Water Research 115 (2017) 138–148

Contents lists available at ScienceDirect

Water Research

journal homepage: www.elsevier.com/locate/watres

Bayesian inference of uncertainty in freshwater quality caused by low-resolution monitoring

Tobias Krueger

IRI THESys, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

ARTICLE INFO

Article history:
Received 14 September 2016
Received in revised form
11 January 2017
Accepted 26 February 2017
Available online 27 February 2017

Keywords:
Water Framework Directive
Multinomial model
Phosphorus
Nitrogen
Oxygen
Tamar

ABSTRACT

Regulatory, low temporal resolution monitoring of freshwater quality does not fully capture the frequency distributions of the requisite parameters, particularly those that are highly skewed and heavy-tailed. Hence the summary statistics ultimately compared to environmental standards are uncertain. Quantifying this uncertainty is crucial for robust water quality assessment and possible remediation, but requires strong assumptions. This paper compares three ways to model the missing data needed to fully characterise a frequency distribution in a Bayesian framework using multi-year/multi-location orthophosphate (arithmetic mean standard), dissolved oxygen (DO; 10th percentile standard) and ammonia (90th percentile standard) data from the Tamar catchment in Southwest England. First, fitting an assumed parametric model of the frequency distribution (lognormal or Weibull), there is appreciable uncertainty around the "best" model fit. Second, Bayesian Model Averaging is more general in accommodating cases where the data are ambiguous with regard to the best model, but does not take into account possibly missing data. Third, a quasi-nonparametric multinomial model of the monitoring process that places some weight on those missing data yields wider and heavier-tailed frequency distributions. One-at-a-time sensitivity analysis suggests that the multinomial model for mean orthophosphate is sensitive to the choice of support range and the prior weights given to the missing data. Sensitivity is lower for 10th percentile DO and 90th percentile ammonia. The resultant probability densities of ecological status under the EU Water Framework Directive span several status classes, meaning ecological status is more uncertain than previously acknowledged. For orthophosphate, the regulatory, empirical determination of ecological status is not only overly precise but also biased.

1. Introduction

Freshwater quality parameters, such as phosphorus, nitrogen and oxygen concentrations, are routinely monitored by environmental regulators to assess the status of surface waters and inform water resources management. In Europe, the legislative driver is currently the Water Framework Directive (WFD; 2000/60/EC). In the US, it is the Clean Water Act (33 U.S.C. ch. 26) through the Impaired Waters and Total Maximum Daily Load Program. The compliance monitoring is typically done at a low temporal resolution, which in the UK, for example, is fortnightly or monthly, so that no more than 12–26 samples per year are available. This sampling pattern does not fully capture the frequency distributions of the parameters, particularly those that are highly skewed and heavy-tailed, such as phosphorus which is characterised by short-term extremes. Johnes (2007) demonstrated this effect for daily data of discharge, suspended solids and total phosphorus, showing how the upper tails of the empirical frequency distributions contracted progressively as the data were sub-sampled to weekly and then monthly resolution. Ferrant et al. (2013) showed based on sub-sampling a 10-min nitrate-N dataset that a fortnightly monitoring scheme would have missed all extreme concentration values. This error is partly due to the operational realities of sample collection, which usually prevent sampling during heavy rainfall events and other extreme conditions that are highly relevant for pollutant mobilisation and transport. The sampling error, which of course remains unknown outside of sub-sampling studies, translates into uncertainty about the statistical moment or percentile which is ultimately compared to an environmental standard or objective. Skeffington et al. (2015) demonstrated based on sub-sampling hourly dissolved oxygen and total reactive phosphorus data how the uncertainty of the WFD classification and the risk of misclassification increased progressively when moving to weekly and

E-mail address: tobias.krueger@hu-berlin.de.

http://dx.doi.org/10.1016/j.watres.2017.02.061

monthly resolution. Despite advances in high-resolution monitoring for research purposes (Campbell et al., 2015; Jordan et al., 2005; Outram et al., 2014), the limitations of the regulatory monitoring are likely to remain that way (Johnes, 2007). There is thus an imperative to understand and work explicitly with the uncertainties associated with these data.

For robust water quality assessment and possible remediation, a quantification of water quality uncertainty becomes crucial (Skeffington et al., 2015). The problem, however, is that this type of uncertainty quantification requires consideration of the data that could have been sampled but have not, an oxymoron really, which will always rely on assumptions. The objective of this paper is to investigate which assumptions about the possibly missing data might reasonably be made and with what consequences by comparing three Bayesian approaches that quantify the uncertainty caused by data limitations probabilistically. The starting point and first approach studied here is Bayesian inference of an assumed parametric model of the frequency distribution from the available data (Gelman et al., 2013). The pivotal assumption here is the parametric model, and the problem becomes one of verifying this. I have not found a straightforward application of this approach to a water quality problem, though Carstensen (2007) fitted lognormal distributions to marine nutrient concentration data, albeit using classic Maximum Likelihood instead of Bayesian inference. Bayesian water quality studies that did infer frequency distributions from the data explicitly augmented this procedure with some process modelling (Patil and Deng, 2011; Qian and Reckhow, 2007).

It will generally be more robust to average over multiple hypotheses of the frequency distribution, known as Bayesian Model Averaging (BMA; Hoeting et al., 1999), which is the second approach analysed in this paper. BMA has, to my knowledge, not been applied to frequency distributions in a water quality context, but there are applications in other fields. Conigliani (2010) used BMA in a clinical cost-effectiveness context to average lognormal, gamma, Weibull and inverse normal models of highly skewed and heavy-tailed patient cost data. In BMA, individual model results are weighted by the model likelihood. However, the model likelihood is still conditional on the available data and not those that have not been sampled. In the case of insufficient sampling, BMA will thus generally under-estimate the true uncertainty.

BMA will be compared with a third approach, Bayesian inference of a quasi-nonparametric model, here the multinomial model of the sampling process (Aitkin, 2010), which does reflect the analyst's prior ignorance of the shape of the frequency distribution. Under a special case of prior that again neglects possibly missing data, the multinomial model is known as Bayesian bootstrap (Rubin, 1981). The Bayesian bootstrap, like the classic bootstrap (Hirsch et al., 2015), considers the available data representative of the population, which may again be unjustified for small samples from skewed and heavy-tailed frequency distributions (Conigliani, 2010). Under the more general multinomial model, as will be seen, the "prior weight" over the support range of the water quality parameter becomes the pivotal assumption. It will be discussed how this assumption can be made in practice. While the multinomial model can deal with any type of summary statistics, including percentiles and means, for percentile standards the quasi-nonparametric binomial model is a more parsimonious choice (McBride and Ellis, 2001; Smith et al., 2001; Solow and Gaines, 1995). Hence, the multinomial model results will be briefly checked for consistency with the binomial model in this paper.

The structure of the paper is as follows. Section 2 describes the methods of Bayesian inference for an assumed parametric model, BMA and the quasi-nonparametric multinomial and binomial models, and how these will be analysed and compared using data from the Tamar catchment in Southwest England. Section 3 presents the results of the analysis by comparing the summary statistics and associated uncertainty distributions resulting from the three approaches for typical moments and percentiles of three selected water quality parameters. The summary statistics will be evaluated against existing water quality standards to illustrate the impact of uncertainty on the assessment of surface waters. A sensitivity analysis of the multinomial model will be carried out. Section 4 discusses the limitations and benefits of the individual approaches, suggests how their assumptions may be best made in practice and draws out common lessons. Section 5 concludes with some general implications.

2. Methods

When assessing a water quality parameter we want to make inference about a population $\boldsymbol{Y}$, i.e. the instances of the parameter in a time window (e.g. a year), using a sample $\boldsymbol{y} = (y_1, \ldots, y_n)$ of size n drawn from the population. We are interested in summary statistics of the population, such as the arithmetic mean and percentiles. In this paper, I compare three methods of estimating these summary statistics probabilistically. Notes on mathematical notation: vectors are in bold face throughout this paper; generic parameter vectors are denoted by $\boldsymbol{\theta}$; super-script $[t]$ denotes the tth realisation of a quantity from a Monte Carlo sample.

2.1. Bayesian inference of assumed parametric model

The description follows Aitkin (2010). In Bayesian theory, assuming a parametric model of the population $f(\boldsymbol{y}|\boldsymbol{\theta})$, the posterior probability distribution of the model parameters $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ is the prior probability distribution $\pi(\boldsymbol{\theta})$ updated by the likelihood function $L(\boldsymbol{\theta}|\boldsymbol{y})$ through Bayes rule:

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{L(\boldsymbol{\theta}|\boldsymbol{y})\pi(\boldsymbol{\theta})}{\int L(\boldsymbol{\theta}|\boldsymbol{y})\ \pi(\boldsymbol{\theta})\ d\boldsymbol{\theta}}. \tag{1}$$

The likelihood function is the probability of the observed data as a function of the model parameters given measurement precision δ, which is generally considered high relative to the variability in the data:

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \left[\prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})\right]\delta^n. \tag{2}$$

Bayesian theory requires that we express any prior information as a probability distribution, although this can be non-informative relative to the information in the data. From the posterior distribution of model parameters, the desired summary statistics of $\boldsymbol{Y}$ (e.g. arithmetic mean and percentiles) can be calculated, generally by simulation, in special cases analytically. Here, I compare two parametric models of the frequency distributions of water quality parameters, the lognormal and the Weibull distribution, which were chosen for their flexible and complementary behaviour (Conigliani, 2010). The lognormal model is right-skewed whereas the Weibull model may be left-skewed, right-skewed or symmetrical, and may thus approximate the normal distribution without negative support. The models are sensible choices for water quality data that cannot be negative, are right-skewed (lognormal) with possibly heavy tails (Weibull) like orthophosphate-phosphorus and ammonia-nitrogen, or occasionally left-skewed (Weibull) like dissolved oxygen. For other data, other models may be chosen based on our theoretical understanding of their behaviour. However, our past experience of what might be suitable distributional forms may be influenced by the very same sample deficiencies that we try to

model with these distributions.

2.1.1. Lognormal distribution

The likelihood function of the lognormal model with parameters μ and σ is

$$L(\mu, \sigma|\boldsymbol{y}) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n \prod_{i=1}^n y_i} \exp\left(\frac{(n-1)s^2 + n(\bar{y}-\mu)^2}{-2\sigma^2}\right), \tag{3}$$

with sample mean $\bar{y} = \frac{1}{n}\sum_{i=1}^n \log(y_i)$ and sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^n (\log(y_i) - \bar{y})^2$ of the log-transformed data. I chose a non-informative prior $\pi(\mu, \sigma) = \frac{1}{\sigma^2}$ (Conigliani, 2010; Gelman et al., 2013). The arithmetic mean of the lognormal model is $m_a = \exp\left(\mu + \frac{\sigma^2}{2}\right)$.

2.1.2. Weibull distribution

The likelihood function of the Weibull model with parameters λ and k is:

$$L(\lambda, k|\boldsymbol{y}) = \frac{k^n \left(\prod_{i=1}^n y_i\right)^{k-1}}{\lambda^{nk}} \exp\left(\frac{\sum_{i=1}^n y_i^k}{-\lambda^k}\right). \tag{4}$$

I chose a non-informative prior $\pi(\lambda, k) = \frac{1}{\lambda}$ (Conigliani, 2010; Lee and Kim, 2008). The arithmetic mean of the Weibull model is $m_a = \lambda\,\Gamma\left(1 + \frac{1}{k}\right)$, where $\Gamma(\cdot)$ is the gamma function.

2.1.3. Sampling from posterior distribution

For each model, the posterior distribution was sampled by Markov Chain Monte Carlo (MCMC) exploiting the proportionality property $\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto L(\boldsymbol{\theta}|\boldsymbol{y})\,\pi(\boldsymbol{\theta})$. 100,000 realisations were generated via the Matlab® Slice sampler with a "burn in" phase of 10,000 samples (algorithm after Neal, 2003). The Slice sampler was chosen for its limited tuning requirements which could be automated relatively easily so as to cycle through a large number of datasets (section 2.4). The only tuning parameters required here were the two scale estimates of the parameter space (Neal, 2003). These were adjusted incrementally to achieve adequate sample coverage of the posterior distribution. For the lognormal model, this method was verified with samples generated from the analytically derived posterior distribution (Gelman et al., 2013). The MCMC sample was used to calculate the posterior distributions of the arithmetic mean and of two percentiles (10th and 90th) of the population as these are the summary statistics required for WFD reporting in the UK for the water quality parameters studied here.

2.2. Bayesian model averaging

The probabilistic treatment of parameters extends to models as a whole in Bayesian theory, so that several candidate models can be compared by their respective probabilities. In our case, either of the two models will be preferred given a dataset, but model probabilities may be very similar. A more general method is Bayesian Model Averaging (BMA), which averages the posterior distributions of the candidate model results weighted by the posterior probabilities of the models. Here, again, we are interested in averaging the posterior distributions of the arithmetic mean and of the two percentiles of the population. In case of strong evidence in the dataset for a particular model, this will dominate the average.

Following Aitkin et al. (2009), I used the posterior *distributions* of the two model likelihoods for averaging, rather than the integrated likelihoods (Conigliani, 2010; Hoeting et al., 1999). The posterior probability of model j follows, again, from Bayes rule, which updates the prior model probability $\pi(M_j)$ by the model likelihood $L(M_j|\boldsymbol{y})$ and normalises for k models:

$$\pi(M_j|\boldsymbol{y}) = \frac{L(M_j|\boldsymbol{y})\,\pi(M_j)}{\sum_k L(M_k|\boldsymbol{y})\,\pi(M_k)}. \tag{5}$$

The prior probabilities of the two models were set equal, $\pi(M_1) = \pi(M_2) = 1$, reflecting prior ignorance. The BMA algorithm then proceeded as follows (Aitkin et al., 2009):

- For each model j, substitute the samples $\boldsymbol{\theta}_j^{[t]}$ from the posterior distribution (section 2.1) into the likelihood function. These are 100,000 independent draws from each posterior likelihood $L^{[t]}(M_j|\boldsymbol{y})$.
- Compute $\pi^{[t]}(M_j|\boldsymbol{y})$ as per Equation (5).
- With probability $\pi^{[t]}(M_j|\boldsymbol{y})$, set average summary statistic $p_{ave}^{[t]} = p_j^{[t]}$, i.e. the summary statistic of model j.

2.3. Bayesian inference of quasi-nonparametric model

2.3.1. Multinomial model

The multinomial model ("sampling with replacement") can be considered an always true model of a population and thus allows nonparametric inference (Aitkin, 2010). It holds for water quality monitoring if this can be treated as random sampling from a finite population with measurement precision δ and thus D discrete values Y_J with counts N_J and proportions $p_J = \frac{N_J}{\sum_J N_J}$. A sample from the population can be expressed through the sample counts n_J at Y_J, most of which will be zero. The randomness assumption holds approximately at fortnightly to monthly resolution since every element of the population has approximately equal chance of being sampled. Water quality parameters also effectively stem from a finite discrete population as the measurement precision δ will always be finite (e.g. the detection limit) and there will be physical limits to the smallest and largest population value Y_1 and Y_D, respectively.

The likelihood function of the multinomial model, constant combinatorial term omitted, is:

$$L(p_1, \ldots, p_D|\boldsymbol{y}) = \prod_{J=1}^D p_J^{n_J}. \tag{6}$$

I chose the natural conjugate Dirichlet prior

$$\pi(p_1, \ldots, p_D) = \frac{\Gamma\left(\sum_{J=1}^D a_J\right)}{\prod_{J=1}^D \Gamma(a_J)} \prod_{J=1}^D p_J^{a_J - 1}, \tag{7}$$

with a "prior weight" of each possible discrete value of $a_J = \varepsilon$ and a "total prior weight" of $a = \sum_{J=1}^D a_J = 1$ (the proper prior of Ericson, 1969), thus non-informative. As discussed by Aitkin (2010), the choice of prior is more important here than in parsimonious parametric models since many of the positive values of n_J will be 1 or a small integer. With my choice, the effective information provided by the prior as a whole is equal to the information provided by *one* sample value. The total prior weight a thus augments the total sample weight $n = \sum_{J=1}^D n_J$ in the posterior. For the choice of the improper Haldane prior with $a_J = 0$ the term "Bayesian bootstrap" has been used (Rubin, 1981). The arithmetic mean of the

multinomial model is $m_a = \sum_{J=1}^{D} p_J\, Y_J$.

The posterior distribution was sampled 100,000 times as follows (Aitkin, 2010):

- For each realisation, generate D independent variables G_J from the Gamma$(n_J + a_J, 1)$ distribution.
- Calculate $p_J = \frac{G_J}{\sum G_J}$.
- Repeat the simulation T times yielding realisations $p_J^{[t]}$.

2.3.2. Binomial model

The binomial model can be considered a special case of the multinomial model where the interest is in the exceedance proportion p of a specific water quality parameter value in a population, not in the exact frequency distribution (McBride and Ellis, 2001; Smith et al., 2001; Solow and Gaines, 1995). This situation is given when we have a percentile standard where a specific parameter value must not be exceeded more than say 10% of the time (the 90th percentile ammonia-N standard of the UK implementation of the WFD) or where a specific parameter value must be exceeded at least say 90% of the time (the 10th percentile dissolved oxygen standard). A sample $\boldsymbol{y}$ from the population can be expressed through the sample size n and the number of exceedances in the sample e.

The likelihood function of the binomial model, again omitting the constant combinatorial term (here the binomial coefficient), is:

$$L(p|\boldsymbol{y}) = p^e\ (1-p)^{n-e}. \tag{8}$$

As prior I compared the uniform prior $\pi(p) = \text{Beta}(1,1)$ (McBride and Ellis, 2001; Smith et al., 2001) and the Jeffreys prior $\pi(p) = \text{Beta}(0.5, 0.5)$ (McBride and Ellis, 2001). The posterior distribution was derived analytically (Gelman et al., 2013).

2.4. Data and analytical steps

The data used to compare the three methods comprise 29 locations in the Tamar catchment, Southwest England, which have been monitored routinely by the Environment Agency of England and Wales (Fig. 1). For each location, I used 17 three-year moving average windows from 1991 to 2007 (±one year) in keeping with

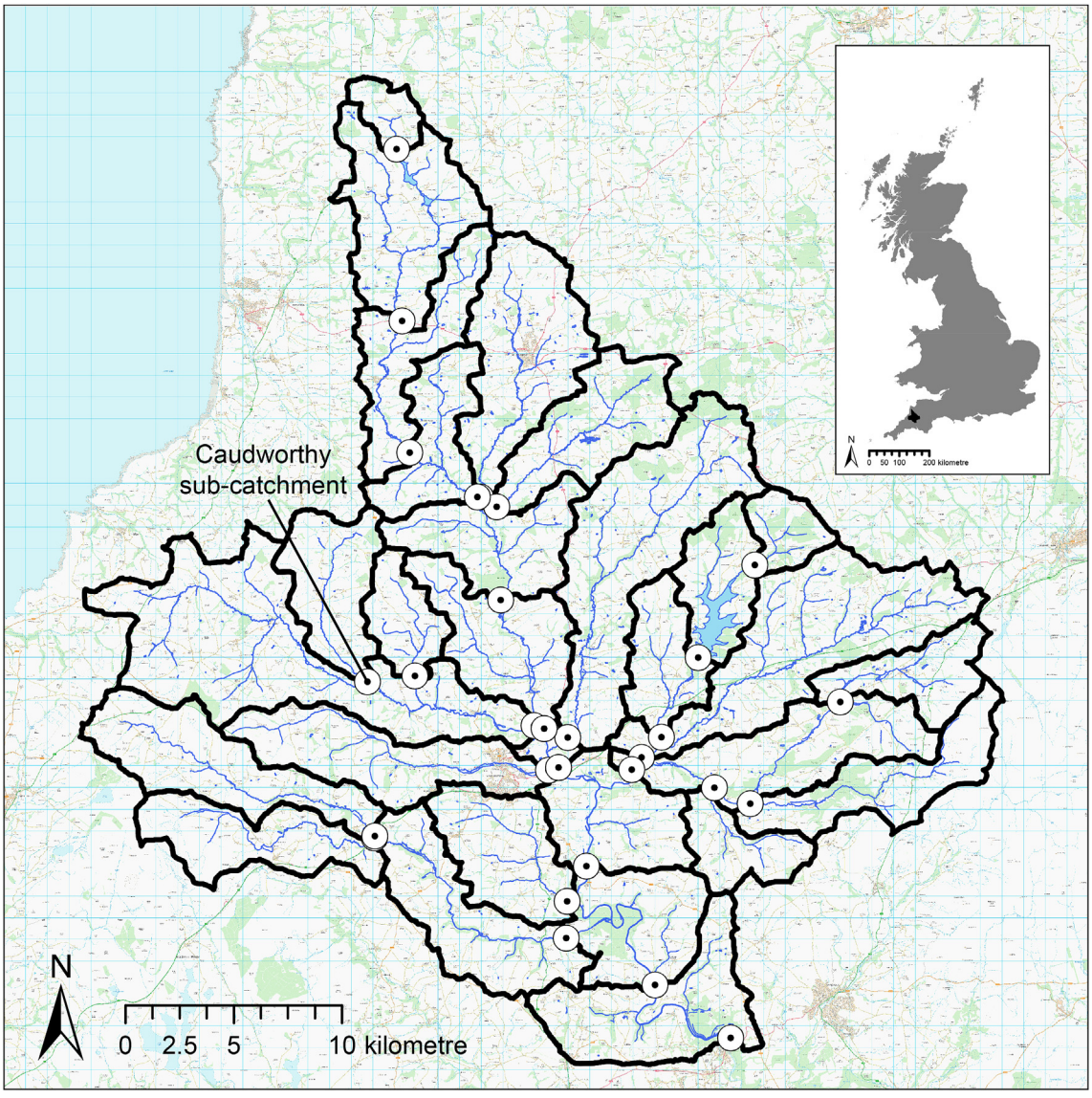

Fig. 1. Tamar catchment with monitoring locations and corresponding sub-catchments. Contains OS data© Crown copyright and database right 2016.

WFD reporting, yielding 493 datasets in total. Summary results will be reported for these datasets while more detailed examples will be given for one location, the Caudworthy Water at the Ottery confluence, for the period 2006–2008.

Three water quality parameters were selected, each requiring a different summary statistic for the determination of ecological status: orthophosphate-phosphorus (arithmetic mean), dissolved oxygen (10th percentile), ammonia-nitrogen (90th percentile). The multinomial model inference was set up as follows: $Y_1 = 0$ for all parameters; $Y_D = 4$ mg l^{-1} for orthophosphate-P, $Y_D = 150$ % for dissolved oxygen, $Y_D = 20$ mg l^{-1} for ammonia-N; $\delta = 0.01$ mg l^{-1} (detection limit) for orthophosphate-P and ammonia-N, $\delta = 1\%$ for dissolved oxygen. The Y_D values are the maxima recorded across all datasets from the Tamar, rounded up.

Since especially the choice of Y_D is difficult to benchmark, and might differ between applications, a one-at-a-time (OAT) sensitivity analysis of the multinomial model with respect to its parameters was carried out for the three selected water quality parameters using the Caudworthy data 2006–2008. The total prior weight a was varied between 0 and 2, with $a = 0$ being the Bayesian bootstrap which neglects possibly missing data. The largest population value Y_D was varied between a value just above the maximum recorded in the Caudworthy data 2006–2008 and a value about double the maximum recorded across all datasets from the Tamar. The measurement precision δ was varied between the detection limit (set to 1% for dissolved oxygen) and a value one order of magnitude higher.

Parametric model preference and resultant BMA for each location and water quality parameter will be summarised by the deviance difference:

$$D_{12} = -2\log[L_1/L_2], \tag{9}$$

with negative D_{12} indicating preference for model 1 (here the lognormal distribution) and positive D_{12} indicating preference for model 2 (here the Weibull distribution).

BMA and the multinomial model will be compared for each location and water quality parameter by the difference in summary statistic. The multinomial model will then be used for further analysis. First, the uncertainties of the summary statistics will be compared using the interquartile range of the probability density divided by the median as measure of uncertainty. Second, it will be illustrated how the probabilistic water quality assessments may be translated into policy relevant metrics, here probability statements of ecological status under the WFD and temporal trends. The uncertainties of ecological status according to the different water quality parameters will be compared using the Shannon entropy of the ecological status histogram as measure of uncertainty:

$$H = -\sum_{i=1}^{5} \pi(s_i|\boldsymbol{\theta})\log_2 \pi(s_i|\boldsymbol{\theta}), \tag{10}$$

with $\pi(s_i|\boldsymbol{\theta})$ being the posterior probability of the ith ecological status class. A value of $H = 0$ means no uncertainty, while greater values of H mean greater uncertainty, i.e. a distribution approaching uniformity. The biases of the empirical ecological status estimates will be compared using the difference between the empirical estimate and the mode of the probability density as measure of bias. Empirical here means calculated straight from the sample. Third, a comparison to the binomial model for the percentile standards will be made using the mean absolute difference of the ecological status distributions of the multinomial and binomial models.

3. Results

When fitting the assumed parametric models, there is appreciable uncertainty around the fit of the "best" model (Fig. 2a–c)

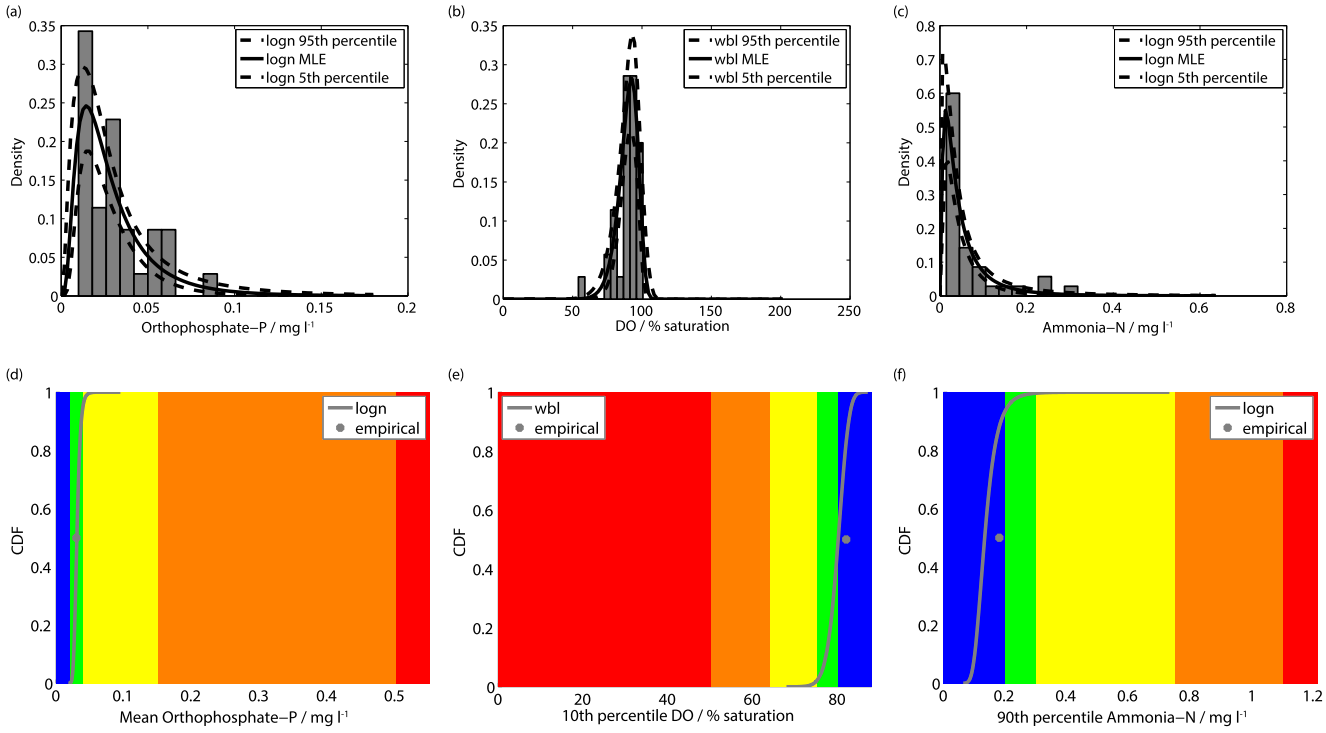

Fig. 2. Caudworthy data 2006–2008. (a–c): Empirical frequency distribution and "best" model fit, lognormal (logn) or Weibull (wbl), for three water quality parameters; 5th and 95th percentiles of Bayesian inference with maximum likelihood estimate (MLE) shown for comparison. (d–f): Empirical summary statistic and cumulative distribution function (CDF) of "best" model result against background of WFD ecological status classes (blue = "high", green = "good", yellow = "moderate", orange = "poor", red = "bad"). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

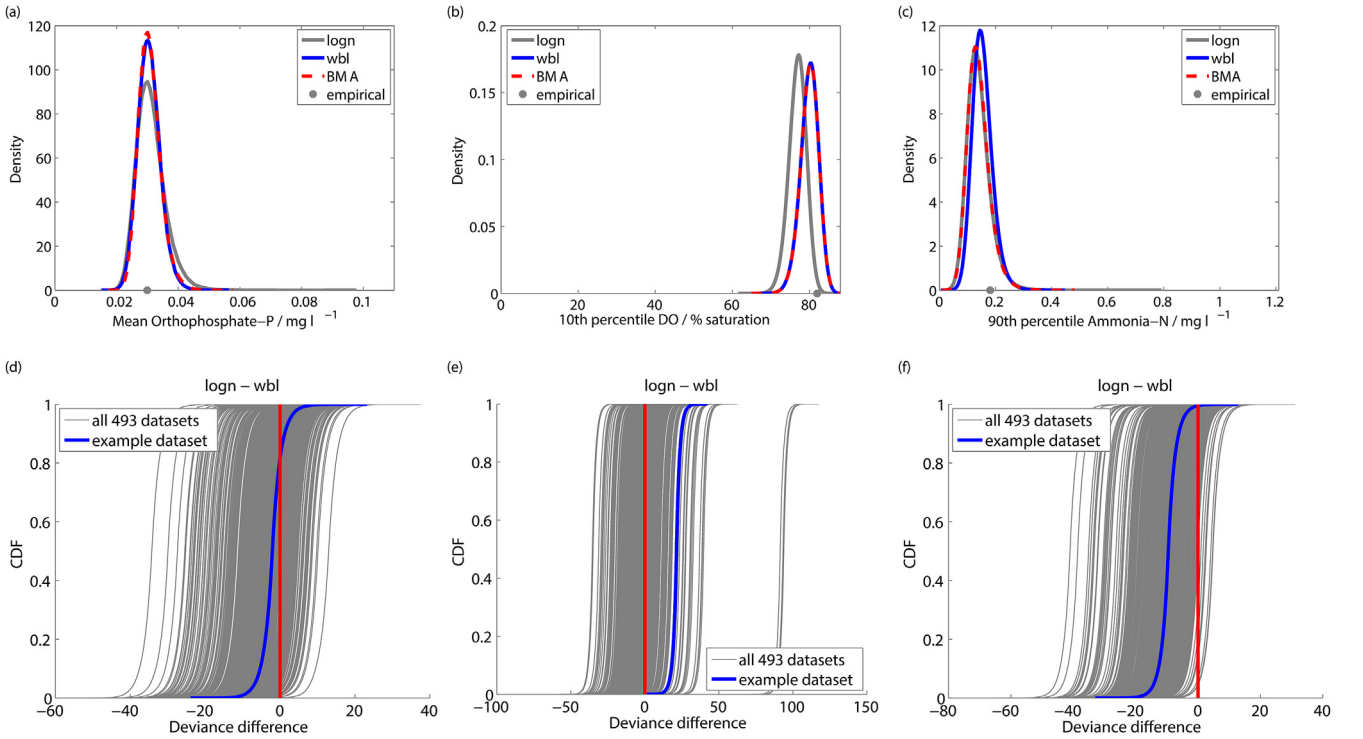

Fig. 3. (a–c): Caudworthy data 2006–2008. Empirical summary statistic and densities of three model results for three water quality parameters: lognormal (logn), Weibull (wbl) and Bayesian Model Averaging (BMA) of the former two. (d–f): All 493 datasets, Caudworthy example highlighted. Cumulative distribution function (CDF) of deviance difference between lognormal and Weibull model.

which translates into uncertainty around the summary statistics of the population (Fig. 2d–f). Across all datasets and parameters, often one model is strongly preferred over the other, although there are instances of greater ambiguity demonstrating the benefit of BMA when an average of both candidate models is suggested by the data (Fig. 3). In Fig. 3b, for example, the Weibull distribution is

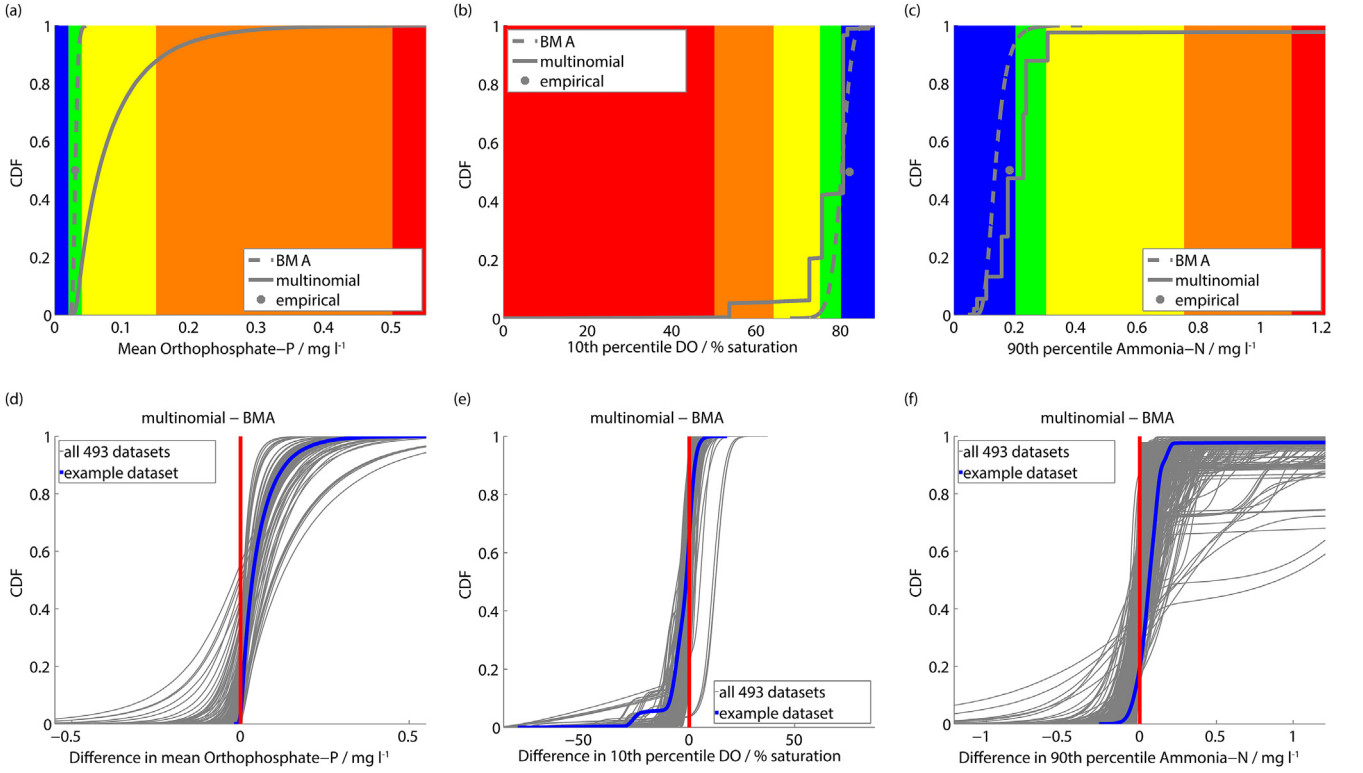

Fig. 4. (a–c): Caudworthy data 2006–2008. Empirical summary statistic and cumulative distribution functions (CDFs) of multinomial model result and Bayesian Model Averaging (BMA) for three water quality parameters against background of WFD ecological status classes (blue = "high", green = "good", yellow = "moderate", orange = "poor", red = "bad"). (d–f): All 493 datasets, Caudworthy example highlighted. CDF of difference in summary statistic between multinomial model result and BMA. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

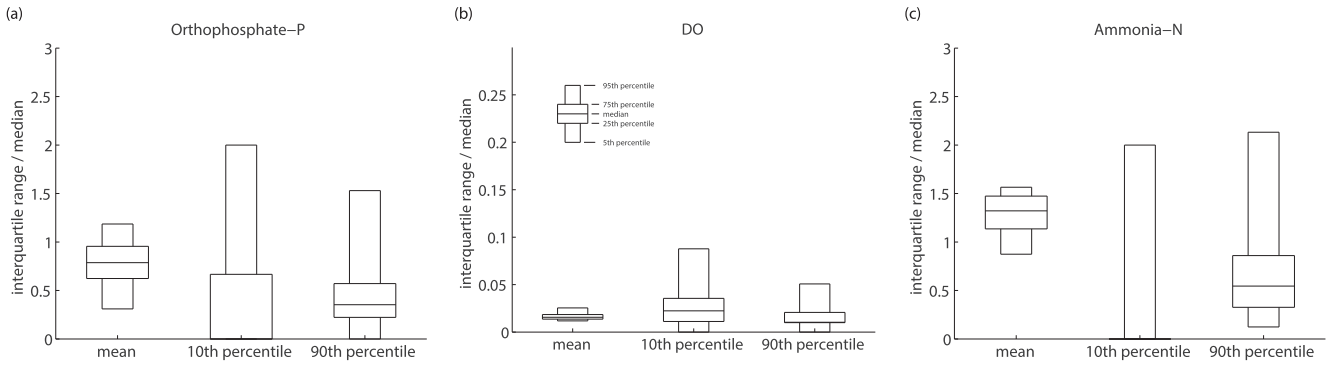

Fig. 5. All 493 datasets. Widths of probability densities of three summary statistics for three water quality parameters summarised as boxplots across all locations and years. Width of density expressed as interquartile range normalised by median.

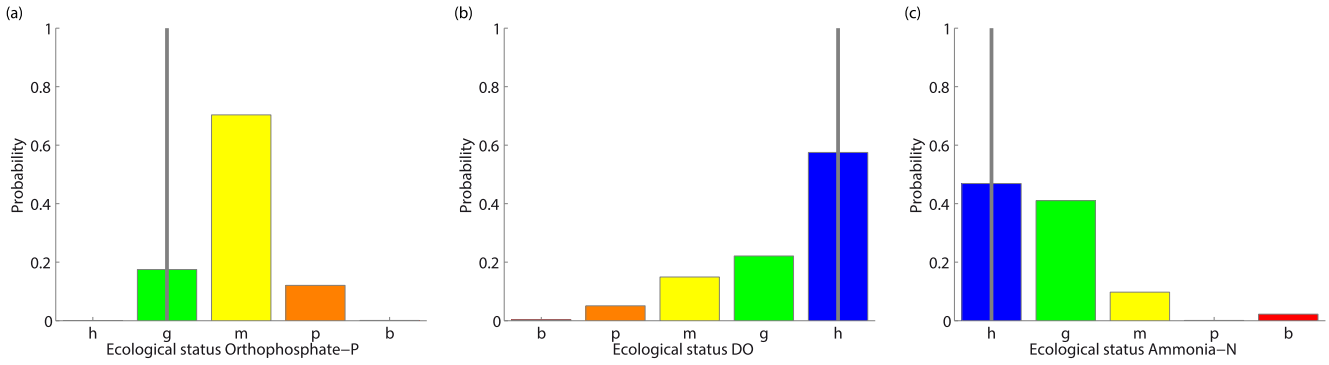

Fig. 6. Caudworthy data 2006–2008. Empirical ecological status (vertical line) and probability of WFD ecological status according to the multinomial model for three water quality parameters: "high" (h), "good" (g), "moderate" (m), "poor" (p),"bad" (b).

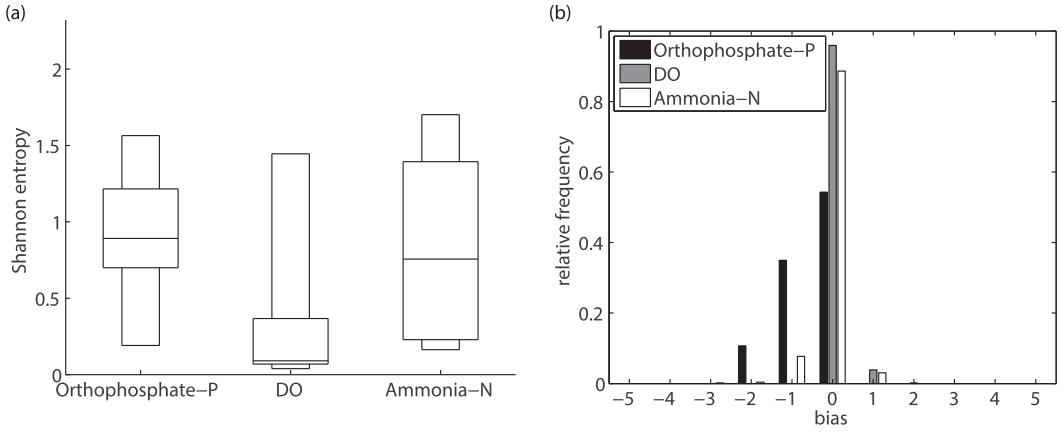

Fig. 7. All 493 datasets. (a): Shannon entropy as a measure of uncertainty of the ecological status distribution for three water quality parameters. The distributions of values across all locations and years are visualised as boxplots, see Fig. 5 for legend. Greater entropy values mean greater uncertainty. (b): Bias of the empirical ecological status with respect to the mode of the probability density for three water quality parameters. The distributions of values across all locations and years are visualised as histograms. Negative values mean the empirical value is to the left of the mode, positive values mean it is to the right.

unambiguously preferred over the lognormal distribution, and so BMA selects the Weibull model. Similarly in Fig. 3c, the lognormal distribution is preferred over the Weibull distribution, and this is again reflected in the BMA result. In Fig. 3a, in contrast, BMA yields a 4:1 average of lognormal:Weibull. For orthophosphate-P and ammonia-N, the lognormal model is preferred in many cases (Fig. 3d, f) as these data are mostly right-skewed, whereas for dissolved oxygen the preference tends towards the Weibull model (Fig. 3e) as these data are mostly left-skewed. Often both models yield appreciable likelihoods. Similar tendencies should be expected for the same parameters in other locations, although model preference is already highly variable among the 493 small-*n* datasets studied here.

Compared to the BMA results, the quasi-nonparametric multinomial model yields distributions that are consistently wider and heavier-tailed, particularly towards lower ecological status (Fig. 4). This behaviour is similar across the parameters. When comparing the summary statistics for orthophosphate-P and ammonia-N, the

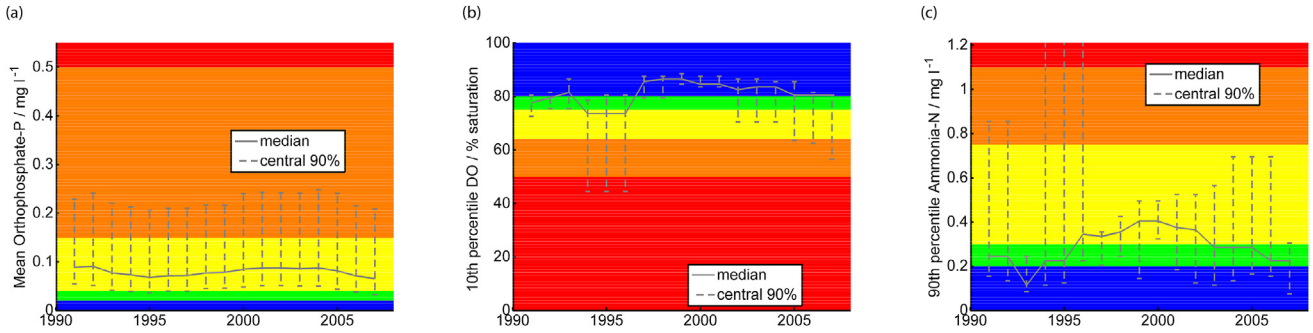

Fig. 8. Caudworthy data 2006–2008. Time series of the probability densities (median and central 90%) of three water quality parameters according to the multinomial model against background of WFD ecological status classes (blue = "high", green = "good", yellow = "moderate", orange = "poor", red = "bad"). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

mean is more uncertain than the 90th percentile, which in turn is more uncertain than the 10th percentile (Fig. 5a, c). The orders of magnitude of uncertainty are comparable between the two parameters. For dissolved oxygen, the uncertainties are an order of magnitude smaller and similar across the summary statistics, with the 10th percentile being slightly more uncertain than the mean and the 90th percentile (Fig. 5b). The uncertainty of dissolved oxygen is smaller than that of the other parameters because dissolved oxygen is bounded by a narrower physical range and the data are less variable.

The probabilistic water quality assessment can be translated into policy relevant metrics. Fig. 6 exemplifies this for probability statements of ecological status under the WFD, based on the multinomial model. Conversely, the probability of site misclassification can be quantified. Ecological status is most uncertain for orthophosphate-P, followed by ammonia-N, with dissolved oxygen being much less uncertain (Fig. 7a). The empirical ecological status for orthophosphate-P is negatively biased with respect to the mode of the probability distribution, i.e. over-predicting ecological status, by one status class in 35% of the cases and by two status classes in 11% of the cases. Ammonia-N and dissolved oxygen are unbiased to 89% and 96%, respectively (Fig. 7b). The results are similar to those resulting from the binomial model for percentile standards when using a uniform prior for ammonia-N and a Jeffreys prior for dissolved oxygen. The mean absolute difference between the two probability densities of ecological status is below 0.01 in 73% of the cases for ammonia-N and in 78% of the cases for dissolved oxygen. The results are not exactly the same because using a non-informative prior on the support of the water quality parameter, as in the multinomial model, is not the same as using a non-informative prior on the exceedance proportions, as in the binomial model. Demonstrating these differences in detail is beyond the scope of this paper. Another policy relevant application is trend analysis allowing probabilistic comparisons of years and locations. Fig. 8 exemplifies this for temporal trends, based on the multinomial model. If the 90% credible intervals in these plots are especially wide then this is caused by a few extreme sample values in those years.

The multinomial model is sensitive to the choices of prior weights of possibly missing data, largest population value and measurement precision (Fig. 9). Based on the OAT sensitivity analysis for the Caudworthy data 2006–2008, mean orthophosphate-P shows high sensitivity to total prior weight a and largest population value Y_D , and low sensitivity to measurement precision δ. Increasing Y_D widens the probability distribution (Fig. 9d), while increasing a widens the distribution as well as shifting its mode to higher values (Fig. 9a). When $a = 0$ (Bayesian bootstrap) then possibly missing data are neglected and hence the resultant distribution is similar to that resulting from BMA (Fig. 4a). Increasing δ by one order of magnitude above the detection limit from 0.01 to 0.1 has little effect (Fig. 9g). Note 0.09 is the largest orthophosphate-P value recorded in the Caudworthy data 2006–2008. Increasing levels of aggregation beyond this point (not shown) shift the mode of the distribution linearly to higher values as the centre of the increasingly wider bin containing all the data increases. The effect of decreasing δ below the detection limit of 0.01 (not shown) is negligible, and physically unrealistic.

The 10th percentile dissolved oxygen and the 90th percentile ammonia-N show lower sensitivities to the multinomial model parameters than mean orthophosphate-P. Both parameters are moderately sensitive to δ (Fig. 9h–i), with increasing values shifting the distributions to lower values, which in the case of dissolved oxygen means lower ecological status. Increasing δ has the effect of increasing the widths of the bins containing the data and hence reassigning the data values to new centre points. Extreme percentiles are more sensitive to these reassignments than the mean, which for the data used here leads to lower values. The effect can be confirmed when calculating the percentiles empirically from the raw histograms using the same variation of bin widths.

The sensitivity of both the 10th percentile dissolved oxygen and the 90th percentile ammonia-N is low with respect to a (Fig. 9b–c) and negligible with respect to Y_D (Fig. 9e–f). Increasing Y_D while keeping a constant has the effect of rebalancing the total prior weight between the two ranges either side of the cluster of observations, assigning a relatively greater weight to the range of values above the observations, while also extending this range to greater values. While this affects the mean, as shown above, the effect on percentiles depends on the previous size of the ranges either side of the observations. For the ammonia-N data used here, the range below the cluster of observations is small. Hence increasing Y_D while keeping a constant does not appreciably change the prior weight of extreme values; only the range of the extreme values increases. While greater values will be sampled, the proportion of values above a certain percentile remains unchanged, resulting in no discernible effect on the 90th percentile ammonia-N (Fig. 9f). For the dissolved oxygen data used here, the range above the cluster of observations is much smaller than the range below, because of the physical limits to the upper range. Hence increasing Y_D rebalances the total prior weight slightly in favour of the upper range, meaning this range is relatively over-sampled and the 10th percentile dissolved oxygen distribution contracts slightly towards greater values (Fig. 9e).

The effect on the percentiles of increasing a is greater than that of increasing Y_D because the increase in total prior weight reinforces the dominance of the larger of the two ranges either side of the observations, leading to relatively more samples being drawn

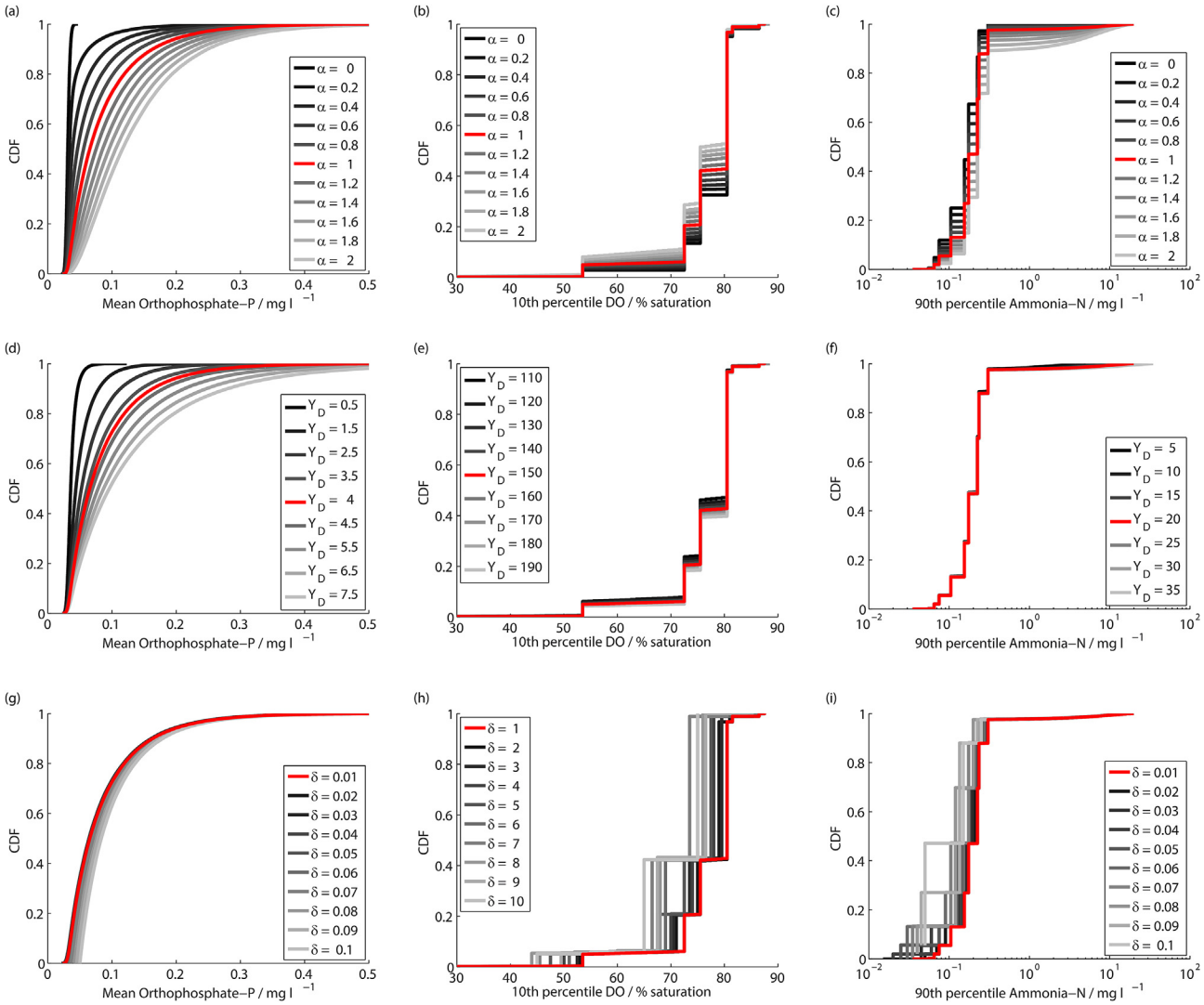

Fig. 9. Caudworthy data 2006–2008. Sensitivity of multinomial model for three water quality parameters to choice of (a–c) total prior weight a, (d–f) largest population value Y_D and (g–i) measurement precision δ. Choices made in the remainder of the paper are in red. The case $a = 0$ (a–c), where possibly missing data get zero weight, is known as the Bayesian bootstrap. Note the log-scale for ammonia-N. Note the CDFs of the percentiles (b, c, e, f, h, i) are step-like because the posterior distributions of percentiles under the multinomial model are discrete (Aitkin, 2010). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

from that side. Hence, for the data used here, the 10th percentiles dissolved oxygen distribution expands towards smaller values (Fig. 9b) while the 90th percentile ammonia-N distribution expands towards greater values (Fig. 9c). Still, for reasons explained above, the effects are smaller than for the mean.

4. Discussion

If we assume a theoretical model of the frequency distribution of a water quality parameter then Bayesian inference provides coherent information about the uncertainty of a summary statistic given the available data and some prior belief about the model parameters; in our case effectively complete uncertainty. However, how do we know the model is right?

A more general approach is Bayesian Model Averaging (BMA), which provides an average probability density for a summary statistic across the candidate models (see Fig. 3a). In case of strong evidence in the available data for a particular model, this will dominate the average (see Fig. 3b, c). Because the individual model results are weighted by the model likelihood, which is still conditional on the available data and not those that have not been sampled, BMA will in the general case of insufficient sampling still under-estimate the true uncertainty. How then to place more emphasis on the data that have not been collected, but could have been?

The quasi-nonparametric multinomial model has the desired property of increasing the spread of the probability density of a summary statistic towards possibly missing data. However, critical assumptions are the largest population value, the measurement precision, and the “prior weights” of the missing data. The smallest population value is uncritical as this can be fixed at zero. For the measurement precision, a physically defensible choice is the detection limit. The OAT sensitivity analysis suggests that the choices of prior weights and largest population value have large effects on mean orthophosphate-P, whereas they have small effects on the 10th percentile dissolved oxygen and the 90th percentile ammonia-N (Fig. 9). Setting the largest population value for each parameter at the maximum value recorded anywhere in the Tamar between 1990 and 2008 was a pragmatic choice, but meant that aspects of the data entered the analysis *a prior*, in principle violating

orthodox Bayesian reasoning. In future analyses it would seem more reasonable to determine physical bounds of the largest population values.

Choosing the prior weights will be more ambiguous. A "total prior weight" of 1 as used here effectively assigns the missing data collectively the weight of one data point. This may be considered a reasonable lower limit. Increasing the total prior weight shifts the probability density towards extreme values (Fig. 9). Certainly for mean orthophosphate-P this shift is rapid and hence values should be chosen with care. The exact choice of the total prior weight as well as its distribution over the discrete prior weights a_J will be best made by expert elicitation, as is commonly done in Bayesian analysis. However, even experts will find it difficult to weigh total prior weight against total sample weight as these are abstract concepts. Moreover, the experts' experiences will again be biased to the data ranges that have been observed in the past, not the extreme values that might have been missed. If the monitoring purpose is compliance regulation, as the WFD posits (Skeffington et al., 2015), then pollution incidents such as industrial or farm yard spills that are commonly at least partially missed by standard monitoring schemes are real concerns that can increase means and percentiles, despite deviating considerably from past experience. Hence expert elicitation will have to find novel ways of elaborating (O'Hagan, 2012) the prior weights.

Perhaps the weighing of evidence that was collected against that which was not, but could have been is more of a value judgement, reflecting one's aversion or not to the possibility of surprise and one's affinity or not to precaution. In this case, the choice of prior weights would be best deliberated by the stakeholders affecting or affected by the management decisions made on the basis of the water quality assessment (Krueger et al., 2012). The process of stakeholder deliberation, just like expert elicitation, would require grappling intensely with the inner workings of the multinomial model so that participants understand the impacts of their choices. However, not many stakeholders will want to engage with such technical problems; and if they do, the process will take considerable time (Krueger et al., 2016). It certainly seems easier to assume a parametric model for which parameter estimation is straightforward, but if the underlying data are biased then any parametric model will be grossly misleading. Hence we should invest the time and effort required to deliberate what assumptions we should make about those possibly missing data.

Important as it is to discuss the appropriateness of uncertainty models, we must remind ourselves that water quality status, such as defined by the WFD, is uncertain – no matter which uncertainty model we choose. This uncertainty has been recognised in other elements of the WFD, such as the biological quality elements and, often forgotten, the boundaries of the ecological status classes themselves (Nõges et al., 2009). In the UK, for example, the class boundaries of orthophosphate-P in rivers have been revised in 2013 (UK Technical Advisory Group on the Water Framework Directive, 2013). My results back up the conclusion of Skeffington et al. (2015) that there is considerable risk of misclassification of ecological status. Moreover, the uncertainty has implications for trend detection (Hirsch et al., 2015), resulting in potentially spurious trends in water quality that are artefacts of the sampling scheme (Skeffington et al., 2015).

Contrary to Skeffington et al. (2015), I found the empirical class estimates for orthophosphate-P based on standard sampling to be frequently (in 46% of the cases) incongruent with the modes of the probability densities; meaning the empirical analysis is often biased, not just overly precise. This phenomenon follows from my parameterisation of the *a priori* ignorance of missing data, which implies an expectation that greater values of orthophosphate-P could have been collected, in line with empirical evidence (Ferrant et al., 2013; Johnes, 2007). The sub-sampling study of Skeffington et al. (2015) shows cases where this expectation holds *a posteriori* for 98th percentile temperature, but not for mean total reactive phosphorus. Their uncertainty distributions for the percentiles are also wider than those for the means, while I found the opposite for orthophosphate-P and ammonia-N, but comparable uncertainties of means and percentiles for dissolved oxygen. The sign and magnitude of the empirical bias and the width of the uncertainty distribution will depend on the water quality parameter, the summary statistic and the pollution dynamics in specific locations. While the statistical methodology is readily transferable to other locations and other applications, the particular setup of the multinomial model used in this paper will require validation against high-resolution data in sub-sampling experiments.

5. Conclusions

Statements of freshwater quality such as ecological status under the EU Water Framework Directive (WFD) are uncertain because standard low-resolution monitoring does not resolve the full frequency distributions of the requisite water quality parameters. Using a Bayesian multinomial model of the monitoring process to quantify this uncertainty yields probability densities of ecological status that span several status classes. The modes of the probability densities can be very different from the empirical summary statistics, such as in the orthophosphate-P datasets studied here where the mode was frequently one status class below the empirical status, occasionally two. This means that standard regulatory practice leads not only to overly precise but occasionally biased results.

The quantification of freshwater quality uncertainty requires strong assumptions about the missing data needed to fully characterise the frequency distribution. Three assumptions that might be made in a Bayesian framework were compared in this paper. First, assuming a parametric model of the frequency distribution is straightforward, but the model is impossible to validate. Second, Bayesian Model Averaging (BMA) alleviates the problem of model validation to some extent by comparing and averaging over multiple candidate models, but yields overly precise results when sampling is insufficient. Third, the multinomial model is a quasi-nonparametric choice that places some weight on the missing data, including the tails of the frequency distribution, which makes it preferable over the other two models. However, crucial assumptions here are the upper bound of the possibly missing parameter values and the prior weights given to these missing data. In limiting cases, the three methods overlap, i.e. BMA selects a single parametric model when the available data are unambiguous and the multinomial model is similar to BMA when possibly missing data are ignored (Bayesian bootstrap).

I have argued that the model assumptions are best deliberated by the stakeholders affecting or affected by the management decision made on the basis of the water quality assessment, as these choices really are value judgements related to surprise and precaution. It will be interesting to discuss whether the conventional precautionary practice of WFD implementation that uses the lowest of several indicators to determine overall status ("one-out-all-out" principle) should be retained in the probabilistic assessment (and how), or whether an average across all indicator distributions should be preferred instead.

Acknowledgements

The initial work reported here was funded by a UK Natural Environment Research Council (NERC) Knowledge Exchange

Fellowship (grant no. NE/J500513/1) and a UK Research Councils Rural Economy and Land Use (RELU) project (grant no. RES-229-25-0009-A) held at the University of East Anglia (UEA). The computations were carried out on the High Performance Computing Cluster supported by the Research Computing Service at UEA and a large-memory server at IRI THESys, Humboldt-Universität zu Berlin which is funded by the German Excellence Initiative. The data were supplied under license by the Environment Agency of England and Wales, but no official endorsement of the results should be inferred. The paper was written at IRI THESys.

References

Aitkin, M., Liu, C.C., Chadwick, T., 2009. Bayesian model comparison and model averaging for small-area estimation. Ann. Appl. Stat. 3 (1), 199–221.
Aitkin, M., 2010. Statistical Inference: an Integrated Bayesian/Likelihood Approach. Chapman and Hall/CRC, Boca Raton.
Campbell, J.M., Jordan, P., Arnscheidt, J., 2015. Using high-resolution phosphorus data to investigate mitigation measures in headwater river catchments. Hydrol. Earth Syst. Sci. 19 (1), 453–464.
Carstensen, J., 2007. Statistical principles for ecological status classification of Water Framework Directive monitoring data. Mar. Pollut. Bull. 55 (1–6), 3–15.
Conigliani, C., 2010. A Bayesian model averaging approach with non-informative priors for cost-effectiveness analyses. Stat. Med. 29 (16), 1696–1709.
Ericson, W.A., 1969. Subjective Bayesian models in sampling finite populations. J. R. statistical Soc. Ser. B Methodol. 31 (2), 195–233.
Ferrant, S., Laplanche, C., Durbe, G., Probst, A., Dugast, P., Durand, P., Sanchez-Perez, J.M., Probst, J.L., 2013. Continuous measurement of nitrate concentration in a highly event-responsive agricultural catchment in south-west of France: is the gain of information useful? Hydrol. Process. 27 (12), 1751–1763.
Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian Data Analysis. Chapman and Hall/CRC, Boca Raton.
Hirsch, R.M., Archfield, S.A., De Cicco, L.A., 2015. A bootstrap method for estimating uncertainty of water quality trends. Environ. Model. Softw. 73, 148–166.
Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Stat. Sci. 14 (4), 382–401.
Johnes, P.J., 2007. Uncertainties in annual riverine phosphorus load estimation: impact of load estimation methodology, sampling frequency, baseflow index and catchment population density. J. Hydrol. 332 (1–2), 241–258.
Jordan, P., Arnscheidt, J., McGrogan, H., McCormick, S., 2005. High-resolution phosphorus transfers at the catchment scale: the hidden importance of non-storm transfers. Hydrol. Earth Syst. Sci. 9 (6), 685–691.
Krueger, T., Page, T., Hubacek, K., Smith, L., Hiscock, K., 2012. The role of expert opinion in environmental modelling. Environ. Model. Softw. 36, 4–18.
Krueger, T., Maynard, C., Carr, G., Bruns, A., Mueller, E.N., Lane, S., 2016. A transdisciplinary account of water research. Wiley Interdiscip. Rev. Water 3, 369–389.
Lee, K.S., Kim, S.U., 2008. Identification of uncertainty in low flow frequency analysis using Bayesian MCMC method. Hydrol. Process. 22 (12), 1949–1964.
McBride, G.B., Ellis, J.C., 2001. Confidence of compliance: a bayesian approach for percentile standards. Water Res. 35 (5), 1117–1124.
Neal, R.M., 2003. Slice sampling. Ann. Stat. 31 (3), 705–741.
Nõges, P., van de Bund, W., Cardoso, A.C., Solimini, A.G., Heiskanen, A.-S., 2009. Assessment of the ecological status of European surface waters: a work in progress. Hydrobiologia 633 (1), 197–211.
O'Hagan, A., 2012. Probabilistic uncertainty specification: overview, elaboration techniques and their application to a mechanistic model of carbon flux. Environ. Model. Softw. 36, 35–48.
Outram, F.N., Lloyd, C.E.M., Jonczyk, J., Benskin, C.M.H., Grant, F., Perks, M.T., Deasy, C., Burke, S.P., Collins, A.L., Freer, J., Haygarth, P.M., Hiscock, K.M., Johnes, P.J., Lovett, A.L., 2014. High-frequency monitoring of nitrogen and phosphorus response in three rural catchments to the end of the 2011–2012 drought in England. Hydrol. Earth Syst. Sci. 18 (9), 3429–3448.
Patil, A., Deng, Z.-Q., 2011. Bayesian approach to estimating margin of safety for total maximum daily load development. J. Environ. Manag. 92 (3), 910–918.
Qian, S.S., Reckhow, K.H., 2007. Combining model results and monitoring data for water quality assessment. Environ. Sci. Technol. 41 (14), 5008–5013.
Rubin, D.B., 1981. The Bayesian bootstrap. Ann. Stat. 9 (1), 130–134.
Skeffington, R.A., Halliday, S.J., Wade, A.J., Bowes, M.J., Loewenthal, M., 2015. Using high-frequency water quality data to assess sampling strategies for the EU water framework directive. Hydrol. Earth Syst. Sci. 19 (5), 2491–2504.
Smith, E.P., Ye, K.Y., Hughes, C., Shabman, L., 2001. Statistical assessment of violations of water quality standards under section 303(d) of the clean water act. Environ. Sci. Technol. 35 (3), 606–612.
Solow, A.R., Gaines, A.G., 1995. An empirical bayes approach to monitoring water quality. Environmetrics 6 (1), 1–5.
UK Technical Advisory Group on the Water Framework Directive, 2013. Updated Recommendations on Phosphorus Standards for Rivers. River Basin Management (2015-2021). Final Report. August 2013, p. 12.