Drone-based imaging to assess the microbial water quality in an irrigation pond: A pilot study

B.J. Morgan, M.D. Stocker, J. Valdes-Abellan, M.S. Kim, Y. Pachepsky

PII: S0048-9697(19)35752-3

DOI: https://doi.org/10.1016/j.scitotenv.2019.135757

Reference: STOTEN 135757

To appear in: *Science of the Total Environment*

Received date: 23 September 2019

Revised date: 22 November 2019

Accepted date: 24 November 2019

Please cite this article as: B.J. Morgan, M.D. Stocker, J. Valdes-Abellan, et al., Drone-based imaging to assess the microbial water quality in an irrigation pond: A pilot study, *Science of the Total Environment* (2019), https://doi.org/10.1016/j.scitotenv.2019.135757

Drone-based imaging to assess the microbial water quality in an irrigation pond: a pilot study

Morgan, B. J.,[1] Stocker, M. D.,[1] Valdes-Abellan, J.,[2] Kim., M. S.,[1] Pachepsky, Y.[1]

[1]USDA-ARS Environmental Microbial and Food Safety Laboratory, Beltsville, MD

[2]Department of Civil Engineering, University of Alicante, Alicante, Spain

Corresponding author: email yakov.pachepsky@usda.gov, phone +1-301-504-7468

ABSTRACT

Microbial water quality datasets are essential in irrigated agricultural practices to detect and inform measures to prevent the contamination of produce. *Escherichia coli* concentrations are commonly used to evaluate microbial water quality. Remote sensing imagery has been successfully used to retrieve several water quality parameters that can be determinants of *E. coli* habitats in waterbodies. In this study, a pilot study was conducted to test the possibility of using imagery from a small unmanned aerial vehicle (sUAV or drone) to improve the estimation of microbial water quality in small irrigation ponds. *In situ* measurements of pH, turbidity, specific conductance, and concentrations of dissolved oxygen, chlorophyll-*a*, phycocyanin, and fluorescent dissolved organic matter were taken at depths of 0–15 cm in 23 locations across a pond in Central Maryland, USA. The pond surface was concurrently imaged using a drone with three modified GoPro cameras, and a multispectral MicaSense RedEdge camera with five spectral bands. The GoPro imagery was decomposed into red, blue, and green components. Mean digital numbers for 1-m radius areas in the images were combined with the water quality data to provide input for a regression tree-based analysis. The accuracy of the regression-tree data description with "only imagery" inputs was the same or better than that of trees constructed with "only water-quality parameters" as inputs. From multiple cross-validation runs with "only imagery" inputs for the regression trees, the average (± SD) determination coefficient and root-mean-squared error of the decimal logarithm of *E. coli* concentrations were 0.793 ± 0.035 and 0.131 ± 0.011, respectively. The results of this study demonstrate the opportunities for using sUAV imagery for obtaining a more accurate delineation of the spatial variation of *E. coli* concentrations in irrigation ponds.

Keywords: UAV; water quality; *E. coli*; irrigation pond; pond habitats; remote sensing

1. Introduction

Microbial water quality is a common focus of monitoring because of the multiple waterborne illnesses affecting human and animal life. Pathogenic *Escherichia coli*-related cases in the USA cost more than $270 million in 2013 alone in terms of productivity losses and medical costs (Economic Research Service ERS-USDA, 2014). Much larger annual loss estimates have been produced for other pathogens (Pachepsky et al., 2011; Scharff, 2010) including $3.6 billion for salmonellosis and $39 billion for all foodborne microbial illnesses in the USA. In response to these risks, research interest in microbial water quality is steadily growing (Alegbeleye et al., 2018; Hellberg and Chu, 2016; Jongman and Korsten, 2018; Olaimat and Holley, 2012; Pachepsky et al., 2018; Rochelle-Newall et al., 2015). Monitoring the microbial quality of irrigation water is an essential step to further improve food safety by informing the application of strategies for minimizing produce contamination (Pachepsky et al., 2011).

E. coli, a common inhabitant of the intestinal tract of warm-blooded animals, is the most commonly used indicator of microbial water quality. Elevated *E. coli* concentrations in water indicate the possible presence of enteric pathogens (Environmental Protection Agency (EPA), 2012). Most regulations throughout the world consider *E.coli* as a useful indicator from which to infer the microbial quality of freshwater irrespective of its use, e.g., recreation or irrigation, or its source, e.g., natural or treated waste-water (Boletin Oficial del Legislación Consolidada, 2007; European Parliament, 2006; Kay et al., 2004; USA Congress, 2011; World Health Organization (WHO), 2006).

E. coli growth and death rates are typically determined by site-specific environmental conditions and how the microorganism adapts to these conditions via its gene expression patterns

(Van Elsas et al., 2011). The environmental controls affecting *E. coli* survival are numerous. Carbon substrate is a key factor in aquatic environments and the lack, with its absence reducing the likelihood of survival and/or promoting dormancy of the bacteria (Franz et al., 2005; Semenov et al., 2008). Other factors such as temperature (both fluctuations and high temperatures have a negative impact on *E. coli* survival), pH (low pH impacts survival negatively and triggers acid-resistance mechanisms in *E.coli*), the status and population structure of planktonic algae (which may affect *E. coli* survival both positively and negatively), the biodiversity in the aquatic environment (with high biodiversity hindering colonization by *E. coli*), and the presence of periphyton and algae mats (which harbor bacterial populations) have been reported to have a strong influence on the fate of *E. coli* in open environments (Ansa et al., 2012; Byappanahalli et al., 2009; Carr et al., 2005; Choi et al., 2000; Semenov et al., 2007).

High spatial variability can lead to substantial uncertainty in *E. coli* concentration estimates in water bodies. For example, Quilliam et al. (2011) reported that different population levels were estimated based on *E. coli* concentrations obtained from two opposite banks of the same river. Pachepsky et al. (2018) detected a stable spatial pattern of *E. coli* concentrations over time in two artificial ponds, with lower concentrations occurring in the pond interiors. Microbial monitoring of surface water sources must, therefore, account for the spatial variability in microbial concentrations; however, study of the spatial variation of microbial concentrations remains a resource-intensive challenge.

The use of remote sensing (RS) imagery from satellites, airplanes, and small unmanned aerial vehicles (sUAVs) offers opportunities to bypass—or at least reduce—the time and budget constraints associated with water quality monitoring performed using grab sampling. In the last few years, RS imagery has been applied in both coastal (Constantin et al., 2017) and inland

environments (Giardino et al., 2014) to infer water quality parameters including turbidity (Duan et al., 2019; Isidro et al., 2018), chlorophyll-*a* (Wang et al., 2018), and algal population attributes (Boddula et al., 2017). A recent review on using sUAVs for algal bloom research (Kislik et al., 2019) demonstrates the variety of approaches that can be adopted with respect to data processing and accounting for weather conditions, and the range of camera and sensor types available.

It has not yet been possible to detect indicator or pathogenic bacteria in water bodies using sUAV imagery. However, given that it is now possible to use RS products to retrieve water quality parameters that may directly or indirectly characterize and control *E. coli* habitats, it seems plausible that the spatial distribution of *E. coli* can be inferred by (a) estimating water quality habitat parameters from the sUAV imagery and (b) using empirically established relationships between these water quality habitat parameters and *E. coli* concentrations. Information from different wavelength ranges of sUAV imagery may also provide an integrative characterization of *E. coli* habitats in addition to providing spatial information on the distribution of *E. coli* without the requirement of *in situ* water quality measurements.

The objective of this work was to evaluate the usability of imagery obtained using a sUAV for improving the estimation of the microbial water quality of small ponds both with and without additional information on the water quality parameters affecting the suitability of aquatic habitats for *E. coli*.

2. Material and methods

2.1 Study site and water quality parameters

The study embankment pond was located on a private farm and, at the time of study, had an area of approximately 3780 m^2. Water is the pond is used exclusively for the irrigation of the

surrounding farmland (Fig. 1) during the summer. The average depth of the pond was 2.7 m. Runoff from the surrounding farmland can enter the pond, and while chemical fertilizers are used on the land, animal manures are not applied. The farm operator decides the irrigation regime, and potential sources of *E. coli* in the pond include wildlife, a creek feeding into the pond, transfers from another nearby pond, and occasional bathers.

Twenty-three sampling locations were established across the surface of the pond in an approximate regular pattern. Twelve of the sampling points were located near to the shore and the other 11 were located in the pond interior (Fig. 1). Water samples were collected at a depth of 0–15 cm over a two-hour period on July 18 and August 1, 2017. Samples from the interior points were obtained using a kayak and the nearshore sampling was done with a 500-mL plastic dipper with a 1.5-m long handle. The water samples were placed on ice shortly after collection and transported to the laboratory before being processed for *E. coli* enumeration within two hours of collection. Measurements of environmental covariates were taken *in situ* and in conjunction with water samples using a handheld YSI Exo-2Sonde (YSI Inc, Yellow Springs, OH) and included temperature (°C), dissolved oxygen (DO, mg L^{-1}), pH, specific conductance (SPC, μS cm^{-1}), fluorescent dissolved organic matter (fDOM, in relative fluorescence units, RFU), cyanobacteria phycocyanin (BGAPC, RFU), *in situ* chlorophyll-*a* content (CHLAF, RFU), and turbidity (TURB, Nephelometric Turbidity Units, NTU). Precipitation and solar radiation data were obtained from a local weather station located less than 3 km from the pond.

Samples collected in the field were immediately placed in an insulated and ice-chilled container. *E.coli* concentrations were determined in the laboratory within 24 h of collection in accordance with EPA standard methods (2002) using membrane thermotolerant *E. coli* (mTEC) agar (Difco, Sparks, MD). *In vitro* determination of chlorophyll-*a* (CHLAS) according to EPA

standard methods (Arar and Collins, 1997) was also performed. Statistical analyses were performed using the R package for statistical computing (R Core Team, 2014) using decimal logarithms of *E. coli* concentrations in CFU $(100\ ml)^{-1}$ (LGCFU).

2.2 Imagery acquisition and processing

The imagery was obtained with a RedEdge-M (MicaSense, Seattle, WA) camera and GoPro cameras (GoPro, San Mateo, CA) mounted on a 3DRSolo® (3DR, Berkeley, CA) drone. The RedEdge-M camera collects reflectance information in the five following narrow wavelength bands (Fig. S1): blue (MS1, centered at 475 nm with a 20 nm bandwidth), green (MS2, centered at 560 nm with a 20 nm bandwidth), red (MS3, centered at 668 nm with a 10 nm bandwidth), red edge (MS4, centered at 717 nm with a 10 nm bandwidth), and near-infrared (MS5, centered at 840 nm with a 40 nm bandwidth). This camera provided five images for each sampling location. Additionally, the drone was flown with three GoPro cameras (Stuntcams Inc., Ada, MI) with lenses modified to reduce distortion and provide a wavelength range wider than the visible spectrum. The first GoPro camera captured the entire range (ER) of wavelengths that included the visible and infrared (IR) range of the spectrum. The second and the third GoPro cameras had filters that provided the visible (VI) only and the IR only ranges of the spectrum, respectively. All flights were done in immediate succession. The spectral analysis for the GoPro lens and filter can be seen in Fig. S2. The images obtained by each of the GoPro cameras were demosaiced (Kimmel, (1999) into red (ER1, ER2, and ER3), green (IR1, IR2, IR3), and blue (VI1, VI2, and VI3) components, respectively. This splitting created nine layers from the GoPro imagery data (i.e., RGB for the three cameras) for each sampling day.

The UAV flight altitude was 400 feet (130 m) on both dates. The sky was clear, and solar radiation was very similar on both sampling days (Fig. S2). The solar zenith angle varied

between 20° and 30° between imaging times. The images were aligned using co-registration to an image of the pond obtained from the United States Geological Survey (USGS) base map using the georeferencing tool in ArcMap 10.5.

To obtain spatial and temporal information from the imagery corresponding to the water quality samples, GPS coordinates acquired during water sampling were imported into ArcMap. A circular 1-m diameter buffer was created around each water sampling location. The nearshore sampling locations might have changed between the sampling days due to fluctuations in the pond depth, changes in vegetation surrounding the pond, and manual image aligning. To account for this, 1-m buffers were created, which were incrementally moved normal to the shoreline towards the interior until they were placed completely inside the water surface area as shown in the image. A Python code was used to facilitate this incremental buffer relocation. After the locations of all buffers (interior and nearshore) were defined, the "clip" tool was used to extract the buffered portions from each of the 28 images and the "summary statistics" tool was used to find the mean digital number (MDN) values in each case. DN distributions inside the clipped areas were normal ($p < 0.05$) after the removal of several outliers, which altered the MDNs by less than 1%. The MDNs of the buffer areas were then used in the analysis along with the water quality data for the same locations.

2.3 Regression trees

Regression trees were used to relate the logarithms of *E. coli* concentrations to the MDNs from all 28 images and the water quality parameters. Regression trees are a machine-learning technique that is particularly useful for establishing the relative importance of a large number of potential predictors and obtaining a transparent predictive or explanatory model (Sorrell et al., 2013). The application of regression trees in environmental sciences is common and has been

employed in several water quality studies (e. g., Jones et al., 2013; Park et al., 2015). In this study, the regression tree technique was applied to assess the suitability of the imagery data for complementing or replacing the water quality data when interring microbial concentrations.

The regression tree algorithm split the data into two groups, then split each of these groups into two subgroups, and then each of these subgroups into two further subgroups, etc. Each split was performed to create the most dissimilar subgroups (with respect to the dependent variable) by maximizing the between-subgroup sum-of-squares as determined using analysis of variance (Therneau and Atkinson, 2018). Splitting was done by sorting the dataset by each of the predictor variables and selecting the value of each predictor variable that provided the two most homogeneous and most dissimilar subgroups. The final split was performed based on a comparison of between-subgroup sum-of-squares for all of the predictors and by selecting the predictor and predictor value that provided the largest between-subgroup sum-of-squares. The splitting process was finished when the subgroups became too small, or when the splitting was no longer effective at creating dissimilar subgroups. This sequential splitting process yields a tree-like structure, where each split creates two branches. Groups that cannot be further split are often called “leaves”. The average of the dependent variable values across all the datasets in a leaf serves as the predicted value for all of the included datasets. The largest increases in the between-subgroup sum-of-squares occur in the early stages of the splitting process, and splitting predictors during these stages are considered to be the most influential. The minimum number of datasets required to perform a split is a control parameter of the algorithm. The optimal minimum number of datasets per branch was set to four, which was determined based on Park et al. (2015).

The output from the regression tree analysis included (a) the root-mean-squared error of the regression, (b) the regression determination coefficient R^2, and (c) the names of the variables that provided the first split and the two subsequent splits. Uncertainty in these outputs was evaluated using the jackknife resampling technique (Kroll et al., 2015). This involved the following steps: (1) data subsamples were created by systematically leaving out a number (*k)* of observations from the original dataset; (2) a regression tree was created using each subsample and the output was recorded; and (3) a histogram for the output values obtained from all of the subsamples was plotted. By applying the jackknife resampling, the risk of obtaining conclusions that are correct based on a single dataset but which might not represent the general pattern across all datasets can be assessed. Here, the entire database included 46 datasets (i.e., 23 sampling locations on two different dates), and each subsample was obtained by removing three observations. The total number of subsamples was 15,180, to which the regression tree technique was applied. After each application, the output variables (i.e., the root-mean-squared error RMSE, R^2, and names of the variables) were recorded. Univariate statistical tests were then applied to identify statistically significant differences in the mean and variance values of RMSE and R^2.

2.4 Data input sets

Each sampling location on each sampling day was characterized by the logarithm of the *E. coli* concentration as the dependent variable and 22 potential predictors including DO, pH, SPC, TURB, fDOM, CHLAF, CHLAS, BGAPC, and 14 image MDNs. The following five input sets were considered: (1) input set 1 included the physicochemical variables DO, SPC, pH, and turbidity; (2) input set 2 consisted of the MDNs for all 14 images, i.e., ER1, ER2, ER3, IR1, IR2, IR3, VI1, VI2, VI3, MS1, MS2, MS3, MS4, and MS5; (3) input set 3 included all the MDNs and

physicochemical variables, i.e., combined input sets 1 and 2; (4) input set 4 included both the physicochemical and biological water quality parameters, consisting of DO, SPC, pH, turbidity, fDOM, BGAPC, CHLAF, and CHLAS; (5) input set 5 combined input set 4 and the imagery parameters from input set 2.

3. Results

3.1. Data overview

Data summaries are presented in the Supplementary Materials. The average water temperature during the observation period on day 1 and 2 was 29.1 ºC and 25.4 ºC, respectively. Summary statistics for water quality parameters on each day are presented in Table S1. All of the water quality parameter distributions were non-normal. The water quality parameters that showed the most distinct differences between the two sampling days were fDOM and SPC; the pond water sampled on the 1st August had relatively low electrolyte concentrations and relatively high concentrations of fDOM on the 18th July. Overall, the imagery and water quality parameters showed more variability on the first sampling date.

Descriptive statistics for water quality parameters and mean digital numbers pooled between sampling days are shown in Table S2. The variability of DO, SPC, and pH was low (coefficient of variation (CV) < 10%). In comparison, the variability of the other water quality parameters, such as total chlorophyll-*a* content and turbidity, was much higher (CV ~ 38%). The highest variability was found for the phycocyanin, chlorophyll-*a*, and fluorescent dissolved organic matter concentrations (CV ranged between 73% and 152%). The logarithm of *E. coli* concentrations had the CV of 17%, and the CV for the MDNs ranged between 10% and 30%, with the exception of the RedEdge-M measurements, which had a CV of 46%.

Pearson correlation coefficients for the two-day dataset are shown in Table 1. For the purposes of comparison, correlations with $|r| > 0.8$, $0.7 < |r| < 0.8$, and $0.6 < |r| < 0.7$ are considered as very strong, strong, and moderate, respectively. There were no strong correlations between the water quality parameters and the mean digital numbers obtained from the imagery with the exception of ER2 MDN, which correlated strongly with both SPC and fDOM. Very strong correlations were found between turbidity, phycocyanin concentrations, and *in situ* chlorophyll-*a* measurements. Strong and very strong correlations were found between GoPro MDNs across the entire spectrum and for the infrared range of wavelengths. RedEdge-M 2 and 3 (blue and red) correlated strongly, RedEdge 5 (infrared) strongly correlated with several other MDNs. No strong correlations were found between the logarithms of *E. coli* concentrations and any of the measured parameters. Additional moderate correlations were found that indicated relationships of potential interest, particularly the correlation between the logarithm of *E. coli* concentration and fDOM. A moderate negative correlation was also found between fDOM and the three GoPro MDNs (ER3, IR1, and VI3).

The regression trees for the logarithms of *E. coli* concentrations are shown in Fig. 2. Five sets of input variables were employed as described in Section 2. The most influential variable in input set 1 was SPC whereas DO and pH were of secondary importance. Values of DO provided the data split at higher SPC values, while lower DO concentrations corresponded to the smallest logarithm values of *E. coli* concentrations. pH was more important when SPC was low, where lower pH defined the data subgroup with the largest logarithm *E. coli* concentrations.

The MDNs of the green component of the image across the entire range (ER2) and the green component of the image from the infrared wavelength range (IR2) were the most influential variables in input sets 2 and 3 (Fig. 2). The green component ER2 across the entire

range was far more influential. Large ER2 values corresponded to lower *E. coli* concentrations. The lowest *E. coli* concentrations (mean logarithm = 1.26) occurred where ER2 was large and the green component of the infrared image was small. The lowest ER2 values corresponded with the highest logarithm concentrations (mean logarithm = 1.92). Combining input sets 1 and 2 into input set 3 resulted in the same regression tree as using input set 2 alone.

Including biological variables in input set 4 resulted in fDOM becoming the most influential variable. Smaller concentrations of *E. coli* were found where the fDOM was large (>2.92). Where fDOM concentrations were lower, and low pH locations had the highest concentrations of *E. coli* (mean logarithm = 1.87). Using input set 5 did not change the regression tree obtained without the imagery data.

3.2 Regression tree performance metrics

Each of the five input variable sets was used in the resampling procedure to evaluate the uncertainty in the regression trees and the selection of the influential variables. For each of the input sets, 7% of the samples were removed and the regression trees were built for the remaining 93% of the data. The total number of evaluated regression trees was 15,180. Histograms of the determination coefficients and the root-mean-square errors are shown in Fig. 3. The average (± standard deviation, SD) determination coefficients were 0.624 ± 0.055, 0.787 ± 0.032, 0.793 ± 0.035, 0.768 ± 0.057, and 0.803 ± 0.046, and the average (± SD) values of the RMSE of the logarithm of *E. coli* concentrations were 0.177 ± 0.014, 0.133 ± 0.009, 0.131 ± 0.011, 0.138 ± 0.018, and 0.128±0.017 for input sets 1 to 5, respectively. Performance was substantially improved after either the imagery data or biological water quality parameters were added to the physicochemical parameters.

The imagery (input set 2) broadly gave the same level of accuracy as the combination of biological and physicochemical parameters (input 4). Combining the imagery data with the biological and physicochemical parameters (input set 5) did not improve the accuracy of the trees; however, some of the most accurate predictions ($R^2 > 0.9$) were from this set, although these were lower in frequency. Imagery-based performance metrics were much less variable than metrics based on the water quality parameters.

3.3. Most influential input variables

Figure 4 shows the frequencies of the measured water quality parameters and MDNs that were the most influential in all of the datasets. The most influential variables were those that formed the first split and the second most important were those forming the second split, etc. For input set 1 (physicochemical variables only, Fig. 4a) the most influential variables were specific conductance (93% of cases) and pH (7% of cases). The second split was formed by pH, followed by turbidity, and DO. For input 2 (imagery only, Fig. 4b), the most influential variable was the MDN of the green component from the entire range (ER2) GoPro images. The second split was most commonly based on the MDN of the green component of the infrared range (IR2) GoPro images. The most influential variables for the trees for input set 3 were similar to those of input set 2 but specific conductance was the most influential and the MDNs of the red component from the visible range (VI1) GoPro images defined the second split (Fig. 4c).

We also tested the hypothesis that the number of samples removed during resampling would affect the relative influence of the different variables and the accuracy of the analysis. For this, resampling was repeated with one or two datasets removed. The results showed that the ranking of variables by their influence was almost the same as shown in Fig. 4 (data not shown).

4. Discussion

There were some notable differences between environmental factors affecting *E. coli* survival in the pond water between the two sampling days. There was only a very small difference in total radiation between the two sampling days (Fig. S3), although this may not reflect any difference in UV irradiation and transmittance fluxes. There was a difference in temperature of approximately 4 °C between the sampling days. *E. coli* survival has been correlated with temperature in previous studies (e.g., Blaustein et al., 2013), with the measured 4 °C difference in temperature potentially changing the inactivation rate by less than 15–30%.

Correlations in the dataset (Table 1) were used to test hypotheses about the suitability of the pong habitats for *E. coli*. Very strong negative correlations between specific conductance and fDOM can be partially explained by the substantial differences in values measured on the two sampling days. This was likely due to the differences in precipitation on the days preceding the sampling days. There were no precipitation days in the week preceding the 18th of July whereas there was a significant rainfall event on the 28th and 29th of July, when 100 mm of rainfall fell within 24 h. This likely caused the dilution seen in the SPC values and the increase in fDOM as a function by increased runoff. However, the observed correlation could also be partially attributed to the action of photolysis on DOM (Bertilsson and Travnik, 2000), which can create ionic products that affect specific conductance. Very strong correlations between turbidity, phycocyanin, and chlorophyll-*a* point to the important role of cyanobacteria in the planktonic algae community in the study pond. Cyanobacteria are commonly present in freshwater ponds and lakes (Oliver and Ganf, 2002).

The moderate positive correlation between the logarithms of *E. coli* concentrations and fDOM may reflect the important influence of fDOM on *E. coli* survival (Gruber, 2007). This may occur via a shielding effect of DOM for *E. coli*, and by DOM providing nutrients to *E. coli* via DOM photolysis. After photolysis has occurred, the natural mixing of pond water columns may help *E. coli* repair DNA damage caused by short exposure to UVB (Huot et al, 2000). Absorption by DOM and water accounted for 37–77% of UVR attenuation measured in the Laurentian Great Lakes, where scattering and absorption by particles accounted for up to 41% and 52% of attenuation at longer UVR wavelengths, respectively (Smith et al., 2004). It is also possible that the *E. coli* concentrations and fDOM correlated because they derive from the same source and have similar decay and die-off kinetics, respectively. This may be of interest to explore in future research.

There are two distinct features of regression tree-based analysis of water quality data. First, regression trees demonstrate the existence of data subsets with different influential predictors. For input set 1, two data subsets with the lowest *E. coli* concentrations were found where SPC was high (Fig. 2). Clear water with low DO had the lowest *E. coli* concentrations. Elevated DO is commonly thought to lead to the *E. coli* inactivation (Ansa et al., 2011). However, an increase in DO in locations where DO and fDOM were generally low, may signify an increase in the photosynthetic activity of algae and a potential increase in the supply of nutrients and shading of *E. coli*. For input set 1, pH was the second most influential variable in the regression tree analysis (Fig. 2). Although pH did not correlate with *E. coli* concentrations across the entire dataset (Table 1), it discriminated between two distinctly different levels of *E. coli* in datasets with high fDOM. It has been found that *E. coli* becomes deactivated by exogenous mechanisms, while endogenous mechanisms have a more muted effect under moderate pH conditions (Davies-

Colley et al., 1999; Liu et al., 2016). This might explain the influence of pH that can be seen in Fig. 2.

The second important feature of regression tree analysis is the need to anticipate and deal with multicollinearity, i.e., the correlation between inputs. The strong correlation between SPC and fDOM led to SPC becoming a surrogate for fDOM. The influential role of SPC in the higher-level splits (Fig. 2a and 2b) is hard to interpret until it is realized that SPC provides a representation of fDOM.

Variability in the accuracy of the regression trees obtained from the resampling indicates that both accuracy metrics were very sensitive to the removal of just 5% of the data. A possible reason for this is that the regression tree algorithm focused not on the accuracy but on creating groups of data that are most dissimilar. It may also be the case that the accuracy may have been affected by the presence of outliers in the data; using Tukey tests, we found that approximately 9% of the data for all water quality variables could be considered outliers (data not shown). However, these data may simply reflect the patchiness that is commonly observed in ponds and reservoirs, and removing outliers is not necessarily justified. Figure 4 indicates that the individual trees shown in Fig. 2 do not represent the overall influence of the different variables. Thus, the resampling results provide a much more complete picture. These results also imply that if the trees are to be used as prediction tools, tree ensembles should be used, such as the random forest methodology (Belgiu and Drăguţ, 2016).

Chlorophyll-*a* was the second most influential variable (Fig. 4d). The influence of the total planktonic biomass on *E. coli* survival can be either negative or positive (Ansa et al., 2011). An increase in total planktonic biomass can increase DO and pH and, thus, enhance die-off. Lysis of

algal cells, however, may enhance coliform survival because algal degradation may provide a source of carbon and energy (Sampson et al., 2006).

The imagery data reflect features of the *E. coli* aquatic habitat and allow suitable locations for survival to be identified. The survival of various microorganisms can vary even with similar water quality conditions, meaning that results for *E. coli* estimation based on imagery data cannot be extended to other bacteria. However, it is possible that the imagery data can be used for other microorganisms to distinguish their local habitats in the same water body. To test this, future comparisons of the suitability of imagery data for several organisms would be a useful exercise.

The MDNs for the RedEdge-M wavelength bands were rarely included in regression trees produced using the resampling (Fig. 4). The multispectral RedEdge-M data allowed for the computation of a large number of indices developed for land-based imagery applications, such as the Normalized Difference Vegetation Index (NDVI), Canopy Chlorophyll Content Index (CCCI), Chlorophyll Index Green (CIG) and Chlorophyll Index Red Edge (CIRDG), Chlorophyll Vegetation Index (CVI), NIR/Green Green Difference Vegetation Index (GDVI), Green-Blue NDVI (GBNDVI), and Green-Red NDVI (GRNDVI). We used the imagery input data with and without indices in input sets 2, 3, and 5. The commonly used indices computed from these bands provided second splits in the trees created by the resampling in a very small number of cases (data not shown). It is plausible that the multispectral indices used to analyze the radiance of terrestrial data sources are not effective for the analysis of freshwater data since water is a strong adsorbent in the spectral range in which vegetation is highly reflective. It is also possible that the structure of the indices can be preserved, but other wavelength bands need to be used.

The results of this work have implications for microbial water quality monitoring in ponds. For example, if the rankings of variables according to their influence on the two sampling days hold, ER2 values could be used to delineate zones with different concentrations of *E. coli*. This could inform the sampling of ponds to obtain a more accurate representation of the spatial distribution of *E. coli*.

5. Conclusion

This study has shown that sUAV imagery provided the same or a better level of accuracy in estimates of *E. coli* concentrations across the study pond compared to *in situ* measurements of physicochemical and biological water quality parameters. This demonstrates the opportunity for using sUAV-based imagery to inform microbial water quality monitoring focusing on the spatial and temporal variability of *E. coli* concentrations in irrigation ponds. Our results suggest that including sUAV imagery as part of the environmental microbial food safety toolbox has significant potential that deserves to be explored further.

6. Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors

7 **References**

Alegbeleye, O. O., Singleton, I., & Sant'Ana, A. S. (2018). Sources and contamination routes of microbial pathogens to fresh produce during field cultivation: a review. Food Microbiology, 73, 177–208.

Ansa, E. D. O., Lubberding, H. J., Ampofo, J. A., & Gijzen, H. J. (2011). The role of algae in the removal of *Escherichia coli* in a tropical eutrophic lake. Ecological Engineering, 37(2), 317–324.

Ansa, E. D. O., Lubberding, H. J., Ampofo, J. A., Amegbe, G. B., & Gijzen, H. J. (2012). Attachment of faecal coliform and macro-invertebrate activity in the removal of faecal coliform in domestic wastewater treatment pond systems. Ecological Engineering, 42, 35–41.

Arar, E. J., & Collins, G. B. (1997). Method 445.0 In Vitro Determination of Chlorophyll *a* and Pheophytin *a* in Marine and Freshwater Algae by Fluorescence. U.S. Environmental Protection Agency, Washington, DC.

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing, 114, 24–31.

Bertilsson, S., & Tranvik, L. J. (2000). Photochemical transformation of dissolved organic matter in lakes. Limnology and Oceanography, 45(4), 753–762.

Blaustein, R. A., Pachepsky, Y., Hill, R. L., Shelton, D. R., & Whelan, G. (2013). *Escherichia coli* survival in waters: temperature dependence. Water Research, 47(2), 569–578.

Boddula, V., Ramaswamy, L., & Mishra, D. (2017). A Spatio-Temporal Mining Approach for Enhancing Satellite Data Availability: A Case Study on Blue Green Algae. In 2017 IEEE

International Congress on Big Data (BigData Congress) (pp. 216–223). Institute of Electrical and Electronics Engineers.

Boletín Oficial Del Estado Legislación Consolidada (2007). RD 1620/2007, de 7 de diciembre, por el que se establece el régimen jurídico de la reutilización de las aguas depuradas. Ministerio de la Presidencia (BOE) núm. 294, de 8 de diciembre de 2007 Referencia: BOE-A-2007-21092.[Official State Bulletin Consolidated Legislation (2007). RD 1620/2007, of December 7, which establishes the legal regulations for the reuse of purified water. Ministry of the Presidency (BOE) no. 294, of December 8, 2007 Reference: BOE-A-2007-21092]

Byappanahalli, M. N., Sawdey, R., Ishii, S., Shively, D. A., Ferguson, J. A., Whitman, R. L., & Sadowsky, M. J. (2009). Seasonal stability of *Cladophora*-associated *Salmonella* in Lake Michigan watersheds. Water Research, 43(3), 806–814.

Carr, G. M., Morin, A., & Chambers, P. A. (2005). Bacteria and algae in stream periphyton along a nutrient gradient. Freshwater Biology, 50(8), 1337–1350.

Choi, S. H., Baumler, D. J., & Kaspar, C. W. (2000). Contribution of dps to acid stress tolerance and oxidative stress tolerance in *Escherichia coli* O157: H7. Applied Environmental Microbiology, 66(9), 3911–3916.

Constantin, S., Constantinescu, Ș., & Doxaran, D. (2017). Long-term analysis of turbidity patterns in Danube Delta coastal area based on MODIS satellite data. Journal of Marine Systems, 170, 10–21.

Davies-Colley, R. J., Donnison, A. M., Speed, D. J., Ross, C. M., & Nagels, J. (1999). Inactivation of faecal indicator micro-organisms in waste stabilisation ponds: interactions of environmental factors with sunlight. Water Research, 33(5), 1220–1230.

Duan, H., Cao, Z., Shen, M., Liu, D., & Xiao, Q. (2019). Detection of illicit sand mining and the associated environmental effects in China's fourth largest freshwater lake using daytime and nighttime satellite images. Science of the Total Environment, 647, 606–618.

Economic Research Service, ERS. USDA. (2014). Cost Estimates of Foodborne Illnesses. Available at: https://www.ers.usda.gov/data-products/cost-estimates-of-foodborne-illnesses/

Environmental Protection Agency. EPA. (2012). Recreational Water Quality Criteria. Available at: https://www.epa.gov/sites/production/files/2015-10/documents/rwqc2012.pdf

European Parliament and the Council of the European Union (2006). Directive 2006/7/EC concerning the management of bathing water quality and repealing Directive 76/160/EEC. Official Journal of the European Union, 4.3, L64/37.

Franz, E., van Diepeningen, A. D., de Vos, O. J., & van Bruggen, A. H. (2005). Effects of cattle feeding regimen and soil management type on the fate of *Escherichia coli* O157: H7 and *Salmonella* enterica serovar *Typhimurium* in manure, manure-amended soil, and lettuce. Applied Environmental Microbiology, 71(10), 6165–6174.

Giardino, C., Bresciani, M., Cazzaniga, I., Schenk, K., Rieger, P., Braga, F., Matta, E., & Brando, V. (2014). Evaluation of multi-resolution satellite sensors for assessing water quality and bottom depth of Lake Garda. Sensors, 14(12), 24116–24131.

Hellberg, R.S., & Chu, E. (2016). Effects of climate change on the persistence and dispersal of foodborne bacterial pathogens in the outdoor environment: A review. Critical Reviews in Microbiology, 42, 548–572.

Huot, Y., Jeffrey, W. H., Davis, R. F., & Cullen, J. J. (2000). Damage to DNA in bacterioplankton: a model of damage by ultraviolet radiation and its repair as influenced by vertical mixing. Photochemistry and Photobiology, 72(1), 62–74.

Isidro, C. M., McIntyre, N., Lechner, A. M., & Callow, I. (2018). Quantifying suspended solids in small rivers using satellite data. Science of the Total Environment, 634, 1554–1562.

Jones, R. M., Liu, L., & Dorevitch, S. (2013). Hydrometeorological variables predict fecal indicator bacteria densities in freshwater: data-driven methods for variable selection. Environmental Monitoring and Assessment, 185(3), 2355–2366.

Jongman, M., & Korsten, L. (2018). Irrigation water quality and microbial safety of leafy greens in different vegetable production systems: A review. Food Reviews International, 34(4), 308–328.

Kay, D., Bartram, J., Prüss, A., Ashbolt, N., Wyer, M.D., Fleisher, J.M., Fewtrell, L., Rogers, A. & Rees, G. (2004). Derivation of numerical values for the World Health Organization guidelines for recreational waters. Water Research, 38(5), 1296–1304.

Kimmel, R. (1999). Demosaicing: image reconstruction from color CCD samples. IEEE Transactions on Image Processing, 8(9), 1221–1228.

Kislik, C., Dronova, I., & Kelly, M. (2018). UAVs in support of algal bloom research: a review of current applications and future opportunities. Drones, 2(4), 35.

Kroll, C. N., Croteau, K. E., & Vogel, R. M. (2015). Hypothesis tests for hydrologic alteration. Journal of Hydrology, 530, 117–126.

Liu, L., Hall, G., & Champagne, P. (2016). Effects of environmental factors on the disinfection performance of a wastewater stabilization pond operated in a temperate climate. Water, 8(1), 5.

Olaimat, A. N., & Holley, R. A. (2012). Factors influencing the microbial safety of fresh produce: a review. Food microbiology, 32(1), 1–19.

Oliver, R. L., & Ganf, G.G. (2002). Freshwater blooms. In: Whitton, B. A., & Potts, M. (Eds.). The ecology of cyanobacteria: their diversity in time and space. Springer Science & Business Media., pp. 149-194.

Pachepsky, Y., Kierzewski, R., Stocker, M., Sellner, K., Mulbry, W., Lee, H., & Kim, M. (2018). Temporal stability of *Escherichia coli* concentrations in waters of two irrigation ponds in Maryland. Applied and Environmental Microbiology, 84(3), e01876-17.

Pachepsky, Y., Shelton, D. R., McLain, J. E., Patel, J., & Mandrell, R. E. (2011). Irrigation waters as a source of pathogenic microorganisms in produce: a review. Advances in Agronomy, 113, 75–141.

Park, Y., Pachepsky, Y. A., Cho, K. H., Jeon, D. J., & Kim, J. H. (2015). Stressor–response modeling using the 2D water quality model and regression trees to predict *chlorophyll-a* in a reservoir system. Journal of Hydrology, 529, 805–815.

Quilliam, R. S., Clements, K., Duce, C., Cottrill, S. B., Malham, S. K., & Jones, D. L. (2011). Spatial variation of waterborne *Escherichia coli* – implications for routine water quality monitoring. Journal of Water and Health, 9(4), 734–737.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org/

Rochelle-Newall, E., Nguyen, T. M., Le, T. P., Sengtaheuanghoung, O., & Ribolzi, O. (2015). A short review of fecal indicator bacteria in tropical aquatic ecosystems: knowledge gaps and future directions. Frontiers in Microbiology, 6, 308.

Sampson, R.W., Swiatnicki, S.A., Osinga, V.L., Supita, J.L., McDermott, C.M., & Kleinheiz, G.T. (2006). Effects of temperature and sand on *E. coli* survival in a northern lake water microcosm. Journal of Water and Health, 4(3), 389–393

Scharff, R.L. 2010. Health-related costs from foodborne illness in the United States. Produce Safety Project, Georgetown University. Available at: http://www.publichealth.lacounty.gov/eh/docs/ReportPublication/HlthRelatedCostsFromFoodborneIllinessUS.pdf

Semenov, A. V., Franz, E., Van Overbeek, L., Termorshuizen, A. J., & Van Bruggen, A. H. (2008). Estimating the stability of *Escherichia coli* O157: H7 survival in manure-amended soils with different management histories. Environmental Microbiology, 10(6), 1450–1459.

Semenov, A. V., Van Bruggen, A. H., Van Overbeek, L., Termorshuizen, A. J., & Semenov, A. M. (2007). Influence of temperature fluctuations *on Escherichia coli* O157: H7 and *Salmonella enterica* serovar *Typhimurium* in cow manure. FEMS Microbiology Ecology, 60(3), 419–428.

Smith, R. E., Allen, C. D., & Charlton, M. N. (2004). Dissolved organic matter and ultraviolet radiation penetration in the Laurentian Great Lakes and tributary waters. Journal of Great Lakes Research, 30(3), 367–380.

Sorrell, B. K., Hawes, I., & Safi, K. (2013). Nitrogen and carbon limitation of planktonic primary production and phytoplankton–bacterioplankton coupling in ponds on the McMurdo Ice Shelf, Antarctica. Environmental Research Letters, 8(3), 035043.

Therneau, T. M., & Atkinson, E. J. 2018. An Introduction to Recursive Partitioning Using the PART Routines. Mayo Foundation. Available at: https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf

United States Congress (2011). FDA Food Safety Modernization Act. Public Law 111-353, 124 STAT. 3885.

United States Environmental Protection Agency. (2002). Method 1603. *Escherichia coli* (*E. coli*) in water by membrane filtration using modified membrane thermotolerant Escherichia coli agar (modified mTEC). EPA 821-R-02-023.

Van Elsas, J. D., Semenov, A. V., Costa, R., & Trevors, J. T. (2011). Survival of *Escherichia coli* in the environment: fundamental and public health aspects. The ISME Journal, 5(2), 173.

Wang, X., Gong, Z., & Pu, R. (2018). Estimation of chlorophyll a content in inland turbidity waters using WorldView-2 imagery: a case study of the Guanting Reservoir, Beijing, China. Environmental Monitoring and Assessment, 190(10), 620.

World Health Organization (2006). *Guidelines for the safe use of wastewater, excreta and greywater* (Vol. 1). World Health Organization.

Table S1. Descriptive statistics of water quality parameters and mean digital numbers for each observation day.

Table S2. Descriptive statistics of water quality parameters and mean digital numbers pooled for both observation days.

Table 1. Correlations between water quality parameters and mean digital numbers obtained from the imagery analysis.

Figure 1. Locations of the sampling points at the studied pond.

Figure 3. Histograms of (a) the coefficient of determination (R^2) and (b) root-mean-squared error (RMSE) of regression trees obtained by the removal of all possible dataset triplets from a total of 46 available datasets.

Figure 4. Relative influence of measured water quality parameters and mean digital numbers estimated from regression trees obtained by the removal of all possible dataset triplets from a total of 46 available datasets: (a) physicochemical parameters (input set 1); (b) imagery data only (input set 2); (c) physicochemical parameters and imagery data (input set 3); (d) physicochemical and biological parameters (input set 4); and (e) physicochemical and biological parameters combined with imagery data (input set 5). The legend applies to panels (a) to (e).

Figure S1. Five bands of the RedEdge-M camera imagery.

Figure S2. Transmittance of GoPro lens filters.

Figure S3. Total solar radiation on the observation days.

Table 1. Correlations between water quality parameters and mean digital numbers from imagery

	DO	SPC	pH	TURB	BGAPC	CHLAF	fDOM	CHLAS	LGCFU	ER1	ER2	ER3	IR1	IR2	IR3	VI1	VI2	VI3	MS1	MS2	MS3	MS4	MS5
DO	1.00				0.62	0.77						0.62											
SPC		1.00					-1.00				0.81	0.61	0.63										
pH			1.00																				
TURB				1.00	0.93	0.91																	
BGAPC	0.62			0.93	1.00	0.96																	
CHLAF	0.77			0.91	0.93	1.00																	
fDOM		-1.00					1.00		0.61		-0.81	-0.63	-0.62					-0.62					
CHLAS								1.00															
LGCFU							0.61		1.00														
ER1										1.00	0.85	0.72	0.80	0.75	0.75	0.76	0.81						0.71
ER2		0.81					-0.81			0.85	1.00	0.78	0.76		0.73		0.67	0.64					
ER3	0.62	0.61					-0.63			0.72	0.78	1.00	0.83	0.77	0.82			0.63					
IR1		0.63					-0.62			0.80	0.76	0.83	1.00	0.92	0.95		0.66						0.62
IR2										0.75		0.77	0.92	1.00	0.92	0.68	0.69					0.63	0.78
IR3										0.75	0.73	0.82	0.95	0.92	1.00	0.61	0.68	0.62					0.66
VI1										0.76				0.68	0.61	1.00	0.96		0.67	0.62	0.69		0.82
VI2										0.81	0.67		0.66	0.69	0.68	0.96	1.00		0.65				0.73
VI3							-0.62				0.64	0.63			0.62			1.00					
MS1																0.67	0.65		1.00	0.65			
MS2																0.62			0.65	1.00	0.90		0.64
MS3																0.69				0.90	1.00		0.64
MS4														0.63								1.00	
MS5										0.71			0.62	0.78	0.66	0.82	0.73			0.64	0.64		1.00

Conflict of interests declaration

Authors of the manuscript “Drone-based imaging to assess the microbial water quality in an irrigation pond: a pilot study” Billie Morgan, Matt Stocker, Javier Valdes-Abellan, Moon Kim, and Yakov Pachepsky declare no conflict of interest.

Graphical abstract

Highlights

- Remote monitoring of microbial water quality in irrigation ponds is challenging.
- UAV-derived imagery was used characterize habitat parameters for *E. coli*.
- Regression trees were constructed using the remote imagery and field-based data.
- An average determination coefficient of 0.793 ± 0.035 was obtained.
- This work demonstrates the potential of UAV-based monitoring of ponds.

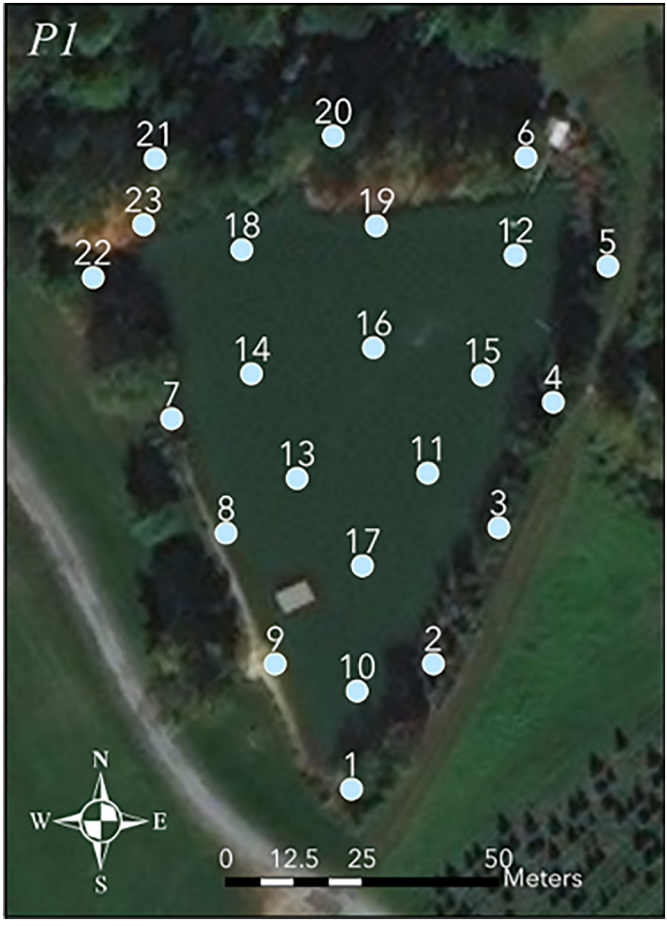
P1
20
21
6
23
19
18
12
5
22
16
14
15
4
7
13
11
8
3
17
9
2
10
1
N
W
E
S
0
12.5
25
50
Meters

Figure 1

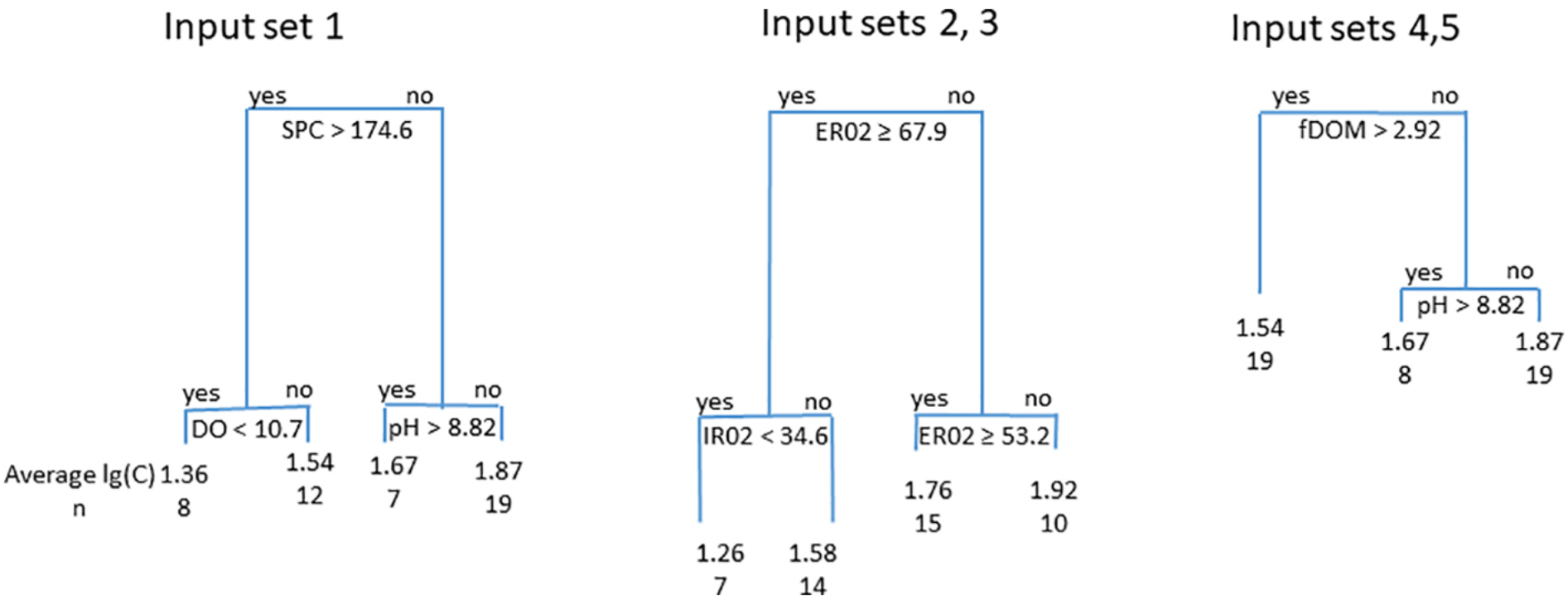
Input set 1
yes
no
SPC > 174.6
yes
no
yes
no
DO < 10.7
pH > 8.82
Average lg(C)
1.36
1.54
1.67
1.87
n
8
12
7
19
Input sets 2, 3
yes
no
ER02 ≥ 67.9
yes
no
yes
no
IR02 < 34.6
ER02 ≥ 53.2
1.26
1.58
1.76
1.92
7
14
15
10
Input sets 4,5
yes
no
fDOM > 2.92
yes
no
pH > 8.82
1.54
1.67
1.87
19
8
19

Figure 2

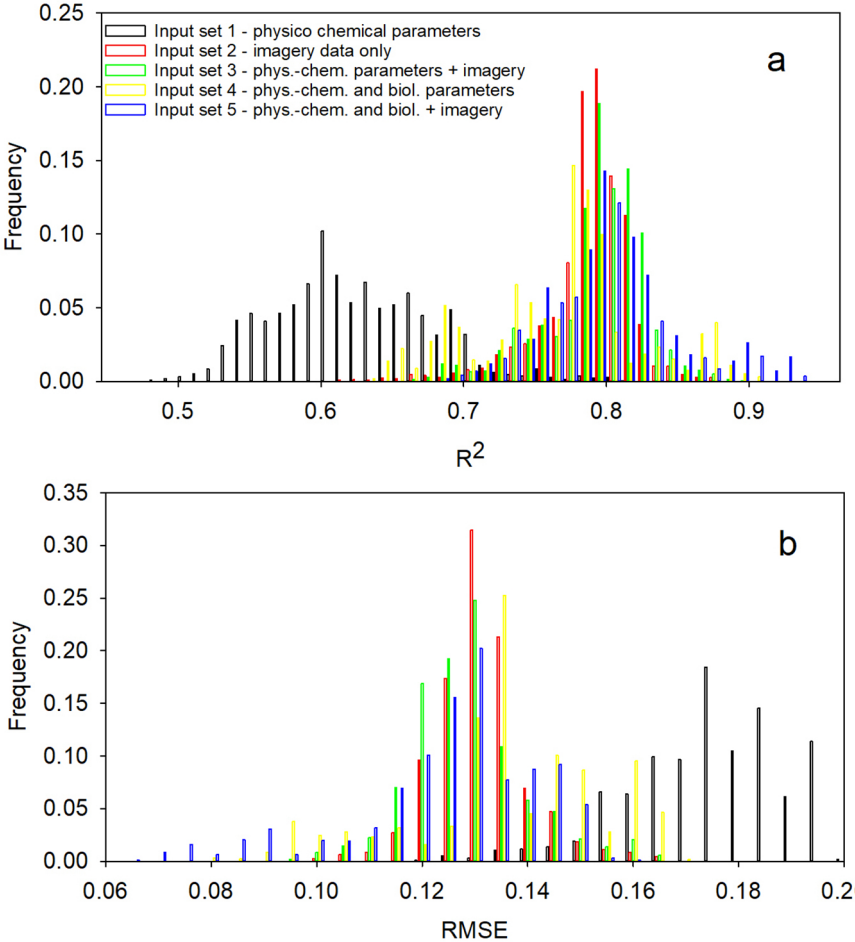
Input set 1 - physico chemical parameters
Input set 2 - imagery data only
Input set 3 - phys.-chem. parameters + imagery
Input set 4 - phys.-chem. and biol. parameters
Input set 5 - phys.-chem. and biol. + imagery
a
Frequency
0.25
0.20
0.15
0.10
0.05
0.00
0.5
0.6
0.7
0.8
0.9
R2
b
Frequency
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.06
0.08
0.10
0.12
0.14
0.16
0.18
0.20
RMSE

Figure 3

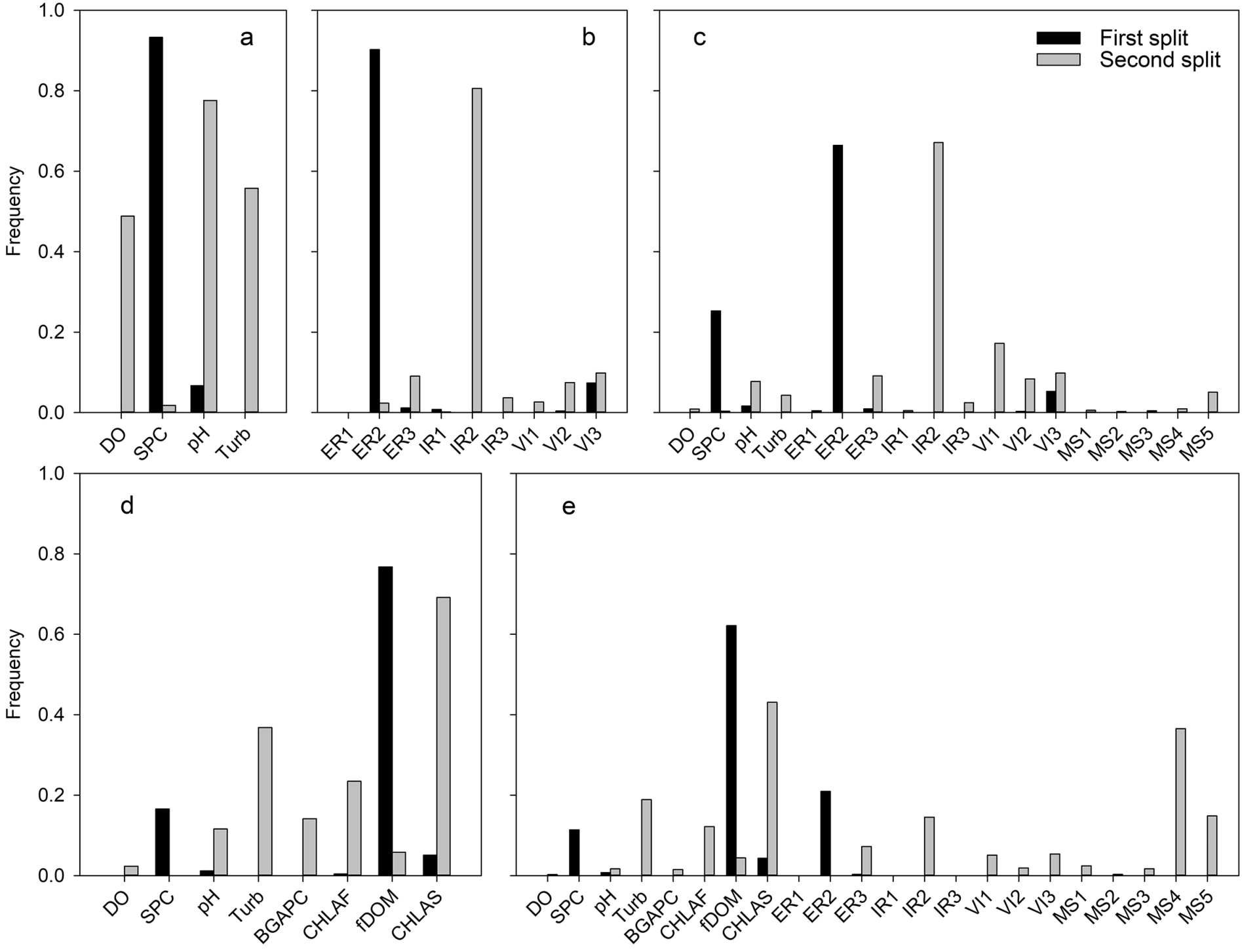
a
b
c
d
e
Frequency
First split
Second split
DO
SPC
pH
Turb
BGAPC
CHLAF
fDOM
CHLAS
ER1
ER2
ER3
IR1
IR2
IR3
VI1
VI2
VI3
MS1
MS2
MS3
MS4
MS5

Figure 4