npj | Clean Water

www.nature.com/npjcleanwater

ARTICLE OPEN

Predicting quality and quantity of water used by urban households based on tap water service

Aurelie Jeandron[1]*, Oliver Cumming[1], Lumami Kapepula[2] and Simon Cousens[3]

Despite significant progress in improving access to safe water globally, inadequate access remains a major public health concern in low- and middle-income countries. We collected data on the bacterial quality of stored drinking water and the quantity of water used domestically from 416 households in Uvira, Democratic Republic of the Congo. An indicator of tap water availability was constructed using invoices from 3685 georeferenced piped water connections. We examined how well this indicator predicts the probability that a household's stored drinking water is contaminated with *Escherichia coli*, and the total amount of water used at home daily, accounting for distance from alternative surface water sources. Probability of drinking water contamination is predicted with good discrimination overall, and very good discrimination for poorer households. More than 80% of the households are predicted to store contaminated drinking water in areas closest to the rivers and with the worst tap water service, where river water is also the most likely reported source of drinking water. A model including household composition predicts nearly two-thirds of the variability in the reported quantity of water used daily at home. Households located near surface water and with a poor tap water service indicator are more likely to use water directly at the source. Our results provide valuable information that supports an ongoing large-scale investment in water supply infrastructure in Uvira designed to reduce the high burden of cholera and other diarrhoeal diseases. This approach may be useful in other urban settings with limited water supply access.

npj Clean Water (2019)2:23 ; https://doi.org/10.1038/s41545-019-0047-9

INTRODUCTION

The public health importance of both drinking water quality and the quantity of water available for domestic consumption has been long recognised.[1] The quality of drinking water influences "waterborne" disease transmission, that is the ingestion of infectious agents via contaminated drinking water, while the quantity of water available domestically influences "water-washed" disease transmission, that is where insufficient water is available to allow adequate hygiene practices.[2,3] Many efforts have been made to quantify the health risks associated with both, with an emphasis on diarrhoeal diseases.[4–13] In turn, the estimated levels of health risk have been used to classify households' water sources, combining water quality characteristics, such as protection from microbiological contamination, and accessibility and availability criteria, such as distance to the source, time needed to fetch water and continuity of water availability.

In 2015, the international community set the ambitious target of "safely managed water for all" by 2030 (Sustainable Development Goals (SDGs)—target 6.1) in an ongoing effort to address the detrimental health and social impacts of poor access to safe water for those without these services.[14] "Safely managed water" is defined by three criteria: (1) water free of faecal and priority contamination; (2) water accessible directly on premises; (3) water available when needed. This definition replaces the dichotomy between unimproved/improved drinking water source previously used for the Millennium Development Goals (MDGs) between 1990 and 2015 that did not capture adequately the levels of accessibility and availability offered by improved sources.[15]

Estimates of the health benefits of improved accessibility and continuously available water sources are rather heterogeneous across studies, but a recent review estimates that moving from an unimproved drinking water source to a continuous piped water supply on the premises could reduce the risk of diarrhoeal diseases by up to 75%.[12] Using quantitative microbial risk assessment methods (QMRA), Bivins et al. estimated that 13,700 disability-adjusted life years (DALY) are attributable to intermittency of piped water supplies in Sub-Saharan Africa (SSA) with 109,000 annual DALY worldwide.[16] The causal pathway for the above health impacts involves a change in both quality and quantity of water used by households for drinking and domestic purposes, but establishing the respective importance of each is challenging. A continuous piped water supply on the premises suggests that households do not require water handling and storage, which are known risk factors for microbial contamination.[17] Continuous tap water service and pressure also reduce the risk of contaminant ingress into the distribution network and water quality deterioration.[18] Access to water on the premises implies minimal time and effort needed from household members to collect water, and doubles or more water consumption in comparison with households using a source outside their premises.[1,8] Water quantity consumed by households was indeed shown to be stable when using a source outside the compound located as far as ~30 min of return collection journey, to decrease as the journey time became longer than 30 min, but to increase sharply when the source was located on the premises. An uninterrupted supply that users trust to perform consistently well also reduces the probability of a household occasionally reverting to other water sources of lesser quality or more demanding to fetch, and the probability of some hygiene practices being temporarily abandoned.[19,20]

[1]Environmental Health Group, Department of Disease Control, Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, WC1E 7HT London, UK. [2]Centre de Recherche en Hydrobiologie (CRH), Département d'Hydrologie, Section Hydrochimie, Uvira, République Démocratique du Congo. [3]Department of Infectious Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, WC1E 7HT London, UK. *email: aurelie.jeandron@lshtm.ac.uk

Published in partnership with King Fahd University of Petroleum & Minerals

npj nature partner journals

To enhance consistency and comparability of SDG estimates globally, emphasis is largely placed on households' main source of water, which overlooks the widespread use of multiple sources of water by households.[21] In addition, SDG estimates are mostly based on standardised national cross-sectional household surveys and do not capture geographical inequalities at a smaller scale, although methodological innovations, such as using spatial and remote-sensing tools and data, have been proposed to provide better coverage estimates at the district level.[22]

The Democratic Republic of the Congo (DRC) is one of a few countries in which access to piped water on premises has markedly declined over the MDG period, by nearly two-thirds for the urban population, from 48 to 17% coverage.[23] In addition to increasing urbanisation, this is most likely a consequence of the deterioration in the operational capacity of the National Water Agency, Regideso, which operates piped water systems in 94 cities and secondary urban centres across the country.[24] None of the piped water supply systems in DRC are considered uninterrupted, and none of the country's population has access to safely managed water.[23] The piped water infrastructure in Uvira, the second largest town of South-Kivu province, is no exception, and is currently being rehabilitated and extended to improve access to piped water services for its ~250,000 inhabitants, in an attempt to better control cholera and other diarrhoeal diseases. Uvira is an identified hotspot for endemic cholera in Eastern DRC, from which the disease regularly spreads to less frequently affected areas of the country.[25–27]

The present analysis draws on data collected as part of an evaluation of the impact of these improvements on households' water-related practices, and the incidence of suspected cholera and other diarrhoeal diseases. Its aim is to develop and assess spatially explicit predictive models for estimating the probability of households storing contaminated drinking water and for estimating the quantity of water used at home for domestic activities, based on piped water access and distance from surface water sources. This research has the potential to provide estimates of water service level coverage at a high spatial resolution, to enable better targeted water infrastructure investments and detection of geographical inequities.

RESULTS

Built-up areas in Uvira cover ~12 km^2, with an estimated population of 254,438 inhabitants in November 2017. Population density estimated at the street level ($n = 205$) ranged from 5248 to 261,342 inhabitants per km^2 (median: 23,835; interquartile range —IQR: 17,565–33,659). In March 2018, 3685 taps were invoiced for a total of 60,925 m^3, with invoices ranging from 1 to 1800 m^3 (Supplementary Fig. S3). This represented 80.6% of the water treated by the Regideso over the same period. The total volume of water treated at the Regideso plant during the 31 days preceding the last survey day (October 12th–November 11th, 2017) amounted to 76,163 m^3, leading to an estimated volume distributed of 61,387 m^3 and a daily average of 1980 m^3. Over all built-up areas of Uvira, the 250-m tap water service indicator ranged from 5×10^{-4} to 99.3 LCPD (Supplementary Fig. S4). The maximum distances from the closest river and from the lake shore, adjusted for slope, were 5429 meters and 1812 meters, respectively, for all built-up areas.

Data from 416 households were included in the data analyses, of which 371 (89.2%) were recruited during the 2016 survey. The tap water service indicator ind250 for interviewed households ranged from 0.03 to 85 LCPD. Maximum distance from the closest river and from the lake shore were 1880 meters and 1800 meters, respectively, for interviewed households (Supplementary Fig. S4). Drinking water contamination with *E. coli* was detected in 273 samples of the 411 analysed (66.4%). In total, 301 (72.4%) households reported having collected their stored drinking water from a tap. In total, 174 (58.6%) drinking water samples collected at a tap and analysed were contaminated with *E. coli*, in comparison with the 99 (87.6%) collected from surface sources, mostly from rivers. The reported total amount of water used by households at home on the day preceding interview ranged from 10 to 460 l (median 145 l, IQR 100–205), and the number of household members present the day preceding the interview ranged from 1 to 19 (median 7, IQR 5–10). In total, 115 (27.6%) households reported having performed water-consuming activities directly at the source.

The training data set contained information on 235 households, and the test data set 181 households (Table 1).

Household drinking water contamination

Drinking water contamination is strongly associated with the reported source of collection, with drinking water collected at a surface water source having five times the odds of contamination (95% CI 2.7–9.1) of drinking water collected at a tap. The odds of a household having collected their drinking water at a tap rather than a surface source are also strongly associated with distance from the nearest river and the tap water service indicator (Supplementary Table S2).

The best-fitting model to predict drinking water contamination included tap water service indicator, distance from the nearest river and a linear interaction between distance from the nearest river and tap water service indicator. There was only weak evidence of miscalibration of the model fit to the training and testing data, and no evidence suggesting that assumption of linear relationships on the logit scale was inappropriate. Figure 1a

Table 1. Characteristics of households included in training and testing data sets.

	Training ($N = 235$) *n* (%) or median (IQR/range)	Testing ($N = 181$) *n* (%) or median (IQR/range)
Drinking water contaminated with *E. coli*	151 (64.3%)	122 (67.4%)
Missing	2 (0.9%)	3 (1.7%)
Total amount of water used in the household the previous day (in l)	145 (100–210/ 10–460)	144 (102–200/ 20–385)
Missing	1 (0.4%)	1 (0.6%)
Number of household members present the previous day		
Total	7 (5–9/1–18)	7 (5–10/1–19)
Adults and children aged 15 and older	3 (2–5/1–12)	3 (2–5/1–12)
Children under 15	4 (2–5/0–11)	4 (2–5/0–10)
Tap water service indicator (ind250) in LCPD[a]	4.5 (1.9–12.8/ 0–85.1)	5.1 (2.2–12.9/ 0–62.9)
Distance to the nearest river in m[a]	672 (355–1091/ 0–1727)	630 (293–1148/ 30–1880)
Distance to the lake in m[a]	539 (272–879/ 0–1680)	607 (345–896/ 30–1799)
Wealth quintile (index range)		
#0 (−2.23 to −1.29)	43 (18.3%)	38 (21%)
−1.26 to −0.7	49 (20.9%)	34 (18.8%)
−0.7 to −0.16	41 (17.4%)	42 (23.2%)
−0.15 to −0.79	51 (21.7%)	33 (18.2%)
#4 (0.8–5.38)	51 (21.7%)	32 (17.7%)
Missing	0 (0%)	2 (1.1%)

[a]At a spatial resolution of 34 m × 42 m

shows the observed versus predicted probability for drinking water contamination by deciles of predicted probability, for model fits to training and testing data.

Discrimination performance of the selected model on the training and test data was fair with AUCs of 0.73 and 0.75, respectively (95% CI 0.66–0.8 and 0.67–0.83, respectively) (Fig. 1b). In the selected model fitted to the training data set, the predicted probability of drinking water contamination decreases as tap water service improves, although this relationship is modified by distance from the nearest river. The largest marginal effect of increased tap water access—i.e. the predicted change in contamination probability for each LCPD unit increment in tap water indicator—is predicted to occur at the shortest distance from the nearest river (−1.6%, 95% CI −0.8% to −2.4%). Model coefficients are reported in Table 2.

When comparing AUC estimates for the entire data set based on training model fit and stratified by household wealth quintile, the discrimination ability of the model decreases noticeably between the two lowest (AUC 0.80; 95% CI 0.71–0.89) and the three highest wealth quintiles (AUC 0.66; 95% CI 0.59–0.74). However, wealth quintile inclusion in the model as an independent variable only improves very slightly its discrimination performance on the training data set (AUC 0.76; 95% CI 0.7–0.82), while reducing it noticeably on the testing data set (AUC 0.68; 95% CI 0.6–0.77).

Unsurprisingly considering the difference in odds of contamination in samples coming from tap and surface sources, the predicted probability of drinking water contamination also discriminates well between the use of a surface water or a tap as drinking water source (AUC 0.86; 95% CI 0.82–0.90).

Figure 2 shows the geographic distribution of predicted probabilities of households storing contaminated drinking water across Uvira. The map highlights the areas surrounding rivers as those at the highest risk of contamination, and the central areas of town, south of the water reservoir, as at the lowest risk.

Quantity of water used at home for domestic activities

The best-fitting model on the training data set includes household composition, tap water service indicator multiplied by the number of household members, distances from the lake and nearest river and linear interaction terms between tap water service and distance from the nearest river, and between distance to the lake and distance from the nearest river. The selected model explains >60% of the variability in reported household water consumption in the training data (adjusted R^2 0.61). The mean difference between reported and predicted values (RMSE) is nonetheless high, and is nearly equivalent to the cumulated daily consumption of two adults and one child (50 litres). There was no evidence that using linear functions for the predictors was inappropriate. Reported and fitted values for both training and testing data sets are shown in Fig. 3.

The model predicts additional reported consumption of 21.8 and 10.6 litres for each extra adult and child present in the household, respectively. Water consumption at home is predicted to increase with improved tap water access, up to a distance of ~1300 m from the nearest river, with the highest marginal effect of 0.32 litres (95% CI 0.16–0.47 litres) increased consumption per capita for households closest to the rivers. Beyond 1300 m from the nearest river, there is little evidence of an effect of improved tap water access over water consumption at home, with wide confidence intervals, including no effect at all. Household water consumption is predicted to decrease, as distance from the lake increases for households >500 m from the nearest river, at a rate of up to −12.8 litres per 100-m distance increment. The predicted effect of increasing distance from the nearest river varies with both distance to the lake and tap water service. The marginal effect is positive for households close to the lake (<500 m) and up to good access to tap water (<20 LPCD, 85th

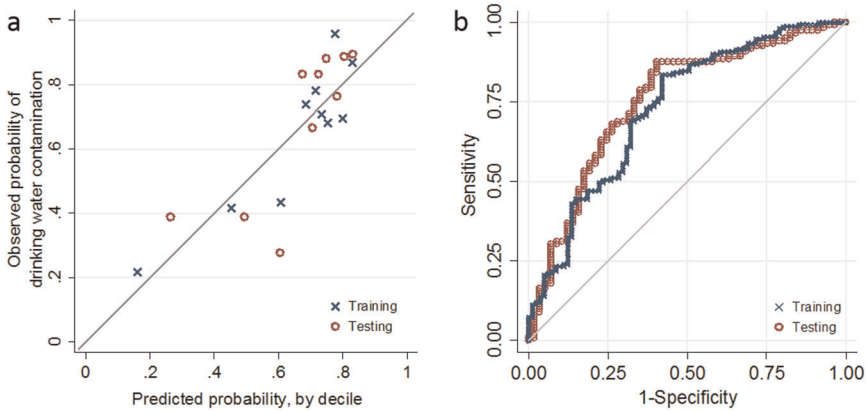

Fig. 1 Calibration and performance of the training and testing models for drinking water contamination with *E. coli*. Observed versus predicted probability for drinking water contamination by deciles of predicted probability, for model fits to training and testing data sets (**a**) and ROC for model fits to training and testing data sets (**b**).

Table 2. Selected logistic regression model for drinking water contamination with *E. coli*.

	Logit coefficient (SE)	OR (95% CI)	*p*-value[a]
Intercept	1.84 (0.41)	–	<0.001
Tap water service indicator (ind250) in LCPD	−0.1 (0.03)	0.9 (0.84–0.96)	0.002
Distance from the nearest river in m	−6.93 (4.83)[b]	0.9993 (0.9984–1.0003)	0.151
Dist. River X ind250	0.48 (0.29)[b]	1.00005 (0.99999–1.00011)	0.097

[a]*p*-value for chi-square of likelihood ratio test
[b]$\times 10^{-4}$
Pregibon's goodness-of-link test: $p = 0.36$; Hosmer–Lemeshow statistic for training and testing data: $p = 0.16$ and $p = 0.09$, respectively

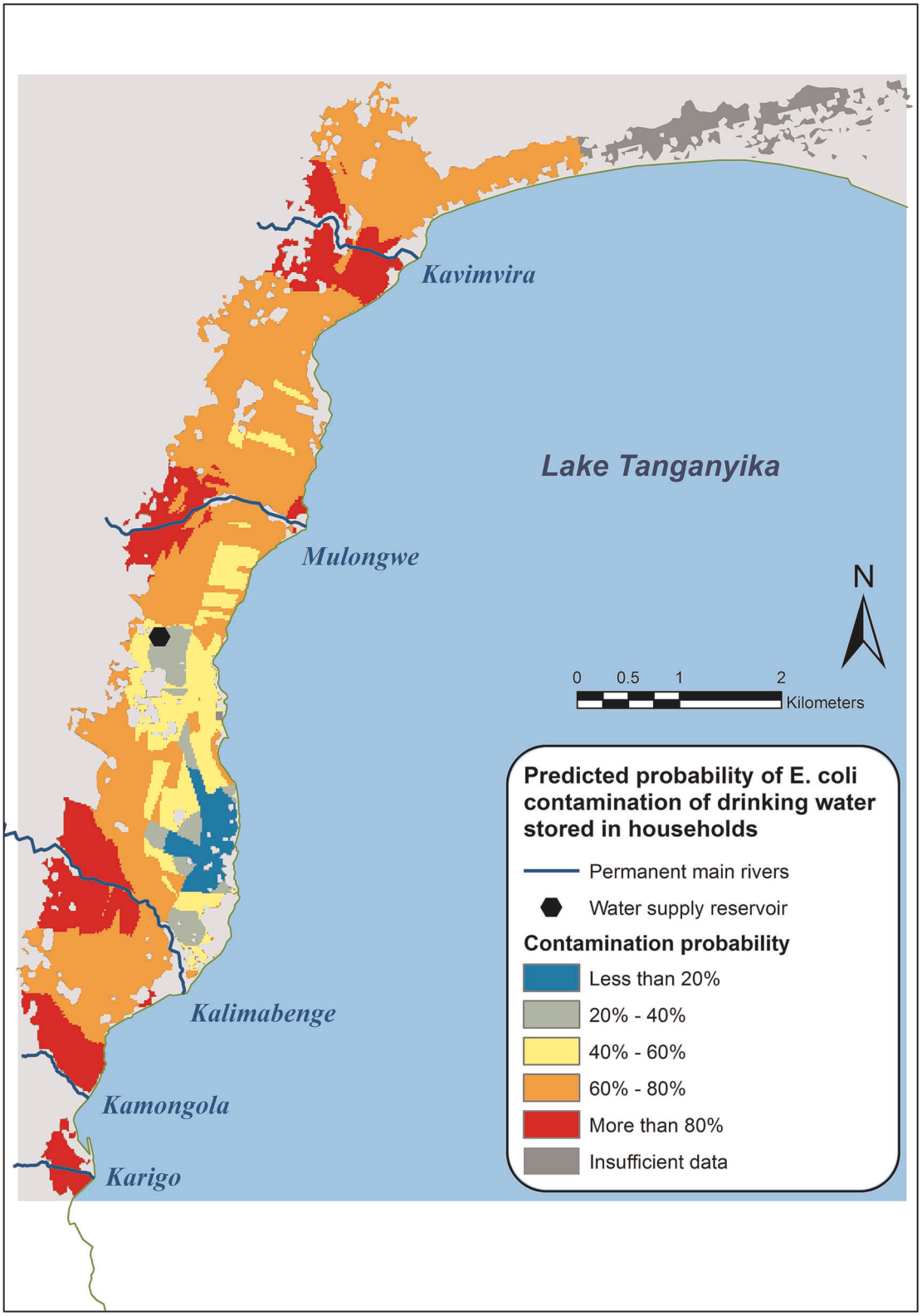

Fig. 2 Geographic distribution of predicted probabilities of households to store contaminated drinking water across Uvira.

percentile over the full data set). In other situations, household water consumption is predicted to decrease as distance from the lake increases, by up to −10.6 litres per 100-m distance increment at the median access to tap water (4.75 LPCD). Regression coefficients are presented in Table 3.

When adding wealth quintile to the selected model, its performance remains unchanged, with no evidence that wealth quintile independently affects the predicted household water consumption. When adding water use at the source, however, the model performance improves slightly (adjusted R^2 0.64; RMSE 48.7)

and the model predicts a reduction in the quantity of water used at home of 34.9 litres (95% CI 19.4–50.3 litres; $p < 0.001$) by households performing water-consuming activities at the source, with little change to other model coefficients. A logistic regression model of household water use at the source suggests that tap water service, distances from the nearest river and from the lake and linear interaction between tap water service and distance from the nearest river are important predictors for such practice, with good discrimination (AUC 0.76). Over all observations in the entire data set, the predicted probability of household water use at source decreases on average by 1%, 2.9% and 2.8% as tap water service and distances from the nearest river and from the lake increase (for 1 LPCD increment and 100-m increments, respectively).

Figure 4a shows the predicted amount of water used for domestic activities performed at home, by a household composed of three adults or children over 15, and four children under 15, across Uvira. Households in areas poorly served by the existing water supply network and distant from both the lake and rivers are predicted to report using <140 litres per day at home. Areas near the rivers and the lake are also highlighted as areas with low predicted water consumption at home. These areas are, however, overlapping with those with a high probability of water use at the source (Fig. 4b).

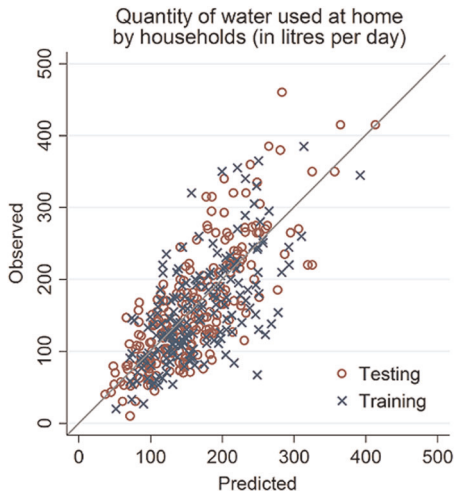

Fig. 3 Reported and predicted daily quantity of water used at home by households in litres.

Sensitivity analysis

When fitting similarly formulated models to the training data with indicators constructed over 500-m or 750-m radii and indicators based on tap location and density without invoicing data at radii 250 m, 500 m and 750 m, predictions of drinking water contamination with *E. coli* and predictions of the quantity of water used by a household show similar performance in comparison with ind250 (Supplementary Table S3).

DISCUSSION

Our study used household location relative to surface water sources and a measure of their access to tap water to predict the probability of microbial contamination of household drinking water and the quantity of water consumed domestically. These predictions were developed at the scale of Uvira, the second largest town of South-Kivu province in DRC and summarised into high-resolution maps of this secondary urban centre. The models were developed using a subset of households' survey data collected in October 2017, and their predictive performance assessed against the remaining portion of survey data.

The drinking water model predicted whether household-stored contaminated drinking water with fair discrimination performed better for the poorest households. Probability of contamination was strongly predicted by tap water service and distance from the nearest river, with the lowest probabilities predicted to occur for households with a better tap water service and further away from the rivers. The same improvement in tap water service was predicted to reduce the probability of contamination in areas <250 m from the nearest river by twice as much as in areas >1250 m from the nearest river. The same model was also shown to predict reasonably well the type of source used by the household for drinking water (surface water or tap), which was associated with very different risks of contamination.

The total quantity of water reported to be used at home by households was predicted by its demographic composition, tap water service and distances from both the lake and the nearest river. An adult member was predicted to report more than double the increase in daily household water consumption than a child, with 21.8 litres versus 10.6 litres, respectively. The marginal effect of tap water service improvement was predicted to be small, and decreased as distance from the nearest river increased. On average over the study households, an increase of one LCPD in tap water service represented only 1.1-litre increase in the reported total household consumption, corresponding to an average of 0.15 litre per household member. The lower amount of water consumed at home by households located close to the rivers and

Table 3. Selected linear regression model for predicting the amount of water used at home by households for domestic activities.

	Coefficient (SE)	95% CI	*p*-value[a]
Intercept	−12.68 (15.26)	−42.74–17.38	0.407
The number of household members aged 15 or more	21.79 (1.92)	18.01–25.58	<0.001
The number of household members under 15	10.62 (1.35)	7.96–13.28	<0.001
Tap water service indicator (ind250) × number of household members (in litres per day)	0.32 (0.08)	0.16–0.47	<0.001
Distance from the nearest river (in m)	0.08 (0.02)	0.04–0.11	<0.001
Distance from the lake (in m)	0.04 (0.02)	0.01–0.08	0.006
Dist. River × Dist. Lake	−0.87 (0.21)[b]	−1.28 to −0.45[b]	<0.001
Dist. River × (ind250 × number of household members)	−2.37 (0.66)[b]	−3.68 to −1.07[b]	<0.001

[a]*p*-value for χ^2 of likelihood ratio test
[b]$\times 10^{-4}$
Pregibon's goodness-of-link test for training and testing data: $p = 0.58$ and $p = 0.74$, respectively. RMSE for training and testing data: 57.8 and 50.6, respectively

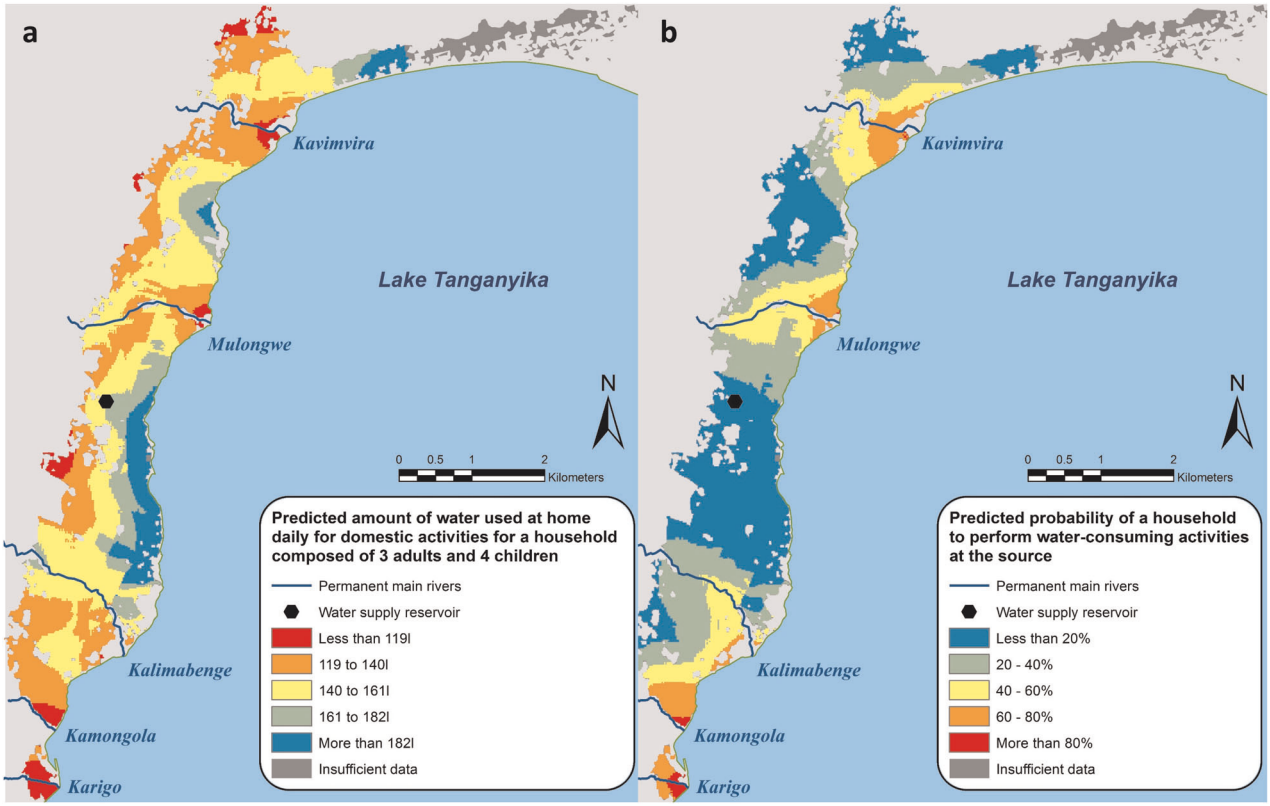

Fig. 4 Geographic distribution of predicted quantity of water used at home for domestic activities and of the predicted probability of performing water-consumingactivities at the water source. Geographic distribution of predicted quantity of water used at home for domestic activities for a standardised household composed of three adults and four children (**a**) and geographic distribution of the predicted probability of performing water-consuming activities at the water source (**b**).

lake appears associated with a higher probability of performing activities directly at the source.

The tap water service indicator used in the present analysis was built upon the assumption that a higher geographical density of water taps in the vicinity of a household represents a higher probability for that household of accessing and using it. By weighing the spatial density with the water volume invoiced and population density, the proxy indicator was then adjusted for a potential limited supply of water at the taps, because of unreliability, intermittency or because of "competition" with other households for access. Within these assumptions, the sensitivity analysis results of the indicator estimated over 500-m and 750-m radii first suggest that household behaviours related to water sources are influenced by spatial density of taps at least up to 750 m. The results obtained with the indicators ignoring the actual water volume invoiced for each tap then suggest that either reliability or intermittency of water supply at the tap has little bearing over these behaviours, or that the information on reliability or intermittency of a tap is already included in the measure of tap density. Taps are indeed more likely to be subscribed to and active in areas where the supply is reliable, which results in a spatial clustering captured by the tap-density measure. The lack of influence of wealth quintile in model predictions may also be explained by the strong correlation between wealth index and tap water service indicator (Pearson's correlation coefficient $r = 0.49$).

The good discrimination ability of the drinking water model indicates that factors exogenous to households play an important role in stored drinking water contamination. Assuming that households do have a preference for tap water as a drinking water source due to quality/health concerns, distance from the nearest river influences how access to a tap actually translates into tap use as a drinking water source, especially when tap access is relatively poor. The contamination OR between drinking water collected at a tap or at surface sources by households in Uvira (0.2; 95% CI 0.11–0.37) relates well to that reported by Bain et al. in their review of faecal contamination of drinking water in low- and middle-income countries.[28] Beyond encouraging its use as drinking water source, access to a closer and more reliable tap could plausibly reduce the risk of post-collection contamination linked to transportation and storage of tap water. Another review indeed determined the odds of contamination for stored household water more than double those of contamination of source water, even for piped water supplies.[29] The better performance of the drinking water model for poorer households implies that these households are even more dependent on exogenous factors, with few coping mechanisms. Even in the case of poor tap water service, wealthier households may have a stronger preference for piped water as a result of higher education achievements or social pressure, and use coping mechanisms—better storage containers, point-of-use water treatment and payment for their water to be collected for them at a tap for example.[20]

Predicting household water consumption at home for domestic activities provides weaker evidence of the relationship between proxy measures for access to tap water or surface water and the quantity of water used. The nonlinear relationship between access to water and water use has been highlighted before, especially outside of rural contexts with few remote sources.[30] When multiple sources are available, water used for different purposes is indeed a heterogeneous good, with characteristics—perceived quality, convenience of access and monetary and opportunity costs—influencing households collection and use behaviours depending on their education and capacity to pay.[31] Even in a setting like Uvira, where the diversity of sources is limited—tap,

rivers or lake—simple measures of access are unlikely to capture enough of the preferences guiding households' choices in source use, and how access to the sources chosen impacts the quantity of water used at home. By attempting to replace in the model the two distance variables by a single one representing distance to the nearest surface water source, the predictive performance of the model decreased sufficiently to highlight different households' behaviours towards the two surface water sources. This could reflect a preference for one or the other surface source at equal distance, either related to the distance measure (e.g. distance not capturing the difficulties to access the lake shore in parts of town) or related to the water characteristics. The predictive model reached here is nevertheless compatible with the concept of a water consumption plateau put forward by Cairncross et al., according to which water consumption of households remains stable when the source is located between 5 and 30 min from the house, and decreases only when the location is further away than 30 min, while increasing massively as the source is brought closer than 5 min (source within premises).[2] Our model suggests that water consumption only decreases with increasing distance when the nearest river or the lake are further away than 500 meters, which roughly corresponds to 30 min (20 min there and back, with 10 min to fill up containers). The results, however, do not reflect the sharp increase in water consumption observed when a source is accessible on household premises, as the tap water indicator is unlikely to distinguish between households with an active tap within their compound and those with an active tap at a short distance but outside their compound, with less convenient access.

Our study has several limitations. The data used in the analysis were collected at a single point in time, during the rainy season. Although seasonality has a limited impact on the availability of surface and tap water, it was previously reported that tap water supply interruptions are more frequent in Uvira during the dry season versus the rainy season due to more frequent power cuts that affect the water treatment plant operations.[32] Heavy rains on the steep slopes overlooking Uvira also cause rivers and lake along the shores to become extremely muddy for several hours, deterring people from using the surface water. In addition, our data do not support the inclusion of areas of Uvira located more than 2 km from a river. A single measure of the outcomes will also not capture daily or weekly differences in activities, which may give rise to important variations in daily water consumption, with some activities not carried out daily. Reported water quantity use is also likely to suffer from a substantial random error. The observational nature of the relationships established by the two models should also be highlighted, and a causal relationship between the predictors and the outcomes should not be assumed, even though some degree of causality is plausible, especially for drinking water. Finally, the predictive nature of our methodology is unsuitable for untangling further the multiple factors affecting water use by households, especially when not only considering drinking water.

Geographical predictions of improved drinking water or improved sanitation coverage within SSA countries were previously performed at the district level or rural community level.[33,34] Multi-country meta-analyses also provide estimates of contamination levels for drinking water collected at different sources or different levels of tap water service.[35] In addition, several methods were developed to estimate and predict water demand for urban piped water systems in low- or middle-income countries, with a focus on engineering and financial planning.[31] However, we could not identify previous attempts to spatially predict drinking water quality or domestic water quantity used at the household level that provides valuable information for water supply improvements targeting health benefits. We believe that our approach and the results obtained here in Uvira warrant a further exploration of their value in other contexts. Should it be generalisable, our approach could allow identification of priority intervention areas for water supply improvements in urban settings, without implementing costly household surveys, by geolocalising taps, possibly along with the volume of water consumed at a single point in time, geolocalising alternative sources of water and population data at a relatively small scale. The present results also support their use to investigate the possible relationship between tap water access and observed time and space patterns of health outcomes incidence, such as suspected and confirmed cholera.

METHODS

Study area

Uvira extends ~10 km along the northernmost shores of Lake Tanganyika, <2 km inland and is crossed by five permanent rivers.

The population, estimated at 254,000 inhabitants in November 2017, relies on both surface water sources and the tap water system managed by the national water agency Regideso. Water from the river Mulongwe is drawn upstream of inhabited areas and treated at the Regideso water treatment plant before being fed into a single 1600 m^3 reservoir, from which it is distributed to private and shared taps by gravity. The current distribution system fails to serve adequately the taps located further away from the reservoir or higher in altitude, and the daily amount of water distributed is irregular. There are no wells or boreholes due to unfavourable geological terrain.

Uvira's water supply infrastructure is undergoing refurbishment and expansion since September 2018 through a project funded by the European Union (EU), the French Development Agency (AFD) and the Veolia Foundation in partnership with the Regideso.

Household data

The household data were collected as part of two surveys of household water-related practices, conducted in October 2016 and October–November 2017.

Recent, reliable data on the location and functionality of Regideso taps were unavailable at the time households were sampled in October 2016. Therefore, to establish a cohort of households representing a wide range of access to tap water in Uvira, and in the absence of a household-sampling frame, a two-stage random spatial sampling method was used based on a piped water availability index. Details of this sampling method are given in Supplementary Information, along with households' selection and enrolment methods. During the second survey implemented in 2017, the buildings sampled and georeferenced during the first survey were revisited. If the household inhabiting the building was different, the same enrolment process was used with the new family.

During both surveys, households were interviewed about water-related practices at home. This included the amount of fresh (as in not recycled from a previous activity) water used the previous day for various domestic activities using a visual aid, the number of adults and children present that day and water use at the source. The amounts of fresh water used at home for bathing, laundry, dishwashing, food and produce rinsing, dwelling cleaning, handwashing and drinking were added up into a total amount of fresh water used at home for domestic activities. Water used at home to prepare food or items for sale or to render a paid-for service was excluded from the total.

During each interview in 2017, a 150-ml sample of stored drinking water was collected in a sterile sample Whirl-pak® ThioBag® (Nasco, Fort Atkinson, WI) containing 30 g of sodium thiosulfate. The participant was requested to provide the water that would be used for drinking at the time of the interview, with the utensils usually used for serving such drinking water and reported where the stored drinking water had been collected.

Samples were brought back daily in cool bags to the Centre de Recherche Hydrobiologique (CRH) in Uvira and analysed within 6 h of collection. Turbidity was measured with a digital turbidimeter, and the volume of water filtered adjusted to aim for a turbidity of 3NTU once diluted with sterile water to reach a volume of 100 ml. Between 5 and 100 ml of each sample was filtered in sterile conditions through a 0.47-µm filter, and the filter was then aerobically incubated on sterile pads saturated with mColiblue24 broth (Hach Co, Frederick, MD) for 24 h at 35 °C in a portable incubator. The number of blue colonies grown on the filter was then counted and multiplied by the appropriate factor to obtain the number of colony-forming units (CFU) *Escherichia coli* per 100 ml. Although there is still debate about the relationship between CFU *E. coli* per 100 ml in drinking water and health risks, we used a single cut-off of

one CFU *E. coli* per 100 ml to define a binary outcome of contaminated/non-contaminated drinking water.[36]

A household wealth index based on ownership of durable items and dwelling characteristics was constructed and classified into quintiles (details in Supplementary Information).

For each household interviewed, the shortest distances from one of the town's five main rivers and from Lake Tanganyika were computed with a slope adjustment based on Tobler's hiking function.[37]

Tap water service indicator

Data on the daily volume of water treated were collected from the register held at the Regideso water treatment plant. Unique identifiers of active tap connections were retrieved from the Regideso customers database, along with the volume invoiced in March 2018, and each of them geolocalised.

Based on these data, an index of tap water availability was constructed. In brief, a kernel density function with a radius of 250 m around each tap was used to combine data on the water volume invoiced for each functional tap with information on population density. This produced a smooth "surface" with a resolution of ~34 m × 42 m of tap water availability across Uvira in litres per capita per day (LCPD). This indicator was extracted at interviewed households' location and used as a continuous variable.

Population data sources, delineation methods for built-up areas and the construction of the tap water availability index are detailed in Supplementary Information.

Statistical methods

Logistic regression was used to assess whether the tap water service indicator, distance to the closest river and distance to the lake shore were predictive of the probability of stored household drinking water being contaminated. The model was developed on a random sample of 60% of the data set records (training), and tested on the remaining 40%. Model selection followed a hierarchical backward strategy, starting with a full model, including the three variables as continuous and all possible two- and three-way linear interactions.[38] Model terms were eliminated starting with the least significant, to identify the model with the lowest value of Akaike's information criterion (AIC).

The Hosmer–Lemeshow statistic and a plot of predicted against observed probabilities by decile of predicted probability were used to assess the model fit and calibration to the training and testing data sets. Lowess plots of standardised residuals and Pregibon's goodness-of-link test were used to assess whether assuming a linear relationship was appropriate.[39] The discrimination of the model was assessed using receiver–operator curves and derived area under the curve (ROC and AUC). The model's discriminative ability was examined by household wealth quintile, to assess whether model performance varied with household wealth. The importance of reported drinking water source as a possible explanatory variable for contamination was also explored. The predicted probability of household drinking water contamination was then mapped using the coefficients of the selected model derived from the training data set.

A similar approach was used to model the total quantity of water used within a household using linear rather than logistic regression and including demographic composition of the household (the number of children, number of adults) as additional covariates. Tap water service being expressed as a quantity of tap water per capita, we used the indicator multiplied by the number of household members present the previous day. Lowess plots of standardised residuals and Pregibon's goodness-of-link test were used to assess whether assuming a linear relationship was appropriate, while adjusted R^2 and root-mean-square error (RMSE) described the model fit to the data. Wealth quintile and reporting of having performed water-consuming activities (laundry, dishwashing and bathing) at the source were added separately as independent variables to the selected model in order to investigate whether they improved the model predictions. We also investigated to what extent water use at the source was predicted by tap water service, distance from the nearest river and distance from the lake shore. The predicted quantity of water used per household, for households having the median household composition, was then mapped.

To avoid undue influence of extreme outliers on model parameters, four records were excluded from the analysis: two households located at more than 4000 meters from the closest river and two households for which more than 20 members were reported present the day preceding the interview. To avoid extrapolating model estimates, these exclusions were taken into account by limiting the mapping to areas less than 2000 meters from the nearest river.

A sensitivity analysis was performed by replacing the tap water service indicator constructed at a 250-m radius (ind250) with indicators constructed at radius 500 m or 750 m (ind500 and ind750). A tap-density indicator at 250 m, 500 m and 750 m (dens250, dens500 and dens750) was also used. Expressed in number of taps per 1000 people, this indicator was constructed with a constant weight applied to all taps, ignoring individual taps invoicing and the possible variations in water availability and tap reliability invoicing data may represent.

Data preparation and tap water service indicator construction were performed with ArcGIS ArcMap 10.3 (ESRI, Aylesbury, UK) and R,[40] in particular the R package "sparr".[41] Data were analysed with STATA 14.2 (StataCorp, College Station, TX).

Ethical considerations

Household interviews were only performed after written consent to participate was obtained, in accordance with study approvals from the ethics committees of the School of Public Health at the University of Kinshasa, Democratic Republic of the Congo (ESP/CE/088c/2017), and of the London School of Hygiene and Tropical Medicine, United Kingdom (No. 10603). The study is part of a broader evaluation of the impacts of tap water supply improvements on cholera and other diarrhoeal diseases in Uvira registered at clinicaltrials.gov (Reference: NCT02928341). All the data were anonymised before analysis.

DATA AVAILABILITY

Anonymised data are available from the corresponding author upon request.

Received: 26 June 2019; Accepted: 14 November 2019;
Published online: 16 December 2019

REFERENCES

1. White, G. F., Bradley, D. J. & White, A. U. *Drawers of Water; Domestic Water Use in East Africa* (University of Chicago Press, 1972).
2. Cairncross, S. & Feachem, R. G. *Environmental Health Engineering in the Tropics: Water, Sanitation and Disease Control* 3rd edn (Taylor & Francis Group, Routledge, 2019).
3. Feachem, R., Bradley, D., Garelick, H. & Mara, D. D. *Sanitation and Disease: Health Aspects of Excreta and Wastewater Management*. (John Wiley & Sons, 1983).
4. Clasen, T., Schmidt, W. P., Rabie, T., Roberts, I. & Cairncross, S. Interventions to improve water quality for preventing diarrhoea: systematic review and meta-analysis. *Br. Med. J.* **334**, 782 (2007).
5. Esrey, S. A., Feachem, R. G. & Hughes, J. M. Interventions for the control of diarrhoeal diseases among young children: improving water supplies and excreta disposal facilities. *Bull. World Health Organ* **63**, 757–772 (1985).
6. Esrey, S. A., Potash, J. B., Roberts, L. & Shiff, C. Effects of improved water supply and sanitation on ascariasis, diarrhoea, dracunculiasis, hookworm infection, schistosomiasis, and trachoma. *Bull. World Health Organ* **69**, 609–621 (1991).
7. Fewtrell, L. et al. Water, sanitation, and hygiene interventions to reduce diarrhoea in less developed countries: a systematic review and meta-analysis. *Lancet Infect. Dis.* **5**, 42–52 (2005).
8. Overbo, A., Williams, A. R., Evans, B., Hunter, P. R. & Bartram, J. On-plot drinking water supplies and health: A systematic review. *Int. J. Hyg. Environ. Health* **219**, 317–330 (2016).
9. Pruss-Ustun, A. et al. Burden of disease from inadequate water, sanitation and hygiene for selected adverse health outcomes: an updated analysis with a focus on low- and middle-income countries. *Int J. Hyg. Environ. Health*. https://doi.org/10.1016/j.ijheh.2019.05.004 (2019).
10. Stelmach, R. D. & Clasen, T. Household water quantity and health: a systematic review. *Int J. Environ. Res. Public Health* **12**, 5954–5974 (2015).
11. Waddington, H., Snilstveit, B., White, H. & Fewtrell, L. *Water, Sanitation and Hygiene Interventions to Combat Childhood Diarrhoea in Developing Countries* (New Delhi: International Initiative for Impact Evaluation, 2009).
12. Wolf, J. et al. Impact of drinking water, sanitation and handwashing with soap on childhood diarrhoeal disease: updated meta-analysis and meta-regression. *Tropical Med. Int. Health* **23**, 508–525 (2018).

13. Wolf, J. et al. Assessing the impact of drinking water and sanitation on diarrhoeal disease in low- and middle-income settings: systematic review and meta-regression. *Trop. Med. Int Health* **19**, 928–942 (2014).
14. World Health Organization & UNICEF. *Progress on drinking water, sanitation and hygiene: 2017 update and SDG baselines*. (World Health Organization, Geneva, 2017).
15. Clasen, T. F. Millennium development goals water target claim exaggerates achievement. *Tropical Med. Int. health* **17**, 1178–1180 (2012).
16. Bivins, A. W. et al. Estimating infection risks and the global burden of diarrheal disease attributable to intermittent water supply using QMRA. *Environ. Sci. Technol.* **51**, 7542–7551 (2017).
17. Wright, J., Gundry, S. & Conroy, R. Household drinking water in developing countries: a systematic review of microbiological contamination between source and point-of-use. *Trop. Med. Int. Health* **9**, 106–117 (2004).
18. Kumpel, E. & Nelson, K. L. Comparing microbial water quality in an intermittent and continuous piped water supply. *Water Res.* **47**, 5176–5188 (2013).
19. Kumpel, E. & Nelson, K. L. Intermittent Water Supply: Prevalence, Practice, and Microbial Water Quality. *Environ. Sci. Technol.* **50**, 542–553 (2016).
20. Majuru, B., Suhrcke, M. & Hunter, P. R. How do households respond to unreliable water supplies? A Systematic Review. *Int. J. Environ. Res. Public Health* **13**, https://doi.org/10.3390/ijerph13121222 (2016).
21. Elliott, M. et al. Addressing how multiple household water sources and uses build water resilience and support sustainable development. *npj Clean. Water* **2**, 6 (2019).
22. Thomas, E., Andrés, L. A., Borja-Vega, C. & Sturzenegger, G. *Innovations in WASH impact measures: water and sanitation measurement technologies and practices to inform the Sustainable Development Goals*. (The World Bank, 2018).
23. World Health Organization & UNICEF. *Progress on drinking water, sanitation and hygiene: 2015 update and MDG assessment*. (World Health Organization, Geneva, 2015).
24. African Ministers' Council on Water (AMCOW). *Water Supply and Sanitation in the Democratic Republic of the Congo: Turning Finance into Services for 2015 and beyond* (Water and Sanitation Program, 2011).
25. Bompangue, D. et al. Lakes as source of cholera outbreaks, Democratic Republic of Congo. *Emerg. Infect. Dis.* **14**, 798–800 (2008).
26. Brecht, I. et al. Recurrent cholera outbreaks, Democratic Republic of the Congo, 2008–2017. *Emerg. Infect. Dis. J.* **25**, 856 (2019).
27. Jeandron, A., Cumming, O., Rumedeka, B. B., Saidi, J. M. & Cousens, S. Confirmation of cholera by rapid diagnostic test amongst patients admitted to the cholera treatment centre in Uvira, Democratic Republic of the Congo. *PLoS ONE* **13**, e0201306 (2018).
28. Bain, R. et al. Fecal contamination of drinking-water in low- and middle-income countries: a systematic review and meta-analysis. *PLOS Med.* **11**, e1001644 (2014).
29. Shields, K. F., Bain, R. E. S., Cronk, R., Wright, J. A. & Bartram, J. Association of supply type with fecal contamination of source water and household stored drinking water in developing countries: a bivariate meta-analysis. *Environ. Health Perspect.* **123**, 1222–1231 (2015).
30. Evans, B. E. et al. *Public Health and Social Benefits of at-house Water Supplies* Vol. 53. (Department for International Development, 2013).
31. Nauges, C. & Whittington, D. Estimation of water demand in developing countries: an overview. *World Bank Res. Obser* **25**, 263–294 (2010).
32. Jeandron, A. et al. Water supply interruptions and suspected cholera incidence: a time-series regression in the Democratic Republic of the Congo. *PLoS Med.* **12**, e1001893 (2015).
33. Oswald, W. E. et al. Prediction of low community sanitation coverage using environmental and sociodemographic factors in Amhara Region, Ethiopia. *Am. J. Trop. Med. Hyg.* **95**, 709–719 (2016).
34. Pullan, R. L., Freeman, M. C., Gething, P. W. & Brooker, S. J. Geographical inequalities in use of improved drinking water supply and sanitation across Sub-Saharan Africa: mapping and spatial analysis of cross-sectional survey data. *PLOS Med.* **11**, e1001626 (2014).
35. Bain, R., Johnston, R., Mitis, F., Chatterley, C. & Slaymaker, T. Establishing sustainable development goal baselines for household drinking water, sanitation and hygiene services. *Water* **10**, 1711 (2018).
36. Gruber, J. S., Ercumen, A. & Colford, J. M. Jr. Coliform bacteria as indicators of diarrheal risk in household drinking water: systematic review and meta-analysis. *PloS ONE* **9**, e107429–e107429 (2014).
37. Tobler, W. *Three Presentations on Geographical Analysis and Modeling: Non- Isotropic Geographic Modeling; Speculations on the Geometry of Geography; and Global Spatial Analysis (93-1)*, https://escholarship.org/uc/item/05r820mz (1993).
38. Kleinbaum, D. G., Klein, M. & Pryor, E. R. *Logistic Regression: A Self-learning Text* 3rd edn (Springer, 2010).
39. Pregibon, D. Goodness of link tests for generalized linear models. *J. R. Stat. Soc. Ser. C. (Appl. Stat.)* **29**, 14–15 (1980).
40. *R: A language and environment for statistical computing*. (R Foundation for Statistical Computing, Vienna, Austria, 2013).
41. Davies, T. M., Hazelton, M. L. & Marshall, J. C. sparr: analyzing spatial relative risk using fixed and adaptive kernel density estimation in R. *J. Stat. Softw.* **1**. https://doi.org/10.18637/jss.v039.i01 (2011).

ACKNOWLEDGEMENTS

The study was conducted with the funding of the Veolia Foundation (PF9233) and the French Development Agency (ERS-EVA 364-2015). We are grateful to the funders for their valuable comments on the final paper. We would like to thank all study participants for their valuable time answering our questions, and to the entire team of interviewers and supervisors in Uvira for their hard work. We are also grateful to the staff of the Centre de Recherche en Hydrobiologie for a fruitful collaboration and of OXFAM GB in DRC for their continuing support in the field.

AUTHOR CONTRIBUTIONS

A.J., O.C. and S.C. conceptualised the household survey methodology. A.J. designed survey tools and curated the data. A.J. and L.K. managed laboratory resources and performed the water quality analyses. A.J. and S.C. developed the model and performed the formal analyses. The paper first draft was prepared by A.J. and substantially reviewed and revised by all authors.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41545-019-0047-9.

Correspondence and requests for materials should be addressed to A.J.

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.