The Safe Water Optimization Tool Artificial Neural Network Analytics for the SWOT v2.0

Revision History

Date	Version Number	Summary of Changes
2021-06-10	0	N/A

Executive Summary

This report presents a summary of the Safe Water Optimization Tool (SWOT) artificial neural network (ANN) analytics and seeks to provide transparency into these analytics. The SWOT-ANN analytics were introduced to address the high levels of uncertainty in post-distribution chlorine decay, which poses a major challenge in modelling household free residual chlorine (FRC) in refugee and internally displaced person (IDP) settlements. To overcome this uncertainty, the SWOT-ANN analytics use ANNs, a type of data driven model, to avoid making assumptions about the uncerlying decay behaviour, and groups these models into an ensemble forecast which models household FRC probabilistically, allowing water system operators to understand the risk of drinking water having insufficient FRC at the household.

One of the key features of the SWOT-ANN version 2 analytics is dynamic input variable selection, where the input variables for the models are determined based on the dataset uploaded by the user. The SWOT-ANN always uses tapstand FRC, elapsed time, and time of collection as input variables, but if a sufficient number of measurements are available, the SWOT-ANN also uses electrical conductivity and water temperature.

The SWOT-ANN version 2 analytics also include new performance diagnostics which investigate the probabilistic performance using the percent capture, confidence interval reliability diagram, rank histogram, and continuous ranked probability score. These scores provide a better indication of the probabilistic performance of the ANN ensembles and are thus better diagnostics of the accuracy of the risk-based FRC targets generated by the SWOT, as copared to deterministic performance measures like R^2. To improve the probabilistic performance, the SWOT-ANN version 2 analytics also feature ensemble post-processing using kernel dressing which is a distribution-free post-processing method, meaning that this method improves the forecasting performance without forcing the forecast to fit a pre-defined distribution

The version 2 analytics of the SWOT-ANN also introduce a scenario analysis feature which generates multiple risk-based FRC targets based on tapstand water quality scenarios, as well as the time of collection which has substantial impact on the post-distribution FRC decay. This provides water system operators with additional information, allowing them to tailor their risk-based FRC targets to site conditions. This also allows water system operators to better understand decay behaviours occurring on site.

This White Paper is intended as a living document, with updates introduced as further refinement of the SWOT-ANN version 2 analytics are applied. We also include appendices that summarize key analyses we performed to select the modelling parameters included in the version 2 analytics, and a functions glossary to summarize the functions include in the SWOT-ANN version 2 code, which is available on GitHub here: https://github.com/safeh2o/swot-python-analysis

Table of Contents

Contents

Figures

Tables

1 Introduction

1.1 Motivation

This white paper presents the details of the second version of the Safe Water Optimization Tool artificial neural network tool (SWOT-ANN). The SWOT-ANN is a probabilistic, data-driven, tool used to generate free residual chlorine (FRC) guidance for water system operators in humanitarian response settings. These analytics were introduced to address two of the major challenges of modelling FRC during the post-distribution phase of collection, transport, storage, and use. The first major challenge is the limited study into the specific phenomena that drive chlorine decay during the post-distribution phase. This leads to a limited understanding of the expected chlorine decay behaviour, especially as this behaviour may change over the course of household storage as new contaminants may be introduced. All of this makes it difficult to select an appropriate decay model for the post-distribution phase. To overcome this, we use ANNs which are a data driven approach that can learn the behaviour from the underlying data without making any prior assumptions about the chlorine decay behavior. The second major challenge of modelling FRC during the post-distribution phase is that the decay behaviour is highly variable and may be impacted by a number of factors, many of which may not be easily quantifiable (e.g., user interactions with the water, frequency of container cleaning, change in water temperature during storage, etc.). In practice, this results in a single set of conditions at the tapstand producing a range of household FRC concentrations, making point predictions of household FRC insufficient. To overcome this challenge, the SWOT-ANN used a probabilistic ensemble modelling approach by grouping the predictions of multiple individual ANNs into a probabilistic ensemble forecast. This forecast quantify the uncertainty in the predicted household FRC concentration and provide information about the distribution of household FRC concentrations. We developed a probabilistic modelling approach by developing an ensemble of ANNs. The SWOT-ANN uses these probabilistic forecasts to generate risk-based tapstand FRC guidance based on the probability of having insufficient FRC (<0.2 mg/L) at the point of consumption in the household. This white paper presents the details of the analytics used to generate this probabilistic FRC guidance in an effort to provide transparency into the analytical approach taken in the SWOT-ANN version 2 analytics. The SWOT-ANN code is available on the SWOT project GitHub page at: https://github.com/safeh2o/swot-python-analysis.

1.2 Included in this Report

Section 1 of this report provides the introduction to the SWOT-ANN and the motivation for this tool as well as this white paper. Section 2 provides a high-level summary of the version 2 SWOT-ANN tool. Sections 3 and 4 provides a summary of the backend tasks included in the SWOT-ANN, with Section 3 covering the importing of the data as well as data pre-processing tasks and input variable selection and Section 4 providing the details of training the ensemble model starting with the model set up and extending to evaluating the model performance and post-processing the results. Finally, Section 5 summarizes how the SWOT-ANN generates a recommendation and provides guidance on using the outputs to determine the FRC target. There is also a glossary of the functions implemented in the SWOT-ANN code and appendices summarizing some of the key decisions that went into preparing version 2 of the SWOT-ANN.

2 Workflow of the Version 2 SWOT ANN Analytics

Figure 1 provides a high-level workflow of the process used to generate risk-based FRC guidance using the SWOT ANN version 2 analytics. The uploaded data set is first pre-cleaned by the SWOT web analytics prior to importing the data into Python. Once imported into Python, additional data pre-processing occurs to ensure that the SWOT-ANN is able to run. This step also includes selecting the appropriate input variable combination. After this, the model is set up and the data is used to train each individual neural network in the ensemble. Third, we evaluate the model performance using the provided data to understand how well the models reproduce the underlying behaviour. At this point we also post-process the ensemble predictions to determine if post-processing improves the model performance. Fourth, we use several sets of fixed inputs to perform a scenario analysis by simulating potential conditions at the tapstand. If post-processing was shown to improve performance in the third step, we also post-process the forecasts on fixed inputs. Finally, we use these forecasts to predict the risk of inadequate household FRC and to generate a recommendation for a series of scenarios.

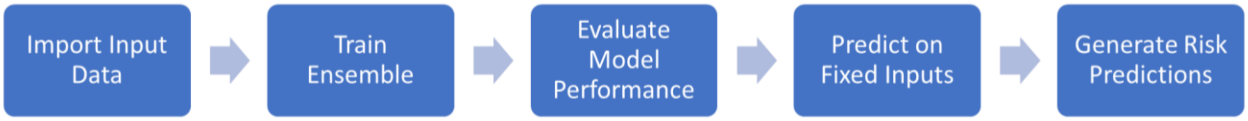


Figure 1: High-level modelling workflow

3 Importing the Data

Data is received by the ANN analytical module following some initial pre-cleaning through the SWOT web tool. At this point, the SWOT ANN analytics import the data as a .csv file and perform the following tasks:

1. The column names are used to identify the input variables and the output variable (household FRC).
2. The tapstand and household timestamps for each sample are used to determine the elapsed time in hours for each sample as well as the time of collection, which is converted into a binary variable for collection before or after noon (AM/PM collection).
3. The input variable set is determined (Section 3.1)

4. For the selected input variables, rows with missing entries for any variable are removed. This step is required for training the ANNs as training will stop if a missing value is encountered.

3.1 Input Variable Selection

Input variable selection for the SWOT ANN analytics is not predefined and instead is a dynamic process with the input variables selection occurring within the SWOT-ANN. All possible input variables, and the rationale for their inclusion, are listed below:

- **Tapstand FRC:** The tapstand FRC is intuitively a crucial variable for modelling household FRC. This is confirmed by a partial correlation analysis by De Santi et al. (2021) that showed that of all routinely collected water quality variables for refugee and IDP settlements, tapstand FRC has the greatest influence on the household FRC.
- **Elapsed time (hours):** The elapsed time here refers to the period of time beginning when water leaves the tapstand and ending at the time of the household FRC measurement. While FRC decay is a time dependent reaction, past studies have shown that elapsed time on its own is not a strong predictor of point-of-consumption FRC, likely due to confounding with other variables, such as the time-of-collection (De Santi et al., 2021). For this reason, we include both elapsed time and time-of-collection in the ANN models to help clarify the influence of elapsed time. The rationale for the selection of time-based variables is provided in Appendix A.
- **Time of Collection (binary, AM/PM):** This variable denotes the time of collection measured when water leaves the tapstand. This time of collection is converted into a binary variable for AM or PM collection (for samples collected before and after 12:00 noon, respectively). This variable is included to help clarify the influence of elapsed time by disaggregating the data into morning collection which includes hotter periods of the day and which typically allows for more user interaction with the water due to daytime storage, and afternoon collection which typically includes overnight storage where water temperatures are cooler and less user interaction. This variable was included based on an investigation into alternative approaches to incorporating time-related data into the ANN model. Further detail on the selection of time of collection as an input variable is included in Appendix A.
- **Tapstand Water Temperature (°C):** Water temperature is measured from the water directly as it leaves the tapstand. We included this variable as water temperature has been shown to impact FRC decay due to the effect of water temperature on the rate of chemical reactions in studies of piped distribution systems (Clark and Sivaganesan, 2002; Fisher et al., 2017; Powell et al., 2000; Warton et al., 2006). Water temperature has also been shown to have an impact on post-distribution FRC decay (De Santi et al., 2021).

- **Electrical Conductivity (µs/L):** Electrical conductivity (EC) is measured from the water directly as it leaves the tapstand. EC is an indicator of dissolved ions and is not a direct measure of chlorine demand in the water (World Health Organization, 2011), though it may provide an indication of inorganic chlorine demand. We have included EC in the model as past studies have found that EC to be strongly associated with FRC decay during the post-distribution phase (Ali et al., 2015; De Santi et al., 2021).

Of the potential input variables, the first three: tapstand FRC, elapsed time, and time of collection are always included in the model as all records will require at least FRC and timestamp information for the tapstand and household. Water temperature and EC, however, may not be available at all sites. For this reason, the SWOT-ANN analytics check for the number of observations missing each of these measurements and if more than 90% of the records are missing a measurement for one of these variables, that variable is removed. The 90% missing measurement threshold was selected as an indicator that a variable was not included in routine water quality monitoring. We use the 90% threshold instead of a 100% threshold in cases where there may be data entry issues, transition between data collection practices, or other anomalies where a very small number of samples have these measurements are included despite these variables not being included in routine monitoring. We based this decision on an analysis of model performance across SWOT sites using different variable combinations which found that the SWOT-ANN models tended to perform best when more water quality variables were included, even when a large percentage of records for those water quality measurements were missing. More details on this analysis and on the selection of the 90% threshold for removing input variables are included in Appendix B.

4 Training the SWOT-ANN model

4.1 Model Set-Up and Architecture

The model architecture used by the SWOT-ANNs is an ensemble of 200 artificial neural networks. The individual neural networks in the ensemble are referred to as the base learners. The base learners used for the SWOT-ANN ensemble are multi-layer perceptrons (MLPs). This type of ANN consists of three types of layers of interconnected nodes: an input layer, one or more hidden layers, and an output layer, as shown in Figure 1. The MLP structure with one hidden layer was selected because it has been shown to outperform other types of ANN architectures and data-driven models for predicting FRC in piped distribution systems, especially when predicting extreme values (Gibbs et al., 2006; Rodriguez and Sérodes, 1998). Additionally, this ANN structure has been demonstrated to be an effective architecture for modelling post-distribution FRC (De Santi et al., 2021). In the MLP, predictor variable data enters the model at the input layer, is fed forward to the hidden layer, and then data from each node of the hidden layer is passed to the output layer. As data move along the connections from one layer to the next, the values are multiplied by a weight specific to that connection. At each node an activation function determines if information will continue to propagate through the network and a numerical bias is added to the value at that node.

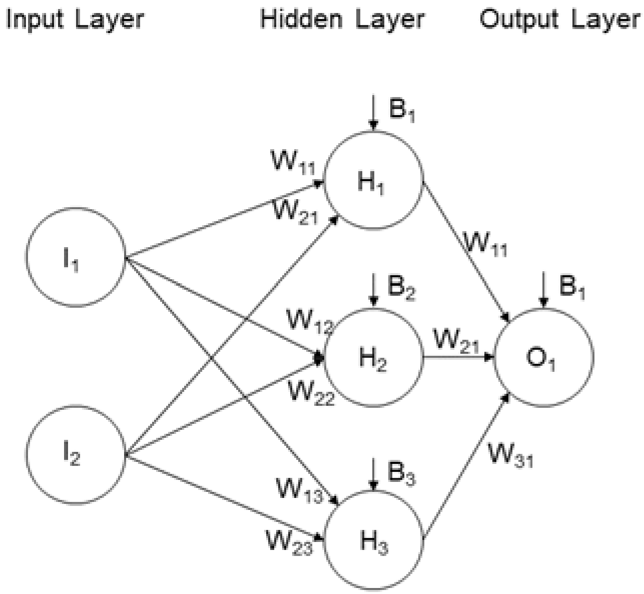


Figure 2: Schematic of an MLP showing flow of data from the input layer to the output layer with weights and biases. The shown MLP with two input nodes and one output node would have two input variables (other water quality parameters, etc.) and one output (household FRC).

The MLP base learners used in the SWOT-ANNs have 1 output node for the single output variable (household FRC), and twelve hidden nodes which was determined via a preliminary analysis of model performance using datasets from three sites actively using the SWOT-ANN analytics. The size of the input layer is not predetermined and is instead selected to match the number of input variables, which is determined during the importing of the input data (c.f. Section 3.1). This is a departure from version 1 of the SWOT-ANN where the model architecture was predefined, but the change is necessary to facilitate a flexible approach to input variable selection. The SWOT-ANN base learners use a hyperbolic tangent activation function on the hidden layer and a linear activation function on the output layer.

4.2 Training the Ensemble

To train and test the ensemble base learners, the imported dataset is first rescaled between -1 and 1 using the SciKit Learn MinMaxScaler package (Pedregosa et al., 2011) to speed up the convergence of the training process and to ensure that all input variables contribute equally to the output variable at the beginning of training. The overall dataset is then divided into two subsets: the training set and the validation set. These subsets are determined by randomly sampling 33.3% of the data for training and 66.7% for validation. The sampling is randomized for each base learner so that the allocation of the dataset into the training and validation subsets is different for each individual ANN. The network is trained by starting with a random set of weights and biases which are then iteratively adjusted to minimize the mean squared error (MSE) of the predictions on the training set using the Nadam backpropagation training algorithm. During training, the MSE on the validation set is also calculated and is used to determine the stopping point when training the base learners. Initially during training, both the training and validation MSE should decrease, indicating improvement, but as training continues these will diverge, with the training MSE continuing to decrease while the validation MSE increases. This indicates that that the model is overfitting (i.e., becoming overly specific to the training data and thus less useful for predicting on new data). When the validation MSE begins to increase, training was stopped using an early stopping procedure. The early stopping procedure in the SWOT-ANN uses a patience of 10 epochs, meaning that after the validation MSE begins to increase, training continues for 10 more epochs (iterations) to see if the validation MSE will decrease again, at which point training resumes as normal. If the validation MSE does begin to decrease again, then the SWOT-ANN restores the weights and biases from the iteration with the lowest validation MSE. At this point training is complete. The model weights and biases are saved, and then the Tensorflow training state is reset for the next base learner in the ensemble. This reset is critical for removing training state data and ensuring that the individual ensemble members are independent of each other.

4.3 Evaluating Model Performance

Once all 200 base learners have been trained, their predictions on the full dataset are used to evaluate the probabilistic performance of the ensemble model. To do this, the trained ensemble is used to predict the household FRC for all observations in the full dataset, and the predictions from all 200 base learners are grouped into a probability density function (pdf) for each observation. This pdf is referred to as the forecast and each forecast and its corresponding observation is a forecast-observation pair. The SWOT-ANN then evaluates the probabilistic performance of these forecasts using 4 performance metrics: the percent capture, the confidence interval (CI) reliability score, the δ-score, and the continuous ranked probability score (CRPS). 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. F refers to the full set of forecasted point-of-consumption FRC concentrations forecasted by the ensembles, where f_i^m is the prediction by the m^{th} base learner in the ensemble on 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 $m = 0$ to $m = M$.

Note, typically the performance of an ensemble forecast is only evaluated on data that has not been used in the training or calibration of the model. However, for the SWOT-ANN this would require either a two-step training process (first with the test set left out, then again with the full set) or it would require some data to be left out from the training process. Thus, the performance evaluation included in the SWOT-ANN analytics is performed using the same data that was used to train and validate the ensemble with the knowledge that the resulting ensemble performance is only an approximation of performance and does not necessarily reflect the performance we would expect on new data.

4.3.1 Percent Capture

Percent capture measures the percentage of observations where the observed household FRC concentration was within the limits of the ensembles forecast. The percent capture is a positively oriented score, meaning that a higher percent capture indicates better performance (more observations capture within the forecast limits) with an upper limit of 100% and a lower limit of 0%. Observation o_i is considered captured if $f_i^0 \leq o_i \leq f_i^M$. When evaluating the ensemble performance, the SWOT-ANN considers both the percent capture of the overall dataset (referred to in this report as PC) as well as the percent capture of observations with point-of-consumption FRC below 0.2 mg/L ($PC_{<0.2}$).

4.3.2 CI Reliability Score

The CI reliability score is derived from the CI reliability diagram which shows the percentage of total observations captured within each ensemble CI within the ensemble plotted against the CI level. This provides a visual indicator of ensemble performance as the ideal model will have all points plotted along the 1:1 line showing that the observed probabilities are equal to the forecasted probabilities. The CI reliability score is calculated as the squared distance between the percent capture within each CI and the ideal percent capture in that CI (De Santi et al., 2021). This was calculated for each CI threshold, $k,$ from 10% to 100% in 10% increments as shown in Equation 1. 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. The SWOT-ANN plots CI reliability diagrams and calculates the CI reliability score for both the overall data set (CI_{score}) and for forecast-observation pairs where the observed point-of-consumption FRC concentration was below 0.2 mg/L ($CI_{score_{<0.2}}$).

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

4.3.3 Rank Histogram δ-score

The Rank Histogram (RH) is another visual tool used to assess the reliability of ensemble forecasts. It is constructed by assigning a rank to each observation based on the observed household FRC value relative to the predicted value of each ensemble member and then making a histogram of these ranks. If the forecast and observed probabilities are the same, then any observation is equally likely to occur in any rank of the ensemble, which would result in a flat rank RH (uniform distribution). If the forecasted and observed probability distributions are different, then the rank histogram will not be flat and may be either u-shaped, indicating underdispersion, arch-shaped, indicating overdispersion; or skewed, indicating bias (Hamill, 2001; Talagrand et al., 1997). The flatness, or degree of uniformity, of the RH is quantified in the δ score which measures the deviations from flatness in the RH (Equation 2). The ideal δ-score is 1 with scores much greater than 1 indicating substantial deviations from flatness and scores less than 1 indicating interdependence between ensemble predictions (Candille and Talagrand, 2005). The SWOT-ANN calculates the δ score for each model both for the overall dataset (δ) and for only those observations where the observed point-of-consumption FRC was below 0.2 mg/L ($\delta_{<0.2}$).

$$\delta = \frac{\Delta}{\Delta_o} \tag{2}$$

The two components of the δ score are shown in Equations 3 and 4 where M is the total number of ensemble members (200 for the SWOT -ANN model), I is the total number of observations, and s_k is the number of elements in the k^{th} bin of the rank histogram (Candille and Talagrand, 2005).

$$\Delta = \sum_{k=1}^{M+1} \left(s_k - \frac{I}{M+1}\right)^2 \tag{3}$$

$$\Delta_o = \frac{I * M}{M+1} \tag{4}$$

4.3.4 Continuous Ranked Probability Score

The continuous ranked probability score (CRPS) measures the area between the forecast cumulative distribution function (cdf) and the observed cdf for each forecast-observation pairing. For a given forecast-observation pair, the cdf of the forecast is calculated from the ensemble forecast pdf. Since each observation is a discrete value, its cdf is represented with the Heaviside function $H\{x \geq x_a\}$; a stepwise function which is 0 for all concentrations of point-of-consumption FRC below the observed FRC and 1 for all concentrations of household FRC above the observed concentration. The calculation of the CRPS is given in Equation 5 where F_i is the cdf of the forecast values for observation o_i and the x axis referenced is the concentrations of point-of-consumption FRC concentration. Note that Equation 5 shows the calculation of CRPS for a single forecast-observation pairing. To evaluate the ensemble models, the average CRPS, $\overline{CRPS}$, is calculated by taking the mean CRPS over all forecast-observation pairs.

$$CRPS = \int_{-\infty}^{\infty} (F_i(x) - H\{x \geq o_i\})^2 dx \tag{5}$$

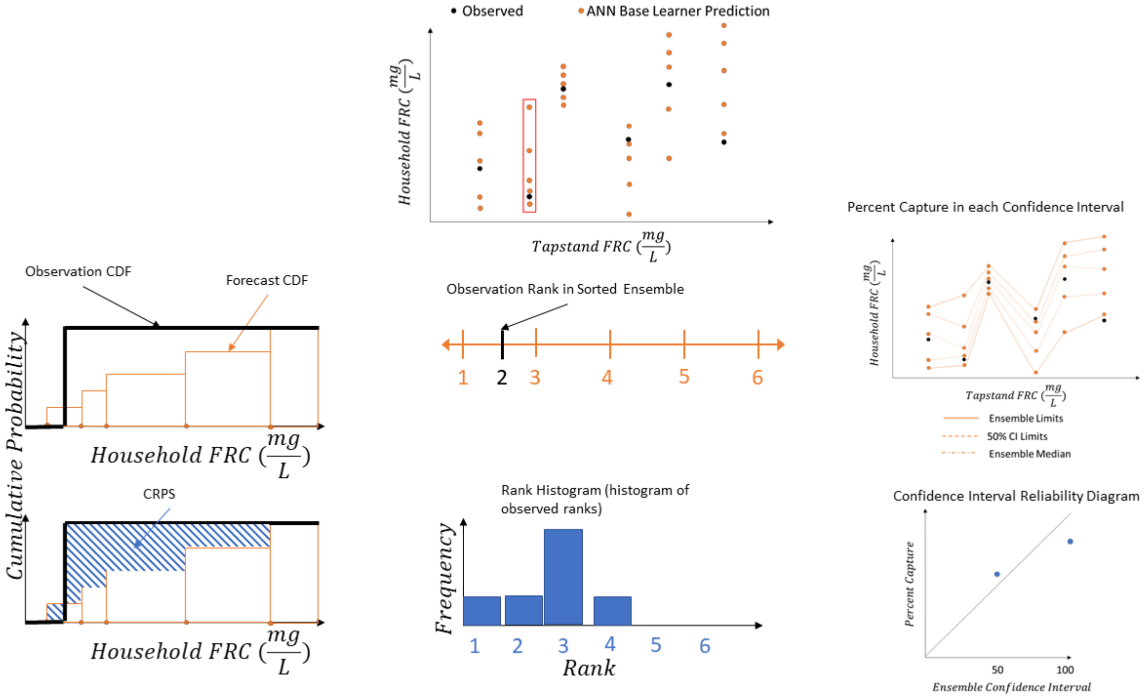
When using post-processed forecasts (Section 4.4) the SWOT-ANN calculates CRPS directly using Equation 5 and take the mean over all forecast-observation pairings. For the raw ensemble, the SWOT-ANN uses the Hersbach (2000) decomposition which treats the forecast cdf as a stepwise continuous function with $N = M + 1$ bins where each bin is bounded at two ensemble forecasts and the value in each bin is the cumulative probability. $\overline{CRPS}$ is calculated using $\overline{g_n}$, 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. Using these values, the $\overline{CRPS}$ for an ensemble can be calculated as:

$$\overline{CRPS} = \sum_{n=1}^{N} \overline{g_n}[(1-\overline{o_n})p_n^2 + \overline{o_n}(1-p_n)^2] \text{ (Hersbach, 2000)} \qquad (6)$$

Where p_n is the probability associated with each bin, $p_n = \frac{n}{N}$ (Hersbach, 2000).

4.3.5 Model Performance Summary and Outputs

Figure 3 below shows the interrelation between the ensemble verification metrics. This figure shows how the CRPS for each forecast-observation pair is calculated from the forecast and observation cdf while RH is obtained through a ranking of the observation within the members of the ensemble and finally how percent capture and CI reliability are both derived from the overall collection of forecasts and observations.



• Observed
• ANN Base Learner Prediction
Household FRC ($\frac{mg}{L}$)
Tapstand FRC ($\frac{mg}{L}$)
Observation CDF
Forecast CDF
Cumulative Probability
Household FRC ($\frac{mg}{L}$)
CRPS
Household FRC ($\frac{mg}{L}$)
Observation Rank in Sorted Ensemble
1 2 3 4 5 6
Rank Histogram (histogram of observed ranks)
Frequency
1 2 3 4 5 6
Rank
Percent Capture in each Confidence Interval
Household FRC ($\frac{mg}{L}$)
Tapstand FRC ($\frac{mg}{L}$)
Ensemble Limits
50% CI Limits
Ensemble Median
Confidence Interval Reliability Diagram
Percent Capture
50
100
Ensemble Confidence Interval

Figure 3: Interrelation of model performance metrics and visual intuition behind their derivation

Figure 4 presents the calibration diagnostic figures included in the SWOT-ANN output. This figure includes a plot of the predicted and forecast point-of-consumption FRC, the CI reliability diagram and the rank histogram. This figure can be a useful tool for understanding the reliability of the ensemble forecasts, and thus, the accuracy of the resulting targets. Ideally, Figure 4a would show all observations captured within the ensemble forecasts, and those forecasts would have a reasonable shape (the forecasts should follow the general trends in the underlying data), Figure 4b would have all points falling on the 1:1 line and Figure 4c would be a completely flat RH. For the example provided below, we see in subplot (a) that there are a number of observations that fall outside of the ensemble forecast range (shown with the error bars). Based on this, we can tell that the forecasts are underdispersed (the forecast spread is less than the spread of the observations). This is confirmed in Figure 4b where the points generally fall below the 1:1 line, and in Figure 4c, where the RH is U-shaped. Both of these are further indicators of underdispersion, and while this indicates a performance challenge for the SWOT-ANN, it is worth noting that underdispersion is common in ensembles, especially ensembles of neural networks. We overcome these challenges in the SWOT-ANN using both forecast post-processing (Section 4.4) and scenario analysis (Section 5.1)

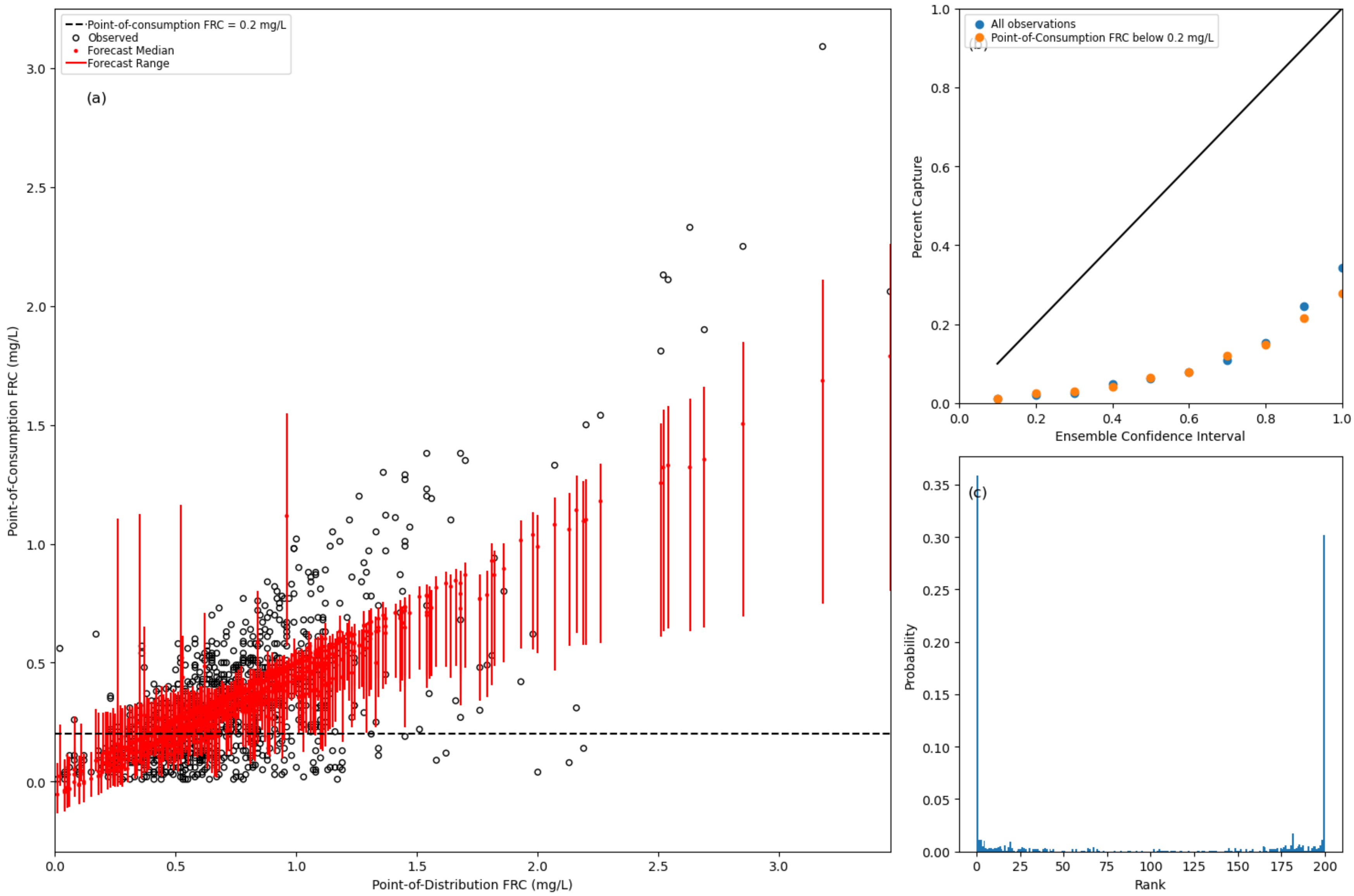


Figure 4: Sample performance evaluation figures, showing underdispersed forecasts, as can be identified via the numerous outlying observations, as well as the points in the CI reliability diagram and the u-shaped RH.

4.4 Post-Processing

In order to generate effective risk-based FRC targets using the SWOT-ANN, it is important that the forecast distribution matches the underlying distribution of the observed data. However, there are often dissimilarities between the forecast distribution and the distribution of the observed data. Ensemble post-processing is used to modify the ensemble forecasts to improve the similarity between the observed and forecast distributions. The SWOT-ANN uses kernel dressing to post-process the raw ensemble forecasts. This method follows a two-step process: first a kernel function is fit centred on each base learner prediction in the forecast for each observation, then each member's kernel is summed together to produce the post-processed pdf which is a non-parametric mixture distribution function. The SWOT-ANN uses a Gaussian kernel function in keeping with past studies (Boucher et al., 2015, 2011; Bröcker and Smith, 2008; Roulston and Smith, 2003), though the selection of the specific kernel function is not critical (Boucher et al., 2015). Kernel dressing is implemented in the SWOT-ANN using the Scipy Kernel Density Estimation (KDE) toolkit (Virtanen et al., 2020) with the kernel bandwidth defined using the Wang and Bishop (2005) method. This approach aims to minimize the difference between the variances of the ensemble forecasts and the observed data (Wang and Bishop, 2005). We selected the Wang and Bishop approach for use in the SWOT-ANN by comparing the ensemble forecasting performance of three different bandwidth determination methods which showed that, for post-distribution FRC data, the Wand and Bishop (2005) method performed best. The full comparison is included in Appendix C. The bandwidth for the kernels is calculated in the SWOT-ANN using Equation 7.

$$\sigma^2_{\kappa_{WB}} = \overline{(\overline{x_i} - y_i)^2} - \left(1 + \frac{1}{N}\right) * \overline{s^2_{x_i}} \qquad (7)$$

Where:

- $\sigma^2_{\kappa_{WB}}$ is the kernel bandwidth estimated using the Wang and Bishop (2005) method.
- $\overline{x_i}$ is the mean of the raw ensemble forecast of the i^{th} observation.
- $\overline{(\overline{x_i} - y_i)^2}$ is the mean error between the forecast mean of the i^{th} observation and the measured value of the i^{th} observations over all i observations.
- $\overline{s^2_{x_i}}$ is the mean variance of the ensemble forecasts.
- N is the number of observations.

The Wang and Bishop (2005) method of kernel bandwidth determination is specifically targeted for underdispersed forecasts where the spread of the predictions is less than the spread of the observations. This means that in some cases, the use of post-processing may not improve the overall ensemble performance and may in fact worsen the ensemble performance. To ensure that the best-performing option between the raw and post-processed ensemble is used, the SWOT-ANN tests the performance of the post-processed ensemble using the percent capture (PC, $PC_{<0.2}$), CI reliability score ($CI_{score}, CI_{score_{<0.2}}$), and $\overline{CRPS}$and compares these scores to those achieved by the raw ensemble (the δ score is not included in this comparison as the post-processed ensemble is a continuous distribution and as such does not have clearly defined "ranks" the way that the raw ensemble does). The SWOT-ANN performs this comparison using a skill score calculation which normalizes the change in performance from a reference baseline between negative infinity and 1 (Equation 8). For comparing the raw and post-processed ensembles, the raw ensemble score is used as the baseline score and the post-processed score is used as the score obtained. The ideal score is dependent on the metric, for percent capture the ideal score is 100%, for the CRPS and CI reliability scores, the ideal score is 0.

$$Skill\ Score = \frac{score\ obtained - baseline}{ideal\ score - baseline} \tag{8}$$

After calculating the skill score for all performance metrics, the SWOT-ANN selects the preferred forecasting approach (raw or post-processed) by taking the sum of the skill scores for each metric. If this sum is greater than zero, than post-processing yields and net performance improvement and the post-processing approach selected. If the sum of the skill scores is equal to or less than 0, the raw ensemble is used.

5 Obtaining a Tapstand FRC Target

The SWOT-ANN generates a risk-based FRC target by using the trained ANN ensemble to forecast the household FRC for tapstand FRC concentrations ranging from 0.2 to 2.0 mg/L in 0.05 mg/L increments. For each tapstand FRC concentration, the predicted risk of insufficient FRC is calculated as the probability of the household FRC being below 0.2 mg/L which is taken from the forecast cdf. If the target is being generated using the raw ensemble, this is obtained directly as the percentage of ensemble members that predicted that the household FRC below 0.2 mg/L. If using the post-processed ensemble, the predicted risk is obtained through s numerical integration of the post-processed pdf.

When generating the risk-based FRC targets, the tapstand FRC is incremented between 0.2 and 2.0 mg/L and all other input variables are held static. The elapsed time used is the user inputted target storage duration. To account for the effect of the remaining input variables (time of collection, EC, water temperature) on the FRC target, the SWOT-ANN implements a scenario analysis approach which considers different time of collection and water quality scenarios to produce a series of risk-based FRC targets.

5.1 Scenario Analysis

The SWOT-ANN uses scenario analysis to account for the influence of both the time of collection and tapstand water quality conditions when generating risk-based FRC targets. The time of collection scenario analysis generates two FRC targets: one for water collected from the tapstand before 12:00 noon ('AM Collection') and one for water collected after 12:00 noon (‘PM Collection’). Since the time-of-collection variable is always included in the input variable combination, the SWOT-ANN always produces a target for both scenarios. If no additional water quality variables (EC, water temperature) are included, then these are the only two scenarios considered by the SWOT-ANN analytics. If at least one of the two additional water quality variables is included in the model, the SWOT-ANN also generates FRC targets for two decay scenarios: an “average case” scenario which uses the median EC and/or water temperature values, and a “worst case” scenario which uses the 95th percentile EC and/or water temperature values. The selection of the 95th percentile as a “worst case” is based both on an empirical understanding of FRC decay behaviour: high water temperature will accelerate chemical reaction kinetics and thus increase the rate of FRC decay, and higher EC is indicative of dissolved inorganics which may indicate chlorine-consuming metals. This theoretical understanding is also supported by the findings of a proof-of-concept evaluation of risk-based FRC targets generated using ensembles of ANNs which showed that in most cases, EC and water temperature are at least moderately negatively correlated with household FRC and higher water temperature and EC values produced more conservative FRC targets (De Santi et al., 2021). The SWOT-ANN considers the decay scenarios in conjunction with the time of collection scenarios. Thus, if either EC or water temperature are included in the model, the SWOT-ANN produces FRC targets for four scenarios:

- “AM Collection” with “average case” decay conditions.
- “AM Collection” with “worst case” decay conditions.
- “PM Collection” with “average case” decay conditions.
- “PM Collection” with “worst case” decay conditions.

5.2 Outputs and Interpretation

When generating the risk-based FRC targets, the SWOT-ANN produces the following outputs:

1. Forecast plots showing the ensemble forecasts of household FRC for all scenarios (Figures 5, 6 and 7)
2. Histograms of the input variables used (Figure 8)
3. Plot of predicted risk against tapstand FRC (Figures 9 and 10), and output tables of the predicted risk of insufficient point-of-consumption FRC (Tables 1 and 2)

The following sections provide a detailed summary of how we recommend interpreting these outputs. A summary is provided in Section 5.2.4.

5.2.1 Interpreting Output 1: Forecast Plots

Figure 5 shows the forecasted household FRC generated by the SWOT-ANN for a site where both EC and water temperature were included in the dataset. The four subplots in figure 5 correspond to the four scenarios identified in Section 5.1. Figure 6 shows the FRC forecasts generated by the SWOT-ANN for the same site when the EC and water temperature measurements are removed from the dataset. These figures are primarily used as a visual diagnostic tool to verify that the model accurately reproduces the underlying trends in the observed data. These figures show the forecast median, forecast range, and the 95th percentile range of the forecast from the forecasts generated by predicting on fixed data as well as the observations used to train and validate the ANN models. When reviewing these plots there are three important factors to check. First, the forecast median should increase as the tapstand FRC concentration increases. Second, the shape of the forecast range and 95th percentile ranges should be acceptable, meaning that the upper and lower bounds of the forecast range and the 95th percentile range should all increase as the tapstand FRC increases. Furthermore, while some over or underprediction is expected, these bounds should be generally reasonable. Third, the forecast range should include most of the observations with household FRC below 0.2 mg/L (observations falling below the dashed line). This third check is often the most difficult to obtain due to forecast underdispersion.

When reviewing Figures 5 and 6 below, the median forecast household FRC increases with increasing tapstand FRC for all scenarios. Additionally, while in some cases the forecast extends below 0 or above the tapstand FRC concentration, there are no substantial outliers and the upper and lower bounds of both the forecast range and the 95th percentile range generally increase as the tapstand FRC increases, all of which indicate that the forecast shape is reasonable. Finally, while none of the forecasts capture all of the observations with household FRC below 0.2 mg/L, the worst-case forecasts in Figure 5 capture most of these observations. This provides a useful reference when selecting a tapstand FRC target from the risk predictions, as discussed in the sections below. It should be noted that neither of the scenarios in Figure 6 meet this final check , indicating that either a factor of safety would need to be applied, or the FRC target generated by the SWOT Engineering Optimization Tool (SWOT-EO) should be used.

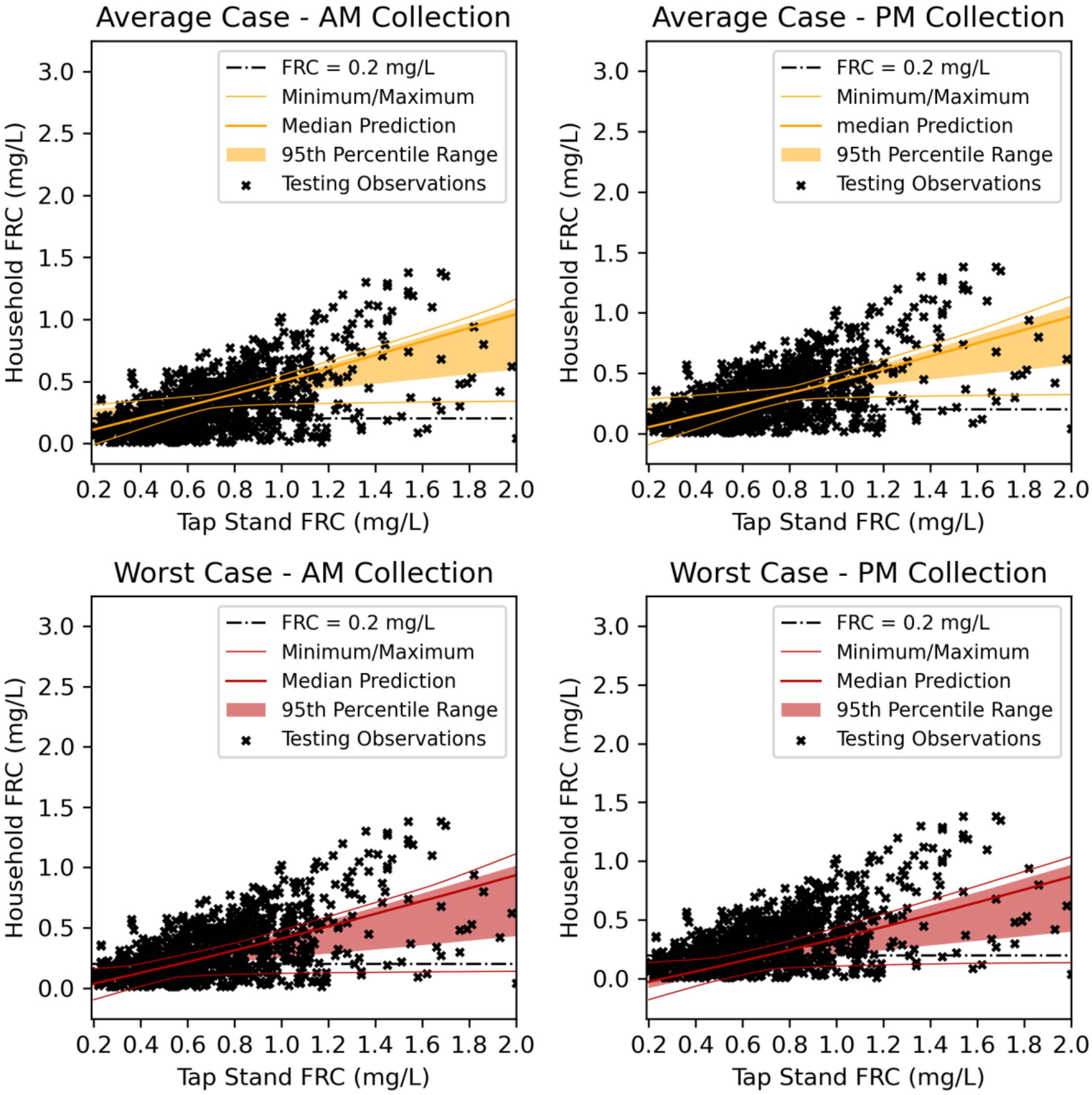


Figure 5: Sample SWOT-ANN output for dataset with additional water quality variables included (EC, water temperature)

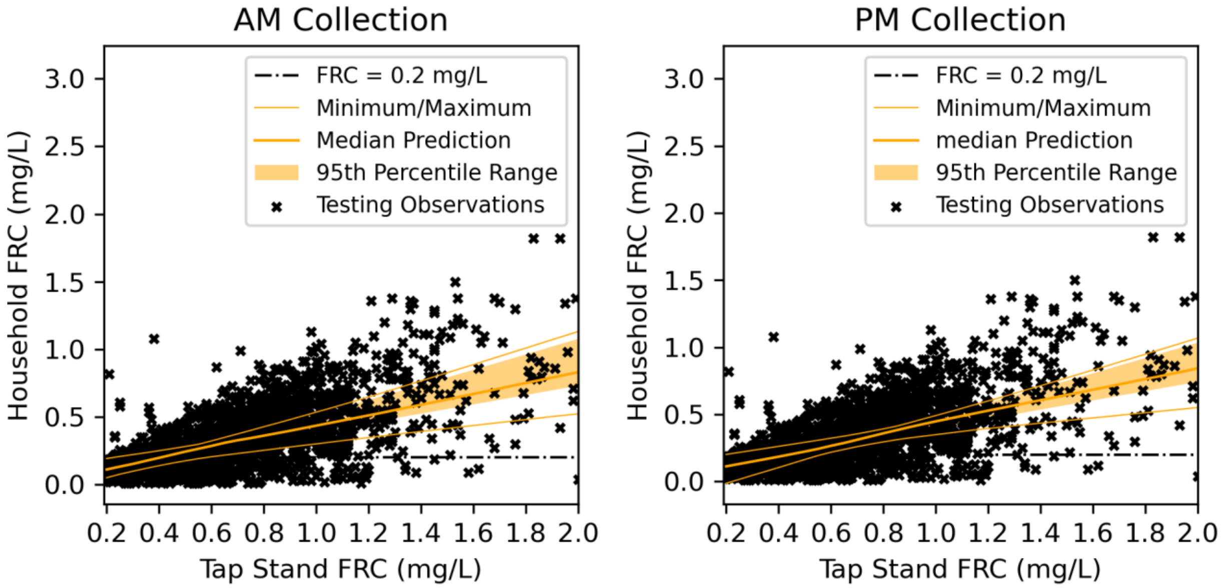


Figure 6: Sample SWOT-ANN output for dataset without additional water quality variables (EC, water temperature)

Figure 7 shows a special case when interpreting the household FRC forecast plots where there is either very sparse data or very limited data. In this figure, none of the forecasts effectively capture the observations with low household FRC. However, due to the sparse dataset, a useful FRC target can be visually identified from the graph by determining the tapstand FRC where there are no observations below the dashed line. In this case, we can see that a tapstand FRC of 0.6 mg/L would be sufficient to ensure sufficient protection at the household based on the data collected.

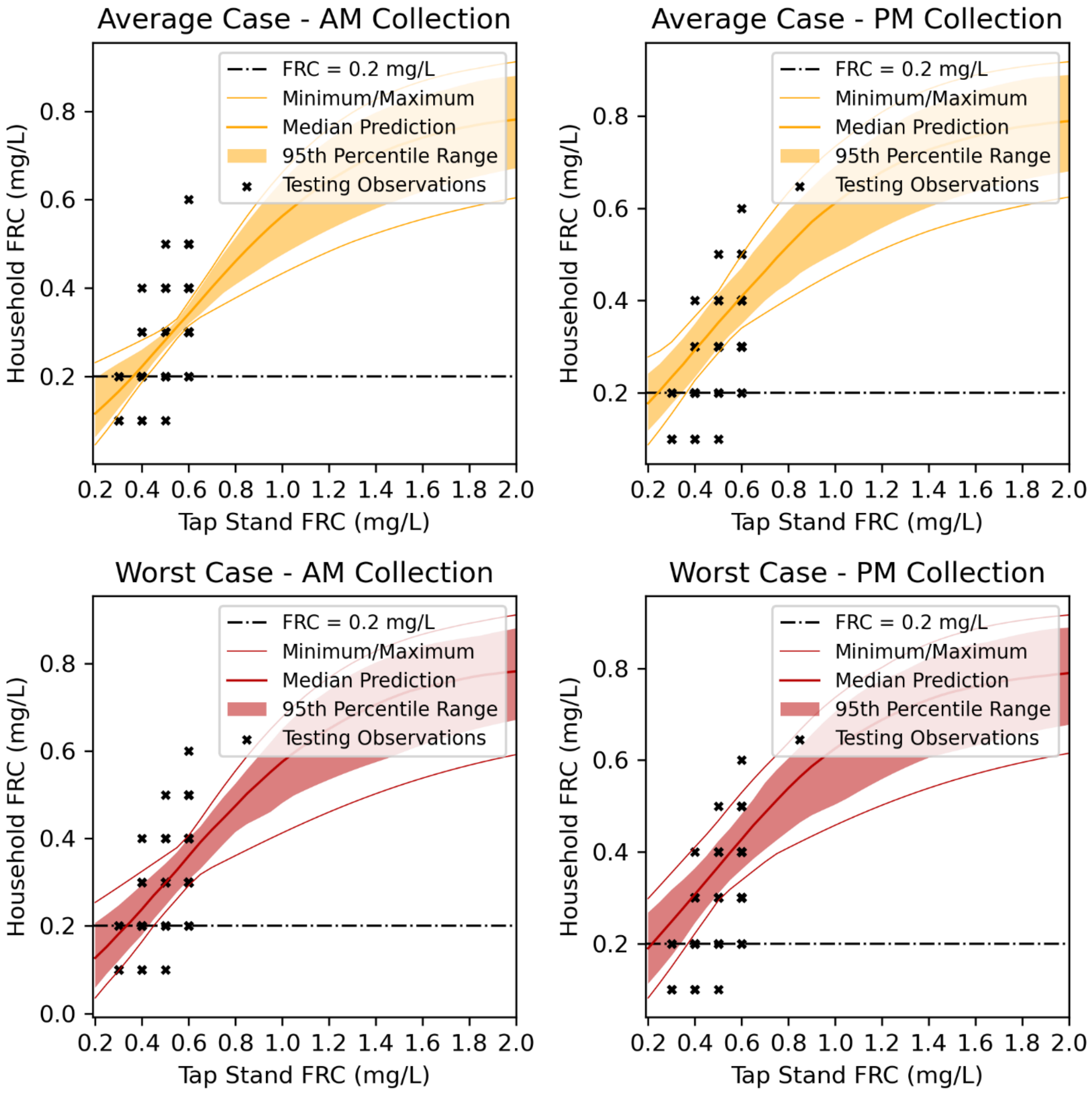


Figure 7: Predictions with sparse data. Note poor coverage of unsafe values but required tapstand FRC can be easily identified visually.

5.2.2 Interpreting Output 2: Input and output variable histograms

The input and output variable histograms are useful tools to understanding the water quality trends on a site. They are also useful for evaluating the parameters selected for the scenario analysis which can be helpful when selecting an FRC target. Specifically, while we always recommend selecting the most conservative tapstand FRC target, the histogram for time of collection can be used to determine if there is a dominant collection time on site (between AM and PM) which can be used to select between the AM and PM risk targets. Additionally, if EC or water temperature are included as input variables, the histograms provide an indication of the distribution of these variables as well as indicating which values were used for the average and worst-case scenarios.

Figure 8 shows the input variable histogram for the dataset used in Figure 2. From this figure we note several important factors. First, both the EC and water temperature observations appear to be from multi-modal distributions, meaning that in practice the EC and water temperature both tend to cluster around certain common values. When generating the scenario analysis, we see that for water temperature, both the average case and worst case values were drawn from the same cluster within the multi-modal distribution, with the median and 95th percentile values only separated by a few degrees Celsius. By contrast, the average and worst case EC values were drawn from different clusters within the multi-modal distribution, and with the worst case value nearly 1.5 times greater than the average case EC value. These are not necessarily good or bad, but these demonstrate interesting trends in the tapstand water quality at this stie. Finally, these histograms show that both times of collection (AM and PM) are well represented, but AM collection is more common.

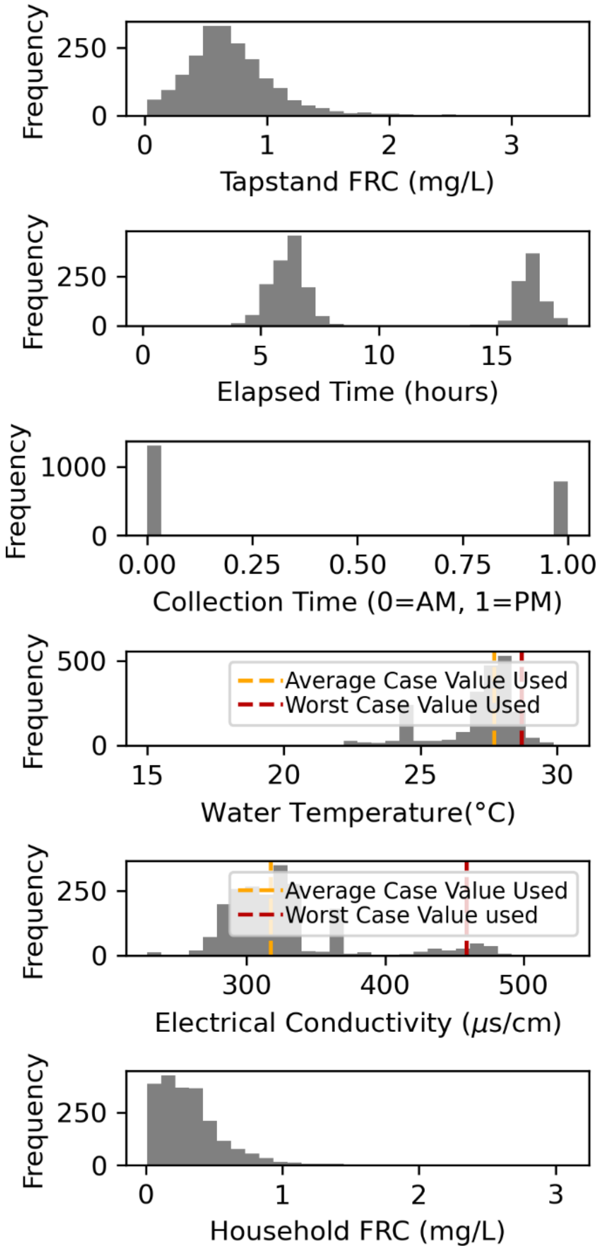


Figure 8: Input and output variable histograms.

5.2.3 Interpreting Output 3: Risk Predictions

After reviewing the predictions and the input and output variable histograms, the final outputs to review are the predicted risk figure and the associated tables. Figure 9 shows the predicted risk corresponding to the predictions shown in Figure 5. This figure shows the predicted risk for all scenarios on the same plot, allowing for a simple comparison of the predicted risk for all scenarios. When reviewing Figure 9, recall that both worst-case scenario models appeared to effectively capture most of the observations with insufficient household FRC. Based on this, we should obtain the tapstand FRC target from one of the two worst-case scenario lines. We recommend always selecting the most conservative FRC target, which in this case would correspond to PM collection. However, from Figure 8 we know that AM collection was more common, which could be a justification for using the AM collection risk targets.

From Figure 9, we can identify that the models predict little or no risk of household FRC below 0.2 g/L when the tapstand FRC is around 1 to 1.10 mg/L. For further accuracy, we can review the risk tables, Tables 1 and 2 below. Table 1 provides the average case targets table and Table 2 provides the worst-case risk predictions. Based on Figure 5 and 9, Table 2 should be used as, for this site, the worst case scenario targets were more conservative. The SWOT-ANN always provides both the average and worst case target tables, so it is important to identify the correct target table. From Table 2, we confirm the PM collection scenario is more conservative, as at each tapstand FRC, the predicted risk in the “PM Collection” column is greater than the predicted risk in the “AM Collection” column. From this table we also see that to obtain 0.000 predicted risk of insufficient household FRC, a tapstand FRC concentration of 1.10 mg/L is required. Note that this 0.0% predicted risk does not mean that there is no risk of having insufficient household FRC. There is always some risk of household FRC being below 0.2 mg/L, a predicted risk of 0 simply means that the model forecasts a very low probability (less that 0.001, or 0.1%) of household FRC below 0.2 mg/L. Also, this prediction should always be taken in context of the model performance (Section 4.3) as well as the actual predictions (shown in Figure 5) as a low predicted risk from a model with poor performance or that does not capture observations with low household FRC is not necessarily accurate.

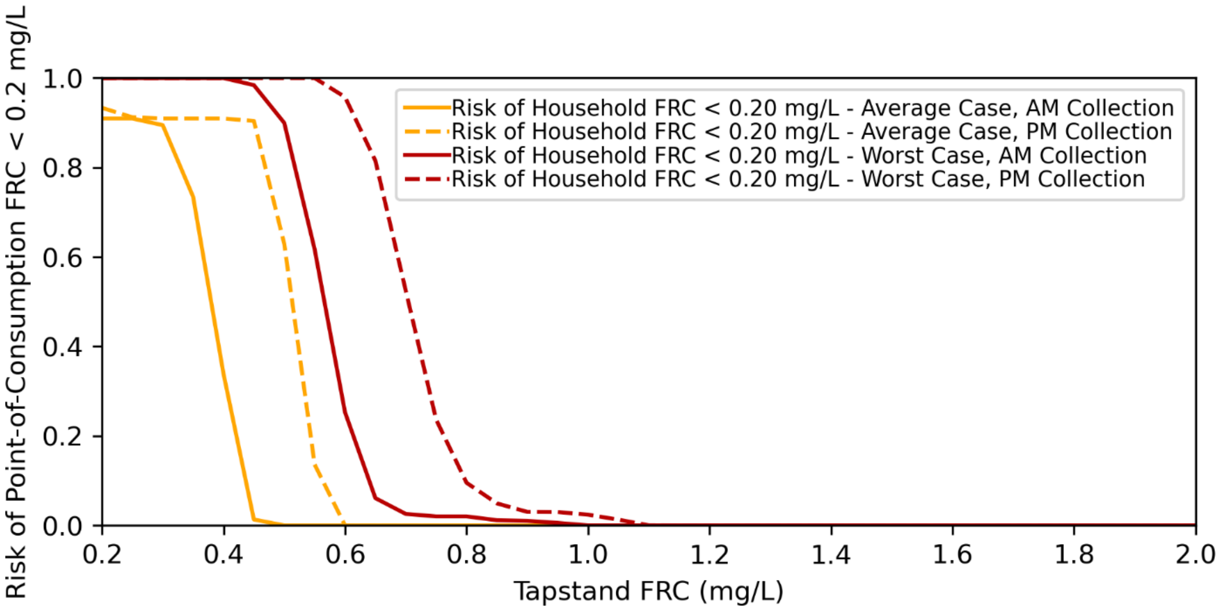


Figure 9: Risk predictions corresponding to the predictions from Figure 5

Table 1: Average case targets table

Input FRC (mg/L)	Storage Duration Target	Water Temperature (C)	Electrical Conductivity (10^-6s/cm)	Median Predicted Household FRC Concentration (mg/L) - AM Collection	Median Predicted Household FRC Concentration (mg/L) - PM Collection	Predicted Risk of Household FRC below 0.20 mg/L - AM Collection	Predicted Risk of Household FRC below 0.20 mg/L - PM Collection
0.20	15	28.9	472.0	0.112	0.055	0.910	0.933
0.25	15	28.9	472.0	0.136	0.077	0.910	0.912
0.30	15	28.9	472.0	0.160	0.1	0.895	0.910
0.35	15	28.9	472.0	0.184	0.124	0.735	0.910
0.40	15	28.9	472.0	0.207	0.147	0.339	0.910
0.45	15	28.9	472.0	0.231	0.17	0.013	0.904
0.50	15	28.9	472.0	0.255	0.193	0.000	0.629
0.55	15	28.9	472.0	0.279	0.216	0.000	0.136
0.60	15	28.9	472.0	0.303	0.241	0.000	0.000
0.65	15	28.9	472.0	0.327	0.265	0.000	0.000
0.70	15	28.9	472.0	0.351	0.289	0.000	0.000
0.75	15	28.9	472.0	0.375	0.313	0.000	0.000
0.80	15	28.9	472.0	0.399	0.336	0.000	0.000
0.85	15	28.9	472.0	0.424	0.359	0.000	0.000
0.90	15	28.9	472.0	0.449	0.382	0.000	0.000
0.95	15	28.9	472.0	0.475	0.407	0.000	0.000
1.00	15	28.9	472.0	0.501	0.432	0.000	0.000
1.05	15	28.9	472.0	0.526	0.458	0.000	0.000
1.10	15	28.9	472.0	0.552	0.483	0.000	0.000
1.15	15	28.9	472.0	0.578	0.51	0.000	0.000
1.20	15	28.9	472.0	0.604	0.536	0.000	0.000
1.25	15	28.9	472.0	0.631	0.562	0.000	0.000
1.30	15	28.9	472.0	0.658	0.589	0.000	0.000
1.35	15	28.9	472.0	0.685	0.615	0.000	0.000

Input FRC (mg/L)	Storage Duration Target	Water Temperature (C)	Electrical Conductivity (10^-6s/cm)	Median Predicted Household FRC Concentration (mg/L) - AM Collection	Median Predicted Household FRC Concentration (mg/L) - PM Collection	Predicted Risk of Household FRC below 0.20 mg/L - AM Collection	Predicted Risk of Household FRC below 0.20 mg/L - PM Collection
1.40	15	28.9	472.0	0.712	0.642	0.000	0.000
1.45	15	28.9	472.0	0.739	0.669	0.000	0.000
1.50	15	28.9	472.0	0.766	0.695	0.000	0.000
1.55	15	28.9	472.0	0.793	0.722	0.000	0.000
1.60	15	28.9	472.0	0.821	0.749	0.000	0.000
1.65	15	28.9	472.0	0.848	0.777	0.000	0.000
1.70	15	28.9	472.0	0.876	0.804	0.000	0.000
1.75	15	28.9	472.0	0.903	0.831	0.000	0.000
1.80	15	28.9	472.0	0.931	0.859	0.000	0.000
1.85	15	28.9	472.0	0.958	0.886	0.000	0.000
1.90	15	28.9	472.0	0.986	0.914	0.000	0.000
1.95	15	28.9	472.0	1.014	0.942	0.000	0.000
2.00	15	28.9	472.0	1.042	0.97	0.000	0.000

Table 2: Worst case targets table

Input FRC (mg/L)	Storage Duration Target	Water Temperature (C)	Electrical Conductivity (10^-6s/cm)	Median Predicted Household FRC Concentration (mg/L) - AM Collection	Median Predicted Household FRC Concentration (mg/L) - PM Collection	Predicted Risk of Household FRC below 0.20 mg/L - AM Collection	Predicted Risk of Household FRC below 0.20 mg/L - PM Collection
0.20	15	27.8	308.0	0.036	-0.026	1.000	1.000
0.25	15	27.8	308.0	0.058	-0.004	1.000	1.000
0.30	15	27.8	308.0	0.081	0.018	1.000	1.000
0.35	15	27.8	308.0	0.103	0.04	1.000	1.000
0.40	15	27.8	308.0	0.126	0.062	1.000	1.000
0.45	15	27.8	308.0	0.149	0.084	0.984	1.000
0.50	15	27.8	308.0	0.171	0.106	0.900	1.000
0.55	15	27.8	308.0	0.193	0.13	0.617	1.000
0.60	15	27.8	308.0	0.216	0.153	0.253	0.958
0.65	15	27.8	308.0	0.24	0.175	0.061	0.815
0.70	15	27.8	308.0	0.263	0.198	0.026	0.525
0.75	15	27.8	308.0	0.288	0.221	0.020	0.236
0.80	15	27.8	308.0	0.312	0.245	0.020	0.095
0.85	15	27.8	308.0	0.336	0.27	0.012	0.049
0.90	15	27.8	308.0	0.36	0.293	0.010	0.030
0.95	15	27.8	308.0	0.385	0.317	0.006	0.030
1.00	15	27.8	308.0	0.409	0.341	0.000	0.024
1.05	15	27.8	308.0	0.434	0.366	0.000	0.013
1.10	15	27.8	308.0	0.459	0.391	0.000	0.000
1.15	15	27.8	308.0	0.485	0.416	0.000	0.000
1.20	15	27.8	308.0	0.51	0.441	0.000	0.000
1.25	15	27.8	308.0	0.536	0.466	0.000	0.000
1.30	15	27.8	308.0	0.562	0.492	0.000	0.000
1.35	15	27.8	308.0	0.589	0.518	0.000	0.000

Input FRC (mg/L)	Storage Duration Target	Water Temperature (C)	Electrical Conductivity (10^-6s/cm)	Median Predicted Household FRC Concentration (mg/L) - AM Collection	Median Predicted Household FRC Concentration (mg/L) - PM Collection	Predicted Risk of Household FRC below 0.20 mg/L - AM Collection	Predicted Risk of Household FRC below 0.20 mg/L - PM Collection
1.40	15	27.8	308.0	0.615	0.544	0.000	0.000
1.45	15	27.8	308.0	0.642	0.57	0.000	0.000
1.50	15	27.8	308.0	0.668	0.596	0.000	0.000
1.55	15	27.8	308.0	0.694	0.623	0.000	0.000
1.60	15	27.8	308.0	0.721	0.649	0.000	0.000
1.65	15	27.8	308.0	0.747	0.676	0.000	0.000
1.70	15	27.8	308.0	0.773	0.704	0.000	0.000
1.75	15	27.8	308.0	0.8	0.731	0.000	0.000
1.80	15	27.8	308.0	0.827	0.758	0.000	0.000
1.85	15	27.8	308.0	0.854	0.786	0.000	0.000
1.90	15	27.8	308.0	0.881	0.813	0.000	0.000
1.95	15	27.8	308.0	0.909	0.841	0.000	0.000
2.00	15	27.8	308.0	0.936	0.868	0.000	0.000

It is also worth noting that the worst-case scenario is derived based on a theoretical understanding of the effect of EC and water temperature on FRC decay, and as such the worst-case scenario may not always be the most conservative. Figure 10 shows the predicted risk for a site where the average case scenario actually leads to more conservative targets than the worst case. Thus, it is always crucial to review the predicted risk plots before choosing a table from which to obtain an FRC target.

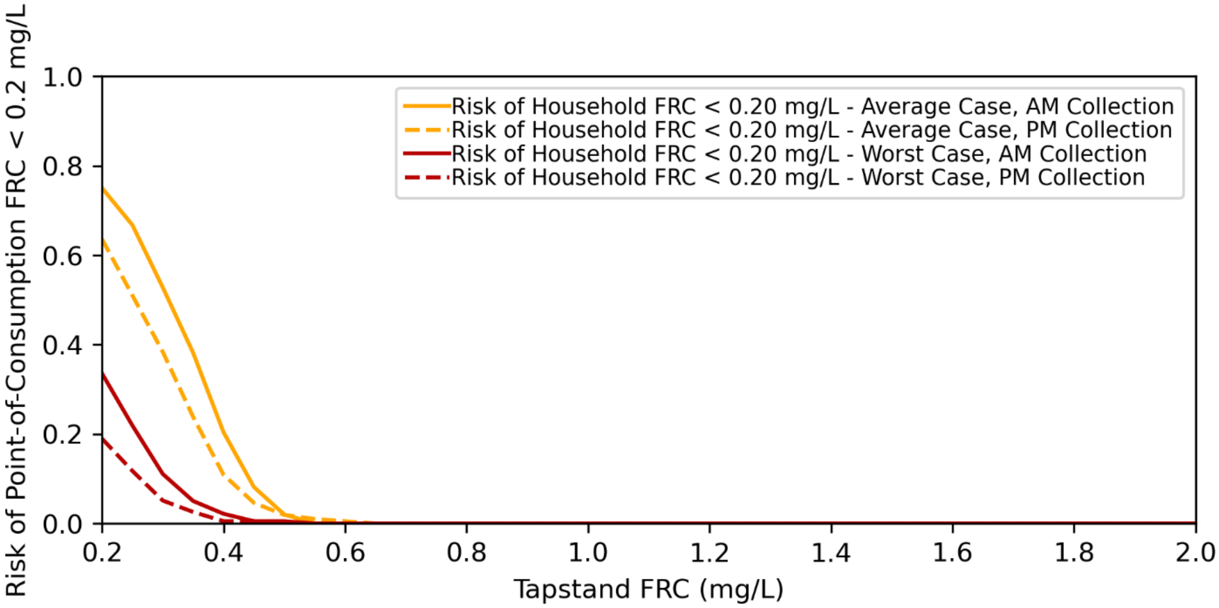


Figure 10: Predicted risk for a site where the worst-case scenario is not the most conservative.

5.2.4 Output Interpretation Summary

The SWOT-ANN produces 3 outputs to help obtain an FRC target:

1. Plot of predictions against point-of-distribution FRC for all scenarios
2. Histograms input and output variables used
3. Plot of predicted risk against tapstand FRC (Figures 9 and 10), and output tables of the predicted risk of insufficient household FRC (Tables 1 and 2)

The recommended order for reviewing these outputs is:

1. Review the plot of predictions against tapstand FRC for the following:
   - Forecast median increases with increasing tapstand FRC
   - Forecast range and forecast 95th percentile range both increase with increasing tapstand FRC, and shape is generally appropriate
   - Adequate coverage of observations with household FRC below 0.2 mg/L:
     - If this is not met, you may want to apply a factor of safety to the target, obtain a target through visual review of the data, or use the physical modelling target.

2. Review input and output variable histograms to identify the water quality variables used and to identify any large dominance of one collection period over the other.
3. Use the predicted risk figures to identify the appropriate table and column from which to obtain the FRC target:
    - This figure should at the very least identify the most conservative scenario (between average and worst case, if applicable).
    - To select between AM and PM collection, either use the most conservative target (recommended) or use the dominant scenario as identified from the histograms above (only recommended if one scenario is much more prevalent than the other.
    - Obtain the tapstand FRC target by reading off the lowest tapstand FRC concentration that produces a predicted risk of 0.000, or another value if you have a specific risk target.

6 Conclusion

This report summarizes the analytics included in version 2 of the SWOT-ANN, providing both the theoretical framework for the analytics as well as providing support for interpreting the SWOT-ANN outputs for the purposes of obtaining a risk-based FRC target. This version of the SWOT-ANN includes many features targeted to produce effective, evidence-based FRC targets, while addressing the complex challenges associated with modelling FRC during the post-distribution period. The SWOT-ANN code is available on the SWOT project GitHub page at: https://github.com/safeh2o/swot-python-analysis.

7 References

Ali, S.I., Ali, S.S., Fesselet, J.-F., 2015. Effectiveness of emergency water treatment practices in refugee camps in South Sudan. Bull. World Health Organ. 93, 550–558. https://doi.org/10.2471/BLT.14.147645

Boucher, M.A., Anctil, F., Perreault, L., Tremblay, D., 2011. A comparison between ensemble and deterministic hydrological forecasts in an operational context. Adv. Geosci. 29, 85–94. https://doi.org/10.5194/adgeo-29-85-2011

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

Bröcker, J., Smith, L.A., 2008. From ensemble forecasts to predictive distribution functions. Tellus, Ser. A Dyn. Meteorol. Oceanogr. 60 A, 663–678. https://doi.org/10.1111/j.1600-0870.2008.00333.x

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

Clark, R.M., Sivaganesan, M., 2002. Predicting chlorine residuals in drinking water: second order model. J. Water Resour. Plan. Manag. 128, 152–161. https://doi.org/10.1061/(ASCE)0733-9496(2002)128:2(152)

De Santi, M., Khan, U.T., Arnold, M., Fesselet, J.-F., Ali, S.I., 2021. Forecasting point-of-consumption chlorine residual in refugee settlements using ensembles of artificial neural networks. npj Clean Water Forthcomin.

Fisher, I., Kastl, G., Sathasivan, A., 2017. A comprehensive bulk chlorine decay model for simulating residuals in water distribution systems. Urban Water J. 14, 361–368. https://doi.org/10.1080/1573062X.2016.1148180

Gibbs, M.S., Morgan, N., Maier, H.R., Dandy, G.C., Nixon, J.B., Holmes, M., 2006. Investigation into the relationship between chlorine decay and water distribution parameters using data driven methods. Math. Comput. Model. 44, 485–498. https://doi.org/10.1016/j.mcm.2006.01.007

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

Hersbach, H., 2000. Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. Weather Forecast. 15, 559–570.

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

Powell, J.C., Hallam, N.B., West, J.R., Forster, C.F., Simms, J., 2000. Factors which control bulk chlorine decay rates. Water Res. 34, 117–126. https://doi.org/10.1016/S0043-1354(99)00097-4

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

Roulston, M.S., Smith, L.A., 2003. Combining dynamical and statistical ensembles. Tellus, Ser. A Dyn. Meteorol. Oceanogr. 55, 16–30. https://doi.org/10.1034/j.1600-0870.2003.201378.x

Talagrand, O., Vautard, R., Strauss, B., 1997. Evaluation of probabilistic prediction systems, in: Proceedings, ECMWF Workshop on Predictability. ECMWF, Shinfield Park, Reading, pp. 1–25.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, Ì., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors, 2020.

SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nat. Methods 17, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wang, X., Bishop, C.H., 2005. Improvement of ensemble reliability with a new dressing kernel. Q. J. R. Meteorol. Soc. 131, 965–986. https://doi.org/10.1256/qj.04.120

Warton, B., Heitz, A., Joll, C., Kagi, R., 2006. A new method for calculation of the chlorine demand of natural and treated waters. Water Res. 40, 2877–2884. https://doi.org/10.1016/j.watres.2006.05.020

World Health Organization, 2011. WHO Guidelines for Drinking-water quality, Fourth. ed. World Health Organization, Geneva, Switzerland.

Glossary of Functions and Explanations

The SWOT Version 2 code is available on GitHub at https://github.com/safeh2o/swot-python-analysis.. This Glossary provides an overview of the functions used in the *NNetwork.py* code (which is the main analytical code for the SWOT-ANN) and provides a brief summary of each function. Note, the code is built using object-oriented programming that builds the “NNetwork” class, and thus the “self” input is included in all of the functions in the code. Note the *NNetwork.py* code is called from the *run_swot_script.py* code which handles the running of the actual SWOT analytics.

Function	Inputs	Description	Outputs
import_data_from_csv	filename (input file name)	Called from “run_swot_script”, imports the data from the uploaded file and checks for all preprocessing rules. This function uses the tapstand and household timestamps to calculate the elapsed time of storage and time of collection. This function also checks the number of missing measurements to define the input variable selection process	Predictor and target DataFrames, count of dropped rows and other rule checks
valid_dates	series	Called from ”import_data_from_csv”. Data pre-processing step: removes observations with invalid dates (blank timestamp information, unknown formatting, dates that cannot be coverted to datetimes)	List of indices to remove from the input file
execute_rule	description, column, matches	Called from ”import_data_from_csv”, executes a data preprocessing rule	Removes observations from the input file based on a given rule
set_up_model		Called from “run_swot_script”, defines the architecture of the Keras model and compiles the model	Compiled Keras MLP model
train_SWOT_network	directory (directory to store saved models)	Called from “run_swot_script”, defines the training parameters for the overall SWOT neural network, saves the trained networks. Note, this function does not train the neural network but calls the “train_network” function	200 saved ensemble models

Function	Inputs	Description	Outputs
train_network	x, t, directory	Called from “train_SWOT_network”, trains the individual neural network	1 trained neural network
calibration_performance_evaluation	filename	Called from “train_SWOT_network”, calculates the performance of the raw ensemble. Performance metrics calculated: Percent Capture (overall and for observations with household FRC below 0.2), CI reliability score (overall and for observations with household FRC below 0.2), delta score (overall and for observations with household FRC below 0.2), CRPS, CRPS reliability term. Also produces diagnostic figures: plot of observations vs forecast range, CI reliability diagram, rank histogram	Performance metrics Performance Diagnostic Figures
post_process_cal		Called from “run_swot_script”, compares the performance of the raw and post-processed ensembles to determine the best-performing method. This is determined in the post_process_check variable which compares the sum of skill scores for the percent capture, percent capture of observations with household FRC below 0.2 mg/L, CI reliability, CI reliability for observations with household FRC below 0.2 mg/L, and the CRPS. If the sum of skill scores is positive, then post-processing is used, if negative, post-processing is not	post_process_check: if True, post-processing is used to generate targets, if False, post-processing is not used
get_bw		Called from “post_process_cal”, calculates the kernel bandwidth used for post-processing	Bandwidth

Function	Inputs	Description	Outputs
post_process_performance_cal	Bandwidth	Called from "post_process_cal", calculates the post-processed ensemble performance for the percent capture, percent capture of observations with household FRC below 0.2 mg/L, CI reliability, CI reliability for observations with household FRC below 0.2 mg/L, and the CRPS.	Post-processed ensemble performance metrics
set_inputs_for_table	storage_target	Called from "run_swot_script", uses the storage target provided to generate the tables used for forecasting the risk-based FRC targets. Note that four tables are produced for different possible scenario analyses	Tables of inputs for generating risk-based FRC targets
import_pretrained_model	directory	Called from "run_swot_script", loads the networks saved in the "train_SWOT_network" function	Loaded networks
predict		Called from "run_swot_script", generates ensemble forecasts of the household FRC for the inputs defined in the "set_inputs_for_table". Note, if post-processing is being used, the forecasts are also post-processed. This function also calculates the risk of having household FRC below 0.2 mg/L	Raw ensemble forecasts, post-processed ensemble forecasts, risk of low household FRC for each scenario
post_process_predictions	results_table_frc	Called from "predict" if the post_process_check is set to True. Post-processes the raw ensemble forecasts used to generate the risk-based FRC targets	Post-processed ensemble forecast
dsplay_results		Called from "run_swot_script", prints the results of the predict function	Printed results
export_results_to_csv	results_file	Called from "run_swot_script", exports the results for each scenario to a csv file	Saved .csv results file for each scenario

Function	Inputs	Description	Outputs
generate_html_report	report file, storage_target	Called from “run_swot_script”, compiles an html report of the key results: Prediction figures Risk figures Input and output variable histograms Risk tables Diagnostic figures Table of skipped rows	HTML report
prepare_table_for_html_report	storage_target	Called from “generate_html_report”, prepares a table of all of the inputs used to generate the risk-based FRC targets and the predicted risk for each scenario	Tables of predicted risk
results_visualization	filename, storage_target	Called from “generate_html_report” generates figures associated with risk-based FRC targets: Prediction figures Risk figures Input and output variable histograms	Prediction figures Risk figures Input and output variable histograms
skipped_rows_html		Called from “generate_html_report”, prepares a table of all rows that were dropped during data preprocessing	Table of skipped rows

Appendix A – Including Time in the ANN Model

A.1 Introduction

The SWOT ANN analytics uses ensembles of artificial neural networks (ANNs) to forecast point-of-consumption FRC in refugee and IDP settlements. The ANN base learners in the ensemble are a type of data driven model, meaning that they do not include any assumptions about the physical behaviour which they are modelling, and instead they learn from the underlying data. This is very useful for modelling FRC in the post-distribution period where the physical processes occurring are hard to quantify, however, this also means that ANN models do not always reflect the physical behaviour underlying the models. In particular, despite FRC decay being a time dependent reaction, elapsed time does not tend to be a strong predictor in ANN models. This was demonstrated in the development of the initial SWOT-ANN version 1 analytics, and was recently documented by De Santi et al. (2021) who showed that time was not a strong predictor of post-distribution FRC when using ensembles of ANNs, and confirmed this using partial correlation which showed that when controlling for the other variables included in the study, there was little correlation between elapsed time and the point-of-distribution FRC concentration.

The findings pose several problems. First, as mentioned above, FRC decay is time dependent and as such elapsed time should have some influence on the model, and the lack of influence of elapsed time may reduce confidence in the models. Second, having time as a predictor is important as it allows the SWOT to incorporate storage time as a variable, but these targets may be compromised if time is a weak predictor as we may end up with unconventional behaviour. Third, the elapsed time is already being collected, so discarding elapsed time as a variable means that we lose the value of the data being collected.

The limited usefulness of elapsed time in the ANN ensemble models may be due to clustering of elapsed time values around a few storage times, leading to confounding with behavioural and environmental factors (De Santi et al., 2021). Longer storage times have been hypothesized to reflect overnight storage when temperatures are cooler and where there is less opportunity for interaction with the water, whereas shorter storage times may reflect daytime storage when the ambient temperatures are higher and there is more potential for interaction with the water (De Santi et al., 2021). This Appendix presents an investigation into the interaction of time with potential confounding variables – specifically elapsed time and storage duration. We begin with an exploratory data analysis to visually identify trends between elapsed time, storage duration, and time of collection and continue with a comparison of model performance for these scenarios. The primary objectives of this study are to understand which variables are confounding the impact of storage time, and then select a modelling approach for the SWOT-ANN that incorporates time-related variables to produce the best performance.

A.2 Methods

A.2.1 Sites and data collection

The data used in this analysis included both the datasets collected from the sites used for the initial development of the SWOT (South Sudan, Jordan (2014), Jordan (2015), Rwanda), as well as data collected through SWOT field trials (Bangladesh, Tanzania, Nigeria).

A.2.2 Ethics

The initial field work in South Sudan received exemption from full ethics review by the Medical Director of Médecins sans Frontières (MSF) (Operational Centre Amsterdam) as data collected was routine for the on-going water supply intervention at the study site. For subsequent field studies in Jordan and Rwanda, ethics approval was obtained from the Committee for Protection of Human Subjects (CPHS) of the Institutional Review Board at the University of California, Berkeley (CPHS Protocol Number: 2014-05-6326). Informed consent was provided throughout all data collection.

The studies in Bangladesh, Tanzania, and Nigeria received approval from Human Participants Review Committee, Office of Research Ethics at York University (Certificate #: 2019-186), The study in Bangladesh also received approval from the MSF Ethical Review Board (ID #: 1932), and the Centre for Injury Prevention and Research Bangladesh (Memo #: CIPRB/Admin/2019/168).

A.2.3 Exploratory Data Analysis

This investigation began with an exploratory analysis to visualize the distribution of the time-based-variables on site (storage duration and hour of collection) for the seven datasets included in this analysis to identify any key patterns between these two explanatory variables that may aid in understanding potential confounding with elapsed time.

A.2.4 Modelling Approach

After the exploratory analysis, we trained and tested ANN ensembles with different approaches to incorporating time into the SWOT-ANN analytics and evaluated the resulting performance. To perform this evaluation, we began with a base model without any time-based-variables and then developed a series of experiments to incorporate time-based-variables.

A.2.4.1 Base Model Set-Up

The baseline models did not include any time-based-variables, though two input variable combinations were considered. The first (IV1) only included point-of-distribution FRC, and the second (IV2) included tapstand FRC, EC, and water temperature (the input variable combination used by the SWOT version 1). The base models used in this investigation were an ensemble of 200 multi-layer perceptrons (MLPs) with a single hidden layer as this is the same ensemble architecture used on the SWOT. The hidden layer used a hyperbolic tangent activation function, and the output layer used a linear activation function. The hidden layer size was selected based on the site and input variable combinations, with the hidden layer size for the IV1 models ranging from 4 to 16 hidden nodes and the hidden layer size for the IV2 models ranging from 8 to 16 hidden nodes. The overall dataset was split into three subsets. 25% of the overall dataset was used for testing, and the remaining 75% of the data was used for calibration. This calibration dataset was subdivided for each base learner with 33% of the calibration set (25% of the overall dataset) used for training and the remaining 66% of the calibration dataset (50% of the overall dataset) used for validating the training process. This validation set was used to trigger an early stopping procedure whereby if performance stopped improving on the validation set during training, training would be halted. This early stopping procedure is used to prevent overfitting of the models during training (i.e., it prevents each base learner becoming overly specific to the training data without being adequately generalizable).

A.2.4.2 Experiments

Nine experiments were proposed to incorporate time-based variables into the baseline model described above. Three time-based-variables were considered: elapsed time as a continuous variable, as used in De Santi et al. (2021); the storage duration as a binary variable representing long and short duration storage (with the cut-off being 12-hour storage); and the time of collection as a binary variable representing AM or PM collection. The two binary variables were intended to correspond to the clustering observed in De Santi et al. (2021) by addressing two potential confounding cases. The binary storage variable simplifies the model into long or short storage, which, as described above may account for differences between daytime and overnight storage. The time of collection variable addresses whether the storage period begins in the morning, thus including the period of daytime storage; or in the evening, representing less potential for storage during the hottest and most active times of day. Note that we took two approaches to using binary variables: we could either add them into the model as additional variables or we could split the model based on these binary variables to create two separate models. The nine approaches we considered are listed below:

1. Include elapsed time only as an input variable
2. Include the time of collection only as an input variable
3. Include the storage duration only as an input variable
4. Include elapsed time and time of collection as input variables
5. Include elapsed time and storage duration as input variables
6. Split the model based on the storage duration
7. Split the model based on the storage duration and add elapsed time as an input variable
8. Split the model based on time of collection
9. Split the model based on time of collection and add elapsed time as an input variable

A.2.5 Performance Metrics

The quality of the probabilistic forecasts was evaluated using the same performance metrics listed in Section 4.3 above:

- Percent capture (for the overall dataset and for observations where the household FRC is below 0.2 mg/L)
    - Described in Section 4.3.1: the percent capture describes the percentage of observations that fall within the forecast range and thus evaluates if the models are underdispersed
- CI reliability score (for the overall dataset and for observations where the household FRC is below 0.2 mg/L)
    - Described in Section 4.3.2: the CI reliability score measures the percentage of observations captured within each ensemble CI and compares them to the ideal percent capture, which would be a capture equal to the CI level. This evaluates the ensemble forecast reliability (i.e., the similarity between the observed and forecasted distributions)
- The rank histogram δ-score (for the overall dataset and for observations where the household FRC is below 0.2 mg/L)
    - Described in Section 4.3.3: the δ-score evaluates the uniformity or flatness of the rank histogram, providing an indication of forecast reliability as a flat rank histogram indicates that an observation is equally likely to appear anywhere within the ensemble forecast range (Candille and Talagrand, 2005; Hamill, 2001).
- The CRPS and CRPS reliability term
    - Described in Section 4.3.4: The CRPS is a probabilistic equivalent of mean absolute error (Ferro, 2014; Hersbach, 2000) which evaluates the forecast sharpness, reliability, and uncertainty. The CRPS reliability term is based on a decomposition of the CRPS for ensemble forecasts and directly evaluates the ensemble reliability (Hersbach, 2000).

A.2.5.1 Evaluation Methods

To compare different approaches for incorporating time-based-variables equally across all sites and input variable combinations, we normalized the score for each experiment by taking the Skill Score (Equation A-1). This converts each numerical score to a normalized score with the range of 1 to negative infinity, with a positive skill score indicating that performance has improved over a reference baseline score, and a negative score indicating a performance decrease. For percent capture and the δ-score, the ideal score is 1, and for the CI reliability score and CRPS and CRPS reliability term, the ideal score is 0.

$$Skill\ Score = \frac{score\ obtained - baseline}{ideal\ score - baseline} \qquad \text{(A-1)}$$

Since the skill score is a normalized indicator of improvement over a reference, we set the baseline score in this case to be the score obtained by the baseline model using no time-based-variables. Thus, each site and variable combination (IV1 vs IV2) has its own baseline score.

When comparing the skill scores across all sites and variable combinations, we used two short hand metrics to simplify the comparison. First, we took the sum of the skill scores for each site and variable combination to derive a Net Improvement Score for each site and variable combination which indicates if an experiment led to an overall improvement or decrease in the ensemble forecasting performance. We then determined the overall magnitude of improvement for a given experiment by taking the sum of all of the net improvement scores for that experiment. The overall magnitude of improvement for an experiment is effective at identifying if the experiment provides large performance improvements, but it may be dominated by a few good performances. Thus we balanced the magnitude of improvement against the consistency of improvement, which we calculated as the count, for each experiment, of all site and variable combinations with positive Net Improvement Scores. This consistency metric provides a useful indication if an experiment typically improves performance, without giving any indication as to whether or the improvements in performance are substantial.

A.3 Results and Analysis

A.3.1 Exploratory Data Analysis

Figure A-1 shows histograms of the storage duration for each site, disaggregated by the time of collection (morning or afternoon). This figure shows that each site tends to have a clear trend of longer or shorter storage based on the time of collection, though this trend may vary from site to site. For example, in South Sudan, afternoon collection primarily corresponds to shorter duration storage than morning collection, though in Bangladesh, the opposite appears to be true. Additionally, it is worth noting that both morning and afternoon collection were practiced at all sites.

Figure A-2 shows histograms of the time of collection for each site, disaggregated by the storage duration (shorter or longer than 12 hours). As with Figure A-1, we see some patterns emerging at each site relating the time of collection with the storage duration. In particular, early collection periods may correspond to either long or short storage durations, however, the later in the day water is collected, the more frequently that water is stored over 12 hours. This shows that the time-of-collection behaviour is clearly interrelated with the storage behaviour of the water. Unlike in Figure A-1 though, both long and short storage are not reflected at all sites, with only 3 observations in Nigeria having storage over 12 hours and no observations in South Sudan having storage over 12 hours. This is critical as these findings make experiments 6 and 7 non-viable because there is insufficient data to develop a separate long-duration storage model for these sites. Thus, these experiments were removed from consideration.

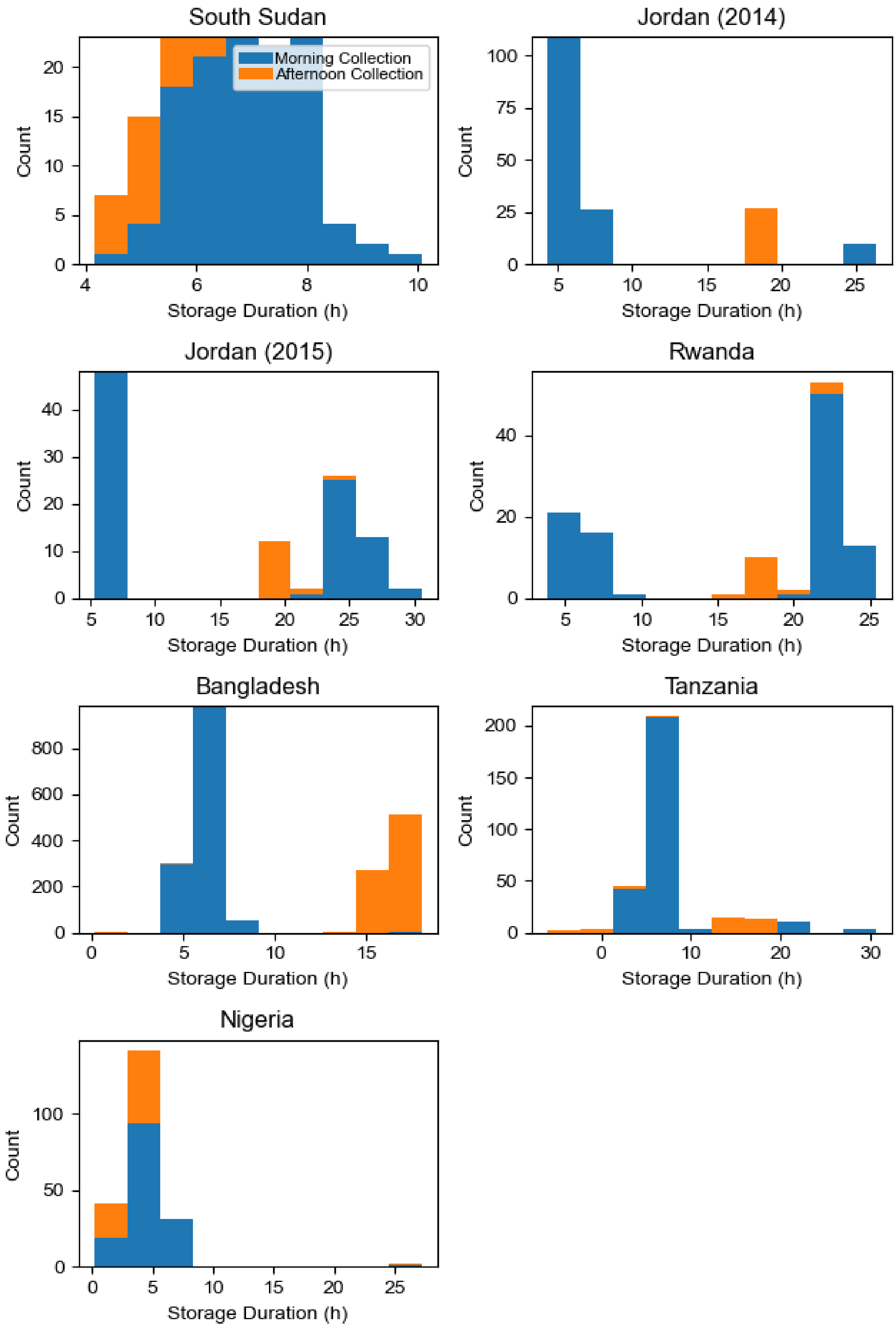


Figure A-1: Storage duration disaggregated by time of collection (morning vs afternoon)

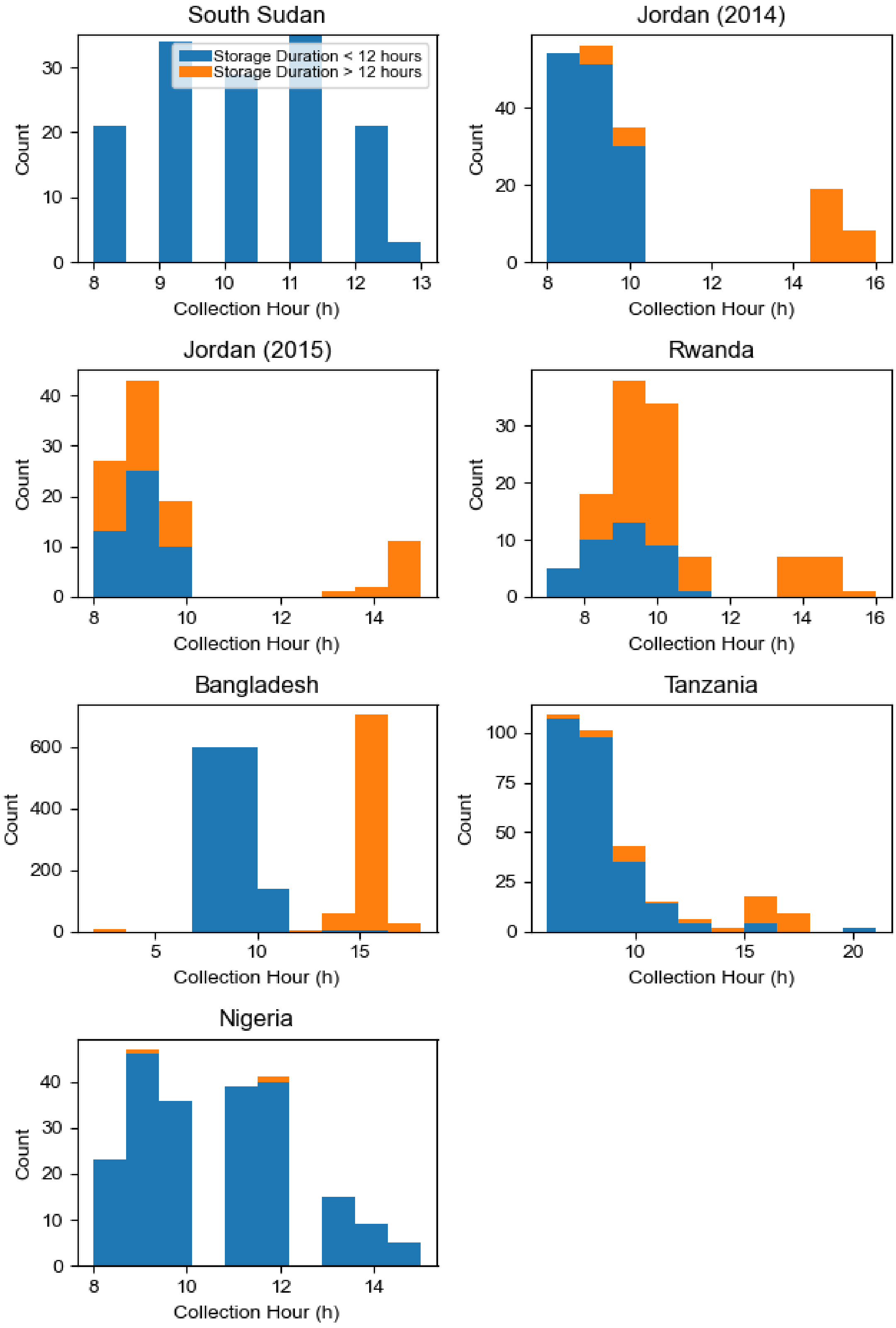


Figure A-2: time of collection disaggregated by storage duration (longer or shorter than 12 hours)

A.3.2 Ensemble Performance

Figure A-3 shows the Net Improvement Score for each experiment at each site and variable combination. The different coloured bars in Figure A-3 show the different site and variable combinations with the Net Improvement Scores grouped by experiment. From this figure we see that the inclusion of time-based-variables always resulted in a net decrease in performance for the South Sudan IV2 model, though this model has been shown to perform anomalously due to the combination of data from many subsites into one model (De Santi et al., 2021). However, this highlights another important trend in Figure A-3: the change in performance with the inclusion of time-based-variables is highly site specific, and the impact of each experiment is not consistent across all sites. However, when we consider the consistency of improvement (shown in Figure A-4) we see that Experiments 4, 5, and 9 all produce the most consistent improvements in model performance, each producing a net improvement in 13 out of a possible 14 cases. These experiments are unique in that they are all experiments that include both one of the categorical variables (storage duration or time of collection) as well as the continuous elapsed time variable. Experiments 4 and 5 are the experiments that directly include these as variables, and Experiment 9 splits the model based on collection time and includes elapsed time as a variable. These three experiments also produce the largest magnitude of improvement (shown in Figure A-4). The largest magnitude of performance improvement was observed in Experiment 4, with Experiment 9 producing the next largest magnitude of improvement.

From these results, we can clearly see that the elapsed time is an important predictor as all of the best models included elapsed time. Furthermore, additional variables to explain the confounding of elapsed time with behavioural and/or environmental parameters are required as the model with elapsed time alone did not perform as well as those models that included storage duration or time of collection. This indicates that the neural network model is able to find patterns relating elapsed time and household FRC within the storage duration or time of collection categories However, while including either storage duration or time of collection yielded improvement over elapsed time alone, the models using time of collection likely performed better because the time of collection provides a different type of information not included in the storage duration. Storage duration is derived from the elapsed time, thus there is some duplication of information between these variables whereas including time of collection allows the ANN models to find interactions between two variables which are more different from each other. Additionally, we found that when handling time of collection either as a binary variable or a model splitting criterion, we found that including it as a variable yielded more performance improvements, indicating that the model derives benefit from quantifying the interactions between these two variables.

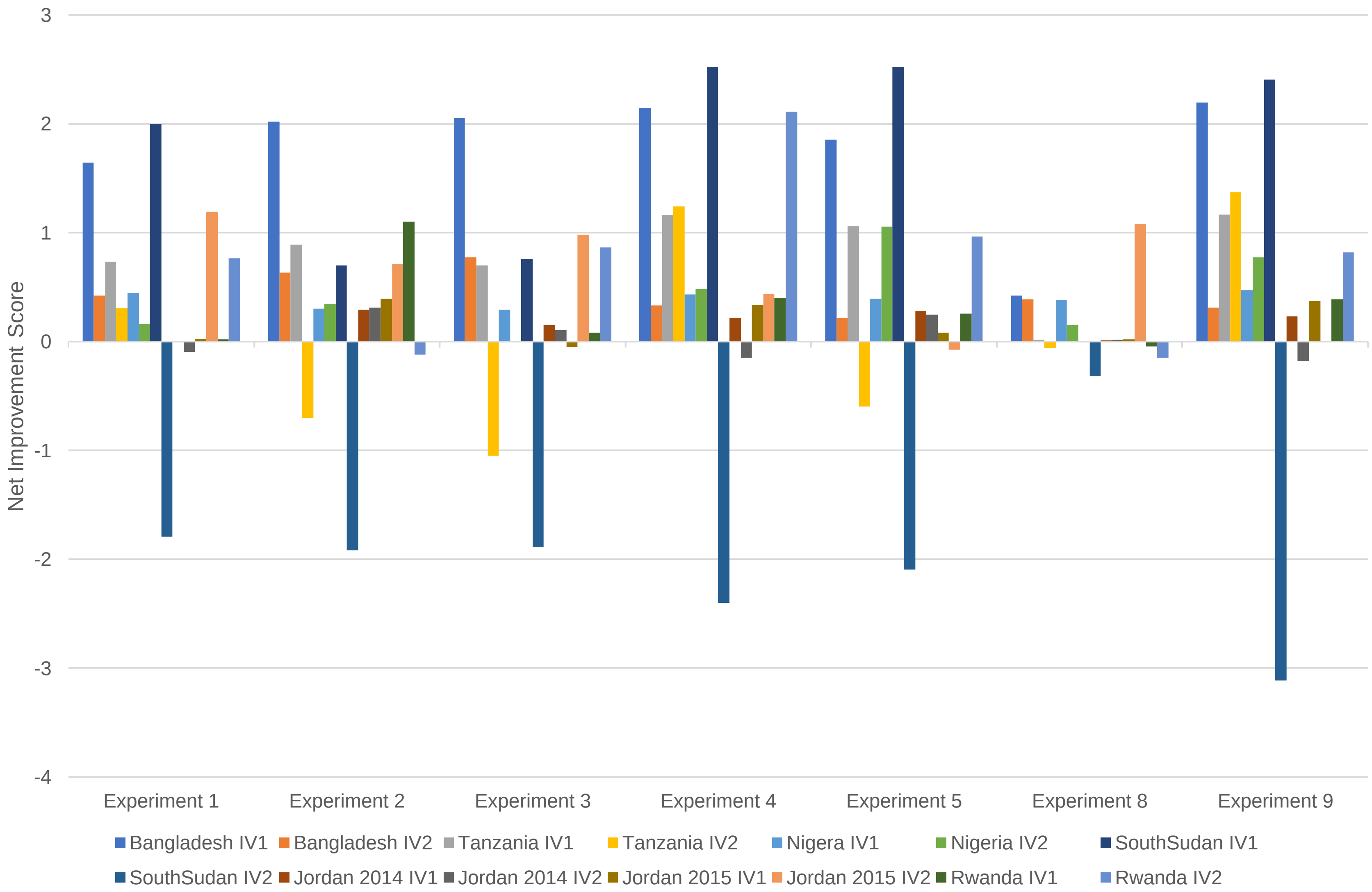


Figure A-3: Net Improvement Score for each site and variable combination, grouped by experiment

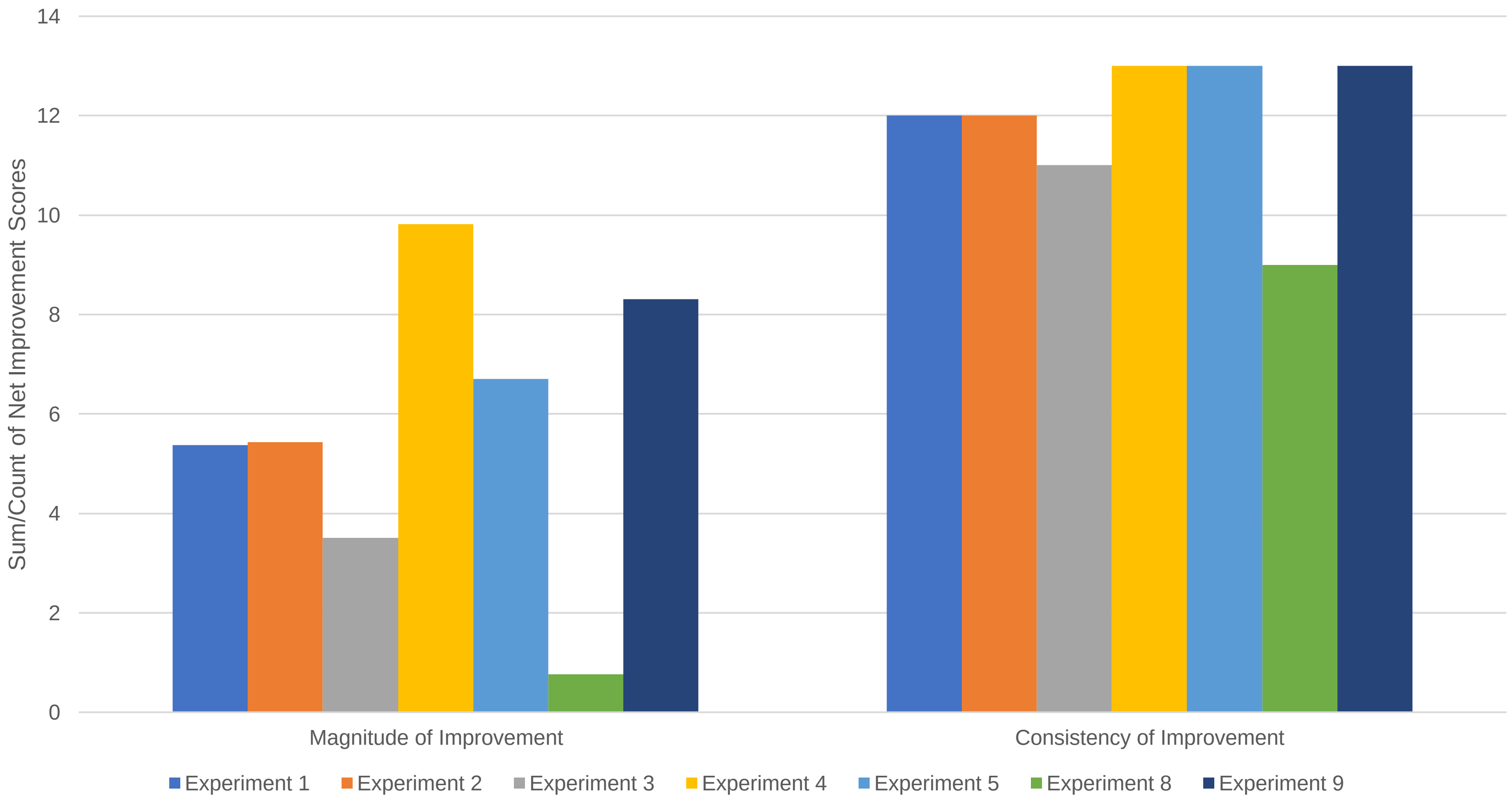


Figure A-4: magnitude and consistency of Net Improvement Scores for each experiment

A.4 Conclusion

Based on the findings presented above, we recommend that the SWOT-ANN analytics include both elapsed time and the time of collection as time-based-variables as this yields the best model performance, and both variables can be derived from the timestamps included in the tapstand and household measurements, and as such will not increase the data collection burden.

A.5 References

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

De Santi, M., Khan, U.T., Arnold, M., Fesselet, J.-F., Ali, S.I., 2021. Forecasting point-of-consumption chlorine residual in refugee settlements using ensembles of artificial neural networks. npj Clean Water Forthcomin.

Ferro, C.A.T., 2014. Fair scores for ensemble forecasts. Q. J. R. Meteorol. Soc. 140, 1917–1923. https://doi.org/10.1002/qj.2270

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

Hersbach, H., 2000. Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. Weather Forecast. 15, 559–570.

Appendix B – Input Variable Selection for the SWOT-ANN

B.1 Introduction

The SWOT ANN analytics uses ensembles of artificial neural networks (ANNs) to forecast point-of-consumption FRC in refugee and IDP settlements. One of the advantages of the ANN based approach is that, unlike process-based models, they can directly accept field water quality measurements of explanatory variables when modelling post-distribution FRC (Bowden et al., 2006; De Santi et al., 2021; Soyupak et al., 2011). However, in an operational context, it is not always possible to collect the additional water quality variables used in the SWOT-ANN analytics, specifically electrical conductivity (EC) and water temperature. This creates a challenge as the ANN base learners cannot accept missing data, so if a measurement is missing, the entire row must be discarded, creating a trade-off between the number of variables included in the model and the number of observations available to train the model. Furthermore, as described in Section 5.1, these additional water quality variables are used for scenario analysis, so if too many variables are removed, then the scenario analysis for different decay scenarios cannot be performed. The SWOT-ANN version 1 analytics filled missing water quality measurements using synthetic measurements, specifically the average conductivity and water temperature, however, this creates two challenges. First, in cases where a large amount of data is missing, there may be more synthetic measurements than real measurements. Second, these additional water quality variables have a strong impact on post-distribution FRC, and replacing them with the mean value leads to the model learning on wrong information and actually reduces the model performance. Thus, for the SWOT-ANN version 2 analytics we no longer replace missing values with synthetic data, but we must now identify the appropriate trade-off between removing observations with missing measurement and removing entire variables. This investigation compares the probabilistic performance of models trained with varying amounts of observations removed as well as varying input variable sets to directly evaluate this trade-off.

B.2 Methods

B.2.1 Sites and data collection

The data used in this analysis included both the datasets collected from the sites used for the initial development of the SWOT (South Sudan, Jordan (2014), Jordan (2015), Rwanda), as well as data collected through SWOT field trials (Bangladesh, Tanzania, Nigeria).

B.2.2 Ethics

The initial field work in South Sudan received exemption from full ethics review by the Medical Director of Médecins sans Frontières (MSF) (Operational Centre Amsterdam) as data collected was routine for the on-going water supply intervention at the study site. For subsequent field studies in Jordan and Rwanda, ethics approval was obtained from the Committee for Protection of Human Subjects (CPHS) of the Institutional Review Board at the University of California, Berkeley (CPHS Protocol Number: 2014-05-6326). Informed consent was provided throughout all data collection.

The studies in Bangladesh, Tanzania, and Nigeria received approval from Human Participants Review Committee, Office of Research Ethics at York University (Certificate #: 2019-186), The study in Bangladesh also received approval from the MSF Ethical Review Board (ID #: 1932), and the Centre for Injury Prevention and Research Bangladesh (Memo #: CIPRB/Admin/2019/168).

B.2.3 Analytical Approach

To analyze the trade-off between including larger input variable sets with fewer observations and smaller input variable sets with more observations we considered four possible input variable combinations. The first, IV1, includes only tapstand FRC, the elapsed time of storage, and the time of collection. This is the smallest feasible input variable set as all of these variables include only the two required measurements for the SWOT-ANN model: tapstand FRC and timestamp data. The second input variable combination (IV2) includes all the variables and IV1 and water temperature. For the datasets used in this study, water temperature tends to be more regularly collected than EC, and as such the IV2 input variable combination will typically have the second most observations. The third input variable combination (IV3) includes the IV1 variables and EC without water temperature. EC tends to be less regularly collected than water temperature, but past research suggests this may be a more informative predictor of household FRC than water temperature (De Santi et al., 2021). Finally, the fourth input variable combination (IV4) includes all potential variables: tapstand FRC, elapsed time, time of collection, water temperature, and EC. This final input variable set includes the most potential input variables, however, will likely have the fewest available observations. We used these four input variable approaches instead of a more systematic approach because each site had very different numbers of observations available for each input variable combination, giving a balanced representation of the possible outcomes.

The performance using each of these four input variables was recorded for each site, and then compared to the following potential explanatory factors:

- Number of observations dropped due to missing measurements for each input variable combination (using IV1 as a reference)
- Percentage of observations dropped due to missing measurements
- Standard deviation of the household FRC for each input variable combination
- Absolute change in the standard deviation as observations dropped due to missing measurements (using IV1 as a reference)
- Percent change in the standard deviation as observations dropped due to missing measurements

B.2.4 Base Model Set-up

The model architecture was kept similar to the approach taken in the previous appendix. The base models used in this investigation were an ensemble of 200 multi-layer perceptrons (MLPs) with a single hidden layer as this is the same ensemble architecture used on the SWOT. The hidden layer used a hyperbolic tangent activation function, and the output layer used a linear activation function. The hidden layer size was selected based on the site and input variable combinations, with the hidden layer size for the IV1 models ranging from 4 to 16 hidden nodes and the hidden layer size for the IV2 through IV4 models ranging from 8 to 16 hidden nodes. The overall dataset was split into three subsets. 25% of the overall dataset was used for testing, and the remaining 75% of the data was used for calibration. This calibration dataset was subdivided for each base learner with 33% of the calibration set (25% of the overall dataset) used for training and the remaining 66% of the calibration dataset (50% of the overall dataset) used for validating the training process. This validation set was used to trigger an early stopping procedure whereby if performance stopped improving on the validation set during training, training would be halted. This early stopping procedure is used to prevent overfitting of the models during training (i.e., it prevents each base learner becoming overly specific to the training data without being adequately generalizable).

B.2.5 Performance Metrics

The quality of the probabilistic forecasts was evaluated using the same performance metrics listed in Section 4.3 above:

- Percent capture (for the overall dataset and for observations where the household FRC is below 0.2 mg/L)
    - Described in Section 4.3.1: the percent capture describes the percentage of observations that fall within the forecast range and thus evaluates if the models are underdispersed
- CI reliability score (for the overall dataset and for observations where the household FRC is below 0.2 mg/L)
    - Described in Section 4.3.2: the CI reliability score measures the percentage of observations captured within each ensemble CI and compares them to the ideal percent capture, which would be a capture equal to the CI level. This evaluates the ensemble forecast reliability (i.e., the similarity between the observed and forecasted distributions)
- The rank histogram δ-score (for the overall dataset and for observations where the household FRC is below 0.2 mg/L)
    - Described in Section 4.3.3: the δ-score evaluates the uniformity or flatness of the rank histogram, providing an indication of forecast reliability as a flat rank histogram indicates that an observation is equally likely to appear anywhere within the ensemble forecast range (Candille and Talagrand, 2005; Hamill, 2001).
- The CRPS and CRPS reliability term

- Described in Section 4.3.4: The CRPS is a probabilistic equivalent of mean absolute error (Ferro, 2014; Hersbach, 2000) which evaluates the forecast sharpness, reliability, and uncertainty. The CRPS reliability term is based on a decomposition of the CRPS for ensemble forecasts and directly evaluates the ensemble reliability (Hersbach, 2000).

To compare the impact of dropping observations to add variables, we normalized the score for the IV2, IV3, and IV4 input variable combinations by taking the Skill Score (Equation B-1). This converts each numerical score to a normalized score with the range of 1 to negative infinity. A positive skill score indicating that performance has improved over a reference baseline score, and a negative score indicating a performance decrease. For percent capture and the δ-score, the ideal score is 1, and for the CI reliability score and CRPS and CRPS reliability term, the ideal score is 0.

$$Skill\ Score = \frac{score\ obtained - baseline}{ideal\ score - baseline} \qquad \text{(A-1)}$$

Since the skill score is a normalized indicator of improvement over a reference, we set the baseline score in this case to be the score obtained by the IV1 model and thus the skill score for each input variable combination indicates the improvement or deterioration in performance resulting from the addition of new input variables and the subsequent loss of observations with missing measurements. We took the sum of the skill scores for each metric for each site and variable combination to derive a Net Improvement Score which indicates the total improvement (or deterioration) of performance relative to the baseline for the IV2, IV3, and IV4 input variable combinations.

B.3 Results

Table B-1 summarizes, for each site and input variable combination, the total number of observations available, the standard deviation, and the net improvement score relative to the IV1 models. From this table we see that only at two sites were negative net improvement scores observed (Bangladesh and South Sudan). At all other sites, regardless of the number of observations removed, the net improvement scores were positive, indicating that removing observations to obtain a larger input variable set, in all but two cases, improved performance. It is also worth noting that in 4 out of 7 sites, the IV4 input variable set produced the best performance. In Bangladesh, the best performance was obtained by the IV1 input variable combination, in Jordan (2014) the best performance was obtained by the IV4 input variable combination, and in South Sudan the best performance was obtained by the IV2 input variable combination. Furthermore, even at Jordan (2014) the decrease in performance from IV3 to IV4 is not substantial.

Table 1: Summary of net improvement for each input variable combination for each site

Site	Input Variable Combination	Total Observations	Standard Deviation of HH FRC	Net Improvement Score
Bangladesh	IV1	2130	0.28	-
	IV2	1964	0.29	-0.056251484
	IV3	974	0.30	-0.655242458
	IV4	974	0.30	-0.498693426
Jordan (2014)	IV1	106	0.33	-
	IV2	106	0.33	0.105118514
	IV3	103	0.32	1.256280783
	IV4	103	0.32	0.919813512
Jordan (2015)	IV1	87	0.15	-
	IV2	87	0.15	0.775603178
	IV3	78	0.15	1.56719527
	IV4	78	0.15	1.813876966
Nigeria	IV1	216	0.11	-
	IV2	216	0.11	0.282824828
	IV3	216	0.11	0.10609025
	IV4	216	0.11	1.276432365
Rwanda	IV1	117	0.23	-
	IV2	94	0.19	1.046354642
	IV3	94	0.19	0.557674369
	IV4	94	0.19	1.027231832
South Sudan	IV1	143	0.37	-
	IV2	142	0.37	0.515658127
	IV3	127	0.36	-0.480462487
	IV4	126	0.36	-3.169370686
Tanzania	IV1	305	0.15	-
	IV2	89	0.20	0.903270956
	IV3	250	0.15	0.857070407
	IV4	89	0.20	1.909282998

To gain a better understanding of why the performance deteriorated with additional variables at some sites and improved at others, compared the Net Improvement Score to the number of observations removed, the percentage of observations removed, the change in standard deviation, and the percentage change in standard deviation. These comparisons are shown in Figure B-1. From this figure we do not observe a discernible trend in the change in performance with any of these factors. It is also worth noting that the sites that had the largest percentage of observations removed and the largest percentage changes in standard deviation of household FRC between input variable combinations (Tanzania) had one of the highest net improvement scores (1.9). In consideration of these finding it does not appear that there is a clear point where it substantially improves model performance to remove an input variable to gain more training observations. Thus, we recommend that whenever possible, the largest possible input variable combination be used, as the IV4 input variable combination was most commonly the best performing alternative.

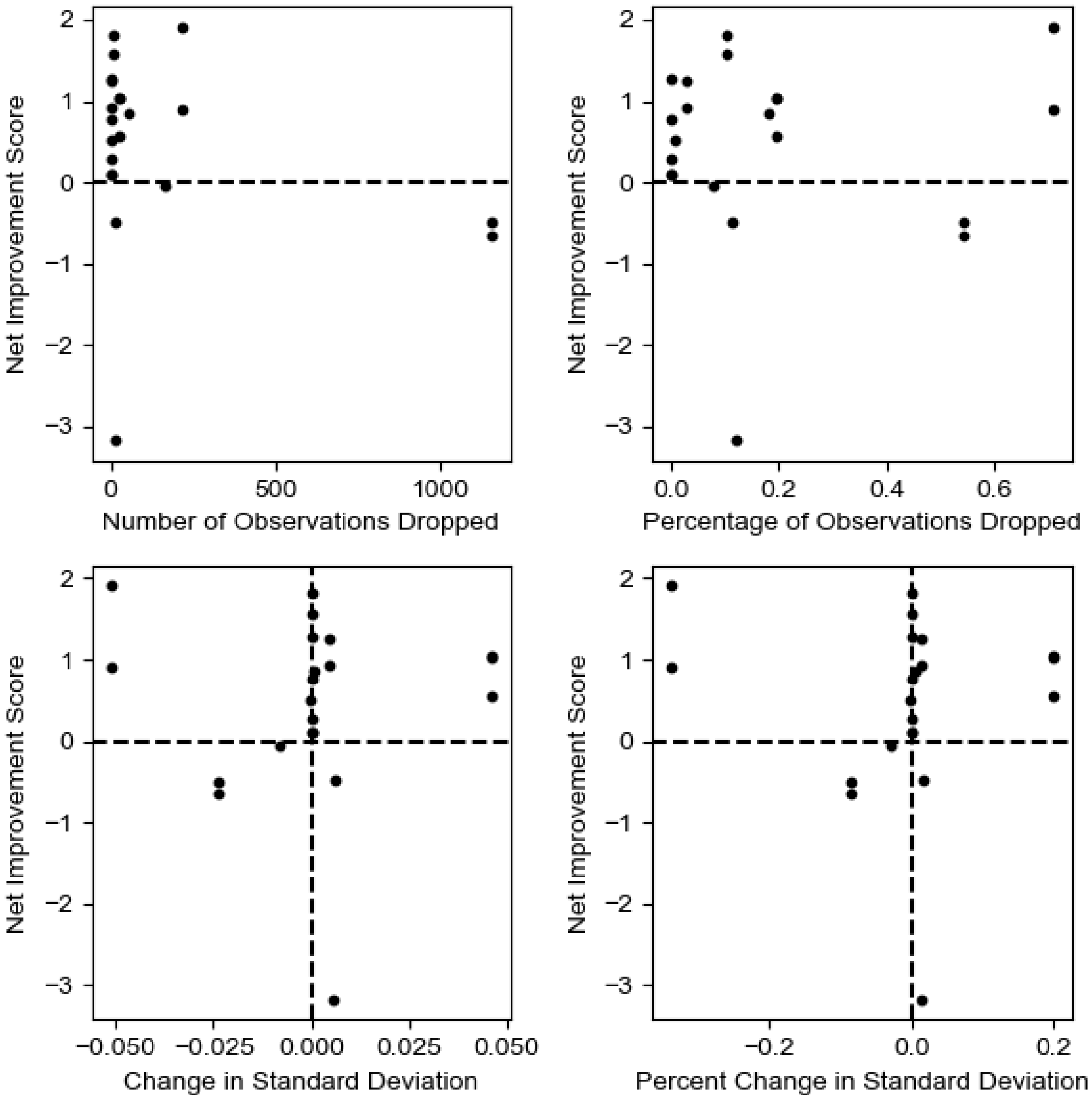


Figure 1: Comparison of Net Improvement Score to potential explanatory factors in the dataset

B.4 Conclusion

As stated above, there does not appear to be a clear point where removing an input variable to gain additional training observations improves ensemble forecasting performance, even when tested for a variety of datasets. This highlights the usefulness of additional water quality variables for explaining post-distribution FRC decay. Thus, we recommend that whenever possible the maximum possible number of input variables be used. For the SWOT-ANN version 2, this means that we recommend that if at least 10% observations have a measurement for a variable, that this variable be included. This 10% threshold was selected to allow for cases where there may be data entry issues, transition between data collection practices, or other anomalies where a very small number of samples have these measurements are included despite these variables not being included in routine monitoring

B.5 References

Bowden, G.J., Nixon, J.B., Dandy, G.C., Maier, H.R., Holmes, M., 2006. Forecasting chlorine residuals in a water distribution system using a general regression neural network. Math. Comput. Model. 44, 469–484. https://doi.org/10.1016/j.mcm.2006.01.006

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

De Santi, M., Khan, U.T., Arnold, M., Fesselet, J.-F., Ali, S.I., 2021. Forecasting point-of-consumption chlorine residual in refugee settlements using ensembles of artificial neural networks. npj Clean Water Forthcomin.

Ferro, C.A.T., 2014. Fair scores for ensemble forecasts. Q. J. R. Meteorol. Soc. 140, 1917–1923. https://doi.org/10.1002/qj.2270

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

Hersbach, H., 2000. Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. Weather Forecast. 15, 559–570.

Soyupak, S., Kilic, H., Karadirek, I.E., Muhammetoglu, H., 2011. 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 60, 51–60. https://doi.org/10.2166/aqua.2011.086

Appendix C – Comparison of Bandwidth Selection Methods for Post-Processing

C.1 Introduction

The SWOT ANN analytics uses ensembles of artificial neural networks (ANNs) to forecast point-of-consumption FRC in refugee and IDP settlements. These models account for the high degree of uncertainty in post-distribution FRC decay by generating probabilistic forecasts of household FRC which the SWOT uses to generated risk-based FRC targets. In order to produce accurate risk-based FRC targets, we need the ensemble forecasts to be reliable; that is, the probability distribution forecasted by the ensemble model should match the underlying distribution of the data. A challenge in achieving this using ensembles of ANNs is that these ensembles tend to be underdispersed, meaning that the spread of the predictions is less than the spread of the observations (De Santi et al., 2021). A common approach to overcoming ensemble underdispersion is to use post-processing methods (Boucher et al., 2015). These methods are applied to an ensemble forecast after the fact to improve the forecast reliability.

De Santi et al. (2021) proposed the use of kernel-dressing for post-processing ensemble forecasts of household FRC. Kernel dressing is a common approach to post-processing that ensemble forecasts where a kernel function (typically a Gaussian distribution) is fit around the prediction of each ensemble member. The ensemble forecast is then generated by taking the sum of each members kernel, which produces a non-parametric mixture distribution (Boucher et al., 2015). The benefits of kernel dressing for post-processing ensemble forecasts include the relatively low computational cost of kernel dressing, the simplicity of the method, and it's benefits specifically for improving underdispersed forecasts. However, a major challenge in implementing kernel based post-processing is selecting the kernel bandwidth. This is functionally the variance of the Gaussian distribution fit around each ensemble member. The selection of an appropriate bandwidth is key to generating reliable ensemble forecasts. The previous study by De Santi et al. (2021) which applied kernel post-processing for forecasting household FRC using ensembles of ANNs implemented the best member error developed by Roulston and Smith (2003). This is a common reference point for kernel post-processing, though in the De Santi et al. (2021) study, this approach improved the ensemble forecasts, but not enough to alleviate the underdispersion of the forecasts. This appendix provides an investigation into alternative bandwidth selection methods for post-processing ensemble forecasts. The objective of this investigation is to determine the bandwidth selection method that produces the best performance for ensembles of ANNs forecasting household FRC.

C.2 Methods

C.2.1 Sites and data collection

The data used in this analysis included both the datasets collected from the sites used for the initial development of the SWOT (South Sudan, Jordan (2014), Jordan (2015), Rwanda), as well as data collected through SWOT field trials (Bangladesh, Tanzania, Nigeria).

C.2.2 Ethics

The initial field work in South Sudan received exemption from full ethics review by the Medical Director of Médecins sans Frontières (MSF) (Operational Centre Amsterdam) as data collected was routine for the on-going water supply intervention at the study site. For subsequent field studies in Jordan and Rwanda, ethics approval was obtained from the Committee for Protection of Human Subjects (CPHS) of the Institutional Review Board at the University of California, Berkeley (CPHS Protocol Number: 2014-05-6326). Informed consent was provided throughout all data collection.

The studies in Bangladesh, Tanzania, and Nigeria received approval from Human Participants Review Committee, Office of Research Ethics at York University (Certificate #: 2019-186), The study in Bangladesh also received approval from the MSF Ethical Review Board (ID #: 1932), and the Centre for Injury Prevention and Research Bangladesh (Memo #: CIPRB/Admin/2019/168).

After the exploratory analysis, we trained and tested ANN ensembles with different approaches to incorporating time into the SWOT-ANN analytics and evaluated the resulting performance. To perform this evaluation, we began with a base model without any time-based-variables and then developed a series of experiments to incorporate time-based-variables.

C.2.3 Modelling Approach

C.2.3.1 Base Model Set-Up

Baseline models were developed for two input variable combinations. The first (IV1) included tapstand FRC, elapsed time, and the time of collection. The second input variable combination (IV2) included all of the same IV1 variables as well as electrical conductivity (EC) and water temperature. The base models used in this investigation were an ensemble of 200 multi-layer perceptrons (MLPs) with a single hidden layer as this is the same ensemble architecture used on the SWOT. The hidden layer used a hyperbolic tangent activation function, and the output layer used a linear activation function. The hidden layer size was selected based on the site and input variable combinations, with the hidden layer size for the IV1 models ranging from 4 to 16 hidden nodes and the hidden layer size for the IV2 models ranging from 8 to 16 hidden nodes. The overall dataset was split into three subsets. 25% of the overall dataset was used for testing, and the remaining 75% of the data was used for calibration. This calibration dataset was subdivided for each base learner with 33% of the calibration set (25% of the overall dataset) used for training and the remaining 66% of the calibration dataset (50% of the overall dataset) used for validating the training process. This validation set was used to trigger an early stopping procedure whereby if performance stopped improving on the validation set during training, training would be halted. This early stopping procedure is used to prevent overfitting of the models during training (i.e., it prevents each base learner becoming overly specific to the training data without being adequately generalizable).

C.2.3.2 Kernel Post Processing

As described above, the kernel dressing method of ensemble post-processing follows a two-step process: first a kernel function is fit centred on the base learner prediction for each observation, then each member’s kernel is summed together to produce the post-processed pdf which is a non-parametric mixture distribution function. We used a Gaussian kernel function in keeping with past studies(Boucher et al., 2015, 2011; Bröcker and Smith, 2008; De Santi et al., 2021; Roulston and Smith, 2003), though the selection of the specific kernel function is not critical (Boucher et al., 2015). The key to this process is the selection of an appropriate bandwidth. Following the example of Boucher et al. (2015), we considered three different bandwidths.

The first method we considered was the best member error approach developed by Roulston and Smith (2003). This approach uses the ensemble to generate a forecast for every observation in the calibration dataset. Then, for each observation, the best member (the member with the smallest error from the observation) is identified, and this member's error is taken as the best member error. The kernel bandwidth is then taken as the variance of all best member errors. This approach is both intuitive and simple to calculate, however, past studies have shown that it is not effective for reproducing the spread of the observed data (Wang and Bishop, 2005). The bandwidth for the Wang and Bishop method is calculated using Equation C-1.

$$\sigma_{\kappa_{WB}}^2 = \overline{(\overline{x_i} - y_i)^2} - \left(1 + \frac{1}{N}\right) * \overline{s_{x_i}^2} \tag{C-1}$$

Where:

- $\sigma_{\kappa_{WB}}^2$ is the kernel bandwidth estimated using the Wang and Bishop (2005) method.
- $\overline{x_i}$ is the mean of the raw ensemble forecast of the i^{th} observation of the calibration dataset.
- $\overline{(\overline{x_i} - y_i)^2}$ is the mean error between the forecast mean of the i^{th} observation in the calibration dataset and the measured value of the i^{th} observations over all i observations.
- $\overline{s_{x_i}^2}$ is the mean variance of the ensemble forecasts.
- N is the number of observations in the calibration dataset.

The third method considered in this investigation is the method developed by Fortin et al. (2006). This is method is also derived from the Roulston and Smith (2003) method. In this method, after forecasting on the calibration dataset, each ensemble forecast is sorted by prediction from low to high and the rank of the best member is determined as well as the best member error. After this is repeated for each calibration observation, a unique bandwidth is selected for each ensemble rank based on the variance of the best member errors for every time the best member error was in that rank. Furthermore, when summing the kernels to form the ensemble forecast, the sum is weighted by the probability of each rank having a best member (Fortin et al., 2006).

C.2.4 Performance Metrics

The quality of the probabilistic forecasts was evaluated using the same performance metrics listed in Section 4.3 above. Note that unlike the previous two appendices, the rank histogram δ-score and the CRPS reliability term are not included as the calculation of those metrics requires discrete ensemble member predictions and not a continuous forecast (which is obtained from the summation of the kernels).

- Percent capture (for the overall dataset and for observations where the household FRC is below 0.2 mg/L)
  - Described in Section 4.3.1: the percent capture describes the percentage of observations that fall within the forecast range and thus evaluates if the models are underdispersed
- CI reliability score (for the overall dataset and for observations where the household FRC is below 0.2 mg/L)

    - Described in Section 4.3.2: the CI reliability score measures the percentage of observations captured within each ensemble CI and compares them to the ideal percent capture, which would be a capture equal to the CI level. This evaluates the ensemble forecast reliability (i.e., the similarity between the observed and forecasted distributions)
- The CRPS
    - Described in Section 4.3.4: The CRPS is a probabilistic equivalent of mean absolute error (Ferro, 2014; Hersbach, 2000) which evaluates the forecast sharpness, reliability, and uncertainty.

The scores listed above were normalized by taking the Skill Score (Equation C-2). This converts each numerical score to a normalized score with the range of 1 to negative infinity. A positive skill score indicating that performance has improved over a reference baseline score, and a negative score indicating a performance decrease. For percent capture, the ideal score is 1, and for the CI reliability score and CRPS, the ideal score is 0. The baseline score was taken as the raw ensemble performance, and as such a positive skill score indicates that the post-processing improved the score, and a negative score indicates that the post-processing made the performance worse. To simplify the comparison of different methods across a large number of sites and variable combinations, we took the sum of the skill scores for all five performance metrics for each site and variable combination to combine into an overall Net Improvement Score, which indicates the total performance improvement or deterioration at a site. A positive net improvement score indicates a performance increase whereas a negative Net Improvement Score indicates that overall, the post-processing method made the performance worse.

$$Skill\ Score = \frac{score\ obtained - baseline}{ideal\ score - baseline} \qquad \text{(C-2)}$$

C.3 Results

Table C-1 shows the net scores for each post-processing method for each site and variable combination. Most notably, this table shows substantial deterioration of performance for at almost all sites when using the Fortin et al. (2006) method. This is surprising, as this method was found to outperform the other two methods listed by Boucher et al. (2015). When reviewing the individual performance metrics for these sites though, it is worth noting that the Fortin et al. (2006) method actually substantially improved the percent capture and CI reliability at most sites, but it also substantially increased the CRPS. This indicates that for our application, the Fortin et al. (2006) method substantially improved the dispersion and reliability of the ensemble forecasts, but at the expense of sharpness. We demonstrate an example of this in Figure C-1 that shows the predictions and observations for the raw ensemble and each post-processed ensemble for the Bangladesh IV1 model. From this figure we see that the Fortin method greatly increases the spread of the ensemble forecast, leading to much better capture, but it also creates substantial overdispersion, leading to substantial overdispersion. Another challenge, not capture in Figure C-1, is that often the best member rank was the same for many observations, or in some cases, there were too few observations for each ensemble rank to be represented, meaning that the Fortin et al. (2006) method does not use the predictions of each ensemble member (as the bandwidth cannot be calculated for a rank with no best members). Thus, the Fortin et al. (2006) method, while having been demonstrated to be highly effective in past studies, is not well suited to the SWOT-ANN where there are often more ensemble members than testing observations.

Table C-1: Comparison of Net Improvement Scores for the three post-processing methods

Site	Input Variable Combination	Best Member Error (Roulston and Smith, 2003)	Wang and Bishop (2005) method	Fortin et al. (2006) method
South Sudan	IV1	0.57	1.52	-4.28
	IV2	0.73	1.45	-3.83
Jordan (2014)	IV1	0.52	1.02	1.21
	IV2	0.46	1.31	-5.29
Jordan (2015)	IV1	-0.87	0.32	-2.37
	IV2	0.05	0.56	-23.55
Rwanda	IV1	-0.01	0.51	-11.34
	IV2	0.01	0.92	-18.53
Bangladesh	IV1	0.48	0.82	-306.25
	IV2	0.59	0.95	-1.28
Tanzania	IV1	-1.93	0.22	-11.07
	IV2	-0.58	0.29	-25.50
Nigeria	IV1	-1.60	0.51	-65.10
	IV2	-1.75	-0.01	-7.13

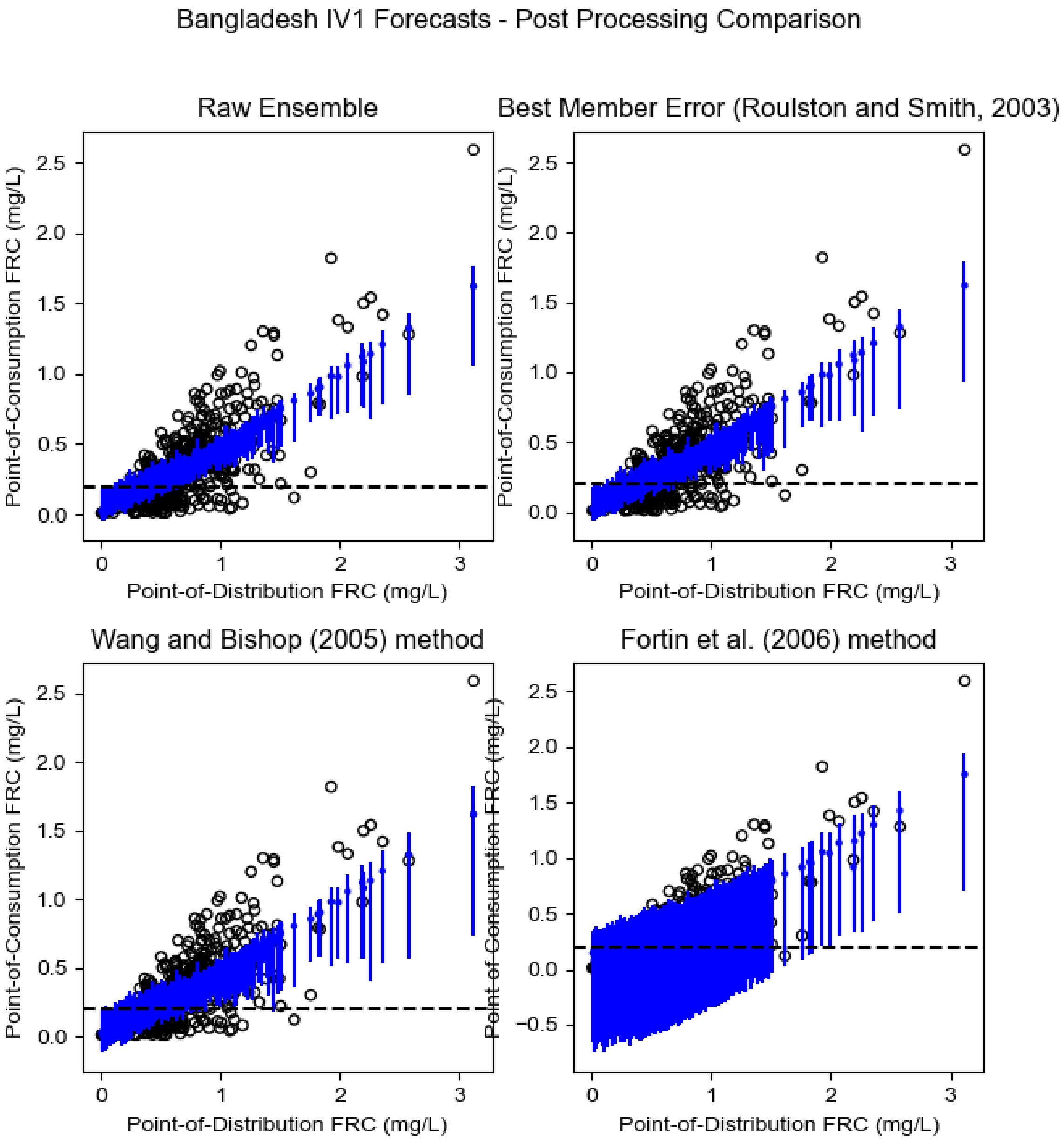


Figure C-1: Comparison of post-processing methods for the Bangladesh IV1

Table C-1 also shows that between the best member error method and the Wang and Bishop (2005) method, the Wang and Bishop (2005) method provides the most consistently positive Net Improvement Score, as well as providing the largest magnitude of improvement. Based on this, we recommend the Wang and Bishop (2005) method for inclusion in the SWOT-ANN version 2 analytics.

C.4 Conclusion

From the above results it is clear that the Wang and Bishop method produces the best performance for the post-processing ensemble forecasts of post-distribution FRC. Thus, we recommend this method for inclusion in the SWOT-ANN version 2 analytics. However, we also note that the Wang and Bishop method does not always lead to improved performance, and as such, the SWOT version 2 analytics should always compare the raw and post-processed performance to ensure that the best performing ensemble is used.

C.5 References

Boucher, M.A., Anctil, F., Perreault, L., Tremblay, D., 2011. A comparison between ensemble and deterministic hydrological forecasts in an operational context. Adv. Geosci. 29, 85–94. https://doi.org/10.5194/adgeo-29-85-2011

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

Bröcker, J., Smith, L.A., 2008. From ensemble forecasts to predictive distribution functions. Tellus, Ser. A Dyn. Meteorol. Oceanogr. 60 A, 663–678. https://doi.org/10.1111/j.1600-0870.2008.00333.x

De Santi, M., Khan, U.T., Arnold, M., Fesselet, J.-F., Ali, S.I., 2021. Forecasting point-of-consumption chlorine residual in refugee settlements using ensembles of artificial neural networks. npj Clean Water Forthcomin.

Ferro, C.A.T., 2014. Fair scores for ensemble forecasts. Q. J. R. Meteorol. Soc. 140, 1917–1923. https://doi.org/10.1002/qj.2270

Fortin, V., Favre, A.C., Saïd, M., 2006. Probabilistic forecasting from ensemble prediction systems: Improving upon the best-member method by using a different weight and dressing kernel for each member. Q. J. R. Meteorol. Soc. 132, 1349–1369. https://doi.org/10.1256/qj.05.167

Hersbach, H., 2000. Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. Weather Forecast. 15, 559–570.

Roulston, M.S., Smith, L.A., 2003. Combining dynamical and statistical ensembles. Tellus, Ser. A Dyn. Meteorol. Oceanogr. 55, 16–30. https://doi.org/10.1034/j.1600-0870.2003.201378.x

Wang, X., Bishop, C.H., 2005. Improvement of ensemble reliability with a new dressing kernel. Q. J. R. Meteorol. Soc. 131, 965–986. https://doi.org/10.1256/qj.04.120