Using high frequency water quality data to assess sampling strategies for the EU Water Framework Directive

R. A. Skeffington[1], S. J. Halliday[1], A. J. Wade[1], M. J. Bowes[2] and M. Loewenthal[3]

[1]Dept. of Geography and Environmental Sciences, University of Reading, Reading, RG6 6DW, UK

[2]Centre for Ecology and Hydrology, Wallingford, Oxon., OX10 8BB, UK

[3]Environment Agency, Fobney Mead, Reading, RG2 0SF, UK

Correspondence to: R. A. Skeffington (r.a.skeffington@reading.ac.uk)

Abstract

The EU Water Framework Directive (WFD) requires that the ecological and chemical status of water bodies in Europe should be assessed, and action taken where possible to ensure that at least "good" quality is attained in each case by 2015. This paper is concerned with the accuracy and precision with which chemical status in rivers can be measured given certain sampling strategies, and how this can be improved. High frequency (hourly) chemical data from four rivers in southern England were subsampled to simulate different sampling strategies for four parameters used for WFD classification: dissolved phosphorus, dissolved oxygen, pH and water temperature. These data sub-sets were then used to calculate the WFD classification for each site. Monthly sampling was less precise than weekly sampling, but the effect on WFD classification depended on the closeness of the range of concentrations to the class boundaries. In some cases, monthly sampling for a year could result in the same water body being assigned to ~~one of~~ 3 or 4 of the WFD classes with 95% confidence, due to random sampling effects, whereas with weekly sampling this was 1 or 2 classes for the same cases. In the most extreme case, ~~random sampling effects could result in~~ the same water body could have been~~being~~ assigned to any of the 5 WFD quality classes. Weekly sampling considerably reduces the uncertainties compared to monthly sampling. The width of the weekly sampled confidence intervals was about 33% that of the monthly for P species and pH, about 50% for dissolved oxygen, and about 67% for water temperature. For water temperature, which is

assessed as the 98th percentile in the UK, monthly sampling biases the mean downwards by about 1°C compared to the true value, due to problems of assessing high percentiles with limited data. Low frequency measurements will generally be unsuitable for assessing standards expressed as high percentiles. Confining sampling to the working week compared to all seven days made little difference, but a modest improvement in precision could be obtained by sampling at the same time of day within a 3-hour time window, and this is recommended. For parameters with a strong diel variation, such as dissolved oxygen, the value obtained, and thus possibly the WFD classification, can depend markedly on when in the cycle the sample was taken. Specifying this in the sampling regime would be a straightforward way to improve precision, but there needs to be agreement about how best to characterise risk in different types of river. These results suggest that in some cases it will be difficult to assign accurate WFD chemical classes or to detect likely trends using current sampling regimes, even for these largely groundwater-fed rivers. A more critical approach to sampling is needed to ensure that management actions are appropriate and supported by data.

1 Introduction

The principal aim of the EU Water Framework Directive (WFD: EU, 2000) is to protect and enhance the status of aquatic ecosystems in the European Union and to prevent their further deterioration. To support this aim, the status of European waters needs to be assessed by a monitoring programme. In relation to surface (fresh) waters, the subject of this paper, the Directive states "The monitoring network shall be designed so as to provide a coherent and comprehensive overview of ecological and chemical status within each river basin and shall permit classification of water bodies into five classes…" (EU, 2000, Annex V, Section 1.3). These classes are designated, in increasing order of quality, "bad", "poor", "moderate", "good" and "high". One specific aim of the Directive is that all waters should be of at least "good" quality by the year 2015, though derogations from this are possible. If waters fail to meet this standard, then action must be taken to remedy the situation. Monitoring of waters and their assignment to quality classes is thus central to the operation of the WFD, though monitoring also has other objectives such as increasing system understanding and designing mitigation options. Because the quality of all waters varies both spatially and temporally, the representativeness of water samples is a crucial issue. There is a large literature on the design of aquatic monitoring programmes which invariably covers sampling problems. For instance, Hunt and Wilson (1986: Chapter 3) reviewed 386 references on water sampling up to 1986,

Dixon and Chiswell (1996) found about 150 up to 1995, and more recently Strobl and Robillard (2008) and Horowitz (2013) have reviewed the subject further. There is general agreement in these references of the importance of defining specific objectives for monitoring. Here the WFD is reasonably specific, defining objectives for three types of monitoring, namely surveillance monitoring to establish the present status; operational monitoring aimed at those water bodies at risk of non-compliance with objectives, and investigative monitoring for establishing the reasons for non-compliance and the magnitude of accidental pollution episodes (EU, 2000, Annex V, Section 1.3). Both the former types have "assessment of change" as a sub-objective. More detailed guidance on sampling objectives is given in various guidance documents (e.g. EU, 2009). These are the result of ~~a lot of~~ much discussion in expert committees, work groups, workshops, etc., but the diversity of surface waters in the EU means these can do little more than state the issues which should be taken into consideration, rather than giving specific guidance.

The WFD also recognises that the variability of surface waters causes problems in classifying them and in trend detection. There is a trade-off between the improved precision and accuracy obtained by sampling more frequently and the increased costs incurred. The issue of sampling frequency is extensively discussed in the reviews quoted above. The WFD states "Frequencies shall be chosen so as to achieve an acceptable level of confidence and precision" (EU, 2000 Annex V, Section 1.3.4). What is acceptable is left open, but estimates of confidence and precision have to be quoted in the River Basin Management Plans which are therefore open to public scrutiny. The WFD specifies that monitoring for physico-chemical determinands should be not less than 3 months, but leaves open the possibility that monitoring frequencies could be greater or smaller depending on expert judgement. The WFD also recognises the need to take seasonal variation into account, but not, apparently, regular variation on shorter timescales such as diurnal variation. This need is, however, well recognised in the wider literature. Hunt and Wilson (1986, p.52), for instance, state that where cyclic variations are of similar size to random variation, sampling *times* "should be chosen so that representative sampling of the cycle is achieved".

The present paper uses high frequency chemical data from four rivers in southern England to assess the accuracy and precision of the WFD classifications applied to them, and to evaluate some strategies for improving accuracy and precision. The data were subsampled to simulate different sampling frequencies, and to simulate a variety of sampling strategies. This approach

has previously been used to evaluate the influence of sampling strategy on stream concentrations (e.g. Kronvang and Bruhn, 1996; Bowes et al., 2009) and estimates of pollutant loading in rivers (e.g. Johnes, 2007; Cassidy and Jordan, 2011) but has not as far as we are aware been applied to WFD classifications. The paper also raises ~~some~~ questions about the conclusions which can legitimately be drawn from current monitoring programmes. ~~in the UK at least.~~

2 Methods

2.1 Study sites

The catchments used for this study are shown in Figure 1, and some relevant hydrological characteristics in Table 1. More detail on each site is given in the papers quoted in this section. All the rivers are affected to some extent by groundwater abstractions and transfers, a common situation in southern England. The effects of these can be clearly seen in Table 1, with reduced specific flows in the Kennet and enhanced flows in The Cut due to water imports.

The upper River Kennet (Fig. 1a) was sampled at Mildenhall, some 2 km E. of Marlborough (Palmer-Felgate et al., 2008). The catchment consists entirely of chalk of Cretaceous age. The river is predominantly groundwater-fed, with a baseflow index of 0.94 (Table 1), hence a damped hydrological response to rainfall. Land use is predominantly arable agriculture with some intensive livestock farming. The town of Marlborough (pop. c.8,400) is the only significant urban settlement. Above Marlborough sewage treatment works (STW), the Water Framework Directive classification is “good” deteriorating to “moderate” below (see http://maps.environment-agency.gov.uk/).

The River Enborne (Fig. 1b) was sampled near the catchment outlet at Brimpton (Halliday et al., 2014). Cretaceous chalk underlies the catchment and outcrops in the upper reaches, but much of the surface geology consists of impervious Tertiary clays. The Enborne is thus more hydrologically responsive than the Kennet. Land use is a mixture of grassland, arable and woodland. The WFD classification is a mixture of “good” and “moderate” depending on the reach (Fig. 1b).

The Cut (Fig 1c) was sampled near its confluence with the River Thames at Bray (Wade et al., 2012; Halliday et al., 2015). The catchment geology is predominantly London Clay and Reading Beds (Palaeocene clays and sands), giving an impermeable catchment with a baseflow index of 0.46. The catchment population is around 190,000, mostly in the large urban centres of Bracknell and Maidenhead. Improved grassland covers 30% of the catchment and 26% is classed as arable, mostly in the northern half, and woodland occupies 15%, mostly in the south. River flows are substantially increased by abstraction from the Thames for drinking water (Halliday et al., 2015) and its subsequent release through the STWs, increasing the specific runoff (Table 1). The WFD classification is mostly "poor", being "moderate" only in the upper reaches above the major conurbations. Note the river is called "The Cut", hence "The" is capitalized throughout.

The River Frome (Fig. 1d) was sampled at East Stoke (Bowes et al., 2005; 2009; 2011). It has been studied for many years as an example of a chalk stream: the geology is mostly chalk but there are other Cretaceous formations in the headwaters, principally the Gault and Upper Greensand formations in the headwaters, and sands, gravels and clays in the lower catchment. Dorchester (pop 27,000) the only significant urban centre. Land use is mainly agricultural, 47% arable, 39% grassland and 9% woodland. There is some aquaculture, mainly watercress growing, affecting the river. The WFD classification is mostly "~~moderate~~poor" but "good" in some side streams.

2.2 High frequency water sampling

Methods for collecting high frequency water chemistry data varied somewhat between rivers: they are summarized here and are described in more detail in the papers cited below. Sampling of the River Enborne is described in Wade et al. (2012) and Halliday et al. (2014). Sampling began on 1 November 2009 and finished on the 29 February 2012. Sampling frequency was hourly. A YSI 6600 multi-parameter sonde was used to measure a standard suite of parameters, including dissolved oxygen, pH and water temperature. A bankside mains-powered instrument, the Systea Micromac C, was used to make hourly measurements of total reactive phosphorus (TRP). The instrument uses the phosphomolybdenum blue complexation method on an unfiltered sample, hence TRP is an operationally defined measurement, predominantly comprised of orthophosphate (PO_4) and readily hydrolysable P species.

The River Kennet at Mildenhall was sampled from January 2004 to November 2006 and used the same instrumental setup as the Enborne, as described by Palmer-Felgate et al. (2008).

The Cut was sampled from April 2010 to February 2012 (Wade et al., 2012; Halliday et al., 2015). Sampling frequency was hourly and measurements of dissolved oxygen, pH and water temperature were made by a YSI multi-parameter sonde as above. Phosphorus species were measured using a Hach Lange Phosphax Sigma which uses phosphomolybdenum blue complexation to measure TRP as above, and also total phosphorus (TP) by acid persulphate digestion after heating to 140 °C, at a pressure of 2.5 bar (359 kPa), followed by phosphomolybdenum blue complexation. There was no filtration step in either analysis.

The River Frome at East Stoke was sampled as described by Bowes et al. (2009) between 1 February 2005 and 31st January 2006, as part of a much longer, lower frequency study (Bowes et al., 2011). Samples of river water (500 ml) were taken from approximately the mid depth of the river using an automatic water sampler (Montec Epic, model 1011). Sampling frequency varied from two to four times per day during dry periods and up to eight samples per day during periods of rainfall. A total of 1358 samples were taken over the one year monitoring period. Total phosphorus was determined in the laboratory by digesting the sample with acidic potassium persulphate in an autoclave at 121°C, then reacting with acidic ammonium molybdate reagent to produce phosphomolybdenum blue complex (Murphy and Riley, 1962). Soluble reactive phosphorus (SRP) was determined by filtering river water samples through a 0.45 μm cellulose nitrate membrane, and analysing for phosphate as above.

2.3 Statistical analysis

As the determination of the WFD status of a water is based on annual means, the datasets were divided into annual subsets: 2010 and 2011 for the Enborne; 2004 and 2005 for the Kennet; 2011 for The Cut and 2005 for the Frome. A standard set of descriptive statistics was then calculated for all the datasets, including those required for WFD determinations in the UK, which are: the mean for P and pH;~~,~~ 10th percentile for dissolved oxygen; and 98th percentile for water temperature. The analysis in this paper is restricted to these four variables. Each of the high-frequency annual datasets was then resampled using two different sampling frequencies and five different sampling strategies, to create a series of ten sampling scenarios. Sampling frequency was either monthly or weekly. Within each of these, the strategies were [with abbreviations in brackets] :

- Sampling at any time [ANY];
- Sampling on any day of the week, but restricted to normal working hours, defined as between 9:00 and 17:59 UTC, [AW9-18];
- Sampling on Monday to Friday only, and also restricted to normal working hours [MF9-18]. This is the commonest sampling approach used by the regulatory agencies;
- Sample collection on any day, but restricted to a 3 hour window between 09:00 and 11:59 UTC [AW9-12];
- Sample collection restricted to Monday to Friday and also restricted to a 3 hour window between 09:00 and 11:59 UTC [MF9-12].

Each of these re-sampling strategies was applied to each dataset using the MATLAB function *datasample* (Mathworks, 2014). This was set up to sample at random from the appropriate hourly time-series using a uniform distribution. Only one sample was taken from a given month or week, to replicate a real sampling programme. The datasets were resampled 1000 times, each generating a secondary dataset which represents a set of samples which might have been collected if the given sampling strategy had been implemented. There are thus 1000 implementations of each sampling strategy, which were used to generate statistics showing the resulting distributions of measurements and the WFD classifications which would have been obtained. In particular, the means and 95% confidence limits on the means were calculated and are used in the following analysis. The 95% confidence limits were calculated as the 2.5th and 97.5th percentiles of the distribution of means generated by the 1000 trials - this is the percentile bootstrap confidence interval (Davison and Hinkley, 1997; Section 5.3) which will simply be referred to in this paper as the confidence interval (CI).

3 Results and Discussion

Figures 2 to 5 show the means and 95% confidence intervals for four determinands – P species, dissolved oxygen, pH and water temperature – given different sampling strategies. The five bars on the left of each graph represent monthly sampling: those on the right, weekly sampling. Within each of these the sampling strategies represent (from left to right) the ANY; AW9-18; MF9-18; AW9-12; and MF9-12 sampling strategies (see previous paragraph). The boundaries between different river quality classes in the UK implementation of the WFD are also shown where appropriate. The statistics plotted are those used in the UK for the WFD:

means for pH and P species; the 10th percentile for dissolved oxygen and the 98th percentile for water temperature.

3.1 Monthly versus weekly sampling

Though it is clear a priori that weekly sampling will give a more precise estimate than monthly sampling, Figures 2 to 5 show that the magnitude of the effect varies between determinands and sites, and even between different years at the same site. The improvement in precision between monthly and weekly sampling is however generally considerable. For instance, the mean TRP in the River Kennet in 2004 for the MF9-18 sampling strategy (Fig. 2) was 103 µg P L^{-1}, with a 95% confidence interval (CI) of 38 – 251 µg P L^{-1}. For weekly sampling the corresponding CI was 74 – 138 µg P L^{-1}; mean, 102 µg P L^{-1}. As can be seen in Fig. 2, the monthly ~~phosphorus~~ TRP CI covers three WFD classes (poor, moderate and good, just missing high), whereas the weekly sampling CI is contained entirely within the moderate class. Similarly, the 95% CI for MF9-18 sampling of TRP on The Cut covers 247 µg P L^{-1} (480 – 727) whereas the corresponding 95% CI for weekly sampling is only 70 µg P L^{-1} (546-616), though all samples are in the "poor" WFD class. The width of the weekly sampled confidence intervals was about 33% that of the monthly for P species and pH (Figs 2 and 4), about 50% for dissolved oxygen (Fig. 3) and about 67% for temperature (Fig. 5). Whether the improvement of precision of weekly sampling makes any difference to the possible range of WFD classes depends on the closeness of the range of concentrations to the class boundaries. ~~f~~For instance, monthly sampling of temperature is less precise than weekly (Fig. 5) but this makes no difference to the WFD classification~~temperatures are classed as "high" whatever the sampling frequency~~ except on The Cut, whereas for P species (Fig. 2) the difference is considerable.

Another way to evaluate the effect of sampling frequency on WFD classification is to calculate the probability that a water body will be allocated to a given class in any one year. This is shown for dissolved oxygen (DO) on The Cut in Fig. 6, and TRP on the Kennet in Fig.7. Monthly sampling at any time could result in The Cut being allocated to *any* of the five WFD classes in any one year due to random sampling effects (with a 0.3% chance of "high" just visible on the diagram). The probability ~~in~~of any one year ~~of~~ being allocated to the correct class for this sampling strategy, which was "poor" according to the high frequency data, was just 47%. In contrast, weekly sampling under the same conditions allocated The Cut to three classes, with a 78% chance of "poor". These results have implications for detecting

trends in the data. For instance, ~~assuming DO concentrations stayed the same for 5 years, then~~ using the most common sampling strategy (MF9-18), the probability of the WFD class being correctly assigned to "good" is 52% for monthly sampling and 89% for weekly sampling (Fig. 6). Assuming DO concentrations stayed the same for 5 years, the probability of the classification being correct in every year is only 4% (0.52^5) with monthly sampling, whereas it is 54% (0.89^5) with weekly sampling. The potential for generating spurious "trends" in the WFD classification due to purely random sampling effects is obvious, if the sampling frequency is not great enough. For TRP on the Kennet (Fig. 7), weekly sampling always produces the correct classification of "good", whereas with monthly sampling the classification is correct only 65-75% of the time. Proportions of other classifications are "moderate", 16-20%; "poor", 5-11%; and "high", 0-2%, indicating the considerable uncertainty and wide range of possible classifications if the sampling frequency is not high enough. These considerations apply when the ~~range of measured~~confidence intervals of the mean re-sampled concentrations crosses one or more WFD class boundaries – inspection of Figs. 2-5 shows where this occurs. For some cases, e.g. pH (Fig. 4), class boundaries are not crossed and any sampling strategy always gives the same classification.

For P species, DO, and pH, the means of the monthly and weekly sampled average values are essentially the same (Figs 2-5). They are also close to the true means calculated from all the high frequency observed data – normally within 1% of the true mean value, with weekly sampling a little more precise. This shows that sampling introduces no systematic bias, and the means shown in Figs 2-5 represent the observed means. It does not follow from this that monthly and weekly sampling would generally give the same mean in *a given year* – only that the mean would be the same if it was possible to continue the sampling for long enough, effectively 1000 years in this case. For the 98th percentile water temperatures, however, the yearly means of monthly sample~~ds means~~ are clearly lower than the weekly means (Fig. 5), and sampling frequency does introduce a systematic bias. Table 2 shows the true and sampled temperatures for each river and sampling strategy, "true" being defined as the temperature calculated from all the ~~appropriate~~ measured data for the particular frequency, strategy and river. Table 2 shows that monthly sampling is underestimating water temperatures by about 1°C, sometimes more, whereas weekly sampling overestimates less consistently by about 0.1°C. These differences arise from the methods used to interpolate the 98th percentile temperature. When there are not many measurements (as in the monthly samples here), a systematic bias is likely as well as wide confidence intervals. The problems

involved in the estimation of percentiles used as water quality standards are extensively discussed by Ellis and Lacey (1980) who note that the confidence limits are likely to be very wide for high (or low) percentiles and depend markedly on the underlying distributions of the measured values. The adoption of a 98th percentile as a standard was probably intended to apply to continuously-measured temperature data where the large number of data points reduces both random error and systematic bias in estimation of the percentile. Use of a high percentile as a standard with spot measurements, which are typically fewer in number, needs to be more critically evaluated.

3.2 Diurnal sampling precision

One aim of this paper is to investigate whether restricting the times at which samples are taken would improve the precision of the estimates for the chemical variables. This can be measured by comparing the height of each bar in Figs 2-5 with the bar corresponding to unrestricted sampling ("ANY"). Table 3 shows a quantitative measure of this, i.e. 95% $CI_{(s)}$/95% $CI_{(Any)}$ expressed as a percentage, where 95% $CI_{(s)}$ is the 95% confidence interval for a particular strategy and 95% $CI_{(Any)}$ is the 95% CI for sampling at any time. Overall, restricting the sampling time improves the precision of the estimates in 71% of cases – those where it does not do so are highlighted in the Table. The most consistent improvements in precision are obtained using the 3-hour sampling strategies (AW9-12 and MF9-12) for TRP, DO and pH with weekly sampling. Monthly sampling shows a similar pattern but is less consistent. In general, the 3-hour strategies improve the precision more than the full working hours strategies (AW9-18 and MF9-18) - the average CI is 88% of unrestricted for the 9-12 strategies versus 95% for the 9-18 strategies. There is no overall difference between the precision of sampling on the AW versus the MF strategies (both 91% of unrestricted). There are differences in response between the rivers, and between the same river in different years, and between weekly and monthly sampling. In spite of these inconsistencies, however, it seems that restricting the sampling time to a 3-hour window would in general give a worthwhile improvement in precision of the estimates of the four chemical variables, and thus a more accurate estimate of the WFD class.

3.3 Different sampling strategies lead to different estimates of variables

It is clear from Figs 2-5 that different sampling strategies give different estimates for the variables being considered. Apart from the differences in water temperature between monthly

and weekly sampling referred to in Section 3.1, these are largely due to diel variations in processes affecting the variables. It is well known that DO has a strong diel variation due to the balance between photosynthesis and respiration, with low DO concentrations at night when there is no photosynthesis and high concentrations during the day when photosynthesis is active. This explains the patterns seen in Fig. 3, when the AW/MF9-18 strategies have higher DO concentrations than the average for the entire 24 hours (ANY), and the AW/MF9-12 strategies are intermediate (as DO concentrations are generally higher in the afternoon). The patterns are most pronounced on The Cut, which has a very strong diel DO cycle (Wade et al., 2012; Halliday et al., 2015) and least on the Enborne, where heavy riparian shading due to deciduous trees restricts a strong diel DO cycle to the early spring (Halliday et al., 2014). The same cycle can be seen in the pH values (Fig. 4), where higher pH in the AW/MF9-18 samples is due to lower carbonic acid concentrations during the day because of photosynthetic uptake of carbon. Likewise, the prevalence of high water temperatures is lower in the morning than for the whole day, or even the full 24 hours (Fig. 5). Phosphorus species have a less obvious pattern (Fig. 2), though there is a suggestion that MF values are slightly higher than AW values, reflecting a different outflow pattern from sewage treatment works between weekday and weekend (see Halliday et al. 2014).

These results raise the question of which sampling strategy generates the best ~~what the correct value for the measured~~ concentration~~s~~ estimates for use in WFD classifications ~~should be~~. The differences between strategies are greatest with dissolved oxygen, and can substantially affect the WFD classification. To take the most extreme example, The Cut has a classification of “poor” if sampled at any time of day (ANY), “good” if sampled at any time during working hours, and “good” but with less certainty if sampled from 9:00 to 11:59. It could be argued that “poor” is the correct classification, since organisms are exposed to conditions throughout the 24 hour period, including low DO concentrations during the night. Conversely it could be argued that since the boundaries between the WFD classes are derived in the UK from statistical associations between chemical parameters and biological quality based on sampling at conventional times, i.e. during working hours, then the correct classification is “good”. Whether “good” is a reasonable representation may depend on the diel dynamics of DO at the particular site. The Cut is a productive stream with both high photosynthesis and respiration rates – DO concentrations can fall to as little as 27% at night (Wade et al., 2012; Halliday et al., 2015). The Enborne in 2011 would also have been classified as “good”, but the magnitude of diel fluctuations is much smaller, with night-time DO concentrations no lower than 60%

(Halliday et al., 2014). Clearly The Cut is much more at risk of deleterious effects due to anoxia than the Enborne, but the daytime sampling regime does not register this difference very strongly (Fig. 3). If the issue is low night-time DO concentrations, and the measurements are available because the site is being continuously monitored, then it would seem more logical to use measurements made at night as the standard. The Cut might however be seen as an extreme case given its high STW load, and comparing the working day and anytime means and CIs on Fig 3 shows that working day sampling is a better representation of the full range of DO concentrations on the Enborne than The Cut, with the Kennet intermediate. Based on this sample of 3 rivers, it may be that daytime sampling for DO is not a good measure of risk for rivers with high respiration rates due to organic loading and/or high rates of primary production. This would need further investigation on more sites. What is not satisfactory, however, is that it is possible to obtain such widely differing WFD classifications because the sampling time is not defined. Defining a sampling time as part of the assessment procedure would be a straightforward process and reduce some of the uncertainty being discussed here, as previously suggested for The Cut by (Halliday et al., 2015).

3.4 Differences between years

The Kennet and Enborne were both assessed for two consecutive years, and it is therefore possible to obtain an indication of the extent to which chemical concentrations and WFD class assignments are stable with time. River pH was essentially the same between years (Fig. 4) but the other determinands show differences. TRP concentrations fell between 2010 and 2011 on the Enborne (Fig. 2), increasing the WFD class from “poor” to “moderate”. If non-overlapping confidence intervals are taken as a measure of a significant difference, this is a significant improvement detectable with weekly sampling, but not with monthly sampling. This is the only significant difference between years evident in the data. DO, in contrast, declined on the Enborne between the same years, and the mean WFD class fell from “high” to “good”. On the Kennet, the mean TRP stayed much the same between years, but TRP had much wider confidence intervals in 2004 than 2005, due to some especially high values. DO was lower on the Kennet in 2005 than 2004, though the WFD classification did not change. The differences between years are likely to be due to hydrological differences rather than any change in management. On the Kennet, flows in 2004 were close to the long-term average, whereas 2005 was a dry year, with flows only 62% of average (UKNRFA, 2014), leading to a higher volume-specific rate of oxygen consumption, which depresses the 10-percentile value .

On the Enborne, 2010 was a wetter year than 2011, with high and variable flows at the beginning of the period, explaining the greater variation in most concentrations in 2010 observable in Figs. 2-5. In general, the range in concentrations is determined by individual flow events which are not apparent in annually aggregated statistics, but this study illustrates that such differences do occur and will add to the variation observed.

4 Wider Discussion

This study shows that for these four rivers, the WFD class cannot be assigned with 95% confidence for a number of variables and sampling strategies. Taking the strategy most commonly used in practice, (MF9-18), the WFD class cannot be assigned for monthly sampling of phosphorus on the Enborne in 2010 and 2011 and the Kennet in 2004; dissolved oxygen on the Enborne in 2011, the Kennet in 2005 and The Cut in 2011; and water temperature on The Cut in 2011. For weekly sampling, the WFD class cannot be assigned for dissolved oxygen on the Enborne in 2011 and The Cut in 2011, and temperature on The Cut in 2011. Clearly, weekly sampling generates less ambiguity, and this matches the conclusions of Johnes (2007) that monthly sampling gave highly uncertain load estimates for a variety of British rivers, including the Enborne. In contrast, the WFD class can be assigned unambiguously for pH on all rivers and temperature in most (all “high”) and phosphorus on The Cut (“poor”), whatever the sampling strategy. Where the sample mean is close to a class boundary (as for dissolved oxygen on the Enborne 2010) then consistent assignment to a single class is unlikely, but this should not be a major issue as long as the potential size of the confidence intervals is realised when drawing conclusions. Of most concern are situations where the confidence interval crosses several classes, as with dissolved oxygen on The Cut, which can be assigned to 4 WFD classes with 95% confidence given monthly sampling as opposed to 2 or 3 classes with weekly sampling. It seems clear that if the aim is to identify WFD classes it would be better to spend limited resources on monitoring dissolved oxygen than pH in these rivers. This sort of judgement should be made in the light of technical knowledge and considering the objectives of the monitoring programme. For instance, all these rivers are fed by well-buffered calcareous groundwater and monitoring shows the pH to be well above the high/good boundary. A change of WFD status for pH is thus unlikely and occasional monitoring (e.g. twice a year) would suffice. The same considerations might apply to P concentrations on The Cut, which are unlikely to drop below “poor” in view of the high P load from sewage treatment works, except that here the WFD objectives specify that P

concentrations should be reduced in an attempt to improve the classification. Hence more frequent monitoring is justified even though the classification is likely to remain "poor" for the foreseeable future, and it becomes relevant that the 95% confidence interval for monthly sampling is around 250 µg P L^{-1} as opposed to 70 µg P L^{-1} for weekly sampling. For detection of likely trends, weekly sampling will be required. This differentiated approach to monitoring is suggested in the WFD. In practice, sampling effort may not be affected much if more frequent samples have to be taken from the same site in any case, but analytical effort may be reduced given that different determinands are analysed using different equipment.

Formatte
Roman, N

Formatte
Roman, N

Formatte
Roman, N

The results show that there is little difference between sampling Monday to Friday or during the whole week. Differences can be seen in Figs 2-5, but they are generally small in magnitude and not consistent in direction. Phosphorus is the determinand for which differences might be most likely, as the pattern of sewage treatment works output differs somewhat between weekdays and weekends (e.g. Halliday et al., 2014) but this is not apparent in Fig.2. On the other hand, restricting sampling to the three hour period between 9:00 and 11:59 leads to an improvement in precision for TRP, dissolved oxygen and pH, especially with weekly sampling (Table 3). The improvement is modest, amounting to a narrowing of the 95% confidence interval by about 13% for P, 20% for dissolved oxygen and 25% for pH, for weekly samples, but it is consistent. For monthly samples the corresponding figures are 6%, 6% and 12% respectively, and the changes are not ~~totally~~ completely consistent in direction. For 98th percentile water temperature, there is no improvement in precision from restricting sampling times. The biggest improvements are shown by the determinands with the strongest diel variation (pH and dissolved oxygen), but are apparent for P as well. These improvements in precision seem worthwhile, so restricting the sampling time to a 3-hour window seems a useful strategy as it would be easy and cheap to implement.

In the case of the 98th percentile water temperature, monthly sampling not only gives wider confidence intervals than weekly sampling, but also biases the mean temperature estimates downwards by 0.7 to 1°C compared to the "true" value, depending on sampling strategy, while weekly sampling biases the means upwards by ~~about 0.0~~up to 0.2 °C – a smaller change but still detectable given the precision of temperature measurement, and potentially significant when calculating limits~~.~~. These biases arise from the method used to estimate percentiles. Estimation of a percentile with limited data requires either an assumption about,

or assessment of, the distribution of values, or use of a distribution-free method which interpolates between values (see Ellis and Lacey, 1980). For monthly sampling (12 values) a 98th percentile cannot be interpolated, and is effectively assumed by the MATLAB function *prctile* to be the maximum sample value. For weekly sampling (52 values) the function interpolates between the two highest values - the bias introduced by this will depend on the behaviour of the extreme end of the distribution. As Ellis and Lacey (1980) state in a similar context, "even if the correct form of the distribution was known without doubt, the uncertainty in the estimate would render it virtually useless", and that calculating confidence limits for percentiles "is of limited value except in emphasizing the statistical hazards in this area". The conclusion for estimating the WFD limits is that the 98th percentile criterion should only be used where there are sufficient values to calculate a percentile, and cannot be done with spot sampled values at frequencies of weekly or greater.

One of the implications of the results in this paper is that the precision of sampling needs to be taken into account when designing mitigation strategies or other management interventions. For instance, managers should be discouraged from basing mitigation plans on noncompliance of one location in one year, in circumstances when the non-compliance could simply be due to sampling error. This will require a critical case-by-case look at each location and sampling strategy.

This study has also shown the need to define more precisely what a sample taken for WFD monitoring is meant to represent. Different WFD classifications can be obtained by regular sampling at different times of day, especially for variables with a strong diel variation, such as dissolved oxygen. This is surely an unsatisfactory situation, and it would be better to define a relatively narrow sampling time range to standardise this. There also needs to be some debate about whether a daytime sample for dissolved oxygen adequately represents the risk of anoxia occurring in all ~~cases~~types of river, given the variety of behaviour exhibited by the Enborne and The Cut. Similar considerations apply to seasonal sampling, though not covered in this paper. For instance, Rozemeijer et al. (2014) criticised the use of summer-only sampling for assessing nutrient losses from agriculture to surface-and groundwater.

Formatte
Roman, 1

This study is based on an illustrative but restricted sample of four rivers, and so must be applied with caution elsewhere. For instance, in the international context, these rivers are rather small (Table 1), though typical of rivers to which the WFD is applied in the UK. The conclusions may not be appropriate for much larger rivers – for instance, Liu et al. (2014)

used an objective method to optimise sampling frequencies on the Xiangjiang River in China, concluding that adequate characterisation could be obtained by sampling at intervals varying between every 2 months and every 6 months. The Xiangjiang River, however, is a major tributary of the Yangtze, draining an area of 85,000 km^2, and sampling less frequently than once a month may be appropriate here as larger rivers will tend to have slower responses. Naddeo et al. (2013) suggested that for some rivers in southern Italy, of about the size of the Frome in this study or slightly larger, sampling frequencies could be reduced in some cases to less than once a month without affecting the WFD classification. However, neither of these studies considered sampling frequencies greater than monthly, assuming implicitly that monthly sampling gives the "correct" value. As shown in the present paper for these English rivers, this is not necessarily the case: a conclusion also supported in the context of load estimation by the work of Johnes (2007). The other relevant characteristic of the four rivers in the present study is their high baseflow index. This will reduce the temporal variability of most variables and hence increase sampling precision for a given sampling frequency. If the present methodology was applied to flashier rivers such as those studied by Cassidy and Jordan (2011), the confidence limits observed would probably be even wider.

5 Conclusions

Overall, a more critical attitude needs to be taken over water sampling in support of the WFD in rivers such as these. For many parameters, routine monthly sampling is unlikely to be able to assign a classification accurately or to detect trends unless they are very large. However for some parameters, such as pH in this case, monthly sampling is unnecessarily frequent and possibly a waste of resources. The wide confidence intervals observed even for weekly sampling in some cases imply that there is a real possibility of identifying deleterious "trends" which do not really exist and wasting resources trying to correct them, or alternatively failing to identify genuine water quality reductions and thus not taking the necessary improvement actions. This is particularly so given differences between years which are most probably driven by varying hydrological conditions. The precision and accuracy of measurements can be improved by specifying a sampling time interval, but a realistic assessment of the uncertainty attached to any given WFD classification seems essential before taking management action.

Formatte
Roman, E

6 Acknowledgements

We would like to thank the Natural Environment Research Council for funding the monitoring of the Rivers Frome and Kennet; the Engineering and Physical Sciences Research Council for funding the LIMPIDS project (EP/G019967/1) as part of which the Enborne and The Cut were monitored; and Liz Palmer-Felgate, Emma Gozzard, Jonathan Newman, Colin Roberts, Linda Armstrong, Sarah Harman, and Heather Wickham for providing the field and laboratory support that produced the Kennet, Cut and Enborne data sets.

7 References

Bowes, M. J., Leach, D. V., and House, W. A.: Seasonal nutrient dynamics in a chalk stream: the River Frome, Dorset, UK, Sci Total Environ, 336, 225-241, http://dx.doi.org/10.1016/j.scitotenv.2004.05.026, 2005.

Bowes, M. J., Smith, J. T., and Neal, C.: The value of high-resolution nutrient monitoring: A case study of the River Frome, Dorset, UK, J Hydrol, 378, 82-96, http://dx.doi.org/10.1016/j.jhydrol.2009.09.015, 2009.

Bowes, M. J., Smith, J. T., Neal, C., Leach, D. V., Scarlett, P. M., Wickham, H. D., Harman, S. A., Armstrong, L. K., Davy-Bowker, J., Haft, M., and Davies, C. E.: Changes in water quality of the River Frome (UK) from 1965 to 2009: Is phosphorus mitigation finally working?, Sci Total Environ, 409, 3418-3430, http://dx.doi.org/10.1016/j.scitotenv.2011.04.049, 2011.

Cassidy, R., and Jordan, P.: Limitations of instantaneous water quality sampling in surface-water catchments: comparison with near-continuous phosphorus time-series data, J Hydrol, 405, 182-193, 2011.

Davison, A. C., and Hinkley, D. V.: Bootstrap Methods and their Applications, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press Cambridge, 1997.

Dixon, W., and Chiswell, B.: Review of aquatic monitoring program design, Water Res, 30, 1935-1948, http://dx.doi.org./10.1016/0043-1354(96)00087-5, 1996.

Ellis, M. A., and Lacey, R. F.: Sampling; defining the task and planning the scheme, Water Pollut Control, 79, 452-467, 1980.

EU: Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy, Official Journal of the European Communities, L327, 1-70, 2000.

EU: Common implementation strategy for the Water Framework Directive (2000/60/EC). Guidance document No. 19: guidance on surface water chemical monitoring under the Water Framework Directive, Luxembourg Technical Report 2009-025, 2009.

Halliday, S. J., Skeffington, R. A., Bowes, M. J., Gozzard, E., Newman, J. R., Loewenthal, M., Palmer-Felgate, E. J., Jarvie, H. P., and Wade, A. J.: The water quality of the River Enborne, UK: observations from high-frequency monitoring in a rural, lowland river system, Water, 6, 150-180, 2014.

Halliday, S. J., Skeffington, R. A., Wade, A. J., Bowes, M. J., Gozzard, E., Newman, J. R., Loewenthal, M., Palmer-Felgate, E. J., and Jarvie, H. P.: High-frequency water quality monitoring in an urban catchment: hydrochemical dynamics, primary production and implications for the Water Framework Directive, Hydrol Process, doi: 10.1002/hyp.10453 ~~In review,~~ 2015.

Horowitz, A. J.: A review of selected inorganic surface water quality-monitoring practices: are we really measuring what we think, and if so, are we doing it right?, Environ Sci Technol, 47, 2471-2486, 2013.

Hunt, D. T. E., and Wilson, A. L.: The chemical analysis of water: general principles and techniques, 2nd ed., Royal Society of Chemistry, London, 1986.

Johnes, P.: Uncertainties in annual riverine phosphorus load estimation: impact of load estimation methodology, sampling frequency, baseflow index and catchment population density, J Hydrol, 332, 241-258, 2007.

Kronvang, B., and Bruhn, A.: Choice of sampling strategy and estimation method for calculating nitrogen and phosphorus transport in small lowland streams, Hydrol Process, 10, 1483-1501, 1996.

Liu, Y., Zheng, B., Wang, M., Xu, Y., and Qin, Y.: Optimization of sampling frequency for routine river water quality monitoring, Science China Chemistry, 57, 772-778, 2014.

MATLAB: http://www.mathworks.co.uk/products/matlab/, access: 20 September 2014, 2014.

Murphy, J., and Riley, J.: A modified single solution method for the determination of phosphate in natural waters, Anal Chim Acta, 27, 31-36, 1962.

Naddeo, V., Scannapieco, D., Zarra, T., and Belgiorno, V.: River water quality assessment: Implementation of non-parametric tests for sampling frequency optimization, Land Use Policy, 30, 197-205, 2013.

Palmer-Felgate, E. J., Jarvie, H. P., Williams, R. J., Mortimer, R. J. G., Loewenthal, M., and Neal, C.: Phosphorus dynamics and productivity in a sewage-impacted lowland chalk stream, J Hydrol, 351, 87-97, http://dx.doi.org/10.1016/j.jhydrol.2007.11.036, 2008.

Rozemeijer, J. C., Klein, J., Broers, H. P., van Tol-Leenders, T. P., and van der Grift, B.: Water quality status and trends in agriculture-dominated headwaters; a national monitoring network for assessing the effectiveness of national and European manure legislation in The Netherlands, Environ Monitor Assess, 186, 8981-8995, 2014.

Strobl, R. O., and Robillard, P. D.: Network design for water quality monitoring of surface freshwaters: A review, J Environ Manage, 87, 639-648, http://dx.doi.org./10.1016/j.jenvman.2007.03.001, 2008.

UK National River Flow Archive: http://www.ceh.ac.uk/data/nrfa/index.html, access: 14 September, 2014.

Wade, A., Palmer-Felgate, E., Halliday, S., Skeffington, R., Loewenthal, M., Jarvie, H., Bowes, M., Greenway, G., Haswell, S., and Bell, I.: Hydrochemical processes in lowland rivers: insights from in situ, high-resolution monitoring, Hydrol Earth Syst Sc, 16, 4323-4342, 2012.

Formatte

Tables

Table 1. Some characteristics of the sampled rivers.

River	Catchment Area (km^2)	Precipitation ($mm\ yr^{-1}$)	[1]Mean flow ($m^3\ s^{-1}$)	Baseflow Index	Population (2011 census)
Kennet	220	770	c. 1.26	0.94	12 800
Enborne	148	790	1.31	0.53	18 300
The Cut	124	676	c. 1.32	0.46	190 000
Frome	414	968	6.65	0.84	46 000

Data from the UK National River flow archive http://www.ceh.ac.uk/data/nrfa/index.html unless otherwise specified. [1]Only the rivers Enborne and Frome are gauged at the sampling point. Flow in the Kennet was estimated from gauging stations located approximately 2 km upstream. Flow in The Cut was estimated from a gauging station at Binfield (gauging 50km^2 of the catchment), plus measured discharges from the sewage treatment works, plus an estimate of discharge from the lower part of the catchment based on that from the upper (Halliday et al., 2015).

Table 2. Sampled and true 98th percentile temperatures for the rivers and sampling strategies.

Temp.	Frequency	Strategy	En10	En11	Ken04	Ken05	Cut11	Mean
True	Monthly	ANY	18.01	17.05	15.20	15.80	19.08	17.03
Sampled	Monthly	ANY	17.28	16.19	14.19	14.51	18.17	16.07
Difference	Monthly	ANY	-0.73	-0.86	-1.01	-1.29	-0.91	-0.96
True	Monthly	AW9-18	18.40	17.16	15.70	16.32	20.01	17.52
Sampled	Monthly	AW9-18	17.59	16.38	14.90	15.14	18.97	16.59
Difference	Monthly	AW9-18	-0.81	-0.78	-0.80	-1.18	-1.04	-0.92
True	Monthly	MF9-18	18.36	17.74	15.50	16.30	20.01	17.58
Sampled	Monthly	MF9-18	17.53	16.38	14.80	15.21	18.89	16.56
Difference	Monthly	MF9-18	-0.83	-1.36	-0.70	-1.09	-1.12	-1.02
True	Monthly	AW9-12	17.88	16.86	14.00	14.40	18.81	16.39
Sampled	Monthly	AW9-12	17.17	16.08	13.67	13.60	17.98	15.70
Difference	Monthly	AW9-12	-0.71	-0.78	-0.33	-0.80	-0.83	-0.69
True	Monthly	MF9-12	17.79	17.38	13.90	14.40	18.98	16.49
Sampled	Monthly	MF9-12	17.14	16.12	13.54	13.65	18.04	15.70
Difference	Monthly	MF9-12	-0.65	-1.26	-0.36	-0.75	-0.94	-0.79
True	Weekly	ANY	18.01	17.05	15.20	15.80	19.08	17.03
Sampled	Weekly	ANY	18.01	17.15	15.24	15.82	19.42	17.13
Difference	Weekly	ANY	0.00	0.10	0.04	0.02	0.34	0.10
True	Weekly	AW9-18	18.40	17.16	15.70	16.32	20.01	17.52
Sampled	Weekly	AW9-18	18.39	17.29	15.84	16.40	20.16	17.62
Difference	Weekly	AW9-18	-0.01	0.13	0.14	0.08	0.15	0.10
True	Weekly	MF9-18	18.36	17.74	15.50	16.30	20.01	17.58
Sampled	Weekly	MF9-18	18.29	17.43	15.63	16.31	20.30	17.59
Difference	Weekly	MF9-18	-0.07	-0.31	0.13	0.01	0.29	0.01
True	Weekly	AW9-12	17.88	16.86	14.00	14.40	18.81	16.39
Sampled	Weekly	AW9-12	17.94	16.95	14.49	14.41	19.13	16.58
Difference	Weekly	AW9-12	0.06	0.09	0.49	0.01	0.32	0.19
True	Weekly	MF9-12	17.79	17.38	13.90	14.40	18.98	16.49
Sampled	Weekly	MF9-12	17.85	17.19	14.30	14.44	19.32	16.62
Difference	Weekly	MF9-12	0.06	-0.19	0.40	0.04	0.34	0.13

Temperatures in °C. Abbreviations for the rivers are, respectively, Enborne 2010, Enborne 2011, Kennet 2004, Kennet 2005, The Cut 2011. Strategy abbreviations: AW9-18, all week, working hours (9:00 – 17:59); MF9-18, Monday to Friday, working hours; AW9-12, all week, 9:00 to 11:59; MF9-12, Monday to Friday, 9:00 to 11:59. The final column is the mean across all the rivers.

Table 3. 95% confidence intervals for each strategy as a percentage of the 95% CI for sampling at any time.

	River	En10	En11	Ken04	Ken05	Cut11
a) TRP						
Monthly	AW9-18	91	84	97	97	116
Monthly	MF9-18	87	83	106	99	105
Monthly	AW9-12	97	93	83	82	112
Monthly	MF9-12	97	94	94	84	107
Weekly	AW9-18	79	86	97	107	96
Weekly	MF9-18	79	78	107	107	95
Weekly	AW9-12	80	89	89	86	91
Weekly	MF9-12	83	82	100	82	87
b) Dissolved Oxygen						
Monthly	AW9-18	93	102	89	102	100
Monthly	MF9-18	92	102	94	108	102
Monthly	AW9-12	91	106	85	97	85
Monthly	MF9-12	93	104	87	102	88
Weekly	AW9-18	81	100	83	107	84
Weekly	MF9-18	72	101	84	109	78
Weekly	AW9-12	82	98	63	88	70
Weekly	MF9-12	77	99	69	79	71
c) pH						
Monthly	AW9-18	105	103	89	105	94
Monthly	MF9-18	104	102	93	104	95
Monthly	AW9-12	88	99	82	95	67
Monthly	MF9-12	87	104	87	102	63
Weekly	AW9-18	98	107	80	90	90
Weekly	MF9-18	102	101	86	90	86
Weekly	AW9-12	86	94	70	82	54
Weekly	MF9-12	81	95	73	77	50
d) Temperature						
Monthly	AW9-18	109	101	107	93	91
Monthly	MF9-18	95	101	84	78	93
Monthly	AW9-12	96	93	102	70	84
Monthly	MF9-12	85	101	100	54	94
Weekly	AW9-18	98	107	87	78	100
Weekly	MF9-18	88	110	88	71	102
Weekly	AW9-12	115	104	108	70	95
Weekly	MF9-12	117	110	108	69	92

Abbreviations for the rivers are, respectively, Enborne 2010, Enborne 2011, Kennet 2004, Kennet 2005, The Cut 2011. AW9-18, all week, working hours (9:00 – 17:59); MF9-18, Monday to Friday, working hours; AW9-12, all week, 9:00 to 11:59; MF9-12, Monday to Friday, 9:00 to 11:59. Percentages greater than 100 are highlighted.

Legends to Figures

Figure 1. The four river catchments used in this study. The rivers are coloured according to their official status under the EU Water Framework Directive (WFD), as calculated by the English Environment Agency (http://maps.environment-agency.gov.uk/). Larger towns are marked by initials: M, Marlborough; Ma, Maidenhead; B, Bracknell; A, Ascot; D, Dorchester.

Figure 2. Means and 95% confidence intervals for phosphorus species generated by resampling from high frequency data. First 5 columns: monthly sampling; remaining 5: weekly sampling. Red bars – at any date or time; green, working hours (9:00 – 17:59) only; blue, 9:00 to 11:59 only. AW – on any day of the week; MF – Monday to Friday only. Horizontal lines represent Water Framework Directive class boundaries where applicable, from the bottom: High/Good; Good/Moderate; Moderate/Poor. Note different scale for The Cut. P species are defined in Section 2.2: TRP – total reactive phosphorus; SRP, soluble reactive phosphorus; TP, total phosphorus.

Figure 3. Mean 10th percentiles and 95% confidence intervals for dissolved oxygen generated by resampling from high frequency data. First 5 columns: monthly sampling; remaining 5: weekly sampling. Horizontal lines represent Water Framework Directive class boundaries – from the top: High/Good; Good/Moderate; Moderate/Poor; Poor/Bad.

Figure 4. Means and 95% confidence intervals for pH generated by resampling from high frequency data. First 5 columns: monthly sampling; remaining 5: weekly sampling. The WFD class is uniformly "high" (pH > 6.60).

Figure 5. Mean 98th percentiles and 95% confidence intervals for water temperature generated by resampling from high frequency data. First 5 columns: monthly sampling; remaining 5: weekly sampling. Horizontal line represents Waters Framework Directive class boundary between "high" (<20°C) and "good".

Figure 6. The probability that sampling dissolved oxygen on The Cut for one year would put the river into a given WFD class, a) monthly sampling, b) weekly sampling. Strategy labels:

Any- at any time; AW9-18, all week, working hours (9:00 – 17:59); MF9-18, Monday to Friday, working hours; AW9-12, all week, 9:00 to 11:59; MF9-12, Monday to Friday, 9:00 to 11:59.

Figure 7. The probability that sampling TRP on the River Kennet for one year would put the river into a given WFD class, a) monthly sampling, b) weekly sampling. Strategy labels: Any- at any time; AW9-18, all week, working hours (9:00 – 17:59); MF9-18, Monday to Friday, working hours; AW9-12, all week, 9:00 to 11:59; MF9-12, Monday to Friday, 9:00 to 11:59.

Anonymous Referee #1

Received and published: 23 February 2015

General comments

This manuscript addresses the consequences of different sampling strategies in relation to classifying the ecological and chemical status of rivers as stated by the Water Framework Directives (WFD). High frequency data of four different parameters (dissolved phosphorus, dissolved oxygen , PH and water temperature) were analysed for four different rivers. Based on the data new sub-series were constructed, to simulate different sampling frequencies and sampling strategies. It was found that both sampling frequencies and strategies can highly influence the process of assigning the streams to the appropriate WFD classes.

The manuscript addresses some very important issues regarding the challenge of balancing sampling frequency and strategy with the desired precision and representativeness as well as the cost of obtaining them. Generally the manuscript is well written and clearly structured, and only minor corrections are suggested below. Hence, it is found that the manuscript addresses main scientific questions relevant for HESS, and that the paper is of general interest for the readership of HESS, specifically relevant for the discussion of monitoring strategies in surface waters to assist EU directives.

Specific comments

In the paper you address the sampling frequencies where you are simulating 1000 years with your data, which gives a very good data set for conducting the analysis of sampling frequency and strategy. However, the analysis of the temporal aspect is also interesting, and it could have been interesting to see some duration curves as well. For instance, how many “years” (simulated re-sampled series) are needed before the mean values seen for TP in fig. 2 are obtained? (assuming steady conditions represented by the limited years of data).This is of interest in terms of classifying streams, as not only sampling frequency but also the length of the period measured plays a role for obtaining a representative picture of the status of the stream. I do recognize that it would extend the focus of this paper, and I find the paper comprehensive enough by just focusing on the sampling strategies. However, this possibility/issue could briefly be mentioned in the discussion.

This is an interesting and relevant question, but our approach is not well-suited to answering it. Because each “year” is a random sample from the existing data, the speed with which the cumulative mean for 1000 realisations is approached is heavily dependent on purely random events in the initial few simulations. Trials show that if there is a high or low percentile value in the first few simulations, then it takes some years for it to be averaged out, and the overall mean is not approximately attained for 50-100 years. If the first few simulations happen to be close to the long-term mean, then approximate convergence is attained more quickly. In these circumstances it would be misleading to draw conclusions about differences between determinands and rivers, as these are likely to be purely random variations.

P. 1280, line 14: You write “to one of 3 or 4 WFD classes”, do you mean that the water body is classified to belong to 3-4 different classes dependent on the sampling strategy? Could you please rephrase the sentence.

This is not the essential point, so we have clarified the text.

P. 1282, line 6: Could you please replace “a lot of” with a more formal phrase?

Done.

P. 1283, line 6-7. I would prefer that the last sentence is deleted or maybe better, rewritten to be more specific, leaving out for instance the words “some” and “at least”.

We have re-written in a more acceptable form.

P. 1288, l. 7: You already defined CI, so no need to repeat it.

We could omit – but sometimes it is helpful to the reader to repeat abbreviations at the start of the results/ discussion section for those who have not read the Methods thoroughly.

P. 1288, l. 12-13: I assume mean TP, rather than P?

TRP – text altered.

P. 1288, l. 12-13: I find it confusing the way you refer to the CI, could you maybe just write the interval itself, it is not necessary to specify the difference between max and min in the CI.

We think that the minimum, maximum and range values are all important. Not quoting the range would require the reader to do mental arithmetic on each set of values.

P. 1288, l. 17-20. I do not find it clear what is meant. I assume you are still referring to temperature, since we just saw that for instance for TP it makes a huge difference to the possible WFD going from monthly to weekly sampling? I suggest that you rephrase this paragraph.

We have tried to explain this better.

P. 1289, line 3: You refer to a period of 5 years, but it is not completely clear, how this data series of 5 years has been created? I assume it is by letting for instance sampling every Monday represent “one year”, and so on, giving you five datasets.

However, it is not clear from the text. Could you please specify this, so that is becomes clear.

It was not done like that. We have been more explicit in the revised MS.

P. 1289, line 7-9: I am not convinced that I understand what is meant in this paragraph. As I understand it, you show the range of yearly average values in your figures 2 – 5. What do you then mean by "the range of measured concentrations"? Do you refer to the different yearly averages based on your constructed data series, or do you refer to the variability in concentrations during the year (the original dataset)? You refer to fig. 2-5 for the reader to see where "the range of measured concentrations crosses one or more WFD class boundaries", but as I understand it, it is not the "range of measured concentrations", but the calculated yearly averages you refer to, or do I misunderstand?

Could you please clarify this in the text.

Referee 1 is right to point out that we are not plotting the "range of measured concentrations" in Figs 2-5, but the means and confidence intervals of the re-sampled data. We have altered the text accordingly.

P. 1289, line 12: You refer to the mean of all yearly averages of monthly and weekly sampled values, right? If yes, could you then clarify this in the text to avoid misunderstandings?

That is right. We have clarified.

P. 1289, line 20: Again, I suppose you refer to the yearly mean of the monthly concentrations, right? I would prefer that you wrote: "yearly mean of monthly samples". I suggest that this is specified throughout the manuscript, to avoid misunderstandings. Also in the figure captions, it could be specified that it is yearly means, percentiles and confidence intervals.

Right again. The text has been altered as suggested.

P. 1289, line 23: What does "appropriate" refer to? The mean value of all measured data over entire measuring period, or? Please clarify in the text.

"Appropriate" means all the data for the particular frequency, strategy and river. We have now spelt this out in the text.

P. 1291, line 22-23: Could you rephrase this sentence, as the correct value for a measured concentration must obviously be the value that is measured (if correctly measured)?

Rephrased to clarify.

P 1295, line 19: You write “biases the means upwards by about 0.0 to 0.2 degrees Celsius”, could you rephrase, for instance: “biases the means upwards by up to 0.2 degrees Celsius”, (0.0 is not an upwards bias).

Done.

P 1295, line 20. Is 0.2 degrees Celsius a significant change? I guess that depends on the measurement precision as well? Could you please comment on this?

Done.

P. 1296, line 12: What is meant by “all cases”? Is it referring to the diurnal variability? Please specify this in the text.

Clarified.

P. 1297, line 5: “most likely” would be appropriate to add to the last sentence.

Added “probably”.

P. 1309, figure 6: You show this informative probability plot only for dissolved oxygen. Is there a specific reason why you did not do the same for phosphorus? I think it could be interesting to have phosphorus plotted in the same way.

We have added this as Fig. 7, together with some text.

<u>Technical corrections</u>

P. 1281, line 1: Write river in plural. **“river” is better grammatically.**

P. 1286, line 8: Word missing between “means” and “these**”? We cannot see anything missing but have re-punctuated to clarify.**

P. 1288, line 26: replace “in” with “of”, delete “of” before “being”. **Done**

P. 1295, line 10: Please delete “totally” or replace with more formal word. **Done**

Figure 3: Add comma after “Mean”. **? Cannot see anywhere where this would make sense.**

Figure 4: Could the different limits between WFD classes be mentioned in the figure caption? **Done**

Figure 5: Add comma after “Mean”. **? See comment for Fig. 3**

Anonymous Referee #2

Received and published: 31 March 2015

This paper applies high-frequency datasets (dissolved P, dissolved oxygen, pH, and temperature) from four rivers to assess the optimal (low-frequency) sampling strategy for WFD-related compliance testing (or WFD classification). The paper is in a good shape. It's well written and easy to follow. The message of the high uncertainty in the WFD-classification (and trend detection) is important.

I have 3 general comments or suggestions for the paper:

1: The paper focuses on monitoring for WFD-classification. This is a good focus which enables to quantify uncertainties. However, it may also suggest that WFD-classification is the only thing that water quality monitoring is aiming at. However, in addition to compliance testing, water quality monitoring also plays a role in the selection and evaluation of mitigation options, which requires system knowledge. These broader monitoring objectives and the reasoning behind the focus on WFD-classification could be added to the introduction.

Some text added along these lines.

2: The paper proposes different sampling strategies for different parameters. In practice however, all these parameters are usually coupled; they are analyzed for the same samples. Therefore, a distinct strategy for each solute may not be realistic. This could be mentioned in the discussion.

Added a little discussion

3: The uncertainty of the WFD-classification for a specific location in a specific year is an important message and conclusion. Can you advise for water quality managers how to deal with this uncertainty? For example: do not base mitigation plans on noncompliance of 1 location in 1 year. The same issue was recently addressed in the discussion of Rozemeijer et al., 2004: Water quality status and trends in agriculture dominated headwaters; a national monitoring network for assessing the effectiveness of national and European manure legislation in The Netherlands,. Environ. Mon. Assess 186, 8981-8995.

This is a good point – short paragraph added to amplify, and reference added.

Some minor comments/suggestions:

#p1-L26-28: Introduce this sentence with e.g.: Weekly sampling considerably reduces the uncertainties compared to monthly sampling.

Done.

#p2, L10-11: 'A more critical approach to sampling: : :' This advice is a bit vague.

This is not meant to be specific advice. The implication is that the points raised in the Abstract and the paper discussion need to be considered

p5, L14: The WFD-classification of the River Frome seems to be “poor” (orange) in figure 1.

True. “Poor” is the correct status – text changed.

p8, L24-26: This sentence could be used in the summary/conclusions to support the conclusion of the high uncertainties in the classification. Maybe a table with these percentages for the other rivers/parameters could be added?

As a matter of style, we think the Conclusions should be bold statements and not include supporting evidence which is extensively discussed in the rest of the paper. We have added an extra Figure at the suggestion of Referee 1 which covers most of the additional uncertainties, so would not want to duplicate this in an extra table.

#p9-L27-29: Another important message. Maybe also add this statement to the conclusions/ abstract?

Added to the Abstract – already in the Discussion

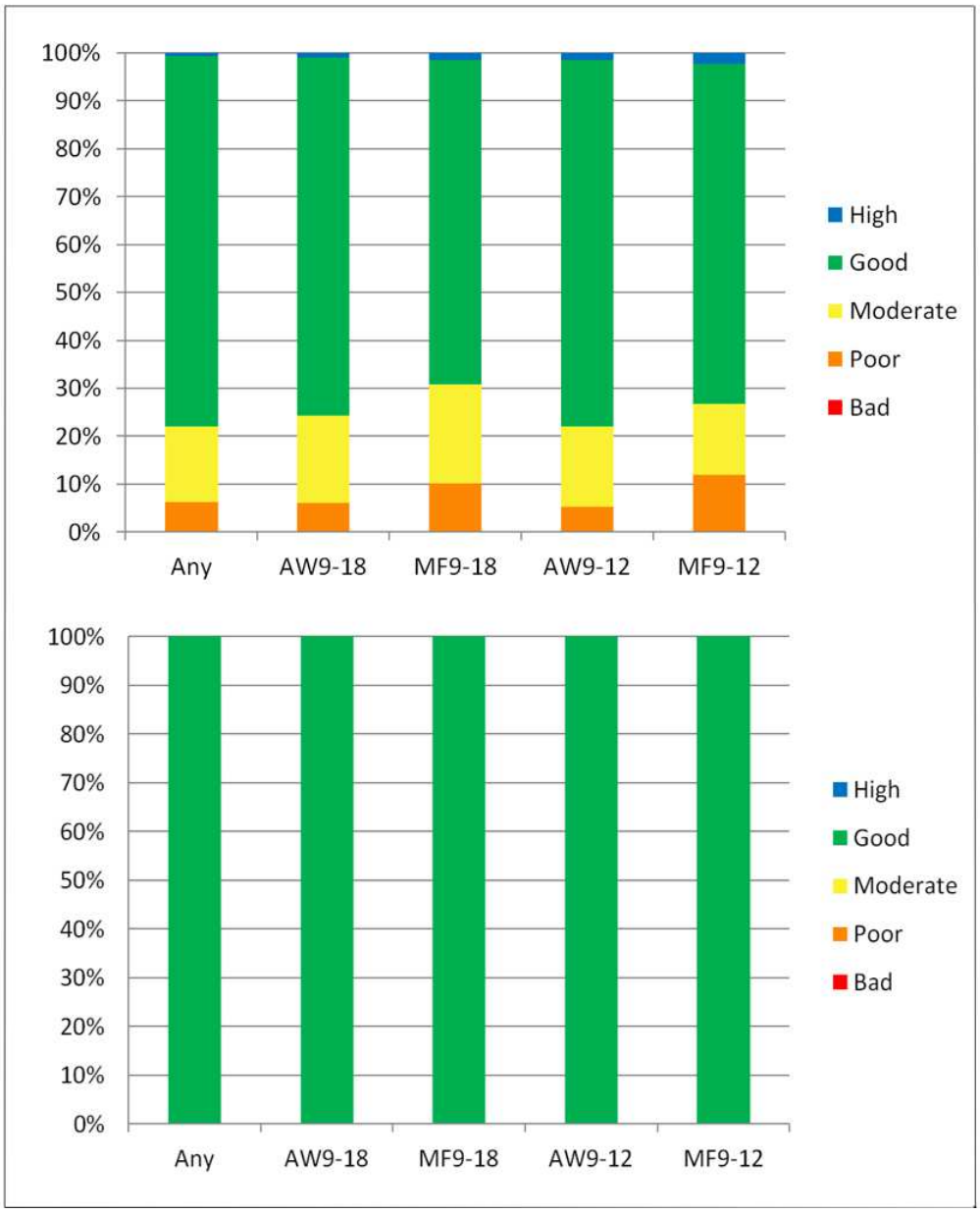
#p10-L14-17: Can you explain why a 3-hours sampling window improves the precision?

This has obviously got something to do with the diel variation, but any reason would have to be speculative – we would sooner leave this as an empirical observation.

#p12-L14-18: You may add the explanation why a low flow leads to lower DO. Less dilution of STW-effluent? Larger biological DO-consumption?

Done. This is a common pattern in these rivers. The second is partly a consequence of the first, but DO consumption from respiration of primary production plays a part as well.

New Fig. 7


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Any
AW9-18
MF9-18
AW9-12
MF9-12
High
Good
Moderate
Poor
Bad
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Any
AW9-18
MF9-18
AW9-12
MF9-12
High
Good
Moderate
Poor
Bad