Hi (researcher name),

I think it is important if we clarify the research question further before we decide which method to use. Here is a list of questions that you need to think about:

What is your reason for choosing a longitudinal study? Do you want to examine the effect of time in the year, or seasonality that affect water quality?

What other exposures are you considering, such as geographical location and/or food eaten? Is there a pathogen of interest and will you have data for lab test results of the water? There can be many reasons that cause diarrhea, water is not necessarily the main risk factor. The number of glasses of water drink is not particularly useful here unless there is a clear relation between water drank and diarrhea onset.

Another issue with the longitudinal design is your sampling frequency: you only plan to do it once per-month and only ask for exposure one day before; in this case you miss out all the other days. The data acquired in this way is not going to be representative.

An alternative is to carry out a large cross-sectional study, and do analysis using logistic regression while adjusting for other exposures. Then you can identify whether the exposures are strongly associated with your outcome. If you insist using a longitudinal design, then you need a mixed model for repeated measurements. Logistic regression is the recommended method for your research though.

The sample size depends a lot on your study design, but generally the more the better, a few hundreds will be a good start. You need to re-think your study design, and then we can come up with a more concrete sample calculation.

Best regards,

Chi