FORECASTING CHLORINE RESIDUAL FOR WATER SAFETY USING ARTIFICIAL NEURAL NETWORKS ENSEMBLES IN HUMANITARIAN WATER SYSTEMS

MICHAEL DE SANTI

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF APPLIED SCIENCE

GRADUATE PROGRAM IN CIVIL ENGINEERING

YORK UNIVERSITY

TORONTO, ONTARIO

August 2021

Abstract

Waterborne illnesses are a leading health concern in refugee and internally displaced person (IDP) settlements where waterborne pathogens often spread through household recontamination of stored water. Ensuring sufficient chlorine residual is important for protecting drinking water against recontamination and ensuring water remains safe up to the point-of-consumption. This thesis investigated the use of ensembles of artificial neural networks (ANNs) to probabilistically forecast the point-of-consumption free residual chlorine (FRC) concentration using water quality data from six refugee and IDP settlements. These models were then used to generate point-of-distribution FRC targets based on the risk of insufficient FRC at the point-of consumption. Overall, the ensemble ANN approach produced accurate risk-based FRC targets, though the ensemble forecasts were underdispersed. Three approaches for overcoming the underdispersion were considered: post-processing ensemble predictions, training the ANNs using cost-sensitive learning, and multi-objective training of the ANNs. Of these approaches, the multi-objective training yielded the best results.

Dedication

This thesis is dedicated to all of my friends and family. To my parents for supporting me and for always believing in me, to Jay, the best elder sibling a little brother could have for always having my back and inspiring to be better, and to Michelle, for being there with me through it all.

Acknowledgements

I would first like to gratefully acknowledge my supervisors, Dr. Usman Khan and Dr. Syed Imran Ali both for their guidance and even more for their friendship.

Next, I would like to extend my gratitude for the support from colleagues in the local refugee population, from MSF, and from UNHCR. From South Sudan, I would like to thank: Simon Juma Choul, Thomas Bashir, Alfaki Yusif, Abdalbagi Madani, Mark Onna, Sebit Khalil and Issa Wallah. In Jordan: Khaled Shapsough, Samer Al- Janadi, Basel Al-Akrad, and the two field data collectors who wish to remain anonymous. In Rwanda: Jérémie Munyarugero, Martin Gashema, Olivier Nugiraneza, Jackson Karumuna, and Pie Migisha. In Tanzania: Anna Hyvaerinen. In Nigeria: Dawn Taylor, Walter Kinyera, Joshua Salia, Daniel Gbaa, Terna Nathaniel Adom, Abraham T Ayangebee, Benedict Demian Tyevejir, Steven Aliegba, Apaa Gemga, Vera Gilbert, Jonathan Francis, and Monia Al-Ra'aini. From MSF OCA: Biserka Pop-Stefanija and Mohammed Ali Omer. From UNCHR: Bernadette Castel-Hollingsworth, Boutros Hijazeen, Murad Al-Shishani, Amin Juzar Bhai, Grace Shaidi Mungwe, Claudia Perlongo, Murray Burt, and Dominique Porteaud.

Additionally, I would like to thank the technical advisors on the SWOT project for their valuable input in helping to direct my research and understand the field challenges that SWOT users face: Jean-François Fesselet, Matthew Arnold, as well as James Orbinski from DIGHR for his advisory support of the SWOT project. I would also like to thank Rahma Shakir, Daliah Adlaer, and Apostolos Vasileiou for their preliminary work on these ANN tools.

I would also like to thank Stephanie Gora for providing feedback on my work and helping me to see the big picture of water safety and helping me gain a deeper understanding of water engineering and Everett Snieder who always had brilliant advice, whether a new idea to try out, a new direction when I was stuck, or just that perfect paper that would crack everything wide open.

Lastly I would like to thank York University, the National Science and Engineering Research Council, the Achmea Foundation, and Grand Challenges Canada for providing funding for my research.

Table of Contents

Abstract ii
Dedication iii
Acknowledgements iv
Table of Contents v
List of Tables xi
List of Figures xiii
List of Abbreviations xx
Glossary xxii

Chapter 1 Introduction 1
1.1 Motivation 1
1.2 Background 1
1.2.1 Chlorination, Waterborne Diseases, and Household Recontamination 1
1.2.2 Modelling Post-Distribution Chlorine Decay 9
1.3 Research Objectives 12
1.4 Thesis Outline 12
1.5 References 15

Chapter 2 Datasets and Data Analysis 20
2.1 Study Site Descriptions 20
2.1.1 South Sudan 22
2.1.2 Jordan 22
2.1.3 Rwanda 22
2.1.4 Bangladesh 23
2.1.5 Tanzania 23
2.1.6 Nigeria 23
2.2 Description of Data Sets 23
2.3 Ethics 26
2.4 Preliminary Data Analysis 26
2.5 References 45

Chapter 3 Proof-of-Concept Study 46
3.1 Chapter Preamble 46
3.2 Abstract 48
3.3 Introduction 48
3.4 Results 53
3.4.1 Ensemble Model Performance 53
3.4.2 South Sudan 57
3.4.3 Jordan (2014) 59
3.4.4 Jordan (2015) 61
3.4.5 Rwanda 62
3.4.6 Partial Correlation Analysis Results 64
3.4.7 Risk-Based FRC Targets 66
3.5 Discussion 71
3.6 Methods 76
3.6.1 Study Sites and Data Collection 76
3.6.2 Ethics 79
3.6.3 Input variable selection 79
3.6.4 Base Learner structure and architecture 80
3.6.5 Data Division 80
3.6.6 Ensemble Model Formation 81
3.6.7 Ensemble Post Processing 82
3.6.8 Ensemble Verification and Performance Evaluation 82
3.6.9 Percent Capture 82
3.6.10 CI Reliability Diagram 83
3.6.11 Continuous Ranked Probability Score 84
3.6.12 Generation of Risk Based Targets 85
3.7 References 87

Chapter 4 Cost-Sensitive Learning 94
4.1 Chapter Preamble 94

4.2 Abstract 96
4.3 Introduction 96
4.4 Methods 101
4.4.1 Description of Study Sites and Data Sets Used 101
4.4.2 Ethics 102
4.4.3 Ensemble Model Building 102
4.4.4 Data Division for Scenario Analysis 104
4.4.5 Cost Functions 112
4.4.6 Cost Weightings 114
4.4.7 Performance Metrics 116
4.5 Results and Discussion 124
4.5.1 Percent Capture Performance 128
4.5.2 CI Reliability Performance 131
4.5.3 RH Performance 134
4.5.4 CRPS and CRPS Reliability 136
4.5.5 Selection of Preferred Model 140
4.5.6 Bangladesh Time-Series Analysis 145
4.6 Conclusion 149
4.7 References 151

Chapter 5 Multi-Objective Training 157
5.1 Chapter Preamble 157
5.2 Abstract 159
5.3 Introduction 159
5.4 Methods 163
5.4.1 Description of Data Sets Used 163
5.4.2 Ethics 163
5.4.3 ANN Ensemble Model Description 163
5.4.4 Approaches to Multi-Objective Training 170
5.4.5 Ensemble Verification Metrics 175
5.5 Results and Discussion 182

5.5.1 Selection of Weights 182
5.5.2 Comparison of Multi-Objective Methods 185
5.5.3 Comparison to Ensemble Post-Processing and Cost-Sensitive Learning 195
5.6 Conclusion 199
5.7 References 201

Chapter 6 Conclusion 206
6.1 Thesis Summary 206
6.2 Opportunities for Future Research 209

Appendices 210
Appendix A. Supplemental Information for Chapter 3 210
Appendix B. Unpublished work: "Predicting drinking water safety in humanitarian crises using artificial neural networks" 251
B.1 Abstract 251
B.2 Acronyms and Abbreviations 251
B.3 Introduction 252
B.4 Methods 255
B.4.1 Study Sites 255
B.4.2 Data Collection 258
B.4.3 Ethics 262
B.4.4 Ensemble Model Building Process 262
B.4.5 Methods for Comparing Local and Global Models 268
B.5 Results and Discussion 269
B.5.1 Model Building 269
B.5.2 Comparison of Local and Global Models 278
B.5.3 Study Limitations and opportunities for future work 283
B.6 Conclusion 285
B.7 Author Contributions 285
Appendix B References 286
Appendix B-1 Data Cleaning Rules 289

Appendix B-2 CNPSA Backwards Selection Pseudocode 290
Appendix B-3 Grid-Search Optimization Pseudocode for Network Optimization 291
Appendix C. SWOT-ANN v2 Analytics White Paper 293
C.1 Executive Summary 293
C.2 Introduction 295
C.3 Workflow of the SWOT-ANN v2 Analytics 296
C.4 Importing Data 297
C.4.1 Input Variable Selection 297
C.5 Training the ANN ensemble models for the SWOT-ANN v2 analytics 299
C.5.1 Model Set-Up and Architecture 299
C.5.2 Training the Ensemble 301
C.5.3 Evaluating Model Performance 302
C.5.4 Model Performance Summary and Outputs 305
C.5.5 Post-Processing 309
C.6 Obtaining a Tapstand FRC Target 310
C.6.1 Scenario Analysis 311
C.6.2 Outputs and Interpretation 312
C.7 Next Steps 327
C.7.1 Planned SWOT-ANN v2 Analytics Updates 327
C.7.2 SWOT-ANN v3 Analytics 327
C.8 Conclusion 328
C.9 References 328
Appendix C-1– Including Time in the ANN Model 330
Introduction 330
Methods 332
Modelling Approach 332
Performance Metrics 334
Results and Analysis 336
Conclusion 343
Appendix C-1 References 343
Appendix C-2 – Input Variable Selection for the SWOT-ANN v2 analytics 343

Introduction 343
Methods 344
Results 348
Conclusion 351
Appendix C-2 References 351
Appendix C-3 – Comparison of Bandwidth Selection Methods for Post-Processing 352
Introduction 352
Methods 353
Results 356
Conclusion 360
Appendix C-3 References 360
Glossary of Functions and Explanations 361
Appendix D. – Supplemental Material for Chapter 4 371
D.1 Data Cleaning Rules 371
D.2 Calculation of weighted cost functions 371
D.2.1 Weighted MSE 372
D.2.2 Weighted NSE 372
D.2.3 Weighting KGE 372
D.2.4 Weighted AI 373
D.3 Supplemental Information 375

List of Tables

Table 1-1: Summary of outbreaks of waterborne diseases in refugee and IDP settlements where household recontamination contributed to the spread of the disease 4
Table 2-1: Water quality parameters collected 25
Table 2-2: Comparison of mean and standard deviation from point-of-distribution to point-of-consumption for historic sites 36
Table 3-1: Ensemble verification metrics for all sites and variable combinations for raw and post-processed ensembles 55
Table 3-2: Partial correlation analysis results between water quality variables and point-of-consumption FRC 65
Table 3-3: Summary of Key Site Characteristics (Médecins Sans Frontières, 2013; PAJER, 2015; UNICEF, 2015) 78
Table 4-1: Input and output variable mean, median, and standard deviations for all sites and input variable combinations for the calibration and testing datasets. Note that the same variable at the same site may have different statistics between the two input variable combinations due to observations being removed for missing. 105
Table 4-2: Input and output variable mean, median, and standard deviations for Bangladesh data in each two-week period 110
Table 4-3: Summary of best performing cost function and weighting combination for each performance metric 125
Table 5-1: Input and output variable mean, median, and standard deviations for all sites and input variable combinations for the calibration and testing datasets. Note that the same variable at the same site may have different statistics between the two input variable combinations due to observations being removed for missing. 165
Table 5-2:Baseline performance for ANN ensembles trained using MSE 181
Table 5-3:Objective weights used for multi-objective training Method 1. Weights are determined as the ratio of each objective's score to the lowest score obtained by an ensemble trained on MSE. This table indicates that the ensembles trained with MSE performed best on the β objective, and worst on the α objective, so the weightings were assigned to counterbalance this. 183

Table 5-4: Summary of skill scores for each ensemble verification metric for multi-objective method as well as the sum of the skill scores (Net column). The baseline used to calculate the skill scores was the MSE performance shown in Table 5-3. 187
Table 5-5: Comparison of approaches to improving ensemble forecast dispersion and reliability. Note for post-processing approach 1, the kernel bandwidth is derived using the best member error method (Roulston & Smith, 2003), and for post-processing approach 2, the kernel is derived using the method proposed by Wang and Bishop (2005)..... 196

List of Figures

Figure 1-1: Typical refugee or IDP settlement water treatment and supply. Since current drinking water quality guidelines only provide sufficient FRC at the point-of-distribution, FRC decay during the post-distribution period (shown in orange) can leave drinking water vulnerable to household recontamination. 3

Figure 1-2: Schematic of an MLP showing flow of data from the input layer to the output layer with weights and biases. Input variable data is accepted in the input layer..... 11

Figure 2-1: SWOT site locations. Development sites (2013-2015) include South Sudan, Jordan, and Rwanda. Implementation sites (2019-Present) include Bangladesh, Tanzania, and Nigeria. 21

Figure 2-2: South Sudan input and output variable histograms and boxplot comparison of point-of-distribution and point-of-consumption FRC. Boxes show the interquartile range, whiskers show the 95th percentile range, and circles represent outliers beyond the 95th percentile range. 28

Figure 2-3:Jordan (2014) input and output variable histograms and boxplot comparison of point-of-distribution and point-of-consumption FRC. Boxes show the interquartile range, whiskers show the 95th percentile range, and circles represent outliers beyond the 95th percentile range. 29

Figure 2-4:Jordan (2015) input and output variable histograms and boxplot comparison of point-of-distribution and point-of-consumption FRC. Boxes show the interquartile range, whiskers show the 95th percentile range, and circles represent outliers beyond the 95th percentile range. 30

Figure 2-5: Rwanda input and output variable histograms and boxplot comparison of point-of-distribution and point-of-consumption FRC. Boxes show the interquartile range, whiskers show the 95th percentile range, and circles represent outliers beyond the 95th percentile range. 31

Figure 2-6: Bangladesh input and output variable histograms and boxplot comparison of point-of-distribution and point-of-consumption FRC Boxes show the interquartile range, whiskers show the 95th percentile range, and circles represent outliers beyond the 95th percentile range. 32

Figure 2-7: Tanzania input and output variable histograms and boxplot comparison of point-of-distribution and point-of-consumption FRC Boxes show the interquartile range, whiskers show the 95th percentile range, and circles represent outliers beyond the 95th percentile range. 33
Figure 2-8: Nigeria input and output variable histograms and boxplot comparison of point-of-distribution and point-of-consumption FRC. Boxes show the interquartile range, whiskers show the 95th percentile range, and circles represent outliers beyond the 95th percentile range. 34
Figure 2-9: South Sudan trends between input variables and point-of-consumption FRC. All variables other than elapsed time and point-of-distribution FRC have negative trends with point-of-consumption FRC. 38
Figure 2-10: Jordan 2014 trends between input variables and point-of-consumption FRC. All variables other than elapsed time and point-of-distribution FRC have negative trends with point-of-consumption FRC. 39
Figure 2-11: Jordan 2015 trends between input variables and point-of-consumption FRC. All variables other than elapsed time, turbidity, and point-of-distribution FRC have negative trends with point-of-consumption FRC. 40
Figure 2-12: Rwanda trends between input variables and point-of-consumption FRC. All variables other than turbidity and pH have positive trends with point-of-consumption FRC. 41
Figure 2-13: Bangladesh trends between input variables and point-of-consumption FRC. All variables other than point-of-distribution FRC have negative trends with point-of-consumption FRC. 42
Figure 2-14: Tanzania trends between input variables and point-of-consumption FRC. All variables have positive trends with point-of-consumption FRC. 43
Figure 2-15: Nigeria trends between input variables and point-of-consumption FRC. Point-of-distribution FRC and water temperature both have positive trends with point-of-consumption FRC. 44
Figure 3-1: Post-distribution period shown in context of overall water supply system for typical refugee or IDP settlement. Water obtained from ground or surface water is centrally treated then conveyed via piped distribution system to the tap stand (point-of-

distribution). The post-distribution period begins when water is collected from the tap stand and continues as it is transported to the household and then stored until use (point-of-consumption). 50
Figure 3-2: Confidence Interval reliability diagrams for all sites. Raw and post-processed CI reliability diagrams for all sites for both the overall dataset (left) and for observations where point-of-consumption FRC is below 0.2 mg/L (right). All ensembles have Percent Capture below the 1:1 line, indicating underdispersion at all CI's, though better reliability is observed for models using the IV2 input variable combination.. 56
Figure 3-3: South Sudan observations and post-processed forecasts of point-of-consumption FRC. The observations and forecasts are shown against (a) point-of-distribution FRC, (b) elapsed time, (c) point-of-distribution EC, (d) point-of-distribution water temperature, (e) point-of-distribution pH, (f) point-of-distribution turbidity. A strong trend between point-of-consumption and point-of-distribution FRC is observed and IV2 forecasts are much more dispersed than IV1 forecasts. 58
Figure 3-4: Jordan (2014) observations and post-processed forecasts of point-of-consumption FRC. The observations and forecasts are shown against (a) point-of-distribution FRC, (b) elapsed time, (c) point-of-distribution EC, (d) point-of-distribution water temperature, (e) point-of-distribution pH, (f) point-of-distribution turbidity. IV1 forecasts show a strong regression to the mean behaviour. Strong trends between point-of-consumption FRC and: point-of-distribution FRC, EC, and water temperature..... 60
Figure 3-5: Jordan (2015) observations and post-processed forecasts of point-of-consumption FRC. The observations and forecasts are shown against (a) point-of-distribution FRC, (b) elapsed time, (c) point-of-distribution EC, (d) point-of-distribution water temperature, (e) point-of-distribution pH, (f) point-of-distribution turbidity. Both IV1 and IV2 forecasts are very flat due to low overall rates of FRC decay at this site. ... 62
Figure 3-6: Rwanda observations and post-processed forecasts of point-of-consumption FRC. The observations and forecasts are shown against (a) point-of-distribution FRC, (b) elapsed time, (c) point-of-distribution EC, (d) point-of-distribution water temperature, (e) point-of-distribution pH, (f) point-of-distribution turbidity. IV2

forecasts tend to be much more dispersed, leading to better overall capture, especially of observations with point-of-consumption FRC below 0.2 mg/L. 63
Figure 3-7: Predicted risk of insufficient point-of-consumption FRC (below 0.2 mg/L). The predicted risk is shown for (a) South Sudan, (b) Jordan (2014), (c) Jordan (2015), and (d) Rwanda. To achieve negligible risk, the ANN ensemble models recommend point of distribution FRC between 0.65 and 0.90 mg/L in South Sudan, between 0.7 and 1.75 mg/L in Jordan (2014), between 0.2 and 0.4 mg/L in Jordan (2015), and between 0.60 and 0.90 mg/L in Rwanda. The upper limit of the recommendation for Jordan (2014) does not ensure negligible risk, as this was never achieved, but represents a plateau in the predicted risk of FRC below 0.2 mg/L. 68
Figure 3-8: Forecasts used to generate risk-based FRC targets. Top row: South Sudan, Second row: Jordan (2014), Third Row: Jordan (2015), Bottom row: Rwanda. Left column: forecasts produced by models using IV1, middle column: forecasts produced by models using IV2 for average case. Right column: forecasts produced by models using IV2 for worst case scenario. 70
Figure 4-1: Ensemble forecasts where ANN base-learners trained with MSE for two input variable combinations (IV1 and IV2, see Section 4.4.3). Forecast-observation pairs shown for: (a) Bangladesh IV1, (b) Bangladesh IV2, (c) Tanzania IV1, (d) Tanzania IV2, (e) Nigeria IV1, and (f) Nigeria IV2 These figures show that ensembles trained with MSE are highly underdispersed as the forecast range does not cover all observations. 99
Figure 4-2: Visualization of the method used to calculate the different performance evaluation metrics used in this study for the CRPS (left), RH (centre) and CI reliability diagram (right). The CRPS is calculated from the difference in area between the forecast cdf and the Heaviside function (observation cdf), the rank histogram is derived from the rank of the observation relative to each model prediction, and the CI reliability diagram is based off the percent capture in each ensemble CI. 122
Figure 4-3: Forecast-observation comparison for all sites showing difference in forecast range and median between the baseline ensemble (trained with unweighted MSE) and ensemble with the highest PC for each site and variable combination. Forecast observation pairs shown for: (a) Bangladesh IV1, (b) Bangladesh IV2, (c) Tanzania

IV1, (d) Tanzania IV2, (e) Nigeria IV1, and (f) Nigeria IV2. In all of the above subplots we see that the best performing ensemble (shown in red) produces a larger forecast range (shaded area) than the baseline ensemble (blue), which is what produced the better Percent Capture. 129
Figure 4-4: Forecast-observation comparison for all sites showing difference in forecast range and median between the baseline ensemble (trained with unweighted MSE) and ensemble with the highest $PC < 0.2$ for each site and variable combination. Forecast observation pairs shown for: (a) Bangladesh IV1, (b) Bangladesh IV2, (c) Tanzania IV1, (d) Tanzania IV2, (e) Nigeria IV1, and (f) Nigeria IV2. In all of the above subplots we see that the best performing ensemble (shown in red) produces a larger forecast range (shaded area) than the baseline ensemble (blue), which is what produced the better Percent Capture. 130
Figure 4-5: CI reliability diagrams for all sites showing difference in forecast range and median between the baseline ensemble (trained with unweighted MSE) and ensemble with the best $CIscore$ score for each site and variable combination. CI reliability diagrams shown for: (a) Bangladesh IV1, (b) Bangladesh IV2, (c) Tanzania IV1, (d) Tanzania IV2, (e) Nigeria IV1, and (f) Nigeria IV2. These figures show that in all cases the baseline ensemble (trained with unweighted MSE) produced underdispersed forecasts, with the capture in all CIs below the 1:1 line. 132
Figure 4-6: CI reliability diagrams for all sites showing difference in forecast range and median between the baseline ensemble (trained with unweighted MSE) and ensemble with the best $CIscore < 0.2$ score for each site and variable combination. CI reliability diagrams shown for: (a) Bangladesh IV1, (b) Bangladesh IV2, (c) Tanzania IV1, (d) Tanzania IV2, (e) Nigeria IV1, and (f) Nigeria IV2. 133
Figure 4-7: RH for Bangladesh with the IV1 variable combination. The RH in (a) shows the results for the models trained with unweighted MSE, and (b) shows the model with the best δ-. Both RHs are underdispersed, as seen from the u-shape of the RH, but the size of the outlier bars, and the difference between the outliers and internal bars are much smaller in (b) than in (a), indicating improved reliability with alternative cost functions and cost function weighting. 135

Figure 4-8: RH of observations with point-of-consumption FRC below 0.2 mg/L for Bangladesh with the IV1 variable combination. The RH in (a) shows the results for the models trained with unweighted MSE, and (b) shows the model with the best $\delta < 0.2$-score. Both RHs are underdispersed, as seen from the u-shape of the RH, but the size of the outlier bars, and the difference between the outliers and internal bars are much smaller in (b) than in (a), indicating improved reliability with alternative cost functions and cost function weighting. 136
Figure 4-9: Forecast-observation comparison for all sites showing difference in forecast range and median between the baseline ensemble (trained with unweighted MSE) and ensemble with the best $CRPS$ for each site and variable combination. Forecast observation pairs shown for: (a) Bangladesh IV1, (b) Bangladesh IV2, (c) Tanzania IV1, (d) Tanzania IV2, (e) Nigeria IV1, and (f) Nigeria IV2. 138
Figure 4-10: Forecast-observation comparison for all sites showing difference in forecast range and median between the baseline ensemble (trained with unweighted MSE) and ensemble with the best $Reli$ for each site and variable combination. Forecast observation pairs shown for: (a) Bangladesh IV1, (b) Bangladesh IV2, (c) Tanzania IV1, (d) Tanzania IV2, (e) Nigeria IV1, and (f) Nigeria IV2. 139
Figure 4-11: Frequency of cost functions and weighting combinations producing the best performance (left) or one of the "top five" performances for a given site and input variable combination for each performance metric and for all performance metrics combined (bottom row). Consistently KGE and IoA with weighting 3 produce the most "best" and "top five" performances, and MSE, particularly unweighted MSE, performs poorly. 141
Figure 4-12: Comparison of daily observed and predicted point-of-consumption FRC concentrations for ensembles with base learners trained using (a) unweighted MSE, (b) IoA with Weighting 3, and (c) KGE with Weighting 3. Increasing observations used for calibration represents increasing data becoming available over time. The MSE forecasts are consistently underdispersed, weighted IoA and KGE both better match the observations. 147
Figure 4-13: Comparison of ensemble verification metrics for Bangladesh time-series analysis. From top: CI Reliability score, Percent Capture, δ-score, CRPS and reliability term.

The models trained using KGE with weighting 3 tend to have the best capture, CI reliability score, and δ-score. The CRPS reliability results are less clear. 148
Figure 5-1: Summary of max and min performance improvement over 4 rounds of random search to determine weights for the multi-objective training Method 3. This figure shows that over 4 rounds of optimization the net improvement at each site increases, as does the number of weighting combinations that produce improvement. This indicates that the successive random search optimization process used effectively navigated the search space towards good alternatives. 184
Figure 5-2: Comparison of net improvement and number of positive skill scores for all multi-objective training methods, showing that consistently multi-objective training Method 2 produces the highest net improvement scores and the most positive skill scores (indicating consistent improvement). 189
Figure 5-3: Predictions, CI reliability diagram, and Rank Histogram for each site and variable combination using multi-objective training Method 2. From this figure we see that there is some overdispersion in the Bangladesh models, and at all sites there are predictions that are impossible (point-of-consumption FRC lower than 0 or higher than the point-of-distribution concentration). While these predictions are physically impossible, they are an indicator of the models reflecting the high degree of uncertainty present at these sites. 193
Figure 5-4: Predictions, CI reliability diagram, and Rank Histogram for each site and variable combination using multi-objective training Method 2, after post processing. In comparison to Figure 5-3, this figure includes no unrealistic predictions, however, there is minimal change in the ensemble performance. This indicates that the post-processing technique effectively constrained the predictions to within the range of possible point-of-consumption FRC concentrations without meaningfully impacting the underlying distribution. 194
Figure 5-5: Comparison of net improvement and number of positive skill scores for all approaches, showing that consistently multi-objective training Method 2 produces the highest net improvement scores and the most positive skill scores (indicating consistent improvement). 199

List of Abbreviations

ANN	Artificial neural network
cdf	Cumulative density function
CI	Confidence interval
CRPS	Continuous ranked probability score
CPHS	Committee for the Protection of Human Subjects
DBP	Disinfection by-product
DIGHR	Dahdaleh Institute for Global Health Research
EC	Electrical conductivity
FRC	Free residual chlorine
IDP	Internally displaced person
IV1	Input variable combination 1
IV2	Input variable combination 2
IVC	Input variable combination
KGE	Kling-Gupta Efficiency
MAE	Mean absolute error
MLP	Multi-layer perceptron
MOGA	Multi-objective genetic algorithm
MSE	Mean squared error
MSF	Médecins sans Frontières
NSE	Nash-Sutcliffe Efficiency
NSGA	Non-dominated Sorting Genetic Algorithm

pdf	Probability density function
RH	Rank histogram
SciPy	Scientific Python software package
SWOT	Safe Water Optimization Tool
UNHCR	United Nations High Commissioner for Refugees
WASH	Water, sanitation, and hygiene

Glossary

Residual Chlorine	Chlorine remaining after initial treatment and disinfection of drinking water used to prevent recontamination of drinking water in the distribution system and during household storage. Also referred to as a "secondary disinfectant"
Process-Based Model	A model that uses chemistry and chemical processes to simulate chlorine decay behaviour
Data-Driven Model	A model that replicates behaviours in observed data without assumptions of the underlying behaviour
Dispersion	A measure of the spread of the individual predictions within an ensemble forecast. An overdispersed forecast has a prediction spread greater than the spread of the observations, and an underdispersed forecast has a prediction spread that is smaller than the spread of the observations
Reliability	A general term for the similarity of a forecast distribution to the underlying distribution. Good reliability means that the probability distribution of a model forecast is very similar to the underlying distribution of the data. Reliability is not a score in and of itself, though it is measured by scores such as the Rank Histogram δ score, the CI Reliability Score, and the $CRPS$ and $CRPS$ reliability score.
Cost Function	The function used during training to optimize the model parameters of a data-driven model. Often common error metrics like mean squared error are used.

Chapter 1 Introduction

1.1 Motivation

Despite decades of effort in developing drinking water quality guidelines for humanitarian response, outbreaks of waterborne illness are still prevalent in refugee and internally displaced person (IDP) settlements. In these settlements, waterborne illnesses may spread through recontamination of drinking water during collection, transport, and storage because the residual chlorine targets in current drinking water quality guidelines for humanitarian response (such as the sector-standard Sphere Handbook (Sphere Association, 2018)) do not account for chlorine decay after water leaves the distribution system. With the growing number of displaced persons worldwide, it is critical to address this key gap in the provision of safe drinking water in refugee and IDP settlements. The Safe Water Optimization Tool (SWOT) project aims to leverage the wealth of water quality data that is routinely collected in refugee and IDP settlements to develop improved, evidence based residual chlorine targets to ensure drinking water is protected against recontamination up to the point-of-consumption.

1.2 Background

1.2.1 Chlorination, Waterborne Diseases, and Household Recontamination

Waterborne illnesses are a leading cause of excess morbidity and mortality in refugee and IDP settlements (Connolly et al., 2004; Cronin et al., 2008; Salama, Spiegel, Talley, Waldman, and Street, 2004). As shown in Table 1-1, recontamination of drinking water during collection and household storage is a major factor in outbreaks of waterborne illnesses and has been identified as a contributing factor to outbreaks of cholera, Hepatitis E, and shigellosis in refugee and IDP settlements in Kenya, Malawi, Sudan, South Sudan, and Uganda, amongst others. A common problem identified in the outbreaks listed in Table 1-1 is inadequate chlorination, particularly a lack of sufficient residual chlorine at the point-of-consumption. This is critical as all of the pathogens identified in Table 1-1 are sensitive to inactivation by chlorine and their transmission in stored drinking water can be prevented by ensuring adequate residual chlorine up to the point of consumption (Girones et al., 2014; WHO, 2017). Typically, a free residual chlorine (FRC) concentration of 0.2 mg/L is sufficient to prevent recontamination by these pathogens so long as this residual is maintained throughout the post-distribution period (CDC, 2012; Girones et al., 2014; Lantagne, 2008; Rashid et al., 2016; Sikder et al., 2020; WHO, 2011). However, current

drinking water quality guidelines for refugee and IDP settlements do not ensure sufficient residual chlorine up to the point-of-consumption. Instead, these guidelines are based on municipal drinking water guidelines and only provide 0.2-0.5 mg/L of FRC up to the end of the piped distribution system, typically a water distribution point. Chlorine decay during the post-distribution period of collection, transport and household storage reduces the amount of residual chlorine available to protect against recontamination to levels below 0.2 mg/L, leaving stored drinking water vulnerable to pathogenic recontamination. In order to ensure that drinking water remains safe up to the point-of-consumption, new FRC guidelines are required for the point-of-distribution that account for post-distribution chlorine decay. Figure 1-1, which is a modified version of Figure 3-1, shows the post-distribution period in the context of the overall water treatment and supply processes in a refugee or IDP settlement.

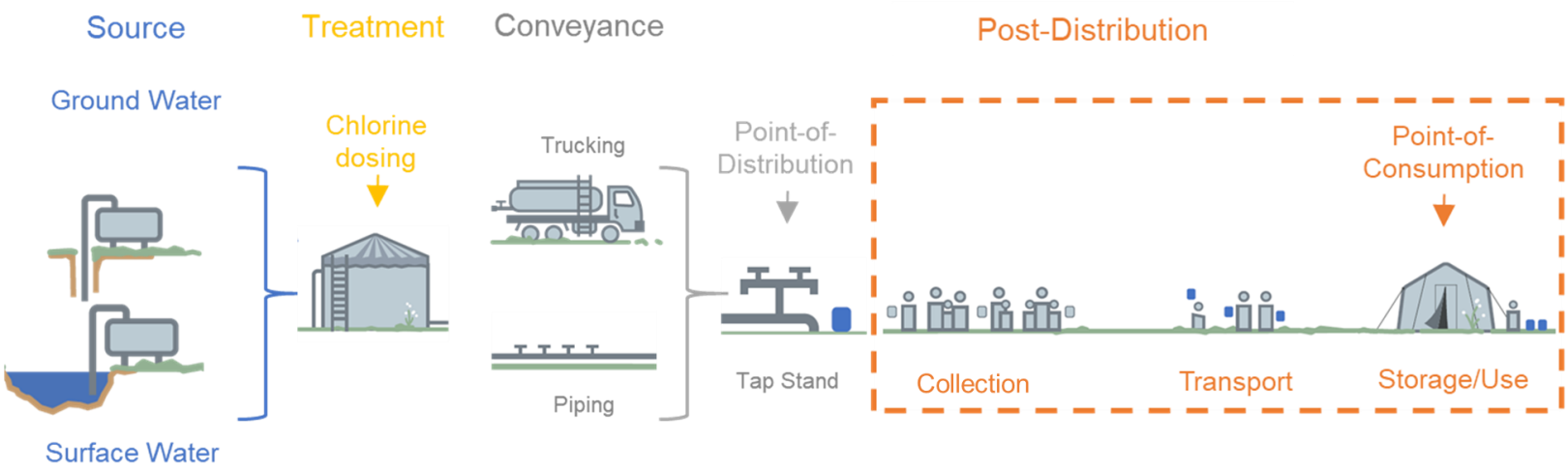

Figure 1-1: Typical refugee or IDP settlement water treatment and supply. Since current drinking water quality guidelines only provide sufficient FRC at the point-of-distribution, FRC decay during the post-distribution period (shown in orange) can leave drinking water vulnerable to household recontamination.

Table 1-1: Summary of outbreaks of waterborne diseases in refugee and IDP settlements where household recontamination contributed to the spread of the disease

Location of Outbreak	**Date of Outbreak**	**Disease (Pathogen)**	**Summary of Findings**
Dabaab Refugee Camp, Garissa County, Kenya	November 2015 to June 2016	Cholera (*Vibrio Cholerae*)	A geospatial analysis found clustering of cases in housing blocks, and some evidence of multiple members of the same household becoming ill, indicating that cholera was spreading at least in part through the household. FRC in the household was below the outbreak recommendation of 0.5 mg/L, suggesting that spread of cholera in the household may have been through contaminated drinking water (Golicha et al., 2018).
Kakuma Refugee Camp, Turkana County, Kenya	April 2005	Cholera (*Vibrio Cholerae*)	A multivariate analysis of case-controlled odds-ratios for different risk factors found that the only significant protective factor was storing drinking water in a sealed/covered container. This indicates that cholera may have spread at least in part through household recontamination of drinking water. While FRC was not measured in the household, the FRC provided at the public distribution point varied between 0.0 mg/L and 5.0 mg/L. Thus, insufficient FRC in stored drinking water may have allowed the disease to spread (Shultz et al., 2009).

Location of Outbreak	Date of Outbreak	Disease (Pathogen)	Summary of Findings
Nyamithuthu Refugee Camp, Nsanje District, Malawi	August 1990	Cholera (*Vibrio Cholerae*)	This paper includes two case-controlled studies in the camp, one of which that found dipping hands in stored household water to be a significant risk factor, indicating transmission may have occurred through post-distribution recontamination. Samples taken from four households found *vibrio cholera* bacteria in household stored drinking water, indicating stored drinking water as a major potential risk factor. Chlorination took place in the camp, however, samples taken from drinking water stored in the household showed no FRC, indicating that FRC decay may have left water stored in the dwelling vulnerable to recontamination (Swerdlow et al., 1997).
Batil, Jamam, and Gendrassa refugee camps, Maban County, South Sudan	Multiple outbreaks from 2012 to 2014	Acute Jaundice Syndrome (AJS) (Hepatitis E Virus [HEV])	Hepatitis E virus spread in the camps during the rainy season when poor drainage in the camps caused latrines to flood (Ali, Ali, and Fesselet, 2015). Despite drinking water at public distribution points meeting or exceeding Sphere standards, chlorine decay during household storage led to no FRC being detectable in 40-58% of households (Ali et al., 2015). An environmental study following the outbreak tested for human adenovirus (HadV) as an

Location of Outbreak	Date of Outbreak	Disease (Pathogen)	Summary of Findings
			indicator of faecal recontamination as the environmental conditions during the study were no longer favourable for HEV. HadV was not detected in any source water or water from public distribution points, but it was detected in household stored drinking water, specifically in houses with no FRC (Guerrero-Latorre, Hundesa, and Girones, 2016). This indicates not only that viral contamination of drinking water spread due to insufficient FRC, but also that the lack of chlorine residual in the household was due to post-distribution FRC decay.
Abou Shouk IDP Camp, Darfur Province, Sudan	May to June 2004	Dysentery/ Shigellosis (*Shigella dysenteriae*)	Source water sampling showed little to no faecal contamination (only one coliform detected in one 100 mL sample from 19 shallow wells), indicating that the spread of shigellosis was likely through contamination occurring during collection or storage. Water was not chlorinated prior to the outbreak, but chlorination began in response to the outbreak, along with a container disinfection program which provided an average chlorine residual of 0.22 mg/L in the household. While this study did not directly measure microbial contamination during storage before or after the

Location of Outbreak	Date of Outbreak	Disease (Pathogen)	Summary of Findings
			outbreak, there was a sharp decline in cases of watery and bloody diarrhea after the disinfection campaign. Thus, providing sufficient FRC during collection and storage likely helped to stop the spread of shigellosis through household recontamination (Walden, Lamond, and Field, 2005).
Odki satellite settlement IDP camp, Kitgum Region, Uganda	May 2006 and July 2007	Cholera (*Vibrio Cholerae*)	Outbreaks of cholera were assumed to be caused by household recontamination as drinking water sources were mostly contaminant free. Containers were disinfected using a sodium hypochlorite solution, however, recontamination occurred after cleaning, resulting in regrowth of pathogens after disinfection had taken place, in some cases with higher total coliforms 3 or 5 days after cleaning than there were before (Steele, Clarke, and Watkins, 2008). This shows the need for chlorination of drinking water to prevent ongoing container contamination.
Madi Opei sub-county IDP camps, Kitgum Region, Uganda	October 2007	AJS (HEV)	A multivariate analysis of risk factors from a case-controlled study found that storage of drinking water in a wide mouthed container and using a communal

Location of Outbreak	Date of Outbreak	Disease (Pathogen)	Summary of Findings
			handwashing basin were significant risk factors for HEV infection. This indicates that the virus spread through contamination of drinking water, especially since the water distributed was not chlorinated. The camp used POU chlorination, however, having chlorine tablets was not a significant protective factor (Howard et al., 2010). This may indicate, that the tablets were not being used, highlighting the need for central chlorination.

1.2.2 Modelling Post-Distribution Chlorine Decay

Ensuring that there will be adequate chlorine residual throughout the post-distribution period requires us to identify a chlorine dose for the water distribution point that consistently provides at least 0.2 mg/L at the point-of-consumption. However, post-distribution chlorine decay is highly specific to the conditions of a local settlement or site, so no single dose is acceptable for all sites. Instead, we need numerical models that accurately predict the point-of-consumption FRC concentration using data available at water distribution points to generate site-specific FRC guidance. The Safe Water Optimization Tool (SWOT) project uses regularly collected water quality data from refugee and IDP settlements to generate models of post-distribution FRC decay that provide improved FRC guidance aimed at ensuring sufficient protection against recontamination up to the point-of-consumption.

A common approach to modelling FRC decay in piped distribution systems is to use process-based models which use generalized chemical reaction models to estimate the rate of FRC decay within the distribution system. This process-based approach has been combined with numerical modelling approaches to predict post-distribution FRC decay in refugee and IDP settlements. However, one of the main challenges in this method is that the nature of chlorine decay outside of piped distribution systems is not well understood, so substantial simplifications are required in modelling the post-distribution decay behaviour (Ali, Ali, and Fesselet, 2021).

Artificial neural networks (ANNs) are a type of data-driven model that have been proposed as an alternative to process-based models for predicting FRC in piped distribution systems (Rodriguez & Sérodes, 1998). As data-driven models, ANNs learn the underlying behaviour from the data instead of assuming the behaviour *a priori*, thus avoiding questions of the best decay model. ANNs can also be trained on data representing a wide range of operating conditions and can be retrained easily with new data, unlike process-based models which require decay parameters to be calibrated to a single set of conditions (Bowden, Nixon, Dandy, Maier, and Holmes, 2006; Soyupak, Kilic, Karadirek, and Muhammetoglu, 2011). ANNs are also effective even when only using data collected through routine monitoring (Gibbs et al., 2003, 2006) as they do not need to quantify the effect of each input variable on the decay parameters, the model simply identifies the observed relation between each input variable and the predicted FRC.

While many types of ANNs exist, the most common ANN type used for modelling FRC is the multi-layer perceptron (MLP). 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 1-2. The MLP structure with one hidden layer 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 & Sérodes, 1998). 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. This process is repeated as data moves from the hidden layer to the output layer, with the final output producing a prediction of the target variable (the FRC concentration at the point where FRC is being predicted). A cost function is then used to quantify the difference between the predicted FRC concentration and the observed concentration, and a training algorithm is used to modify the weights and biases (which are equivalent to calibration parameters in physical models) to minimize the value of this cost function.

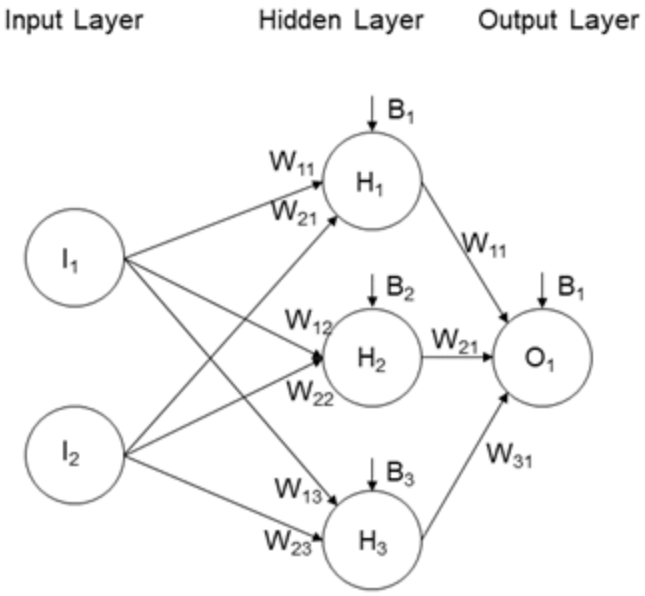

Figure 1-2: Schematic of an MLP showing flow of data from the input layer to the output layer with weights and biases. Input variable data is accepted in the input layer

One of the limitations of both the process-based approach and the data-driven approach are that these models output deterministic (i.e., point) predictions. However, due to the numerous quantifiable and unquantifiable factors influencing chlorine decay, post-distribution FRC decay tends to be highly variable. As an open system, this variability is driven by differences in water quality, environmental conditions, and user interactions. This leads to a high degree of uncertainty that cannot be communicated using a deterministic model. A common approach for quantifying model uncertainty is to use probabilistic ensemble modelling where the predictions of multiple individual models are grouped into a probability density function to form a probabilistic forecast (Boucher, Anctil, Perreault, and Tremblay, 2011; Boucher, Perreault, and Anctil, 2009).This approach is most common in atmospheric sciences where weather forecasts are often obtained using ensembles of physical models, though there has been some limited use ensembles of ANNs for forecasting hydrological variables (Boucher et al., 2011, 2009). This approach is fundamentally different from the typical use for ensembles of ANNs which are typically used for improving the mean prediction by producing a more robust mean prediction or quantifying the uncertainty in the mean prediction. Probabilistic ANN ensembles operate from a different perspective: instead of trying to identify the prediction that best fits the data, probabilistic ensemble models produce forecasts that account for the uncertainty in the process

as a whole (Boucher et al., 2009). This paradigm is particularly promising for the application of modelling post-distribution FRC concentrations due to the high degree of uncertainty inherent in the process. In this thesis we present an investigation into the use of probabilistic ANN ensembles for forecasting point-of-consumption FRC and extend this to generating risk-based FRC guidance.

1.3 Research Objectives

This research presents the first attempt to use ensembles of ANNs to probabilistically forecast point-of-consumption FRC during the post-distribution period. The objectives for this research are to:

1. *Generate and evaluate ANN ensembles to produce probabilistic forecasts of point-of-consumption FRC using regularly collected water quality data from refugee and IDP settlements*
2. *Develop a technique for using probabilistic forecasts obtained from ANN ensembles to generate site-specific FRC recommendations based on the risk of inadequate FRC at the point-of-consumption*
3. *Investigate alternate approaches to training and generating ANN base learners to improve the probabilistic performance of ANN ensembles*

1.4 Thesis Outline

This thesis follows the format of a manuscript-based thesis, with Chapters 3, 4, and 5 all including modified versions of published or proposed manuscripts. As such, each chapter has its own references section, and Chapters 3, 4, and 5 each have their own abstract and introduction, which may lead to some repeated information.

The structure of the thesis is as follows. Chapter 1 provides a brief introduction to the need to ensure sufficient chlorine residual up to the point of consumption in refugee and IDP settlements and presents past approaches to modelling chlorine residual decay. Chapter 2 provides a summary of the datasets used in this study as well as exploratory analysis of this data. The following three chapters are modified versions of published or proposed journal articles. Chapter 3 presents a proof-of-concept study demonstrating the effectiveness of the ensemble ANN approach for modelling post-distribution FRC in refugee and IDP settlements. This study

developed ensemble models using water quality datasets from three refugee settlements in South Sudan, Jordan, and Rwanda. Both the raw and post-processed ensembles were evaluated and a risk-based method was developed to convert forecasts into FRC recommendations. This proof-of-concept study identified the key challenge of underdispersion in the ANN ensemble forecasts, which limits the ability of the ANN ensemble forecasts to capture extreme events (such as those with low household FRC) which reduces the accuracy of the risk-based FRC targets. Underdispersion is a common challenge of probabilistic ensemble modelling, especially when using ensembles of ANNs (Boucher et al., 2009; Boucher, Perreault, Anctil, and Favre, 2015), so solving this challenge became the primary focus of the remainder of the research. In Chapter 3, ensemble post-processing via kernel dressing, a common solution to forecast underdispersion, was used to attempt to improve the dispersion and reliability of the ensemble forecasts. This study was published in *npj Clean Water* in June, 2021.

While post-processing is a common approach to improving forecasting in ensemble modelling, other machine learning applications highlight that addressing challenges in model behaviour during training tends to be more effective than post-processing model outputs (Dress, Lessmann, and von Mettenheim, 2018). Chapters 4 and 5 investigate two alternative approaches to training the ANN base learners of the ensemble models to improve the dispersion and reliability of the ensemble forecasts. Chapter 4 investigates cost-sensitive learning which modifies the training of an ANN to prioritize certain regions of the output space. In cost-sensitive learning, cost functions are weighted either to prioritize a specific behaviour (such as prioritizing samples with low household FRC to improve accuracy in that part of the output space) or as an approach to rebalancing the dataset (typically used when one region of the output space is much more common than another). These have been proposed and demonstrated to be effective for deterministic regression and for classification approaches (Crone, Lessmann, and Stahlbock, 2005; Elkan, 2001; Ling & Sheng, 2008; Toth, 2016; Zhou & Liu, 2010), but they have not been extended to probabilistic ensembles. Chapter 5 investigates the use of multi-objective training. This approach trains an ANN to optimize multiple objectives at once when training the ANN base learners. This method is very common for regression problems, including ensemble modelling (Abbass, 2003; Albuquerque Teixeira, Braga, Takahashi, and Saldanha, 2000; de Vos & Rientjes, 2008; Taormina & Chau, 2015), but prior to this research it has not been extended to probabilisti*c* ensemble modelling. Chapter 5 concludes with a comparison of the three

approaches to overcoming forecast underdispersion (post-processing, cost-sensitive learning, multi-objective training) to identify the ideal approach for training ensembles of ANNs for generating risk-based FRC targets in humanitarian response. This is discussed further in Chapter 6 which summarizes the major conclusions of each chapter and identifies recommendations for implementation of the research findings as well as directions for future research.

There are four appendices for this thesis. Appendix A includes the Supplemental Information that was published with Chapter 3. Appendix B includes an exploratory analysis into the development of a deterministic, ANN-based modelling approach which partially informed the need for a probabilistic modelling approach. Appendix C includes a white paper summarizing the work done to incorporate the findings of Chapter 3 into the online SWOT analytics, including additional research into dynamic input variable selection, including time-based variables into the SWOT-ANN, and a comparison of approaches to bandwidth determination for the kernel post-processing. Finally, Appendix D includes the Supplemental Information for Chapters 4 and 5.

1.5 References

Abbass, H. A. (2003). Pareto Neuro-Evolution: Constructing Ensemble. *Congress on Evolutionary Computation, 3,* 2074–2080.

Albuquerque Teixeira, R. de, Braga, A. P., Takahashi, R. H. C., & Saldanha, R. R. (2000). Improving generalization of MLPs with multi-objective optimization. *Neurocomputing, 35*(1–4), 189–194. https://doi.org/10.1016/S0925-2312(00)00327-1

Ali, S. I., Ali, S. S., & Fesselet, J.-F. (2015). Effectiveness of emergency water treatment practices in refugee camps in South Sudan. *Bulletin of the World Health Organization, 93*(8), 550–558. https://doi.org/10.2471/BLT.14.147645

Boucher, M. A., Perreault, L., & Anctil, F. (2009). Tools for the assessment of hydrological ensemble forecasts obtained by neural networks. *Journal of Hydroinformatics, 11*(3–4), 297–307. https://doi.org/10.2166/hydro.2009.037

Boucher, M. A., Anctil, F., Perreault, L., & Tremblay, D. (2011). A comparison between ensemble and deterministic hydrological forecasts in an operational context. *Advances in Geosciences, 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. *Hydrological Processes, 29*(6), 1141–1155. https://doi.org/10.1002/hyp.10234

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. *Mathematical and Computer Modelling, 44*(5–6), 469–484. https://doi.org/10.1016/j.mcm.2006.01.006

Centre for Diseases Control (CDC). (2012). Chlorine Residual Testing. Retrieved from http://www.cdc.gov/safewater/chlorine-residual-testing.html

Connolly, M. A., Gayer, M., Ryan, M. J., Salama, P., Spiegel, P., & Heymann, D. L. (2004). Communicable diseases in complex emergencies: impact and challenges. *The Lancet, 364*(9449), 1974–1983. https://doi.org/https://doi.org/10.1016/S0140-6736(04)17481-3

Crone, S. F., Lessmann, S., & Stahlbock, R. (2005). Utility based data mining for time series

analysis - Cost-sensitive learning for neural network predictors. *Proceedings of the 1st International Workshop on Utility-Based Data Mining, UBDM '05,* 59–68. https://doi.org/10.1145/1089827.1089835

Cronin, A. A., Shrestha, D., Cornier, N., Abdalla, F., Ezard, N., & Aramburu, C. (2008). A review of water and sanitation provision in refugee camps in association with selected health and nutrition indicators - the need for integrated service provision. *Journal of Water and Health*, *6*(1), 1–13. https://doi.org/10.2166/wh.2007.019

de Vos, N. J., & Rientjes, T. H. M. (2008). Multiobjective training of artificial neural networks for rainfall-runoff modeling. *Water Resources Research*, *44*(8), 1–15. https://doi.org/10.1029/2007WR006734

Dress, K., Lessmann, S., & von Mettenheim, H. J. (2018). Residual value forecasting using asymmetric cost functions. *International Journal of Forecasting*, *34*(4), 551–565. https://doi.org/10.1016/j.ijforecast.2018.01.008

Elkan, C. (2001). The Foundations of Cost-Sensitive Learning The Foundations of Cost-Sensitive Learning. *In Proceedings ofthe Seventeenth International Joint Conference on Artificial Intelligence.* Seattle, Washington. 973–978

Gibbs, M. S., Morgan, N., Maier, H. R., Dandy, G. C., Holmes, M., & Nixon, J. B. (2003). Use of Artificial Neural Networks for Modelling Chlorine Residuals in Water Distribution Systems. *In MODSIM 2003 International Congress on Modelling and Simulation: Integrative Modelling of Biophysical, Social, and Economic Systems for Resource Management Solutions*. 789–794.

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. *Mathematical and Computer Modelling*, *44*(5–6), 485–498. https://doi.org/10.1016/j.mcm.2006.01.007

Girones, R., Carratalà, A., Calgua, B., Calvo, M., Rodriguez-Manzano, J., & Emerson, S. (2014). Chlorine inactivation of hepatitis e virus and human adenovirus 2 in water. *Journal of Water and Health*, *12*(3), 436–442. https://doi.org/10.2166/wh.2014.027

Golicha, Q.,Shetty, S.,Nasiblov, O., Hussein, A., Wainaina, E., Obonyo, M., Macharia, D., Musyoka, R. N., Abdille, H., Ope, M., Joseph, R., Kabugi, W., Kiogora, J., Said, M., Boru, W., Galgalo, T., Lowther, S. A., Juma, B., Mugoh, R.,...Burton, J.W. (2018). Cholera outbreak in Dadaab Refugee camp, Kenya — November 2015–June 2016. *Morbidity and Mortality Weekly Report*, *67*(34), 958–961. https://doi.org/10.15585/mmwr.mm6734a4

Guerrero-Latorre, L., Hundesa, A., & Girones, R. (2016). Transmission Sources of Waterborne Viruses in South Sudan Refugee Camps. *Clean - Soil, Air, Water*, *44*(7), 775–780. https://doi.org/10.1002/clen.201500358

Howard, C. M., Handzel, T., Hill, V. R., Grytdal, S. P., Blanton, C., Kamili, S., Drobeniuc, J., Hu, D., & Teshale, E. (2010). Novel Risk Factors Associated with Hepatitis E Virus Infection in a Large Outbreak in Northern Uganda: Results from a Case-Control Study and Environmental Analysis. *American Journal of Tropical Medicine and Hygiene*, *83*(5), 1170–1173. https://doi.org/10.4269/ajtmh.2010.10-0384

Lantagne, D. S. (2008). Sodium hypochlorite dosage for household and emergency water treatment. *Journal of American Water Works Association*, *100*(8), 106–114. https://doi.org//10.1002/j.1551-8833.2008.tb09704.x

Ling, C. X., & Sheng, V. S. (2008). Cost-Sensitive Learning and the Class Imbalance Problem. Encyclopedia of Machine Learning, 231–235. Retrieved from http://www.springer.com/computer/ai/book/978-0-387-30768-8%5Cnhttp://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.164.4418andrep=rep1andtype=pdf

Rashid, M.-u., George, C. M., Monira, S., Mahmud, T., Rahman, Z., Mustafiz, M., Parvin, T., Bhuyian, S. I., Zohura, F., Begum, F., Biswas, S. K., Akhter, S., Zhang, X., Sack, D., Sack, R. B., & Alam, M. (2016). Chlorination of Household Drinking Water among Cholera Patients ' Households to Prevent Transmission of Toxigenic Vibrio cholerae in Dhaka , Bangladesh : CHoBI7 Trial. *American Journal of Tropical Medicine and Hygiene*, *95*(6), 1299–1304. https://doi.org/10.4269/ajtmh.16-0420

Salama, P., Spiegel, P., Talley, L., Waldman, R., & Street, G. (2004). Lessons learned from complex emergencies over past decade. *Correspondence to : Lancet, 364*, 1801–1813.

Shultz, A., Omollo, J. O., Burke, H., Qassim, M., Ochieng, J. B., Weinberg, M., Feikin, D. R., & Breiman, R. F. (2009). Cholera outbreak in Kenyan Refugee Camp: Risk Factors for Illness and Importance of Sanitation. *American Journal of Tropical Medicine and Hygiene*, 80(4), 640–645. https://doi.org/10.4269/ajtmh.2009.80.640

Sikder, M., String, G., Kamal, Y., Farrington, M., Rahman, A. S., & Lantagne, D. (2020). Effectiveness of water chlorination programs along the emergency-transition-post-emergency continuum: Evaluations of bucket, in-line, and piped water chlorination programs in Cox's Bazar. *Water Research*, *178*, 115854. https://doi.org/10.1016/j.watres.2020.115854

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. *Journal of Water Supply: Research and Technology - AQUA, 60*(1), 51–60. https://doi.org/10.2166/aqua.2011.086

Steele, A., Clarke, B., & Watkins, O. (2008). Impact of jerry can disinfection in a camp environment - Experiences in an IDP camp in Northern Uganda. *Journal of Water and Health, 6*(4), 559–564. https://doi.org/10.2166/wh.2008.072

Swerdlow, D.L. Malenga, G., Begkoyian, G., Nyangulu, D., Toole, M., Waldman, R. J., Puhr, D. N. D., & Tauxe, R. V. (1997). Epidemic cholera among refugees in Malawi, Africa: treatment and transmission. *Epidemiology and Infection, 118*(3), 207–214. https://doi.org/https://doi.org/10.1017/S0950268896007352

Taormina, R., & Chau, K. W. (2015). Neural network river forecasting with multi-objective fully informed particle swarm optimization. *Journal of Hydroinformatics, 17*(1), 99–113. https://doi.org/10.2166/hydro.2014.116

Toth, E. (2016). Estimation of flood warning runoff thresholds in ungauged basins with asymmetric error functions. *Hydrology and Earth System Sciences, 20*(6), 2383–2394. https://doi.org/10.5194/hess-20-2383-2016

Walden, V. M., Lamond, E. A., & Field, S. A. (2005). Container contamination as a possible source of a diarrhoea outbreak in Abou Shouk camp, Darfur province, Sudan. *Disasters,*

29(3), 213–221. https://doi.org/10.1111/j.0361-3666.2005.00287.x

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

WHO. (2017). Guidelines for Drinking-water Quality (Fourth). Geneva. Retrieved from https://apps.who.int/iris/bitstream/handle/10665/254637/9789241549950-eng.pdf;jsessionid=264F04037908425EA63C6659366C9AA4?sequence=1

Zhou, Z.-H., & Liu, X.-Y. (2010). On Multi-Class Cost-Sensitive Learning. *Computational Intelligence*, *26*(3), 232–257.

Chapter 2 Datasets and Data Analysis

2.1 Study Site Descriptions

This research used data collected from six refugee settlements between 2013 and 2020. Broadly, these sites have been subdivided into two categories: development sites (South Sudan, Jordan, Rwanda) whose data was used in the initial development of the SWOT project, and implementation sites (Bangladesh, Tanzania, Rwanda), sites whose data was collected as part of the SWOT field trials. For this research, the development sites were used in the proof-of-concept study included in Chapter 3, and data from the implementation sites were used in the cost sensitive learning and multi-objective studies (Chapters 4 and 5 respectively). The development and implementation site locations are shown on Figure 2-1 below and are discussed further in Section 2.2. The following subsections provide a description of each study site.

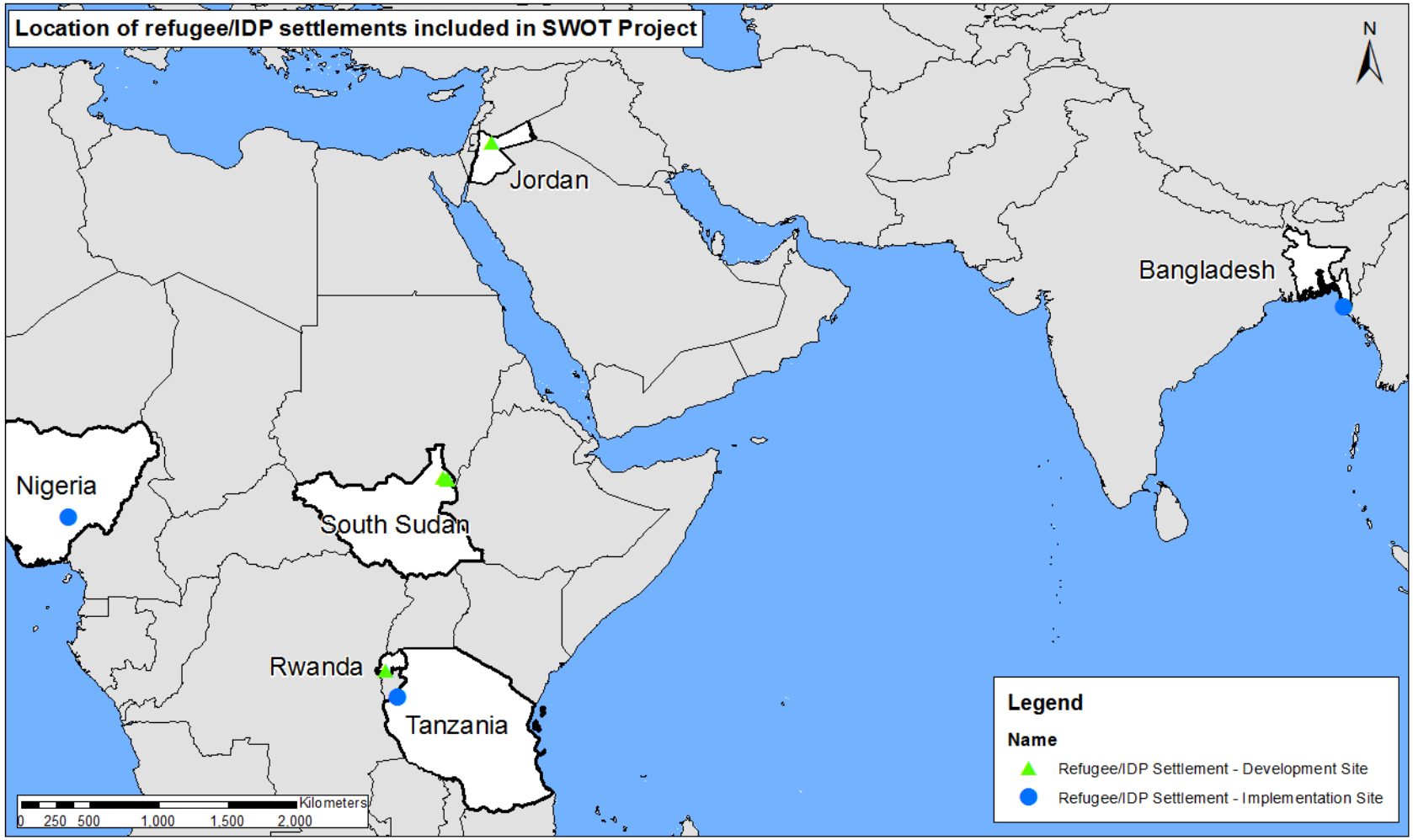

Figure 2-1: SWOT site locations. Development sites (2013-2015) include South Sudan, Jordan, and Rwanda. Implementation sites (2019-Present) include Bangladesh, Tanzania, and Nigeria.

2.1.1 South Sudan

The South Sudan site was made up of three separate refugee settlements: Batil, Jamam, and Gendrassa, all located in Maban County in Upper Nile State of South Sudan. Data was collected by MSF at the South Sudan site beginning in March 2013 and continued for six weeks until April 2013. Unlike the other development sites, data collected in South Sudan was part of expanded routine water quality monitoring in the aftermath of a Hepatitis E outbreak in the three settlements. The population of the three settlements included in the South Sudan site ranged from over 15,000 to over 37,000, with a combined population of over 68,000. At all three settlements, drinking water was obtained via groundwater boreholes and the only treatment was inline chlorination with calcium hypochlorite.

2.1.2 Jordan

The Jordan datasets was collected by UNHCR from the Azraq settlement, located near the town of Azraq in eastern Jordan. Unlike the other study sites included in this research, two datasets were obtained from Jordan: the first was collected between July and August of 2014 and the second was obtained nine months later between March and April of 2015. There was a substantial change in population between these two periods, with the population in 2014 being recorded as 7,470 and the population in 2015 recorded as 14,797, and the temperature also varied substantially between the 2014 dataset which was collected in the summer where the temperatures experienced ranged from 27 to 43 degrees Celsius, and the 2015 dataset which was collected in the late winter/early spring where temperatures ranged from 15 to 29 degrees Celsius. Drinking water at the Azraq site was obtained from groundwater boreholes which was then treated via reverse osmosis due to salinity concerns, and then chlorinated via inline chlorination with chlorine gas. Unlike the other development sites, drinking water was then trucked to site before piped distribution.

2.1.3 Rwanda

The Rwanda dataset was collected by UNHCR in the Kigeme refugee settlement in the Nyamagabe District of Rwanda. Data was collected at this site from June to July of 2015. The site population at the time of data collection was over 18,000. Drinking water at this site was obtained from a local stream, making this the only dataset with a surface water source. Water

treatment included conventional flocculation, sedimentation, and filtration and disinfection was performed with calcium hypochlorite.

2.1.4 Bangladesh

The Bangladesh dataset was collected by MSF and UNHCR from the Kutapalong-Balukali Extension Site 1 refugee settlement located in Cox's Bazaar in Bangladesh. This was the first of the implementation field trials which was supported by researchers from the Dahdaleh Institute for Global Health Research (DIGHR) at York University. The data collection for this site took place over a period of 6 months from June to December of 2019. Thus, the Bangladesh dataset covers the longest period of time of all datasets used in this research. The population of the site was 80,000 at the time of data collection. Drinking water at this site was obtained from groundwater boreholes and disinfected with inline chlorination with calcium hypochlorite.

2.1.5 Tanzania

The Tanzania dataset was collected by UNHCR and the Norwegian Refugee Council from the Nyaragusu refugee settlement in the Kigoma region. Data collection at this site occurred from December 2019 to January 2020, and data collection remains ongoing. The population of the Nyaragusu settlement during this period was 132,000. Drinking water at this site was obtained from groundwater boreholes.

2.1.6 Nigeria

The Nigeria dataset was collected by MSF from the Mbawa IDP settlement located in Benue State. The data used from this site was collected between February and July, 2020, though data collection remains ongoing. The population of the Mbawa settlement was 6,600 during the period of data collection. Drinking water in the Mbawa settlement is obtained from groundwater boreholes and is treated via batch chlorination in tanks using calcium hypochlorite.

2.2 Description of Data Sets

The dataset for each site includes FRC as well as other water quality parameters which are routinely collected in humanitarian water systems operation. For the development sites this includes electrical conductivity (EC), water temperature, turbidity, and pH, whereas at the implementation sites the only additional water quality variables used were EC and water temperature. Data was collected using paired sampling whereby the same unit of water was sampled at multiple points along the post-distribution water supply chain (recall the point-of-

distribution and point-of-consumption shown in Figure 1-1). For the development sites water quality was measured at four points: from the tap at the point of distribution, in the container immediately after collection, in the container immediately after transport to the dwelling, and after a follow-up period of storage in the household. For the implementation sites, data collection was streamlined to only include measurement from the tapstand and a follow-up during household storage. Thus, all of the studies in Chapters 3, 4, and 5 only consider data collected at these two points in the post-distribution water supply chain.

Table 2-1 summarizes the water quality variables that were collected and used in the models for each site. Additional variables were collected at each site but are not included in Table 2-1 either because they are not commonly recommended for collection in refugee and IDP settlements, or because they were not available at all sites included in a specific modelling study. The latter was particularly common for our Bangladesh site where more water quality variables were available than at the Nigeria and Tanzania sites, so for consistency, the same variables were used for the modelling at each site. Thus, turbidity and pH were not included in the Bangladesh models, despite these variables being available.

Table 2-1: Water quality parameters collected

Parameter	**Sites where Parameter was Measured**	**Collection Method**
Free residual chlorine	All	Colorimetric method using Palintest PTH 7091 compact chlorometer and Wagtech 7100 photometer with Palintest DPD1/DPD3 reagents (Palintest Ltd., Tyne and Wear, UK) Pool testers (Nigeria)
Turbidity	South Sudan, Jordan Rwanda, Bangladesh	Nephelometric method using Palintest PTH 090 compact turbidimeter (Palintest Ltd., Tyne and Wear, UK)
Water temperature	All	All recorded via potentiometric method using Eijkelkamp 18.21 multimeter (Eijkelkamp Agrisearch Equipment, Giesbeek, Netherlands), Hanna Instruments HI 98311 EC/TDS/temperature multi-meter (HANNA Instruments, Woonsocket, RI, USA), or Hach sensION+ multi-meter (Hach Instruments, USA). Note: instrument used varied between sites; multiple instruments were not used at the same site.
Conductivity	All	
pH	South Sudan, Jordan Rwanda, Bangladesh	

2.3 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 the 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 the 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).

2.4 Preliminary Data Analysis

A preliminary data analysis was undertaken to achieve four goals. First, it was used to gain an understanding of the input and output variables through visualization and by providing summary statistics (number of observations, mean, median, standard deviation). This provided a reference for understanding the variables collected. Second, this analysis was used to determine if there was a statistically significant difference between the point-of-distribution and point-of-consumption FRC concentrations. This serves to reinforce the need for modelling FRC beyond the point-of-distribution up to the point-of-consumption. Third, this analysis was used to check to see if the point-of-consumption FRC follows a common distribution at all sites. This would indicate if we could approach probabilistic modelling by estimating distribution parameters, or if we need to use a non-parametric approach. Finally, this analysis provides visualizations of the relationship between each input variable and point-of-consumption FRC which provides a visual understanding of the relationship between the input variables used in this study and point-of-consumption FRC.

Figures 2-2, 2-3, 2-4, and 2-5 show histograms of the input and output variables for the South Sudan, Jordan (2014), Jordan (2015), and Rwanda datasets, respectively, and Figures 2-6, 2-7, and 2-8 show input variable histograms for the implementation sites Bangladesh, Tanzania, and

Nigeria, respectively). These figures also show boxplots comparing the point-of-consumption and point-of-distribution FRC concentrations. The data cleaning rules used to prepare these figures are documented in Chapter 3 for the development sites and in Chapter 4 for the implementation sites. These figures show first that the distributions of FRC concentrations between the point-of-distribution and point-of-consumption are very different and the point-of-consumption FRC distributions were all right-skewed to varying degrees, except for the Jordan (2015) site, indicating a very high frequency of observations with low FRC. This is visually also confirmed from the boxplots which show that the point-of-consumption FRC distributions tend to be grouped around lower values of household FRC than the point-of-distribution FRC. These figures also show that the range of observed water quality measurements vary substantially from site to site. This is critical as it highlights potential challenges in training a global FRC decay model: the water quality parameters from one site to the next tend to be very different. Finally, these figures show that the distribution of typical follow up times for the point-of-distribution measurement vary from site to site. These histograms loosely reflects typical storage times, but it should be noted that the follow up time was also impacted by logistics of sample collection and water availability. For all sites but South Sudan and Nigeria there tend to be two clusters of storage times: a long storage time cluster (up to around 24 hours) and a short storage time cluster (up to around 10 to 15 hours). In Bangladesh these clusters tend to occur around 15 hours and between 5-10 hours but the bimodal distribution is still present.

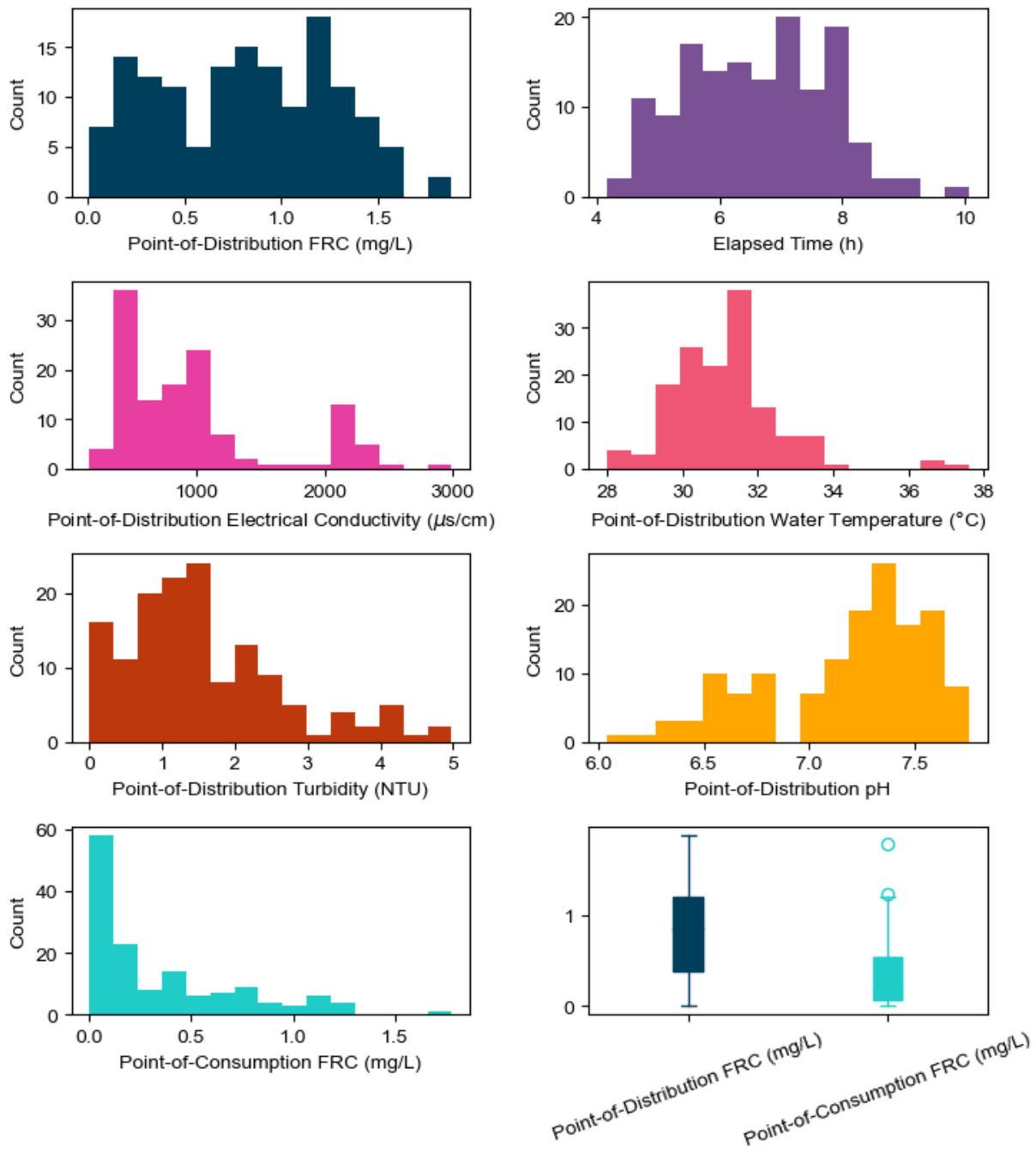

Figure 2-2: South Sudan input and output variable histograms and boxplot comparison of point-of-distribution and point-of-consumption FRC. Boxes show the interquartile range, whiskers show the 95th percentile range, and circles represent outliers beyond the 95th percentile range.

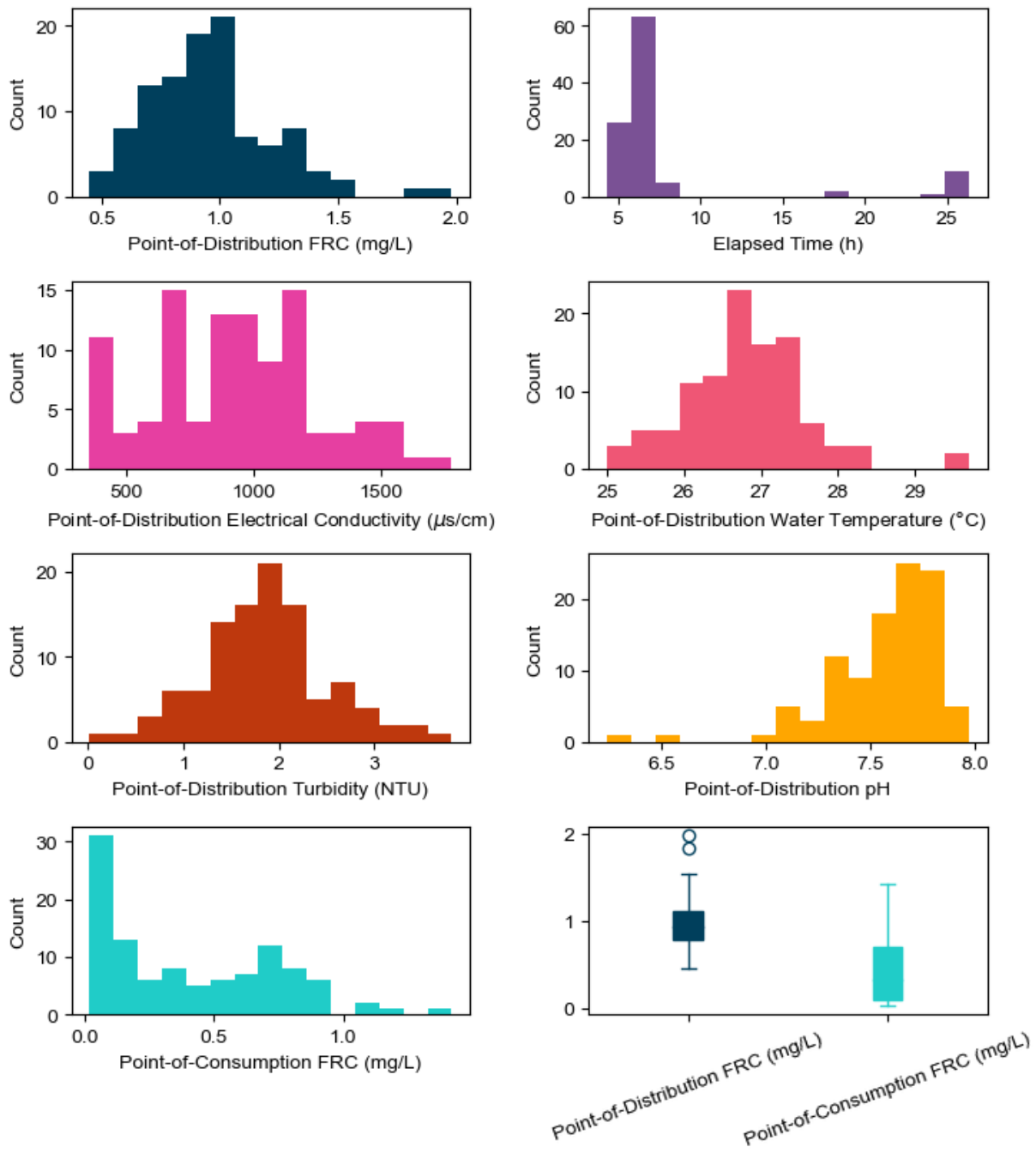

Figure 2-3:Jordan (2014) input and output variable histograms and boxplot comparison of point-of-distribution and point-of-consumption FRC. Boxes show the interquartile range, whiskers show the 95th percentile range, and circles represent outliers beyond the 95th percentile range.

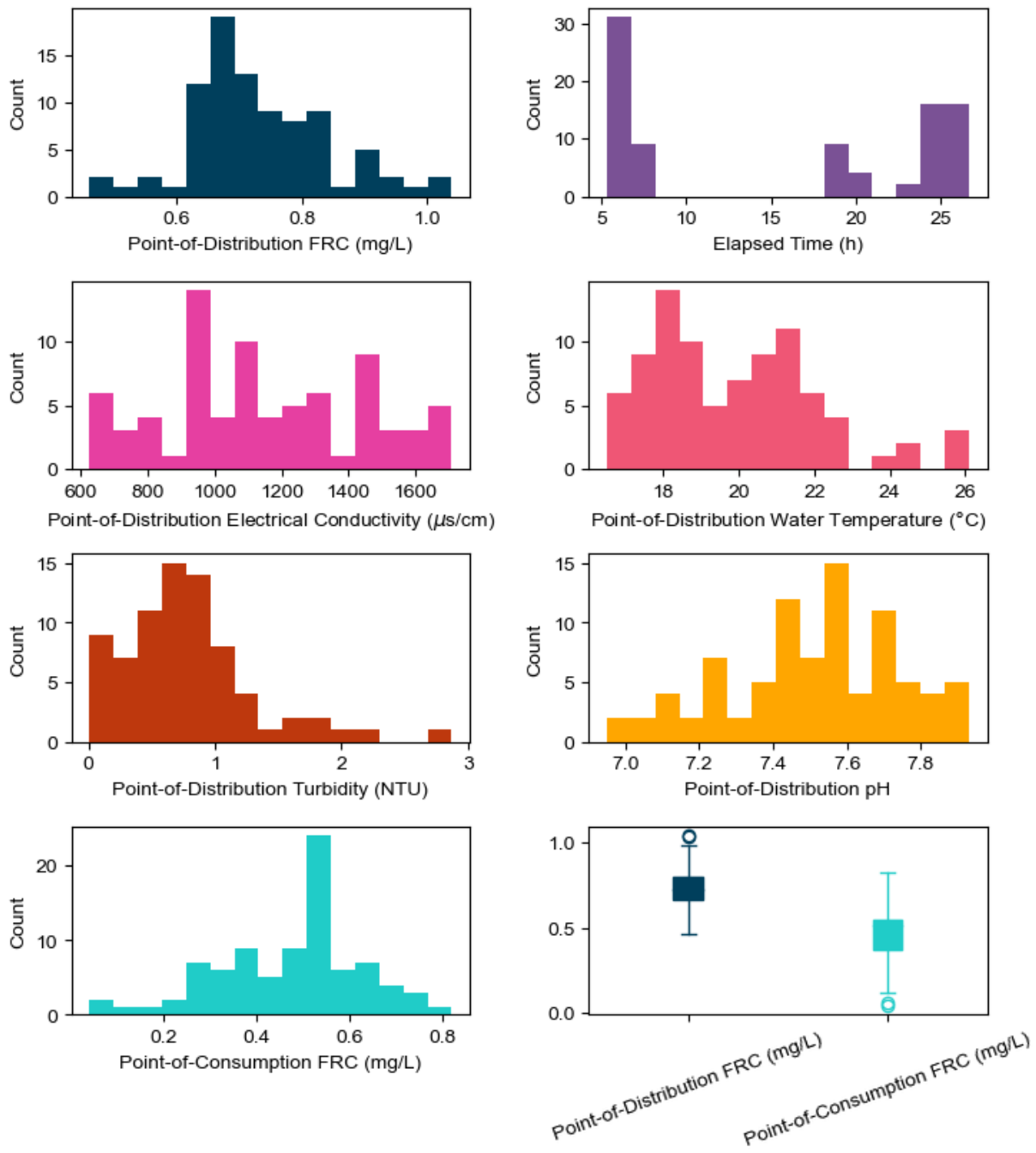

Figure 2-4:Jordan (2015) input and output variable histograms and boxplot comparison of point-of-distribution and point-of-consumption FRC. Boxes show the interquartile range, whiskers show the 95th percentile range, and circles represent outliers beyond the 95th percentile range.

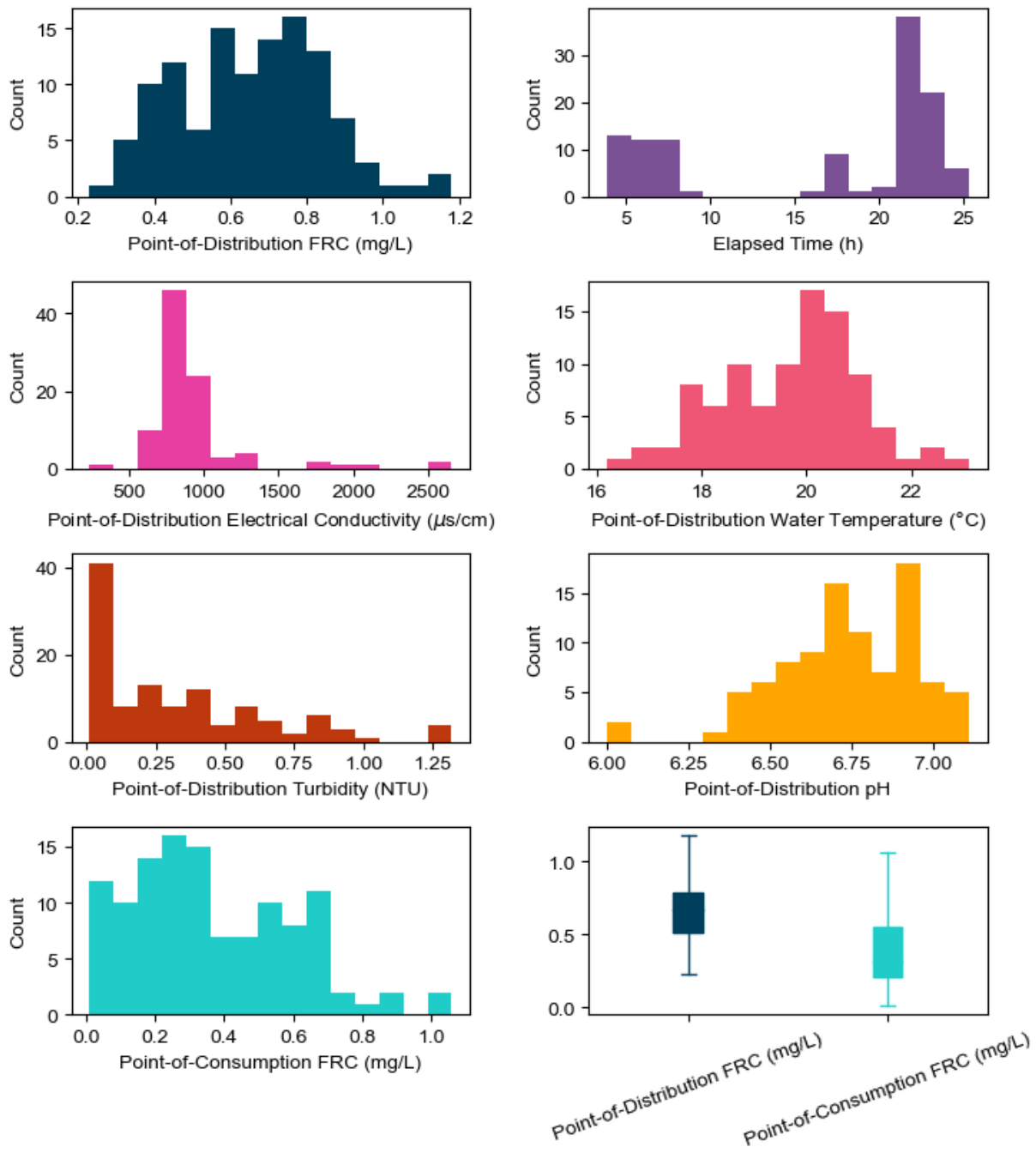

Figure 2-5: Rwanda input and output variable histograms and boxplot comparison of point-of-distribution and point-of-consumption FRC. Boxes show the interquartile range, whiskers show the 95th percentile range, and circles represent outliers beyond the 95th percentile range.

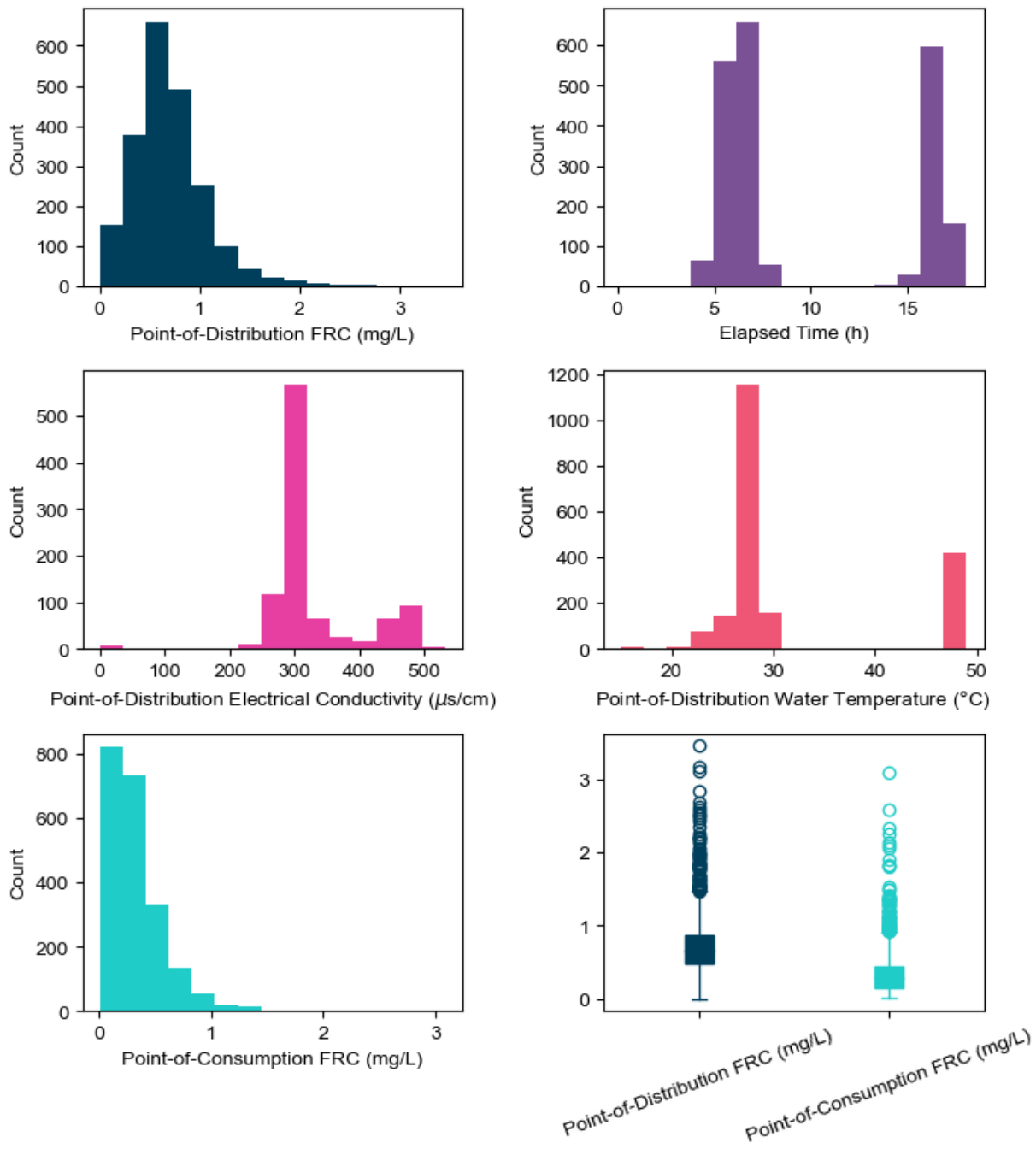

Figure 2-6: Bangladesh input and output variable histograms and boxplot comparison of point-of-distribution and point-of-consumption FRC Boxes show the interquartile range, whiskers

show the 95th percentile range, and circles represent outliers beyond the 95th percentile range.

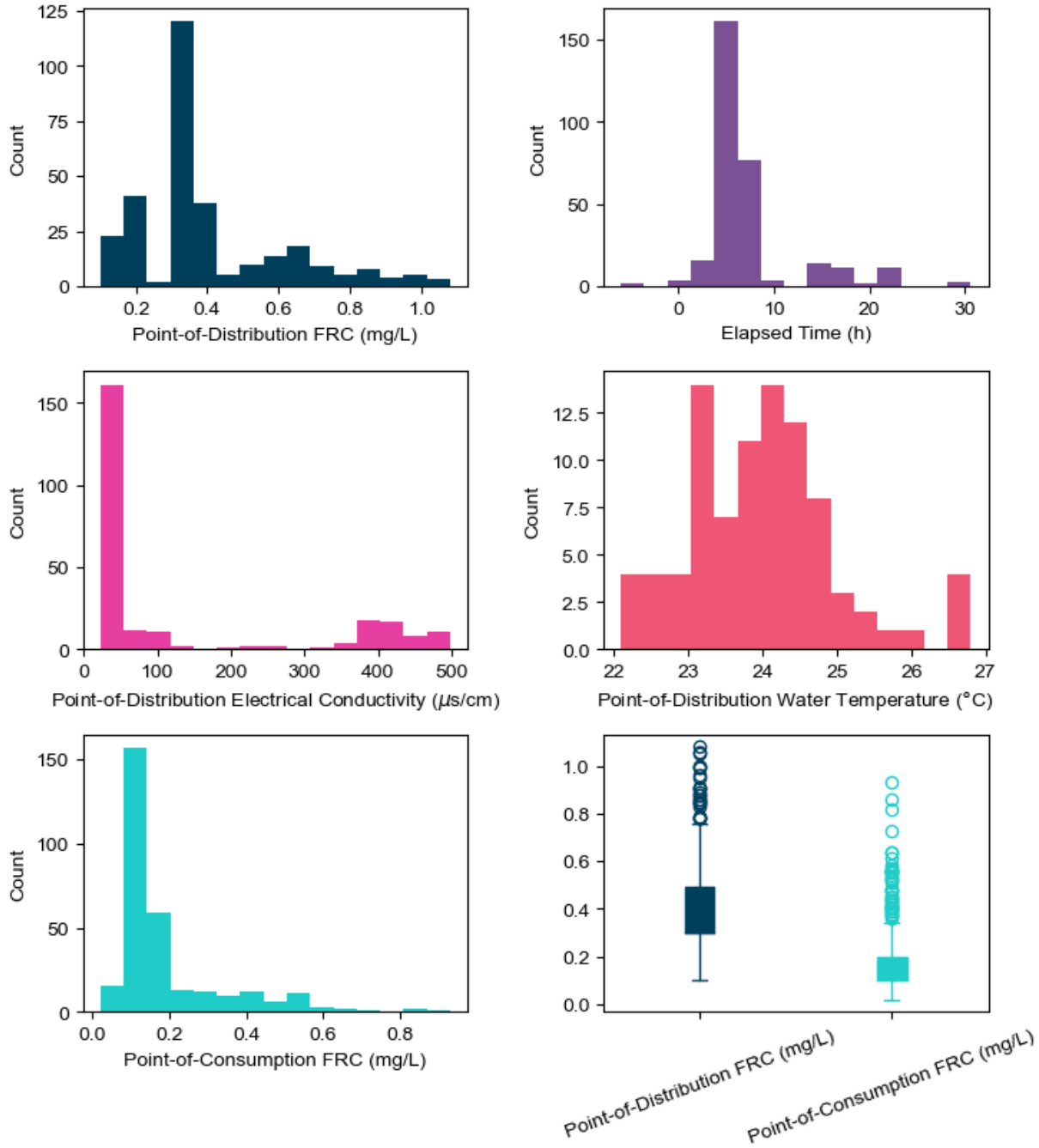

Figure 2-7: Tanzania input and output variable histograms and boxplot comparison of point-of-distribution and point-of-consumption FRC Boxes show the interquartile range, whiskers show

the 95th percentile range, and circles represent outliers beyond the 95th percentile range.

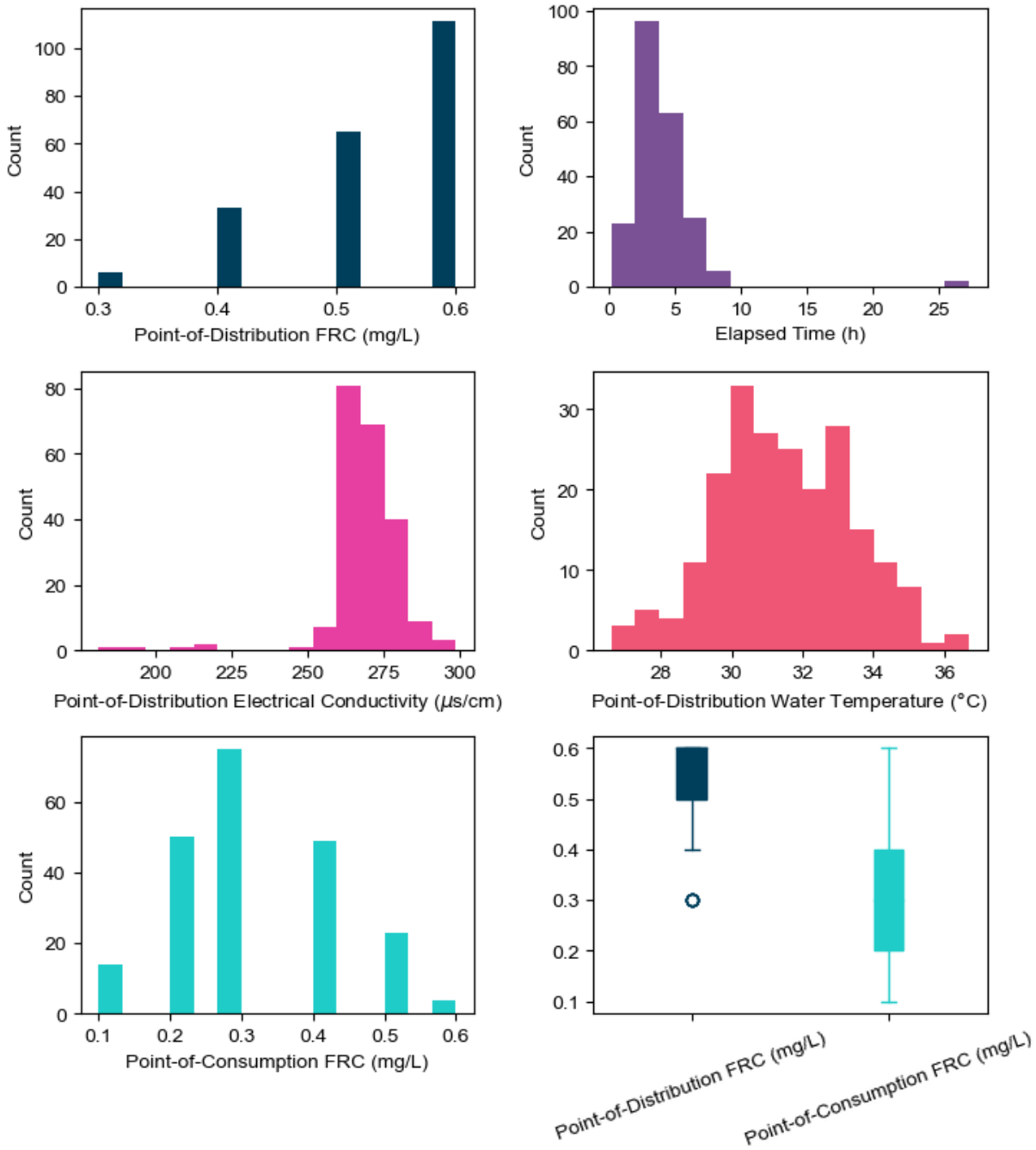

Figure 2-8: Nigeria input and output variable histograms and boxplot comparison of point-of-distribution and point-of-consumption FRC. Boxes show the interquartile range, whiskers show the 95th percentile range, and circles represent outliers beyond the 95th percentile range.

As stated above, a primary observation from Figures 2-2 through 2-5 is that the mean point-of-consumption FRC appear to be much lower than the point-of-distribution FRC. We confirmed this using a Kruskal-Wallis location test, which is a non-parametric test of the similarity of means between two distributions. The null hypothesis of the Kruskal-Wallis test is that the two samples used have the same mean, and at low *p*-values (typically less than 0.05) we reject this hypothesis (i.e., a lower *p*-value suggests a lower probability that the means are drawn from the same distribution). When comparing the means of the distributions with the Kruskal-Wallis test, we obtained *p*-values ranging between $1x10^{-17}$ to $1.1x10^{-287}$. This shows that it is improbable that the means of the point-of-distribution and point-of-consumption FRC concentrations are the same. This strongly supports the need to develop FRC targets that account for post-distribution FRC decay as the dissimilarity of these means strongly supports that FRC decay continues after chlorine leaves the piped distribution system. To further reinforce this, we show the mean and standard deviation of the point-of-distribution and point-of-consumption FRC concentrations in Table 2-2. This table shows that across sites there is a substantial drop in the mean FRC from the point-of-distribution to the point-of-consumption, a drop that is shown to be significant by the Kruskal-Wallis test (also shown in Table 2-2). Table 2-2 also shows that the Nigeria dataset has the smallest standard deviations, likely due to the clustering of observed values seen in Figure 2-8. This apparent clustering is due to FRC data being collected on site with pool testers which have a lower resolution than the colorimetric methods used at the other sites.

Table 2-2: Comparison of mean and standard deviation from point-of-distribution to point-of-consumption for historic sites

Site	Point-of-Distribution		Point-of-Consumption		Kruskal-Wallis Location Test	
	Mean	Standard Deviation	Mean	Standard Deviation	Test Statistic	*p*-value
South Sudan	0.82	0.45	0.35	0.37	77	$1.5x10^{-18}$
Jordan (2014)	0.96	0.27	0.41	0.33	100	$1.4x10^{-23}$
Jordan (2015)	0.73	0.11	0.47	0.15	95	$2.3x10^{-22}$
Rwanda	0.65	0.19	0.37	0.23	73	$1.1x10^{-17}$
Bangladesh	0.71	0.38	0.34	0.28	1314	$1.1x10^{-287}$
Tanzania	0.39	0.22	0.19	0.15	187	$1.4x10^{-42}$
Nigeria	0.53	0.08	0.31	0.11	239	$5.4x10^{-54}$

Figures 2-2 through 2-8 also showed that the point-of-consumption FRC followed a right-skewed (positive) distribution, and especially at South Sudan and for Jordan (2014), appeared to potentially follow an exponential distribution. To test if the point-of-consumption FRC followed any particular distribution, we used the Scientific Python (SciPy) package in Python (Virtanen et al., 2020) to test 97 continuous distributions to see if they matched the observed data. We performed this testing by first fitting the distribution parameters to the point-of-consumption FRC data and then comparing this distribution against the observed data using a Kolmogorov-Smirnov test, which is a non-parametric test to determine if two samples are drawn from the same distribution. The null hypothesis of the Kolmogorov-Smirnov test is that the two samples are from the same population, and we reject the null hypothesis at low *p*-values (typically less than 0.05). In only one case was the *p*-value greater than 0.05: using the Lévy distribution for the Jordan (2015) dataset, where the *p*-value was 0.10. Based on this, we conclude that there is not a single representative distribution for representing point-of-consumption FRC, and instead we should prioritize non-parametric probabilistic estimation methods.

To investigate the impact of different input variables on the point-of-consumption FRC, Figures 2-9, 2-10, 2-11, and 2-12 plot point-of-consumption FRC against each input variable used from the historic sites for South Sudan, Jordan (2014), Jordan (2015), and Rwanda, respectively, and Figures 2-13, 2-14, and 2-15 plot these figures for the implementation sites (Bangladesh, Tanzania, and Nigeria, respectively). These figures also show the linear regression line of best fit for each input variable. From these figures, we see that at all sites except Jordan (2015), there is a strong positive relationship between point-of-distribution FRC and point-of-consumption FRC, meaning that if more FRC is available at the point-of-distribution, there will be more FRC available at the point-of-consumption. While this is intuitive, it is useful to know that there is such a strong relationship between these two variables. When considering the other water quality variables (electrical conductivity, water temperature, turbidity, and pH for the historical sites, electrical conductivity and water temperature for the implementation sites), we see that for the most part, there are negative trends between these variables and point-of-consumption FRC. We also see that there are some sites, notably Tanzania, where these trends are reversed, indicating that higher conductivity and water temperature values actually correspond to higher point-of-consumption FRC concentrations. This may indicate that there are additional variables confounded with these input variables that are not included in the dataset. This highlights a key benefit of the ANN approach: ANNs learn the behaviour from the underlying data on a site-by site basis, so the models are not restricted by assumed relationships between any of the input variables and point-of-consumption FRC, and instead they will reproduce the existing relationships as observed in the data.

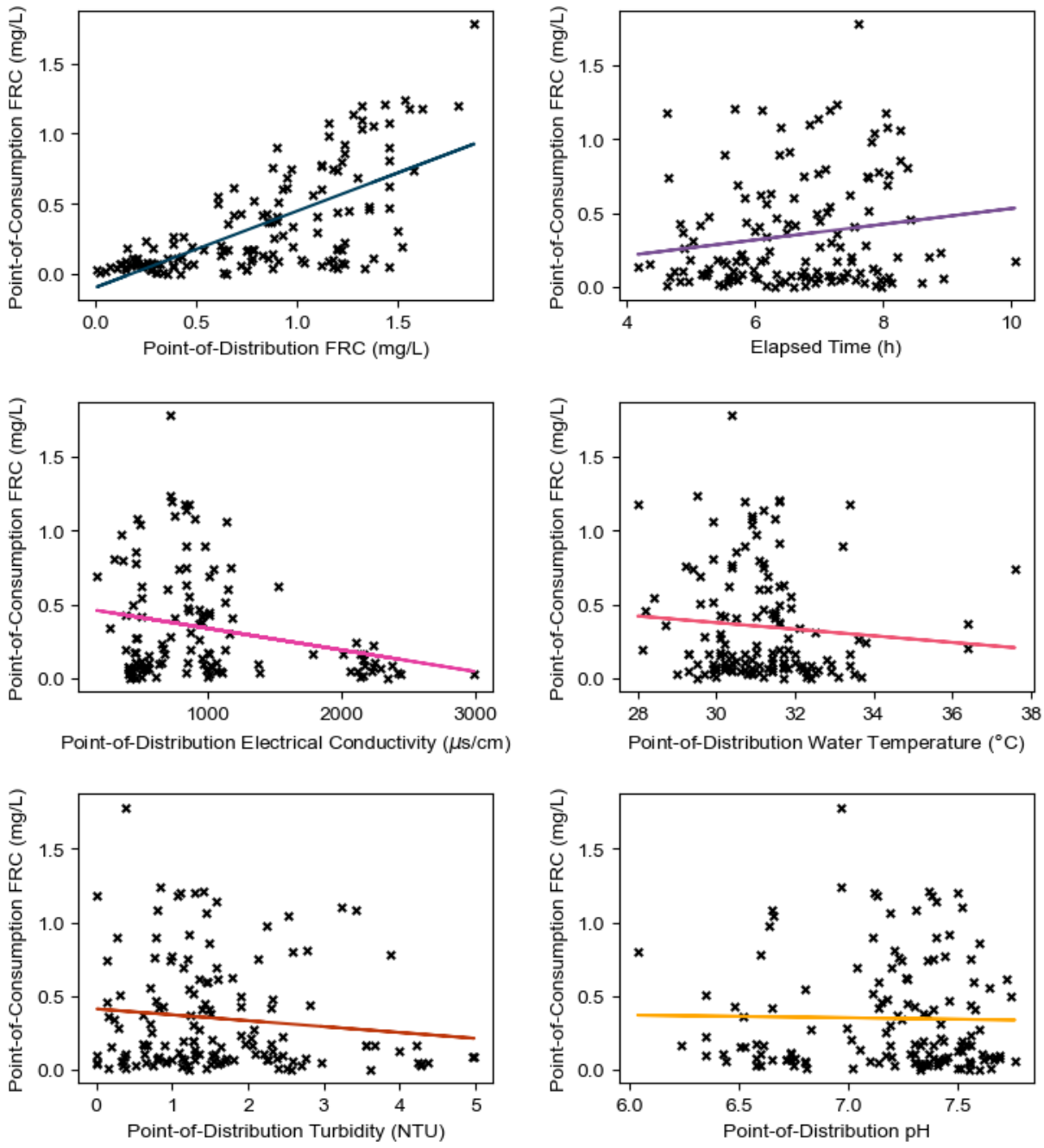

Figure 2-9: South Sudan trends between input variables and point-of-consumption FRC. All variables other than elapsed time and point-of-distribution FRC have negative trends with point-of-consumption FRC.

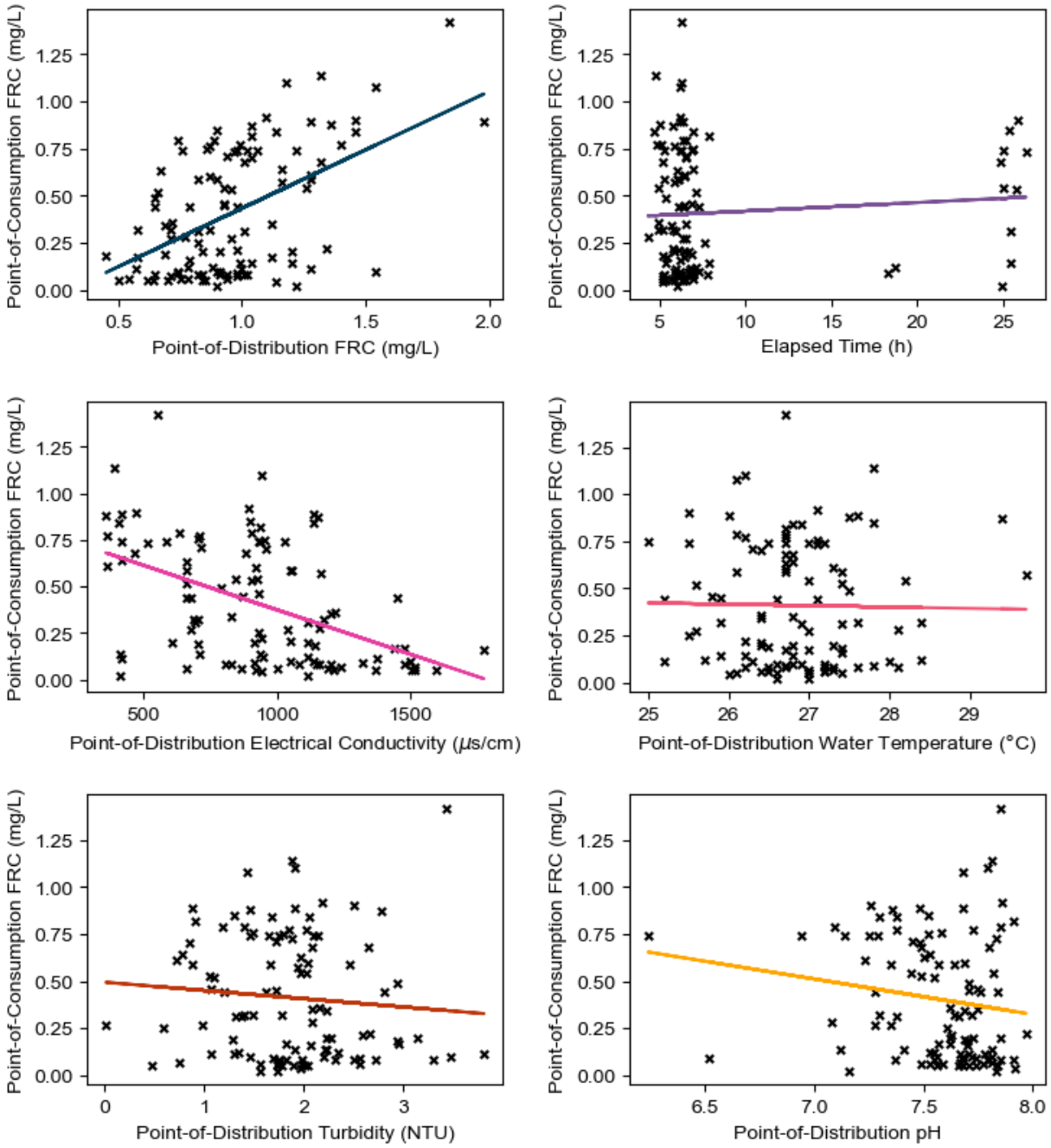

Figure 2-10: Jordan 2014 trends between input variables and point-of-consumption FRC. All variables other than elapsed time and point-of-distribution FRC have negative trends with point-of-consumption FRC.

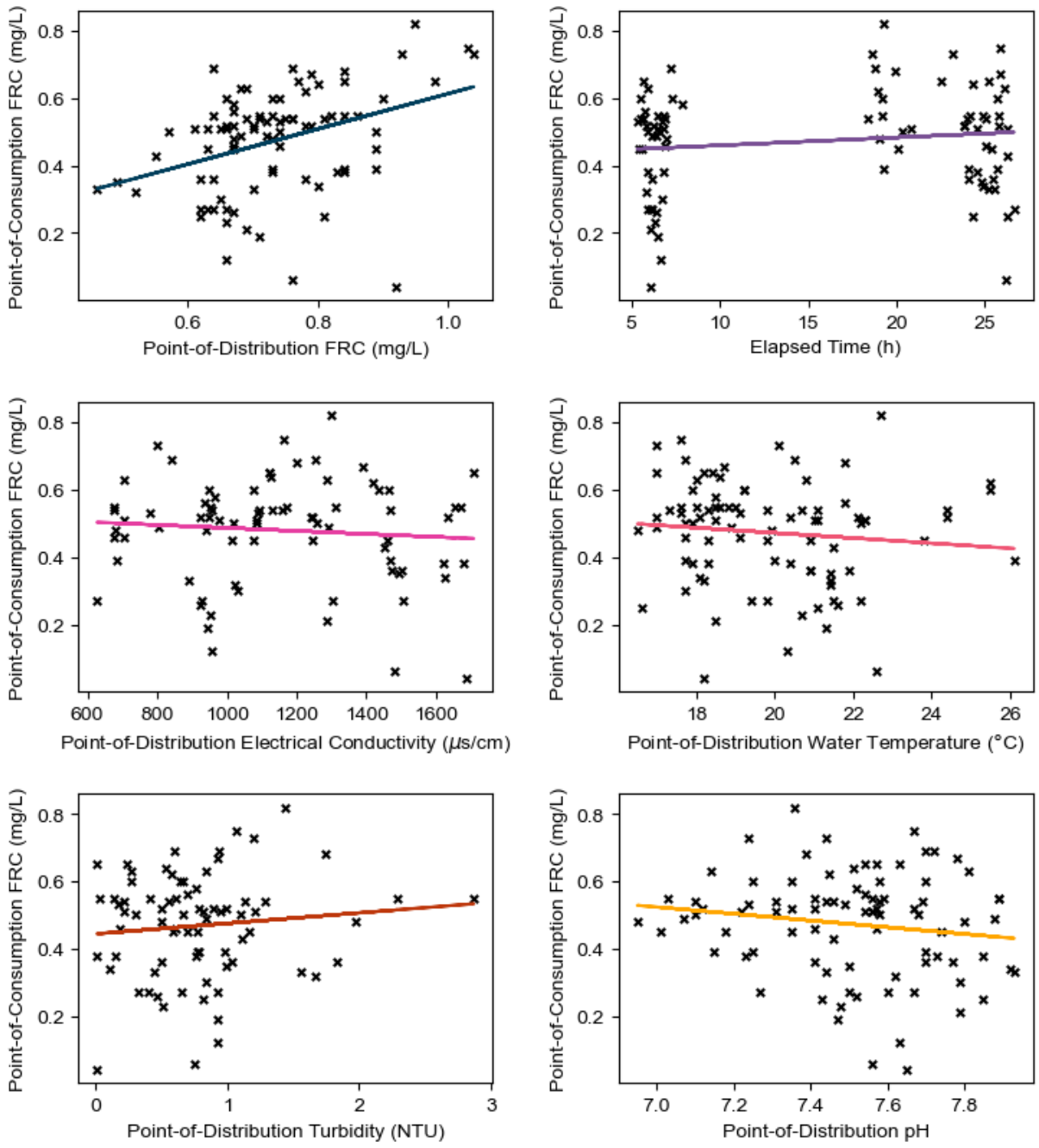

Figure 2-11: Jordan 2015 trends between input variables and point-of-consumption FRC. All variables other than elapsed time, turbidity, and point-of-distribution FRC have negative trends with point-of-consumption FRC.

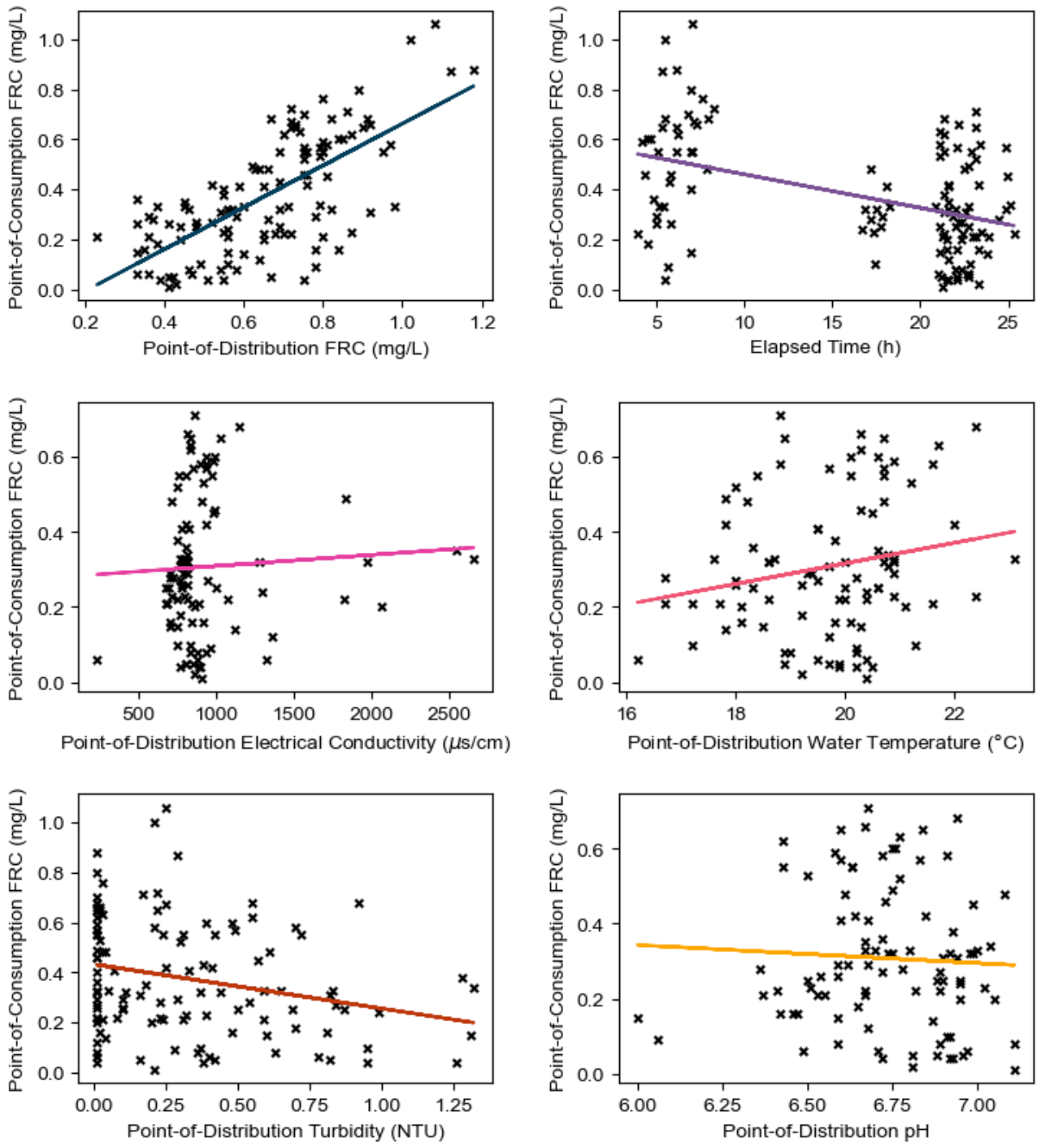

Figure 2-12: Rwanda trends between input variables and point-of-consumption FRC. All variables other than turbidity and pH have positive trends with point-of-consumption FRC.

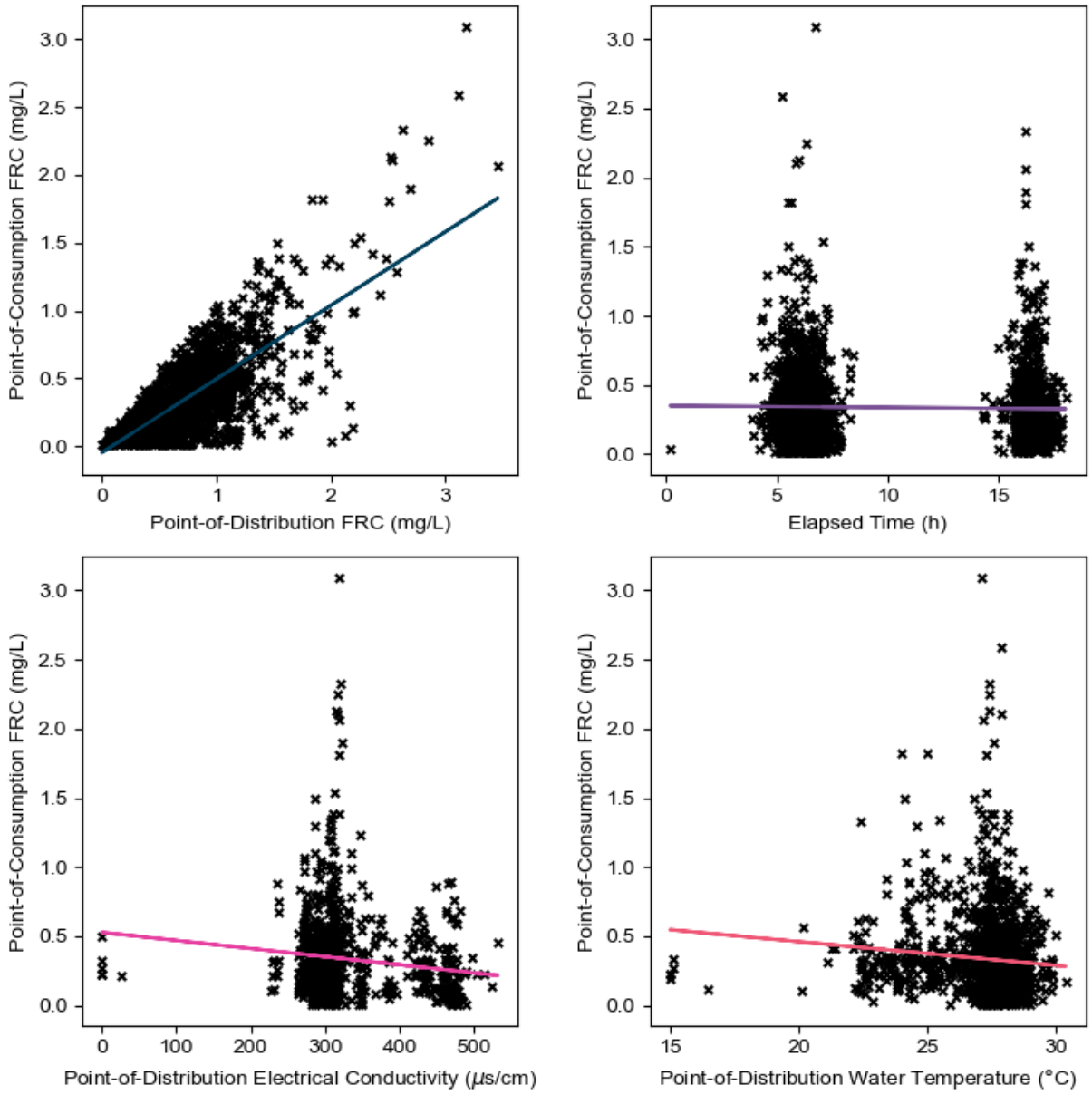

Figure 2-13: Bangladesh trends between input variables and point-of-consumption FRC. All variables other than point-of-distribution FRC have negative trends with point-of-consumption FRC.

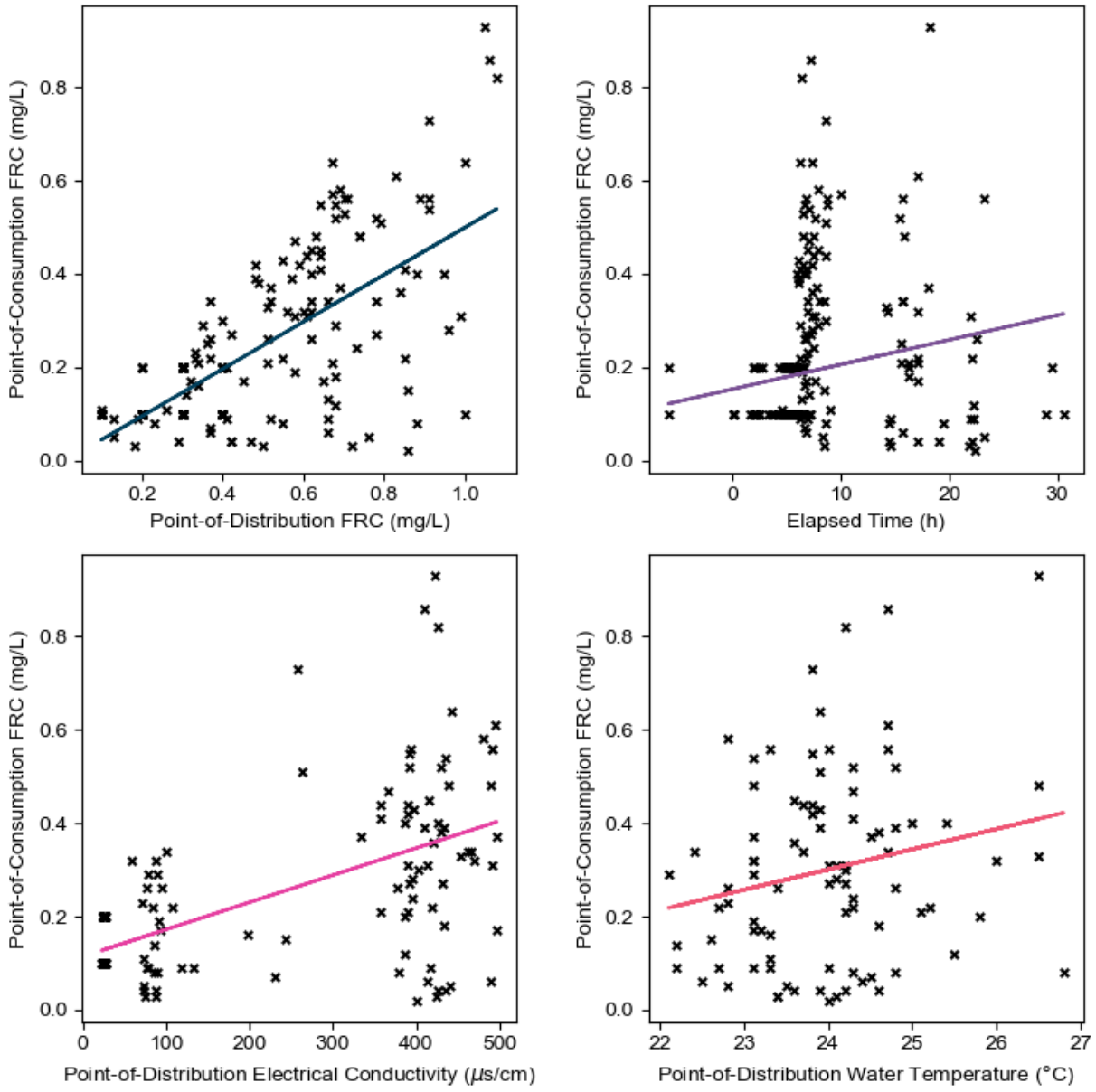

Figure 2-14: Tanzania trends between input variables and point-of-consumption FRC. All variables have positive trends with point-of-consumption FRC.

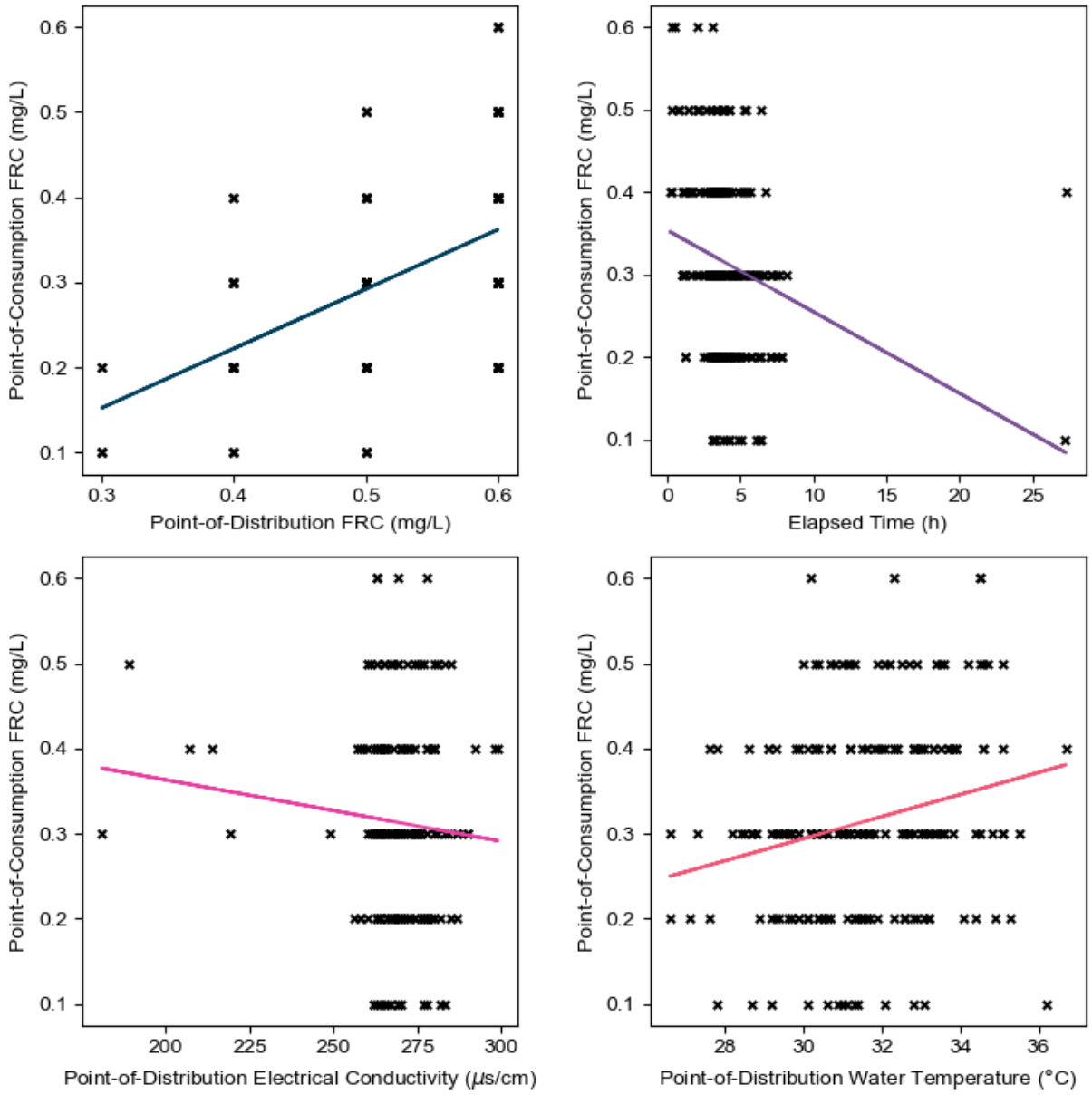

Figure 2-15: Nigeria trends between input variables and point-of-consumption FRC. Point-of-distribution FRC and water temperature both have positive trends with point-of-consumption FRC.

2.5 References

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.,...van Mulbregt, P. & SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods, 17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Chapter 3 Proof-of-Concept Study

3.1 Chapter Preamble

This chapter presents the initial proof-of-concept study used to validate the ANN ensemble approach for modelling post-distribution FRC. A modified version of this chapter was published in *npj Clean Water* on June 25, 2021. Note that, as a Nature journal, *npj Clean Water* includes the Methods after the Results and Discussion sections, which is maintained in Chapter 3, though Chapters 4 and 5 retain a more conventional structure with Methods presented before the Results and Discussion. The citation for the original article is:

De Santi, M., Khan, U.T., Arnold, M. *et al.* Forecasting point-of-consumption chlorine residual in refugee settlements using ensembles of artificial neural networks. *npj Clean Water* **4,** 35 (2021). https://doi.org/10.1038/s41545-021-00125-2

The Supplementary Information published with this article are included in Appendix A.

This proof-of-concept study primarily explored the first two objectives of this thesis: first, it sought to generate and evaluate the effectiveness ANN ensembles for forecasting point-of-consumption FRC and it presented a method for generating risk-based FRC targets from the ensemble forecasts. This study also examined the effectiveness of kernel dressing as a simple post-processing method for improving the reliability of these forecasts. This proof-of-concept study was in part motivated by an exploratory analysis (c.f. Appendix B) which used a deterministic approach which we found was not appropriate for producing accurate risk-based FRC targets. The study presented in this chapter found that the ANN ensembles were able to generate accurate FRC guidance, with the predicted risk for three out of four sites agreeing very well with the observed data. However, we also found that the models tended to be underdispersed, even after post-processing. This may have been due to the use of mean squared error (MSE) for training the ensemble base learners. Thus, one of the most important take-aways from this study was the need to address underdispersion during the training of the ANNs in the ensemble models, instead of addressing underdispersion after the fact with post-processing.

These findings were incorporated into the version 2 ANN of the SWOT ANN analytics which will be released in August 2021**.** A white paper will also be publisedh with the launch of the SWOT version 2 that will summarize the open source version of these analytics as well as

updates that were made to address operational concerns from the proof-of-concept study, in particular investigating additional post-processing approaches as well as developing approaches to better incorporate time in the model. This white paper is included in Appendix C.

As the lead author I was responsible for the conception of the study presented in this chapter, as well as all model development and analysis. I was also responsible for preparing the manuscript. Dr. Usman Khan was responsible for modelling supervision and manuscript preparation. Dr. Syed Imran Ali was responsible for data collection, coordination of partners, securing funding, and manuscript review. Jean-François Fesselet was responsible for coordination of partners, securing funding, and manuscript review. Matt Arnold was responsible for coordination of partners and manuscript review.

3.2 Abstract

Waterborne illnesses are a leading health concern in refugee and internally displaced person (IDP) settlements where waterborne pathogens often spread through household recontamination of stored water. Ensuring sufficient chlorine residual is important for protecting drinking water against recontamination and ensuring water remains safe up to the point-of-consumption. We used ensembles of artificial neural networks (ANNs) to probabilistically forecast the point-of-consumption free residual chlorine (FRC) concentration and to develop point-of-distribution FRC targets based on the risk of insufficient FRC at the point-of consumption. We built ANN ensemble models using data from three refugee settlements and found that the risk-based FRC targets generated by the ensemble models were consistent with an empirical water safety evaluation, indicating that the models accurately predicted the risk of low point-of-consumption FRC despite all ensemble forecasts being underdispersed even after post-processing. This demonstrates the usefulness of ANN ensembles for generating risk-based point-of-distribution FRC targets to ensure safe drinking water in humanitarian operations.

3.3 Introduction

Waterborne diseases are a leading cause of morbidity and mortality in refugee and internally displaced person (IDP) settlements, so providing safe drinking water is critical for ensuring the health of displaced persons during humanitarian responses (Connolly et al., 2004; Cronin et al., 2008; Salama et al., 2004; Toole & Waldman, 1997). Recontamination of previously safe drinking water remains a major challenge in these settlements, having been identified as contributing factor in outbreaks of cholera, hepatitis E, and shigellosis in refugee and IDP settlements in Kenya (Golicha et al., 2018; Shultz et al., 2009), Malawi (Swerdlow et al., 1997), Sudan (Walden et al., 2005), South Sudan (Ali et al., 2015; Guerrero-Latorre et al., 2016), and Uganda (Howard et al., 2010; Steele et al., 2008). Residual chlorine protects against household recontamination by inactivating waterborne pathogens as they are introduced. According to globally used guidelines and past studies, this requires a free residual chlorine (FRC) concentration of at least 0.2 mg/L to be maintained up to the point-of-consumption (CDC, 2012; Girones et al., 2014; Lantagne, 2008; Rashid et al., 2016; Sikder et al., 2020; WHO, 2011). Current humanitarian drinking water quality guidelines, such as the sector-standard Sphere Handbook, do not ensure sufficient chlorine residual at the point-of-consumption as they fail to account for FRC decay during the *post-distribution* period, which begins when treated water

leaves the central water distribution point (tap stand) and ends at the point-of-consumption. Thus, the post-distribution period includes collection, transport, and household storage, as depicted in Figure 3-1 which shows the post-distribution period in the context of the overall water treatment and distribution system for the sites included in this study.

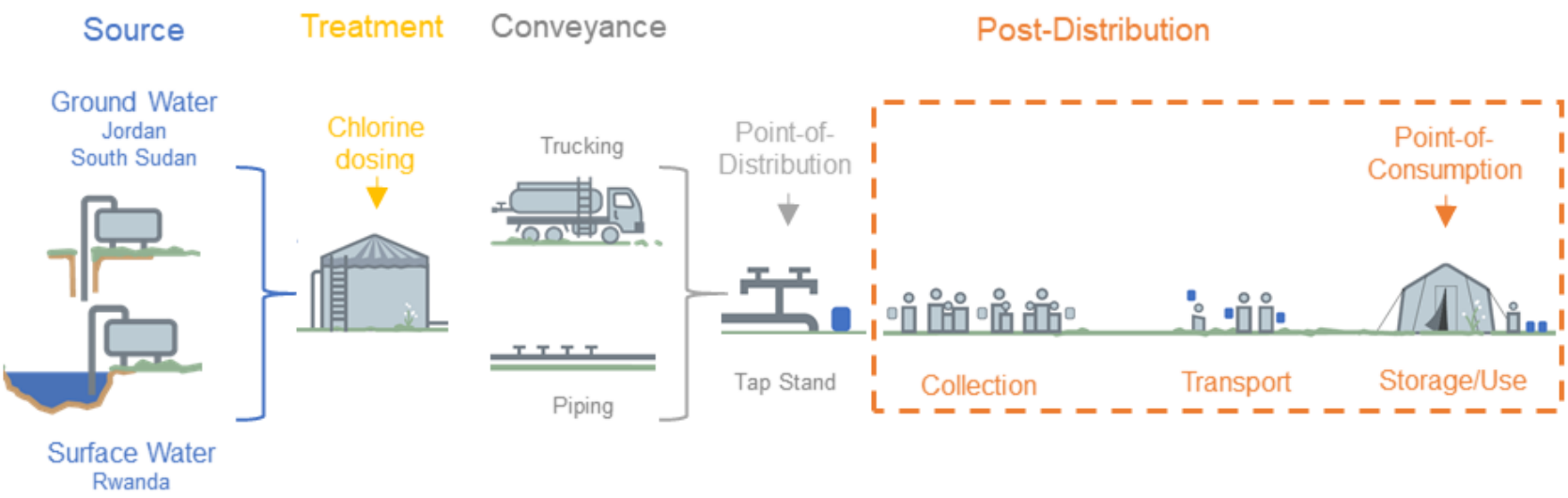

Figure 3-1: Post-distribution period shown in context of overall water supply system for typical refugee or IDP settlement. Water obtained from ground or surface water is centrally treated then conveyed via piped distribution system to the tap stand (point-of-distribution). The post-distribution period begins when water is collected from the tap stand and continues as it is transported to the household and then stored until use (point-of-consumption).

To ensure that there will be adequate chlorine residual throughout the post-distribution period, water system operators must select a chlorine dose during treatment for the point-of-distribution that provides 0.2 mg/L at the point-of-consumption (refer to Figure 3-1 for a summary of water treatment and distribution infrastructure for the sites included in this study). To achieve this, they need models that accurately predict the point-of-consumption FRC concentration using data available at water distribution points. Since post-distribution FRC decay is impacted by a number of quantifiable and unquantifiable factors, ranging from other water quality parameters to contaminants introduced through user interaction with water, modelling approaches must also account for the high degree of variability and uncertainty when modelling post-distribution FRC decay. Past studies have used numerical modelling based on fundamental chemical rate relationships to generate overall empirical kinetic models of post-distribution FRC decay for multiple refugee settlements that predict point-of-consumption FRC (Ali, Ali, and Fesselet, 2021). This process-based modelling approach accounted for uncertainty in post-distribution FRC decay by calibrating the rate and order of chlorine decay based on observed data and by implementing a confidence region estimation of decay parameters. However, these models only produced point predictions of household FRC that cannot quantify the model uncertainty. Furthermore, these process-based models only utilized FRC and time as explanatory variables, and cannot directly incorporate other water quality parameters (e.g., turbidity) which may contribute to chlorine decay.

In this study we developed ensembles of artificial neural networks (ANNs) to produce probabilistic forecasts of point-of-consumption FRC using data collected from the point of distribution as an alternative to process-based modelling of FRC decay. While ANNs have not previously been used for modelling post-distribution FRC, they have been demonstrated to be an effective alternative to process-based models for predicting FRC in piped water distribution systems (Bowden et al., 2006; Gibbs et al., 2006; Rodriguez & Sérodes, 1998; Soyupak et al., 2011) . As a data-driven model, ANNs learn the underlying behaviour from the data instead of assuming the behaviour *a priori*, which is particularly beneficial for modelling post-distribution FRC where the decay behaviour is not well understood. ANNs can also be trained on data representing a wide range of operating conditions and can be retrained easily with new data, unlike process-based models which require decay parameters to be calibrated to a single set of conditions (Bowden et al., 2006; Soyupak et al., 2011). ANNs are also effective even when only

using data collected through routine monitoring (Gibbs et al., 2003, 2006), which is particularly beneficial in humanitarian settings where detailed lab-based water quality evaluations may not be available (Kotlarz, Lantagne, Preston, and Jellison, 2009). In grouping multiple ANNs into an ensemble, we are able to quantify model uncertainty by combining the predictions of multiple ANNs into a probabilistic forecast (Boucher et al., 2011, 2009), providing an important improvement in contrast to past, deterministic, studies attempts to model post-distribution FRC decay. Since ensemble models, including ensembles of ANNs, often produce *underdispersed* forecasts where the spread of the ensemble predictions is less than the spread of the observed data (Boucher et al., 2009, 2015), we also used kernel dressing to post-process the ensemble forecasts to obtain a better match between the forecasted and observed distributions. While this type of post-processing has been used in a variety of contexts for physical models, especially atmospheric models, our study presents an investigation into the effectiveness of post-processing for improving underdispersion of ANN ensemble forecasts of FRC in drinking water.

In developing these ANN ensemble models, our study had two objectives. First, we sought to evaluate the performance of raw and post-processed ANN ensembles for forecasting post-distribution FRC concentrations. Second, we sought to use these models to generate FRC targets for public water distribution points in refugee settlements based on the risk of having insufficient FRC at the point of consumption while also quantifying model uncertainty for water system operators. We generated the ANN ensembles using four datasets from three refugee settlements in South Sudan, Jordan, and Rwanda (two separate datasets were obtained in Jordan, one from 2014 and one from 2015). For each site, we used two input variable combinations using data collected from the point-of-distribution (refer to Figure 3-1): the first (IV1), included only point-of-distribution FRC and the elapsed time of collection, transport, and storage between when water is obtained from the central distribution point and the point-of-consumption, which represents the minimum amount of water quality data that would be reliably available in humanitarian response. The second variable combination (IV2) included all water quality variables recommended for routine monitoring in humanitarian response: point-of-distribution FRC, water temperature, electrical conductivity (EC), turbidity, and pH, as well as elapsed time between the time of collection at the water distribution point and the point-of-consumption(Frazier, 2008; Médecins Sans Frontières, 2010; Sphere Association, 2018). The data-driven approach taken in this study presents an important step in prioritizing evidence-based

solutions for public health engineering in humanitarian response, as well as shifting the paradigm away from searching for a "perfect" model and towards communicating model uncertainty.

3.4 Results

3.4.1 Ensemble Model Performance

Table 3-1 summarizes the performance of both the raw and post-processed ensembles for each variable combination at each site. To prioritize model performance in an operationally acceptable range, we removed observations with water quality parameters outside of the acceptable ranges identified in humanitarian drinking water guidelines as these may represent either atypical values or measurement errors. Specifically, observations were removed if FRC was greater than 2 mg/L, if turbidity was greater than 5 NTU, or if pH was less than 6 or greater than 8 (Médecins Sans Frontières, 2010; Sphere Association, 2018; UNHCR, 2020). From Table 3-1, the Percent Capture of all models is below 100%, ranging from 27% to 65% for the overall dataset and from 0% to 58% for observations with point-of-consumption FRC below 0.2 mg/L, indicating underdispersion, even after post-processing. Figure 3-2, which shows the confidence interval (CI) reliability diagram for each site for the raw and post-processed ensembles, confirms this, showing that the Percent Capture for each ensemble CI fell below the 1:1 line, indicating that at all CI's the models captured less than the optimal number of values, another indication that the forecasts were underdispersed. While the post-processed forecasts were underdispersed, post-processing improves both the dispersion and reliability of the ensembles. The improved dispersion is seen in the higher percentage of values captured, with all models having equal or greater Percent Capture after post-processing. Furthermore, post-processing improved the CI reliability score for both the overall dataset and for observations with point-of-consumption FRC below 0.2 mg/L for all sites except Rwanda. Figure 3-2 shows that this improvement was primarily at the very high ensemble CIs (90-99% CI), and that post-processing did not substantially impact Percent Capture for the lower ensemble CIs. The impact of post-processing on the Continuous Ranked Probability Score (CRPS), which measures the forecast sharpness, reliability, and uncertainty, was less consistent, with the South Sudan and Jordan (2014) models showing improved CRPS with post-processing, and the Jordan (2015) and Rwanda models showing a decrease. This is likely because post-processing improves the underdispersion, which improves the reliability component of CRPS, but also widens the forecast range which produces a worse score for the sharpness component of CRPS.

The ensemble models using the larger IV2 input variable combination typically had better dispersion and reliability, except in South Sudan where the IV1 input variable combination produced lower Percent Capture, but better reliability as shown in Table 3-1. Figure 3-2 also shows that for all sites other than South Sudan, the models using the IV2 variable combination produced forecast with better capture across multiple CIs, leading to a substantial improvement in reliability that is reflected in the CI reliability scores documented in Table 3-1.

Table 3-1: Ensemble verification metrics for all sites and variable combinations for raw and post-processed ensembles

Site	Input Variables	Raw/Post-Processed	Percent Capture [%]	Percent Capture (FRC below 0.2 mg/L [%])	CI Reliability Score	CI Reliability Score (FRC below 0.2 mg/L)	CRPS
South Sudan	IV1	Raw	36	45	1.58	1.15	0.26
		Post-Processed	44	50	1.48	1.10	0.18
	IV2	Raw	47	47	1.85	1.73	0.32
		Post-Processed	56	58	1.76	1.64	0.20
Jordan (2014)	IV1	Raw	30	10	2.65	3.66	0.30
		Post-Processed	37	20	2.55	3.49	0.22
	IV2	Raw	60	45	1.65	2.41	0.27
		Post-Processed	60	45	1.63	2.41	0.19
Jordan (2015)	IV1	Raw	27	0	2.40	3.85	0.11
		Post-Processed	27	0	2.48	3.85	0.17
	IV2	Raw	33	0	2.27	3.85	0.12
		Post-Processed	33	0	2.15	3.85	0.15
Rwanda	IV1	Raw	30	0	2.25	3.85	0.16
		Post-Processed	30	0	2.32	3.85	0.19
	IV2	Raw	65	17	0.77	3.27	0.16
		Post-Processed	65	17	0.89	3.03	0.23

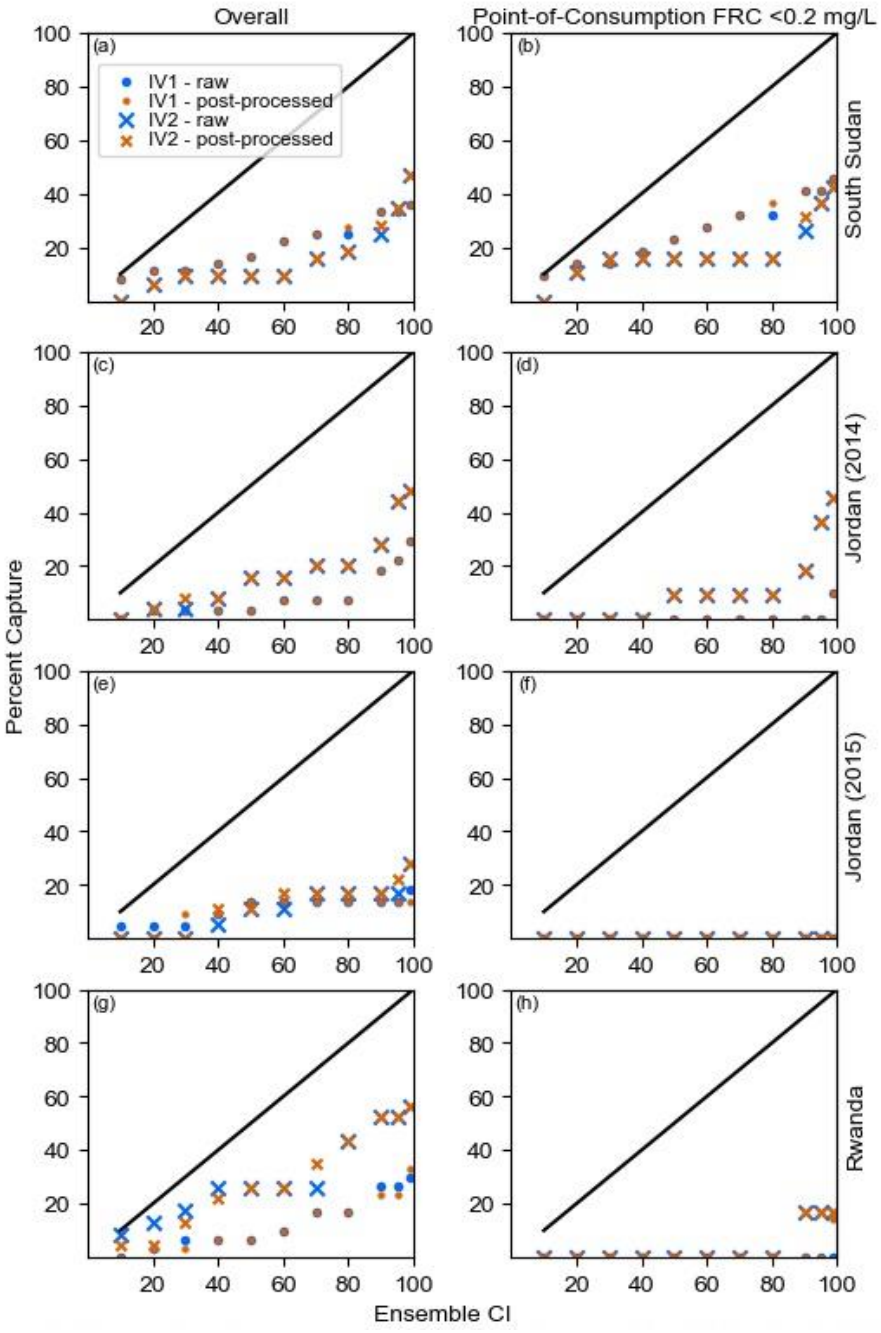

Figure 3-2: Confidence Interval reliability diagrams for all sites. Raw and post-processed CI reliability diagrams for all sites for both the overall dataset (left) and for observations where point-of-consumption FRC is below 0.2 mg/L (right). All ensembles have Percent Capture below the 1:1 line, indicating underdispersion at all CI's, though better reliability is observed for models using the IV2 input variable combination.

The following sections provide the modelling results for the post-processed ensembles for each site and variable combination. Only post-processed results are shown in this section as the post-processing consistently provided better performance. The raw ensemble results are included in Figures A-1 through A-4 in Appendix A.

3.4.2 South Sudan

Figure 3-3 shows the observed and post-processed ensemble forecasts of point-of-consumption FRC against the IV1 and IV2 input variables for South Sudan. The ensemble forecasts generally follow the same trends as the observations, though there are several observations lying outside of the ensemble forecast range, confirming that the ensembles are underdispersed. The ensembles using the IV2 input variable combination produced much wider forecasts, which explains the higher Percent Capture for the IV2 models documented in Table 3-1.

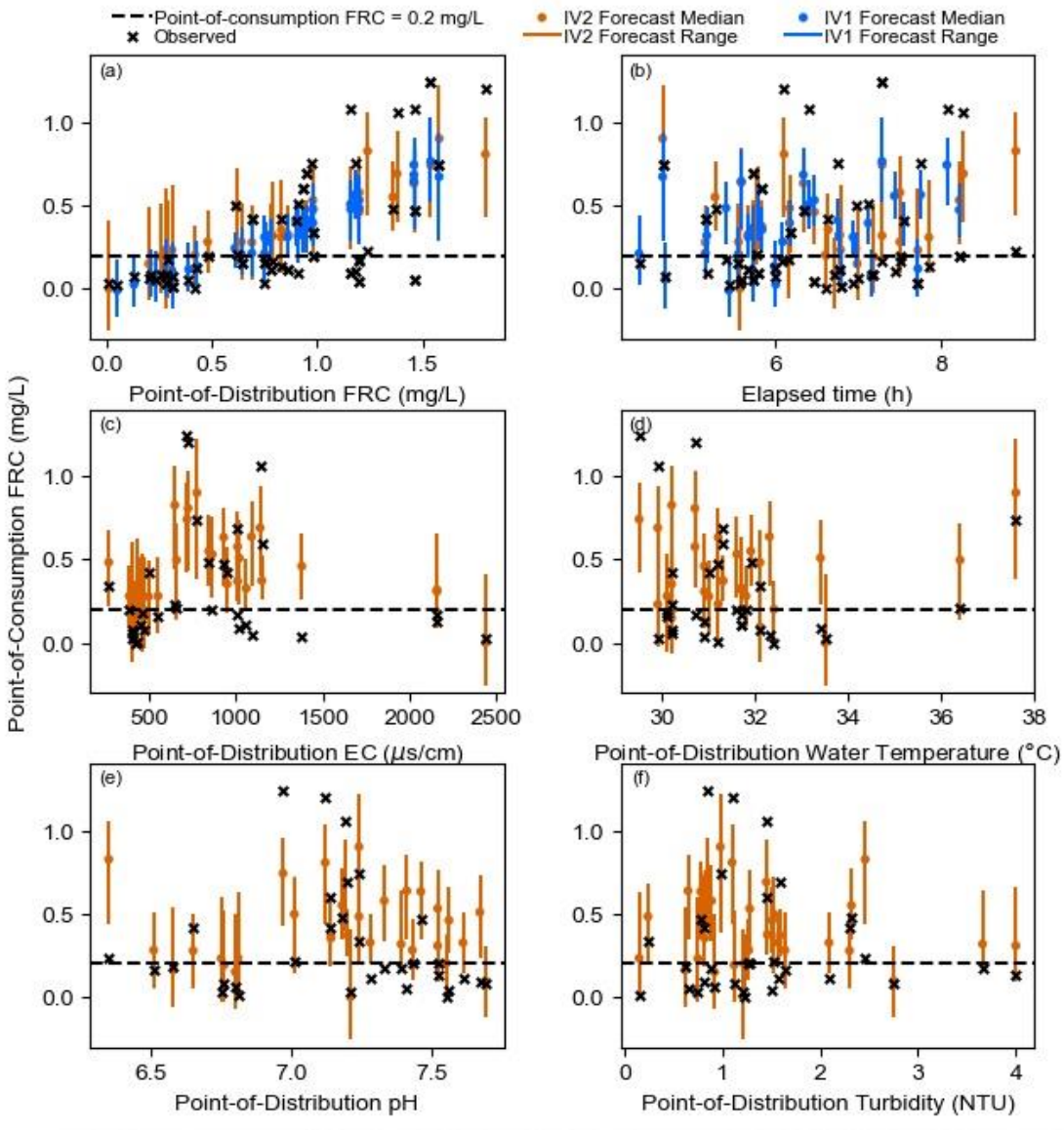

Figure 3-3: South Sudan observations and post-processed forecasts of point-of-consumption FRC. The observations and forecasts are shown against (a) point-of-distribution FRC, (b) elapsed time, (c) point-of-distribution EC, (d) point-of-distribution water temperature, (e) point-of-distribution pH, (f) point-of-distribution turbidity. A strong trend between point-of-consumption and point-of-distribution FRC is observed and IV2 forecasts are much more dispersed than IV1 forecasts.

The clearest trend between the observed and forecasted point-of-consumption FRC and the input variables shown in Figure 3-3 was with the point-of-distribution FRC. There was very little evidence of a trend between the elapsed time and the point-of-consumption FRC. Figure 3-3 also

shows negative trends between the forecasted and observed point-of-consumption FRC and water temperature and turbidity. The trend between point-of-consumption FRC and EC is less clear as at low conductivities; there appears to be a positive trend, but at high conductivities there appears to be a negative trend. Finally, there was not a strong trend between pH and the point-of-consumption FRC.

3.4.3 Jordan (2014)

Figure 3-4 shows the Jordan (2014) forecast-observation pairs against the IV1 and IV2 input variables for the post-processed ensemble forecasts. The ensembles using the IV1 input variable combination produced substantially narrower forecasts, especially in regions of the output space where there is a large density of observations with behaviour resembling that of a linear regression where the ensemble predictions regress to the mean in locations where there is a high density of data. By contrast, the ensemble models using the IV2 input variable combination produced much wider forecasts, leading to the better Percent Capture documented in Table 3-1.

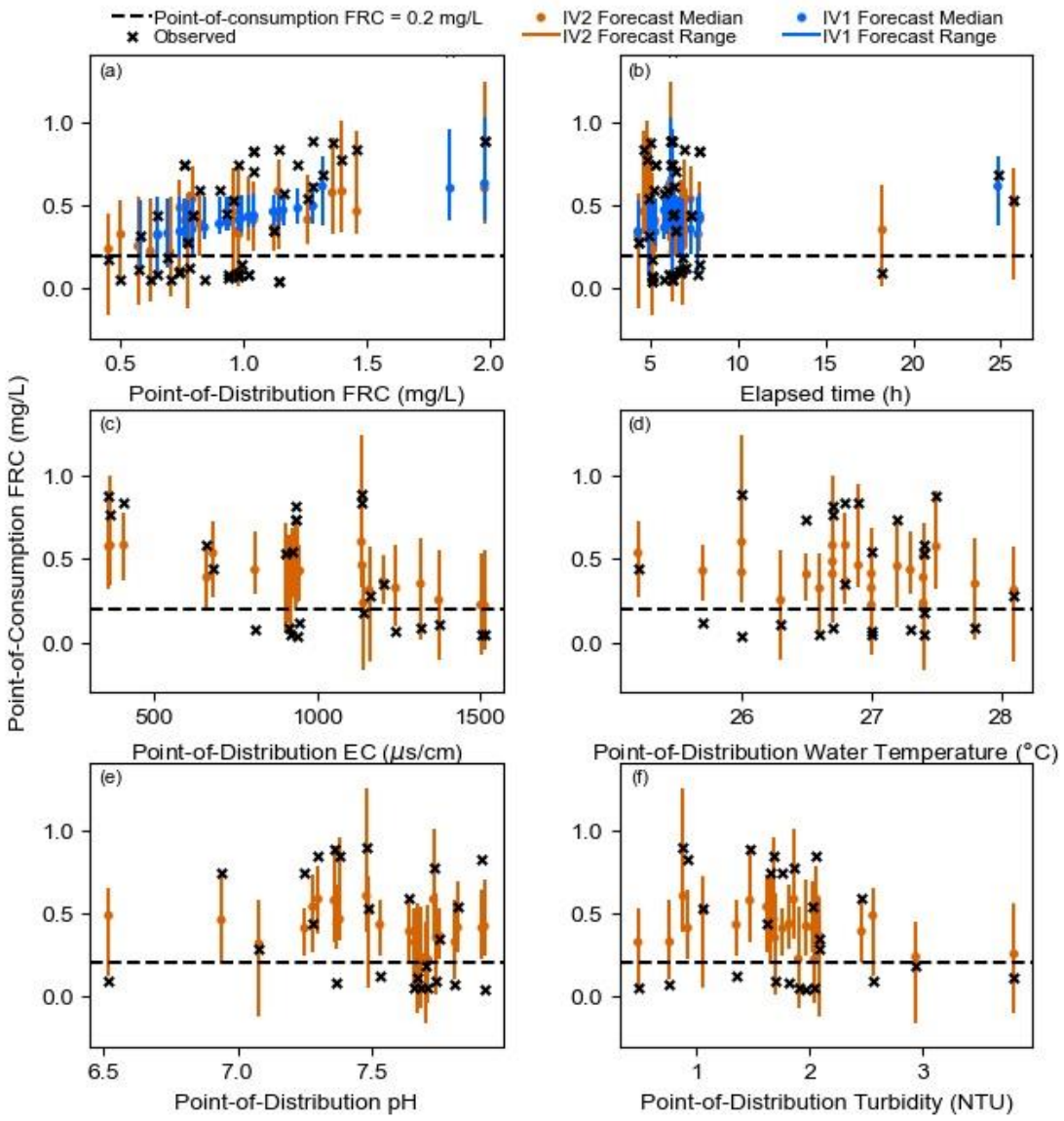

Figure 3-4: Jordan (2014) observations and post-processed forecasts of point-of-consumption FRC. The observations and forecasts are shown against (a) point-of-distribution FRC, (b) elapsed time, (c) point-of-distribution EC, (d) point-of-distribution water temperature, (e) point-of-distribution pH, (f) point-of-distribution turbidity. IV1 forecasts show a strong regression to the mean behaviour. Strong trends between point-of-consumption FRC and: point-of-distribution FRC, EC, and water temperature.

As in South Sudan, the forecast point-of-consumption FRC for both input variable combinations followed similar trends as the observed data, with the clearest trend between input and output variables between point-of-consumption FRC and point-of-distribution FRC. There was little

evidence of a trend between elapsed time and observed or forecasted point-of-consumption FRC. There were also clear negative trends between the observed and forecasted point-of-consumption FRC concentration and EC, water temperature, and turbidity, indicating that as the values of these water quality parameters increase, point of consumption FRC decreases. There was not a strong trend observed with pH.

3.4.4 Jordan (2015)

Figure 3-5 shows the Jordan (2015) forecast-observation pairs against the IV1 and IV2 input variables for the post-processed ensembles forecasts. As with the Jordan (2014) model, the Jordan (2015) ensembles using IV2 produce wider ensemble forecasts than the models using IV1, however these were not wide enough to capture the only observation where the point-of-consumption FRC concentration was below 0.2 mg/L, as it was a very distant outlier. There was little observed trend between the observed point-of-consumption FRC and the IV1 and IV2 input variables, and the resulting forecasts showed little variability in the forecast point-of-consumption FRC.

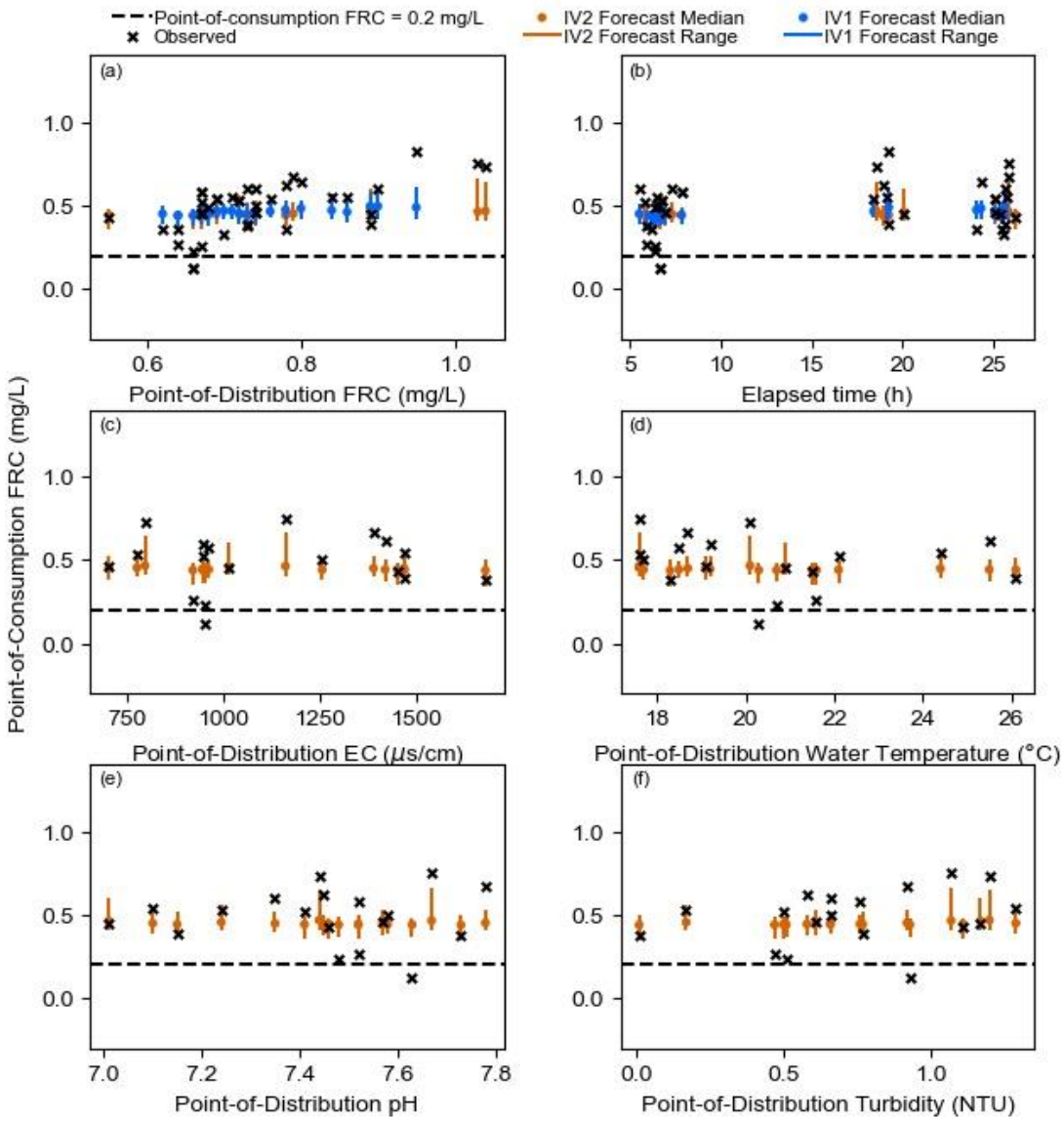

Figure 3-5: Jordan (2015) observations and post-processed forecasts of point-of-consumption FRC. The observations and forecasts are shown against (a) point-of-distribution FRC, (b) elapsed time, (c) point-of-distribution EC, (d) point-of-distribution water temperature, (e) point-of-distribution pH, (f) point-of-distribution turbidity. Both IV1 and IV2 forecasts are very flat due to low overall rates of FRC decay at this site.

3.4.5 Rwanda

Figure 3-6 shows the forecast-observation pairs for Rwanda against the IV1 and IV2 input variables for the post-processed ensemble forecasts. As with the Jordan (2014) model, the ensemble models using the IV1 input variable combination produce forecast behaviour

resembling a regression to the mean where the forecast range decreases where large numbers of observations are present. This narrowing of the forecast range resulted in no forecasts capturing observations with point-of-consumption FRC below 0.2 mg/L, as documented in Table 3-1. The models using the IV2 input variable combination produced forecasts that matched the spread of the observations much better, which lead to the improved Percent Capture for these models documented in Table 3-1.

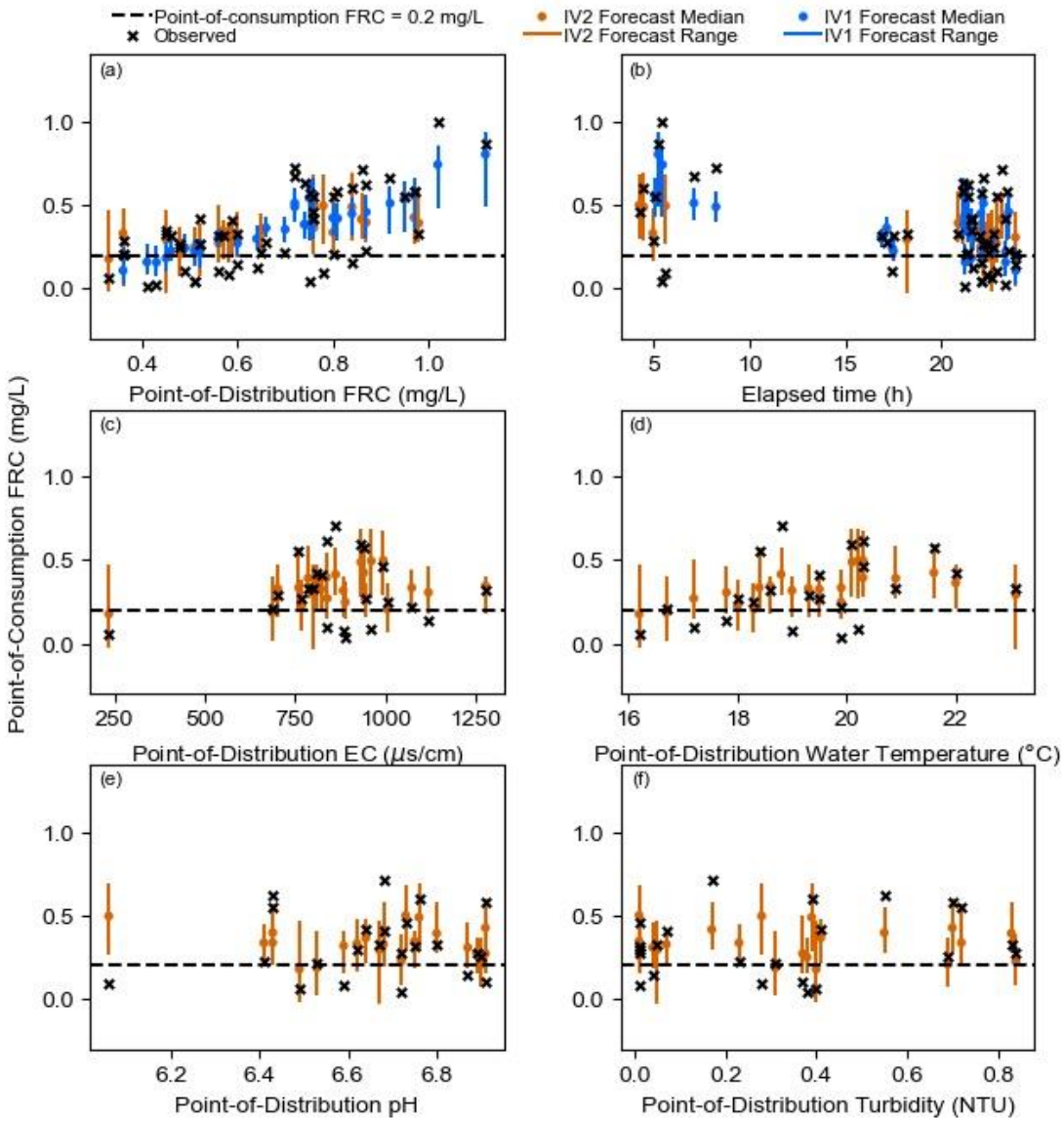

Figure 3-6: Rwanda observations and post-processed forecasts of point-of-consumption FRC. The observations and forecasts are shown against (a) point-of-distribution FRC, (b) elapsed time, (c) point-of-distribution EC, (d) point-of-distribution water temperature, (e) point-of-

distribution pH, (f) point-of-distribution turbidity. IV2 forecasts tend to be much more dispersed, leading to better overall capture, especially of observations with point-of-consumption FRC below 0.2 mg/L.

From Figure 3-6, we see that the forecast point-of-consumption FRC tends to follow the same trends as the observations and that the clearest trends were between the forecasted point-of-consumption FRC and the point-of-distribution FRC and elapsed time. This latter trend had not been strong at the other sites. Furthermore, the remaining water quality variables did not display clear trends with the forecasted point-of-consumption FRC, despite their inclusion substantially improving model performance.

3.4.6 Partial Correlation Analysis Results

Table 3-2 presents the results of a partial correlation analysis that was performed for each site to provide additional details on the trends shown between point-of-consumption FRC and the six input variables. The partial correlation between each input variable and the observed point-of-consumption FRC is shown for each site and for all four datasets together ("Combined" column in Table 2). Table 3-2 shows that point-of-distribution FRC had the strongest partial correlation with point-of-consumption FRC at all sites. The other water quality variables had mostly consistent negative partial correlations with point-of-consumption FRC, indicating that point-of-consumption FRC decreases as the magnitudes of these parameters increase, with the strength of the partial correlation varying from site to site. The generally negative partial correlation, as well as the variability of the magnitude of the partial correlation, coheres with the trends shown visually in Figures 3-3 through 3-6. Additionally, the negative correlations between FRC and water temperature and turbidity conform with the findings of past studies of FRC decay both within piped distribution systems and for household stored drinking water (Clark & Sivaganesan, 2002; Fisher, Kastl, and Sathasivan, 2017; Lantagne, 2008; M. W. LeChevallier, Evans, and Seidler, 1981; James C. Powell, West, Hallam, Forster, and Simms, 2000; Warton, Heitz, Joll, and Kagi, 2006). The relationship between point-of-consumption FRC and elapsed time, however, was less consistent with half the sites having positive partial correlations between point-of-consumption FRC and elapsed time, and the other half having negative partial correlations.

Table 3-2: Partial correlation analysis results between water quality variables and point-of-consumption FRC

Point-of-Distribution Water Quality Variable	South Sudan	Jordan (2014)	Jordan (2015)	Rwanda	Combined
FRC	0.66	0.43	0.31	0.63	0.59
Elapsed Time	0.10	-0.09	0.20	-0.26	-0.01
EC	-0.07	-0.34	-0.08	-0.04	-0.10
Water Temperature	0.00	-0.06	-0.10	-0.13	-0.15
pH	-0.10	-0.09	-0.14	0.07	-0.01
Turbidity	-0.01	-0.03	0.05	-0.20	-0.04

3.4.7 Risk-Based FRC Targets

We generated point-of-distribution FRC targets for each site by forecasting the point-of-consumption FRC for a range of point-of-distribution FRC concentrations (from 0.2 mg/L to 2.0 mg/L). We selected this range considering both the experience of water system operators and point-of-distribution FRC recommendations in drinking water quality guidelines from refugee and IDP settlements(Médecins Sans Frontières, 2010; Sphere Association, 2018; UNHCR, 2020). Following this, the risk of point-of-consumption FRC being below 0.2 mg/L was determined for each point-of-distribution FRC concentration from the forecast cumulative density function (cdf). We selected the FRC target as the lowest point-of-distribution FRC concentration that produced negligible risk. We consider negligible risk to be a 0% predicted risk of low point-of-consumption FRC. While this risk can never truly be non-existent, 0% predicted risk indicates that the predicted risk is too small to be stored as a floating-point number. The use of 0% predicted risk in this study is meant to be illustrative, and in practice the target FRC could be selected for any level of protection. While the below section presents the recommendations required to achieve negligible risk, we present recommended targets for 5% and 15% risk thresholds in Table A-1 in Appendix A. These higher risk thresholds may be needed in sites with high FRC decay or low chlorine taste and odour acceptability.

We used a storage duration of 24 hours for all sites and datasets except in South Sudan where we used a storage duration of 10 hours in keeping with past studies that have shown that long storage durations were not practiced at this site and that it is difficult to maintain a chlorine residual over long storage durations at this site due to high FRC decay rates that have been attributed to very hot temperatures and poor overall water, sanitation, and hygiene (WASH) conditions (Ali et al., 2015, 2021). For the IV2 models, which include additional water quality variables, we simulated two scenarios: an “average case” scenario which used the median values of EC, water temperature, pH, and turbidity, and a “worst-case” scenario where we simulated water quality conditions that would be unfavourable for maintaining a chlorine residual. From the partial correlation analysis presented above, as well as the trends shown in Figure 3-3 through 3-6, we determined that higher values for the four water quality parameters (EC, water temperature, turbidity, and pH) would produce the least favourable conditions, so we used the 95^{th} percentile value observed in each dataset for the “worst case” scenario. Thus the “worst case” scenario reflects only the values observed during the data collection period, and do not

account for seasonal factors such as flooding or monsoon seasons that occurred outside of the period of data collection.

The predicted risk of point-of-consumption FRC below 0.2 mg/L for each site for all three cases (IV1, IV2 average case, IV2 worst case) are presented in Figure 3-7. To achieve negligible risk of point-of-consumption FRC below 0.2 mg/L in South Sudan (Figure 3-7a), the recommended point-of-distribution FRC concentration ranges from 0.70 mg/L (IV2 “worst case”) to 0.95 mg/L (IV1), with 0.75 mg/L recommended for the IV2 “average case” scenario. In Jordan (2014) (Figure 3-7b) the recommended point of distribution FRC using the IV1 model is 0.70 mg/L and is 1.05 mg/L for the “average case” scenario using the IV2 model. No point-of-distribution FRC concentration was able to achieve negligible risk of point-of-distribution FRC below 0.2 mg/L in the “worst case” scenario for the IV2 model though there is very little change in the predicted risk for point-of-distribution FRC concentrations between 1.75 mg/L and 2.0 mg/L. Thus, any point-of-distribution FRC concentration between 1.75 mg/L and 2.0 mg/L would achieve similar risk of having point-of-consumption FRC below 0.2 mg/L. Therefore, we recommend using the lowest FRC concentration within this range (1.75 mg/L) for the “worst case” scenario to reduce the potential for disinfection by-product (DBP) formation or taste and odour concerns. In Jordan (2015) (Figure 3-7c) a point-of-distribution FRC concentration of 0.2 mg/L is recommended for the IV1 and IV2 “average case” scenarios, and 0.4 mg/L for the IV2 “worst case” scenario. In Rwanda, (Figure 3-7d) the recommended point-of-distribution FRC concentration ranges from 0.60 mg/L (IV1 and IV2 “average case”) to 0.90 mg/L (IV2 “worst case”).

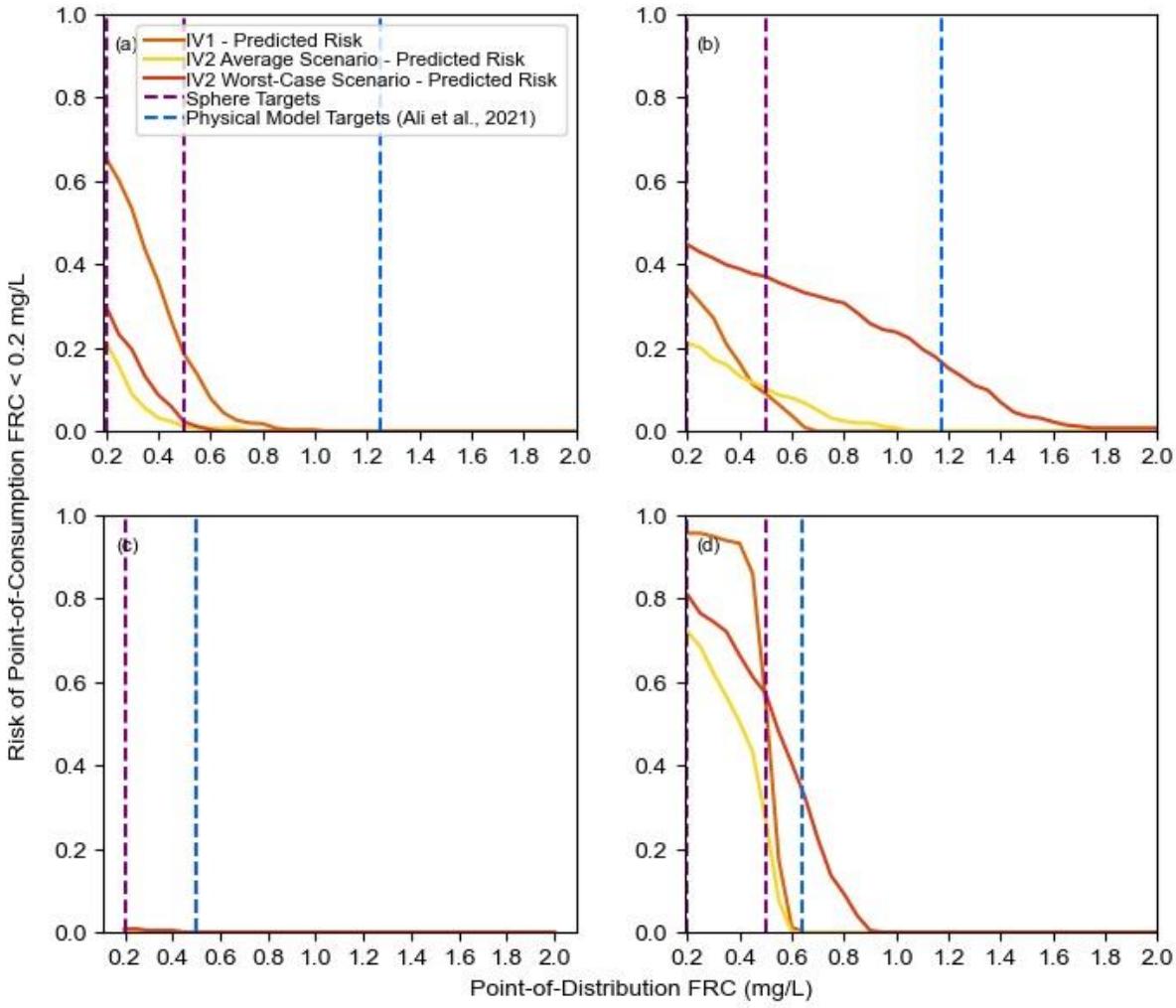

Figure 3-7: Predicted risk of insufficient point-of-consumption FRC (below 0.2 mg/L). The predicted risk is shown for (a) South Sudan, (b) Jordan (2014), (c) Jordan (2015), and (d) Rwanda. To achieve negligible risk, the ANN ensemble models recommend point of distribution FRC between 0.65 and 0.90 mg/L in South Sudan, between 0.7 and 1.75 mg/L in Jordan (2014), between 0.2 and 0.4 mg/L in Jordan (2015), and between 0.60 and 0.90 mg/L in Rwanda. The upper limit of the recommendation for Jordan (2014) does not ensure negligible risk, as this was never achieved, but represents a plateau in the predicted risk of FRC below 0.2 mg/L.

To provide additional context for the risk predictions, Figure 3-8 shows the forecast range at each point-of-distribution FRC concentration for the three scenarios as well as the recorded observations for similar storage durations (6-12 hours for South Sudan, 20-28 hours for all other sites). This figure shows that the ANN ensemble forecasts reflect uncertainty well, with wider forecasts where there are fewer observations (and hence greater uncertainty), and narrower

forecasts where there are more observations. However, at all sites except Rwanda (bottom row) this leads to an overprediction of point-of-consumption FRC at low point-of-distribution FRC concentrations. While these forecasts are unrealistic, they could easily be corrected with further post-processing. Figure 3-8 also shows that the forecasts produced by the ensemble models using the IV2 input variable combination (shown in the middle column for the "average case" scenario and in the right column for the "worst case" scenario) tended to produce wider forecast ranges for all sites except South Sudan (top row). Additionally, we see that the forecasts produced by the IV2 model for the "worst-case" scenario in Jordan (2014) (Figure 3-8f) and for Rwanda (Figure 3-8l) captured all of the observations with point-of-consumption FRC below 0.2 mg/L and very effectively reproduced the behaviour of observations with low point-of-consumption FRC.

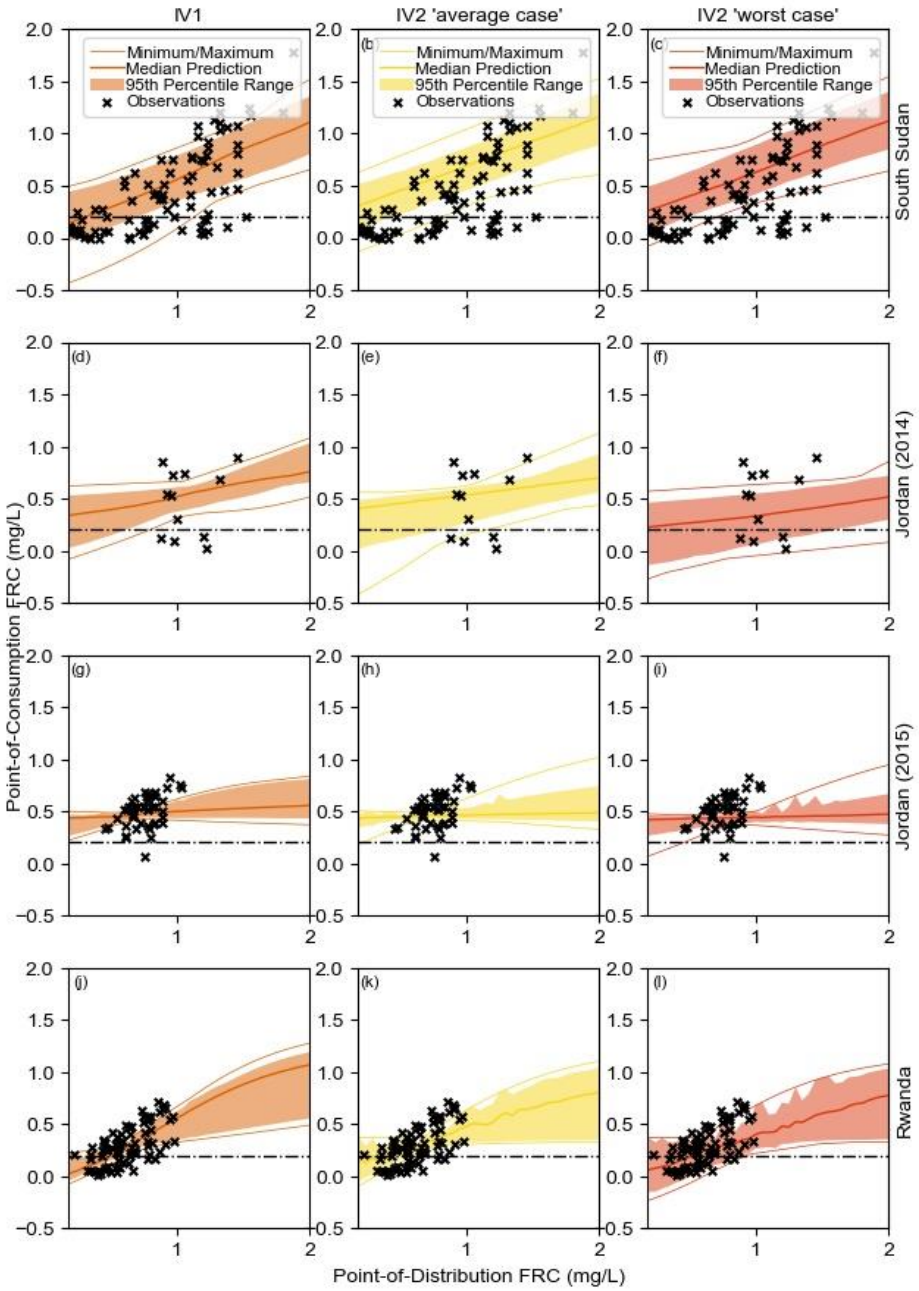

Figure 3-8: Forecasts used to generate risk-based FRC targets. Top row: South Sudan, Second row: Jordan (2014), Third Row: Jordan (2015), Bottom row: Rwanda. Left column: forecasts produced by models using IV1, middle column: forecasts produced by models using IV2 for average case. Right column: forecasts produced by models using IV2 for worst case scenario.

3.5 Discussion

The ensemble performance metrics listed in Table 3-1, as well as the results shown in Figures 3-2 through 3-6, highlight that the forecasts produced by the ANN ensembles were underdispersed. This problem has also been identified when using ANN ensembles to forecast hydrological variables (Boucher et al., 2011, 2009). However, these previous studies did not implement post-processing of the ensemble forecasts. While the post-processing implemented in this study generally improved the ensemble reliability and dispersion, it did not lead to full capture of the observations, nor did it substantially improve the reliability of the ensemble forecasts. Future study should investigate opportunities to improve the raw ensemble forecasting performance, as well as alternative ensemble formation techniques and other machine learning models to reduce the dependence on post-processing. Future study should also investigate more sophisticated post-processing methods which have been proposed and validated in the literature (Boucher et al., 2015; Bröcker & Smith, 2008; Fortin, Favre, & Saïd, 2006; Wang & Bishop, 2005). In particular, considering the regression-to-the-mean style behaviour shown for some of the models, the use of mean squared error (MSE) as the cost function for training the base learners may be contributing to the forecast underdispersion, as this cost function tends to reward clustering near the centre of the distribution of observed values. Future studies should investigate alternative cost functions and training options to avoid this type of behaviour.

The models using the IV2 input variable combination tended to produce better dispersion and reliability than those using the IV1 input variable combination. This shows that including additional water quality variables allowed the models to better reproduce the observed variability and match the distribution of the observed values of point-of-consumption FRC. This is particularly important as all of these water quality variables can be monitored directly in the field and are often part of routine water quality monitoring programs in humanitarian response settings. Of the water quality variables included in this study, water temperature and EC had the most consistent relationship with point-of-consumption FRC, as shown in the trends in Figures 3-3 through 3-6, and in the partial correlation results (Table 3-2). This reflects the findings of past studies which show that water temperature has an important impact on FRC decay within distribution systems as it impacts the rate of the decay reactions (Clark & Sivaganesan, 2002; Fisher et al., 2017; J. C. Powell, Hallam, West, Forster, & Simms, 2000; Warton et al., 2006). The relationship between EC and FRC is not as well documented as EC is a bulk indicator that

may correspond to many compounds such as salts, metals, and dissolved organics, and only some of these are likely to exert chlorine demand (WHO, 2011). However, past studies have shown that EC had a significant effect on post-distribution FRC decay in South Sudan (Ali et al., 2015). The relationship between turbidity and point-of-distribution FRC was less consistent, likely because turbidity is also a bulk indicator that does not reflect any individual compound. In some cases, turbidity-causing compounds, especially oxidizable organic material or suspended metals can exert a large chlorine demand (Gallandat, Stack, String, and Lantagne, 2019; Kotlarz et al., 2009; M. W. LeChevallier et al., 1981), but other turbidity-causing compounds, such as clays which are a common source of turbidity in groundwater, do not exert strong chlorine demand (Gallandat et al., 2019; Wu & Dorea, 2020). The weaker observed trends and partial correlation between point-of-consumption FRC and pH are interesting as pH has been shown to be an important factor in FRC decay (Clark & Sivaganesan, 2002). In this study, the pH range is rather small, (between 6 and 8), which may explain the limited trend with pH as this neutral range is typically associated with the highest rate of FRC decay (Adam & Gordon, 1999). Interestingly, the input variable that displayed the weakest trends and partial correlation with point-of-consumption FRC was elapsed time, despite FRC decay being a time dependent reaction. This may indicate that elapsed time is confounded with other factors, especially considering that elapsed time tended to cluster around a few different values at each site, (c.f., Figures A-5 through A-12 in the Appendix A). Longer storage durations include periods of overnight storage when temperatures are cooler, and where there is less opportunity for user interaction with the water, which may lead to a lower overall rate of decay. Conversely, shorter storage durations tend to reflect water collected in the morning and stored during the day when temperatures are warmer and when water is being used more frequently, both of which may contribute to higher rates of FRC decay. Thus, while FRC is a time dependent reaction, other time-dependent factors may confound the effect of elapsed time on post-distribution FRC. Additionally, while the inclusion of additional water quality variables improved model performance, we did not quantify the impact of individual water quality variables on water quality performance, nor did we implement any evidence-based input variable selection techniques. Future study should focus on identifying which variables are most important to model performance to streamline data collection, and in particular, should seek to clarify the influence of elapsed time on FRC decay.

Unlike the other sites, in South Sudan, models using the IV1 combination had better reliability than those using IV2. The poorer ensemble reliability exhibited for South Sudan using the IV2 variable combination may be due to the nature of water supply at this site. The South Sudan site was comprised of three subcamps, each with their own water distributions systems. Since chlorine decay behaviour is specific to the distribution system (Vasconcelos, Rossman, Grayman, Boulos, & Clark, 1997), the impact of these other water quality parameters may have varied between the three subcamps, so a consistent behaviour for these additional variables could not be identified during the training of the ANN base learners. Future work should investigate the possibility of developing individual models for each subcamp to identify if this behaviour is observed even with distribution system-specific models. This may be challenging with the current dataset however due to the relatively small number of observations available at each subcamp.

The risk-based FRC targets produced in this study varied substantially from site to site, and in the case of Jordan, varied over time as well. This highlights a key shortcoming of current humanitarian drinking water quality guidelines: they are universal and static, recommending the same range of point-of-distribution FRC concentrations for all sites at all times. The results of this research highlight that this is not effective for ensuring adequate FRC levels at the point-of-consumption as for all sites except Jordan (2015), the ANN ensembles predicted a substantial risk of insufficient point-of-consumption FRC when using the Sphere-recommended 0.2-0.5 mg/L FRC concentration at the point-of-distribution. This is reinforced by a previous study that used process-based models of FRC decay that also found the Sphere recommendations would not provide sufficient FRC at any of these sites except Jordan (2015). Furthermore, the authors of the previous study identified that the Sphere guidelines were only effective in Jordan (2015) due to very low FRC decay rates which resulted from low temperatures and very good overall site hygiene. However, since the Jordan (2014) model showed substantial risk of unsafe drinking water using the Sphere guidelines, it is unclear if these favourable conditions would be long-lasting.

The risk-based FRC targets generated in this study also showed interesting relationships with the FRC targets generated in a previous study through process-based modelling. The process-based models recommended point-of-distribution FRC concentrations to ensure 0.2 mg/L at the point

of consumption, with 1.25 mg/L recommended for South Sudan, 1.17 mg/L for Jordan (2014), 0.5 mg/L for Jordan (2015), and 0.64 mg/L for Rwanda (Ali et al., 2021). These are largely in line with the IV1 model recommendations, and the IV2 "average case" scenario. Moreover, the process-based study also included an empirical water safety evaluation using the primary field data to assess how many dwellings had adequate point-of-consumption FRC using the recommendations from the process-based models. They found that, using the FRC targets generated by the process-based models, listed above, 71% of dwellings in South Sudan 82% of dwellings in Jordan (2014), 100% of dwellings in Jordan (2015), and 68% of dwellings in Rwanda had point of consumption FRC above 0.2 mg/L (Ali et al., 2021). In Jordan and Rwanda, this coheres with the risk of point-of-consumption FRC below 0.2 mg/L predicted by the worst-case scenario which predicted a 17% risk of insufficient point-of-consumption FRC for the process-based recommendation in Jordan (2014), negligible risk in Jordan (2015), and 32% risk in Rwanda. This shows that for these sites, the "worst-case" scenario for the models using the IV2 variable combination provides very accurate predictions of the risk of insufficient FRC.

The exception to this is South Sudan where all model scenarios predicted negligible risk of insufficient point-of-consumption FRC for the point-of-distribution FRC target recommended by the process-based model. This may be due to differences in data preparation between this study and the previous study as we removed observations where the point-of-distribution water quality parameters exceeded guideline values, but the previous study did not and the South Sudan dataset had numerous observations with large differences in FRC between distribution and consumption where the point-of-distribution water quality did not meet guideline values. By removing these observations to prioritize model performance in operationally acceptable ranges, we may have created models which were overly optimistic, especially when compared to previous studies that did not omit these values. Additionally, for all of these targets, the scenarios were generated using data only from a short period of data collection and do not represent long-term "average" or "worst case" scenarios. However, this highlights an advantage of the ANN modelling approach: it is very simple to retrain the models, allowing them to adapt to potentially dynamic water quality conditions in refugee and IDP settlements and at the same time, it is also very simple to track a long-term "average case" and "worst case" set of water quality conditions, if needed, for generating FRC targets. Future studies should investigate the advantages and

drawbacks of using long and short-term datasets for both training ANN ensembles and for generating FRC targets.

By accurately predicting the risk of insufficient FRC, the ANN ensemble models not only provide FRC targets which can provide better confidence for water system operators, it also allows water system operators to balance the risk of insufficient FRC against other concerns such as DBP formation or taste and odour concerns, both of which increase as the chlorine residual increases. In particular, taste and odour concerns can be problematic as they may result in water users turning to unsafe drinking water sources(WHO, 2011). Attitudes towards chlorine taste and odour tend to be both site specific and dynamic, though the reported average chlorine taste and odour acceptability threshold from studies in Bangladesh, Ethiopia, and Zambia ranges from 1.25 mg/L to 2.0 mg/L (Crider et al., 2018; Lantagne, 2008), which indicates that the "worst-case" scenario recommendation for Jordan (2014) could cause taste and odour concerns. Future work should seek to quantify and link the risks of taste and odour concerns and DBP formation on a site-by-site basis in conjunction with analytics presented in this study to further inform the selection of an appropriate point-of-distribution FRC target.

Operationally, a major advantage of the probabilistic approach to generating FRC targets used in this study is that, by communicating the predicted risk that FRC will be below 0.2 mg/L at the point-of-consumption, we allow water system operators to balance the trade-offs between water safety risks and DBP and taste and odour risks. Thus, the ensemble ANN approach allows operators to select a point-of-distribution FRC concentration based on the allowable risk of low FRC at the point-of-consumption. Furthermore, we defined low FRC using a threshold point-of-consumption FRC concentration of 0.2 mg/L based on humanitarian drinking water quality guidelines(Médecins Sans Frontières, 2010; Sphere Association, 2018; UNHCR, 2020) and on past studies that show this is effective for protecting against pathogenic recontamination both in piped distribution systems and in water stored in dwellings (CDC, 2012; Girones et al., 2014; Lantagne, 2008; Rashid et al., 2016; Sikder et al., 2020; WHO, 2011). However, operationally, any threshold value of FRC could be used with the ensemble ANN approach. This is especially important as many of the water quality parameters included in this study not only impact FRC decay, but also the disinfection effectiveness of chlorination (Mark W. LeChevallier, Welch, & Smith, 1996; WHO, 2011).

This study demonstrated the benefits of using a probabilistic, ANN ensemble-based approach for modelling post-distribution FRC and generating risk-based FRC targets. These models used routinely collected water quality data to generate probabilistic, evidence-based FRC targets which showed good agreement with other studies in these settlements, while providing additional benefits by communicating uncertainty and risk. To facilitate the adoption of this probabilistic approach for developing risk-based FRC targets, the analytics presented here have been made freely available to support water system operators in refugee and IDP settlements through the new web-based *Safe Water Optimization Tool* (https://safeh2o.app).

3.6 Methods

3.6.1 Study Sites and Data Collection

The data used for this study was obtained from a previous multi-site study on post-distribution FRC decay collected from refugee settlements in South Sudan, Jordan, and Rwanda(Ali et al., 2021). This dataset was selected as process-based models have been used to produce FRC targets for these sites which provide a useful comparison to the risk-based targets generated in this study. Details of the data collected at these sites, as well as important site characteristics are included in Table 3-3. Two datasets were collected from Jordan: one from the summer of 2014 and one nine months later from the late winter of 2015. The original study treated these as two separate datasets due to differences in environmental conditions between the two datasets (10 °C difference in average temperature) and amount of time between the two datasets (Ali et al., 2021). To ensure a consistent comparison with the original study, we have also treated the 2014 and 2015 data from Jordan as two distinct datasets.

The dataset for each site includes FRC as well as other water quality parameters which are routinely collected in humanitarian water systems operation including total residual chlorine, EC, water temperature, turbidity, and pH. Data was collected using paired sampling whereby the same unit of water was sampled at the following points along the post-distribution water supply chain:

- From the tap at the point of distribution
- In the container immediately after collection
- In the container immediately after transport to the dwelling

- After a follow-up period of storage in the household

Table 3-3: Summary of Key Site Characteristics (Médecins Sans Frontières, 2013; PAJER, 2015; UNICEF, 2015)

Site Country	Name of Refugee Settlement(s)	Ambient Air Temperature (°C)	Population	Water Source	Drinking Water Treatment	Data Collection Period	Number of Paired Samples Collected
South Sudan	Batil Gendrassa Jamam	Average: 35.3 (Min: 28.3; Max: 45.7)	37,199 15,810 15,670	Groundwater (boreholes)	In-line chlorination with calcium hypochlorite	March-April, 2013	69 76 75
Jordan (2014)	Azraq	Average: 32.7 (Min: 27.1; Max: 43.3)	7,470	Groundwater (boreholes)	Reverse osmosis; in-line chlorination with chlorine gas	July-August, 2014	199
Jordan (2015)	Azraq	Average: 21.7 (Min: 14.5; Max: 29.3)	14,797	Groundwater (boreholes)	Reverse osmosis; in-line chlorination with chlorine gas	March-April, 2015	140
Rwanda	Kigeme	Average: 22.2 (Min: 18.3; Max: 31.0)	18,569	Surface water (stream source)	Flocculation, filtration, and chlorination with calcium hypochlorite	June-July, 2015	134

This study only used the measurements at the point-of-distribution and point-of-consumption to reflect data collection practices that are more feasible for humanitarian operations. In preparing the dataset, observations were removed if the point-of-distribution water quality did not meet humanitarian drinking water quality guidelines. Table A-2 in Appendix A includes the full list of data cleaning steps that were used to prepare the data for use in the ANN models.

3.6.2 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.

3.6.3 Input variable selection

Two input variable combinations were considered for predicting the output variable, the point-of-consumption FRC concentration. The variables considered are all variables that are routinely monitored in humanitarian water system operations. The first input variable combination (IV1) included FRC at the water point-of-distribution and the elapsed time between the measurement at the point-of-distribution and the point-of-consumption. This input variable combination represents the minimum number of variables that would be regularly collected under current humanitarian drinking water quality guidelines(Sphere Association, 2018). Additionally, these are the only two variables included in the process-based model developed in a past study for these sites(Ali et al., 2021), so this input variable combination allows for a direct comparison of the ANN ensemble models with the process-based models. The second input variable combination (IV2) included the variables from IV1 as well as additional water quality variables measured from the point-of-distribution (directly after water had left the water distribution point): EC, water temperature, pH, and turbidity. These additional variables are recommended for collection in some humanitarian drinking water quality guidelines(Frazier, 2008; Médecins Sans Frontières, 2010; Sphere Association, 2018), and as such, may also be available in humanitarian response settings. This larger input variable set allowed us to investigate the

usefulness of additional water quality variables for forecasting point-of-consumption FRC concentrations.

3.6.4 Base Learner structure and architecture

The ensemble base learners (the individual ANNs in the ensemble models) were built as multi-layer perceptrons (MLPs) with a single hidden layer using the Keras 2.3.0 package(Cholette, 2015) in Python v3.7(Python Software Foundation, 2019). This structure was selected because it has been shown to outperform other data-driven models and ANN architectures for predicting FRC in piped distribution systems(Gibbs et al., 2006; Rodriguez & Sérodes, 1998). The weights and biases of the base learners were optimized to minimize mean squared error (MSE) using the Nadam algorithm with a learning rate of 0.1. An early stopping procedure with a patience of 10 epochs was used to prevent overfitting.

The hidden layer size of the base learners was determined through an exploratory analysis by consecutively doubling the hidden layer size until performance decreased or ceased to improve substantially from one iteration to the next. Based on this analysis, we selected a hidden layer size of four hidden neurons at all sites for the models using the IV1 variable combination for all sites. For the models using the IV2 input variable combination, we selected a hidden layer size of 16 hidden nodes for South Sudan and Jordan (2015), and a hidden layer size of eight hidden nodes for Jordan (2014) and Rwanda. The full results of the exploratory analysis into hidden layer size are included in Figures A-13 through A-20 in Appendix A.

3.6.5 Data Division

The full dataset for each site and variable combination was divided into calibration and testing subsets, with the calibration subset further subdivided into training and validation data. The testing subset was obtained by randomly sampling 25% of the overall dataset. The same testing subset was used for all base-learners so that each base-learner's testing predictions could be combined into an ensemble forecast. The training and validation data was obtained by randomly resampling from the calibration subset, with a different combination of training and validation data for each base learner to promote ensemble diversity. The ratio of data from the calibration set used for training and validation respectively was selected to avoid both overfitting and underfitting through an exploratory analysis using a grid search process.In all but two cases, we selected a validation set that was twice the size of the training set, for an overall training-

validation-testing split of 25%-50%-25%. The two exceptions to this were for the Jordan (2014) model when using the IV1 input variable combination where we found that a training-validation-testing split of 50-25-25 produced better performance, and for the Jordan (2015) model when using the IV1 input variable combination where a training-validation-testing split of 30-45-25 performed substantially better. The full results of the exploratory analysis for data division are included in Figures A-21 through A-28 in Appendix A. Descriptive statistics for the calibration and testing datasets are included in Tables A-3 and A-4 in Appendix A, and histograms of the input and output variables are provided in Figures A-5 through A-12 in Appendix A to provide context of the range and patterns in the data used to train the ANN base learners.

3.6.6 Ensemble Model Formation

The ensemble models in this study were used to generate probabilistic forecasts of post-distribution FRC by combining the predictions of each base learner into a probability density function (pdf). Thus, for each observation of FRC at the point-of-consumption, the ensemble model outputs a pdf representing the predicted probability of point-of-consumption FRC concentrations. This pdf can then be used to identify ensemble confidence intervals (CIs) for the expected point-of-consumption FRC concentration. To ensure a good representation of the full output space in the final pdfs, two approaches were taken to ensure ensemble diversity. First, as discussed above, the data used to train the base-learner ANNs was randomly sampled from the calibration set, so each ANN was trained on a different subset of the data. Second, the initial weights and biases were randomized for each base learner in a random-start process. Both of these are implicit approaches to ensuring ensemble diversity as they do not directly create diversity and instead the diversity arises through the randomization of the training data and the weights and biases(Brown, Wyatt, Harris, and Yao, 2005). The benefit of implicit approaches is that the differences between the base-learners are derived from randomness in the data(Brown et al., 2005).

The ensemble size (number of base learners included in the ensemble) was also determined through an exploratory analysis using a grid search procedure This exploratory analysis showed that in general, performance increased with larger ensemble sizes, but improvements in performance plateaued at ensemble sizes ranging from 50 members to 250 members. Based on this, a standard ensemble size of 250 members was selected for all sites and variable

combinations. The full results of the exploratory analysis for ensemble size are included in Figures A-29 through A-36 in Appendix A.

3.6.7 Ensemble Post Processing

We used ensemble post-processing to attempt to improve the forecasts generated by the raw ensembles. We used the kernel dressing method to post-process ensemble predictions (Roulston & Smith, 2003). This method 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., 2011, 2015; Bröcker & Smith, 2008; Roulston & Smith, 2003), though the selection of the specific kernel function is not critical (Boucher et al., 2015). The kernel bandwidth was defined using the best member error method where the bandwidth for all kernels is the variance of the absolute error of the prediction that is closest to each observation in the calibration dataset (Roulston & Smith, 2003).

3.6.8 Ensemble Verification and Performance Evaluation

We used ensemble verification metrics to evaluate the performance of the raw and post-processed ensembles for each site and variable combination. Ensemble verification metrics differ from traditional measures of performance (e.g., Nash Sutcliffe Efficiency, MSE, etc.) as they assess the performance of the probabilistic forecasts of an ensemble whereas traditional measures typically evaluate the average performance of an ensemble model or the predictions of a deterministic model(Hamill, 2001). Throughout the following section, O refers to the full set of observed FRC concentrations at the point-of-consumption and o_i refers to the i^{th} observation, where there are I total observations. F refers to the full set of probabilistic forecasts for point-of-consumption FRC, where F_i is the probabilistic forecast corresponding to observation o_i and f_i^m is the prediction by the m^{th} base learner in the ensemble on the i^{th} observation. 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 $m = M$.

3.6.9 Percent Capture

Percent Capture measures the percentage of observations which are captured within the ensemble forecast and provides a useful indication of how well the model can reproduce the full range of

observed values, and, as such, can indicate if a model is underdispersed. For a raw ensemble forecast, the i^{th} observation is captured if $f_i^0 \leq o_i \leq f_i^M$. For a post-processed forecast, the i^{th} observation is captured if the probability of o_i in the mixture distribution is greater than 0. While not commonly used for ensemble verification, a similar metric has been used for evaluating other probabilistic or possibilistic models, especially neurofuzzy networks, referred to either as the Percent Capture or the percent of coverage (Alvisi & Franchini, 2011, 2012; Khan & Valeo, 2016, 2017). The Percent Capture was calculated both for the overall set of observations, as well as for observations with point-of-consumption FRC below 0.2 mg/L. The latter is a useful indicator of how well the model can predict if water will have sufficient FRC at the point-of-consumption, which is an important indicator of the degree of confidence we have in the risk-based targets generated using these ensemble models.

3.6.10 CI Reliability Diagram

Reliability diagrams are visual indicators of ensemble reliability, where reliability refers to the similarity between the observed and forecasted probability distributions with the ideal model having all observations plotted along the 1:1 line showing that the observed probabilities are equal to the forecasted probabilities. These diagrams plot the observed relative frequency of events against the forecast probability of that event, though the reliability diagram has been adapted in past studies as the CI reliability diagram which compares the frequency of observed values within the corresponding CI of the ensemble. For raw ensembles, the CIs are derived from the sorted forecasts of the base learners (for example, the ensemble 90% CI would include all of the forecasts between $f^{0.05M}$ and $f^{0.95M}$) and for post-processed ensembles, the CIs are calculated directly from the probability distribution. In this study, we extended the CI reliability diagram further by plotting the Percent Capture of each CI within the ensemble against the CI level. For each ensemble model we plotted the CI reliability for the 10% to 100% CI levels at 10% intervals as well as at the 95% and 99% CI. We used this to develop a numerical score for the CI reliability diagram which is calculated as the squared distance between the percentage of observations captured within each CI and the ideal Percent Capture in that CI. This was calculated for each CI threshold, k, from 10% to 100% in 10% increments as shown in Equation 3-1.

$$CI\ Reliability\ Score = \sum_{k=0.1}^{1}(k - Percent\ Capture\ in\ CI_k)^2 \quad (3\text{-}1)$$

The CI reliability score measures the horizontal distance between the Percent Capture and the 1:1 line for each CI. The ideal value for this score would be 0, indicating all points fall on the 1:1 line. The worst possible score will depend on the number of CI's included in the calculation of the score; for this study the worst score is 3.9, which would only occur if no observations were captured in any CI of the ensembles. The CI reliability score was calculated for both the overall dataset and for forecast-observation pairs where the observed household FRC concentration was below 0.2 mg/L.

3.6.11 Continuous Ranked Probability Score

The Continuous Ranked Probability Score (CRPS) is a common metric for evaluating probabilistic forecasts that evaluates the difference between the predicted and observed probabilities of continuous variables and is equivalent to the mean absolute error of a deterministic forecast(Ferro, 2014; Hersbach, 2000). The CRPS measures not only model reliability but also sharpness, which is an indicator of how closely the ensemble predictions are clustered around the observed values. Thus, the CRPS can be a useful measure of overdispersion and can provide an indication if improvements in reliability are being obtained at the expense of excess overdispersion. The CRPS is measured as the area between the forecast cumulative distribution function (cdf) and the observed cdf for each forecast observation pairing (Hersbach, 2000). Since each observation is a discrete value, the observation cdf is represented with the Heaviside function $H\{x \geq x_a\}$ which is a stepwise function with a value of 0 for all point-of-consumption FRC concentrations below the observed concentration and 1 for all point-of-consumption FRC concentrations above the observed concentration. The equation for calculating the CRPS of a single forecast observation pair is given in Equation 3-2. Note that Equation 3-2 shows the calculation of CRPS for a single forecast-observation pair. To evaluate the ensemble models, the average CRPS, $\overline{CRPS}$, is calculated by taking the mean CRPS over all forecast-observation pairs.

$$CRPS = \int_{-\infty}^{\infty} (F_i(x) - H\{x \geq o_i\})^2 dx \tag{3-2}$$

For the post-processed probability distributions, we calculated CRPS directly from Equation 3-2 using numerical integration. For the raw ensemble, we treated 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 (Hersbach, 2000). $\overline{CRPS}$ 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(Hersbach, 2000). Using these values, the $\overline{CRPS}$ for an ensemble can be calculated as:

$$\overline{CRPS} = \sum_{n=1}^{N} \overline{g_n}[(1-\overline{o_n})p_n^2 + \overline{o_n}(1-p_n)^2] \text{ (Hersbach, 2000)} \quad (3\text{-}3)$$

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

3.6.12 Generation of Risk Based Targets

To generate the risk-based FRC targets, the trained ensembles of ANNs were used to forecast the point-of-consumption FRC for a series of point-of-distribution FRC concentrations from 0.2 to 2 mg/L in 0.05 mg/L increments. For each point-of-distribution FRC concentration, the predicted risk of insufficient FRC was calculated from the forecast pdf as the cumulative probability of FRC at the point-of-consumption being below 0.2 mg/L. Using this predicted risk, the target FRC concentration for the point-of-distribution was then selected as the lowest FRC concentration at the water point-of-distribution that provides the desired level of protection. For this study we selected the FRC concentration that resulted in negligible risk of FRC being below the 0.2 mg/L threshold (i.e., the lowest FRC concentration where the predicted risk is 0), though operationally any level of protection could be used and the risk of insufficient FRC at the point-of-consumption should be balanced against risks associated with high FRC concentrations, such as DBP formation and taste and odour concerns.

For comparison with the previously published results, we used a storage duration of 10 hours when generating the FRC targets for South Sudan, and 24 hours for all other sites(Ali et al., 2021). Since the IV2 model also requires values for EC, water temperature, pH, and turbidity, two scenarios were considered. First, an “average” scenario was used where the median observed value for all other water quality parameters were selected. The second scenario considered was a “worst-case” scenario, where we simulated a scenario where water quality conditions were unfavourable for maintaining chlorine residual. A partial correlation analysis, which assesses the correlation between an input variable and the output variable while controlling for the impacts of other input variables, was used to determine the least favourable conditions for each input variable. The partial correlation analysis is performed by first developing multiple linear

regression predictions of both the output variable (point-of-consumption FRC) and the input variable of interest using the remaining input variables as the predictors to the linear regression models and then taking the Pearson correlation coefficient of the residuals between the two regression models. Partial correlation was used to assess the directionality of the effect of the additional water quality variables included in IV2 to assess whether high or low values of these inputs would create a worst-case scenario. Once the directionality of the impact of the different variables had been established, the 95^{th} or 5^{th} percentile observed value of that variable was used at each site to simulate the worst-case scenario.

3.7 References

Adam, L. C., & Gordon, G. (1999). Hypochlorite Ion Decomposition: Effects of Temperature, Ionic Strength, and Chloride Ion. *Inorganic Chemistry, 38*(6), 1299–1304. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11670917

Ali, S. I., Ali, S. S., & Fesselet, J.-F. (2015). Effectiveness of emergency water treatment practices in refugee camps in South Sudan. *Bulletin of the World Health Organization, 93*(8), 550–558. https://doi.org/10.2471/BLT.14.147645

Ali, S. I., Ali, S. S., & Fesselet, J. (2021). Evidence-based chlorination targets for household water safety in humanitarian settings: Recommendations from a multi-site study in refugee camps in South Sudan, Jordan, and Rwanda. *Water Research, 189*(116642), 1–17. https://doi.org/https://doi.org/10.1016/j.watres.2020.116642

Alvisi, S., & Franchini, M. (2011). Fuzzy neural networks for water level and discharge forecasting with uncertainty. *Environmental Modelling and Software, 26*(4), 523–537. https://doi.org/10.1016/j.envsoft.2010.10.016

Alvisi, S., & Franchini, M. (2012). Grey neural networks for river stage forecasting with uncertainty. *Physics and Chemistry of the Earth, 42–44*, 108–118. https://doi.org/10.1016/j.pce.2011.04.002

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

Boucher, M. A., Perreault, L., & Anctil, F. (2009). Tools for the assessment of hydrological ensemble forecasts obtained by neural networks. *Journal of Hydroinformatics, 11*(3–4), 297–307. https://doi.org/10.2166/hydro.2009.037

Boucher, M. A., Perreault, L., Anctil, F., & Favre, A. C. (2015). Exploratory analysis of statistical post-processing methods for hydrological ensemble forecasts. *Hydrological Processes, 29(*6), 1141–1155. https://doi.org/10.1002/hyp.10234

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.

Mathematical and Computer Modelling, *44*(5–6), 469–484. https://doi.org/10.1016/j.mcm.2006.01.006

Bröcker, J., & Smith, L. A. (2008). From ensemble forecasts to predictive distribution functions. *Tellus, Series A: Dynamic Meteorology and Oceanography*, *60 A*(4), 663–678. https://doi.org/10.1111/j.1600-0870.2008.00333.x

Brown, G., Wyatt, J., Harris, R., & Yao, X. (2005). Diversity creation methods: A survey and categorisation. *Information Fusion*, *6*(1), 5–20. https://doi.org/10.1016/j.inffus.2004.04.004

CDC. (2012). Chlorine Residual Testing. Retrieved from http://www.cdc.gov/safewater/chlorine-residual-testing.html

Cholette, F. (2015). Keras. Retrieved from https://keras.io

Clark, R. M., & Sivaganesan, M. (2002). Predicting chlorine residuals in drinking water: second order model. *Journal of Water Resources Planning and Management*, *128*(2), 152–161. https://doi.org/10.1061/(ASCE)0733-9496(2002)128:2(152)

Connolly, M. A., Gayer, M., Ryan, M. J., Salama, P., Spiegel, P., & Heymann, D. L. (2004). Communicable diseases in complex emergencies: impact and challenges. *The Lancet*, *364*(9449), 1974–1983. https://doi.org/https://doi.org/10.1016/S0140-6736(04)17481-3

Crider, Y., Sultana, S., Unicomb, L., Davis, J., Luby, S. P., & Pickering, A. J. (2018). Can you taste it? Taste detection and acceptability thresholds for chlorine residual in drinking water in Dhaka, Bangladesh. *Science of the Total Environment*, *613–614*, 840–846. https://doi.org/10.1016/j.scitotenv.2017.09.135

Cronin, A. A., Shrestha, D., Cornier, N., Abdalla, F., Ezard, N., & Aramburu, C. (2008). A review of water and sanitation provision in refugee camps in association with selected health and nutrition indicators - the need for integrated service provision. *Journal of Water and Health*, *6*(1), 1–13. https://doi.org/10.2166/wh.2007.019

Ferro, C. A. T. (2014). Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, *140*(683), 1917–1923. https://doi.org/10.1002/qj.2270

Fisher, I., Kastl, G., & Sathasivan, A. (2017). A comprehensive bulk chlorine decay model for

simulating residuals in water distribution systems. *Urban Water Journal, 14*(4), 361–368. https://doi.org/10.1080/1573062X.2016.1148180

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. *Quarterly Journal of the Royal Meteorological Society, 132*(617), 1349–1369. https://doi.org/10.1256/qj.05.167

Frazier, C. (2008). Water, sanitation and hygiene in emergencies. In E. C. Rand (Ed.), The Johns Hopkins and Red Cross Red Crescent health guide Public in emergencies (2nd ed., pp. 372–441). Geneva: International Federation of Red Cross and Red Crescent Societies. Retrieved from www.ifrc.org

Gallandat, K., Stack, D., String, G., & Lantagne, D. (2019). Residual maintenance using sodium hypochlorite, sodium dichloroisocyanurate, and chlorine dioxide in laboratory waters of varying turbidity. *Water, 11*(6). https://doi.org/10.3390/w11061309

Gibbs, M. S., Morgan, N., Maier, H. R., Dandy, G. C., Holmes, M., & Nixon, J. B. (2003). Use of Artificial Neural Networks for Modelling Chlorine Residuals in Water Distribution Systems. *In MODSIM 2003 International Congress on Modelling and Simulation: Integrative Modelling of Biophysical, Social, and Economic Systems for Resource Management Solutions* 789–794.

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. *Mathematical and Computer Modelling, 44*(5–6), 485–498. https://doi.org/10.1016/j.mcm.2006.01.007

Girones, R., Carratalà, A., Calgua, B., Calvo, M., Rodriguez-Manzano, J., & Emerson, S. (2014). Chlorine inactivation of hepatitis e virus and human adenovirus 2 in water. *Journal of Water and Health, 12*(3), 436–442. https://doi.org/10.2166/wh.2014.027

Golicha, Q.,Shetty, S.,Nasiblov, O., Hussein, A., Wainaina, E., Obonyo, M., Macharia, D., Musyoka, R. N., Abdille, H., Ope, M., Joseph, R., Kabugi, W., Kiogora, J., Said, M., Boru, W., Galgalo, T., Lowther, S. A., Juma, B., Mugoh, R.,...Burton, J.W. (2018). Cholera

outbreak in Dadaab Refugee camp, Kenya — November 2015–June 2016. *Morbidity and Mortality Weekly Report*, *67*(34), 958–961. https://doi.org/10.15585/mmwr.mm6734a4

Guerrero-Latorre, L., Hundesa, A., & Girones, R. (2016). Transmission Sources of Waterborne Viruses in South Sudan Refugee Camps. *Clean - Soil, Air, Water*, *44*(7), 775–780. https://doi.org/10.1002/clen.201500358

Hamill, T. M. (2001). Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Monthly Weather Review*, *129*(3), 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 and Forecasting*, *15*(5), 559–570.

Howard, C. M., Handzel, T., Hill, V. R., Grytdal, S. P., Blanton, C., Kamili, S., Drobeniuc, J., Hu, D., & Teshale, E. (2010). Novel Risk Factors Associated with Hepatitis E Virus Infection in a Large Outbreak in Northern Uganda: Results from a Case-Control Study and Environmental Analysis. *American Journal of Tropical Medicine and Hygiene*, *83*(5), 1170–1173. https://doi.org/10.4269/ajtmh.2010.10-0384

Khan, U. T., & Valeo, C. (2016). Dissolved oxygen prediction using a possibility theory based fuzzy neural network. *Hydrology and Earth System Sciences*, *20*, 2267–2293. https://doi.org/10.5194/hess-20-2267-2016

Khan, U. T., & Valeo, C. (2017). Comparing a Bayesian and fuzzy number approach to uncertainty quantification in short-term dissolved oxygen prediction. *Journal of Environmental Informatics*, *30*(1), 1–16. https://doi.org/10.3808/jei.201700371

Kotlarz, N., Lantagne, D., Preston, K., & Jellison, K. (2009). Turbidity and chlorine demand reduction using locally available physical water clarification mechanisms before household chlorination in developing countries. *Journal of Water and Health*, *7*(3), 497–506. https://doi.org/10.2166/wh.2009.071

Lantagne, D. S. (2008). Sodium hypochlorite dosage for household and emergency water treatment. *Journal of American Water Works Association*, *100*(8), 106–114. https://doi.org//10.1002/j.1551-8833.2008.tb09704.x

LeChevallier, M. W., Evans, T. M., & Seidler, R. J. (1981). Effect of turbidity on chlorination efficiency and bacterial persistence in drinking water. *Applied and Environmental Microbiology, 42*(1), 159–167. https://doi.org/10.1128/aem.42.1.159-167.1981

LeChevallier, Mark W., Welch, N. J., & Smith, D. B. (1996). Full-scale studies of factors related to coliform regrowth in drinking water. *Applied and Environmental Microbiology, 62*(7), 2201–2211. https://doi.org/10.1128/aem.62.7.2201-2211.1996

Médecins Sans Frontières. (2010). Public Health Engineering In Precarious Situations. (J. V. D. Noortgate and P. Maes, Eds.) (2nd ed.). Brussels: Médecins Sans Frontières.

Médecins Sans Frontières. (2013). Maban County, South Sudan WASH Coordination Report (Week 11 and 12). Amsterdam.

PAJER. (2015). Kigeme, Rwanda WASH Monthy Updates (June-July). Kigali.

Powell, J. C., Hallam, N. B., West, J. R., Forster, C. F., & Simms, J. (2000). Factors which control bulk chlorine decay rates. *Water Research, 34*(1), 117–126. https://doi.org/10.1016/S0043-1354(99)00097-4

Powell, James C., West, J. R., Hallam, N. B., Forster, C. F., and Simms, J. (2000). Performance of various kinetic models for chlorine decay. Journal of Water Resources Planning and Management, 126(1), 13–20.

Python Software Foundation. (2019). Python v3.7.4. Python Software Foundation. Retrieved from https://www.python.org/

Rashid, M.-u., George, C. M., Monira, S., Mahmud, T., Rahman, Z., Mustafiz, M., Parvin, T., Bhuyian, S. I., Zohura, F., Begum, F., Biswas, S. K., Akhter, S., Zhang, X., Sack, D., Sack, R. B., & Alam, M. (2016). Chlorination of Household Drinking Water among Cholera Patients ' Households to Prevent Transmission of Toxigenic Vibrio cholerae in Dhaka , Bangladesh : CHoBI7 Trial. *American Journal of Tropical Medicine and Hygiene, 95*(6), 1299–1304. https://doi.org/10.4269/ajtmh.16-0420

Rodriguez, M. J., & Sérodes, J. B. (1998). Assessing empirical linear and non-linear modelling of residual chlorine in urban drinking water systems. *Environmental Modelling and Software, 14*(1), 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, Series A: Dynamic Meteorology and Oceanography*, *55*(1), 16–30. https://doi.org/10.1034/j.1600-0870.2003.201378.x

Salama, P., Spiegel, P., Talley, L., Waldman, R., & Street, G. (2004). Lessons learned from complex emergencies over past decade. *Correspondence to : Lancet*, 364, 1801–1813.

Shultz, A., Omollo, J. O., Burke, H., Qassim, M., Ochieng, J. B., Weinberg, M., Feikin, D. R., & Breiman, R. F. (2009). Cholera outbreak in Kenyan Refugee Camp: Risk Factors for Illness and Importance of Sanitation. *American Journal of Tropical Medicine and Hygiene, 80*(4), 640–645. https://doi.org/10.4269/ajtmh.2009.80.640

Sikder, M., String, G., Kamal, Y., Farrington, M., Rahman, A. S., & Lantagne, D. (2020). Effectiveness of water chlorination programs along the emergency-transition-post-emergency continuum: Evaluations of bucket, in-line, and piped water chlorination programs in Cox's Bazar. *Water Research, 178,* 115854. https://doi.org/10.1016/j.watres.2020.115854

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. *Journal of Water Supply: Research and Technology - AQUA, 60*(1), 51–60. https://doi.org/10.2166/aqua.2011.086

Sphere Association. (2018). The Sphere Handbook: Humanitarian Charter and Minimum Standards in Humanitarian Response (4th ed.). Geneva: Practical Action Publishing. Retrieved from www.spherestandards.org/handbook

Steele, A., Clarke, B., & Watkins, O. (2008). Impact of jerry can disinfection in a camp environment - Experiences in an IDP camp in Northern Uganda. Journal of Water and Health, 6(4), 559–564. https://doi.org/10.2166/wh.2008.072

Swerdlow, D.L. Malenga, G., Begkoyian, G., Nyangulu, D., Toole, M., Waldman, R. J., Puhr, D. N. D., & Tauxe, R. V. (1997). Epidemic cholera among refugees in Malawi, Africa: treatment and transmission. *Epidemiology and Infection, 118*(3), 207–214. https://doi.org/https://doi.org/10.1017/S0950268896007352

Toole, M. J., & Waldman, R. J. (1997). The Public Health Aaspects of Complex Emergencies and Refugee Situations. Annual Review of Public Health, 18, 283–312.

UNHCR. (2020). WASH Manual - Pratical Guidance for Refugee Settings.

UNICEF. (2015). Azraq, Jordan WASH Monitoring Reports 2014 and 2015. Amman.

Vasconcelos, J. J., Rossman, L. A., Grayman, W. M., Boulos, P. F., and Clark, R. M. (1997). Kinetics of chlorine decay. Journal of the American Water Works Association, 89(7), 54–65. https://doi.org/10.1002/j.1551-8833.1997.tb08259.x

Walden, V. M., Lamond, E. A., and Field, S. A. (2005). Container contamination as a possible source of a diarrhoea outbreak in Abou Shouk camp, Darfur province, Sudan. Disasters, 29(3), 213–221. https://doi.org/10.1111/j.0361-3666.2005.00287.x

Wang, X., and Bishop, C. H. (2005). Improvement of ensemble reliability with a new dressing kernel. Quarterly Journal of the Royal Meteorological Society, 131(607), 965–986. https://doi.org/10.1256/qj.04.120

Warton, B., Heitz, A., Joll, C., and Kagi, R. (2006). A new method for calculation of the chlorine demand of natural and treated waters. Water Research, 40(15), 2877–2884. https://doi.org/10.1016/j.watres.2006.05.020

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

Wu, H., and Dorea, C. C. (2020). Towards a predictive model for initial chlorine dose in humanitarian emergencies. Water (Switzerland), 12(5). https://doi.org/10.3390/w12051506

Chapter 4 Cost-Sensitive Learning

4.1 Chapter Preamble

This chapter presents an investigation into the use of alternate cost functions and cost-sensitive learning for improving the dispersion of ANN ensemble forecasts of point-of-consumption FRC. This was informed by the findings of the proof-of-concept paper which found ANN ensemble forecasts of point-of-consumption FRC to be underdispersed which was attributed to the use of MSE as a training function (see Chapter 3 for more details).

A modified version of this chapter is currently being prepared for publication. The supplemental information and appendices for this chapter are included in Appendix D. Cost-sensitive learning is an approach to modifying the cost function for a data-driven model to prioritize performance in specific regions of the output space. While cost-sensitive learning is relatively common for data-driven approaches to classification, we investigated the use of cost sensitive learning primarily with the intent of improving the dispersion and reliability of ANN ensemble forecasts.

The research presented in this chapter found that ANN ensembles trained using the Kling-Gupta Efficiency (KGE) weighted with inverse frequency weighting performed best across multiple ensemble verification metrics at multiple sites using multiple variable combinations. This is likely due to the KGE cost function improving probabilistic performance as it evaluates distribution properties directly, which is aided by inverse frequency weighting, which prioritizes equal performance in all regions of the output space. Thus, in this probabilistic ensemble modelling aspect, where the objective was to produce models that reproduce the full distribution of the underlying observations, using KGE with inverse frequency weighting produced the best performance because the cost function and weighting both promote model behaviour that appropriately cover the output space. This highlights both a useful approach for modelling post-distribution FRC, and also highlights that for ANN applications in general, it is important to select a cost function that aligns with the intended model behaviour. However, while training the ANNs in the ensemble with KGE with inverse frequency weighting was a substantial improvement over the baseline models trained with MSE, the resulting models were still underdispersed, highlighting the need for additional improvement.

As the lead author I was responsible for the conception of the study presented in this chapter, as well as all model development and analysis. I was also responsible for preparing the manuscript. Dr. Usman Khan was responsible for modelling supervision and manuscript preparation. Dr. Syed Imran Ali was responsible for supporting data collection at all sites, coordination of partners, securing funding, and manuscript review. Jean-François Fesselet was responsible for coordination of partners, securing funding, and manuscript review. Matt Arnold was responsible for leading data collection in Bangladesh and supporting data collection in Tanzania and Nigeria, coordination of partners and manuscript review. Dawn Taylor was responsible for leading data collection in Nigeria and manuscript review. Anne Hyvaerinen was responsible for leading data collection in Tanzania and manuscript review.

4.2 Abstract

The Safe Water Optimization Tool (SWOT) uses ensembles of artificial neural networks (ANNs) to produce probabilistic, risk-based free chlorine residual (FRC) targets to prevent household recontamination of drinking water in refugee and internally displaced person (IDP) settlements. This probabilistic approach is taken to account for the high degree of variability and uncertainty in chlorine decay during the post-distribution period of collection, transport, and household storage. However, the typical error metrics used to train these ensembles produce underdispersed forecasts which reduce the reliability of ensemble forecasts and leads to ensembles underpredicting the risk of having insufficient point-of-consumption FRC. Alternative cost functions and cost function weightings are common approaches used to overcome the shortcomings of typical error metrics for regression and classification problems in machine learning, however these techniques have not been applied for training members of probabilistic ensembles. This research investigated the effectiveness of alternative cost functions and cost function weightings for improving probabilistic forecasts of ensembles of ANNs, like those used on the SWOT. We investigated the impact of using four cost functions and three weighting schemes on ensemble forecast dispersion and reliability using three datasets from refugee settlements in Bangladesh, Tanzania, and Nigeria. We found that training the ANN base learners with Kling Gupta Efficiency weighted with inverse frequency weighting produced the best performance across multiple sites, variable combinations, and performance metrics, leading to substantial improvements in dispersion and reliability. Incorporating these findings into the SWOT can substantially improve forecasts of point-of-consumption FRC which can in turn lead to more effective water chlorination targets that better protect water against household recontamination in refugee and internally displaced person (IDP) settlements.

4.3 Introduction

Providing safe drinking water is critical in humanitarian response as waterborne illnesses are a leading cause of excess morbidity and mortality in refugee and internally displaced person (IDP) settlements (Cronin et al., 2008). In these settings, water users typically do not have drinking water piped into the household; instead, they typically collect water from public water distribution points (the point-of-distribution) which they then transport and store in the dwelling.

Providing safe drinking water in these contexts can be challenging as waterborne pathogens can spread through recontamination of previously safe drinking water during the post-distribution period of collection, transport, and household storage and use. Recontamination of previously safe drinking water has been identified as contributing factor in outbreaks of cholera, hepatitis E, and shigellosis in refugee and IDP settlements in Kenya (Golicha et al., 2018; Shultz et al., 2009), Malawi (Swerdlow et al., 1997), Sudan (Walden et al., 2005), South Sudan (Ali et al., 2015; Guerrero-Latorre et al., 2016), and Uganda (Howard et al., 2010; Steele et al., 2008).

Residual chlorine protects drinking water against pathogenic recontamination during the post-distribution period by inactivating pathogens as they are introduced to the water. Globally used guidelines and past research both recommend a free residual chlorine (FRC) concentration of at least 0.2 mg/L to prevent pathogenic recontamination of drinking water (CDC, 2012; Girones et al., 2014; Lantagne, 2008; Rashid et al., 2016; Sikder et al., 2020; WHO, 2011), but current drinking water quality guidelines for humanitarian response only provide sufficient FRC at the point-of-distribution and do not account for the loss of chlorine residual through post-distribution chlorine decay (Ali et al., 2015, 2021). Loss of chlorine residual during the post-distribution period leads to an increased risk of recontamination while water is stored and used in the dwelling (Ali et al., 2015, 2021; De Santi, Khan, Arnold, Fesselet, & Ali, 2021; Wu & Dorea, 2020). Thus, to protect drinking water against recontamination, water system operators must provide a residual chlorine concentration at the point-of-distribution that can ensure adequate (i.e., at least 0.2 mg/L) FRC at the point-of-consumption. To determine this point-of-distribution FRC concentration, we need models that can predict the point-of-consumption FRC concentration using data collected at the point-of-distribution.

Modelling point-of-consumption FRC can be challenging due to the high degree of variability and uncertainty in post-distribution FRC decay. This uncertainty stems from the numerous quantifiable and unquantifiable factors that may impact post-distribution FRC decay, ranging from water quality factors (e.g., water temperature, reactant concentrations), to water handling factors which impact the degree of user interactions with the water (e.g., container type, drawing method). Thus, for any set of conditions at the point-of-distribution, a range of point-of-consumption FRC concentrations is possible. Given this, deterministic models of post-distribution FRC decay are inadequate as they cannot quantify the uncertainty associated with

the predicted FRC concentration. Instead, probabilistic forecasts of the point-of-consumption FRC are needed as they can account for the uncertainty in FRC decay behaviour. These probabilistic forecasts can then be extended to develop risk-based FRC targets for the point-of-distribution based on the probabilities of having sufficient or insufficient FRC at the point of consumption (De Santi et al., 2021).

A common approach to generating probabilistic forecasts is through ensemble modelling. Ensemble models group the predictions of multiple deterministic models into a probability distribution (the forecast). Probabilistic ensemble forecasting often use physical or process-based models, however, ensembles of artificial neural networks (ANNs), a type of data-driven model, have also been used for probabilistic forecasting in hydrology (Boucher et al., 2009). This probabilistic ensemble ANN approach is also used to model point-of-consumption FRC concentrations in refugee and IDP settlement by the Safe Water Optimization Tool (SWOT), an analytical tool which generates evidence-based, site-specific FRC guidance for water system operators in humanitarian response settings.

A challenge in using ensemble models, especially ensembles of ANNs, for probabilistic forecasting is that the resulting forecasts tend to be underdispersed, meaning that the spread of the forecast is smaller than the spread of the observed data (Boucher et al., 2009; De Santi et al., 2021). For ANN ensembles, this may be due to the use of typical error metrics, such as mean squared error (MSE) as cost functions. ANNs trained with MSE tend to produce predictions that regress to the mean (Crone et al., 2005; Toth, 2016). This regression to the mean behaviour tends to lead to predictions clustering near the centre of the distribution of observed values, leading to the forecasts being underdispersed, as was identified in De Santi et al. (2021) (Chapter 3). This underdispersion is shown in Figure 4-1, which shows forecasts of point-of-consumption FRC for ensembles of ANNs trained with MSE. The ensemble forecasts in Figure 4-1 are clustered towards the centre of the distribution and fail to capture a large number of observations, including many observations with insufficient point-of-consumption FRC. This means both that the ensemble forecasts do not match the underlying distribution of the observed data, and that the model performs poorly on the observations with the greatest health risk (observations with low or insufficient FRC). Together these reduce the accuracy and utility of the risk-based FRC targets generated from the forecasts of these ANN ensembles.

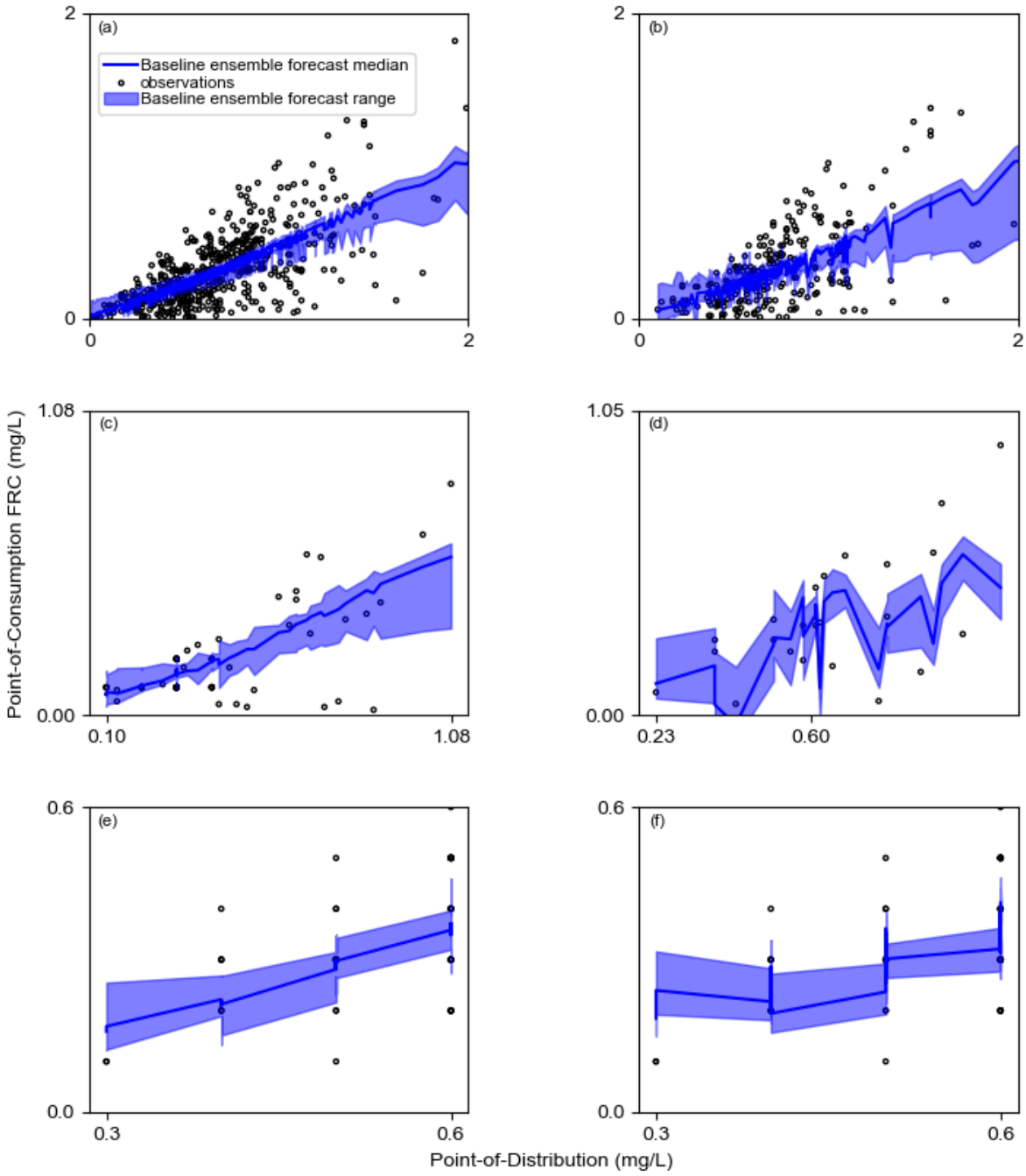

Figure 4-1: Ensemble forecasts where ANN base-learners trained with MSE for two input variable combinations (IV1 and IV2, see Section 4.4.3). Forecast-observation pairs shown for: (a) Bangladesh IV1, (b) Bangladesh IV2, (c) Tanzania IV1, (d) Tanzania IV2, (e) Nigeria IV1, and (f) Nigeria IV2 These figures show that ensembles trained with MSE are highly underdispersed as the forecast range does not cover all observations.

Instead of using typical error metrics, the cost function used to train ANN models should be selected to best reflect the intended function of the model (Crone et al., 2005). There is little research, to the author's knowledge, into the selection of appropriate cost functions for training probabilistic ensembles of ANNs. For deterministic, regression-type machine learning problems, common alternatives to typical error metrics as cost functions are alternative performance metrics (de Vos & Rientjes, 2008), asymmetric cost functions (Crone et al., 2005; Toth, 2016), or multiplying the cost function by a weighting factor to prioritize performance for specific regions of the output space (Almeida, Castel-Branco, & Falcão, 2002; Kneale, See, & Smith, 2001). This weighting approach is also common in classification machine-learning problems where it is typically referred to as cost-sensitive learning which weighs the cost-functions to prioritize model performance on high-importance classes. The weightings can either be sample-based, which associates a specific weight to each sample based on its output value, or class-based weightings, which weight each sample based on a defined output class (Krawczyk, 2016; Zhou & Liu, 2010). Rescaling is an extension of class-based weighting which is used for imbalanced classification problems (problems with very unequal distributions of data between classes) by assigning weights to each class to counteract the data imbalance the classes to ensure that all classes are equally prioritized (Ling & Sheng, 2008; Liu & Zhou, 2006; McCarthy, Zabar, & Weiss, 2005; Zhou & Liu, 2010).

In this study we sought to identify appropriate cost functions and cost function weighting approaches for training ANN ensembles to overcome underdispersion and improve the reliability of probabilistic forecasts of point-of-consumption FRC. We achieved this by evaluating the probabilistic performance of ANN ensemble forecasts using a variety of cost-functions and cost-sensitive learning approaches, using water quality data collected through the SWOT project from three refugee settlements in Bangladesh, Nigeria, and Tanzania. The objectives of this study were to:

- Investigate the impact of alternative cost functions and cost weighting on forecast underdispersion and the ability of ensemble forecasts to capture the full range of observed concentrations of point-of-consumption FRC, especially observations where point-of-consumption FRC is below the 0.2 mg/L, the FRC threshold required to protect against recontamination, evaluated using Percent Capture.

- Investigate the impact of alternative cost functions and cost weighting on forecast reliability (similarity between observed and forecast distributions) using the confidence interval (CI) reliability score, rank histogram δ-score, Continuous Ranked Probability Score (CRPS) and the CRPS reliability component.

- Identify a preferred cost function and weighting combination alternative that best meets the above two objectives. The preferred alternative should demonstrate consistently good performance for a range of sites and variable combinations to ensure the broader applicability of the selected alternative in future sites and with a range of input variable combinations.

To ensure that the above objectives are applicable to practical application for the SWOT project, we tested them using multiple input variable combinations and using multiple testing approaches to simulate different operational conditions (c.f. Sections 4.4.1 and 4.4.3). Achieving these objectives will contribute to improved probabilistic forecasts for use in the SWOT which, in turn, will assist water system operators in refugee and IDP settlements in ensuring that FRC remains sufficient to prevent pathogenic recontamination up to the point-of-consumption.

4.4 Methods

4.4.1 Description of Study Sites and Data Sets Used

This study used data from three refugee and IDP settlements in Bangladesh, Tanzania, and Nigeria, which were collected through the SWOT project by field teams from Médecins sans Frontières (MSF) and the United Nations High Commissioner for Refugees (UNHCR) and the Norwegian Refugee Council (NRC). All three sites used groundwater that was centrally treated and then piped to public water distribution points, typically a tap stand. At each site, water quality was measured directly from the public water distribution point (point-of-distribution), immediately prior to collection by a water user. The water quality parameters collected at the point-of-distribution were FRC, electrical conductivity (EC), and water temperature. The FRC for that same unit of water was measured again in the dwelling (point-of-consumption) after a follow up period of ranging from a few hours to over 24 hours later. Note that we refer to a unit of water instead of a specific container as in some cases water may be collected in one container and then transferred to a different container for storage. Thus, each observation collected consists

of two paired water quality measurements from the point-of-distribution and point-of-consumption, and timestamps provided with each measurement allowed us to calculate the elapsed time between measurements. Appendix D.1 provides the data cleaning rules that were used to prepare the dataset for use by the ANN models and histograms of the input variables are provided in Figures D-1 through D-3 in Appendix D.

4.4.2 Ethics

The studies in Bangladesh, Tanzania, and Nigeria received approval from the 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). All water quality samples were collected only when informed consent was provided from the water user.

4.4.3 Ensemble Model Building

The ensemble base learners in this study are the individual ANN models which use the multilayer perceptron (MLP) structure with one hidden layer as this type of ANN has been shown to outperform other ANN architectures for predicting extreme values of FRC in piped distribution systems (Gibbs et al., 2006; Rodriguez & Sérodes, 1998). For each site, two input variable combinations were considered: the first input variable combination (IV1) included only point-of-distribution FRC and elapsed time, representing the minimum water quality data which are regularly collected in humanitarian response settings, and the second input variable combination (IV2) included point-of-distribution FRC, EC, water temperature, and elapsed time, representing a wider set of water quality parameters that may be collected in humanitarian settings. We included two input variable combinations as the variables available for forecasting point-of-consumption FRC may vary between sites. In particular, the smaller IV1 input variable combination may be more feasible in sites with constraints on time or equipment for data collection. To ensure the broader operational applicability of this study, we needed to understand if the ability of different cost functions and weighting combinations to meet this study's objectives depends on the number of input variables.

The hyperparameters of the base learners were selected through an exploratory analysis informed by the processes documented in De Santi et al. (2021). The hidden layer size was selected by

successively doubling the hidden layer size and then selecting hidden layer size where performance began to plateau or decrease. From this exploratory analysis, the selected hidden layer size for the Tanzania and Nigeria IV1 models was four hidden nodes and eight hidden nodes for the IV2 models, respectively. In Bangladesh, we selected a hidden layer size of 16 hidden nodes for both the IV1 and IV2 models. The results of this exploratory analysis are presented in Figures D4 through D9 in Appendix D. The hyperbolic tangent activation function was used in the hidden layer and a linear activation function was used for the output layer. The *Nadam* training algorithm with a learning rate of 0.1 was used to optimize the weights and biases. To prevent overfitting, training was stopped using an early stopping procedure with a patience of ten epochs.

The ensemble size (number of base learners in an ensemble) was also selected through an exploratory grid search analysis where ensemble sizes of 50 to 500 were investigated, the results of which are included in Figures D-10 through D-15 in Appendix D. Performance typically increased with increasing ensemble size, though the performance tended to plateau for ensemble sizes between 100 and 200. Thus, for all sites and variable combinations, an ensemble size of 200 ANNs was selected to ensure that the ensemble size did not constrain performance at any site while avoiding the additional computational time associated with larger ensemble sizes.

When developing an ensemble model for probabilistic forecasting it is critical that the base learners are sufficiently distinct from each other so that the resulting forecast accurately quantifies the uncertainty in the underlying behaviour (Bröcker, 2012; Hamill, 2001). In ANN ensembles, this difference between base learners is referred to as ensemble diversity, which can represent a variety of differences ranging from differences in the model parameter (weights and biases) to differences in the model structure or hyperparameters, and even inclusion of different types of models within an ensemble (Brown et al., 2005). In this study, ensemble diversity was only introduced in the weights and biases of the individual ANNs and was promoted using implicit techniques in two ways. First, the initial weights and biases were randomized so that each base learner was trained starting at a different location on the error surface. Additionally, each base learner was trained on a different subset of the overall dataset data each time (c.f. Section 4.4.4). These approaches allowed us to obtain a diverse ensemble while ensuring that the base learners remain independent of each other.

4.4.4 Data Division for Scenario Analysis

We used two different data-division approaches to evaluate the ensemble model performance – the first using conventional random sampling, and the second to simulate time-series analysis. We used two separate approaches to evaluating model performance for several reasons. First, having two separate approaches allowed us to evaluate impact of the different cost function and weighting combinations on model performance under difference circumstances, allowing for a more robust evaluation. Secondly, while the time-series analysis represents a use of the model that is much closer to the operational use, each time interval constitutes a new test dataset making rapid comparison of different cost function and weighting alternatives very complex. By contrast, the first approach, which uses conventional random sampling, produces a single test dataset, which makes a rapid evaluation of the models much simpler. Thus, the first approach was needed to quickly identify high-performing cost function and weighting combinations that were worth investigating in more detail with the time-series analysis.

The first approach to data division was a conventional random sampling approach where we used 25% of the dataset for testing and used the rest for calibration. We further subdivided the calibration dataset into training and validation datasets, with 66.7% of the calibration data (50% of the overall dataset) used for training, and 33.3% of the calibration data (25% of the overall dataset) used for validation. The testing dataset was the same for all base learners, though the division of the calibration set into training and validation subsets was randomized for each base learner. This allowed us to maintain a static division of data between calibration and testing while still promoting ensemble diversity by training each base learner on a different random subset of the data. Using this first approach, we identified a list of the highest-performing cost function and weighting combinations. Targeted resampling techniques such as over/under-sampling were not used in this study as this leads to the ensemble predictions no longer representing a random variable which violates the assumptions of some of the ensemble verification metrics listed in Section 4.4.7 (Ferro, 2014; Hamill, 2001). A summary of the data, including descriptive statistics, for the input variables for calibration and testing data are compared in Table 4-1. The descriptive statistics for an input variable may change between input variable combinations because records with missing data were removed for each model and in some cases more records had to be removed for the IV2 input variable combination due to the greater chance of encountering missing data with a larger set of parameters.

Table 4-1: Input and output variable mean, median, and standard deviations for all sites and input variable combinations for the calibration and testing datasets. Note that the same variable at the same site may have different statistics between the two input variable combinations due to observations being removed for missing.

		Calibration				Testing			
		Number of Observations	**Mean**	**Median**	**Standard Deviation**	**Number of Observations**	**Mean**	**Median**	**Standard Deviation**
Bangladesh IV1	Point-of-distribution FRC (mg/L)	1,597	0.71	0.66	0.38	533	0.70	0.64	0.38
	Elapsed Time (h)		10.02	6.70	5.04		9.66	6.67	4.93
	Point-of-consumption FRC (mg/L)		0.34	0.28	0.28		0.34	0.28	0.28
Bangladesh IV2	Point-of-distribution FRC (mg/L)	728	0.74	0.67	0.38	244	0.77	0.69	0.41
	Elapsed Time (h)		10.27	6.80	5.07		10.18	6.88	5.00
	EC (μs/cm)		329	308	68.85		334	310	64.08

		Calibration				Testing			
		Number of Observations	**Mean**	**Median**	**Standard Deviation**	**Number of Observations**	**Mean**	**Median**	**Standard Deviation**
	Water Temperature (°C)		27.51	27.80	1.49		27.52	27.70	1.29
	Point-of-consumption FRC (mg/L)		0.33	0.27	0.28		0.35	0.26	0.36
Tanzania IV1	Point-of-distribution FRC (mg/L)	228	0.39	0.30	0.22	77	0.39	0.30	0.21
	Elapsed Time (h)		7.35	5.65	4.96		7.18	5.60	5.51
	Point-of-consumption FRC (mg/L)		0.20	0.10	0.15		0.18	0.10	0.15
Tanzania IV2	Point-of-distribution FRC (mg/L)	66	0.60	0.61	0.23	23	0.65	0.62	0.20
	Elapsed Time (h)		11.52	8.19	5.70		11.75	8.53	5.94

		Calibration				Testing			
		Number of Observations	**Mean**	**Median**	**Standard Deviation**	**Number of Observations**	**Mean**	**Median**	**Standard Deviation**
	EC (µs/cm)		325	393	150		295	390	170
	Water Temperature (°C)		24.03	24.05	0.97		23.83	23.80	1.07
	Point-of-consumption FRC (mg/L)		0.29	0.28	0.20		0.34	0.31	0.22
Nigeria IV1	Point-of-distribution FRC (mg/L)	162	0.53	0.55	0.08	54	0.54	0.60	0.09
	Elapsed Time (h)		4.08	3.53	3.08		3.91	3.73	1.76
	Point-of-consumption FRC (mg/L)		0.31	0.30	0.11		0.33	0.30	0.12
Nigeria IV2	Point-of-distribution FRC (mg/L)	162	0.53	0.50	0.08	54	0.54	0.60	0.09

	Calibration				Testing			
	Number of Observations	**Mean**	**Median**	**Standard Deviation**	**Number of Observations**	**Mean**	**Median**	**Standard Deviation**
Elapsed Time (h)		4.08	3.53	3.02		3.91	3.73	1.76
EC (µs/cm)		270	270	11.43		266	267	17.06
Water Temperature (°C)		31.39	31.30	1.91		31.85	31.80	1.96
Point-of-consumption FRC (mg/L)		0.31	0.30	0.11		0.33	0.30	0.12

In the second approach, we performed a time-series analysis for the high-performance cost function and weighting combinations using the Bangladesh data where we retrained and retested the performance of the models with progressively more data to simulate the use of these models in an operational setting where additional data would be collected over time. For this analysis, we divided the Bangladesh dataset into two-week increments. The ANN ensembles were initially trained with the data from the first two-week increment and tested on the data from the second increment. The data from the second two-week increment was then added to the calibration subset, the model was retrained on this larger calibration set and was then tested using the data from the third two-week increment. This process was repeated, increasing the size of the calibration set in two week increments and testing the models on the upcoming two-week period, until all but the last two-week increment was included in the calibration dataset. The two-week increment was selected to align with reporting requirements for water system operators at the sites included in this study. As with the first approach, 66.7% of the calibration dataset was used for training and 33.3% was used for validation, with the division of the calibration set into training and validation subsets randomized for each base learner to promote ensemble diversity. Table 4-2 summarizes the characteristics of the test data set for the two input variables in Bangladesh for each two-week period.

Table 4-2: Input and output variable mean, median, and standard deviations for Bangladesh data in each two-week period

			Week from Start of Data Collection										
			2	**4**	**6**	**8**	**10**	**12**	**14**	**16**	**18**	**20**	**22**
IV1	Samples collected in two-week period		179	199	210	208	214	208	214	101	179	192	53
	Mean	Point-of-distribution FRC (mg/L)	0.63	0.67	0.61	0.78	0.64	0.60	0.72	0.89	0.86	0.72	0.68
		Elapsed Time (h)	9.51	10.01	10.14	10.06	9.90	9.96	9.58	9.27	9.66	10.02	11.57
		Point-of-consumption FRC (mg/L)	0.30	0.34	0.29	0.35	0.29	0.27	0.34	0.43	0.44	0.31	0.30
	Median	Point-of-distribution FRC (mg/L)	0.67	0.65	0.60	0.68	0.61	0.56	0.62	0.84	0.85	0.68	0.65
		Elapsed Time (h)	6.45	6.85	6.94	6.78	6.77	6.70	6.48	5.83	6.08	6.87	15.97
		Point-of-consumption FRC (mg/L)	0.28	0.26	0.25	0.29	0.23	0.23	0.26	0.38	0.34	0.30	0.28
	Standard Deviation	Point-of-distribution FRC (mg/L)	0.30	0.34	0.31	0.45	0.30	0.32	0.45	0.36	0.42	0.30	0.20
		Elapsed Time (h)	4.74	4.75	4.92	4.94	5.07	5.07	4.97	5.38	5.08	5.23	5.24
		Point-of-consumption FRC (mg/L)	0.22	0.26	0.23	0.37	0.25	0.18	0.25	0.26	0.35	0.16	0.13
IV2	Samples collected in two week period		121	102	107	104	106	95	71	48	46	43	13
	Mean	Point-of-distribution FRC (mg/L)	0.70	0.70	0.67	0.87	0.58	0.72	0.77	0.90	0.93	0.79	0.58
		Elapsed Time (h)	9.89	10.3	10.2	10.1	10.0	10.2	8.4	12.4	10.4	10.4	16.4
		EC (µs/cm)	346	332	318	327	338	333	317	297	330	320	348
		Water Temperature (°C)	28.1	28.1	27.6	28.1	27.8	28.1	27.7	26.8	25.2	24.3	22.4
		Point-of-consumption FRC (mg/L)	0.33	0.34	0.32	0.39	0.20	0.27	0.29	0.34	0.47	0.42	0.32
	Median	Point-of-distribution FRC (mg/L)	0.69	0.66	0.62	0.70	0.58	0.69	0.69	0.85	0.92	0.68	0.53
		Elapsed Time (h)	6.72	7.03	6.95	6.78	7.02	6.93	6.25	16.46	7.44	6.33	16.85
		EC (µs/cm)	312	308	309	309	309	312	305	277	277	293	418
		Water Temperature (°C)	28.1	28.2	27.7	28.0	27.8	28.30	27.80	27.05	25.15	24.30	22.30
		Point-of-consumption FRC (mg/L)	0.33	0.26	0.28	0.25	0.15	0.25	0.24	0.29	0.47	0.32	0.31
	Standard Deviation	Point-of-distribution FRC (mg/L)	0.29	0.31	0.30	0.57	0.21	0.24	0.36	0.40	0.32	0.36	0.19
		Elapsed Time (h)	4.76	4.73	4.86	4.87	5.16	5.20	4.42	5.46	5.24	5.78	3.10

	Week from Start of Data Collection										
	2	**4**	**6**	**8**	**10**	**12**	**14**	**16**	**18**	**20**	**22**
EC (μs/cm)	68.5	63.8	55.3	58.6	65.2	83.5	62.0	67.8	86.1	72.4	123.6
Water Temperature (°C)	0.57	0.54	1.28	0.61	3.53	1.50	0.51	1.24	1.28	1.28	0.50
Point-of-consumption FRC (mg/L)	0.23	0.28	0.24	0.13	0.13	0.15	0.16	0.22	0.32	0.24	0.15

4.4.5 Cost Functions

This section describes the different cost functions used for training the base learners of the ensemble models. The cost function used to train an ANN determines the behaviour that the ANN learns from the underlying data and as such, should be selected to align with the priorities of the modelling task (Crone et al., 2005). As such, we selected cost functions that prioritize matching the spread or distribution of the underlying data in an attempt to alleviate underdispersion.

Throughout this section the notation O and P refer to the set of observed and predicted point-of-consumption FRC concentrations, respectively, and o_i and p_i refers to the i^{th} observed and predicted point-of-consumption FRC concentration, respectively. Note that in this section, a prediction refers to the output of one base learner in the ensemble.

Mean Squared Error

MSE is a measure of the average squared error between each predicted value and the corresponding observation. It is negatively oriented (lower scores are preferable) with a lower bound of 0 and no upper bound. It is primarily used in this study as a reference benchmark against which other cost functions and weightings are assessed. MSE is calculated using Equation 4-1:

$$MSE = \frac{\sum_{i=1}^{N}(p_i - o_i)^2}{N} \tag{4-1}$$

Nash Sutcliffe Efficiency

The Nash Sutcliffe Efficiency (NSE) is a common model performance metric for hydrological models. Functionally, NSE is the MSE normalized about the variance of the observations and can be understood as the amount of observed variance explained by the model (Gupta, Kling, Yilmaz, and Martinez, 2009). NSE is positively oriented (higher scores are preferable) with an upper limit of 1 and no lower bound. A naïve model which simply estimates the mean observation will have an NSE of 0. NSE is calculated using Equation 4-2. Since the *Nadam* optimizer can only find the minimum of a function, when the NSE was used to train the base learners of the ensemble, it was multiplied by -1 to convert it to a negatively oriented score with a lower bound of -1 and no upper bound.

$$NSE = 1 - \frac{\sum_{i=1}^{N}(p_i - o_i)^2}{\sum_{i=1}^{N}(o_i - \overline{O})^2} \quad (4\text{-}2)$$

Kling Gupta Efficiency

Kling Gupta Efficiency arose out of a decomposition of NSE into three components: correlation (r), the ratio of the variance of predictions to the variance of the observations (α), and the variance of the mean of predictions and the mean of the observations (β) (Gupta et al., 2009). These three components are calculated using Equations 4-3, 4-4, and 4-5 below:

$$r = \frac{\mathrm{cov}(O,P)}{\sqrt{\mathrm{cov}(O,O) * \mathrm{cov}(P,P)}} \quad (4\text{-}3)$$

$$\alpha = \frac{\sigma_P}{\sigma_O} \quad (4\text{-}4)$$

$$\beta = \frac{\overline{P}}{\overline{O}} \quad (4\text{-}5)$$

The r, α, and β terms are used to calculate the Euclidean distance from predictions to the ideal model result which would have a correlation coefficient of 1 and identical means and standard deviations as the observed data (as shown in Equation 4-6 below). KGE is then calculated by subtracting the Euclidean distance from 1 (Equation 4-7 from (Gupta et al., 2009)). Thus, KGE is a measure of the correlation between the predictions and observations, as well as a measure of the similarity of the first and second moments of the distributions of the observations and predictions. We included KGE as a cost function in this study as one of our objectives is to identify cost functions that can produce forecasts with similar distributions to the underlying data and KGE can directly promote this in each base learner. As with NSE, the KGE score is positively oriented, with higher scores representing shorter Euclidean distances from the model being tested and the ideal model result, with an upper limit of 1 and no lower limit. To convert KGE into a negatively oriented score for training the base learners, the KGE score was multiplied by -1.

$$ED = \sqrt{(r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \quad (4\text{-}6)$$

$$KGE = 1 - ED \quad (4\text{-}7)$$

Index of Agreement

The Index of Agreement (IoA) is a modified version of the NSE with a revised denominator, as shown below in Equation 4-8. The IoA specifies the degree of similarity of the deviations about the mean between the observed and predicted values (Willmott, 1981). We included IoA as a cost function in this study for this ability to assess the deviations about the mean as a way of better addressing the challenge of forecast underdispersion where the predictions tend to be less dispersed about the mean than the observations. Like NSE, IoA is positively oriented with an upper limit of 1 and a lower bound of 0. We converted IoA into a negatively oriented by multiplying the calculated score by -1.

$$IoA = 1 - \frac{\sum_{i=1}^{N}(p_i - o_i)^2}{\sum_{i=1}^{N}(|p_i - \overline{O}| + |o_i - \overline{O}|)^2} \quad (4\text{-}8)$$

4.4.6 Cost Weightings

Six weighting approaches were applied to the cost functions outlined in Section 4.4.5. These weightings were used to prioritize model performance in different regions of the outcome space. Most of the weighting schemes were applied to prioritize model performance for observations with lower point-of-consumption FRC as the lower the FRC at the point-of-consumption, the greater the risk of household recontamination. This section describes the three weighting approaches which are modified from existing approaches that have been documented in literature. Details of calculating the weighted version of the cost functions described in Section 4.4.5 are provided in Appendix D.2. In the following section, O represents the set of observed point-of-consumption FRC concentrations, o_i represents the i^{th} observed point-of-consumption FRC concentration, and w_i represents the weighting applied to the cost function for the i^{th} prediction-observation pairing.

Weighting 1

The first weighting function uses an instance-based approach to prioritize observations with low point-of-consumption FRC with the weights are taken as the inverse of the observed point-of-consumption FRC. This weighting approach was developed in considering a weighted approach to scoring flood forecasts where high flows are more critical than low flows. Kneale et al. (2021) proposed multiplying common scoring metrics by the observed flow rate as an approach to better evaluating the testing performance of flood forecasting neural network models. Since low FRC

concentrations are more critical than high FRC concentrations, in this study we modified this approach by multiplying the score by the inverse of the observed FRC concentration. This leads to a reciprocal increase in the weight as FRC decreases, as shown in Equation 4-9.

$$w_i = \frac{1}{o_i} \tag{4-9}$$

Equation 4-9 was modified for implementation in training the base learners of the ANN ensemble to account for the input and output data being normalized between -1 and 1. Using Equation 4-9 with these normalized inputs would produce two asymptotes at the median observed point-of-consumption FRC concentration. To avoid this, a fixed constant, 1.1, is added to the normalized observed value, as shown in Equation 4-10.

$$w_i = \frac{1}{o_{i_{norm}} + 1.1} \tag{4-10}$$

Weighting 2

The second weighting function used class-based weighting which is a common approach to prioritizing specific classes in cost-sensitive learning for classification problems (Krawczyk, 2016; Zhou & Liu, 2010). We developed the weighting by classifying observations based on their observed household FRC concentration and assigning a unique weight to each class. The classes were developed considering both drinking water quality guidelines for humanitarian response as well as past literature, and are listed below:

- FRC between 0 mg/L and 0.2 mg/L – these are the highest risk observations since they have insufficient FRC to prevent recontamination at the point-of-consumption.

- FRC between 0.2 mg/L and 0.5 mg/L – these are considered moderate risk as they meet drinking water quality guidelines but recontamination may still occur from bacteria with moderate chlorine tolerance, or if there are other favourable water quality parameters such as high turbidity and high organic content (M. W. LeChevallier et al., 1981; Mark W. LeChevallier et al., 1996).

- FRC concentration between 0.5 mg/L and 1.0 mg/L – this range is typically recommended during waterborne illness outbreaks and the risk of recontamination is considered low (Médecins Sans Frontières, 2010).

- FRC above 1.0 mg/L – this is beyond the range recommended even during waterborne illness outbreaks so we consider the risk of recontamination with point-of-consumption FRC above 1.0 mg/L is very low.

The weights assigned to each class were determined based on the risk of household recontamination to prioritize model performance on observations with the greatest risk. Thus, we assigned a weight of 1.0 to the highest priority class (point-of-consumption FRC below 0.2 mg/L) and halved the weight for each subsequent class.

$$w_i = \begin{cases} 1.0 \; if \; 0 \leq o_i < 0.2 \\ 0.5 \; if \; 0.2 \leq o_i < 0.5 \\ 0.25 \; if \; 0.5 \leq o_i < 1.0 \\ 0.125 \; if \; o_i \geq 1.0 \end{cases} \quad (14)$$

Weighting 3

The third weighting approach used a modification of class-based learning called rescaling, where the weights are assigned to counteract data imbalances to ensure each class is equally prioritized by the model (Ling & Sheng, 2008; Liu & Zhou, 2006; McCarthy et al., 2005; Zhou & Liu, 2010). To achieve this we set the weights for each class as the inverse of the frequency of observations in that category. This is referred to as probabilistic rescaling, or inverse frequency weighting (Ling & Sheng, 2008; Liu & Zhou, 2006; McCarthy et al., 2005; Zhou & Liu, 2010). The inverse frequency weight for the j^{th} category was calculated as:

$$w_j = \frac{number \; of \; observations \; in \; category \; j}{total \; number \; of \; observations}^{-1} \quad (16)$$

4.4.7 Performance Metrics

This section details the metrics used to evaluate the probabilistic ensemble forecasts based on the objectives listed in Section 4.3. Probabilistic forecasts were derived from the ensemble by weighting the predictions of each base learner and then combining these predictions into a probability density function (pdf). In this study each base learner was equally weighted, so the weight assigned was equal to $\frac{1}{M}$ where M is the number of base learners in the ensemble. For this study all ensembles had $M = 200$ base learners.

Since the output of the models were probabilistic forecasts, deterministic scores such as the MSE or NSE are not useful for evaluating the performance of the forecasts (Boucher et al., 2009; Hamill, 2001). Instead, scores and performance metrics were selected which could evaluate the probabilistic forecasts of each ensemble according to the objectives listed in Section 4.3. Each of the ensemble verification metrics listed below were evaluated for the test dataset only (see Section 4.4.4) to verify each ensemble's ability to forecast on new data. 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 $m = M$.

Percent Capture

Percent Capture is a measure that has been commonly used to evaluate the effectiveness of probabilistic and possibilistic models (Alvisi & Franchini, 2011, 2012; Khan & Valeo, 2016, 2017). The Percent Capture is the percentage of observations where the observed point-of-consumption FRC concentration was within the limits of the ensemble forecast. We used the Percent Capture to accomplish the first objective of this study (identify cost functions and weighting combinations that produce ensemble forecasts that capture the full range of observed concentrations of point-of-consumption FRC).

The Percent Capture is a positively oriented score, with an upper limit of 100% and a lower limit of 0%. To calculate Percent Capture, observation o_i is considered captured if $f_i^0 \leq o_i \leq f_i^M$. When evaluating the ensemble performance, we considered both the Percent Capture of the overall dataset (referred to throughout as PC) as an indicator of underdispersion, as well as the Percent Capture of observations with point-of-consumption FRC below 0.2 mg/L ($PC_{<0.2}$) which provides an indication of how well the model can accurately predict if there will be sufficient FRC at the point-of-consumption.

Reliability Metrics for Ensemble Verification

Ensemble reliability is a term commonly used in atmospheric modelling that refers to the similarity between the observed and forecasted probability distributions. These verification metrics were used to evaluate the second objective listed in Section 4.3. Numerous reliability metrics have been developed in the literature, this study uses three reliability metrics which are commonly used in atmospheric sciences and which have been applied outside of atmospheric sciences for hydrological applications (Boucher et al., 2009).

CI Reliability Diagram

The first reliability metric used for ensemble verification was the reliability diagram. The reliability diagram plots observed relative frequency of events against the forecast probability of that event. Boucher et al. (2009) adapted this diagram for ensemble modelling as the confidence interval (CI) reliability diagram which compares the frequency of observed values with the corresponding CI of the ensemble, where the ensemble CIs are derived from the sorted forecasts of the base learners (for example, the ensemble 90% CI would include all of the forecasts between $f^{0.05M}$ and $f^{0.95M}$). We extended this further by plotting the Percent Capture of each CI within the ensemble against the CI level. For each ensemble model we plotted the CI reliability for the 10% to 100% CI levels at 10% intervals as well as at the 95% and 99% CI.

The reliability diagram and CI reliability diagram are visual indicators of ensemble reliability with the ideal model having all observations plotted along the 1:1 line showing that the observed probabilities are equal to the forecasted probabilities. De Santi et al. (2021) developed a numerical score for the CI reliability diagram which calculated the squared distance between the Percent Capture within each CI and the ideal Percent Capture in that CI. This was calculated for each CI threshold, k, from 10% to 100% in 10% increments as shown in Equation 4-11. The CI Reliability Score measures the horizontal distance between the Percent Capture and the 1:1 line for each CI. 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. CI Reliability diagrams were plotted and the CI reliability score calculated for both the overall data set (referred to throughout as CI_{score}) and for forecast-observation pairs where the observed point-of-consumption FRC concentration was below 0.2 mg/L ($CI_{score_{<0.2}}$).

$$CI\ Reliability\ Score = \sum_{k=0.1}^{1}\left(j - Percent\ Capture\ in\ CI_j\right)^2 \quad (4\text{-}11)$$

Rank Histograms

The Rank Histogram (RH) is another visual tool used to assess the reliability of ensemble forecasts. To construct the rank histogram, for each forecast-observation pair, the observation o_i is added to the sorted vector of forecast values F_i, with the new vector having $M + 1$ members. A rank is assigned to the observed value based on where in the set of forecasted values it falls. This is then repeated for each forecast-observation pair. The RH is the histogram of the ranks assigned to each observation, o_i. If the forecast and observed probabilities are the same, then any observation is equally likely to occur in any of the $M + 1$ ranks, which would result in a flat rank histogram. 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, Vautard, and Strauss, 1997).

The RH is a visual tool, but Candille and Talagrande (2005) proposed a numerical score, the δ score which measures the deviations from flatness (Equation 4-12). The δ score only measures deviations from flatness, and cannot be used to diagnose over/under dispersion or flatness. The ideal score is 1 with scores much greater than 1 indicating substantial deviations from flatness and scores less than 1 indicating interdependence between ensemble predictions (Candille & Talagrand, 2005). The δ score was calculated for each model both for the overall dataset (referred to throughout as δ) and for only those observations where the observed point-of-consumption FRC was below 0.2 mg/L ($\delta_{<0.2}$).

$$\delta = \frac{\Delta}{\Delta_o} \tag{4-12}$$

The two components of the δ score are shown in Equations 4-13 and 4-14 where M is the total number of ensemble members, 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 & Talagrand, 2005).

$$\Delta = \sum_{k=1}^{M+1} \left(s_k - \frac{I}{M+1} \right)^2 \tag{4-13}$$

$$\Delta_o = \frac{I * M}{M+1} \tag{4-14}$$

Continuously Ranked Probability Score

The continuously ranked probability score (CRPS) is a commonly used metric for evaluating probabilistic forecasts. It represents a continuous version of the Brier Score (Brier, 1950) and measures the area between the forecast cumulative distribution function (cdf) and the observed cdf for each forecast-observation pairing. The CRPS measures not only model reliability but also sharpness and uncertainty which has been shown to correspond to the MAE of a deterministic forecast (Ferro, 2014; Hersbach, 2000). For a given forecast-observation pair, the cdf of the forecast is calculated from the probability distribution of the predictions of the base learners of the ensembles. Since each observation is a discrete value, it 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 predicted concentrations of point-of-consumption FRC above the observed concentration. The calculation of the CRPS is given in Equation 4-15 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 4-15 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.

$$CRPS = \int_{-\infty}^{\infty}(F_i(x) - H\{x \geq o_i\})^2 dx \tag{4-15}$$

Hersbach (2000) derived a calculation of CRPS for ensemble models that 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:

$$\overline{CRPS} = \sum_{n=1}^{N} \overline{g_n}[(1 - \overline{o_n})p_n^2 + \overline{o_n}(1 - p_n)^2] \tag{4-16}$$

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

Hersbach (2000) also decomposed the ensemble CRPS calculation into its reliability, resolution, and uncertainty scores. The reliability term is of particular interest as it reflects the similarity

between the observed and forecast probability distributions. The average reliability term ($\overline{Reli}$) is calculated as shown in Equation 4-17 (Hersbach, 2000).

$$\overline{Reli} = \sum_{n=1}^{N} \overline{g_n} \, (\overline{o_n} - p_n)^2 \tag{4-17}$$

The reliability component of the CRPS is considered by Hersbach (2000) to be a better representation of reliability than the RH δ-score because the CRPS reliability term also considers the bin widths whereas the δ-score only considers the vertical deviations from a flat RH. The Brier score reliability term (from which the CRPS reliability term is derived) can also be considered as the vertical/horizontal distance from the 1:1 line on the reliability diagram (Atger, 2004). We directly calculate this distance using Equation 4-11. However, the reliability term of the CRPS considers all probability levels of the ensemble instead of the predefined CI levels used in the CI reliability score. The relationships between these three measures of ensemble reliability are shown in Figure 4-2.

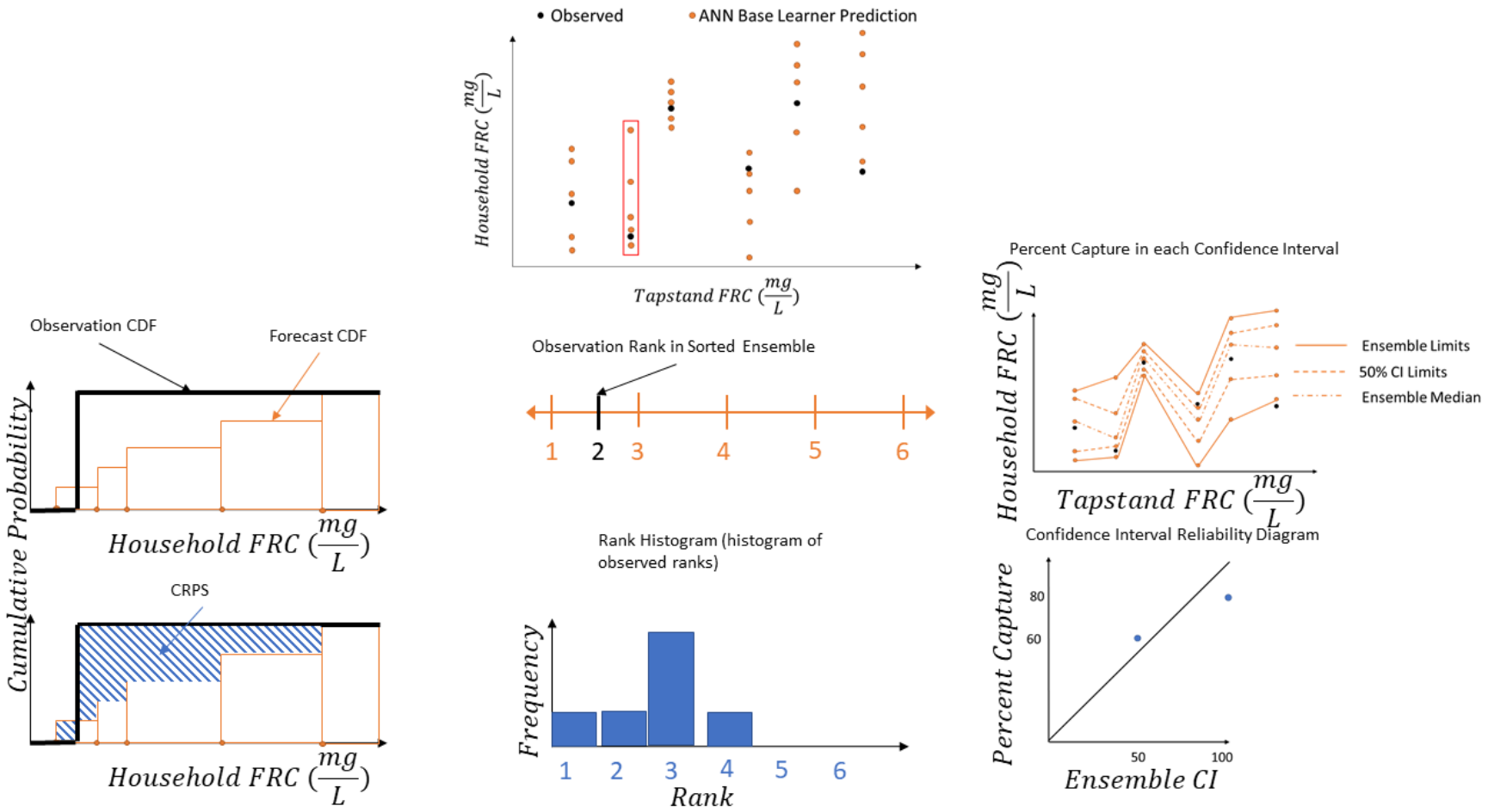

Figure 4-2: Visualization of the method used to calculate the different performance evaluation metrics used in this study for the CRPS (left), RH (centre) and CI reliability diagram (right). The CRPS is calculated from the difference in area between the forecast cdf and the Heaviside function (observation cdf), the rank histogram is derived from the rank of the observation relative to each model prediction, and the CI reliability diagram is based off the percent capture in each ensemble CI PC (not shown) would measure the

number of observations captured within the forecast bounds, whereas the RH and CI reliability diagram both provide indicators of reliability (similarity between observed and forecast distribution), whereas CRPS considers both reliability and sharpness.

Skill Scores

While in many cases, the score obtained for the performance metrics above are highly informative (e.g., Percent Capture provides clear insight into the amount of observations captured, the CI Reliability Score provides a measurable distance from the ideal model), it can also be helpful to normalize scores, both to facilitate comparison of scores between sites, and to allow for a comparison of the relative improvement between metrics with different scales. Skill scores provide a means of normalizing a performance metric using a baseline and an ideal score. For our case, we used the ensemble verification metrics obtained by the ensemble trained with unweighted MSE as the baseline for developing the skill score as MSE is a common industry standard and is currently in use on the SWOT project. The ideal score for Percent Capture is 100% capture, the ideal CI reliability score, CRPS, and CRPS reliability term are all 0, and the ideal δ-score is 1. Any performance metric can be converted to a skill score using Equation 4-18. The skill score is normalized between negative infinity and 1, with 1 meaning that the score obtained is the ideal score and a positive score indicating improvement over baseline. A skill score of 0 means that there is no difference between the score obtained and the baseline, and a negative score indicates that the score obtained is worse than the baseline.

$$Skill\ Score = \frac{score\ obtained - baseline}{ideal\ score - baseline} \tag{4-18}$$

4.5 Results and Discussion

The following sections describe and review the performance of the ensemble models trained with each cost function and weighting combination at each site with each input variable combination. Table 4-3 summarizes the cost function and weighting combinations that produced the best performance for each site with each variable combination. While the subsequent sections provide further detail on the individual performance metrics and on the identification of preferred cost function and weighting alternatives, Table 4-3 shows that the use of alternative cost functions and weighting functions led to performance improvements, with the best performance for nearly every metric being obtained by alternatives that used both alternative cost functions as well as cost weighting. The scores and skill scores obtained by the cost function and weighting combinations listed in Table 4-3, and for all other combinations, are included in Table D-1 in Appendix D.

Table 4-3: Summary of best performing cost function and weighting combination for each performance metric

Site and Input Variable combination	$\boldsymbol{PC}$	$\boldsymbol{PC_{<0.2}}$	$\boldsymbol{CI_{score}}$	$\boldsymbol{CI_{score_{<0.2}}}$	$\boldsymbol{\delta}$	$\boldsymbol{\delta_{<0.2}}$	$\boldsymbol{\overline{CRPS}}$	$\boldsymbol{\overline{Reli}}$
Bangladesh IV1	KGE weighting 3	KGE weighting 3	KGE weighting 3	KGE weighting 3	KGE weighting 3	KGE weighting 3	MSE weighting 1	MSE weighting 1
Bangladesh IV2	KGE weighting 3	KGE weighting 3	KGE weighting 3	KGE weighting 3	KGE weighting 3	KGE weighting 3	Unweighted IoA	Unweighted IoA
Tanzania IV1	IoA weighting 3	KGE weighting 1 IoA weighting 3[1]	KGE weighting 3	KGE weighting 3	KGE weighting 3	IoA Weighting 1	IoA weighting 1	IoA weighting 1
Tanzania IV2	KGE weighting 3	MSE weighting 1 NSE weighting 1[1]	KGE weighting 3	KGE weighting 1	KGE weighting 3	MSE weighting 1 NSE weighting 1 KGE weighting 1[1]	Unweighted NSE	IoA weighting 3

Site and Input Variable combination	$\boldsymbol{PC}$	$\boldsymbol{PC_{<0.2}}$	$\boldsymbol{CI_{score}}$	$\boldsymbol{CI_{score_{<0.2}}}$	$\boldsymbol{\delta}$	$\boldsymbol{\delta_{<0.2}}$	$\boldsymbol{\overline{CRPS}}$	$\boldsymbol{\overline{Reli}}$
Nigeria IV1	IoA weighting 3	MSE weightings 1 and3 NSE unweighted and weighting 1 KGE (all weightings) IoA (all weightings)[1,2]	IoA weighting 3	NSE weighting 3	IoA weighting 3	MSE weightings 1 and3 NSE unweighted and weighting 1 KGE (all weightings) IoA (all weightings)[1,2]	NSE weighting 2	NSE weighting 2
Nigeria IV2	IoA weighting 3	MSE weighting 3 NSE weighting 3 KGE weightings 1, 2, 3 IoA weightings 1, 2, 3[2]	IoA weighting 3	IoA weighting 3	IoA weighting 3	MSE weighting 3 NSE weighting 3 KGE weightings 1, 2, 3	NSE weighting 1	KGE weighting 1

Site and Input Variable combination	$\boldsymbol{PC}$	$\boldsymbol{PC_{<0.2}}$	$\boldsymbol{CI_{score}}$	$\boldsymbol{CI_{score_{<0.2}}}$	$\boldsymbol{\delta}$	$\boldsymbol{\delta_{<0.2}}$	$\boldsymbol{\overline{CRPS}}$	$\boldsymbol{\overline{Reli}}$
						IoA weightings 1, 2, 3[2]		

Notes:

[1]For some metrics at some sites, multiple ensembles achieved the same performance, so multiple best performances are possible.

[2] All of a specific cost function means that all weighting combinations for that cost function, including unweighted, achieved the same score

Sections 4.5.1 through 4.5.4 present a metric-by-metric evaluation of the impact of the alternative cost functions and cost weighting on ensemble performance, presenting both the absolute performance as well as a comparison to the baseline performance obtained by the ensembles trained using unweighted MSE. Based on these metric-by-metric evaluations, Section 4.5.5 identifies preferred cost function and weighting combinations using a frequency analysis to identify cost function and weighting combinations that produced consistently good performance across all performance metrics at each site and with each variable combination. Section 4.5.6 uses the preferred cost-function and weighting combinations identified in Section 4.5.5 to perform a time-series analysis using the Bangladesh dataset to simulate the use of the preferred combinations in an operational setting.

4.5.1 Percent Capture Performance

PC and $PC_{<0.2}$ were used to assess the first objective listed in Section 4.3, which was to investigate the impact of alternative cost functions and cost weighting on improving forecast underdispersion, with an emphasis on observations with insufficient point-of-consumption FRC (< 0.2 mg/L). The use of alternative cost functions and cost function weighting produced substantial improvements in these scores over the baseline condition of unweighted MSE, as shown in Figures 4-3 and 4-4, which show the observed and forecasted point-of-consumption FRC concentrations for the models trained with unweighted MSE compared to the model with the highest PC and $PC_{<0.2}$ at each site (Table 4-3). Figure 4-3 shows this comparison for PC and Figure 4-4 shows this comparison for $PC_{<0.2}$score. From these figures we see that the models trained with unweighted MSE converge towards the centre of the range of observations whereas the models with better PC and $PC_{<0.2}$ have a wider range of predictions. This wider spread of predictions indicates that the use of alternative cost functions and cost weighting to train the base learners of the ensembles can overcome the challenge of underdispersion that arises from using MSE as the cost function.

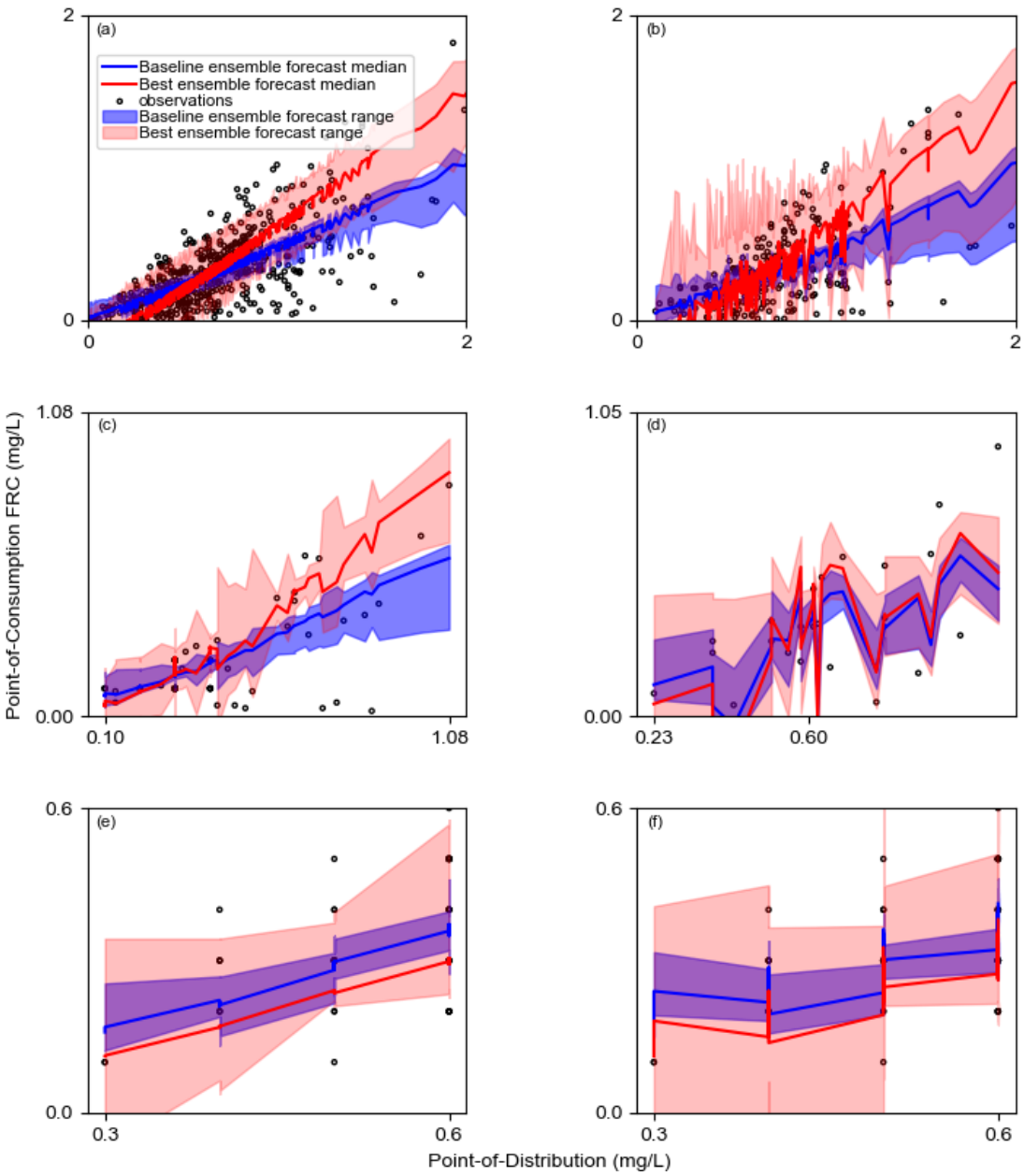

Figure 4-3: Forecast-observation comparison for all sites showing difference in forecast range and median between the baseline ensemble (trained with unweighted MSE) and ensemble with the highest PC for each site and variable combination. Forecast observation pairs shown for: (a) Bangladesh IV1, (b) Bangladesh IV2, (c) Tanzania IV1, (d) Tanzania IV2, (e) Nigeria IV1, and (f) Nigeria IV2. In all of the above subplots we see that the best performing ensemble (shown in red) produces a larger forecast range (shaded area) than the baseline ensemble (blue), which is what produced the better Percent Capture.

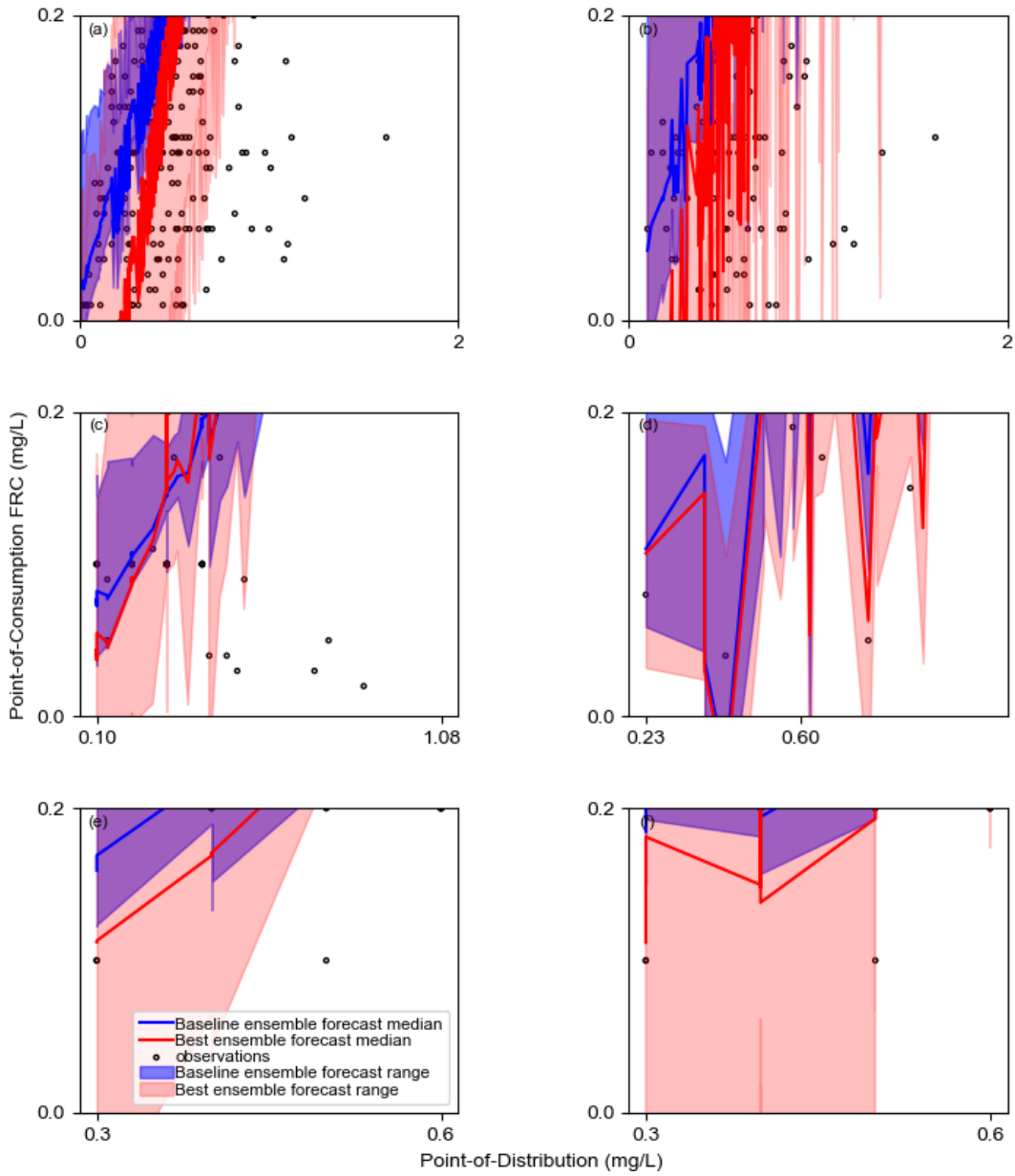

Figure 4-4: Forecast-observation comparison for all sites showing difference in forecast range and median between the baseline ensemble (trained with unweighted MSE) and ensemble with the highest $PC_{<0.2}$ for each site and variable combination. Forecast observation pairs shown for: (a) Bangladesh IV1, (b) Bangladesh IV2, (c) Tanzania IV1, (d) Tanzania IV2, (e) Nigeria IV1, and (f) Nigeria IV2. In all of the above subplots we see that the best performing ensemble (shown in red) produces a larger forecast range (shaded area) than the baseline ensemble (blue), which is what produced the better Percent Capture.

The ANN ensembles trained with unweighted MSE produced highly underdispersed forecasts, with the Percent Capture ranging from 22% to 35% for the PC and from 0% to 50% for the $PC_{<0.2}$. Using alternative cost functions and cost function weighting, the best PC at each site ranged from 61% to 85% and the best $PC_{<0.2}$ ranged from 67% to 84%. This corresponds to a maximum skill score between 0.40 and 0.80 for PC and between 0.67 and 0.79 for $PC_{<0.2}$. The improvement in Percent Capture can be seen in Figures 4-3 and 4-4. Operationally, these improvements will enhance the accuracy of the risk-based FRC targets produced by the SWOT as the ensembles trained with alternative cost functions and cost weighting, particularly the models trained with KGE weighted by inverse frequency weighting, are better able to capture observations with insufficient point-of-consumption FRC, and as such as are better able to accurately determine the point-of-distribution FRC required to ensure sufficient point-of-consumption FRC.

4.5.2 CI Reliability Performance

The CI_{score} and $CI_{score_{<0.2}}$ were used to assess the second objective listed in Section 4.3, which was to investigate the impact of cost functions and weighting schemes on the forecast reliability for the overall dataset and for observations with point-of-consumption FRC below 0.2 mg/L. As with the Percent Capture, the use of alternative cost functions and cost function weighting produced substantial improvements in both the The CI_{score} and $CI_{score_{<0.2}}$ over the baseline condition of unweighted MSE, as shown in Figures 4-5 and 4-6. Figure 4-5 compares the CI reliability diagrams using the full dataset for the baseline ensemble and for the ensemble with the lowest CI_{score} (Table 4-3) and Figure 4-6 compares the CI reliability diagram only for observations with point-of-consumption FRC below 0.2 mg/L for the baseline ensemble and for the models with the lowest $CI_{score_{<0.2}}$ (Table 4-3). From these two figures, the percentage of observations captured in each CI is consistently below the 1:1, indicating that even the best models are underdispersed. However, both Figure 4-5 and especially Figure 4-6 show that the models with lower CI_{score} and $CI_{score_{<0.2}}$ are much closer to the theoretical capture value of each confidence interval, showing that improvement in model dispersion is occurring at multiple CI's and not just for the extreme ranges of the ensemble forecasts, which means that the overall reliability is improving.

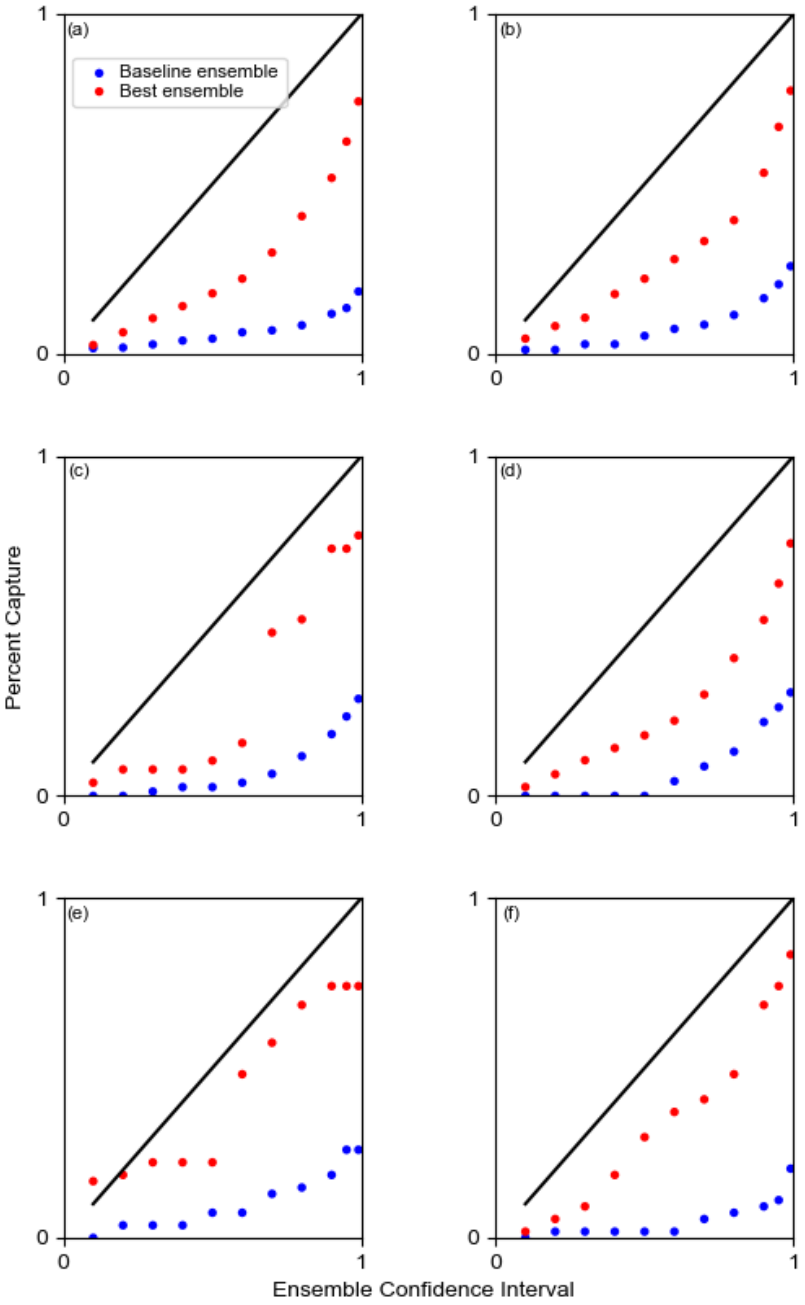

Figure 4-5: CI reliability diagrams for all sites showing difference in forecast range and median between the baseline ensemble (trained with unweighted MSE) and ensemble with the best CI_{score} score for each site and variable combination. CI reliability diagrams shown for: (a) Bangladesh IV1, (b) Bangladesh IV2, (c) Tanzania IV1, (d) Tanzania IV2, (e) Nigeria IV1, and (f) Nigeria IV2. These figures show that in all cases the baseline ensemble (trained with unweighted MSE) produced underdispersed forecasts, with the capture in all CIs below the 1:1 line.

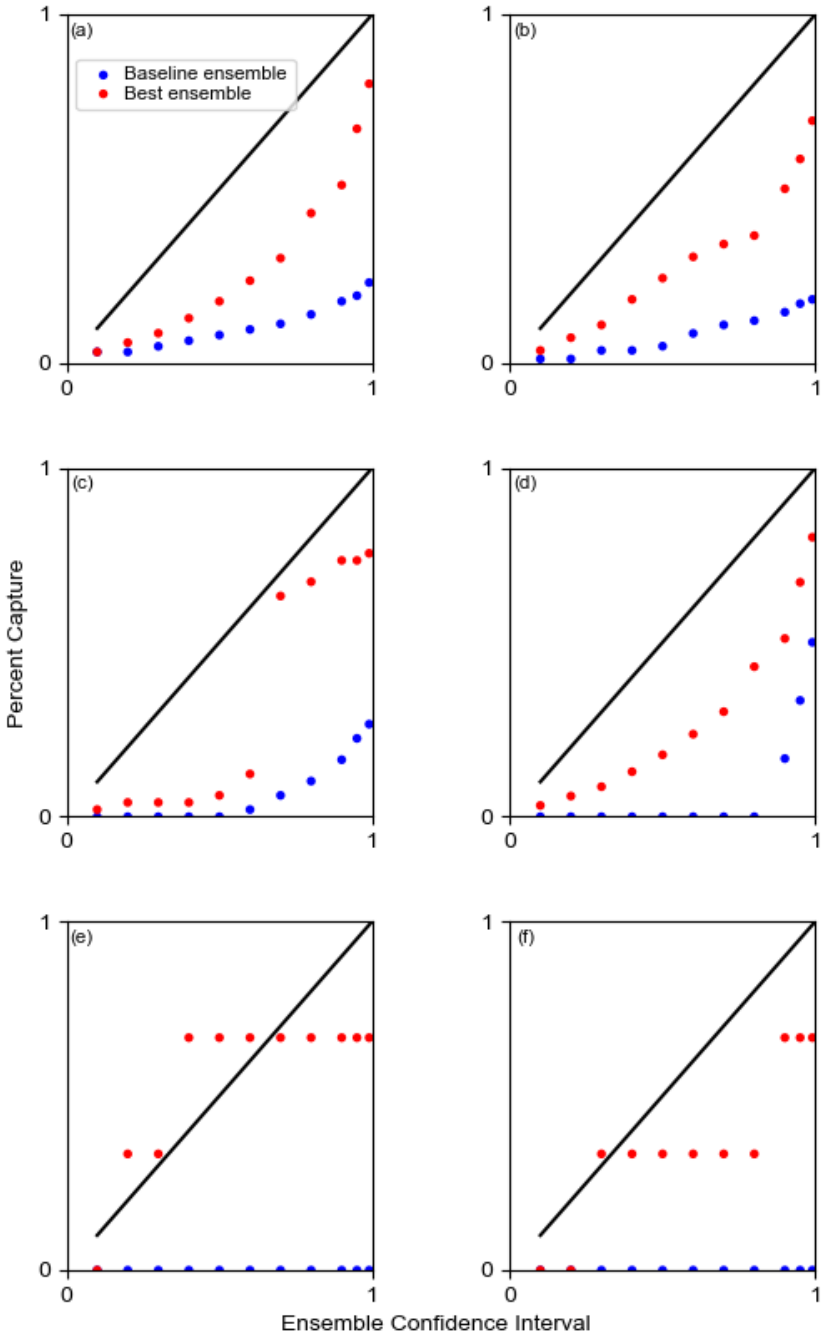

Figure 4-6: CI reliability diagrams for all sites showing difference in forecast range and median between the baseline ensemble (trained with unweighted MSE) and ensemble with the best $CI_{score_{<0.2}}$ *score for each site and variable combination. CI reliability diagrams shown for: (a) Bangladesh IV1, (b) Bangladesh IV2, (c) Tanzania IV1, (d) Tanzania IV2, (e) Nigeria IV1, and (f) Nigeria IV2.*

The CI_{score} for the ANN ensembles trained with unweighted MSE ranged from 2.51 to 3.01 for with the $CI_{score_{<0.2}}$ ranging from. By contrast, the best CI_{score} obtained using alternative cost functions and cost function weighting at each site ranged from 0.26 to 1.25, and the best $CI_{score_{<0.2}}$ ranged from 0.32 to 0.85. This corresponds to a skill score between 0.51 and 0.90 for the CI_{score} and between 0.66 and 0.92 for the $CI_{score_{<0.2}}$. This shows that the use of alternative cost functions and cost weightings greatly improve the reliability of the ensemble forecasts meaning that the predicted risk, which is based on the probability of having insufficient point-of-consumption FRC, will be more accurate. Furthermore, the CIs used in the CI reliability score also correspond to risk thresholds, so improvements in the CI reliability score indicate an improved ability to accurately forecast specific risk thresholds.

4.5.3 RH Performance

The RH and δ-score were also used to evaluate the impact of cost functions and weighting schemes on the forecast reliability. The use of alternative cost functions and cost weighting led to substantial improvements in the δ-score over the baseline condition of ensembles trained with unweighted MSE. The δ-scores for the ensembles trained with unweighted MSE ranged from 5.09 to 153 and the $\delta_{<0.2}$ ranged from 1.98 to 97 for observations with point-of-consumption FRC below 0.2 mg/L. Using alternative cost functions and cost function weightings, the best δ-score ranged from 1.94 to 14.4 and the best $\delta_{<0.2}$ scores ranged from 0.98 to 5.81. This corresponds to maximum skill scores between 0.57 and 0.93 for the δ score and between 0.83 and 1 for the $\delta_{<0.2}$ scores. However, while the δ score improved substantially through the use of alternative cost functions and cost function weighting, even the best models tended to score far from the ideal value of 1.

The δ-score is itself only a measure of deviation from a uniform RH, and alone does not indicate reasons for deviations from uniformity. Figures 4-7 and 4-8 show the Bangladesh IV1 RHs for the models trained using unweighted MSE models and the models with the lowest δ and $\delta_{<0.2}$ scores, respectively (Table 4-3). The RHs for the remaining sites are included in Figures D-16 through D-25 in Appendix D. In all cases the RHs have noticeable U-shapes, reinforcing that the ensemble forecasts are underdispersed even when using alternative cost functions and cost function weighting. However, this U-shape is less pronounced in the RHs for the models with the lowest δ-scores, and the count of observations at the extremes of the ensemble tend to be lower

in the RHs for the models with the lowest δ-scores. This indicates that the improvements in the δ-score obtained using alternative cost functions and cost function weightings occurred primarily due to the reduction in underdispersion, highlighting that by improving underdispersion, we also improve the reliability.

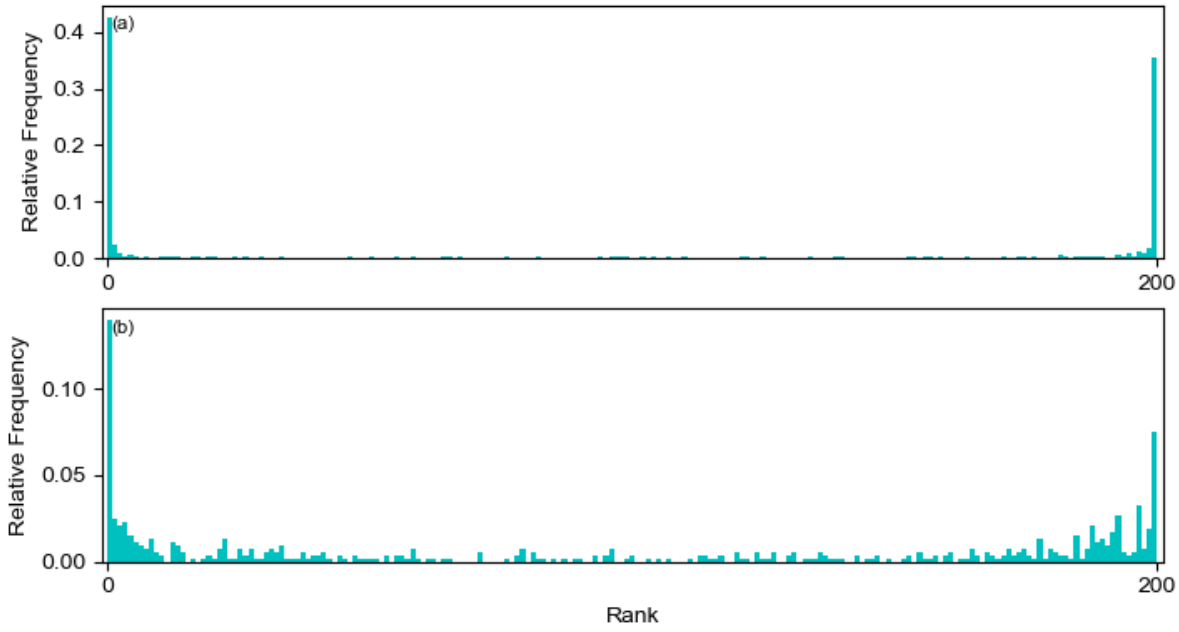

Figure 4-7: RH for Bangladesh with the IV1 variable combination. The RH in (a) shows the results for the models trained with unweighted MSE, and (b) shows the model with the best δ-. Both RHs are underdispersed, as seen from the u-shape of the RH, but the size of the outlier bars, and the difference between the outliers and internal bars are much smaller in (b) than in (a), indicating improved reliability with alternative cost functions and cost function weighting.

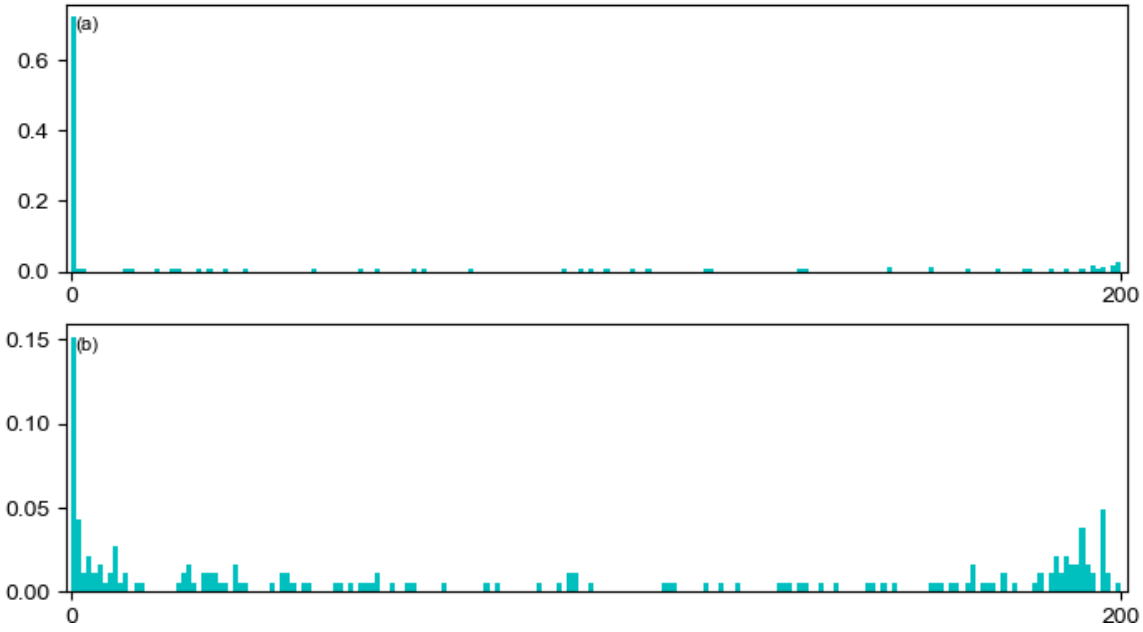

Figure 4-8: RH of observations with point-of-consumption FRC below 0.2 mg/L for Bangladesh with the IV1 variable combination. The RH in (a) shows the results for the models trained with unweighted MSE, and (b) shows the model with the best $\delta_{<0.2}$-score. Both RHs are underdispersed, as seen from the u-shape of the RH, but the size of the outlier bars, and the difference between the outliers and internal bars are much smaller in (b) than in (a), indicating improved reliability with alternative cost functions and cost function weighting.

4.5.4 CRPS and CRPS Reliability

The CRPS and CRPS reliability component were both used to evaluate the impact of alternative cost functions and cost weighting on forecast reliability. As with the previous measures, the use of alternative cost functions and cost function weighting produced substantial improvements in both the CRPS and the CRPS reliability term over the baseline condition of unweighted MSE. The $\overline{CRPS}$ obtained by the ANN ensembles trained with unweighted MSE ranged from 0.099 to 0.21 and the $\overline{Reli}$ ranged from 0.048 to 0.11. By contrast, the best $\overline{CRPS}$ obtained using alternative cost functions and cost function weighting ranged from 0.074 to 0.15 and the best $\overline{Reli}$ ranged from 0.027 to 0.061. This corresponds to maximum skill scores ranging from 0.06 to 0.35 for $\overline{CRPS}$ and 0.17 to 0.54 for $\overline{Reli}$. This improvement is shown in in Figures 4-9 and 4-10. Figure 4-9 shows the forecast-observation pairs for the overall dataset for the models trained

with unweighted MSE and for the ensembles trained using the cost function and weighting scheme which produced the best $\overline{CRPS}$ for each site and variable combination (Table 4-3). Figure 4-10 shows the forecast-observation pairs for the ensembles trained using the cost function and weighting scheme which produced the best $\overline{Reli}$ for each site and variable combination. Despite the $\overline{CRPS}$ prioritizing sharpness, and often being dominated by the sharpness term (Ferro, 2014; Hersbach, 2000), the ensembles trained with the cost function and weighting combinations that produced the forecasts with the lowest $\overline{CRPS}$ and $\overline{Reli}$ tended to have wider forecast ranges than those trained with unweighted MSE. This highlights that where alternative cost functions and cost function weightings improve the $\overline{CRPS}$, this is likely being driven by improvement in the reliability term. Additionally, the cost functions and weighting schemes which produced the best $\overline{CRPS}$ also tended to produce the best $\overline{Reli}$, again indicating that the improvements in $\overline{CRPS}$ are being driven by improvements in forecast reliability, not sharpness.

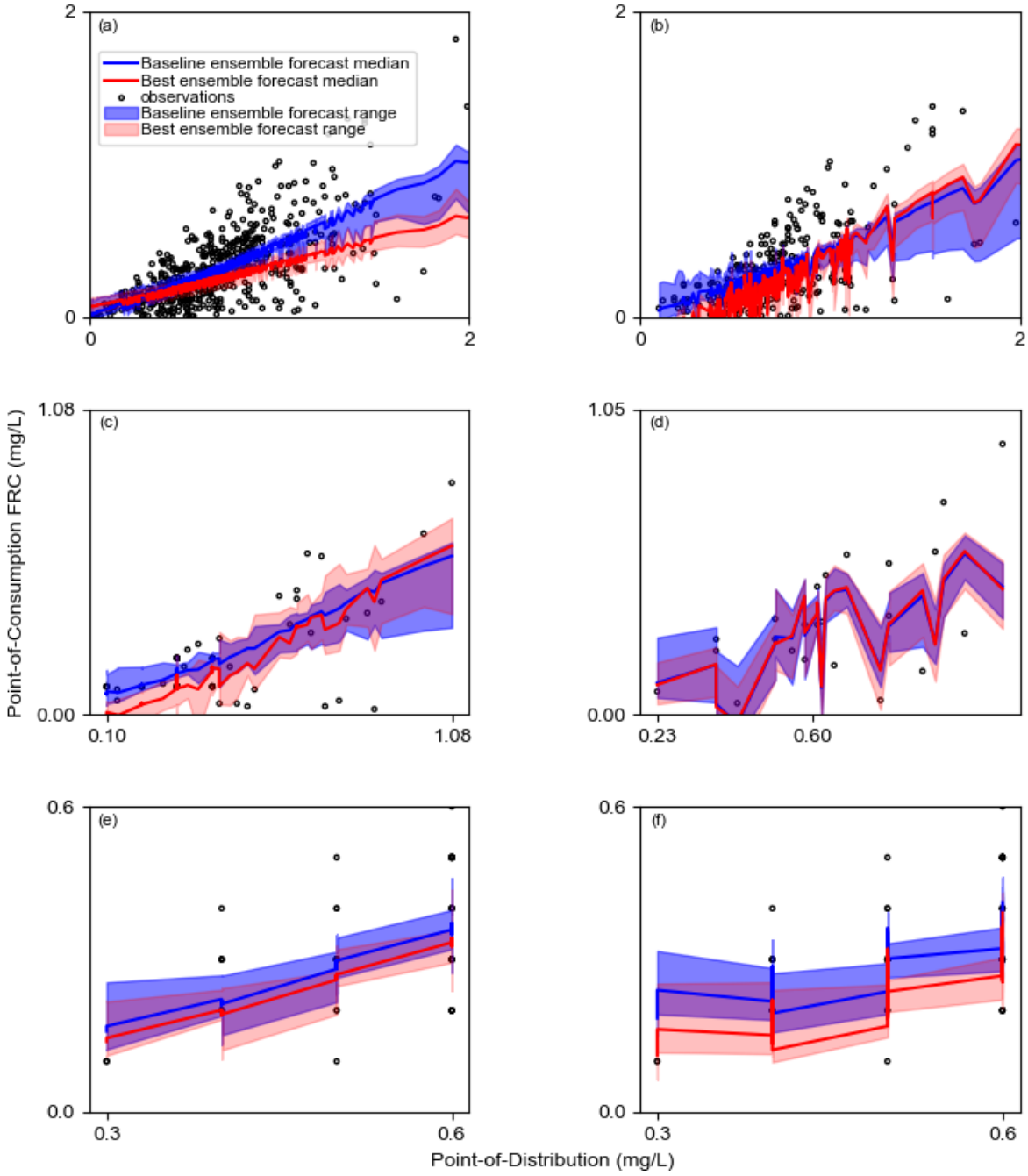

Figure 4-9: Forecast-observation comparison for all sites showing difference in forecast range and median between the baseline ensemble (trained with unweighted MSE) and ensemble with the best $\overline{CRPS}$ *for each site and variable combination. Forecast observation pairs shown for: (a) Bangladesh IV1, (b) Bangladesh IV2, (c) Tanzania IV1, (d) Tanzania IV2, (e) Nigeria IV1, and (f) Nigeria IV2.*

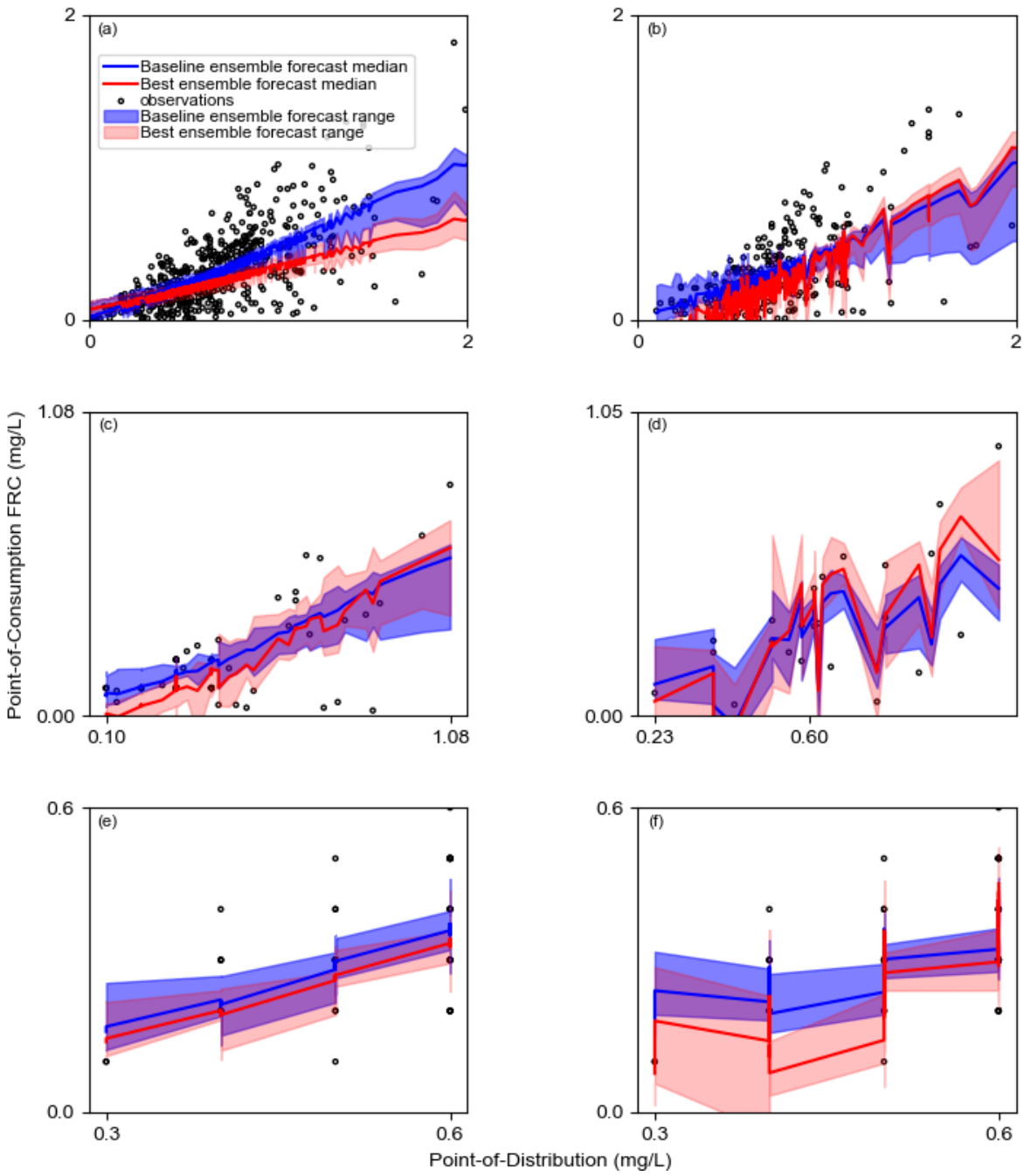

Figure 4-10: Forecast-observation comparison for all sites showing difference in forecast range and median between the baseline ensemble (trained with unweighted MSE) and ensemble with the best $\overline{Reli}$ *for each site and variable combination. Forecast observation pairs shown for: (a) Bangladesh IV1, (b) Bangladesh IV2, (c) Tanzania IV1, (d) Tanzania IV2, (e) Nigeria IV1, and (f) Nigeria IV2.*

4.5.5 Selection of Preferred Model

As discussed in the introduction to Section 4.5, the preferred alternative was selected using a frequency analysis to identify which alternatives were most frequently either the best performing alternative ("best") or one of the five best performing alternatives ("top five") for each ensemble verification metric at each site and with each variable combination. This approach was taken in consideration of the third objective of this study, which was to identify preferred alternatives that demonstrate consistently good performance across the of sites and variable combinations, to increase the likelihood that the selected alternative will be applicable at future sites. In conducting this frequency analysis, we found that for certain ensemble verification metrics multiple cost functions and weighting combinations produced the same performance, so while each cost function and weighting combination can only be the best performing alterative or one of the five best performing alternatives a maximum of 48 times, there are a total of 88 "best" and 283 "top five" cost function and weighting combinations. Figure 4-11 shows the frequency with which each cost function and weighting combination appeared as the "best" or in the "top five" alternatives across all sites and variables both disaggregated by performance metric and for all performance metrics combined.

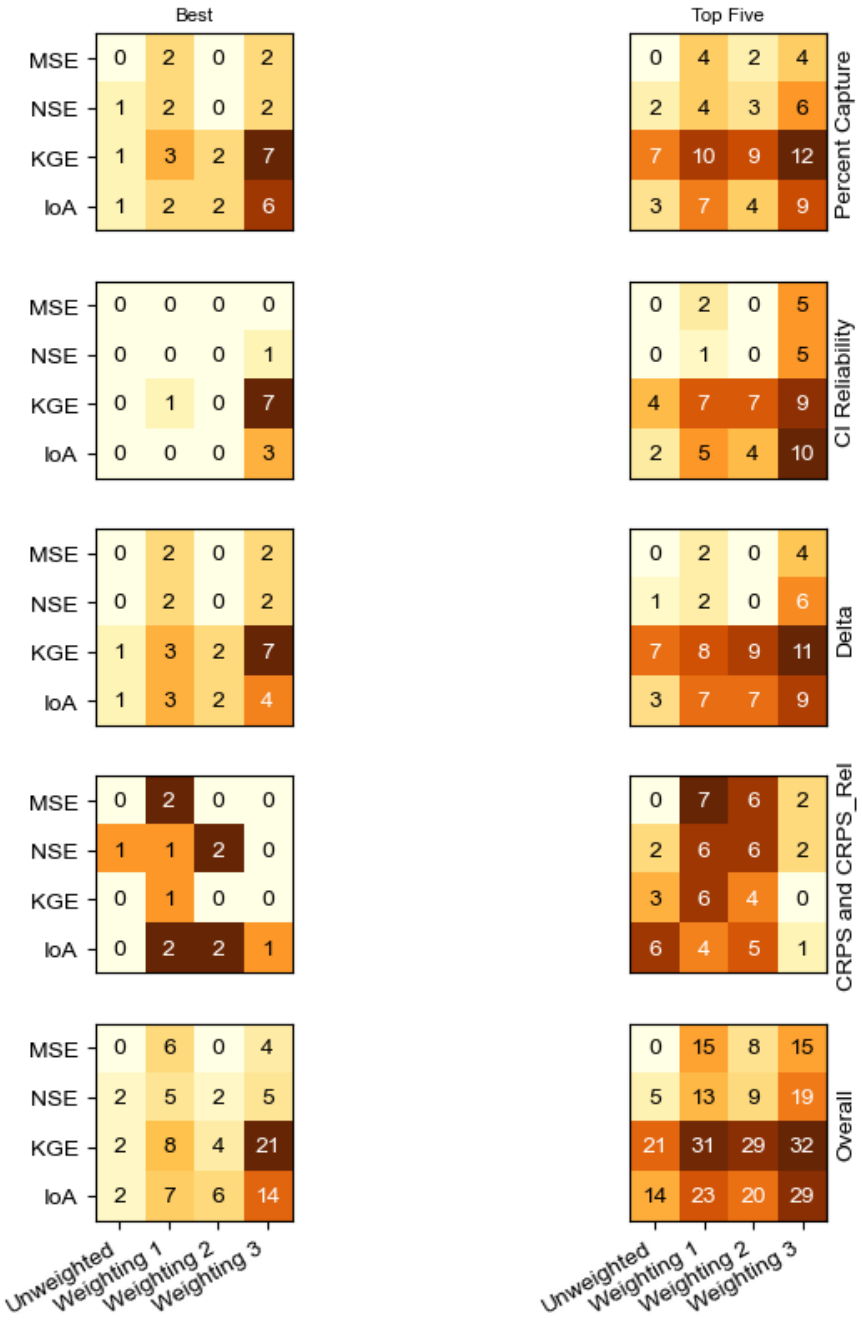

Figure 4-11: Frequency of cost functions and weighting combinations producing the best performance (left) or one of the "top five" performances for a given site and input variable combination for each performance metric and for all performance metrics combined (bottom row). Consistently KGE and IoA with weighting 3 produce the most "best" and "top five" performances, and MSE, particularly unweighted MSE, performs poorly.

Figure 4-11 shows that using alternative cost functions and cost weighting for training the base learners of the ensembles consistently improved performance of ensemble forecasts. 82 of the 88 "best" cost function and weighting combinations (93%) incorporated some kind of cost weighting, and cost weighting was incorporated into 243 of the 283 (86%) of the five cost function and weighting combinations. Additionally, only 10 cost function and weighting combinations that included MSE produced the best performance (11%), and only 38 of the 283 "top five" cost function and weighting combinations (13%) included MSE. Furthermore, unweighted MSE was never the best or one of the "top five" cost function and weighting combinations. This shows that the use of alternative cost functions and cost weighting results in better ensemble forecasts as measured by a variety of metrics across multiple sites, using multiple variable combinations, as is expected based on the success of cost function weighting approaches used in past studies (Crone et al., 2005; Elkan, 2001; Toth, 2016; Zhou & Liu, 2010). These findings also highlight that, despite being a common cost function, MSE (and by extension, other typical error functions) is not always the best choice for training ANNs for practical and engineering applications. Instead, the cost function, and if needed, cost weighting approach, should be selected to match the intended behaviour of the model. Another interesting finding is that the improvements in performance obtained through cost function weighting observed in this study were much greater than those reported by De Santi et al. (2021) when using ensemble post-processing techniques to improve probabilistic forecasts of point-of-consumption FRC using ensembles of ANNs. While that study used data from different sites, this may suggest that the selection of an appropriate cost function and weighting combination may be more useful that post-processing predictions, a concept that has been proposed in other applications as well (Dress et al., 2018).

Figure 4-11 also shows that the cost function and weighting combinations which resulted in the best performance across the various ensemble verification metrics most frequently were KGE and IoA with the inverse frequency weighting (weighting 3). KGE with weighting 3 produced the best performance in 21 of a possible 48 cases, and was among the five best cost functions in 32 of a possible 48. IoA with weighting 3 produced the best performance in 14 of a possible 48 cases, and was among the five best cost functions in 29 of a possible 48 cases. While IoA with weighting 3 was less common overall, it was more commonly preferred model for in Nigeria where it was the best cost function and weighting combination in 11 of a possible 16 cases and

was in the top five cost function and weighting combinations in 12 of a possible 16 cases. Additionally, when taking the sum of the skill scores for each metric at each site, which can be thought of as a summary of the net performance improvement over baseline for the site, KGE with weighting 3 produced the largest improvement for Bangladesh and for Tanzania, and IoA with weighting 3 produced the largest improvement in Nigeria. Based on these results, these two cost function and weighting combinations were selected for more detailed analysis.

Not only did KGE and IoA with weighting 3 consistently produce high performance across multiple metrics, all of the weighted and unweighted forms of these two cost functions performed well. Figure 4-11 shows that of the 88 "best" cost functions, 29 included IoA and 35 included KGE, meaning that 73% of the "best" cost functions and weighting combinations used one of these two cost functions. Furthermore, 86 of the 283 "top five" cost function and weighting combinations used IoA, and 113 used KGE, meaning that 70% of the "top five" cost function and weighting combinations used one of these two cost functions. These two cost functions likely performed well because they directly evaluate the difference between the observed and predicted distributions, instead of the error between each prediction and observation pair. KGE explicitly evaluates the difference between the observed and predicted mean and variance (the first two moments of a probability distribution) as well as the correlation (Gupta et al., 2009), while IoA measures the differences between the observed and predicted means and variances (Willmott, 1981). Thus, both KGE and IoA prioritize matching the observed distribution while training ANNs and thus it is understandable that ensembles of ANNs trained using these cost functions produce better capture through prioritizing a matching of variances and better reliability by evaluating the overall similarity between the observed and forecast distributions. Interestingly, both of these cost functions are related to NSE, with KGE having been formulated out of a decomposition of NSE and IoA being considered a modification of NSE (Kneale et al., 2001), and yet the ensembles with base learners trained to optimize NSE did not perform nearly as well as those trained with KGE or IoA. This is likely because the aspects of matching distribution parameters which are explicit in KGE and IoA are only implicit in NSE. This once again highlights the importance of selecting a cost function that directly rewards the intended behaviour for the base learner.

The prevalence of inverse frequency weighting (weighting 3) in both the "best" and "top five" cost function and weighting combinations is likely due to the unique nature of this weighting scheme as a rescaling term as opposed to the other cost functions which predominantly aimed to prioritize performance for high-priority samples. Instead of prioritizing samples with low point-of-consumption FRC, as was done by the other five weighting schemes, the inverse frequency weighting prioritizes regions of the output space that are sparsely populated to ensure that the model equally prioritizes all regions. Thus, instead of predictions clustering in regions with higher densities of observations, as is observed when using MSE to train the neural networks, the model represents the full output space, leading to ensemble forecasts that better represent the overall distribution of the data. Interestingly, while the class with point-of-consumption FRC between 0 and 0.2 mg/L was typically well populated with observations, the use of inverse frequency weighting still improved performance in this range, meaning that without deliberately prioritizing performance for that class, by creating a model that covered the whole outcome space, we were able to improve performance for our priority class.

The inverse frequency weighting approach we used was first used for classification machine learning approaches. Thus, it is somewhat surprising that using this approach to train ANNs for a regression problem produced the best results. One reason for this is that classification problems are inherently probabilistic, with the selected class being based on the selection of the class with the highest probability of being true (Elkan, 2001), and while the base learners in the ensembles were regression based, the overall ensemble in this approach was used for probabilistic forecasting. This highlights a potential avenue for future research into the integration of classification techniques in the training of probabilistic ensemble models, even if the base learners in these ensembles are regression-based.

Finally, comparing the performance of the ANN ensembles using the IV1 and IV2 input variable combinations showed that the ANN ensembles using the IV2 input variable combination consistently perform better than those using the IV1 combination, despite in some cases large proportions of observations being removed due to missing conductivity or water temperature data. This reinforces the findings of De Santi et al. (2021) which found that including water quality variables beyond point-of-distribution FRC and elapsed time helped ANN ensembles explain variability of point-of-consumption FRC concentration. It should still be noted that while

the ANN ensembles using IV2 tended to perform better than those using IV1, we still see substantial improvements for both variable combinations when implementing alternative cost functions and cost function weighting, which is critical as input variables beyond point-of-distribution FRC and elapsed time may not always be available. This comparison also shows that, for each site, the same cost function and weighting combinations tended to produce the best performance. This is important as the availability of EC and water temperature may vary over time, but, operationally, this would not necessitate a change in the cost function and weighting used to train the ANN ensembles.

4.5.6 Bangladesh Time-Series Analysis

Since KGE and IoA with inverse frequency weighting (weighting 3) performed substantially better than the other cost function and weighting combinations, both of these alternatives were selected to proceed to the more detailed time-series analysis using the Bangladesh data. This analysis was used to provide a more detailed evaluation of model performance while also testing the models in a scenario that is closer to the way the models would be used operationally. Additionally, since KGE produced better performance at sites with large data volumes and IoA produced better performance at sites with smaller data volumes, this analysis was used to test if there is a change in the relative performance of these two cost functions using an increasing amount of data at the same site. Since the Bangladesh ensemble models using the IV2 input variable combination consistently outperformed the models using the IV1 input variable combination, this analysis was only performed using the models trained with IV2.

Figure 4-12 shows the daily observed and forecasted point-of-consumption FRC concentrations for the two selected cost function and cost weighting combinations, as well as for unweighted MSE which was retained only as a baseline for comparison. From Figure 4-12, the model trained with unweighted MSE tended to miss many of the outlying observations, whereas these tended to be captured better by the ensembles using IoA and KGE with weighting 3. However, the model trained with KGE with weighting 3 tended to match the variability of the observations better than the model trained with IoA using weighting 3. This is confirmed in Figure 4-13, which compares the performance of the ensembles trained using unweighted MSE to the models trained using IoA and KGE with weighting 3. Figure 4-13 shows that the ensembles trained using IoA and KGE with weighting 3 both perform better than the model trained with unweighted MSE, showing that the improvements obtained from alternative cost functions and cost weightings still hold true in an operational context. This figure also shows that for the most part the forecasts produced by the ANN ensemble models trained using KGE with weighting 3 captured more overall observations, as well as more observations with point-of-consumption FRC below 0.2 mg/L.

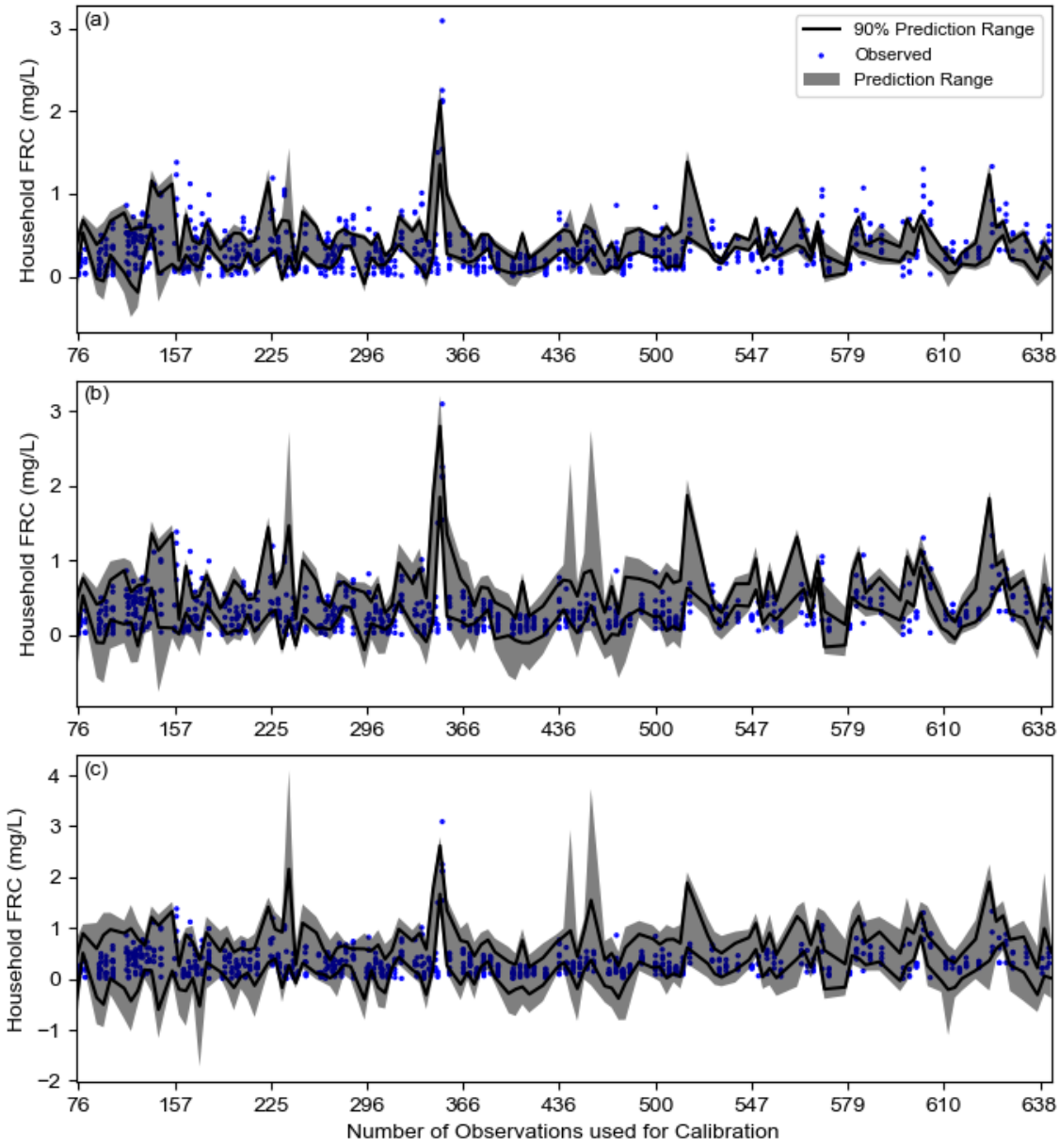

Figure 4-12: Comparison of daily observed and predicted point-of-consumption FRC concentrations for ensembles with base learners trained using (a) unweighted MSE, (b) IoA with Weighting 3, and (c) KGE with Weighting 3. Increasing observations used for calibration represents increasing data becoming available over time. The MSE forecasts are consistently underdispersed, weighted IoA and KGE both better match the observations.

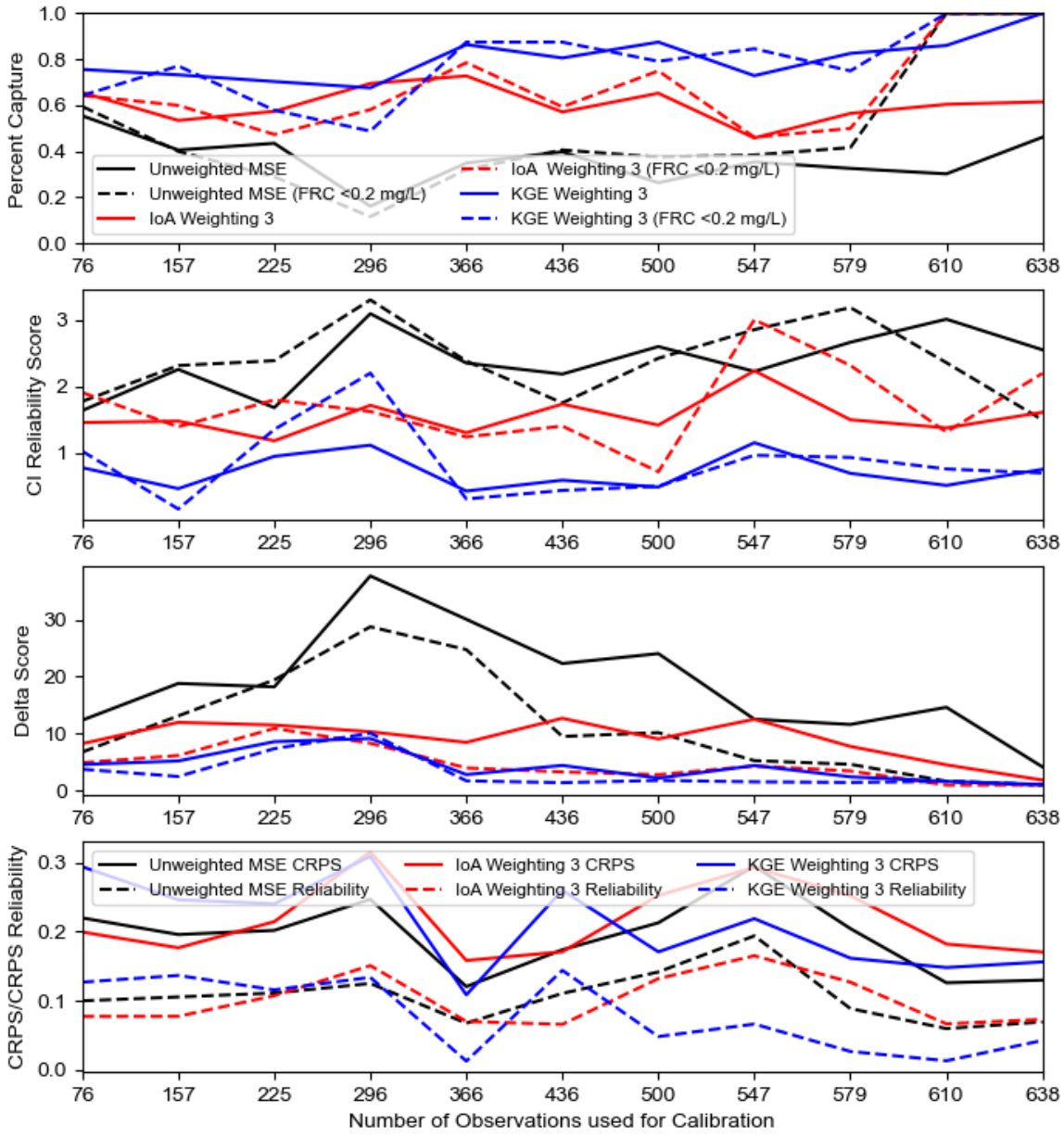

Figure 4-13: Comparison of ensemble verification metrics for Bangladesh time-series analysis. From top: CI Reliability score, Percent Capture, δ-score, CRPS and reliability term. The models trained using KGE with weighting 3 tend to have the best capture, CI reliability score, and δ-score. The CRPS reliability results are less clear.

Furthermore, the forecasts produced by the models trained with KGE with weighting 3 consistently performed better for the CI reliability score as well as the δ-score, both for the overall dataset and for observations with point-of-consumption FRC below 0.2 mg/L. While the difference in performance between the ANN ensembles trained with these two different cost

function and weighting combination were not as distinct for the $\overline{CRPS}$or the $\overline{Reli}$, the results presented in Figure 4-12 show that training the ANN ensembles using KGE with weighting 3 not only produced better capture of the observed data, but also more reliable forecasts. Implementing these improvements in the SWOT project would produce ANN ensembles that can better reproduce the underlying distributions of the observed data, improving the tool's ability to predict the risk of insufficient FRC at the point-of-consumption, and ultimately leading to better informed FRC targets for water system operators. It should be noted that Figure 4-2 shows that KGE with Weighting 3 produced the best performance regardless of the amount of data used for calibration, indicating perhaps that IoA with Weighting 3 did not perform best in Nigeria due to the data volume, and instead the difference may be due to other artefacts of the underlying data. Further study should investigate why IoA performed best in Nigeria as opposed to KGE.

4.6 Conclusion

This study investigated alternative cost functions and cost function weighting for training base learners in ensembles of ANNs to improve the dispersion and reliability of probabilistic forecasts of point-of-consumption FRC in refugee settlements. Producing reliable ensemble forecasts which match the dispersion of the observed data is critical for developing risk-based FRC targets for humanitarian response. This study found that training ANNs with weighted, alternative cost functions substantially improved forecasting performance across multiple indicators of dispersion and reliability. Training the ANN base learners using weighted forms of IoA and KGE consistently produced ensembles whose forecasts achieved the best dispersion and reliability, particularly when these cost functions were weighted using an inverse frequency weighting scheme. While weighted forms of IoA tended to perform better than KGE in sites with smaller data volumes, a time-series analysis where the model was retrained with continuously larger data volumes showed that KGE weighted with inverse frequency weighting produced the best performance, regardless of the size of the calibration dataset. While none of the cost functions or error weighting approaches led to perfect capture or reliability, they represent a clear improvement over the baseline unweighted MSE model that is currently in use in the Safe Water Optimization Tool (SWOT). Based on the findings of this study, we recommend implementing weighted cost functions for training base learners of the ANN ensembles as they greatly improve the model's ability to predict the risk of insufficient chlorine residual at the point-of-

consumption and as such will provide much better FRC recommendations for water system operators using the SWOT, helping them ensure water safety in humanitarian settings.

4.7 References

Ali, S. I., Ali, S. S., & Fesselet, J.-F. (2015). Effectiveness of emergency water treatment practices in refugee camps in South Sudan. *Bulletin of the World Health Organization, 93*(8), 550–558. https://doi.org/10.2471/BLT.14.147645

Ali, S. I., Ali, S. S., & Fesselet, J. (2021). Evidence-based chlorination targets for household water safety in humanitarian settings: Recommendations from a multi-site study in refugee camps in South Sudan, Jordan, and Rwanda. *Water Research, 189*(116642), 1–17. https://doi.org/https://doi.org/10.1016/j.watres.2020.116642

Almeida, A. M., Castel-Branco, M. M., & Falcão, A. C. (2002). Linear regression for calibration lines revisited: Weighting schemes for bioanalytical methods. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences, 774*(2), 215–222. https://doi.org/10.1016/S1570-0232(02)00244-1

Alvisi, S., & Franchini, M. (2011). Fuzzy neural networks for water level and discharge forecasting with uncertainty. *Environmental Modelling and Software, 26*(4), 523–537. https://doi.org/10.1016/j.envsoft.2010.10.016

Alvisi, S., & Franchini, M. (2012). Grey neural networks for river stage forecasting with uncertainty. *Physics and Chemistry of the Earth, 42–44*, 108–118. https://doi.org/10.1016/j.pce.2011.04.002

Atger, F. (2004). Estimation of the reliability of ensemble-based probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society, 130*(597B), 627–646. https://doi.org/10.1256/qj.03.23

Boucher, M. A., Perreault, L., & Anctil, F. (2009). Tools for the assessment of hydrological ensemble forecasts obtained by neural networks. *Journal of Hydroinformatics, 11*(3–4), 297–307. https://doi.org/10.2166/hydro.2009.037

Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review, 78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2

Bröcker, J. (2012). Evaluating raw ensembles with the continuous ranked probability score.

Quarterly Journal of the Royal Meteorological Society, 138(667), 1611–1617. https://doi.org/10.1002/qj.1891

Brown, G., Wyatt, J., Harris, R., & Yao, X. (2005). Diversity creation methods: A survey and categorisation. *Information Fusion, 6*(1), 5–20. https://doi.org/10.1016/j.inffus.2004.04.004

Candille, G., & Talagrand, O. (2005). Evaluation of probabilistic prediction systems for a scalar variable. *Quarterly Journal of the Royal Meteorological Society, 131*(609), 2131–2150. https://doi.org/10.1256/qj.04.71

CDC. (2012). Chlorine Residual Testing. Retrieved from http://www.cdc.gov/safewater/chlorine-residual-testing.html

Crone, S. F., Lessmann, S., & Stahlbock, R. (2005). Utility based data mining for time series analysis - Cost-sensitive learning for neural network predictors. *Proceedings of the 1st International Workshop on Utility-Based Data Mining, UBDM '05*, 59–68. https://doi.org/10.1145/1089827.1089835

Cronin, A. A., Shrestha, D., Cornier, N., Abdalla, F., Ezard, N., & Aramburu, C. (2008). A review of water and sanitation provision in refugee camps in association with selected health and nutrition indicators - the need for integrated service provision. *Journal of Water and Health, 6*(1), 1–13. https://doi.org/10.2166/wh.2007.019

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, 4*(35), 1–16. https://doi.org/10.1038/s41545-021-00125-2

de Vos, N. J., & Rientjes, T. H. M. (2008). Multiobjective training of artificial neural networks for rainfall-runoff modeling. *Water Resources Research, 44*(8), 1–15. https://doi.org/10.1029/2007WR006734

Dress, K., Lessmann, S., & von Mettenheim, H. J. (2018). Residual value forecasting using asymmetric cost functions. *International Journal of Forecasting, 34*(4), 551–565. https://doi.org/10.1016/j.ijforecast.2018.01.008

Elkan, C. (2001). The Foundations of Cost-Sensitive Learning The Foundations of Cost-Sensitive Learning. *In Proceedings ofthe Seventeenth International Joint Conference on*

Artificial Intelligence, Seattle, Washington, 973–978

Ferro, C. A. T. (2014). Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society, 140*(683), 1917–1923. https://doi.org/10.1002/qj.2270

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. *Mathematical and Computer Modelling*, *44*(5–6), 485–498. https://doi.org/10.1016/j.mcm.2006.01.007

Girones, R., Carratalà, A., Calgua, B., Calvo, M., Rodriguez-Manzano, J., & Emerson, S. (2014). Chlorine inactivation of hepatitis e virus and human adenovirus 2 in water. *Journal of Water and Health*, *12*(3), 436–442. https://doi.org/10.2166/wh.2014.027

Golicha, Q.,Shetty, S.,Nasiblov, O., Hussein, A., Wainaina, E., Obonyo, M., Macharia, D., Musyoka, R. N., Abdille, H., Ope, M., Joseph, R., Kabugi, W., Kiogora, J., Said, M., Boru, W., Galgalo, T., Lowther, S. A., Juma, B., Mugoh, R.,...Burton, J.W. (2018). Cholera outbreak in Dadaab Refugee camp, Kenya — November 2015–June 2016. *Morbidity and Mortality Weekly Report*, *67*(34), 958–961. https://doi.org/10.15585/mmwr.mm6734a4

Guerrero-Latorre, L., Hundesa, A., & Girones, R. (2016). Transmission Sources of Waterborne Viruses in South Sudan Refugee Camps. *Clean - Soil, Air, Water*, *44*(7), 775–780. https://doi.org/10.1002/clen.201500358

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*, 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Hamill, T. M. (2001). Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Monthly Weather Review*, *129*(3), 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 and Forecasting*, *15*(5), 559–570.

Howard, C. M., Handzel, T., Hill, V. R., Grytdal, S. P., Blanton, C., Kamili, S., Drobeniuc, J., Hu, D., & Teshale, E. (2010). Novel Risk Factors Associated with Hepatitis E Virus

Infection in a Large Outbreak in Northern Uganda: Results from a Case-Control Study and Environmental Analysis. *American Journal of Tropical Medicine and Hygiene, 83*(5), 1170–1173. https://doi.org/10.4269/ajtmh.2010.10-0384

Khan, U. T., & Valeo, C. (2016). Dissolved oxygen prediction using a possibility theory based fuzzy neural network. *Hydrology and Earth System Sciences, 20*, 2267–2293. https://doi.org/10.5194/hess-20-2267-2016

Khan, U. T., & Valeo, C. (2017). Comparing a Bayesian and fuzzy number approach to uncertainty quantification in short-term dissolved oxygen prediction. *Journal of Environmental Informatics, 30*(1), 1–16. https://doi.org/10.3808/jei.201700371

Kneale, P., See, L., & Smith, A. (2001). Towards Defining Evaluation Measures for Neural Network Forecasting Models. *In Proceedings of the Sixth International Conference on GeoComputation*. University of Queensland, Brisbane, Australia. 1-11 http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.163.4057

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence, 5*(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

Lantagne, D. S. (2008). Sodium hypochlorite dosage for household and emergency water treatment. *Journal of American Water Works Association, 100*(8), 106–114. https://doi.org//10.1002/j.1551-8833.2008.tb09704.x

LeChevallier, M. W., Evans, T. M., & Seidler, R. J. (1981). Effect of turbidity on chlorination efficiency and bacterial persistence in drinking water. *Applied and Environmental Microbiology, 42*(1), 159–167. https://doi.org/10.1128/aem.42.1.159-167.1981

LeChevallier, Mark W., Welch, N. J., & Smith, D. B. (1996). Full-scale studies of factors related to coliform regrowth in drinking water. *Applied and Environmental Microbiology, 62*(7), 2201–2211. https://doi.org/10.1128/aem.62.7.2201-2211.1996

Ling, C. X., & Sheng, V. S. (2008). Cost-Sensitive Learning and the Class Imbalance Problem. *Encyclopedia of Machine Learning*, 231–235. Retrieved from http://www.springer.com/computer/ai/book/978-0-387-30768-

8%5Cnhttp://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.164.4418andrep=rep1andtype=pdf

Liu, X. Y., & Zhou, Z. H. (2006). The influence of class imbalance on cost-sensitive learning: An empirical study. *In IEEE International Conference on Data Mining, ICDM,* 970–974. https://doi.org/10.1109/ICDM.2006.158

McCarthy, K., Zabar, B., & Weiss, G. (2005). Does cost-sensitive learning beat sampling for classifying rare classes? *Proceedings of the 1st International Workshop on Utility-Based Data Mining, UBDM '05, 69–77*. https://doi.org/10.1145/1089827.1089836

Médecins Sans Frontières. (2010). Public Health Engineering In Precarious Situations. (J. V. D. Noortgate and P. Maes, Eds.) (2nd ed.). Brussels: Médecins Sans Frontières.

Rashid, M.-u., George, C. M., Monira, S., Mahmud, T., Rahman, Z., Mustafiz, M., Parvin, T., Bhuyian, S. I., Zohura, F., Begum, F., Biswas, S. K., Akhter, S., Zhang, X., Sack, D., Sack, R. B., & Alam, M. (2016). Chlorination of Household Drinking Water among Cholera Patients ' Households to Prevent Transmission of Toxigenic Vibrio cholerae in Dhaka , Bangladesh : CHoBI7 Trial. *American Journal of Tropical Medicine and Hygiene, 95*(6), 1299–1304. https://doi.org/10.4269/ajtmh.16-0420

Rodriguez, M. J., & Sérodes, J. B. (1998). Assessing empirical linear and non-linear modelling of residual chlorine in urban drinking water systems. *Environmental Modelling and Software, 14*(1), 93–102. https://doi.org/10.1016/S1364-8152(98)00061-9

Shultz, A., Omollo, J. O., Burke, H., Qassim, M., Ochieng, J. B., Weinberg, M., Feikin, D. R., & Breiman, R. F. (2009). Cholera outbreak in Kenyan Refugee Camp: Risk Factors for Illness and Importance of Sanitation. *American Journal of Tropical Medicine and Hygiene, 80*(4), 640–645. https://doi.org/10.4269/ajtmh.2009.80.640

Sikder, M., String, G., Kamal, Y., Farrington, M., Rahman, A. S., & Lantagne, D. (2020). Effectiveness of water chlorination programs along the emergency-transition-post-emergency continuum: Evaluations of bucket, in-line, and piped water chlorination programs in Cox's Bazar. *Water Research, 178,* 115854. https://doi.org/10.1016/j.watres.2020.115854

Steele, A., Clarke, B., & Watkins, O. (2008). Impact of jerry can disinfection in a camp environment - Experiences in an IDP camp in Northern Uganda. *Journal of Water and Health*, *6*(4), 559–564. https://doi.org/10.2166/wh.2008.072

Swerdlow, D.L. Malenga, G., Begkoyian, G., Nyangulu, D., Toole, M., Waldman, R. J., Puhr, D. N. D., & Tauxe, R. V. (1997). Epidemic cholera among refugees in Malawi, Africa: treatment and transmission. Epidemiology and Infection, 118(3), 207–214. https://doi.org/https://doi.org/10.1017/S0950268896007352

Talagrand, O., Vautard, R., & Strauss, B. (1997). Evaluation of probabilistic prediction systems. *In Proceedings, ECMWF Workshop on Predictability*. Shinfield Park, Reading: ECMWF, 1-25. Retrieved from https://www.ecmwf.int/node/12555

Toth, E. (2016). Estimation of flood warning runoff thresholds in ungauged basins with asymmetric error functions. *Hydrology and Earth System Sciences*, *20*(6), 2383–2394. https://doi.org/10.5194/hess-20-2383-2016

Walden, V. M., Lamond, E. A., & Field, S. A. (2005). Container contamination as a possible source of a diarrhoea outbreak in Abou Shouk camp, Darfur province, Sudan. *Disasters*, *29*(3), 213–221. https://doi.org/10.1111/j.0361-3666.2005.00287.x

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

Willmott, C. J. (1981). On the Validation of Models. *Physical Geography, 2*, 184–194. https://doi.org/10.1080/02723646.1981.10642213

Wu, H., & C. Dorea, C. C. (2020). Towards a Predictive Model for Initial Chlorine Dose in Humanitarian Emergencies. *Water*, *12*(5), 1506. https://doi.org/10.3390/w12051506

Zhou, Z.-H., & Liu, X.-Y. (2010). On Multi-Class Cost-Sensitive Learning. *Computational Intelligence*, *26*(3), 232–257.

Chapter 5 Multi-Objective Training

5.1 Chapter Preamble

This chapter presents a study into the use of multi-objective training to improve the dispersion and reliability of ANN ensemble forecasts of point-of-consumption FRC. A modified version of this chapter is currently being prepared for publication. Since the same datasets were used for this chapter as Chapter 5, the supplemental information for this chapter is included in Appendix D.

Multi-objective training is a process where an ANN is trained to optimize performance over multiple objectives. This can either be achieved using *preference-based* methods where multiple objectives are combined into a single score using multiplication or a weighted sum, or using *no-preference* methods that use metaheuristic optimization approaches, such as genetic algorithms or particle swarm optimization, to train the ANNs to simultaneously optimize many objectives at once. It is important to note that multi-objective cost functions and multi-objective training use these approaches to set the weights and biases of a base learner ANN in an ensemble. This is different from multi-objective optimization of neural network architecture which uses similar processes to optimize the design of the neural network model (number of nodes, activation functions, input variable selections etc.). Multi-objective training of neural networks has been investigated in several disciplines and in general multi-objective neural networks tend to outperform neural networks trained on a single objective, especially those using common error metrics, however, these have never been applied for probabilistic forecasting and have not previously been used for forecasting post-distribution FRC.

Based on the findings of Chapters 3 and 4, our goal in this study was to use multi-objective training to improve the dispersion and reliability of ANN ensemble forecasts. This informed the selection of the training objectives which were measures of the similarity between the distribution of the training data and the predictions of each base learner ANN. Following the usefulness of KGE as a cost function identified in Chapter 4, three of the training objectives we selected to characterize the similarity of the observed and predicted distributions were derived from the KGE formula. We used both preference-based methods which combine multiple objectives into a single objective, and no-preference methods that use metaheuristic training

approaches to simultaneously optimize multiple objectives. The study presented in this chapter found that ANNs trained using multi-objective training outperformed the baseline ANNs trained with MSE (Chapter 3) and cost sensitive learning approaches (Chapter 4). Furthermore, of the multi-objective training approaches considered, the best performance was obtained using an objective combining approach that assigned the weights to each objective using a Das-Dennis grid, which populates a balanced set of solutions in the objective space. The use of this approach led to substantial improvements in forecast dispersion and reliability across multiple sites and variable combinations, likely because it directly prioritized a balanced exploration of the objective space.

As the lead author I was responsible for the conception of the study presented in this chapter, as well as all model development and analysis. I was also responsible for preparing the manuscript. Dr. Usman Khan was responsible for modelling supervision and manuscript preparation. Dr. Syed Imran Ali was responsible for supporting data collection at all sites, coordination of partners, securing funding, and manuscript review. Jean-François Fesselet was responsible for coordination of partners, securing funding, and manuscript review. Matt Arnold was responsible for leading data collection in Bangladesh and supporting data collection in Tanzania and Nigeria, coordination of partners and manuscript review. Dawn Taylor was responsible for leading data collection in Nigeria and manuscript review. Anne Hyvaerinen was responsible for leading data collection in Tanzania and manuscript review.

5.2 Abstract

The Safe Water Optimization Tool (SWOT) uses ensembles of artificial neural networks (ANNs) to produce probabilistic, risk-based free chlorine residual (FRC) targets to prevent household recontamination of drinking water in refugee and internally displaced person (IDP) settlements. To ensure that these risk-based targets are accurate, the ensemble models must produce forecasts which closely match the distribution of the observed data. When typical error metrics are used as cost functions to train the ANNs in these ensembles, the forecasts tend to be underdispersed and do not match the underlying distribution of the data. This research presents an investigation into the use of multi-objective training for ANN ensembles where the objectives used directly evaluate the ability of the ANN models to reproduce the underlying distribution of the data and to evaluate the risk of having insufficient drinking water. Four multi-objective training methods were considered in this study: three preference-based methods, where multiple objectives were combined into a single-objective cost function via a weighted sum, and one no-preference method which used the Non-Dominated Sorting Genetic Algorithm III (NSGA-III) algorithm to simultaneously optimize multiple objectives. Using drinking water quality datasets from refugee settlements in Bangladesh, Tanzania, and Nigeria to test these training methods, we found that multi-objective training consistently outperformed the baseline model trained using mean squared error, highlighting the importance of selecting appropriate cost functions and training methods for ANN applications. The best performance was obtained using a preference-based method that assigned the weights to each objective using a Das-Dennis grid which populates a balanced set of solutions in the objective space. Unlike the other preference-based methods included in this research, the objectives in this method were weighted differently for each ensemble member, leading to an increased degree of ensemble diversity for this method over the other preference-based methods. Implementing this method for training the ANN ensembles in the SWOT can substantially improve forecasts of point-of-consumption FRC, which in turn can lead to FRC targets that better protect drinking water against household recontamination in refugee settlements.

5.3 Introduction

Providing safe drinking water in refugee and internally displaced person (IDP) settlements is critical as waterborne illnesses are a leading cause of excess morbidity and mortality in these settings. Many of these settlements have drinking water systems where water is treated and piped

to central distribution points from which water is collected and then stored for an extended period of time in households. Recontamination of drinking water during this post-distribution period of collection, transport, storage, and use is a common contributor to the spread of waterborne diseases in refugee and IDP settlements, having been linked to outbreaks of cholera, hepatitis E, and shigellosis in refugee and IDP settlements in Kenya (Golicha et al., 2018; Shultz et al., 2009), Malawi (Swerdlow et al., 1997), Sudan (Walden et al., 2005), South Sudan (Ali et al., 2015; Guerrero-Latorre et al., 2016), and Uganda (Howard et al., 2010; Steele et al., 2008). Residual chlorine is effective for preventing recontamination during the post-distribution period as it can inactivate pathogens as they are introduced into stored drinking water. A free residual chlorine (FRC) concentration of 0.2 mg/L is effective for preventing recontamination by common pathogens responsible for outbreaks of waterborne illness in refugee and IDP settlements (CDC, 2012; Girones et al., 2014; Lantagne, 2008; Rashid et al., 2016; Sikder et al., 2020; WHO, 2011). However, FRC targets in current drinking water quality guidelines for humanitarian response (e.g., Sphere Handbook) do not provide adequate protection against recontamination as they do not account for chlorine decay during this post-distribution period, leading to a loss of protection over time and an increased risk of recontamination of drinking water during household storage (Ali et al., 2015, 2021).

To ensure that there is adequate protection against recontamination up to the point-of-consumption, revised targets for chlorine residual are required at the point-of-distribution that can account for post-distribution chlorine decay. Developing these targets can be challenging as post-distribution chlorine decay is highly variable, and there is substantial uncertainty in the processes driving this decay. This uncertainty arises due to the numerous quantifiable and unquantifiable factors that can impact the rate and nature of post-distribution chlorine decay, including water quality and environmental parameters that may contribute to the rate of decay, as well as water handling factors, such as user interactions with the water that may introduce contaminants that consume the chlorine residuals (Ali et al., 2021). Thus, for a single set of conditions at the water distribution point, a range of residual chlorine concentrations are possible at the point-of-consumption. Due to this level of uncertainty, deterministic FRC predictions are inadequate for developing chlorine residual targets as they cannot communicate the uncertainty in the predicted FRC concentration. Instead, when generating FRC targets for the water distribution point probabilistic forecasts of the point-of-consumption FRC should be used as they

can quantify the uncertainty in the model predictions and convey the risk of having insufficient FRC at the point of consumption.

A common approach to generating probabilistic forecasts is through ensemble modelling. Ensemble models group the predictions of multiple deterministic models into a probability distribution (the forecast). Probabilistic ensemble forecasting often use physical or process-based models, however, ensembles of artificial neural networks (ANNs), a type of data-driven model, have also been used for probabilistic forecasting in hydrology (Boucher et al., 2009). Using data-driven models for probabilistic forecasting has the added advantage that these models are not limited by *a priori* assumptions about underlying behaviour, making them better for modelling highly uncertain process. The Safe Water Optimization Tool (SWOT) is an analytical tool which generates evidence-based, site-specific FRC target guidance for water system operators in humanitarian response settings using ensembles of ANNs for probabilistic forecasting.

A common challenge in ensemble modelling, especially when using ensembles of ANNs, is that the resulting forecasts tend to be underdispersed, meaning that the spread of the forecasts tends to be lower than the spread of the underlying data (Boucher et al., 2009). This reduces the reliability of the forecasts as the forecast distribution will not match the observed data, and it means that the model may fail to capture extreme events, including observations with low point-of-consumption FRC where the health risk is highest. For ANNs, this underdispersion may be due to the use of common error metrics, such as mean squared error (MSE) as cost functions. Studies from other disciplines, primarily in business management applications, but also water resources, have noted that common error metrics are not appropriate for real-world applications as the regression-to-the-mean behaviour that they produce does not represent the needs of users, and instead ANNs should be trained using cost functions that promote the desired behaviour (Crone et al., 2005; Toth, 2016). Past attempts to overcome underdispersion for ANN ensemble forecasts of post-distribution FRC include ensemble post-processing, which uses non-parametric methods to increase the dispersion of the forecast (De Santi et. al, 2021 [Chapter 3]), as well as cost-sensitive learning (Chapter 4). An alternative approach to overcoming the common challenges of typical error metrics for ANNs is to use multi-objective training. Multi-objective training has been demonstrated as an effective tool for training ANNs for modelling chlorine residual in piped distribution systems, as well as for other engineering applications ranging from

structural engineering to hydrology (Chatterjee, Sarkar, Dey, et al., 2017; Chatterjee, Sarkar, Hore, et al., 2017; de Vos & Rientjes, 2008). In particular, these studies have shown that ANNs trained using multi-objective optimization perform particularly well for predicting outlier events. Broadly, multi-objective training can be divided into two methods: first, multiple objectives can be combined into a single objective through multiplication or using a weighted sum and then the neural network can be trained using conventional means such as backpropagation. These are referred to as *preference-based* methods as the use of a weighted sum to combine objectives allows certain objectives to be preferred or prioritized over others. Alternatively, the second approach is to train the neural network model parameters using *no-preference* methods which use metaheuristic approaches (a type of higher-level search procedure) search procedures such as a multi-objective genetic algorithm (MOGA) or particle swarm optimization (PSO) to simultaneously optimize all objectives when training the model weights and biases (de Vos & Rientjes, 2008). This latter method also has the benefit for ensemble learning that the final set of solutions (or a subset thereof) can be used to form an ensemble (Abbass, 2003; Chen & Yao, 2010).

This study investigated the use multi-objective training ensembles of ANNs for probabilistic forecasting of point-of-consumption FRC in refugee and IDP settlements. While multi-objective training of ANNs and ANN ensembles has previously been applied to many contexts, this is the first attempt to use these methods for probabilistic ensemble forecasting and is also the first use of multi-objective training for data-driven models of post-distribution FRC. To accomplish this, we selected a novel combination of objectives that were specifically targeted to evaluate different aspects of the similarity between the forecasts and the distribution of the observations. The goal of this research was to assess the effectiveness of using multi-objective training to improve the dispersion and reliability of ANN ensemble forecasts of point-of-consumption FRC. Thus, we had three primary aims for this investigation. First, we sought to produce and evaluate ensemble forecasts of point-of-consumption FRC using ANNs trained using preference-based and no-preference multi-objective training and identify the best multi-objective training method. Secondly, we sought to compare the ensemble forecasting performance of ANN ensembles trained using multi-objective training methods against the baseline of ensembles of ANNs trained using MSE. Finally, we sought to compare the multi-objective training approach to other

approaches that have been investigated for improving the performance of the SWOT ANNs, namely ensemble post-processing and cost-sensitive learning.

5.4 Methods

5.4.1 Description of Data Sets Used

This study used data from three refugee settlements in Bangladesh, Tanzania, and Nigeria, which were collected through the SWOT Project. At each site, water quality changes from the point-of-distribution to the point-of-consumption were documented using paired sampling wherein the same unit of water was sampled at both the water distribution point and at the point-of-consumption. At each site, data on the same five parameters were collected:

1. FRC concentration at the water distribution point
2. Electrical conductivity (EC) at the water distribution point
3. Water temperature at the water distribution point
4. FRC concentration at the point-of-consumption
5. Elapsed time from the measurement at the water distribution point to the measurement in the point-of-consumption

5.4.2 Ethics

The studies in Bangladesh, Tanzania, and Nigeria received approval from the 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).

5.4.3 ANN Ensemble Model Description

The ANN base learners for the ensembles in this study use the multilayer perceptron (MLP) structure with one hidden layer as this type of ANN has been shown to outperform other ANN architectures for predicting extreme values of FRC in piped distribution systems (Gibbs et al., 2006; Rodriguez & Sérodes, 1998). For each site, two input variable combinations were

considered: the first input variable combination (IV1) included only point-of-distribution FRC and elapsed time, representing the minimum water quality data which are regularly collected in humanitarian response settings, and the second input variable combination (IV2) included point-of-distribution FRC, EC, water temperature, and elapsed time, representing commonly available water quality data in humanitarian settings. Two input variable combinations were considered to determine if different multi-objective training approaches perform better with larger or smaller variable combinations. In particular, maximizing the performance of ANN ensembles using the minimalist input variable combinations can help reduce the data collection burden for water system operators in humanitarian settings. For each variable combination, the overall dataset was subdivided into a calibration set used for training and validating the models, and a testing set used to evaluate model performance. We used 25% of the overall dataset for testing, 25% for validation, and 50% for training. We provide descriptive statistics for the input and output variables in each variable combination for the testing and calibration dataset in Table 5-1. Note that we do not subdivide the calibration dataset into validation and training in Table 5-1 because for each base learner ANN, the training and validation datasets were randomly sampled from within the calibration dataset.

Table 5-1: Input and output variable mean, median, and standard deviations for all sites and input variable combinations for the calibration and testing datasets. Note that the same variable at the same site may have different statistics between the two input variable combinations due to observations being removed for missing.

		Calibration				Testing			
		Number of Observations	**Mean**	**Median**	**Standard Deviation**	**Number of Observations**	**Mean**	**Median**	**Standard Deviation**
Bangladesh IV1	Point-of-distribution FRC (mg/L)	1,597	0.71	0.66	0.38	533	0.70	0.64	0.38
	Elapsed Time (h)		10.02	6.70	5.04		9.66	6.67	4.93
	Point-of-consumption FRC (mg/L)		0.34	0.28	0.28		0.34	0.28	0.28
Bangladesh IV2	Point-of-distribution FRC (mg/L)	728	0.74	0.67	0.38	244	0.77	0.69	0.41
	Elapsed Time (h)		10.27	6.80	5.07		10.18	6.88	5.00
	EC (μs/cm)		329	308	68.85		334	310	64.08

		Calibration				Testing			
		Number of Observations	**Mean**	**Median**	**Standard Deviation**	**Number of Observations**	**Mean**	**Median**	**Standard Deviation**
	Water Temperature (°C)		27.51	27.80	1.49		27.52	27.70	1.29
	Point-of-consumption FRC (mg/L)		0.33	0.27	0.28		0.35	0.26	0.36
Tanzania IV1	Point-of-distribution FRC (mg/L)	228	0.39	0.30	0.22	77	0.39	0.30	0.21
	Elapsed Time (h)		7.35	5.65	4.96		7.18	5.60	5.51
	Point-of-consumption FRC (mg/L)		0.20	0.10	0.15		0.18	0.10	0.15
Tanzania IV2	Point-of-distribution FRC (mg/L)	66	0.60	0.61	0.23	23	0.65	0.62	0.20
	Elapsed Time (h)		11.52	8.19	5.70		11.75	8.53	5.94

		Calibration				Testing			
		Number of Observations	**Mean**	**Median**	**Standard Deviation**	**Number of Observations**	**Mean**	**Median**	**Standard Deviation**
	EC (µs/cm)		325	393	150		295	390	170
	Water Temperature (°C)		24.03	24.05	0.97		23.83	23.80	1.07
	Point-of-consumption FRC (mg/L)		0.29	0.28	0.20		0.34	0.31	0.22
Nigeria IV1	Point-of-distribution FRC (mg/L)	162	0.53	0.55	0.08	54	0.54	0.60	0.09
	Elapsed Time (h)		4.08	3.53	3.08		3.91	3.73	1.76
	Point-of-consumption FRC (mg/L)		0.31	0.30	0.11		0.33	0.30	0.12
Nigeria IV2	Point-of-distribution FRC (mg/L)	162	0.53	0.50	0.08	54	0.54	0.60	0.09

	Calibration				Testing			
	Number of Observations	**Mean**	**Median**	**Standard Deviation**	**Number of Observations**	**Mean**	**Median**	**Standard Deviation**
Elapsed Time (h)		4.08	3.53	3.02		3.91	3.73	1.76
EC (µs/cm)		270	270	11.43		266	267	17.06
Water Temperature (°C)		31.39	31.30	1.91		31.85	31.80	1.96
Point-of-consumption FRC (mg/L)		0.31	0.30	0.11		0.33	0.30	0.12

The hyperparameters of the base learners were selected through an exploratory analysis informed by the processes documented in De Santi et al. (2021). The hidden layer size was selected by successively doubling the number of hidden nodes and then selecting the size where performance improvements slowed or where performance decreased. The selected hidden layer size for the Tanzania and Nigeria IV1 models was four hidden nodes and eight hidden nodes for the IV2 models. In Bangladesh, we selected a hidden layer size of 16 hidden nodes for both the IV1 and IV2 models. The results of this exploratory analysis are presented in Appendix D as they are the same as those used in Chapter 4. The hyperbolic tangent activation function was used in the hidden layer and a linear activation function was used for the output layer.

The ensemble size (number of base learners in an ensemble) was also selected through an exploratory grid search analysis where ensemble sizes of 50 to 500 were investigated. Performance typically increased with increasing ensemble size, though the improvement as size increased tended to decrease for ensemble sizes between 100 and 200, so for all sites and variable combinations, an ensemble size of 200 ANNs was selected to ensure that the ensemble size did not constrain performance at any site while avoiding the additional computational time associated with larger ensemble sizes. Note that some of the multi-objective training approaches required slight changes to this ensemble size. These changes have been documented in the description of the different multi-objective training approach sections.

When developing ensemble models for probabilistic forecasting it is critical that the ensemble members or base learners are sufficiently distinct from each other so that the resulting forecast accurately quantifies the uncertainty in the underlying behaviour (Bröcker, 2012; Hamill, 2001). In ANN ensembles, this difference between base learners is referred to as ensemble diversity, which can represent a variety of differences ranging from differences in the model parameter (weights and biases) to differences in the model structure or hyperparameters, and even inclusion of different types of models within an ensemble (Brown et al., 2005). We promoted ensemble diversity in this study using implicit techniques in two ways. First, the initial weights and biases were randomized so that each base learner was trained starting at a different location on the error surface. Additionally, the calibration data was randomly sampled into training and validation subsets so that all of the calibration data was used during training but this data was randomly allocated to training or validation. By using random processes to initialize the models and to

select the training and validation data, we obtain a diverse ensemble while ensuring that the base learners remain independent of each other. In some cases, additional diversity was introduced through the training approach, which has been noted in the descriptions of the different training approaches.

5.4.4 Approaches to Multi-Objective Training

As discussed in Section 5.3, there are two main categories of multi-objective training methods for ANNs. The first are preference-based methods which combine many objectives into a single score – either by multiplying the scores for each objective or by combining each the scores in a weighted sum. In a preference-based method, since there is a single score, the ANNs can be trained using typical backpropagation training. The second category are no-preference methods which simultaneously optimize all objectives using a metaheuristic optimization procedure such as MOGAs or PSO.

Objectives Considered

Many different objectives and combinations of objectives have been proposed in past studies of multi-objective ANN training. When using multi-objective training for modelling natural systems, it is important to select cost functions that meaningfully assess the desired behaviour (de Vos & Rientjes, 2008). Since the ensembles in this study are being used for probabilistic forecasting, and will be assessed based on the similarity between the forecast distribution and the underlying distribution, we selected objectives that evaluate the difference between the predictions of each model and the underlying distribution.

The objectives we used are adapted from the Kling Gupta Efficiency (KGE) cost function and are used to assess the similarity between the observed and predicted distributions. We selected this cost function as a basis for selecting training objectives as it has been shown in previous studies to produce very good performance when used to train ensembles of ANNs, likely due to its prioritization of matching key distribution parameters (Chapter 4). This cost function was developed by decomposing common performance metrics into three components: α, the ratio of the observed and predicted standard deviations; β, the ratio of observed and predicted means; and r the Pearson's correlation coefficient between the observed and predicted data. We converted these into error scores using the same approach taken by Gupta et al. (2009): we took

the squared difference between each component and the ideal value (1 in each case). Thus, the first three objectives we used were:

$$Objective\ 1: \alpha\ score = (1-\alpha)^2 \tag{5-1}$$

$$Objective\ 2: \beta\ score = (1-\beta)^2 \tag{5-2}$$

$$Objective\ 3: r\ score = (1-r)^2 \tag{5-3}$$

Each of these scores are negatively oriented (lower value is better) ranging from 0 to infinity.

One of the unique challenges of using probabilistic ensembles to forecast point-of-consumption FRC is that, while FRC at the point-of-consumption is a continuous variable, the risk-based targets operate on a binary classification approach to determine if there is sufficient FRC at the point-of-consumption. In this study, we included two objectives that directly measure the classification performance of the model: recall and precision. Both describe the probability of misclassifying if there will be adequate FRC at the point-of-consumption.

Recall is measured as the ratio of true positives to the sum of true positives and false negatives. For this study, a true positive is defined as a sample where the observed and predicted point-of-consumption FRC are both below 0.2 mg/L (i.e., a sample where the base learner model correctly predicts water will have insufficient FRC at the point-of-consumption). A false negative occurs when the observed point of consumption FRC is below 0.2 mg/L but the predicted FRC is above 0.2 mg/L (i.e., the model incorrectly predicts that there will be adequate FRC at the point-of-consumption). Thus, recall can be thought of as the probability that the model will correctly predict if point-of-consumption FRC will be below 0.2 mg/L.

$$Objective\ 4: Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} = \frac{True\ Positives}{Total\ Samples\ with\ Observed\ FRC\ below\ 0.2\frac{mg}{L}} \tag{5-4}$$

Precision is measured as the ratio of true positives to the sum of true positives and false positives. The definition of true positive is the same as for recall, and a false positive occurs when the model predicts that point-of-consumption FRC will be below 0.2 mg/L, but the measured point-of-consumption FRC is above 0.2 mg/L (i.e., the model incorrectly predicts that there will be inadequate FRC at the point-of-consumption). Thus, precision can be thought of as

the probability that point-of-consumption FRC will actually be above 0.2 mg/L when the model predicts that point-of-consumption FRC will be below 0.2 mg/L.

$$Objective\ 5: Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} = \frac{True\ Positives}{Total\ Samples\ with\ predicted\ FRC\ below\ 0.2\frac{mg}{L}} \quad (5\text{-}5)$$

Recall and precision are both positively oriented scores (higher score is better) ranging from 0 to 1.

Preference Based Methods

We developed three preference-based methods for multi-objective training where we combined multiple objectives into a single objective using a weighted sum. We used a weighted sum instead of multiplication for two reasons. First, we found that the backpropagation algorithm did not converge effectively when multiplying more than two objectives together. While this problem has not been described in past studies, it was likely a problem due to discontinuities in the error surface arising from binary classification objectives (recall and precision) combined with differences in potential scale of the other objectives. The second reason for choosing a weighted sum approach is that it allowed us to prioritize the objectives differently. However, a challenge with using a weighted sum is finding appropriate weights for the different objectives. We developed three methods to weighting the various objectives, described below.

Method 1

The first preference-based method used a weighted sum of only the first three objectives and sought to rebalance the underdispersion that we observed when training ensembles with MSE. To obtain the weights for each of the three objectives we first trained an ensemble of ANNs using mean squared error (MSE). We then evaluated the mean performance for each objective ($\alpha\ score, +\beta\ score, r\ score$) obtained by the ensemble trained with MSE. We set the weights for each objective inversely proportional to the score obtained by the model trained using MSE. To achieve this, we set the weight for the objective with the best (lowest) score to 1 and set the weights for the other objectives as the ratio of each objective's score to the value of the objective with the lowest score, using the scores obtained by the model trained with MSE. Thus, if, for

example, the beta score was the objective with the lowest score, the combined multi-objective cost function would be:

$$Cost = \frac{\alpha\ score_{MSE}}{\beta\ score_{MSE}} * \alpha\ score + \beta\ score + \frac{r\ score_{MSE}}{\beta\ score_{MSE}} * r\ score \quad (5\text{-}6)$$

where the $\alpha, \beta, r\ score_{MSE}$ values refer to the scores for those objectives obtained by the ensemble model trained with MSE. The cost function in Equation 5-6 was then used to train a new ensemble of 200 base learners. Only the first three objectives were included as these objectives are derived from the components of the MSE score, whereas recall and precision cannot be derived directly from the MSE score.

Method 2

This preference-based method used all five objectives, however, unlike the other preference-based methods, a single weighted sum was not used. Instead, each base learner was provided with a different weighting for each objective with the intention of optimizing each base-learner to a different location in the objective space, thus ensuring a balanced and full exploration of the output space. The weights assigned to each objective for each base learner were obtained using a Das-Dennis grid which is an *m*-dimensional grid with *p*-partitions, where *m* is the number of objectives, and *p* is selected based on the desired number of solutions. For this case, *m* is five since there are five objectives, and *p* was selected to be 6 as this yields a grid of 210 possible weights, which is the closest to the ensemble size of 200 selected in Section 5.4.3 above. With six partitions, the possible weights for each objective are {0, 0.167, 0.333, 0.5, 0.666, 0.8333, 1}, with the Das-Dennis grid providing a matrix of all possible combinations of weights for each objective.

Method 3

The third preference-based method used a single weighted sum of the five objectives, as shown in Equation 5-7.

$$Cost = \alpha_{weight} * \alpha\ score + \beta_{weight} * \beta\ score + r_{weight} * r\ score - Recall_{weight} * Recall - Precision_{weight} * Precision \quad (5\text{-}7)$$

To determine the weights assigned to each objective, successive random searches were used to narrow the search space to an optimal weight combination. In the initial search, all weights were

randomly sampled between 5 and 100 in intervals of 5. The performance of the resulting ensembles were then used to narrow the search space until an optimal weighting combination could be found. Unlike Methods 1 and 2, this method selects the weights based on the performance of the resulting ensembles and not a pre-defined assumption of what should produce good performance. There are however two challenges with this method. First, several rounds of random search are required to narrow the search space, which is time consuming. Second, to ensure that we select a weighting combination that is generalizable, we need to identify a combination of weights for the five objectives that balances performance across multiple sites.

No-Preference Methods

Method 4

The fourth multi-objective training method was a no-preference training method using the NSGA-III MOGA developed by Deb and Jain (2014) to set the weights and biases. This algorithm expands the popular NSGA-II algorithm, which has been previously demonstrated to perform well for training ANNs (Chatterjee, Sarkar, Dey, et al., 2017; Chatterjee, Sarkar, Hore, et al., 2017; de Vos & Rientjes, 2008), for "many-objective" optimization (multi-objective optimization with more than two objectives).

As a genetic algorithm, the NSGA-III MOGA adheres to the following general process. An initial set of solutions are generated, their "fitness" or performance on the training objectives is evaluated, and genetic operators are used to generate offspring solutions from these initial parents. Then, the fitness of the combined set of parent and offspring solutions is evaluated, and the most-fit solutions from the combined set of parents and offspring are used to generate the next generation of offspring. This procedure then continues for a fixed number of iterations or until some other stopping criteria is met.

The main innovations of the NSGA MOGAs are their approaches to evaluating fitness of solutions for multi-objective problems. The NSGA-III algorithm uses a fast, elitist sorting mechanism to sort the solution set for each generation of the genetic algorithm to identify and prioritize Pareto-optimal or non-dominated solutions through the formation of Pareto frontiers. Pareto-optimal solutions are solutions where no other solution exists that could perform better on one of the training objectives (in this case either $\alpha\ score$, $\beta score$, $r\ score$, $Recall$, or

$Precision$), without decreasing performance on another objective. Pareto frontiers are used to group solutions on their level of domination. The first Pareto frontier contains all non-dominated solutions, the second frontier includes all solutions that are non-dominated after the first frontier is removed, and so on. Both the NSGA-II and NSGA-III algorithms sort solutions by their Pareto frontier to prioritize including non-dominated or less-dominated solutions in each subsequent generation.

The difference between the NSGA-II and NSGA-III algorithms is how they ensure a diverse set of solutions. NSGA-II used a crowding operator to select for solutions that provide a balanced exploration of the Pareto-front. However, as the number of objectives increases, this becomes computationally inefficient so the NSGA-III algorithm uses a niching operator to associate solutions to different locations in the objective space using the same Das-Dennis grid described for Method 2. To reduce the optimization time required, we provided an initial set of weights and biases by training an ensemble of 212 base learners (number of solutions recommended by Deb and Jain (2014) of the NSGA-III algorithm when using five objectives) and using the trained weights and biases as the starting point for NSGA-III optimization. We allowed for 1000 generations of the genetic algorithm, and implemented a regularization procedure whereby, if the mean scores for all objectives had stopped improving for 20 generations, the best, previous set of weights and biases would be restored and the magnitude of mutation would be reduced to focus the search near this previous, best set of solutions. This was a simple regularization approach inspired by early stopping procedures used to train ANNs. We then used the final set of solutions as the weights and biases of the ensemble model. We found that this was necessary as after many generations without regularization, the performance for some objectives would become very poor.

5.4.5 Ensemble Verification Metrics

Ensemble verification metrics are measures used to assess the probabilistic forecasting performance of an ensemble model. Unlike deterministic scores like MSE or NSE, these scores evaluate the ensemble output as a distribution. Probabilistic forecasts were derived from the ensemble by weighting the predictions of each base learner and then combining these predictions into a probability density function (pdf). In this study each base learner was equally weighted, so the weight assigned was equal to $1/M$ where M is the number of base learners in the ensemble.

Since the output of the models were probabilistic forecasts, deterministic scores such as the MSE or NSE are not useful for evaluating the performance of the forecasts (Boucher et al., 2009; Hamill, 2001). Instead, scores and performance metrics were selected which could evaluate the probabilistic forecasts of each ensemble according to the goals listed in Section 5.3. Each of the ensemble verification metrics listed below were evaluated for the test dataset only (see Section 5.4.3) to verify each ensemble's ability to forecast on future data. 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. F refers to the full set of forecasted point-of-consumption FRC concentrations forecasted by the ensembles, F_i refers to the forecast associated with the i^{th} observation, and f_i^m refers to the prediction by the m^{th} base learner in the ensemble on the i^{th} observation 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 $m = M$.

Percent Capture

Percent Capture is a measure that has been commonly used to evaluate the effectiveness of probabilistic and possibilistic models (Alvisi & Franchini, 2012, 2011; Khan & Valeo, 2017, 2016). The Percent Capture is the percentage of observations where the observed point-of-consumption FRC concentration was within the limits of the ensembles forecast. Observation o_i is considered captured if $f_i^0 \leq o_i \leq f_i^M$. The Percent Capture is a positively oriented score, with an upper limit of 100% and a lower limit of 0%. We calculated the Percent Capture for both the overall dataset and for only observations with point-of-consumption FRC below 0.2 mg/L. This approach is common in literature of probabilistic forecasting with data-driven models as an approach to assess both the overall dispersion as well as the model's skill in capturing values outside of an acceptable threshold (Alvisi & Franchini, 2012, 2011; Khan & Valeo, 2017, 2016)

Reliability Metrics for Ensemble Verification

Ensemble reliability refers to the similarity between the observed and forecasted probability distributions. Numerous reliability metrics have been developed in the literature, this study uses three reliability metrics which are commonly used in atmospheric sciences and which have been applied outside of atmospheric sciences for hydrological applications (Boucher et al., 2009).

CI Reliability Diagram

The first reliability metric used for ensemble verification was the reliability diagram. The reliability diagram plots observed relative frequency of events against the forecast probability of that event. Boucher et al. (2009) adapted this diagram for ensemble modelling as the confidence interval (CI) reliability diagram which compares the frequency of observed values with the corresponding CI of the ensemble, where the ensemble CIs are derived from the sorted forecasts of the base learners (for example, the ensemble 90% CI would include all of the forecasts between $f^{0.05M}$ and $f^{0.95M}$). We extended this this further by plotting the Percent Capture of each CI within the ensemble against the CI level. For each ensemble model we plotted the CI reliability for the 10% to 100% CI levels at 10% intervals as well as at the 95% and 99% CI.

The reliability diagram and CI reliability diagram are visual indicators of ensemble reliability with the ideal model having all observations plotted along the 1:1 line showing that the observed probabilities are equal to the forecasted probabilities. De Santi et al. (2021) developed a numerical score for the CI reliability diagram which calculated the squared distance between the Percent Capture within each CI and the ideal Percent Capture in that CI. This was calculated for each CI threshold, k, from 10% to 100% in 10% increments as shown in Equation 5-8. The CI Reliability Score measures the horizontal distance between the Percent Capture and the 1:1 line for each CI. 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. CI Reliability diagrams were plotted and the CI reliability score calculated for both the overall data set and for forecast-observation pairs where the observed point-of-consumption FRC concentration was below 0.2 mg/L.

$$CI\ Reliability\ Score = \sum_{k=0.1}^{1}\left(j - Percent\ Capture\ in\ CI_j\right)^2 \quad (5\text{-}8)$$

Rank Histograms

The Rank Histogram (RH) is another visual tool used to assess the reliability of ensemble forecasts. To construct the rank histogram, for each forecast-observation pair, the observation o_i is added to the sorted vector of forecast values F_i, with the new vector having $M + 1$ members. A rank is assigned to the observed value based on where in the set of forecasted values it falls. This is repeated for each forecast-observation pair. The RH is the histogram of the ranks assigned to each observation, o_i. If the forecast and observed probabilities are the same, then any observation is equally likely to occur in any of the M+1 ranks, which would result in a flat rank

histogram. 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 RH is a visual tool, but Candille and Talagrande (2005) proposed a numerical score, the δ score which measures the deviations from flatness (Equation 5-9). The ideal score is 1 with scores much greater than 1 indicating substantial deviations from flatness and scores less than 1 indicating interdependence between ensemble predictions (Candille & Talagrand, 2005). This latter case is very uncommon. The δ score was calculated for each model both for the overall dataset and for only those observations where the observed point-of-consumption FRC was below 0.2 mg/L.

$$\delta = \frac{\Delta}{\Delta_o} \tag{5-9}$$

The two components of the δ score are shown in Equations 5-10 and 5-11 where M is the total number of ensemble members, 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 & Talagrand, 2005).

$$\Delta = \sum_{k=1}^{M+1} \left(s_k - \frac{I}{M+1} \right)^2 \tag{5-10}$$

$$\Delta\_o = \frac{I*M}{M+1} \tag{5-11}$$

Continuously Ranked Probability Score

The continuously ranked probability score (CRPS) is a commonly used metric for evaluating probabilistic forecasts. It measures the area between the forecast cumulative distribution function (cdf) and the observed cdf for each forecast observation pairing. The CRPS measures not only model reliability but also sharpness and uncertainty, where the sharpness penalizes excess spread of the ensemble forecast, and uncertainty relates to the uncertainty in the observed data (Ferro, 2014; Hersbach, 2000). For a given forecast-observation pair, the cdf of the forecast is calculated from the probability distribution of the predictions of the base learners of the ensembles. Since each observation is a discrete value, it is represented with the Heaviside function $H\{x \geq x_a\}$ which is 0 for all concentrations of point-of-consumption FRC below the observed FRC and 1 for all predicted concentrations of point-of-consumption FRC above the observed concentration.

The calculation of the CRPS is given in Equation 5-12 where F_i is the cdf of the forecast values for observation o_i. Note that Equation 5-12 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.

$$CRPS = \int_{-\infty}^{\infty}(F_i(x) - H\{x \geq o_i\})^2 dx \quad (5\text{-}12)$$

Hersbach (2000) derived a calculation of CRPS for ensemble models that 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:

$$\overline{CRPS} = \sum_{n=1}^{N} \overline{g_n}[(1 - \overline{o_n})p_n^2 + \overline{o_n}(1 - p_n)^2] \quad (5\text{-}13)$$

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

Hersbach (2000) also decomposed the ensemble CRPS calculation into its reliability, resolution, and uncertainty scores. The reliability term is of particular interest as it reflects the similarity between the observed and forecast probability distributions. The average reliability term $\overline{Reli}$ is calculated as shown in Equation 5-14 (Hersbach, 2000).

$$\overline{Reli} = \sum_{n=1}^{N} \overline{g_n}\,(\overline{o_n} - p_n)^2 \quad (5\text{-}14)$$

Skill Scores

While in many cases, the score obtained for the performance metrics above are highly informative (e.g., Percent Capture provides clear insight into the amount of observations captured, the CI Reliability Score provides a measurable distance from the ideal model), it can also be helpful to normalize scores, both to facilitate comparison of scores between sites, and to allow for a comparison of the relative improvement between metrics with different scales. Skill scores provide a means of normalizing a performance metric using a baseline and an ideal score. For our case, we used the scores obtained by an ensemble of 200 neural networks trained with MSE as the baseline for developing the skill score. The ideal score for Percent Capture is 100%

capture, the ideal CI reliability score, CRPS, and CRPS reliability term are all 0, and the ideal δ-score is 1. Any performance metric can be converted to a skill score using Equation 5-15. The skill score is normalized between negative infinity and 1, with 1 meaning that the score obtained is the ideal score and a positive score indicating improvement over baseline. A skill score of 0 means that there is no difference between the score obtained and the baseline, and a negative score indicates that the score obtained is worse than the baseline.

$$Skill\ Score = \frac{score\ obtained - baseline}{ideal\ score - baseline} \tag{5-15}$$

The baseline scores for the ensembles trained on MSE are presented in Table 5-2.

Table 5-2:Baseline performance for ANN ensembles trained using MSE

Site and Variable Combination	PC	$PC_{<0.2}$	CI_{score}	$CI_{score_{<0.2}}$	δ	$\delta_{<0.2}$	$\overline{CRPS}$	$\overline{Reli}$
Bangladesh IV1	22	25	2.86	2.50	153	97	0.16	0.087
Bangladesh IV2	33	31	2.48	2.48	52	42	0.21	0.11
Tanzania IV1	31	31	2.64	2.78	19	24	0.10	0.057
Tanzania IV2	30	50	2.69	2.61	5.92	1.98	0.15	0.065
Nigeria IV1	30	0	2.36	3.85	14.31	4.00	0.10	0.056
Nigeria IV2	33	25	2.28	3.41	12.86	2.49	0.11	0.063

5.5 Results and Discussion

The following section presents the performance of the models trained using the four multi-objective training methods listed in Section 5.4.4 and compares the performance obtained to the baseline performance obtained for ensembles trained with MSE, as well as comparing the performance obtained through multi-objective training to other approaches that have been investigated for improving the performance of ANN ensembles used to probabilistically forecast point-of-consumption FRC, and to generate risk-based FRC targets, namely ensemble post-processing and cost-sensitive learning.

5.5.1 Selection of Weights

Prior to comparing the different multi-objective training methods, the following sub-section discusses the selection of objective weights for Methods 1 and 3. The objective weights for Method 2 are pre-determined using the Das-Dennis grid.

Method 1

The weights for the three objectives ($\alpha\ score, \beta\ score, r\ score$) were determined in Method 1 by first training an ensemble of 200 ANNs using MSE as the cost function and calculating the mean scores for these objectives and then using inverse weighting to train a new ensemble. For each site and variable combination, the weight for the objective with the lowest (best) score was set to 1 and the weights for the other two objectives were taken as the ratio of the score for that objective to the objective with the best score. Table 5-3 shows the weights assigned to each objective for each site and variable combination. Table 5-3 shows that in all cases the β score was the lowest score for the models trained with MSE by a very substantial margin (the $\beta\ score$ was between 50 and 200 times better than the other scores). This indicates that the ANN models trained with only MSE are very proficient at reproducing the mean of the observed data, reinforcing our understanding that training ensembles of ANNs with MSE tends to produce a regression to the mean behaviour (De Santi et al., 2021). Furthermore, the very poor α score reinforces that these models are not proficient at reproducing the spread of the observations, thus explaining the poor dispersion and reliability.

Table 5-3:Objective weights used for multi-objective training Method 1. Weights are determined as the ratio of each objective's score to the lowest score obtained by an ensemble trained on MSE. This table indicates that the ensembles trained with MSE performed best on the β objective, and worst on the α objective, so the weightings were assigned to counterbalance this.

Site and Variable Combination	**α weight**	**β weight**	**r weight**
Bangladesh IV1	123	1	92
Bangladesh IV1	72	1	50
Tanzania IV1	140	1	79
Tanzania IV2	126	1	82
Nigeria IV1	52	1	26
Nigeria IV2	48	1	31

Method 3

The training function in Method 3 was a weighted sum of the $\alpha\ score, \beta\ score, r\ score$, $Recall$, and $Precision$. The weights for these five objectives were obtained using successive tests where the weights for each objective were randomly sampled and the ensemble performance was evaluated using the ensemble verification metrics to successively narrow the search space through multiple rounds of random search. In each round of random search, 100 possible weighting combinations were considered, making it challenging to evaluate all alternatives for each site, variable combination, and performance metric. To simplify the comparison of the alternatives we calculated the net improvement achieved by each alternative at each site for each variable combination by taking the sum of the skill scores. Thus, for each alternative we had six net improvement scores, one for each site and variable combination, as an indicator of the magnitude of improvement for each weighting combination. We also considered the number of positive skill scores for each site and variable combination, as an indicator of the consistency of improvement for each weighting. Figure 5-1 shows the minimum and maximum net improvement scores at each site for each round, as well as the sum of the net improvement scores and boxplots of the number of positive net improvement scores in each round. This figure shows that while in many cases the end of the optimization did not yield the highest maximum scores, the optimization process effectively eliminated many of the worst scores. We see this with the

increasing minimum values as well as the convergence of the boxplots towards the higher positivity values, indicating that more solutions produced a positive net improvement score for more sites and variable combinations.

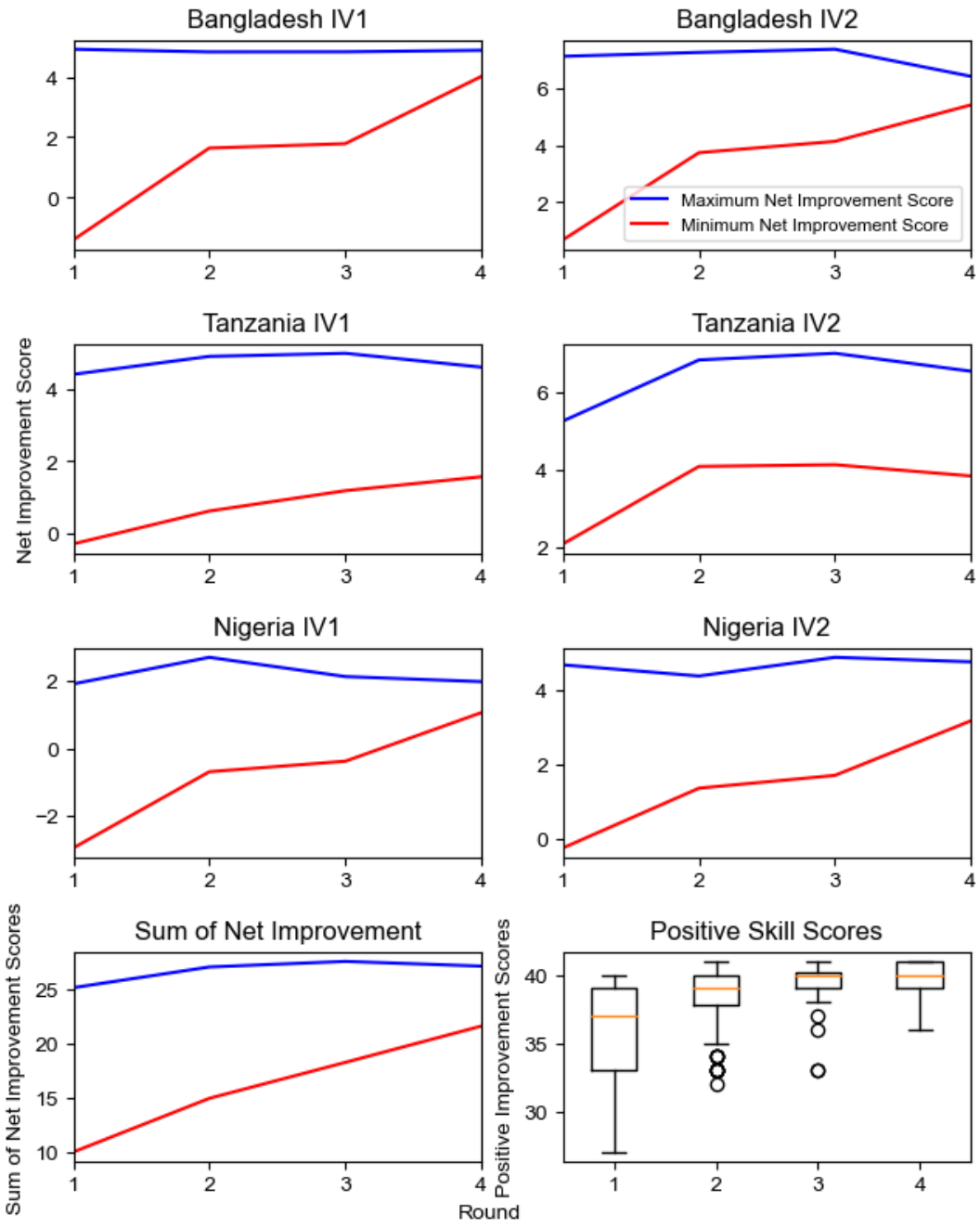

Figure 5-1: Summary of max and min performance improvement over 4 rounds of random search to determine weights for the multi-objective training Method 3. This figure shows that

over 4 rounds of optimization the net improvement at each site increases, as does the number of weighting combinations that produce improvement. This indicates that the successive random search optimization process used effectively navigated the search space towards good alternatives.

The weight range for each objective in the first round of random search was between 5 and 100. using the evaluation methods above, we constrained the search space as follows in the subsequent rounds:

- Round 2: Constrain the α and recall weights between 50 and 100
- Round 3: Constrain the α weight between 60 and 100, constrain the β weight between 5 and 65, constrain the r weight between 5 and 60.
- Round 4: Constrain the α weight between 80 and 100, constrain the β weight between 20 and 35, constrain the r weight between 20 and 35, constrain the recall weight between 75 and 85, constrain the precision weight between 35 and 50.

At the end of the fourth round, we identified a preferred weight combination of:

- α weight: 85
- β weight: 20
- r weight: 20
- Recall weight: 80
- Precision weight: 50

From these weights, we see that the α and recall objectives are most critical for obtaining the best ensemble verification performance. The importance of the α score objective is likely in part because the high variability in post-distribution FRC decay makes matching standard deviations critical. The importance of the recall objective is likely a result of having three performance objectives only evaluated on observations where the point-of-consumption FRC is below 0.2 mg/L as recall prioritizes performance on these samples.

5.5.2 Comparison of Multi-Objective Methods

Table 5-4 shows the scores obtained for each ensemble verification metric at each site with each variable combination across the four multi-objective training methods. Figure 5-2 shows the sum

of the normalized skill scores for each of the four multi-objective training methods as the "Net Improvement Score", as well as a tally of the number of positive skill scores achieved by each multi-objective training method at each site for each variable combination. Together Table 5-4 and Figure 5-2 show that all of the multi-objective training methods produced ensembles with better forecast dispersion and reliability than the baseline ensembles trained with only MSE, as shown by the Percent Capture, CI Reliability, and δ scores in Table 5-4, as well as the high net improvement scores and large number of positive skill scores obtained across all methods. This is consistent with findings of past studies that have also found that multi-objective training of neural networks tends to produce better performance than using a single error metric, especially in complex engineering applications (Chatterjee, Sarkar, Dey, et al., 2017; Chatterjee, Sarkar, Hore, et al., 2017; de Vos & Rientjes, 2008). Those studies tended to train deterministic models, but the results in Tables 5-4 and in Figure 5-2 highlight that multi-objective training is effective for training ANNs for probabilistic ensembles. Interestingly, Table 5-4 shows that in many cases the ensemble forecasts were able to capture 100% of the observations, including all observations with point-of-consumption FRC below 0.2 mg/L, meaning that the ensembles produced using multi-objective training are not underdispersed. This represents a substantial improvement over both post-processing (Chapter 3) and cost-sensitive learning approaches (Chapter 4) that have been used to address underdispersion in past research which were able to reduce, but not eliminate, underdispersion. The improvement in the CI Reliability score and the δ-score indicate that multi-objective training also improves the reliability, indicating that the distributions of observed and predicted data become more similar with multi-objective training. This is critical for generating probabilistic FRC recommendations, as an underlying assumption of these targets is that the model accurately reproduces the underlying distribution of the observed data.

Table 5-4: Summary of skill scores for each ensemble verification metric for multi-objective method as well as the sum of the skill scores (Net column). The baseline used to calculate the skill scores was the MSE performance shown in Table 5-3.

Site and Variable Combination	Method	PC	$PC_{<0.2}$	CI_{score}	$CI_{score_{<0.2}}$	δ	$\delta_{<0.2}$	$\overline{CRPS}$	$\overline{Reli}$
Bangladesh IV1	1	96%	100%	0.04	0.03	2.59	1.72	0.20	0.07
	2	100%	100%	0.61	0.77	2.93	2.24	0.15	0.05
	3	74%	93%	0.29	0.22	37.87	7.15	0.20	0.09
	4	88%	87%	0.06	0.06	8.09	4.32	0.27	0.17
Bangladesh IV2	1	99%	100%	0.11	0.070	1.30	1.08	0.13	0.014
	2	100%	100%	0.43	0.26	2.24	1.55	0.17	0.07
	3	92%	100%	0.04	0.01	3.08	1.31	0.23	0.12
	4	73%	54%	1.64	2.42	23.02	19.32	0.76	0.64
Tanzania IV1	1	68%	59%	0.90	1.15	4.93	5.04	0.15	0.059
	2	100%	100%	0.07	1.85	1.32	2.51	0.12	0.01
	3	78%	100%	0.93	1.17	6.76	5.11	0.15	0.066
	4	79%	76%	0.23	0.47	7.47	8.07	0.067	0.014
Tanzania IV2	1	83%	89%	0.56	0.18	1.33	0.96	0.16	0.08
	2	100%	100%	0.15	0.40	0.98	0.96	0.10	0.010
	3	83%	89%	0.69	0.41	1.26	1.18	0.15	0.069
	4	74%	50%	0.60	0.83	1.52	1.98	0.22	0.12
Nigeria IV1	1	70%	33%	1.30	3.29	7.14	3.32	0.10	0.052

Site and Variable Combination	Method	PC	$PC_{<0.2}$	CI_{score}	$CI_{score_{<0.2}}$	δ	$\delta_{<0.2}$	$\overline{CRPS}$	$\overline{Reli}$
	2	96%	67%	1.42	2.96	5.06	1.65	0.08	0.036
	3	63%	0%	0.81	3.85	9.10	6.00	0.08	0.039
	4	91%	100%	0.17	0.85	2.18	0.99	0.16	0.083
Nigeria IV2	1	59%	33%	1.28	1.87	9.52	2.99	0.081	0.029
	2	100%	100%	0.49	2.58	1.68	1.31	0.066	0.021
	3	76%	17%	0.30	3.27	3.86	4.32	0.065	0.015
	4	74%	67%	0.73	0.36	4.40	0.99	0.24	0.14

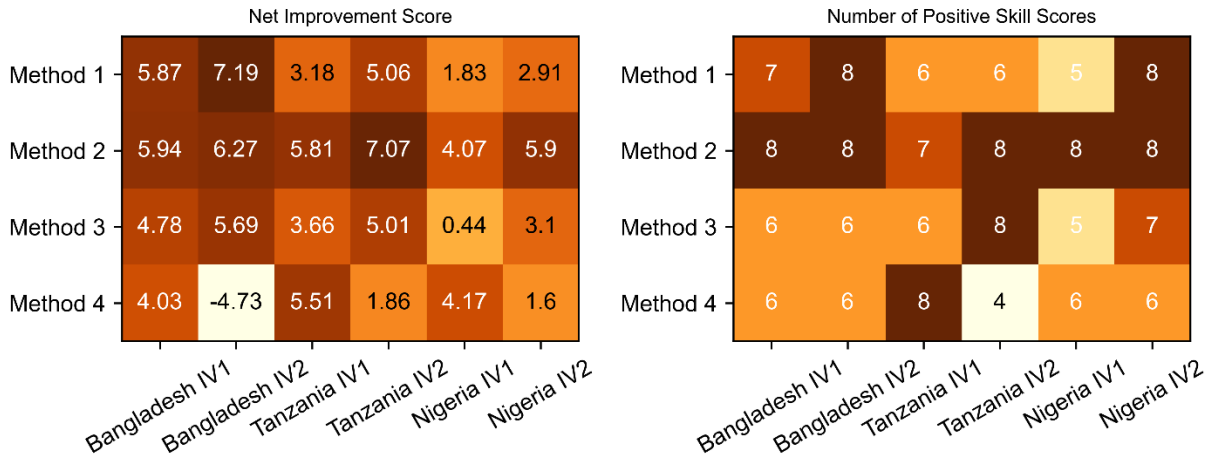

Figure 5-2: Comparison of net improvement and number of positive skill scores for all multi-objective training methods, showing that consistently multi-objective training Method 2 produces the highest net improvement scores and the most positive skill scores (indicating consistent improvement).

When comparing the multi-objective training methods, we see that while all of the multi-objective training methods improve the Percent Capture, CI Reliability score, and δ-score, the second multi-objective training score improved all eight ensemble verification metrics for all model configurations except for the Tanzania IV1 model. No other method was able to yield such consistent improvements in performance over the baseline. This is also critical as CRPS tends to penalize sharpness and overdispersion, so this shows that the improvement in forecast reliability obtained by training the ensemble base learners using Method 2 outweighed any resulting loss in sharpness or resolution. Furthermore, the multi-objective training Method 2 produced the highest net improvement score for four out of six site and variable combinations. In the two cases where Method 2 did not produce the highest net performance increase (Nigeria IV1 and Bangladesh IV2), it produced the second highest net improvement score. Together, these indicators show that, more than any other training method, the second multi-objective training method produces consistent, substantial improvement across many ensemble verification metrics across all sites and input variable combinations.

The reason for the greater performance produced by Method 2, as opposed to the preference-based methods is likely due to the increased and targeted ensemble diversity obtained through this method. In the other preference-based methods, ensemble diversity arises only from the randomization of initial weights and biases and the randomization of the calibration data into training and validation datasets. All of these are considered implicit approaches to creating ensemble diversity, meaning that we introduce some randomness in the hope that it will create sufficiently diverse ensembles (Brown et al., 2005). Method 2 retains these implicit diversity creation methods and adds additional diversity by varying the cost function for each base learner. Additionally, because the diversity in the cost functions is controlled by the Das-Dennis grid, which is not randomized and instead deliberately searches different areas of the objective space, providing a balance between the objectives used. This diversity introduced through the cost function can be thought to exist somewhere between implicit and explicit diversity. It is not entirely explicit, as we do not directly use the results of the other ensemble members to inform the training of the rest of the ensemble in the way that explicit diversity generating algorithms such as negative correlation or boosting, however, it is much more structured in its exploration of the search space than a typical implicit diversity generating approach (Brown et al., 2005). It is likely this increased diversity, and the ability of each objective to effectively align with the goals of the multi-objective training process

Method 2 also outperformed the no-preference multi-objective training method, Method 4. This is surprising as a past study that showed that multi-objective training using NSGA-II, a related algorithm, outperformed preference-based methods (de Vos & Rientjes, 2008), and many multi-objective studies have shown that no-preference methods tend to outperform single-objective training (Albuquerque Teixeira et al., 2000; Chatterjee, Sarkar, Dey, et al., 2017; Chatterjee, Sarkar, Hore, et al., 2017; Taormina & Chau, 2015). However, there are several key differences between those past studies and the current research. First, of the above studies only the de Vos and Rientjes study considered preference-based multi-objective training methods. Second, the current research used multi-objective training to form probabilistic ensembles; not to train a deterministic model, and as such we are not optimizing to find a single best solution, we are instead trying to match the distribution of the observed data. Finally, the de Vos and Rientjes (2008) used a multiplicative preference-based method. By multiplying objectives together, that previous study could not prioritize different objectives. That Method 2 outperformed Method 4 is

also interesting as the Das-Dennis grid used to set the objective weights in Method 2 is also used by the NSGA-III algorithm in Method 4 as part of its fitness selection. However, in the NSGA-III algorithm, the Das-Dennis grid only used after selecting Pareto-optimal solutions. Additionally, the Das-Dennis grid is not used to inform the selection of model parameters directly, it instead informs a niching operation in the NSGA-III algorithm that prioritizes diverse solutions amongst the selected solutions which is only used to select from the solutions on the last Pareto front if there are too many solutions in that front to obtain the correct number of parents for the next generation. Furthermore, this niching operation associates solutions to the nearest grid point, however, this does not mean that a solution is close to that grid point in absolute terms. When examining these considerations together, we see that the Das-Dennis grid, while used in the NSGA-II algorithm is used only secondarily to the elitist sorting operation, and is only ever applied indirectly, and cannot be used to actually ensure that solutions are diverse. In contrast, Method 2 uses the Das-Dennis gird to train each base learner to a different location in the objective-space obtains better model performance, which produces a more balanced, and diverse set of solutions.

An additional benefit of Method 2 shown in Table 5-4 is that there is less of a difference in performance between the IV1 and IV2 models for a given site as compared to the difference seen when using the baseline model. The IV2 models still tend to outperform the IV1 models, highlighting the benefits of incorporating additional water quality variables, however, the magnitude of this difference is much smaller when using the multi-objective training Method 2 than for the baseline. This is critical as additional water quality variables are not always available, which is why improving the performance of the IV1 model and reducing the difference in model performance between the IV1 and IV2 models was a key area of future improvement identified by De Santi et al. (2021). Past attempts to improve model performance (post-processing and cost-sensitive learning) still saw a large disparity between the IV1 and IV2 ensemble performances, so the ability of Method 2 to reduce this difference and achieve similar performance for both input variable combinations is a crucial improvement that will enable better performance in real-world sites where additional water quality variables may not be consistently available.

Figure 5-3 shows the predictions, CI reliability diagram, and rank histogram for model configurations using Method 2. From this Figure we see substantial differences in behaviour as the data volume changes. In Nigeria and Tanzania, the sites with the lower data volumes, we see some remaining underdispersion, though overall the models very effectively reproduce the underlying data. Conversely, the Bangladesh model forecasts, which have the largest data volumes, are overdispersed. While this is not ideal, this can be easily mitigated by capping the maximum amount of data used to train the models. One additional possible concern is that the forecast range tends to be unrealistically large, with forecast point-of-consumption FRC values above the point-of-distribution concentrations or below 0. However, this is merely an indicator of the models reproducing the underlying uncertainty and can be easily rectified with post-processing to a value within possible limits. The results of this post-processing step are shown in Figure 5-4**.** The post-processing was achieved simply by rewriting impossible values to be within possible limits so as not to affect the underlying distribution of the data. From Figure 5-4 there is no substantial change in the ensemble performance diagrams but all predictions have been restricted to within possible values.

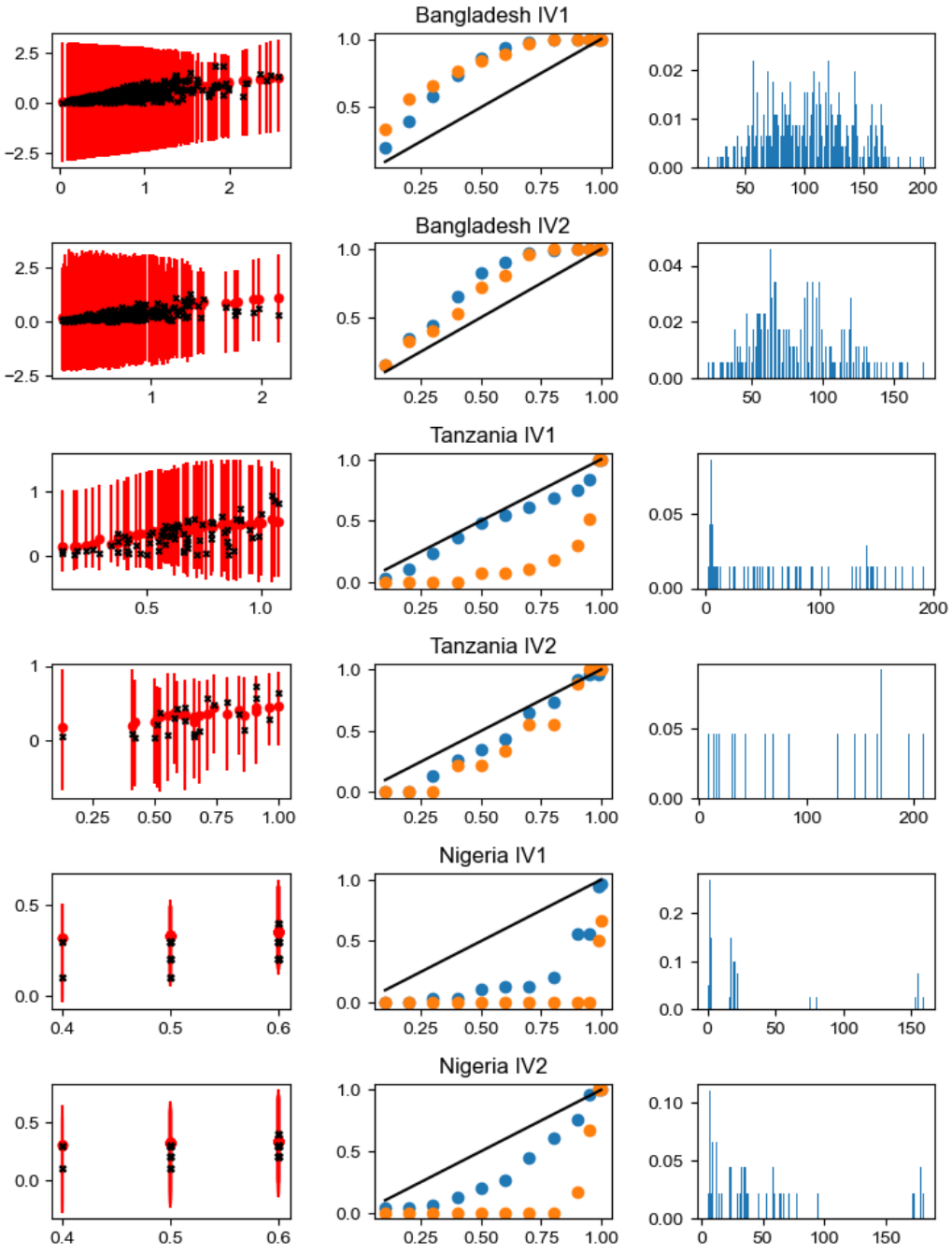

Figure 5-3: Predictions, CI reliability diagram, and Rank Histogram for each site and variable combination using multi-objective training Method 2. From this figure we see that there is some overdispersion in the Bangladesh models, and at all sites there are predictions that are impossible (point-of-consumption FRC lower than 0 or higher than the point-of-distribution concentration). While these predictions are physically impossible, they are an indicator of the models reflecting the high degree of uncertainty present at these sites.

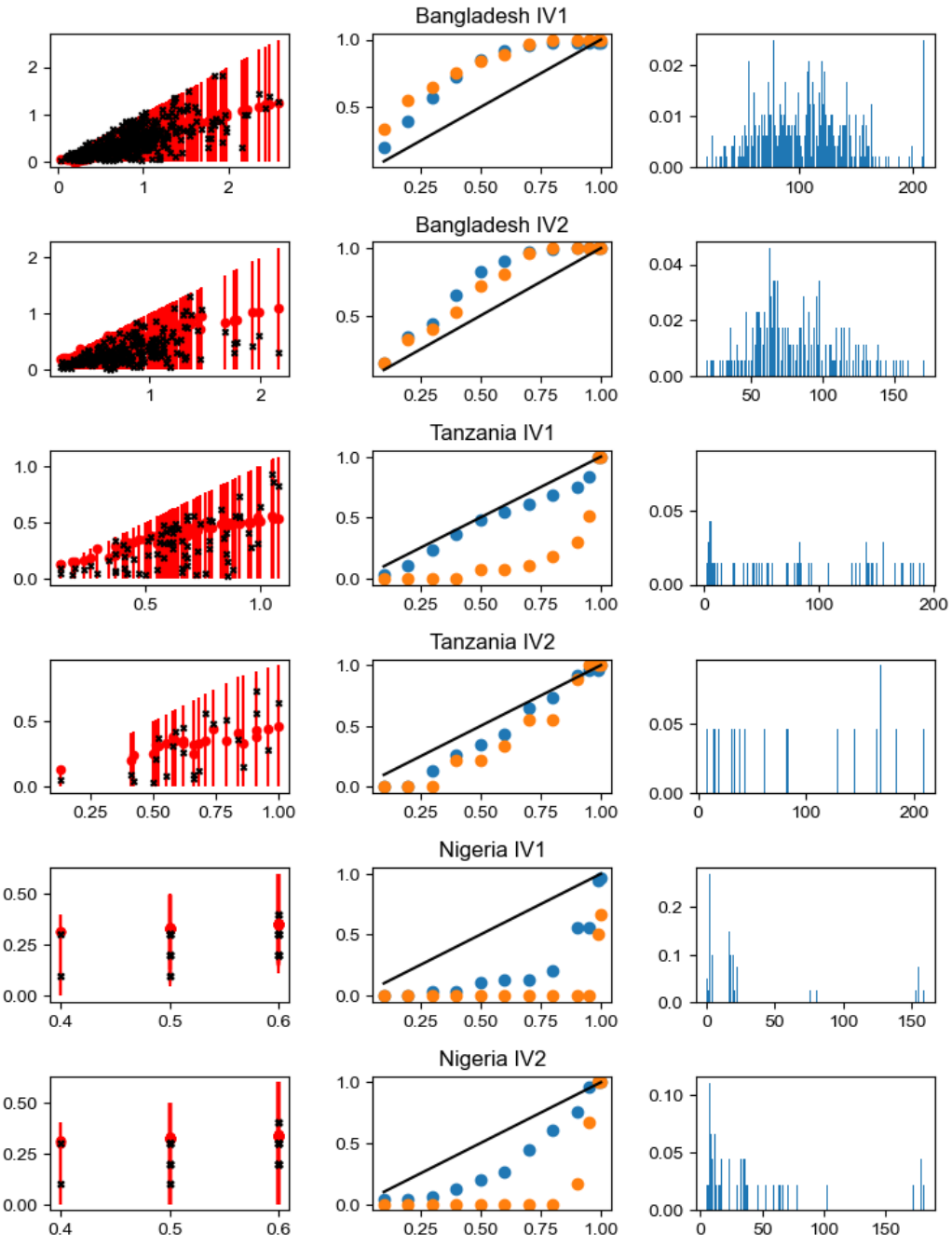

Figure 5-4: Predictions, CI reliability diagram, and Rank Histogram for each site and variable combination using multi-objective training Method 2, after post processing. In comparison to Figure 5-3, this figure includes no unrealistic predictions, however, there is minimal change in the ensemble performance. This indicates that the post-processing technique effectively constrained the predictions to within the range of possible point-of-consumption FRC concentrations without meaningfully impacting the underlying distribution.

5.5.3 Comparison to Ensemble Post-Processing and Cost-Sensitive Learning

Table 5-5 compares the skill scores obtained using multi-objective training Method 2 to three other approaches that have been used to improve forecast dispersion and reliability for the ANN ensemble forecasts of point-of-consumption FRC. The first two approaches use ensemble post-processing to improve the ensemble dispersion and reliability. Both use kernel dressing to improve the ensembles forecasts, with the first approach deriving the kernel bandwidth using the Roulston and Smith (2003) method which was used in Chapter 3 (De Santi et al, 2021) and the second approach deriving the kernel bandwidth using the Wang and Bishop (2005) method which is currently in use for the SWOT version 2 (Appendix C). The third approach compared is the cost-sensitive learning approach described in Chapter 4. For consistency of comparison, in all cases the cost-sensitive learning approach used was KGE with inverse frequency weighting (weighting 3 from Chapter 4), as this cost function and weighting combination produced the best performance most consistently. The comparison in Table 5-5 includes all SWOT sites, not just the three implementation sites, as this allowed for a better evaluation of the generalization of the approaches considered. Note that Table 5-5 only shows the skill scores for five performance metrics are included as the δ scores and CRPS reliability term require discrete ensemble members and are not easily calculated for continuous probability distributions obtained from kernel post-processing. Furthermore, the sum of the skill scores for each approach at each site with each variable combination are shown in Figure 5-5 along with a tally of the number of positive skill scores. From Table 5-5 and Figure 5-5, the multi-objective training Method 2 is the only approach which produced improvement across all five performance metrics for all sites and variable combinations. Furthermore, multi-objective training Method 2 produced the largest net score (largest total magnitude of improvement) for all sites and variable combinations. Based on these two observations, multi-objective training using Method 2 is the most effective method for improving ensemble diversity and reliability of ensemble forecasts of point-of-consumption FRC. Future study should prioritize further optimization and operationalization of this approach over those presented in Chapters 3 and 4 as the superior dispersion and reliability makes it ideal for generating risk-based point-of-distribution FRC targets for refugee and IDP settlements.

Table 5-5: Comparison of approaches to improving ensemble forecast dispersion and reliability. Note for post-processing approach 1, the kernel bandwidth is derived using the best member error method (Roulston & Smith, 2003), and for post-processing approach 2, the kernel is derived using the method proposed by Wang and Bishop (2005).

Site and Variable Combination	Approach	PC	$PC_{<0.2}$	CI_{score}	$CI_{score_{<0.2}}$	$\overline{CRPS}$
Bangladesh IV1	Post Processing approach 1	0.065	0.099	0.034	0.065	0.234
	Post Processing approach 2	0.132	0.198	0.059	0.107	0.294
	Cost Sensitive Learning (KGE Weighting 3)	0.839	0.863	0.673	0.601	-0.075
	Multi-Objective (Method 2)	**1.000**	**1.000**	**0.701**	**0.732**	**0.073**
Bangladesh IV2	Post Processing approach 1	0.103	0.102	0.047	0.049	0.248
	Post Processing approach 2	0.200	0.271	0.078	0.096	0.285
	Cost Sensitive Learning (KGE Weighting 3)	0.818	0.797	0.821	0.806	-0.034
	Multi-Objective (Method 2)	**1.000**	**1.000**	**0.826**	**0.955**	**0.128**
Jordan (2014) IV1	Post Processing approach 1	0.053	0.100	0.019	0.049	0.272
	Post Processing approach 2	0.263	0.300	0.085	0.132	0.272
	Cost Sensitive Learning (KGE Weighting 3)	0.421	0.300	0.777	0.434	0.156
	Multi-Objective (Method 2)	**0.947**	**0.900**	**0.947**	**0.447**	**0.392**
Jordan (2014) IV2	Post Processing approach 1	0.000	0.000	-0.007	0.000	0.434
	Post Processing approach 2	0.067	0.000	0.036	0.039	0.434
	Cost Sensitive Learning (KGE Weighting 3)	0.467	0.500	0.299	0.286	0.215
	Multi-Objective (Method 2)	**1.000**	**1.000**	**0.956**	**0.645**	**0.497**

Site and Variable Combination	Approach	PC	$PC_{<0.2}$	CI_{score}	$CI_{score_{<0.2}}$	$\overline{CRPS}$
Jordan (2015) IV1	Post Processing Approach 1	0.000	0.000	0.016	0.000	-0.777
	Post Processing Approach 2	0.053	0.000	0.026	0.000	-0.095
	Cost Sensitive Learning (KGE Weighting 3)	0.789	1.000	0.638	0.260	-0.557
	Multi-Objective (Method 2)	**1.000**	**1.000**	**0.867**	**0.260**	**0.161**
Jordan (2015) IV2	Post Processing Approach 1	0.000	N/A	0.038	N/A	0.077
	Post Processing Approach 2	0.182	N/A	0.051	N/A	0.204
	Cost Sensitive Learning (KGE Weighting 3)	0.909	N/A	0.866	N/A	0.074
	Multi-Objective (Method 2)	**1.000**	**N/A**	**0.922**	**N/A**	**0.323**
Nigeria IV1	Post Processing Approach 1	0.000	0.000	-0.004	0.000	-3.992
	Post Processing Approach 2	0.000	0.000	-0.031	0.000	-0.653
	Cost Sensitive Learning (KGE Weighting 3)	0.395	0.667	0.751	0.831	-0.010
	Multi-Objective (Method 2)	**1.000**	**1.000**	**0.919**	**0.745**	**0.430**
Nigeria IV2	Post Processing Approach 1	-0.068	0.000	0.011	0.000	-3.786
	Post Processing Approach 2	0.023	0.000	0.025	0.000	-1.266
	Cost Sensitive Learning (KGE Weighting 3)	0.432	0.333	0.691	0.271	-0.018
	Multi-Objective (Method 2)	**1.000**	**1.000**	**0.866**	**0.456**	**0.412**
Rwanda IV1	Post Processing Approach 1	0.000	0.000	-0.029	0.000	0.102
	Post Processing Approach 2	0.000	0.000	-0.009	0.000	0.325
	Cost Sensitive Learning (KGE Weighting 3)	0.632	0.286	0.423	0.189	0.252
	Multi-Objective (Method 2)	**0.947**	**0.857**	**0.820**	**0.254**	**0.371**
Rwanda IV2	Post Processing Approach 1	0.000	0.000	-0.051	0.000	-0.088
	Post Processing Approach 2	0.063	0.143	0.024	0.056	0.250
	Cost Sensitive Learning (KGE Weighting 3)	0.563	0.429	0.664	0.333	0.016
	Multi-Objective (Method 2)	**1.000**	**1.000**	**0.846**	**0.685**	**0.303**
South Sudan IV1	Post Processing Approach 1	0.120	0.143	0.054	0.062	0.330
	Post Processing Approach 2	0.200	0.286	0.085	0.115	0.331
	Cost Sensitive Learning (KGE Weighting 3)	0.560	0.357	0.429	0.317	-0.543

Site and Variable Combination	Approach	PC	$PC_{<0.2}$	CI_{score}	$CI_{score_{<0.2}}$	$\overline{CRPS}$
	Multi-Objective (Method 2)	**1.000**	**1.000**	**0.964**	**0.889**	**0.393**
South Sudan IV2	Post Processing Approach 1	0.077	0.067	0.055	0.068	0.303
	Post Processing Approach 2	0.308	0.333	0.134	0.158	0.303
	Cost Sensitive Learning (KGE Weighting 3)	0.577	0.467	0.724	0.620	0.353
	Multi-Objective (Method 2)	**1.000**	**1.000**	**0.975**	**0.952**	**0.473**
Tanzania IV1	Post Processing Approach 1	0.000	0.000	0.056	0.069	-2.823
	Post Processing Approach 2	0.038	0.057	0.057	0.117	-0.352
	Cost Sensitive Learning (KGE Weighting 3)	0.660	0.657	0.549	0.587	-0.178
	Multi-Objective (Method 2)	1.000	1.000	0.907	0.581	0.174
Tanzania IV2	Post Processing Approach 1	0.000	0.000	0.000	0.072	-1.057
	Post Processing Approach 2	0.000	0.000	0.003	0.000	0.055
	Cost Sensitive Learning (KGE Weighting 3)	0.647	0.750	0.566	0.555	-0.600
	Multi-Objective (Method 2)	**1.000**	**1.000**	**0.995**	**0.901**	**0.346**

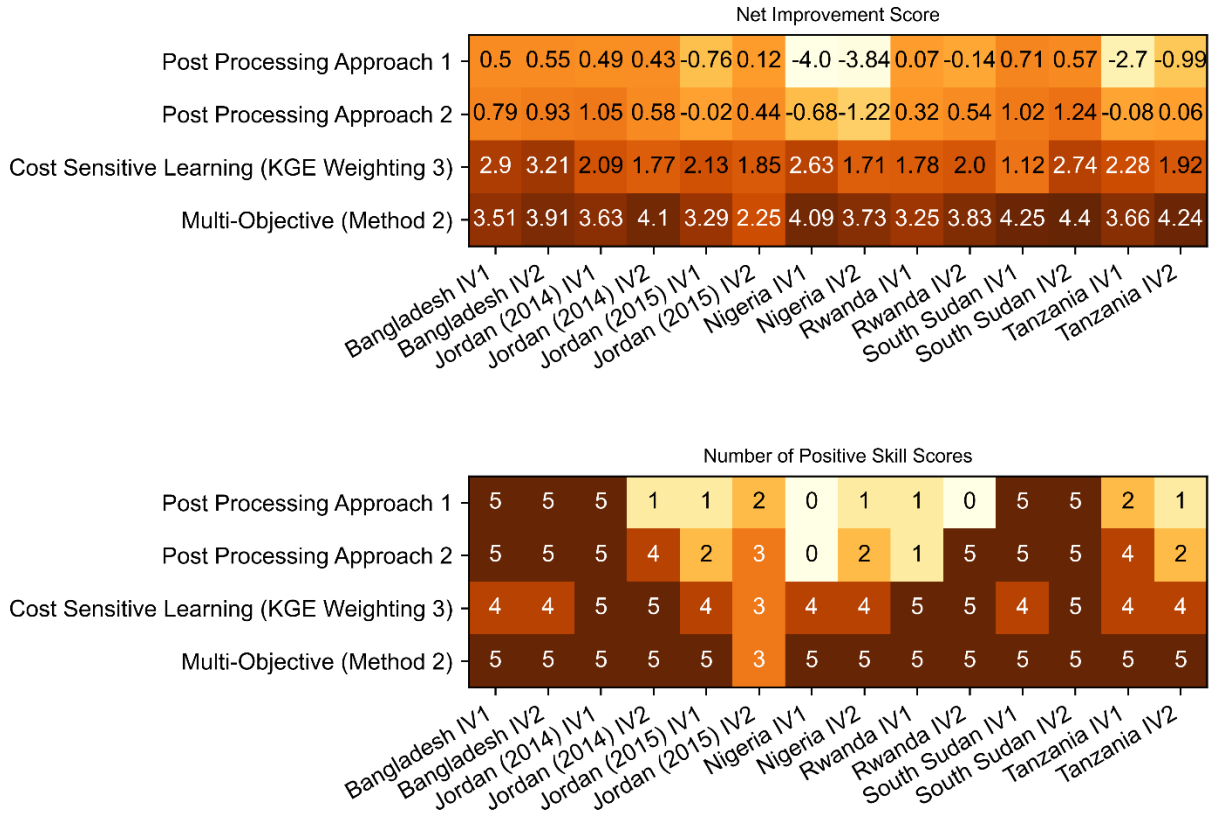

Figure 5-5: Comparison of net improvement and number of positive skill scores for all approaches, showing that consistently multi-objective training Method 2 produces the highest net improvement scores and the most positive skill scores (indicating consistent improvement).

5.6 Conclusion

This research investigated the effectiveness of multi-objective training for improving the dispersion and reliability of ANN ensemble forecasts of point-of-consumption FRC in humanitarian response. We found that in almost all cases, multi-objective training improved the ensemble forecasting performance over the baseline ensembles trained using MSE. In addition, we developed a preference-based multi-objective training approach that combines multiple objectives using a weighted sum with a unique weighting for each base learner determined using a Das-Dennis grid. This method produced the most consistent improvement across multiple sites and variable combinations. When comparing the preferred multi-objective training approach to other approaches to improving ensemble forecasts (ensemble post-processing and cost sensitive learning), we found that the multi-objective approach outperformed these other approaches

across multiple sites and variable combinations. Furthermore, we also found that, as compared to ensembles trained with MSE or using cost-sensitive learning, the difference in performance between models trained with different variable combinations is smaller using multi-objective training, which is critical for use in sites where there is limited capacity to collect additional water quality variables. These findings present an important step in improving probabilistic forecasts of point-of-consumption FRC which will provide a critical benefit for the Safe Water Optimization Tool project and will in turn allow water system operators to have a greater degree of confidence when selecting point-of-distribution FRC targets. This also highlights the potential benefits of incorporating multi-objective training into other data-driven modelling applications where typical error metrics do not adequately reflect the intended outcomes of the models, especially models of complex phenomena.

5.7 References

Abbass, H. A. (2003). Pareto Neuro-Evolution: Constructing Ensemble. *Congress on Evolutionary Computation, 3*, 2074–2080.

Albuquerque Teixeira, R. de, Braga, A. P., Takahashi, R. H. C., & Saldanha, R. R. (2000). *Improving generalization of MLPs with multi-objective optimization. Neurocomputing, 35*(1–4), 189–194. https://doi.org/10.1016/S0925-2312(00)00327-1

Ali, S. I., Ali, S. S., & Fesselet, J.-F. (2015). Effectiveness of emergency water treatment practices in refugee camps in South Sudan. *Bulletin of the World Health Organization, 93*(8), 550–558. https://doi.org/10.2471/BLT.14.147645

Ali, S. I., Ali, S. S., & Fesselet, J. (2021). Evidence-based chlorination targets for household water safety in humanitarian settings: Recommendations from a multi-site study in refugee camps in South Sudan, Jordan, and Rwanda. *Water Research, 189*(116642), 1–17. https://doi.org/https://doi.org/10.1016/j.watres.2020.116642

Alvisi, S., & Franchini, M. (2011). Fuzzy neural networks for water level and discharge forecasting with uncertainty. *Environmental Modelling and Software, 26*(4), 523–537. https://doi.org/10.1016/j.envsoft.2010.10.016

Alvisi, S., & Franchini, M. (2012). Grey neural networks for river stage forecasting with uncertainty. *Physics and Chemistry of the Earth, 42–44*, 108–118. https://doi.org/10.1016/j.pce.2011.04.002

Boucher, M. A., Perreault, L., & Anctil, F. (2009). Tools for the assessment of hydrological ensemble forecasts obtained by neural networks. *Journal of Hydroinformatics, 11*(3–4), 297–307. https://doi.org/10.2166/hydro.2009.037

Bröcker, J. (2012). Evaluating raw ensembles with the continuous ranked probability score. *Quarterly Journal of the Royal Meteorological Society, 138*(667), 1611–1617. https://doi.org/10.1002/qj.1891

Brown, G., Wyatt, J., Harris, R., & Yao, X. (2005). Diversity creation methods: A survey and categorisation. *Information Fusion, 6*(1), 5–20. https://doi.org/10.1016/j.inffus.2004.04.004

CDC. (2012). Chlorine Residual Testing. Retrieved from http://www.cdc.gov/safewater/chlorine-

residual-testing.html

Candille, G., & Talagrand, O. (2005). Evaluation of probabilistic prediction systems for a scalar variable. *Quarterly Journal of the Royal Meteorological Society*, *131*(609), 2131–2150. https://doi.org/10.1256/qj.04.71

Chatterjee, S., Sarkar, S., Dey, N., Sen, S., Goto, T., & Debnath, N. C. (2017). Water Quality Prediction: Multi Objective Genetic Algorithm couple Artificial Neural Network based approach. *In 2017 IEEE 15th International Conference on Industrial Informatics (INDIN) (pp. 963–968)*. Emden, Germany: IEEE. https://doi.org/10.1109/INDIN.2017.8104902

Chatterjee, S., Sarkar, S., Hore, S., Dey, N., Ashour, A. S., Shi, F., & Le, D. N. (2017). Structural failure classification for reinforced concrete buildings using trained neural network based multi-objective genetic algorithm. *Structural Engineering and Mechanics*, *63*(4), 429–438. https://doi.org/10.12989/sem.2017.63.4.429

Chen, H., & Yao, X. (2010). Multiobjective neural network ensembles based on regularized negative correlation learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(12), 1738–1751. https://doi.org/10.1109/TKDE.2010.26

Crone, S. F., Lessmann, S., & Stahlbock, R. (2005). Utility based data mining for time series analysis - Cost-sensitive learning for neural network predictors. *In Proceedings of the 1st International Workshop on Utility-Based Data Mining, UBDM '05*, 59–68. https://doi.org/10.1145/1089827.1089835

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*, *4*(35), 1–16. https://doi.org/10.1038/s41545-021-00125-2

de Vos, N. J., & Rientjes, T. H. M. (2008). Multiobjective training of artificial neural networks for rainfall-runoff modeling. *Water Resources Research*, *44*(8), 1–15. https://doi.org/10.1029/2007WR006734

Deb, K., & Jain, H. (2014). An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, Part I: Solving problems with box constraints. *IEEE Transactions on Evolutionary Computation*, *18*(4), 577–601.

https://doi.org/10.1109/TEVC.2013.2281535

Ferro, C. A. T. (2014). Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society, 140*(683), 1917–1923. https://doi.org/10.1002/qj.2270

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. *Mathematical and Computer Modelling, 44*(5–6), 485–498. https://doi.org/10.1016/j.mcm.2006.01.007

Girones, R., Carratalà, A., Calgua, B., Calvo, M., Rodriguez-Manzano, J., & Emerson, S. (2014). Chlorine inactivation of hepatitis e virus and human adenovirus 2 in water. *Journal of Water and Health, 12*(3), 436–442. https://doi.org/10.2166/wh.2014.027

Golicha, Q.,Shetty, S.,Nasiblov, O., Hussein, A., Wainaina, E., Obonyo, M., Macharia, D., Musyoka, R. N., Abdille, H., Ope, M., Joseph, R., Kabugi, W., Kiogora, J., Said, M., Boru, W., Galgalo, T., Lowther, S. A., Juma, B., Mugoh, R.,...Burton, J.W.. (2018). Cholera outbreak in Dadaab Refugee camp, Kenya — November 2015–June 2016. *Morbidity and Mortality Weekly Report, 67*(34), 958–961. https://doi.org/10.15585/mmwr.mm6734a4

Guerrero-Latorre, L., Hundesa, A., & Girones, R. (2016). Transmission Sources of Waterborne Viruses in South Sudan Refugee Camps. *Clean - Soil, Air, Water, 44*(7), 775–780. https://doi.org/10.1002/clen.201500358

Hamill, T. M. (2001). Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Monthly Weather Review, 129*(3), 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 and Forecasting, 15*(5), 559–570.

Howard, C. M., Handzel, T., Hill, V. R., Grytdal, S. P., Blanton, C., Kamili, S., Drobeniuc, J., Hu, D., & Teshale, E. (2010). Novel Risk Factors Associated with Hepatitis E Virus Infection in a Large Outbreak in Northern Uganda: Results from a Case-Control Study and Environmental Analysis. *American Journal of Tropical Medicine and Hygiene, 83*(5), 1170–1173. https://doi.org/10.4269/ajtmh.2010.10-0384

Khan, U. T., & Valeo, C. (2016). Dissolved oxygen prediction using a possibility theory based fuzzy neural network. *Hydrology and Earth System Sciences, 20*, 2267–2293. https://doi.org/10.5194/hess-20-2267-2016

Khan, U. T., & Valeo, C. (2017). Comparing a Bayesian and fuzzy number approach to uncertainty quantification in short-term dissolved oxygen prediction. *Journal of Environmental Informatics, 30*(1), 1–16. https://doi.org/10.3808/jei.201700371

Lantagne, D. S. (2008). Sodium hypochlorite dosage for household and emergency water treatment. *Journal of American Water Works Association, 100*(8), 106–114. https://doi.org//10.1002/j.1551-8833.2008.tb09704.x

Rashid, M.-u., George, C. M., Monira, S., Mahmud, T., Rahman, Z., Mustafiz, M., Parvin, T., Bhuyian, S. I., Zohura, F., Begum, F., Biswas, S. K., Akhter, S., Zhang, X., Sack, D., Sack, R. B., & Alam, M. (2016). Chlorination of Household Drinking Water among Cholera Patients ' Households to Prevent Transmission of Toxigenic Vibrio cholerae in Dhaka , Bangladesh : CHoBI7 Trial. *American Journal of Tropical Medicine and Hygiene, 95*(6), 1299–1304. https://doi.org/10.4269/ajtmh.16-0420

Rodriguez, M. J., & Sérodes, J. B. (1998). Assessing empirical linear and non-linear modelling of residual chlorine in urban drinking water systems. *Environmental Modelling and Software, 14*(1), 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, Series A: Dynamic Meteorology and Oceanography, 55*(1), 16–30. https://doi.org/10.1034/j.1600-0870.2003.201378.x

Shultz, A., Omollo, J. O., Burke, H., Qassim, M., Ochieng, J. B., Weinberg, M., Feikin, D. R., & Breiman, R. F. (2009). Cholera outbreak in Kenyan Refugee Camp: Risk Factors for Illness and Importance of Sanitation. *American Journal of Tropical Medicine and Hygiene, 80*(4), 640–645. https://doi.org/10.4269/ajtmh.2009.80.640

Sikder, M., String, G., Kamal, Y., Farrington, M., Rahman, A. S., & Lantagne, D. (2020). Effectiveness of water chlorination programs along the emergency-transition-post-emergency continuum: Evaluations of bucket, in-line, and piped water chlorination

programs in Cox's Bazar. *Water Research, 178,* 115854. https://doi.org/10.1016/j.watres.2020.115854

Steele, A., Clarke, B., & Watkins, O. (2008). Impact of jerry can disinfection in a camp environment - Experiences in an IDP camp in Northern Uganda. *Journal of Water and Health, 6*(4), 559–564. https://doi.org/10.2166/wh.2008.072

Swerdlow, D.L. Malenga, G., Begkoyian, G., Nyangulu, D., Toole, M., Waldman, R. J., Puhr, D. N. D., & Tauxe, R. V. (1997). Epidemic cholera among refugees in Malawi, Africa: treatment and transmission. *Epidemiology and Infection, 118*(3), 207–214. https://doi.org/https://doi.org/10.1017/S0950268896007352

Taormina, R., & Chau, K. W. (2015). Neural network river forecasting with multi-objective fully informed particle swarm optimization. *Journal of Hydroinformatics, 17*(1), 99–113. https://doi.org/10.2166/hydro.2014.116

Toth, E. (2016). Estimation of flood warning runoff thresholds in ungauged basins with asymmetric error functions. *Hydrology and Earth System Sciences, 20*(6), 2383–2394. https://doi.org/10.5194/hess-20-2383-2016

Walden, V. M., Lamond, E. A., & Field, S. A. (2005). Container contamination as a possible source of a diarrhoea outbreak in Abou Shouk camp, Darfur province, Sudan. *Disasters, 29*(3), 213–221. https://doi.org/10.1111/j.0361-3666.2005.00287.x

Wang, X., & Bishop, C. H. (2005). Improvement of ensemble reliability with a new dressing kernel. *Quarterly Journal of the Royal Meteorological Society, 131*(607), 965–986. https://doi.org/10.1256/qj.04.120

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

Chapter 6 Conclusion

Waterborne illnesses remain a leading health risk to displaced populations living in refugee and IDP settlements. The spread of waterborne pathogens through household recontamination contributes to outbreaks of these illnesses, despite being preventable through the provision of adequate residual chlorine which protects against recontamination, as current FRC targets do not account for post-distribution FRC decay. However, generating new FRC targets that account for post-distribution FRC decay can be challenging due to the high degree of uncertainty underlying the FRC decay processes after water leaves the piped distribution system. To account for the uncertainty in post-distribution FRC decay, this thesis sought to develop probabilistic, data-driven models of post-distribution FRC decay with the objectives of evaluating the base performance of these models, developing an approach to generating FRC targets from these models, and then investigating approaches to improve these models.

6.1 Thesis Summary

The first two objectives of this thesis are addressed in Chapter 3 which presented a proof-of-concept investigation. Chapter 3 showed that ANN ensembles can develop very accurate risk-based FRC targets using only routine water quality monitoring data. However, this study also identified the primary challenge of this approach: forecast underdispersion. The study in Chapter 3 presented a simple attempt to address this challenge through ensemble post-processing, however, the models remained underdispersed. This motivated the studies included in Chapters 4 and 5 which investigated new cost functions for training the base learner ANNs. The cost function became the primary focus of these investigations because the cost function determines what model behaviours are rewarded during training and which are penalized. We initially considered this because we observed that ensembles with base learners trained using MSE produced regression-to-the-mean behaviour (MSE is functionally identical to the cost function minimized by ordinary least squares regression). This was further reinforced through reading numerous studies that identified MSE and common error metrics in general as poor cost functions for real-world approaches as they tend to prioritize the wrong elements of the output space. A challenge in applying these techniques is that they have not been widely applied to ANN ensembles used for probabilistic forecasting (as opposed to regression or classification).

Chapter 4 investigated the use of cost-sensitive learning and alternative cost functions for training the base learners of the ensemble. This research showed that the best performing alternatives used cost functions which directly prioritized matching the distributions of the underlying data combined with weightings that rebalance the cost function to equally prioritize the entire output space. Of the cost functions and weightings tested in Chapter 4, KGE with inverse frequency weighting produced the best performance. Both KGE as a cost function and inverse frequency weighting tend to prioritize an equal prioritization of the entire output space, so it is unsurprising that combining KGE and inverse frequency weighting produced the best ensemble forecast dispersion and reliability. However, it does reinforce the importance of selecting a cost function and weighting combination that aligns with the intended model behaviour.

Chapter 5 investigated the use of multi-objective training where we considered the similarity between aspects of the predicted and observed distributions (mean, variance, correlation) as well as classification parameters (recall, precision) as their own objectives. Based on the findings of Chapter 4, the objectives used to measure the similarity between the predicted and observed distributions were taken from the KGE cost function formula. Chapter 5 showed that the multi-objective training approach was highly effective overall and that the most effective multi-objective training approach used a Das-Dennis grid to weight each objective for each base learner, so that each base learner was trained using the same objectives, but with a different weighting assigned to each objective. Chapter 5 concluded with a comparison of the different approaches to overcoming model dispersion considered in this study (post-processing, cost-sensitive learning, multi-objective training) which showed that across all sites and variable combinations, both development and implementation, the multi-objective training approach produced the largest magnitude of improvement over the reference baseline, as well as being most effective at improving multiple probabilistic performance metrics. Based on this, we recommend the multi-objective approach for adoption into future versions of the Safe Water Optimization Tool. The improved dispersion and reliability of the forecasts produced using this method should provide much more accurate risk-based FRC targets, allowing water system operators to make better informed, evidence-based chlorination decisions. An additional benefit of the multi-objective training approach as compared to the other approaches presented in this thesis was the ability to minimize the difference in performance between models using different

input variable combinations. For all of the other approaches, there was a clear decrease in performance when using a smaller input variable set as compared to a larger input variable set, but with the multi-objective training, the performance difference was much smaller. This is of crucial importance as it reduces the data collection burden on water system operators and ensures good model performance even at sites where collecting additional water quality variables is infeasible.

As a whole, the research in this thesis presents an important first step into the use of machine learning and data-driven modelling for forecasting point-of-consumption FRC. These models address a crucial need by developing improved FRC targets for humanitarian response that ensure water remains safe to drink from the time it is collected to the time it is consumed by protecting against the spread of waterborne pathogens. This research developed a process for quantifying the uncertainty in post-distribution FRC decay using ensembles of ANNs in a way that avoids many of the pitfalls of deterministic modelling. Furthermore, this research developed a means of converting these forecasts into risk-based FRC targets that allow water treatment decisions to be made on the basis of risk, which communicates the uncertainty in post-distribution FRC decay in a way that deterministic FRC targets cannot. Finally, this study applied advanced techniques (post-processing, cost-sensitive learning, multi-objective training) from other fields to improve the forecasting performance of these ANN ensembles, overcoming the underdispersion that occurs when using standard modelling techniques. Many of these improvements are had not previously been applied in the context of data-driven, probabilistic modelling and were adapted to suit the modelling challenge presented in forecasting point-of-consumption FRC. These modelling techniques greatly improved the reliability and dispersion of the ensemble models, and this improved performance translates into more accurate FRC targets for water system operators in humanitarian response, providing operators with a greater ability to protect water against household recontamination. Furthermore, all of these models used only water quality data that is routinely collected in refugee and IDP settlements, meaning that the modelling approaches developed in this thesis can be applied by water system operators around the world without needing additional data collection tools or equipment.

The modelling techniques and approaches developed in this thesis are also being made available through the Safe Water Optimization Tool (SWOT) (https://www.safeh2o.app/) to ensure that

they are available to water system operators in a free, open-source tool. The modelling approach presented in Chapter 3 has already been incorporated into the SWOT-ANN version 2 analytics, and the Chapter 5 findings will be incorporated into the SWOT-ANN version 3. This ensures that the modelling techniques developed here will be easily accessible to help water system operators provide safe drinking water in refugee and IDP settlements around the world.

6.2 Opportunities for Future Research

Based on the findings presented in Chapter 5, an immediate next step should be identifying approaches to operationalize the multi-objective training approach to investigate approaches for generating FRC targets using these models. In particular, the scenario analysis approach presented in Chapter 3 was effective for overcoming an underdispersed model, however, since the models developed using the Das-Dennis grid for objective combining are not underdispersed, a new approach for generating FRC targets may be needed.

A second important area for future study is the challenge of input variable selection. The input variable sets used in this study were very small, and as such we did not attempt to reduce these, however, investigating the usefulness of different water quality variables can help prioritize data collection to minimize the data collection burden for water system operators. A further approach for reducing the data collection burden would be to implement solutions that do not require paired data samples (samples from the same unit of water).

Finally, this research used relatively standard machine learning models and ensemble forming techniques for forecasting point-of-consumption FRC, with the prime focus on identifying cost function and training approaches that produced reliable ensemble forecasts. However, the field of machine learning and data-driven modelling is incredibly broad, and there are many new and emerging modelling approaches that may be beneficial, and as such, future research should investigate alternative ensemble formation approaches, machine learning models, and prediction-interval generation approaches. In particular, ensemble formation techniques such as bagging, boosting, and negative correlation training may all produce benefits over the ensemble techniques used in this study. Furthermore, there is emerging research into ensemble techniques that combine multiple types of machine learning models into a single ensemble, as well as investigations into combining process-based and machine learning models into hybrid models that leverage the advantages of both data-driven and process-based modelling approaches.

Appendix A. Supplemental Information for Chapter 3

Table A-1: Comparison of point-of-distribution FRC recommendations with different allowable risk of low FRC at the point-of-consumption. Three risk thresholds are included (negligible risk, 5% risk, 15% risk). From this table we see that the sensitivity of the FRC recommendation to the risk threshold varies from site to site based on the distribution of the underlying data.

Site	Negligible Risk			5% Risk			15% Risk		
	IV1 Target (mg/L)	IV2 Average Target (mg/L)	IV2 Worst Case Target (mg/L)	IV1 Target (mg/L)	IV2 Average Target (mg/L)	IV2 Worst Case Target (mg/L)	IV1 Target (mg/L)	IV2 Average Target (mg/L)	IV2 Worst Case Target (mg/L)
South Sudan	0.95	0.75	0.70	0.65	0.40	0.50	0.55	0.30	0.40
Jordan 2014	0.75	1.05	*1.75*[1]	0.60	0.70	1.50	0.40	0.40	1.25
Jordan 2015	0.20	0.20	0.40	0.20	0.20	0.20	0.20	0.20	0.20
Rwanda	0.65	0.65	0.95	0.55	0.60	0.85	0.55	0.55	0.75

Notes:

[1]No point-of distribution FRC concentration produced negligible risk for the Jordan (2014) worst-case scenario analysis. FRC target of 1.75 mg/L selected as predicted risk does not substantially improve for higher point-of-distribution FRC concentrations

Table A-2: Summary of number of data cleaning steps, rationale, and observations removed

Step	Rationale	Observations Removed			
		South Sudan	Jordan (2014)	Jordan (2015)	Rwanda
1 – Remove fields with personal identifiers	Protect anonymity of data collectors and refugee settlement water users	-	-	-	-
2 – Convert timestep measurements to elapsed time	Ensure data is useable for ANN models	-	-	-	-
3 – Remove records where difference FRC at point-of-distribution is greater than FRC at point-of-consumption by >0.06 mg/L	Maximum measurement error was 0.03 mg/L, so an increase in FRC from distribution to consumption of more than twice this error indicates one or both measurements was inaccurate	6	4	3	0
4 – Remove records with measurements outside of guideline values at point-of-distribution (FRC>2 mg/L, Turbidity >5 NTU, pH<6 or >8)	Prioritizes performance for typical operating range; also eliminates erroneous values	45	12	2	1
5 – Remove records with missing measurements	Required for ANN functionality. Note – more records removed for IV2 because more possible variables with missing measurements	IV1: 9 IV2: 37	IV1: 11 IV2: 17	IV1: 8 IV2: 31	IV1: 16 IV2: 41

Step	Rationale	Observations Removed			
		South Sudan	**Jordan (2014)**	**Jordan (2015)**	**Rwanda**
6 – Remove records where water was stored in the sun	Jordan only – removed for consistency with past studies as behaviour is not typical	-	66	15	-
Initial Number of Observations		220	199	114	134
Remaining Observations		IV1: 160 IV2: 132	IV1: 106 IV2: 100	IV1: 86 IV2: 63	IV1: 117 IV2: 92

Table A-3: Median, mean, and standard deviation for IV1 input variables. Generally similar characteristics of between calibration and testing datasets, except the mean storage duration in Jordan (2014) is 2 hours longer in the calibration dataset than the testing and in Rwanda the mean storage duration in the calibration dataset is nearly 3 hours shorter than in the testing

		Calibration		**Testing**	
Site		FRC	Elapsed Time	FRC	Elapsed Time
South Sudan	Median	0.83	6.78	0.91	6.13
	Mean	0.81	6.72	0.84	6.29
	Standard Deviation	0.46	1.18	0.42	0.98
Jordan (2014)	Median	0.91	6.33	0.98	6.32
	Mean	0.94	8.60	1.02	6.90
	Standard Deviation	0.24	6.38	0.32	3.62
Jordan (2015)	Median	0.72	18.75	0.71	19.21
	Mean	0.73	15.19	0.74	16.71
	Standard Deviation	0.11	8.78	0.09	8.74
Rwanda	Median	0.65	21.13	0.72	21.36
	Mean	0.64	15.99	0.69	18.28
	Standard Deviation	0.19	7.89	0.19	6.37

Table A-4: Median, mean, and standard deviation for IV2 input variables. Generally similar characteristics of between calibration and testing datasets, except see large differences between the median and mean EC in South Sudan between the calibration and training dataset but otherwise the water quality variable statistics were similar for both calibration and testing.

Site		Calibration						Testing					
		FRC	Elapsed Time	EC	Water Temperature	pH	Turbidity	FRC	Elapsed Time	EC	Water Temperature	pH	Turbidity
South Sudan	Median	0.79	6.77	847	31	7.33	1.56	0.84	6.53	751	31	7.21	1.25
	Mean	0.79	6.68	1020	31	7.21	1.81	0.87	6.56	871	32	7.16	1.44
	Standard Deviation	0.46	1.21	646	1.22	0.37	1.16	0.46	1.00	527	1.71	0.35	0.86
Jordan (2014)	Median	0.91	6.33	904	27	7.63	1.89	0.98	6.17	937	27	7.64	1.82
	Mean	0.94	8.64	900	27	7.59	1.88	0.97	7.22	975	27	7.51	1.80
	Standard Deviation	0.23	6.37	319	0.70	0.26	0.66	0.34	4.57	310	0.65	0.32	0.70
Jordan (2015)	Median	0.71	19	1127	20	7.54	0.75	0.73	7.58	989	20	7.47	0.71
	Mean	0.73	15	1147	20	7.51	0.79	0.75	14	1127	21	7.45	0.74
	Standard Deviation	0.11	8.79	314	2.06	0.25	0.57	0.12	8.18	282	2.55	0.21	0.34
Rwanda	Median	0.62	22	820	20	6.75	0.31	0.59	22	862	20	6.68	0.31
	Mean	0.61	19	965	20	6.75	0.40	0.64	19	866	19	6.66	0.33
	Standard Deviation	0.18	6.35	402	1.22	0.21	0.36	0.19	6.64	190	1.63	0.20	0.28

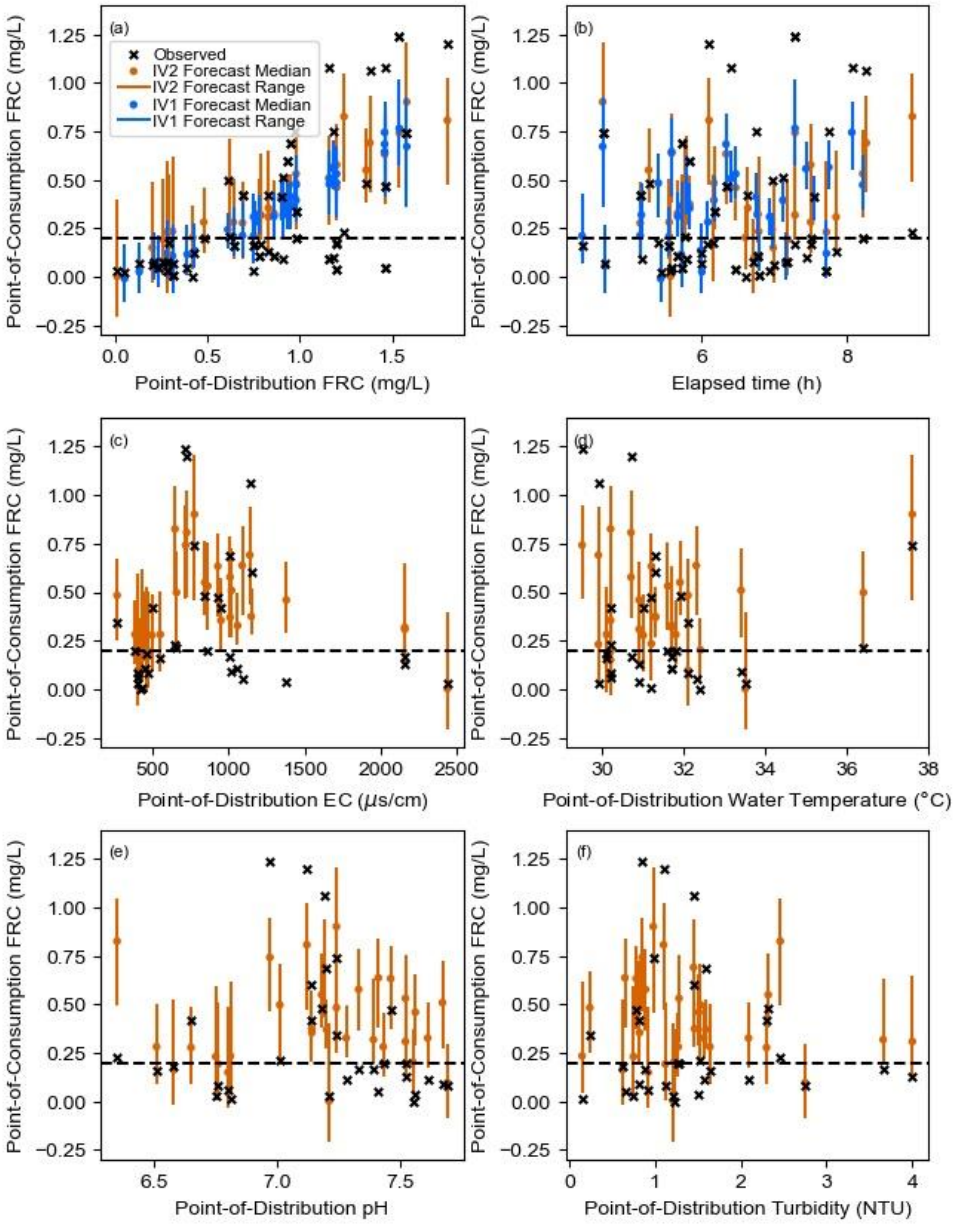

Figure A-1: South Sudan observations and raw ensemble forecasts of point-of-consumption FRC against (a) point-of-distribution FRC, (b) elapsed time, (c) point-of-distribution EC, (d) point-of-distribution water temperature, (e) point-of-distribution pH, (f) point-of-distribution turbidity. Similar trends between input variables and observations and forecasts as compared to post-processed ensembles, though raw ensemble forecast width is narrower, leading to fewer observations captured, especially observations where point-of-consumption FRC is below 0.2 mg/L

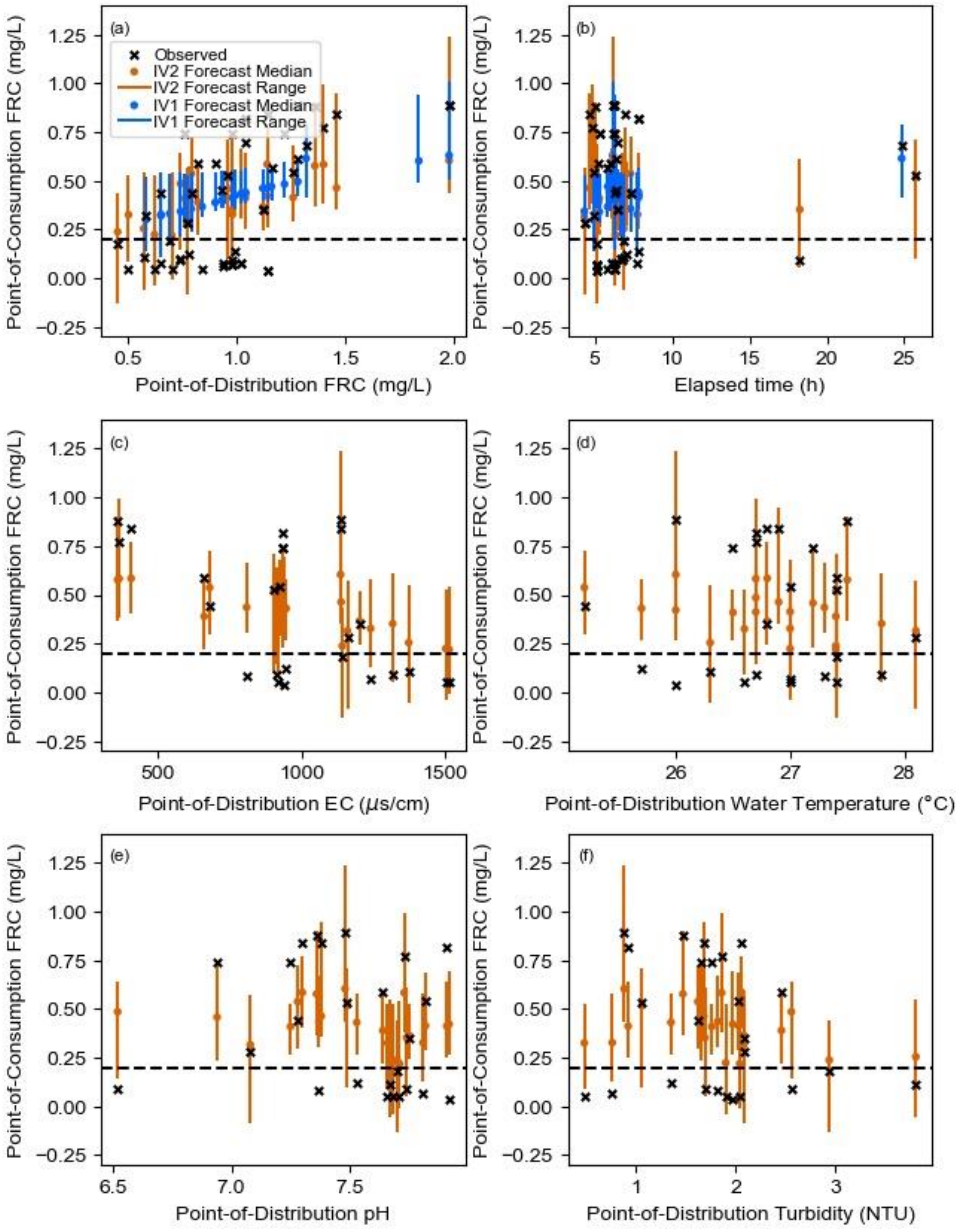

Figure A-2: Jordan (2014) observations and raw ensemble forecasts of point-of-consumption FRC against (a) point-of-distribution FRC, (b) elapsed time, (c) point-of-distribution EC, (d) point-of-distribution water temperature, (e) point-of-distribution pH, (f) point-of-distribution turbidity. Similar trends between input variables and observations and forecasts as compared to post-processed ensembles, though raw ensemble forecast width is narrower, leading to fewer observations captured, especially observations where point-of-consumption FRC is below 0.2 mg/L

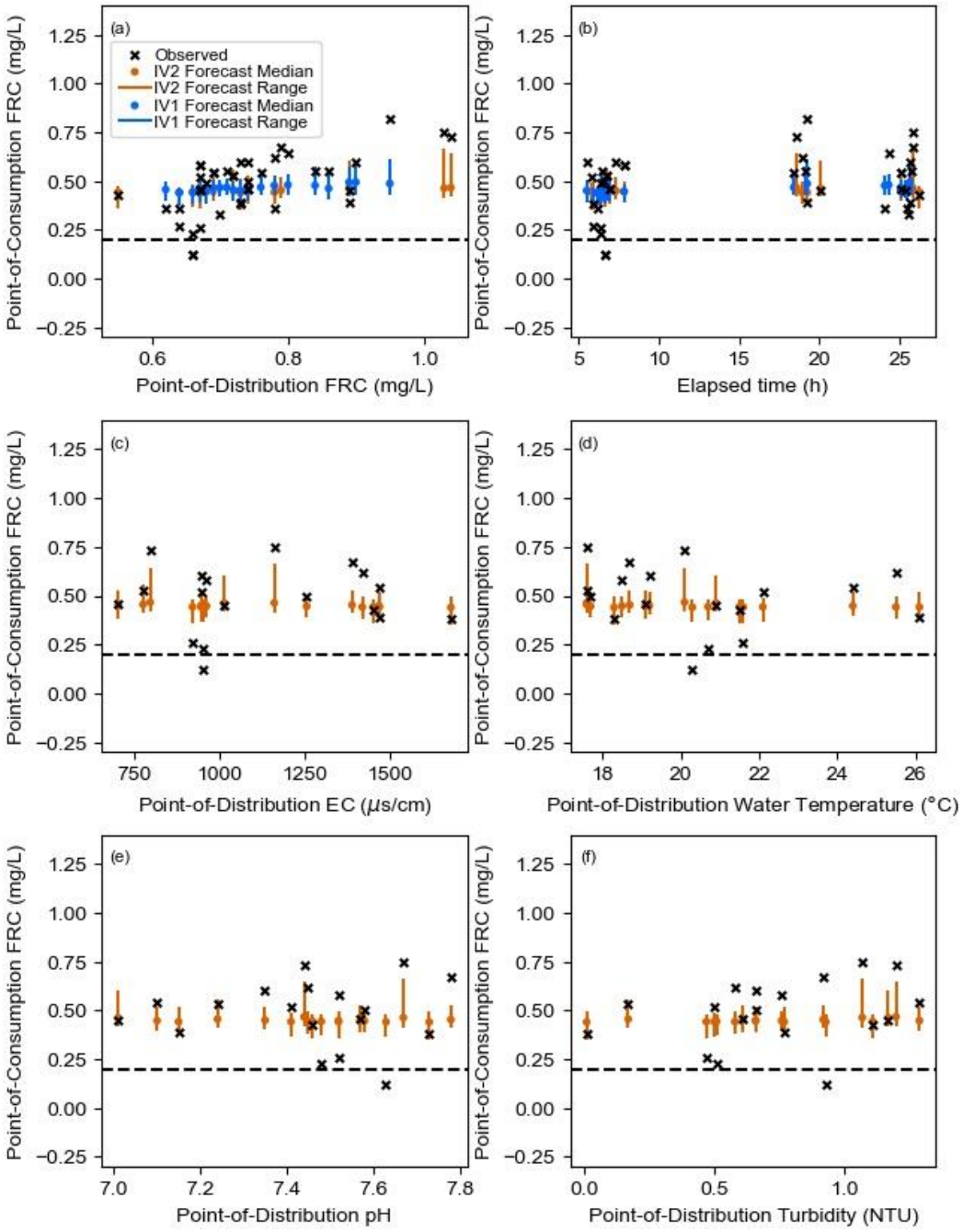

Figure A-3: Jordan (2015) observations and raw ensemble forecasts of point-of-consumption FRC against (a) point-of-distribution FRC, (b) elapsed time, (c) point-of-distribution EC, (d) point-of-distribution water temperature, (e) point-of-distribution pH, (f) point-of-distribution turbidity. IV1 and IV2 forecasts are both flat and narrow.

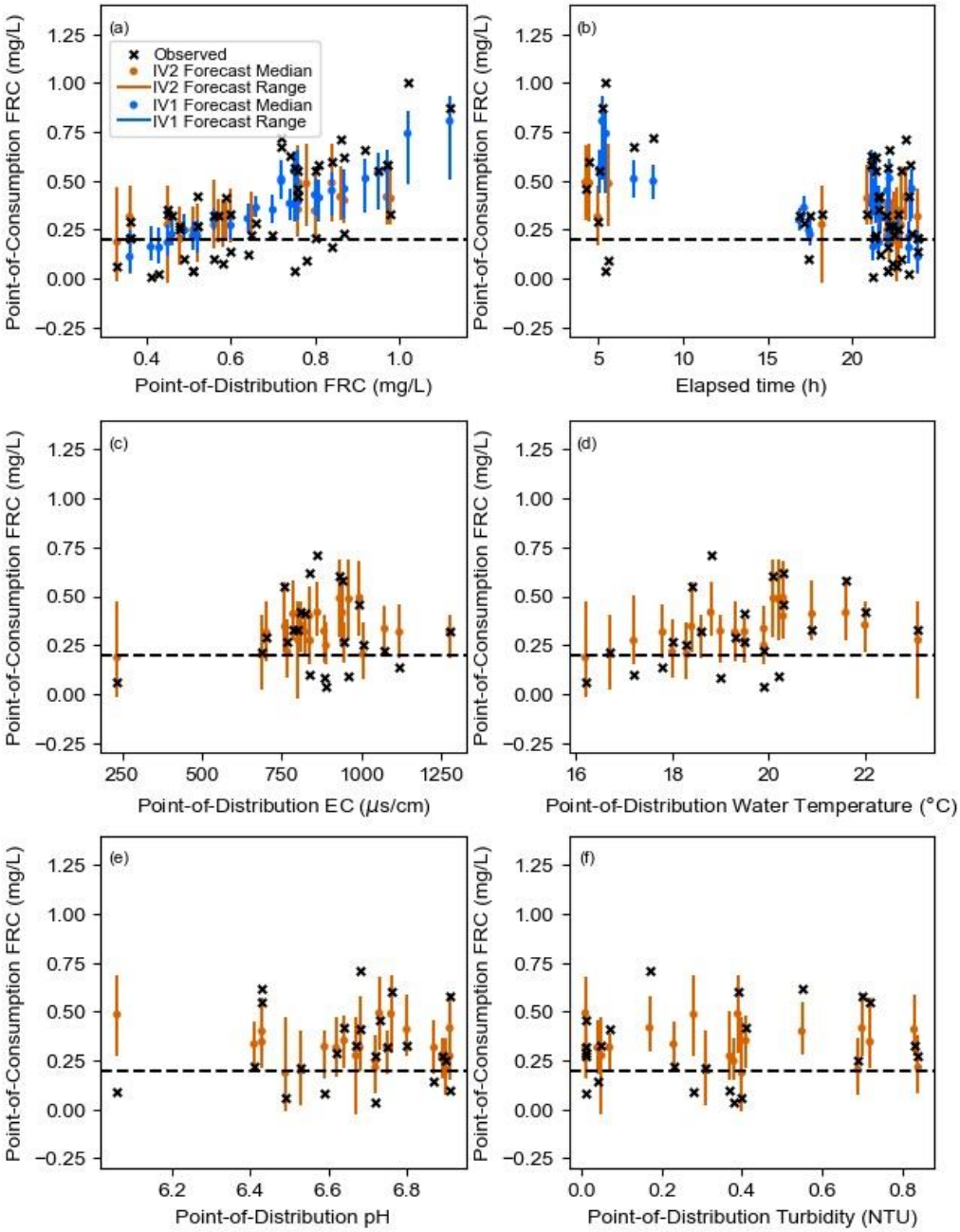

Figure A-4: Rwanda observations and raw ensemble forecasts of point-of-consumption FRC against (a) point-of-distribution FRC, (b) elapsed time, (c) point-of-distribution EC, (d) point-of-distribution water temperature, (e) point-of-distribution pH, (f) point-of-distribution turbidity. Forecast range is narrower than post-processed forecasts, but this does not impact percentage of observations captured within the forecast range.

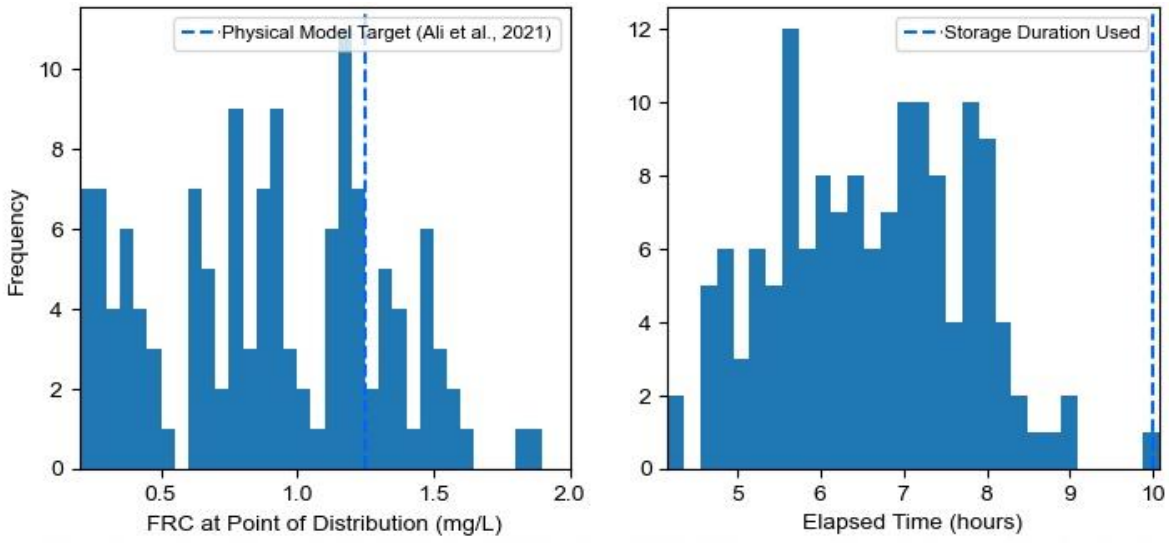

Figure A-5: South Sudan IV1 input variable histograms. Target storage duration is at the upper end of elapsed time histogram

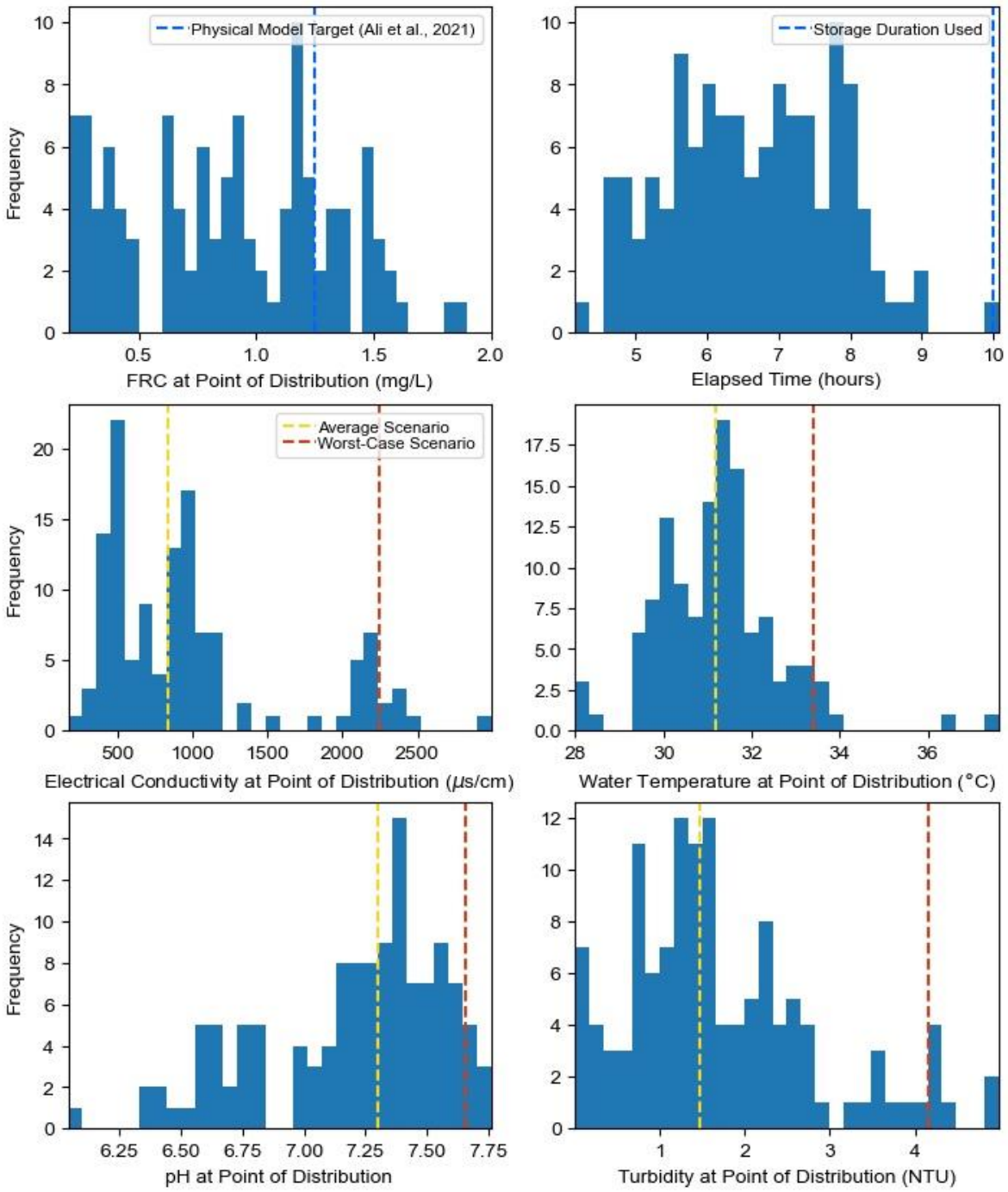

Figure A-6: South Sudan IV2 input variable histograms. Most input variables display multimodal distributions, but most notably EC and water temperature, likely due to differences in water sources.

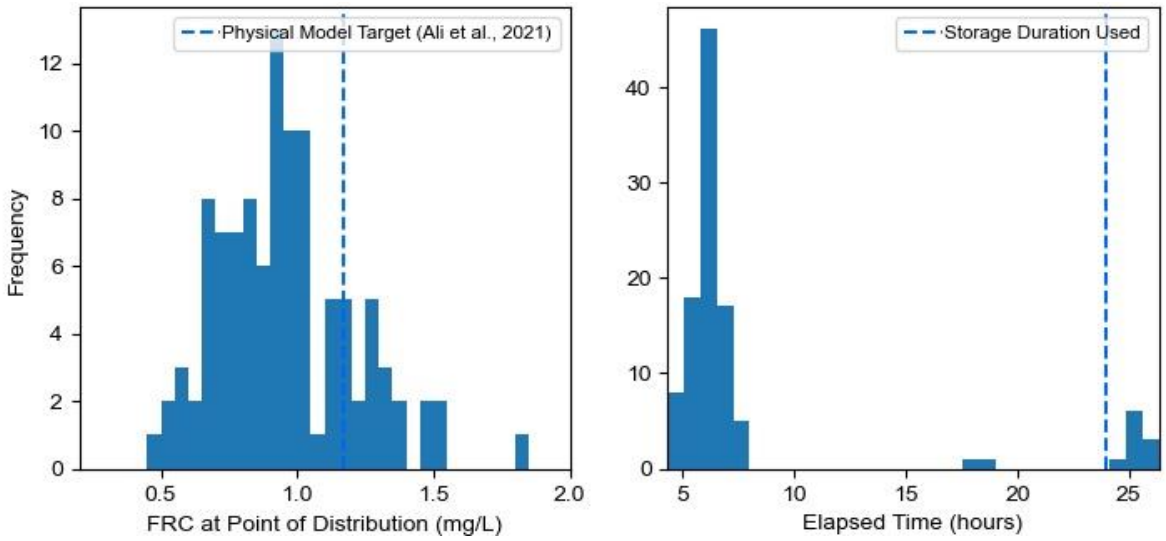

Figure A-7: Jordan (2014) IV1 input variable histograms. Observed storage appears to be predominantly under 10 hours, which is substantially lower than the target storage duration.

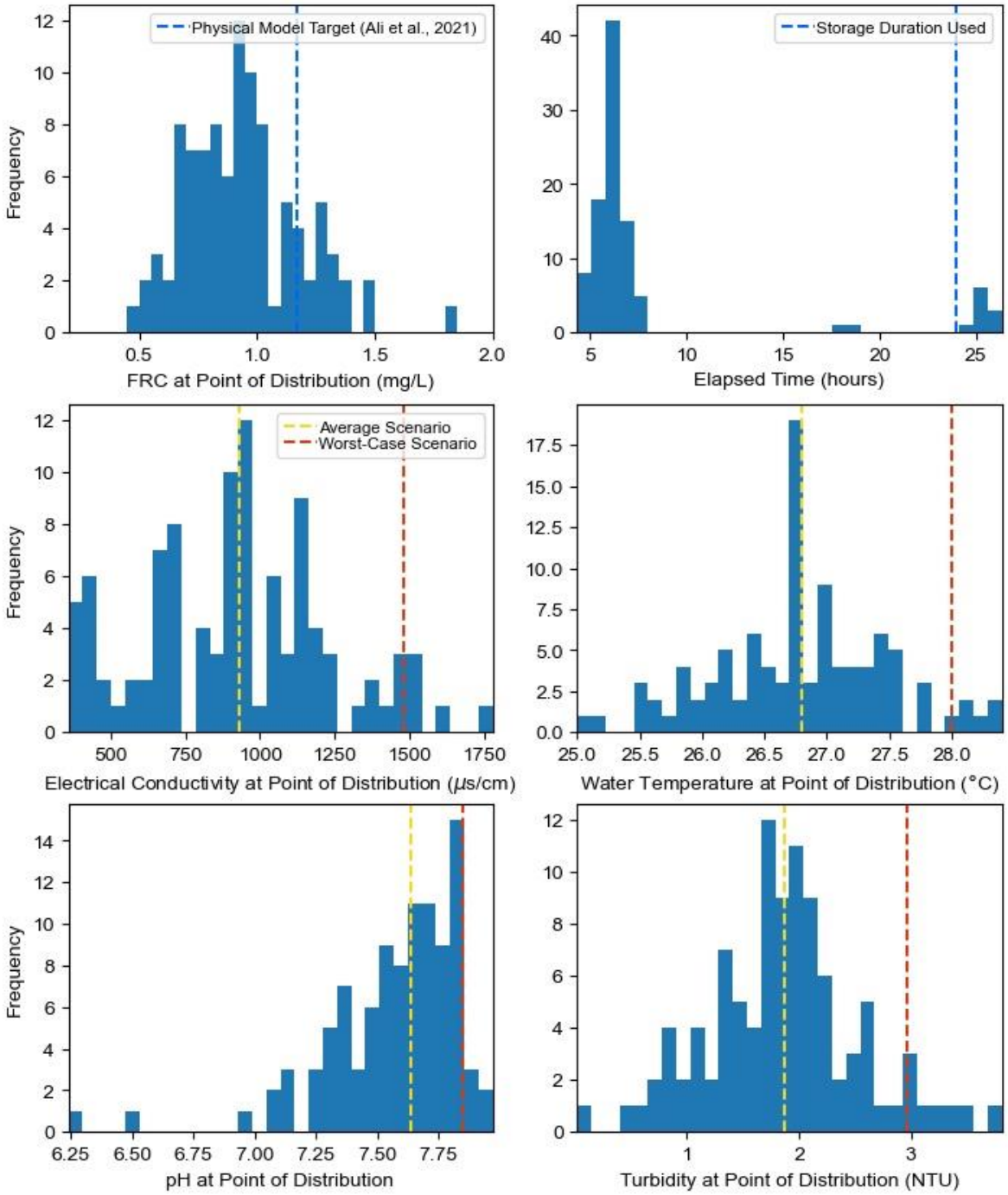

Figure A-8: Jordan (2014) IV2 input variable histograms.

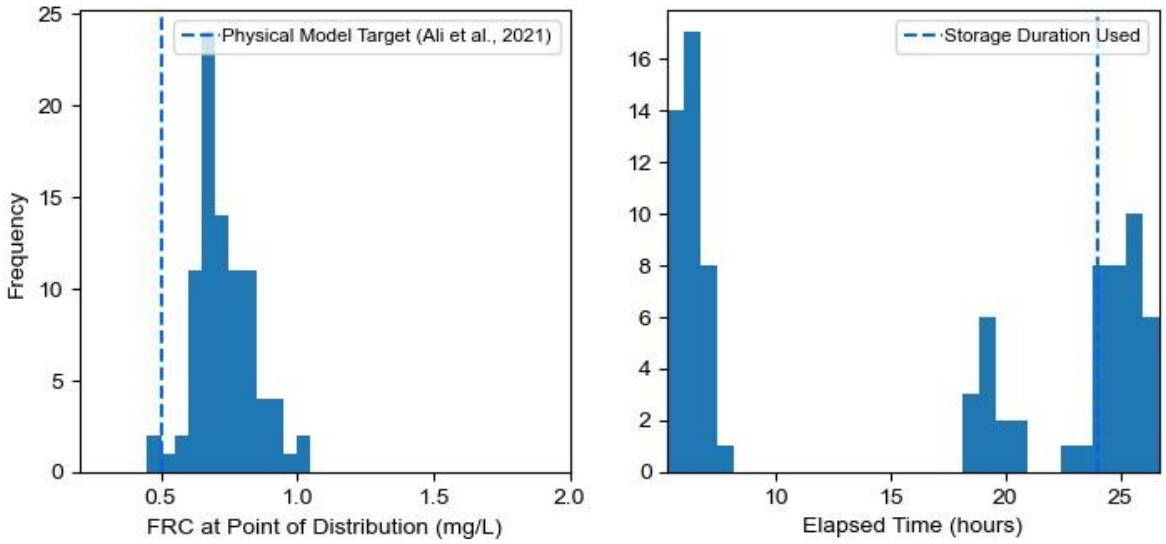

Figure A-9: Jordan (2015) IV1 input variable histograms. Storage durations are much longer than the Jordan (2014) dataset.

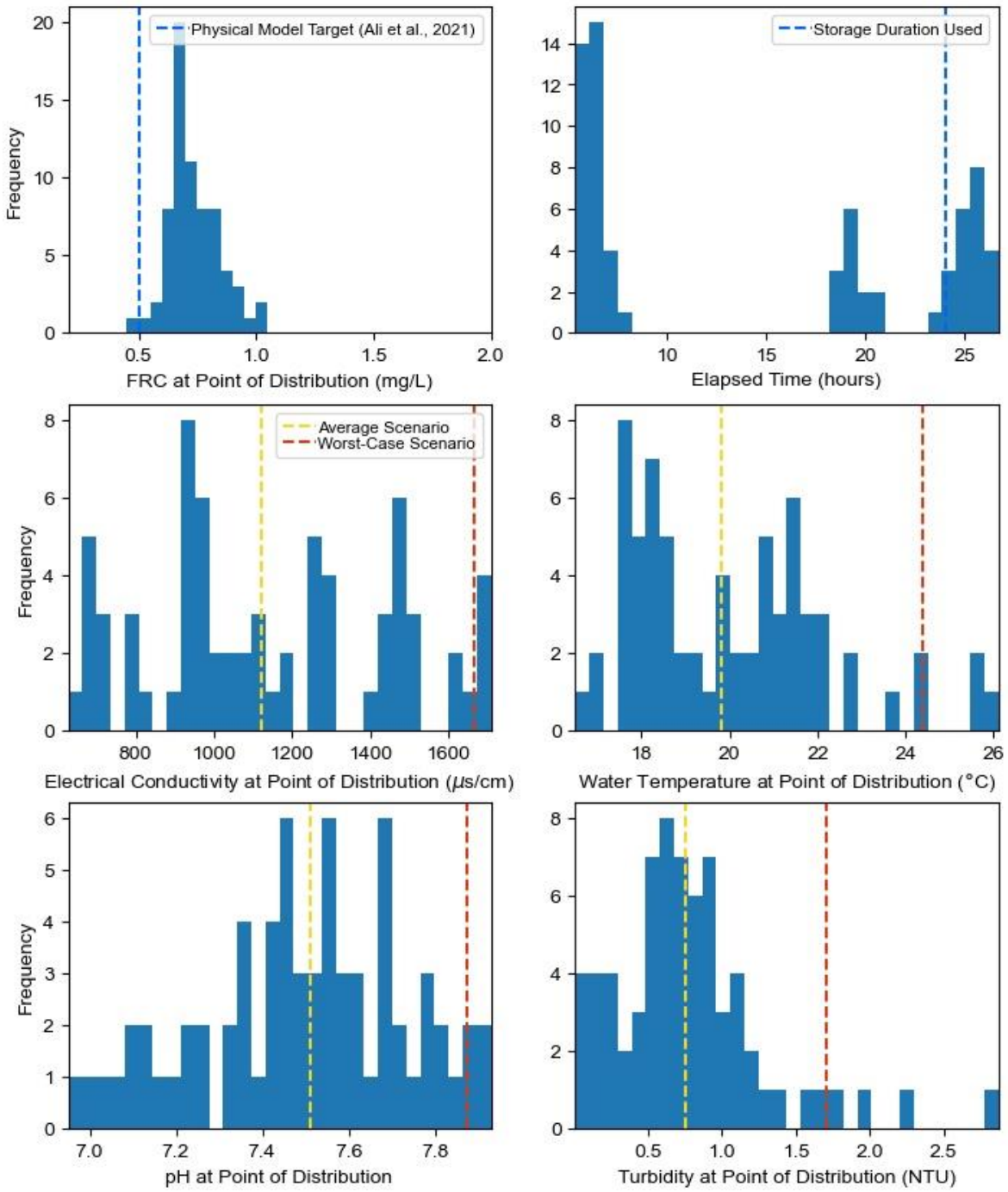

Figure A-10: Jordan (2015) IV2 input variable histograms

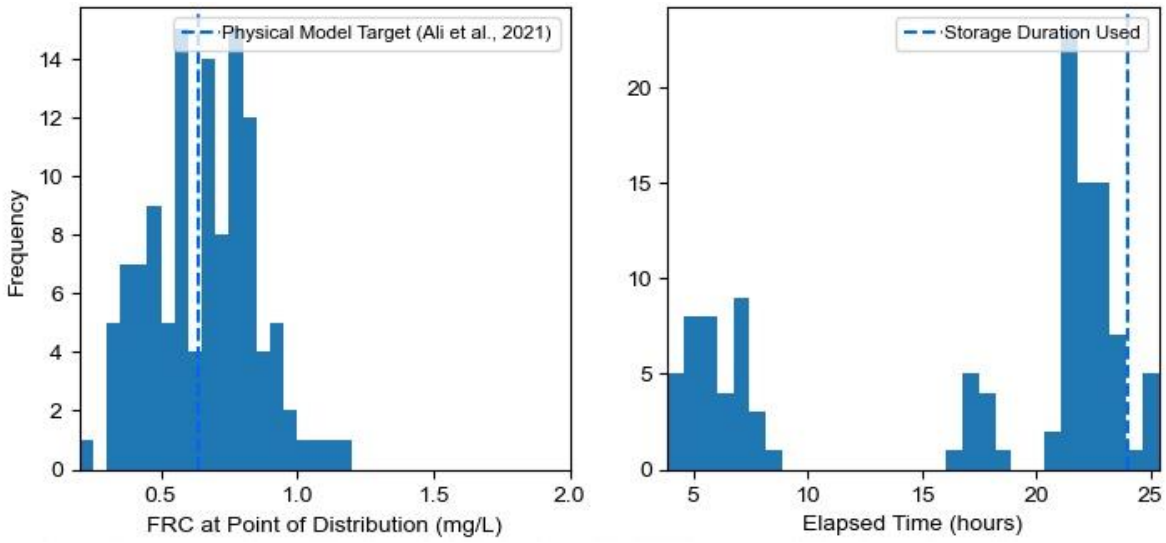

Figure A-11: Rwanda IV1 input variable histograms. Note clustering of storage durations into three distinct regions.

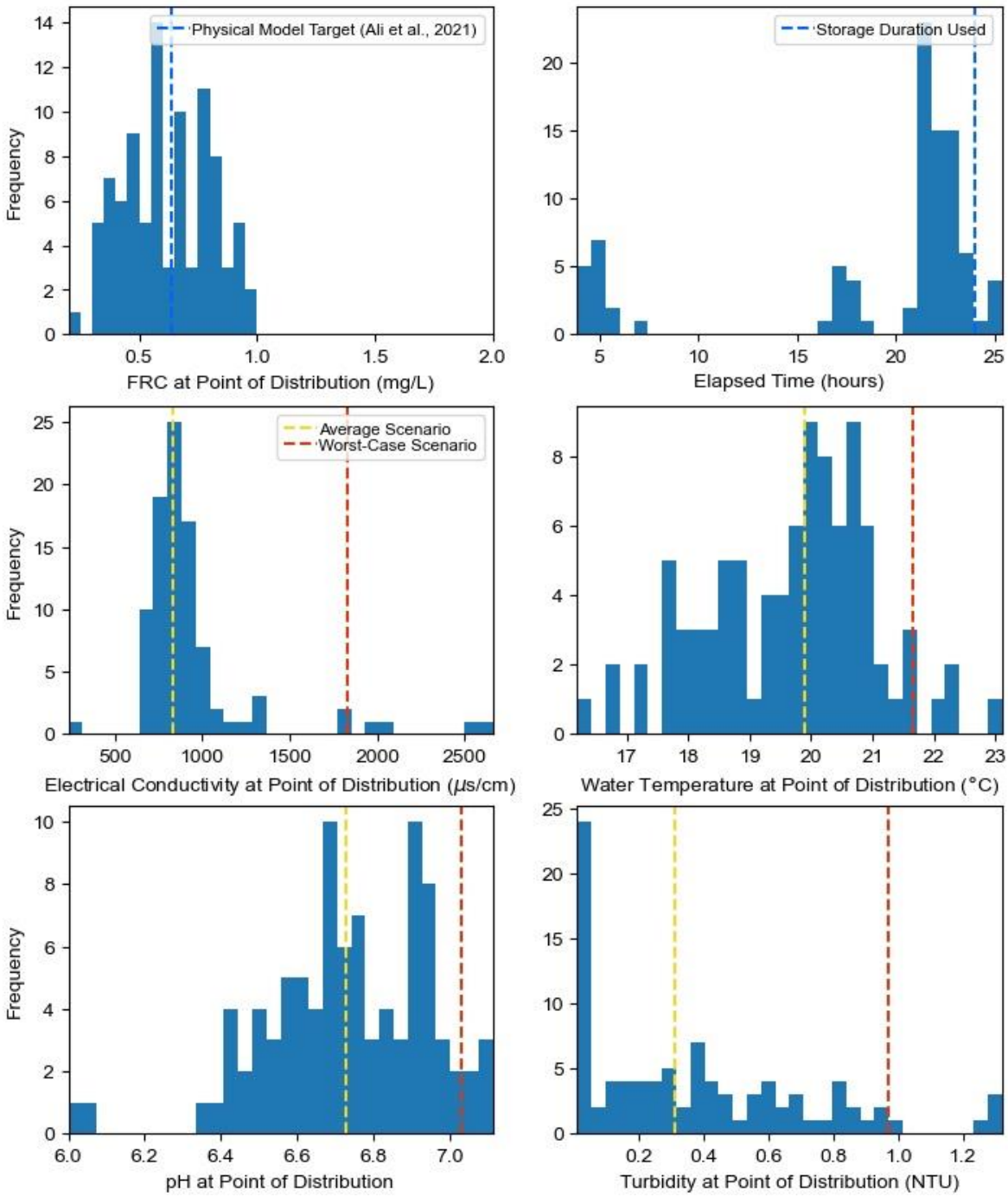

Figure A-12: Rwanda IV2 input variable histograms.

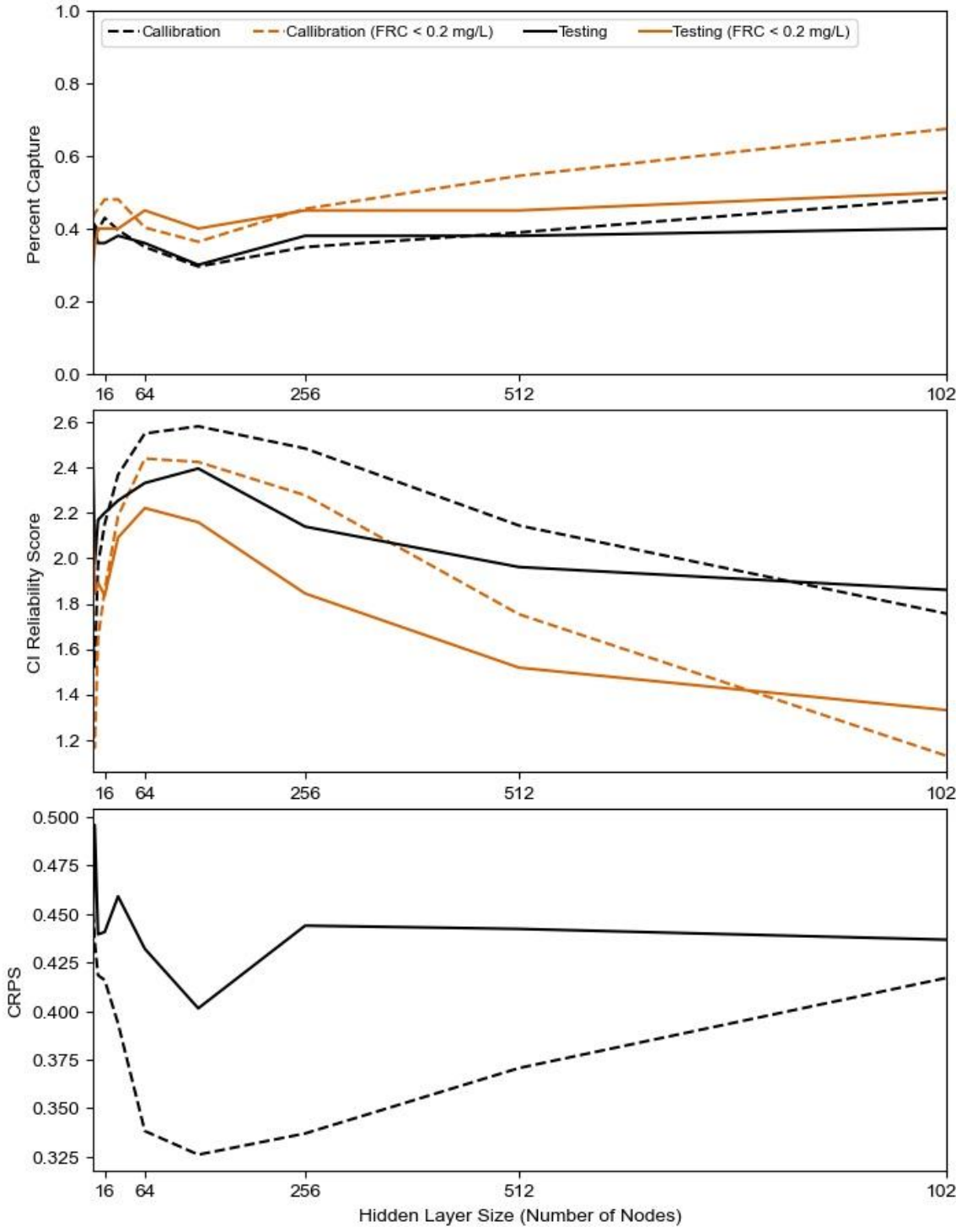

Figure A-13: South Sudan IV1 hidden layer size selection. Rapid improvement when hidden layer size increases from 2 to 4, improvement stops or slows after so we select a hidden layer size of 4 hidden nodes.

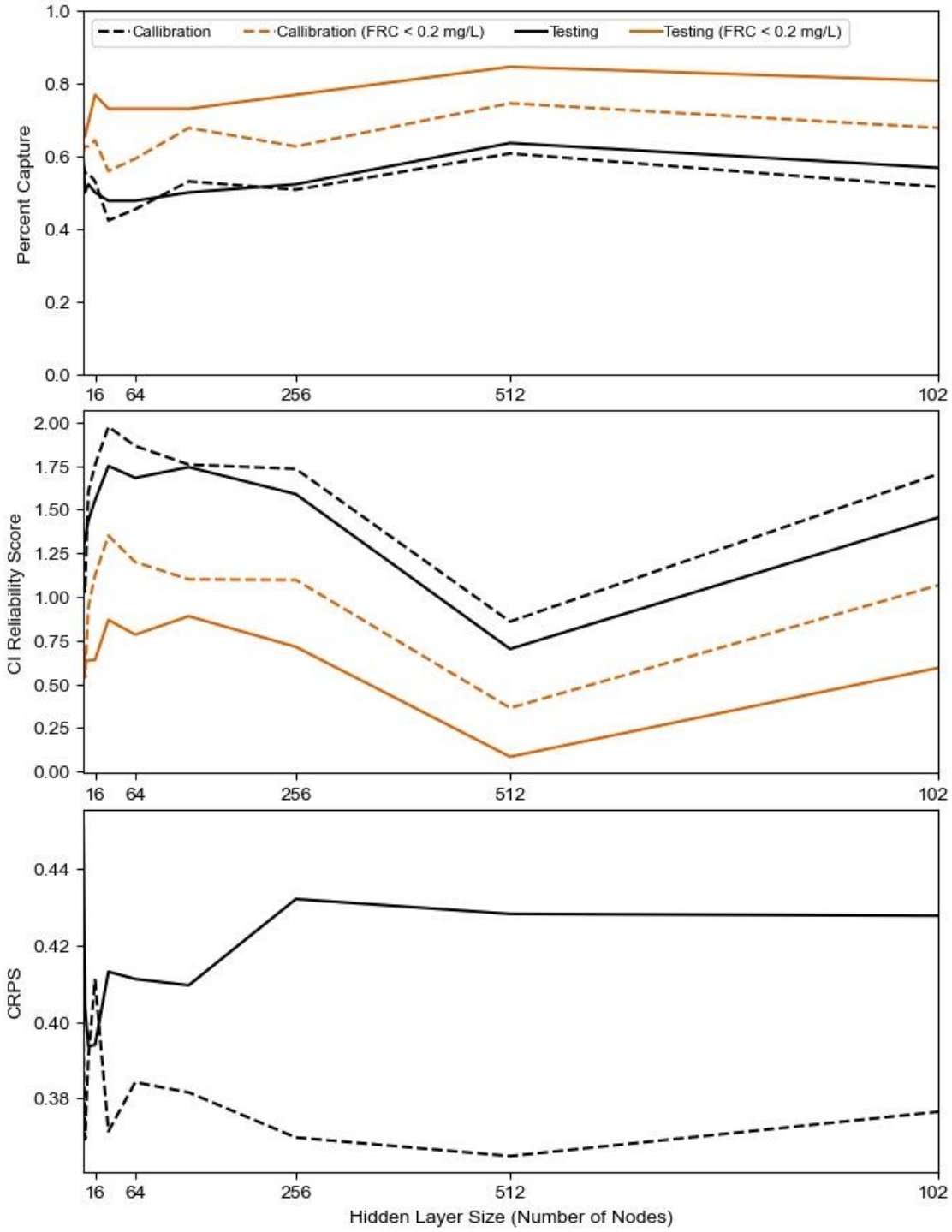

Figure A-14: South Sudan IV2 hidden layer size selection. Best Percent Capture at hidden layer size of 16 nodes, also good CI Reliability for unsafe values and low CRPS at hidden layer size of 16 hidden nodes so we select a hidden layer size of 16 hidden nodes.

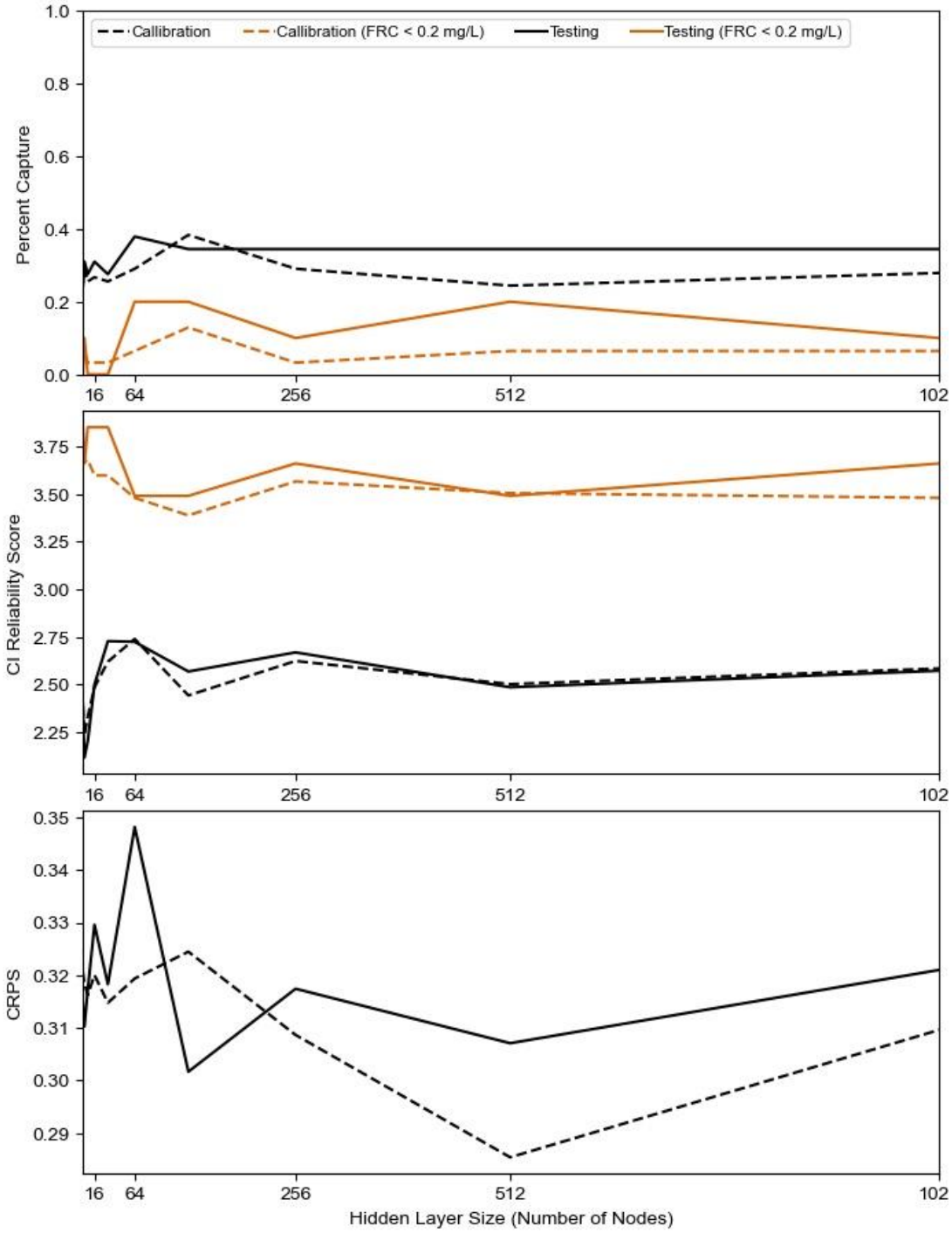

Figure A-15: Jordan (2014) IV1 hidden layer size selection. Rapid improvement when hidden layer size is increased from 2 to 4 hidden nodes. Further improvements at 64 hidden nodes, but not commensurate with increase in hidden layer size so we select a hidden layer size of 4 hidden nodes.

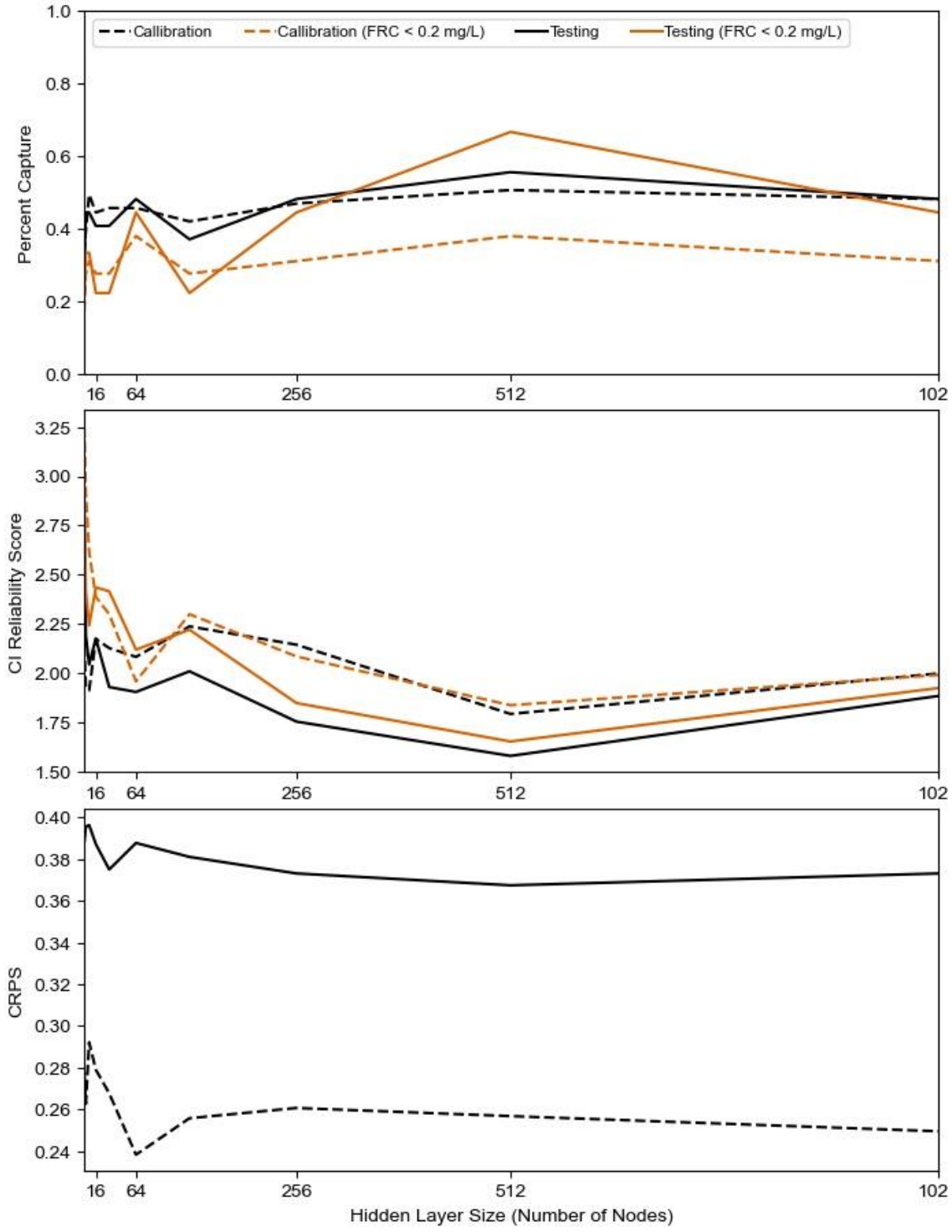

Figure A-16: Jordan (2014) IV2 hidden layer size selection. Substantial improvements in performance up to a hidden layer size of 8 nodes. Further improvements occur but not commensurate with the increase in layer size so we select a hidden layer size of 8 hidden nodes,

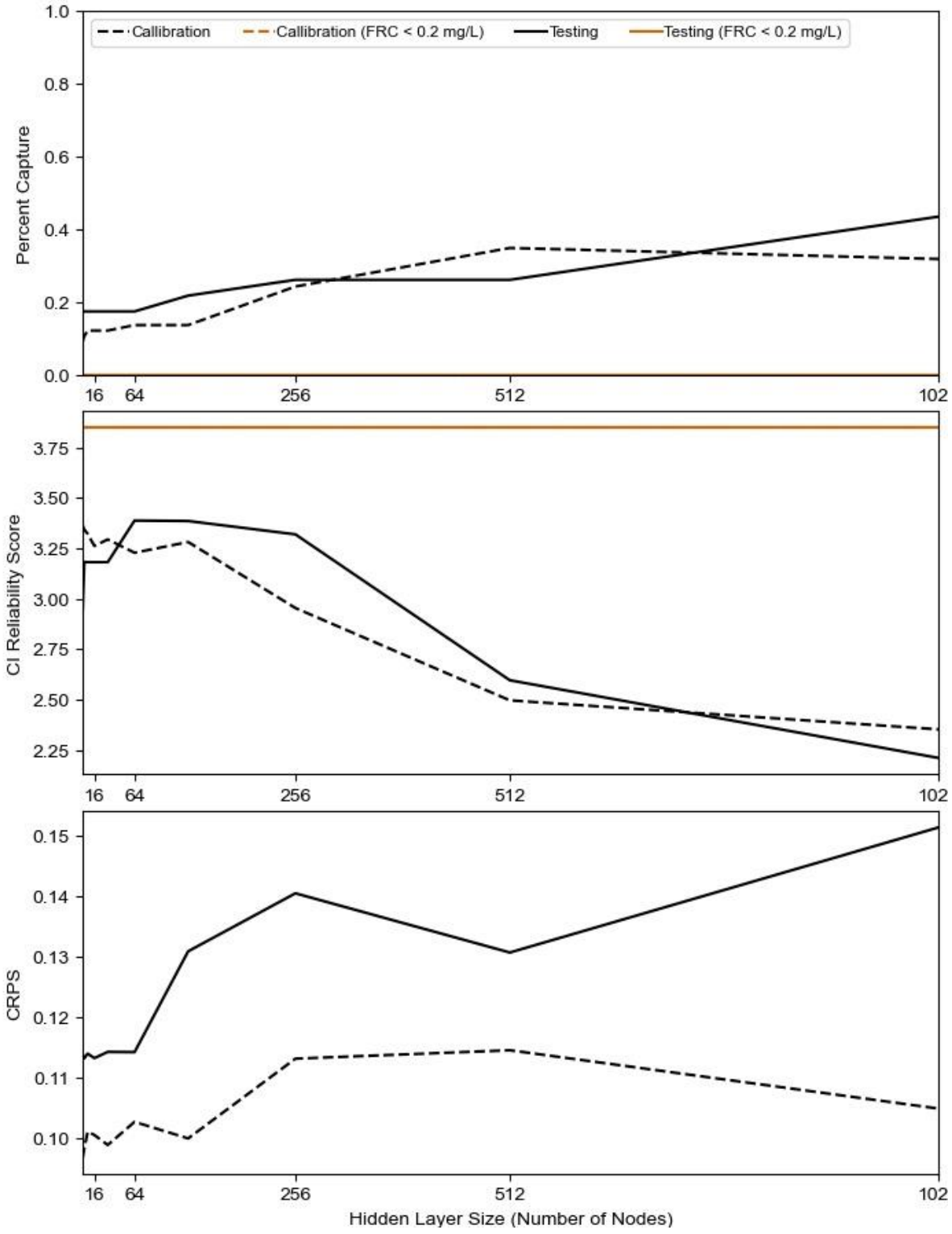

Figure A-17: Jordan (2015) IV1 hidden layer size selection. No clear best point so hidden layer size of 4 nodes selected based on experience of other sites.

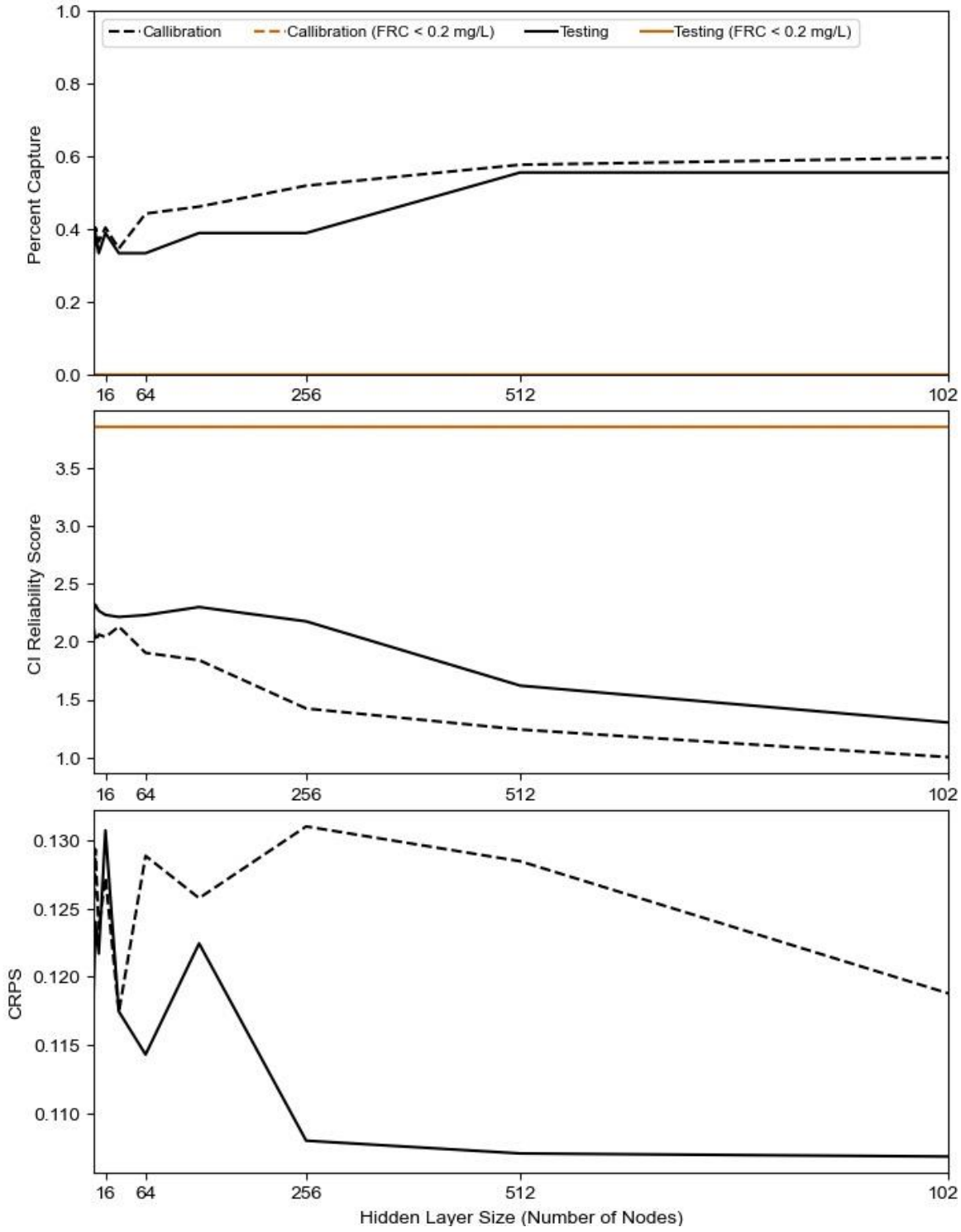

Figure A-18: Jordan (2015) IV2 hidden layer size selection. Hidden layer size of 16 simultaneously gives best performance across all metrics relative to hidden layer size so we select a hidden layer size of 16.

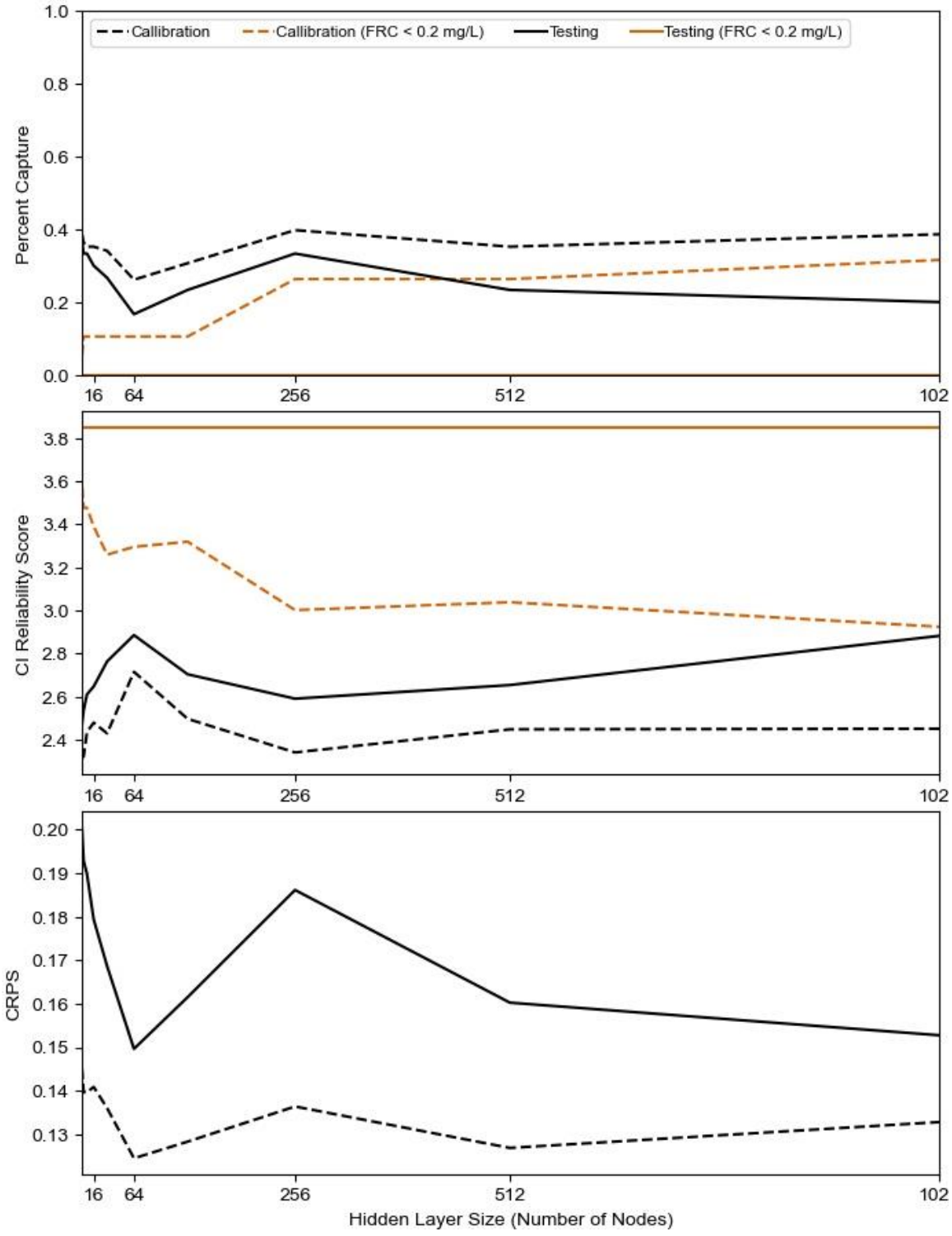

Figure A-19: Rwanda IV1 hidden layer size selection. Hidden layer size of 4 selected to avoid performance decreases.

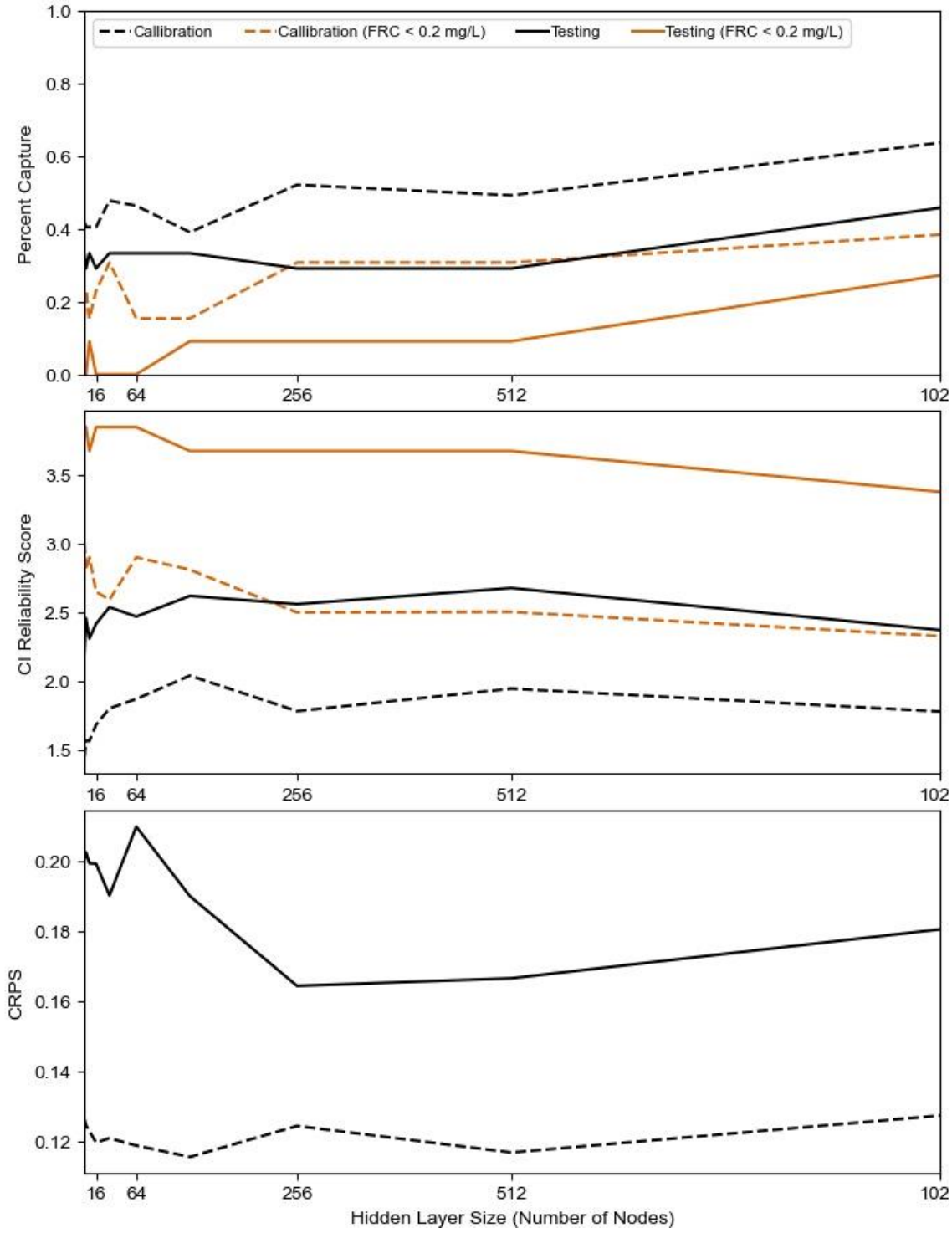

Figure A-20: Rwanda IV2 hidden layer size. Hidden layer size of 8 selected to allow for good overall, good capture good capture of observations with point-of-consumption FRC below 0.2 mg/L, as well as good CI reliability and CRPS.

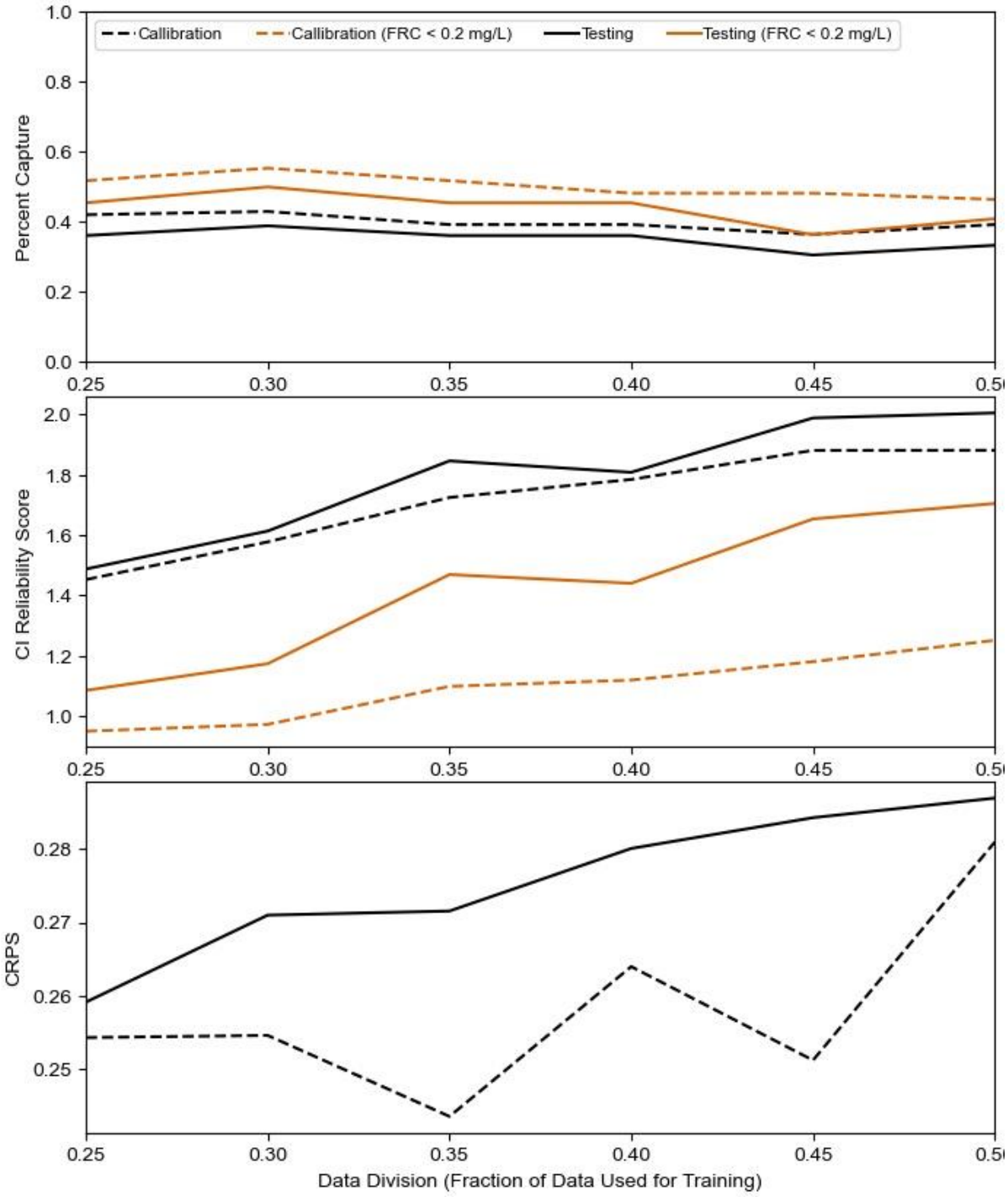

Figure A-21: South Sudan IV1 data division selection. Both testing and calibration performance decreases as the percent of data used for training increases, so a training fraction of 25% was selected, for an overall training-validation-testing data division of 25%-50%-25%.

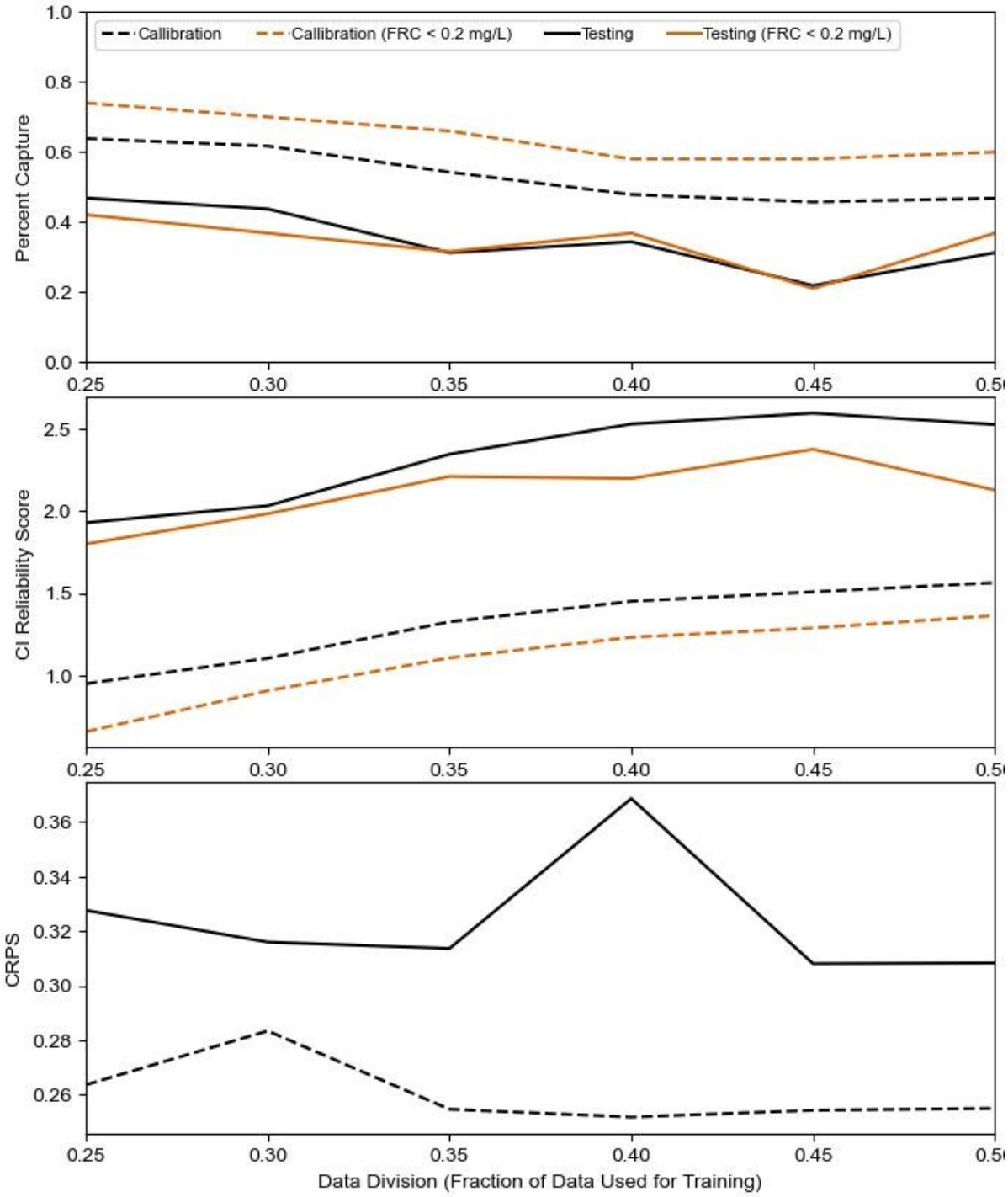

Figure A-22: South Sudan IV2 data division selection. Both testing and calibration performance decreases as the percent of data used for training increases, so a training fraction of 25% was selected, for an overall training-validation-testing data division of 25%-50%-25%.

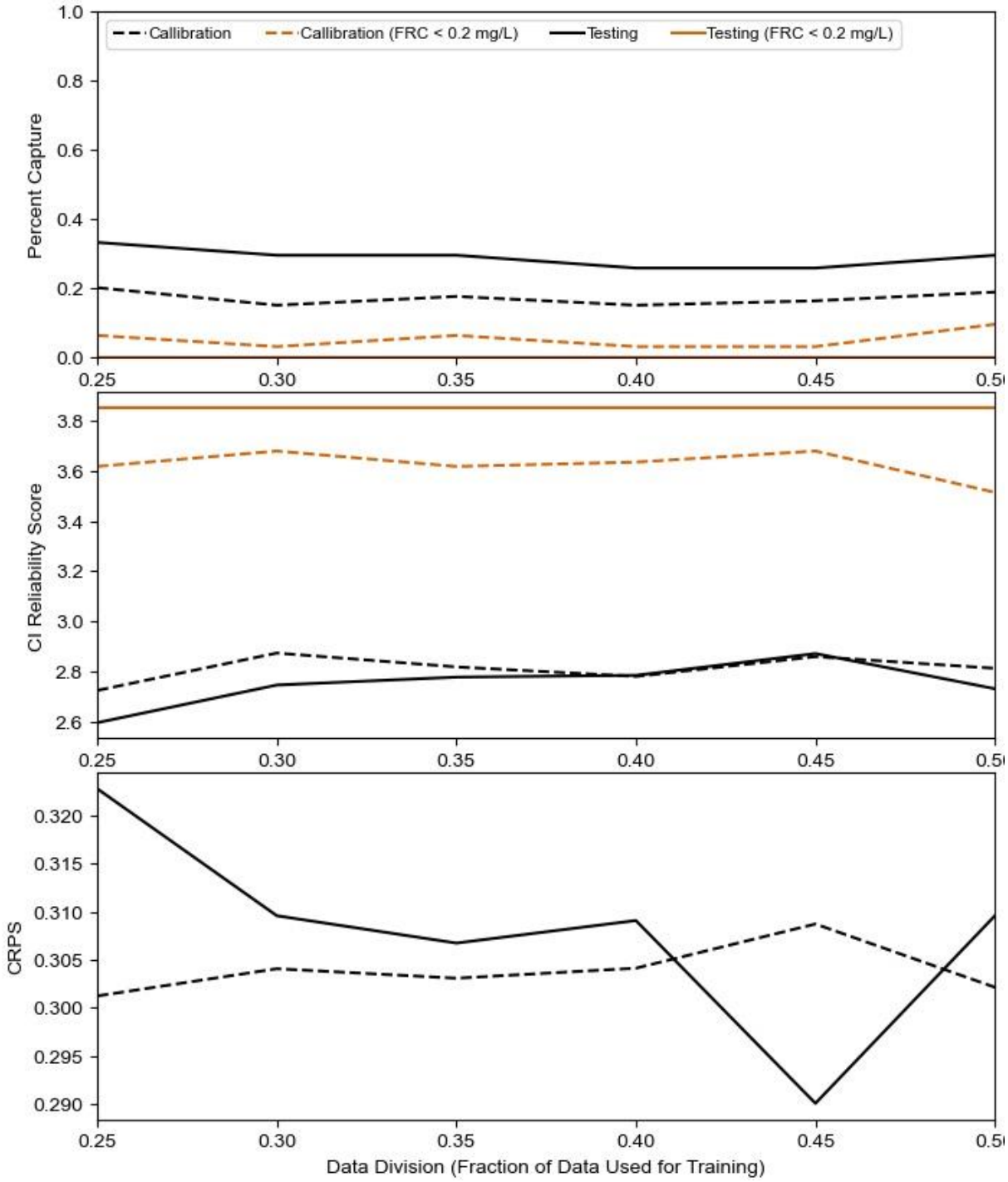

Figure A-23: Jordan (2014) IV1 data division selection. Flat performance except CRPS which improves with increasing training data, so a training fraction of 50% is selected, for an overall training-validation-testing data division of 50%-25%-25%.

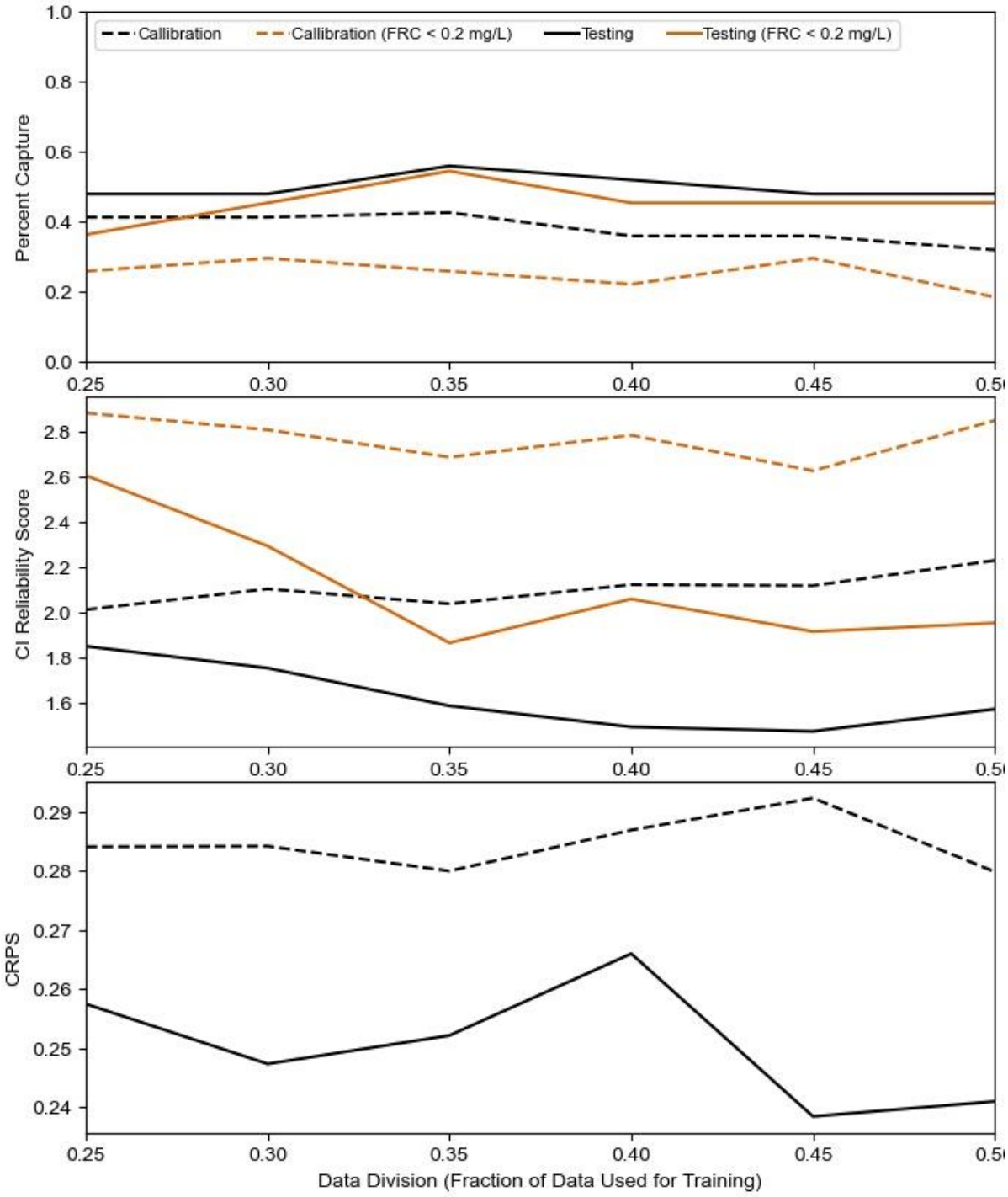

Figure A-24: Jordan (2014) IV2 data division selection. Both testing and calibration performance decreases as the percent of data used for training increases, indicating a 25% data fraction was selected, for an overall training-validation-testing data division of 25%-50%-25%.

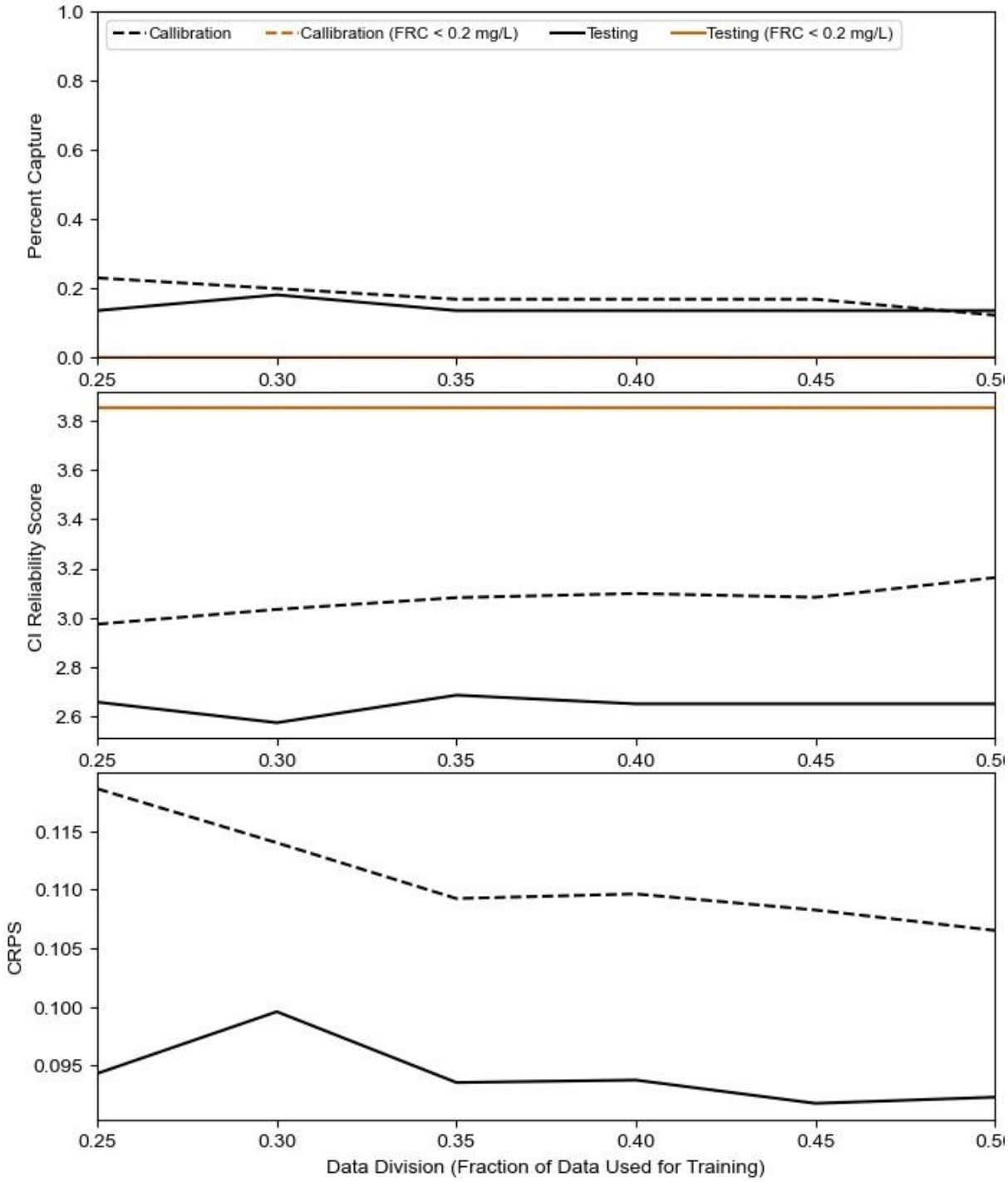

Figure A-25: Jordan (2015) IV1 data division selection. Calibration and testing performance diverge when the training fraction increases beyond 30% so this is selected as the training fraction, for an overall training-validation-testing data division of 30%-45%-25%.

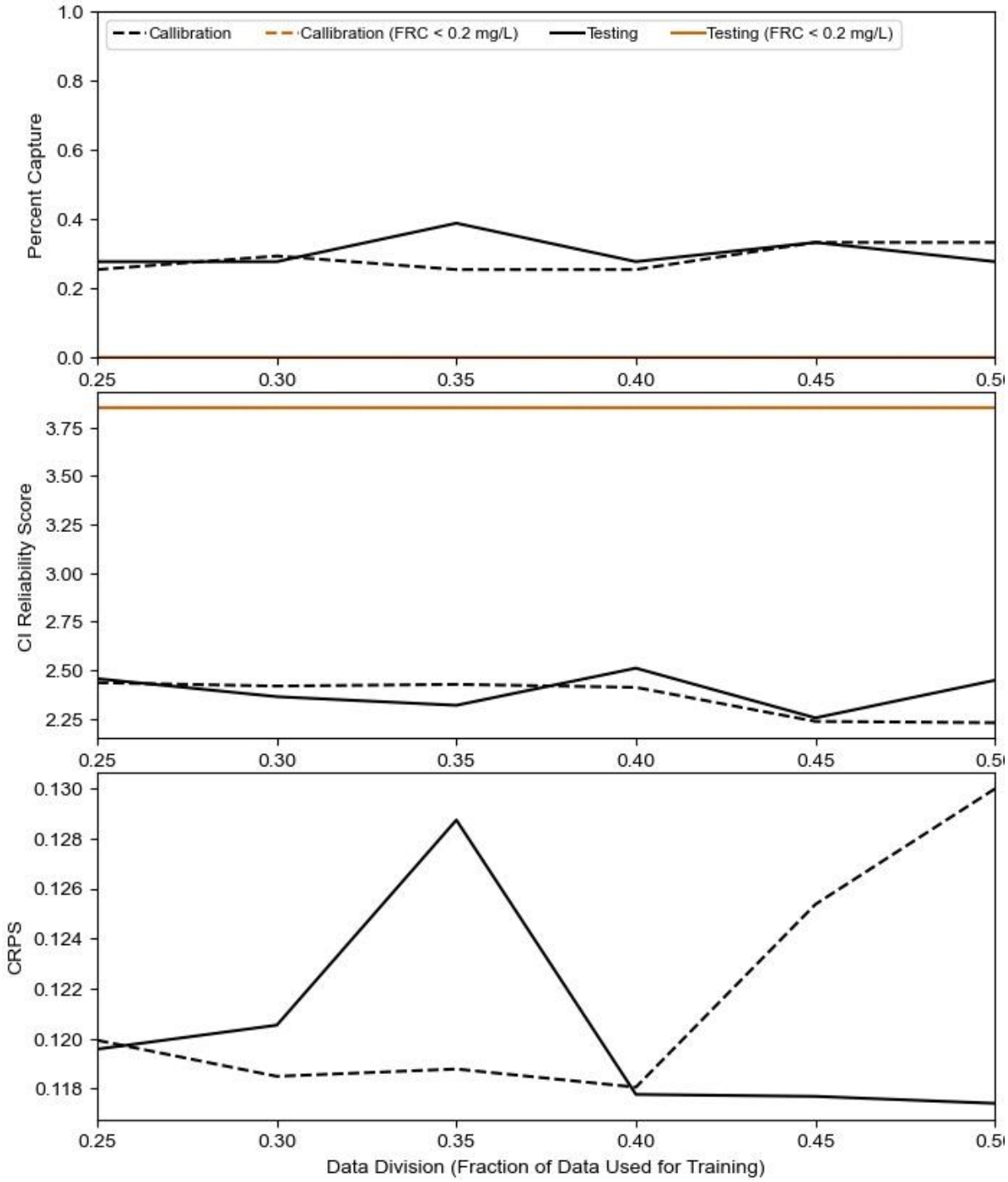

Figure A-26: Jordan (2015) IV2 data division selection. Best testing and calibration performance for Percent Capture and CI reliability occur when the training fraction is 25%, for an overall training-validation-testing data division of 25%-50%-25%.

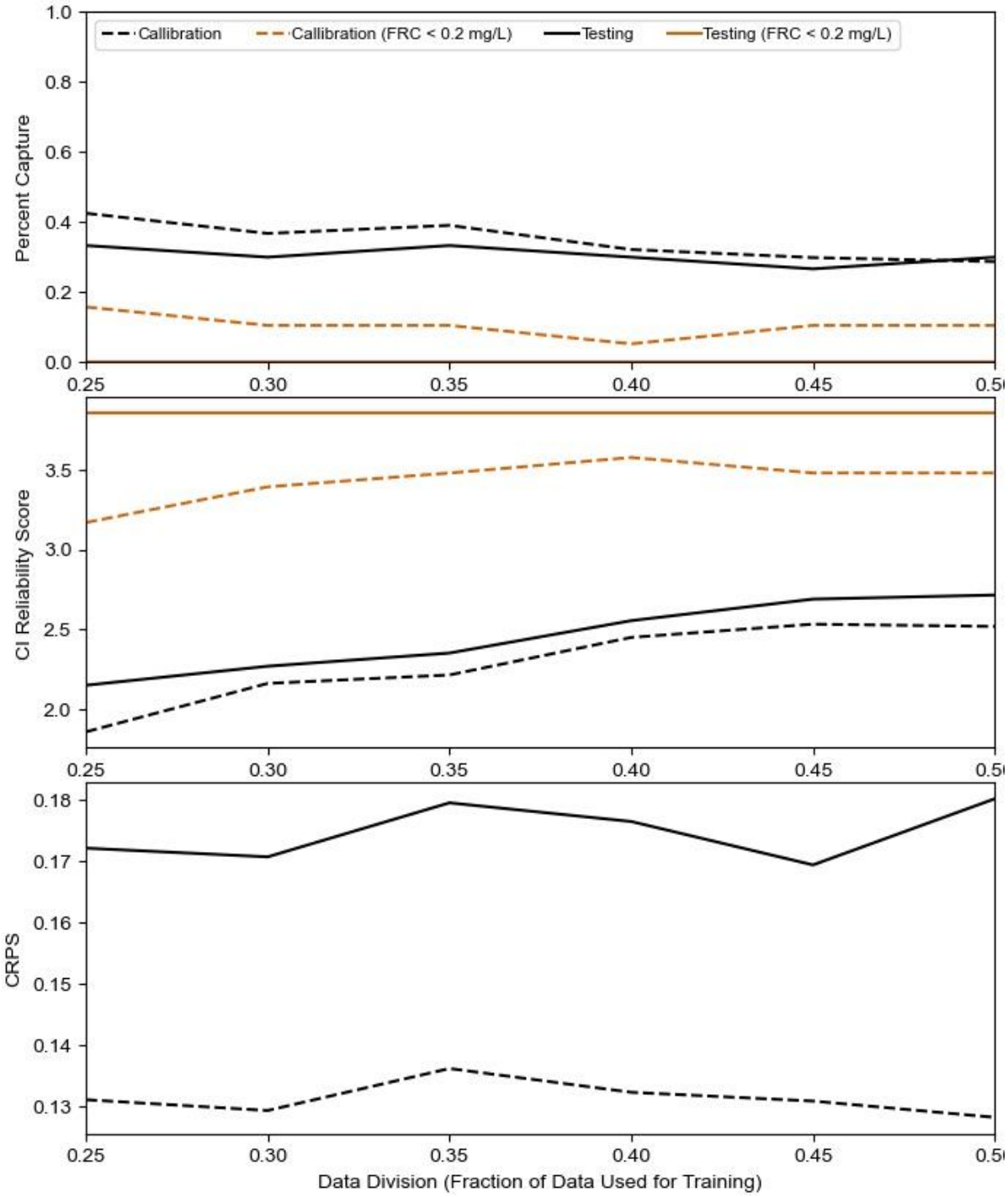

Figure A-27: Rwanda IV1 data division selection. Flat performance except CRPS which worsens with increasing training data, so a training fraction of 25% is selecte, for an overall training-validation-testing data division of 25%-50%-25%.

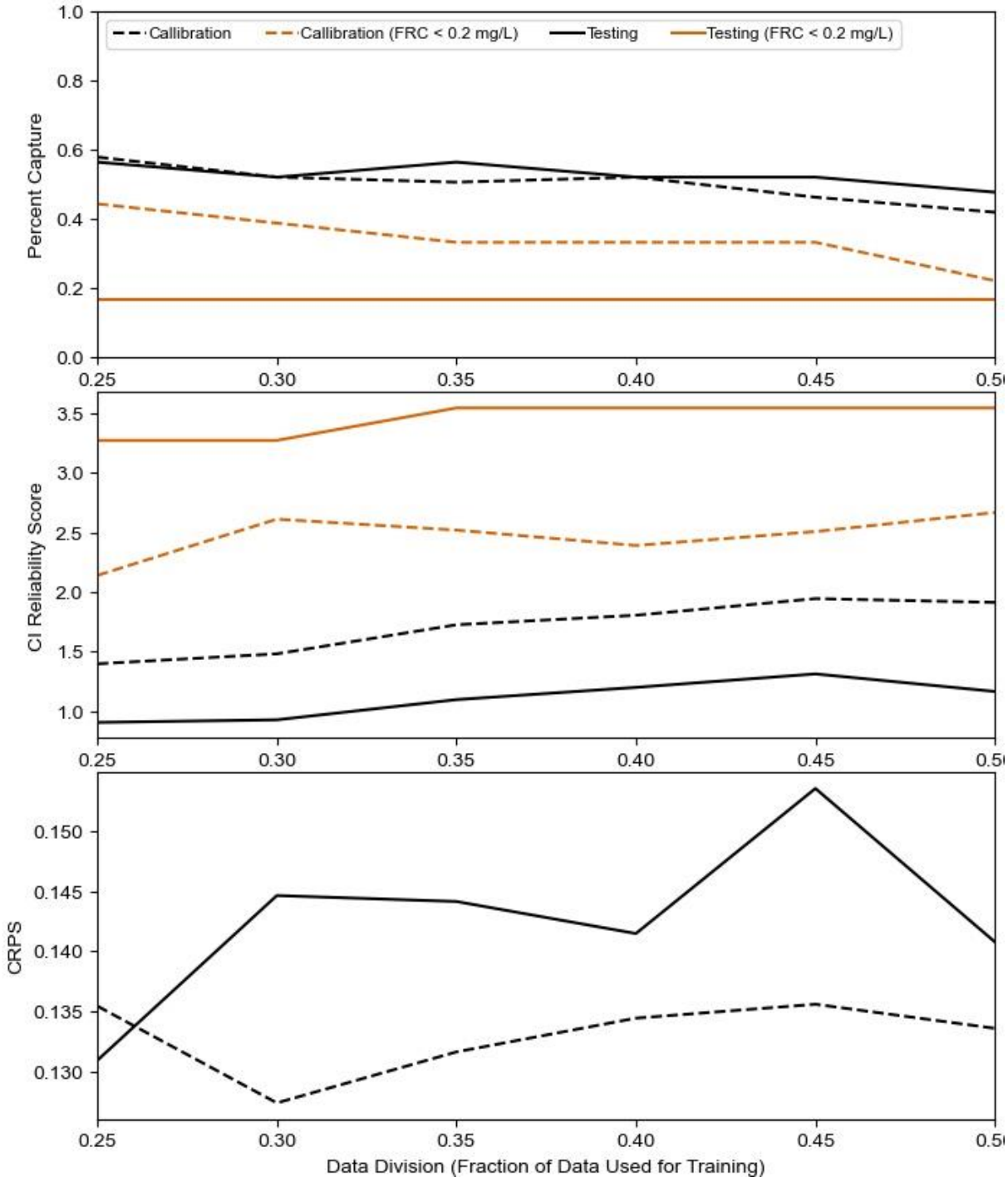

Figure A-28: Rwanda IV2 data division selection. Both testing and calibration performance decreases as the percent of data used for training increases, so we select a training fraction of 25%, for an overall training-validation-testing data division of 25%-50%-25%.

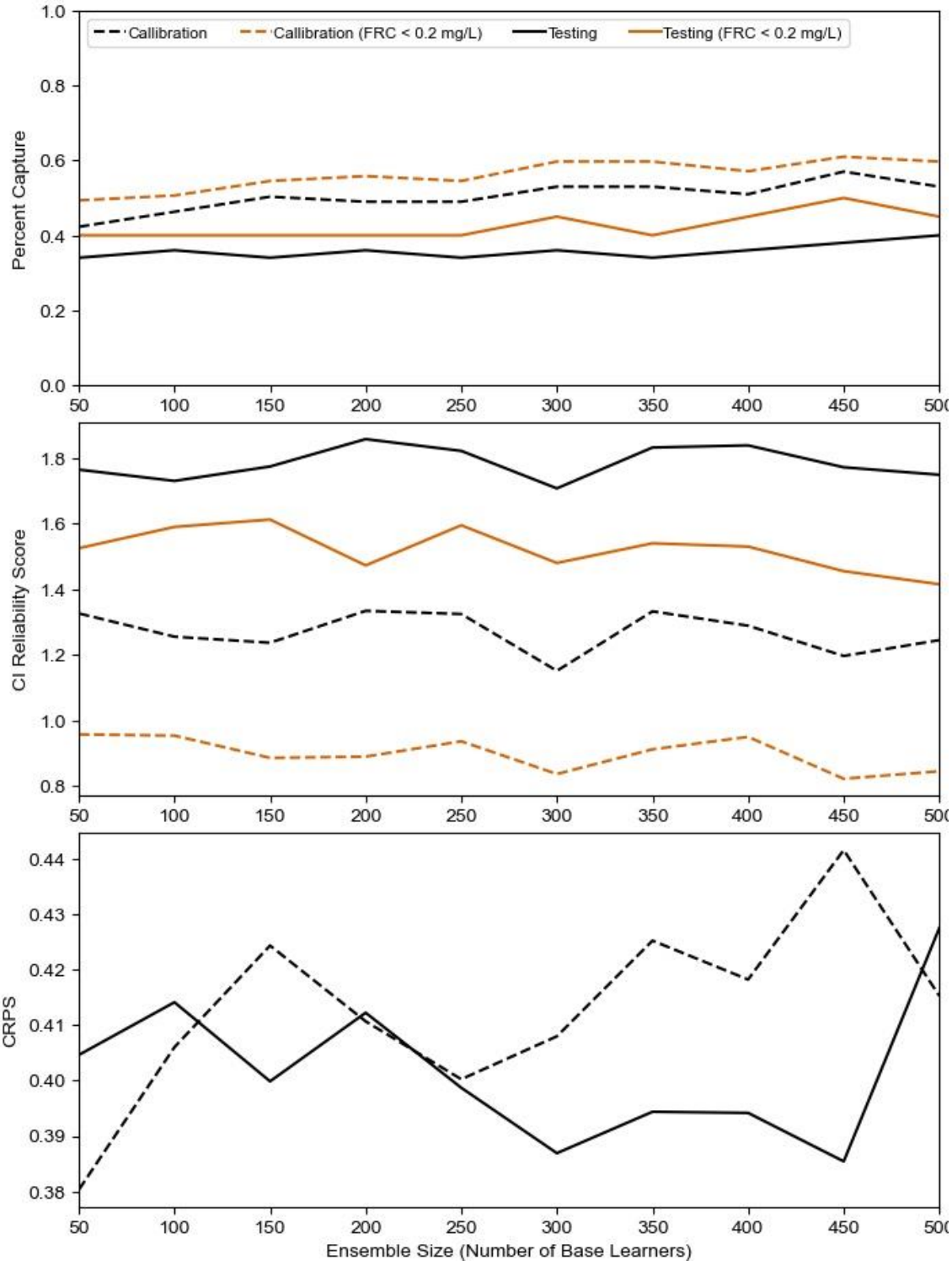

Figure A-29: South Sudan IV1 ensemble size selection. Only minimal improvements in performance as ensemble size increases, so an ensemble size of 50 members is preferred.

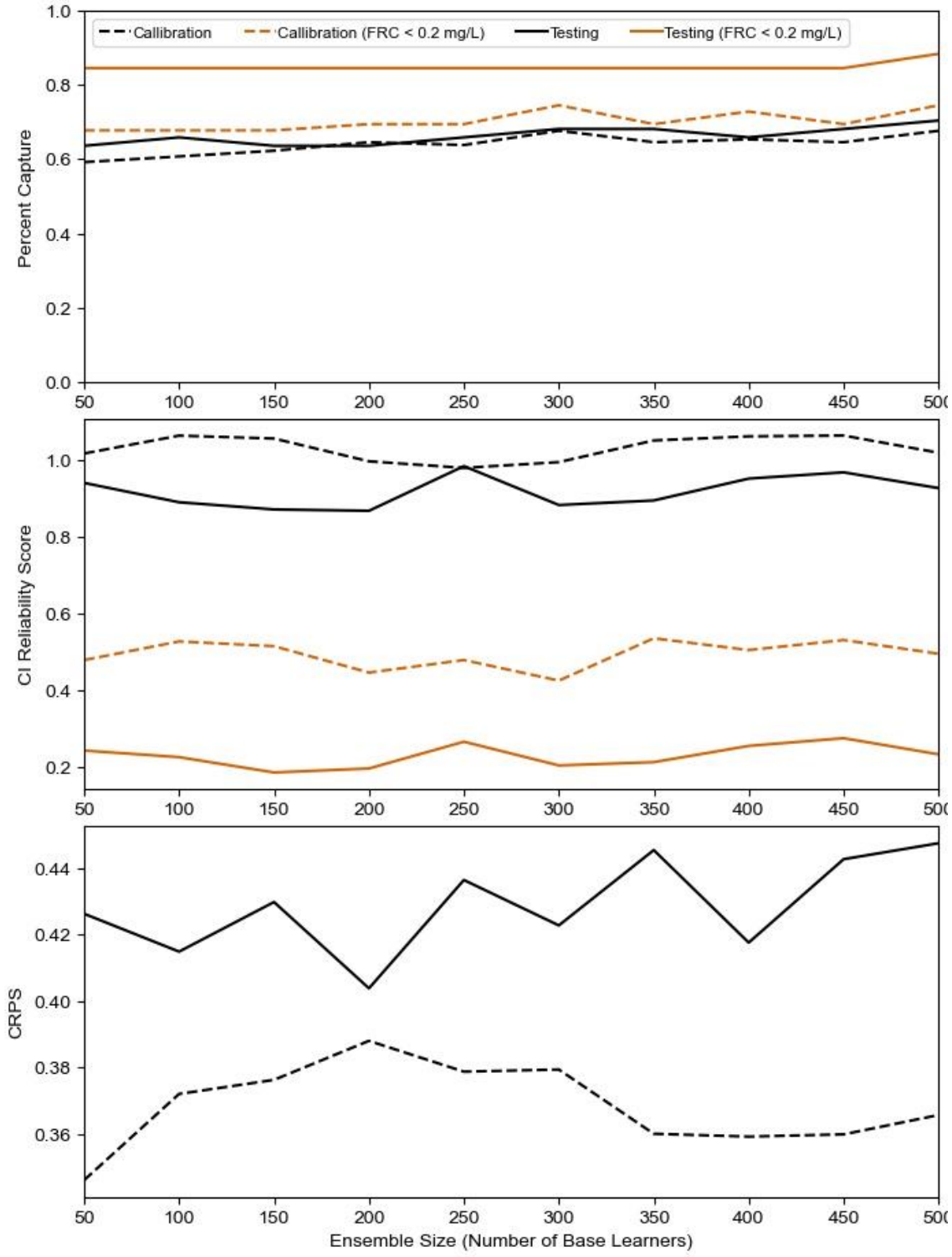

Figure A-30: South Sudan IV2 ensemble size selection. Only minimal improvements in performance as ensemble size increases, so an ensemble size of 50 members is preferred.

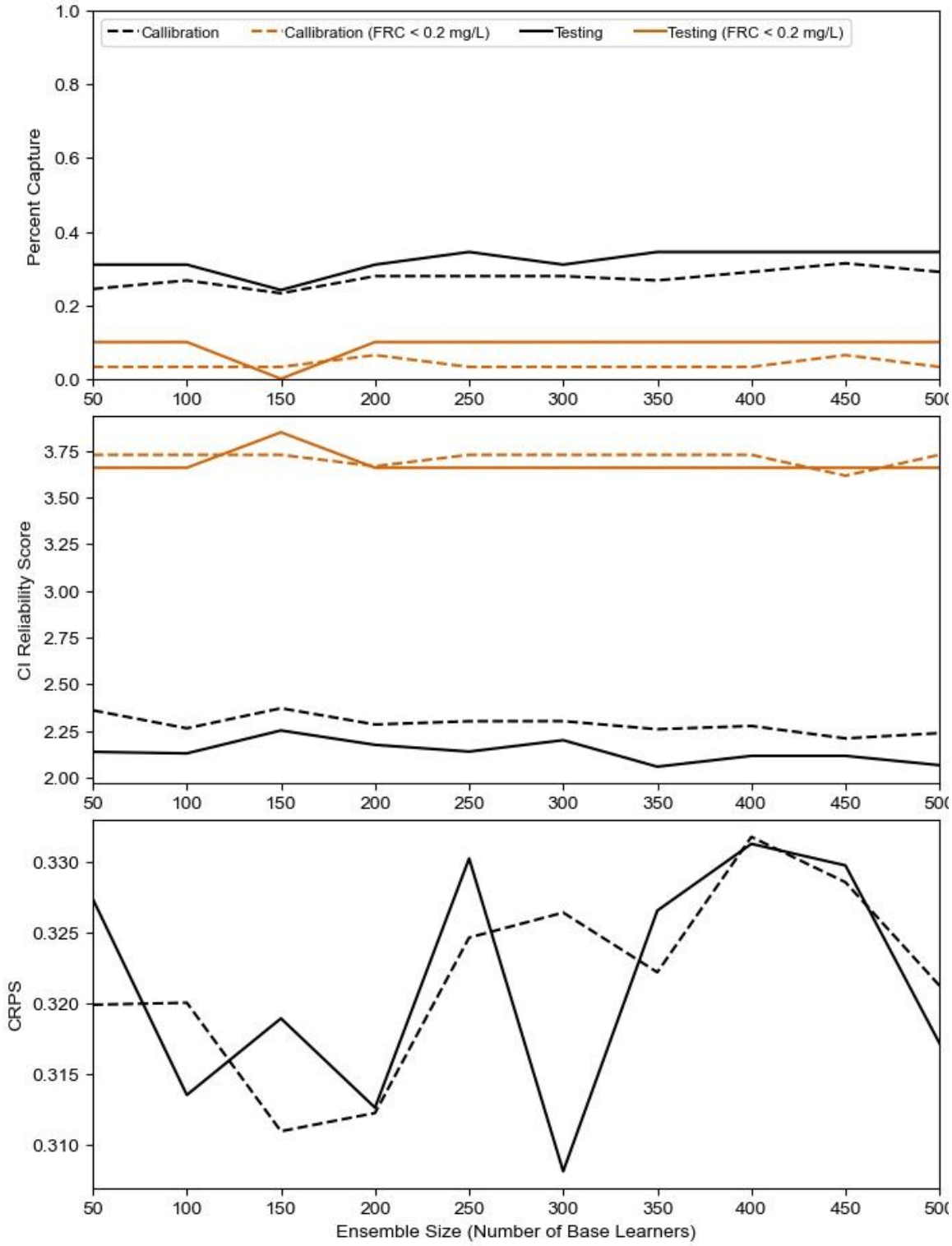

Figure A-31: Jordan (2014) IV1 ensemble size selection. Ensemble size of 200 provides best simultaneous Percent Capture and CI reliability, particularly for unsafe observations.

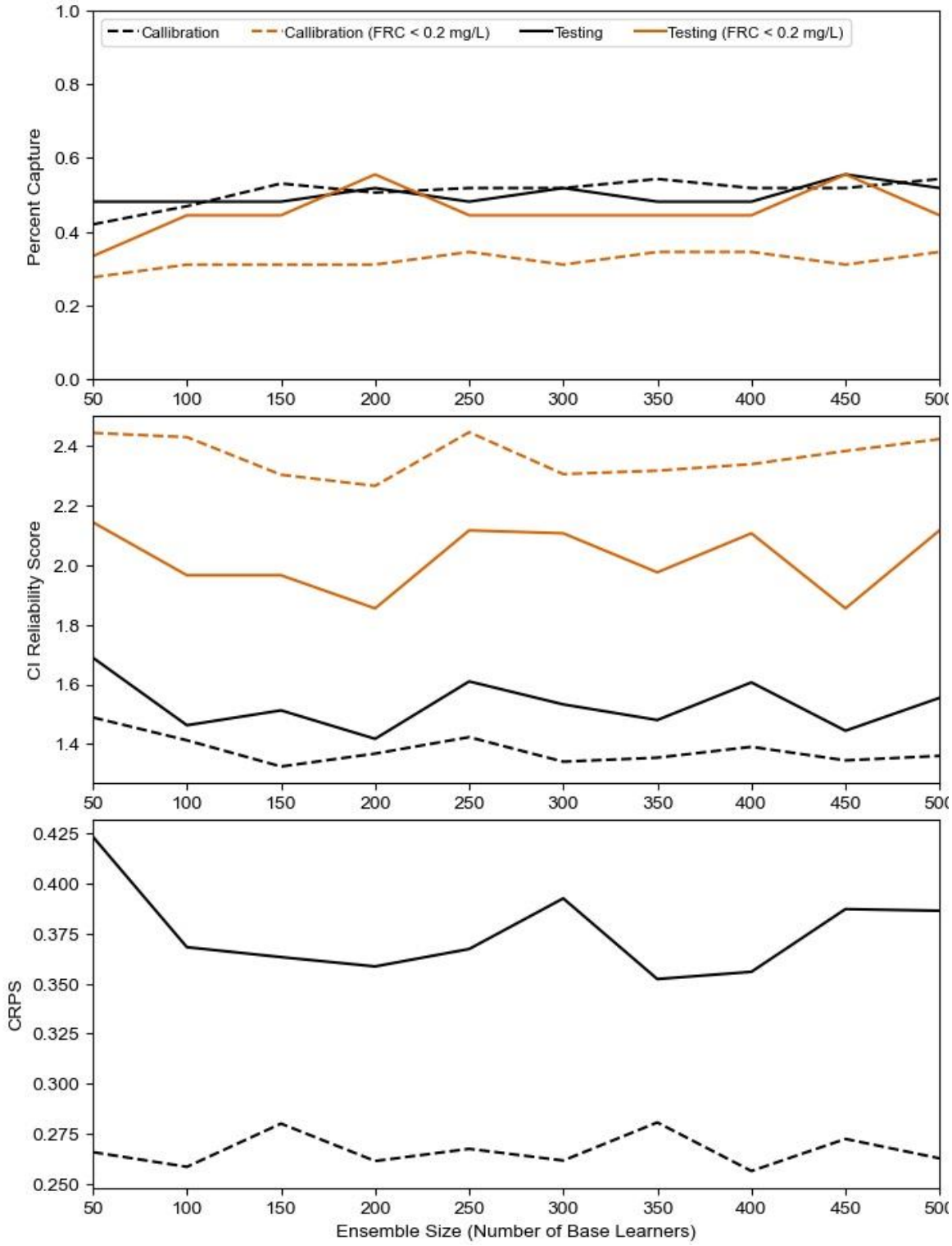

Figure A-32: Jordan (2014) IV2 ensemble size selection. An ensemble size of 200 members provides the best capture of unsafe values.

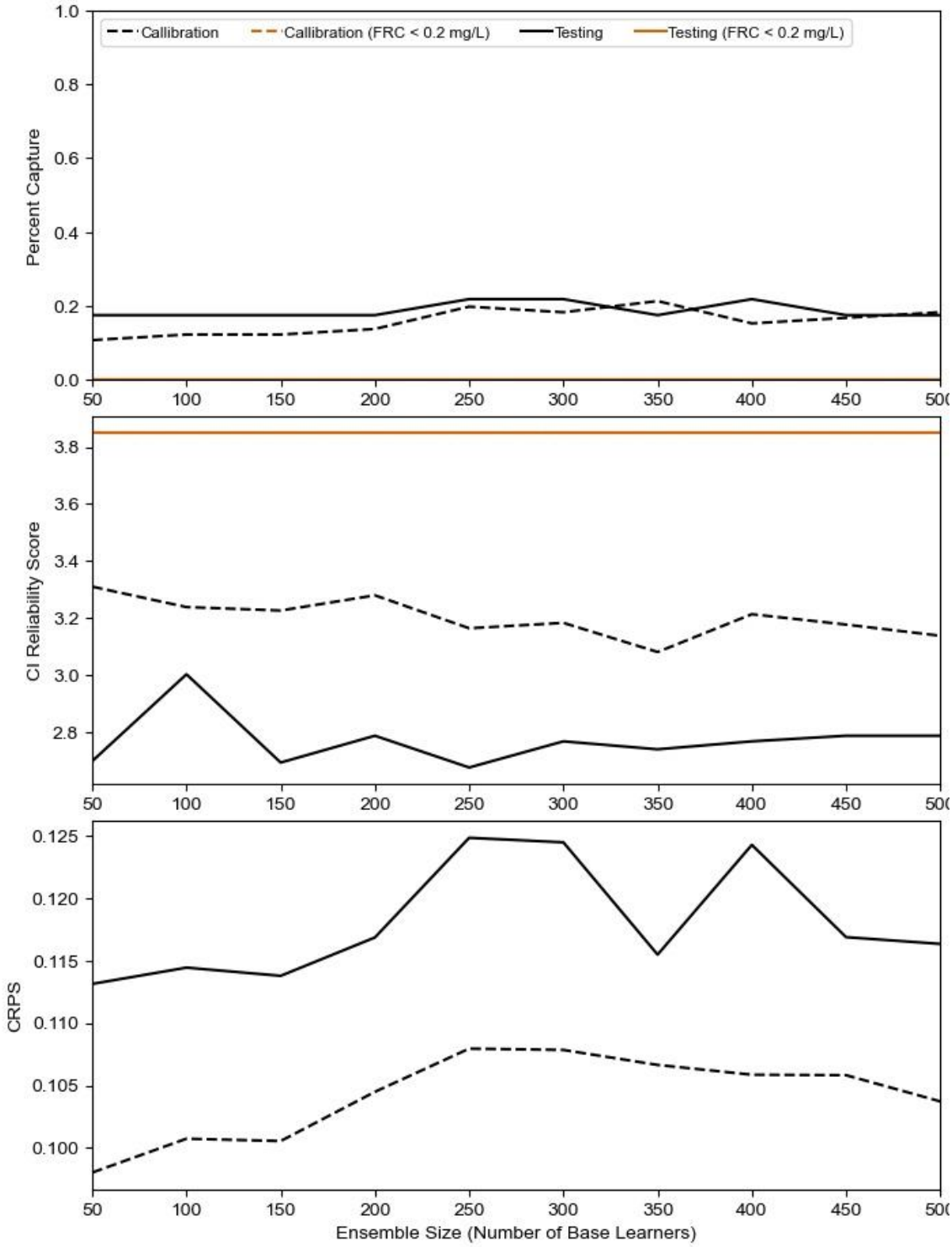

Figure A-33: Jordan (2015) IV1 ensemble size selection. Ensemble size of 250 provides best capture as well as best CI reliability.

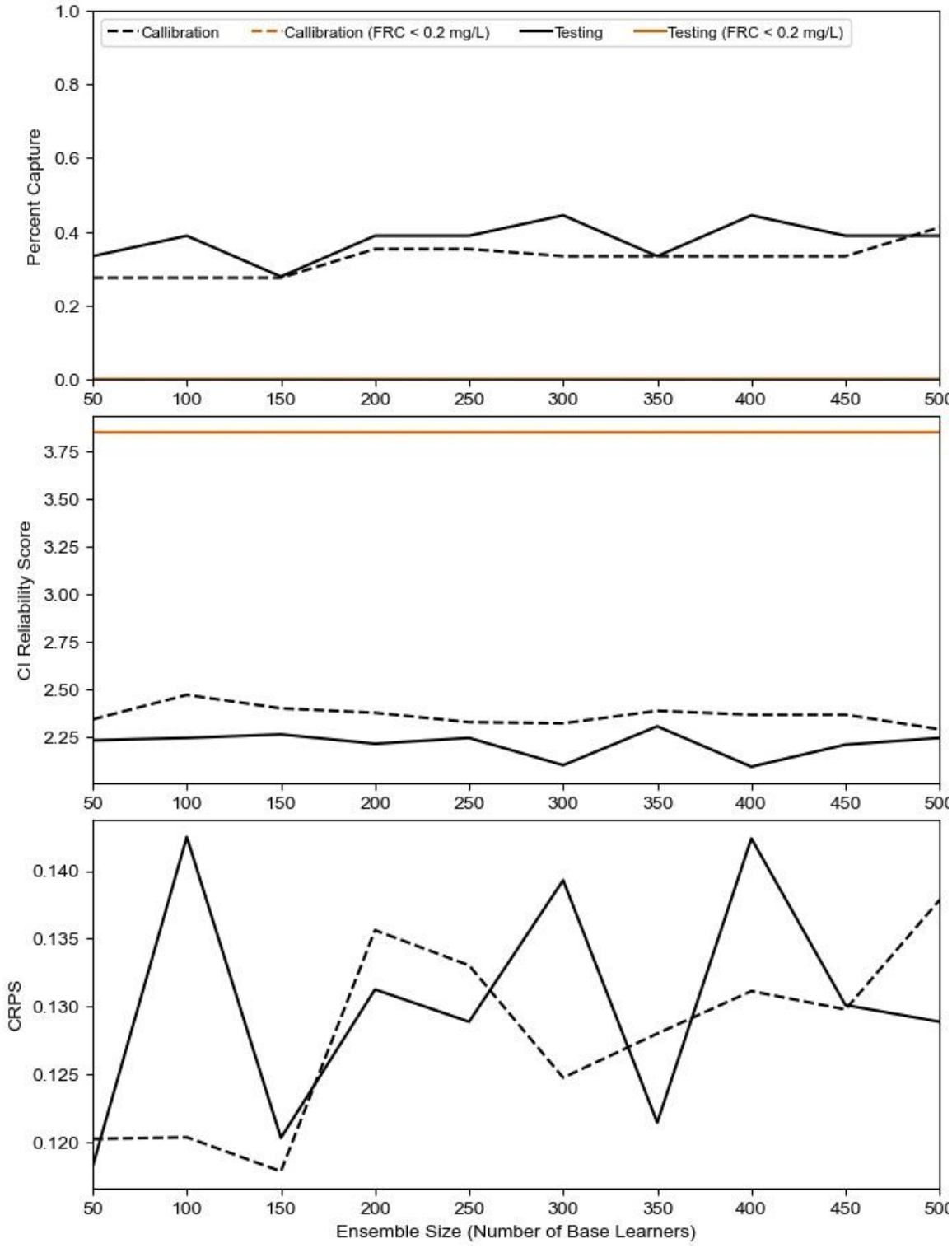

Figure A-34: Jordan (2015) IV2 ensemble size selection. Good capture at ensembles sizes of 100 and 300, ensemble size of 100 is preferred for computational efficiency.

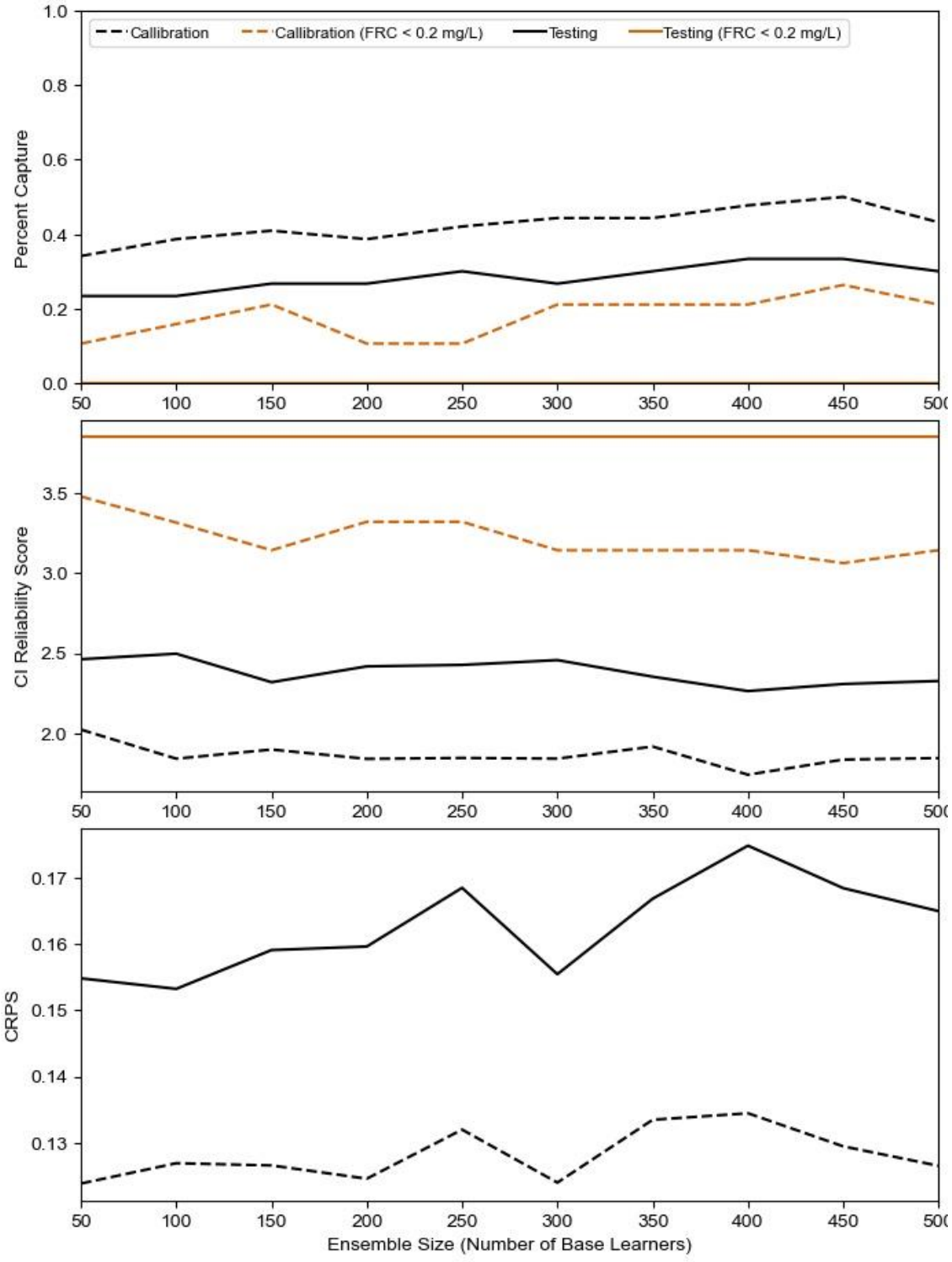

Figure A-35: Rwanda IV1 ensemble size selection. Slight improvement in performance for Percent Capture and CI reliability starting at ensemble size of 250 members.

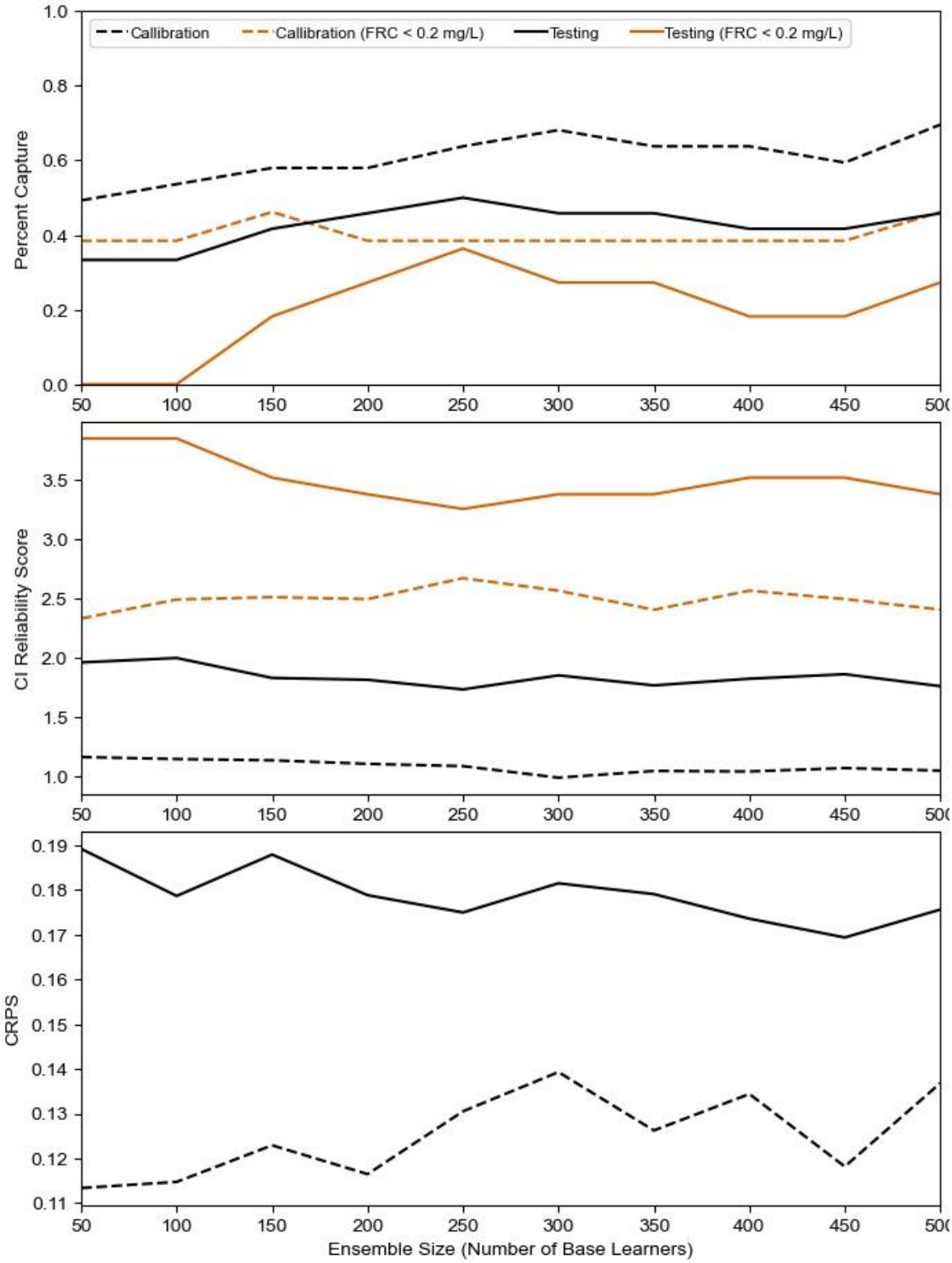

Figure A-36: Rwanda IV2 ensemble size selection. Ensemble size of 250 members provides best performance for unsafe values.

Appendix B. Unpublished work: “Predicting drinking water safety in humanitarian crises using artificial neural networks”

B.1 Abstract

Free residual chlorine (FRC) protects against pathogenic recontamination of drinking during collection, transport, and household storage in refugee and internally displaced person (IDP) settlements. This study used ensembles of artificial neural networks (ANNs) to predict FRC at the point of consumption and generate risk-based FRC targets for six refugee settlements. Input variables configurations were selected using expert opinion and a novel variation of combined network pathway strength analysis, an ANN-specific input variable selection method. The network architecture for the individual ANNs (hidden layer size and division of data into training, validation and testing subsets) was optimized via grid search. Both site specific and global ensemble models were developed and the performance of these two types of models were compared. The ensemble models were also used to develop risk-based FRC targets for a refugee settlement in Bangladesh.

At most sites the variables selected through combined network pathway strength analysis outperformed the configurations selected through expert opinion and the network architecture optimization typically selected a small fraction of the data for training and large hidden layers. We also found that local models outperformed the global model in all cases, showing that local models trained with a small amount data achieves better performance than a global model with a large amount of data. The FRC targets developed for the Bangladesh refugee settlement using the ANN ensembles varied over time but were consistently higher than current drinking water quality guidelines, indicating that the static FRC targets in drinking water guidelines are not appropriate for the dynamic context of refugee and IDP settlements. This research presents the first use of ANNs for modelling drinking water quality in refugee and IDP settlements and is a crucial step in developing evidence-based, effective FRC guidance for humanitarian response.

B.2 Acronyms and Abbreviations

AIC	Akaike Information Criterion
ANN	Artificial neural network
CNPSA	Combined network pathway strength analysis

CPHS	Committee for Protection of Human Subjects
DIGHR	Dahdaleh Institute for Global Health Research
EC	Electrical conductivity
EQR	Ensemble interquartile range
EWR	Ensemble weight range
FRC	Free residual chlorine
IDP	Internally displaced person
IVS	Input variable selection
MLP	Multi-layer perceptron
MSE	Mean squared error
MSF	Médecins sans Frontières
pmf	probability mass function
R^2	Coefficient of determination
RMSE	Root mean squared error
SWOT	Safe Water Optimization Tool
TRC	Total residual chlorine
UNHCR	United Nations High Commissioner for Refugees

B.3 Introduction

Recontamination of drinking water while it is stored in the household contributes to the spread of waterborne diseases in refugee and internally displaced person (IDP) settlements. Household recontamination of drinking water has been identified as a contributing factor in outbreaks of waterborne diseases in refugee and IDP settlements in Kenya (Golicha et al., 2018; Shultz et al., 2009), Malawi (Swerdlow et al., 1997), Sudan (Walden et al., 2005), South Sudan (Ali et al., 2015; Guerrero-Latorre et al., 2016), and Uganda (Howard et al., 2010; Steele et al., 2008). Free

residual chlorine (FRC) is used to prevent household recontamination in refugee and IDP settlements and is used as an indicator of water safety against pathogenic recontamination in humanitarian response as FRC ≥0.2 mg/L has been shown to prevent recontamination by *Escherichia Coli* in humanitarian response contexts. Humanitarian drinking water guidelines recommend maintaining FRC at public water distribution points (also called "tapstands") between 0.2 and 0.5 mg/L (Sphere Association, 2018). These recommendations are based on guidelines for piped water distribution systems in nonemergency settings and provide insufficient protection in the context of refugee and IDP settlements (Ali et al., 2020b). In these settlements, water-users do not usually have piped connections to the household and instead water is centrally treated and piped to water distribution points where water users collect drinking water which is then stored and used over a period of up to 24 hours. Post-distribution FRC decay during collection, transport and household storage can leave drinking water vulnerable to pathogenic recontamination, increasing the risk of spreading waterborne diseases. Thus, water that meets the guidelines at the distribution point may become unsafe to drink when it is consumed (Ali et al., 2015; Ali et al., 2020b).

The Safe Water Optimization Tool (SWOT) is a new initiative led by the Dahdaleh Institute for Global Health Research (DIGHR) and Médecins sans Frontières (MSF) responding to the challenge of ensuring drinking water safety in refugee and IDP settlements. This project aims to use regularly-collected water quality data from refugee and IDP settlements to generate evidence-based FRC targets for public water distribution points that protect drinking water against recontamination up to the point of consumption. Earlier work on this project developed distribution point FRC targets using a physical FRC decay model based on overall empirical reaction kinetics to model post-distribution FRC decay (Ali et al., 2000b).

Studies on the use of physical models of FRC decay in piped distribution systems have shown that FRC decay is highly complex and dependent on interactions with other water quality parameters in the bulk phase (the main unit of water) as well as physical parameters of the distribution system (Rossman, Clark, and Grayman, 1994). Additionally, FRC decay behaviour is specific to the water distribution system and can vary over time and space within the system (Vasconcelos et al., 1997). Due to the complexity and variability of FRC decay, physical models can be impractical in highly dynamic environments as they require frequent recalibration to account for changes in other variables influencing FRC decay, such as other water quality

parameters (Bowden et al., 2006; Soyupak et al., 2011). To overcome these challenges, several studies have used artificial neural networks (ANNs) as an alternative to physical models for predicting FRC in piped distribution systems (Bowden et al., 2006; Gibbs et al., 2006; Rodriguez & Sérodes, 1998; Soyupak et al., 2011). ANNs are a type of data-driven model that approximate functions based on the relationship between input and output data instead of assuming a behaviour *a priori* (Solomatine and Ostfeld, 2008). This allows them to accurately represent complex FRC decay behaviour. Unlike physical models ANNs can also directly incorporate other variables, such as water quality variables into the model, so they do not require frequent recalibration for different scenarios. Additionally, ANNs can be quickly and easily retrained, allowing them to adapt to the spatial and temporal variability of FRC decay (Gibbs et al., 2006; Khan, He, and Valeo, 2018).

While ANNs offer advantages over physical models, they still only output discrete predictions which cannot account for the within-site variability of FRC decay. Due to the variability of FRC decay in refugee and IDP settlements, any FRC concentration at the water distribution point can produce a range of FRC concentrations in the household. Thus, an approach is needed that can produce probabilistic predictions of FRC at the point of consumption which can be used to identify the risk of unsafe drinking water. Ensembles of ANNs can be used to produce probabilistic predictions by grouping the predictions of multiple ANNs to form a probability mass function (pmf) from their predictions (Boucher et al., 2009). This pmf can then be used to predict, for any concentration of FRC at the water distribution point, the probability that FRC at the point of consumption will be below a given threshold which can be used to generate risk based FRC targets for the distribution point.

This study used ensembles of ANNs to predict point of consumption FRC using data from six refugee settlements. Model inputs were identified through input variable selection (IVS) using both expert opinion and a novel form of combined network pathway strength analysis (CNPSA). This was used to identify the most useful predictors of point of consumption FRC as these were not evaluated in previous studies (Ali et al., 200b). The network architecture of individual ANNs within the ensemble was optimized using a grid search method described in Khan and Valeo (2018). The IVS and network architecture optimization procedures were used to design local ensemble models for each of the six sites and a global ensemble model which combined data from multiple sites. The local and global models were compared to assess which of these two

approaches produce the best ensemble model performance. Finally, the ensemble models were used to develop new FRC recommendations for one of the sites in this study, which were compared to existing drinking water quality guidelines. The results of this investigation will be used to inform future development of the SWOT project which is currently using these ensemble ANN models to develop risk-based FRC targets for humanitarian response.

B.4 Methods

B.4.1 Study Sites

Model development used an existing dataset from a multi-site study on post-distribution FRC decay collected from six refugee settlements in South Sudan, Jordan, Rwanda, and Bangladesh between 2013 and 2019 (Ali et al., 2020a, Ali et al., 2020b). Table B-1 summarizes important characteristics of each settlement, including factors such as population, climate information, and water treatment processes at each site as well as the duration of data collection and number of samples collected at each settlement. From Table B-1, the volume of data collected in Bangladesh is larger than the remaining sites combined (2,095 vs 693). Due to this imbalance, the Bangladesh data set was not included in the global model development described in the following sections.

Table B-1: Summary of Key Site Characteristics (adapted from Ali et al., 2021)

Site Country	Name of Refugee Settlement	Ambient Air Temperature (°C)	Population	Water Source	Drinking Water Treatment	Data Collection Period	Number of Paired Samples Collected
South Sudan	Batil	Average: 35.3 (Min: 28.3; Max: 45.7)	37,199[1]	Groundwater (boreholes)[1]	In-line chlorination with calcium hypochlorite[1]	March-April, 2013	69
	Gendrassa		15,810[1]				76
	Jamam		15,670[1]				75
Jordan	Azraq (Summer)	Average: 32.7 (Min: 27.1; Max: 43.3)	7,470[2]	Groundwater (boreholes)[2]	Reverse osmosis; in-line chlorination with chlorine gas[2]	July-August, 2014	199
	Azraq (Winter)	Average: 21.7 (Min: 14.5; Max: 29.3)	14,797[2]			March-April, 2015	140
Rwanda	Kigeme	Average: 22.2 (Min: 18.3; Max: 31.0)	18,569[3]	Surface water (stream source)[3]	Flocculation, filtration, and chlorination with calcium hypochlorite[3]	June-July, 2015	134
Bangladesh		Average: 32.6	83,000[4]				2095

Kutapalong-Balukhali, Camp 1	(Min: 15.0; Max: 49.0)	Groundwater (boreholes) and handpumps for non-potable purposes[5]	In-line chlorination with calcium hypochlorite (handpumps untreated)[5]	July-December 2019

[1] Data source: *Médecins Sans Frontières, Maban County, South Sudan WASH Coordination Report (Weeks 11 and 12, 2013).*

[2] Data source: *UNICEF, Azraq, Jordan WASH Monitoring Reports (2014 and 2015).*

[3] Data source: *PAJER, Kigeme, Rwanda WASH Monthly Updates (June-July 2105).*

[4] Data source: *UNHCR Population data and key demographical indicator (Block Level) - 15 August 2019*

[5] Data source: *REACH/UNICEF 2019, Water, Sanitation and Hygiene (WASH) Household Dry Season Follow-up Assessment (May 2019), All Camps, Ukhia and Teknaf Upazilas, Cox's Bazar District, Bangladesh. Available at: https://www.humanitarianresponse.info/sites/www.humanitarianresponse.info/files/2019/07/1905_REACH_WASH_Assessment_Dry_Season_May2019.pdf. (Note that this value is averaged across the entire area of the Cox's Bazar Refugee Settlement)*

B.4.2 Data Collection

Water quality, water handling and environmental variables were collected as potential input variables to predict FRC at the point of consumption. Water quality factors, such as dissolved salts (represented through electrical conductivity, EC), as well as organic content and minerals (represented through turbidity), have been shown to influence FRC decay in the bulk phase in studies of FRC decay in piped distribution systems (Rossman et al., 1994). Environmental factors, specifically air temperature, were found to significantly impact post-distribution FRC decay rates in three South Sudanese refugee settlements (Ali et al., 2015). Water handling variables have been shown to influence FRC decay in water stored in the household (Meierhofer, Wietlisbach, and Matiko, 2019). For each site's dataset, changes in water quality from the water distribution point to the point of consumption (also referred to as the household) were documented using paired sampling, so the same unit of water was sampled at both the water distribution point and the household, after a predetermined period of time. The water storage duration varied from site to site depending on overall water supply availability, so the follow-up time for the household measurement was varied from 4 hours up to 24 hours. Table B-2 summarizes the data collected at each site and Appendix A includes the data cleaning steps that were used to prepare the data for use in the ANN models.

Table B-2: Summary of variables and data collection methods in post-distribution water quality dataset.

Category	**Parameter**	**Location where Parameter Measured**	**Sites where Parameter was Measured**	**Collection Method**
General Information	Date and time of each observation	Tapstand Household	All	
	Free residual chlorine	Tapstand Household	All	Colorimetric method using Palintest PTH

Category	Parameter	Location where Parameter Measured	Sites where Parameter was Measured	Collection Method
	Total residual chlorine	Tapstand Household	All	7091 compact chlorometer and Wagtech 7100 photometer with Palintest DPD1/DPD3 reagents (Palintest Ltd., Tyne and Wear, UK)
Physical and chemical water parameter	Turbidity	Tapstand Household	All	Nephelometric method using Palintest PTH 090 compact turbidimeter (Palintest Ltd., Tyne and Wear, UK)
	Water temperature	Tapstand	All	Potentiometric method using Eijkelkamp 18.21 multimeter (Eijkelkamp Agrisearch Equipment, Giesbeek, Netherlands), Hanna
	Conductivity	Household	All	
	pH	Tapstand	All	
Environmental parameters	Air temperature	Tapstand Household	All	

Category	Parameter	Location where Parameter Measured	Sites where Parameter was Measured	Collection Method
				Instruments HI 98311 EC/TDS/temperature multi-meter (HANNA Instruments, Woonsocket, RI, USA), or a Hach sensION+ multi-meter (Hach Instruments, USA)
Water handling and behavioral parameters	Cleanliness of the container	Tapstand Household	All	Spot check or self reporting
	Type of container	Tapstand Household	All	
	If the container is covered	Tapstand Household	All	
	Drawing method	Household	All	
	If water was stored in the sun or the shade	Household	Jordan Rwanda Bangladesh	

Category	Parameter	Location where Parameter Measured	Sites where Parameter was Measured	Collection Method
	If the water had been mixed	Household	All	
Water handling and behavioral parameters	If the water had been transferred to a different vessel	Household	Batil Gendrassa Jamam Jordan Rwanda	Spot check by water system operator or self reporting by water user
	If the water was used in the household	Household	Batil Gendrassa Jamam Jordan Rwanda	
	Percentage of water remaining	Household	Batil Gendrassa Jamam Jordan Rwanda	

B.4.3 Ethics

The initial field work in South Sudan received exemption from full ethics review by the Medical Director of 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). The study in Bangladesh received approval from Human Participants Review Committee, Office of Research Ethics at York University (Certificate #: 2019-186), the MSF Ethical Review Board (ID #: 1932), and the Centre for Injury Prevention and Research Bangladesh (Memo #: CIPRB/Admin/2019/168).

B.4.4 Ensemble Model Building Process

The individual ANNs in the local and global ensemble models were built as multi-layer perceptrons (MLPs) using the Keras 2.3.0 package (Cholette, 2015) in Python v3.7 (Python Software Foundation, 2019). 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 B-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 (Bowden et al., 2006; Gibbs et al., 2006; Rodriguez & Sérodes, 1998). 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. A hyperbolic tangent activation function was used in the hidden layer and a linear activation function on the output layer. The hidden layer size was determined via grid search as described in Section 2.4.3.

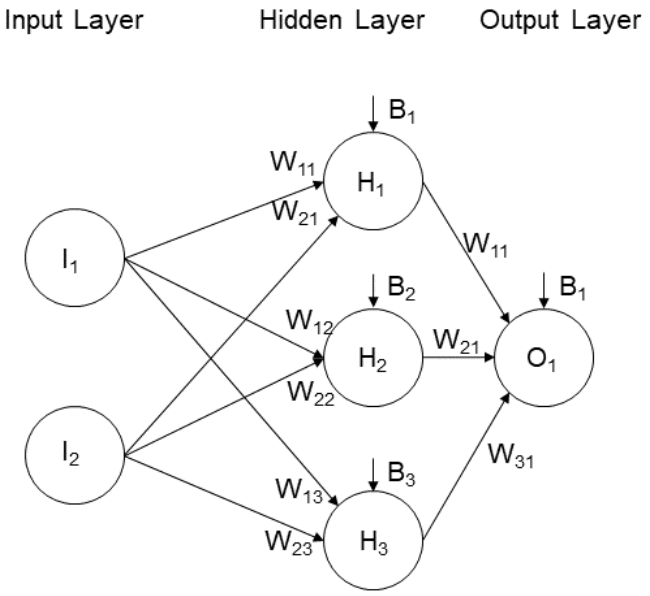

Figure B-1: Schematic of an MLP showing flow of data from the input layer to the output layer with weights and biases. The shown MLP with two input nodes and one output node would have two input variables (other water quality parameters, environmental conditions, water handling behaviours) and one output (household FRC).

To train and test the individual ANNs in the ensemble models, the global dataset and each local dataset were divided into three separate subsets: the training, validation, and testing sets. The network is trained by iteratively adjusting the weights and biases (which are equivalent to calibration parameters in physical models) to minimize the mean squared error (MSE) of the predictions on the training set using the Nadam backpropagation training algorithm. The validation set is used during training to assess MSE on data that is independent of the training data. Initially during training, the training and validation MSE should both decrease, but as training continues the validation MSE will increase, indicating that the model is overfitting (i.e., overly specific to the training data). When the validation MSE begins to increase, training was stopped by a processed called “early stopping”. We used an early stopping procedure with 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. If the validation MSE does not decrease, then the model weights and biases are restored to the iteration

that resulted in the lowest validation MSE. The model performance metrics (see Section 2.4.1) are then evaluated using the testing set. Since the testing set is not included in the training process, the testing performance indicates the model's ability to generalize (Solomatine and Ostfeld, 2008). The procedure for optimizing the division of data into training, validation, and testing subsets is described in Section 2.4.3.

To generate risk-based FRC targets for the water distribution point, multiple ANNs were grouped into an ensemble model. Ensemble models are used to predict probability in ANN applications by grouping predictions from a diverse group of ANNs to form a pmf (Boucher et al., 2009). The diversity of the ensemble arises from the differences between the individual ANNs in the ensemble (Brown, 2011). In this study, diversity was ensured using a multi-start approach where the initial weights and biases were randomized for each ANN so the training algorithm for each ANN started at a different location on the error surface. Multi-start optimization allows for some individual models to converge to locally optimal solutions, but should ensure that the ensemble as a whole is globally optimized. Ensemble diversity was further encouraged by using random sampling to create the training, validation, and testing subsets, so that each individual model is trained, validated, and tested on a different subset of the data. For this study, an ensemble size of 200 ANNs was selected.

B.4.4.1 Performance Indicators

Throughout the design of the local and global ANN models, four indicators were used to measure model performance on the testing subset of the data. Note that for all of the following performance indicators, N is the number of samples, $\bar{y}$ is the mean observed household FRC, and y_i^{obs} and $y_i^{predicted}$ are the observed and predicted household FRC concentrations, respectively.

The first indicator, root mean squared error (RMSE), is a measure of the average error in the same units as the output variable. For this project, the units of RMSE are in the same units as household FRC (i.e., mg/L). The formula for RMSE is shown in Equation B-1.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i^{obs} - y_i^{predicted}\right)^2}{N}} \quad \text{(B-1)}$$

The second indicator, coefficient of determination (R^2) is a measure of how much of the observed variance is explained by the model. The formula for R^2 used is shown in Equation B-2.

$$R^2 = \frac{\sum_{i=1}^{N}\left(y_i^{obs} - y_i^{predicted}\right)^2}{\sum_{i=1}^{N}\left(\bar{y} - y_i^{obs}\right)^2} \quad \text{(B-2)}$$

The third indicator, Akaike Information Criterion (AIC), is an indicator of model parsimony that is, how well the model minimizes MSE with as little complexity as possible. The formula for AIC is shown in Equation B-3 where K is the number of model parameters.

$$AIC = N * \ln\left(\frac{\sum_{i=1}^{N}\left(y_i^{obs} - y_i^{predicted}\right)^2}{N}\right) + 2 * K \quad \text{(B-3)}$$

For a linear regression model, the number of parameters, K, is equal to the number of input variables and interaction terms included in the regression equation. In ANNs, model complexity comes from both the number of input variables and the number of weights and biases so K becomes:

$$K = N_i N_h + N_h N_o + N_h + N_o + 1 \quad \text{(B-4)}$$

Where N_i is the number of input nodes (equal to the number of input variables), N_h is the number of hidden nodes, and N_o is the number of output nodes (equal to the number of output variables).

The fourth indicator, recall, was used as an indicator of how well the model predicts if there will be sufficient FRC at the point of consumption. Recall is defined as the ratio of true positives to the sum of true positives and false negatives as shown in Equation B-5. For this study, water is considered safe if the household FRC concentration is greater than 0.2 mg/L as microbiological recontamination typically does not occur above this concentration of FRC. A true positive was defined as a sample where the measured household FRC concentration was below 0.2 mg/L and the model predicted household FRC below 0.2 mg/L. A false negative was defined as a sample where the measured household FRC concentration was below 0.2 mg/L but the model predicted household FRC above 0.2 mg/L.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad \text{(B-5)}$$

Of the selected metrics, RMSE and R^2 are both common indicators of goodness-of-fit, and AIC is a common metric model design as it quantifies the trade-off between increased model performance and increased model complexity. Recall is a metric primarily intended for classification models (where the output is a distinct class), and is not commonly used for ANN models of FRC decay. Recall was included in this study to evaluate how well the model can discern if the FRC concentration will be above or below the target FRC concentration of 0.2 mg/L in the household and as such is a useful indicator of the reliability of FRC targets generated by the ensemble model.

B.4.4.2 Input Variable Selection

Input variable selection (IVS) was used to select a preferred input variables combination from the full set of candidate variables collected at each site (Table B-2). Selecting a preferred input variable combination through IVS eliminates noise added by low information variables (Snieder, Shakir, and Khan, 2020). IVS also helps to prioritize data collection by identifying the most useful predictors of household FRC. IVS was performed using judgement of water system operators with experience in humanitarian response as well as using quantitative methods. Water system operators identified input variables that are known to be physical determinants of FRC decay that are easily and commonly collected. Two pre-defined input variable combinations were proposed through expert judgement. The first included tapstand FRC and the elapsed time between the tapstand and household FRC measurements. These are the only two water safety variables required by current humanitarian water quality guidelines (Sphere Association, 2018). The second input variable combination defined through expert opinion included tapstand FRC and elapsed time as well as EC and water temperature. EC and water temperature have been shown to influence FRC decay in refugee settlements in South Sudan; EC as an indicator of inorganic FRC demand, and water temperature due to the impact of temperature on reaction rates (Ali et al., 2015). Some drinking water quality guidelines for refugee and IDP settlements recommend monitoring EC and water temperature, meaning that they may already be collected regularly at some sites (MSF, 2010).

Combined network pathway strength analysis (CNPSA) was used in addition to expert judgement as a quantitative IVS method. CNPSA is an ANN-specific IVS technique first proposed by Duncan (2014) and further refined by Snieder et al (2020). The pathway strength, W_{IO}, for an input variable is the dot product of the matrix of weights connecting that variable's input node to each hidden node (W_{IH}) and the matrix of weights from each hidden node to the output layer (W_{HO}), as shown in Equation B-6 (Duncan, 2014).

$$W_{IO} = W_{IH} \cdot W_{HO} \tag{B-6}$$

In the original method proposed by Duncan in 2014, the pathway strength is calculated for each input variable using an ensemble of ANNs. The ensemble quartile range (EQR) for each input variable is then calculated using the distribution of pathway strengths from each ANN in the ensemble. The formula for EQR is shown in Equation B-7 where $Q1$ and $Q3$ are the first and third quartile values of W_{IO}. In the original method proposed by Duncan, all variables with EQR above 0 were selected for inclusion in the model.

$$EQR = \min(|Q1|, |Q3|)/\max(|Q1|, |Q3|) * sign(Q1) * sign(Q3) \tag{B-7}$$

This study made two changes to the Duncan's method. First, instead of using the quartiles to calculate EQR, the 95th and 5th percentiles were used to account for more extreme values of W_{IO}. This range is referred to henceforth as the ensemble weight range (EWR). Second, instead of using the calculated EWR to select the final set of variables, EWR was used to rank input variables and a backwards elimination algorithm was applied to iteratively eliminate the input variable with the lowest EWR. This allowed the model to be retrained without the weakest predictors of household FRC, allowing the model to assign higher weights to the more useful input variables. At each iteration the performance indicators from Section B.4.4.1 were evaluated, and the preferred set of input variables was selected by identifying the iteration with the best performance. The pseudocode for the novel CNPSA is included in Appendix B.

B.4.4.3 Network Architecture Optimization

Optimizing network architecture must strike a balance between complexity and simplicity. Overly complex network architecture can lead to overfitting or a loss of generalization (i.e., the ability to predict accurately when presented new data) but network architecture that is too simple

may lead to underfitting as the model is unable to reproduce the input-output relationships. Two aspects of the network architecture, the division of data into training, validation, and testing sets and the hidden layer size, were simultaneously optimized using a grid search approach described by Khan and Valeo (2018). This process involves training an ensemble of ANNs to test all possible combinations of hidden layer sizes and data divisions within a predefined range. For this study, the percentage of data used for training was varied from 25% to 60% in 5% increments with the remaining data divided evenly between the validation and testing. The hidden layer size was varied from 1 to 30 in increments of 1. For each combination of hidden layer size and data division, the performance indicators from Section 2.4.1 were evaluated, and the preferred network architecture was selected by identifying the iteration with the best performance. The pseudocode for the grid search architecture optimization is included in Appendix C.

B.4.5 Methods for Comparing Local and Global Models

The local and global ensemble models were compared to identify the conditions where each type of model performed best, with the assumption that in a new refugee or IDP settlement a global model would be used until there is sufficient site-specific data to develop a local model. To simulate the performance of the global model at a new site, the global model was trained and validated with all but one site's data and then tested using the remaining site. The performance of the global model was then compared to the performance of the test site's local model. This process was repeated using each of the sites except Bangladesh. Due to the large volume of data available from Bangladesh, both the local and global models were recursively retrained using additional data to represent the evolution of model performance over time and to identify the amount of data required for the local model to outperform the global model. For this test, the training dataset was increased in two-week increments to align with reporting requirements for water system operators in humanitarian settings and the models were tested using the upcoming two-week interval.

B.4.5.1 Developing Risk-Based FRC Targets using Ensemble Models

The local and global ensemble models were used with the Bangladesh dataset to develop risk-based FRC targets for each two-week interval. The ensemble model was retrained in two-week increments as described in Section 2.5 and then this ensemble was used to predict household

FRC for tapstand FRC concentrations ranging from 0.2 mg/L to 2.0 mg/L in 0.05 mg/L increments. For each tapstand FRC concentration, the probability of household FRC below 0.2 mg/L was calculated as the percentage of ANNs in the ensemble that predicted household FRC ≤0.2 mg/L. For example, if 15 of 200 models predicted that household FRC would be below 0.2 mg/L when the tapstand FRC was 0.5 mg/L, then there is a 7.5% probability (p=0.075) that drinking water will have insufficient FRC in the household. For this study, the target FRC for each two-week period was selected as the lowest tapstand FRC concentration where no models predicted that household FRC would be below 0.2 mg/L.

B.5 Results and Discussion

B.5.1 Model Building

B.5.1.1 CNPSA

For both the local and global models, the preferred combination of variables was selected by reviewing both the ensembles performance for the metrics listed in Section B.4.4.1 at each iteration of the backwards elimination process. This is shown for the Jordan (2014) model in Figure B-2 to demonstrate the decision-making process used to select the preferred variable configuration. The bold line in Figure B-2 shows the ensemble median performance while the shaded areas represent the 95^{th} percentile range of performance for each indicator. Figure B-2 shows that the RMSE and R^2 performance on the testing dataset continued to improve up to iteration 22 (after 22 variables had been eliminated). However, the ensemble's median recall performance decreases sharply beyond iteration 20. Thus, iteration 20 was selected as the preferred iteration. At this iteration, the variables included in the model were: tapstand TRC, tapstand FRC, tapstand EC, sun exposure, water use, and using a storage container of type "other" (e.g., cup, bottle, thermos, pot, etc.).

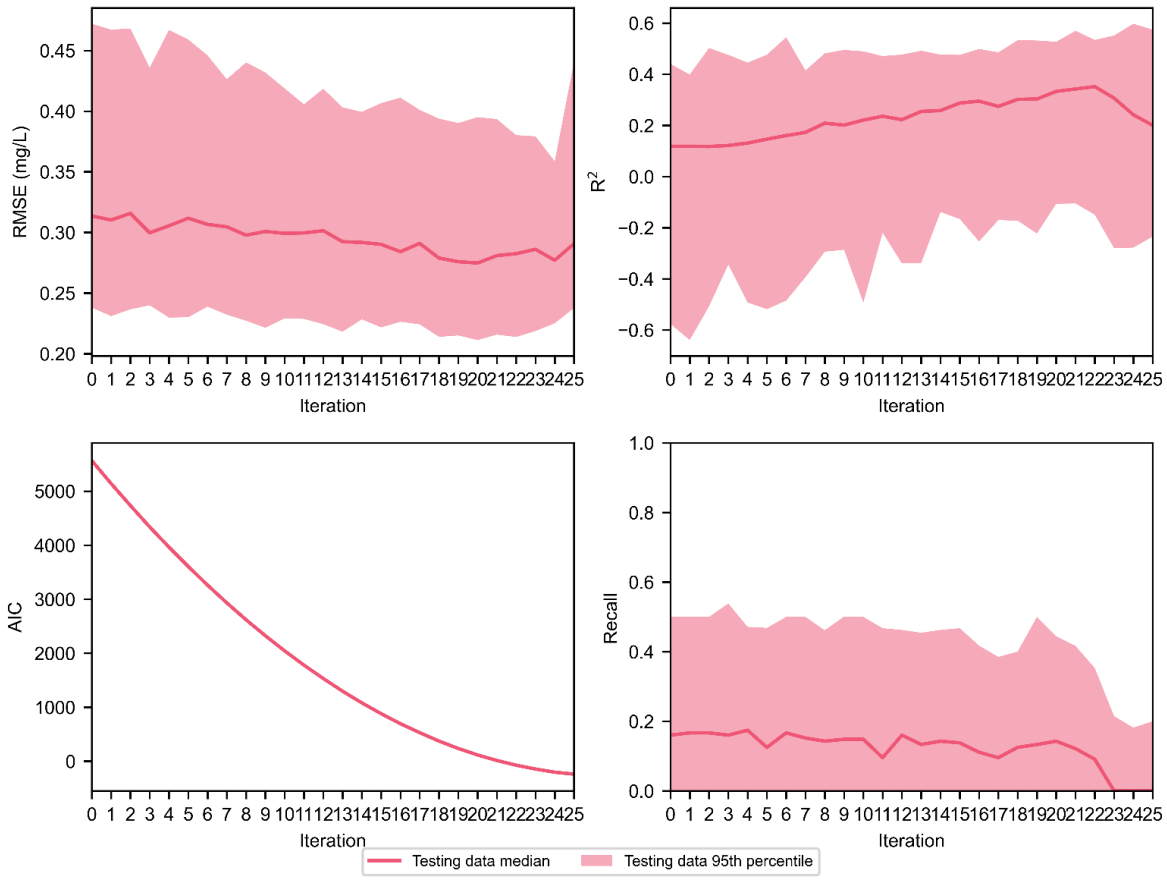

Figure B-2: Jordan (2014) CNPSA performance with each iteration representing an additional variable removed through backwards elimination. From the upper left moving clockwise the plots show RMSE, R^2, recall, and AIC. Iteration 20 was selected as beyond this iteration recall drops off steeply.

Table B-3 lists the variables that were selected at the preferred iteration for each site in order of descending EWR. From Table B-3 it is clear that the optimal variable configuration is site-specific, with no two sites having the same set of selected variables. Even in Jordan, the selected set of variables are different between 2014 and 2015. This may have been due to the difference in weather conditions between the 2014 and 2015 period resulting in different FRC decay behaviour or changes in household conditions and infrastructure due to the population doubling between 2014 and 2015. This indicates that in addition to FRC decay being variable over time, the influence of external factors on FRC decay also change over time.

Table B-3: Summary of variables selected for local and global models using the CNPSA backwards selection algorithm

Site	Variables Selected
Global	Tapstand TRC
Batil	Tapstand TRC, tapstand FRC, storage container type (jerrycan)
Gendrassa	Tapstand FRC, tapstand TRC
Jamam	Tapstand TRC, tapstand FRC, collection container type (oil container)
Jordan (2014)	Tapstand TRC, tapstand FRC, tapstand conductivity, sun exposure, water use, storage container type (other)
Jordan (2015)	Sun exposure, tapstand FRC, mixing of water sources
Rwanda	Tapstand TRC, tapstand FRC, storage container type (jerrycan), drawing method (pouring), drawing method (dipping), collection container type (jerrycan), water transfer, collection period (morning), collection period (overnight), percentage of water remaining, tapstand conductivity, collection container hygiene (unclean)
Bangladesh	Tapstand TRC, tapstand FRC, tapstand pH, drawing method (dipping), air temperature, storage container type (jug), tapstand conductivity, elapsed time, water mixing, collection container covering, tapstand water temperature

Despite the differences shown in Table B-3, there are important commonalities across the models. First, either tapstand FRC or TRC were present in all model configurations, and in six of the eight models, both FRC and TRC were present. This is despite these two variables being highly correlated, with the correlation coefficient between tapstand FRC and TRC ranging between 0.937 and 0.996 depending on the site. This confirms that the most important predictor of household FRC is the upstream residual chlorine concentration, which justifies the inclusion of tapstand FRC in the two expert opinion variable configurations.

Table B-3 also shows that all models except the global model and the Gendrassa model included water handling variables, though these were rarely the variables with the highest EWR. These findings, as well as the findings regarding the impact of chlorination, coheres with findings of investigations into the impact of water handling variables on microbiological contamination, namely that water quality at the water distribution point has a greater impact on water quality at

the point of consumption than water handling behaviour (Meierhofer et al., 2019; Trevett, Carter, and Tyrrel, 2004).

B.5.1.2 Network Architecture Optimization

The grid search procedure outlined in Section 2.4.3 was used to optimize the network architecture for each site with both the CNPSA-selected input variables (Table B-3) and the expert-defined input variable configurations. Figure B-3 shows an example of the change in testing performance with increasing hidden layer size for the Batil model using the CNPSA selected variables with 35% of the data used for training. Figure B-4 shows the change in performance for the same site and input variable configurations, with 25 nodes in the hidden layer, as an increasing proportion of the data is used for training. From Figures B-3 and B-4, the percentage of training data had a much greater effect on model performance, especially the range of model performance, than the size of the hidden layer. Most noticeably, as the percentage of data used for training exceeded 35%, the ensemble range of RMSE and R^2 performance increased, indicating that the models were losing ability to generalize effectively and potentially overfitting. While the hidden layer size had less of an impact on RMSE and R^2, it substantially effects model recall, with ensemble median recall above 0 only occurring consistently with at least 15 hidden nodes. The need for such a large hidden layer indicates that the relationships between the input variables and household FRC are complex and highly non-linear, which has also been identified in studies using ANNs to model FRC in piped distribution systems (Gibbs et al., 2006; Rodriguez & Sérodes, 1998).

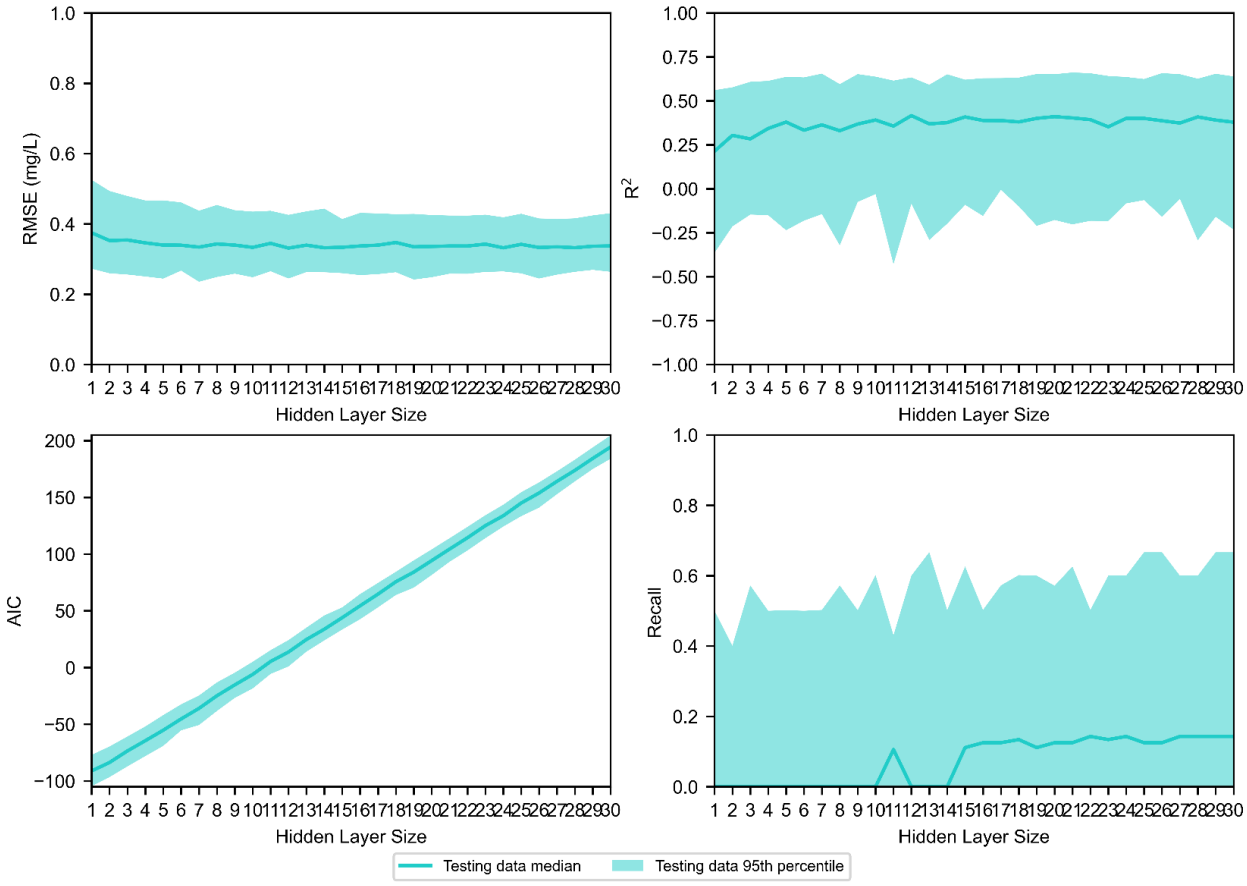

Figure B-3: Batil model performance for different nodes at 35%/32.5%/32.5% data division. From the upper left moving clockwise the plots show RMSE, R^2, recall, and AIC. There is minimal change in RMSE or R^2 performance as the hidden layer size increases, but a minimum of 15 nodes are needed for median recall to increase above 0 consistently.

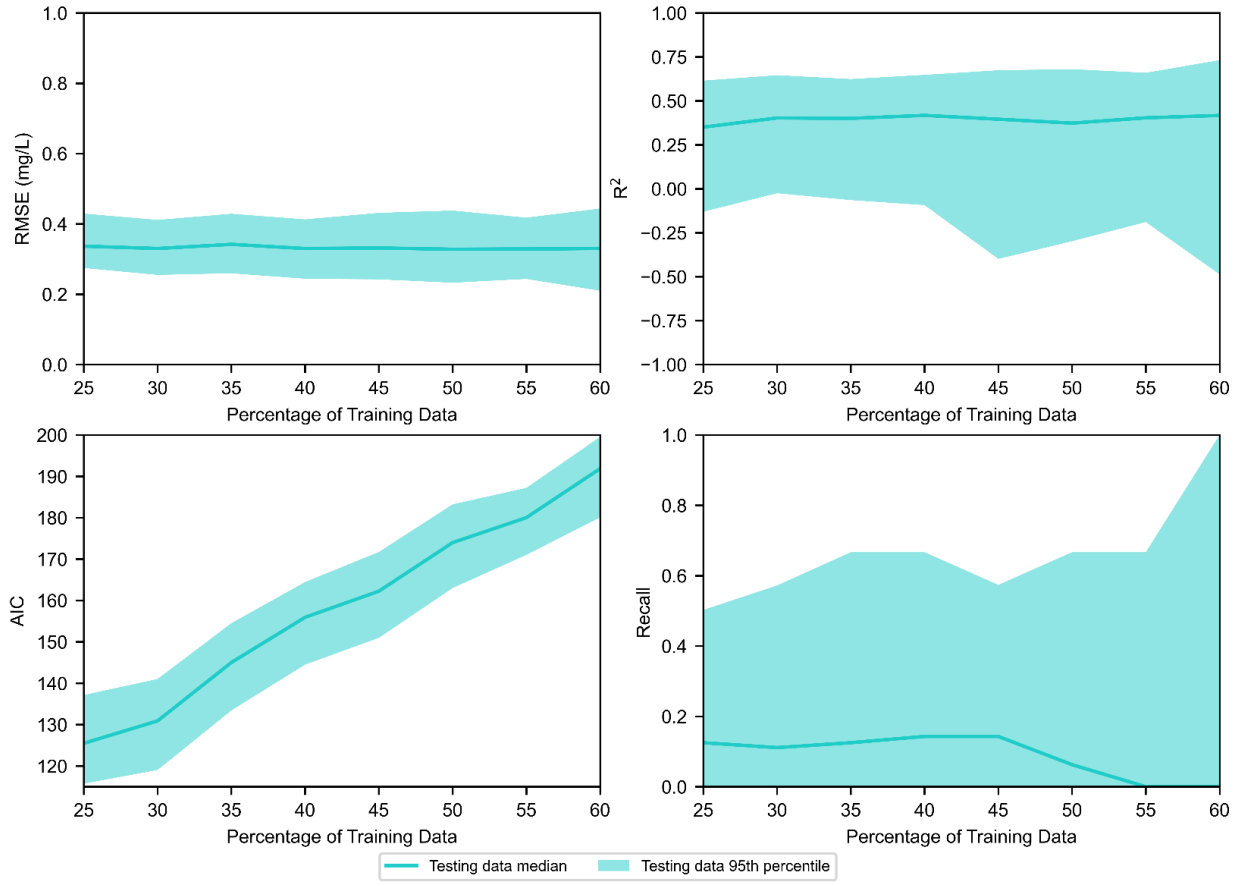

Figure B-4: Batil CNPSA selected model performance for different data division for 25 nodes. From the upper left moving clockwise the plots show RMSE, R^2, recall, and AIC. This figure shows that the range of ensemble performance increases as the training data increases.

Table B-4 summarizes the selected network architecture for each variable combination at each site. While there was no single data division ratio that was selected for all models, 67% (16 of 24) of the models performed best with a training fraction between 25% and 35%, and 88% (21 of 24) performed best with a training fraction between 25% and 50%. This indicates that models of household FRC may be prone to overfitting when using large amounts of data for training. Knowing this, future model development on the SWOT can use a smaller search space for determining the optimal data division. Similarly, there was no consistent pattern in the selected hidden layer size, but 79% (19 of 24) of the models had a hidden layer size between 10 and 30 nodes and 88% (21 of 24) of the models had a hidden layer size between 5 and 30 neurons,

indicating that post-distribution FRC decay is highly non-linear. These findings can also be used to constrain the search space for future model optimization, including the development of new operational models for the SWOT.

Table B-4: Summary of selected network architecture for all sites. The best performing model at each site is shown bolded. The "FRC and Time" variable configuration and the "FRC, Time, Conductivity, Water Temperature" configurations were determined using expert opinion as discussed in Section 2.4.2. The CNPSA selection refers to the variables listed in Table B-3.

Site	Variable Configuration	Data Division (Training/Validation/Testing)	Number of Hidden Nodes
Global	FRC and Time	50/25/25	1
	FRC, Time, Conductivity, Water Temperature	**25/37.5/37.5**	**21**
	CNPSA Selection	45/27.5/27.5	5
Batil	FRC and Time	35/32.5/32.5	26
	FRC, Time, Conductivity, Water Temperature	35/32.5/32.5	30
	CNPSA Selection	**35/32.5/32.5**	**25**
Gendrassa	FRC and Time	25/37.5/37.5	28
	FRC, Time, Conductivity, Water Temperature	30/35/35	18
	CNPSA Selection	**35/32.5/32.5**	**27**
Jamam	FRC and Time	25/37.5/37.5	29
	FRC, Time, Conductivity, Water Temperature	30/35/35	30
	CNPSA Selection	**35/32.5/32.5**	**7**
Jordan (2014)	FRC and Time	25/37.5/37.5	19

Site	Variable Configuration	Data Division (Training/Validation/Testing)	Number of Hidden Nodes
	FRC, Time, Conductivity, Water Temperature	35/32.5/32.5	29
	CNPSA Selection	**25/37.5/37.5**	**22**
Jordan (2015)	FRC and Time	60/20/20	4
	FRC, Time, Conductivity, Water Temperature	60/20/20	25
	CNPSA Selection	**55/22.5/22.5**	**27**
Rwanda	FRC and Time	35/32.5/32.5	29
	FRC, Time, Conductivity, Water Temperature	45/27.5/27.5	12
	CNPSA Selection	**40/30/30**	**22**
Bangladesh	**FRC and Time**	**25/37.5/37.5**	**2**
	FRC, Time, Conductivity, Water Temperature	25/37.5/37.5	10
	CNPSA Selection	45/27.5/27.5	10

B.5.1.3 Comparison of IVS methods

We compared the optimized ensemble model performance for each variable configuration to select a preferred input variable configuration at each site, shown bolded in Table B-4. For all models except the Bangladesh and global models, the optimized models using the input variables identified through CNPSA resulted in the best performance, especially for recall. Figure B-5 shows this for the Jamam models where the most substantial difference in performance between the optimized models was in recall. Both of the models using expert-defined variable configurations have lower ensemble median recall as well as lower recall scores at the quartiles and the 95th percentiles of the ensemble. This shows that the variables included in the model through CNPSA are critical for enabling the ANN models to predict if there will be sufficient

FRC at the point of consumption. One of the challenges of operationalizing these findings is that the CNPSA procedure included many variables which are not commonly collected, such as TRC and water handling variables. This indicates that in order to obtain the best possible model performance, expanding data collection to include additional parameters beyond only those specified in drinking water quality guidelines may be necessary, though, the benefits of expanding data collection should be balanced against the additional effort required in collecting the data. Alternatively, future modelling approaches should investigate methods of obtaining the best performance from ANNs with only those variables that are most commonly collected, such as FRC and elapsed time.

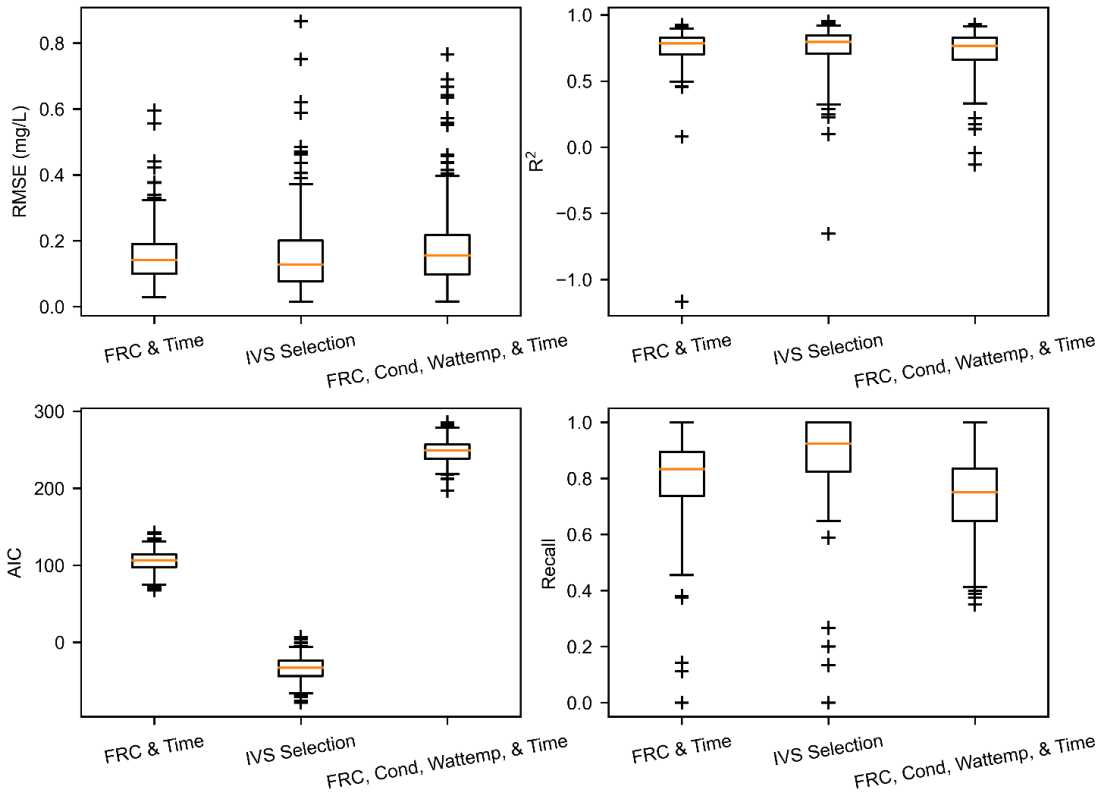

Figure B-5: Comparison of performance between the optimized CNPSA model and the pre-defined variable configurations for Jamam. From the upper left moving clockwise the plots show

RMSE, R^2, recall, and AIC. The orange line on each boxplot shows the median, the boxes show the interquartile range, the whiskers show the 95th percentile range, and the crosses show outliers. This figure shows that all models have similar RMSE and R^2 performance but the model using CNPSA-selected variables had much better recall.

While the CNPSA selected models achieved the best model performance at most sites, the global model performed best using the second pre-defined variable configuration (FRC, elapsed time, conductivity, and water temperature), and the Bangladesh model performed best using only FRC and elapsed time. This may be due to the spatial and temporal variability of FRC decay. In a model that includes data from multiple sites, such as the global model, or a model with data collected over a long period of time, such as the Bangladesh model, there may not be consistent set of ideal input variables, so the variables selected through expert opinion provide better performance as these variables are related to the FRC decay reaction.

B.5.2 Comparison of Local and Global Models

B.5.2.1 Short-Term Collection Sites

The local and global ensemble models were compared to determine if training the models with a large amount of data obtained by combining data from multiple sites produces better performance, or if, due to the spatial variability of FRC decay, better performance is achieved using a smaller amount of local data. Figure B-6 shows the testing RMSE and recall for Batil, Jamam, Gendrassa, Jordan (2014 and 2015), and Rwanda using both the local and global optimized models. The local models typically had better median performance for both metrics, indicating that even a small amount of local data can result in better model performance than a large amount of global data. The global ensemble model had a smaller range of performance, indicating that the predictions of the individual ANNs in the ensemble were more similar. Since the nature of FRC decay is highly site-specific, training the global model with multiple sites may have reduced model diversity as the unique conditions of each individual site were lost, resulting in each individual ANN within the global ensemble converging to a solution representing the average behaviour of all training sites. This shows that post-distribution FRC decay is a highly site-specific phenomenon, so that even models trained on a very small amount of local data can outperform a model trained on large amounts of data from multiple sites. Thus, field

implementation of the SWOT should prioritize local model development to produce site-specific FRC targets.

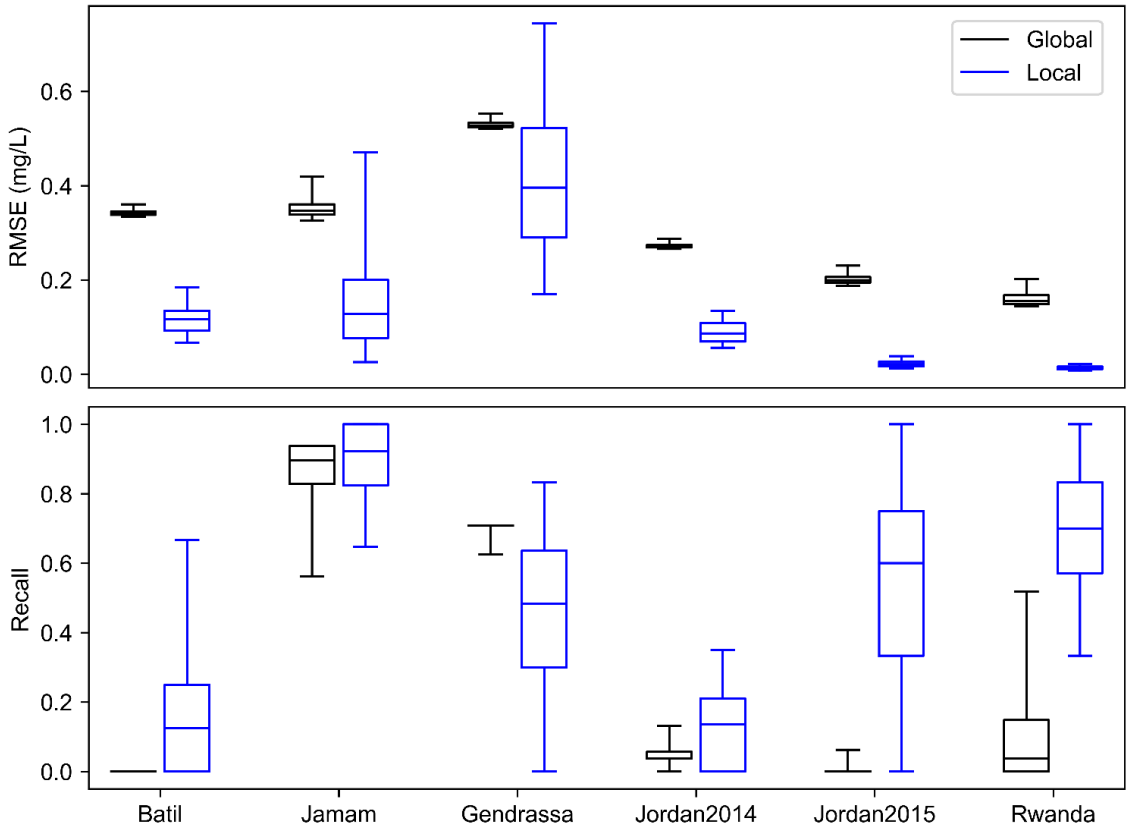

Figure B-6: Comparison of RMSE performance using optimized local and global models showing better median performance for the local models and lower range in performance from the global model.

While all local models outperformed the global model, the local models for sites with larger data volumes (Jordan and Rwanda) performed better than the models with small data volumes (Batil, Gendrassa, and Jamam), indicating that model performance improves with additional data collection. Thus, while local models trained even on a small amount of site-specific data outperform the global model, increased data collection improves model performance so data collection should remain an ongoing process.

B.5.2.2 Bangladesh Local-Global Model Comparison

The optimized global and Bangladesh models (model characteristics shown bolded in Table B-4) were compared using the procedure described in Section B.4.5 to evaluate the evolution of local and global model performance over time and with increasing data. Figure B-7 compares observed household FRC concentration to the range of household FRC predicted by the local and global ensemble models. Both models tended to predict a narrower range of household FRC concentrations than what was observed, though the local model predicted a larger range than the global model. This indicates that while neither model completely reproduced the observed behaviour, the local model reproduced the variability of FRC decay better than the global model. Figure B-8 shows the RMSE and recall respectively for the optimized Bangladesh and global models. From these figures, the Bangladesh local model had similar median RMSE performance to the global model, with no clear point beyond which the local model outperformed the global model. The local model also had a larger range of RMSE scores, perhaps due to the wider range of predictions shown in Figure B-7. The local model consistently achieved higher ensemble median recall and higher 5^{th} and 95^{th} percentile recall scores, indicating that the local ensemble as a whole continued to classify more effectively if water stored in the household would have sufficient FRC. The difference between the recall and RMSE performance is likely linked to the observation from Figure B-7 that the Bangladesh model predicted a wider range of values, and therefore was able capture more observations with lower household FRC. In summary, both the local and global models predicted similarly near the centre of the range of observed values, resulting in similar RMSE, but the broader range of predictions obtained by the local model resulted in better capture of observations with low household FRC. Since the local model had better recall performance than the global model even when the amount of data available for training is small, these results confirm the findings in Section B.5.2.1 that local models with small amounts of data outperform global models. Thus, future model development for the SWOT project should exclusively develop local models.

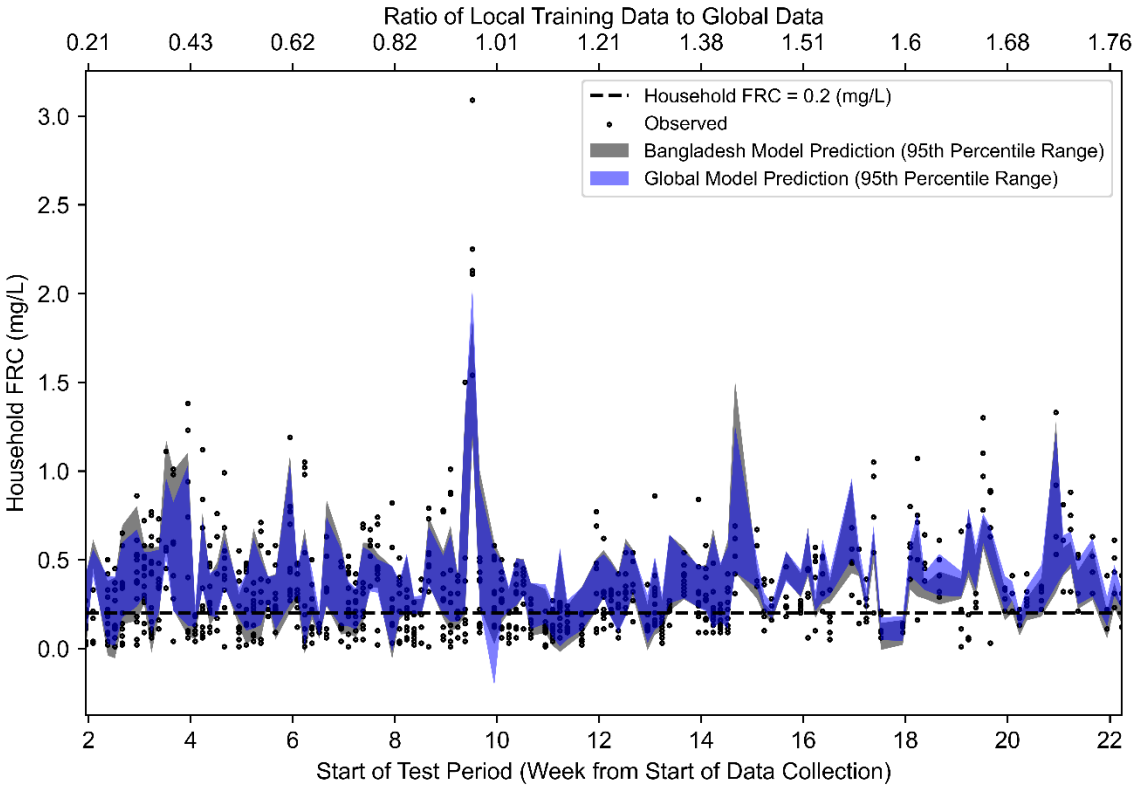

Figure B-7: Comparison of predicted household FRC using both the local and global models compared to observed household FRC showing that both local and global models predicted a narower range of household FRC concentrations than what was observed.

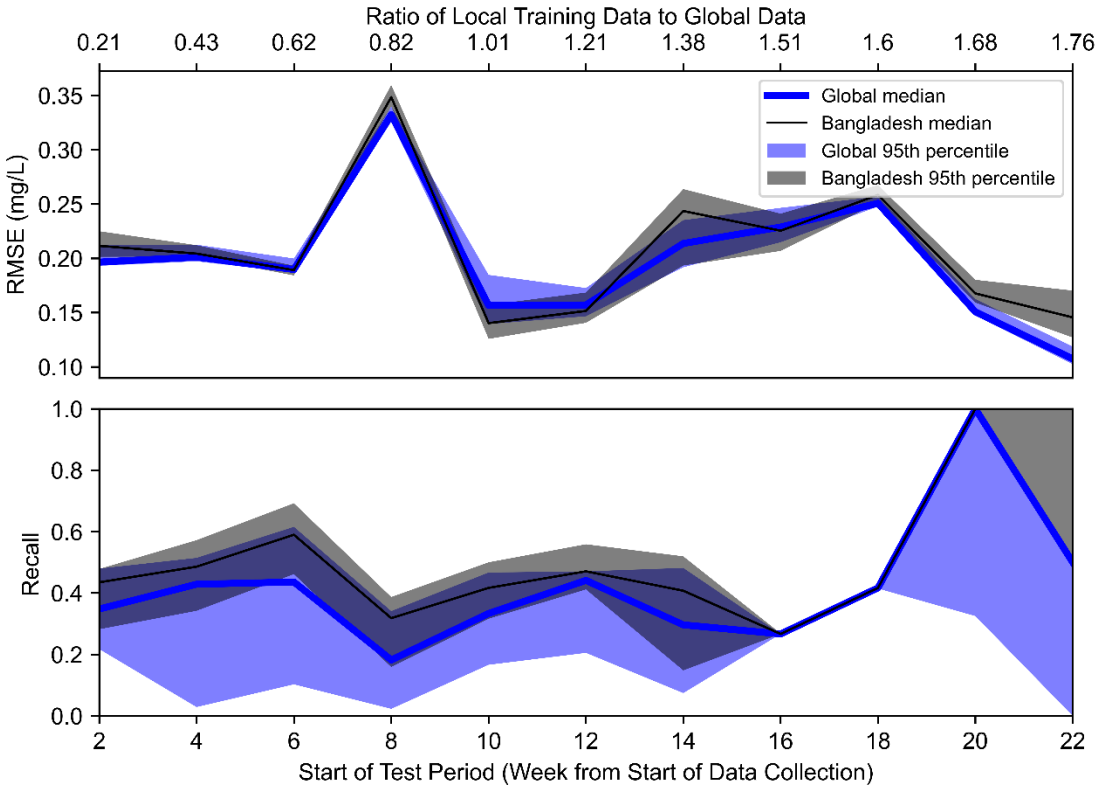

Figure B-8: Testing RMSE and Recall performance for the local and global models using the Bangladesh field trial data, showing that RMSE performance was similar between the local and global models but the local model had consistently equal or better recall throughout.

The local and global models were also used to produce a risk-based FRC target for each two-week period as described in Section 2.5.1. The target was defined as the lowest FRC required at the water distribution point that would ensure that all ANNs in the ensemble would predict household FRC above 0.2 mg/L after 15 hours of storage. This storage time was selected based as 15 hours was the maximum storage time in the Bangladesh refugee settlement (Ali et al., 2020a). Figure B-9 shows the evolution of the targets as the model was retrained. Both the local and global models recommended that FRC at the water distribution point should be above 0.5 mg/L. This confirms that current humanitarian sector drinking water quality guidelines do not

provide sufficient FRC to protect water up to the point of consumption (Ali et al., 2015, Ali et al., 2021). Additionally, the recommendations from both models changed over time, indicating that static recommendations are not appropriate for the dynamic nature of refugee and IDP settlements, highlighting the need for the adaptive targets provided through the SWOT project. Finally, the local model consistently recommended a higher FRC target, likely due to the broader range of predicted household FRC concentrations better reproducing the variability of FRC decay.

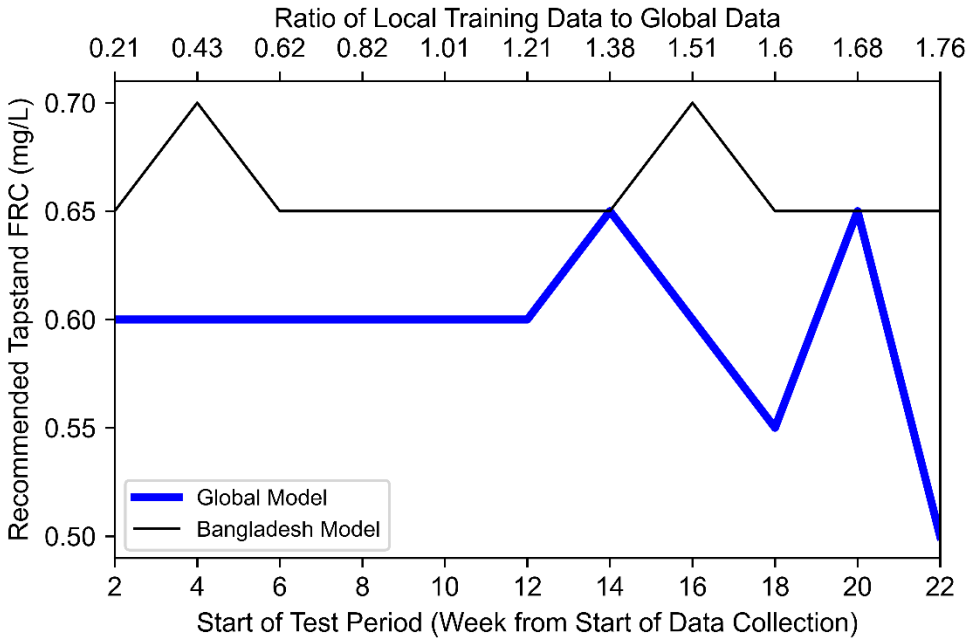

Figure B-9: Target tapstand FRC to ensure that household FRC remains above 0.2 mg/L after 15 hours of storage showing that the both models consistently recommend FRC targets higher than those provided in current drinking water quality guidelines, with the local model providing more conservative recommendations than the global model.

B.5.3 Study Limitations and opportunities for future work

An important limitation of this study is the imbalanced number of samples from each site. This led to the Bangladesh data set being excluded from the global model (except as described in Section 2.5.1) and to an imbalanced contribution from each site in the global model. In future, alternative sampling approaches could be taken to ensure each dataset is represented equally.

Additionally, the local models were developed for individual settlements but FRC decay is specific to the distribution network (Vasconcelos et al., 1997). Future work should determine if settlement specific models are sufficient or if a separate model is needed for each water distribution network within a settlement.

A further limitation of this study arises from the model-based nature of CNPSA. This method identifies useful predictors of FRC, but it does not quantify the statistical effect of different input variables and as such cannot identify if any of these variables have a causal impact on FRC decay. For example, despite FRC decay being a time-dependent reaction, it was only selected through CNPSA for the Bangladesh model. Additionally, the backwards elimination algorithm eliminated variables based on low EWR, which is caused by low or inconsistent W_{IO}. Thus, while the preferred iteration was selected on ensemble performance, the variable eliminated at each iteration was not. Future work should investigate either model-based IVS methods that incorporate model performance, such as input omission, or model-free IVS methods which use statistical relationships between variables to select a preferred set of input variables.

The backwards elimination CNPSA also selected a different number of variables selected for each model. Future work should incorporate sensitivity analysis to determine the importance of each variable's contribution to the model. Sensitivity analysis should also be included in network optimization to identify if unique network architecture is necessary for every model or if the data division and/or hidden layer size can be standardized. This could alleviate the need to optimize network architecture for future site implementation of the SWOT.

The risk-based targets were developed for static values of all variables except for tapstand FRC. This ignores that storage durations may change with water availability and EC and water temperature also fluctuate over time and between water distribution networks. Additionally, using ensemble models to develop probabilistic predictions assumes that the distribution of ensemble predictions matches the output variable distribution (household FRC) (Boucher et al., 2009). If these distributions are not the same, the model may under or over-predict the risk of unsafe drinking water.

B.6 Conclusion

Ensuring that drinking water remains safe up to the point of consumption is critical for protecting public health among the growing number of refugees and IDPs worldwide. This study is a first of its kind application of ANNs for predicting the FRC concentration in drinking water stored in the household. The model building process showed that for sites with smaller data sets, the novel variation of CNPSA developed for this study selected variable combinations which achieved better model performance than the variable combinations defined through expert opinion. The ANN ensemble models developed for this study were used to generate risk-based FRC targets for a refugee settlement in Bangladesh, improving upon the typical discrete targets provided by physical models. This study demonstrated that:

- Post-distribution decay is very site specific, evidenced by each site-specific model having a unique optimal preferred variable combination, data division, and hidden layer size.

- Site-specific models consistently outperformed global models, even for the sites with the smallest data volumes. Site-specific models with larger data volumes also tended to outperform those with small data volumes, highlighting the importance of continued data collection.

- The FRC required at public water distribution points in Bangladesh was shown to be higher than the currently recommended range in the humanitarian drinking water quality guidelines. Additionally, the required FRC changes over time highlighting a key shortcoming of the static FRC recommendation in the current humanitarian drinking water quality guidelines.

These lessons will be useful for the continued ANN implementation on the SWOT project and for future studies of post-distribution FRC decay in refugee and IDP settlements.

B.7 Author Contributions

Michael De Santi: ANN modelling, data analysis, manuscript preparation. Dr. Usman T Khan: manuscript review, modelling supervision. Dr. Syed Imran Ali: data collection (South Sudan, Jordan 2014/2015, Rwanda, Bangladesh), coordination of partners, securing funding, manuscript

review. Jean-François Fesselet: coordination of partners, securing funding, manuscript review. Matthew Arnold: data collection (Bangladesh), coordination of partners, manuscript review.

Appendix B References

Ali, S. I., Ali, S. S., and Fesselet, J.-F. (2015). Effectiveness of emergency water treatment practices in refugee camps in South Sudan. Bulletin of the World Health Organization, 93(8), 550–558. https://doi.org/10.2471/BLT.14.147645

Ali, S. I., Ali, S. S., and Fesselet, J. (2021). Evidence-based chlorination targets for household water safety in humanitarian settings: Recommendations from a multi-site study in refugee camps in South Sudan, Jordan, and Rwanda. Water Research, 189(116642), 1–17. https://doi.org/https://doi.org/10.1016/j.watres.2020.116642

Ali, S. I., Fesselet, J.-F., Arnold, M., Khan, U., Ali, S. S., Spendlove, M., … Orbinski, J. (2020). Development of a machine-learning enabled Safe Water Optimization Tool for humanitarian response. F1000Research, 9(442). https://doi.org/10.7490/F1000RESEARCH.1117910.1

Boucher, M. A., Perreault, L., and Anctil, F. (2009). Tools for the assessment of hydrological ensemble forecasts obtained by neural networks. Journal of Hydroinformatics |, 11(3–4), 297–307. https://doi.org/10.2166/hydro.2009.037

Bowden, G. J., Nixon, J. B., Dandy, G. C., Maier, H. R., and Holmes, M. (2006). Forecasting chlorine residuals in a water distribution system using a general regression neural network. Mathematical and Computer Modelling, 44(5–6), 469–484. https://doi.org/10.1016/j.mcm.2006.01.006

Brown, G. (2011). Ensemble Learning. In C. Sammut and G. I. Webb (Eds.), Encyclopedia of Machine Learning (pp. 312–320). Boston, MA: Springer. https://doi.org/10.1007/978-0-387-30164-8

Cholette, F. (2015). Keras. Retrieved from https://keras.io

Duncan, A. (2014). The Analysis and Application of Artificial Neural Networks for Early Warning Systems in Hydrology and the Environment. University of Exeter. Retrieved from

http://files/78/Duncan_2014_The Analysis and Application of Artificial Neural Networks for Early Warning.pdf

Gibbs, M. S., Morgan, N., Maier, H. R., Dandy, G. C., Nixon, J. B., and Holmes, M. (2006). Investigation into the relationship between chlorine decay and water distribution parameters using data-driven methods. Mathematical and Computer Modelling, 44(5–6), 485–498. https://doi.org/10.1016/j.mcm.2006.01.007

Golicha, Q., Shetty, S., Nasiblov, O., Hussein, A., Wainaina, E., Obonyo, M., … Burton, J. W. (2018). Cholera outbreak in Dadaab Refugee camp, Kenya — November 2015–June 2016. Morbidity and Mortality Weekly Report, 67(34), 958–961. https://doi.org/10.15585/mmwr.mm6734a4

Guerrero-Latorre, L., Hundesa, A., and Girones, R. (2016). Transmission Sources of Waterborne Viruses in South Sudan Refugee Camps. Clean - Soil, Air, Water, 44(7), 775–780. https://doi.org/10.1002/clen.201500358

Howard, C. M., Handzel, T., Hill, V. R., Grytdal, S. P., Blanton, C., Kamili, S., … Teshale, E. (2010). Novel Risk Factors Associated with Hepatitis E Virus Infection in a Large Outbreak in Northern Uganda: Results from a Case-Control Study and Environmental Analysis. American Journal of Tropical Medicine and Hygiene, 83(5), 1170–1173. https://doi.org/10.4269/ajtmh.2010.10-0384

Khan, U. T., He, J., and Valeo, C. (2018). River flood prediction using fuzzy neural Networks: An investigation on automated network architecture. Water Science and Technology, 2017(1), 238–247. https://doi.org/10.2166/wst.2018.107

Médecins Sans Frontières. (2010). Public Health Engineering In Precarious Situations. (J. V. D. Noortgate and P. Maes, Eds.) (2nd ed.). Brussels: Médecins Sans Frontières.

Meierhofer, R., Wietlisbach, B., and Matiko, C. (2019). Influence of container cleanliness, container disinfection with chlorine, and container handling on recontamination of water collected from a water kiosk in a Kenyan slum. Journal of Water and Health, 17(2), 308–317. https://doi.org/10.2166/wh.2019.282

Python Software Foundation. (2019). Python v3.7.4. Python Software Foundation. Retrieved from https://www.python.org/

Rodriguez, M. J., and Sérodes, J. B. (1998). Assessing empirical linear and non-linear modelling of residual chlorine in urban drinking water systems. Environmental Modelling and Software, 14(1), 93–102. https://doi.org/10.1016/S1364-8152(98)00061-9

Rossman, L. A., Clark, R. M., and Grayman, W. M. (1994). Modeling chlorine residuals in drinking-water distribution systems. Journal of Environmental Engineering, 120(4), 803–820. https://doi.org/10.1061/(ASCE)0733-9372(1994)120:4(803)

Shultz, A., Omollo, J. O., Burke, H., Qassim, M., Ochieng, J. B., Weinberg, M., … Breiman, R. F. (2009). Cholera outbreak in Kenyan Refugee Camp: Risk Factors for Illness and Importance of Sanitation. American Journal of Tropical Medicine and Hygiene, 80(4), 640–645. https://doi.org/10.4269/ajtmh.2009.80.640

Snieder, E., Shakir, R., and Khan, U. T. (2020). A comprehensive comparison of four input variable selection methods for artificial neural network flow forecasting models. Journal of Hydrology, 583. https://doi.org/10.1016/j.jhydrol.2019.124299

Solomatine, D. P., and Ostfeld, A. (2008). Data-driven modelling: Some past experiences and new approaches. Journal of Hydroinformatics, 10(1), 3–22. https://doi.org/10.2166/hydro.2008.015

Soyupak, S., Kilic, H., Karadirek, I. E., and Muhammetoglu, H. (2011). On the usage of artificial neural networks in chlorine control applications for water distribution networks with high quality water. Journal of Water Supply: Research and Technology - AQUA, 60(1), 51–60. https://doi.org/10.2166/aqua.2011.086

Sphere Association. (2018). The Sphere Handbook: Humanitarian Charter and Minimum Standards in Humanitarian Response (4th ed.). Geneva: Practical Action Publishing. Retrieved from www.spherestandards.org/handbook

Steele, A., Clarke, B., and Watkins, O. (2008). Impact of jerry can disinfection in a camp environment - Experiences in an IDP camp in Northern Uganda. Journal of Water and

Health, 6(4), 559–564. https://doi.org/10.2166/wh.2008.072

Swerdlow, D. L., Malenga, G., Begkoyian, G., Nyangulu, D., Toole, M., Waldman, R. J., … Tauxe, R. V. (1997). Epidemic cholera among refugees in Malawi, Africa: treatment and transmission. Epidemiology and Infection, 118(3), 207–214. https://doi.org/https://doi.org/10.1017/S0950268896007352

Trevett, A. F., Carter, R. C., and Tyrrel, S. F. (2004). Water quality deterioration: a study of household drinking water quality in rural Honduras. International Journal of Environmental Health Research, 14(4), 273–283. https://doi.org/10.1080/09603120410001725612

Vasconcelos, J. J., Rossman, L. A., Grayman, W. M., Boulos, P. F., and Clark, R. M. (1997). Kinetics of chlorine decay. Journal of the American Water Works Association, 89(7), 54–65. https://doi.org/10.1002/j.1551-8833.1997.tb08259.x

Walden, V. M., Lamond, E. A., and Field, S. A. (2005). Container contamination as a possible source of a diarrhoea outbreak in Abou Shouk camp, Darfur province, Sudan. Disasters, 29(3), 213–221. https://doi.org/10.1111/j.0361-3666.2005.00287.x

Appendix B-1 Data Cleaning Rules

Several data cleaning steps were employed both to avoid including data points with high measurement error and to ensure functionality of the ANN code.

First, any personal identifiers in the data were removed (data collector's name, GPS coordinates of households, etc.)

Second, the timestep measurements were converted into measurements of elapsed time

Third, any measurements where the FRC increased from the water distribution point to the point of consumption by more than 0.06 mg/L were removed. The 0.06 mg/L threshold was selected as the measurement error of the chlorometer used was 0.03 mg/L so this allows for no FRC decay to occur and both measurements to be off by the maximum measurement error

Fourth all categorical variables (hygiene, container type, covering, drawing method, etc.) were converted into binaries. So instead of having a variable *container type* with factors *collapsible*

jerrycan, jerrycan, and *oil container*, three separate variables would be created (*container type_collapsible_jerrycan, container type_jerrycan, container type_oil container*) as the ANN can only accept numerical data.

Additionally, within each ANN code, if any records were missing a measurement for any one of the variables, the entire row was removed.

Appendix B-2 CNPSA Backwards Selection Pseudocode

Import all candidate input variables as matrix X and let Tapstand FRC be vector Y	
Scale all variables between (-1,1)	
for j=1:n	Where *n* is the number of candidate input variables
Initialize the neural network	
for i=1:200	200 is the ensemble sized used to evaluate model performance
Divide data into training, validation, and testing subsets	Data division occurs within this loop to ensure all ensemble members are trained on different data to ensure ensemble diversity
Randomize weights and biases	Weights and biases randomized so that training begins at a different location to ensure ensemble diversity (multi-start search)
Train network with training subset, using early stopping with validation subset with a patience of 10 epochs	

Evaluate performance metrics on the testing subset (RMSE, R^2, Recall, AIC)	
Calculate pathway strength for all variables	Take the dot product of the input layer weights for each input variable and the output weights
end	
Calculate EWR	
Rank variables by EWR	
Eliminate variable with the lowest EWR	
end	

Appendix B-3 Grid-Search Optimization Pseudocode for Network Optimization

Import the predictor variables as matrix X and let Tapstand FRC be vector Y	X can be determined through expert opinion or through CNPSA as shown in Appendix B
Scale all variables between (-1,1)	
for n=1:30	Where *n* is the number of hidden nodes ranging between 1 and 30
Initialize the neural network with an input layer size equal to the number of input variables, a hidden layer size equal to n *and an output layer size of 1*	
for d=25:60 (increments of 5)	Where *d* is the percentage of data used for training

for i=1:200	Use an ensemble size of 200 individual ANNs
Divide data into training, validation, and testing subsets where the training subset uses d *percent of the total data and the validation and testing sets use (100-d)/2 percent of the data*	Data division occurs within this loop to ensure all ensemble members are trained on different data to ensure ensemble diversity
Randomize weights and biases	Weights and biases randomized so that training begins at a different location to ensure ensemble diversity (multi-start search)
Train network with training subset, using early stopping with validation subset with a patience of 10 epochs	
Evaluate performance metrics on the testing subset (RMSE, R^2, Recall, AIC	
end	Take the dot product of the input layer weights for each input variable and the output weights
end	
end	

Appendix C. SWOT-ANN v2 Analytics White Paper

C.1 Executive Summary

This report presents a summary of the Safe Water Optimization Tool artificial neural network version 2 (SWOT-ANN v2) analytics and seeks to provide transparency into how recent research findings were operationalized into these analytics. The SWOT-ANN v2 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 v2 analytics use ANNs, a type of data-driven model, to avoid making assumptions about the underlying 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 v2 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 v2 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 v2 models also use electrical conductivity and water temperature as predictor variables. The inclusion of these additional variables has been demonstrated to improve model performance, increase accuracy of risk-based FRC targets, and enables more in-depth FRC guidance through analysis of different water quality scenarios.

The SWOT-ANN v2 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, as compared 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 SWOT-ANN v2 analytics also introduce a scenario analysis feature which generates multiple risk-based FRC targets based on different 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, providing them a better understanding of decay behaviours on site and allowing them to tailor their risk-based FRC targets to site conditions.

This white paper is intended as a living document, with updates introduced as further refinement of the SWOT-ANN v2 analytics are applied. We also include appendices that summarize key analyses we performed to select the modelling parameters included in the analytics, and a functions glossary to summarize the functions include in the SWOT-ANN v2 code, which is available on GitHub here: https://github.com/safeh2o/swot-python-analysis

C.2 Introduction

This white paper presents the details of the second version of the Safe Water Optimization Tool artificial neural network (SWOT-ANN v2) analytics. The SWOT-ANN v2 analytics use a probabilistic, data-driven, modelling 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 artificial neural networks (ANNs), 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 means that for a single set of conditions at the tapstand, a wide range of household FRC concentrations are possible, making point predictions of household FRC insufficient. To overcome this challenge, the SWOT-ANN v2 used a probabilistic ensemble modelling approach by grouping the predictions of multiple individual ANNs into a probabilistic ensemble forecast. This forecast quantifies the uncertainty in the predicted household FRC concentration and provides information about the distribution of household FRC concentrations. The SWOT-ANN v2 analytics use 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 analytics used to generate this probabilistic FRC guidance to provide transparency into the analytical approach taken in the SWOT-ANN v2 analytics. The full code for these analytics is available on the SWOT project GitHub page at: https://github.com/safeh2o/swot-python-analysis.

Section 1 of this report provides the introduction to the SWOT-ANN v2 and the motivation for this white paper. Section C.4 provides a high-level summary of the SWOT-ANN v2 analytics. Sections C.5 and C.6 provides a summary of the backend tasks included in the SWOT-ANN v2, with Section C.5 covering the importing of the data as well as data pre-processing tasks and input variable selection and Section C.6 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 C.7 summarizes how the SWOT-ANN v2 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 v2 code and appendices summarizing some of the key decisions that went into operationalizing the SWOT-ANN v2 analytics.

C.3 Workflow of the SWOT-ANN v2 Analytics

Figure C-1 provides a high-level workflow of the process used to generate risk-based FRC guidance using the SWOT-ANN v2 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 v2 analytics are able to run. This step also includes selecting the appropriate input variable combination. After this, ANN models are set up and the data is used to train each individual ANN 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 improves 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 C-1: High-level modelling workflow of the SWOT-ANN v2 analytics

C.4 Importing 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 v2 analytics import the data as a .csv file and perform the following tasks:

1. The column names are used to identify the input and output variable columns (tapstand and household FRC, tapstand and household timestamps, tapstand water temperature, tapstand electrical conductivity [EC]).

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 C.4.1)

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.

C.4.1 Input Variable Selection

Input variable selection for the SWOT-ANN v2 analytics is not predefined and instead is a dynamic process with the input variables selection occurring within the analytics. The list of candidate input variables was selected based on the findings De Santi et al. (2021) who found that tapstand FRC, water quality, and EC are all strong predictors of household FRC. While the De Santi et al. (2021) study did not find elapsed time to be a strong predictor of household FRC, both elapsed time and time of collection were included based on the findings of an investigation into the effectiveness of time-based-variables for forecasting household FRC. presented in Appendix C-1. 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 & Sivaganesan, 2002; Fisher et al., 2017; J. C. 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):** 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 (WHO, 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 v2 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. 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 included despite these variables not being included in routine monitoring. We based this on an analysis included in Appendix C-2 that showed that the ANN ensemble models used in the SWOT-ANN v2 analytics tend to perform best when more water quality variables were included, even when a large percentage of records for those water quality measurements were missing. For the full details of this investigation, refer to Appendix C-2.

C.5 Training the ANN ensemble models for the SWOT-ANN v2 analytics

C.5.1 Model Set-Up and Architecture

The model architecture used in the SWOT-ANN v2 analytics is an ensemble of 200 ANNs. The individual ANNs in the ensemble are referred to as the base learners. The ensemble base learners used for the SWOT-ANN v2 analytics 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 C-2. 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 & 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. The base learners in the SWOT-ANN v2 analytics use a hyperbolic tangent activation function on the hidden layer and a linear activation function on the output layer.

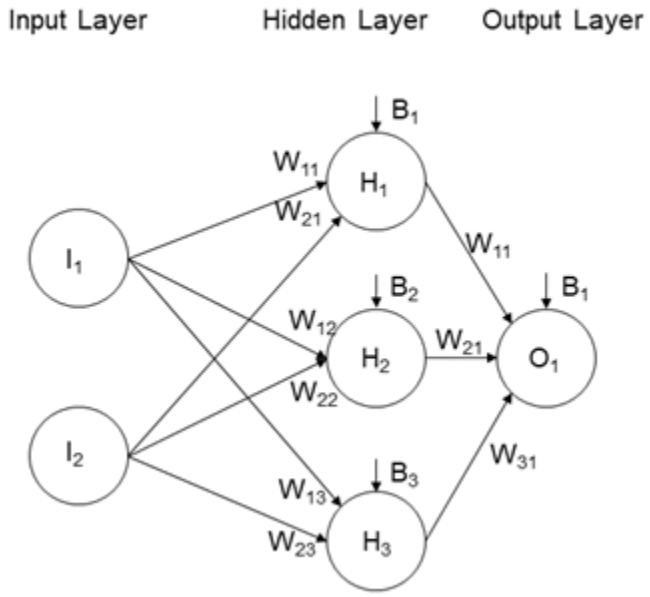

Figure C-2: Schematic of an MLP showing flow of data from the input layer to the output layer with weights and biases. The shown MLP with two input nodes and one output node would have two input variables (other water quality parameters, etc.) and one output (household FRC).

The MLP base learners used in the SWOT-ANN v2 analytics have one 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 v2 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 the SWOT-ANN v1 analytics where the model architecture was predefined with models having three input nodes, five hidden nodes, and one

output node, but the change is necessary to facilitate a flexible approach to input variable selection.

C.5.2 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 v2 analytics 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 model 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.

C.5.3 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 v2 analytics then evaluate the probabilistic performance of these forecasts using four performance metrics: 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 v2 analytics 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 v2 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.

C.5.3.1 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 v2 analytics calculates 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 ($PC_{<0.2}$).

C.5.3.2 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 reliability, where reliability is defined as the similarity between the distribution of the forecast and the underlying distribution of the data. The CI reliability diagram for an 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 C-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 v2 analytics evaluate 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 ($CI_{score_{<0.2}}$).

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

C.5.3.3 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 C-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 & Talagrand, 2005). The SWOT-ANN v2 analytics 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 ($\delta_{<0.2}$).

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

The two components of the δ score are shown in Equations C-3 and C-4 where M is the total number of ensemble members (200), 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 & Talagrand, 2005).

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

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

C.5.3.4 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 C-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 C-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.

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

When using post-processed forecasts (Section C.5.5) the SWOT-ANN v2 analytics calculate CRPS directly using Equation C-5 and take the mean over all forecast-observation pairings. For the raw ensemble, the SWOT-ANN v2 analytics use 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:

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

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

C.5.4 Model Performance Summary and Outputs

Figure C-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. Specifically it shows how the CRPS (left) is calculated from the difference between the observed and forecast CDF, how the RH (centre) is constructed from the rank of each observation relative to the sorted ensemble predictions, and how the Percent Capture at each CI level is used to create the CI reliability diagram (right).

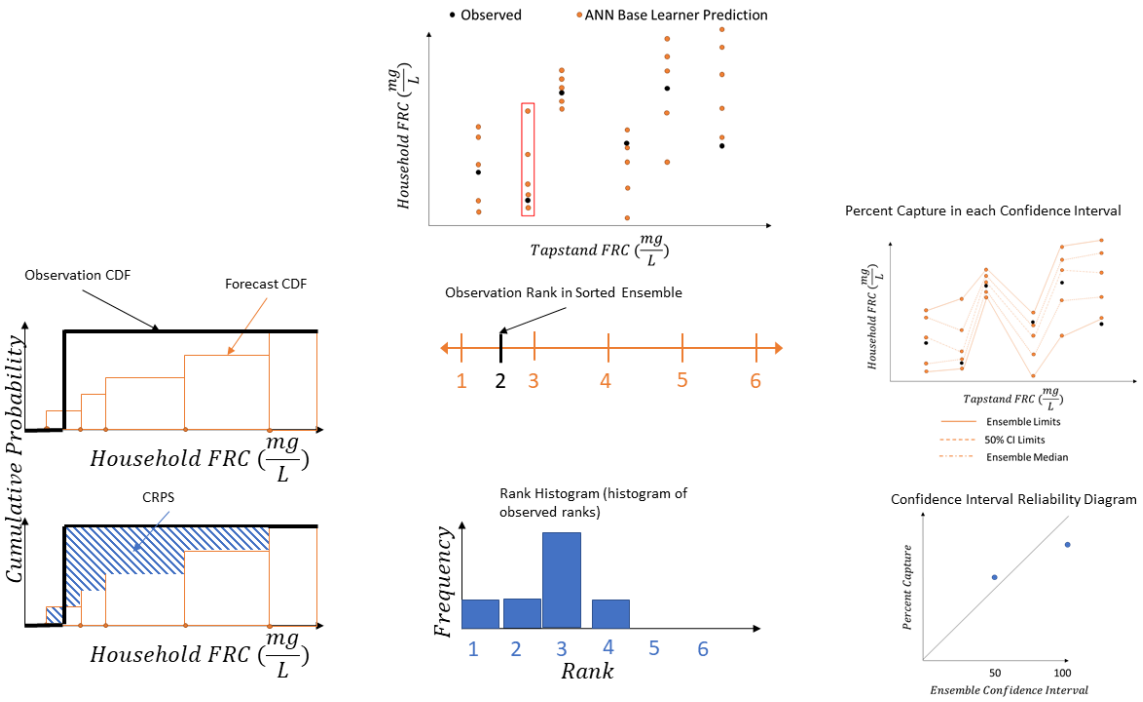

Figure C-3: Interrelation of model performance metrics and visual intuition behind their derivation for the CRPS (left), RH (centre) and CI reliability diagram (right).

Figure C-4 presents the calibration diagnostic figures included in the SWOT-ANN v2 analytics 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 C-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 C-4b would have all points falling on the 1:1 line and Figure C-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 C-4b where the points generally fall below the 1:1 line, and in Figure C-4c, where the RH is U-shaped. Figure C-4c also shows that in this case, the model tends to slightly overpredict the point-of-consumption FRC, as the leftmost bar in the RH is taller than the rightmost bar, indicating more of the uncaptured observations fell below the ensemble forecast range (i.e., there were more low-outliers than high-

outliers). Both of these are further indicators of underdispersion, and while this indicates a performance challenge for the SWOT-ANN v2 analytics, it is worth noting that underdispersion is common in ensembles, especially ensembles of ANNs (Boucher et al., 2015. De Santi et al, 2021). The SWOT-ANN v2 analytics address these challenges using both forecast post-processing (Section C.6.5) and scenario analysis (Section C.7.1)

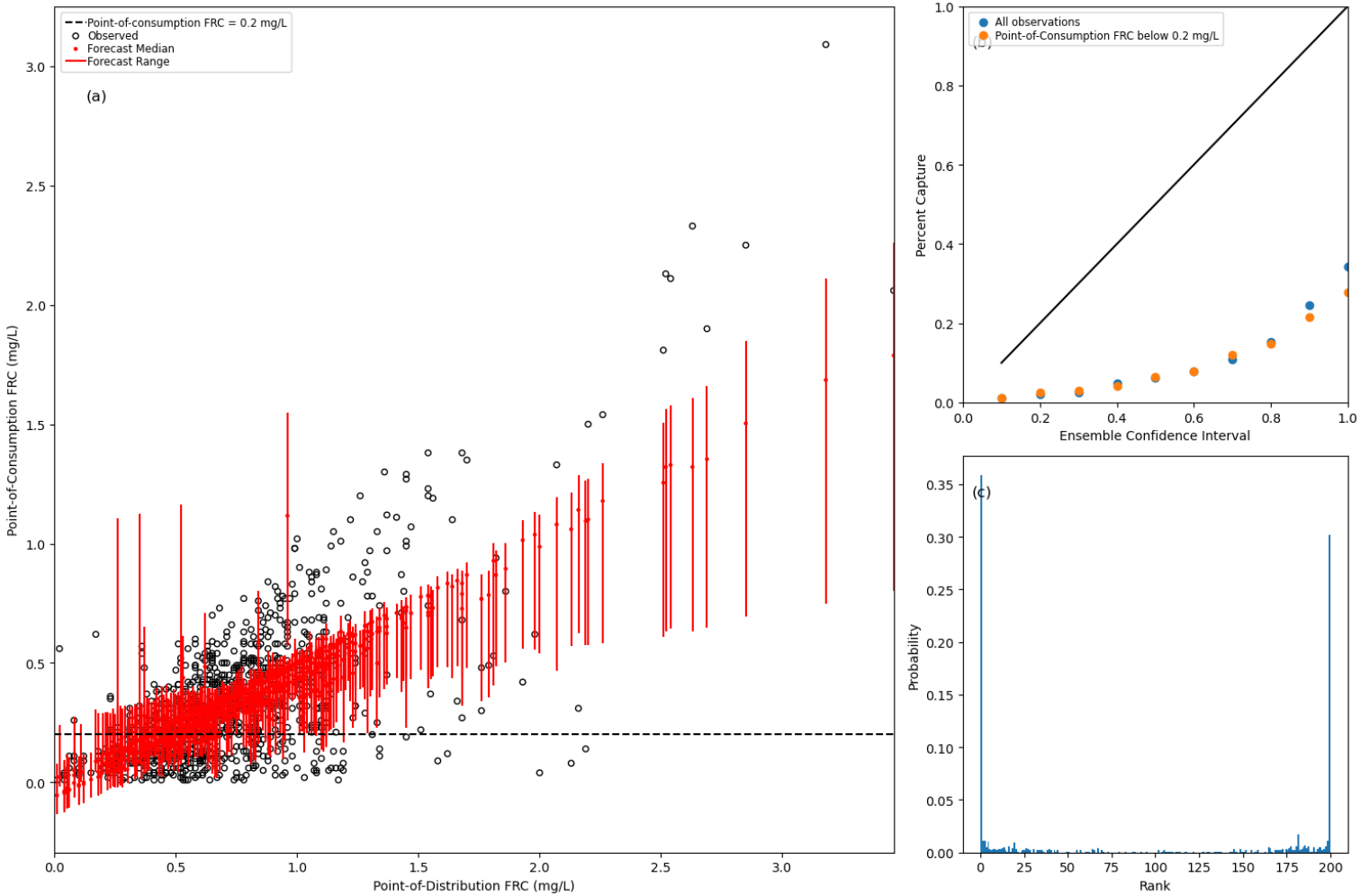

Figure C-4: Sample performance evaluation figures, showing underdispersed forecasts, as can be identified via the numerous outlying observations, as well as the points in the CI reliability diagram and the u-shaped RH.

C.5.5 Post-Processing

In order to generate effective risk-based FRC targets using the SWOT-ANN v2 analytics, 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 v2 analytics use 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 v2 analytics use a Gaussian kernel function in keeping with past studies (Boucher et al., 2011, 2015; Bröcker & Smith, 2008; Roulston & Smith, 2003), though the selection of the specific kernel function is not critical (Boucher et al., 2015). Kernel dressing is implemented 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). The Wang and Bishop approach was selected by comparing the ensemble forecasting performance of three different bandwidth determination methods which showed that, for post-distribution FRC data, the Wang and Bishop (2005) method performed best. The full comparison is included in Appendix C-3. The bandwidth for the kernels is calculated using Equation C-7.

$$\sigma_{\kappa_{WB}}^2 = \overline{(\overline{x_i} - y_i)^2} - \left(1 + \frac{1}{N}\right) * \overline{s_{x_i}^2} \tag{C-7}$$

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.
- $\overline{(\overline{x_i} - y_i)^2}$ 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.

- $\overline{s_{x_i}^2}$ 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 v2 analytics evaluate 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). This comparison is performed using a skill score calculation which normalizes the change in performance from a reference baseline between negative infinity and 1 (Equation C-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.

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

After calculating the skill score for all performance metrics, the SWOT-ANN v2 analytics 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, this indicates that post-processing yields a net performance improvement and post-processing is used when generating risk-based FRC targets. If the sum of the skill scores is equal to or less than 0, the raw ensemble is used.

C.6 Obtaining a Tapstand FRC Target

The SWOT-ANN v2 analytics generate risk-based FRC targets 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 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 v2 analytics implement a scenario analysis approach which considers different time of collection and water quality scenarios to produce a series of risk-based FRC targets.

C.6.1 Scenario Analysis

The SWOT-ANN v2 analytics use 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, targets are always produced for both of these scenarios. If no additional water quality variables (EC, water temperature) are included, then these are the only two scenarios considered by the SWOT-ANN v2 analytics. If at least one of the two additional water quality variables is included in the model, the SWOT-ANN v2 analytics also generate 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 95^{th} percentile EC and/or water temperature values. The selection of the 95^{th} 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 could increase the rate of FRC decay, and higher EC may be 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 v2 analytics consider the decay scenarios in conjunction with the time of collection scenarios. Thus, if either EC or water temperature are included in the model, targets are produced 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.

Currently the SWOT-ANN v2 analytics only include these four scenarios based on time-of-collection, water temperature, and EC. However, this scenario analysis approach can also be expanded to incorporate other water quality or environmental paramters into the decay conditions (pH, turbidity, air temperature) if they are routinely available. Additionally, new scenarios can be considered to incorporate additional water handling factors (container covering, drawing method, etc.) if they are part of routine monitoring.

C.6.2 Outputs and Interpretation

When generating the risk-based FRC targets, the SWOT-ANN v2 analytics produce the following outputs:

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

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

C.6.2.1 Interpreting Output 1: Forecast Plots

Figure C-5 shows the forecasted household FRC generated by the SWOT-ANN v2 analytics for a refugee settlement in Bangladesh using data collected between June through December 2019 where both EC and water temperature were included in the dataset. The four subplots in Figure C-5 correspond to the four scenarios identified in Section 7.1. Figure C-6 shows the FRC forecasts generated by the SWOT-ANN v2 analytics 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 reasonably follow any visually apparent trends in the data. 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 C-5 and C-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 generally follows the apparent trends in the observed data. 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 C-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 C-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.

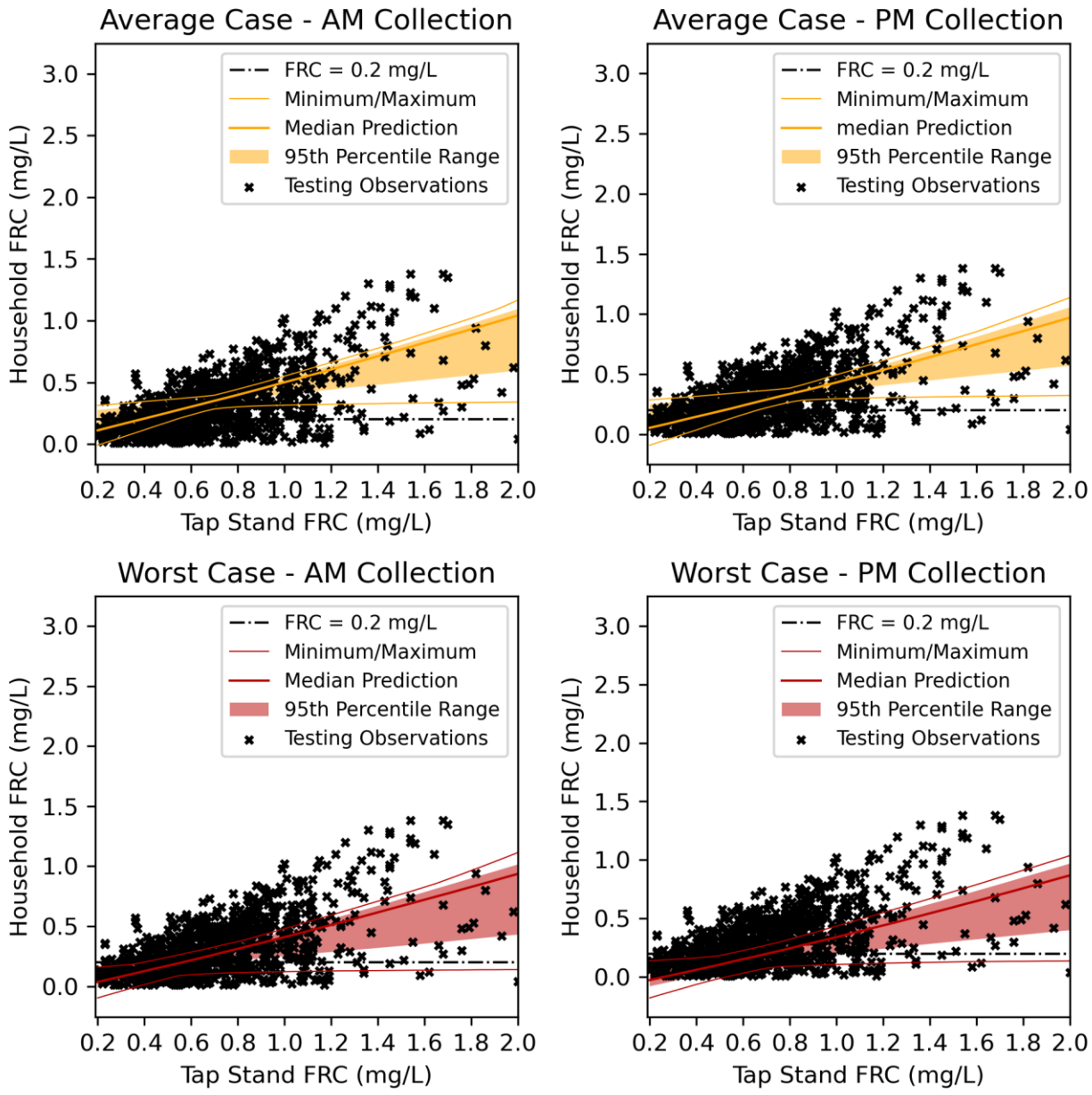

Figure C-5: Sample SWOT-ANN v2 analytics output for dataset with additional water quality variables included (EC, water temperature)

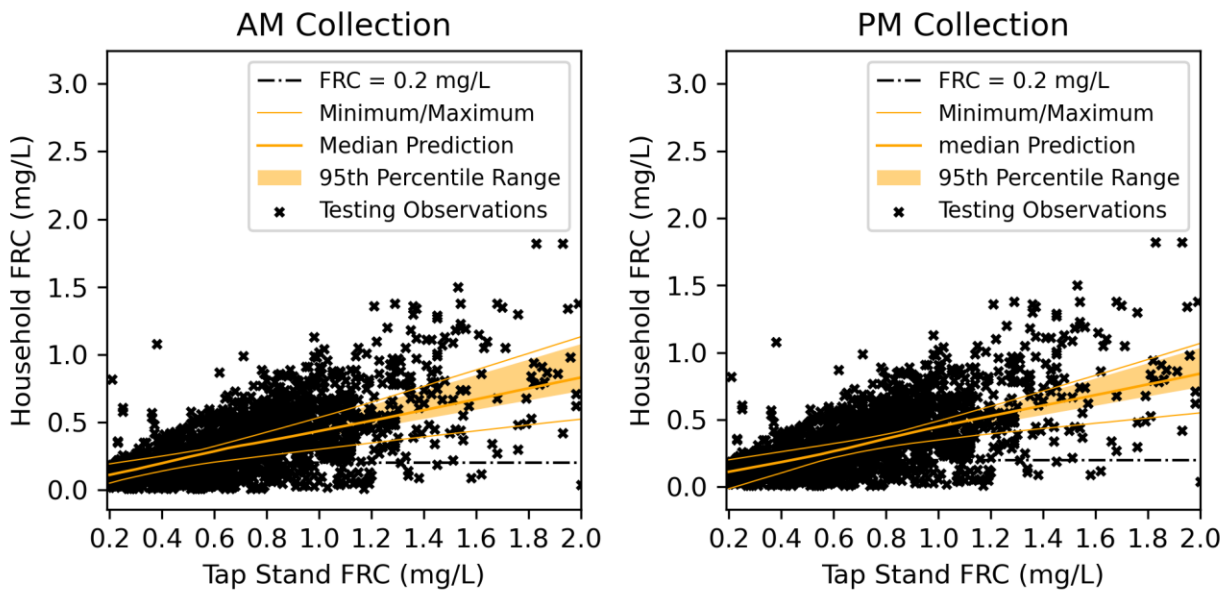

Figure C-6: Sample SWOT-ANN v2 analytics output for dataset without additional water quality variables (EC, water temperature)

Figure C-7 shows a special case from an IDP settlement in Nigeria where there was both limited and sparse data. The dataset used to generate Figure C-7 was obtained early in a field trial so the volume of data is low. Additionally, FRC was measured in this settlement using 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.

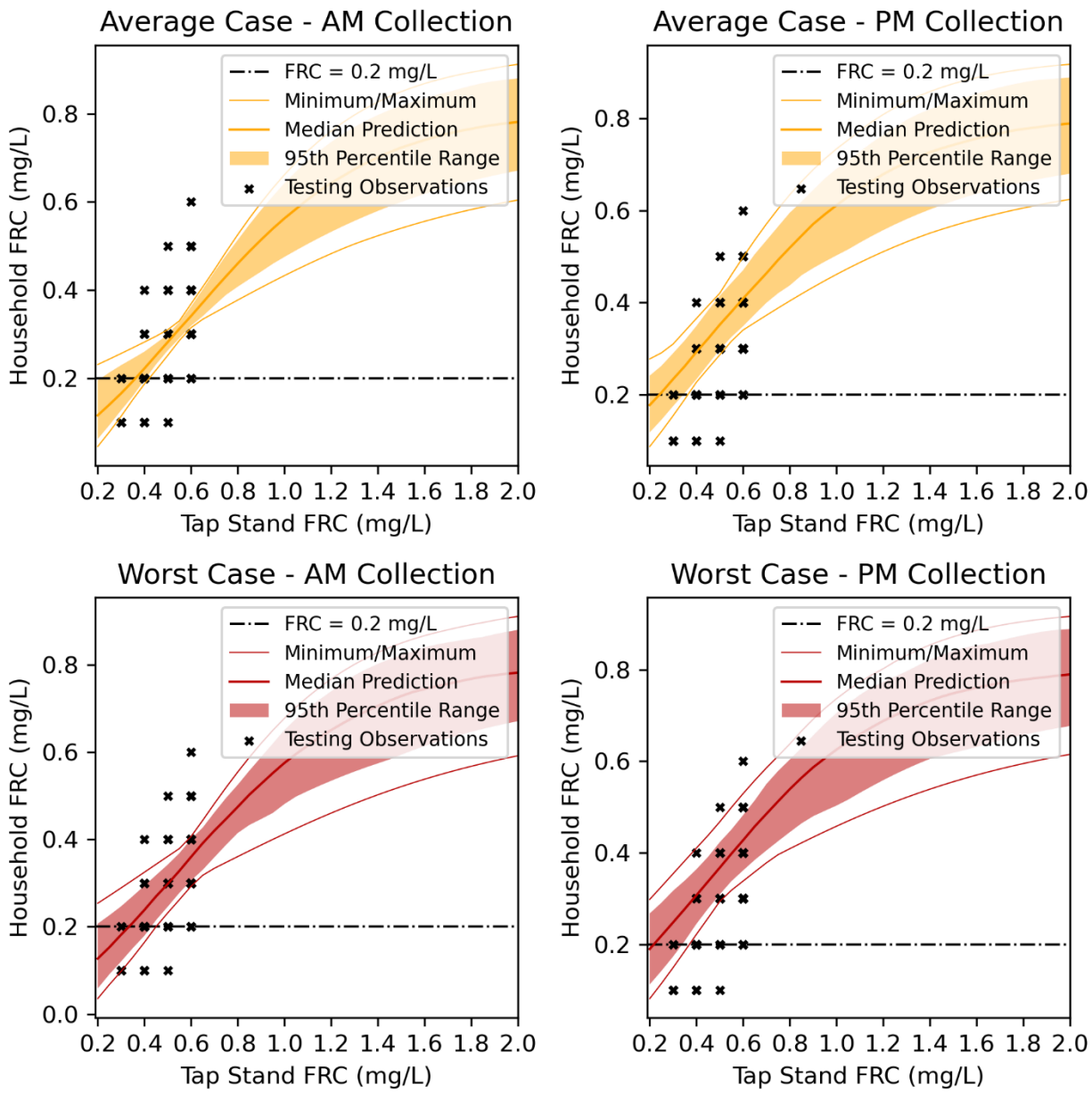

Figure C-7: Predictions with sparse data. Note poor coverage of unsafe values but required tapstand FRC can be easily identified visually.

C.6.2.2 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 produced across the multiple

scenarios, 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. However, if this histogram is used to select a target for a specific collection time, it is also important to ensure that the most frequent time for collecting tapstand water quality samples also matches actual water handling behaviour. 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 C-8 shows the input variable histogram for Bangladesh dataset used to generate Figure C-5. 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 tapstand sampling for AM collection is more common, though again, it is important to determine if this sampling pattern reflects actual water usage behaviour on site.

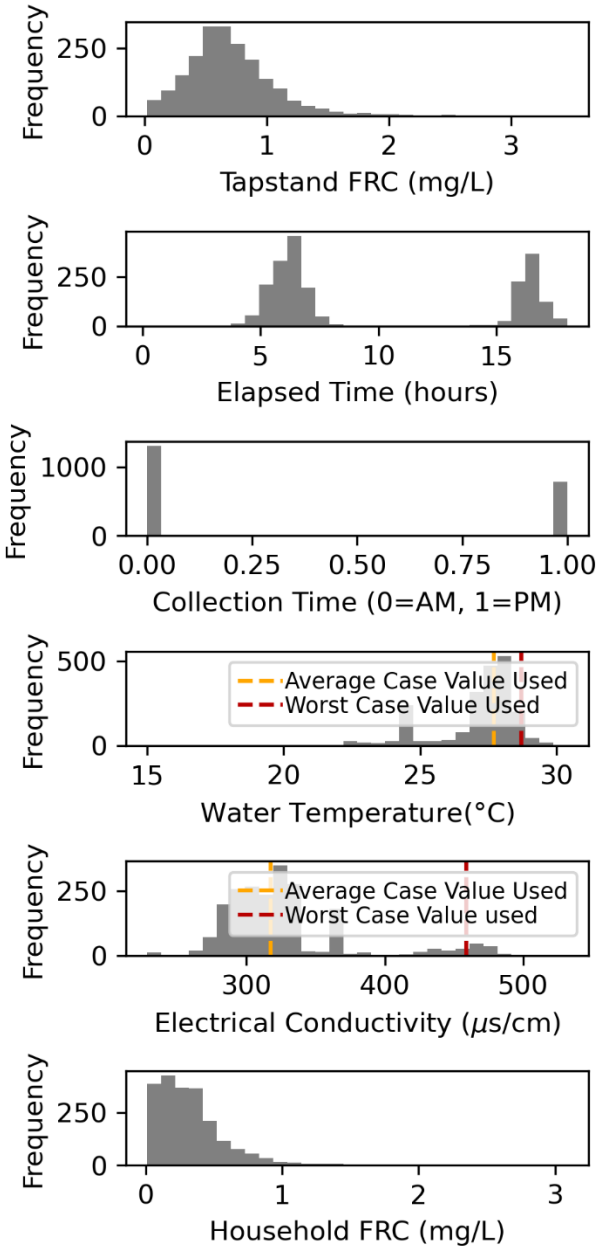

Figure C-8: Input and output variable histograms.

C.6.2.3 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 C-9 shows the predicted risk corresponding to the predictions shown in Figure C-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 C-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 C-8 we know that AM collection was more common, which could be a justification for using the AM collection risk targets.

From Figure C-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 1.0 to 1.1 mg/L. For further accuracy, we can review the risk tables, Tables C-1 and C-2 below. Table C-1 provides the average case targets table and Table C-2 provides the worst-case risk predictions. Based on Figure C-5 and C-9, Table C-2 should be used as, for this site, the worst case scenario targets were more conservative. The SWOT-ANN v2 analytics always provides both the average and worst case target tables, so it is important to identify the correct target table. From Table C-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 based on the available data provided to the model. 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 C-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.

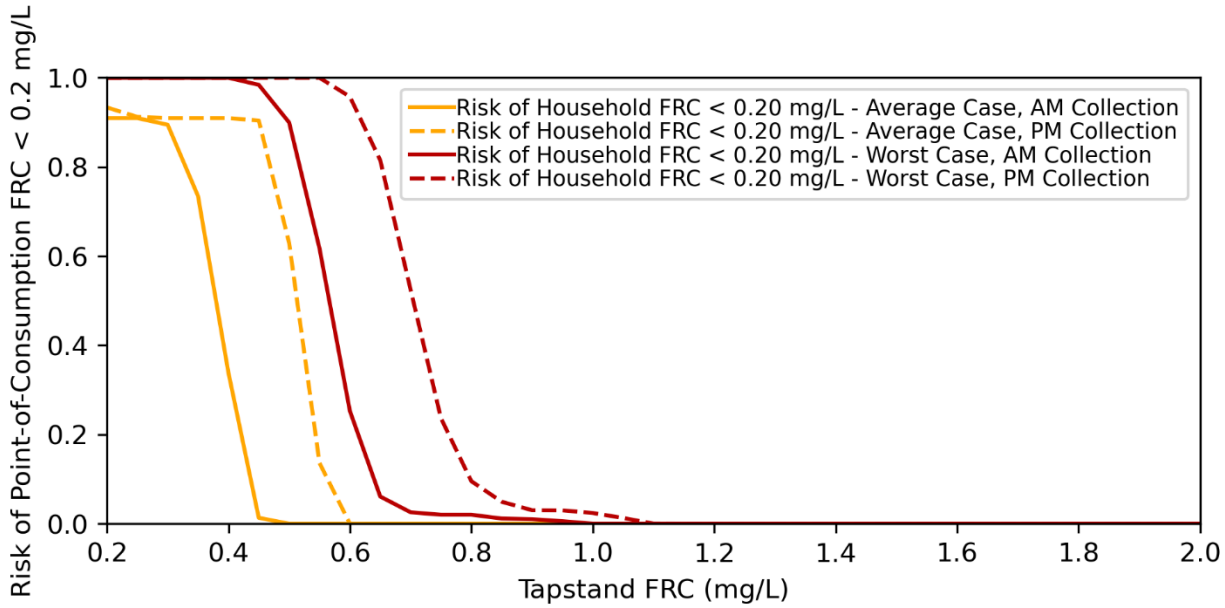

Figure C-9: Risk predictions corresponding to the predictions from Figure C-5

Table C-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

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
1.35	15	28.9	472.0	0.685	0.615	0.000	0.000
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 C-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

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
1.35	15	27.8	308.0	0.589	0.518	0.000	0.000
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 C-10 shows the predicted risk for a site in Tanzania 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.

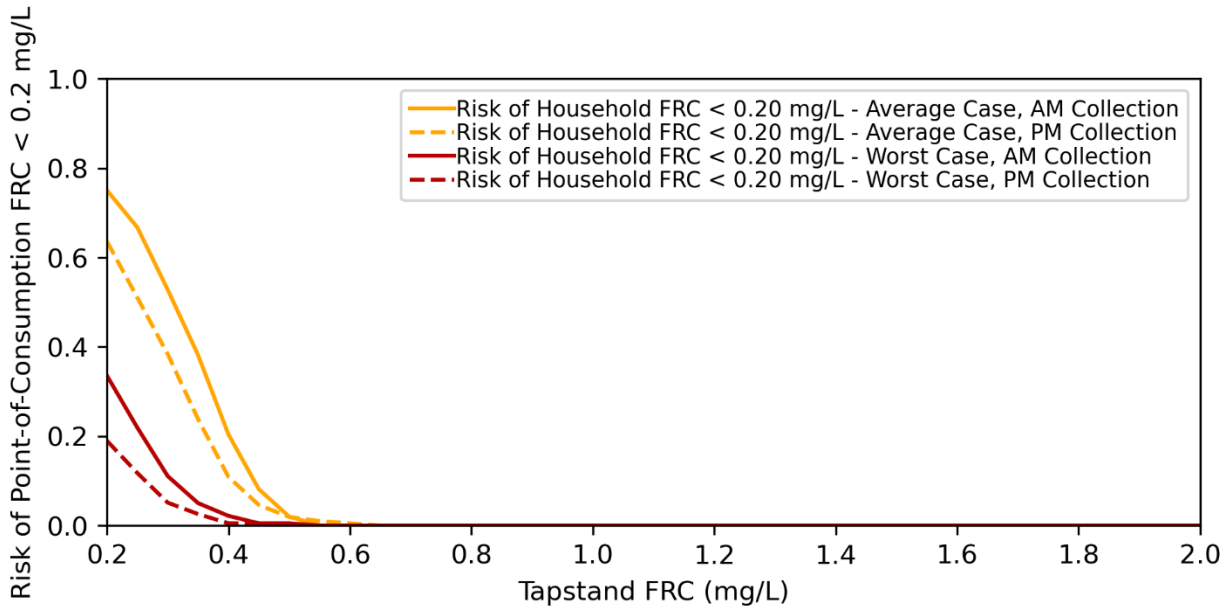

Figure C-10: Predicted risk for a site where the worst-case scenario is not the most conservative.

C.6.2.4 Alternate Interpretation Approach

The above approaches describe means of interpreting the predicted risk figures and tables to select a tapstand FRC target based on allowable risk. However, these figures and tables can also be used to prospectively assess expected risk of having insufficient household FRC for a predefined target. To do this, locate the proposed tapstand FRC target on the x-axis of the predicted risk figure, and then trace up to the predicted risk lines for each scenario. Alternatively, locate the proposed tapstand FRC target in the predicted risk tables in the leftmost column ("Input FRC") and read off the predicted risk from the two rightmost columns in each table. If water quality variables are included in the model, there will be two tables to consider (one for

High Temp and EC, one for Average Temp and EC). Note that when using this alternative interpretation approach to prospectively assess the water safety risk associated with an FRC target, it is important to still review the forecast plots and input and output variable histograms to identify which scenario predictions best match the observed data.

C.6.2.5 Output Interpretation Summary

The SWOT-ANN v2 analytics produce three outputs to help obtain an FRC target:

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

The recommended order for reviewing these outputs is:

1. Review the plot of predictions against tapstand FRC to check 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 matches any patterns in the observed data
    - 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.

C.7 Next Steps

C.7.1 Planned SWOT-ANN v2 Analytics Updates

The following updates are planned in future revisions of the SWOT-ANN v2 Analytics

- Provide automated diagnostics of the forecasts produced by the SWOT-ANN v2 Analytics to identify when the SWOT-ANN targets are likely to underpredict household safety risk. This will replace or supplement the current approach of visually checking the forecast plots (Section C.6.2.1)

- Automate selecting a tapstand FRC targets based on user-defined allowable risk thresholds to supplement the user defined approach described in Section C.6.2.4

- Automate providing prospective risk assessments for user-defined tapstand FRC targets to supplement the user defined approach described in Section C.6.2.5

- Integrate the SWOT-EO recommendations to provide a prospective assessment of household water safety risk based on the SWOT-EO targets. Further assess if these targets meet a user-defined risk threshold, and where risk produced by SWOT-EO targets is too high, propose alternative tapstand FRC target.

C.7.2 SWOT-ANN v3 Analytics

The next major version of the SWOT-ANN analytics will include an update to the training approach to incorporate multi-objective training, which preliminary investigations have shown to improve the forecast dispersion. The next version will also further refine the dynamic IVS process to incorporate additional water quality and water handling variables.

C.8 Conclusion

This report summarizes the analytics included in the SWOT-ANN v2 analytics, providing both the theoretical framework for the analytics as well as providing support for interpreting the outputs for the purposes of obtaining a risk-based FRC target. This version of the SWOT-ANN analytics 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 v2 analytics code is available on the SWOT project GitHub page at: https://github.com/safeh2o/swot-python-analysis.

C.9 References

Ali, S. I., Ali, S. S., and Fesselet, J.-F. (2015). Effectiveness of emergency water treatment practices in refugee camps in South Sudan. Bulletin of the World Health Organization, 93(8), 550–558. https://doi.org/10.2471/BLT.14.147645

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

Boucher, M. A., Perreault, L., Anctil, F., and Favre, A. C. (2015). Exploratory analysis of statistical post-processing methods for hydrological ensemble forecasts. Hydrological Processes, 29(6), 1141–1155. https://doi.org/10.1002/hyp.10234

Bröcker, J., & Smith, L. A. (2008). From ensemble forecasts to predictive distribution functions. Tellus, Series A: Dynamic Meteorology and Oceanography, 60 A(4), 663–678. https://doi.org/10.1111/j.1600-0870.2008.00333.x

Candille, G., and Talagrand, O. (2005). Evaluation of probabilistic prediction systems for a scalar variable. Quarterly Journal of the Royal Meteorological Society, 131(609), 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. *Journal of Water Resources Planning and Management*, *128*(2), 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, 4*(35), 1–16. https://doi.org/10.1038/s41545-021-00125-2

Fisher, I., Kastl, G., & Sathasivan, A. (2017). A comprehensive bulk chlorine decay model for simulating residuals in water distribution systems. *Urban Water Journal, 14*(4), 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. *Mathematical and Computer Modelling, 44*(5–6), 485–498. https://doi.org/10.1016/j.mcm.2006.01.007

Hamill, T. M. (2001). Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Monthly Weather Review, 129*(3), 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 and Forecasting, 15*(5), 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. *Journal of Machine Learning Research, 12*(85), 2825–2830. Retrieved from http://jmlr.org/papers/v12/pedregosa11a.html

Powell, J. C., Hallam, N. B., West, J. R., Forster, C. F., & Simms, J. (2000). Factors which control bulk chlorine decay rates. *Water Research, 34*(1), 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. *Environmental Modelling and Software, 14*(1), 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,*

Series A: Dynamic Meteorology and Oceanography, *55*(1), 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.* Shinfield Park, Reading: ECMWF, 1-25. Retrieved from https://www.ecmwf.int/node/12555

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.,...van Mulbregt, P. & SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods, 17*(3), 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. *Quarterly Journal of the Royal Meteorological Society, 131*(607), 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 Research, 40*(15), 2877–2884. https://doi.org/10.1016/j.watres.2006.05.020

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

Appendix C-1– Including Time in the ANN Model

Introduction

The Safe Water Optimization Tool artificial neural network version 2 (SWOT-ANN v2) analytics use 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 v1 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-ANN v2 analytics 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 v2 analytics that incorporates time-related variables to produce the best performance.

Methods

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).

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).

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.

Modelling Approach

After the exploratory analysis, we trained and tested ANN ensembles with different approaches to incorporating time into the SWOT-ANN v2 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.

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-ANN v1 analytics). 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 proposed for the SWOT-ANN v2 analytics. 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).

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

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 C.5.3.1: 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 C.5.3.2: 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 C.5.3.3: 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 & Talagrand, 2005; Hamill, 2001).
- The CRPS and CRPS reliability term
    - Described in Section C.5.3.4: 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).

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 C-8 in Section C.5.5). 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.

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.

Results and Analysis

Exploratory Data Analysis

Figure C-11 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 C-12 shows histograms of the time of collection for each site, disaggregated by the storage duration (shorter or longer than 12 hours). As with Figure C-11, 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 C-11 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.

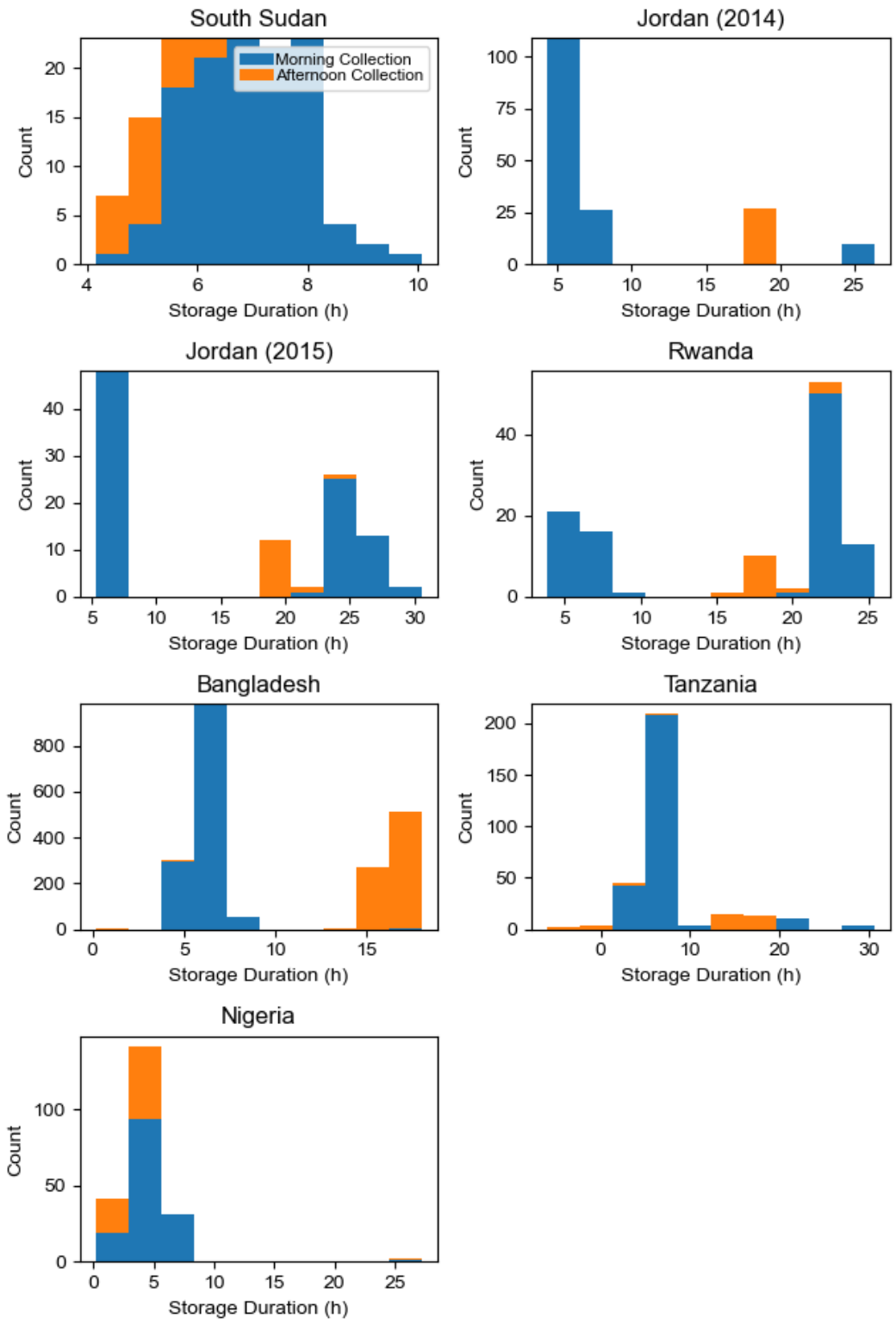

Figure C-11: Storage duration disaggregated by time of collection (morning vs afternoon)

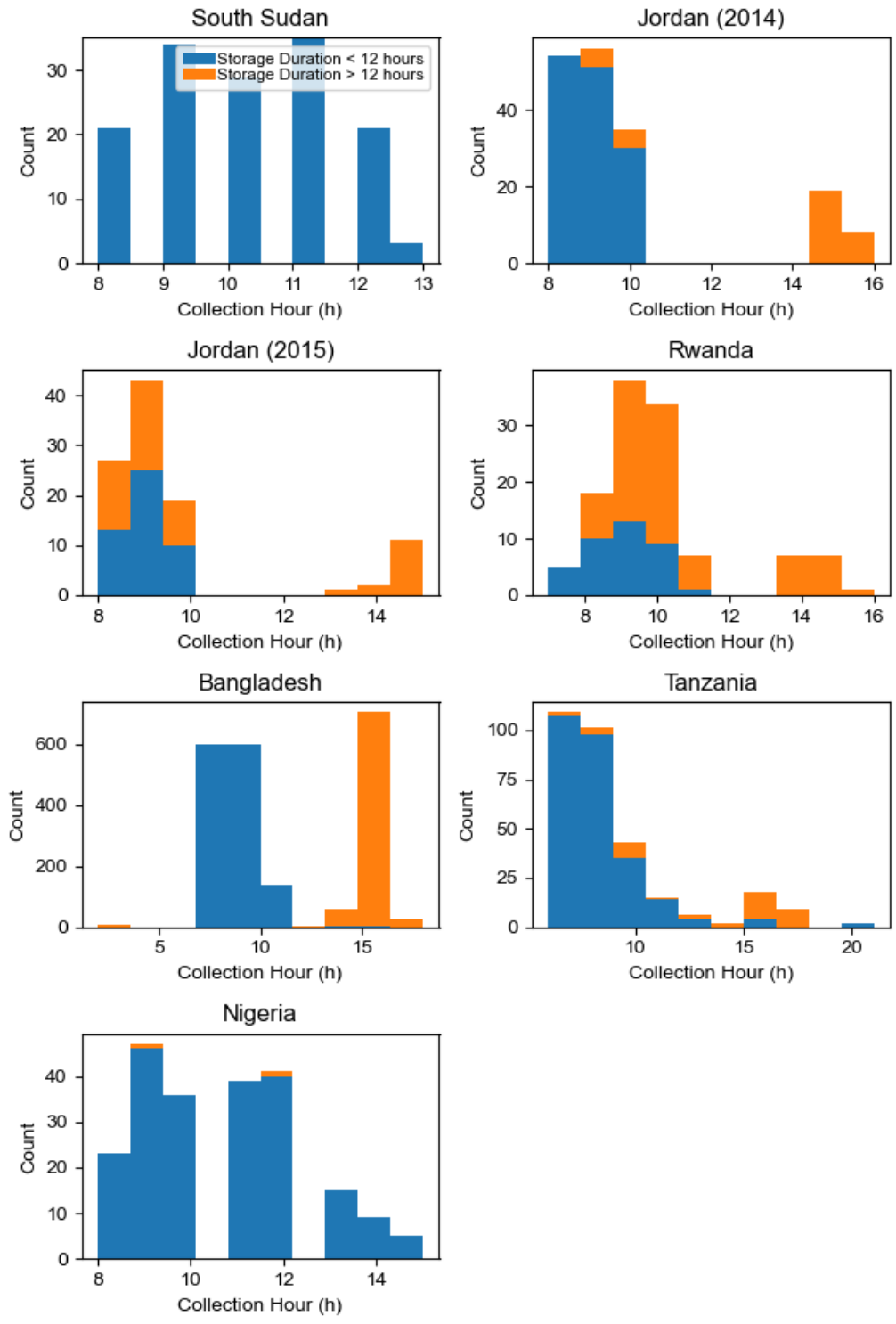

Figure C-12: Time of collection disaggregated by storage duration (longer or shorter than 12 hours)

Ensemble Performance

Figure C-13 shows the Net Improvement Score for each experiment at each site and variable combination. The different coloured bars in Figure C-13 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 C-13: 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 C-14) 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 C-14). 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.

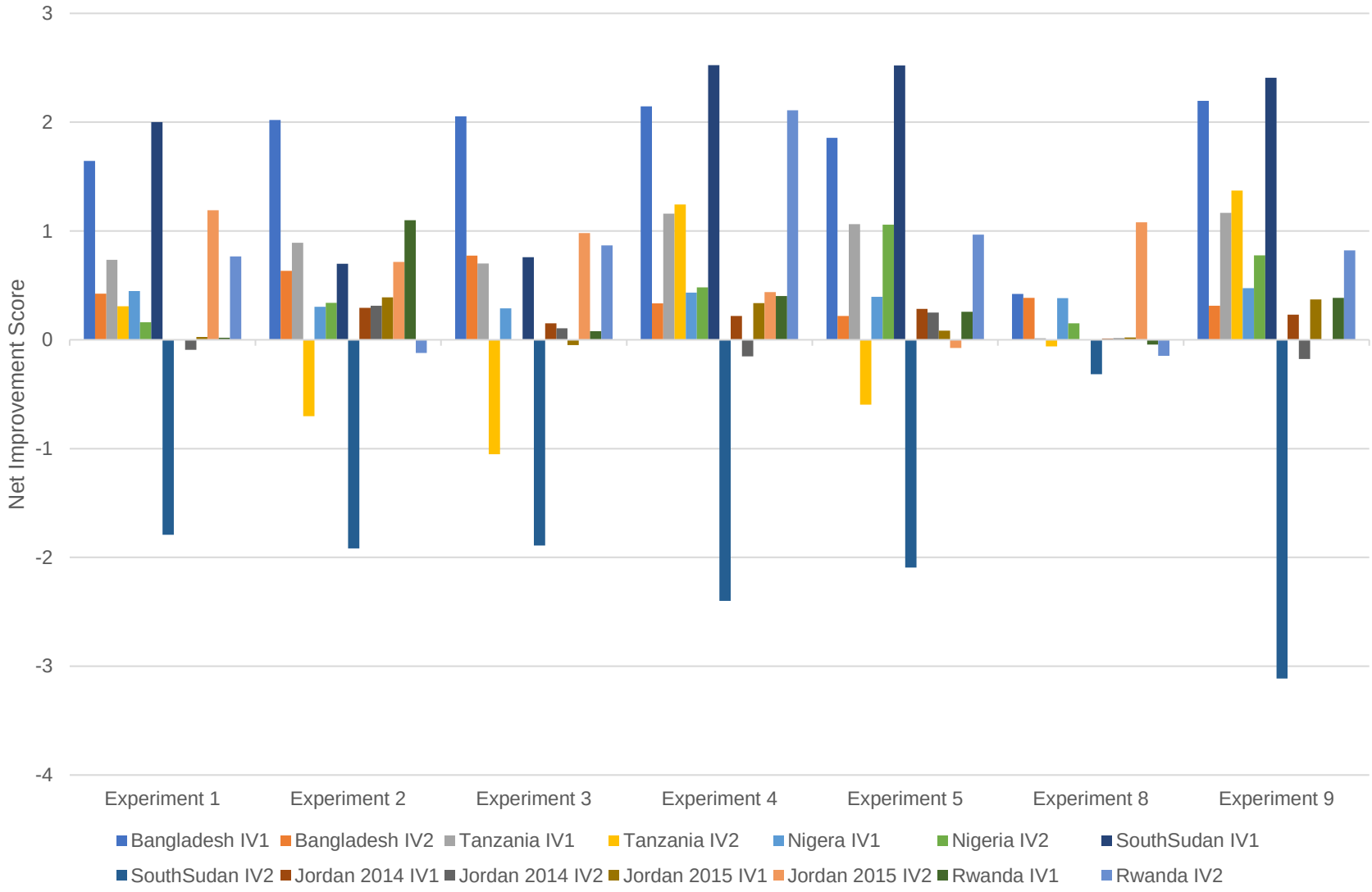

Figure C-13: Net Improvement Score for each site and variable combination, grouped by experiment

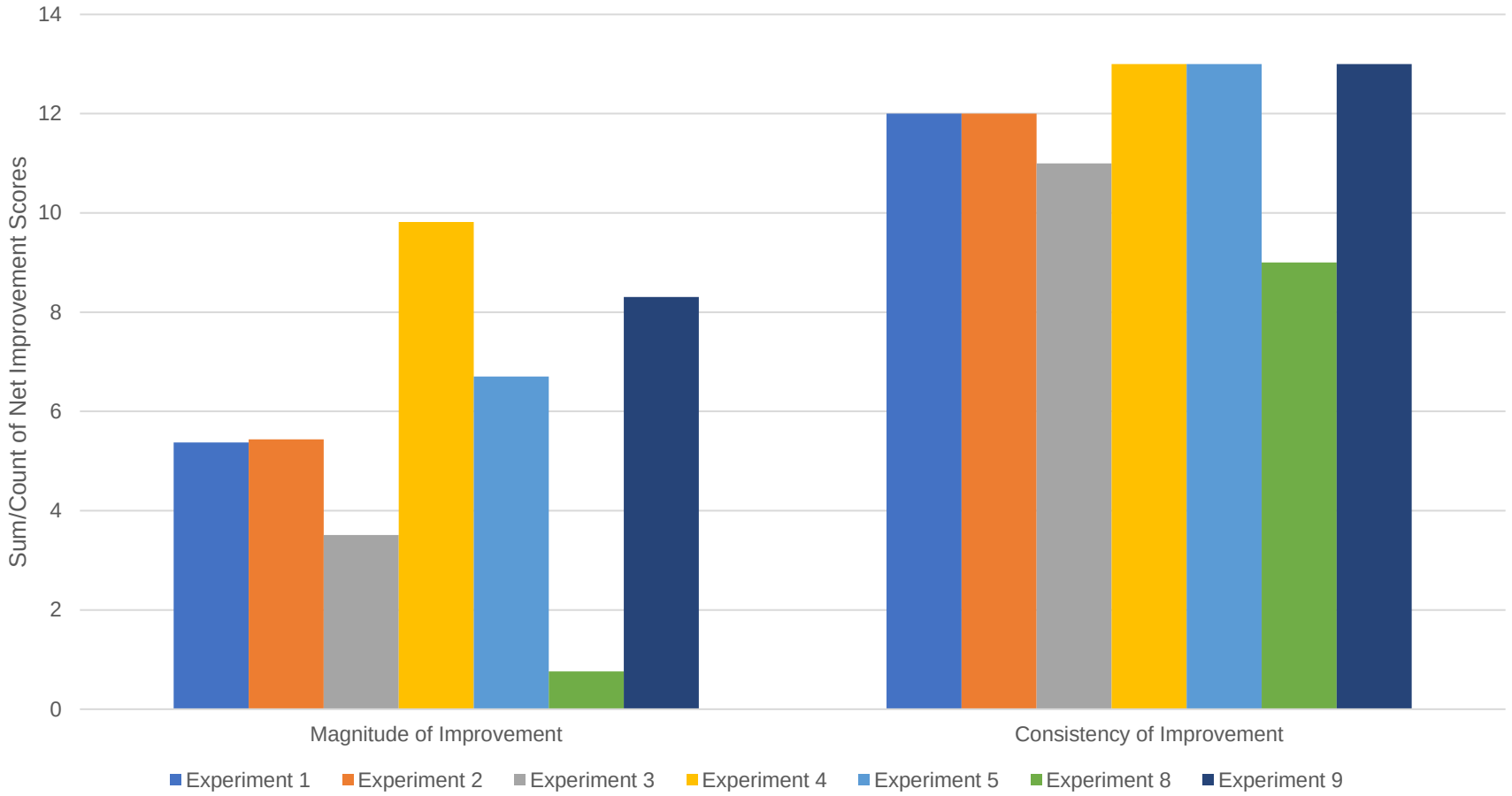

Figure C-14: magnitude and consistency of Net Improvement Scores for each experiment

Conclusion

Based on the findings presented above, we recommend that the SWOT-ANN v2 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.

Appendix C-1 References

Candille, G., and Talagrand, O. (2005). Evaluation of probabilistic prediction systems for a scalar variable. Quarterly Journal of the Royal Meteorological Society, 131(609), 2131–2150. https://doi.org/10.1256/qj.04.71

De Santi, M., Khan, U. T., Arnold, M., Fesselet, J.-F., and Ali, S. I. (2021). Forecasting point-of-consumption chlorine residual in refugee settlements using ensembles of artificial neural networks. Npj Clean Water, 4(35), 1–16. https://doi.org/10.1038/s41545-021-00125-2

Ferro, C. A. T. (2014). Fair scores for ensemble forecasts. Quarterly Journal of the Royal Meteorological Society, 140(683), 1917–1923. https://doi.org/10.1002/qj.2270

Hamill, T. M. (2001). Interpretation of Rank Histograms for Verifying Ensemble Forecasts. Monthly Weather Review, 129(3), 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 and Forecasting, 15(5), 559–570.

Appendix C-2 – Input Variable Selection for the SWOT-ANN v2 analytics

Introduction

The Safe Water Optimization Tool artificial neural network version 2 (SWOT-ANN v2) analytics use 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 additional water quality variables. 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 C.7.1, 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 v1 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 v2 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.

Methods

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).

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).

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 v2 analytics: 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

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 proposed for the SWOT-ANN v2 analytics. 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).

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 C.5.3.1: 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 C.5.3.2: 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 C.5.3.3: 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 & Talagrand, 2005; Hamill, 2001).
- The CRPS and CRPS reliability term
    - Described in Section C.5.3.4: 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 C-8 in Section C.5.5). 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.

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.

Results

Table C-3 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 C-3: 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	-0.056
	IV3	974	0.30	-0.66
	IV4	974	0.30	-0.50
Jordan (2014)	IV1	106	0.33	-
	IV2	106	0.33	0.11
	IV3	103	0.32	1.26
	IV4	103	0.32	0.92
Jordan (2015)	IV1	87	0.15	-
	IV2	87	0.15	0.78
	IV3	78	0.15	1.57
	IV4	78	0.15	1.81
Nigeria	IV1	216	0.11	-
	IV2	216	0.11	0.28
	IV3	216	0.11	0.11
	IV4	216	0.11	1.28
Rwanda	IV1	117	0.23	-
	IV2	94	0.19	1.05
	IV3	94	0.19	0.56
	IV4	94	0.19	1.03
South Sudan	IV1	143	0.37	-
	IV2	142	0.37	0.52
	IV3	127	0.36	-0.48
	IV4	126	0.36	-3.17
Tanzania	IV1	305	0.15	-
	IV2	89	0.20	0.90
	IV3	250	0.15	0.86
	IV4	89	0.20	1.91

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 C-15. 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.

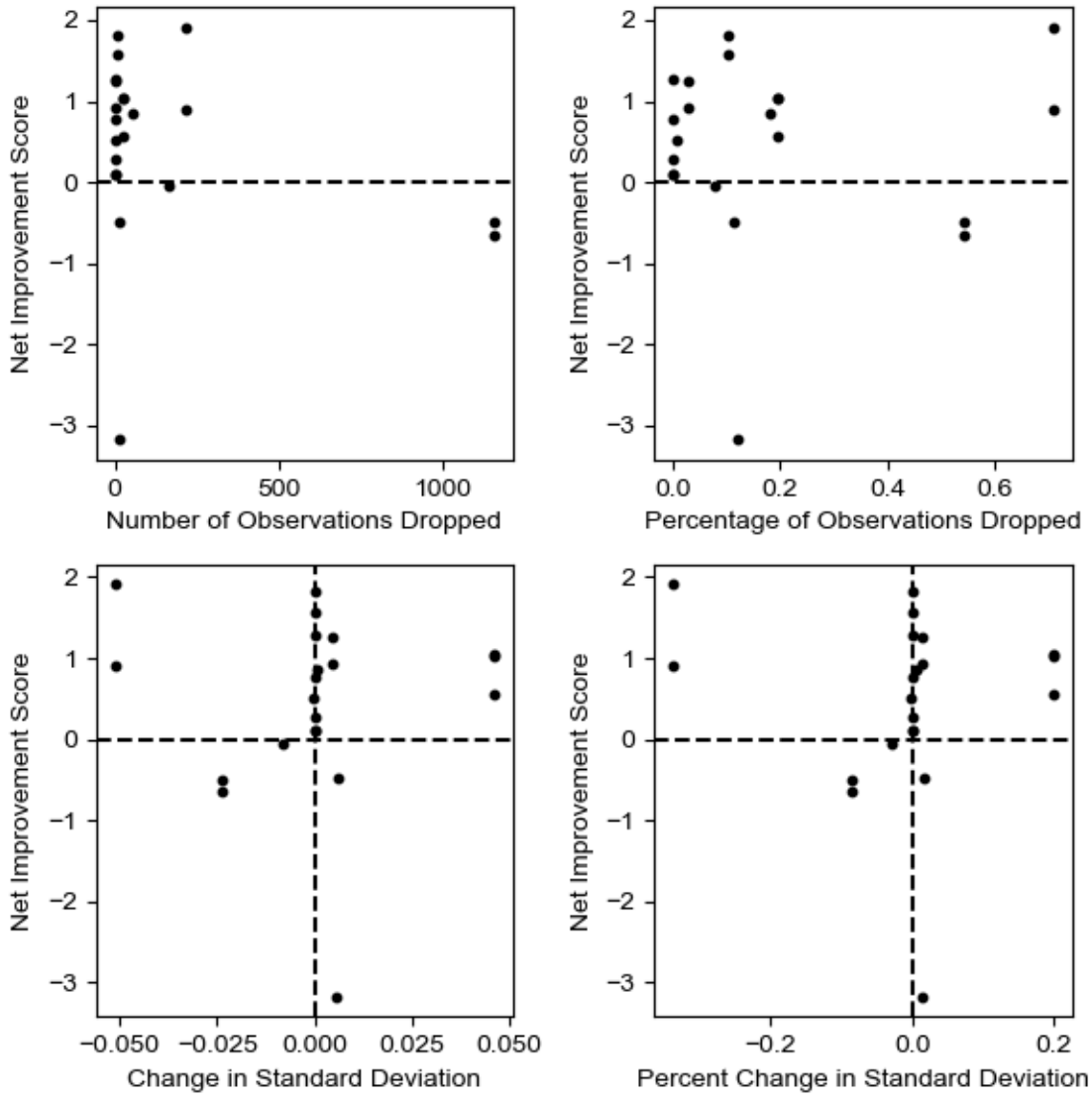

Figure C-15: Comparison of Net Improvement Score to potential explanatory factors in the dataset

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 v2 analytics, 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.

Appendix C-2 References

Bowden, G. J., Nixon, J. B., Dandy, G. C., Maier, H. R., and Holmes, M. (2006). Forecasting chlorine residuals in a water distribution system using a general regression neural network. Mathematical and Computer Modelling, 44(5–6), 469–484. https://doi.org/10.1016/j.mcm.2006.01.006

Candille, G., and Talagrand, O. (2005). Evaluation of probabilistic prediction systems for a scalar variable. Quarterly Journal of the Royal Meteorological Society, 131(609), 2131–2150. https://doi.org/10.1256/qj.04.71

De Santi, M., Khan, U. T., Arnold, M., Fesselet, J.-F., and Ali, S. I. (2021). Forecasting point-of-consumption chlorine residual in refugee settlements using ensembles of artificial neural networks. Npj Clean Water, 4(35), 1–16. https://doi.org/10.1038/s41545-021-00125-2

Ferro, C. A. T. (2014). Fair scores for ensemble forecasts. Quarterly Journal of the Royal Meteorological Society, 140(683), 1917–1923. https://doi.org/10.1002/qj.2270

Hamill, T. M. (2001). Interpretation of Rank Histograms for Verifying Ensemble Forecasts. Monthly Weather Review, 129(3), 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 and Forecasting, 15(5), 559–570.

Soyupak, S., Kilic, H., Karadirek, I. E., and Muhammetoglu, H. (2011). On the usage of artificial neural networks in chlorine control applications for water distribution networks with high quality water. Journal of Water Supply: Research and Technology - AQUA, 60(1), 51–60. https://doi.org/10.2166/aqua.2011.086

Appendix C-3 – Comparison of Bandwidth Selection Methods for Post-Processing

Introduction

The Safe Water Optimization Tool artificial neural network version 2 (SWOT-ANN v2) analytics use 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-ANN v2 analytics use 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. However, ensembles of ANNs 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.

Methods

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).

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).

Modelling Approach

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 proposed for the SWOT-ANN v2 analytics. 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).

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., 2011, 2015; Bröcker & Smith, 2008; De Santi et al., 2021; Roulston & 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-7 from Section C.5.5.

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).

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-8 in Section C.5.5). 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.

Results

Table C-4 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-16 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-16, 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 v2 analytics where there are often more ensemble members than testing observations.

Table C-4: Comparison of Net Improvement Scores for the three post-processing methods

Site	Input Variable Combination	Best Member Error (Roulston & 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

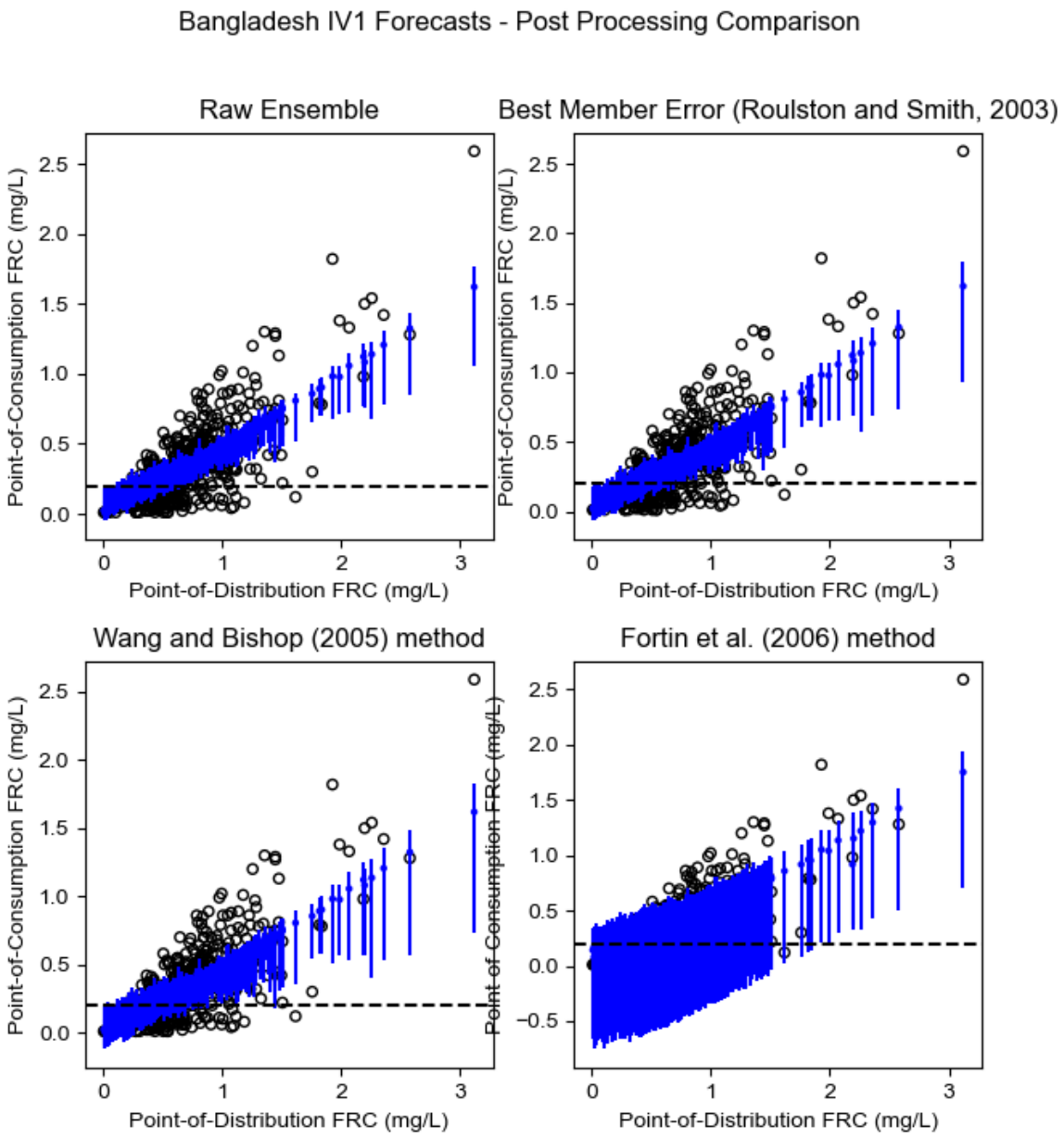

Figure C-16: Comparison of post-processing methods for the Bangladesh IV1

Table C-4 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 v2 analytics.

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 v2 analytics. However, we also note that the Wang and Bishop method does not always lead to improved performance, and as such, the SWOT-ANN v2 analytics should always compare the raw and post-processed performance to ensure that the best performing ensemble is used.

Appendix C-3 References

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

Boucher, M. A., Perreault, L., Anctil, F., and Favre, A. C. (2015). Exploratory analysis of statistical post-processing methods for hydrological ensemble forecasts. Hydrological Processes, 29(6), 1141–1155. https://doi.org/10.1002/hyp.10234

Bröcker, J., & Smith, L. A. (2008). From ensemble forecasts to predictive distribution functions. Tellus, Series A: Dynamic Meteorology and Oceanography, 60 A(4), 663–678. https://doi.org/10.1111/j.1600-0870.2008.00333.x

De Santi, M., Khan, U. T., Arnold, M., Fesselet, J.-F., and Ali, S. I. (2021). Forecasting point-of-consumption chlorine residual in refugee settlements using ensembles of artificial neural networks. Npj Clean Water, 4(35), 1–16. https://doi.org/10.1038/s41545-021-00125-2

Ferro, C. A. T. (2014). Fair scores for ensemble forecasts. Quarterly Journal of the Royal Meteorological Society, 140(683), 1917–1923. https://doi.org/10.1002/qj.2270

Fortin, V., Favre, A. C., and 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. Quarterly Journal of the Royal Meteorological Society, 132(617),

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 and Forecasting, 15(5), 559–570.

Roulston, M. S., & Smith, L. A. (2003). Combining dynamical and statistical ensembles. *Tellus, Series A: Dynamic Meteorology and Oceanography*, *55*(1), 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. *Quarterly Journal of the Royal Meteorological Society, 131*(607), 965–986. https://doi.org/10.1256/qj.04.120

Glossary of Functions and Explanations

The Safe Water Optimization Tool artificial neural network version 2 (SWOT-ANN v2) analytis 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 v2 analytics) 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-ANN v2 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	Predictor and target DataFrames, count of dropped rows and other rule checks

Function	Inputs	Description	Outputs
		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	
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

Function	Inputs	Description	Outputs
			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
train_network	x, t, directory	Called from "train_SWOT_network", trains the individual neural network	1 trained neural network

Function	Inputs	Description	Outputs
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

Function	Inputs	Description	Outputs
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	post_process_check: if True, post-processing is used to generate targets, if False, post-processing is not used

Function	Inputs	Description	Outputs
		is used, if negative, post-processing is not	
get_bw		Called from "post_process_cal", calculates the kernel bandwidth used for post-processing	Bandwidth
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	Tables of inputs for generating risk-based FRC targets

Function	Inputs	Description	Outputs
		the tables used for forecasting the risk-based FRC targets. Note that four tables are produced for different possible scenario analyses	
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	Raw ensemble forecasts, post-processed ensemble forecasts, risk of low household FRC for each scenario

Function	Inputs	Description	Outputs
		having household FRC below 0.2 mg/L	
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
generate_html_report	report file, storage_target	Called from “run_swot_script”, compiles an html report of the key results: Prediction figures	HTML report

Function	Inputs	Description	Outputs
		Risk figures Input and output variable histograms Risk tables Diagnostic figures Table of skipped rows	
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

Function	Inputs	Description	Outputs
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 D. – Supplemental Material for Chapter 4

D.1 Data Cleaning Rules

Data cleaning rules were used to remove measurements with obvious errors to avoid impacting the ANN analytics. Three data cleaning rules were implemented, removing observations for the following criteria:

- Incorrect timestamp: a time stamp was deemed incorrect if the data recorded occurred before the field trial began. This led to the removal of 4 observations in Bangladesh

- Erroneous elapsed time: samples with negative elapsed times or elapsed times greater than 48 hours were removed. This led to the removal of 5 observations in Bangladesh and 2 observations in Tanzania

- Erroneous water temperature: samples with water temperature above 50 were removed. This led to the removal of 1 observation in Nigeria. Additionally, 420 observations in Bangladesh had water temperatures of 49 degrees Celsius which data collectors had used as an indicator that the water temperature probe was malfunctioning. These were recoded to null measurements.

  Increase in FRC from point-of-distribution to point-of-consumption by more than 0.06 mg/L. The measurement error for the FRC measurements was 0.03 mg/L so any increase in FRC from the point-of-distribution to point-of-consumption greater than 0.06 mg/L (double the measurement error reflecting maximum error in both measurements) is impossible. This led to the removal of 8 samples in Bangladesh and 2 in Tanzania.

D.2 Calculation of weighted cost functions

This appendix shows how the weighted form of each of the cost functions listed in Section 2.5 was calculated. The weighting factor w_i shown in the following calculation refers to the weight applied to the i^{th} observation.

D.2.1 Weighted MSE

The weighted form of MSE is calculated as a weighted average. Thus, the numerator, shown in Equation D-1 becomes the weighted sum of squared errors, with the squared error of each sample multiplied by the sample weight. Instead of using the number of observations, the denominator also becomes the sum of the weights for all samples.

$$MSE_w = \frac{\sum_{i=1}^{N} w_i (P_i - O_i)^2}{\sum_{i=1}^{N} w_i} \quad \text{(D-1)}$$

D.2.2 Weighted NSE

To calculate the weighted NSE, we use the form of the NSE shown in Gupta et al. (2009) which describes the NSE as the MSE normalized about the variance of the observations, subtracted from one. Thus, to calculated weighted NSE, we use the weighted MSE (Equation D-1) as well as the weighted standard deviation, shown in Equation D-2:

$$\sigma_{y_{O_w}}^2 = \frac{\sum_{i=1}^{N} w_i (O_i - \overline{O_w})^2}{\sum_{i=1}^{N} w_i} \quad \text{(D-2)}$$

In Equation D-2 $\overline{O_w}$ is the weighted average of the observations. This is calculated as:

$$\overline{\mathrm{O_w}} = \frac{\sum_{i=1}^{N} w_i O_i}{\sum_{i=1}^{N} w_i} \quad \text{(D-3)}$$

Finally, the weighted NSE is calculated using the Gupta et al. (2009) decomposition as:

$$NSE_w = 1 - (MSE_w)/\sigma_{O_w}^2 \quad \text{(D-4)}$$

D.2.3 Weighting KGE

As discussed in Section 2.5.3, the KGE score is composed from three terms which evaluate the correlation between the observations and the predictions (r), the ratio between the observed and predicted variance (α), and the ratio between the observed and predicted means (β). To calculate these terms first the weighted means and variances of the observations and predictions were calculated. For the observations these are shown in Equation D-2 for the weighted variance, and D-3 for the weighted mean. The weighted mean and variance of the predictions ($\overline{P_w}$ and σ_{P_W}, respectively) are calculated using the same approach as for the observations. Using these, the

weighted covariance is calculated in Equation D-5 for the weighted covariance between the observations and predictions, and in Equations D-6 and D-7 for the weighted covariance of the observations and predictions, respectively.

Using this, the weighted covariances are calculated as Second calculate the weighted covariance:

$$\mathrm{cov}(O,P;w) = \frac{\sum_{i=1}^{N} w_i\left(O_i - \overline{\mathrm{O_w}}\right)\left(P_i - \overline{\mathrm{P_w}}\right)}{\sum_{i=1}^{N} w_i} \quad \text{(D-5)}$$

$$\mathrm{cov}(O,O;w) = \frac{\sum_{i} w_i\left(O_i - \overline{\mathrm{O_w}}\right)^2}{\sum_{i=1}^{N} w_i} \quad \text{(D-6)}$$

$$\mathrm{cov}(P,P;w) = \frac{\sum_{i} w_i\left(P_i - \overline{\mathrm{P_w}}\right)^2}{\sum_{i=1}^{N} w_i} \quad \text{(D-7)}$$

Using these covariances, the weighted correlation is calculated in Equation D-8. This provides the weigthed version of the r term, r_W.

$$\mathrm{r_w} = \frac{\mathrm{cov}(O,P;w)}{\sqrt{cov(O,O;w)cov(P,P;w)}} \quad \text{(D-8)}$$

The weighted versions of the α and β terms are calculated as the ratio of the weighted variances and weighted means, respectively, as shown in Equations D-9 and D-10.

$$\alpha_w = \frac{\sigma_{P_w}}{\sigma_{O_w}} \quad \text{(D-9)}$$

$$\beta_w = \frac{\overline{\mathrm{P_w}}}{\overline{\mathrm{O_w}}} \quad \text{(D-10)}$$

The weighted KGE is then calculated using the same Euclidean distance calculation as was shown in Section 2.5.3 but replacing the r, α, and β terms with their weighted forms (r_W, α_W, β_W).

$$KGE_w = 1 - \sqrt{(r_w - 1)^2 + (\alpha_w - 1)^2 + (\beta_w - 1)^2} \quad \text{(D-11)}$$

D.2.4 Weighted AI

Since the index of agreement is a modification of the NSE a similar approach was taken for calculating weighted AI. Weighted MSE (Equation D-1) was used as the numerator, and the

weighted variance in the denominator was calculated analogously to NSE, with the resulting calculation as:

$$d = 1 - \frac{MSE_w}{\sum_{i=1}^{N}(w_i|P_i - \overline{O_w}| + w_i|O_i - \overline{O_w}|)^2 / \sum_{i=1}^{N} w_i} \quad \text{(D-12)}$$

D.3 Supplemental Information

Site and Input Variable Combination	Cost Function and Weighting Combination	PC	$PC_{<0.2}$	CI_{score}	$CI_{score_{<0.2}}$	δ	$\delta_{<0.2}$	$\overline{CRPS}$	$\overline{Reli}$
Bangladesh IV1	Unweighted MSE	0.22	0.25	2.86	2.50	153	97	0.16	0.09
	MSE Weighting 1	0.20	0.26	2.84	2.53	170	98	0.14	0.06
	MSE Weighting 2	0.24	0.30	2.76	2.46	156	90	0.14	0.06
	MSE Weighting 3	0.37	0.41	2.22	1.97	119	57	0.20	0.11
	Unweighted NSE	0.16	0.23	2.99	2.49	178	98	0.16	0.08
	NSE Weighting 1	0.24	0.26	2.77	2.48	152	102	0.15	0.07
	NSE Weighting 2	0.23	0.25	2.76	2.48	158	102	0.15	0.07
	NSE Weighting 3	0.33	0.39	2.32	2.03	126	60	0.18	0.10
	Unweighted KGE	0.40	0.40	2.30	2.31	94	36	0.16	0.07
	KGE Weighting 1	0.65	0.81	1.75	1.55	45	14	0.18	0.09
	KGE Weighting 2	0.51	0.58	1.83	1.62	63	28	0.18	0.09
	KGE Weighting 3	0.79	0.84	0.89	0.85	14	5.8	0.19	0.09
	Unweighted IoA	0.27	0.30	2.83	2.69	137	51	0.15	0.07
	IoA Weighting 1	0.50	0.71	2.33	1.83	87	28	0.17	0.09
	IoA Weighting 2	0.52	0.75	2.35	1.82	85	17	0.17	0.09
	IoA Weighting 3	0.50	0.51	2.00	1.98	77	37	0.20	0.10
Bangladesh IV2	Unweighted MSE	0.32	0.23	2.58	2.68	54	49	0.21	0.11
	MSE Weighting 1	0.39	0.35	2.24	2.15	47	35	0.23	0.12
	MSE Weighting 2	0.37	0.32	2.33	2.39	49	39	0.22	0.11
	MSE Weighting 3	0.49	0.38	1.92	2.36	37	32	0.24	0.13
	Unweighted NSE	0.36	0.22	2.62	2.75	48	50	0.23	0.12
	NSE Weighting 1	0.35	0.32	2.33	2.18	52	39	0.22	0.11
	NSE Weighting 2	0.34	0.32	2.43	2.45	53	39	0.22	0.11
	NSE Weighting 3	0.53	0.40	1.84	2.26	40	30	0.25	0.14
	Unweighted KGE	0.52	0.48	2.02	1.85	27	16	0.20	0.09

Site and Input Variable Combination	Cost Function and Weighting Combination	PC	$PC_{<0.2}$	CI_{score}	$CI_{score_{<0.2}}$	δ	$\delta_{<0.2}$	$\overline{CRPS}$	$\overline{Reli}$
	KGE Weighting 1	0.63	0.71	1.41	1.03	21	7.4	0.20	0.10
	KGE Weighting 2	0.57	0.57	1.59	1.29	23	11	0.20	0.09
	KGE Weighting 3	0.82	0.74	0.75	0.80	6.2	6.2	0.24	0.12
	Unweighted IoA	0.38	0.44	2.45	2.16	46	16	0.18	0.08
	IoA Weighting 1	0.43	0.66	2.46	2.04	56	8.0	0.18	0.08
	IoA Weighting 2	0.28	0.41	2.66	2.29	85	14	0.14	0.05
	IoA Weighting 3	0.51	0.51	1.41	1.42	31	17	0.21	0.10
	Unweighted MSE	0.31	0.31	2.67	2.81	19	24	0.10	0.05
	MSE Weighting 1	0.23	0.35	2.73	2.23	22	21	0.09	0.05
	MSE Weighting 2	0.18	0.27	3.03	2.65	25	27	0.09	0.05
	MSE Weighting 3	0.44	0.31	1.91	1.96	20	24	0.12	0.07
	Unweighted NSE	0.32	0.33	2.30	2.38	18	23	0.10	0.06
	NSE Weighting 1	0.18	0.29	2.76	2.22	25	24	0.09	0.05
	NSE Weighting 2	0.22	0.33	3.04	2.66	23	24	0.09	0.05
	NSE Weighting 3	0.39	0.29	2.10	2.07	24	25	0.12	0.07
	Unweighted KGE	0.56	0.67	2.32	2.53	9.3	6.5	0.09	0.04
	KGE Weighting 1	0.62	0.80	2.04	1.91	12	8.4	0.08	0.03
	KGE Weighting 2	0.61	0.76	1.99	1.92	10	6.6	0.08	0.03
	KGE Weighting 3	0.78	0.78	0.73	0.75	5.9	5.0	0.12	0.06
	Unweighted IoA	0.47	0.53	2.45	2.84	14	9.1	0.08	0.03
	IoA Weighting 1	0.48	0.63	1.89	1.32	15	4.7	0.07	0.03
	IoA Weighting 2	0.47	0.59	1.90	1.41	15	8.0	0.08	0.03
Tanzania IV1	IoA Weighting 3	0.79	0.80	1.67	1.54	7.5	8.4	0.11	0.05
	Unweighted MSE	0.35	0.50	2.58	2.83	5.1	2.0	0.16	0.07
	MSE Weighting 1	0.43	0.83	1.79	0.94	5.5	0.98	0.16	0.07
	MSE Weighting 2	0.43	0.67	2.57	2.31	4.9	1.6	0.17	0.08
Tanzania IV2	MSE Weighting 3	0.35	0.50	1.92	1.63	5.2	2.0	0.16	0.07

Site and Input Variable Combination	Cost Function and Weighting Combination	PC	$PC_{<0.2}$	CI_{score}	$CI_{score_{<0.2}}$	δ	$\delta_{<0.2}$	$\overline{CRPS}$	$\overline{Reli}$
	Unweighted NSE	0.35	0.67	2.35	2.24	5.4	1.3	0.15	0.06
	NSE Weighting 1	0.43	0.83	1.87	1.18	5.5	0.98	0.16	0.07
	NSE Weighting 2	0.43	0.67	2.61	2.31	4.9	1.6	0.16	0.08
	NSE Weighting 3	0.39	0.67	1.86	1.34	4.7	1.3	0.16	0.07
	Unweighted KGE	0.39	0.67	1.63	1.16	4.7	1.3	0.17	0.07
	KGE Weighting 1	0.43	0.67	1.57	0.47	5.5	0.98	0.16	0.06
	KGE Weighting 2	0.43	0.50	1.69	0.74	4.8	1.3	0.17	0.07
	KGE Weighting 3	0.61	0.67	1.25	1.27	2.8	1.3	0.23	0.12
	Unweighted IoA	0.39	0.67	2.10	1.51	4.7	1.3	0.17	0.07
	IoA Weighting 1	0.43	0.50	2.03	1.13	4.8	1.3	0.18	0.08
	IoA Weighting 2	0.30	0.50	2.01	0.72	6.3	1.3	0.16	0.07
	IoA Weighting 3	0.48	0.67	1.74	2.23	3.4	1.3	0.16	0.06
	Unweighted MSE	0.26	0.00	2.51	3.85	14	3.00	0.10	0.05
	MSE Weighting 1	0.33	0.67	1.98	0.43	16	0.99	0.09	0.04
	MSE Weighting 2	0.30	0.00	2.95	3.85	16	3.00	0.08	0.04
	MSE Weighting 3	0.50	0.67	1.18	0.49	7.7	0.99	0.10	0.04
	Unweighted NSE	0.30	0.67	2.53	2.96	13	1.7	0.10	0.05
	NSE Weighting 1	0.33	0.67	1.90	0.34	16	0.99	0.09	0.04
	NSE Weighting 2	0.30	0.00	2.90	3.85	16	3.00	0.08	0.03
	NSE Weighting 3	0.52	0.67	1.20	0.32	7.0	0.99	0.10	0.04
	Unweighted KGE	0.70	0.67	1.50	1.23	6.3	0.99	0.12	0.05
	KGE Weighting 1	0.44	0.67	3.06	2.96	12	0.99	0.11	0.05
	KGE Weighting 2	0.41	0.67	2.75	2.47	10	0.99	0.13	0.07
	KGE Weighting 3	0.56	0.67	0.67	0.96	7.7	1.7	0.11	0.05
	Unweighted IoA	0.44	0.67	1.84	0.36	8.7	0.99	0.12	0.06
	IoA Weighting 1	0.37	0.67	3.06	2.96	11	0.99	0.12	0.06
Nigeria IV1	IoA Weighting 2	0.26	0.67	2.69	1.41	15	0.99	0.10	0.05

Site and Input Variable Combination	Cost Function and Weighting Combination	PC	$PC_{<0.2}$	CI_{score}	$CI_{score_{<0.2}}$	δ	$\delta_{<0.2}$	$\overline{CRPS}$	$\overline{Reli}$
	IoA Weighting 3	0.74	0.67	0.26	0.47	5.2	0.99	0.11	0.05
	Unweighted MSE	0.24	0.00	3.01	3.85	15	3.00	0.10	0.05
	MSE Weighting 1	0.33	0.33	2.63	1.74	14	1.7	0.09	0.04
	MSE Weighting 2	0.24	0.00	2.73	3.85	16	3.00	0.09	0.04
	MSE Weighting 3	0.65	0.67	1.55	0.76	5.0	0.99	0.11	0.05
	Unweighted NSE	0.19	0.00	3.15	3.85	17	3.00	0.10	0.05
	NSE Weighting 1	0.31	0.33	2.53	1.52	15	1.7	0.09	0.04
	NSE Weighting 2	0.28	0.33	2.65	3.29	15	1.7	0.09	0.04
	NSE Weighting 3	0.59	0.67	1.64	0.94	5.7	0.99	0.11	0.04
	Unweighted KGE	0.44	0.33	2.56	1.29	9.7	1.7	0.13	0.07
	KGE Weighting 1	0.30	0.67	2.13	2.05	14	0.99	0.09	0.04
	KGE Weighting 2	0.56	0.67	2.05	1.18	6.9	0.99	0.12	0.06
	KGE Weighting 3	0.67	0.67	0.96	1.41	4.9	0.99	0.11	0.04
	Unweighted IoA	0.35	0.33	2.70	1.29	11	1.7	0.12	0.06
	IoA Weighting 1	0.57	0.67	1.75	2.05	5.8	0.99	0.12	0.06
	IoA Weighting 2	0.30	0.67	2.33	2.47	13	0.99	0.10	0.04
Nigeria IV2	IoA Weighting 3	0.85	0.67	0.47	0.67	1.9	0.99	0.11	0.05

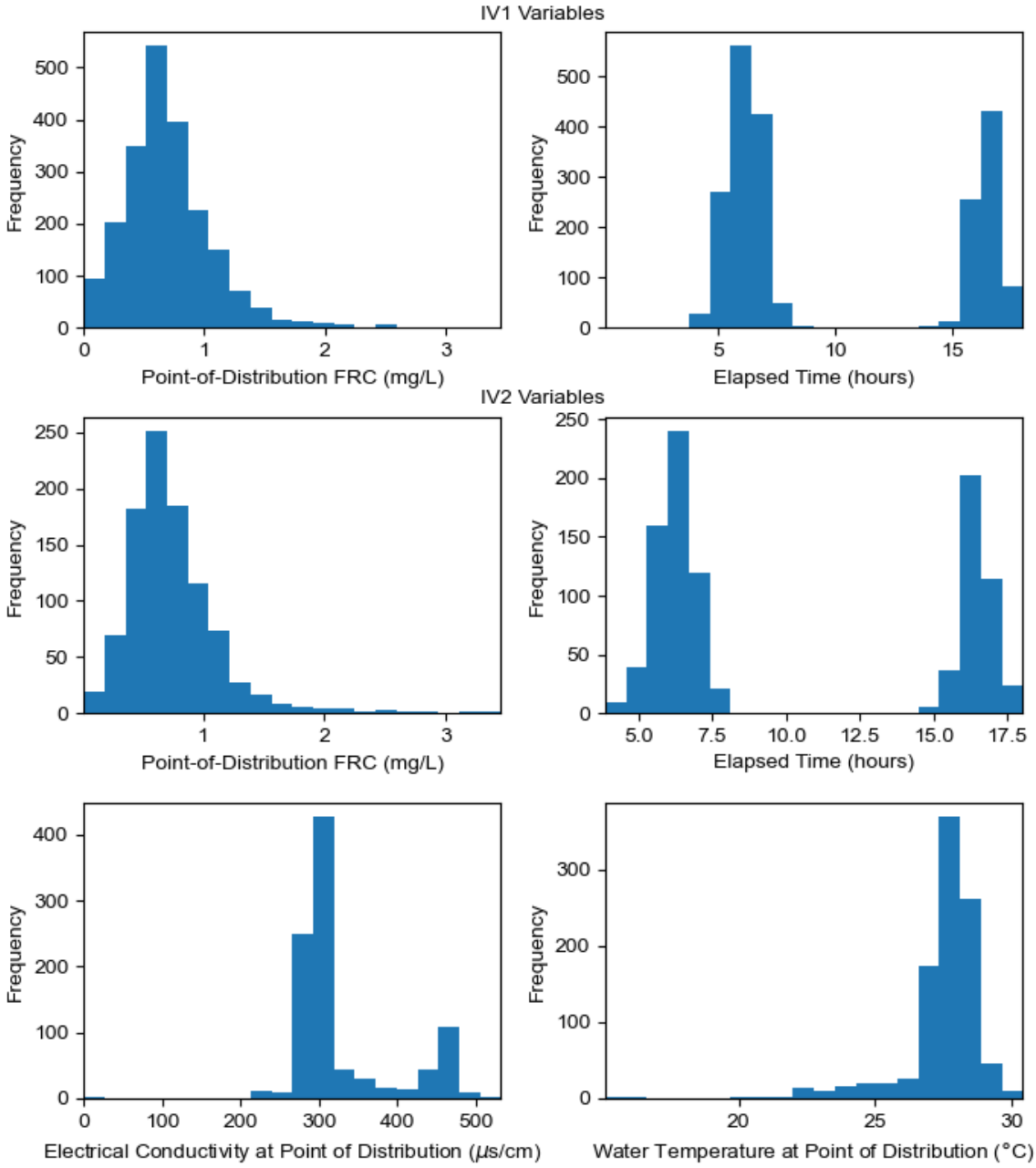

Figure D-1: Input variable histograms for Bangladesh showing IV1 and IV2 input variables

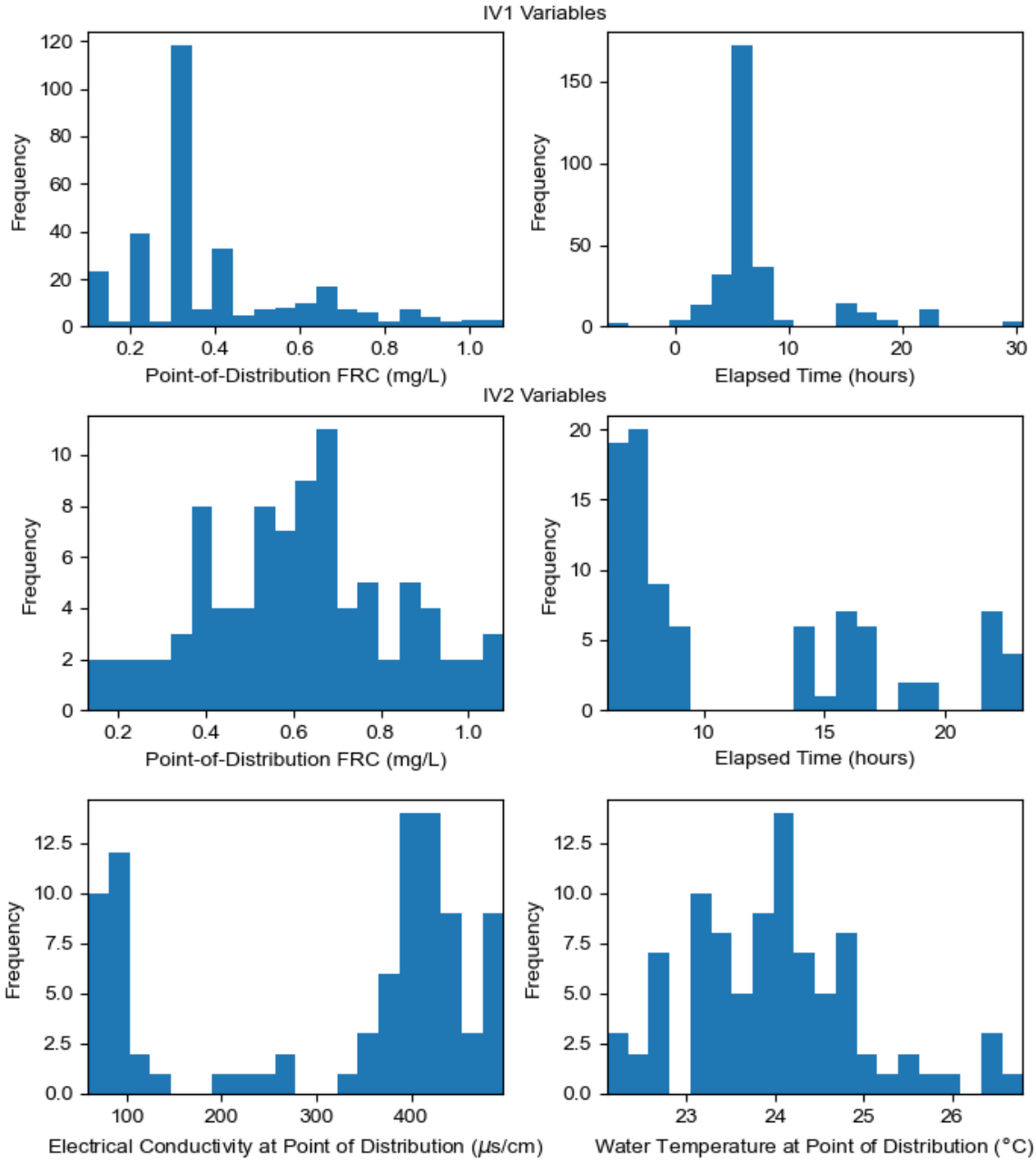

Figure D-2: Input variable histograms for Tanzania showing IV1 and IV2 input variables

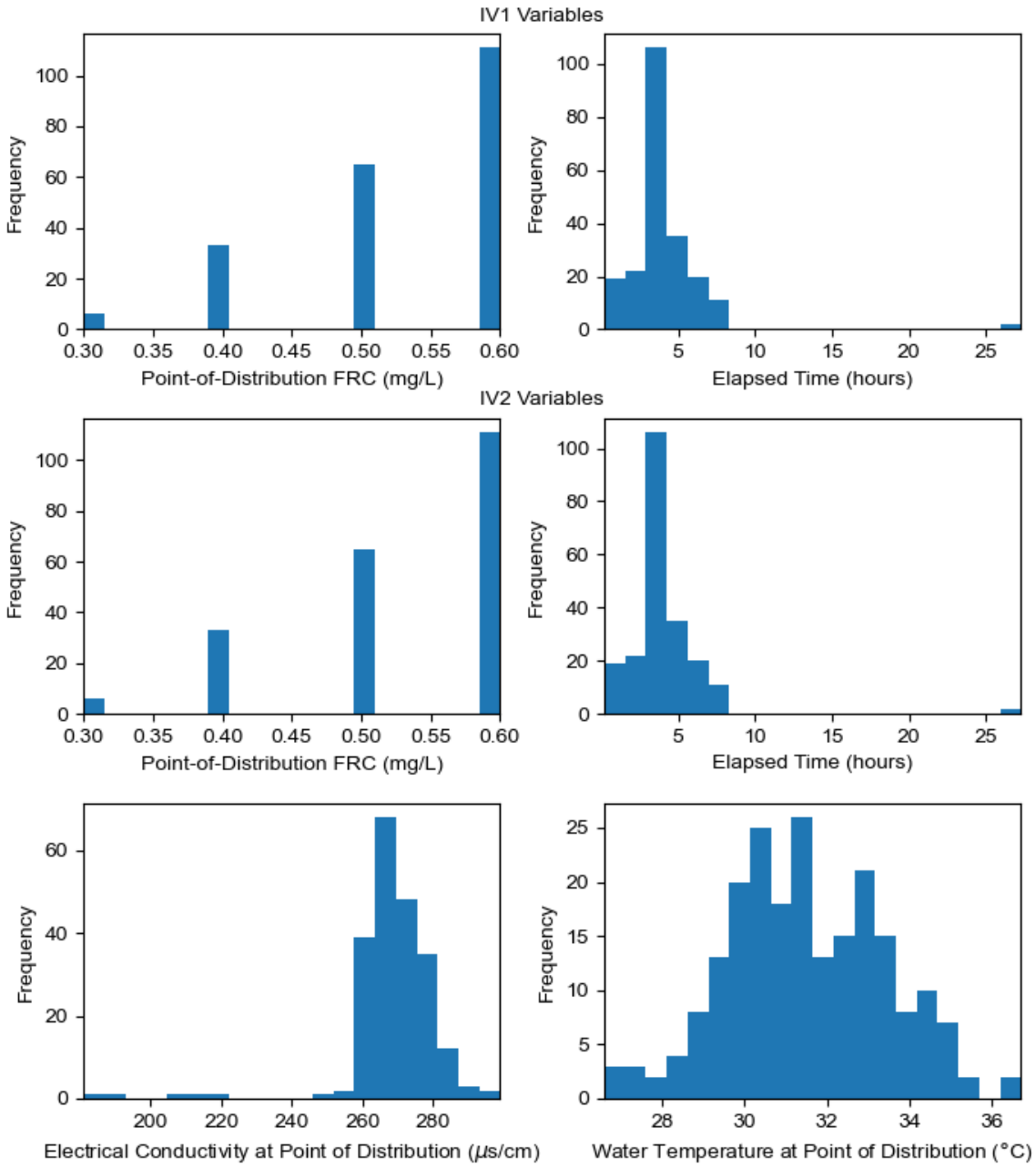

Figure D-3: Input variable histograms for Nigeria showing IV1 and IV2 input variables

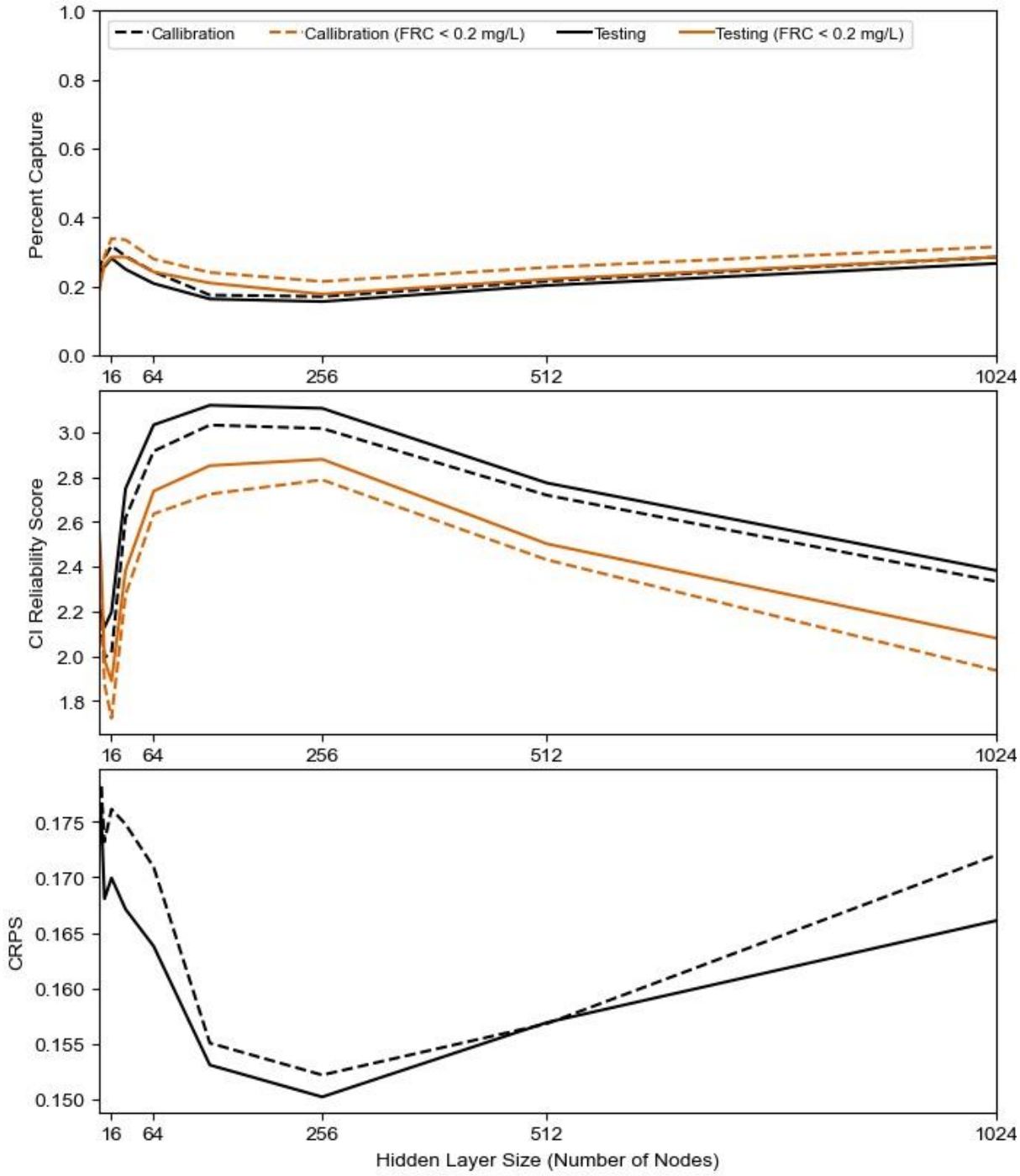

Figure D-4: Bangladesh IV1 hidden layer size selection. Hidden layer size of 16 selected as it simultaneously delivers best Percent Capture and CI reliability score.

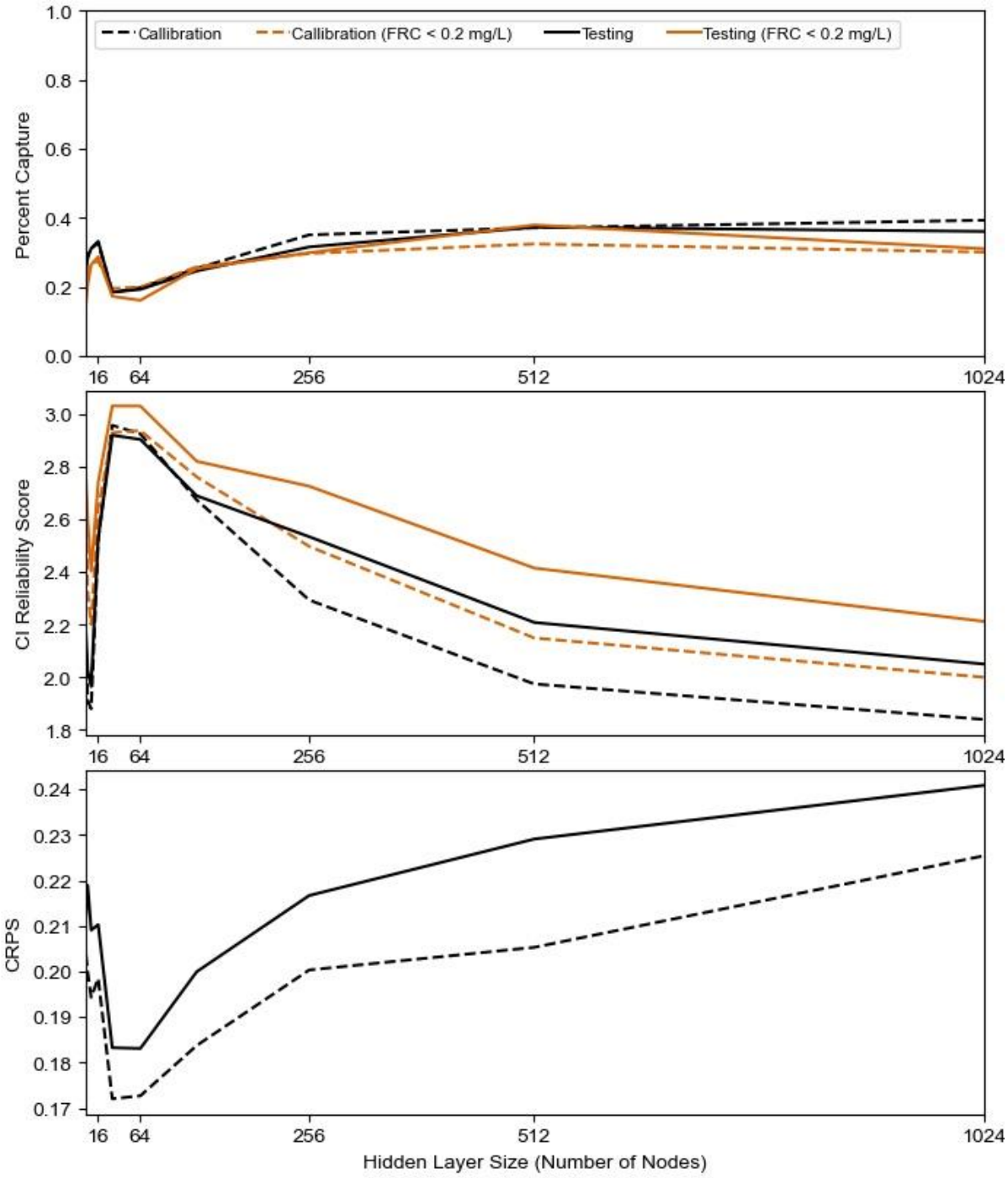

Figure D-5: Bangladesh IV2 hidden layer size selection. Hidden layer size of 16 selected as it simultaneously delivers best Percent Capture and CI reliability score.

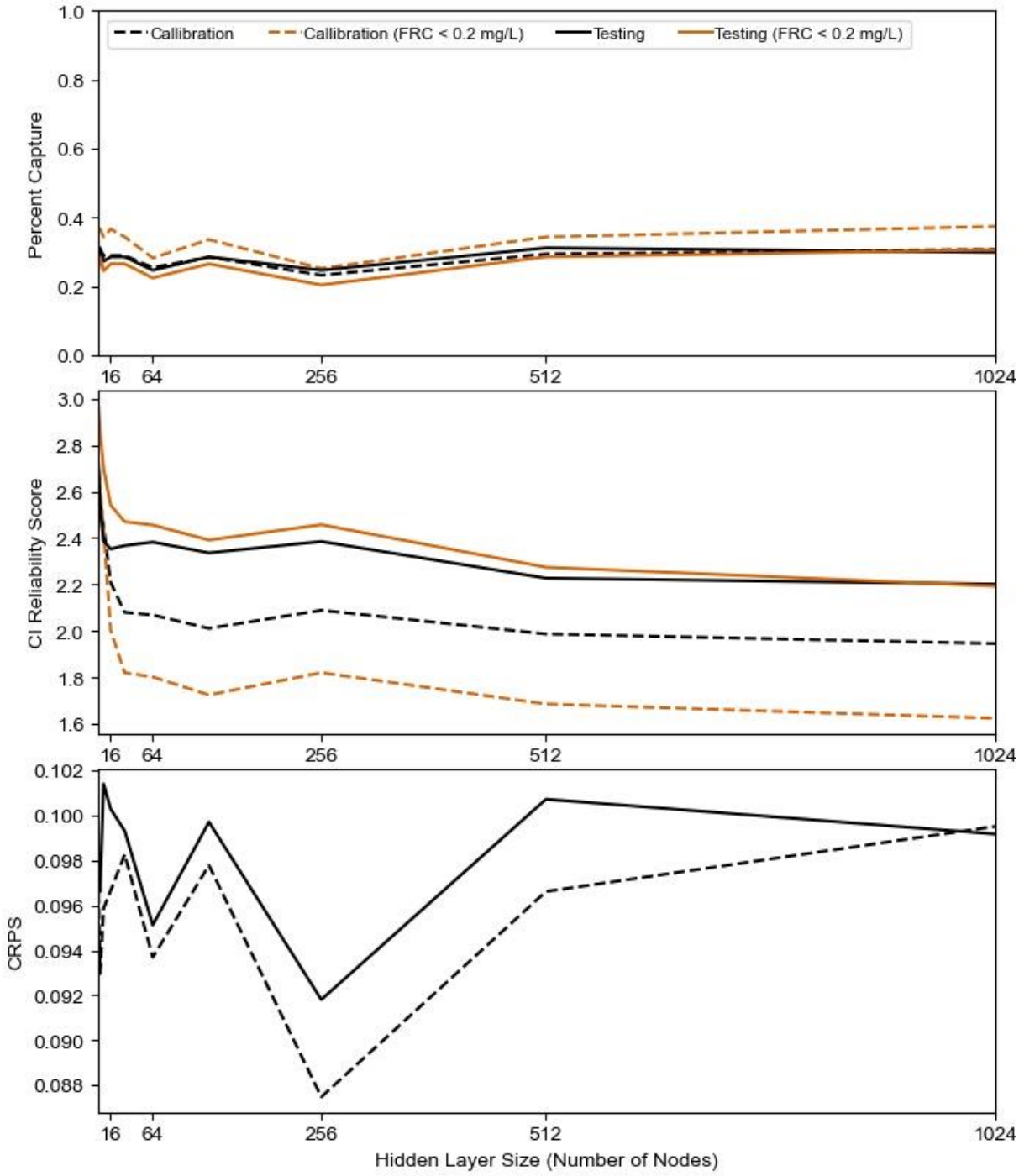

Figure D-6: Tanzania IV1 hidden layer size selection. Hidden layer size of 4 selected for good CI reliability performance without compromising capture and CRPS performance.

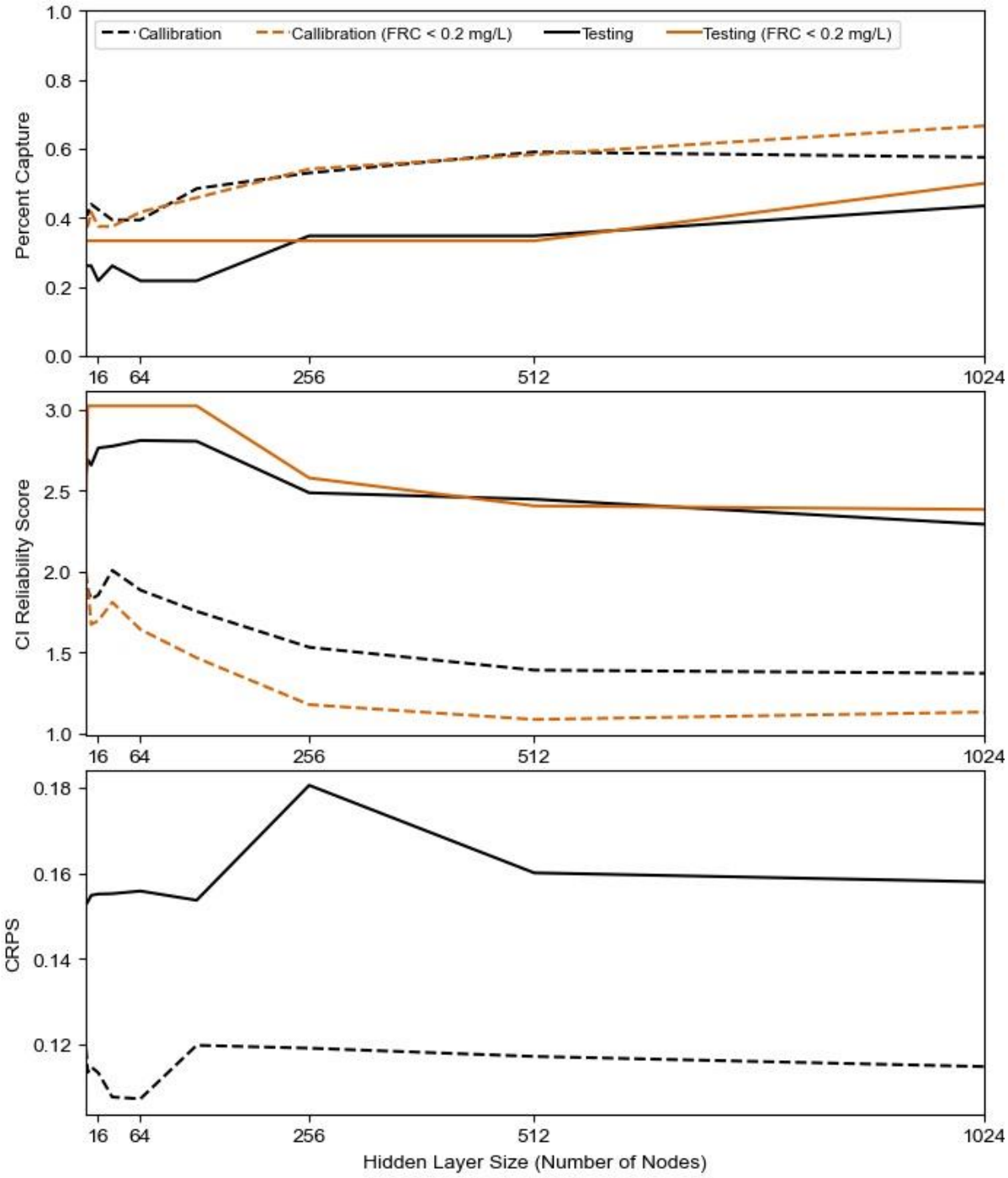

Figure D-7: Tanzania IV2 hidden layer size selection. Hidden layer size of 8 selected for best CI reliability and Percent Capture relative to hidden layer size.

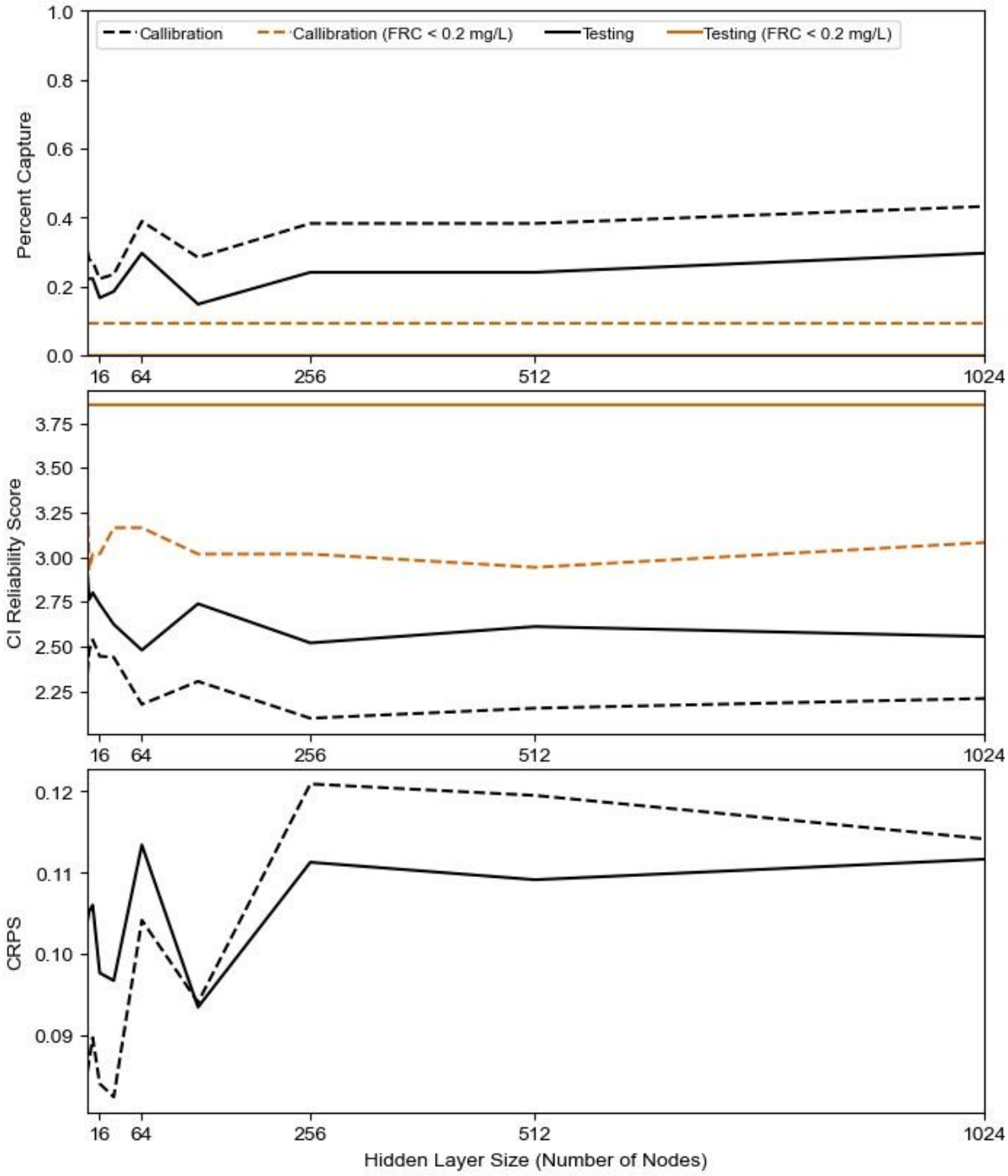

Figure D-8: Nigeria IV1 hidden layer size selection. Hidden layer size of 4 selected for best performance before performance drop.

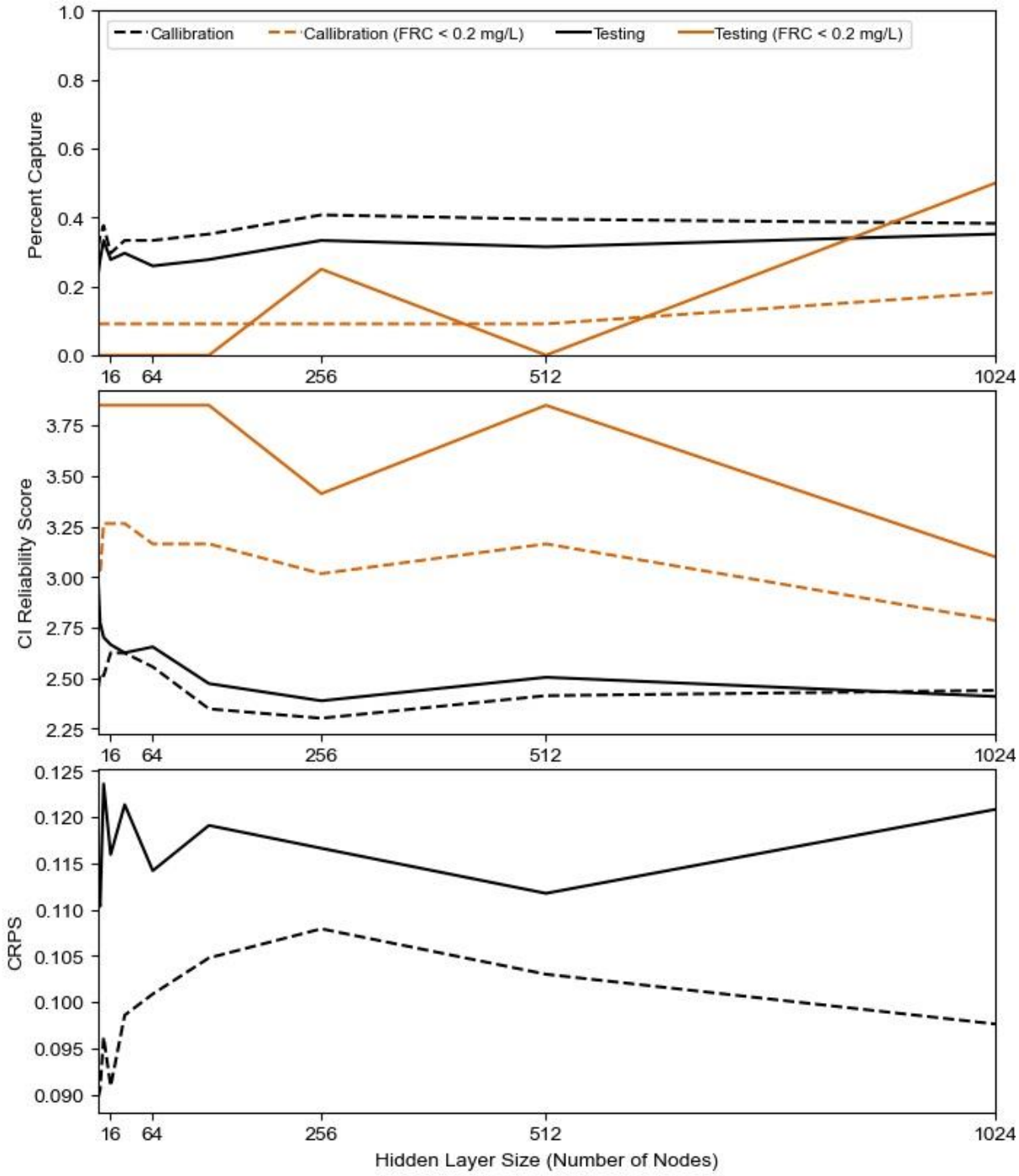

Figure D-9: Nigeria IV2 hidden layer size selection. Hidden layer size of 8 selected for best performance relative to hidden layer size.

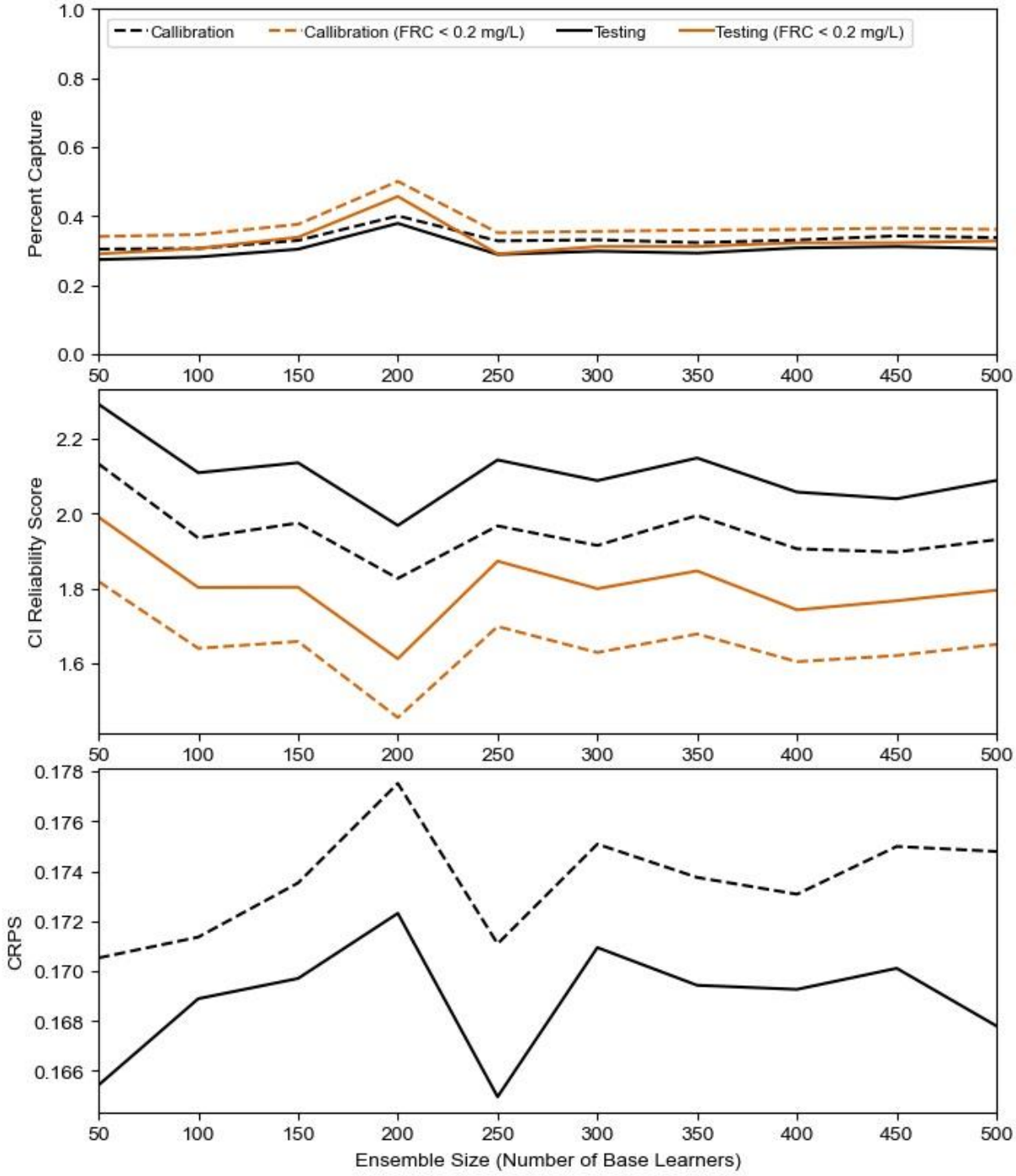

Figure D-10: Bangladesh IV1 ensemble size selection. Best performance at ensemble size of 200, so ensemble size of at least 200 required.

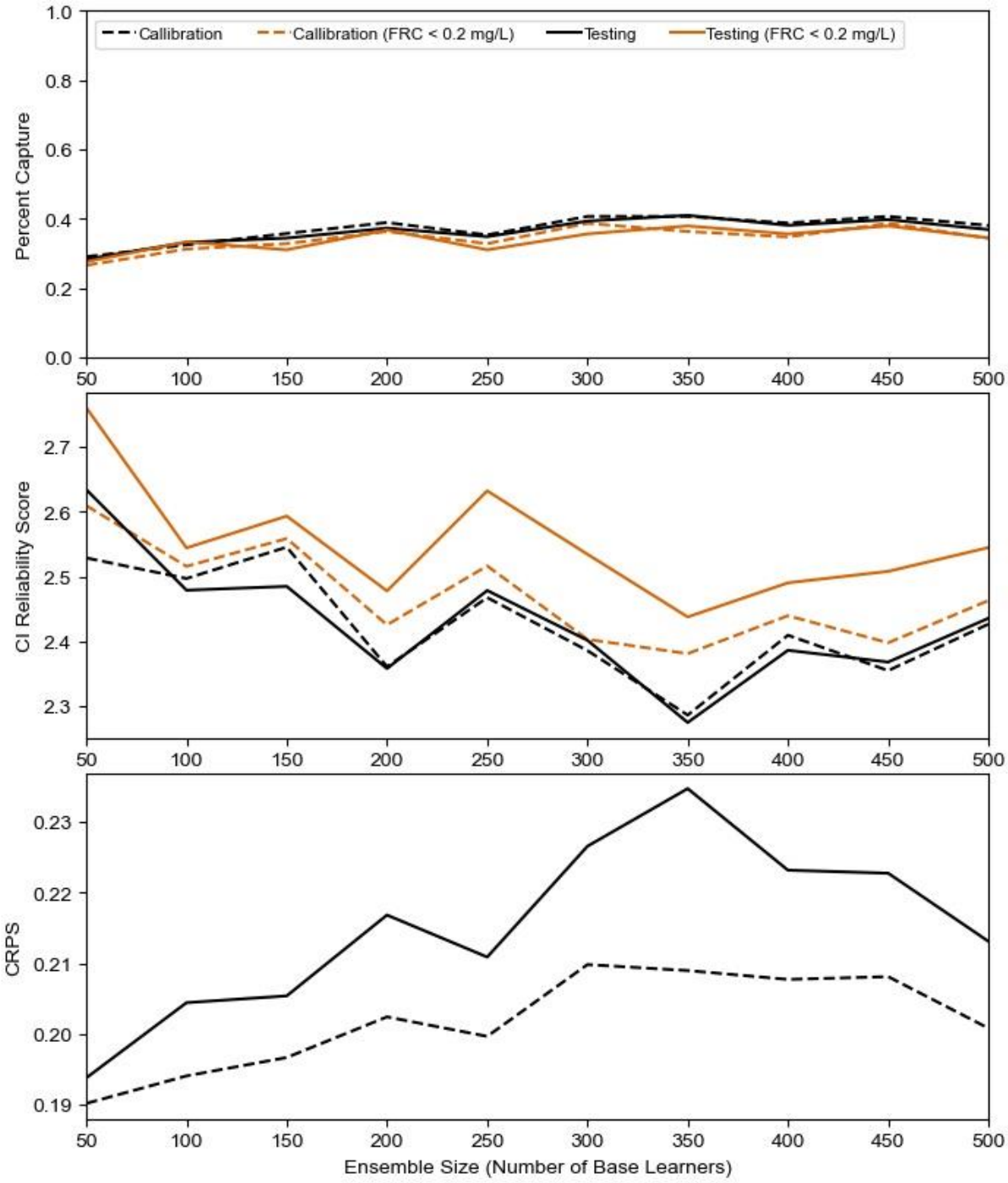

Figure D-11: Bangladesh IV2 ensemble size selection. Largest performance increase is up to 200 ensemble members, beyond this CI reliability and Percent Capture do not improve substantially, but CRPS worsens substantially.

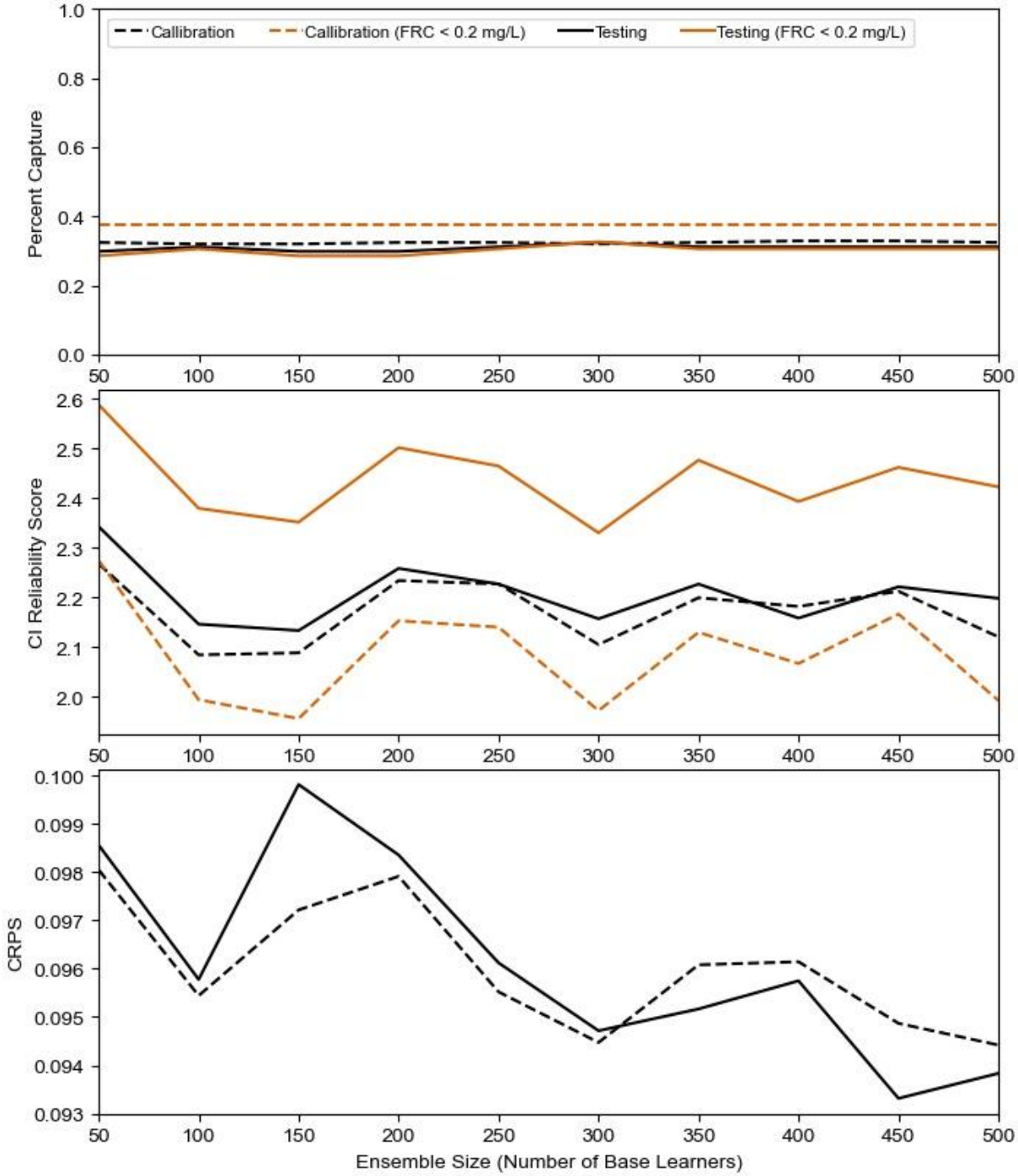

Figure D-12: Tanzania IV1 ensemble size selection. Substantial performance improvement when ensemble size is greater than 150, though beyond that, performance is highly variable.

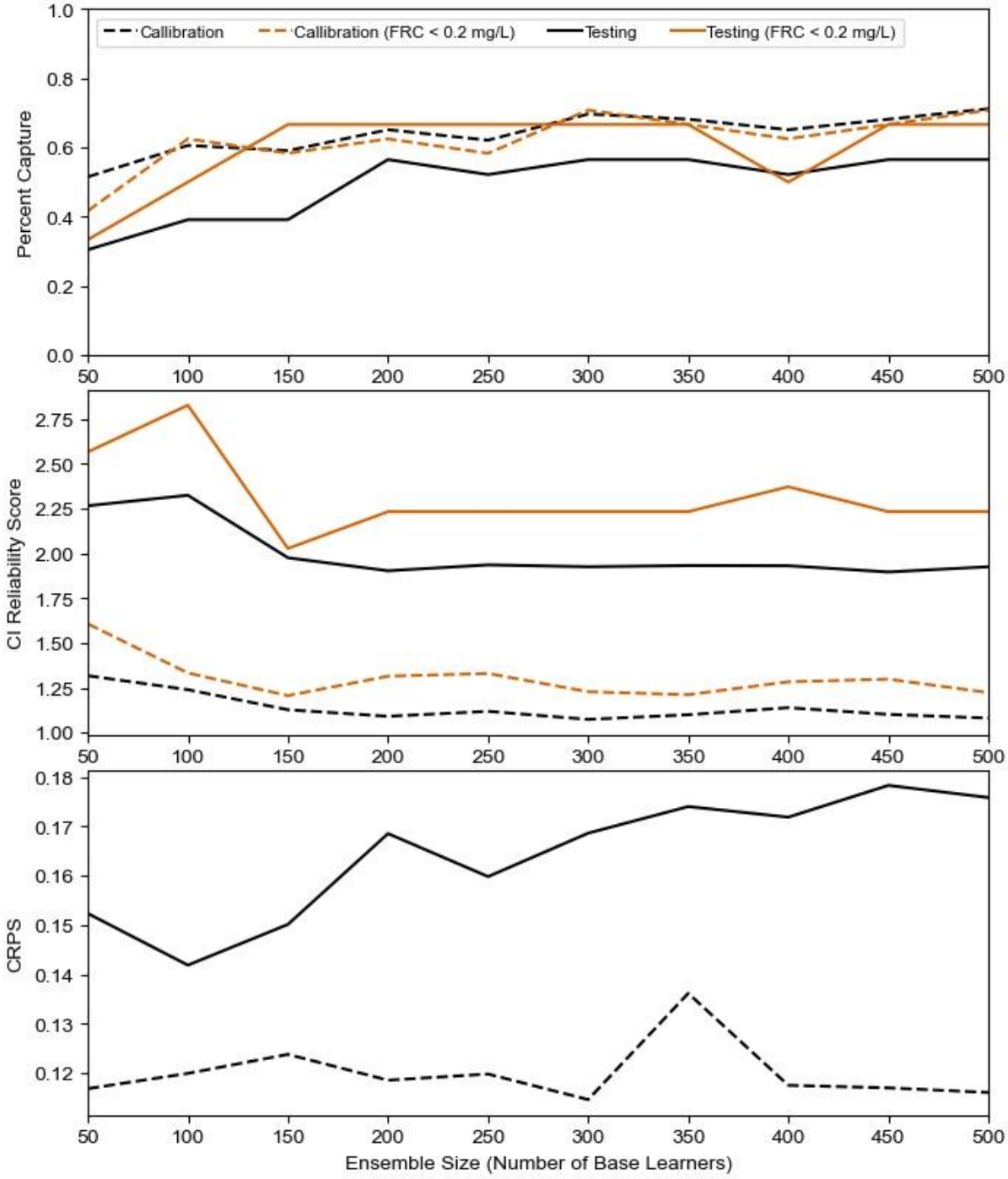

Figure D-13: Tanzania IV2 ensemble size selection. Substantial performance improvement up to ensemble size of 150, beyond this CRPS worsens without increasing capture or CI reliability.

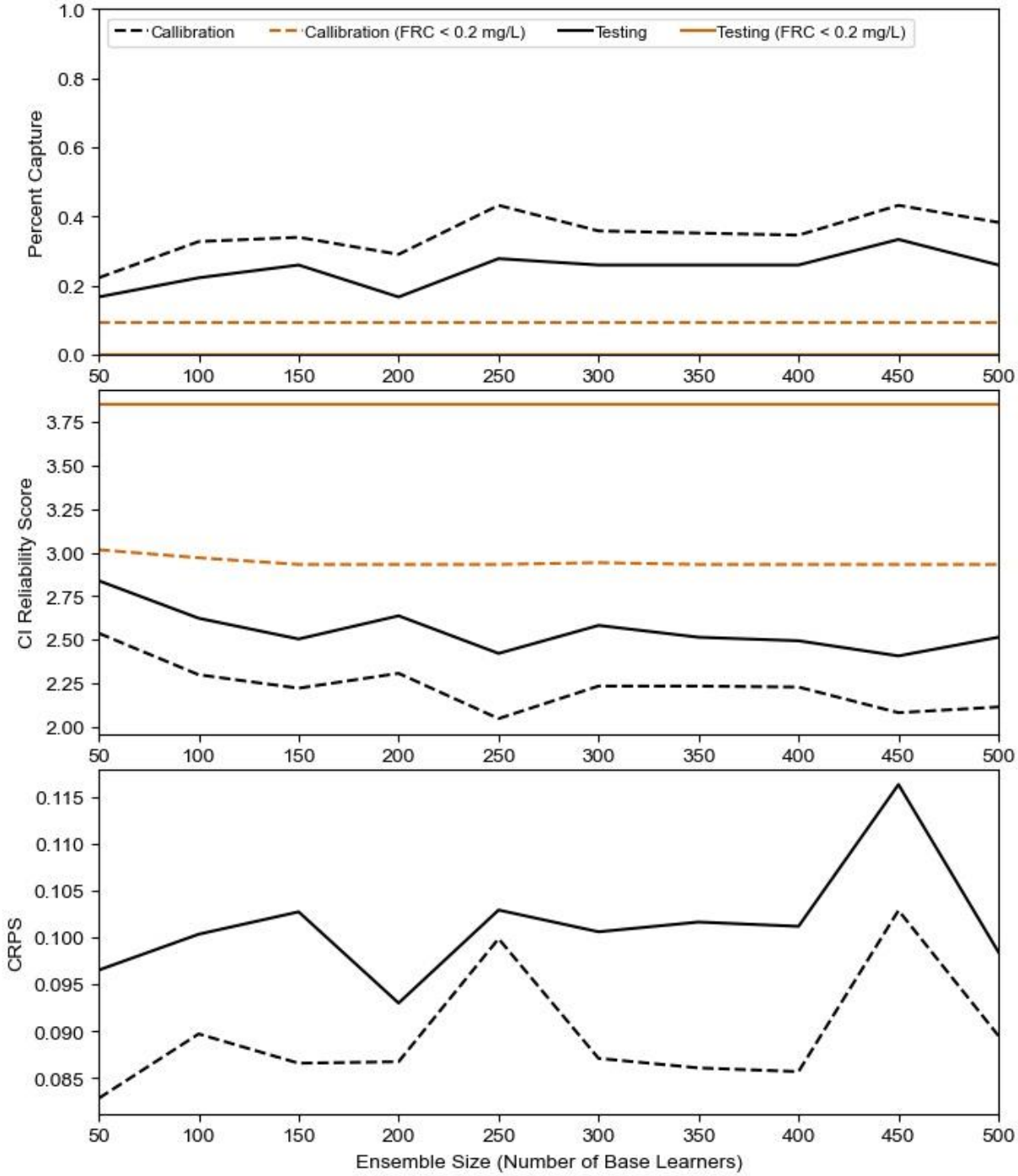

Figure D-14: Nigeria IV1 ensemble size selection. Performance relatively stable, very gradual improvement with increasing ensemble size for CI reliability and Percent Capture, but CRPS tends to worse with increasing ensemble size. Local best found at 200 members.

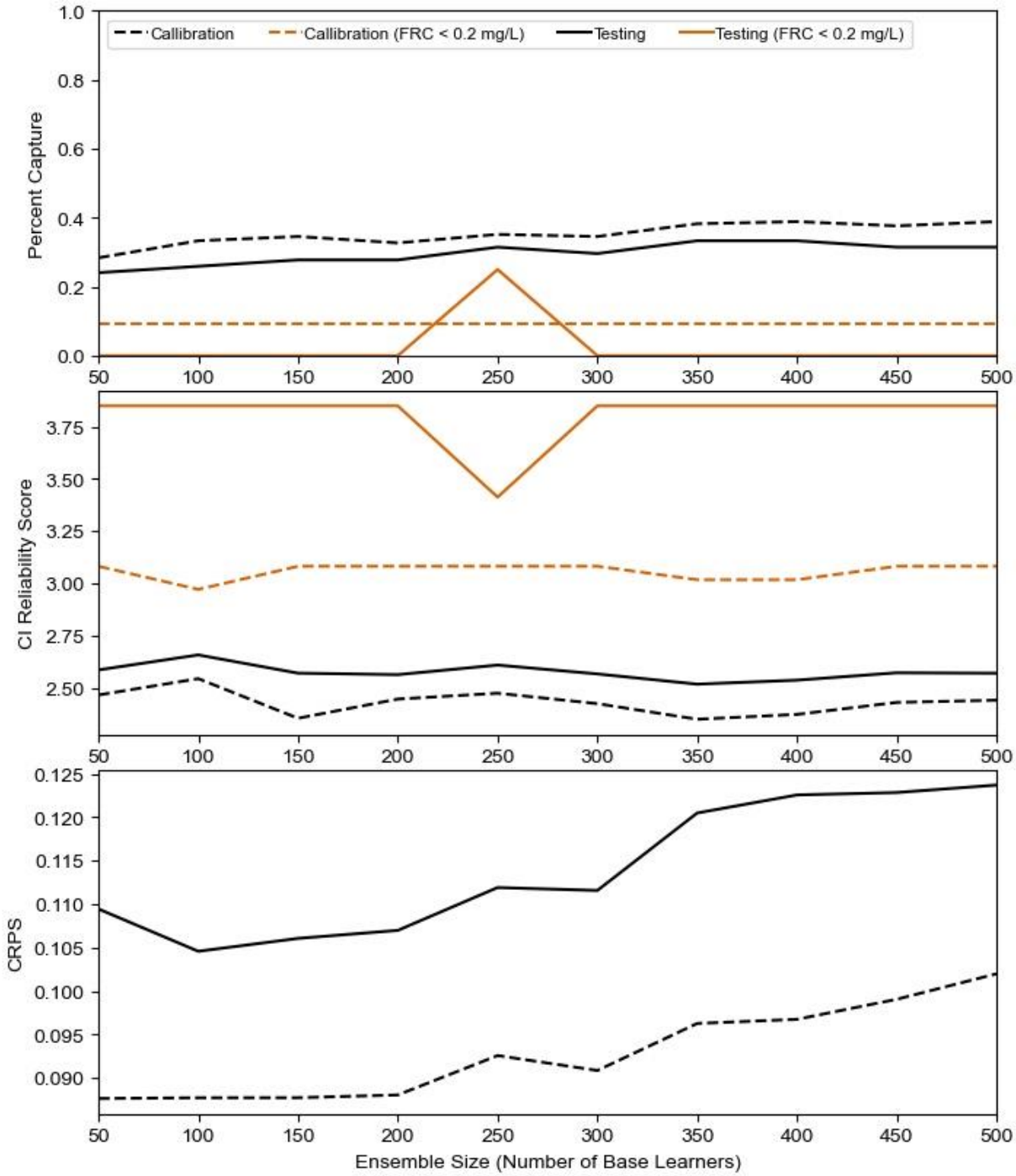

Figure D-15: Nigeria IV2 ensemble size selection. Performance relatively stable, though CRPS drastically worsens above ensemble size of 200 members.

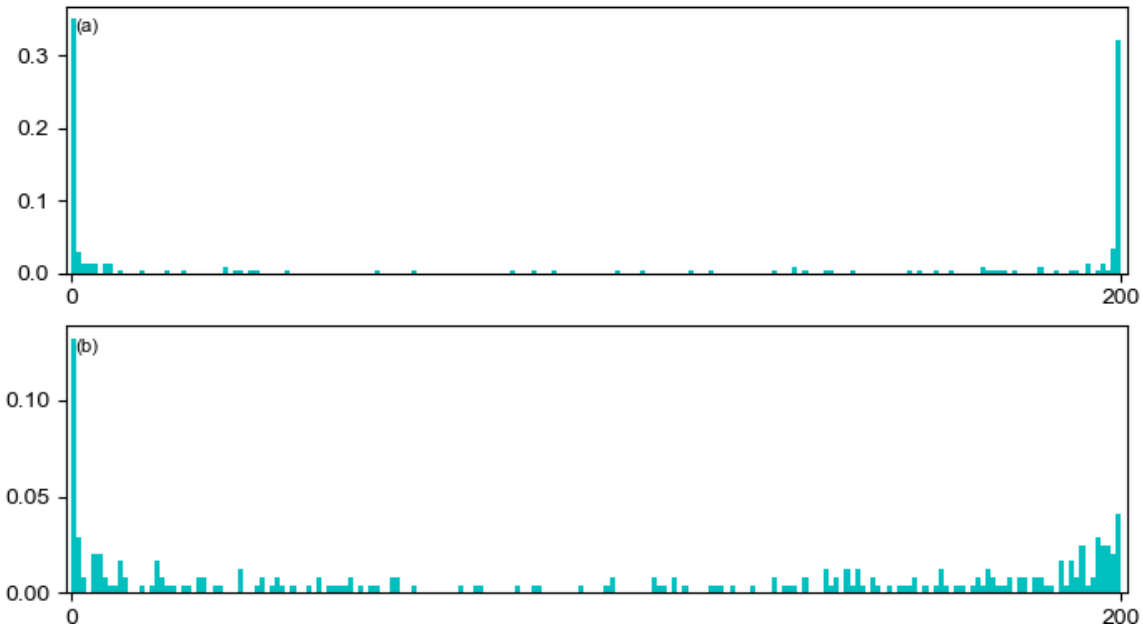

Figure D-16: Rank histogram (RH) for Bangladesh with the IV2 variable combination. The RH in (a) shows the results for the models trained with unweighted MSE, and (b) shows the model with the best δ-score. Both RHs are underdispersed, as seen from the u-shape of the RH, but the size of the bars at the extreme end are smaller in (b) and the inner bars are larger, indicating improved reliability with alternative cost functions and cost function weighting.

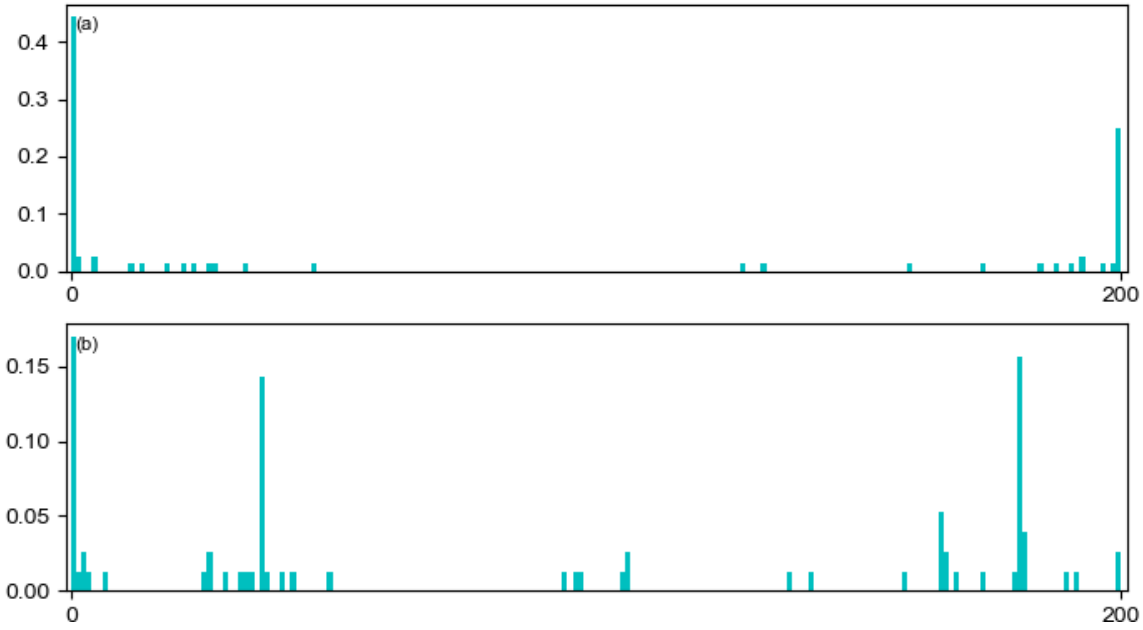

Figure D-17: RH for Tanzania with the IV1 variable combination. The RH in (a) shows the results for the models trained with unweighted MSE, and (b) shows the model with the best δ-score. Both RHs are underdispersed, as seen from the u-shape of the RH, but the size of the bars at the extreme end are smaller in (b) and the inner bars are larger, with some even comparable to the outer bars, indicating improved reliability with alternative cost functions and cost function weighting. Due to the smaller number of observations, many ranks have no bars.

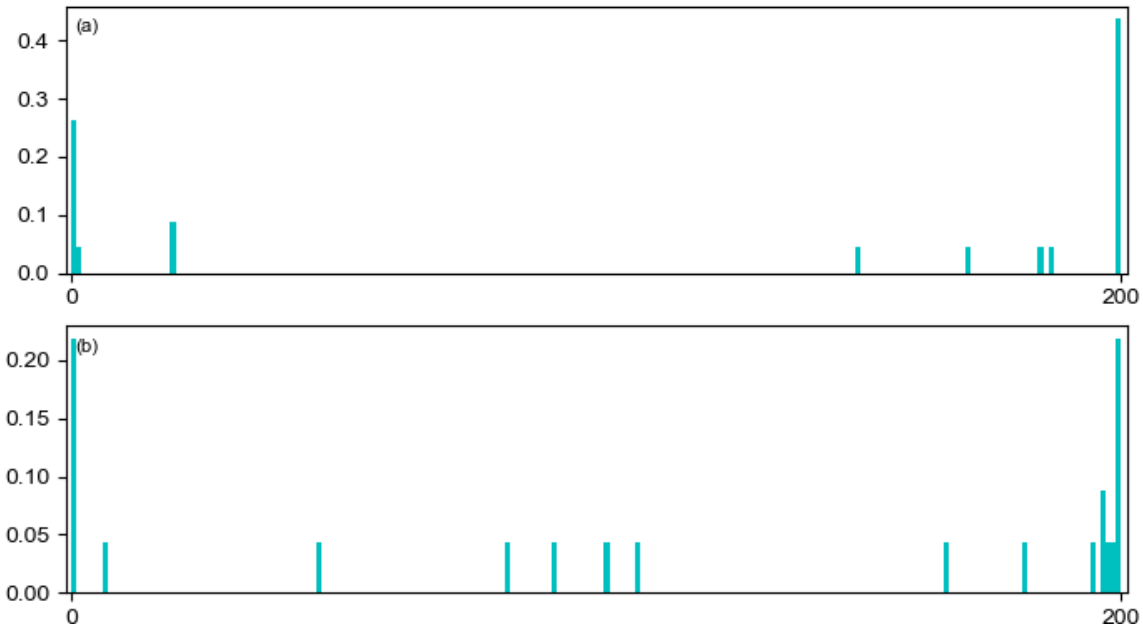

Figure D-18: RH for Tanzania with the IV2 variable combination. The RH in (a) shows the results for the models trained with unweighted MSE, and (b) shows the model with the best δ-score. Both RHs are underdispersed, as seen from the u-shape of the RH, but the RH in (b) shows less underdispersion, as well as less skewness, with the extreme bars appearing more even.

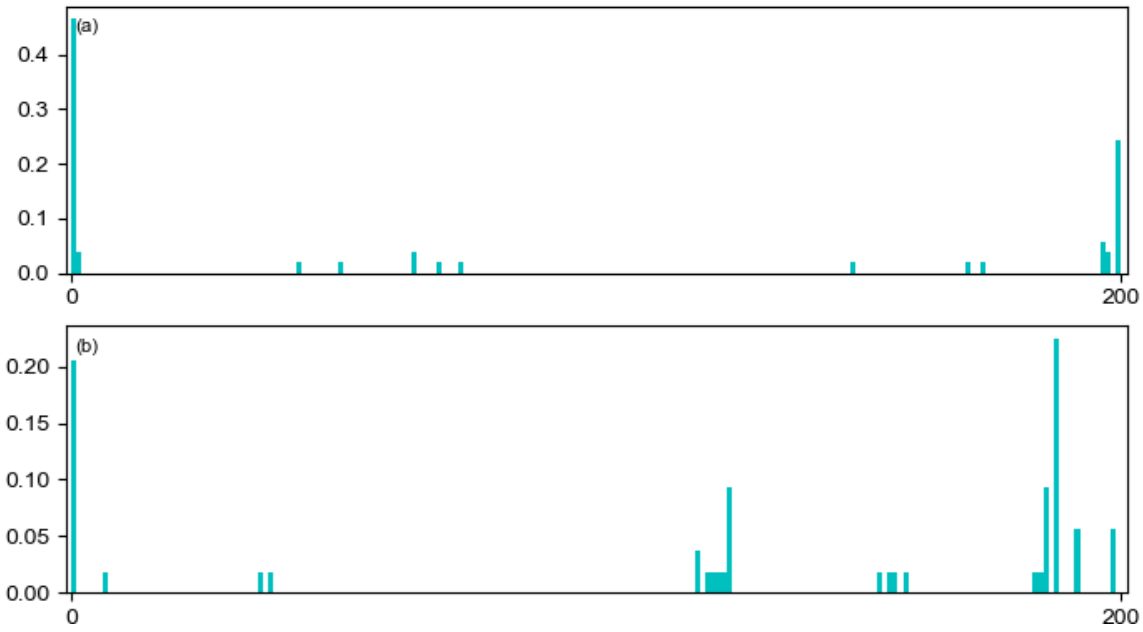

Figure D-19: RH for Nigeria with the IV1 variable combination. The RH in (a) shows the results for the models trained with unweighted MSE, and (b) shows the model with the best δ-score. Both RHs are underdispersed, as seen from the u-shape of the RH, but the RH in (b) shows less underdispersion, as well as less skewness, with the extreme bars appearing more even.

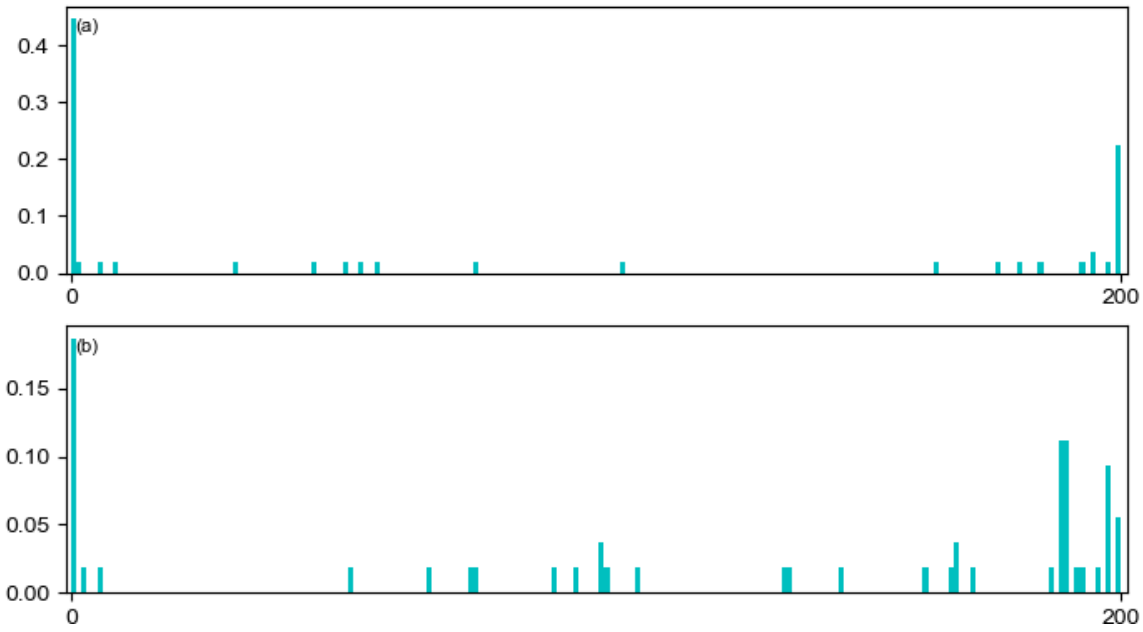

Figure D-20: RH for Nigeria with the IV2 variable combination. The RH in (a) shows the results for the models trained with unweighted MSE, and (b) shows the model with the best δ-score. Both RHs are underdispersed, as seen from the u-shape of the RH, but the RH in (b) shows less underdispersion, though the results remain skewed.

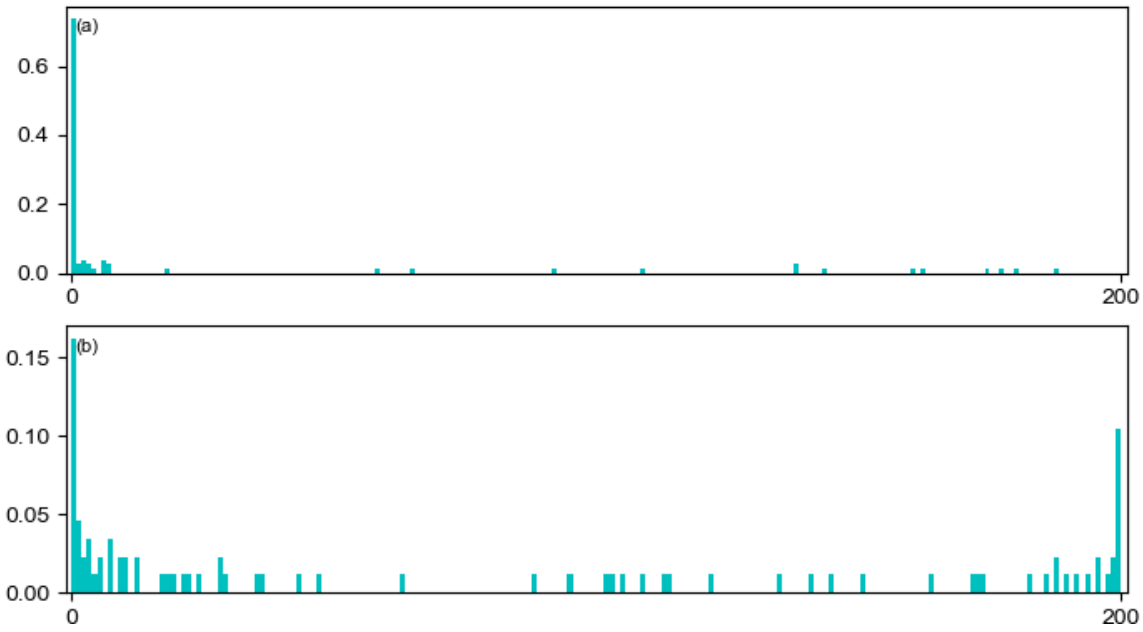

Figure D-21: Rank histogram (RH) for observations with point-of-consumption FRC below 0.2 mg/L for Bangladesh with the IV2 variable combination. The RH in (a) shows the results for the models trained with unweighted MSE, and (b) shows the model with the best $\delta_{<0.2}$-score. As with the overall dataset, both RHs are underdispersed but the RH shown in (b) is less underdispersed than that in (a), indicating improved reliability with alternative cost functions and cost function weighting.

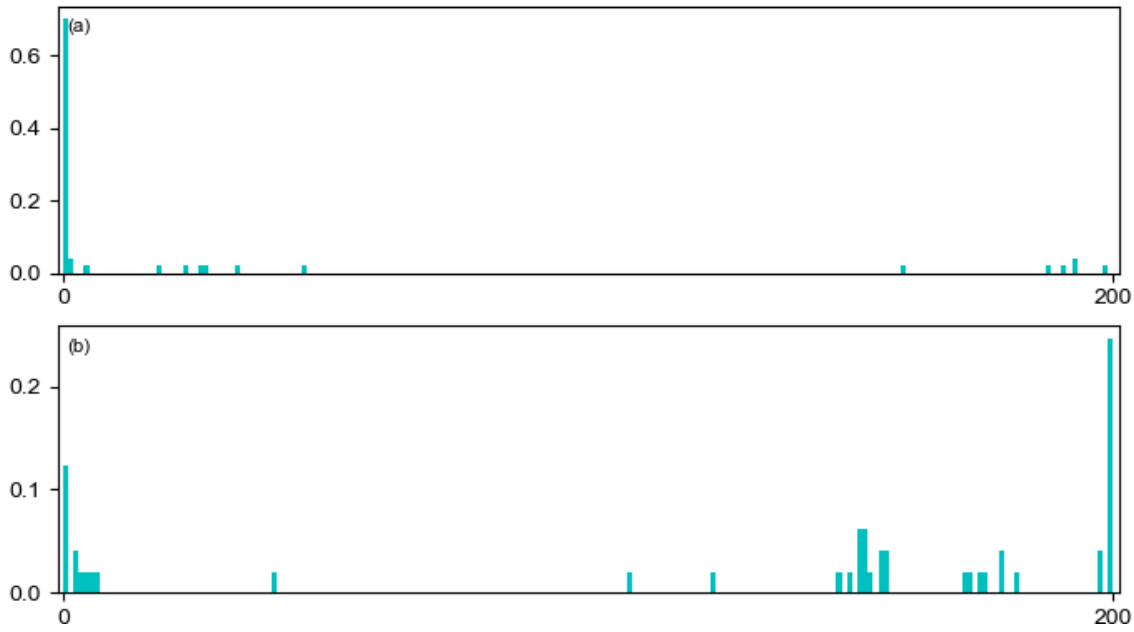

Figure D-22: Rank histogram (RH) for observations with point-of-consumption FRC below 0.2 mg/L for Tanzania with the IV1 variable combination. The RH in (a) shows the results for the models trained with unweighted MSE, and (b) shows the model with the best $\delta_{<0.2}$-score. Both RHs are underdispersed, but the RH in (a) is also heavily skewed towards overprediction, as shown with the large bar at the 0 rank. The use of alternate cost functions and weighting both improves the dispersion and the skewness, as shown in (b).

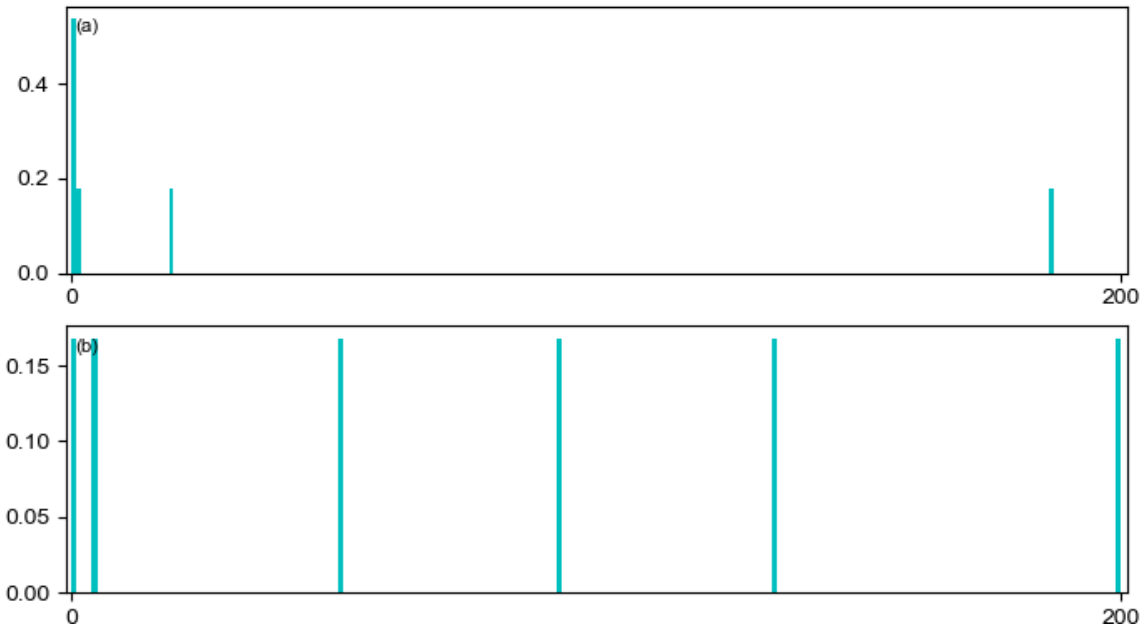

Figure D-23: Rank histogram (RH) for observations with point-of-consumption FRC below 0.2 mg/L for Tanzania with the IV2 variable combination. The RH in (a) shows the results for the models trained with unweighted MSE, and (b) shows the model with the best $\delta_{<0.2}$-score. The RH in (a) is both underdispersed and skewed but the use of alternate cost functions and weighting both improves the dispersion and the skewness, as shown in (b).

Figure D-24: Rank histogram (RH) for observations with point-of-consumption FRC below 0.2 mg/L for Nigeria with the IV1 variable combination. The RH in (a) shows the results for the models trained with unweighted MSE, and (b) shows the model with the best $\delta_{<0.2}$-score. Since the baseline model did not capture any observations with household FRC below 0.2 mg/L, all observations are in the 0 bar for the RH in (a), but better dispersion is shown in (b).

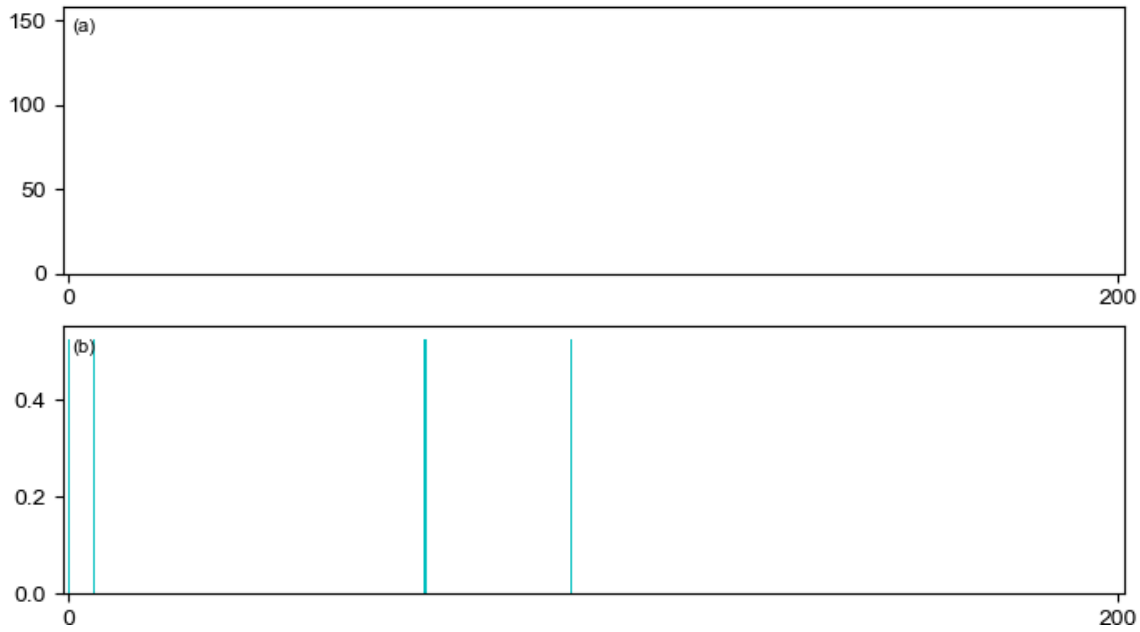

Figure D-25: Rank histogram (RH) for observations with point-of-consumption FRC below 0.2 mg/L for Nigeria with the IV2 variable combination. The RH in (a) shows the results for the models trained with unweighted MSE, and (b) shows the model with the best $\delta_{<0.2}$-score. Since the baseline model did not capture any observations with household FRC below 0.2 mg/L, all observations are in the 0 bar for the RH in (a), but better dispersion is shown in (b).