PLOS WATER

RESEARCH ARTICLE

Modelling point-of-consumption residual chlorine in humanitarian response: Can cost-sensitive learning improve probabilistic forecasts?

Michael De Santi[1,2]*, **Syed Imran Ali**[1,2], **Matthew Arnold**[2], **Jean-François Fesselet**[3], **Anne M. J. Hyvärinen**[4], **Dawn Taylor**[3], **Usman T. Khan**[1]

1 Department of Civil Engineering, Lassonde School of Engineering, York University, Toronto, Canada, **2** Dahdaleh Institute for Global Health Research, York University, Toronto, Canada, **3** Public Health Department, Médecins Sans Frontières, Amsterdam, The Netherlands, **4** Division of Resilience and Solutions, United Nations High Commissioner for Refugees, Genève, Switzerland

* desantim@yorku.ca

Check for updates

OPEN ACCESS

Citation: De Santi M, Ali SI, Arnold M, Fesselet J-F, Hyvärinen AMJ, Taylor D, et al. (2022) Modelling point-of-consumption residual chlorine in humanitarian response: Can cost-sensitive learning improve probabilistic forecasts? PLOS Water 1(9): e0000040. https://doi.org/10.1371/journal.pwat.0000040

Editor: Manuel Herrera, University of Cambridge, UNITED KINGDOM

Received: May 16, 2022

Accepted: July 29, 2022

Published: September 6, 2022

Copyright: © 2022 De Santi et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The code to train, test, and evaluate the ANN-EFS is available on the SWOT Project's GitHub at the following link: https://github.com/safeh2o/swot_research_cost_sensitive. This repository also includes the cleaned Tanzania and Bangladesh datasets.

Funding: Field data collection was supported by the Achmea Foundation (https://www.achmea.nl/en/foundation/) (SIA & JF, grant no: 2018.001). Research funding was provided by Natural

Abstract

Ensuring sufficient free residual chlorine (FRC) up to the time and place water is consumed in refugee settlements is essential for preventing the spread of waterborne illnesses. Water system operators need accurate forecasts of FRC during the household storage period. However, factors that drive FRC decay after water leaves the piped distribution system vary substantially, introducing significant uncertainty when modelling point-of-consumption FRC. Artificial neural network (ANN) ensemble forecasting systems (EFS) can account for this uncertainty by generating probabilistic forecasts of point-of-consumption FRC. ANNs are typically trained using symmetrical error metrics like mean squared error (MSE), but this leads to forecast underdispersion forecasts (the spread of the forecast is smaller than the spread of the observations). This study proposes to solve forecast underdispersion by training an ANN-EFS using cost functions that combine alternative metrics (Nash-Sutcliffe efficiency, Kling Gupta Efficiency, Index of Agreement) with cost-sensitive learning (inverse FRC weighting, class-based FRC weighting, inverse frequency weighting). The ANN-EFS trained with each cost function was evaluated using water quality data from refugee settlements in Bangladesh and Tanzania by comparing the percent capture, confidence interval reliability diagrams, rank histograms, and the continuous ranked probability. Training the ANN-EFS using the cost functions developed in this study produced up to a 70% improvement in forecast reliability and dispersion compared to the baseline cost function (MSE), with the best performance typically obtained by training the model using Kling-Gupta Efficiency and inverse frequency weighting. Our findings demonstrate that training the ANN-EFS using alternative metrics and cost-sensitive learning can improve the quality of forecasts of point-of-consumption FRC and better account for uncertainty in post-distribution chlorine decay. These techniques can enable humanitarian responders to ensure sufficient FRC more reliably at the point-of-consumption, thereby preventing the spread of waterborne illnesses.

Sciences and Engineering Research Council of Canada (NSERC, https://www.nserc-crsng.gc.ca/index_eng.asp) (UTK, grant no: RGPIN-2017-05661) and by ELRHA's Humanitarian Innovation Fund/(https://www.elrha.org/programme/hif/) (SIA, grant no: WASH Evidence Challenge 50642). The Safe Water Optimization Tool (SWOT) is supported by Creating Hope in Conflict: A Humanitarian Grand Challenge (https://humanitariangrandchallenge.org/): a partnership of USAID, the UK Government, the Ministry of Foreign Affairs of the Netherlands, and Global Affairs Canada, with support from Grand Challenges Canada (SIA, grant no: R-HGC-POC-1803-22449). MDS received graduate research funding from York University (https://www.yorku.ca/) and an NSERC Canada Graduate Scholarship – Masters The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The views, opinions and policies expressed do not necessarily reflect the views, opinions, and policies of funding partners.

Competing interests: The authors have declared that no competing interests exist.

1 Introduction

Waterborne illnesses are one of the leading causes of infectious disease outbreaks in humanitarian response settings [1]. In refugee and internally displaced persons (IDP) settlements, water-users typically do not have drinking water piped to their premises; instead, they collect water from public distribution points (tapstands) which they then transport, store, and use over time in their dwellings. Recontamination of previously safe drinking water during this post-distribution period of collection, transport, and storage is an important factor in waterborne illness outbreaks, having been linked to outbreaks of cholera, hepatitis E, and shigellosis in refugee and IDP settlements in Kenya, Malawi, Sudan, South Sudan, and Uganda [2–9]. To prevent outbreaks in refugee and IDP settlements, drinking water needs to be protected against pathogenic recontamination until the end of the household storage period when the final cup is consumed. Global drinking water quality guidelines recommend providing at least 0.2 mg/L of free residual chlorine (FRC) throughout the post-distribution period to prevent recontamination, and past research has identified that this is sufficient to prevent recontamination by priority pathogens in humanitarian settings such as cholera and hepatitis E [10–15]. Thus, water system operators must determine how much FRC is needed at the point-of-distribution to ensure that there is still at least 0.2 mg/L of FRC at the point-of-consumption. To do this, they require models that can accurately predict FRC concentrations throughout the household storage period.

Recent studies have developed deterministic, process-based models of FRC decay during household storage that output point predictions of the post-distribution FRC concentration [6, 16, 17]. However, deterministic models cannot quantify the uncertainty in post-distribution FRC decay. Water stored in the household is essentially an open system, so chlorine decay can be influenced by a range of factors including environmental, water quality, and water handling factors. This leads to a high degree of variability and uncertainty when modelling post-distribution FRC decay as a single set of conditions at the point-of-distribution can produce a range of FRC concentrations at the point-of-consumption [18]. In this context, point predictions produced by deterministic models are not appropriate, and probabilistic modelling approaches are required that can predict the distribution of probable point-of-consumption FRC concentrations. However, probabilistic methods are not commonly used to model chlorine decay, and when they are, they are typically used to improve the robustness of the model calibration process, not to output probabilistic predictions of chlorine decay [19, 20].

Ensemble forecasting systems (EFSs) are a common type of probabilistic model that groups together point predictions from multiple models into a probability distribution [21]. Whereas deterministic models seek to find a single best prediction, an EFS aims to reproduce the underlying distribution of the observed data and quantify the uncertainty in the modelled processes [21]. While EFSs are often formed from a collection of physical process-based models, they can also be formed using data-driven models [21–23]. Data-driven modelling, including machine learning or artificial intelligence, is increasingly being used to predict and monitor a range of drinking water treatment and distribution processes [24–28]. Recent research has used data-driven modelling for a complex range of tasks, e.g., controlling dosing of chlorine [29] and other oxidants [30], predicting disinfection by-product formation [31], optimizing cyanide removal [32], and detecting bacterial growth in water samples using image analysis [33]. These models have been used for over two decades to model chlorine residuals in distribution systems, either as standalone models [34–38] or as part of a hybrid data-driven and process-based modelling system [19]. One of the most common and effective branches of data-driven models used in drinking water–especially for modelling chlorine residuals–are artificial neural networks (ANNs) [27, 30, 34–36, 38], though none of these previous studies

have modelled chlorine residuals in the post-distribution period. ANNs have been used for probabilistic modelling in an EFS [21, 22], though we are not aware of any ANN-EFS being used in drinking water quality modelling, beyond our previous work which used an ANN-EFS to generate risk-based FRC targets by predicting the probability of water users having insufficient point-of-consumption FRC [18]. This modelling approach was incorporated into the Safe Water Optimization Tool (SWOT [39]), a web-based modelling tool that generates evidence-based, site-specific FRC guidance for water system operators in humanitarian response settings. A limitation of this earlier approach is that the probabilistic forecasts were underdispersed: the spread of the ensemble forecast was smaller than the spread of the observations. This decreased the forecast reliability (the similarity between the forecast probability distribution and the underlying distribution of the observations) and the model's ability to predict high-risk events when the point-of-consumption FRC was below 0.2 mg/L, reducing the accuracy of risk-based FRC targets [18].

The underdispersion observed in this earlier work may have been at least, in part, due to the use of mean squared error (MSE) as the cost function to train the ANN-EFS, as this produced a regression to the mean-type behaviour for the ensemble forecast [18]. An ANN's cost function measures the difference between the model predictions and the true values of the target variable. During training, the model is calibrated to minimize this difference [40]. While symmetrical error metrics like MSE or MAE are common cost functions for ANNs, they prioritize performance at the average (mean or median, depending on the metric) of the observations, not for the whole output space [41, 42]. For an EFS, the predictions of the individual models should form a representative sample of the whole distribution of the observations [43], not just the average. Thus, alternative cost functions that prioritize prediction of the whole distribution of observations and not just the average, are needed for training an ANN-EFS.

There are two main approaches to overcoming the limitations of symmetrical error metrics when training ANNs. One is to train the ANN using alternative error metrics [44]. The other is through cost-sensitive learning. Cost-sensitive learning encompasses multiple approaches used to alter the training of machine learning models to prioritize a specific region of the output space or a specific behaviour. Common cost-sensitive learning approaches involve either resampling from high-priority classes, changing a decision threshold in classification models, or reweighting the cost function itself to reflect the cost of misprediction [45, 46]. In cost-sensitive learning, the cost function becomes the combination of an error metric, symmetrical or otherwise, and a weighting. Alternative error metrics and cost-sensitive learning have been applied for regression and classification modelling to predict rare or high-priority events [47] such as flooding [45, 48, 49]; fraudulent credit card purchases [50]; fault detection in machinery [51]; cholera cases [52] and for differentiating between benign and malignant cysts for cancer detection [53]. They have even been applied for anomaly detection and compliance monitoring in water treatment [54, 55]. However, these methods have not been applied to probabilistic models like EFS.

This study sought to investigate whether modifying the cost function used to train an ANN-EFS by combining alternative error metrics and cost-sensitive learning techniques could resolve the problem of underdispersion and improve the reliability of point-of-consumption FRC forecasts. Our first objective was to evaluate the effect of training an ANN-EFS using alternative error metrics and cost-sensitive learning on the model's probabilistic performance. Our second objective was to identify the cost function that produced the best performance when forecasting point-of-consumption FRC in humanitarian response settings. This is the first study, to the authors' knowledge, to use these approaches when modelling FRC during the post-distribution period. Achieving these objectives will improve the reliability of point-of-consumption FRC forecasts and, thus, the accuracy of risk-based chlorination guidance

provided by the SWOT. This, in turn, will help humanitarian responders ensure that water remains protected against pathogenic recontamination up to when the final cup is consumed.

2 Materials and methods

The following section provides an overview of the datasets used in our modelling and the model development procedures. Additionally, we describe the alternative error metrics and cost-sensitive learning approaches selected for investigation in this study, as well as the metrics we used to evaluate the forecasting performance of the ANN-EFS.

2.1 Ethics statement

Field data collection for the datasets used in this study received approval from the Human Participants Review Committee, Office of Research Ethics at York University (Certificate #: 2019–186). Data collection in Bangladesh also received approval from the MSF Ethics Review Board (ID #: 1932), and the Centre for Injury Prevention and Research Bangladesh (Memo #: CIPRB/Admin/2019/168). All water quality samples were collected only when informed consent was provided by the water user.

2.1.1 Inclusivity in global research. Additional information regarding the ethical, cultural, and scientific considerations specific to inclusivity in global research is included in the S1 Checklist.

2.2 Description of study sites and data sets

This study used routine water quality monitoring data from two refugee settlements in Bangladesh and Tanzania collected through the SWOT Project The Bangladesh dataset was collected by Médecins Sans Frontières (MSF) from Camp 1 of the Kutupalong-Balukhali Extension Site, Cox's Bazaar, where 2,130 samples were collected between June and December 2019. At the time of data collection, the site hosted 83,000 Rohingya refugees from neighbouring Myanmar. This site used groundwater obtained from 14 boreholes equipped with inline chlorination using high-test calcium hypochlorite (HTH). The Tanzania dataset was collected by the United Nations High Commissioner for Refugees (UNHCR) and the Norwegian Refugee Council (NRC) at the Nyaragusu Refugee Settlement, where 305 samples were collected between December 2019 and January 2020. This settlement hosted over 130,000 refugees from Burundi and the Democratic Republic of Congo at the time of data collection. Water was obtained from both groundwater and surface water sources subject to inline chlorination using HTH.

At both sites, FRC was measured at the point-of-distribution immediately before collection and then again in the same unit of water at the point-of-consumption after a follow-up period ranging from 1 to 30 hours. Thus, each observation consisted of two paired water quality measurements from the point-of-distribution and point-of-consumption. The elapsed time for each observation was calculated from timestamps for the two measurements. In addition to FRC, at the Bangladesh site, total residual chlorine, electrical conductivity (EC), pH, turbidity, water temperature, and water handling behaviours were collected both at the point-of-distribution and the point-of-consumption. At the Tanzania site, only FRC, EC, and water temperature were collected at the point-of-distribution and only FRC was collected at the point-of consumption. The main type of error observed in the collected data was incomplete records, where one or more water quality parameters were missing at the point-of-distribution. This could have arisen due to equipment malfunction or lack of reagents. At both sites, more than half of the samples were missing measurements for one water quality parameter other than FRC (1,513 incomplete records in Bangladesh and 216 in Tanzania).

Since the paired water quality samples were collected as a part of the overall water system operations at each site, there was not a fixed water quality sampling schedule. In Bangladesh, there were 2,130 samples collected over the six months, averaging 355 samples per month, with the number of samples collected per month ranging from 72 in July to 471 in October. In Tanzania, there were 305 samples collected over two months, with 199 collected in December 2019 and 106 collected in January 2020.

2.3 Model description

This study developed an ANN-EFS to forecast point-of-consumption FRC using inputs collected at the point-of-distribution. The following sections describe the architecture of the base learners (i.e., the individual ANNs within the ANN-EFS) and the approach to generating the ANN-EFS from these base learners.

2.3.1 Base learners. Many model types are included in the ANN branch of machine learning. This study used the multilayer perceptron (MLP) type with one hidden layer for the base learner as this ANN-type has previously been used in an ANN-EFS to forecast FRC during the post-distribution period [18], and because it has consistently outperformed other models and ANN types for predicting chlorine residual [28, 34, 35]. The base learners were built using Python version 3.7.4 [56] and the Keras package [57]. Table 1 summarizes the hyperparameters of the base learners. Hyperparameter selection is discussed below for the input variables, hidden layer size, and data division.

The ability of ANNs to incorporate routine water quality variables other than just upstream residual chlorine is an advantage ANNs possess over process-based models of FRC decay [37, 38]. In humanitarian response, water quality data may be limited by constraints on data collection, limited water quality analysis capacities, or lack of reagents for field monitoring [58, 59]. This can be seen even in the current study where additional water quality data was collected, but equipment issues led to large numbers of incomplete measurements. Thus, to ensure the transferability of the modelling approach developed in this study, we only used the minimum number of input variables that can be expected in a humanitarian response setting: point-of-distribution FRC and elapsed time. S1 Appendix provides the data cleaning rules that were used to prepare the dataset. Histograms of the input and output variables are provided in S1 and S2 Figs. We also considered a second input variable set with two additional water quality variables: water temperature and electrical conductivity; however, the findings from this

Table 1. Summary of ANN base learner hyperparameters.

Model Architecture	MLP
Number of hidden layers	1
Input variables	Point-of-distribution FRC
	Elapsed time (from point-of-distribution to point-of-consumption)
Output variable	Point-of-consumption FRC
Hidden layer size	Bangladesh: 16 nodes
	Tanzania: 4 nodes
Hidden layer activation function	Hyperbolic tangent
Output layer activation function	Linear
Training function	*Nadam*
	Learning Rate: 0.1
Data division	Training set: 50%
	Validation set: 25%
	Test set: 25%

https://doi.org/10.1371/journal.pwat.0000040.t001

analysis were largely the same as those using only point-of-distribution FRC and elapsed time, so these findings are not discussed in the main body (for more, see S2 Table).

The hidden layer size of the MLPs was selected by successively doubling the number of nodes in the hidden layer and then selecting the hidden layer size where the performance began to plateau or when the training performance began to exceed the testing performance, indicating overfitting. The full results of this exploratory analysis are presented in S3 and S4 Figs.

The full dataset for each site was divided into calibration and testing subsets, with the calibration subset further subdivided into training and validation data. The testing subset was obtained by randomly sampling 25% of the overall dataset. The same testing subset was used for all base learners so that each base learner's testing predictions could be combined into an ensemble forecast. The training and validation data were obtained by randomly resampling from the calibration subset, with a different combination of training and validation data for each base learner to promote ensemble diversity, with 66.7% of the calibration data (50% of the overall dataset) used for training and 33.3% of the calibration (25% of the original dataset) used for validation. The network is trained by iteratively adjusting the weights and biases of the base learner to minimize the difference between the predictions and observations for the training set as measured by the cost function. The validation set is used during training to assess the cost function on data that is independent of the training set. Initially during training, the cost function for the training and validation should both decrease, but as training continues the validation cost will increase, indicating that the model is overfitting (i.e., overly specific to the training set). To prevent overfitting, we used a procedure called "early stopping" to end training. The early stopping procedure ends training if the validation cost increases for a fixed number of iterations called the patience. This study used a patience of 10 epochs.

A summary of the data, including the size and descriptive statistics for the calibration and testing datasets, is provided in Table 2. Importantly, Table 2 shows a large decrease in FRC from the point-of-distribution to the point-of-consumption for both the Bangladesh and Tanzania calibration and testing datasets, indicating that post-distribution FRC decay was substantial at both sites.

2.3.2 ANN-EFS formation. Each base learner was trained individually, and the ensemble forecast was formed by combining the predictions of each base learner into a probability density function (PDF). The ensemble size was selected via grid search by testing all ensemble sizes between 50 and 500 base learners in increments of 50. The results of this grid search are included in S5 and S6 Figs. An ensemble size of 200 base learners was selected as this was the

Table 2. Input and output variable descriptive statistics for calibration and testing datasets.

		Calibration				Testing			
		Number of Observations	Mean	Median	Standard Deviation	Number of Observations	Mean	Median	Standard Deviation
Bangladesh	Point-of-distribution FRC (mg/L)	1,597	0.71	0.66	0.38	533	0.70	0.64	0.38
	Elapsed Time (hours)		10.02	6.70	5.04		9.66	6.67	4.93
	Point-of-consumption FRC (mg/L)		0.34	0.28	0.28		0.34	0.28	0.28
Tanzania	Point-of-distribution FRC (mg/L)	228	0.39	0.30	0.22	77	0.39	0.30	0.21
	Elapsed Time (hours)		7.35	5.65	4.96		7.18	5.60	5.51
	Point-of-consumption FRC (mg/L)		0.20	0.10	0.15		0.18	0.10	0.15

https://doi.org/10.1371/journal.pwat.0000040.t002

smallest size that could ensure optimal performance while avoiding the additional computational time needed for larger ensembles.

When developing an EFS, the base learners must be sufficiently different from each other so that the resulting forecast accurately quantifies the uncertainty in the underlying behaviour [60, 61]. This study achieved this by varying the weights and biases between the base learners using two techniques. First, the initial weights and biases were randomized, so no two base learners started the training process with the same parameters. Second, as discussed in Section 2.3.1, each base learner was trained on a different subset of the calibration dataset by randomly sampling the training data and validation data. This provides variation in the base learner parameters by optimizing each base learner to a different subset of the calibration data.

2.4 Error metrics

During training, an ANN's weights and biases are calibrated to minimize the difference between the predictions and the observed data. The cost function determines how this difference is measured, meaning the cost function determines the behaviour that the ANN learns during training [41]. In this study, we generated cost functions by combining an error metric with a cost-sensitive learning technique. Since the main limitation of past applications of ANN-EFSs for forecasting point-of-distribution FRC was underdispersion leading to poor reliability [18], the error metrics evaluated in this study all measure the similarity of the spread or distribution of the predictions with the observed data. Details for each error metric are provided below.

Throughout this section O and P refer to the full set of observed and predicted point-of-consumption FRC concentrations, respectively; o_i and p_i refer to the i^{th} observed and predicted point-of-consumption FRC concentration, respectively; and N refers to the total number of observations. Note that in this section, a prediction refers to the output of one base learner in the ANN-EFS.

2.4.1 Mean squared error. MSE (Eq 1) is a symmetrical error metric that is commonly used as a cost function in machine learning [41]. It is negatively oriented, meaning that lower scores are preferable, with a lower limit of 0 and no upper bound. Past research has shown that an ANN-EFS trained using MSE produced underdispersed forecasts of point-of-consumption FRC which may be because MSE prioritizes performance near the mean of the distribution of the observations [18]. This study used MSE as a benchmark for comparison with the other error metrics considered.

$$MSE = \frac{\sum_{i=1}^{N} (p_i - o_i)^2}{N} \tag{1}$$

2.4.2 Nash Sutcliffe Efficiency. The Nash Sutcliffe Efficiency (NSE) measures the amount of observed variance explained by the model and can be obtained by normalizing the MSE about the variance of the observations (Eq 2) [62]. While NSE does not explicitly measure the similarity of the spread or distribution between a base learner's predictions and the observations, it does implicitly account for the spread of the observations in the cost by including the variance of the observations in the cost calculation. NSE is positively oriented, meaning that higher scores are preferable, with an upper limit of 1 and no lower limit. Since the *Nadam* optimizer can only find the minimum of a function, NSE was multiplied by -1 to convert it to a negatively oriented score with a lower limit of -1 and no upper bound.

$$NSE = 1 - \frac{\sum_{i=1}^{N} (p_i - o_i)^2}{\sum_{i=1}^{N} (o_i - \overline{O})^2} \tag{2}$$

2.4.3 Kling-Gupta Efficiency. Kling-Gupta Efficiency (KGE) arose out of a decomposition of NSE by Gupta et al [62] into three components (Eqs 3–5, respectively): correlation (r), the ratio of the variance of the predictions to the variance of the observations (α), and the ratio of the mean of the predictions to the mean of the observations (β).

$$r = \frac{\mathrm{cov}(O, P)}{\sqrt{\mathrm{cov}(O, O) * \mathrm{cov}(P, P)}} \tag{3}$$

$$\alpha = \frac{\sigma_P}{\sigma_O} \tag{4}$$

$$\beta = \frac{\overline{P}}{\overline{O}} \tag{5}$$

The Euclidean distance is then calculated between the r, α, and β scores obtained by the model and the scores for the ideal model, which would have a value of 1 for all three of the above as the ideal correlation coefficient is 1 and the ideal model would produce means and standard deviations equal to those of the observed data [62]. Eq 6 shows the calculation of the Euclidean distance in the square root term. KGE is then calculated by subtracting the Euclidean distance from 1 (Eq 6). This study included KGE because it explicitly penalizes differences between the first and second moments of the distributions of the predictions and the observations. This may lead to each base learner better reproducing the underlying distribution of the observations which could improve the reliability of the ANN-EFS as a whole. As with NSE, KGE is positively oriented, with higher scores representing shorter Euclidean distances from the ideal model. KGE has an upper limit of 1 and no lower limit. KGE was multiplied by -1 to convert it into a negatively oriented score for training the base learners.

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \tag{6}$$

2.4.4 Index of Agreement. The Index of Agreement (IoA) is a modified version of the NSE with a revised denominator (Eq 7). IoA measures the difference between the deviations about the mean of the predictions and the observations [63]. This study included IoA for this ability to prioritize similar spread about the mean since this could help prevent forecast underdispersion. Like NSE, IoA is positively oriented with an upper limit of 1 and no lower limit. IoA was converted into a negatively oriented score by multiplying the calculated score by -1.

$$IoA = 1 - \frac{\sum_{i=1}^{N} (p_i - o_i)^2}{\sum_{i=1}^{N} (|p_i - \overline{O}| + |o_i - \overline{O}|)^2} \tag{7}$$

2.5 Cost-sensitive learning weightings

There are several techniques to implement cost-sensitive learning in data-driven modelling. These include resampling techniques that address data imbalances through synthetic data or strategic over/under-sampling; by modifying a classification model's decision threshold; or by weighting samples in the model's cost function [45, 46, 64, 65]. We took this latter approach as it integrates well with the use of alternative error metrics. Thus, during training, the error metric determines how the difference between predictions and observations is measured, and the cost-sensitive learning approach weights the error metric to prioritize performance in a certain region of the output space.

This study evaluated three weightings, described below. In the following sections, O is the set of observed point-of-consumption FRC concentrations, o_i is the i^{th} observed point-of-consumption FRC concentration, and w_i is the weighting applied to the error metric for the i^{th} prediction-observation pairing. S2 Appendix shows the approach for calculating the cost functions when each error metric is weighted with a cost-sensitive learning approach.

2.5.1 Weighting 1: Inverse FRC weighting. The first cost-sensitive learning approach, inverse FRC weighting, uses a sample-based approach where the weight assigned to each observation is based on that sample's household FRC measurement [50, 65]. We multiplied each observation by the inverse of its point-of-consumption FRC concentration to prioritize high-risk observations (i.e., those with the lowest point-of-consumption FRC).

$$w_i = \frac{1}{o_i} \tag{8}$$

Eq 8 was modified for training the base learners of the ANN-EFS to account for the input and output data being normalized between -1 and 1. Using Eq 8 with these normalized inputs would produce two asymptotes at the median observed point-of-consumption FRC concentration. To avoid this, we added a fixed constant, 1.1 to the normalized observed value, as shown in Eq 9, where $_{i_{norm}}$ is the i^{th} normalized observation.

$$w_i = \frac{1}{o_{i_{norm}} + 1.1} \tag{9}$$

2.5.2 Weighting 2: Class-based weighting by FRC. The second weighting, class-based weighting by FRC, also prioritizes observations with low household FRC, however, in this case, observations were first grouped into classes based on their household FRC value and then a weight was assigned to each class, instead of to each observation. Class-based weighting is a common cost-sensitive learning approach for classification models when prioritizing specific classes [45, 54, 55, 62, 65]. The thresholds used to group the observations into classes were selected based on groupings used in literature and water quality guidelines for humanitarian response [66–68]:

- FRC between 0 mg/L and 0.2 mg/L–observations with FRC in this range are considered high risk since they have insufficient FRC to prevent recontamination [67].
- FRC between 0.2 mg/L and 0.5 mg/L–observations with FRC in this range are considered moderate risk. This is sufficient to prevent recontamination under normal circumstances, though it may be insufficient during a waterborne illness outbreak or when conditions favour recontamination [67, 69].
- FRC between 0.5 mg/L and 1.0 mg/L–observations with FRC in this range are considered low risk. This range is typically recommended to prevent recontamination during outbreaks of waterborne illness [67].
- FRC above 1.0 mg/L–observations with FRC in this range are considered very low risk as this is above even the range recommended during outbreaks of waterborne illness [67]. If recontamination occurs with FRC above 1.0 mg/L, there may be factors other than insufficient chlorine residual driving recontamination [69].

The weights assigned to each class were determined based on the risk of household recontamination to prioritize performance on observations with the greatest risk. The highest priority class, point-of-consumption FRC below 0.2 mg/L, was assigned a weight of 1.0. This weight

was then halved for each subsequent class (Eq 10).

$$w_i = \begin{cases} 1.0 \; if \; 0 \leq o_i < 0.2 \\ 0.5 \; if \; 0.2 \leq o_i < 0.5 \\ 0.25 \; if \; 0.5 \leq o_i < 1.0 \\ 0.125 \; if \; o_i \geq 1.0 \end{cases} \tag{10}$$

2.5.3 Weighting 3: Inverse frequency weighting. The third weighting used a special type of class-based weighting called inverse frequency weighting, where the weights are assigned to counteract data imbalances, ensuring each class is equally prioritized during training [27, 54, 55, 65, 70–72]. To achieve this, the weights for each class were calculated as the inverse of the relative frequency of observations in that class. Using the same classes as Weighting 2, the inverse frequency weight for the j^{th} class was calculated as:

$$w_j = \frac{number \; of \; observations \; in \; category \; j}{total \; number \; of \; observations}^{-1} \tag{11}$$

2.6 Ensemble verification metrics and model selection

Since the ANN-EFS predicts point-of-consumption FRC as a probability distribution and not a point prediction, performance metrics for point predictions, such as MSE or NSE, cannot be used to evaluate the EFS [21, 61]. Instead, this study evaluated the ANN-EFS using ensemble verification metrics which measure the *probabilistic* performance of the EFS. Probabilistic forecasts are typically evaluated on two criteria: reliability and sharpness [73]. Reliability refers to the similarity between the probability distributions of the forecast and the observations, and sharpness refers to the narrowness of the forecast spread around a given observation. The first priority when evaluating ensemble forecasts is reliability, but a sharper forecast is preferable over a less sharp forecast if the reliabilities are the same [61, 73]. EFSs are evaluated for their ability to generalize on new data, so we only used these metrics to evaluate performance on the test dataset.

The following sections describe the ensemble verification metrics used in this study. Throughout the following section, O refers to the full set of observed point-of-consumption FRC concentrations, and o_i refers to the i^{th} observation, where there are I total observations in the test dataset. F refers to the full set of forecasted point-of-consumption FRC concentrations, where f_i^m is the prediction by the mt^h base learner in the ANN-EFS for the i^{th} observation and F_i refers to the ensemble forecast for the i^{th} observation. Thus, for each observation, there is a corresponding probabilistic forecast. Together these are referred to as a forecast-observation pair. For the following metrics, it is assumed that the predictions of each base learner in the ensemble are sorted from low to high for each observation such that $f_i^m \leq f_i^{m+1}$ from $m = 1$ to $m = M$.

2.6.1 Percent capture. Percent capture is the percentage of observations where the observed point-of-consumption FRC concentration was within the limits of the ensemble's forecast. While percent capture does not directly evaluate reliability or sharpness, it does indicate the degree to which the forecasts are underdispersed. The percent capture is a positively oriented score, with an upper limit of 100% and a lower limit of 0%. To calculate percent capture, observation o_i is considered captured if $f_i^1 \leq o_i \leq f_i^M$. When evaluating the ensemble forecasts, we used the percent capture of the overall dataset (PC) as an indicator of underdispersion and the percent capture of observations with point-of-consumption FRC below 0.2 mg/L ($PC_{<0.2}$) to indicate how well the ANN-EFS can capture high-risk observations.

2.6.2 Rank Histograms. The Rank Histogram (RH) is a visual tool that assesses the reliability of ensemble forecasts. The RH is constructed by adding each observation, o_i to the sorted ensemble forecast F_i and determining the observation's rank within the ensemble (i.e., the corresponding m if it were a base learner prediction). The RH is thus simply the histogram of the rank assigned to each o_i. If the forecast and observed probabilities are the same, then any observation is equally likely to occur in any of the M+1 ranks, which would result in a flat RH [61, 74]. The more dissimilar the forecasted and observed probability distributions are, the farther from flat the RH will be. Candille & Talagrande [75] proposed a numerical score, the δ score (Eq 12), to measure deviations from flatness in an RH. The ideal score is 1, with scores much greater than 1 indicating substantial deviations from flatness and scores less than 1 indicating interdependence between ensemble predictions. A δ score other than 1 only indicates deviations from flatness, not the reason for the deviation (i.e., dispersion, skew, etc.) which must be determined from visual inspection [75].

$$\delta = \frac{\Delta}{\Delta_o} \tag{12}$$

The two components of the δ score are shown in Eqs 13 and 14 where M is the total number of base learners, I is the total number of observations, and s_k is the number of elements in the k^{th} bin of the RH [75].

$$\Delta = \sum_{k=1}^{M+1} \left(s_k - \frac{I}{M+1} \right)^2 \tag{13}$$

$$\Delta_o = \frac{I * M}{M+1} \tag{14}$$

The δ score was calculated for both the overall dataset (referred to throughout as δ) and for only those observations where the observed point-of-consumption FRC was below 0.2 mg/L ($\delta_{<0.2}$).

2.6.3 Confidence interval reliability diagram. Reliability diagrams are plots of the observed relative frequency of events against the forecast probability of that event occurring [76]. This diagram has been adapted for ANN-EFS modelling as the confidence interval (CI) reliability diagram which compares the frequency of observed values with the corresponding CI of the ensemble, where the ensemble CIs are derived from the sorted forecasts of the base learners (for example, the forecast 90% CI would include the range between $f^{0.05M}$ and $f^{0.95M}$) [21]. We extended this further by plotting the percent capture within each CI against the CI level.

The CI reliability diagram is a visual indicator of forecast reliability. The ideal model would have percent capture in all CIs plotted along the 1:1 line; showing that the forecasted probabilities at each level are equal to the observed probabilities. We previously developed a numerical score for the CI reliability diagram, the CI reliability score, which calculates the sum of the squared vertical distance between the percent capture within each CI and the 1:1 line [18]. Since a smaller absolute distance means that each point is closer to the 1:1 line, this score is negatively oriented with a minimum value of 0. We calculated this score using CI thresholds, k, from 10% to 100% in 10% increments (Eq 15) for both the overall data set (CI_{score}) and for only those observations where the observed point-of-consumption FRC was below 0.2 mg/L ($CI_{score_{<0.2}}$).

$$CI\ Reliability\ Score = \sum_{k=0.1}^{1} (j - Percent\ Capture\ in\ CI_j)^2 \tag{15}$$

2.6.4 Continuously ranked probability score. The continuously ranked probability score (CRPS) is the mean integrated square difference between the forecast cumulative distribution function (CDF) and the observed CDF for each forecast-observation pairing. CRPS simultaneously measures the reliability, sharpness, and uncertainty of a forecast [76, 77]. The calculation of the CRPS is given in Eq 16 where F_i is the CDF of the forecast values for observation o_i and the x axis referenced is the point-of-consumption FRC concentration. Since each observation is a discrete value, its CDF is represented with the Heaviside function $H(x \geq x_a)$ which is a stepwise function: 0 for all concentrations of point-of-consumption FRC below the observed concentration and 1 for all concentrations above the observed. Eq 16 shows the calculation of CRPS for a single forecast-observation pairing. To evaluate the ANN-EFS, the average CRPS, $\overline{CRPS}$, is calculated by taking the mean CRPS over all forecast-observation pairs.

$$CRPS = \int_{-\infty}^{\infty} (F_i(x) - H\{x \geq o_i\})^2 dx \tag{16}$$

Hersbach [77] derived a calculation of CRPS for EFS that treats the forecast CDF as a stepwise continuous function with $N = M+1$ bins where each bin is bounded at two ensemble forecasts. $\overline{CRPS}$ is calculated from $\overline{g_n}$, the average width of bin n (average difference in FRC concentration between forecast values m and $m+1$) and $\overline{o_n}$, the likelihood of the observed value being in bin n. The $\overline{CRPS}$ can be calculated as:

$$\overline{CRPS} = \sum_{n=1}^{N} \overline{g_n} [(1 - \overline{o_n}) p_n^2 + \overline{o_n} (1 - p_n)^2] \tag{17}$$

Where p_n is the probability associated with each bin, $p_n = \frac{n}{N}$ [77].

2.6.5 Skill scores. Skill scores evaluate improvement over a baseline model by normalizing the score obtained for an ensemble verification metric using a baseline score and an ideal score. Any score can be converted to a skill score using Eq 18. The skill score values range from $-\infty$ to 1, with 1 indicating that the score obtained by the model being evaluated is the ideal score and a positive score indicating improvement over the baseline. A skill score of 0 means that there is no difference between the score for the model being evaluated and the baseline, and a negative score indicates that the score obtained is worse than the baseline. In this study, the scores obtained using the ANN-EFS trained with unweighted MSE were used as the baseline, and all of the models tested were the same (ANN-EFS with the same size and base learner architecture), with the exception of the cost function. Therefore, the skill score indicates the effect of using each cost function for training the ANN-EFS for forecasting point-of-consumption FRC relative to the baseline performance obtained when the ANN-EFS is trained with unweighted MSE.

$$Skill\ Score = \frac{score\ obtained - baseline}{ideal\ score - baseline} \tag{18}$$

2.7 Code and data availability

The code to train, test, and evaluate the ANN-EFS is available on the SWOT Project's GitHub at the following link: https://github.com/safeh2o/swot_research_cost_sensitive. This repository also includes the cleaned Tanzania and Bangladesh datasets.

3 Results and discussion

In the following sections, we first present the performance of the ANN-EFS trained with the baseline cost function: unweighted MSE. Second, we evaluate the performance of the ANN-EFS when trained using the alternative error metrics and cost-sensitive learning techniques presented in Sections 2.4 and 2.5. Third, we select the best cost function for training an

ANN-EFS to forecast point-of-consumption FRC using the performance metrics outlined in Section 2.6. Fourth, we compare the ANN-EFS performance when trained with the selected cost function against the baseline performance. Finally, we discuss the implications of these findings for practitioners in humanitarian response.

3.1 Baseline ANN-EFS performance

Figs 1 and 2 show the baseline performance of the ANN-EFS for the Bangladesh and Tanzania datasets, respectively. These figures show the ensemble forecasts and observations (Figs 1A and 2A), RHs (Figs 1B and 2B), and CI reliability diagrams (Figs 1C and 2C) for each site. For both datasets, the forecasts produced by the ANN-EFS trained with the baseline cost function are highly underdispersed: the forecast spread is much smaller than the spread of the observations; the CI reliability diagram has all points well below the 1:1 line; and the RH has a pronounced U-shape. This underdispersion is also reflected in the low percent capture (below 50% for both the overall dataset and observations with FRC below 0.2 mg/L).

The underdispersion also led to poor reliability. This is best reflected in the RH and associated δ scores. An ideal RH would be flat, reflecting a uniform distribution [61, 74, 75]. The U-shaped RHs shown in Figs 1B and 2B then indicate not only underdispersion but also poor reliability, which is also reflected in the δ scores between 19 to 153, which are substantially larger than 1 indicating poor reliability. This poor reliability is also shown in the CI reliability diagram where, for all CI levels, the percent capture at each CI level was much lower than the ideal, shown on the 1:1 line. Together these demonstrate that the baseline ANN-EFS, trained using a conventional cost function (unweighted MSE) produced highly underdispersed forecasts with poor reliability, despite unweighted MSE being a common cost function for ANNs

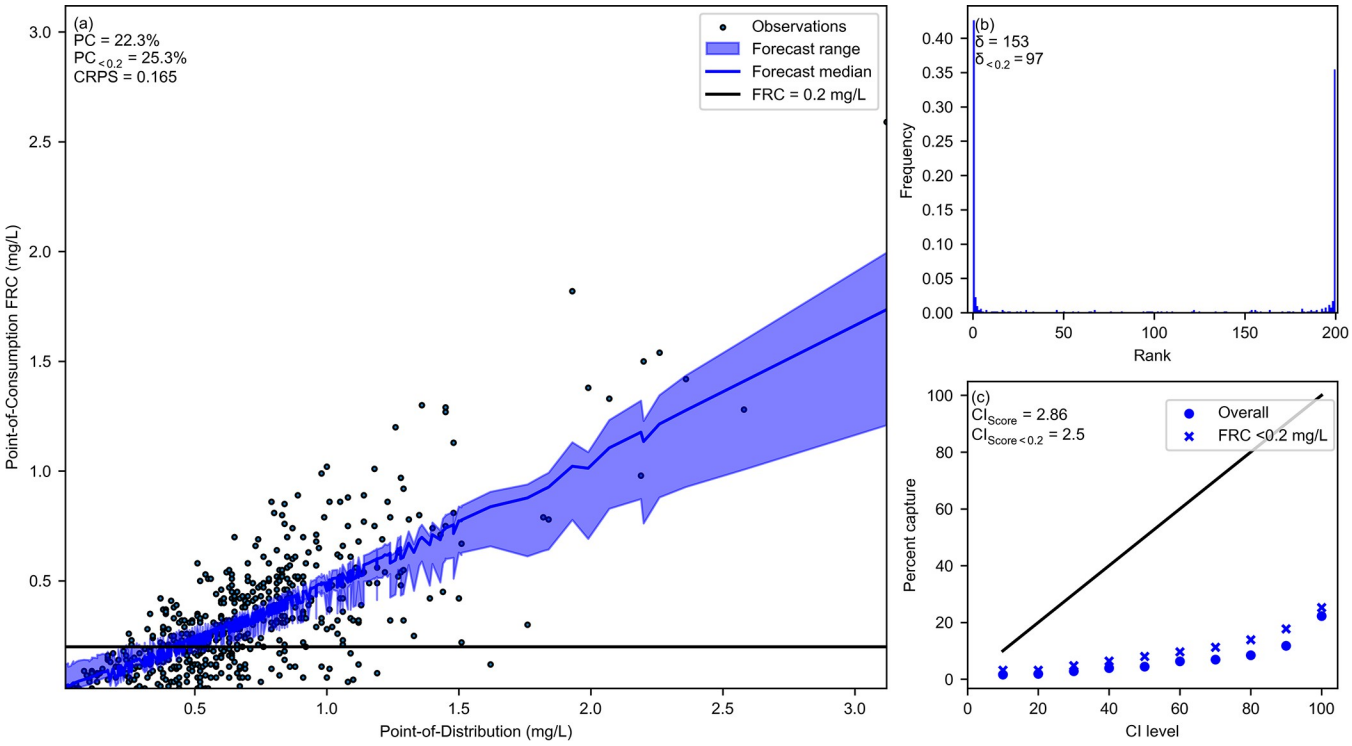

Fig 1. Baseline ANN-EFS performance for Bangladesh dataset. Forecast shown with true observations in (a), rank histogram shown in (b), and CI reliability diagram shown in (c).

https://doi.org/10.1371/journal.pwat.0000040.g001

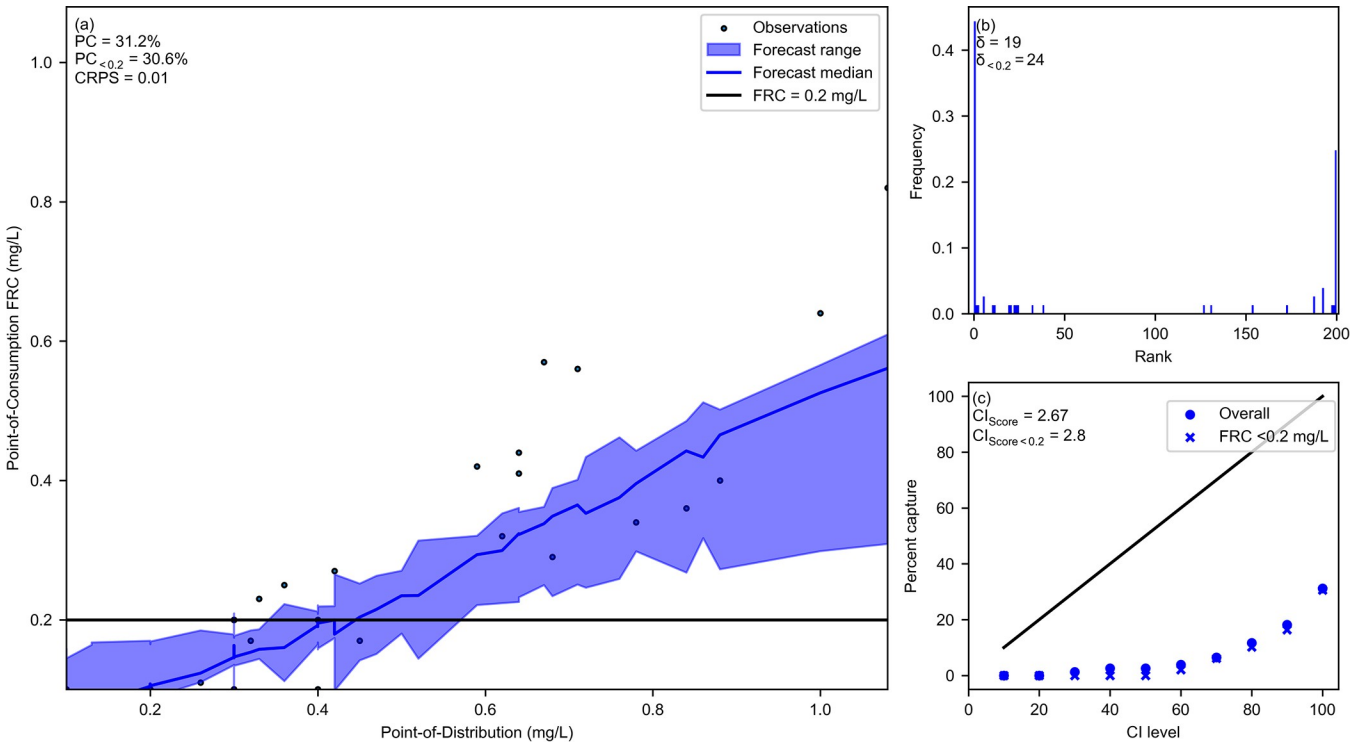

Fig 2. Baseline ANN-EFS performance for Tanzania dataset. Forecast shown with true observations in (a), rank histogram shown in (b), and CI reliability diagram shown in (c).

https://doi.org/10.1371/journal.pwat.0000040.g002

[41]. This supports past findings that show that symmetrical error metrics like MSE are not effective cost functions unless the desired model behaviour is convergence to the mean of the distribution of the observations [41, 42]. While both Crone et al. [41] and Toth [42] considered this in the context of regression-based modelling where the cost-sensitive behaviour was asymmetric (i.e., the cost of overprediction errors was different from underprediction in errors). Our findings show that the limitations of MSE are also important for probabilistic modelling using EFS.

3.2 Comparison of alternative error metrics and cost-sensitive learning

Fig 3 compares the skill scores obtained for the ensemble verification metrics listed in Section 2.6 when the ANN-EFS was trained with each cost function considered in this study. The raw scores are provided in S1 Table. As discussed in Section 2.3, we evaluated a second input variable combination, which included electrical conductivity and water temperature in addition to point-of-distribution FRC and elapsed time, the results for which are provided in S2 Table. Fig 3 shows that for each ensemble verification metric there was some combination of an alternative error metric and cost-sensitive learning technique that yielded a positive skill score, indicating that alternative error metrics and cost-sensitive learning could always be combined to improve performance over the baseline.

Fig 3 shows that the forecast dispersion, measured using percent capture, improved substantially when the ANN-EFS was trained with the cost functions that combine alternative error metrics and cost-sensitive learning, compared to baseline training with MSE. The highest skill scores for *PC* ranged from 0.698 in Tanzania to 0.725 in Bangladesh, and the highest skill scores for $PC_{<0.2}$ ranged from 0.706 in Tanzania to 0.791 in Bangladesh. This indicates that

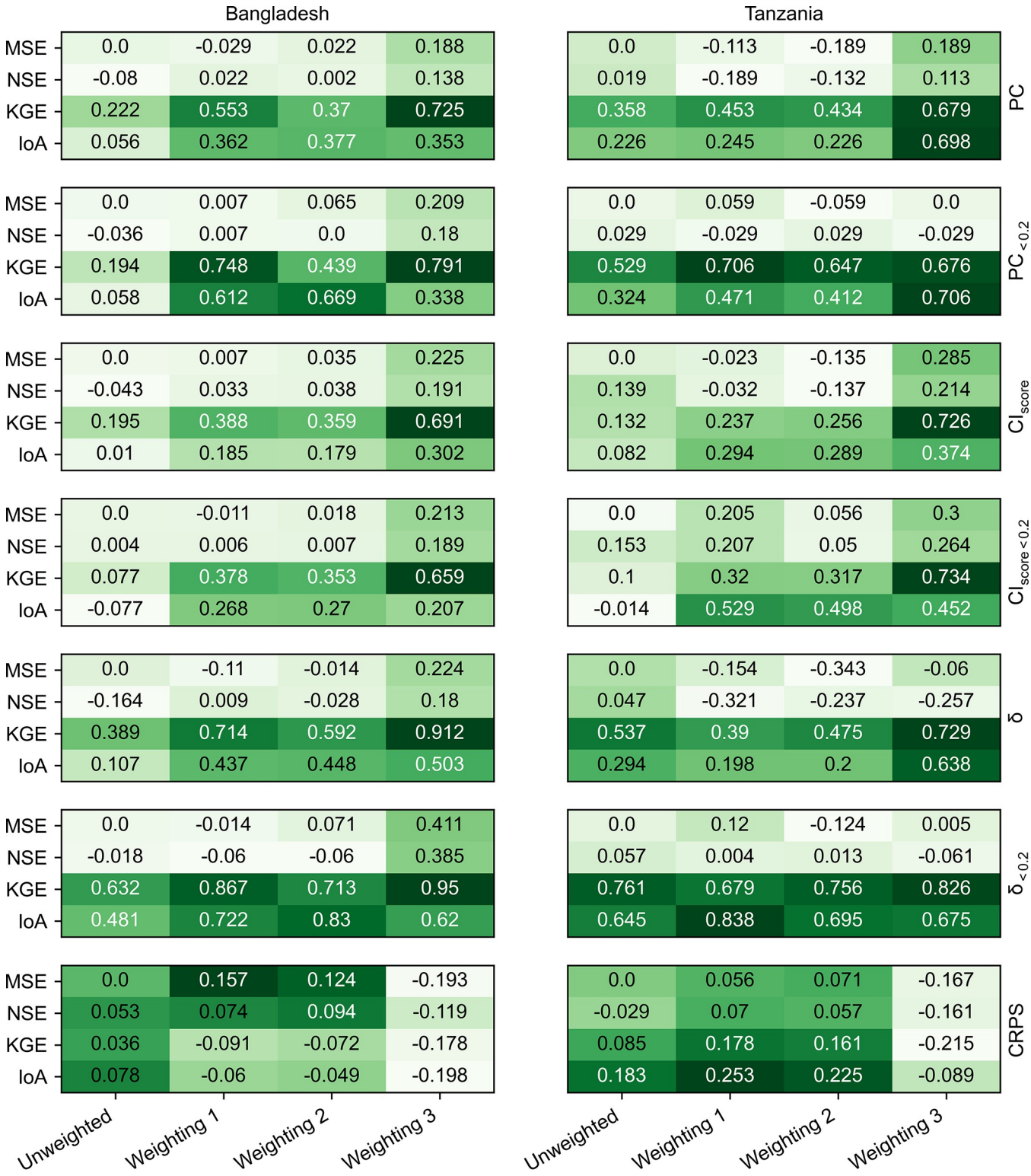

Fig 3. Skill scores for each cost function considered. Left column: Bangladesh, right column: Tanzania. Skill scores shown in rows, from the top: *PC*, $PC_{<0.2}$, CI_{score}, $CI_{score_{<0.2}}$, δ, $\delta_{<0.2}$ and CRPS.

https://doi.org/10.1371/journal.pwat.0000040.g003

training the ANN-EFS with alternative error metrics and cost-sensitive learning led to a 70% improvement in forecast dispersion relative to the baseline performance. The largest improvement in percent capture was produced when the error metric used in the cost function was either KGE or IoA. This is likely because the scores for these error metrics improve as the spread of the base learner's predictions becomes more similar to the spread of the observations. KGE's β term measures this as the similarity between the variance of the predictions and the observations [61]. IoA measures this as the similarity of the deviations about the mean for the predictions and observations [62]. Training the ANN-EFS with NSE did not consistently improve the percent capture as NSE uses the ratio of absolute differences normalized about the variance of the observations but does not include the predicted variance. Without including the predicted variance, NSE cannot explicitly ensure that the spread of the predictions matches the spread of the observations and thus, training the ANN-EFS with NSE does not improve forecast dispersion.

In addition to the alternative error metrics, cost-sensitive learning also improved forecast dispersion. Fig 3 shows that when the ANN-EFS was trained using KGE or IoA combined with any of the three cost-sensitive learning approaches, the model achieved higher skill scores for percent capture than when the model was trained using the cost-insensitive (unweighted) form of the error metric. The best overall percent capture (*PC*) at both sites was obtained when the cost function used to train the ANN-EFS included Weighting 3 (inverse frequency weighting). Inverse frequency weighting even led to improvements in *PC* when combined with MSE or NSE. This is likely because inverse frequency weighting rebalances the error metric to equally prioritize the full output space, leading to better predictions in regions of the output space that have fewer observations [45]. When considering only observations with point-of-consumption FRC below 0.2 mg/L ($PC_{<0.2}$), Weightings 1 and 2 typically produced better performance, likely because these approaches prioritize performance on observations with lower point-of-consumption FRC. Despite this, in Bangladesh, the ANN-EFS trained using KGE with inverse frequency weighting produced the best capture even of these high-risk observations.

Both the CI reliability score and the RH δ score followed similar patterns to the percent capture shown in Fig 3, with alternative error metrics and cost-sensitive learning producing substantial improvements in these scores. The highest CI_{score} at each site was 0.691 in Bangladesh and 0.726 in Tanzania and the highest $CI_{score_{<0.2}}$ was 0.659 in Bangladesh and 0.734 in Tanzania. The improvements were even higher for the δ score with skill scores ranging from 0.729 to 0.912 for the overall dataset and between 0.838 and 0.95 for observations with household FRC below 0.2 mg/L. These results demonstrate that the use of alternative error metrics and cost-sensitive learning can improve forecast reliability when modelling point-of-consumption FRC with an ANN-EFS. The highest skill scores, reflecting the largest improvement, were obtained when the ANN-EFS was trained with KGE. This is likely because KGE measures the actual similarity of the first two moments of the distribution (mean and standard deviation) between the predictions and observations.

The CRPS also improved when the ANN-EFS was trained with alternative error metrics and cost-sensitive learning, though not when trained using KGE or IoA with Weighting 3. This is likely because CRPS tends to be dominated by the sharpness term [76, 77] so the improvements in dispersion achieved when the ANN-EFS was trained using these cost functions may have also led to the CRPS becoming worse as the forecast spread become larger. As discussed in Section 2.6, the first priority when evaluating an ensemble forecast must be reliability, and sharpness should only be considered once adequate reliability has been obtained [73].

These findings highlight that training an ANN-EFS using alternative error metrics and cost-sensitive learning substantially improves the dispersion and reliability of ensemble forecasts of point-of-consumption FRC. The improvement over the baseline performance was obtained by changing only a single hyperparameter of the base learners of the ANN-EFS: the cost function. This is consistent with findings from several other fields including inventory management [41], flood modelling [42, 49], fraud detection [50], epidemiology [52], and drinking water quality modelling [54, 55], all of which have shown that changing the error metric and implementing cost-sensitive learning is much more effective than using standard symmetrical error metrics and cost insensitive learning. However, the present study shows this for the first time when using probabilistic EFS. This is an important distinction because the performance of a regression or classification model can be evaluated using its cost function, and thus the desired behaviour can be more easily specified for the model. For example, Olowookere et al. [50] developed a cost-sensitive learning framework for detecting fraudulent credit card purchases where the cost of misprediction was derived from the amount of money spent in the fraudulent transaction, and Crone et al. [41] defined a novel, asymmetric error metric, based on the actual cost of over and understocking shelves in a warehouse. Thus, the desired behaviour can be directly integrated into the ANN training. By contrast, the ensemble verification metrics used to evaluate the ANN-EFS in this study cannot be used to train the base learners since ensemble verification metrics require an ensemble forecast. For example, using KGE as the error metric only evaluates the similarity between the distributions at the base learner level and is not directly a measure of the ANN-EFS's overall probabilistic performance. Thus, it is an important finding that training the ensemble base learners with this cost function translated into improved reliability when the base learner predictions were combined into an ensemble forecast. In consideration of the first aim of this study, which was to investigate the effect of alternative error metrics and cost-sensitive learning on the probabilistic forecasting performance of ANN-EFS, we see that by selecting alternative error metrics and cost-sensitive learning approaches that reflect the intended behaviour, the ANN-EFS performance vastly outperforms a standard cost function (unweighted MSE) when forecasting point-of-consumption FRC. It is also worth noting that training the base learners using alternative cost functions and cost-sensitive learning yielded greater improvements in reliability and dispersion over the baseline ANN-EFS. than were obtained in an earlier ANN-EFS study which used post-statistical processing [18]. Statistical post-processing is a common approach to improving the reliability and dispersion of process-based EFS [78], but for ANN-EFS changing the cost function appears to be more effective. This is also consistent with findings from regression modelling that determined that altering the cost function is more effective than post-processing for obtaining a desired model behaviour [79].

3.3 Selection of preferred cost function

This study used a ranking approach to select the preferred cost function. The skill scores presented in Fig 3 were used to determine how often the ANN-EFS trained with a given cost function (i.e., the combination of an alternative error metric and cost-sensitive learning approach) produced either the best score for an ensemble verification metric ("best") or one of the five best scores ("top five"). Fig 4 shows the results of this ranking approach, identifying the frequency with which each cost function was either the "best" or one of the "top five" for each ensemble verification metric at each site.

Fig 4 shows that in all cases, the "best" cost functions incorporated cost-sensitive learning, and 15 of the 17 "best" cost functions (89%) used an error metric other than MSE. Similarly, 71 of the 80 "top five" cost functions (89%) incorporated cost-sensitive learning, and 76 of the

80 "top five" cost functions (95%) used an error metric other than MSE. Furthermore, the ANN-EFS trained with the baseline cost function, unweighted MSE, never produced the "best" or one of the "top five" scores for any ensemble verification metric. This supports the finding that unweighted MSE is not appropriate for training an ANN-EFS to probabilistically forecast point-of-consumption FRC and that combining alternative error metrics with cost-sensitive learning to train the ANN-EFS leads to better probabilistic performance. This also reinforces that training an ANN-EFS with alternative error metrics and cost-sensitive learning improves the probabilistic performance of the ensembles, as demonstrated through improved dispersion and reliability of the ensemble forecasts.

Of the cost functions considered in this study, Fig 4 shows that combining KGE with Weighting 3 (inverse frequency weighting) consistently outperformed the other cost functions. This combination was the "best" cost function in 9 of a possible 16 cases and was one of the "top five" in 12 of a possible 16 cases. The high performance of this cost function is likely due to the explicit way in which KGE measures reliability, and the ability of inverse frequency weighting to promote performance throughout the output space. KGE promotes improved reliability by explicitly evaluating the difference between the observed and predicted mean and variance (the first two moments of a probability distribution) and the correlation for each base learner in the ANN-EFS [61]. This combines well with inverse frequency weighting which ensures an equal prioritization throughout the output space by more heavily weighting the most sparsely populated output classes. Thus, when combined, KGE with inverse frequency weighting ensures similarity between the distribution of each base learner's predictions and the observations across all regions of the output space, equally. Interestingly, inverse frequency weighting was developed to overcome data imbalances in classification machine learning problems, but the base learners in this study performed regression, not classification. One reason that this weighting was effective may be that classification problems are inherently probabilistic; classification models typically select the class with the highest probability of being true [45]. Thus, while the base learners of the ANN-EFS in this study were regression-based, the overall ensembles were probabilistic, and hence, a probabilistic classification-based cost-sensitive learning approach was most effective for training the base learners. This highlights a potential avenue for future research into the integration of classification techniques in the training of probabilistic EFSs, even if the base learners in these models are regression-based.

3.4 Performance comparison: Baseline vs selected cost function

This section compares the performance of the ANN-EFS used to forecast point-of-consumption FRC when trained with the selected cost function (KGE with inverse frequency weighting) to the baseline performance. Fig 5 compares the forecasted and observed point-of-consumption FRC for the ANN-EFS trained with both the selected cost function and the baseline cost function. This figure shows that when the ANN-EFS is trained with the selected cost function, the forecasts (shown in red) better match the spread of the observations than the baseline (shown in blue). This leads to better capture of the observed household FRC concentrations with PC increasing from 22.3% to 78.6% in Bangladesh and from 31.2% to 77.9%. and $PC_{<0.2}$ increasing from 25.3% to 84.4% in Bangladesh and from 30.6% to 77.6% in Tanzania when using the selected cost function. Thus, when trained using the selected cost function, the ANN-EFS captures over 70% of overall and high-risk observations, whereas when trained using the baseline, the model failed to capture even half of the observations.

These improvements in forecast dispersion also improve ensemble reliability. This is reflected in the RHs for each site shown in Fig 6 for Bangladesh and in Fig 7 for Tanzania. While the RHs produced from the forecasts of the ANN-EFS trained using the selected cost

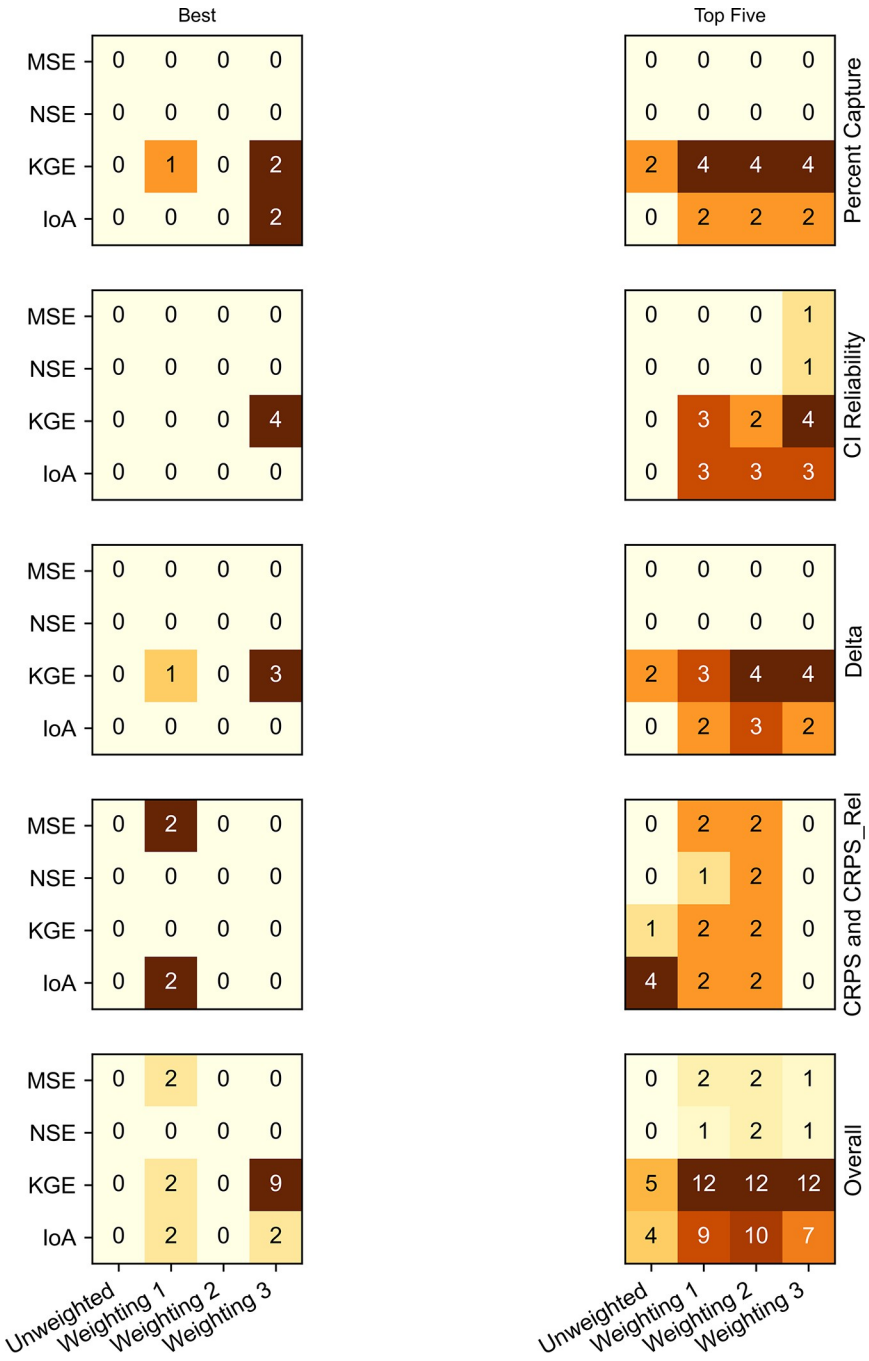

Fig 4. Combined frequency for each cost function as either best (left) or "top five" (right) for each ensemble verification metric and all metrics combined (bottom row).

https://doi.org/10.1371/journal.pwat.0000040.g004

function are still U-shaped, indicating underdispersion, the height of the outlier bars (bins 0 and 200 which indicate under- and over- outliers, respectively) are much lower, in some cases by a factor of 5. Thus, the RHs produced when the ANN-EFS is trained with the selected cost function are much closer to the ideal than those produced when the ANN-EFS was trained with the baseline cost function. This demonstrates that the ANN-EFS forecasts are more reliable when the model is trained using the selected cost function, which in turn means that

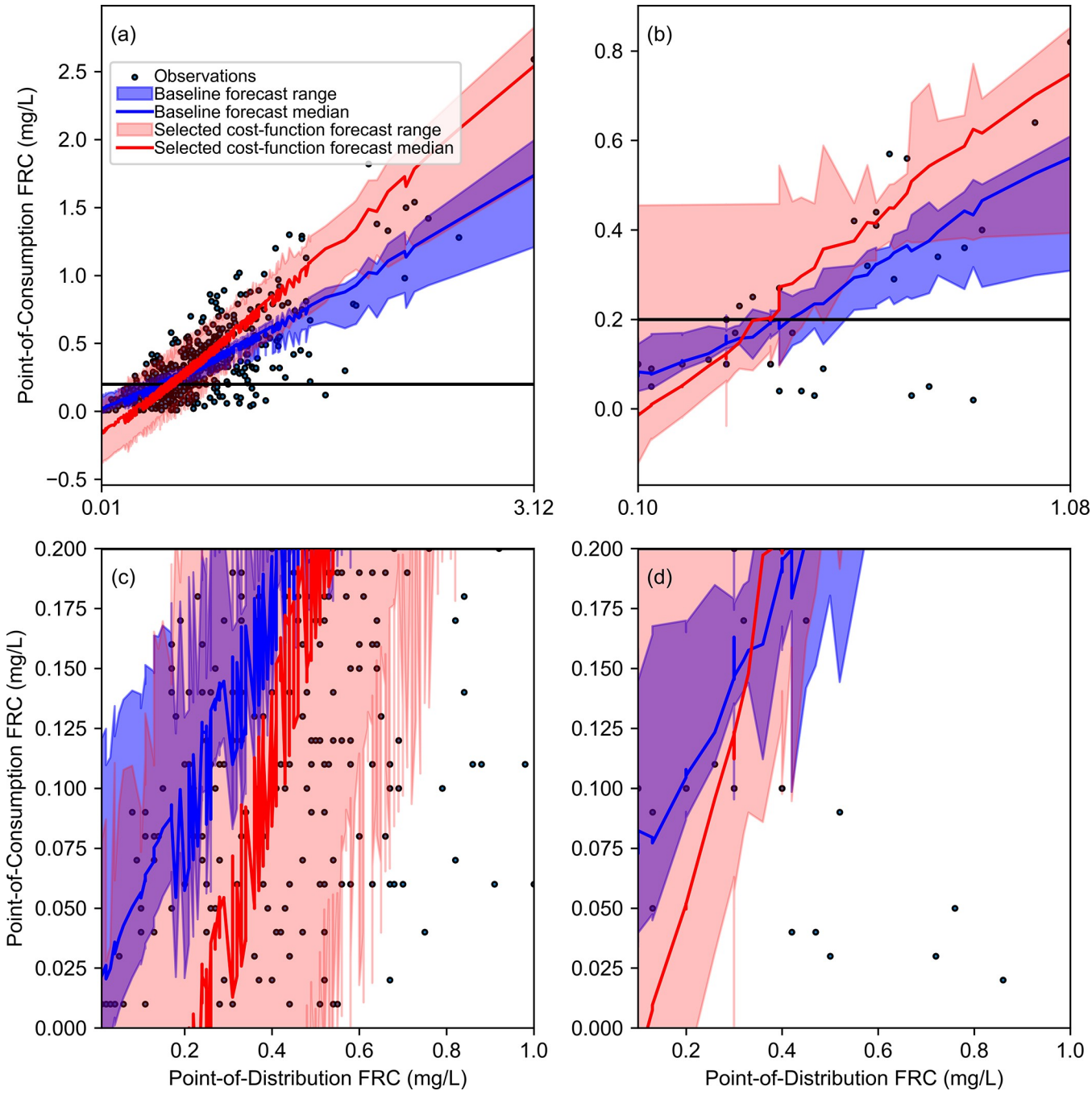

Fig 5. Forecast-observation comparison using baseline and selected cost function. Forecast observation pairs shown for: (a) Bangladesh–all observations, (b) Tanzania–all observations, (c) Bangladesh–observations with FRC below 0.2 mg/L, and (d) Tanzania–observations with FRC below 0.2 mg/L.

https://doi.org/10.1371/journal.pwat.0000040.g005

predicted probabilities obtained from the ANN-EFS (e.g., the probability of FRC being below 0.2 mg/L for a given point-of-distribution target) are much closer to the true probabilities when trained with the selected cost function as opposed to the baseline.

This improved reliability is also reflected in the CI reliability diagrams. Fig 8 shows the CI reliability diagrams for both the Bangladesh and Tanzania datasets for the overall dataset and for observations with point-of-consumption FRC below 0.2 mg/L. Fig 8 shows that at both

sites, the ANN-EFS trained with the selected cost function had capture values much closer to the ideal value at each CI level than when trained using the baseline cost function. This indicates that the improvement in reliability observed is not just due to improved dispersion at the upper and lower limits of the forecast, but that the predicted and true probabilities are much closer at every CI level of the forecast. Thus, training the ANN-EFS with the selected cost function led to an overall improvement in reliability

The CI reliability diagram can also demonstrate the impact of improved reliability on the quality of risk-based FRC targets produced by the ANN-EFS. For example, consider a humanitarian responder who seeks a point-of-distribution FRC target to ensure only a 10% risk of users would have insufficient FRC at the point-of-consumption. They would obtain this target from the lower bound of the 80^{th} percentile CI (this CI is bounded by the 10^{th} and 90^{th} percentiles of the forecast). To ensure the validity of the target for this risk level, the 80^{th} percentile CI should capture 80% of the observations. When the ANN-EFS was trained with unweighted MSE, the 80^{th} percentile CI only captured 8% of the observations in Bangladesh and 12% in Tanzania. Thus, an FRC target generated from this CI is unlikely to produce the desired level of safety since the forecast probability distribution is very different from the true distribution of the data. By contrast, when the ANN-EFS is trained with the selected cost function, the 80^{th} percentile of the forecast captured 41% of observations in Bangladesh and 52% in Tanzania. While these are still underdispersed, they are much closer to the ideal capture, meaning that the forecast distribution is closer to the true distribution of the observations and thus the targets generated by this model are more likely to produce the desired level of safety.

The findings presented above show that training an ANN-EFS using the selected cost function, KGE with inverse frequency weighting, produces better dispersed and more reliable probabilistic forecasts of point-of-consumption FRC than when the ANN-EFS is trained with unweighted MSE. These improvements can lead to improved risk-based FRC targets since the ANN-EFS can better reproduce the underlying distribution of the observed data when predicting the probability of high-risk events occurring, thus giving operators the tools to mitigate these high-risk events. Based on these findings, we recommend that the selected cost function, KGE with inverse frequency weighting, replace unweighted MSE as the cost function used in the SWOT. The findings of this section also demonstrate the importance of selecting an appropriate cost function when training ANNs. In drinking water research, two recent studies have proposed ANN model building frameworks [80, 81], however, neither of these studies include guidance on the selection of an appropriate cost function. Based on the substantial improvements in performance obtained in this study when training the ANN-EFS with the selected cost function as opposed to a default, as well as the improvements over cost-insensitive training obtained in other drinking water studies [54, 55], we recommend that future model development frameworks for drinking water modelling should also include consideration for selection of an appropriate cost function.

3.5 Implications for practitioners

As discussed in Section 1, the SWOT uses an ANN-EFS to predict the risk of having insufficient FRC in drinking water at the point-of-consumption in refugee and IDP settlements. In these settings, risk-based FRC targets help water system operators understand how household water safety risks may change when adjusting chlorination levels, allowing operators to balance this risk against other concerns such as disinfection by-product formation or chlorine taste and odour acceptance. The selected cost function (KGE with inverse frequency weighting) produces substantial improvements in forecast reliability compared to the current approach for training ANN-EFS used by the SWOT (unweighted MSE). Implementing the selected cost

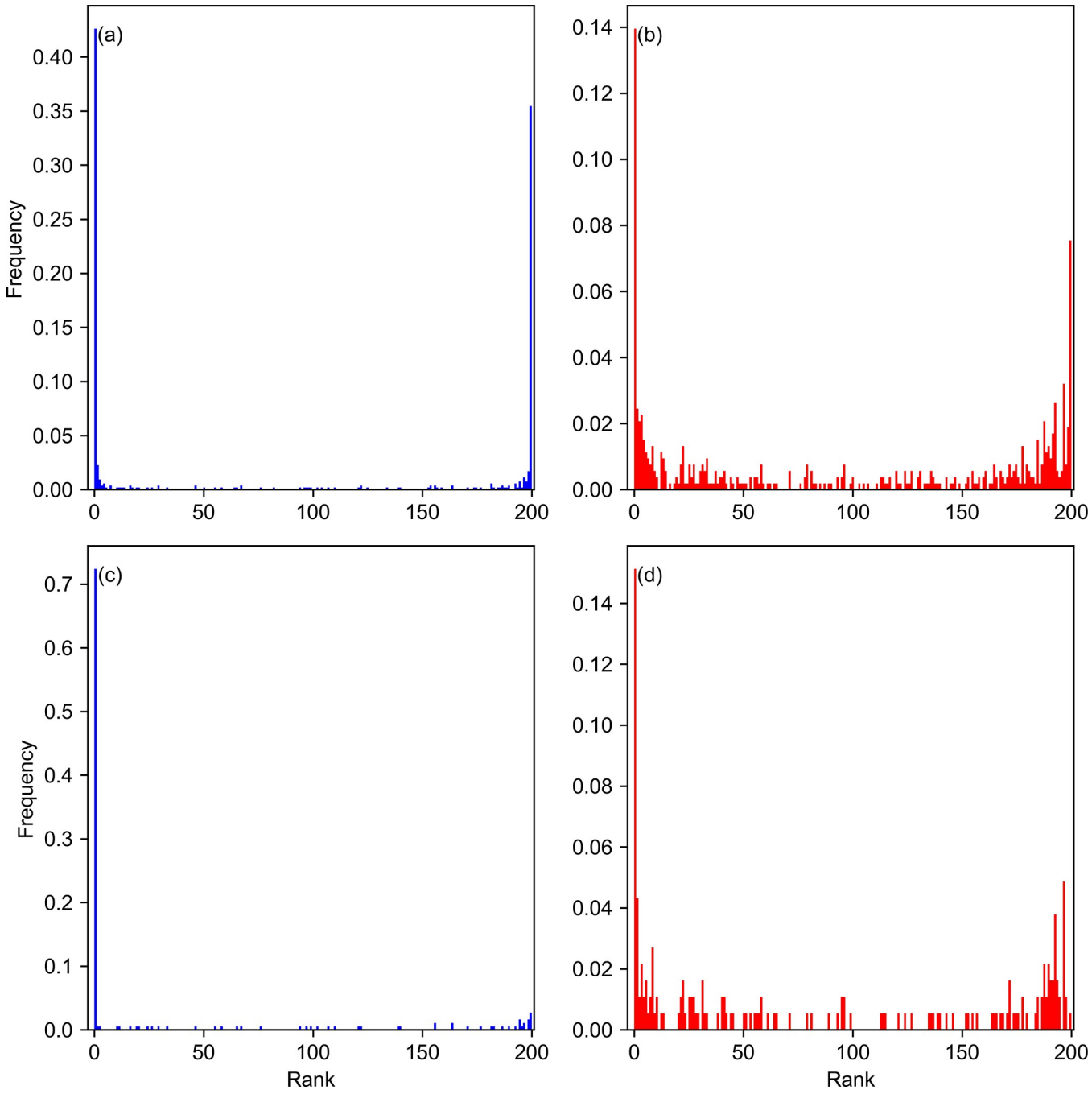

Fig 6. RHs for Bangladesh: (a) baseline model–all observations (b) selected cost function–all observations (c) baseline model,–observations with FRC below 0.2 mg/L and (d) selected cost function,–observations with FRC below 0.2 mg/L.

https://doi.org/10.1371/journal.pwat.0000040.g006

function to train the SWOT ANN-EFS would result in improved forecasts of point-of-consumption FRC that more accurately reflect the true distribution of observed household FRC data. This would result in predictions of the risk of having insufficient FRC at the point-of-consumption that are closer to the true risk, providing operators greater confidence in the predictive forecasting offered by the SWOT and enabling them to better manage the risks of both under- and over-chlorination in emergency water supply. Additionally, while the SWOT

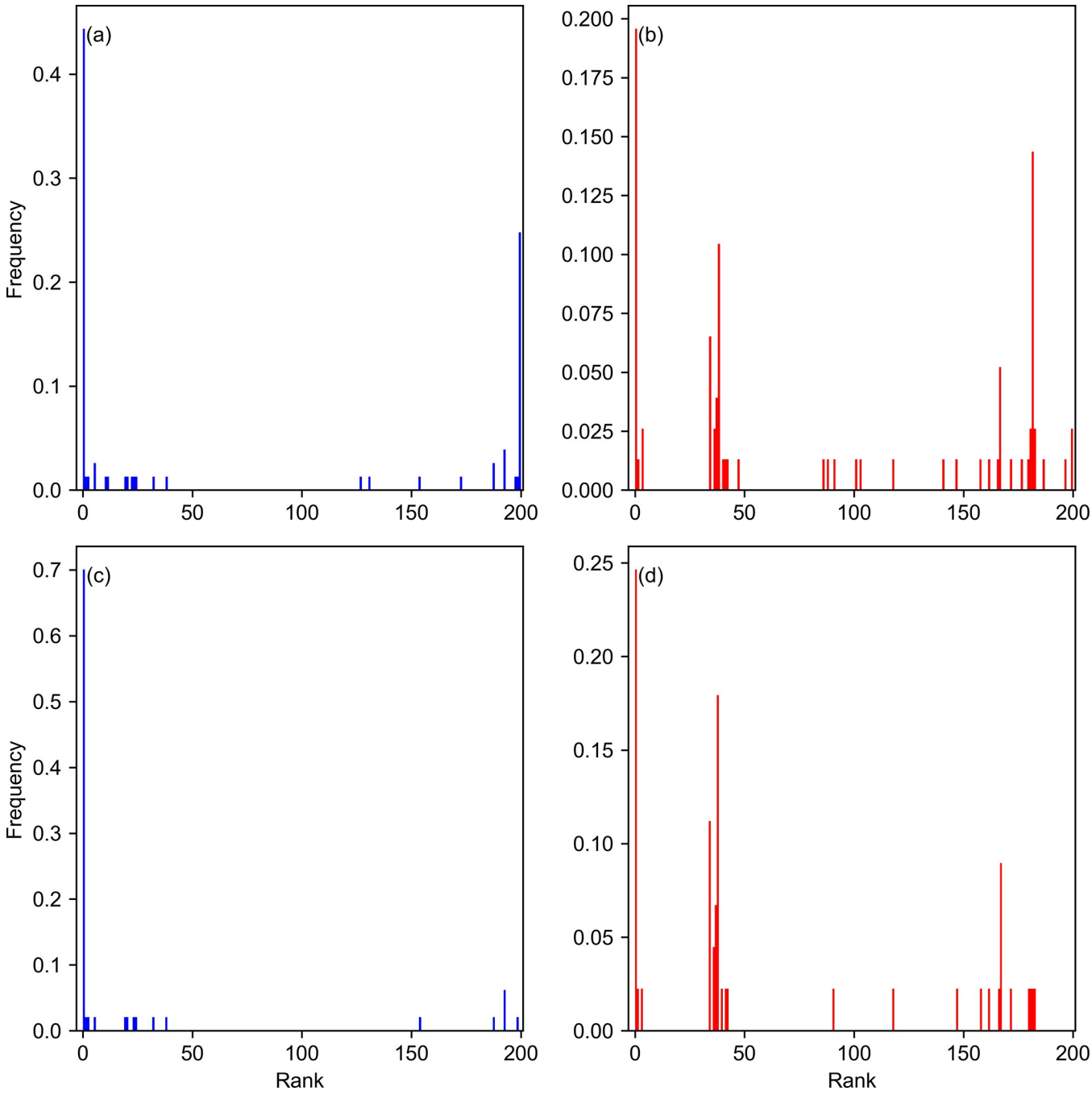

Fig 7. RHs for Tanzania: (a) baseline model–all observations (b) selected cost function–all observations (c) baseline model,–observations with FRC below 0.2 mg/L and (d) selected cost function,–observations with FRC below 0.2.

https://doi.org/10.1371/journal.pwat.0000040.g007

requires paired FRC measurements from the point-of-distribution and the point-of-consumption, implementing the cost-sensitive learning approach presented in this study would not require any changes from the existing data collection protocol recommended for the SWOT. Thus, the improvements in FRC forecasting ability can be implemented in the SWOT with no additional time investment from the user.

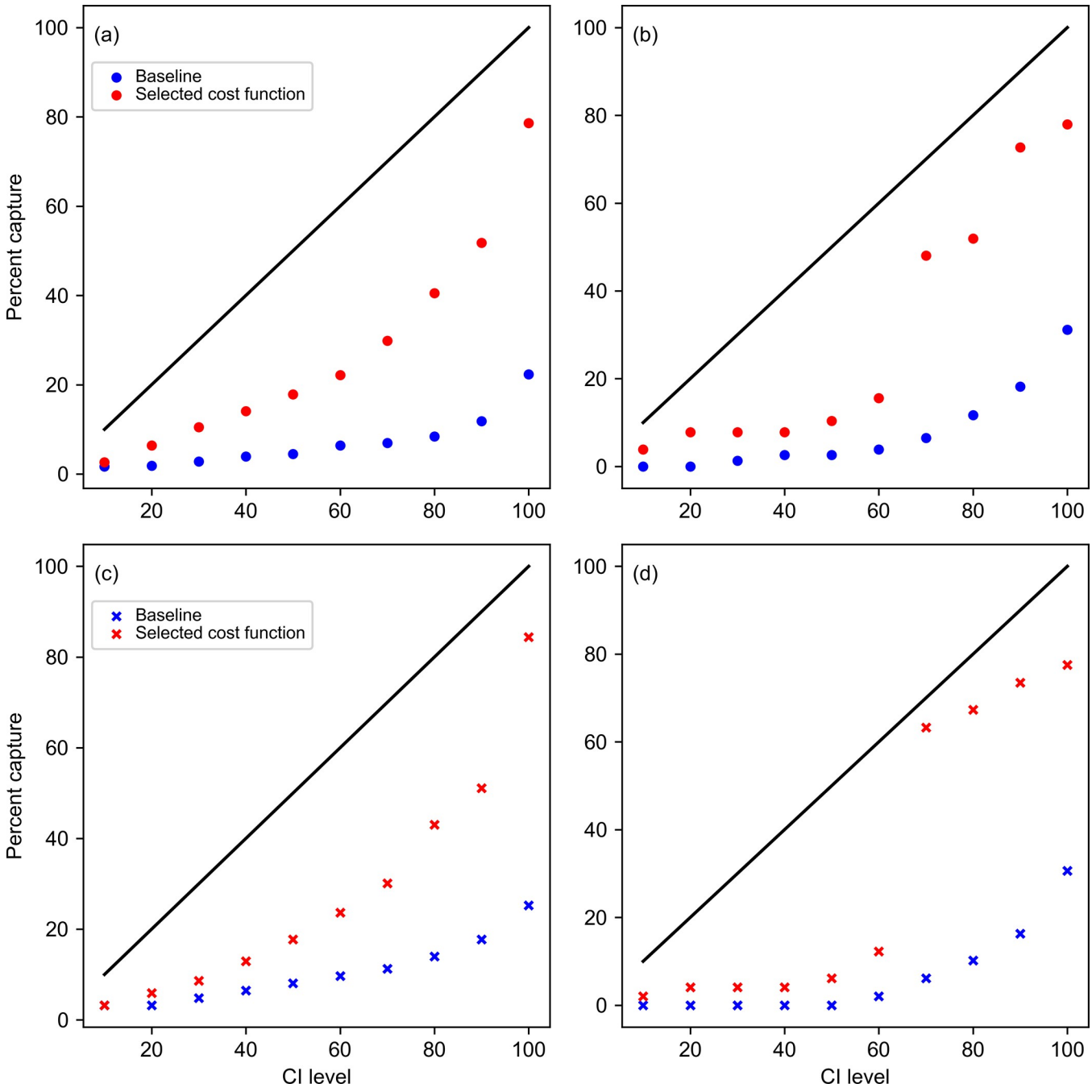

Fig 8. CI reliability diagrams for the baseline and selected cost function: (a) Bangladesh–all observations, (b) Tanzania–all observations, (c) Bangladesh–observations with FRC below 0.2 mg/L, and (d) Tanzania–observations with FRC below 0.2 mg/L.

https://doi.org/10.1371/journal.pwat.0000040.g008

4 Conclusion

Accurate forecasts of point-of-consumption FRC help water system operators in humanitarian response settings prevent the spread of waterborne illnesses. A major challenge in modelling FRC outside of the distribution system is the high degree of uncertainty in post-distribution chlorine decay. To account for this, probabilistic models like ANN-EFS are needed. This study

used alternative error metrics and cost-sensitive learning to train an ANN-EFS for forecasting post-distribution FRC in two refugee settlements in Bangladesh and Tanzania. We found that using these alternative error metrics and cost-sensitive learning techniques to train the ANN-EFS improved both the forecast dispersion and reliability relative to the ANN-EFS trained using the baseline cost function, unweighted MSE. We also selected KGE with inverse frequency weighting as the preferred cost function as it produced the best probabilistic performance for forecasting point-of-consumption FRC. This cost function should be used for forecasting point-of-consumption FRC in refugee and internally displaced person settlements and can be implemented in the Safe Water Optimization Tool to improve the reliability of the ANN-EFS forecasts and to improve the risk-based chlorination targets for water system operators in these settlements.

Supporting information

S1 Checklist. Checklist on inclusivity in global research.
(DOCX)

S1 Appendix. Data cleaning rules.
(DOCX)

S2 Appendix. Calculation of weighted cost functions.
(DOCX)

S1 Fig. Input and output variable histograms for the Bangladesh dataset.
(PNG)

S2 Fig. Input and output variable histograms for the Tanzania dataset.
(PNG)

S3 Fig. Bangladesh hidden layer size selection. Hidden layer size of 16 selected as it simultaneously delivers best Percent Capture and CI reliability score.
(PNG)

S4 Fig. Tanzania hidden layer size selection. Hidden layer size of 4 selected for good CI reliability performance without compromising capture and CRPS performance.
(PNG)

S5 Fig. Bangladesh ensemble size selection. Best performance at ensemble size of 200, so ensemble size of at least 200 required.
(PNG)

S6 Fig. Tanzania ensemble size selection. Substantial performance improvement when ensemble size is greater than 150, though beyond that, performance is highly variable.
(PNG)

S1 Table. Raw scores for each cost function (alternative error metric and cost sensitive learning combination) for Bangladesh and Tanzania using only FRC and elapsed time as input variables.
(XLSX)

S2 Table. Raw scores for each cost function (alternative error metric and cost sensitive learning combination) for Bangladesh and Tanzania using FRC, elapsed time, electrical conductivity, and water temperature as input variables.
(XLSX)

Acknowledgments

We would like to extend our gratitude for the support of our colleagues from the local refugee population, Médecins Sans Frontières (MSF), the United Nations High Commissioner for Refugees (UNHCR), and the Norwegian Refugee Council (NRC) in Bangladesh and Tanzania. We would also like to gratefully acknowledge James Orbinski of the Dahdaleh Institute for Global Health Research (DIGHR) for his advisory support on the SWOT project. We would also like to thank Everett Snieder for his input on this manuscript.

Author Contributions

Conceptualization: Michael De Santi, Usman T. Khan.

Data curation: Syed Imran Ali, Matthew Arnold, Anne M. J. Hyvärinen.

Funding acquisition: Syed Imran Ali, Jean-François Fesselet.

Methodology: Michael De Santi.

Project administration: Syed Imran Ali, Matthew Arnold, Jean-François Fesselet, Anne M. J. Hyvärinen, Dawn Taylor.

Software: Michael De Santi.

Supervision: Syed Imran Ali, Usman T. Khan.

Validation: Michael De Santi.

Visualization: Michael De Santi.

Writing – original draft: Michael De Santi.

Writing – review & editing: Syed Imran Ali, Matthew Arnold, Jean-François Fesselet, Anne M. J. Hyvärinen, Dawn Taylor, Usman T. Khan.

References

1. Altare C, Kahl V, Ngwa M, Goldsmith A, Hering H, Burton A, et al. Infectious disease epidemics in refugee camps: a retrospective analysis of UNHCR data (2009–2017). J Glob Health Report. 2019; 3: e2019064. https://doi.org/10.29392/joghr.3.e2019064
2. Golicha Q, Shetty S, Nasiblov O, Hussein A, Wainaina E, Obonyo M, et al. Cholera outbreak in Dadaab Refugee camp, Kenya—November 2015–June 2016. MMWR Morb Mortal Wkly Rep. 2018; 67: 958–961. https://doi.org/10.15585/mmwr.mm6734a4 PMID: 30161101
3. Shultz A, Omollo JO, Burke H, Qassim M, Ochieng JB, Weinberg M, et al. Cholera outbreak in Kenyan Refugee Camp: Risk Factors for Illness and Importance of Sanitation. Am J Trop Med Hyg. 2009; 80: 640–645. https://doi.org/10.4269/ajtmh.2009.80.640 PMID: 19346392
4. Swerdlow DL, Malenga G, Begkoyian G, Nyangulu D, Toole M, Waldman RJ, et al. Epidemic cholera among refugees in Malawi, Africa: treatment and transmission. Epidemiol Infect. 1997; 118: 207–214. https://doi.org/10.1017/s0950268896007352 PMID: 9207730
5. Walden VM, Lamond EA, Field SA. Container contamination as a possible source of a diarrhoea outbreak in Abou Shouk camp, Darfur province, Sudan. Disasters. 2005; 29: 213–221. https://doi.org/10.1111/j.0361-3666.2005.00287.x PMID: 16108988
6. Ali SI, Ali SS, Fesselet J-F. Effectiveness of emergency water treatment practices in refugee camps in South Sudan. Bull World Health Organ. 2015; 93: 550–558. https://doi.org/10.2471/BLT.14.147645 PMID: 26478612
7. Guerrero-Latorre L, Hundesa A, Girones R. Transmission Sources of Waterborne Viruses in South Sudan Refugee Camps. Clean (Weinh). 2016; 44: 775–780. https://doi.org/10.1002/clen.201500358
8. Howard CM, Handzel T, Hill VR, Grytdal SP, Blanton C, Kamili S, et al. Novel Risk Factors Associated with Hepatitis E Virus Infection in a Large Outbreak in Northern Uganda: Results from a Case-Control Study and Environmental Analysis. Am J Trop Med Hyg. 2010; 83: 1170–1173. https://doi.org/10.4269/ajtmh.2010.10-0384 PMID: 21036857

9. Steele A, Clarke B, Watkins O. Impact of jerry can disinfection in a camp environment—Experiences in an IDP camp in Northern Uganda. J Water Health. 2008; 6: 559–564. https://doi.org/10.2166/wh.2008.072 PMID: 18401121

10. Centres for Disease Control and Prevention (CDC). Chlorine Residual Testing. CDC. 2022 Jan 10 [cited 11 May 2022]. In: Global Water, Sanitation, & Hygiene (WASH). Atlanta: CDC. Available from: http://www.cdc.gov/safewater/chlorine-residual-testing.html.

11. Girones R, Carratalà A, Calgua B, Calvo M, Rodriguez-Manzano J, Emerson S. Chlorine inactivation of hepatitis e virus and human adenovirus 2 in water. J Water Health. 2014; 12: 436–442. https://doi.org/10.2166/wh.2014.027 PMID: 25252347

12. Lantagne DS. Sodium hypochlorite dosage for household and emergency water treatment. J–Am Water Works Assoc. 2008; 100: 106–114. https://doi.org/10.1002/j.1551-8833.2008.tb09704.x

13. Rashid M-u, George CM, Monira S, Mahmud T, Rahman Z, Mustafiz M, et al. Chlorination of Household Drinking Water among Cholera Patients ' Households to Prevent Transmission of Toxigenic Vibrio cholerae in Dhaka, Bangladesh: CHoBI7 Trial. Am J Trop Med Hyg. 2016; 95: 1299–1304. https://doi.org/10.4269/ajtmh.16-0420 PMID: 27698273

14. Sikder M, String G, Kamal Y, Farrington M, Rahman AS, Lantagne D. Effectiveness of water chlorination programs along the emergency-transition-post-emergency continuum: Evaluations of bucket, in-line, and piped water chlorination programs in Cox's Bazar. Water Res. 2020; 178: 115854. https://doi.org/10.1016/j.watres.2020.115854 PMID: 32361348

15. World Health Organization (WHO). WHO Guidelines for Drinking-water quality, 4th ed. Geneva, Switzerland. WHO; 2011.

16. Ali SI, Ali SS, & Fesselet J-F. Evidence-based chlorination targets for household water safety in humanitarian settings: Recommendations from a multi-site study in refugee camps in South Sudan, Jordan, and Rwanda. Water Res. 2021; 189: 116642. https://doi.org/10.1016/j.watres.2020.116642 PMID: 33246215

17. Wu H, Dorea C. Evaluation and application of chlorine decay models for humanitarian emergency water supply contexts. Environ Technol. 2021; 1–10. https://doi.org/10.1080/09593330.2021.1920626 PMID: 33880970

18. De Santi M, Khan UT, Arnold M, Fesselet J-F, Ali SI. Forecasting point-of-consumption chlorine residual in refugee settlements using ensembles of artificial neural networks. NPJ Clean Water. 2021; 4: 1–16. https://doi.org/10.1038/s41545-021-00125-2

19. Rajakuma AG, Kumar M, Amrutur B, Kapelan Z. Real-Time Water Quality Modeling with Ensemble Kalman Filter for State and Parameter Estimation in Water Distribution Networks. J Water Resour Plann Manage. 2019; 145: 04019049. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001118

20. Huang JJ, McBean EA. Using Bayesian statistics to estimate the coefficients of a two- component second-order chlorine bulk decay model for a water distribution system. Water Res. 2007; 41: 287–294. https://doi.org/10.1016/j.watres.2006.10.027 PMID: 17169396

21. Boucher M-A, Perreault L, Anctil F. Tools for the assessment of hydrological ensemble forecasts obtained by neural networks. J Hydroinformatics. 2009; 11: 297–307. https://doi.org/10.2166/hydro.2009.037

22. Boucher M-A, Laliberté J-P, Anctil F. An experiment on the evolution of an ensemble of neural networks for streamflow forecasting. Hydrol Earth Syst Sci. 2010; 603–612. https://doi.org/10.5194/hess-14-603-2010

23. Shrestha DL, Solomatine DP. Data-driven approaches for estimating uncertainty in rainfall-runoff modelling. Int J River Basin Manag. 2008; 6: 109–122. https://doi.org/10.1080/15715124.2008.9635341

24. Aliashrafi A, Zhang Y, Groenwegen H, Peleato NM. A review of data-driven modelling in drinking water treatment. Rev Environ Sci Biotechnol. 2021; 20:985–1009. https://doi.org/10.1007/s11157-021-09592-y

25. Dogo EM, Nwulu NI, Twala B, Aigbavboa C. A survey of machine learning methods applied to anomaly detection on drinking-water quality data. Urban Water J. 2019; 16: 235–248. https://doi.org/10.1080/1573062X.2019.1637002

26. Li L, Rong S, Wang R, Yu S. Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: A review. Chem Eng J. 2021; 405: 126673. https://doi.org/10.1016/j.cej.2020.126673

27. O'Reilly G, Bezuidenhout CC, Bezuidenhout JJ. Artificial neural networks: applications in the drinking water sector. 2018. Water Supply. 2018; 18: 1869–1887. https://doi.org/10.2166/ws.2018.016

28. Wadkar DV, Nangare P, Wagh MP. Evaluation of water treatment plant using Artifcial Neural Network (ANN) case study of Pimpri Chinchwad Municipal Corporation (PCMC). Sustain Water Resour Manag. 2021; 7:1–14. https://doi.org/10.1007/s40899-021-00532-w

29. Wang D, Shen J, Zhu S, Jiang G. Model predictive control for chlorine dosing of drinking water treatment based on support vector machine model. Desalination Water Treat. 2020; 173: 133–141. https://doi.org/10.5004/dwt.2020.2414

30. Godo-Pla L, Emiliano P, Valero F, Poch M, Sin G, Monclus H. Predicting the oxidant demand in full-scale drinking water treatment using an artificial neural network: Uncertainty and sensitivity analysis. Process Saf Environ Prot. 2019; 125: 317–327. https://doi.org/10.1016/j.psep.2019.03.017

31. Deng Y, Zhou X, Shen J, Xiao G, Hong H, Lin H, et al. New methods based on back propagation (BP) and radial basis function (RBF) artificial neural networks (ANNs) for predicting the occurrence of haloketones in tap water. Sci Total Environ. 2021; 772: 145534. https://doi.org/10.1016/j.scitotenv.2021.145534 PMID: 33571763

32. Gheibi M, Eftekhari M, Akrami M, Emrani N, Hajiaghaei-Kesheli M, Fatollahi-Fard A, et al. A Sustainable Decision Support System for Drinking Water Systems: Resiliency Improvement against Cyanide Contamination. Infrastructures. 2022; 7: 88. https://doi.org/10.3390/infrastructures7070088

33. Wang H, Koydemir HC, Qiu Y, Bai B, Zhang Y, Jin Y, et al. Early detection and classification of live bacteria using time-lapse coherent imaging and deep learning. Light Sci Appl. 2020; 9: 118. https://doi.org/10.1038/s41377-020-00358-9 PMID: 32685139

34. Rodriguez MJ, Sérodes JB. Assessing empirical linear and non-linear modelling of residual chlorine in urban drinking water systems. Environ Model Softw. 1998; 14: 93–102. https://doi.org/10.1016/S1364-8152(98)00061-9

35. Gibbs MS, Morgan N, Maier HR, Dandy GC, Nixon JB, Holmes H. Investigation into the relationship between chlorine decay and water distribution parameters using data driven methods. Mathematical and Computer Modelling. 2006; 44: 485–498. https://doi.org/10.1016/j.mcm.2006.01.007

36. Bowden GJ, Nixon JB, Dandy GC, Maier HR, Holmes, M. Forecasting chlorine residuals in a water distribution system using a general regression neural network. Mathematical and Computer Modelling. 2006; 44:469–484. https://doi.org/10.1016/j.mcm.2006.01.006

37. Soyupak S, Kilic H, Karadirek IE, Muhammetoglu H. On the usage of artificial neural networks in chlorine control applications for water distribution networks with high quality water. J Water Supply: Res Technol–AQUA. 2011; 60: 51–60. https://doi.org/10.2166/aqua.2011.086

38. Oyuntha C, Kwio-Tamale JC. Modelling chlorine residuals in drinking water: a review. Int J Environ Sci Technol. 2022. https://doi.org/10.1007/s13762-022-03924-3

39. Safe Water Optimization Tool. Safe Water Optimization Tool [Internet]. Toronto (CAN): Dahdaleh Institute for Global Health Research. 2019. [cited– 2022 July 10]. Available from: https://www.safeh2o.app/

40. Solomatine DP, Ostfeld A. Data-driven modelling: Some past experiences and new approaches. J Hydroinformatics. 2008; 10: 3–22. https://doi.org/10.2166/hydro.2008.015

41. Crone SF, Lessmann S, Stahlbock R. Utility based data mining for time series analysis—Cost-sensitive learning for neural network predictors. In: Weiss G, Saar-Tsechanksy M, Zadrozny B, editors. UBDM '05: Proceedings of the 1st International Workshop on Utility-Based Data Mining; 21 Aug 2005; Chicago, United States of America. Association for Computing Machiney, New York, United States of America. pp 59–68. https://doi.org/10.1145/1089827.1089835

42. Toth E. Estimation of flood warning runoff thresholds in ungauged basins with asymmetric error functions. Hydrol Earth Syst Sci. 2016; 20: 2383–2394. https://doi.org/10.5194/hess-20-2383-2016

43. Thiboult A, Anctil F, Boucher M-A. Accounting for three sources of uncertainty in ensemble hydrological forecasting. Hydrol Earth Syst Sci. 2016; 20: 1809–1825

44. de Vos NJ, Rientjes THM. Multiobjective training of artificial neural networks for rainfall-runoff modeling. Water Resour Res, 2008; 44: 1–15. https://doi.org/10.1029/2007WR006734

45. Elkan C. The Foundations of Cost-Sensitive Learning. In: Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence; August 4–10, 2001; Seattle, Washington. Morgan Kaufmann Publishers Inc, San Francisco, United States of America. pp 973–978. Available from https://cseweb.ucsd.edu//~elkan/rescale.pdf

46. Petrides G, Verbeke W. Cost-sensitive ensemble learning: a unifying framework. Data Min Knowl Discov. 2022; 36:1–28. https://doi.org/10.1007/s10618-021-00790-4

47. Almeida AM, Castel-Branco MM, Falcão AC. Linear regression for calibration lines revisited: Weighting schemes for bioanalytical methods. J Chromatogr B: Anal Technol Biomed Life Sci, 2002; 774: 215–222.

48. Kneale P, See L, Smith A. Towards Defining Evaluation Measures for Neural Network Forecasting Models. In: Proceedings of the Sixth International Conference on GeoComputation; 24–26 Sep 2001;

Brisbane, Australia. GeoComputation: Leeds, United Kingdom, 2001. Available from: http://www.geocomputation.org/2001/papers/kneale.pdf

49. Snieder E, Abogadil K, Khan UT. Resampling and ensemble techniques for improving ANN-based high-flow forecast accuracy. Hydrol Earth Syste Sci. 2021; 25: 2543–2566. https://doi.org/10.5194/hess-25-2543-2021

50. Olowookere TA, Adewale OS. A framework for detecting credit card fraud with cost-sensitive meta-learning ensemble approach. Sci Afr. 2020; 8: e00464. https://doi.org/10.1016/j.sciaf.2020.e00464

51. Xu Q, Lu S, Jia W, Jiang C. Imbalanced fault diagnosis of rotating machinery via multi-domain feature extraction and cost-sensitive learning. J Intell Manuf. 2020; 31: 1467–1481. https://doi.org/10.1007/s10845-019-01522-8

52. Leo J, Luhanga E, Michael K. Machine Learning Model for Imbalanced Cholera Dataset in Tanzania. Sci World J. 2019; 2019: 9397578. https://doi.org/10.1155/2019/9397578 PMID: 31427903

53. Zhang L, Huang J, Liu L. Improved Deep Learning Network Based in combination with Cost-sensitive Learning for Early Detection of Ovarian Cancer in Color Ultrasound Detecting System. J Med Syst. 2019; 43: 251. https://doi.org/10.1007/s10916-019-1356-8 PMID: 31254110

54. Chen X, Liu H, Liu F, Huang T, Shen R, Deng Y, et al. Two novelty learning models developed based on deep cascade forest to address the environmental imbalanced issues: A case study of drinking water quality prediction. Environ Pollut. 2021; 291: 118153. https://doi.org/10.1016/j.envpol.2021.118153 PMID: 34534828

55. Chen X, Liu H, Xu X, Zhang L, Lin T, Zuo M, et al. Identification of Suitable Technologies for Drinking Water Quality Prediction: A Comparative Study of Traditional, Ensemble, Cost-Sensitive, Outlier Detection Learning Models and Sampling Algorithms. Environ Sci Technol Water. 2021; 1: 1676–1685. https://doi.org/10.1021/acsestwater.1c00037

56. Python Software Foundation. Python v3.7.4. [Internet]. 2019. Available from: https://www.python.org/downloads/release/python-374/.

57. Chollet F. Keras. [Internet]. 2015. Available from: https://keras.io.

58. Kotlarz N, Lantagne D, Preston K, Jellison K. Turbidity and chlorine demand reduction using locally available physical water clarification mechanisms before household chlorination in developing countries. J Water Health. 2009; 7: 497–506. https://doi.org/10.2166/wh.2009.071 PMID: 19491500

59. Garcìa R, Naves A, Anta J, Ron M, Molinero J. Drinking water provision and quality at the Sahrawi refugee camps in Tindouf (Algeria) from 2006 to 2016. Sci Total Environ. 2021; 780: 146504. https://doi.org/10.1016/j.scitotenv.2021.146504 PMID: 34030293

60. Bröcker J. Evaluating raw ensembles with the continuous ranked probability score. Q J R Meteorol Soc. 2012; 138: 1611–1617. https://doi.org/10.1002/qj.1891

61. Hamill TM. Interpretation of Rank Histograms for Verifying Ensemble Forecasts. Mon Weather Rev. 2001; 129: 550–560. https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2

62. Gupta HV, Kling H, Yilmaz KK, Martinez GF. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J Hydrol. 2009; 377: 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

63. Willmott CJ. On the Validation of Models. Phys Geogr, 1981; 2: 184–194. https://doi.org/10.1080/02723646.1981.10642213

64. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell, 2016; 5: 221–232. https://doi.org/10.1007/s13748-016-0090

65. Zhou Z-H, Liu X-Y. On Multi-Class Cost-Sensitive Learning. Comput Intell. 2010; 26: 232–257. https://doi.org/10.1111/j.1467-8640.2010.00358.x

66. Kumpel E, Nelson KL. Comparing microbial water quality in an intermittent and continuous piped water supply. Water Res. 2013; 47: 5176–5188. https://doi.org/10.1016/j.watres.2013.05.058 PMID: 23866140

67. Médecins Sans Frontières (MSF). Public Health Engineering In Precarious Situations (2nd ed.). Brussels, Belgium. MSF: 2014.

68. United Nations High Commissioner for Refugees (UNHCR). *UNHCR WASH Manual*, *Practical Guidance for Refugee Settings (* 7th ed.*)*. Geneva, Switzerland. UNHCR: 2020.

69. LeChevallier MW, Welch NJ, Smith DB. Full-scale studies of factors related to coliform regrowth in drinking water. Appl Environ Microbiol. 1996; 62: 2201–2211. https://doi.org/10.1128/aem.62.7.2201-2211.1996 PMID: 8779557

70. Ling CX, Sheng VS. Cost-Sensitive Learning and the Class Imbalance Problem. In: Sammut C, Webb GI editors. Encyclopedia of Machine Learning. Springer; 2010 pp. 231–235. https://doi.org/10.1007/978-0-387-30164-8_181

71. Liu XY, Zhou ZH. The influence of class imbalance on cost-sensitive learning: An empirical study. In: ICDM '06: Proceedings on the Sixth International Conference on Data Mining; 18–22 Dec 2006; Hong Kong, China. IEEE Computer Society, Washington, DC, UInited States. pp. 970–974. https://doi.org/10.1109/ICDM.2006.158

72. McCarthy K, Zabar B, Weiss G. Does cost-sensitive learning beat sampling for classifying rare classes? In: Weiss G, Saar-Tsechanksy M, Zadrozny B, editors. UBDM '05: Proceedings of the 1st International Workshop on Utility-Based Data Mining; 21 Aug 2005; Chicago, United States of America. Association for Computing Machiney, New York, United States of America. pp. 69–77. https://doi.org/10.1145/1089827.1089836

73. Gneiting T, Balabdaoui F, Raftery AE. Probabilistic forecasts, calibration and sharpness, J R Statist Soc B. 2007; 69: pp 243–268. https://doi.org/10.1111/j.1467-9868.2007.00587.x

74. Talagrand O, Vautard R, Strauss B. Evaluation of probabilistic prediction systems. In Proceedings, ECMWF Workshop on Predictability; 20–22 Oct 1997; Shinfield Park, Reading, United Kingdom. European Centre for Medium-Range Weather Forecasts (ECMWF). pp. 1–25. Available from: https://www.ecmwf.int/node/12555

75. Candille G, Talagrand O. Evaluation of probabilistic prediction systems for a scalar variable. Q J R Meteorol Soc. 2005; 131: 2131–2150. https://doi.org/10.1256/qj.04.71

76. Ferro CAT. Fair scores for ensemble forecasts. Q J R Meteorol Soc. 2014; 140: 1917–1923. https://doi.org/10.1002/qj.2270

77. Hersbach H. Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. Weather Forecast. 2000; 15: 559–570. https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2

78. Boucher MA, Perreault L, Anctil F, Favre AC. Exploratory analysis of statistical post-processing methods for hydrological ensemble forecasts. Hydrol Process. 2015; 29: 1141–1155. https://doi.org/10.1002/hyp.10234

79. Dress K, Lessmann S, von Mettenheim HJ. Residual value forecasting using asymmetric cost functions. Int J Forecast. 2018; 34: 551–565. https://doi.org/10.1016/j.ijforecast.2018.01.008

80. Wu W, Dandy GC, Maier HR. Protocol for developing ANN models and its application to the assessment of the quality of the ANN model development process in drinking water quality modelling. Environ Model Softw. 2014; 54: 108–127. https://doi.org/10.1016/j.envsoft.2013.12.016

81. Paulino R, Bérubé P. A framework for the use of artificial neural networks for water treatment: development and application. Wat Supply. 2020; 20: 3301–3317. https://doi.org/10.2166/ws.2020.205