Science of the Total Environment 717 (2020) 137249

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Potential and limitations of a pilot-scale drinking water distribution system for bacterial community predictive modelling

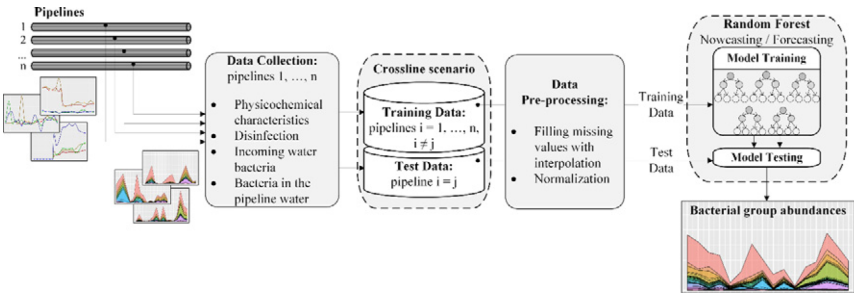
Christina Brester [a,*], Ivan Ryzhikov [a], Sallamaari Siponen [a], Balamuralikrishna Jayaprakash [b], Jenni Ikonen [b], Tarja Pitkänen [b], Ilkka T. Miettinen [b], Eila Torvinen [a], Mikko Kolehmainen [a]

[a] Department of Environmental and Biological Sciences, University of Eastern Finland, P.O. Box 1627, FI-70211 Kuopio, Finland
[b] Department of Health Security, Expert Microbiology Unit, National Institute for Health and Welfare, P.O. Box 95, FI-70701 Kuopio, Finland

HIGHLIGHTS

- Data was collected in a pilot-scale drinking water distribution system.
- Random forest regression was trained to predict abundances of bacterial groups.
- The pipe material and disinfection matter when selecting samples for modelling.
- In 1-week forecasting, *Rhizobiales* and *Pseudanabaenales* were accurately predicted.
- Water pH, temperature, and copper concentration were the most important predictors.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 10 December 2019
Received in revised form 9 February 2020
Accepted 9 February 2020
Available online 11 February 2020

Editor: Damia Barcelo

Keywords:
Nowcasting
Forecasting
Bacterial abundance
Water pipeline
Absolute read count
Random forest

ABSTRACT

Waterborne disease outbreaks are a persistent and serious threat to public health according to reported incidents across the globe. Online drinking water quality monitoring technologies have evolved substantially and have become more accurate and accessible. However, using online measurements alone is unsuitable for detecting microbial regrowth, potentially including harmful species, ahead of time in the distribution systems. Alternatively, observational data could be collected periodically, e.g. once per week or once per month and it could include a representative set of variables: physicochemical water characteristics, disinfectant concentrations, and bacterial abundances, which would be a valuable source of knowledge for predictive modelling that aims to reveal pathogen-related threats. In this study, we utilised data collected from a pilot-scale drinking water distribution system. A data-driven random forest model was used for predictive modelling and was trained for nowcasting and forecasting abundances of bacterial groups. In all the experiments, we followed the realistic crossline scenario, which means that when training and testing the models the data is collected from different pipelines. In spite of the more accurate results of the nowcasting, the 1-week forecasting still provided accurate predictions of the most abundant bacteria, their rapid increase and decrease. In the future predictive modelling might be used as a tool in designing control measures for opportunistic pathogens which are able to multiply in the favourable conditions in drinking water distribution systems (DWDS). Eventually, the forecasting information will be able to produce practically helpful data for controlling the DWDS regrowth.

[*] Corresponding author at: P.O. Box 1627, FI-70211 Kuopio, Finland.
E-mail address: kristina.brester@uef.fi (C. Brester).

https://doi.org/10.1016/j.scitotenv.2020.137249
0048-9697/© 2020 Elsevier B.V. All rights reserved.

1. Introduction

Monitoring drinking water quality aims to protect public health, prevent waterborne disease outbreaks (WBOs), and provide a sustainable water supply (WMO, 2013; Borden and Roy, 2015). Despite a variety of legislation imposed to regulate water quality management, multiple recent reports on WBOs have revealed the evident inadequacy of these measures in preventing illnesses (Figueras and Borrego, 2010). Benedict et al. (2017) summarized 42 reports on WBOs in the U.S. comprising 1006 cases of illness and 13 deaths for 2013–2014. Moreira and Bondelind (2017) reviewed 66 articles reporting incidents of WBOs in Europe, North America and New Zealand for the period of 2000–2014 and found that the highest number of WBOs (more than 25 outbreaks) were associated with distribution network failures, which confirms the importance of water quality surveillance in the drinking water distribution systems (DWDS) (Bridle et al., 2015). While the treated drinking water goes through a DWDS, its quality changes due to complex physicochemical processes and microbial interactions (Ikonen et al., 2013). In particular, water characteristics (temperature, pH, conductivity), environmental factors, pipe material, and microbial communities influence the drinking water quality (Inkinen et al., 2018). Since the existing laboratory methods applied to characterise microbiota are time-consuming, expensive and require highly trained personnel for the analysis, sampling and testing all combinations of all possible factors is not possible (Douterelo et al., 2014). Therefore, cutting-edge studies promote the design of data-driven models which enable knowledge extraction, forecasting and risk assessment (Wu and Rahman, 2017; De Clercq et al., 2018; Muharemi et al., 2019).

Overviews of recent water-related studies infer that online-monitoring, nowcasting and forecasting are the central trends in microbial community modelling (Pachepsky et al., 2018). Nowcasting focuses on estimating the present microbial numbers, usually based on the current environmental data (Zhang et al., 2018), whereas forecasting aims to predict the amount of microorganisms contained in the water at some time in the future (Frick et al., 2008). Forecasting may allow the prevention of the water quality deterioration by implementing timely interventions. Nevertheless, for both nowcasting and forecasting the results are dependent on the quality of the data collected and the relevancy of the predictor set, since the models applied are data driven.

The data extracted from bulk water samples and biofilms, which are bacterial growths on surfaces (e.g. the inner surface of a pipe), is the main source of knowledge of the investigated phenomenon for all water research. Small-scale laboratory experiments enable monitoring some preselected species under the controlled conditions, whereas full-scale experiments represent the dynamics and diversity of the microbial community in the distribution system (Douterelo et al., 2016; Fish et al., 2016; Pachepsky et al., 2018). Large DWDS-simulation facilities comprising tanks, pumps, and many meters of pipes are able to reproduce sophisticated microbial interactions affected by the pipe material, disinfectants and dynamic environment (Fish et al., 2015; Fish and Boxall, 2018).

Making accurate predictions of microorganism abundances requires data which describes the whole microbial community, such as a dataset generated by 16S rRNA gene amplicon sequencing which reveals significant microbial interactions by the learning algorithm automatically (Asgari et al., 2019; Gilfillan et al., 2018). Additionally, the number of potential predictors depends on the taxonomic level chosen for the modelling. The lower the taxonomic level of the characterised microorganisms, the more microbial groups are distinguished in the community, and more predictor variables are presented in the data (Ridenhour et al., 2017). Environmental and water characteristics, as well as the pipe material and disinfectants all affect microbial abundances (Palamuleni and Akoth, 2015). As a result, the high-dimensional set of potential predictors complicates the modelling and becomes an obstacle for many learning algorithms (Guyon and Elisseeff, 2003; Zheng and Casari, 2018). After conducting a full-scale experiment and collecting the relevant data carefully, the appropriate learning algorithm is chosen and applied. To train an adequate predictive model, the algorithm should be capable of handling lots of predictors, multicollinearity, and complex non-linear dependencies.

A learning algorithm that is widely used in predictive modelling, even when the number of variables exceeds the sample size greatly, is the random forest (RF) algorithm (Tyralis and Papacharalampous, 2017). It has already been applied successfully in various microbiological studies to solve classification problems (Baudron et al., 2013; Peters et al., 2007), for nowcasting (Vincenzi et al., 2011) and forecasting (Mohammed and Seidu, 2019; Parkhurst et al., 2005). The studies revealed the potential of RF for microbial predictive modelling and this explains the choice of the learning algorithm for the current study.

This paper presents the results of a data-driven modelling study which aimed to predict abundances of bacteria in microbial communities based on the data collected from a large pilot-scale DWDS-simulation system (Kuopio, Finland). This experiment was performed based on the knowledge extracted from a previous full-scale experiment (Ikonen et al., 2017; Inkinen et al., 2019). Utilising the data from the pilot experiment, we evaluated the quality of the collected data in terms of predictive modelling and investigated the performance of the trained models and their suitability to support real-world microbial risk evaluation procedures. This study aims to model the bacterial community and the major focus of the modelling is on crossline predicting bacterial abundances, i.e. the data collected from one line is used to train a model which is then applied to make predictions for another line. This approach is practically valuable since it simulates a real-world scenario when models trained on data gathered from a few DWDS are used in other DWDS.

2. Materials and methods

2.1. The DWDS-simulation system

A DWDS-simulation system with four study lines employing copper and crosslinked polyethylene (PEX) pipes with sodium hypochlorite or chloramine disinfection was built for the study (Table 1). The DWDS pipelines consisted of two copper lines and two PEX lines with a length of about 57 m and an inner diameter of 10 mm. The pipeline characteristics are shown in Table 1.

Water for the system was supplied from a pilot-scale drinking water treatment plant using surface water. The water treatment included coagulation, flotation, sand filtration, and alkalisation for pH adjustment. Continuous disinfection with an inlet chlorine concentration of 0.3 mg Cl_2/l for all four pipelines was supplied by pumping a sodium hypochlorite solution in two lines, and sodium hypochlorite and ammonium solutions that formed chloramine when mixed in the pipe in the other two lines. The hypochlorite disinfected line was divided into one copper line and one PEX line, and similarly the chloramine disinfected line was divided into two lines with two different pipe materials. The water flow was restricted to 250 ml/min (0.053 m/s).

Weekly sampling for microbial studies and physicochemical measurements were gathered for seven weeks before starting disinfection on 2.8.2017. After the disinfection was initiated, weekly sampling continued for ten weeks. A total of seven samples from each four study lines without disinfection and eleven samples with disinfection were taken over the time period of 21.6.2017–11.10.2017.

Table 1
The DWDS-simulation system pipelines: materials and disinfectants.

Pipeline number	Pipe material	Disinfectant
Line 1	Copper	Hypochlorite
Line 2	Copper	Chloramine
Line 3	PEX	Hypochlorite
Line 4	PEX	Chloramine

Physicochemical parameters such as the temperature, pH, electric conductivity, copper and iron concentrations were measured as previously described by Ikonen et al. (2017). Free and total chlorine, chloramine, and ammonium concentrations were measured immediately after sampling. Bacterial communities were analysed using Illumina MiSeq high-throughput amplicon sequencing targeting the V4 region of the 16S rRNA gene using 341F/785R primers (Klindworth et al., 2013).

One litre of water was filtered on polyethersulfone (PES) membrane filters with pore size of 0.22 μm (Express Plus Membrane, Merck Millipore, Ireland) after which the filters were stored at −75 °C or lower. Total nucleic acids were extracted as previously described by Inkinen et al. (2019). In brief, Chemagic DNA Plant Kit (Perkin Elmer, Waltham, MA, USA) was used and RNA was further purified using Ambion Turbo DNA-free DNase kit (Life Technologies, Carlsbad, CA, USA). cDNA was synthesized using Invitrogen Superscript IV VILO system (Thermo Fisher Scientific, Waltham, MA, USA) and used in the 16S rRNA analysis. The high-throughput Illumina MiSeq amplicon libraries produced by LGC Genomics (LGC Genomics GmbH, Berlin, Germany) were processed and amplicon sequence variants (ASVs) (Callahan et al., 2017) analysed using QIIME (Quantitative Insights Into Microbial Ecology) (Caporaso et al., 2010) to define bacterial groups and their abundances in water samples. The characteristics of the bacterial community in relation to the measured water quality parameters will be presented in more detail by Siponen et al. (2020, unpublished results).

2.2. Data pre-processing

The data used in the modelling is comprised of four groups of variables. The physicochemical water characteristics measured in pipelines include temperature (°C), pH, electric conductivity (μS/cm), Cu (mg/l), and Fe (mg/l). Next, there is the quantity of raw chemical disinfectants entering the pipelines, namely chlorine (mg Cl_2/l) and ammonium (mg/l). The set of variables also includes absolute abundance of ASVs as read counts measured in the incoming and pipeline water.

Sequencing read counts were summed up to the order level in the taxonomy. The most abundant groups were selected for modelling based on their median values. 20 ASVs were chosen from the bacteria in the incoming water, other less abundant bacteria were combined in one more group. Then, 25 of the most abundant ASVs were selected from the bacteria in the pipeline water, and again other bacteria were pooled in one group. Other microbes than bacteria are not included in the analysis of this study. Table 2 and Supplementary material 1 provide more details on the set of variables.

Before applying any data transformation, there was a need for the imputation of missing values. To fill the gaps, a monotone piecewise cubic interpolation method was applied (Fritch and Carlson, 1980). This produced continuous functions which enabled us to generate additional sample points between the collected measurements, which also resulted in a higher time resolution in the dataset: a multivariate time series with a 6-hour interval was derived for training the model.

With more frequent data points, it became possible to vary the prediction horizon, so that we could test 1-, 7- and 30-day prediction horizons. Then, the data was normalised into the interval $[0,1]$ and was applied to all the variables.

2.3. Random forest regression

Decision trees (DTs) are capable of identifying complicated nonlinear interconnections in the data with no assumptions concerning variable distributions (Quinlan, 1986). DTs map the set of m predictors $\mathbf{x} = \{x_1, x_2, \ldots, x_m\}$, which may contain both continuous and categorical variables, to the response variable y, which is categorical in the case of classification and continuous in the case of regression: $y = f(\mathbf{x})$. A DT model is a white box model, which is advantageous for interpreting impacts of predictors (inputs) on the response variable (output) (Breiman et al., 1984).

Despite multiple benefits of the DT model, it appears to be sensitive to small changes in the data and vulnerable to overfitting (Hastie et al., 2009). To reduce the variance and make the model more stable and to prevent overfitting, Breiman proposed an ensemble learning meta-algorithm called the random forest (RF) method (Breiman, 2001). The RF method generates an ensemble of DTs using bootstrap aggregating (also known as bagging). Each i-th DT is trained on the subsample D_i of size $\tilde{N}$ produced from the initial training data D of size N with a replacement, $i = \overline{1, K}$, where K is an ensemble size. If the sampling is uniform and $\tilde{N} = N$ then every subsample D_i contains about two thirds of unique examples from the initial training data D. When the number of DTs in the ensemble is sufficient, averaging their outputs leads to higher stability, robustness to the outliers and lower variance in comparison to a single DT, which varies a lot while training on different subsamples (Hastie et al., 2009).

When training an ensemble of DTs, the learning algorithm evaluates the performance of the i-th tree on the set $D \backslash D_i$ acting as a validation set. The use of this out-of-bag estimate helps to avoid overfitting which makes RF more preferable compared to a single tree. Moreover, training DTs includes a variable selection step, in which the best split for each node is searched for within L randomly selected variables. In this study, the mean squared error estimates the quality of a split (the decrease in the impurity). Due to the built-in variable selection, RF is applicable even if the number of predictors m is much higher than the sample size N (Díaz-Uriarte and Alvarez de Andrés, 2006).

In contrast to a single tree, RF is a black box model since the interpretation of the model structures becomes impossible when the ensemble comprises hundreds or even thousands of trees. However, the influence of predictor variables on the response is measured with special metrics such as the mean decrease in impurity (MDI) or mean decrease in accuracy (MDA) (Genuer et al., 2010; Louppe, 2014). Then MDI, which is used in this work, evaluates the decrease in the impurity caused by splitting with a particular variable. This is weighted with the number of samples in the node and averaged over the ensemble of trees (Nembrini et al., 2018).

RF has a number of meta-parameters, which need tuning: the number of trees in the ensemble K, the maximum tree depth d, and the number of variables selected L while looking for the best split (Goldstein et al., 2011). Typically, some recommended values are acceptable because RF is not very sensitive in this respect. Several hundred is a commonly used value for K, otherwise overly small values lead to overfitting, whereas overly high values make the training process time-consuming. Pruning trees, i.e. reducing d, works well for noisy data and also prevents overfitting. The following rule is applied to define L: $L = \sqrt{m}$, which makes this parameter adaptive.

In this work, the response variable, which is the absolute bacterial abundance (i.e., the read counts), is continuous, therefore, we use the random forest regression method implemented in the scikit-learn library (Pedregosa et al., 2011). The results of the modelling are evaluated with the following metrics:

- The index of agreement (IA) is the relative measure of the model performance and calculated as follows (Willmott, 1981):

$$IA = 1 - \left(\frac{SSE}{\sum_{i=1}^{n} \left(|\hat{y}_i - \overline{y}| + |y_i - \overline{y}| \right)^2} \right),$$

where $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$, y_i is an observed value of the response variable, $\hat{y}_i$ is a predicted value, i.e. a model outcome, $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$, $i = \overline{1, n}$, n is the number of observations. $IA \in [0,1]$, higher values of IA correspond to better models.

Table 2
The set of variables used in the modelling.
The statistical characteristics are estimated for the data collected from all the lines.

Group of variables	Variable	Median	Mean	Std	Max	Min
Physico-chemical	Temperature (°C)	19.4	19.0	1.2	20.6	15.9
	pH	8.0	8.0	0.1	8.2	7.8
	Electric conductivity (μS/cm)	218	219	14	264	199
	Cu (mg/l)	0.11	0.17	0.18	0.69	0.01
	Fe (mg/l)	0.09	0.10	0.04	0.28	0.02
Disinfection	Chlorine (mg/l)	0.23	0.18	0.16	0.33	0.00
	Ammonium (mg/l)	0.07	0.05	0.05	0.10	0.00
Bacteria in the incoming water (abundance as read counts)	Proteobacteria - Gammaproteobacteria - Betaproteobacteriales	5886	9106.34	9738.51	36,015	272
	Proteobacteria - Alphaproteobacteria - Rhizobiales	803	1109.28	1093.11	3573	0
	Proteobacteria - Alphaproteobacteria - Acetobacterales	284	2254.89	3057.86	9551	100
	Chloroflexi - Jg30 - Kf - Cm66 - Na	230	1516.74	2174.75	6772	9
	Planctomycetes - Planctomycetacia - Gemmatales	196	366.59	379.03	1182	11
	Proteobacteria - Alphaproteobacteria - Rhodobacterales	145	122.15	84.64	262	0
	Proteobacteria - Alphaproteobacteria - Caulobacterales	144	110.38	74.84	194	0
	Planctomycetes - Planctomycetacia - Isosphaerales	133	125.72	65.73	214	8
	Cyanobacteria - Oxyphotobacteria - Synechococcales	102	102.94	44.19	222	22
	Bacteroidetes - Bacteroidia - Chitinophagales	95	277.63	469.22	1712	0
	Actinobacteria - Actinobacteria - Frankiales	91	159.45	220.33	819	0
	Proteobacteria - Deltaproteobacteria - Myxococcales	87	89.82	73.54	288	14
	Acidobacteria - Acidobacteriia - Solibacterales	82	352.40	455.94	1417	0
	Bacteroidetes - Bacteroidia - Sphingobacteriales	80	109.62	114.52	440	0
	Proteobacteria - Alphaproteobacteria - Sphingomonadales	77	85.37	67.53	272	0
	Acidobacteria - Subgroup - 6 - Na	68	113.18	122.28	383	0
	Chloroflexi - Olb14 - Na	65	63.03	41.49	116	6
	Planctomycetes - Phycisphaerae - Phycisphaerales	60	58.12	33.42	116	16
	Verrucomicrobia - Verrucomicrobiae - Pedosphaerales	44	47.93	28.00	98	12
	Gemmatimonadetes - Gemmatimonadetes - Gemmatimonadales	42	109.05	132.47	439	8
	Other less abundant bacteria	1200	2158.86	2151.47	7459	329
Bacteria in the pipeline water (abundance as read counts)	Proteobacteria - Gammaproteobacteria - Betaproteobacteriales	3297	6468.70	8224.29	43,834	0
	Proteobacteria - Alphaproteobacteria - Rhizobiales	936	1209.17	1144.00	4660	32
	Proteobacteria - Alphaproteobacteria - Acetobacterales	602	1408.48	1877.87	10,034	0
	Planctomycetes - Planctomycetacia - Gemmatales	183	471.18	614.71	2838	0
	Planctomycetes - Planctomycetacia - Isosphaerales	167	387.67	487.41	2160	0
	Cyanobacteria - Oxyphotobacteria - Synechococcales	152	269.03	369.89	1792	0
	Chloroflexi - Jg30 - Kf - Cm66 - Na	131	1516.33	4299.48	23,627	0
	Acidobacteria - Acidobacteriia - Solibacterales	114	337.98	619.81	3937	0
	Actinobacteria - Actinobacteria - Frankiales	112	203.08	247.24	1052	0
	Bacteroidetes - Bacteroidia - Chitinophagales	109	299.06	512.67	2911	0
	Proteobacteria - Deltaproteobacteria - Myxococcales	107	185.88	230.80	987	0
	Bacteroidetes - Bacteroidia - Sphingobacteriales	104	167.95	207.44	1105	0
	Proteobacteria - Alphaproteobacteria - Caulobacterales	85	311.48	621.69	2799	0
	Verrucomicrobia - Verrucomicrobiae - Pedosphaerales	81	231.32	394.90	2220	0
	Proteobacteria - Deltaproteobacteria - Oligoflexales	74	115.57	127.36	570	0
	Planctomycetes - Planctomycetacia - Planctomycetales	64	647.04	2505.69	17,091	0
	Proteobacteria - Alphaproteobacteria - Sphingomonadales	63	360.41	831.25	3943	0
	Proteobacteria - Alphaproteobacteria - Rhodospirillales	59	296.37	507.16	2190	0
	Acidobacteria - Subgroup_6 - Na	58	175.54	248.30	1157	0
	Cyanobacteria - Oxyphotobacteria - Pseudanabaenales	56	988.07	2401.04	12,604	0
	Cyanobacteria - Oxyphotobacteria - Chloroplast	54	157.59	317.55	2306	0
	Verrucomicrobia - Verrucomicrobiae - Methylacidiphilales	53	134.02	184.97	942	0
	Chloroflexi - Olb14 - Na	53	148.27	208.86	808	0
	Planctomycetes - Phycisphaerae - Phycisphaerales	53	167.43	227.34	1119	0
	Proteobacteria - Alphaproteobacteria - Reyranellales	49	175.83	339.88	2237	0
	Other less abundant bacteria	2620	5819.03	7563.28	32,191	319

- The root mean square error (*RMSE*) is a non-negative scale-dependent measure (Willmott et al., 1985):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}.$$

Lower *RMSE* values indicate a higher model accuracy. The *RMSE* is measured in the units of the response variable and used to compare models within one dataset. This metric is highly sensitive to outliers. In the study, since the values of the read counts vary greatly for different species, the *RMSE* is applied to the normalised values of the response variable to compare the model predictive ability for the species within the community.

In all the experiments, to estimate the RF performance and its ability to make generalisations, the *IA* and *RMSE* were calculated on the test data, whereas the training data was used only to learn the model.

3. Results

3.1. Nowcasting

The nowcasting model predicts the abundance of the *i*-th bacterium from the pipeline water community at time *t* using physicochemical variables, disinfection, bacterial abundances in the incoming water, and abundances of other bacteria in the pipeline water from the same moment *t* as inputs.

Let $\mathbf{u} = \{u_1, u_2, \ldots, u_p\}$ denote the set of predictors including physicochemical variables, disinfection, and read counts in the incoming water and $\mathbf{x} = \{x_1, x_2, \ldots, x_m\}$ denote read counts in the pipeline water. For each *i*-th bacterium $i = \overline{1, m}$, the following model was trained: $x_i^t = f_i(u_1^t, \ldots, u_p^t, x_1^t, \ldots x_j^t, \ldots, x_m^t | \; i \neq j)$.

We conducted a series of experiments wherein the original weekly collected measurements from each pipeline were used as the test data, while models were trained on the data from other pipelines. For example, if the model was tested on the data from Line 1, the training data was generated from samples collected in Lines 2, 3, and 4. In our experiments, we investigated different cases of crossline modelling when the training data was taken from any single pipeline or several pipelines. The training data included the original measurements as well as the interpolated data. After several trials, the main RF parameters were defined as follows: $d = 10$, $K = 250$, $L = \sqrt{m}$.

Fig. 1 presents the experimental results which have been obtained for each bacterium from the community while testing models on different pipelines. The results contain the highest achieved *IA* for every test line (the first four bars) and the aggregated value of *IA* averaged over test lines for each bacterium (the fifth bar). The bars are labelled with the training data that led to the highest *IA*.

This experiment revealed differences in the achieved *IA* for the investigated bacteria, which might be explained by the absence of informative predictors (factors that affect some particular bacteria) or with inaccuracy of measurements. *IA* values which are higher that 0.70–0.80 indicate adequate models which could potentially be applicable in practice.

For many cases, the training data that provided the best *IA* was a set combined from several pipelines. Since there are no identical pipes made of the same material and with the same disinfection, the use of data from different lines tends to make the training set more representative. Therefore, we investigated the influence of these two factors (the pipe material and disinfectant) on the model quality. We compared cases when the training and test samples were taken from pipes of the same material and from pipes with the same disinfection. Fig. 2 illustrates this comparative analysis. Four pairs of tables with *IA* and *RMSE* values correspond to test samples from Lines 1, 2, 3, and 4. Columns in the tables represent different training data. Based on the point estimate given in the last row (the median value), in three out of four cases better values of *IA* and *RMSE* were achieved when the training and test data were taken from pipelines with the same disinfection. The *IA* density functions below also prove this as they are shifted closer to 1.0 in the case of the same disinfection. The *RMSE* densities also differ but not much. Nevertheless, merging two training sets together, representing the same pipe material and disinfection, involves more information in the learning process, and consequently it leads to higher model performance, which is illustrated distinctively with median values and *IA* density functions in all four cases (Fig. 2).

In Supplementary material 2, we provide more results for each bacterium. There are more *IA* values reached with models trained on different sets. This includes the case when the training and test data were taken from the same line (the original measurements and interpolated points were used as the test and training data, correspondingly) and we could gain 0.99–1.0 *IA*, which was possible due to using the training data from the high time resolution.

3.2. Forecasting

While nowcasting reveals existing interconnections in the data, forecasting is more useful for real-world applications since it enables risk estimation in advance. In case of forecasting, the model predicts bacterial abundances at time $t + 1$ based on measurements collected at time *t*: $x_i^{t+1} = f_i(u_1^t, \ldots, u_p^t, x_1^t, \ldots, x_m^t)$.

To evaluate the potential of forecasting in terms of crossline modelling, we repeated the same experiments as we did for nowcasting. The prediction horizon was set to 7 days, which equalled the interval between the original measurements. The results of the forecasting are shown in Fig. 3 (also in Supplementary material 3).

In comparison with the *IA* values shown by the nowcasting, the ones in the forecast are much lower for many bacteria. Nevertheless, some of the trained models provided acceptable *IA* values higher than 0.70–0.80. In many cases the best *IA* values again were obtained with the models trained on the data containing samples from several lines.

While analysing the influence of the pipe material and disinfection on the results of the forecasting, we found that better results were achieved when the training and test data were taken from pipes made of the same material, which was in contrast to what we had found for the nowcasting (Fig. 4). Thus, when the models were trained to make predictions for the future, the pipe material became a more significant factor.

Meanwhile, the use of the joint samples positively affects the model quality and leads to better results, which is consistent with the conclusions made for nowcasting (Fig. 4). Comparing the associated *IA* values obtained for different lines (third columns in the '*IA*' tables), we should note that many of the bacteria predictions made for pipelines 1 and 2 are more accurate than the ones made for lines 3 and 4. The median values show the same.

The predicted abundances of the whole community are illustrated in Fig. 5. In almost all cases, the models captured and forecasted an increase in the amount of bacteria as well as a dramatic decrease. However, the sudden increase in the bacterial numbers was underestimated for pipe 3 on 9.8.2017 (order *Betaproteobacteriales*, class *Gammaproteobacteria*, phylum *Proteobacteria*), for pipe 4 on 27.9.2017 (order *Pseudanabaenales*, class *Cyanobacteria*, phylum *Oxyphotobacteria*; an unassigned ASV from the phylum *Chloroflexi*), and the abrupt reduction of all the ASVs was not predicted for pipe 4 on 16.8.2017. In some cases, bacterial abundances were overestimated as it was with an unassigned ASV from the phylum *Chloroflexi* in line 1 or underestimated as it occurred with *Planctomycetales* (class *Planctomycetacia*, phylum *Planctomycetes*) in line 3. Nevertheless, the most abundant bacteria were mainly predicted correctly.

Analysing the importance of predictor variables in 7-day forecasting, we took into account all the trained crossline models and found that the physicochemical variables, particularly the pH, temperature, and Cu, had the highest importance (Supplementary material 4). The incoming water bacteria such as *Myxococcales* (class *Deltaproteobacteria*, phylum *Proteobacteria*), one unassigned ASV from the phylum *Chloroflexi*, and *Chitinophagales* (class *Bacteroidia*, phylum *Bacteroidetes*) were also found within the most important variables. Other important predictors were related to the bacterial abundances in the pipeline water. These included: *Pseudanabaenales* (class *Oxyphotobacteria*, phylum *Cyanobacteria*), an unassigned ASV from the phylum *Chloroflexi*, and *Sphingomonadales* (class *Alphaproteobacteria*, phylum *Proteobacteria*).

Interestingly, the bacteria in the incoming and pipeline water selected as the most important predictors had medium abundances (not the most abundant ones). Disinfection variables were the least important: the bacterial abundances already reflect this information indirectly. Despite several commonly important predictors, shown as outliers (black dots above the boxplots) in the picture (Fig. 1, Supplementary material 4) prove that for predicting the amount of a particular bacterium some other specific variables are required.

In additional experiments, we varied the prediction horizon and tested models making 1-day and 30-day forecasts. Supplementary material 5 contains the results of these experiments. 1-day predictions appeared to be the most accurate because the abundances x_i^{t+1} do not differ much from x_i^t included in the set of predictors. 30-day predictions were much less accurate but still in some cases the *IA* values reached 0.70–0.80. For the most abundant bacteria *Betaproteobacteriales* (class *Gammaproteobacteria*, phylum *Proteobacteria*), *Rhizobiales* and

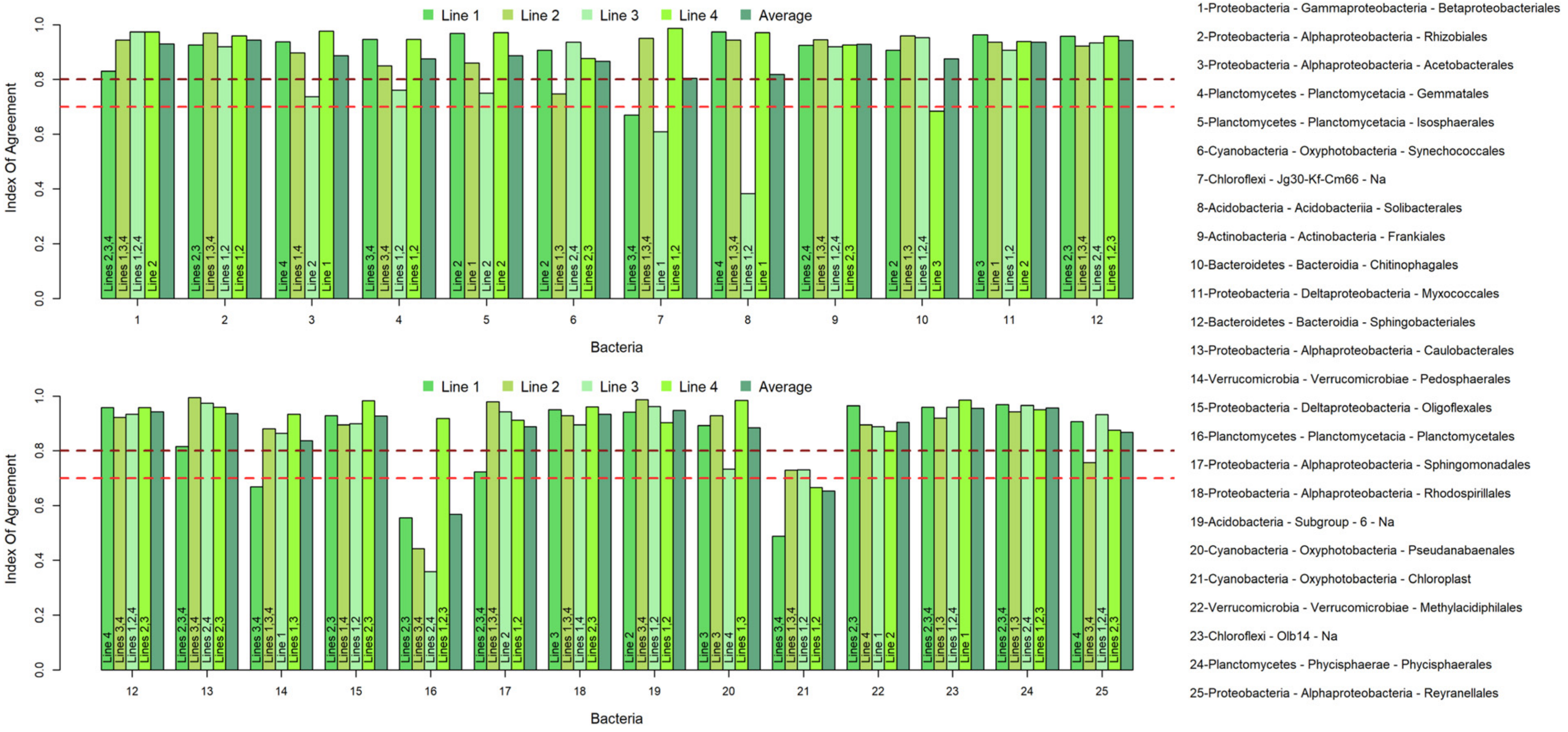


Fig. 1. Crossline nowcasting. The experimental results show the highest *IA* achieved for each bacterium. The bar label corresponds to the training data used and its colour refers to the test data. Dashed lines indicate 0.7 and 0.8 levels of *IA* which correspond to practically acceptable models. The bacteria are listed in descending order of their abundance as read counts.

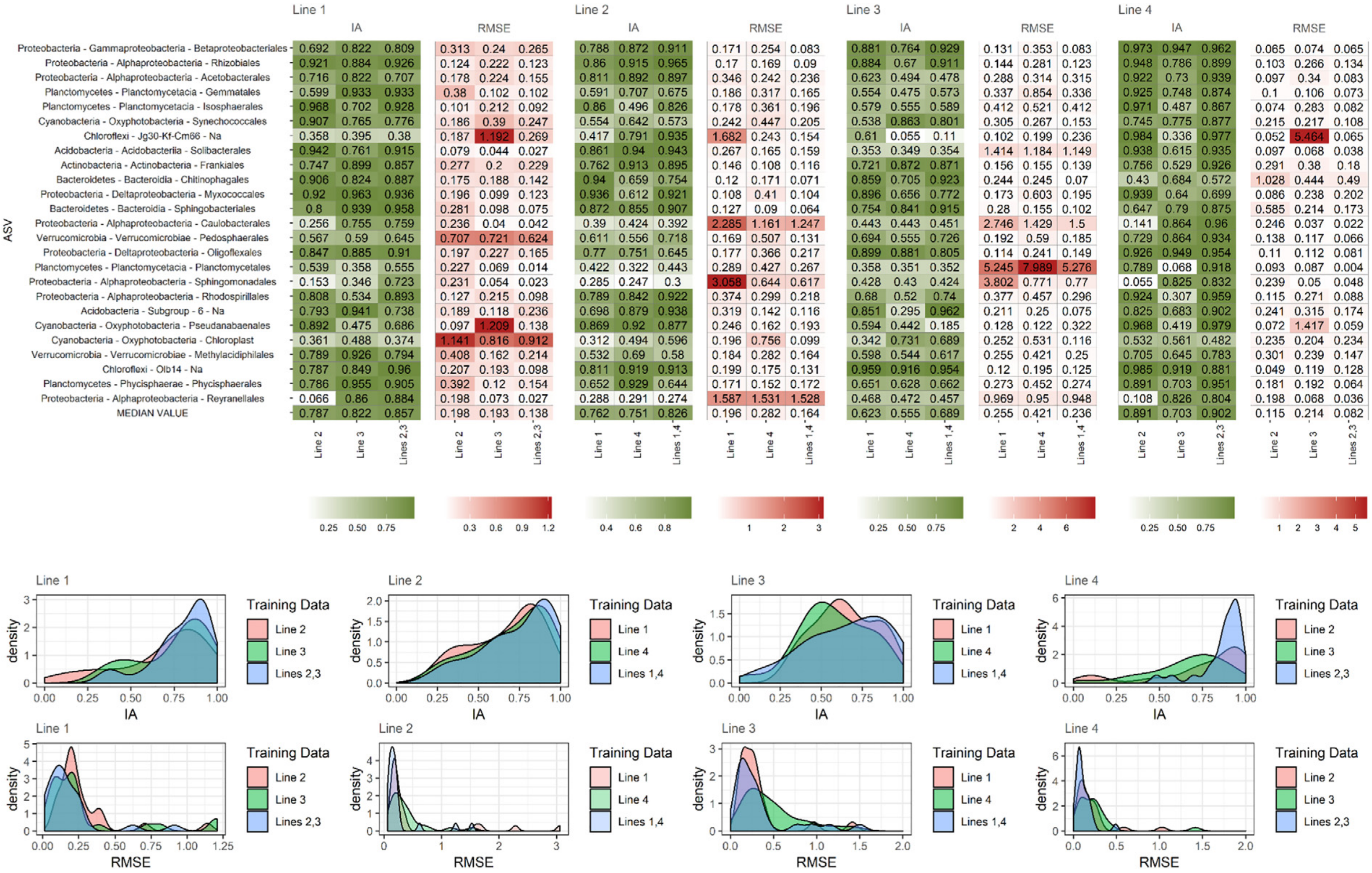


Fig. 2. Influence of the pipe material and disinfection on the model performance in the crossline nowcasting.

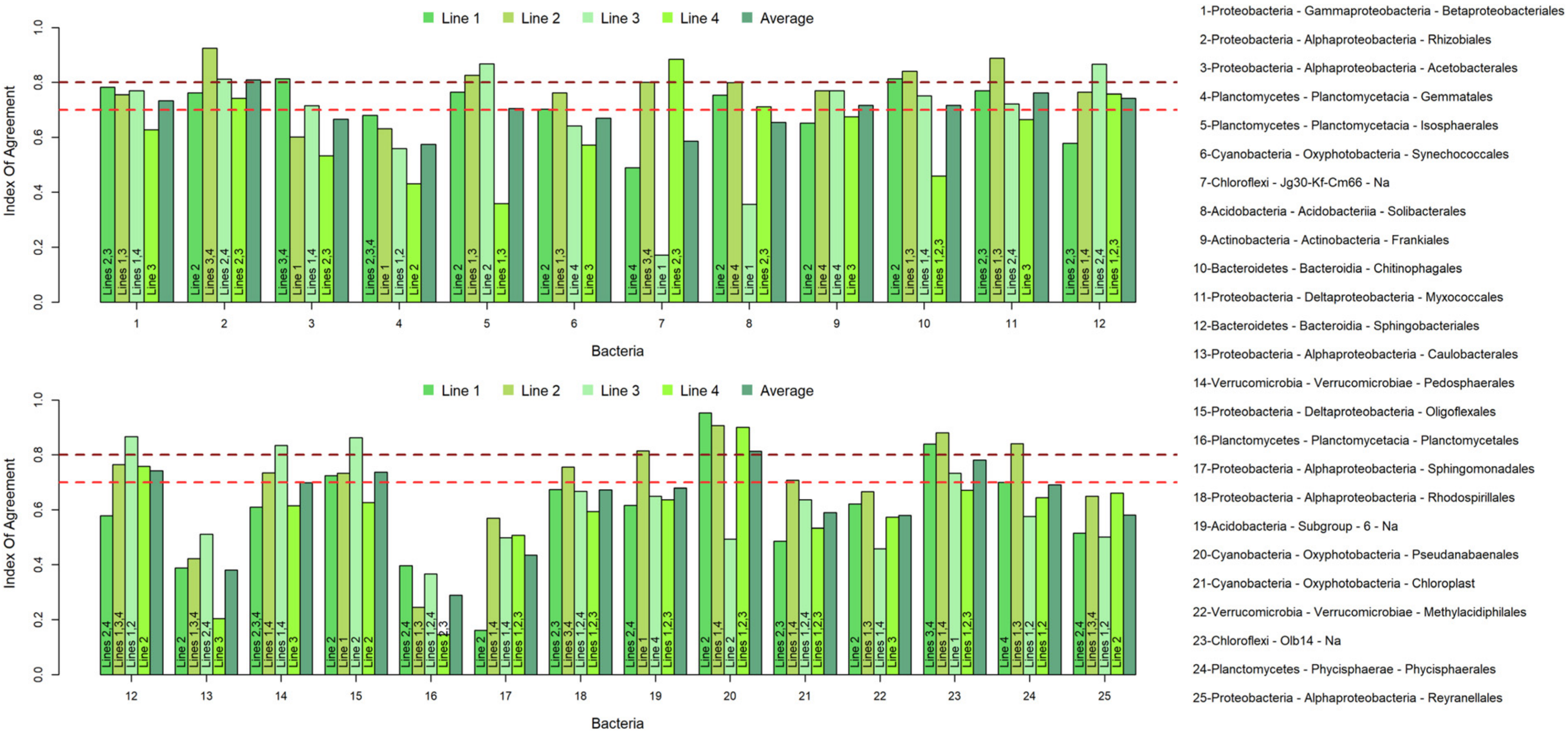
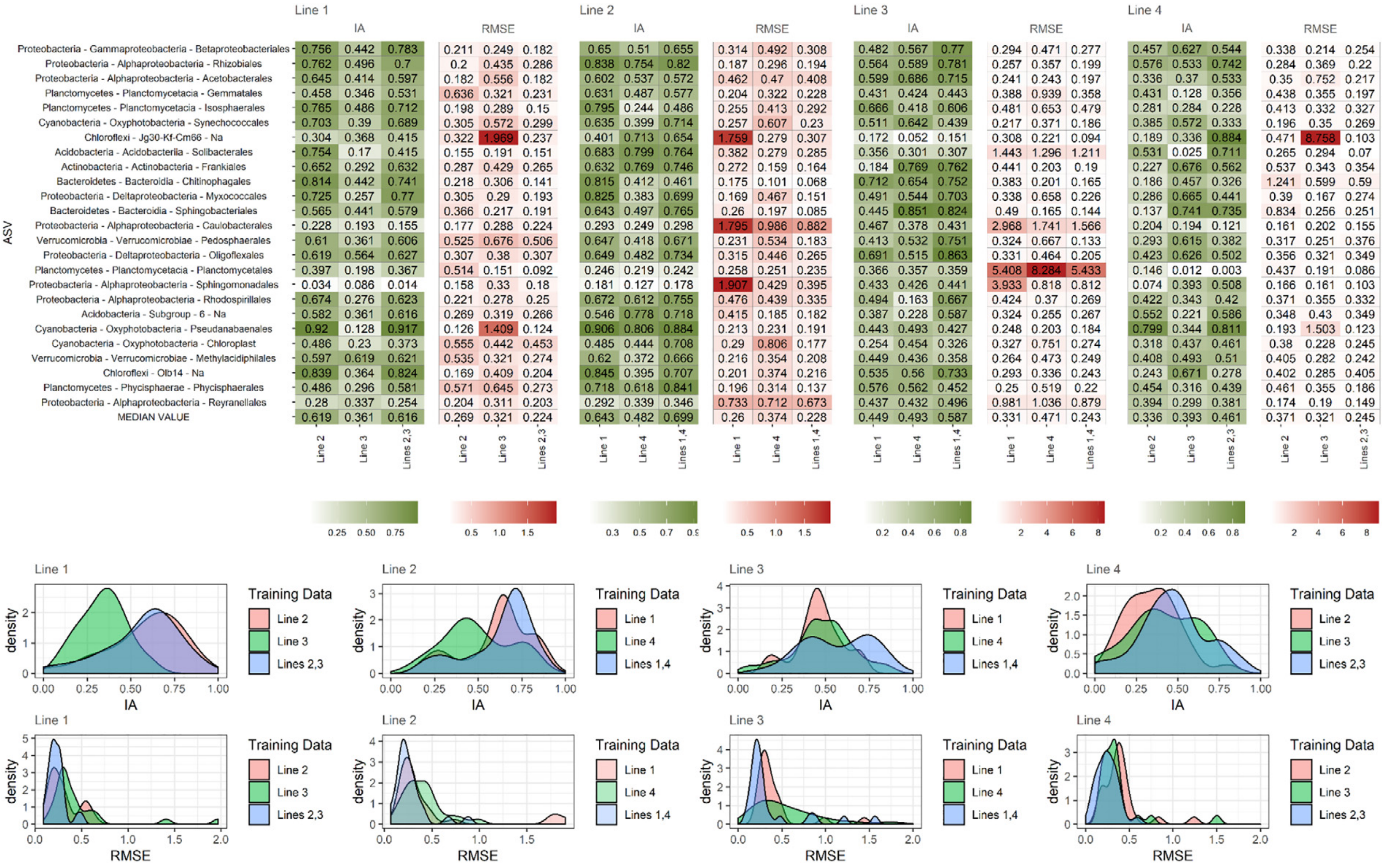


Fig. 3. 7-Day crossline forecasting. The experimental results show the highest *IA* achieved for each bacterium. The bar label corresponds to the training data used and its colour refers to the test data. Dashed lines indicate 0.7 and 0.8 levels of *IA* which correspond to practically acceptable models. The bacteria are listed in descending order of their abundance as read counts.



Line 1

ASV	IA			RMSE		
	Line 2	Line 3	Lines 2,3	Line 2	Line 3	Lines 2,3
Proteobacteria - Gammaproteobacteria - Betaproteobacteriales	0.756	0.442	0.783	0.211	0.249	0.182
Proteobacteria - Alphaproteobacteria - Rhizobiales	0.762	0.496	0.7	0.2	0.435	0.286
Proteobacteria - Alphaproteobacteria - Acetobacterales	0.645	0.414	0.597	0.182	0.556	0.182
Planctomycetes - Planctomycetacia - Gemmatales	0.458	0.346	0.531	0.636	0.321	0.231
Planctomycetes - Planctomycetacia - Isosphaerales	0.765	0.486	0.712	0.198	0.289	0.15
Cyanobacteria - Oxyphotobacteria - Synechococcales	0.703	0.39	0.689	0.305	0.572	0.299
Chloroflexi - Jg30-Kf-Cm66 - Na	0.304	0.368	0.415	0.322	1.969	0.237
Acidobacteria - Acidobacteriia - Solibacterales	0.754	0.17	0.415	0.155	0.191	0.151
Actinobacteria - Actinobacteria - Frankiales	0.652	0.292	0.632	0.287	0.429	0.265
Bacteroidetes - Bacteroidia - Chitinophagales	0.814	0.442	0.741	0.218	0.306	0.141
Proteobacteria - Deltaproteobacteria - Myxococcales	0.725	0.257	0.77	0.305	0.29	0.193
Bacteroidetes - Bacteroidia - Sphingobacteriales	0.565	0.441	0.579	0.366	0.217	0.191
Proteobacteria - Alphaproteobacteria - Caulobacterales	0.228	0.193	0.155	0.177	0.288	0.224
Verrucomicrobia - Verrucomicrobiae - Pedosphaerales	0.61	0.361	0.606	0.525	0.676	0.506
Proteobacteria - Deltaproteobacteria - Oligoflexales	0.619	0.564	0.627	0.307	0.38	0.307
Planctomycetes - Planctomycetacia - Planctomycetales	0.397	0.198	0.367	0.514	0.151	0.092
Proteobacteria - Alphaproteobacteria - Sphingomonadales	0.034	0.086	0.014	0.158	0.33	0.18
Proteobacteria - Alphaproteobacteria - Rhodospirillales	0.674	0.276	0.623	0.221	0.278	0.25
Acidobacteria - Subgroup - 6 - Na	0.582	0.361	0.616	0.269	0.319	0.266
Cyanobacteria - Oxyphotobacteria - Pseudanabaenales	0.92	0.128	0.917	0.126	1.409	0.124
Cyanobacteria - Oxyphotobacteria - Chloroplast	0.486	0.23	0.373	0.555	0.442	0.453
Verrucomicrobia - Verrucomicrobiae - Methylacidiphilales	0.597	0.619	0.621	0.535	0.321	0.274
Chloroflexi - Olb14 - Na	0.839	0.364	0.824	0.169	0.409	0.204
Planctomycetes - Phycisphaerae - Phycisphaerales	0.486	0.296	0.581	0.571	0.645	0.273
Proteobacteria - Alphaproteobacteria - Reyranellales	0.28	0.337	0.254	0.204	0.311	0.203
MEDIAN VALUE	0.619	0.361	0.616	0.269	0.321	0.224

Line 2

ASV	IA			RMSE		
	Line 1	Line 4	Lines 1,4	Line 1	Line 4	Lines 1,4
Proteobacteria - Gammaproteobacteria - Betaproteobacteriales	0.65	0.51	0.655	0.314	0.492	0.308
Proteobacteria - Alphaproteobacteria - Rhizobiales	0.838	0.754	0.82	0.187	0.296	0.194
Proteobacteria - Alphaproteobacteria - Acetobacterales	0.602	0.537	0.572	0.462	0.47	0.408
Planctomycetes - Planctomycetacia - Gemmatales	0.631	0.487	0.577	0.204	0.322	0.228
Planctomycetes - Planctomycetacia - Isosphaerales	0.795	0.244	0.486	0.255	0.413	0.292
Cyanobacteria - Oxyphotobacteria - Synechococcales	0.635	0.399	0.714	0.257	0.607	0.23
Chloroflexi - Jg30-Kf-Cm66 - Na	0.401	0.713	0.654	1.759	0.279	0.307
Acidobacteria - Acidobacteriia - Solibacterales	0.683	0.799	0.764	0.382	0.279	0.285
Actinobacteria - Actinobacteria - Frankiales	0.632	0.769	0.746	0.272	0.159	0.164
Bacteroidetes - Bacteroidia - Chitinophagales	0.815	0.412	0.461	0.175	0.101	0.068
Proteobacteria - Deltaproteobacteria - Myxococcales	0.825	0.383	0.699	0.169	0.467	0.151
Bacteroidetes - Bacteroidia - Sphingobacteriales	0.643	0.497	0.765	0.26	0.197	0.085
Proteobacteria - Alphaproteobacteria - Caulobacterales	0.293	0.249	0.298	1.795	0.986	0.882
Verrucomicrobia - Verrucomicrobiae - Pedosphaerales	0.647	0.418	0.671	0.231	0.534	0.183
Proteobacteria - Deltaproteobacteria - Oligoflexales	0.649	0.482	0.734	0.315	0.446	0.265
Planctomycetes - Planctomycetacia - Planctomycetales	0.246	0.219	0.242	0.258	0.251	0.235
Proteobacteria - Alphaproteobacteria - Sphingomonadales	0.181	0.127	0.178	1.907	0.429	0.395
Proteobacteria - Alphaproteobacteria - Rhodospirillales	0.672	0.612	0.755	0.476	0.439	0.335
Acidobacteria - Subgroup - 6 - Na	0.546	0.778	0.718	0.415	0.185	0.182
Cyanobacteria - Oxyphotobacteria - Pseudanabaenales	0.906	0.806	0.884	0.213	0.231	0.191
Cyanobacteria - Oxyphotobacteria - Chloroplast	0.485	0.444	0.708	0.29	0.806	0.177
Verrucomicrobia - Verrucomicrobiae - Methylacidiphilales	0.62	0.372	0.666	0.216	0.354	0.208
Chloroflexi - Olb14 - Na	0.845	0.395	0.707	0.201	0.374	0.216
Planctomycetes - Phycisphaerae - Phycisphaerales	0.718	0.618	0.841	0.196	0.314	0.137
Proteobacteria - Alphaproteobacteria - Reyranellales	0.292	0.339	0.346	0.733	0.712	0.673
MEDIAN VALUE	0.643	0.482	0.699	0.26	0.374	0.228

Line 3

ASV	IA			RMSE		
	Line 1	Line 4	Lines 1,4	Line 1	Line 4	Lines 1,4
Proteobacteria - Gammaproteobacteria - Betaproteobacteriales	0.482	0.567	0.77	0.294	0.471	0.277
Proteobacteria - Alphaproteobacteria - Rhizobiales	0.564	0.589	0.781	0.257	0.357	0.199
Proteobacteria - Alphaproteobacteria - Acetobacterales	0.599	0.686	0.715	0.241	0.243	0.197
Planctomycetes - Planctomycetacia - Gemmatales	0.431	0.424	0.443	0.388	0.939	0.358
Planctomycetes - Planctomycetacia - Isosphaerales	0.666	0.418	0.606	0.481	0.653	0.479
Cyanobacteria - Oxyphotobacteria - Synechococcales	0.511	0.642	0.439	0.217	0.371	0.186
Chloroflexi - Jg30-Kf-Cm66 - Na	0.172	0.052	0.151	0.308	0.221	0.094
Acidobacteria - Acidobacteriia - Solibacterales	0.356	0.301	0.307	1.443	1.296	1.211
Actinobacteria - Actinobacteria - Frankiales	0.184	0.769	0.762	0.441	0.203	0.19
Bacteroidetes - Bacteroidia - Chitinophagales	0.712	0.654	0.752	0.383	0.201	0.165
Proteobacteria - Deltaproteobacteria - Myxococcales	0.491	0.544	0.703	0.338	0.658	0.226
Bacteroidetes - Bacteroidia - Sphingobacteriales	0.445	0.851	0.824	0.49	0.165	0.144
Proteobacteria - Alphaproteobacteria - Caulobacterales	0.467	0.378	0.431	2.968	1.741	1.566
Verrucomicrobia - Verrucomicrobiae - Pedosphaerales	0.413	0.532	0.751	0.324	0.667	0.133
Proteobacteria - Deltaproteobacteria - Oligoflexales	0.691	0.515	0.863	0.331	0.464	0.205
Planctomycetes - Planctomycetacia - Planctomycetales	0.366	0.357	0.359	5.408	8.284	5.433
Proteobacteria - Alphaproteobacteria - Sphingomonadales	0.433	0.426	0.441	3.933	0.818	0.812
Proteobacteria - Alphaproteobacteria - Rhodospirillales	0.494	0.163	0.667	0.424	0.37	0.269
Acidobacteria - Subgroup - 6 - Na	0.387	0.228	0.587	0.324	0.255	0.267
Cyanobacteria - Oxyphotobacteria - Pseudanabaenales	0.443	0.493	0.427	0.248	0.203	0.184
Cyanobacteria - Oxyphotobacteria - Chloroplast	0.254	0.454	0.326	0.327	0.751	0.274
Verrucomicrobia - Verrucomicrobiae - Methylacidiphilales	0.449	0.436	0.358	0.264	0.473	0.249
Chloroflexi - Olb14 - Na	0.535	0.56	0.733	0.293	0.336	0.243
Planctomycetes - Phycisphaerae - Phycisphaerales	0.576	0.562	0.452	0.25	0.519	0.22
Proteobacteria - Alphaproteobacteria - Reyranellales	0.437	0.432	0.496	0.981	1.036	0.879
MEDIAN VALUE	0.449	0.493	0.587	0.331	0.471	0.243

Line 4

ASV	IA			RMSE		
	Line 2	Line 3	Lines 2,3	Line 2	Line 3	Lines 2,3
Proteobacteria - Gammaproteobacteria - Betaproteobacteriales	0.457	0.627	0.544	0.338	0.214	0.254
Proteobacteria - Alphaproteobacteria - Rhizobiales	0.576	0.533	0.742	0.284	0.369	0.22
Proteobacteria - Alphaproteobacteria - Acetobacterales	0.336	0.37	0.533	0.35	0.752	0.217
Planctomycetes - Planctomycetacia - Gemmatales	0.431	0.128	0.356	0.438	0.355	0.197
Planctomycetes - Planctomycetacia - Isosphaerales	0.281	0.284	0.228	0.413	0.332	0.327
Cyanobacteria - Oxyphotobacteria - Synechococcales	0.385	0.572	0.333	0.196	0.35	0.269
Chloroflexi - Jg30-Kf-Cm66 - Na	0.189	0.336	0.884	0.471	8.758	0.103
Acidobacteria - Acidobacteriia - Solibacterales	0.531	0.025	0.711	0.265	0.294	0.07
Actinobacteria - Actinobacteria - Frankiales	0.227	0.676	0.562	0.537	0.343	0.354
Bacteroidetes - Bacteroidia - Chitinophagales	0.186	0.457	0.326	1.241	0.599	0.59
Proteobacteria - Deltaproteobacteria - Myxococcales	0.286	0.665	0.441	0.39	0.167	0.274
Bacteroidetes - Bacteroidia - Sphingobacteriales	0.137	0.741	0.735	0.834	0.256	0.251
Proteobacteria - Alphaproteobacteria - Caulobacterales	0.204	0.194	0.121	0.161	0.202	0.155
Verrucomicrobia - Verrucomicrobiae - Pedosphaerales	0.293	0.615	0.382	0.317	0.251	0.376
Proteobacteria - Deltaproteobacteria - Oligoflexales	0.423	0.626	0.502	0.356	0.321	0.349
Planctomycetes - Planctomycetacia - Planctomycetales	0.146	0.012	0.003	0.437	0.191	0.086
Proteobacteria - Alphaproteobacteria - Sphingomonadales	0.074	0.393	0.508	0.166	0.161	0.103
Proteobacteria - Alphaproteobacteria - Rhodospirillales	0.422	0.343	0.42	0.371	0.355	0.332
Acidobacteria - Subgroup - 6 - Na	0.552	0.221	0.586	0.348	0.43	0.349
Cyanobacteria - Oxyphotobacteria - Pseudanabaenales	0.799	0.344	0.811	0.193	1.503	0.123
Cyanobacteria - Oxyphotobacteria - Chloroplast	0.318	0.437	0.461	0.38	0.228	0.245
Verrucomicrobia - Verrucomicrobiae - Methylacidiphilales	0.408	0.493	0.51	0.405	0.282	0.242
Chloroflexi - Olb14 - Na	0.243	0.671	0.278	0.402	0.285	0.405
Planctomycetes - Phycisphaerae - Phycisphaerales	0.454	0.316	0.439	0.461	0.355	0.186
Proteobacteria - Alphaproteobacteria - Reyranellales	0.394	0.299	0.381	0.174	0.19	0.149
MEDIAN VALUE	0.336	0.393	0.461	0.371	0.321	0.245

Fig. 4. The influence of the pipe material and disinfection on the model performance in the crossline 7-day forecasting.

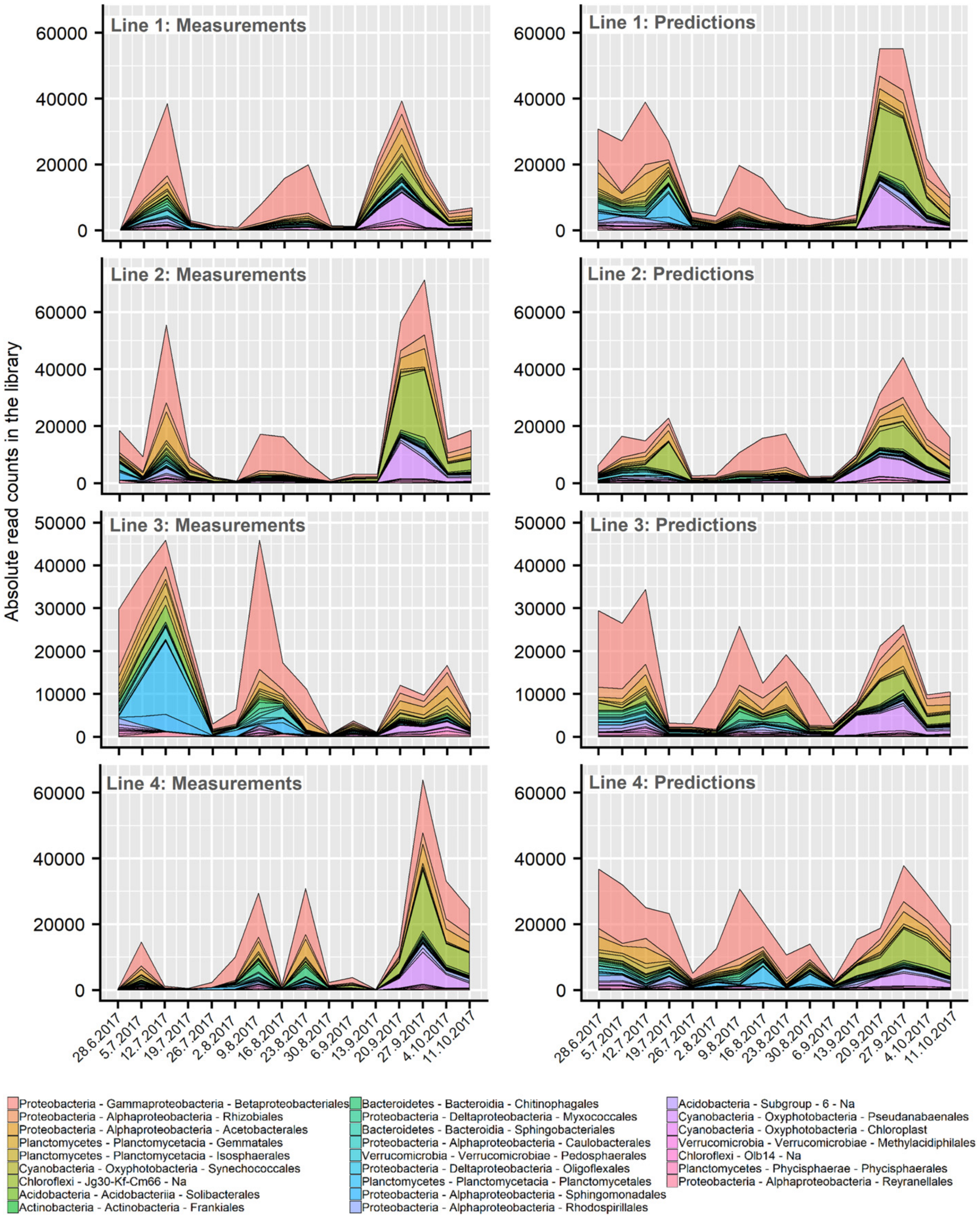


Fig. 5. Comparison of the observed bacterial abundances and the obtained 7-day forecasts.

Acetobacterales (class *Alphaproteobacteria*, phylum *Proteobacteria*), the *IA* values averaged over all the lines were higher than 0.7, which proved the potential for making even long-term predictions.

4. Discussion

Advanced learning algorithms should be involved in drinking water monitoring and risk assessment systems to train adequate data-driven models to predict microbial abundances. Such intellectual systems require reliable and representative data which describes the dynamics within the microbial community (Bersanelli et al., 2016). Learning algorithms extract substantial knowledge and patterns from collected samples, and then make generalisations and forecasts based on the set of predictor variables (Kuang et al., 2016). The model performance depends on the data quality, its accuracy and completeness, since it should reflect all crucial interconnections in the microbial community and how external factors influence it (Zuñiga et al., 2017). Therefore, large DWDS-simulation systems are needed to collect the training data properly, use it in the predictive modelling and, then, to transfer the acquired knowledge into practice.

To build the predictive models for this study, we used data collected from a pilot-scale experiment conducted in Finland and a simulated DWDS. The data included the physicochemical water characteristics, the type and amount of the disinfectants pumped, and microbial abundances in the incoming and pipeline water. In the experiments, we operated with absolute abundance as read counts since an increase or decrease in relative abundance does not necessarily mean the same change in absolute abundance (Props et al., 2017). The predictive models were designed for the bacterial community in the pipeline water. Despite the thorough set of predictor variables and regular measurements, we admit that the main limitation of the data collected is the short observational period (less than four months), which also affects the results of modelling. Thereby, the interpolated points were used not only to fill missing values in the data but were also involved in training the models to enlarge the sample size and make experiments with different prediction horizons possible.

The central idea behind this study was to test a crossline modelling scenario that complies with the real-world application of intellectual monitoring systems in different DWDS. Four pipelines with different characteristics (pipe materials and disinfection) and gathered measurements enabled us to perform this crossline modelling and investigate significant factors which influenced the model quality. The trained models were applied for nowcasting the bacterial abundances as well as for 1-, 7-, and 30-day forecasting. In all the experiments, an RF algorithm was used as the learning algorithm because of its beneficial properties for the microbial predictive modelling and due to multiple existing examples of its successful application in water research (Roguet et al., 2018; Mohammed et al., 2018). Moreover, the bacterial community presented in this study includes many groups which are identical to those revealed in full-scale DWDS (Lührig et al., 2015; Perrin et al., 2019; Wang et al., 2018b). This proves that we fulfilled the modelling under conditions close to reality.

While analysing the results of nowcasting, we found many accurate predictions of the bacterial abundances, which confirmed the presence of interconnections within the bacterial community members and relationship between the environmental factors and bacteria. For the most abundant bacteria, the trained models demonstrated high *IA* values, even reaching 0.90. Interestingly, for *Solibacterales* (class *Acidobacteriia*, phylum *Acidobacteria* (Martiny et al., 2005)) the model could make accurate predictions for lines 1, 2, and 4, whereas for line 3 the *IA* value dropped to 0.40 (Fig. 1). For the period of 5.7.2017–19.7.2017, rapid growth in the amount of *Solibacterales* was observed in line 3, in contrast to other lines where there was no sudden or large increase in the amount of this bacterium for the entire observational period (Fig. 5). This might be because of faulty measurements gathered in line 3 or due to ignoring some influential factors which could cause this rapid change. The reason for this is that the ignored factors were not included in the set of potential predictors or that they were not chosen by the learning algorithm as relevant predictors. Poor *IA* values were also obtained for two other less abundant groups: *Planctomycetales* (class *Planctomycetacia*, phylum *Planctomycetes*) and *Chloroplast* (class *Oxyphotobacteria*, phylum *Cyanobacteria*). Although *Chloroplasts* are parts of plants and do not relate to bacteria, they are often extracted when sequencing. While *Chloroplasts* are typically removed from the analysis, in our study, we deliberately retained this group since it might be an informative predictor for the amount of bacteria in the community (Wang et al., 2018a). Indeed, this predictor has a moderate importance based on the analysis we carried out (Supplementary material 4). However, the predicted abundance of *Chloroplasts* is not as accurate as the predicted abundances of bacteria.

The other essential conclusion concerning nowcasting is that the disinfectant pumped is a more significant factor than the pipe material regarding the model performance. This implies that for nowcasting bacterial abundances in a newly monitored DWDS, models trained on data collected from pipelines with the same disinfection should be applied. The possible way to enhance the accuracy of predictions is to involve data describing the bacterial community in pipelines made of the same material. The combination of two training samples leads to higher *IA* values (Fig. 2). Nevertheless, from the practical point of view, nowcasting does not allow estimations of bacterial abundances for the future, which is necessary to control risks and prevent WBOs, therefore, forecasting is required.

Although the quality of 7-day predictions appeared to be worse than for nowcasting (Fig. 3), there were still some bacteria, for which the RF model could achieve 0.70–0.80 of *IA*. For one of the most abundant bacteria, namely *Rhizobiales*, the best *IA* values varied from 0.74 to 0.92 for different lines. It was shown in another study that the *Rhizobiales* order tends to dominate other bacteria in copper pipelines (Inkinen et al., 2016). Wang et al. (2018b) also reported that *Rhizobiales* was the most abundant group in samples taken from a full-scale DWDS in eastern China.

For *Pseudanabaenales* the highest *IA* value averaged over the lines surpassed the level of 0.80. However, a further analysis revealed that the *IA* was equal to 0.95, 0.91, 0.49, and 0.90 for lines 1, 2, 3, and 4, correspondingly. The reason for the poor *IA* for line 3 was that the model overestimated the bacterial abundance heavily: in other lines, much more dramatic increases in the amount of *Pseudanabaenales* were observed (Fig. 5). Another study reported that the *Pseudanabaenales* order was one of the most abundant bacteria in water samples from Lake Zurich used as a source of drinking water (Monchamp et al., 2016).

For seven more bacteria, the averaged *IA* value exceeded 0.70. The least accurate predictions were made for *Planctomycetales* (class *Planctomycetacia*, phylum *Planctomycetes*) and *Caulobacterales* (class *Alphaproteobacteria*, phylum *Proteobacteria*): for which the *IA* did not surpass the level of 0.40 (Li et al., 2010). However, by making many quite accurate predictions, we have demonstrated the potential of forecasting in DWDS and have highlighted the existing limitations related to the data quality and its completeness.

As opposed to nowcasting, the pipe material becomes a more significant factor for crossline modelling than disinfection. This should be taken into account first while selecting the training samples. Using data which matches both criteria, the same pipe material and the same disinfection, is even more beneficial. This conclusion also corresponds to the one drawn for nowcasting. However, these conclusions should be investigated deeply with data from full-scale experiments where the disinfection is usually applied constantly for a long time over years. In real world cases the pipelines also operate for many years so the pipe effect may vanish over time.

Furthermore, the analysis of the most important predictors for the 7-day forecasting showed that the physicochemical variables had the highest importance on average. This supports many existing studies which discuss the influence of pH (Ratzke and Gore, 2018), temperature

(Jin et al., 2018), and Cu (Ladomersky and Petris, 2015) on the microbial population (Stanish et al., 2016). Nevertheless, due to the complex interactions in the community, the set of important predictors for each bacterium includes a number of specific variables, which are relevant only for a particular species (see in Supplementary material 4, Fig. 1). Apparently, knowledge-based preselection of possible predictors might create limitation for the learning algorithm and add bias to the model (Tomperi and Leiviskä, 2018).

Accurate long-term predictions are practically helpful in managing risks of contamination. They allow less frequent measurements and consequently they reduce the costs of monitoring. However, increasing the prediction horizon complicates the forecasting problem and affects the accuracy of predictions (see in Supplementary material 5, Fig. 2). Therefore, another approach based on a chain of short-term forecasts should be tested as an alternative to the presented one: the predicted bacterial abundances might be used as model inputs to make "next-step" predictions.

To summarize our findings with respect to the practical employment of the RF model, we would like to emphasize the following points:

- In crossline modelling, training data is recommended to include samples from pipelines made of the same material and with the same disinfection as the pipeline, to which the model will be applied.
- In addition to bacterial abundances, physicochemical variables such as pH, Cu, and temperature should be considered as the most important predictors.
- 7 days is the reasonable prediction horizon for RF. The 30-day forecasting does not estimate the bacterial abundances that accurately.

In further experiments, biofilm samples should be involved in the modelling to describe the investigated bacterial community thoroughly (Liu et al., 2012). Furthermore, a longer observational period should enable more data to be collected and would make the training samples more informative.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2020.137249.

CRediT authorship contribution statement

Christina Brester: Methodology, Software, Writing - original draft. **Ivan Ryzhikov:** Methodology, Software, Data curation, Writing - review & editing. **Sallamaari Siponen:** Investigation, Resources, Writing - review & editing. **Balamuralikrishna Jayaprakash:** Data curation. **Jenni Ikonen:** Investigation, Resources. **Tarja Pitkänen:** Conceptualization, Supervision, Writing - review & editing. **Ilkka T. Miettinen:** Project administration, Conceptualization. **Eila Torvinen:** Funding acquisition, Supervision, Writing - review & editing. **Mikko Kolehmainen:** Conceptualization, Supervision, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank the staff of the Department of Environmental Engineering, at Savonia University of Applied Sciences for planning and building the experimental DWDS, and Mr. Robert Ciszek (A.I. Virtanen Institute for Molecular Sciences, UEF) for participating in the design of the experiments. We are grateful for the laboratory assistance of Mr. Matti Ruokolainen (Department of Environmental and Biological Sciences, UEF) and the staff of the Department of Health Security, National Institute for Health and Welfare. We acknowledge the financial support of the Academy of Finland [277639] and the Finnish Water Utilities Association [D9H23FDLF7HG].

References

Asgari, E., Garakani, K., McHardy, A.C., Mofrad, M.R.K., 2019. MicroPheno: predicting environments and host phenotypes from 16S rRNA gene sequencing using a k-mer based representation of shallow sub-samples. Bioinformatics 34 (13), i32–i42. https://doi.org/10.1093/bioinformatics/bty296.

Baudron, P., Alonso-Sarría, F., García-Aróstegui, J.L., Cánovas-García, F., Martínez-Vicente, D., Moreno-Brotóns, J., 2013. Identifying the origin of groundwater samples in a multi-layer aquifer system with random forest classification. J. Hydrol. 499, 303–315. https://doi.org/10.1016/j.jhydrol.2013.07.009.

Benedict, K.M., Reses, H., Vigar, M., Roth, D.M., Roberts, V.A., Mattioli, M., Cooley, L.A., Hilborn, E.D., Wade, T.J., Fullerton, K.E., Yoder, J.S., Hill, V.R., 2017. Surveillance for waterborne disease outbreaks associated with drinking water – United States, 2013–2014. MMWR Morb. Mortal. Wkly Rep. 66, 1216–1221. https://doi.org/10.15585/mmwr.mm6644a3.

Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G.C., Milanesi, L., 2016. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics 17, S15. https://doi.org/10.1186/s12859-015-0857-9.

Borden, C., Roy, D., 2015. Water quality monitoring system design. http://www.iisd.org/sites/default/files/publications/water-quality-monitoring-system-design.pdf, Accessed date: 19 June 2019.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA 978-0-412-04841-8.

Bridle, H., Balharry, D., Gaiser, B., Johnston, H., 2015. Exploitation of nanotechnology for the monitoring of waterborne pathogens: state-of-the-art and future research priorities. Environ. Sci. Technol. 49 (18), 10762–10777. https://doi.org/10.1021/acs.est.5b01673.

Callahan, B.J., McMurdie, P.J., Holmes, S.P., 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J 11, 2639–2643. https://doi.org/10.1038/ismej.2017.119.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J., Knight, R., 2010. QIIME allows analysis of high-throughput community sequencing data. Nat. Methods 7 (5), 335–336. https://doi.org/10.1038/nmeth.f.303.

De Clercq, D., Smith, K., Chou, B., Gonzalez, A., Kothapalle, R., Li, C., Dong, X., Liu, S., Wen, Z., 2018. Identification of urban drinking water supply patterns across 627 cities in China based on supervised and unsupervised statistical learning. J. Environ. Manag. 223, 658–667. https://doi.org/10.1016/j.jenvman.2018.06.073.

Díaz-Uriarte, R., Alvarez de Andrés, S., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7, 3. https://doi.org/10.1186/1471-2105-7-3.

Douterelo, I., Boxall, J.B., Deines, P., Sekar, R., Fish, K.E., Biggs, C.A., 2014. Methodological approaches for studying the microbial ecology of drinking water distribution systems. Water Res. 65, 134–156. https://doi.org/10.1016/j.watres.2014.07.008.

Douterelo, I., Jackson, M., Solomon, C., Boxall, J., 2016. Microbial analysis of in situ biofilm formation in drinking water distribution systems: implications for monitoring and control of drinking water quality. Appl. Microbiol. Biotechnol. 100, 3301–3311. https://doi.org/10.1007/s00253-015-7155-3.

Figueras, M.J., Borrego, J.J., 2010. New perspectives in monitoring drinking water microbial quality. Int. J. Environ. Res. Public Health 7 (12), 4179–4202.

Fish, K.E., Boxall, J.B., 2018. Biofilm microbiome (re)growth dynamics in drinking water distribution systems are impacted by chlorine concentration. Front. Microbiol. 9, 2519. https://doi.org/10.3389/fmicb.2018.02519.

Fish, K.E., Collins, R., Green, N.H., Sharpe, R.L., Douterelo, I., Osborn, A.M., Boxall, J.B., 2015. Characterisation of the physical composition and microbial community structure of biofilms within a model full-scale drinking water distribution system. PLoS One 10 (2), e0115824. https://doi.org/10.1371/journal.pone.0115824.

Fish, K., Osborn, A.M., Boxall, J., 2016. Characterising and understanding the impact of microbial biofilms and the extracellular polymeric substance (EPS) matrix in drinking water distribution systems. Environ. Sci.: Water Res. Technol. 2 (4), 614–630. https://doi.org/10.1039/C6EW00039H.

Frick, W.E., Ge, Z., Zepp, R.G., 2008. Nowcasting and forecasting concentrations of biological contaminants at beaches: a feasibility and case study. Environ. Sci. Technol. 42 (13), 4818–4824. https://doi.org/10.1021/es703185p.

Fritch, F.N., Carlson, R.E., 1980. Monotone piecewise cubic interpolation. SIAM J. Numer. Anal. 17 (2), 238–246. https://doi.org/10.1137/0717021.

Genuer, R., Poggi, J.M., Tuleau-Malot, C., 2010. Variable selection using random forests. Pattern Recogn. Lett. 31 (14), 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014.

Gilfillan, D., Hall, K., Joyner, T.A., Scheuerman, P., 2018. Canonical variable selection for ecological modelling of fecal indicators. J. Environ. Qual. 47, 974–984. https://doi.org/10.2134/jeq2017.12.0474.

Goldstein, B.A., Briggs, F.B.S., Polley, E.C., 2011. Random forests for genetic association studies. Stat. Appl. Genet. Mol. Biol. 10, 1–34. https://doi.org/10.2202/1544-6115.1691.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. JMLR Special Issue on Variable and Feature Selection 3, 1157–1182.

Hastie, T., Tibshirani, R., Friedman, J.H., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag New York Inc.

Ikonen, J., Pitkänen, T., Miettinen, I.T., 2013. Suitability of optical, physical and chemical measurements for detection of changes in bacterial drinking water quality. Int. J. Environ. Res. Public Health 10 (11), 5349–5363. https://doi.org/10.3390/ijerph10115349.

Ikonen, J.M., Hokajärvi, A.M., Heikkinen, J., Pitkänen, T., Ciszek, R., Kolehmainen, M., Pursiainen, A., Kauppinen, A., Kusnetsov, J., Torvinen, E., Miettinen, I.T., 2017. Drinking water quality in distribution systems of surface and ground water works in Finland. J. Water Secur. 3. https://doi.org/10.15544/jws.2017.004 jws2017004.

Inkinen, J., Jayaprakash, B., Santo Domingo, J.W., Keinänen-Toivola, M.M., Ryu, H., Pitkänen, T., 2016. Diversity of ribosomal 16S DNA- and RNA-based bacterial community in an office building drinking water system. J. Appl. Microbiol. 120, 1723–1738. https://doi.org/10.1111/jam.13144.

Inkinen, J., Jayaprakash, B., Ahonen, M., Pitkänen, T., Mäkinen, R., Pursiainen, A., Santo Domingo, J.W., Salonen, H., Elk, M., Keinänen-Toivola, M.M., 2018. Bacterial community changes in copper and PEX drinking water pipeline biofilms under extra disinfection and magnetic water treatment. J. Appl. Microbiol. 124 (2), 611–624. https://doi.org/10.1111/jam.13662.

Inkinen, J., Jayaprakash, B., Siponen, S., Hokajärvi, A.-M., Pursiainen, A., Ikonen, J., Ryzhikov, I., Täubel, M., Kauppinen, A., Paananen, J., Miettinen, I.T., Torvinen, E., Kolehmainen, M., Pitkänen, T., 2019. Active eukaryotes in drinking water distribution systems of ground and surface waterworks. Microbiome 7 (1), 99. https://doi.org/10.1186/s40168-019-0715-5.

Jin, D., Kong, X., Cui, B., Jin, S., Xie, Y., Wang, X., Deng, Y., 2018. Bacterial communities and potential waterborne pathogens within the typical urban surface waters. Sci. Rep. 8, 13368. https://doi.org/10.1038/s41598-018-31706-w.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., Glöckner, F.O., 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucl. Acids Res. 41 (1), e1. https://doi.org/10.1093/nar/gks808.

Kuang, J., Huang, L., He, Z., Chen, L., Hua, Z., Jia, P., Li, S., Liu, J., Li, J., Zhou, J., Shu, W., 2016. Predicting taxonomic and functional structure of microbial communities in acid mine drainage. ISME J 10 (6), 1527–1539. https://doi.org/10.1038/ismej.2015.201.

Ladomersky, E., Petris, M.J., 2015. Copper tolerance and virulence in bacteria. Metallomics 7 (6), 957–964. https://doi.org/10.1039/c4mt00327f.

Li, D., Li, Z., Yu, J., Cao, N., Liu, R., Yang, M., 2010. Characterization of bacterial community structure in a drinking water distribution system during an occurrence of red water. Appl. Environ. Microbiol. 76 (21), 7171–7180. https://doi.org/10.1128/AEM.00832-10.

Liu, R., Yu, Z., Guo, H., Liu, M., Zhang, H., Yang, M., 2012. Pyrosequencing analysis of eukaryotic and bacterial communities in faucet biofilms. Sci. Total Environ. 435–436, 124–131. https://doi.org/10.1016/j.scitotenv.2012.07.022.

Louppe, G., 2014. Understanding Random Forests: From Theory to Practice. PhD Thesis. University of Liege https://doi.org/10.13140/2.1.1570.5928.

Lührig, K., Canbäck, B., Paul, C.J., Johansson, T., Persson, K.M., Rådström, P., 2015. Bacterial community analysis of drinking water biofilms in southern Sweden. Microbes Environ. 30 (1), 99–107. https://doi.org/10.1264/jsme2.ME14123.

Martiny, A.C., Albrechtsen, H.J., Arvin, E., Molin, S., 2005. Identification of bacteria in biofilm and bulk water samples from a nonchlorinated model drinking water distribution system: detection of a large nitrite-oxidizing population associated with Nitrospira spp. Appl. Environ. Microbiol. 71 (12), 8611–8617. https://doi.org/10.1128/AEM.71.12.8611-8617.2005.

Mohammed, H., Seidu, R., 2019. Climate-driven QMRA model for selected water supply systems in Norway accounting for raw water sources and treatment processes. Sci. Total Environ. 660, 306–320. https://doi.org/10.1016/j.scitotenv.2018.12.460.

Mohammed, H., Hameed, I.A., Seidu, R., 2018. Comparative predictive modelling of the occurrence of faecal indicator bacteria in a drinking water source in Norway. Sci. Total Environ. 628–629, 1178–1190. https://doi.org/10.1016/j.scitotenv.2018.02.140.

Monchamp, M.E., Walser, J.C., Pomati, F., Spaak, P., 2016. Sedimentary DNA reveals cyanobacterial community diversity over 200 years in two perialpine lakes. Appl. Environ. Microbiol. 82 (21), 6472–6482. https://doi.org/10.1128/AEM.02174-16.

Moreira, N.A., Bondelind, M., 2017. Safe drinking water and waterborne outbreaks. J. Water Health 15 (1), 83–96. https://doi.org/10.2166/wh.2016.103.

Muharemi, F., Logofătu, D., Leon, F., 2019. Machine learning approaches for anomaly detection of water quality on a real-world data set. Journal of Information and Telecommunication 3 (3), 294–307. https://doi.org/10.1080/24751839.2019.1565653.

Nembrini, S., König, I.R., Wright, M.N., 2018. The revival of the Gini importance? Bioinformatics 34 (21), 3711–3718. https://doi.org/10.1093/bioinformatics/bty373.

Pachepsky, Y.A., Allende, A., Boithias, L., Cho, K., Jamieson, R., Hofstra, N., Molina, M., 2018. Microbial water quality: monitoring and modelling. J. Environ. Qual. 47 (5), 931–938. https://doi.org/10.2134/jeq2018.07.0277.

Palamuleni, L., Akoth, M., 2015. Physico-chemical and microbial analysis of selected borehole water in Mahikeng, South Africa. Int. J. Environ. Res. Public Health 12 (8), 8619–8630. https://doi.org/10.3390/ijerph120808619.

Parkhurst, D.F., Brenner, K.P., Dufour, A.P., Wymer, L.J., 2005. Indicator bacteria at five swimming beaches-analysis using random forests. Water Res. 39 (7), 1354–1360. https://doi.org/10.1016/j.watres.2005.01.001.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. JMLR 12, 2825–2830.

Perrin, Y., Bouchon, D., Delafont, V., Moulin, L., Héchard, Y., 2019. Microbiome of drinking water: a full-scale spatio-temporal study to monitor water quality in the Paris distribution system. Water Res. 149, 375–385. https://doi.org/10.1016/j.watres.2018.11.013.

Peters, J., De Baets, B., Verhoest, N., Samson, R., Degroeve, S., De Becker, P., Huybrechts, W., 2007. Random forests as a tool for ecohydrological distribution modelling. Ecol. Model. 207 (2–4), 304–318.

Props, R., Kerckhof, F.M., Rubbens, P., De Vrieze, J., Hernandez Sanabria, E., Waegeman, W., Monsieurs, P., Hammes, F., Boon, N., 2017. Absolute quantification of microbial taxon abundances. ISME J 11 (2), 584–587. https://doi.org/10.1038/ismej.2016.117.

Quinlan, J.R., 1986. Induction of decision trees. Mach. Learn. 1, 81–106.

Ratzke, C., Gore, J., 2018. Modifying and reacting to the environmental pH can drive bacterial interactions. PLoS Biol. 16 (3), e2004248. https://doi.org/10.1371/journal.pbio.2004248.

Ridenhour, B.J., Brooker, S.L., Williams, J.E., Van Leuven, J.T., Miller, A.W., Dearing, M.D., Remien, C.H., 2017. Modelling time-series data from microbial communities. ISME J 11 (11), 2526–2537. https://doi.org/10.1038/ismej.2017.107.

Roguet, A., Eren, A.M., Newton, R.J., McLellan, S.L., 2018. Fecal source identification using random forest. Microbiome 6 (1), 185. https://doi.org/10.1186/s40168-018-0568-3.

Siponen, S., Ikonen, J., Hokajärvi, A.-M., Ruokolainen, M., Jayaprakash, B., Ryzhikov, I., Inkinen, J., Pitkänen, T., Paananen, J., Kolehmainen, M., Miettinen, I., Torvinen, E., 2020. Effect of the Used Pipe Material and Disinfection Chemical on Activity and Structure of Bacterial Communities in Pilot-Scale Drinking Water Distribution System unpublished results.

Stanish, L.F., Hull, N.M., Robertson, C.E., Harris, J.K., Stevens, M.J., Spear, J.R., Pace, N.R., 2016. Factors influencing bacterial diversity and community composition in municipal drinking waters in the Ohio River basin, USA. PLoS One 11, e0157966. https://doi.org/10.1371/journal.pone.0157966.

Tomperi, J., Leiviskä, K., 2018. Utilizing variable selection methods in modelling the potable water quality. Water Supply 19 (4), 1187–1194. https://doi.org/10.2166/ws.2018.173.

Tyralis, H., Papacharalampous, G., 2017. Variable selection in time series forecasting using random forests. Algorithms 10 (4), 114. https://doi.org/10.3390/a10040114.

Vincenzi, S., Zucchetta, M., Franzoi, P., Pellizzato, M., Pranovi, F., De Leo, G.A., Torricelli, P., 2011. Application of a Random Forest algorithm to predict spatial distribution of the potential yield of Ruditapes philippinarum in the Venice lagoon, Italy. Ecol. Model. 222 (8), 1471–1478. https://doi.org/10.1016/j.ecolmodel.2011.02.007.

Wang, F., Men, X., Zhang, G., Liang, K.C., Xin, Y., Wang, J., Li, A., Zhang, H., Liu, H., Wu, L., 2018a. Assessment of 16S rRNA gene primers for studying bacterial community structure and function of aging flue-cured tobaccos. AMB Express 8 (1), 182. https://doi.org/10.1186/s13568-018-0713-1.

Wang, F.H., Li, W., Li, Y.H., Zhang, J., Chen, J., Zhang, W., Wu, X., 2018b. Molecular analysis of bacterial community in the tap water with different water ages of a drinking water distribution system. Front. Environ. Sci. Eng. 12, 1–10. https://doi.org/10.1007/s11783-018-1020-4.

Willmott, C.J., 1981. On the validation of models. Phys. Geogr. 2, 184–194. https://doi.org/10.1080/02723646.1981.10642213.

Willmott, C.J., Ackleson, S.G., Davis, R.E., Feddema, J.J., Klink, K.M., Legates, D.R., O'Donnell, J., Rowe, C.M., 1985. Statistics for the evaluation and comparison of models. J. Geophys. Res. 900 (C5), 8995–9005. https://doi.org/10.1029/JC090iC05p08995.

World Meteorological Organization, 2013. Planning of water quality monitoring systems. Tech. Rep. Ser. – No.3. WMO 2013 (WMO, No. 1113). Geneva, Switzerland https://library.wmo.int/pmb_ged/wmo_1113_en.pdf, Accessed date: 19 June 2019.

Wu, Z.Y., Rahman, A., 2017. Optimized deep learning framework for water distribution data-driven modelling. Procedia Eng 186, 261–268. https://doi.org/10.1016/j.proeng.2017.03.240.

Zhang, J., Qiu, H., Li, X., Niu, J., Nevers, M.B., Hu, X., Phanikumar, M.S., 2018. Real-time nowcasting of microbiological water quality at recreational beaches: a wavelet and artificial neural network-based hybrid modelling approach. Environ. Sci. Technol. 52 (15), 8446–8455. https://doi.org/10.1021/acs.est.8b01022.

Zheng, A., Casari, A., 2018. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472 978-1-491-95324-2.

Zuñiga, C., Zaramela, L., Zengler, K., 2017. Elucidation of complexity and prediction of interactions in microbial communities. Microb. Biotechnol. 10 (6), 1500–1522. https://doi.org/10.1111/1751-7915.12855.