Community Data – where does the value lie? Assessing confidence limits of community collected water quality data.

EMILY NICHOLSON, JANE RYAN[#] & DAVID HODGKINS[*]
Waterwatch Victoria, c/o Department of Natural Resources & Environment, 1/250 Victoria Pde, East Melbourne 3002 (jane.f.ryan@nre.vic.gov.au)
* Goulburn Broken Waterwatch, c/o Goulburn Valley Water, PO Box 85, Shepparton 3630 (davidh@gvwater.vic.gov.au)

Abstract

Waterwatch is a national community-based monitoring network that aims to involve community groups and individuals in the protection and management of waterways. Waterwatch Victoria has the dual objectives of catchment education and water quality monitoring. The educational outcomes are evident, with the Waterwatch programme facilitating over 9,000 students to monitor more than 2000 sites in waterways in 2000. This paper aims to assess the scientific value of community-collected data, through examining differences between Waterwatch data and professionally collected data. The study looked at all aspects of volunteer data collection, including data confidence protocols, equipment, and data analysis. All professional data was collated by the Victorian Water Quality Monitoring Network (VWQMN). The parameters examined in this study were turbidity, electrical conductivity (EC), pH and total phosphorus.

The level of agreement between community-collected data and professional data varied temporally and spatially. Waterwatch data for EC and pH appeared to be very similar to professionally collected data. Equipment used by Waterwatch volunteers for turbidity and total phosphorus appeared to be limited in accuracy to moderate ranges. Overall the VWQMN professionally collected data showed less variance, suggesting greater variability, potentially due to inaccuracies, in volunteer collected data.

Keywords

Community monitoring, water quality, data confidence, Waterwatch, environmental monitoring,

Introduction

The importance of community involvement in environmental issues and the educative value of environmental monitoring by citizens have long been recognised (e.g. Dvornich *et al.*, 1995; Lathrop & Markowitz, 1995; Saris *et al.*, 1996; Barrett *et al.*, 1999; Bjorkland & Pringle, 2001). Freshwater monitoring in particular has long been seen as means of educating the general public and involving them in the care of local waterways (Lathrop & Markowitz, 1995; Bjorkland & Pringle, 2001; Middleton, 2001). The scientific value of the data produced by such participation has however rarely been assessed (with the notable exception of Reynoldson *et al.* (1986), Heiskary *et al.* (1994), Mattson *et al.* (1994), and Fore *et al.* (2001)).

Boulton (1999) described the important role that non-scientists can play in monitoring river health, and suggested that ideal river indicators would be quick, cheap and easy to use, and the data readily assessed and understood. The current Victorian model of stream health assessment, the Index of Stream Condition, developed by (Ladson *et al.*, 1999) comprises five components of stream condition. One element in particular corresponds with Boulton's (1999) description: water quality based on an assessment of phosphorus, turbidity, electrical conductivity and pH is easy to measure, and can be done cheaply and quickly by untrained field workers. These four parameters are amongst those commonly measured by community monitoring groups such as Waterwatch Victoria.

However Boulton (1999) iterates the oft-repeated scientific line, when relating to volunteer-collected data, that “the data must be of high quality” to be useful in measuring river health. While suggesting that non-scientists have a role in monitoring waterway health, he does not discuss the role of the non-scientific community other than with reference to how ecologists can relate their findings to them.

Volunteer does not equate to unskilled, as training and protocols in community based water quality monitoring programmes are carefully planned and written (Lathrop & Markowitz, 1995); Waterwatch Victoria has a protocol and methods manual (Hodgkins & McCoy, 1999), a data quality assurance manual (McCoy, 2000), and runs several training courses per year. In the USA, physical and chemical data collected by volunteers has been used in water resource quality assessment by the US Environmental Protection Agency (USEPA) (Lathrop & Markowitz, 1995). Few quantitative studies on the quality of volunteer-collected physical and chemical data are however available in scientific literature (Heiskary *et al.*, 1994; Mattson *et al.*, 1994), although several have been published on the success of volunteer monitoring of macro-invertebrates in waterways (Reynoldson *et al.*, 1986; Lathrop & Markowitz, 1995; Fore *et al.*, 2001) and of birds in rural and urban landscapes (e.g. Saris *et al.*, 1996; Barrett *et al.*, 1999).

The monitoring activities and methods of community groups often differ from those of professional scientists, due to more complex and multiple aims of monitoring programmes. While the principle aim of scientific monitoring is data accuracy, community groups seek primarily to educate and inform the volunteers involved (Heiskary *et al.*, 1994; Kishbaugh, 1994; Bjorkland & Pringle, 2001; Hudson, 2001). The accuracy of equipment used by community groups is constrained by cost of purchase and use (such as tubes or chemical products). A further constraint is the educative value of the equipment: understanding the mechanisms involved in measuring a parameter can result in a greater comprehension of the issues involved. For example, a turbidity tube allows the user to see directly how turbid the water is, and thus better understand the meaning of higher levels of turbidity than by simply recording the electronic reading from a meter.

Both professional and volunteer monitoring groups have limited resources. While the equipment of choice for a community group may constrain the accuracy of the data, the number of sites that can be monitored is restricted only by the number of volunteers involved. By contrast, scientific data collection is limited not in equipment but in the number of sites that can be monitored. For example, in Victoria, community groups monitor water quality in over 2000 waterway sites, whereas professional scientists monitor only 120 sites (VWQMN, 2001; Waterwatch Victoria, 2001). The greater site coverage provided by community groups has the potential to broaden water quality monitoring of waterways. The scientific value of the data however needs to be assessed.

Case study:

Waterwatch Victoria (www.vic.waterwatch.org.au) is part of a national community-based monitoring network with the goal of involving community groups and individuals in the protection and management of waterways. During the early 1990's, it became evident that a series of community water quality monitoring programmes were being developed independently by the states. The national Waterwatch programme was formally established in 1992 with the federal government's allocation of $2.9 million over three years. The programme raises awareness and understanding about water and catchment health issues to stimulate activities to achieve the vision of 'healthy waterways'. The programme is run in all 10 catchment regions in Victoria, with Commonwealth funding being matched at a regional level by local natural resource management agencies. These networks of volunteer water quality monitors comprise community groups and schools acting in partnership with the resource management agencies in their catchment to monitor water quality and share data throughout the catchment. All parties are keen to detect environmental problems and develop appropriate action jointly, and to assess and review catchment-based management plans and activities.

The programme has grown from around 200 groups participating in 1994, to over 800 groups in 2000. In 2001, there were over 12,000 Waterwatch participants monitoring over 2000 sites throughout Victoria.

Four parameters for assessing water quality (Ladson *et al.*, 1999) were examined in this study: turbidity, electrical conductivity (EC), pH and total phosphorus. Three of the four parameters (turbidity, EC and pH) are those most commonly measured by Waterwatch. In addition to these field measurements, this study assessed the success of a laboratory-based parameter. Few Waterwatch groups measure total phosphorus, due to lack of access to the laboratory equipment necessary for analysis. The groups whose data was used in this study had regular access to the equipment required to measure total phosphorus. As a lab-based parameter, total phosphorus is more complex for the non-scientist than basic field measures such as turbidity, but is common in scientific monitoring by professionals.

The Victorian Water Quality Monitoring Network (VWQMN) (www.vicwaterdata.net) was established by the Australian Water Resources Council in 1975 to collect water quality data for all major streams and their tributaries in Victoria. In 2001, 170 river and stream sites were monitored for a variety of physico-chemical parameters, including pH, dissolved oxygen, salinity, turbidity, total and reactive phosphorus. The data is collected by government and private-sector scientists, and analysed either in the field (turbidity, EC, and pH) or in accredited scientific laboratories (total phosphorus).

This study aimed to compare Waterwatch water quality data, collected by volunteers and Waterwatch co-ordinators from the Goulburn Broken catchment in northern Victoria, with parallel data collected by VWQMN scientists. Previous evaluation of Waterwatch data confidence comprised a subjective analysis of initial trials (Clark, 1995). Only recently have the historical datasets been collated to allow a more complex analysis. In addition, water quality monitoring by community groups has since evolved, with a substantial increase in resources, co-ordination and participation. The study sought to look beyond the educational merits of the programme and assess the scientific value of community monitoring, and to examine how the potential role Waterwatch can expand.

Methods

Protocol and equipment

The equipment and the protocols used by Waterwatch and VWQMN were compared, as defined in Waterwatch Victoria's methods and data confidence manuals (Hodgkins & McCoy, 1999; McCoy, 2000), and in the VWQMN manual (DNRE, 1999), both available on the internet from the groups' websites.

Water quality data

In order to compare the data generated by Waterwatch with professionally collected data from VWQMN, data was collated from five sites at which both Waterwatch groups and VWQMN have monitored for some time. The sites were Broken Creek at Katamatite (hereafter referred to as Katamatite), Broken River at Gowangardie (Gowangardie), Goulburn River at Balaclava Road in Shepparton (Shepparton), Goulburn River at Murchison (Murchison), Acheron River at Taggerty (Acheron), and the Delatite River at Pirie's Bridge and at Tonga's Bridge (Delatite). The sites were selected to show a range of water quality levels within the catchment, in order to assess the effectiveness of Waterwatch sampling with varying water quality. All tables and figures are listed from highest turbidity (Katamatite) to lowest (Delatite). Each site had between 1 and 5 years of data, from 1995 to 2000. Each year was treated separately to minimise the effects of varying rainfall and other sources of variation. The number of times Waterwatch monitored varied on a year to year basis. The number of parameters also varied, with sometimes as little as one parameter being measured on a single day.

A minimum of four Waterwatch data points per year was required to include a year's data in the study. VWQMN usually collected data once a month. In cases where Waterwatch data were not collected for several months (i.e. more than 3 consecutive months), the VWQMN data from the same period were excluded from comparison, and *vice versa*. Data from periods when data was collected using inappropriate equipment or equipment that is no longer used were also excluded, in order to make the results relevant to current Waterwatch sampling procedure.

T-tests were performed to compare the means of Waterwatch and VWQMN data for each site per year. F-tests were also executed to assess heterogeneity of variance in the datasets.

The statistical power of the T-tests and F-tests was also calculated, using the software Gpower (Faul & Erdfelder, 1992) for T-tests and πface (Lenth, 1996) for F-tests. Statistical power is defined as the probability of detecting a difference, or change, when a difference truly exists (Thomas, 1997). Statistical power is related to β, the likelihood of making a type II error, or of not detecting a difference where it exists (Power = 1-β) (Thomas, 1997; Goudey, 1999). Low statistical power means that only extreme differences will be statistically significant (Thomas, 1997; Fore *et al.*, 2001); in this case, there must be a very large disparity between the volunteer and professional data in order to detect a difference.

As Waterwatch sample sizes per year were expected to be relatively low in the majority of cases ($n<12$), and that statistical power was also likely to be low, a visual analysis of the difference between the means, with 95% confidence intervals, was performed in addition to formal statistical analyses. This allowed trends in difference between the datasets to be assessed and analysed. The difference between the means of the datasets, and the confidence interval, were shown as a percentage of the VWQMN mean, to demonstrate Waterwatch data similarity to or dissimilarity from VWQMN data. The confidence intervals were approximate, and were determined as double the combined standard error for the two data sets. The standard error was calculated using a pooled standard deviation, s_{pooled} (Sokal & Rohlf, 1981):

$$s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

where s_1 and s_2 are the standard deviations for datasets 1 and 2 respectively, and n_1 and n_2 are the sample sizes for datasets 1 and 2 respectively.

Results

Protocol and equipment

The data collection protocols followed by Waterwatch and VWQMN were the same where the equipment used were comparable. Where the equipment differed, the protocols were equivalent in terms of such areas as cleanliness, where samples were to be collected in the stream, and calibration. Disparity between Waterwatch and VWQMN methods and protocols occurred where the equipment used differed greatly.

While VWQMN used a turbidity meter, Waterwatch groups used either a turbidity meter or, in the majority of cases, turbidity tubes. Turbidity tubes aid in making a subjective assessment of colour, using lines in log scale on the tube wall to estimate the limiting depth of water to which markings on the bottom of the tube can be seen. The range of the turbidity tube is 10 to 400 NTU, but is estimated with increasing difficulty above 200, due to the proximity of the lines. As the measure is subjective, the tubes cannot be calibrated.

Waterwatch equipment for measuring total phosphorus in the laboratory differed in detection limit. VWQMN used equipment able to detect phosphorus concentration to 3 decimal points, with a minimum concentration of 0.005 mg/L (DNRE, 1999). Waterwatch used the Hach DR 700, accurate to 2 decimal points, with a minimum of 0.01 mg/L.

Details of the methods and equipment can be found in (DNRE, 1999; Hodgkins & McCoy, 1999; McCoy, 2000), and on the Waterwatch and VWQMN websites.

Water quality data

Three of the 23 annual datasets recorded showed statistically significant different results for turbidity between Waterwatch data and VWQMN data ($P<0.05$) (table 1a.). These datasets were all from the same site. Two of these annual datasets had larger

Waterwatch sample sizes than many of the other sites, and thus higher statistical power and likelihood of detecting a difference if it existed. There were however sites with higher Waterwatch sample sizes (up to 31 at Acheron) with very similar means showing no significant difference, suggesting that many non-significant results were not simply a matter of low statistical power (see below for further discussion of statistical power). The third significant result had a similar Waterwatch sample size to other datasets, but relatively low variability.

In the majority of cases, the annual mean for turbidity from the Waterwatch data was less than 50% different from the VWQMN mean (figure 1a.). The 95% confidence intervals of nearly half (48%) of the datasets included no difference between means. For two datasets, both at Delatite, the Waterwatch means were more than 100% higher than VWQMN means, though the 95% confidence intervals still reached below 100% difference. Due to the high variability and low Waterwatch sample size (7 and 9), these results were, despite the great difference, not statistically significant (P=0.296, P=0.351). In almost all cases, Waterwatch means were higher than the VWQMN means.

Waterwatch turbidity measurement appeared to be most inaccurate at the highest and lowest turbidity levels (figure 1a.). The three significantly different results were from the site with the highest turbidity levels (Katamatite); the greatest difference between the means existed at Delatite, where the turbidity was lowest (less than 10 NTU, compared with VWQMN means of up to and above 100 NTU at other sites [table 1a.]).

The middle range of turbidity showed, with VWQMN means of between approximately 10 and 50 NTU, showed greatest correspondence between Waterwatch and VWQMN data. The sites in this range were Gowangardie, Shepparton, Murchison and Acheron. Eleven of 14 cases were within 40% difference, with the confidence intervals reaching zero difference in 9 and less than 3% difference in the remaining two. The 3 annual datasets from the sites of medium levels of turbidity that exhibited the greatest difference (Gowangardie 1999, and Shepparton 1998 and 2000) each had one or two unusually high samples that had great influence on the mean. For example, the Waterwatch turbidity values range for Shepparton 1998 range from 18 to 74, with one measure of 150 NTU. When this data point is included, the mean is 46.9, with a variance of over 1700, as opposed to a mean of 34, and variance of 306, when excluded. While excluding the value makes Waterwatch dataset more similar to the VWQMN range (9 to 51), mean (25.3) and variance (208), the reading of 150 was not inconceivably high; VWQMN data over the 5 years at Shepparton ranged from 9 to 135, and the high sample may have been due to a large rain event. It may however have been an extraneous sample, taken to assess the turbidity of the river when at its most extreme: all other samples in 1998 were taken between the 15th and 21st of the month, whereas the unusually high sample was measured on the 3rd.

Table 1: t-tests of comparing Waterwatch annual data per site with VWQMN annual data per site

A. Turbidity

Site		**Year**	Waterwatch			VWQMN			s_{pooled}	*P*
			n	**μ**	**s**	*n*	**μ**	**s**		
Broken Creek, Katamatite	*	1997	21	181.8	88.8	8	109.3	40.7	79.2	0.006
	§	1998	31	145.5	92.5	12	134.8	47.4	82.8	0.620
	§	1999	23	177.4	95.0	11	106.5	43.7	79.0	0.005
	#	2000	8	211.4	62.3	12	138.6	35.0	39.9	0.013
Broken River, Gowangardie	#	1997	11	34.0	20.4	12	41.2	35.5	29.3	0.555
	#	1998	9	68.1	32.7	12	52.6	35.4	34.2	0.313
	§	1999	7	74.6	50.9	8	39.6	12.8	35.9	0.123
	#	2000	7	65.1	56.7	12	46.6	20.8	37.6	0.432
Goulburn River, Balaclava Rd, Shepparton	*	1996	24	37.5	14.1	10	33.8	15.6	14.5	0.526
	#	1997	10	34.9	10.7	12	31.6	22.0	17.8	0.653
	#	1998	9	46.9	42.0	12	25.3	14.4	29.4	0.172
	§	1999	8	51.3	26.6	12	43.7	39.2	34.8	0.612
	#	2000	7	49.4	29.9	12	30.4	15.1	21.5	0.155
Goulburn River, Murchison	#	1997	7	17.1	7.3	7	14.0	10.7	9.2	0.535
	#	1998	12	16.8	13.3	6	12.8	9.0	12.1	0.474
	§	1999	11	16.9	10.2	8	13.4	9.6	9.9	0.455
	#	2000	8	22.6	17.0	9	18.0	14.5	15.7	0.559
Acheron, Taggerty	*	2000	31	13.6	15.0	12	13.9	13.3	11.5	0.951
Delatite River, Pirie's Bridge & Tonga's Bridge	*	1995	7	14.0	9.6	12	7.2	6.0	7.5	0.129
	*	1996	9	17.9	27.8	12	5.5	3.5	18.2	0.296
	§	1997	10	4.8	5.2	12	3.1	0.6	3.5	0.338
	§	1998	7	11.3	21.6	12	3.0	0.8	12.8	0.351
	#	2000	5	13.2	14.5	12	8.6	10.6	11.7	0.547

* Turbidity measured predominantly with turbidity tubes
\# Turbidity measured predominantly with turbidity meter
§ Turbidity measured with turbidity tubes and turbidity meter approximately equal number of times throughout the year

The Waterwatch and VWQMN variance for turbidity was statistically significantly different in 48% of cases (table 2a.), despite low statistical power. In all but one of the cases where a significant difference was detected, Waterwatch data had the higher variance. In 70% of cases, Waterwatch variance was higher than VWQMN variance.

Waterwatch data for turbidity were measured with either turbidity tubes or turbidity meters. The predominant means of measuring turbidity was noted (table 1a.), though for a substantial proportion of the data, the equipment used was not identified. There did not appear to be a substantial difference between the performance of the two pieces of equipment, although the results may be confounded with other factors, such as experience of the user and the level of turbidity.

B. Electrical Conductivity

Site	Year	Waterwatch			VWQMN			s_{pooled}	*P*
		n	μ	s	*n*	μ	s		
Broken Creek, Katamatite	1997	6	195.0	61.6	7	177.4	41.4	51.5	0.139
	1998	31	174.8	70.4	12	233.3	118	85.9	0.130
	1999	23	153.7	68.6	12	177.5	42.9	61.2	0.217
	2000	8	191.3	53.3	12	188.3	52.2	52.6	0.905
Broken River, Gowangardie	1997	10	197.0	39.7	11	193.6	43.4	41.7	0.855
	1998	9	222.2	52.9	12	238.3	55.4	54.4	0.510
	1999	7	156.7	55.5	12	174.2	41.7	47.0	0.487
	2000	7	134.6	32.1	12	165.8	40.3	37.6	0.083
Goulburn River, Balaclava Rd, Shepparton	1996	12	177.5	55.3	11	164.9	52.1	53.8	0.581
	1997	9	193.3	45.3	12	191.3	57.8	52.9	0.930
	1998	9	156.7	38.4	12	192.5	44.5	42.1	0.063
	1999	8	183.8	64.4	12	170.0	30.7	46.8	0.587
	2000	8	180.0	33.0	12	156.7	41.3	43.4	0.557
Goulburn River, Murchison	1997	7	148.6	55.2	7	143.4	46.3	51.0	0.853
	1998	12	123.2	41.1	6	159.5	50.9	44.4	0.166
	1999	11	103.6	38.3	10	123.1	26.2	33.1	0.188
	2000	9	137.8	57.4	9	140.3	44.2	51.2	0.917
Acheron, Taggerty	2000	21	34.8	9.6	12	39.2	9.0	9.4	0.211
Delatite River, Pirie's Bridge & Tonga's Bridge	1995	7	48.6	16.8	12	64.4	18.6	18.0	0.055
	1996	9	55.3	12.9	12	61.9	9.7	11.2	0.220
	1997	10	75.0	20.1	12	82.0	15.8	17.9	0.384
	1998	7	66.9	26.1	12	83.4	31.3	29.6	0.236

C. pH

Site	Year	Waterwatch			VWQMN			s_{pooled}	*P*
		n	μ	s	*n*	μ	s		
Broken Creek, Katamatite	1997	4	6.93	0.13	9	6.67	0.22	0.20	0.023
	1998	7	6.79	0.59	12	7.13	0.28	0.42	0.181
	1999	4	7.08	0.36	12	6.96	0.28	0.30	0.197
	2000	6	7.08	0.22	12	6.82	0.18	0.19	0.015
Broken River, Gowangardie	1997	7	7.00	0.36	12	6.92	0.11	0.23	0.571
	1998	8	7.06	0.50	12	6.80	0.33	0.41	0.220
	1999	5	7.12	0.66	12	7.01	0.28	0.42	0.733
Goulburn River, Balaclava Rd, Shepparton	1996	11	6.65	0.47	11	6.62	0.41	0.44	0.849
	1997	7	6.91	0.37	12	6.91	0.16	0.25	0.969
	1998	7	6.80	0.40	12	6.56	0.26	0.32	0.187
	1999	6	7.18	0.42	12	7.03	0.18	0.28	0.407
Goulburn River, Murchison	1997	7	6.67	0.20	7	6.79	0.26	0.23	0.374
	1998	10	6.71	0.40	6	6.58	0.34	0.38	0.515
	1999	7	7.33	0.43	10	6.95	0.4	0.41	0.089
	2000	6	7.22	0.26	9	6.83	0.29	0.28	0.020
Delatite River, Pirie's Bridge & Tonga's Bridge	1996	14	6.88	0.46	11	6.87	0.15	0.36	0.965
	1997	8	6.95	0.21	12	6.98	0.40	0.34	0.858

D. Tot phos

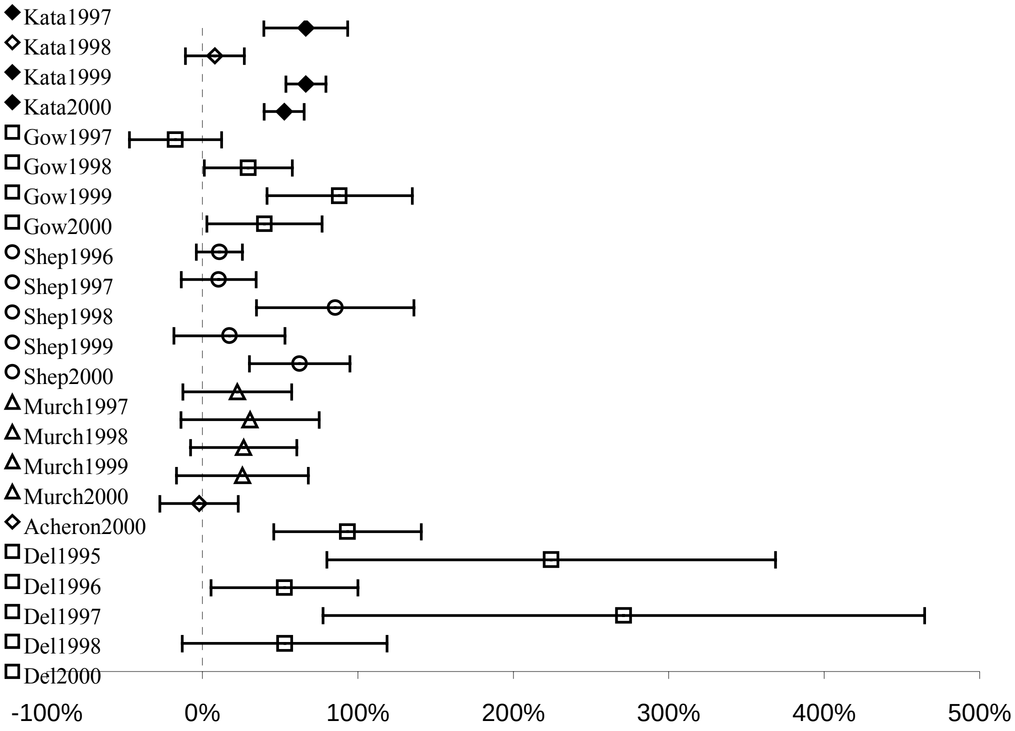
Site	Year	Waterwatch			VWQMN			s_{pooled}	*P*
		n	μ	s	*n*	μ	s		
Broken Creek, Katamatite	1998	7	0.230	0.157	12	0.274	0.113	0.130	0.529
	1999	6	0.242	0.088	12	0.213	0.084	0.085	0.512
	2000	8	0.240	0.053	12	0.268	0.123	0.102	0.505
Broken River, Gowangardie	1997	10	0.129	0.070	12	0.101	0.039	0.055	0.278
	1998	7	0.169	0.038	12	0.129	0.054	0.049	0.079
	1999	5	0.150	0.081	12	0.120	0.047	0.058	0.476
	2000	6	0.118	0.049	12	0.112	0.049	0.049	0.286
Goulburn River, Balaclava Rd, Shepparton	1997	10	0.096	0.043	12	0.082	0.044	0.043	0.453
	1998	9	0.112	0.075	12	0.079	0.045	0.060	0.261
	1999	8	0.101	0.038	12	0.079	0.026	0.031	0.174
Acheron, Taggerty	2000	19	0.073	0.043	12	0.042	0.043	0.043	0.059
Delatite River, Pirie's Bridge & Tonga's Bridge	1995	5	0.042	0.033	12	0.022	0.012	0.020	0.247
	1996	7	0.057	0.041	12	0.022	0.007	0.025	0.064
	1997	9	0.024	0.010	12	0.014	0.007	0.008	0.016
	1998	7	0.036	0.039	12	0.013	0.005	0.023	0.167
	2000	5	0.058	0.016	12	0.029	0.032	0.028	0.026

There was little difference between the data for electrical conductivity (EC) measured by Waterwatch and VWQMN. None of the datasets were statistically significantly different (table 1b.), though Waterwatch means tended to be lower than VWQMN means (figure 1b.). All but one of the Waterwatch means were less than 25% different from the VWQMN means, and 50% were within 10%. All confidence intervals fell within 25%, with 40% including no difference. The difference in variance was likewise small. Two of the 22 annual datasets showing statistically significant difference: in one case Waterwatch had the higher variance; in the other VWQMN variance was greater. VWQMN variance was higher than Waterwatch variance in 50% of cases.

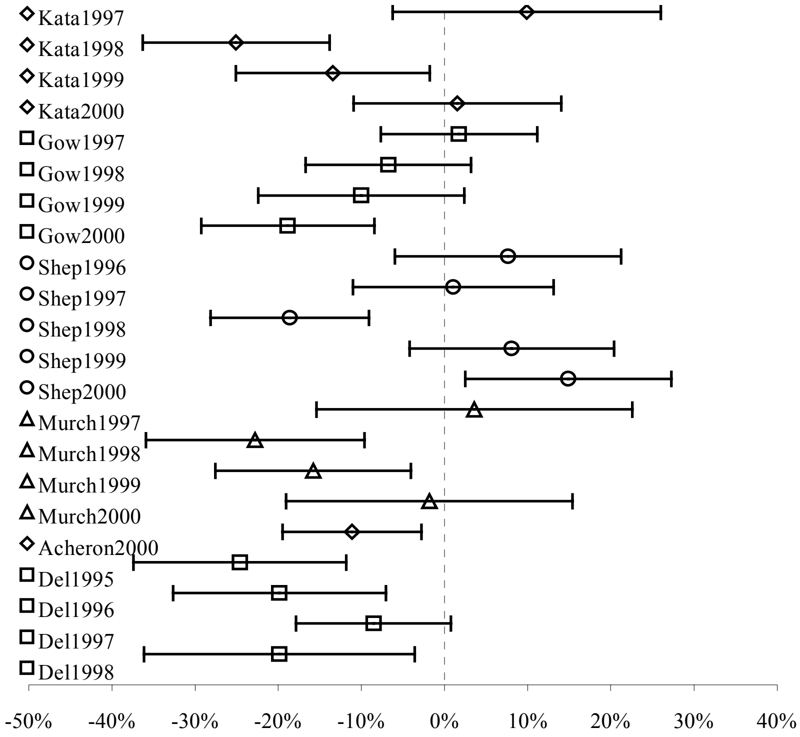
Waterwatch and VWQMN data for pH showed little difference. In three out of seventeen cases, statistically significant differences occurred (table 1c.). All Waterwatch means were less than 6% different to VWQMN means, and over half were less than 2% different. All 95% confidence intervals fell within less than 3.5% difference. Over a third (35%) of the annual datasets however showed statistically significant higher Waterwatch variance than VWQMN variance (table 2c.). In all but four cases, Waterwatch data had higher variance in pH measurement than VWQMN.

Figure 1: the differences between the means ($\text{mean}_{\text{Waterwatch}}$ - $\text{mean}_{\text{VWQMN}}$) expressed as a percentage of $\text{mean}_{\text{VWQMN}}$, with 95% confidence interval calculated using a pooled standard deviation. Coloured symbols refer to those datasets that were statistically significantly difference. (Legend: Kata (Broken Creek, Katamatite); Gow (Broken River, Gowangardie); Shep (Goulburn River, Shepparton); Murch (Goulburn River, Murchison); Acheron (Acheron, Taggerty); Del (Delatite River, Pirie's Bridge & Tonga's Bridge).)

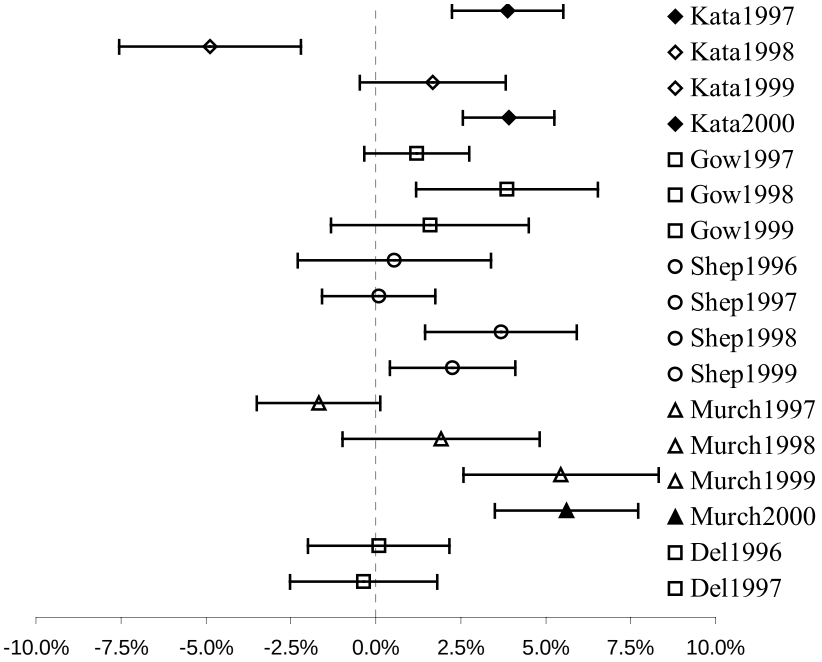
A. Turbidity



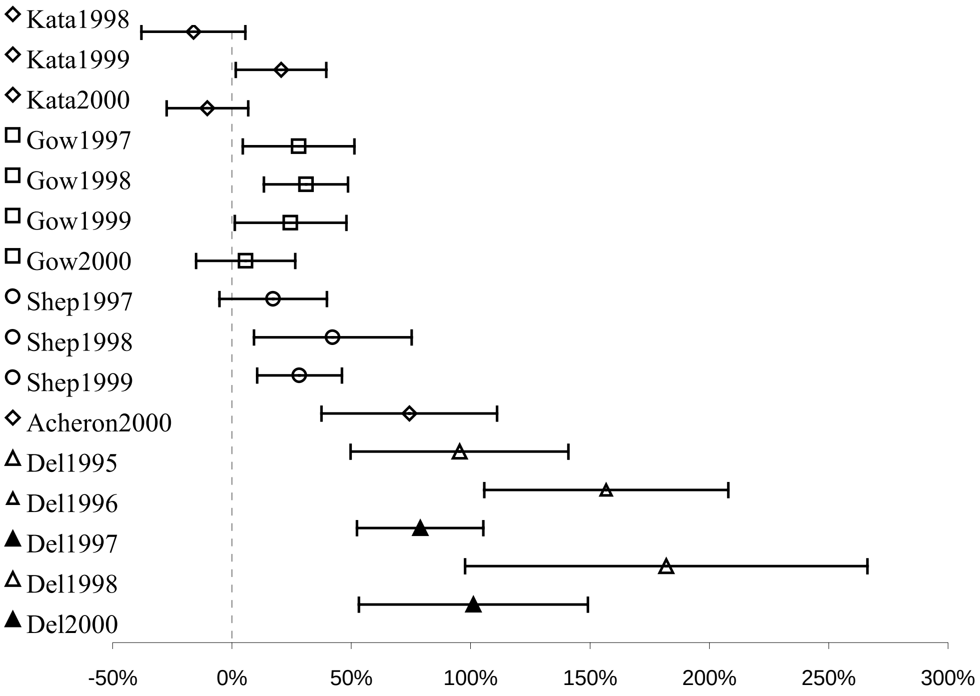
B. Electrical Conductivity



C. pH



D. Total Phosphorus



The results for Total Phosphorus were similar to the results for turbidity, with majority under 50% accurate (figure 1d.). Two datasets out of 16 showed statistically significant difference between Waterwatch and VWQMN data (table 1d.). Waterwatch means tended to be higher than VWQMN means (figure 1d.). The sites with higher total phosphorus concentration (Katamatite, Gowangardie and Shepparton) were all within 50%, with an average difference of 22% from the VWQMN mean. At the low levels of phosphorus (Acheron and, in particular, Delatite), the accuracy of the measure appeared to decrease, resulting in overestimations; the average difference for Acheron and Delatite was 115%. Given the difference in equipment used by Waterwatch and VWQMN, the reduction in accuracy was not surprising. The detection limit of Waterwatch equipment was 0.01 mg/L, with accuracy to one hundredth of a milligram per litre, compared with a limit of 0.005 mg/L, and accuracy to one thousandth of a milligram per litre, for VWQMN. The VWQMN means for Delatite all lay 0.013 and 0.029 mg/L (table 1d.), with some data points as low as 0.005, well below the lowest detectable concentration for Waterwatch. The mean difference between the two statistically different results, both at Delatite, were 0.01 and 0.03 mg/L.

Waterwatch data for total phosphorus showed higher variance than VWQMN in approximately 63% of cases, and were statistically significantly so in three cases (table 2d.), all at Delatite. Once again, this result could be attributed to the limitations of the Waterwatch equipment. VWQMN variance was statistically significantly higher in one case. When Waterwatch variance was higher, the difference was more likely to be much larger, giving larger F-ratio values than when VWQMN variance was greater (table 2d.).

Statistical Power

In the majority of cases, the statistical power was so low as to give little confidence in the ability to detect difference if it existed. This was due to the sample sizes being too small given the high variability (figure 2.). The comparison of differences with 95% confidence intervals is consequently of greater interest than whether the difference is statistically significant. If Waterwatch were to collect data once a month, as does VWQMN, the statistical power would be only 65% to detect a 50% difference, with a pooled standard deviation equal to the difference between the two data sets. This would give an effect size equal to 1, which is optimistic, given the average pooled SD is over 100% of the VWQMN mean for turbidity. With the average pooled SD of 100% of the VWQMN mean (double the difference), the power becomes 22% (figure 2b.). It is not until the difference, or effect size, is 120% (with pooled SD of 100%) that the power crosses the conventionally accepted threshold of 80% (figure 2b.).

In seeking a difference and pooled standard deviation of the same magnitude (effect size of 1), each dataset must have a sample size of 17 per year for the statistical power to reach 80% (figure 2a). VWQMN collects data however a maximum of 12 times per year. With this constraint, Waterwatch must collect data upwards of 28 times per year for a power of 80% in detecting a difference (figure 2a.). Even if Waterwatch were to collect data weekly, the statistical power would only be 87%. Consequently there would be a 13% likelihood of not detecting a difference that exists. Given that the majority of the Waterwatch datasets had fewer than 10 samples per year, the statistical power was less than 60%, and the majority of statistical significance tests were meaningless. The few that were statistically significant ($P<0.05$) had either large Waterwatch sample sizes ($n>20$), or larger differences than pooled SD.

The statistical power of the F-tests of heterogeneity of variance followed similar patterns to the t-test power analyses, but with lower power (Figure 3a.). The sample size required for the conventionally accepted power of 80% is 32 sample a year for Waterwatch, with VWQMN sample size restricted to 12 per annum (figure 3a). If both Waterwatch and VWQMN datasets have sample sizes of 12, the power would reach 80% only if the F-ratio ($\text{variance}_1/\text{variance}_2$) were almost 2.5 (figure 3b). Yet despite the relatively large F-ratio required for high power, in the turbidity data in particular there were many statistically significant results.

Discussion

Waterwatch Victoria’s water quality monitoring programme has both educative and scientific value. The parameters examined in this study showed no overall pattern or bias, but varied from parameter to parameter: some were more accurate and reliable than others when compared with professionally collected data. The differences in results may partly be attributed to volunteers not following the protocols and methods manuals as carefully as professional scientists. Disparity in turbidity and total phosphorus may also be due to the difference in equipment used by the two groups. The results allow Waterwatch to analyse ways in which to complement scientific monitoring, particularly in regard to increasing spatial and temporal scales of water quality monitoring.

While the protocols for monitoring were not found to differ greatly, how strictly they are adhered to both by amateurs and by professionals cannot be quantified or even measured. While local and regional Waterwatch coordinators may be well trained, the volunteers with whom they monitor may be less careful or may simply have misheard or misunderstood instructions. This is the likely cause of the higher variance in the Waterwatch data, which is greater than the variance for VWQMN data in the majority of cases.

Turbidity was the parameter that showed the greatest inaccuracies, in particular at low and high turbidity levels. The increasing inaccuracy at more extreme values can be attributed largely to the limitations of the turbidity tubes used by Waterwatch groups. While the range is 10 to 400 NTU, the ideal range for use for turbidity tubes is 10 to 100 NTU, as after approximately 100, the log scale lines become so close together as to make the estimation of the value difficult and inaccurate. For example, the lines for 300 and 400 NTU are less than 1cm apart on the tube, compared with 23 cm between 10 and 20 NTU. This gives an enormous difference in the relative error that can be made in the estimation. While highly turbid samples could possibly be diluted, estimation of the value would then become more difficult due to the log scale involved, and is therefore not used. Under 8 NTU, which can be crudely estimated using the end of the turbidity tube, all turbidity tube values are best 'guesstimates'. The inaccuracy of turbidity tubes out of the ideal range becomes evident in the Waterwatch results from Katamatite (VWQMN means of 100 to 140 NTU) and Delatite (VWQMN means of less than 9) showing the greatest disparity from VWQMN, and the middle range of turbidity showing reasonably accurate results. Thus it is not just the subjectivity of the measure that can cause problems when using a turbidity tube, but also the limited range of values that can be estimated with much reliability.

The use of turbidity tubes has further limitations other than range. The subjectivity of the measure and the inability to calibrate are easily interpreted as unscientific. To try and combat these problems, many long-term Waterwatch volunteers regularly 'calibrate' their estimating prowess against a turbidity meter, with extraordinarily precise results. The inability to calibrate turbidity tubes may be the cause of the consistent overestimation of turbidity by Waterwatch. As tubes become older, they may become scratched or dirty, which can only be assessed qualitatively, rather than quantitatively.

A comparison of the data produced by volunteers using turbidity tubes and turbidity meters could not be done thoroughly in this study. The origin of the turbidity data were not always known, nor were they measured on the same days or even in the same season. While generally the turbidity tube data did not appear to be any less accurate, particularly when turbidity levels were within the best range, a comprehensive study comparing the volunteer tubes and meter data with professional data needs to be done.

The two parameters that showed the greatest proportionate disparity between Waterwatch and VWQMN data were the most variable parameters, turbidity and total phosphorus. VWQMN means for turbidity, for example, from subsequent years showed a difference of over 50%. This variability is particularly apparent after rain events, when run-off increases the suspended solids and the phosphorus levels in the waterways. The effects of rain events on turbidity and phosphorus becomes more significant when the impetus for monitoring is examined. VWQMN monitor waterways once a month, usually on a set date. While Waterwatch generally monitors monthly or fortnight, depending on the group, some Waterwatch groups are encouraged to monitor immediately after a rain event, as well as regularly, to assess water quality at extreme values. The motivation for sampling may explain Waterwatch's consistent overestimation of turbidity in particular. As demonstrated above, using the dataset from Shepparton 1998 as an example, regular monitoring is, on occasion, augmented with unusual, between-monitoring samples that can greatly influence the resultant mean and variance values.

Waterwatch tended to overestimate total phosphorus concentration in comparison to VWQMN data. The average difference was 0.018 mg/L. Recent examination of the equipment found that the blank specimens were giving a reading of 0.02, suggesting that one or more of the chemicals used may be contaminated. This contamination may account for the bias towards overestimation by Waterwatch. For sites with low concentrations, such as Delatite, this error may be the cause of significant difference, given that 0.02 mg/L in some cases would double the mean. The problem has since been corrected in more recent data.

While the sites with higher total phosphorus levels were reasonably accurate, those with lower concentrations appeared to be more inaccurate, most likely due to the detection limits of the Waterwatch equipment. Most Waterwatch groups use the Merck Oxisolve/Aquaquant test kit, in which the phosphorus concentration is subjectively assessed using colour cards. As a result it is potentially more inaccurate, but is considered more educational, as the effects of higher concentration can be seen by colour change, rather than simply read on a meter. The Merck kit, accurate within a 0.01 to 0.12-mg/L range, and gives greater accuracy at low concentrations than the Hach equipment (0.01-0.9 mg/L range) used in these datasets. Where the phosphorus concentrations are low, as was the case at the Delatite site, the Merck kit could be used in conjunction with the Hach equipment to ensure better accuracy at lower levels. Given that variance appeared no higher for Waterwatch data than VWQMN data, with correction for contamination and appropriate equipment use, Waterwatch should be able to produce reliable data for total phosphorus.

The relatively small differences between the data for electrical conductivity and pH suggest that Waterwatch data for these parameters may be reliable. Waterwatch data for EC in particular showed neither increased variance nor biases such as consistent overestimation. PH likewise showed little disparity, although Waterwatch variance was in most cases greater. (Mattson *et al.*, 1994) similarly found minimal difference between professional and volunteer data in pH measurement.

This study examined the difference of the Waterwatch data from the VWQMN data, and was thus based on the assumption that the professionally collected data are more accurate or correct. While such a comparison was the principle question of interest, the validity of the assumption also needs to be tested. Recent research has shown that large disparities come also from professional analysis (S. Minchin, Water Resource Management, Department of Natural Resources and Environment, unpublished data). Five samples were sent to six laboratories and analysed for total phosphorus and total nitrogen. The results demonstrated that professional and accredited laboratory data can vary widely, with up to a twentyfold difference in concentration of total phosphorus. The results from one sample ranged from 0.02 to 0.5, with an average of 0.3 mg/L. Another sample divided in two and sent to the same lab under separate labels quadrupled in concentration. Given the degree of variation between accredited laboratories, it may not be correct to assume that Waterwatch data are more fallible than professional.

The four parameters discussed in this study are recommended as water quality indicators by (Ladson *et al.*, 1999). Rather than provide all the information on water quality, these parameters may indicate where further water quality information may be required, and more in depth monitoring may be done (Ladson *et al.*, 1999). Such a role of long-term monitoring that may alert waterway managers of larger problems can be and in many cases is already played by Waterwatch and other volunteer groups in local rivers and streams. Waterwatch is able to measure waterways at a much greater spatial scale than VWQMN, monitoring at over 2000 sites as opposed to 170 sites.

Scientific databases are seldom long enough to identify trends and changes in river health (Puckridge *et al.*, 1998; Boulton, 1999), particularly in waterways with high flows and high variability (Puckridge *et al.*, 1998; Boulton, 1999). Waterwatch long-term monitoring programmes have the capacity to fulfil such a role, with many sites having up to 8 years of data.

The power analyses demonstrated the importance of frequent monitoring for the resultant data to be of use for detecting changes and trends in water quality. Goudey (1999) raised doubts in the ability of VWQMN monitoring data to assess compliance of thresholds for physico-chemical parameters, largely due to low statistical power produced by monthly sampling. The feasibility of increasing VWQMN site distribution and monitoring frequency are limited in by resources (Goudey, 1999). In many areas, Waterwatch groups already monitor fortnightly, with the possibility to increase monitoring frequency. In order to be able to produce data that are capable of detecting changes, Waterwatch needs to ensure at a minimum monthly monitoring, with the option in some areas (depending on groups' availability) to monitor more frequently.

Waterwatch is also planning to expand the parameters monitored to include biotic measures, including *E. coli* and macro-invertebrates. Macro-invertebrate assemblages is an indicator of waterway health advocated by many researchers (e.g. Boulton, (1999), Ladson *et al.* (1999)). Volunteer performance collecting data on macro-invertebrate communities has been measured (Fore *et al.*, 2001), and has been found to be quite successful, but limited by the resolution of identification (volunteers are limited to higher taxonomic rank identification) and by the amount of time required for identification of groups (up to hundreds of hours (Fore *et al.*, 2001)). While Waterwatch Victoria does collect data on macro-invertebrate assemblages in some regions, this data is currently limited to order level, with the goal of attaining family-level identification (as in Ladson *et al.* (1999)). More detailed classification can however become difficult for non-scientists, requiring more complex language and identification levels, and most importantly, more time required of volunteers for training and sampling.

Conclusion

It must be noted that the datasets used in this study were amongst the best available within the Waterwatch Victoria programme. The groups had access to pH meters, turbidity meters and laboratories for total phosphorus analysis. Many groups are restricted to using less sophisticated tools, such as litmus papers for pH measurement. These data are however an indication of the type of data the Waterwatch programme is readily capable of collecting, and indeed many of the groups do collect.

Many Waterwatch groups already have the capacity to use a synthetic approach to measuring waterway health, such as the Index of Stream Condition (Ladson *et al.*, 1999). Monitoring by Waterwatch volunteers adds to the database on water quality in areas that are not measured by professionals, with reasonable accuracy. Increasing spatial and temporal scales of waterway monitoring around the country can only add to scientific understanding of waterways and their background levels and fluctuation, measurable by volunteer groups. Just as importantly, if not more so, the involvement of the local community in monitoring water quality raises awareness of waterway health issues, and plays an important educative role. The potential to expand the roles of volunteer groups should be recognised by the scientific and broader communities.

The scientific and decision-making community has long faced the data-adequacy quandary: should monitoring be constrained to only spatially and temporally limited scientific data, or broadened to accept potentially scientifically limited volunteer data? This dilemma is best summarised by (Frankel & Soule, 1981): "conservationists cannot afford the luxury of methodological elegance. We are soldiers in a war and soldiers must be pragmatists. Thus it is our tenet that crude initiatives based on rough guidelines are better than the paralysis of procrastination induced in some scientists by the fear of inadequate data".

Acknowledgments

A big thankyou to all the volunteers that have contributed to the datasets used in this study, in particular Peter Kerr, Katamatite Primary School, Maurice Pawsey, Marie Rosier and Bill Probst of the Goulburn Valley Environment Group. The authors would like to thank Michael McCarthy and the University of Melbourne for the assistance with statistics, and John McCoy for commenting on various drafts.

References

Barrett, G., Davidson, I., G. (1999). Community monitoring of woodland habitats - the Birds on Farms Survey. In *Temperate Eucalypt Woodlands in Australia: Biology, Conservation, Management and Restoration*, eds. R. J. Hobbs & C. J. Yates, Surrey Beatty & Sons. Sydney, pp. 382-399.

Bjorkland, R. & Pringle, C. M. (2001). Educating our communities and ourselves about conservation of aquatic resources through environmental outreach. *Bioscience*, **51**: 279-282.

Boulton, A. J. (1999). An overview of river health assessment: philosophies, practice, problems and prognosis. *Freshwater Biology*, **41**: 469-479.

Clark, T. (1995). Integrated watering strategy community monitoring program. Department of Conservation & Natural Resources.

DNRE. (1999). Victorian Water Quality Monitoring Network and State Biological Monitoring Programme Manual of Procedures. Department of Natural Resources and Environment. 182/97.

Dvornich, K. M., Tudor, M. & Grue, C. E. (1995). NatureMapping: assisting management of natural resources though public education and public participation. *Wildlife Society Bulletin*, **23**: 609-614.

Faul, F. & Erdfelder, E. (1992). GPOWER: a priori, post-hoc, and compromise power analyses for MS-DOS. 2.0 ed, Bonn University, Dept. of Psychology. Bonn.

Fore, L. S., Paulsen, K. & O'Laughlin, K. (2001). Assessing the performance of volunteers in monitoring streams. *Freshwater Biology*, **46**: 109-123.

Frankel, O. H. & Soule, M. E. (1981). *Conservation and evolution*. Cambridge University Press, Cambridge.

Goudey, R. (1999). Assessing water quality objectives: discussion paper. Environmental Protection Authority.

Heiskary, S., Lindbloom, J. & Wilson, C. B. (1994). Detecting water quality trends with citizen volunteer data. *Lake and Reservoir Management*, **9**: 4-9.

Hodgkins, D. & McCoy, J. (1999). Waterwatch Victoria Methods Manual. Waterwatch Victoria.

Hudson, S. J. (2001). Challenges for environmental education: issues and ideas for the 21st century. *Bioscience*, **51**: 283-288.

Kishbaugh, S. A. (1994). Applications and limitations of qualitative lake assessment data. *Lake and Reservoir Management*, **9**: 17-23.

Ladson, A. R., L.J.White, Doolan, J. A., Finlayson, B. L., Hart, B. T., Lake, P. S. & Tilleard, J. W. (1999). Development and testing of an Index of Stream Condition for waterway management in Australia. *Freshwater Biology*, **41**: 453-468.

Lathrop, J. E. & Markowitz, A. (1995). Monitoring water resource quality using volunteers. In *Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making*, eds. W. S. Davis & T. P. Simon, CRC Press. Boca Raton, pp. 303-314.

Lenth, R. V. (1996). Piface: Interface for power and sample-size determination, University of Iowa. Iowa City.

Mattson, M. D., Walk, M.-F., Kerr, P. A., Slepskie, A. M., Zajicek, A. T. & Godfrey, P. J. (1994). Quality assurance testing for a large scale volunteer monitoring program: the acid rain monitoring project. *Lake and Reservoir Management*, **9**: 10-13.

McCoy, J. (2000). Waterwatch Victoria Data Confidence Manual. Waterwatch Victoria.

Middleton, J. V. (2001). The Stream Doctor project: community-driven stream restoration. *Bioscience*, **51**: 293-296.

Puckridge, J. T., Sheldon, F., Walker, K. F. & Boulton, A. J. (1998). Flow variability and the ecology of large rivers. *Marine and Freshwater Ecology*, **49**: 55-72.

Reynoldson, T., Hampel, L. & Martin, J. (1986). Biomonitoring networks operated by school children. *Environmental Pollution (Series A)*, **41**: 363-380.

Saris, F., Groot, H., Hagemeijer, W. & Dijk, A. v. (1996). Monitoring the Dutch agricultural bird community (a choice between a data-death because of the fear of inadequate data and a practical monitoring programme). *Ekologia-Bratislava*, **15**: 87-96.

Sokal, R. & Rohlf, F. (1981). *Biometry: the principles and practice of statistics in biological research*. *Second ed*. WH Freeman and Co, New York.

Thomas, L. (1997). Retrospective power analysis. *Conservation Biology*, **11**: 276–280.

VWQMN. (2001). http://www.vicwaterdata.net.

Waterwatch Victoria. (2001). http://www.vic.waterwatch.org.au.