npj | Clean Water

www.nature.com/npjcleanwater

ARTICLE OPEN

Check for updates

Forecasting point-of-consumption chlorine residual in refugee settlements using ensembles of artificial neural networks

Michael De Santi[1], Usman T. Khan[1], Matthew Arnold[2], Jean-François Fesselet[3] and Syed Imran Ali[1,2,3]✉

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.

npj Clean Water (2021)4:35 ; https://doi.org/10.1038/s41545-021-00125-2

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[1–4]. 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[5,6], Malawi[7], Sudan[8], South Sudan[9,10], and Uganda[11,12]. Residual chlorine protects against household recontamination by inactivating waterborne pathogens as they are introduced into stored drinking water. 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[13–18]. 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 Fig. 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.

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 Fig. 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[19]. 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[20–23]. 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[22,23]. ANNs are also effective even when only using data collected through routine monitoring[21,24], which is particularly

[1]Department of Civil Engineering, Lassonde School of Engineering, York University, Toronto, ON, Canada. [2]Dahdaleh Institute for Global Health Research, York University, Toronto, ON, Canada. [3]Public Health Department, Médecins Sans Frontières, Amsterdam, The Netherlands. ✉email: siali@yorku.ca

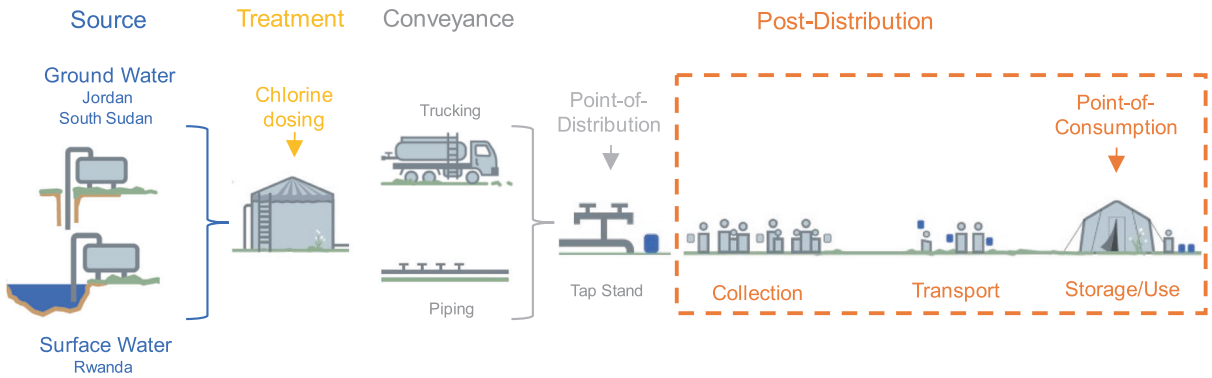

Fig. 1 Post-distribution period shown in context of overall water supply system for typical refugee or IDP settlement. Water obtained from ground or surface water is centrally treated then conveyed via piped distribution system to the tap stand (point-of-distribution). The post-distribution period begins 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).

beneficial in humanitarian settings where detailed lab-based water quality evaluations may not be available[25]. 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[26,27], providing an important improvement in contrast to past, deterministic, 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[26,28], 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 Fig. 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[29–31]. 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.

RESULTS

Ensemble model performance

Table 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[30–32]. From Table 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. Fig. 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 percentage of observations, 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. Fig. 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 1. Figure 2 also shows that for all sites other than South Sudan, the models using the IV2 variable combination produced forecasts with better capture across multiple CIs, leading to a

Table 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

substantial improvement in reliability that is reflected in the CI reliability scores documented in Table 1.

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 Supplementary Figs 1–4 in the Supplementary Information.

South Sudan

Figure 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 1.

The clearest trend between the observed and forecasted point-of-consumption FRC and the input variables shown in Fig. 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 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.

Jordan (2014)

Figure 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, producing 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 1.

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.

Jordan (2015)

Figure 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.

Rwanda

Figure 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 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 1.

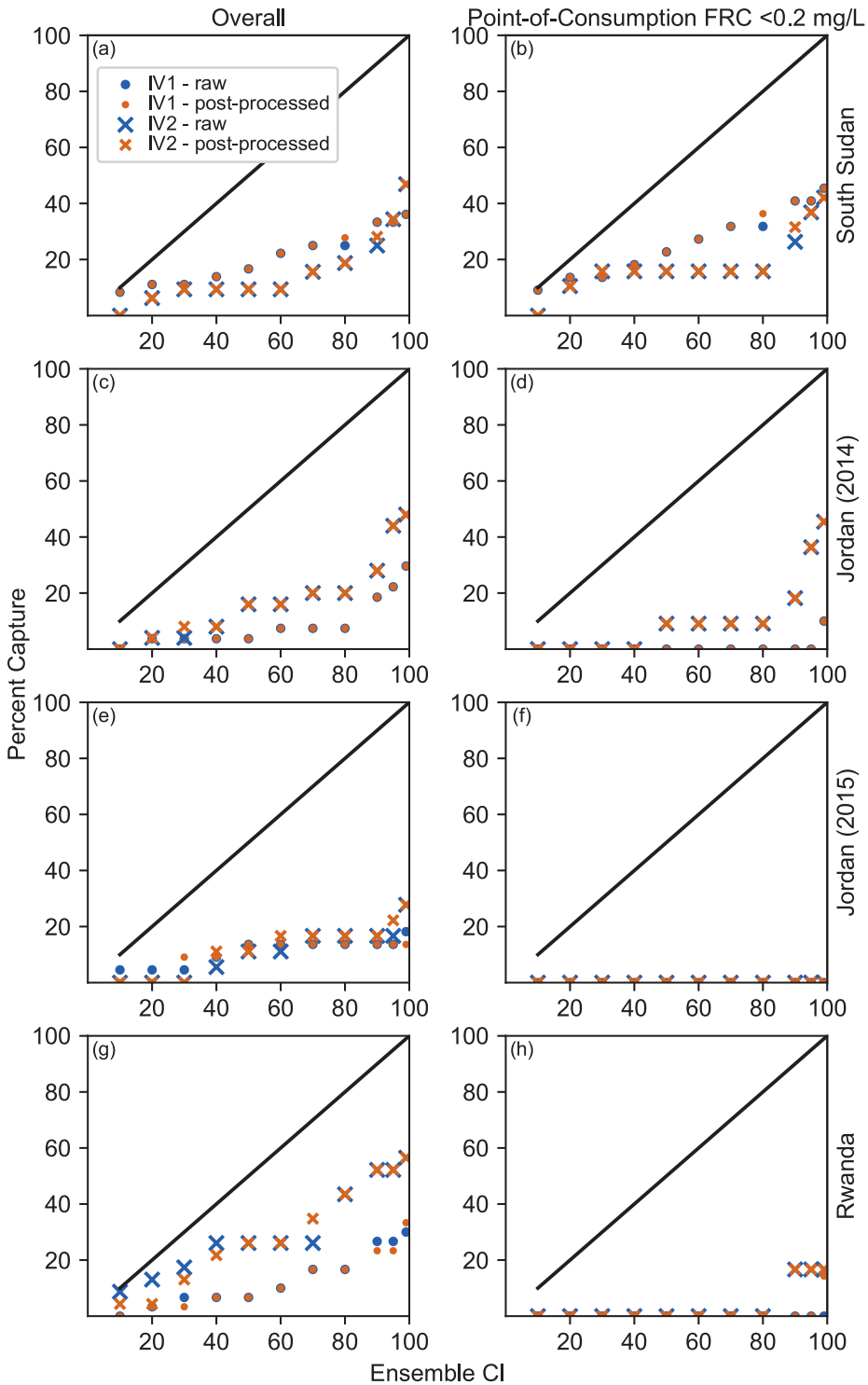

Fig. 2 Confidence interval reliability diagrams for all sites. Raw and post-processed CI reliability diagrams for all sites for both the overall dataset (**a** South Sudan, **c** Jordan (2014), **e** Jordan (2015), **g** Rwanda) and for observations where point-of-consumption FRC is below 0.2 mg/L (**b** South Sudan, **d** Jordan (2014), **f** Jordan (2015), **h** Rwanda). 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.

From Fig. 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.

Partial correlation analysis results

Table 2 presents the results of a partial correlation analysis that was performed for each site to provide additional details on the

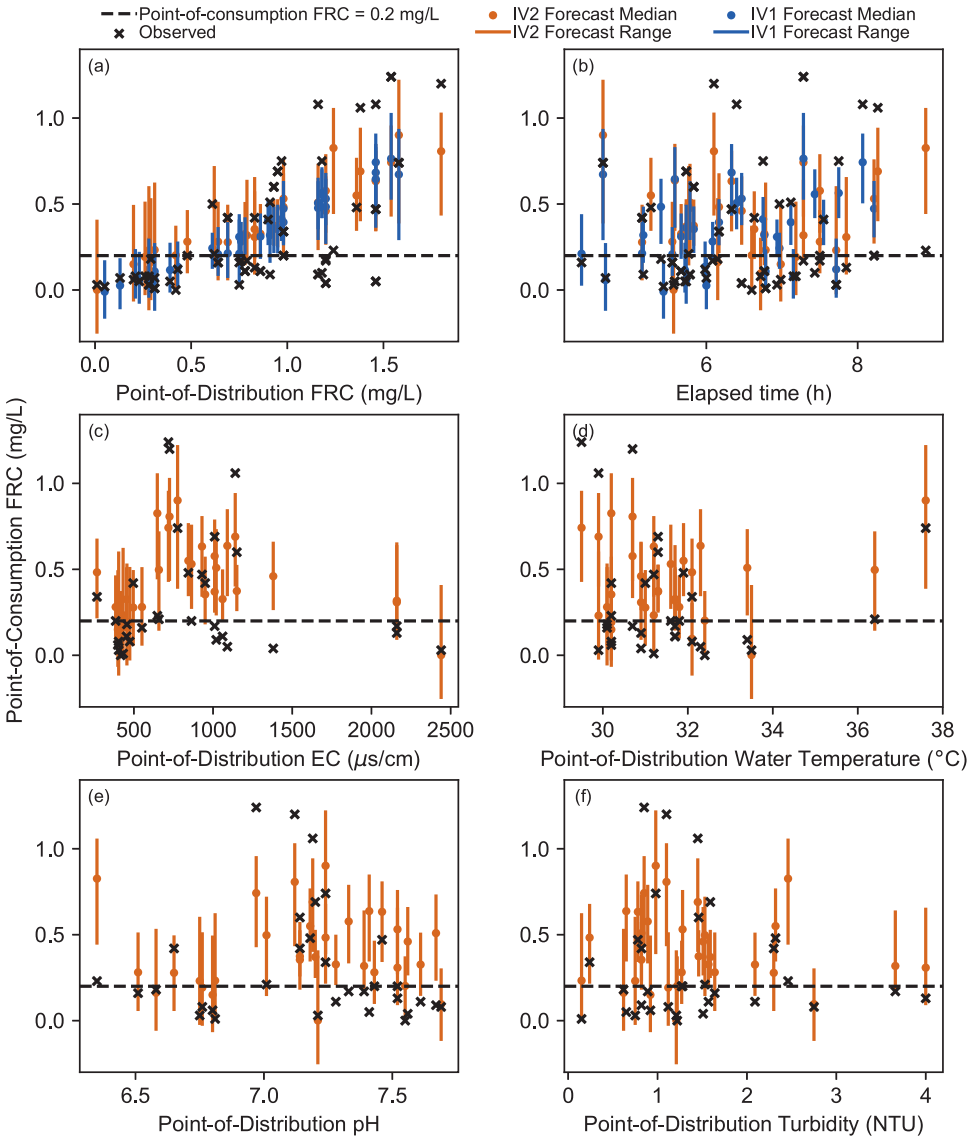

Fig. 3 South Sudan observations and post-processed forecasts of point-of-consumption FRC. The observations and forecasts are shown against **a** point-of-distribution FRC, **b** elapsed time, **c** point-of-distribution EC, **d** point-of-distribution water temperature, **e** point-of-distribution pH, **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.

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 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. 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 Figs 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[15,33–37]. 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.

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[30–32]. 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

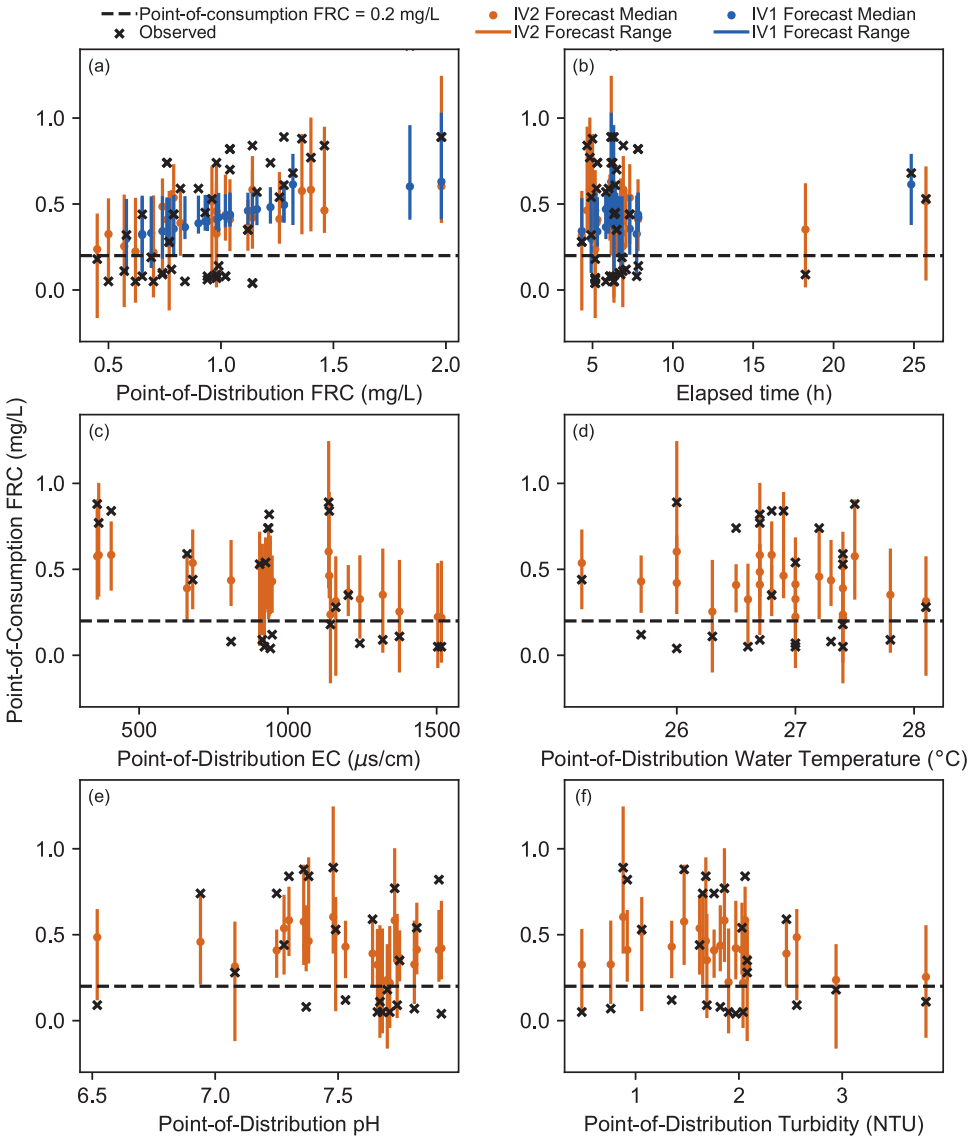

Fig. 4 Jordan (2014) observations and post-processed forecasts of point-of-consumption FRC. The observations and forecasts are shown against **a** point-of-distribution FRC, **b** elapsed time, **c** point-of-distribution EC, **d** point-of-distribution water temperature, **e** point-of-distribution 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.

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 Supplementary Table 1 in the Supplementary Information. 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 h for all sites and datasets except in South Sudan where we used a storage duration of 10 h 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[9,19]. 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 Figs 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 95th 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 Fig. 7. To achieve negligible risk of point-of-consumption FRC below 0.2 mg/L in South Sudan (Fig. 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) (Fig. 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

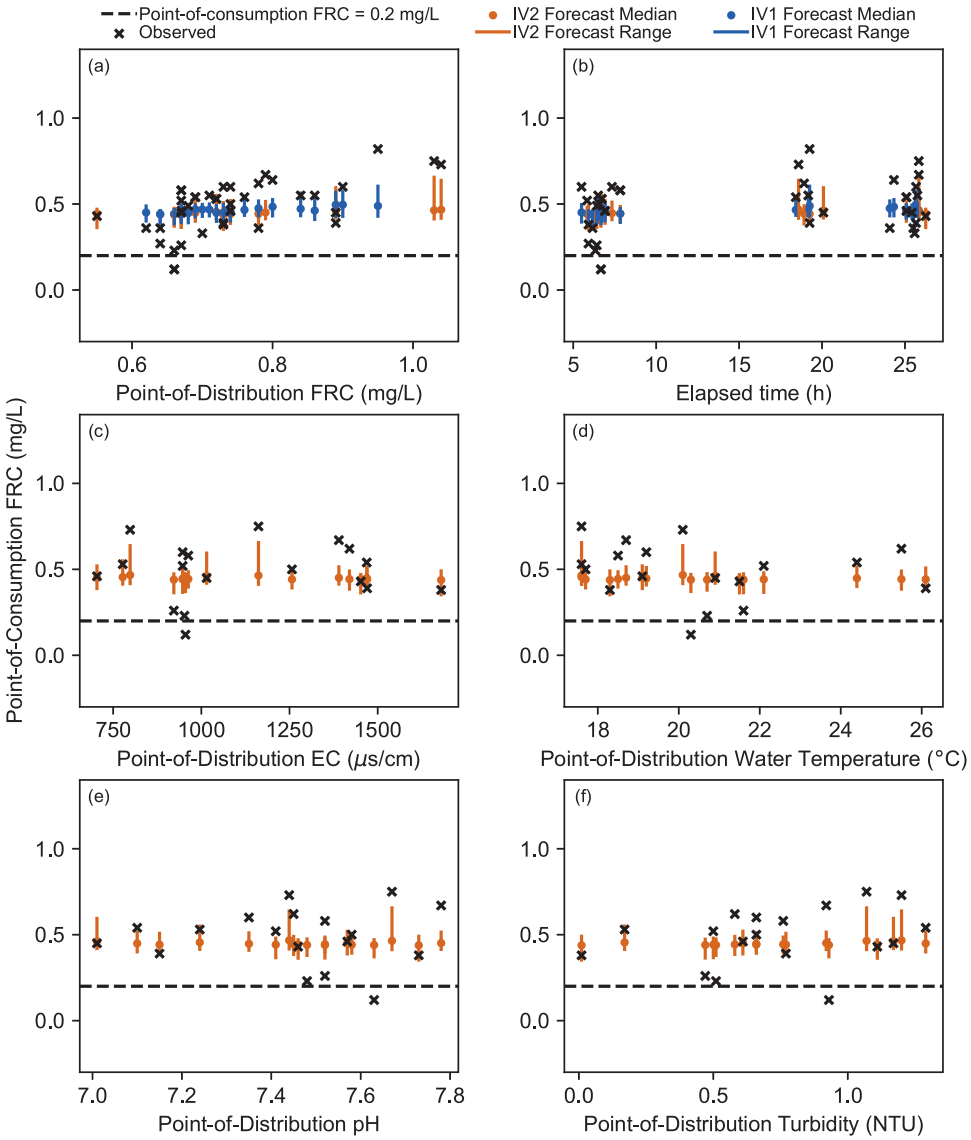

Fig. 5 Jordan (2015) observations and post-processed forecasts of point-of-consumption FRC. The observations and forecasts are shown against **a** point-of-distribution FRC, **b** elapsed time, **c** point-of-distribution EC, **d** point-of-distribution water temperature, **e** point-of-distribution 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.

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) (Fig. 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, (Fig. 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").

To provide additional context for the risk predictions, Figure 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 h for South Sudan, 20–28 h 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 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) (Fig. 8f) and for Rwanda (Fig. 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.

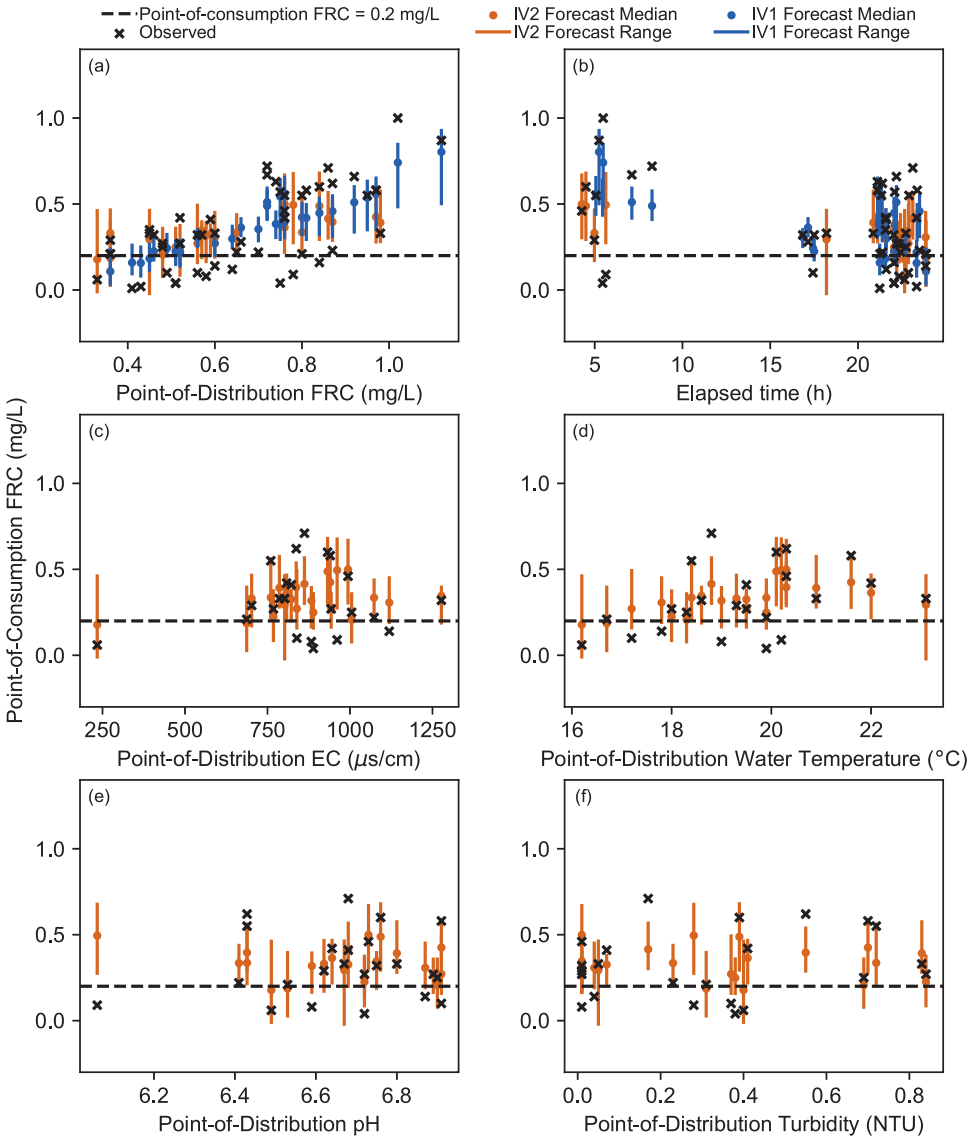

Fig. 6 Rwanda observations and post-processed forecasts of point-of-consumption FRC. The observations and forecasts are shown against **a** point-of-distribution FRC, **b** elapsed time, **c** point-of-distribution EC, **d** point-of-distribution water temperature, **e** point-of-distribution pH, **f** 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.

Table 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

DISCUSSION

The ensemble performance metrics listed in Table 1, as well as the results shown in Figs 2–6, 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[26,27]. 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 neither 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

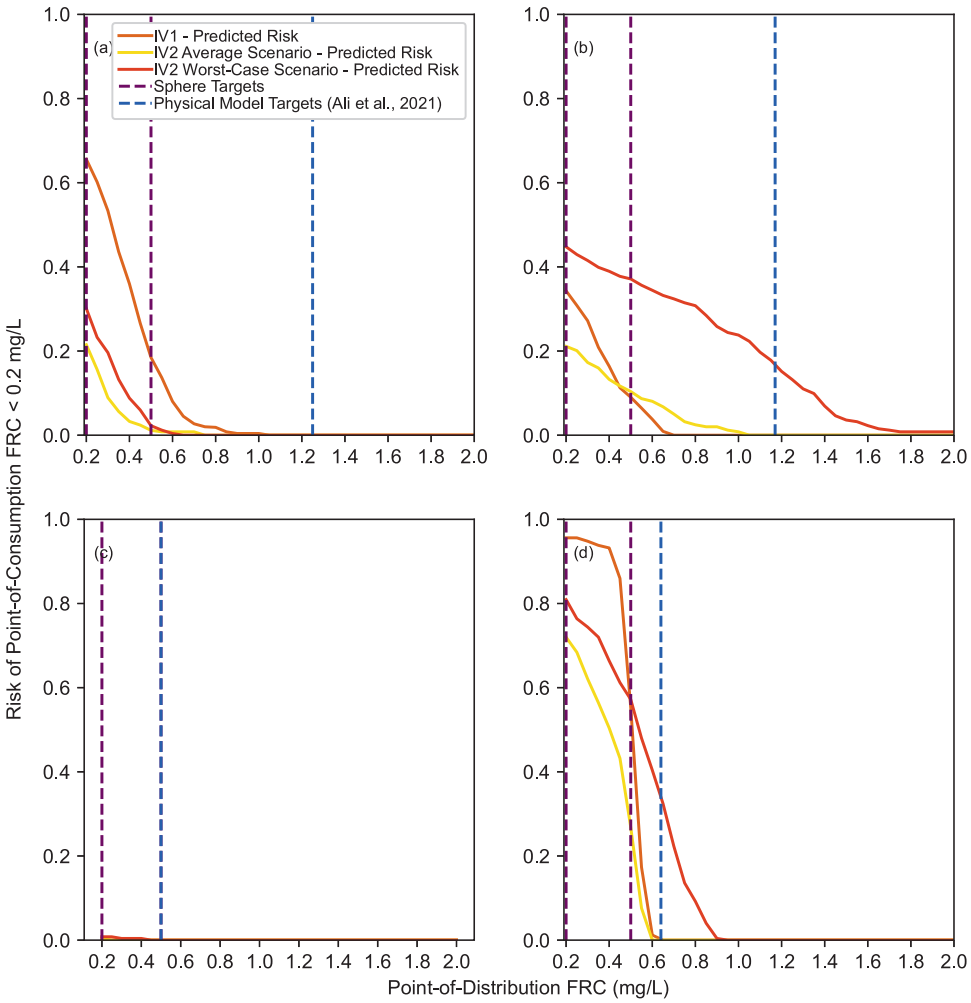

Fig. 7 Predicted risk of insufficient point-of-consumption FRC (below 0.2 mg/L). The predicted risk is shown for **a** South Sudan, **b** Jordan (2014), **c** Jordan (2015), and **d** Rwanda. To achieve negligible risk, the ANN ensemble models recommend point-of distribution FRC between 0.65 and 0.90 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.

dependence on post-processing. Future study should also investigate more sophisticated post-processing methods, which have been proposed and validated in the literature[28,38–40]. 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 predicting 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 Figs 3–6, and in the partial correlation results (Table 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[35–37,41]. 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[18]. However, past studies have shown that EC had a significant effect on post-distribution FRC decay in South Sudan[9]. 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[25,33,42], but other turbidity-causing compounds, such as clays which are a common source of turbidity in groundwater, do not exert strong chlorine demand[42,43]. 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[35]. In this study, the pH range was 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[44]. 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

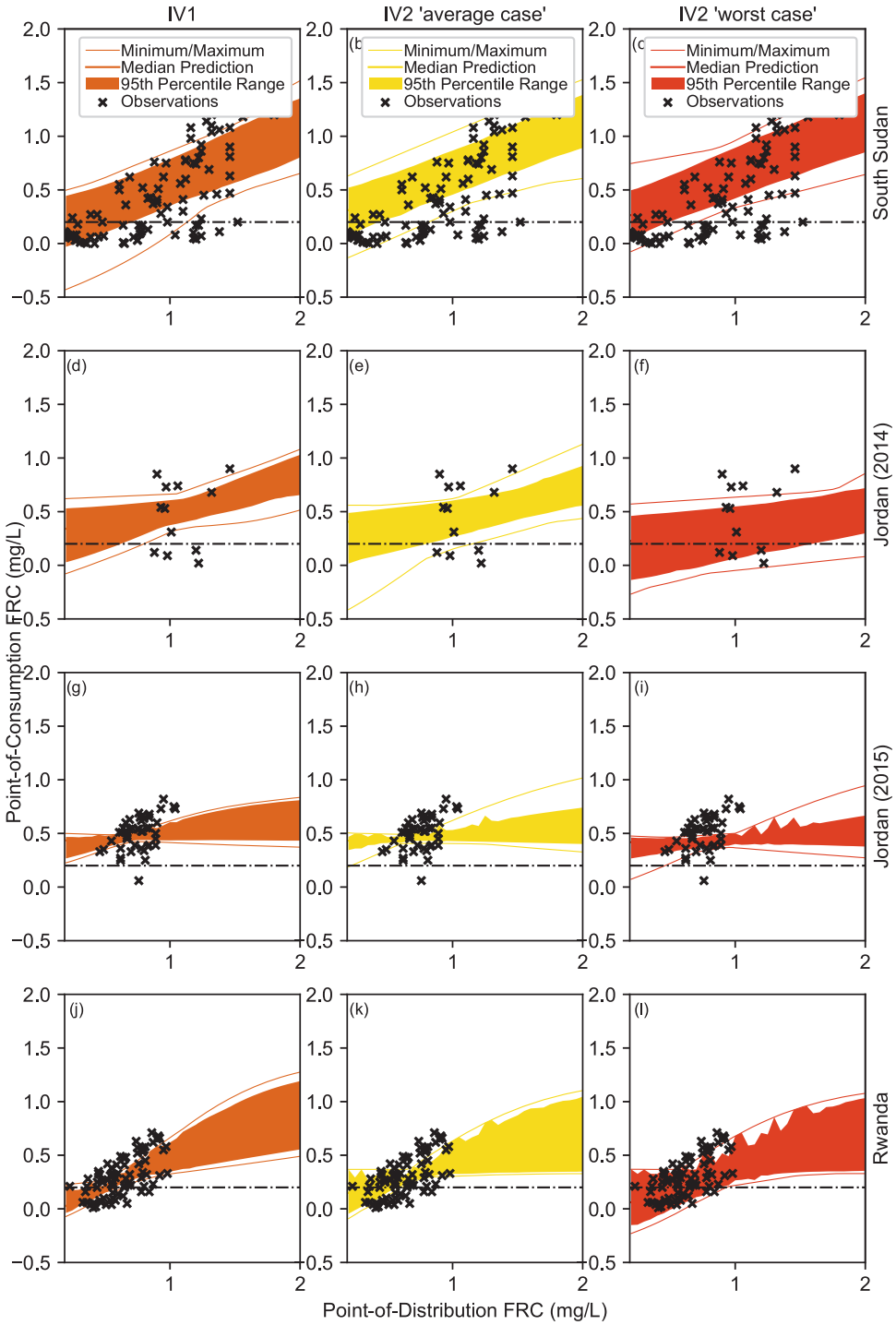

Fig. 8 Forecasts used to generate risk-based FRC targets. Forecasts shown by site and scenario, South Sudan (**a** IV1, **b** IV2 "average case", **c** IV2 "worst case"), Jordan (2014) (**d** IV1, **e** IV2 "average case", **f** IV2 "worst case"), Jordan (2015) (**g** IV1, **h** IV2 "average case", **i** IV2 "worst case"), and Rwanda (**j** IV1, **k** IV2 "average case", **l** IV2 "worst case").

confounded with other factors, especially considering that elapsed time tended to cluster around a few different values at each site, (c.f., Supplementary Figs 5–12 in the Supplementary Information). 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 neither 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 combination. The poorer ensemble reliability exhibited for South Sudan using the IV2 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[45], 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)[19]. 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[19]. 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[19]. 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[18]. 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 to 2.0 mg/L[15,46], 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[30–32] 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[13,15,16,47]. 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[18,47].

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).

METHODS

Study sites and data collection

The data used for this study were obtained from a previous multi-site study on post-distribution FRC decay collected from refugee settlements in South Sudan, Jordan, and Rwanda[19]. 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. Two datasets were collected from Jordan: one from the summer of 2014 and one 9 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[19]. 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 were 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

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. Supplementary Table 2 in the Supplementary Information includes the full list of data cleaning steps that were used to prepare the data for use in the ANN models.

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.

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[31]. Additionally, these are the only two variables included in the process-based model developed in a past study for these sites[19], 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[29–31], and as such, may also be available in

Table 3. Summary of Key Site Characteristics[19,59–61].

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	Average: 35.3 (Min: 28.3; Max: 45.7)	37,199	Groundwater (boreholes)	In-line chlorination with calcium hypochlorite	March–April, 2013	69
	Gendrassa		15,810				76
	Jamam		15,670				75
Jordan (2015)	Azraq	Average: 32.7 (Min: 27.1; Max: 43.3)	7470	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

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.

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[48] in Python v3.7[49]. 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[20,21]. 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 Supplementary Figs 13–20 in the Supplementary Information.

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 were obtained by randomly resampling from the calibration subset, with a different combination of training and validation data for each base learner to promote ensemble diversity. 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 Supplementary Figs 21–28 in the Supplementary Information. Descriptive statistics for the calibration and testing datasets are included in Supplementary Tables 3 and 4 of the Supplementary Information, and histograms of the input and output variables are provided in Supplementary Figs 5–12 in the Supplementary Information to provide context of the range and patterns in the data used to train the ANN base learners.

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[50]. The benefit of implicit approaches is that the differences between the base learners are derived from randomness in the data[50].

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 Supplementary Figs 29–36 in the Supplementary Information.

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[51]. 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[27,28,38,51], though the selection of the specific kernel function is not critical[28]. 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[51].

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[52]. 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 $m = M$.

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 $f_i^0 \leq o_i \leq_i^M$. 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[53–56]. 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.

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 10–100% 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 Eq. 1.

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

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.

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[57,58]. 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[58]. 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 Eq. 2. Note that Eq. 2 shows the calculation of CRPS for a single forecast-observation pair. To evaluate the ensemble models, the average CRPS, $\overline{\text{CRPS}}$, is calculated by taking the mean CRPS overall forecast-observation pairs.

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

For the post-processed probability distributions, we calculated CRPS directly from Eq. 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[58]. $\overline{\text{CRPS}}$ 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[58]. Using these values, the $\overline{\text{CRPS}}$ for an ensemble can be calculated as[58]:

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

Where p_n is the probability associated with each bin, $p_n = \frac{n}{N}$[58].

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 h when generating the FRC targets for South Sudan, and 24 h for all other sites[19]. 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 95th or 5th percentile observed value of that variable was used at each site to simulate the worst-case scenario.

DATA AVAILABILITY

The raw data used in this study were obtained from a previous study available at: https://doi.org/10.17632/twdv4bcwst.1. The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

CODE AVAILABILITY

The analytical code used in this study is available from the authors upon reasonable request.

Received: 25 February 2021; Accepted: 3 June 2021;
Published online: 25 June 2021

REFERENCES

1. Cronin, A. A. et al. A review of water and sanitation provision in refugee camps in association with selected health and nutrition indicators - the need for integrated service provision. *J. Water Health* **6**, 1–13 (2008).
2. Salama, P., Spiegel, P., Talley, L., Waldman, R. & Street, G. Lessons learned from complex emergencies over past decade. *Lancet* **364**, 1801–1813 (2004).
3. Toole, M. J. & Waldman, R. J. The public health aaspects of complex emergencies and refugee situations. *Annu. Rev. Public Health* **18**, 283–312 (1997).
4. Connolly, M. A. et al. Communicable diseases in complex emergencies: impact and challenges. *Lancet* **364**, 1974–1983 (2004).
5. Golicha, Q. et al. Cholera outbreak in Dadaab Refugee camp, Kenya — November 2015–June 2016. *Morb. Mortal. Wkly. Rep.* **67**, 958–961 (2018).
6. Shultz, A. et al. Cholera outbreak in Kenyan Refugee Camp: risk factors for illness and importance of sanitation. *Am. J. Trop. Med. Hyg.* **80**, 640–645 (2009).
7. Swerdlow, D. L. et al. Epidemic cholera among refugees in Malawi, Africa: treatment and transmission. *Epidemiol. Infect.* **118**, 207–214 (1997).
8. Walden, V. M., Lamond, E. A. & Field, S. A. Container contamination as a possible source of a diarrhoea outbreak in Abou Shouk camp, Darfur province, Sudan. *Disasters* **29**, 213–221 (2005).
9. Ali, S. I., Ali, S. S. & Fesselet, J.-F. Effectiveness of emergency water treatment practices in refugee camps in South Sudan. *Bull. World Health Organ.* **93**, 550–558 (2015).
10. Guerrero-Latorre, L., Hundesa, A. & Girones, R. Transmission sources of waterborne viruses in South Sudan Refugee Camps. *Clean. Soil Air Water* **44**, 775–780 (2016).
11. Howard, C. M. et al. Novel risk factors associated with hepatitis E virus infection in a large outbreak in Northern Uganda: results from a case-control study and environmental analysis. *Am. J. Trop. Med. Hyg.* **83**, 1170–1173 (2010).

12. Steele, A., Clarke, B. & Watkins, O. Impact of jerry can disinfection in a camp environment—experiences in an IDP camp in Northern Uganda. *J. Water Health* **6**, 559–564 (2008).
13. Rashid, M.-U. et al. Chlorination of household drinking water among cholera patients' households to prevent transmission of toxigenic Vibrio cholerae in Dhaka, Bangladesh: CHoBI7 Trial. *Am. J. Trop. Med. Hyg.* **95**, 1299–1304 (2016).
14. Girones, R. et al. Chlorine inactivation of hepatitis e virus and human adenovirus 2 in water. *J. Water Health* **12**, 436–442 (2014).
15. Lantagne, D. S. Sodium hypochlorite dosage for household and emergency water treatment. *J. Am. Water Work. Assoc.* **100**, 106–114 (2008).
16. Sikder, M. et al. Effectiveness of water chlorination programs along the emergency-transition-post-emergency continuum: evaluations of bucket, in-line, and piped water chlorination programs in Cox's Bazar. *Water Res*. https://doi.org/10.1016/j.watres.2020.115854 (2020).
17. CDC. *Chlorine Residual Testing*. http://www.cdc.gov/safewater/chlorine-residual-testing.html. (2012).
18. World Health Organization. *WHO Guidelines for Drinking-water quality* (World Health Organization, 2011).
19. Ali, S. I., Ali, S. S. & Fesselet, J. Evidence-based chlorination targets for household water safety in humanitarian settings: recommendations from a multi-site study in refugee camps in South Sudan, Jordan, and Rwanda. *Water Res.* **189**, 1–17 (2021).
20. Rodriguez, M. J. & Sérodes, J. B. Assessing empirical linear and non-linear modelling of residual chlorine in urban drinking water systems. *Environ. Model. Softw.* **14**, 93–102 (1998).
21. Gibbs, M. S. et al. Investigation into the relationship between chlorine decay and water distribution parameters using data driven methods. *Math. Comput. Model.* **44**, 485–498 (2006).
22. Soyupak, S., Kilic, H., Karadirek, I. E. & Muhammetoglu, H. On the usage of artificial neural networks in chlorine control applications for water distribution networks with high quality water. *J. Water Supply Res. Technol. AQUA* **60**, 51–60 (2011).
23. Bowden, G. J., Nixon, J. B., Dandy, G. C., Maier, H. R. & Holmes, M. Forecasting chlorine residuals in a water distribution system using a general regression neural network. *Math. Comput. Model.* **44**, 469–484 (2006).
24. Gibbs, M. S. et al. 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 (2003).
25. Kotlarz, N., Lantagne, D., Preston, K. & Jellison, K. Turbidity and chlorine demand reduction using locally available physical water clarification mechanisms before household chlorination in developing countries. *J. Water Health* **7**, 497–506 (2009).
26. Boucher, M.-A., Perreault, L. & Anctil, F. Tools for the assessment of hydrological ensemble forecasts obtained by neural networks. *J. Hydroinformatics* **11**, 297–307 (2009).
27. Boucher, M. A., Anctil, F., Perreault, L. & Tremblay, D. A comparison between ensemble and deterministic hydrological forecasts in an operational context. *Adv. Geosci.* **29**, 85–94 (2011).
28. Boucher, M. A., Perreault, L., Anctil, F. & Favre, A. C. Exploratory analysis of statistical post-processing methods for hydrological ensemble forecasts. *Hydrol. Process.* **29**, 1141–1155 (2015).
29. Frazier, C. In *The Johns Hopkins and Red Cross Red Crescent health guide Public in emergencies* (ed. Rand, E. C.) 372–441 (International Federation of Red Cross and Red Crescent Societies, 2008).
30. Médecins Sans Frontières. *Public Health Engineering In Precarious Situations* (Médecins Sans Frontières, 2010).
31. Sphere Association. *The Sphere Handbook: Humanitarian Charter and Minimum Standards in Humanitarian Response* (Practical Action Publishing, 2018).
32. UNHCR. *WASH Manual—Pratical Guidance for Refugee Settings* (UNHCR, 2020).
33. LeChevallier, M. W., Evans, T. M. & Seidler, R. J. Effect of turbidity on chlorination efficiency and bacterial persistence in drinking water. *Appl. Environ. Microbiol.* **42**, 159–167 (1981).
34. Powell, J. C., West, J. R., Hallam, N. B., Forster, C. F. & Simms, J. Performance of various kinetic models for chlorine decay. *J. Water Resour. Plan. Manag.* **126**, 13–20 (2000).
35. Clark, R. M. & Sivaganesan, M. Predicting chlorine residuals in drinking water: second order model. *J. Water Resour. Plan. Manag.* **128**, 152–161 (2002).
36. Warton, B., Heitz, A., Joll, C. & Kagi, R. A new method for calculation of the chlorine demand of natural and treated waters. *Water Res.* **40**, 2877–2884 (2006).
37. Fisher, I., Kastl, G. & Sathasivan, A. A comprehensive bulk chlorine decay model for simulating residuals in water distribution systems. *Urban Water J.* **14**, 361–368 (2017).
38. Bröcker, J. & Smith, L. A. From ensemble forecasts to predictive distribution functions. *Tellus, Ser. A Dyn. Meteorol. Oceanogr.* **60**, 663–678 (2008).
39. Wang, X. & Bishop, C. H. Improvement of ensemble reliability with a new dressing kernel. *Q. J. R. Meteorol. Soc.* **131**, 965–986 (2005).
40. Fortin, V., Favre, A. C. & Saïd, M. Probabilistic forecasting from ensemble prediction systems: improving upon the best-member method by using a different weight and dressing kernel for each member. *Q. J. R. Meteorol. Soc.* **132**, 1349–1369 (2006).
41. Powell, J. C., Hallam, N. B., West, J. R., Forster, C. F. & Simms, J. Factors which control bulk chlorine decay rates. *Water Res.* **34**, 117–126 (2000).
42. Gallandat, K., Stack, D., String, G. & Lantagne, D. Residual maintenance using sodium hypochlorite, sodium dichloroisocyanurate, and chlorine dioxide in laboratory waters of varying turbidity. *Water (Switzerland)* **11**, 1309 (2019).
43. Wu, H. & Dorea, C. C. Towards a predictive model for initial chlorine dose in humanitarian emergencies. *Water (Switzerland)* **12**, 1506 (2020).
44. Adam, L. C. & Gordon, G. Hypochlorite ion decomposition: effects of temperature, ionic strength, and chloride ion. *Inorg. Chem.* **38**, 1299–1304 (1999).
45. Vasconcelos, J. J., Rossman, L. A., Grayman, W. M., Boulos, P. F. & Clark, R. M. Kinetics of chlorine decay. *J. Am. Water Work. Assoc.* **89**, 54–65 (1997).
46. Crider, Y. et al. Can you taste it? Taste detection and acceptability thresholds for chlorine residual in drinking water in Dhaka, Bangladesh. *Sci. Total Environ.* **613–614**, 840–846 (2018).
47. Lechevallier, M. W., Welch, N. J. & Smith, D. B. Full-scale studies of factors related to coliform regrowth in drinking water. *Appl. Environ. Microbiol.* **62**, 2201–2211 (1996).
48. Cholette, F. Keras. (2015). https://keras.io. Accessed on 14 June 2021.
49. Python Software Foundation. *Python v3.7.4*. (2019). https://www.python.org/downloads/release/python-374/.
50. Brown, G., Wyatt, J., Harris, R. & Yao, X. Diversity creation methods: a survey and categorisation. *Inf. Fusion* **6**, 5–20 (2005).
51. Roulston, M. S. & Smith, L. A. Combining dynamical and statistical ensembles. *Tellus Ser. A Dyn. Meteorol. Oceanogr.* **55**, 16–30 (2003).
52. Hamill, T. M. Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Weather Rev.* **129**, 550–560 (2001).
53. Khan, U. T. & Valeo, C. Dissolved oxygen prediction using a possibility theory based fuzzy neural network. *Hydrol. Earth Syst. Sci.* **20**, 2267–2293 (2016).
54. Khan, U. T. & Valeo, C. Comparing a Bayesian and fuzzy number approach to uncertainty quantification in short-term dissolved oxygen prediction. *J. Environ. Inform.* **30**, 1–16 (2017).
55. Alvisi, S. & Franchini, M. Fuzzy neural networks for water level and discharge forecasting with uncertainty. *Environ. Model. Softw.* **26**, 523–537 (2011).
56. Alvisi, S. & Franchini, M. Grey neural networks for river stage forecasting with uncertainty. *Phys. Chem. Earth* **42–44**, 108–118 (2012).
57. Ferro, C. A. T. Fair scores for ensemble forecasts. *Q. J. R. Meteorol. Soc.* **140**, 1917–1923 (2014).
58. Hersbach, H. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecast.* **15**, 559–570 (2000).
59. Médecins Sans Frontières. *Maban County, South Sudan WASH Coordination Report (Week 11 and 12)*. (Médecins Sans Frontières, Amsterdam, Netherlands, 2013).
60. United Nations International Children's Emergency Fund. *Azraq, Jordan WASH Monitoring Reports 2014 & 2015*. (United Nations International Children's Emergency Fund, Amman, Jordan, 2015).
61. Parlement des Jeunes Rwandais. *Kigeme, Rwanda WASH Monthy Updates (June–July)*. (Parlement des Jeunes Rwandais, Kigali, Rwanda, 2015).

ACKNOWLEDGEMENTS

We would like to extend our gratitude for the support of our colleagues from the local refugee population, MSF, and UNHCR in South Sudan, Jordan, and Rwanda. We would also like to gratefully acknowledge James Orbinski of DIGHR for his advisory support on the SWOT project. We would also like to express our gratitude to Rahma Shakir for her preliminary work on IVS development for the SWOT ANN models, and Apostolos Vasileiou for his work developing the initial version of the open-source ANN analytics for the SWOT. We would also like to thank Dr. Stephanie Gora and Everett Snieder for their input on this manuscript. We would also like to gratefully acknowledge Ngqabutho Zondo for his aid preparing Fig. 1. Field data collection work was supported by MSF, Amsterdam, Netherlands; UNHCR, Division of Programme Support and Management, Geneva, Switzerland; ELRHA/Humanitarian Innovation Fund, London, UK; and the Development Impact Lab, USAID Higher Education Solutions Network (USAID Cooperative Agreement AID-OAA-A-13-00002). We also acknowledge the Achmea Foundation, Zeist, The Netherlands; the Natural Science and Engineering Research Council, Ottawa, Canada; and York University, Toronto, Canada for additional funding support for this research.

AUTHOR CONTRIBUTIONS

M.D.S.: ANN modelling, data analysis, manuscript preparation. U.T.K.: manuscript review, modelling supervision. S.I.A.: data collection (South Sudan, Jordan 2014/2015, Rwanda), coordination of partners, securing funding, manuscript review. J.-F.F.: coordination of partners, securing funding, manuscript review. M.A.: coordination of partners, manuscript review.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41545-021-00125-2.

Correspondence and requests for materials should be addressed to S.I.A.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.