scientific reports

Check for updates

OPEN

# A Bayesian Belief Network model to link sanitary inspection data to drinking water quality in a medium resource setting in rural Indonesia

D. Daniel✉, Widya Prihesti Iswarani, Saket Pande & Luuk Rietveld

**Assessing water quality and identifying the potential source of contamination, by Sanitary inspections (SI), are essential to improve household drinking water quality. However, no study link the water quality at a point of use (POU), household level or point of collection (POC), and associated SI data in a medium resource setting using a Bayesian Belief Network (BBN) model. We collected water samples and applied an adapted SI at 328 POU and 265 related POC from a rural area in East Sumba, Indonesia. Fecal contamination was detected in 24.4 and 17.7% of 1 ml POC and POU samples, respectively. The BBN model showed that the effect of holistic—combined interventions to improve the water quality were larger compared to individual intervention. The water quality at the POU was strongly related to the water quality at the POC and the effect of household water treatment to improve the water quality was more prominent in the context of better sanitation and hygiene conditions. In addition, it was concluded that the inclusion of extra "external" variable (fullness level of water at storage), besides the standard SI variables, could improve the model's performance in predicting the water quality at POU. Finally, the BBN approach proved to be able to illustrate the interdependencies between variables and to simulate the effect of the individual and combination of variables on the water quality.**

Water quality has a prominent place in the Sustainable Development Goal 6.1[1], because it has been recognised that unsafe drinking water is responsible for high numbers of diarrheal morbidity and mortality among children below the age of five[1]. Water quality analysis becomes important because supplied water, especially in low and middle-income countries (LMICs), is often contaminated, even though it is categorised as an improved water source[2]. Groundwater, which is considered safer than surface waters, is also found contaminated in many locations[3]. In Addition, high levels of contamination has been found at the household level in LMICs and water quality often deteriorates after collection[4–6].

To tackle this, the World Health Organization (WHO) and International Water Association (IWA) launched a Water Safety Plan (WSP) concept, which is a comprehensive risk assessment and management covering all steps in water supply from catchment to consumers[7]. The goal is to minimise the risk of contamination and provide safe drinking water to people. Identifying potential sources of contamination is part of the risk assessment and one of the critical elements in WSP.

In order to assess potential sources of contamination in a water supply system, systematic observation, called sanitary inspections (SI), are performed. SI variables record potential sources of contamination based on "on-site inspection and evaluation by qualified individuals of all conditions, devices, and practices in the water-supply system that pose an actual or potential danger to the health and well-being of the consumer"[8]. SI have the advantage to be easy to implement, not expensive, can be adapted to the local context, and can give a quick snapshot of potential causes or pathways of contamination. However, SI are not a substitute for drinking water quality testing, but identify contamination source in the system, especially in the context of risk management, and can be used to design appropriate actions to change the situation[9]. Therefore, it has been recommended to accompany drinking water quality testing with SI[10].

Department of Water Management, Delft University of Technology, Delft, The Netherlands. ✉email: d.daniel@tudelft.nl

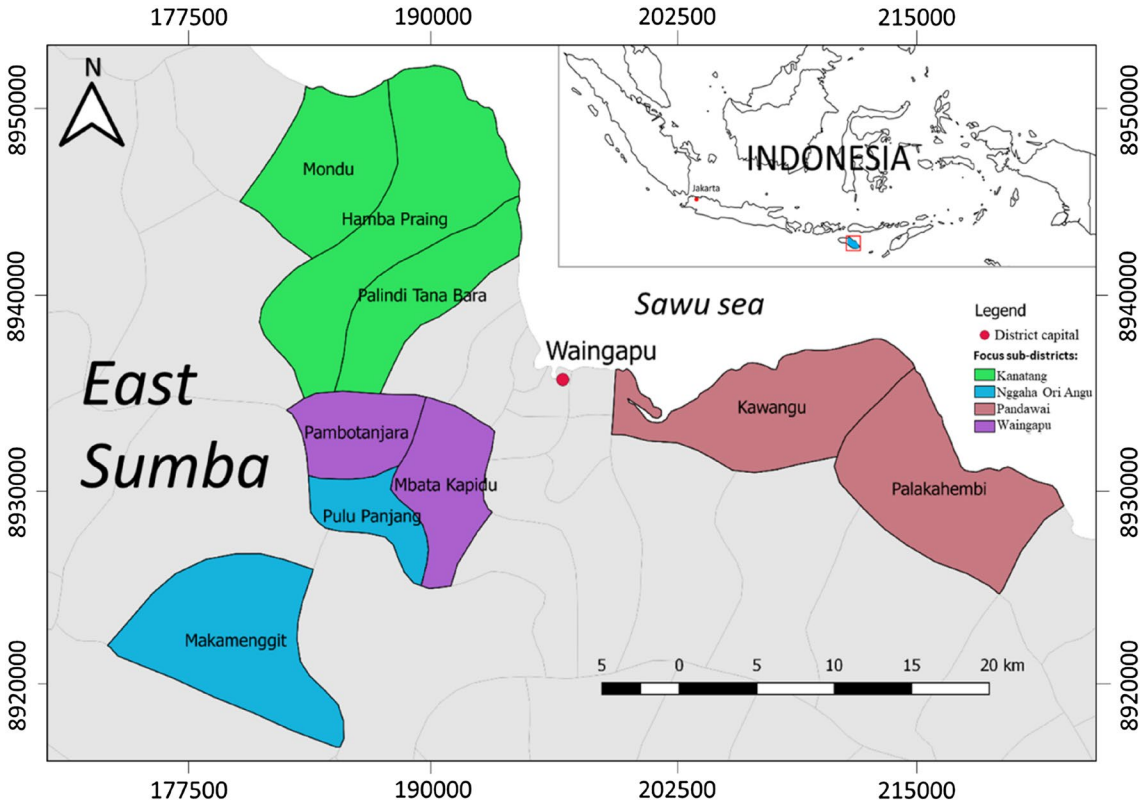


**Figure 1.** Map of the study location. There were nine villages visited in four sub-districts. The map is drawn using QGIS[24].

Conducting drinking water quality testing in LMICs, however, can be challenging, especially because of limited resources such as laboratory facilities or infrastructure[11]. Bain et al.[12] summarised all available microbial water quality tests for low and medium resource settings and they classified the resource settings into low, medium, and high resource settings. A low resource setting has been characterised as having no laboratory equipment and 24 h electricity; the medium one having at least a basic laboratory or clean space with 24 h electricity; while the high resource setting is equipped with reliable 24 h electricity and a modern laboratory. Researchers are able to choose relevant water quality tests according to local context or situation.

Attempts have been made to link SI data to drinking water quality in order to be able to judge the reliability of the system. The most common approach has been to analyse the SI and drinking water quality by using statistical analyses, e.g., bivariate correlation or regression analyses, especially in high resource settings[6,10,13–16].

Bayesian Belief Network (BBN) is another alternative to analyse factors responsible for the water quality[17,18]. BBN offers benefits compared to other statistical methods, such as the ability to integrate quantitative and qualitative information in the model and an intuitive visualisation of the hypothetical causal relationships that can aid stakeholders with less technical knowledge in understanding the system[19].

However, the application of BBN in analysing water quality at the household level [mentioned as a point of use (POU)] and at water source or point of collection (POC) is very limited. Hall and Le[20] utilised BBN to predict the faecal contamination of drinking water by household's socio-economic characteristics as predictor variables, however not using SI variables. To the authors' knowledge , the present study is the first to link drinking water contamination at the POU with a *combination* of water quality at POC, the hygiene conditions in the household, water handling, and household water treatment (HWT) practices in a medium resource setting. This study aims to delineate the microbial water quality and general sanitary conditions in POC and POU in the district of East Sumba, Indonesia.

## Methods

**Study setting.** A cross-sectional study was conducted in July–August 2019 in the district of East Sumba, Province East Nusa Tenggara, Indonesia (Fig. 1). This study is the continuation of a previous household water treatment study conducted in the same area[21]. A total of 328 households in nine villages in four sub-districts were revisited during this study. This area is known as one of the poorest areas in Indonesia where open defecation is still common and there is high prevalence of children's malnutrition[22]. The topography of the area is hilly. Furthermore, about 40% the total populations in East Sumba relied on wells as their main water source and only 18% had access to piped distribution system in 2017[23]. No water treatment is conducted in the rural piped distribution systems in this area.

| Point of collection (POC)[b] | Surrounding environment–hygiene condition | Water storage condition and HWT |
|---|---|---|
| *Type of POC* [Which source do you use for drinking water purpose right now?][b] | *Still practise open defecation* [What types of toilet do you have?] | *Storage covered* [Is the water storage being covered (at that time)?] |
| *Livestock nearby* [Is there livestock near the point of collection (POC), 10 m?] | *Livestock nearby* [Is there livestock around the house?] | *Storage cracked* [Is the container cracked?][a] |
| *Prone to erosion* [Is the area uphill from the source visibly eroded or prone to erosion?] | *Floor cleanliness* [How is the cleanliness of the house floor?][a] | *Place of storage* [When not in use, is the storage container kept in a place where it may become contaminated? E.g., can be reached by animal easily; open space (risk by flies), etc.] |
| *Excreta / garbage nearby* [Is excreta or garbage found within 10 m of the tap stand/water source?] | *Faeces around* [Is there human or animal faeces in the yard (or even inside the house)?][a] | *fullness level of water at storage* [How full is the water storage?][a,c] |
| *Proper fencing* [Is there proper fencing or a barrier around the well to prevent contact with animals?] | *Garbage around* [Is there garbage around the house?] | Household water treatment [Is the water in the storage treated?] |
| *Latrine within 10 m* [Distance to the nearest latrine (m)] | *Flies around* [Could you see flies around the water storage container?] | |
| *Cracked structure* [Are there any damages/cracks in the system/source?] | | |
| *E. coli detected at POC/well*[b] | | |

**Table 1.** Information used for the analysis. [a]The sentence inside the [ ] were the questions in the sanitary inspection and the italic words were the variable/node name in the BBN. [b]Based on water quality testing. [c]External variable besides standard SI variables.

Approximately 100 ml of drinking water sample, i.e., from the drinking water storage container, was taken at each household. The households were asked to give water in the same way as for drinking water. The water samples were put in Nasco Whirl–Pak bags and kept inside a thermos during the transport to the field lab. All the samples were analysed within six hours after collection. We only analysed the microbial water quality and used *E. coli* as an indicator bacteria for fecal contamination in water[25]. We took 1 ml of sample using a 1 ml sterile pipette and placed it on a Nissui Compact dry EC plate (CDP) and incubated for 24 h at 35 ± 2 °C[26]. After incubation, we counted the colony forming units (CFU) of *E. coli* in the CDP and reported in concentration units (CFU/1 ml). The process was conducted as sterile as possible to prevent contamination from sample processing, e.g., using hand gloves and sterile pipette tips when processing the sample, avoid touching the inside of the whirl-pack bag when collecting and processing the sample, and working in a stable and clean space. The sample processing was conducted by two master students from Delft University of Technology who were familiar with microbial water quality analyses. According to the classification of Bain et al.[12], our analysis was categorised as medium resource setting, e.g., there was neither distilled water and proper disinfection for laboratory equipment. Data were collected during the dry season with temperature in that area ranging from 25 to 26 °C.

For the SI, we used the Open Data Kit (ODK) software on a smartphone, and the data were transferred to a computer for analysis. We did SI at POCs and POUs. Information taken at a POC and POU can be found in Table 1. Participation was voluntary and a written informed consent was obtained from all participants. The study was approved by the Human Research Ethic Committee of Delft University of Technology and the Agency for Promotion, Investment, and One-Stop Licensing Service at the district level. All experiments were conducted in accordance with relevant guidelines and regulations.

**Bayesian Belief Network (BBN).** BBN is a directed acyclic graph showing a hypothetical causal relationship between "causal" variables (where the arrow start; called "parent nodes" in BBN) and an "affected" variable (called "child node")[27]. The strength of the relationship between parent and child node is shown by the values in the Conditional Probability Tables (CPT) of the child node. The CPT values are showing the probability of a child node will be in a particular state or category, given all possible combination of the states of its parent nodes. The CPT values can be obtained from expert or stakeholder judgment or elicitation, the output of other models or calculations, or by direct measurement. Cain[19] provides a good and clear explanation of using a BBN in the water sector.

**Data analysis.** A BBN's structure is often inspired by a conceptual theory or framework or by consensus between experts in that field[28]. There are some conceptual frameworks from previous water, sanitation, and hygiene (WASH) studies that can be adapted into a BBN's structure[29,30], including the well-known F-diagram [31]. According to those frameworks, there are four main clusters of determinants of water quality at POU: (1) Surrounding environment–hygiene condition, (2) HWT, (3) (the water quality at) POC, and (4) the water storage conditions (see Fig. 2). All variables for these four cluster are often included in a standard SI form[8].

However, Navab-Daneshmand et al.[29] argues that fecal contamination at the household level in LMICs is complex. This implies that there might be other variables, besides SI variables, that could correlate with the household drinking water quality, such as container material, duration of storing water, inappropriate extraction water from storage, etc[32–34]. However, all these "external" factors are not included in the standard SI form[8].

Based on the above mentioned literature, we created a conceptual model of potential factors that could influence the water quality at the household level (Fig. 2). The conceptual model includes multiple contamination

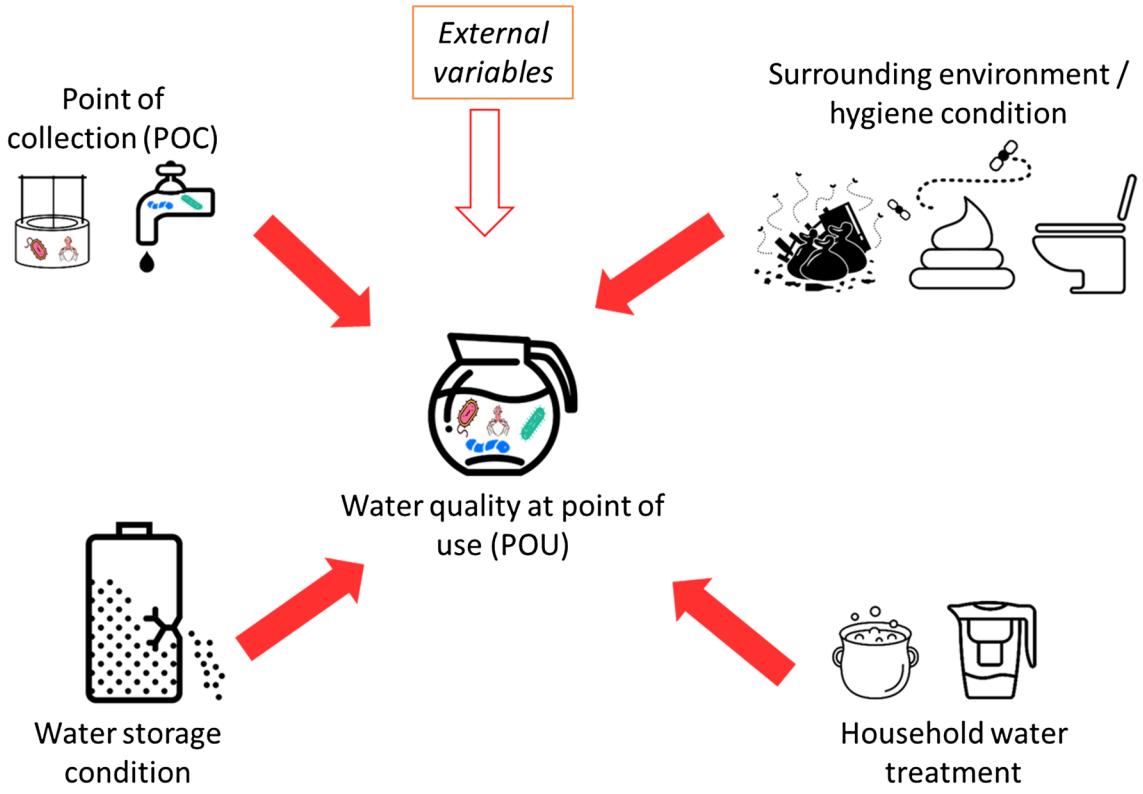


**Figure 2.** The conceptual model of five clusters of the determinants of water quality at a point of use (POU). Red arrows indicate that the variables are often included in a standard SI form and white arrow is not included in the standard SI form.

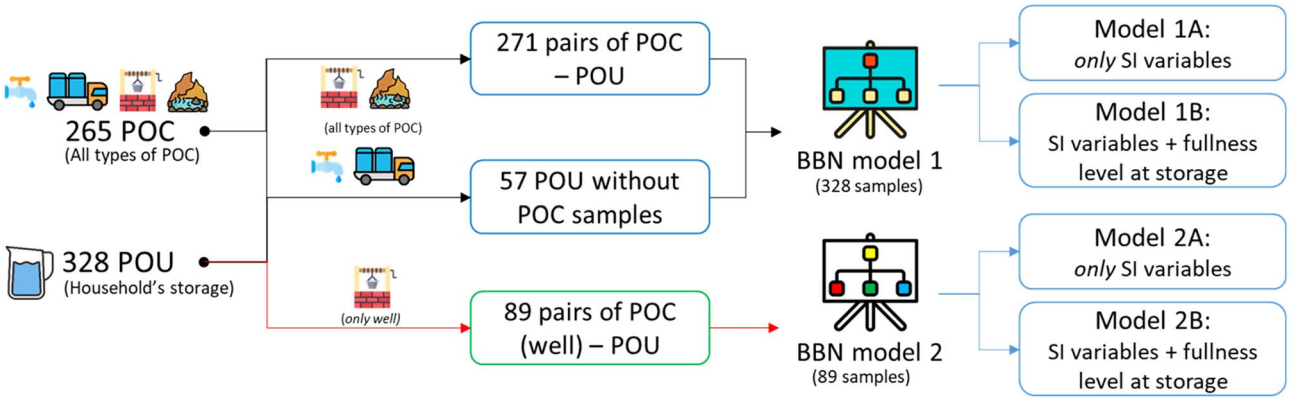


**Figure 3.** Overview of the datasets and analysis.

pathways in a system[35] and was used to create the BBN's structure by clustering SI variables based on those five clusters.

Because some houses used the same POC, we could make pairs of 271 POCs–POUs (Fig. 3). 49 POU did not have POC samples, i.e., POC samples were not taken, mostly due to long distance walk (> 30 min return trip). However, these 49 POU samples were included in the BBN analysis, since the EM algorithm compensated for the missing information with the available data[36].

Four BBN models of the water quality at the POU were created (Fig. 3). BBN model 1 (A and B) and 2 (A and B) differ in terms of the variables used in the cluster of POC. For BBN model 1 we added node *Type of POC* as a parent node for *E. coli detected at POC* (Figs. 4, 5). But for BBN model 2 we used information of the SI at the POC as parent nodes of *E. coli detected at POC*, but we modelled only one type of POC: well (Figs. 6, 7). That is because the SI information that we collected at POC were only relevant to the well's characteristics. For BNN model 1, we had in total of 328 samples and for BNN model 2 was only 89 well samples (Fig. 3).

In addition, we added one extra variable, *fullness level of water at storage*, on top of both models and compared the model's performance, i.e., BBN model 1A vs 1B and model 2A vs 2B. This variable could indicate the duration of storing water, because water quality could deteriorate over time[4]. Thus, BBN model 1A and 2A were

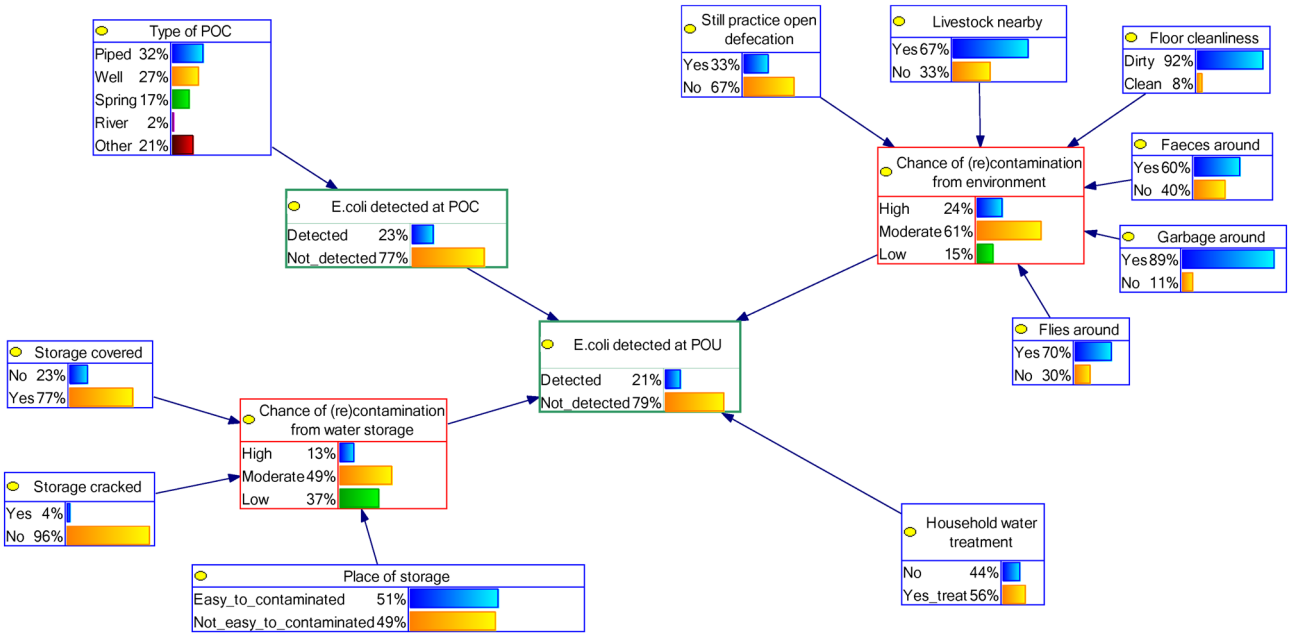


**Figure 4.** BBN model 1A (type of POC as a parent node of "*E. coli* detected at POC"). Blue nodes: data obtained from SI; green nodes: data obtained from water quality testing; red nodes: intermediate nodes were obtained by summation of the value in the outer nodes. The percentages in each node indicate the probability of a node being in a certain state, e.g., 56% of the households perform household water treatment.

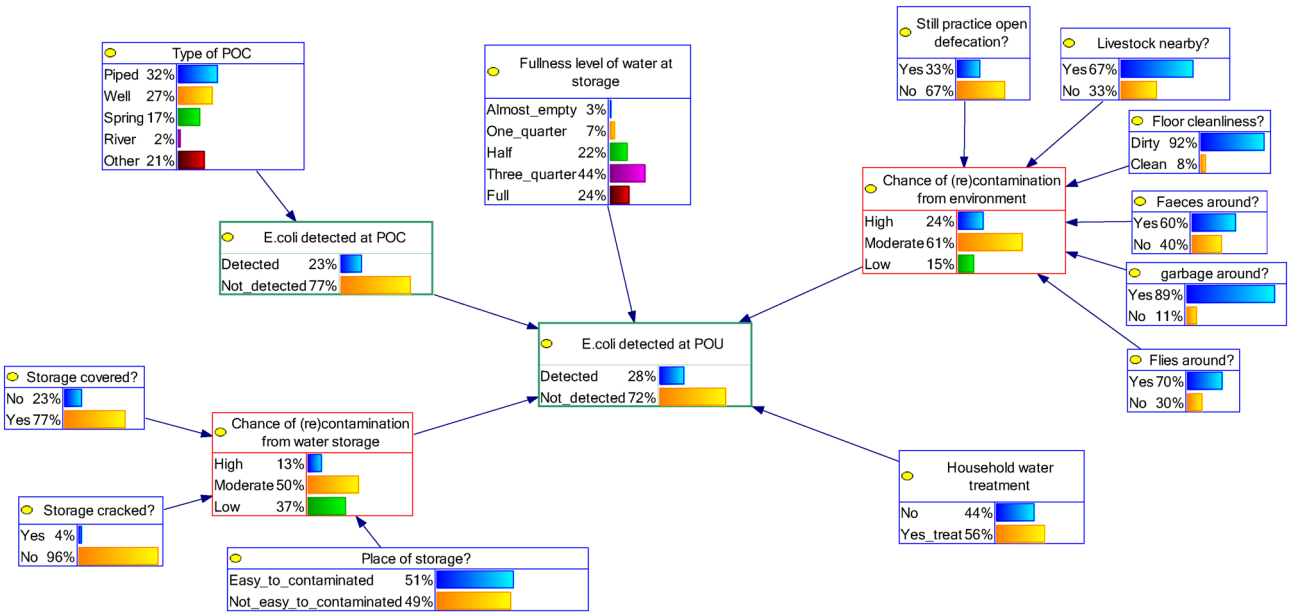


**Figure 5.** BBN model 1B (type of POC as a parent node of "*E. coli* detected at POC" and adding node "fullness of water at storage" as one of the parent nodes of "*E. coli* detected at POC").

the BBN models with SI variables *only* and BBN model 1B and 2B were the BBN models with SI variables plus variable *fullness level of water at storage*. The results of validation tests, i.e., AUC value, indicated the model's performance. The predictive inference tests were then conducted using BBN models with the best performance.

Moreover, Since it is not recommended to have many parent nodes in BBN[19], we needed to reduce the BBN structure as much as possible. Clustering the SI variables reduces the parent nodes of the outcome node, e.g. water quality at the POC. All variables in the SI for POC were grouped as one cluster and the variables in the SI related to water storage were grouped as another cluster. In the latter case, e.g., three variables related to the condition of the water storage, *Storage covered*, *Storage cracked*, and *Place of storage*, were connected to an intermediate node *Chance of (re)contamination from water storage* (red node in Fig. 4).

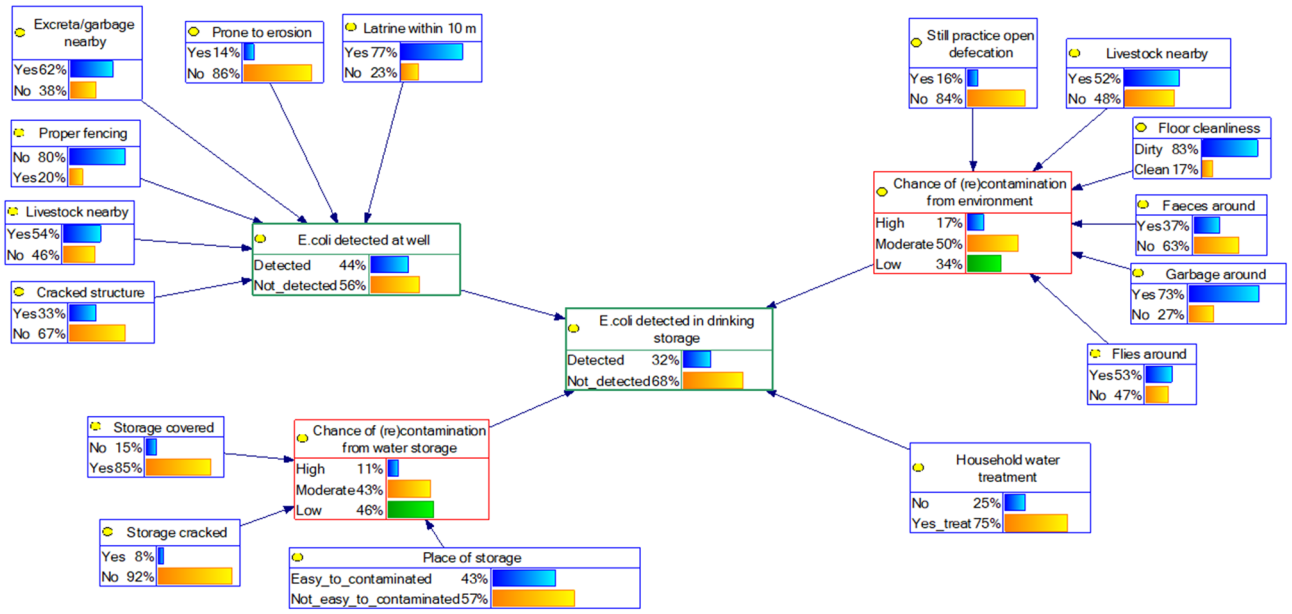


**Figure 6.** BBN model 2A (SI variables at well as parent nodes of "*E. coli* detected at POC").

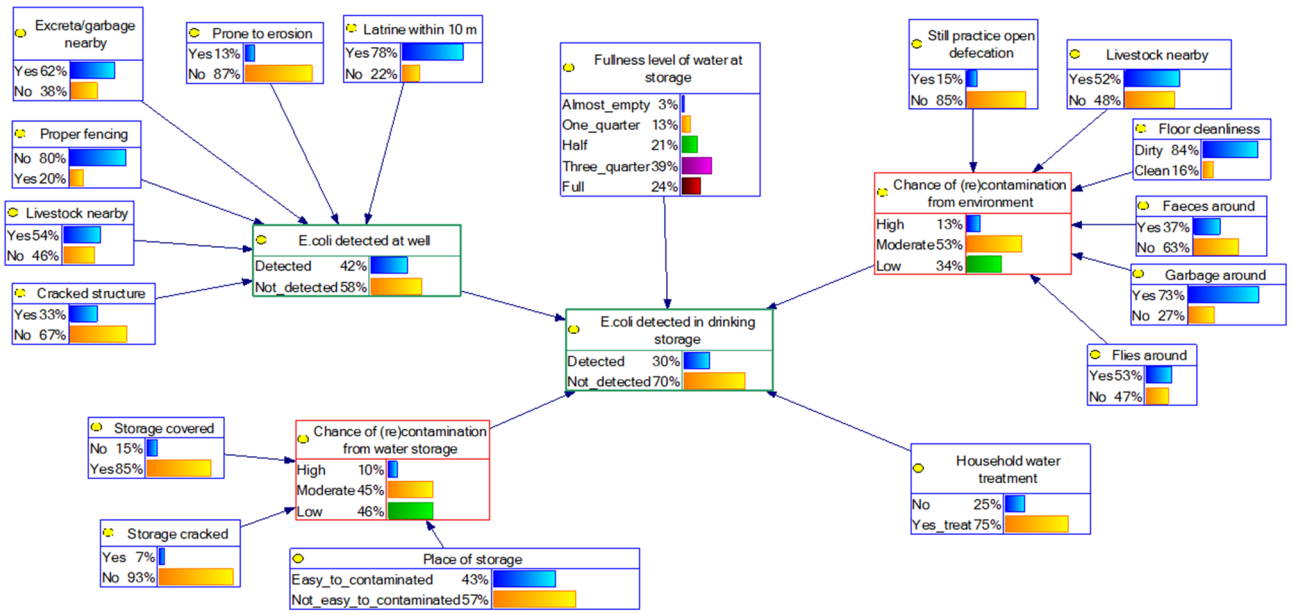


**Figure 7.** BBN model 2B (SI variables at well as parent nodes of "*E. coli* detected at POC" and adding node "fullness of water at storage" as one of the parent nodes of "*E. coli* detected at POC").

Since we did not have the information on intermediate nodes in our datasets, the CPT corresponding to this node was populated manually. First, we gave score 1 to the best situation in each variable, e.g., score 1 if "yes" in variable *storage covered* and score 1 if "no" in variable *storage cracked*. Then we created a simple index by summing all the scores of the three parent nodes. Finally, we categorised it as "low" if the total score was 0–1, "moderate" if the total score was 2, and "high" if the total score was 3. In the same way, another intermediate node *Chance of (re)contamination from environment* was created by six variables (six parent nodes of this variable, see Fig. 4). We categorised it as "low" if the total score was 0–2, "moderate" if the total score was 3–4, and "high" if the total score was 5–6. Different from the other intermediate nodes, we used the results of water quality testing to fill the information of node *E. coli detected at POC* (see Fig. 4; green nodes). BBN requires discrete or categorical information for the analysis. Therefore, we discretised and categorised the number of *E. coli* into *E. coli* detected or non-detected.

We used software GeNIe 2.2 (https://www.bayesfusion.com) to perform the BBN analysis. The software uses the Expectation Maximization (EM) algorithm to estimate the CPT values[36]. We performed validation

tests using the same software to assess the model's performance. We used the ten-fold cross-validation and the performance was reflected by the value of area under the ROC curve (AUC): AUC of 0.5 indicates poor model, AUC between 0.5 and 0.7 is a "less accurate" model, $0.7 < AUC \leq 0.9$ is a "moderately accurate", $0.9 < AUC < 1$ is a "highly accurate" model, and AUC = 1 is a perfect model[37].

We also conducted a "predictive inference" in BBN, to find influential nodes that help us to prioritise actions to improve the water quality of POU in that area. We performed that by setting the state of a specific node to 100% and observe the updated probability in the output node. For example, if we wanted to observe the influence of HWT on POU's water quality, we set the probability of node *Household water treatment* being "yes_treat" to 100% and observed the updated probability of *E. coli detected at POU* being "detected". We did that to all states in all nodes.

Finally, we simulated the "best scenario", i.e., targeting all SI variables or potential source of contaminations in the system, by setting the best situation of all SI variables (outer nodes) at all clusters, including node *Household water treatment* being "yes_treat" and node *E. coli detected at POC* being "not_detected". By setting node *E. coli detected at POC* being "not_detected", we assumed that all types of water source that household use are safe.

## Results

**Socio-demographic characteristics of the respondents.** When asked about the education of the household's head, 12.5% of them had no formal education, and 57.3%, 11.9%, and 18.3% finished primary, secondary, and higher school, respectively. In terms of housing condition, 87.6% did not have permanent walls, e.g., wood or bamboo, 7.5% did not have a permanent roof, i.e., straw, and 71.4% still had a natural floor, i.e., compacted soil. Moreover, 45.3% of the respondents had no electricity. About 32.7% of the respondents practised open defecation. Based on observations, households either had simple pit latrines or pour-flush latrines, some were communal and some were in respective households. Tap water (from a small-scale distribution network) was used by 31.8% of the respondents, followed by wells 27.2%, water trucks 19.6%, and spring water 17.4%, respectively. Remaining respondents used river water, rainwater, or refill potable water stations. Boiling was used to treat the drinking water.

**Description of the sanitary inspection and water quality results.** The general hygiene situation of the respondents is depicted in the BBN model, i.e. the outer nodes in Fig. 4 (in blue colour). For example, 23% of the respondents did not cover their drinking storage and only 30% of the respondent's houses were free from flies. From the cluster of *surrounding environment–hygiene condition*, we found that 66.7% of the respondents kept their livestock near the house, resulting in 60% of the respondents had animal faeces around the house. In addition, 89% and 70% of the respondents had garbage and flies around the water storage or house, respectively. These conditions led to only 15% respondents had low chance of contamination from the surrounding environment and hygiene condition.

The general condition of the cluster *water storage condition* indicated that 37% of the respondents had a low chance of contamination from "bad condition of water storage", i.e., comply to all three criteria: storage with cover, without cracking, and proper-safe place. About 77% and 96% of the storages were found to be covered and without cracking, but 51% of the storages were put in a place that can be prone to (re)contamination, e.g. on the floor.

Of all the POU samples, 56.5% of the respondents claimed to treat water at the time of visit. 75% of households who abstracted water from river treated their drinking water, followed by 68.5% and 59.4% from households who used well and piped system, respectively.

Of all the POU samples, 56.3% of our respondents claimed to treat water at the time of the visit. For the water quality, we did not detect *E. coli* in the 1 ml samples in 195 (75.6%) of the POC samples and 270 (82.3%) of the POU samples. *E. coli* was not detected in almost 90% of the piped and spring samples, while 42% and 83% of well and river samples, respectively, were detected with *E. coli*.

**Comparison of the BBN models' performance.** The four BBN models are shown in Figs. 4, 5, 6 and 7. We first compared the performance of BBN models with SI variables *only* and SI variables plus extra variable *fullness level of water at storage*. The validation tests of these four BBN models gave AUC value: 0.55, 0.69, 0.71, and 0.84 for model 1A (Fig. 4), 1B (Fig. 5), 2A (Fig. 6), and 2B (Fig. 7), respectively. According to the classification of Greiner et al.[37], model 1A and 1B were classified as "less accurate" and model 2A and 2B as "moderately accurate".

The addition of variable *fullness level of water at storage*, which is not part of "standard" SI variables, improved the model's performance. Therefore, we decided to use BBN model 1B (Fig. 5) and 2B (Fig. 7) for further BBN analyses, because model 1 and 2 differ in structure (Fig. 3).

**Predictive inference of the BBN models.** Node *E. coli detected at POC* was the most influential node (see $\Delta P = 21$ in Table 2—left) for the model 1B (type of POC as one of the outer nodes), i.e., the better the water quality at POC, the better the water quality at the household level or POU. Node *Type of POC* and *Fullness level of water at storage* appeared as the second most influential nodes ($\Delta P = 17$ in Table 2—left). The intermediate node *Chance of (re)contamination from the water storage* was the third most influential node ($\Delta P = 10$ in Table 2—left).

The probability of not detected *E. coli* at POU was 75% for households who used both *Piped* and *Spring*, considering other information in the BBN model. The fuller the level of water in the storage, the better the water quality at POU was: the probability of *E. coli* contamination at POU was 58% for *Almost empty* compared to 74% for *Full*. Among all three outer nodes in the cluster *(re)contamination from water storage*, node *storage covered* ($\Delta P = 5$ in Table 2–left) was the most influential node.

| BBN model 1B: with type of POC as one of the outer nodes | | | | | | | BBN mode 2B: with SI at well as one of the outer nodes | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Variable | Probability of *E. coli* not-detected at POU (%) | | | | | ΔP[a] | Variable | Probability of *E. coli* not-detected at POU (%) | | | ΔP |
| **Point of collection** | | | | | | | **Point of collection** | | | | |
| Type of POC | Piped | Well | Spring | River | Other | 17 | Cracked structure | Yes | No | | 2 |
| | 75 | 69 | 75 | 58 | 72 | | | 69 | 71 | | |
| *E. coli* detected at POC | Yes | | No | | | 21 | Livestock nearby | Yes | No | | 1 |
| | 56 | | 77 | | | | | 70 | 71 | | |
| **Household water treatment** | | | | | | | Proper fencing | Yes | No | | 0 |
| Household Water treatment | No | | Yes | | | 6 | | 70 | 70 | | |
| | 69 | | 75 | | | | Excreta/garbage nearby | Yes | No | | 0 |
| **(re)contamination from environment–hygiene condition** | | | | | | | | 70 | 70 | | |
| Still practise open defecation | Yes | | No | | | 2 | Prone to erosion | Yes | No | | 1 |
| | 71 | | 73 | | | | | 71 | 70 | | |
| Livestock nearby | Yes | | No | | | 1 | Latrine within 10 m | Yes | No | | 0 |
| | 72 | | 73 | | | | | 70 | 70 | | |
| Floor cleanliness | Dirty | | Clean | | | 1 | *E. coli* detected at POC | Yes | No | | 19 |
| | 72 | | 71 | | | | | 59 | 78 | | |
| Faeces around | Yes | | No | | | 1 | **Household water treatment** | | | | |
| | 72 | | 73 | | | | Household water treatment | No | Yes | | 13 |
| Garbage around | Yes | | No | | | 0 | | 60 | 73 | | |
| | 72 | | 72 | | | | **(re)contamination from environment–hygiene condition** | | | | |
| Flies around | Yes | | No | | | 0 | Still practise open defecation | Yes | No | | 4 |
| | 72 | | 72 | | | | | 67 | 71 | | |
| Chance of contamination from the environment | High | | Moderate | | Low | 7 | Livestock nearby | Yes | No | | 5 |
| | 68 | | 75 | | 70 | | | 68 | 73 | | |
| **(re)contamination from water storage** | | | | | | | Floor cleanliness | Yes | No | | 2 |
| Storage covered | Yes | | No | | | 5 | | 70 | 72 | | |
| | 74 | | 69 | | | | Faeces around | Yes | No | | 5 |
| Storage cracked | Yes | | No | | | 4 | | 67 | 72 | | |
| | 69 | | 73 | | | | Garbage around | Yes | No | | 1 |
| Place of storage | Easy to contaminated | | Not easy to contaminated | | | 3 | | 70 | 71 | | |
| | 71 | | 74 | | | | Flies around | Yes | No | | 1 |
| Chance of contamination from water storage | High | | Moderate | | Low | 10 | | 70 | 71 | | |
| | 64 | | 74 | | 74 | | Chance of contamination from the environment | High | Moderate | Low | 22 |
| **Fullness level of water at storage** | | | | | | | | 57 | 67 | 79 | |

Continued

| BBN model 1B: with type of POC as one of the outer nodes | | | | | | |
|---|---|---|---|---|---|---|
| Variable | Probability of *E. coli* not-detected at POU (%) | | | | | ΔP[a] |
| Fullness level of water at storage | Almost empty | One quarter | Half | Three quarter | Full | 17 |
| | 58 | 64 | 70 | 75 | 74 | |

| BBN mode 2B: with SI at well as one of the outer nodes | | | | | | |
|---|---|---|---|---|---|---|
| Variable | Probability of *E. coli* not-detected at POU (%) | | | | | ΔP |
| **(re)contamination from water storage** | | | | | | |
| Storage covered | Yes | | No | | | 0 |
| | 70 | | 70 | | | |
| Storage cracked | Yes | | No | | | 0 |
| | 70 | | 70 | | | |
| Place of storage | Easy to contaminated | | Not easy to contaminated | | | 2 |
| | 71 | | 69 | | | |
| Chance of contamination from water storage | High | Moderate | Low | | | 4 |
| | 68 | 72 | 68 | | | |
| **Fullness level of water at storage** | | | | | | |
| Fullness level of water at storage | Almost empty | One quarter | Half | Three quarter | Full | 4 |
| | 70 | 73 | 69 | 70 | 71 | |

**Table 2.** Predictive inference, measuring the effect of changes in the states of each node on the output node of BNN models: *E. coli detected at POU* (drinking water storage). The value under each category corresponding to a node as displayed in the first column is the updated probability of the output node being "Not_detected" given that all households maintain this state. The left side of the table was for the BBN model 1A (Fig. 5) and the right side was for BBN model 2B (Fig. 7). [a]ΔP is the difference between the lowest and highest value of the updated probability of output node: *E. coli detected at POU* being "Not_detected", in %. Examples of how to read the table: (a) row 4–5 *BBN model 1B*: if the type of POC is piped, the Probability of *E. coli* not-detected at POU (%) is 75%; (b) row 6–7 *BBN model 1B*: if *E. coli* is detected at POC ("yes"), the Probability of *E. coli* not-detected at POU (%) is 56%; (c) row 4–5 *BBN model 2B*: if there is a cracked in the structure ("yes"), the Probability of *E. coli* not-detected at POU (%) is 69%.

The households who claimed to do HWT have a higher chance of not to be contaminated by *E. coli* than households who claimed not doing HWT, i.e., $P_{Not\_detected} = 75\%$, $P_{Not\_detected} = 69\%$, respectively.

In model 2B, intermediate node *Chance of (re)contamination from the environment* was the most influential node among households who used a well as their water source (ΔP = 22 in Table 2—right). Node *E. coli detected at POC* was the second most influential nodes (ΔP = 19 in Table 2—right), followed by node *Household water treatment* (ΔP = 13 in Table 2—right). In addition, the influence of node *Fullness level of water at storage* and the intermediate node *Chance of (re)contamination from the water storage* was not large, compared to model 1B (both had ΔP = 4 in Table 2—right).

The effect of HWT to improve the water quality was larger in model 2B (ΔP = 13 in Table 2—right), compared to model 1B (all types of POC; ΔP = 6 in Table 2—left). If we compare the situation of intermediate nodes *Chance of (re)contamination from the environment* and *Chance of (re)contamination from the environment* in model 1B (Fig. 5) and 2B (Fig. 7), the hygiene situation was better in model 2B. The probability of being "high" in both intermediate nodes in model 2B was lower than in model 1B, e.g., 24% in model 1B compared to 13% in model 2B for the intermediate node *Chance of (re)contamination from the environment*.

Furthermore, keeping the house free from livestock ($P_{Not\_detected} = 73\%$) and faeces ($P_{Not\_detected} = 72\%$) seemed important to reduce the probability of fecal contamination at the household storage among households who used a well as their water source. Respondents who practiced open defecation had a larger probability of fecal contamination at the POU than they who did not, i.e., $P_{Not\_detected} = 67\%$, $P_{Not\_detected} = 71\%$, respectively (ΔP = 4). The influence of HWT to reduce the chance of contamination was prominent in model 2B, i.e., $P_{Not\_detected} = 73\%$ for households who treated their drinking water and $P_{Not\_detected} = 69\%$ for not treating water.

The ΔP of intermediate nodes in both model 1B and 2B were bigger than their outer (parent) nodes. For example, in model 2B, the ΔP of 6 outer nodes in the cluster of *surrounding environment–hygiene condition* had less variation (range ΔP = 1–5) compared to the intermediate node *Chance of (re)contamination from the environment* (ΔP = 22), whereas the intermediate nodes were the sum of the values in outer nodes.

For simulating the best scenario, i.e., combination of variables, model 2B was used to simulate all respondents (Fig. 8). The updated probability of outcome node *E. coli detected at POU* being "not_detected" was 91%, compared to the 70% in the baseline situation (Fig. 7). Given the same scenario in model 1B, the updated probability of the outcome node was 92%, compared to the 72% in the baseline (Fig. 5), which suggests the same pattern as model 2B.

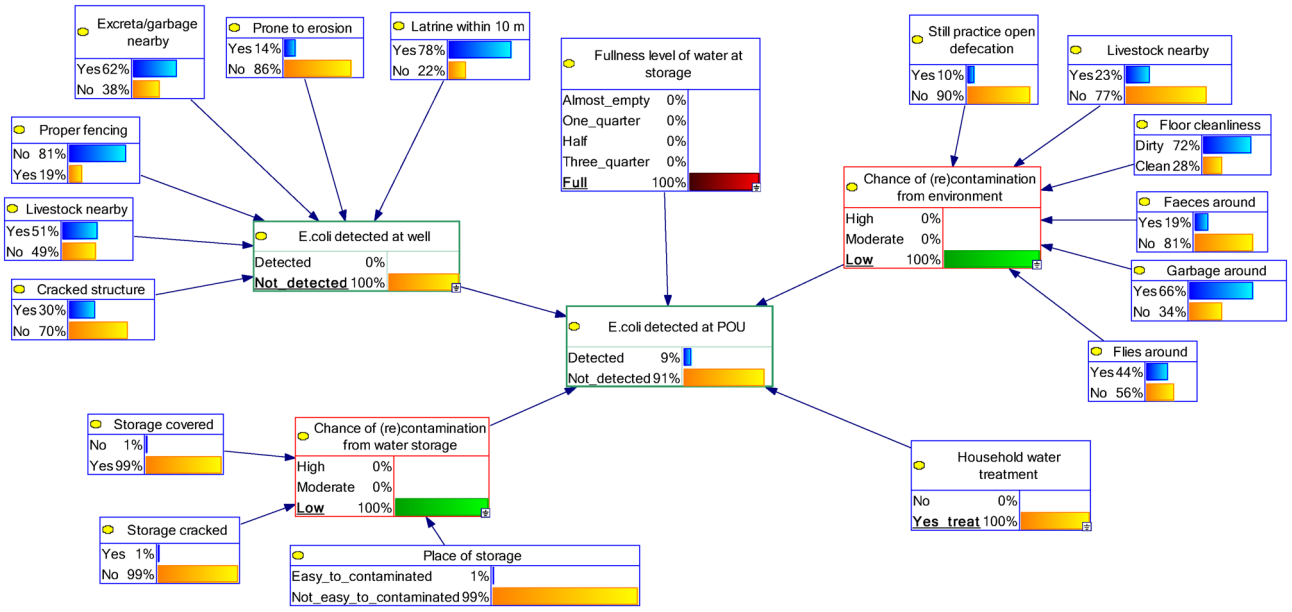


**Figure 8.** The best scenario of water and hygiene management at households level using BBN model 2B (SI at well as one of the outer nodes, SI variables, and *fullness of water at storage*).

## Discussion

**BBN model's performance.** Since there is no BBN study which links SI and water quality data, we compared our models' performance with statistical analysis. Snoad et al.[13] utilized logistic regression to predict the fecal contamination by SI and their AUC values were low (range 0.41–0.64). Other authors also used multiple statistical analyses and found that SI variables could not explain well the water quality[10,14,16], which imply that our models (with AUC values of 0.69 and 0.84) were slightly better in predicting the water quality at POU, using SI data.

However, we found that an "external" factor, besides standard SI variables, increased the model's performance, in our case we used the level of water *fullness inside the storage*, as also found to be relevant in other studies[32–34], suggesting the need to extend the standard SI with external factors for better model performance. In addition, BBN models with SI variables at well (AUC for model 2A and 2B are 0.71 and 0.84, respectively) perform better than BBN models with different types of POC (AUC for model 1A and 1B are 0.55 and 0.69, respectively). Since the same type of POC, e.g., well, can have varying conditions, detailed information of the POC conditions can better explain the water quality than the information on the type of POC itself. This may explain why BBN models with SI variables as explanatory variables perform better than BBN models with types of POCs as explanatory variables.

**Sanitary inspection, water quality, and BBN predictive inferences.** To the authors' knowledge, this is the first study that links SI data with water quality in a medium resource setting. The BBN approach allowed the inclusion of all factors influencing the water quality at POU and grouping them in relevant clusters and pathways, as implied by other conceptual frameworks[29–31]. Furthermore, we were able the analyse the water quality at POU by considering not only the water management and hygiene situation at home, but also the broader scope, such as the situation at the water source. Moreover, the conventional statistical analysis methods, e.g., bivariate correlation or regression analyses, often quantify the effect of the individual variable on water quality, but not a combination of variables or pathways[6,10,16]. The BBN approach was able to simulate both the effects in one model and can then help to prioritise the interventions that improve the water quality at household level, i.e., either targeting one variable or combination of multiple variables.

The BBN approach also enabled the portrayal of interdependencies vividly among variables, while this interdependency have attracted the attention of WASH practitioners and experts over the past years[35]. For example, SI results revealed that there were some hygiene challenges related to livestock ownership. The majority of the respondents (67%) kept livestock in the surroundings of the house, which could be the reason why many flies (70%) and faeces (60%) were detected in our respondents' houses (see Fig. 5 cluster *(re)contamination from environment–hygiene condition*). A study of Ercumen et al.[38] found that the presence of animals is related to fecal contamination, and the presence of animal faeces is associated with diarrhea and stunting[39]. This could be the reason why this area was reported as one of the locations with the highest stunting levels in Indonesia[40]. To tackle these conditions is challenging, since in East Sumba livestock is a symbol of social status[41].

Our BBN models (1B and 2B) showed that the water quality at POCs critically affected the water quality at the POU in the study area, which has also been found by others[6,42]. We also found that types of water source used by the households determine the drinking water quality that they have at home, similar to the findings in rural Honduras[43]. These data suggest that the fecal contamination at POU due to poor water quality at the water

source, especially wells, is a serious problem in East Sumba, i.e., 40% the total populations in East Sumba used well as their main water source[23].

Since we found that the effect of HWT to improve the water quality was larger in model 2B (POC = well only) compared to model 1B (all types of POC), we argue that the effect of HWT to improve the water quality is prominent in the case of better sanitation and hygiene conditions, i.e., the overall condition in model 2B was "more hygienic" than in model 1B. This result has also been suggested by a previous study[44].

Model 1B showed that storage with full water had a better water quality than (almost) empty storage. The explanation could be that the water inside the empty storage was stored for a longer period than a fuller storage, resulting in larger risks for recontamination[4] and permitting bacteria regrowth[45].

Furthermore, we found that the ΔP (the difference between the lowest and highest value of the updated probability of output node: *E. coli detected at POU* being "Not_detected" given the specific condition of a specific node) of intermediate nodes are larger than the influence of their outer (parent) nodes. This implies that collective information of the specific cluster was more meaningful, i.e., more sensitive, to predict the water quality than individual information of specific node or variable. Additionally, it suggests that our simple index, by summing the scores of the parent nodes to populate the CPT in some intermediate nodes, was "acceptable", i.e. simplifying the BBN structure and the intermediate nodes were related to the output node.

A previous WASH study found that a combined HWT, sanitation, handwashing, and house's cleanliness intervention have the same effect as with HWT intervention alone in reducing fecal contamination in household drinking water[46]. In contrast to their study, we found that a combined improvement, targeting all potential contamination sources from the water source until house, had a larger effect in reducing the chance of fecal contamination in the water storage rather than the improvement of one single condition. This suggests that a holistic approach or multi-barrier prevention are needed to minimise drinking water contamination at the POU in rural households[7,47]. However, considering the costs and time constraint, based on the results on impact of water quality at POU, it can be suggested to prioritize the improvement of the water quality at the water source, based on e.g. BNN modelling. Afterwards, WASH behavioural change promotion, e.g., promoting the correct and sustained use of HWT and safe storage container, could be conducted.

Future water quality studies in that area should analyze and include other external factors that may influence the water quality at POC and POU, e.g., type and depth of the well and the types of water containers used by households. This can improve our understanding of water quality in this area.

## Conclusion

This paper introduces an application of BBN to analyse how water quality at the point of use is related to the water quality at the point of collection and associated sanitary inspection data in the medium resource settings in low-middle income countries. The model simulations showed that holistic—combined interventions improved the water quality considerably compared to individual interventions. Moreover, the results demonstrate that water quality at the POC was, as expected, related to the water quality at the POU and (correct and regular) household water treatment had a larger effect of improving the storage water quality in the case of better sanitation and hygiene conditions. We also found that the BBN model performance increased by adding an external variable besides standard SI variables, suggesting that the current SI form should accommodate more (relevant) variables. Additionally, *E. coli* was detected in 24.4 and 17.7% of POC and POU samples, respectively, and there was a hygiene issue related to the ownership and presence of livestock surround the house. Based on the water quality analysis, tap and spring water are relatively cleaner than other types of water sources and, therefore, should be prioritised by the households as main drinking water sources. In order to improve the drinking water quality in this area, reducing the contamination risk at the water source and promoting correct and regular household water treatment are suggested. From the study it can finally be concluded that the BBN approach could be considered as an alternative for conventional statistics to link sanitary inspection and water quality data in low-middle income countries.

Received: 22 May 2020; Accepted: 16 October 2020
Published online: 02 November 2020

## References

1. Prüss-Ustün, A. *et al.* Burden of disease from inadequate water, sanitation and hygiene in low- and middle-income settings: A retrospective analysis of data from 145 countries. *Trop. Med. Int. Heal.* **19**, 894–905 (2014).
2. 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).
3. Podgorski, J. & Berg, M. Global threat of arsenic in groundwater. *Science (80-).* **368**, 845–850 (2020).
4. Levy, K., Nelson, K. L., Hubbard, A. & Eisenberg, J. N. Following the water: A controlled study of drinking water storage in northern coastal Ecuador. *Environ. Heal. Perspect* **116**, 1533–1540 (2008).
5. 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. Heal.* **9**, 106–117 (2004).
6. Daniel, D., Diener, A., van de Vossenberg, J., Bhatta, M. & Marks, S. J. Assessing drinking water quality at the point of collection and within household storage containers in the hilly rural areas of mid and far-western Nepal. *Int. J. Environ. Res. Public Health* **17**, 2172 (2020).
7. WHO. Water Safety Planning for Small Community Water Supplies: Step-by-step Risk Management Guidance for Drinking-Water Supplies in Small Communities (WHO, 2012).
8. WHO. *Surveillance and Control of Community Supplies, Guidelines for Drinking-Water Quality* Vol. 3 (WHO, New York, 1997).
9. Howard, G. *et al.* Identification and management of microbial contaminations in a surface drinking water source. *J. Water Health* **5**, 67–79 (2007).
10. Misati, A. G., Ogendi, G., Peletz, R., Khush, R. & Kumpel, E. Can sanitary surveys replace water quality testing? Evidence from Kisii, Kenya. *Int. J. Environ. Res. Public Health* **14**, 152–164 (2017).

11. Diener, A. *et al.* Adaptable drinking-water laboratory unit for decentralised testing in remote and alpine regions. in 1–7 (40th WEDC International Conference, 2017).
12. Bain, R. *et al.* A summary catalogue of microbial drinking water tests for low and medium resource settings. *Int. J. Environ. Res. Public Health* **9**, 1609–1625 (2012).
13. Snoad, C., Nagel, C., Bhattacharya, A. & Thomas, E. The effectiveness of sanitary inspections as a risk assessment tool for thermotolerant coliform bacteria contamination of rural drinking water: A review of data from West Bengal, India. *Am. J. Trop. Med. Hyg.* **96**, 976–983 (2017).
14. Robinson, D. T. *et al.* Assessing the impact of a risk-based intervention on piped water quality in rural communities: The case of mid-western Nepal. *Int. J. Environ. Res. Public Health* **15**, 1616–1639 (2018).
15. Dey, N. C. *et al.* Microbial contamination of drinking water from risky tubewells situated in different hydrological regions of Bangladesh. *Int. J. Hyg. Environ. Health* **220**, 621–636 (2017).
16. Ercumen, A. *et al.* Can sanitary inspection surveys predict risk of microbiological contamination of groundwater sources? Evidence from shallow tubewells in rural Bangladesh. *Am. J. Trop. Med. Hyg.* **96**, 561–568 (2017).
17. Tang, C., Yi, Y., Yang, Z. & Sun, J. Risk analysis of emergent water pollution accidents based on a Bayesian Network. *J. Environ. Manage.* **165**, 199–205 (2016).
18. Bertone, E., Sahin, O., Richards, R. & Roiko, A. Extreme events, water quality and health: A participatory Bayesian risk assessment tool for managers of reservoirs. *J. Clean. Prod.* **135**, 657–667 (2016).
19. Cain, J. *Planning Improvements in Natural Resources Management* Vol. 44 (UK Centre for Ecology & Hydrology, Wallingford, 2001).
20. Hall, D. C. & Le, Q. B. Use of Bayesian networks in predicting contamination of drinking water with *E. coli* in rural Vietnam. *Trans. R. Soc. Trop. Med. Hyg.* **111**, 270–277 (2017).
21. Daniel, D., Pande, S. & Rietveld, L. The effect of socio-economic characteristics on the use of household water treatment via psychosocial factors: A mediation analysis. *Hydrol. Sci. J.* https://doi.org/10.1080/02626667.2020.1807553 (2020).
22. Sungkar, S. *et al.* Heavy burden of intestinal parasite infections in Kalena Rongo village, a rural area in South West Sumba, eastern part of Indonesia: A cross sectional study. *BMC Public Health* **15**, 1–6 (2015).
23. BPS Statistics of East Sumba Regency. Persentase Rumah Tangga menurut Sumber Air Utama yang Digunakan Untuk Minum di Kabupaten Sumba Timur, 2015–2017. *Statistics of Sumba Timur Regency*. https://sumbatimurkab.bps.go.id/dynamictable/2018/11/12/50/persentase-rumah-tangga-menurut-sumber-air-utama-yang-digunakan-untuk-minum-di-kabupaten-sumba-timur-2015-2017.html (2018).
24. QGIS Development Team. QGIS Geographic Information System ver. 2.18.4. https://download.qgis.org (2017).
25. WHO. *Guidelines for Drinking-Water Quality: Fourth Edition Incorporating The First Addendum* Vol. 1 (World Health Organization, New York, 2017).
26. Nissui Pharmaceutical Co. Ltd. CompactDry "Nissui" EC Illustration Manual. https://www.nissui-pharm.co.jp/english/pdf/products/global/illustration-manual_/CompactDryNissuiEC-IllustrationManual.pdf (n.d.).
27. Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference* (Morgan Kaufmann Publishers Inc., San Mateo, 1988).
28. Nadkarni, S. & Shenoy, P. P. A causal mapping approach to constructing Bayesian networks. *Decis. Support Syst.* **38**, 259–281 (2004).
29. Navab-Daneshmand, T. *et al.* Escherichia coli contamination across multiple environmental compartments (soil, hands, drinking water, and handwashing water) in urban Harare: Correlations and risk factors. *Am. J. Trop. Med. Hyg.* **98**, 803–813 (2018).
30. Cohen, A. *et al.* Microbiological evaluation of household drinking water treatment in rural China shows benefits of electric kettles: A cross-sectional study. *PLoS ONE* **10**, 1–16 (2015).
31. Wagner, E. G., Lanoix, J. N. & World Health Organization. *Excreta Disposal for Rural Areas and Small Communities. Monograph Series* Vol. 39 (World Health Organization, New York, 1958).
32. Boateng, D., Tia-Adjei, M. & Adams, E. A. Determinants of household water quality in the Tamale Metropolis, Ghana. *J. Environ. Earth Sci.* **3**, 70–77 (2013).
33. Elala, D., Labhasetwar, P. & Tyrrel, S. F. Deterioration in water quality from supply chain to household and appropriate storage in the context of intermittent water supplies. *Water Sci. Technol. Water Supply* **11**, 400 (2011).
34. Brick, T. *et al.* Water contamination in urban south India: Household storage practices and their implications for water safety and enteric infections. *Int. J. Hyg. Environ. Health* **207**, 473–480 (2004).
35. Eisenberg, J. N. S., Trostle, J., Sorensen, R. J. D. & Shields, K. F. Toward a systems approach to enteric pathogen transmission: From individual independence to community interdependence. *Annu. Rev. Public Health* **33**, 239–257 (2012).
36. Do, C. B. & Batzoglou, S. What is the expectation maximization algorithm?. *Nat. Biotechnol.* **26**, 897–899 (2008).
37. Greiner, M., Pfeiffer, D. & Smith, R. D. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Prev. Vet. Med.* **45**, 23–41 (2000).
38. Ercumen, A. *et al.* Animal feces contribute to domestic fecal contamination: Evidence from *E. coli* measured in water, hands, food, flies, and soil in Bangladesh. *Environ. Sci. Technol.* **51**, 8725–8734 (2017).
39. Penakalapati, G. *et al.* Exposure to animal feces and human health: A systematic review and proposed research priorities. *Environ. Sci. Technol.* **51**, 11537–11552 (2017).
40. Local Burden of Disease Child Growth Failure Collaborators. Mapping child growth failure across low- and middle-income countries. *Nature* **577**, 231–234 (2020).
41. Bamualim, A. Livestock production and fire management in East Nusa Tenggara. in *Fire and Sustainable Agricultural and Forestry Development in Eastern Indonesia and Northern Australia. Proceedings of an international workshop held at Northern Territory University, Darwin, Australia, 13–15 April 1999*, 69–72 (2000).
42. Cronin, A. A., Breslin, N., Gibson, J. & Pedley, S. Monitoring source and domestic water quality in parallel with sanitary risk identification in Northern Mozambique to prioritise protection interventions. *J. Water Health* **4**, 333–345 (2006).
43. Trevett, A. F., Carter, R. C. & Tyrrel, S. F. Water quality deterioration: A study of household drinking water quality in rural Honduras. *Int. J. Environ. Health Res.* **14**, 273–283 (2004).
44. Esrey, S. A. & Habicht, J. Epidemiologic evidence for health benefits from improved water and sanitation in developing countries. *Epidemiol. Rev.* **8**, 117–128 (1986).
45. Mellor, J. E., Smith, J. A., Samie, A. & Dillingham, R. A. Coliform sources and mechanisms for regrowth in household drinking water in Limpopo, South Africa. *J. Environ. Eng. (United States)* **139**, 1152–1161 (2013).
46. Pickering, A. *et al.* Can individual and integrated water, sanitation, and handwashing interventions reduce fecal contamination in the household environment? Evidence from the WASH Benefits cluster-randomized trial in rural Kenya. https://www.biorxiv.org/content/https://doi.org/10.1101/731992v1.full.pdf (2019) https://doi.org/10.1101/731992.
47. Gundry, S. *et al.* A systematic review of the health outcomes related to household water quality in developing countries. *J. Water Health* **2**, 1–14 (2004).


## Acknowledgements

We thank all respondents in the study, all interviewers, and LKP Anugerah Anak Sumba for the support in data collection. We thank Kirsten van Linden, Ilias Machairas, Dennis Djohan for the hard work during the data collection. We also thank Dr. Doris van Halem, from TU Delft Global Drinking Water, and Armand Middeldorp

to support us with the field water quality test equipment. The first author receives a PhD research funding from Indonesia Endowment Fund for Education (LPDP) and field logistics and from the Delft University of Technology. The second author received a travel fund from TU Delft Global Initiative for the data collection.

## Author contributions

D.D. and W.P.I. contributed to the experimental design. W.P.I. contributed to the sample collection and processing. D.D., W.P.I., S.P., and L.R. contributed to data analysis and validation. S.P. and L.R. supervised the project. D.D. prepared the first draft. All authors reviewed and edited the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to D.D.

**Reprints and permissions information** is available at www.nature.com/reprints.

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