta$ |
| Air temperature | 0.10**** | 0.02 | 0.04** | 0.02 | 0.07**** | 0.03 |
| Wind speed      | -0.75**** | 0.21 | -0.51** | 0.22 | -0.66**** | 0.14 |
| Rainfall intensity | -5.98**** | 2.22 | -10.10**** | 2.46 | -7.96**** | 2.05 |
| Weekday         | 3.34**** | 0.43 | 2.90**** | 0.46 | 3.32**** | 0.35 |
| LDV lags ($k$)  |          |      |          |      |          |      |
| $\alpha_1$ (1 d) | 1.39**** | 0.22 |      |      |          |      |
| $\alpha_5$ (5 d) | 0.70*** | 0.23 |      |      |          |      |
| $\alpha_7$ (7 d) | 0.64*** | 0.22 |      |      |          |      |
| $\alpha_{14}$ (14 d) | 0.60*** | 0.22 |      |      |          |      |
| GLARMA coeff.   |          |      |          |      |          |      |
| $\phi_1$ (1 d AR) |          |      | 0.58**** | 0.05 |          |      |
| $\phi_5$ (5 d AR) |          |      | 0.14*** | 0.05 |          |      |
| $\phi_7$ (7 d MA) |          |      | 0.16* | 0.09 |          |      |
| $\phi_{14}$ (14 d MA) |          |      | 0.27**** | 0.07 |          |      |
| Constant        | -4.38**** | 0.61 | -4.45**** | 0.65 | -3.98**** | 0.74 |
| Observations    | 740 |      | 740 |      | 740 |      |
| Log likelihood  | -338.75 |      | -301.53 |      | -311.07 |      |
| Akaike Inf. Crit. | 687.49 |      | 621.07 |      | 640.14 |      |
| McFadden $R^2$  | 0.25 |      | 0.33 |      | 0.31 |      |
significance level in the form of the p-value (p) (see, e.g., Berthouex and Brown 2002; Quinn and Keough 2002; Ross 2021). The PAR data did not meet the criteria for robust t-test analysis and was, therefore, not included.

Results
The analysis showed a statistically significant difference between the mean of the subset and the population mean in the entire water column for the water temperature data, the DO data, and for the 2 m chlorophyll sensor (Table 3). The 6 m chlorophyll sensor shows no significant difference. We ascribe the statistical difference between all automated collected data and the emulated manual sampled data to the effect of the weather bias on the collection of the manual WQ data.

Water temperature measurements from the manual emulated were, on average and depending upon depth, between 0.4°C and 1.2°C higher than the total measurements gathered by the automated buoy, while annual DO measurements record between −0.8 and −0.4 mgL$^{-1}$ lower values. Chlorophyll at depth 2 m is measured at 2.0 μgL$^{-1}$ higher. These findings make intuitive sense where warmer, less windy, and sunnier “manually biased” weather yields higher water temperature, which would encourage algae growth as well as decreased capacity to hold DO.

Examining the difference between the subset mean and population mean $\mu_{\text{sub}} - \mu_{\text{true}}$ depending on the depth, we find a stratified depth dependency where the consistent magnitude of these WQ parameter differences abruptly changes. For water temperature, there is an approximate 1.1°C difference above the 9 m depth, and an approximate 0.4°C difference below. This phenomenon is investigated further in the next section.

**Effect of weather sampling bias on water temperature using automated buoy and manually collected measurements**

With a weather bias effect upon WQ data having been demonstrated using emulated data, this section investigates the bias effect using the actual manually collected water temperature data to confirm our findings and examine the depth effect noted in the Results section above. The manual and automatic sensors were not jointly calibrated. However, according to the manufacturer, all of the temperature sensors involved in this analysis require no periodic calibration and do not drift. Despite the manufacturer’s calibration assurance,

| Variable and depth | $\mu_{\text{sub}} - \mu_{\text{true}}$ | 95% CI       | t     | df  | p   |
|--------------------|--------------------------------------|--------------|-------|-----|-----|
| Chlorophyll (μgL$^{-1}$) |                                       |              |       |     |     |
| 2 m                 | 2.04                                 | 0.04 to 4.04 | 2.009 | 167 | 0.046 |
| 6 m                 | 8.54                                 | −15.23 to 32.31 | 0.712 | 115 | 0.478 |
| Dissolved oxygen (mgL$^{-1}$) |                                   |              |       |     |     |
| 2 m                 | −0.35                                | −0.52 to −0.18 | −4.035 | 159 | 0.000 |
| 6 m                 | −0.53                                | −0.98 to −0.09 | −2.386 | 121 | 0.019 |
| 9.6 m               | −0.77                                | −1.24 to −0.30 | −3.243 | 149 | 0.001 |
| 16 m                | −0.78                                | −1.22 to −0.34 | −3.496 | 161 | 0.001 |
| Temperature (°C)    |                                       |              |       |     |     |
| 1 m                 | 1.18                                 | 0.65 to 1.71 | 4.370 | 160 | 0.000 |
| 3 m                 | 1.01                                 | 0.49 to 1.54 | 3.795 | 160 | 0.000 |
| 5 m                 | 1.01                                 | 0.48 to 1.53 | 3.800 | 161 | 0.000 |
| 7 m                 | 1.19                                 | 0.64 to 1.74 | 4.281 | 146 | 0.000 |
| 9 m                 | 0.74                                 | 0.29 to 1.18 | 3.282 | 160 | 0.001 |
| 11 m                | 0.43                                 | 0.16 to 0.71 | 3.115 | 160 | 0.002 |
| 13 m                | 0.43                                 | 0.28 to 0.58 | 5.591 | 161 | 0.000 |
| 15 m                | 0.46                                 | 0.34 to 0.57 | 7.930 | 161 | 0.000 |
| 17 m                | 0.44                                 | 0.32 to 0.56 | 7.222 | 137 | 0.000 |
| 19 m                | 0.46                                 | 0.35 to 0.57 | 8.192 | 161 | 0.000 |
we use a paired t-test to demonstrate that the measurements from the different temperature sensors were statistically equivalent in the epilimnion, where the temperature gradient was small, and thus no sensor lag was to be expected. While this does not constitute definitive proof, we consider it sufficient evidence of sensor equivalence between the buoy and the manual probe temperature sensors. DO sensors, the other WQ parameter that the automated data buoy and the manual Aqua Troll 400 have in common, on the other hand, are notoriously fickle and require regular check calibration. While the Aqua Troll DO sensors are calibrated weekly, not knowing the detailed history of how all the DO sensors had been treated and calibrated during the sampling period, this analysis was limited to temperature data.

Data and methods

For our 4-yr automated buoy sampling period (see the Results section) on Great Pond, we have 351,306 valid automatic temperature measurements over 422 d and 14,260 valid manual temperature samples taken on 51 d. Supplementary Table S1 shows descriptive statistics for the fully automated and manually collected temperature data set. The data used in this analysis are historical. No contemporary checks were conducted to assure that the temperature sensors on the automated buoy measure the same temperature as the sensor on the manual Aqua-Troll 400. However, based on the manufacturer’s declaration that the sensors are factory calibrated, do not drift, and require no periodic recalibration, we assume the sensors measure the same temperature value. Furthermore, seeing that the boat used during manual sampling is attached to the automated data buoy, it is assumed that the different sensors are colocated during the manual cast and subject to the same conditions during sampling. The Comparison of matched measurements section details the investigation into whether these assumptions are valid. The Sensor lag correction section details investigation into sensor lag and why, at depths below 9 m, some of our assumptions may be violated.

To investigate whether the manual sensor probe indeed measures the same temperatures as the automatic data buoy sensors, we matched these two data sources according to time and depth. We found 46 d with both automatic and manual sampling on Great Pond, constituting the data set for the analysis of matched pairs. Note that the raw temperature data were collected at depths \( D = \{ 1, 3, 5, 7, 9, 11, 13, 15, 17, 19 \} \) (meters) for the automatic measurements and continuously from depth \( z \in [0,20] \) (meters) for the manual measurements. Hence, to pair the automatic and manual data, the two data sources required harmonization, to which end the manual measurements were aggregated according to their proximity to the automatic depths \( D \).

The automatic measurements (superscript \((a)) were aggregated in time, using all temperature measurements from each day, so that the average automatic temperature measurement for day \( t \) at depth \( d \in D \) was

\[
T_{a(t),d} = \sum_{n=1}^{N_{a(t),d}} \frac{T_{a(t),n},d}{N_{a(t),d}}
\]

where \( N_{a(t),d} \) is the number of automatic measurements on day \( t \) at depth \( d \). We averaged the automatic measurements using all measurements from each day, as opposed to using only daytime measurements, since it was found that the within-day variability...
water temperatures did not vary greatly. Moreover, using only daytime measurements did not produce different results from those presented. Supporting Information Figure S1 shows that all day (24 h) averaged temperature data are perfectly correlated ($R^2 = 1, p < 0.001$) and operationally equivalent to working day (08:00–16:00, 8 h) averaged data.

Adversely, the manual measurements (superscript (m)) were collected continuously at various depths $z$ as the probe was lowered. To match these measurements with the fixed, discrete depths of the automatic buoy, we aggregated the manual measurements according to their distance to sampling depths of the automatic measurements, giving more influence to manual samples closer to the automatic buoy depths $D$, to obtain distance-weighted average manual temperatures

$$T_{t,d}^{(m)} = \sum_{n=1}^{N_{t,d}^{(m)}} w_{n,t,d}^{(m)} T_{n,t,d}^{(m)} / \sum_{n=1}^{N_{t,d}^{(m)}} w_{n,t,d}^{(m)}$$ (4)

with weights $w_{n,t,d}^{(m)} = 1 - |d_z - z_n|$, where $|d_z - z_n|$ is the distance between the depth of the manual measurement $z_n$ and the nearest automatic buoy depth $d_z$, and $N_{t,d}^{(m)}$ is the number of manual measurements within 1 m of the buoy. Figure 4 depicts the paired automatic and manual aggregates, while summary statistics can be found in Supplementary Table S1.

Inspection of the matching data (see the Comparison of matched measurements section for statistical validation) revealed that both the automatic and manual data exhibited some manner of bimodal distribution (Fig. 5, upper panel). We successfully separated the two distributions (Fig. 5, lower panels) into distinct, unimodal distributions by filtering the matching data according to depth. Specifically, we separated data points from the epilimnion (1–7 m) from the layers below, that is, the thermocline and hypolimnion (9–19 m). This bimodal, depth-dependent distribution is the same as identified in the Results section above. In the following

---

**Fig. 5.** Density histograms and kernel density estimates (bandwidth [bw] of the kernel density estimate = 1) showing the distributions of paired automatic temperature measurements (blue) and manual temperature measurements (red), respectively. The top plot shows all the paired temperature data, while the bottom two plots show the data from depths in the epilimnion and below, respectively.
analyses, we consider all three cases, namely (1) the entire data set, (2) data from the epilimnion, and (3) data from the thermocline and hypolimnion regions.

### Results

In this section, we (1) analyze if the automated buoy and the manual temperature sensors measure the same water temperature values (see the Comparison of matched measurements section) and investigate the bimodal depth effect which we ascribe to sensor lag, (2) quantify and correct for the sensor lag (see the Sensor lag correction section) and, (3) analyze the water temperature data sets measured by automated and, sensor-lag corrected, manual means (see the Weather bias estimation section). Using both frequentist and Bayesian statistics, we achieve distinct statistical analyses which constitute robust methods for comparing manual and automated temperature data. A clear difference in overall water temperature is shown between manual and automated collection methods. We attribute these differences to weather bias on the collection of manual WQ samples, confirming the earlier analysis in the Effect of weather sampling bias on WQ parameters using only automated buoy measurements section.

#### Comparison of matched measurements

We tested for a difference between the matched manual and automatic temperature samples using the two methods of a paired student $t$-test (Ross 2021) and an analogous Bayesian analysis for paired samples (Rouder et al. 2009). To evaluate the significance of the Bayesian results, we report the Region of Practical Equivalence (ROPE), which is broadly the Bayesian equivalent of a $p$-value (Kruschke 2018). Additionally, we compute the discrete Jensen–Shannon (JS) divergence between manual and automatic samples, using a bin width of 1°C. The JS divergence is a symmetric measure of the divergence (or “dissimilarity”) between two distributions (Lin 1991).

When comparing all data, it did not appear that automatic and manual sampling measured the same ($t[435] = 14.65, p < 0.001, \mu_{\text{diff}} = 0.73$°C). However, when filtered according to depth, we discovered that for epilimnion depths (1–7 m), the two sampling methods agreed ($t[176] = 1.37, p = 0.173, \mu_{\text{diff}} = 0.05$°C), while for thermocline and hypolimnion depths (9–19 m), they disagreed ($t[258] = 18.18, p < 0.001, \mu_{\text{diff}} = 1.19$°C). Both the Bayesian posterior estimates and the JS divergences confirmed this impression. Please refer to Supplementary Table S2 for further details. We hypothesize that the effect in the hypolimnion can be attributed to sensor-lag related to the manual sensor experiencing greater temperature variability as it passes through the thermocline. This is despite the manual sampling procedure to slowly lower the Aqua-Troll sensor to help eliminate sensor lag. The manual probe was, on average, found to measure higher temperatures than the automated buoy; as it is lowered from generally warmer water, through the thermocline layer where temperature gradients are larger, to cooler water, this temperature increase is attributed to the sensor response time being insufficient to compensate. The violin plot (Fig. 6) illustrates this phenomenon.

#### Sensor lag correction

Having identified a temperature sensor lag that is predominant in the thermocline/hypolimnion this section quantifies and corrects for the lag throughout the entire water column of the Great Pond data. To isolate the weather bias from the identified sensor-lag bias, we first computed the expected manual sensor-lag bias for each automatic sampling depth $d \in D$, using the paired measurements ($N = 436$) from the previous analysis. The resulting sensor-lag biases and relevant statistics are shown in Supplementary Table S3. We then corrected every manual temperature measurement ($N = 14,260$) using the sensor-lag bias corresponding to the nearest automatic buoy depth. We refer to these data as sensor-lag corrected manual measurements, and use them in place of the raw measurements in the analysis below.

#### Weather bias estimation

In this section, having corrected for any sensor lag, we compare all automated and manual water temperature measurements to show the effect of the weather bias in the WQ data from Great Pond. In line with the paired, aggregated data, we found a bimodal distribution in the raw, unaggregated data, which persisted after correcting for the sensor lag bias (Fig. 7). Hence, we filtered the unaggregated data using the
Table 4. Two-sample $t$-tests, Bayesian posteriors and JS divergences comparing unaggregated automatic and sensor-lag corrected manual samples. $\mu_{\text{man}} - \mu_{\text{aut}}$ is the difference between the overall automated and the manual temperature populations and in effect represents the overall average difference of manual compared to automated temperature sampling which we have ascribed to a weather bias effect.

| Depth  | $\mu_{\text{man}} - \mu_{\text{aut}}$ (°C) | 95% CI (°C) | t   | df   | p-value |
|--------|----------------------------------|-------------|-----|------|---------|
| 0–20 m | 1.01                             | 0.92–1.10   | 21.46 | 15,474.0 | <0.001  |
| 0–8 m  | 0.98                             | 0.90–1.07   | 22.49 | 7256.9 | <0.001  |
| 8–20 m | 0.53                             | 0.47–0.60   | 16.07 | 8636.9 | <0.001  |

Bayesian posterior

| Depth  | $\mu_{\text{man}} - \mu_{\text{aut}}$ (°C) | 95% HDI (°C) | ROPE | JS divergence |
|--------|----------------------------------|-------------|------|---------------|
| 0–20 m | 1.01                             | 0.92–1.10   | <0.001 | 0.032        |
| 0–8 m  | 0.98                             | 0.86–1.09   | <0.001 | 0.044        |
| 8–20 m | 0.53                             | 0.47–0.59   | <0.001 | 0.044        |

Fig. 7. Density histograms and kernel density estimates (bw = 1) showing the distributions of the raw, unaggregated automatic temperature measurements (blue) and sensor-lag corrected manual temperature measurements (red), respectively. The top plot shows all the data, while the bottom two plots show the data from depths above and below 7 m, respectively.
same criteria, and treated data from the epilimnion and the hypolimnion separately.

Using both a two-sample $t$-test (Ross 2021) and the analogous Bayesian analysis for two-sample designs (Rouder et al. 2009), we estimate the difference in means between the whole population of the sensor-lag corrected manual and automatic temperature samples as shown in Table 4. Both the $t$-tests and the Bayesian analyses show that there is a difference in the means for all three subsets. For epilimnion depths 0–8 m, the two-sample $t$-test predicts a significant difference of the means of 1.0°C ($t[7256.9] = 22.49, p < 0.001$). For hypolimnion depths 8–20 m, the two-sample $t$-test predicts a significant difference of the means of 0.5°C ($t[8636.9] = 16.07, p < 0.001$). These findings are confirmed by the Bayesian analysis for two-sample designs. We hypothesize that this smaller difference at greater depths can be attributed to the higher lake temperature stability as the depth increases.

As established in the Influence of weather on manual sampling section, “good weather” (high air temperature, low rainfall, and low wind speed) results in a higher probability of manual sampling taking place (Fig. 3). Conversely, “bad weather” (low air temperature, high rainfall, and high wind speed) has a lower probability of manual sampling taking place. “Good weather” is associated with typically higher lake water temperatures and “bad weather” with typically lower lake water temperatures. While further validation on other lakes is required, this result indicates that weather-biased manual water temperature measurements will tend to be higher than their corresponding automated temperature measurements. These findings imply that manual sampling, typically subject to weather bias, will tend to record higher lake water temperatures on average than the “true” average lake water temperature. In this case study, this temperature difference was found to be between 1.0°C and 0.5°C, depending upon depth.

Table 5. Examples of the RMSE temperature values of four lake models taken from recent literature (Dissanayake et al. 2019; Baracchini et al. 2020; Gasca-Ortiz et al. 2020; Man et al. 2021). None of these studies accounted for the effect of weather bias on their manually sampled WQ data. If our findings are representative of these studies and the water temperature weather bias was accounted for, then this would very significantly alter the reported lake water temperature values and/or the RMSE values reported.

| Date  | Data collection method                  | Model         | RMSE (°C) | Temperature sensor             | Accuracy/resolution (°C) |
|-------|----------------------------------------|---------------|-----------|--------------------------------|-------------------------|
| 2021  | Manual CTD profilers                   | SI3D-AED3     | 2.28      | YSI: ProODO                    | ±0.2/not given           |
|       |                                        |               |           | Seacat: SBE 19                 | ±0.01/±0.0001            |
| 2020  | Manual profilers                        | Delft3D       | 1.9       | Unnamed CTD                    | ±0.002/not given         |
| 2020  | Manual CTD Auto thermistor chains      | Delft3D       | 0.82      | Unnamed HOBO logger            | ±0.2/0.024 (typical)     |
|       |                                        |               |           | RBR: XR-620                    | ±0.002/max 0.00005       |
| 2019  | Manual profilers                        | Delft3D       | 1.25      | RBR: TR-1060                   | ±0.002/max 0.00005       |
Comparison of the estimated weather biases

Using two different data sets and both frequentist and Bayesian statistics, we have robustly shown that manual sampling records statistically significantly different lake WQ values as compared to continuous, automated sampling. We have demonstrated this bias for water temperature, and, albeit with less independent validation, for DO and chlorophyll. As shown in Fig. 8, two different approaches for estimating the water temperature bias arrive at broadly the same result. In this figure, the close alignment of the actual and emulated manual water temperature data can be seen, indicating measurement equivalence of the manual and automatic sensors. However, this only holds true in the epilimnion—from 9 m depth and deeper, we see a sensor lag effect in the actual manual measurements which disappears in sensor lag corrected measurements. Importantly, all of these manual measurements are markedly offset from the automated temperature values, visualizing the weather bias in manual sampling.

Conclusions

Our case study of the lake monitoring program across seven lakes in Maine, USA, has revealed a manual sampling weather bias. This weather bias, based on air temperature, rainfall intensity, and wind speed, was found to have a statistically significant effect on the probability of manual WQ sampling taking place. In simple terms, “good” weather promotes manual WQ sampling while “bad” weather inhibits it.

Using measurements from a single automated data buoy and emulating manual sampling data by filtering by the day and time of actual manual sampling, we have shown a measurable effect of this weather bias on the water temperature, DO, and to some extent, chlorophyll measurements. Compared to the continuous buoy measurements, the emulated manual measurements were found to be, on average, statistically significantly different and consequently biased. These mean effects ranged, depending upon depth, between 0.4°C and 1.2°C higher for water temperature, between −0.8 and −0.4 mgL⁻¹ lower for DO, and 2.0 μgL⁻¹ higher for chlorophyll.

Despite the monitoring programs procedure to slowly lower the manual probe to eliminate it, sensor lag was still apparent in the manually collected data. This finding is confirmed by both t-test and Bayesian analysis. In the epilimnion (0–8 m), the automated and manual temperature measurements align. However, in the thermocline and hypolimnion (8–20 m), there is an offset between manual and automated measurements (Fig. 6) which we ascribe to the average effect of the manual temperature probe displaying sensor lag transiting the thermocline layer. This sensor lag effect was calculated and used as a correction factor in the hypolimnion layer. This finding reiterates the care required when using any manual collected data to examine and correct for any sensor lag effects, regardless of the manual collection procedures used to eliminate such a lag.

To reinforce the emulated data bias analysis findings, the effects of the weather bias have been demonstrated by comparing actual manually collected water temperature with that collected by an automated data buoy. In the epilimnion (1–7 m), manual methods measure, on average, 1.0°C higher temperatures than measured by automated methods. In the hypolimnion (8–20 m), once the effect of thermocline/sensor lag has been removed, manual methods measure, on average, 0.5°C higher temperatures than measured by automated methods. The difference between these depth-reliant values can be explained by the increased stability of water temperature (and DO) at lower depths, and their concomitant lower susceptibility to weather bias.

The observed influence of weather-induced sampling bias on lake water temperature, DO, and chlorophyll strongly suggests that this effect may be a general finding applicable to other lakes. More manual sampling programs need examining; however, if this finding is representative, calculating and accounting for weather bias when using manually obtained WQ sampling data should become standard practice.

The effect of weather-induced manual sampling bias is expected to affect other manually measured WQ parameters other than water temperature, chlorophyll, and DO. Further research is required to examine this hypothesis. The literature, for example British Standards (2006), recommends that manual sampling programs, looking for systematic changes over time, be conducted with regular periodicity to avoid detecting variations in WQ that are not of interest. While not specifically concerned with our novel weather bias findings, adherence to such guidelines through strict regular WQ measurement periodicity, would, it is believed, go some way to addressing the effect of the manual sampling weather bias identified. It may be possible to account for any weather bias developed by a manual monitoring program through examination of weather records and any deviations from a strict periodicity of measurement. Ultimately, the use of automated WQ data collection is the solution to the issue of weather bias reported in the paper.

The identified weather bias promotes manual WQ sampling taking place during warmer, dryer, and less windy weather conditions. Conversely, it reduces the probability of manual WQ sampling taking place during periods of colder, wetter, and windier weather. In short, the weather bias promotes manual sampling during “good” weather and reduces manual sampling in “bad” weather, which in turn results in sampling overall warmer water temperatures. Warmer water holds less DO and encourages algal growth and therefore higher chlorophyll; as it is hypothesized that the weather bias relationship observed for water temperature will generally hold true for other WQ parameters, corresponding assessment of overall WQ and ecosystem health will be affected.
The magnitude of errors in lake water temperature due to the manual collection weather-induced bias are similar to typical lake model root mean square errors (RMSE) reported in the literature (Table 5). If our weather bias temperature values were applied to these models, they would significantly alter the lake model temperature inputs with corresponding changes to the lake model temperature outputs and/or the RMSE values. That these weather bias temperature values, have similar magnitudes to model RMSE values, and demonstrate the significance of our findings.

To further illustrate the significance of our findings, when examining the warming of North Eastern, North American lakes due to climate change, Richardson et al. (2017) report a mean warming of lake surface temperatures of 0.05°C yr⁻¹. Much of the data used by Richardson et al. (2017) will inevitably have been obtained through manual sampling. If our case study finding is representative of other manual monitoring programs, then a potential weather bias in the epilimnion of the lake in the State of Maine (State of Maine Legislature 2021).

Our main takeaway conclusion is that users of lake WQ data obtained from manual collection programs should be aware that they may contain weather-induced data bias. It is advised that records of the manual WQ data collection timings and relevant weather records be examined to determine if the presence of weather bias is likely. The methodologies presented in this paper can be used to assessed and correct for such WQ weather-induced biases.

**Data availability statement**

The analysis data and code are available at [https://github.com/MirjamOdile/weather-bias-in-water-quality-data](https://github.com/MirjamOdile/weather-bias-in-water-quality-data).

**References**

Achen, C. H. 2000. Why lagged dependent variables can suppress the explanatory power of other independent variables. Presented at the Annual Meeting of Political Methodology.

Austin, J. 2012. Resolving a persistent offshore surface temperature maximum in lake superior using an autonomous underwater glider. Aquat. Ecosyst. Health Manage. 15: 316–321. doi:10.1080/14634988.2012.711212

Ban, X., Q. Wu, B. Pan, Y. Du, and Q. Feng. 2014. Application of composite water quality identification index on the water quality evaluation in spatial and temporal variations: A case study in Honghu lake, China. Environ. Monit. Assess. 186: 4237–4247. doi:10.1007/s10661-014-3694-9

Baracchini, T., S. Hummel, M. Verlaan, A. Cimatoribus, A. Wüest, and D. Bouffard. 2020. An automated calibration framework and open source tools for 3d lake hydrodynamic models. Environ Model Softw. 134: 104787. doi:10.1016/j.envsoft.2020.104787

Berthouex, P. M., and L. C. Brown. 2002. Statistics for environmental engineers, 2nd Edition. Lewis Publishers. doi:10.2113/8.3.244

British Standards. 2006. BS EN ISO 5667-1: Water quality—sampling—part 1: Guidance on the design of sampling programmes and sampling techniques, BS EN ISO 5667-1.

Dhanak, M. R., and N. I. Xiros. 2016. Springer handbook of ocean engineering, 1st Edition. Springer International Publishing.

Dissanayake, P., H. Hofmann, and F. Peeters. 2019. Comparison of results from two 3d hydrodynamic models with field data: Internal seiches and horizontal currents. Inland Waters 9: 239–260. doi:10.1080/20442041.2019.1580079

Dunsmuir, W. T. 2015. Generalized linear autoregressive moving average models, p. 51–76. In R. A. Davis, S. H. Holan, R. Lund, and N. Ravishanker [eds.], Handbook of discrete-valued time series. CRC Press.

Dunsmuir, W. T. M., and D. J. Scott. 2015. The glarma package for observation-driven time series regression of counts. J. Stat. Softw. 67: 1–36.

Garson, G. D. 2016. Logistic regression: Binomial and multinomial, 2016 Edition. Statistical Associates Publishing.

Gasca-Ortiz, T., D. A. Pantoja, A. Filonov, F. Domínguez-Mota, and J. Alcocer. 2020. Numerical and observational analysis of the hydro-dynamical variability in a small lake: The case of lake zirahuén, México. Water (Basel) 12: 1658.

Hosmer, D. W., S. Lemeshow, and R. X. Sturdivant. 2013. Applied logistic regression, Wiley series in probability and statistics, 3rd Edition. Wiley. doi:10.1002/9781118548387

Keele, L., and N. J. Kelly. 2006. Dynamic models for dynamic theories: The ins and outs of lagged dependent variables. Polit. Anal. 14: 186–205. doi:10.1093/pan/mpj006

Kruschke, J. K. 2018. Rejecting or accepting parameter values in bayesian estimation. Adv. Methods Pract. Psychol. Sci. 1: 270–280. doi:10.1177/2515245918771304

Lin, J. 1991. Divergence measures based on the Shannon entropy. IEEE Trans. Inf. Theory 37: 145–151. doi:10.1109/18.61115

Man, X., C. Lei, C. C. Carey, and J. C. Little. 2021. Relative performance of 1-d versus 3-d hydrodynamic, water-quality models for predicting water temperature and oxygen in a shallow, eutrophic, managed reservoir. Water (Basel) 13: 88. doi:10.3390/w13010088

McFadden, D. 1979. Quantitative methods for analysing travel behaviour of individuals. In: D. Hensher and P. Stopher, (Eds.), Behavioural Travel Modelling, Croom Helm, pp. 279–318. doi:10.4324/97810030156055

Menard, S. 2010. Logistic regression: From introductory to advanced concepts and applications. SAGE Publications. doi:10.4135/9781483348964
Metcalfe, A. V., and P. S. Cowpertwait. 2009. Introductory time series with R. Springer.

Nyberg, H. 2010. Dynamic probit models and financial variables in recession forecasting. J. Forecast. 29: 215–230. doi: 10.1002/for.1161

Quinn, G. P., and M. J. Keough. 2002. Experimental design and data analysis for biologists. Cambridge Univ. Press. doi: 10.3390/w9060442

Richardson, C. D., and others. 2017. Transparency, geomorphology and mixing regime explain variability in trends in lake temperature and stratification across northeastern north america (1975–2014). Water (Basel) 9: 442.

Ross, S. M. 2021. Introduction to probability and statistics for engineers and scientists, 6th Edition. Academic Press.

Rouder, J. N., P. L. Speckman, D. Sun, R. D. Morey, and G. Iverson. 2009. Bayesian t tests for accepting and rejecting the null hypothesis. Psychon. Bull. Rev. 16: 225–237. doi: 10.3758/PBR.16.2.225

Stanley, E. H., S. M. Collins, N. R. Lottig, S. K. Oliver, K. E. Webster, K. S. Cheruvelil, and P. A. Soranno. 2019. Biases in lake water quality sampling and implications for macroscale research. Limnol. Oceanogr. 64: 1572–1585. doi: 10.1002/lno.11136

State of Maine Legislature, 2021. Title 38, Chapter 3, Standards for classification of fresh surface waters. §465.

Webster, P. 1997. Instruction manual for baseline water quality sampling. State of Maine Department of Environmental Protection.

Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council in the UK via grant EP/L016214/1 awarded for the Water Informatics: Science and Engineering (WISE) Centre for Doctoral Training and by the Economic and Social Research Council in the UK via grant ES/P000630/1 awarded for the South West Doctoral Training Partnership (SWDTP). Both of these funding sources are gratefully acknowledged.

Submitted 01 November 2021
Revised 02 March 2022
Accepted 11 March 2022

Associate editor: John P. Smol