For submission to Food and Environmental Virology as an Original Research Paper

Title

Bayesian modeling of enteric virus density in wastewater using left-censored data

Tsuyoshi Kato[1,2], Takayuki Miura[3], Satoshi Okabe[3] and Daisuke Sano[3]*

Each author's affiliation and mailing address:

[1]Department of Computer Science, Graduate School of Engineering, Gunma University, Tenjinmachi 1-5-1, Kiryu, Gunma 376-8515, Japan

[2]Center for Informational Biology, Ochanomizu University, Otsuka 2-1-1, Tokyo 112-8610, Japan

[3]Division of Environmental Engineering, Faculty of Engineering, Hokkaido University, North 13, West 8, Kita-ku, Sapporo, Hokkaido 060-8628, Japan

***Corresponding Author.**

Division of Environmental Engineering, Faculty of Engineering, Hokkaido University, North 13, West 8, Kita-ku, Sapporo, Hokakaido 060-8628, Japan. Telephone/Fax: +81-11-706-7597.

E-mail: dsano@eng.hokudai.ac.jp

Abstract

Stochastic models are used to express pathogen density in environmental samples for performing microbial risk assessment with quantitative uncertainty. However, enteric virus density in water often falls below the quantification limit (non-detect) of the analytical methods employed, and it is always difficult to apply stochastic models to a dataset with a substantially high number of non-detects, i.e., left-censored data. We applied a Bayesian model that is able to model both the detected data (detects) and non-detects to simulated left-censored datasets of enteric virus density in wastewater. One hundred paired datasets were generated for each of the 39 combinations of a sample size and the number of detects, in which three sample sizes (12, 24 and 48) and the number of detects from 1 to 12, 24 and 48 were employed. The simulated observation data were assigned to one of two groups, i.e., detects and non-detects, by setting values on the limit of quantification to obtain the assumed number of detects for creating censored datasets. Then, the Bayesian model was applied to the censored datasets, and the estimated mean and standard deviation were compared to the true values by root mean square deviation. The difference between the true distribution and posterior predictive distribution was evaluated by Kullback-Leibler (KL) divergence, and it was found that the estimation accuracy was strongly affected by the number of detects. It is difficult to describe universal criteria to decide which level of accuracy is enough, but eight or more detects are required to accurately estimate the posterior predictive distributions when the sample size is 12, 24 or 48. The posterior predictive distribution of virus removal efficiency with a wastewater treatment unit process was obtained as the log ratio posterior distributions between the posterior predictive distributions of enteric viruses in untreated wastewater and treated wastewater. The KL divergence between the true distribution and posterior predictive distribution of virus removal efficiency also depends on the number of detects, and eight or more detects in a dataset of treated wastewater are required for its accurate estimation.

Keywords

Bayesian model; enteric virus density; left-censored data; non-detects; predictive distribution; wastewater.

53 **Introduction**

54 Pathogenic microorganisms pose a significant risk of waterborne infectious disease
55 (Shuval 2003; Soller et al. 2010; Dorevitch et al. 2012). Infectious risks of pathogens in water
56 need to be accurately estimated by quantitative microbial risk assessment (QMRA) for proper
57 management of water utilization (Teunis et al. 2010). QMRA comprises four tasks: hazard
58 identification, exposure assessment, dose response assessment and risk characterization (Haas
59 et al. 1999). Exposure assessment entails quantification of pathogens of concern, in which the
60 pathogen concentration in water is required to be expressed with an appropriate probability
61 density function (PDF) (Crainiceanu et al. 2003; Smeets et al. 2007; Smeets et al. 2008;
62 Emelko et al. 2010; Schmidt et al. 2010) because the variability of pathogen density elicited
63 by a variety of factors such as sample inhomogeneity and unstable analytical recovery of
64 pathogens (Morales-Morales et al. 2003) should be included in the risk calculation.

65 Enteric viruses, such as norovirus and sapovirus, constitute a group of important
66 waterborne pathogens (Bosch et al. 2008), and quantitative data on enteric virus occurrence in
67 water and wastewater samples are increasingly available (Rutjes et al. 2006; Haramoto et al.
68 2007; Sano et al. 2011; Perez-Sautu et al. 2012). Virus density in environmental water is often
69 below the quantification limit, but large quantities are sporadically observed. Large variations
70 in virus density may be partially explained by inhomogeneity of enteric virus particles in
71 water bodies, owing mainly to the formation of aggregates or binding to suspended solids (da
72 Silva et al. 2008). Because spatial and temporal variation in virus density in water samples is
73 inevitable, enteric virus density in water often falls below the quantification limit (non-detect)
74 of the analytical methods employed. Consequently, datasets with a substantially high number
75 of non-detects, i.e., left-censored datasets (Helsel 2006), are commonly obtained.

76 Since exposure assessment in QMRA requires the determination of parametric
77 distributions for simulating virus density in a batch of water samples (Petterson et al. 2007), it

78 is critical to employ a proper stochastic model for adequately describing enteric virus density
79 in environmental water based on left-censored data. Statistic models for left-censored data
80 have been developed in various fields including the residue analysis of pesticides in food
81 (Kennedy and Hart, 2009; Kennedy, 2010; EFSA, 2010). The main discussion point in the
82 modeling of left-censored data is how to deal with the non-detects. Substitution of the non-
83 detect data with specific values such as the limit of detection and zero has been a classical
84 approach for dealing with non-detects; however, it has been proposed that the substitution
85 approach should not be employed because it ruins the results of the prediction (Helsel, 2006).
86 Alternatively, the Bayesian approach adapted for left-censored data has been applied to the
87 modeling of residue concentrations in food, in which the log-normal distribution is employed
88 to describe the positive concentrations (Paulo et al., 2005).
89 In this study, the Bayesian approach adapted for left-censored data (Paulo et al.,
90 2005) was applied to model the enteric virus density in wastewater samples with a slight
91 modification, in which the occurrence of the real zero of virus density is not assumed. Virus
92 density is assumed to follow a truncated lognormal distribution. The lognormal distribution is
93 one of the probabilistic distributions previously modeled for enteric virus density in water
94 (Tanaka et al. 1998). One hundred paired datasets were generated for each of the 39
95 combinations of a sample size and the number of detects, in which three sample sizes (12, 24
96 and 48) and 14 values of detects (from 1 to 12, 24 and 48) were employed. The simulated
97 observation data were assigned to one of two groups, i.e., detects and non-detects, by setting
98 values on the limit of quantification to obtain one of the numbers of detects. Then, the
99 Bayesian model was applied to the censored data. The estimated mean and standard deviation
100 were compared with true values by calculating root mean square deviation (RMSD), and the
101 influence of the sample size and positive rate value on the accuracy of the posterior
102 distribution estimation is discussed. Furthermore, a log ratio posterior distribution was

obtained by dividing one posterior predictive distribution by the other to express the fold change between two samples, which can be used for expressing the virus removal efficiency when the posterior predictive distributions of enteric virus density in untreated and treated wastewater were used. The accuracy of the distribution estimation of the fold change was evaluated by Kullback-Leibler (KL) divergence.

Materials and Methods

Generation of left-censored data and model definition

A pair of datasets from untreated wastewater and treated wastewater, X_{pre} and X_{post}, respectively, were generated artificially with the model parameters, $(\mu_*, \beta_*) = (1,1)$ and $(\mu_*, \beta_*) = (4,1)$. The simulated observation data were assigned to one of two groups, i.e., detects and non-detects, by setting values of the limit of quantification θ_v to obtain the assumed number of detects in the treated wastewater samples for creating censored data. One hundred of the paired datasets were generated for each of the 39 combinations of a sample size (12, 24 and 48) and the number of detects (1 to 12, 24 and 48).

This study employs the model presented by Paulo et al. (2005), assuming the concentration data are distributed according to a lognormal distribution. In the model, detected observations follow the truncated lognormal distribution of which the probabilistic density function is expressed as

$$\mathrm{TLN}(c; \mu, \beta^{-1}, \theta_v) = \frac{\sqrt{\beta}}{\left(1-\phi\left(\sqrt{\beta}(\theta_v-\mu)\right)\right)\sqrt{2\pi}\ln(10)c} \exp\left(-\frac{\beta}{2}(\log_{10}(c)-\mu)^2\right). \quad (1)$$

Paulo et al. (2005) use the natural logarithm, but this study employs the common logarithm because the use of the common logarithm makes discussion about fold change easier, which is addressed in the section of log ratio posterior afterward. It is readily seen that the probability of failing to detect the data drawn according to this truncated lognormal distribution is

$\emptyset\left(\sqrt{\beta}(\theta_v-\mu)\right)$, leading to the fact that the probability of n_0 non-detect data being included in n samples is given by $\mathrm{Bin}\left(n_0; n, \emptyset\left(\sqrt{\beta}(\theta_v-\mu)\right)\right)$, where $\mathrm{Bin}(\cdot;\cdot,\cdot)$ is the probabilistic mass function of the binomial distribution defined as

$$\mathrm{Bin}(n_0; n, \rho) = \binom{n}{n_0}\rho^{n_0}(1-\rho)^{n-n_0}. \qquad (2)$$

Thus, the probabilistic model used in this study includes two unknown model parameters, μ and β.

We are now ready to express the likelihood function of the model parameters. Let us denote the detect data by $c_1, c_2, \cdots, c_{n_v}$ where $n_v = n - n_0$. The likelihood function for a given dataset X including n_v detect data, $c_1, c_2, \cdots, c_{n_v}$, and n_0 non-detect data gathered with a limit value of quantification θ_v can be written as

$$p(X|\mu,\beta) = \mathrm{Bin}\left(n_0; n, \emptyset\left(\sqrt{\beta}(\theta_v-\mu)\right)\right)\prod_{i=1}^{n_v}\mathrm{TLN}(c_i;\mu,\beta^{-1},\theta_v). \qquad (3)$$

Bayesian inference algorithm

In this study, we adopt the Bayesian analysis to infer the model parameters. Bayesian analysis offers inference results in the form of probabilistic distributions, which differs from point estimation. Inferred probabilistic distributions provide information about how much a certain value can be believed to be estimated well. For this reason, recent studies of water engineering have supported the use of Bayesian analysis (e.g., Petterson et al. 2010). A prior distribution of model parameters is necessary for Bayesian analysis to infer the model parameters in the form of the posterior distribution. Following Paulo et al. (2005), we employ the same prior distribution $\mu\sim N(0,100)$ and $\beta\sim\mathrm{Gam}(0.01,0.01)$ where $N(m,v)$ denotes the normal distribution with mean m and variance v, and $\mathrm{Gam}(a,b)$ denotes the Gamma distribution with shape parameter a and rate parameter b. The statistical independence between the two model parameters is assumed in the prior distribution.

In this model, the posterior distribution of the two model parameters, $p(\mu, \beta|X)$, cannot be represented with an explicit form. In Bayes' theorem, to compute the posterior distribution, the property that the posterior density function is proportional to the product of the likelihood function and the prior density function is used. Marginal posterior of each model parameter, $p(\mu|X)$ and $p(\beta|X)$, is occasionally more handy to see the inferred value of the model parameter. For example, from the marginal posteriors, we can compute the mean and the standard deviation (SD) of the model parameters. The posterior mean and the posterior SD of the model parameter μ are given by

$$\bar{\mu} = \int \mu \; p(\mu|X) \, d\mu \qquad (4)$$

and

$$s_\mu = \sqrt{\int (\mu - \bar{\mu})^2 \; p(\mu|X) \, d\mu} \; , \qquad (5)$$

respectively. The statistics of β can be defined similarly, but we compute the posterior mean and the posterior SD of

$$\log_{10} \sigma = \log_{10}\left(1/\sqrt{\beta}\right), \qquad (6)$$

instead of β itself. When one needs a single estimated value of model parameters, the posterior mean can be used as the expected value of the inferred model parameters. The posterior SD indicates a confidence level of inference; a smaller SD corresponds to higher confidence and vice versa.

When the true values of μ and β are known, the root mean square deviation (RMSD) can be computed from the marginal posterior distribution as

$$\mathrm{RMSD}_\mu = \sqrt{\int (\mu - \mu_*)^2 \; p(\mu|X) \, d\mu} \qquad (7)$$

and

$$\mathrm{RMSD}_\sigma = \frac{1}{2}\sqrt{\int (\log_{10} (\beta/\beta_*))^2 \; p(\beta|X) \, d\beta}, \qquad (8)$$

where μ_* and β_* are the true values of μ and β, respectively. The criteria, RMSDs, evaluate

estimation errors more directly than posterior mean and posterior SD.

Posterior predictive distribution

Posterior predictive distribution is the distribution of the future observations given the dataset. This distribution accounts for the remaining uncertainty in the model parameters. The posterior predictive distribution in our analysis is the distribution of the common logarithm of pathogen concentration data, say c_{log}, based on the model parameter inferred from a given dataset X, and its probabilistic densities are given by

$$p_{pred}(c_{log}|X) = \int p(\mu, \beta|X)p(c_{log}|\mu, \beta)d\mu d\beta \qquad (9).$$

It is ideal when the posterior predictive distribution is close to the underlying distribution generating pathogen concentration data. Fortunately, because artificially generated datasets are used in this study as described above, we know the true distribution as $N(\mu_*, 1/\beta_*)$, and we can thereby compute some criteria to evaluate how accurate the posterior predictive distributions are. We employ the KL divergence as a criterion, expressed as

$$KL = \int N(c_{log}; \mu_*, 1/\beta_*)\ln N(c_{log}; \mu_*, 1/\beta_*)/p(c_{log}|X)\ dc_{log} \qquad (10)$$

More accurate posterior predictive distributions have a smaller KL divergence from the true distribution.

Log ratio posterior

The removal efficiency of enteric viruses in a unit process of wastewater treatment can be estimated by the ratio of the concentration in untreated wastewater to that of treated wastewater. Our Bayesian analysis offers the inference results in the form of distribution of the common logarithm of the ratio. We refer to this distribution as the log ratio posterior. If we denote two datasets from untreated wastewater and from treated wastewater, respectively, by X_{pre} and X_{post}, the log ratio posterior is given by the integration of the product of the two

posterior predictive distributions, written as

$$p(y|X_{pre}, X_{post}) = \int p_{pred}(c + y|X_{pre})\, p_{pred}(c|X_{post})\, dc \qquad (11)$$

where y is a log ratio. As described above, two datasets from untreated wastewater and treated wastewater, X_{pre} and X_{post}, respectively, were generated artificially from log10 normals with the model parameters, $(\mu_*, \beta_*) = (1,1)$ and $(\mu_*, \beta_*) = (4,1)$. From them, the true distribution of the log ratio is derived as $N(3, 2)$. We use the KL divergence again to compare the log ratio posterior with the true distribution. The KL divergence describes the inference ability of this model; the divergence comes close to zero as the inferred distribution approaches the true distribution.

Numerical Issues

In Bayesian inference, typical statistics are the expected values of something represented in an integral form, as described in Eqs (4), (5), (7), (8), (9), (10), (11). Monte Carlo simulation methods could be employed; however, because of the low-dimensional parameter space a quadrature method was judged to be more efficient. We employed the mixture of uniform distributions to interpolate a two-dimensional distribution, where the uniform distributions are placed in a grid.

Software

Software implementing the algorithm developed for inferring posterior distributions and virus removal efficiency is available upon request to the corresponding author.

Results

Posterior predictive distribution and accuracy of parameter estimation

Hundred datasets, $X_{1,post}$, $X_{2,post}$, …$X_{100,post}$, simulating datasets of enteric virus density

in treated wastewater, were prepared for each case (n, n_v), where each of the datasets includes n_v detects and n_0 $(= n - n_v)$ non-detects. The red lines in panels (a) and (d) in Figs. 1S to 39S show the parental distributions for creating datasets of treated and untreated wastewater, and the red and blue dots on the horizontal axis are detects and non-detects, respectively. Since Fig. 1S shows the calculation results with (n, n_v) = (12, 1), one red dot and eleven blue dots are observed in panel (a). A posterior predictive distribution was inferred from each dataset, and those from the first five datasets of treated and untreated wastewater are shown in black lines in panels (a) and (d) in Figs. 1S to 39S. Posterior distributions of μ and β from the first five datasets of treated wastewater are shown in panels (b) and (c), respectively, and those of untreated wastewater are in panels (e) and (f) in Figs. 1S to 39S. The relationships between μ and β are indicated in panels (g) and (h) in Figs. 1S to 39S. From 100 independent values of μ and $\log_{10}\sigma$ $(\sigma = \frac{1}{\beta})$ obtained at each dataset, posterior mean and posterior standard deviation (SD) of those 100 values of μ of treated wastewater were computed, and the quartiles are shown in panels (a) and (c) in Figs. 1 to 3, and those of $\log_{10}\sigma$ are in panels (b) and (d) in Figs. 1 to 3. The identical parameters for untreated wastewater are not shown, because the number of detects in untreated wastewater is always large in this study.

The sample size is 12 when enteric virus density in a wastewater sample taken from a sampling site is surveyed once a month for one year. As shown in Figs. 1(a) and 1(b), the median values of the posterior mean of μ and $\log_{10}\sigma$ asymptotically approach the true values, i.e., μ = 1 and $\log_{10}\sigma = 0$. The estimation accuracy is also improved by increasing the number of detects, as shown by the posterior SD of μ and $\log_{10}\sigma$ (Figs. 1(c) and 1(d)). RMSDs of μ and $\log_{10}\sigma$, which are theoretically equal to the averages of square deviations of samples from the true values, where the samples are infinitely generated according to the posteriors, are shown in Figs 1(e) and 1(f), implying better estimation with the larger number

of detect data. It is difficult to determine the number of detects required for estimating the posterior predictive distribution at a significant accuracy, because it is exactly like a riddle with no answer to determine the criteria for how accurate is accurate in the parameter estimation. However, it may be possible to estimate the posterior predictive distribution with an accuracy level comparable to the dataset with a 100% positive rate (i.e., (n, n_v) = (12, 12)). In that sense, it seems that eight detects may be at least required, because RMSDs values of μ and $\log_{10}\sigma$ at (n, n_v) = (12, 8) are not very different from those at (n, n_v) = (12, 12).

We also calculated KL divergence between the true distribution and posterior predictive distribution (Fig. 1(g)). Preferably, the posterior predictive distributions (black lines in panel (a) in Figs. 1S to 39S) should be close to the true distribution N(1,1) (red lines in panel (a) in Figs. 1S to 39S). We computed KL divergence for the posterior predictive distribution of each dataset, in order to investigate how much the inferred distribution diverges from the true distribution. Then, we obtained 100 KL divergences for each case of (n, n_v), and some statistics such as the first, second and third quartiles were indicated in panel (g) in Fig. 1. These statistics suggest that a larger number of detects is favorable for the accurate estimation of parameters. It is likely that more than 8 detects out of 12 samples are required to achieve the similar level of accuracy with 12 detects out of 12 samples, judging by the level of KL divergence indicated in Fig. 1(g).

The sample size turns to be 24 when water samples are taken twice a month (every two weeks on average) for one year. As in the case with the sample size of 12 (Fig. 1), median values of a posterior mean of μ and $\log_{10}\sigma$ asymptotically approach the true values (μ = 1 and $\log_{10}\sigma = 0$) (Fig. 2(a) and 2(b)). The improved accuracy of parameter estimation is also shown by the posterior SD and RMDS of μ and $\log_{10}\sigma$ (Figs. 2(c) to 2(f)). KL divergences between the true distribution and the posterior predictive distribution in Fig. 2(g) indicate that improvement of accuracy is substantial when the number of detects is increased from 7 to 8. It

is impossible to determine which level of KL divergence is adequate as discussed above, but the detected sample number of 8 out of 24 samples must be required to give a relatively more accurate estimation of the posterior predictive distribution.

When the sample size is 48 (e.g., water samples are taken four times per month for one year), the accuracy of parameter estimation is also improved when the number of detects is increased (Fig. 3(a) to (f)). KL divergences between the true distribution and the posterior predictive distribution in Fig. 3(g) indicate that improvement of accuracy is also substantial when the number of detects is increased from 7 to 8 as well when the sample size is 24 (Fig. 2(g)). The detected sample number of 8 or larger must be required for accurately estimating the posterior predictive distribution when the sample size is 48.

Log ratio posteriors

One of the useful applications of the posterior predictive distribution of enteric virus density in wastewater is to estimate the removal efficiency of viruses in water and wastewater treatment processes. The removal efficiency is obtained as the log ratio posterior distributions between the posterior predictive distributions of enteric viruses in untreated and treated wastewater, which are given by eq. 11. Box plots in Fig. 4 show the quartiles of KL divergence values between the log ratio posteriors and the true distributions, where panels (a), (b) and (c) are those obtained when the sample size is 12, 24, and 48, respectively. It is clear that a greater number of detects always gives a more accurate estimation, but the important point is how few detects are allowable for the estimation of the log ratio posterior distribution. It was very difficult to obtain an accurate prediction of the log ratio when the number of detects was less than 7, but it appears that a relatively good estimation is achieved when the number of detects was more than 8 when the sample size is 12, 24 or 48. These results imply that prediction accuracy depends on the number of detects rather than the positive rate, and at

300 least 8 detected samples are required when the sample size is between 12 and 48.
301
302 **Discussion**
303 Analytical uncertainty in quantitative results, considering the forms of analytical
304 variance and spatiotemporal variations in pathogen occurrence in water, must be implicated in
305 exposure assessment in QMRA (Petterson et al. 2007; Teunis et al. 2010). Stochastic models
306 have been proposed for describing pathogen density in water (Crainiceanu et al. 2003;
307 Emelko et al. 2010). However, virus density is often expected to fall below the quantification
308 limit as already discussed, particularly in treated wastewater samples, which makes it
309 extremely difficult to apply some stochastic models to describe the virus density. The statistic
310 model employed in this study is the one developed for describing the residues of pesticides in
311 food (Paulo et al. 2005). In this model, it is assumed that there is an unknown proportion of
312 batches with zero residues, and the number of non-detects and the distribution of positive
313 residues are separately modeled depending on the calibration parameters (Kennedy and Hart,
314 2009). The sole modification in this study is not to assume the real zero of enteric virus
315 density in wastewater. The results of parameter estimation indicated in this study showed that
316 it was possible to apply the modified Bayesian model to left-censored data, when the number
317 of detects was 8 or greater out of a sample size of 12, 24 and 48. Again, it is exactly like a
318 riddle with no answer to determine criteria for how accurate is accurate in the parameter
319 estimation. However, it is necessary to repeat the investigation until the positive sample
320 number is accumulated to be 8 or greater to obtain the precise posterior predictive distribution.
321 The virus removal efficiency of water and wastewater treatment processes is
322 indispensable information for the management of microbiologically safe drinking water
323 (Schijven et al. 2011), particularly when performance targets are employed as health-based
324 targets in water safety plans (World Health Organization 2011). However, data acquisition of

virus density before and after treatment is difficult because of antecedent reasons, including the low density of pathogens in treated wastewater. In some cases, it is easily expected that there may be difficulties in accumulating significant amounts of detected samples of enteric viruses in treated wastewater, particularly when membrane-based technologies such as membrane bioreactors are applied to the wastewater treatment, or strong disinfection processes such as ozonation are employed. The statistical treatment of such datasets with a detected sample number less than 8 must be a critical issue in future studies.

The ultimate goal for water and public health practitioners is to manage the usage of water and treated wastewater in a coherent manner and QMRA gives a reasonable solution (World Health Organization 2011). The proposed stochastic modeling is applicable to any enteric viruses, including human noroviruses, when a dataset of viral genome density in wastewater is available. Virus density simulation based on the predictive distribution in treated wastewater can be performed to reproduce data, implying that the posterior predictive distributions of enteric virus density in water inferred by the proposed approach would be highly compatible with QMRA and that virus occurrence probability could be estimated based on monitoring data. It is noteworthy that the lognormal distribution might not always be the true distribution, in general. Therefore, the results presented here, based on simulated data generated from a lognormal, will be an idealized situation. In reality the shape of the distribution might lead to different performance of the method. Future studies should investigate the effect of the shape of distribution, by using the other distributions such as gamma distribution, on the required number of positive samples for obtaining appropriate accuracy in the estimation.

Conclusions

A Bayesian model was employed for estimating posterior predictive distributions of

enteric virus density in wastewater samples using left-censored data. The accuracy of the parameter estimation was significantly dependent on the number of detects, rather than the positive rate, and it is recommended that at least 8 detects be accumulated to precisely estimate the posterior predictive distribution.

Acknowledgments

This work was supported by the Japan Science and Technology Agency through a Grant-in-Aid for Core Research for Evolutionary Science and Technology (CREST), and the Japan Society for the Promotion of Science through Grant-in-Aid for Young Scientist (A) (22686049) and Scientific Research (A) (24246089).

References

Bosch, A., Guix, S., Sano, D., & Pinto, R. M., (2008). New tools for the study and direct surveillance of viral pathogens in water. *Current Opinion in Biotechnology*, 19(3), 295-301.

Crainiceanu, C. M., Stedinger, J. R., Ruppert, D., & Behr, C. T. (2003). Modeling the U.S. national distribution of waterborne pathogen concentrations with application to Cryptosporidium parvum. *Water Resources Research*, 39(9), 1235-1249.

Dorevitch, S., Pratap, P., Wroblewski, M., Hryhorczuk, D. O., Li, H., Liu, L. C., & Scheff, P. A. (2012). Health risks of limited-contact water recreation. *Environmental Health Perspectives*, 120(2), 192-197.

Rubinstein, R. Y., & Kroese, D. P. (2004). The Cross-Entropy Method, Springer.

Emelko, M. B., Schmidt, P. J., & Reilly, P. M. (2010). Particle and microorganism enumeration data: enabling quantitative rigor and judicious interpretation. *Environmental Science and Technology*, 44(5), 1720-1727.

375 Haas, C. N., Rose, J. B., & Gerba, C. P. (1999). *Quantitative Microbial Risk Assessment.* New
376 York: Wiley.
377 Haramoto, E., Katayama, H., Oguma, K., & Ohgaki, S. (2007). Quantitative analysis of
378 human enteric adenoviruses in aquatic environments. *Journal of Applied*
379 *Microbiology*, 103(6), 2153-2159.
380 Helsel, D. R. (2006) Fabricating data: How substituting values for nondetects can ruin results,
381 and what can be done about it. *Chemosphere*, 65, 2434-2439.
382 Kennedy, M. C. (2010). Bayesian modeling of long-term dietary intakes from multiple
383 sources. *Food and Chemical Toxicology*, 48, 250-263.
384 Kennedy, M., & Hart, A. (2009) Bayesian modeling of measurement errors and pesticide
385 concentration in dietary risk assessments. *Risk Analysis*, 29(10), 1427-1442.
386 Morales-Morales, H. A., Vidal, G., Olszewski, J., Rock, C. M., Dasgupta, D., Oshima, K. H.,
387 & Smith, G. B. (2003). Optimization of a reusable hollow-fiber ultrafilter for
388 simultaneous concentration of enteric bacteria, protozoa, and viruses from water.
389 *Applied and Environmental Microbiology*, 69(7), 4098-4102.
390 Paulo, M. J., van der Voet, H., Jansen, M. J. W., ter Braak, C. J. F., & van Klaveren J. D.
391 (2005) Risk assessment of dietary exposure to pesticides using a Bayesian method.
392 *Pest Management Science*, 61, 759-766.
393 Perez-Sautu, U., Sano, D., Guix, S., Georg, K., Pinto, R. M., & Bosch, A. (2012). Human
394 norovirus occurence and diversity in the Llobregat river catchment, Spain.
395 *Environmental Microbiology*, 14(2), 494-502.
396 Petterson, S. R., Signor, R. S., & Ashbolt, N. J. (2007). Incorporating method recovery
397 uncertainties in stochastic estimates of raw water protozoan concentrations for
398 QMRA. *Journal of Water and Health*, 5(S1), 51-65.
399 Rutjes, S. A., van den Berg, H. H. J. L., Lodder, W. J., & de Roda Husman, A. M. (2006).

Real-time detection of noroviruses in surface water by use of a broadly reactive nucleic acid sequence-based amplification assay. *Applied and Environmental Microbiology*, 72(8), 5349-5358.

Sano, D., Perez, U., Guix, S., Pinto, R. M., Miura, T., Okabe, S., & Bosch, A. (2011). Quantification and genotyping of human sapoviruses in the Llobregat River catchment, Spain. *Applied and Environmental Microbiology*, 77(3), 1111-1114.

Schijven, J. F., Teunis, P. F. M., Rutjes, S. A., Bouwknegt, M., & de Roda Husman, A. M. (2011). QMRAspot: A tool for quantitative microbioal risk assessment from surface water to potable water. *Water Research*, 45(17), 5564-6676.

Schmidt, P. J., Emelko, M. B., & Reilly, P. M. (2010). Quantificaiton of analytical recovery in particle and microorganism enumeration methods. *Environmental Science and Technology*, 44(5), 1705-1712.

Shuval, H. (2003). Estimating the global burden of thalassogenic diseases: Human infectious diseases caused by wastewater pollution of the marine environment. *Journal of Water and Health*, 1(2), 53-64.

da Silva, A. K., Le Guyader, F. S., Le Saux, J.-C., Pommepuy, M., Montgomery, M. A., & Elimelech, M. (2008). Norovirus removal and particle association in a waste stabilization pond. *Environmental Science and Technology*, 42(24), 9151-9157.

Smeets, P. W. M. H., van Dijk, J. C., Stanfield, G., Rietveld, L. C., & Medema, G. J. (2007). How can the UK statutory *Cryptosporidium* monitoring be used for quantitative risk assessment of *Cryptosporidium* in drinking water? *Journal of Water and Health*, 5(S1), 107-118.

Smeets, P. W. M. H., Dullemont, Y. J., van Gelder, P. H. A. J. M., van Dijk, J. C., & Medema, G. J. (2008). Improved methods for modelling drinking water treatment in quantitative microbioal risk assessment; A case study of *Campylobacter* reduction

425 by filtration and ozonation. *Journal of Water and Health*, 6(3), 301-314.
426 Soller, J. A., Schoen, M. E., Bartrand, T., Ravenscroft, J. E., & Ashbolt, N. J. (2010).
427 Estimated human health risks from exposure to recreational waters impacted by
428 human and non-human sources of faecal contamination. *Water Research,* 44(16),
429 4674-4691.
430 Tanaka, H., Asano, T., Schroeder, E. D., & Tchobanoglous, G. (1998). Estimating the safety of
431 wastewater reclamation and reuse enteric virus monitoring data. *Water Environment*
432 *Research*, 70(1), 39-51.
433 Teunis, P. F. M., Xu, M., Freming K. K., Yang, J., Moe, C. L., & Lechevallier, M. W. (2010).
434 Enteric virus infection risk from intrusion of sewage into a drinking water
435 distribution network. *Environmental Science and Technology*, 44(22), 8561-8566.
436 World Health Organization (2011). *Guidelines for drinking-water quality*. Geneva,
437 Switzerland.
438
439

List of Figures

Figure 1. Accuracies of Bayesian inference for treated wastewater datasets including n = 12 samples. We considered 12 cases: each case has a different number of detected samples. For each case, 100 datasets were generated. Posterior means, posterior standard deviations and root-mean-square-deviation of μ and $\log_{10}\sigma$, respectively, were obtained from the posterior distribution of model parameters for each dataset, where $\sigma = 1/\beta$. The quartiles of the 100 values for each statistic are depicted with box plots in (a) to (f). The posterior predictive distribution, which is inferred from a single dataset, is evaluated with the KL divergence from the true distribution. The subplot (g) depicts the quartiles of these 100 KL divergences.

Figure 2. Accuracies of Bayesian inference for treated wastewater datasets including n = 24 samples. We considered 13 cases: each case has a different number of detected samples. For each case, 100 datasets were generated. Posterior means, posterior standard deviations and root-mean-square-deviation of μ and $\log_{10}\sigma$, respectively, were obtained from the posterior distribution of model parameters for each dataset, where $\sigma = 1/\beta$. The quartiles of the 100 values for each statistic are depicted with box plots in (a) to (f). The posterior predictive distribution, which is inferred from a single dataset, is evaluated with the KL divergence from the true distribution. The subplot (g) depicts the quartiles of these 100 KL divergences.

Figure 3. Accuracies of Bayesian inference for treated wastewater datasets including n = 48 samples. We considered 14 cases: each case has a different number of detected samples. For each case, 100 datasets were generated. Posterior means, posterior standard deviations and root-mean-square-deviation of μ and $\log_{10}\sigma$, respectively, were obtained from the posterior distribution of model parameters for each dataset, where $\sigma = 1/\beta$. The quartiles of the 100

values for each statistic are depicted with box plots in (a) to (f). The posterior predictive distribution, which is inferred from a single dataset, is evaluated with the KL divergence from the true distribution. The subplot (g) depicts the quartiles of these 100 KL divergences.

Figure 4. Divergences between the log ratio posteriors and the true distributions. We had 100 couples of two datasets of untreated wastewater and treated wastewater: ($X_{1,pre}$, $X_{1,post}$), …., ($X_{100,pre}$, $X_{100,post}$)). A log ratio posterior was evaluated with the KL divergence from the true distribution. From the resultant 100 KL divergences, the quartiles are depicted in box plots.

Figure 1

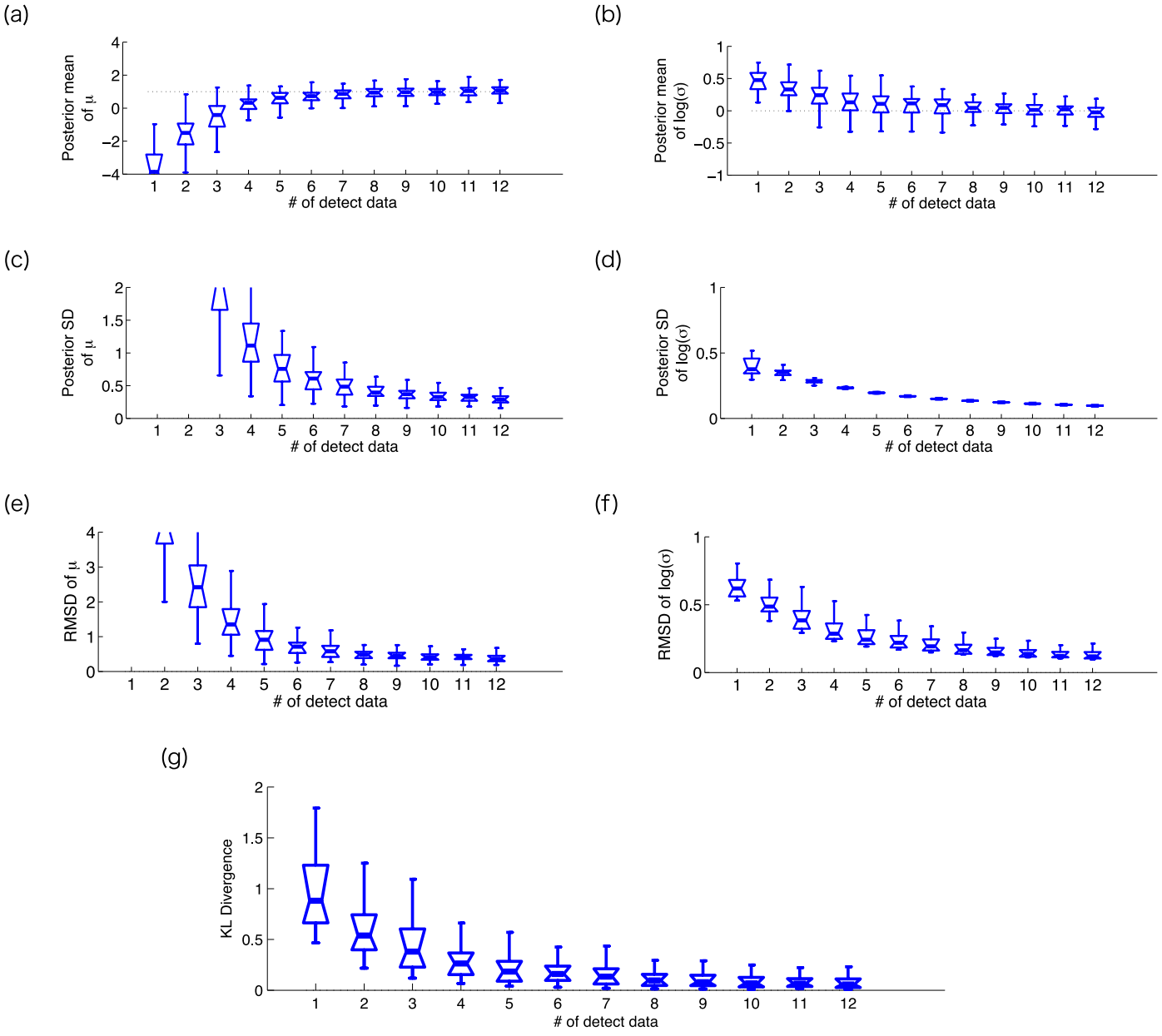

(a)
Posterior mean of μ
# of detect data
(b)
Posterior mean of log(σ)
# of detect data
(c)
Posterior SD of μ
# of detect data
(d)
Posterior SD of log(σ)
# of detect data
(e)
RMSD of μ
# of detect data
(f)
RMSD of log(σ)
# of detect data
(g)
KL Divergence
# of detect data

Figure 2

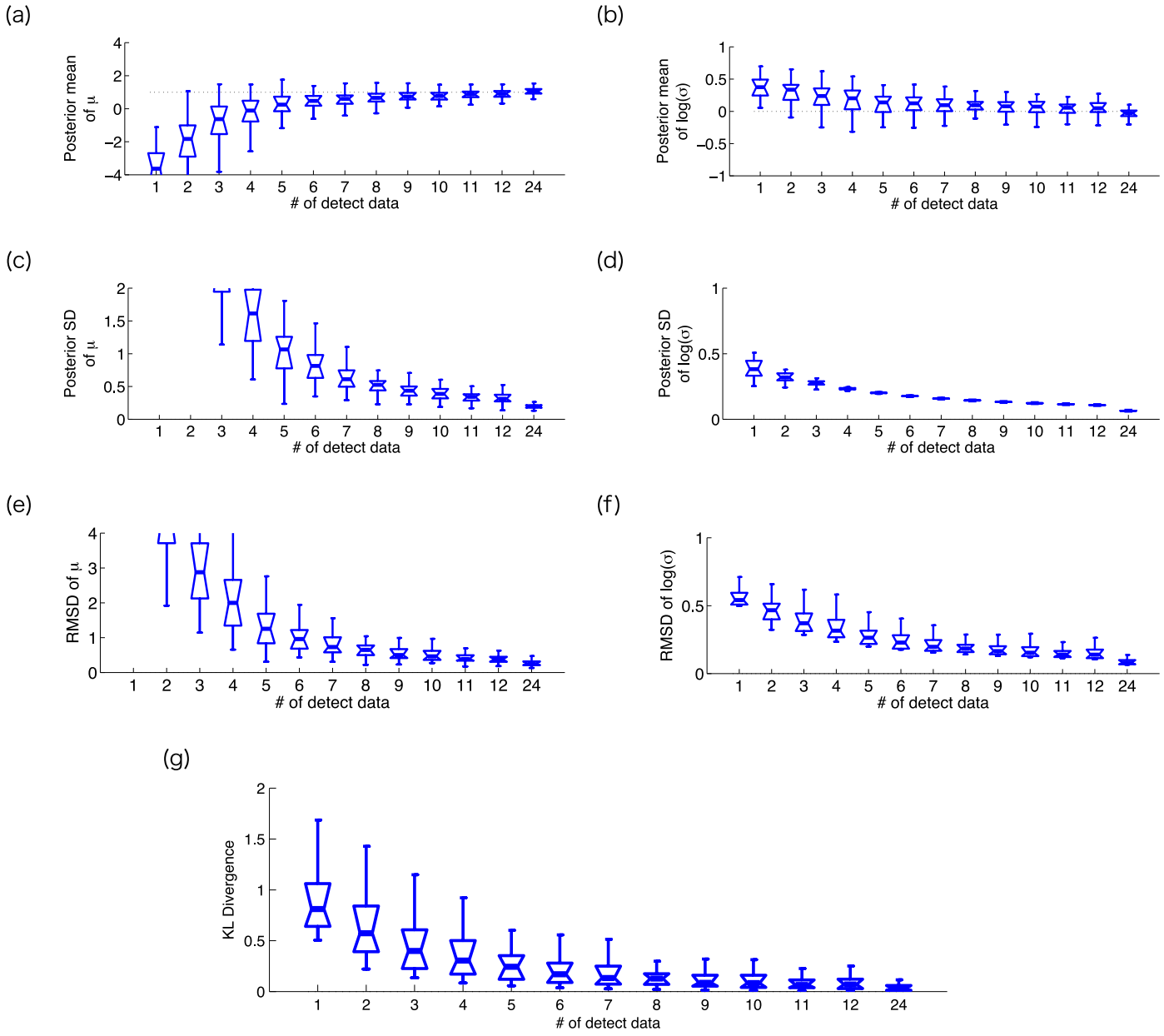

(a)
Posterior mean of μ
# of detect data
(b)
Posterior mean of log(σ)
# of detect data
(c)
Posterior SD of μ
# of detect data
(d)
Posterior SD of log(σ)
# of detect data
(e)
RMSD of μ
# of detect data
(f)
RMSD of log(σ)
# of detect data
(g)
KL Divergence
# of detect data

Figure 3

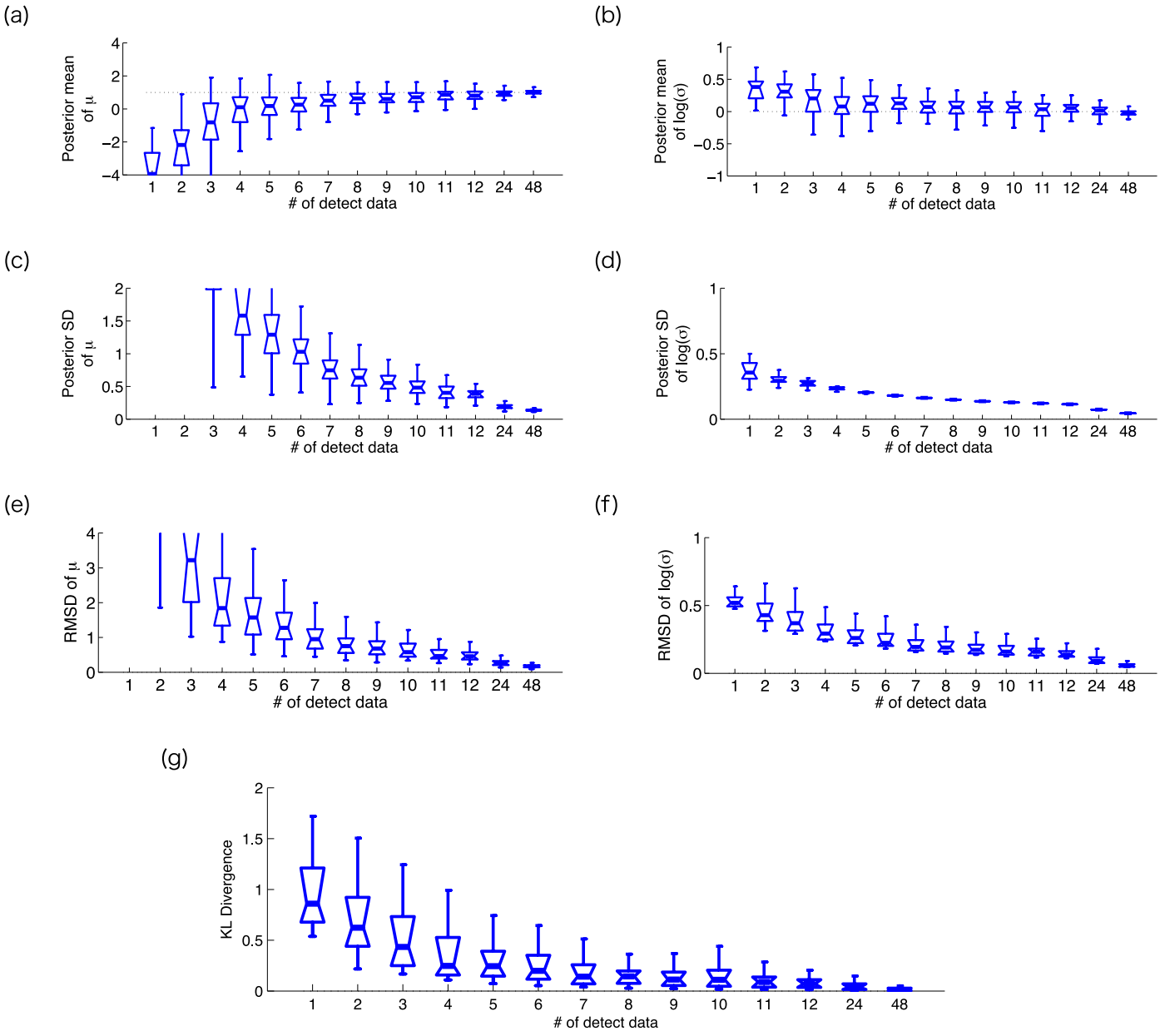
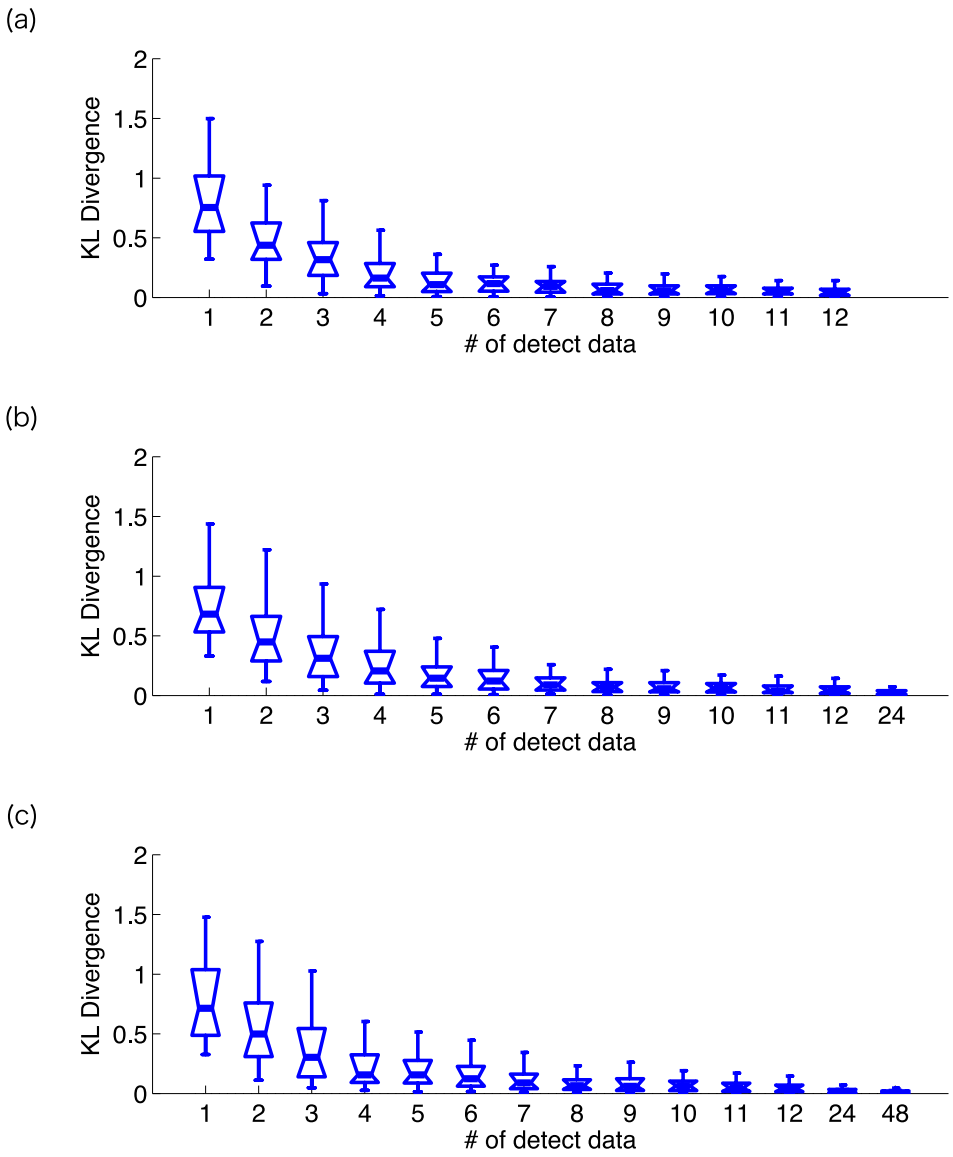

(a)
Posterior mean of μ
# of detect data
(b)
Posterior mean of log(σ)
# of detect data
(c)
Posterior SD of μ
# of detect data
(d)
Posterior SD of log(σ)
# of detect data
(e)
RMSD of μ
# of detect data
(f)
RMSD of log(σ)
# of detect data
(g)
KL Divergence
# of detect data
1 2 3 4 5 6 7 8 9 10 11 12 24 48

Figure 4


(a)
KL Divergence
2
1.5
1
0.5
0
1 2 3 4 5 6 7 8 9 10 11 12
# of detect data
(b)
KL Divergence
2
1.5
1
0.5
0
1 2 3 4 5 6 7 8 9 10 11 12 24
# of detect data
(c)
KL Divergence
2
1.5
1
0.5
0
1 2 3 4 5 6 7 8 9 10 11 12 24 48
# of detect data