Hydrol. Earth Syst. Sci., 11(1), 623–633, 2007
www.hydrol-earth-syst-sci.net/11/623/2007

Hydrology & Earth System Sciences

Towards benchmarking an in-stream water quality model

David B. Boorman

Centre for Ecology and Hydrology, Wallingford, Oxfordshire, OX10 8BB, U.K.

Email: dbb@ceh.ac.uk

Abstract

A method of model evaluation is presented which utilises a comparison with a benchmark model. The proposed benchmarking concept is one that can be applied to many hydrological models but, in this instance, is implemented in the context of an in-stream water quality model. The benchmark model is defined in such a way that it is easily implemented within the framework of the test model, i.e. the approach relies on two applications of the same model code rather than the application of two separate model codes. This is illustrated using two case studies from the UK, the Rivers Aire and Ouse, with the objective of simulating a water quality classification, general quality assessment (GQA), which is based on dissolved oxygen, biochemical oxygen demand and ammonium. Comparisons between the benchmark and test models are made based on GQA, as well as a step-wise assessment against the components required in its derivation. The benchmarking process yields a great deal of important information about the performance of the test model and raises issues about *a priori* definition of the assessment criteria.

Keywords: water quality, model, benchmark, Water Framework Directive, Aire, Ouse

Introduction

A long-standing concern with all hydrological modelling is whether the model used in a particular application is suitable for the purpose intended. Addressing this concern involves specific characteristics of the application (e.g. physical characteristics, spatial and temporal scales of relevance, data availability), the purpose that underlies the modelling exercise (e.g. river basin planning, scientific research, investigation of observed impacts), and some understanding of what constitutes suitability (e.g. achievable within budget, most accurate simulation of key variables, usable by the personnel available). However, the use of models continues as they are generally thought to be the best way to advance both catchment management and science. Furthermore, the requirements of Directive 2000/60/EC of the European Union (2000), the so-called 'Water Framework Directive', are likely to generate an increase in the use of models.

The project *Benchmark Models for the Water Framework Directive* (BMW, project website address http://www.environment.fidefault.asp?contentid=61465&lan=en), aims to assist potential modellers in their selection of a 'suitable' modelling tool and, then, to provide a means for assessing the quality of the model for the particular application. This latter process is 'benchmarking'. The BMW project considers models, primarily of water quality, categorised into a number of domains (e.g. river, lake, groundwater), in addition to integrated models that represent two or more domains. This paper applies this benchmarking process in the context of the river domain, using an in-stream water quality model.

Defining a benchmark

To be useful, a benchmark must represent a test, or methodology, that is freely available and, preferably, easy to apply. One approach would be to say that a particular goodness-of-fit statistic should exceed a specified 'benchmark' value. The problem would be in setting the value prior to applying the model. If the application is a case study with numerous high quality data, then a high standard would be expected, whereas in a case study with few or poor data, a lower standard would be appropriate. One way around this problem is to say that the benchmark goodness-of-fit value is that achieved on the case study application using a particular model (the benchmark model). If the model under consideration (the test model) matches or exceeds the benchmark model's goodness-of-fit, it passes the benchmark evaluation. Such an approach would allow for the differences in data quality between case studies, but it has two difficulties: firstly, it requires a choice of benchmark model, and secondly, two models would have to be applied in each case study. While the latter has obvious resource implications, the former is more difficult to address since it would require the agreement of many modellers on which model should be the benchmark model. In practice, it is hard to see such an approach to benchmarking being usable.

To overcome these problems, in this paper the same model code is used to implement both the test and benchmark models. The benchmark model is the least complex (or most simple) implementation of the model code that can address the issue in the case study. This corresponds to the recommendation of Crabtree *et al*. (1986) that "using the simplest model to yield adequate results" is a key aspect of good modelling. The test model is the modeller's attempt to improve the quality of the model application by introducing greater complexity, or realism. The increase in complexity could be in terms of, for example, spatial and temporal resolution, and addition or representation of processes. This approach obviates the requirement for modellers to agree on a single benchmark model: the benchmark model can change with time; only one model code has to be applied; and the goodness-of-fit is dependent on the particular case study.

However, for this comparison to be helpful, certain criteria should be met. Firstly, the model code must be one that all those involved in the case study agree is appropriate, based, say, on its characteristics and prior usage. Secondly, the benchmark model must be capable of addressing the objective of the case study. Thirdly, the goodness-of-fit measure should be agreed prior to applying the model. There is, of course, no guarantee that the test model will perform better than the benchmark; this will depend on the particular characteristics of the case study. If this is indeed the outcome, then the results from the benchmark model, which all parties have agreed is suitable for the case study, can be used directly and the reasons for the failure of the test model to be an improvement over the benchmark will have to be explored by the modeller.

This concept is illustrated in an application of an in-stream water quality model. The difference between the test and the benchmark models is in the representation of in-stream bio-chemical processes. Thus, while the transportation of water, chemical determinands and heat is the same in both models, the benchmark model represents only the mixing of inputs, whereas the test model includes bio-chemical processes.

Methodology

This assessment of the proposed benchmarking process requires:

(i) an objective for the modelling exercise,
(ii) case study rivers,
(iii) a test model, and
(iv) a protocol for the assessment.

These are described in the following sections.

OBJECTIVE

The objective selected was to represent the General Quality Assessment (GQA) classification within a river network. In the 1990s, GQA was used by the England and Wales National Rivers Authority (NRA) (now the England and Wales Environment Agency). The GQA system has six classes from A ('good') to F ('bad') based on percentiles of dissolved oxygen (DO expressed as a percentage of saturation, DO%), biochemical oxygen demand (BOD), and ammonium (NH_4). It is clear from the values defining the classification that the system is not linear in terms of chemical components (Fig. 1).

The means of estimating the percentiles from data is prescribed as the method of moments, which requires an assumption about the underlying distribution of the data. For DO% this is assumed to be normal while for BOD and NH_4, a log-normal distribution is assumed. To ensure that enough samples were available from their routine sampling programme to give reliable estimates of the distribution percentiles, the NRA usually based the calculation of the GQA class on three-year periods (see Table 1).

This type of quality classification assessment is relevant to the WFD since the general chemical status it requires will probably be based on a system of this type, although it may well contain different parameters.

CASE STUDY RIVERS

Two English rivers, the Aire (catchment area 1932 km^2 above the tidal limit) and Ouse (3315 km^2), were selected for the study; both drain to the top of the Humber estuary on the north-east coast of England. The Aire contains major urban areas, notably Leeds, as well as many traditional (i.e. older and dirtier) industrial sites, while the Ouse has less urbanisation and more extensive areas under agriculture. Both rivers rise in the Pennines, which are largely covered by moorland with rough grazing.

Both rivers have been used in other water quality modelling studies, notably in the Land Ocean Interaction

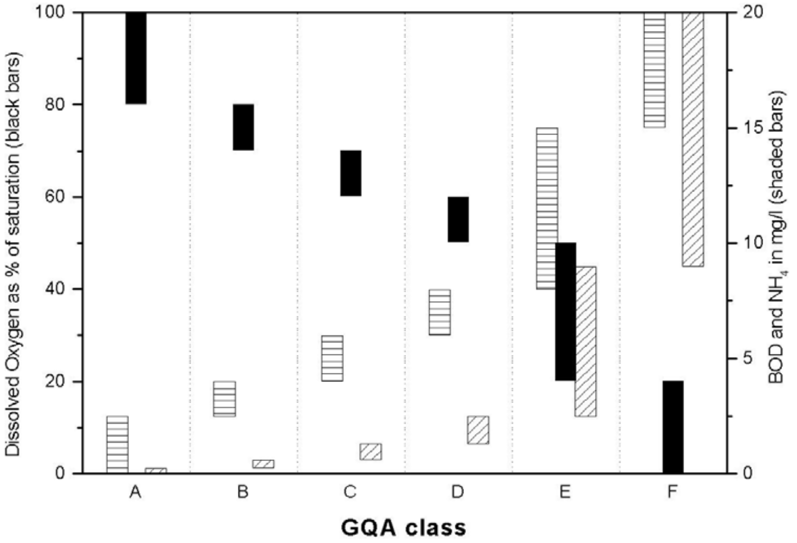

Fig. 1. *Diagrammatic representation of the GQA classification indicating its non-linear nature (DO- solid black bars; BOD – horizontal shading; NH_4 – diagonal shading).*

Table 1. GQA chemical grading for rivers and canals, reproduced from National Rivers Authority, 1991.

Water quality	*Grade*	*Dissolved oxygen (% saturation)*	*Biochemical Oxygen Demand (ATU[1]) (mg l^{-1})*	*Ammonium (mgN l^{-1})*
		10-percentile	*90-percentile*	*90-percentile*
Good	A	80	2.5	0.25
	B	70	4	0.6
Fair	C	60	6	1.3
	D	50	8	2.5
Poor	E	20	15	9.0
Bad	F[2]	–	–	–

[1] as suppressed by adding allyl thio-urea

[2] i.e. quality which does not meet the requirements of grade E in respect of one or more determinands.

Study of the UK Natural Environment Research Council, NERC (Proctor *et al.,* 1999; Tappin *et al*., 2002; Boorman, 2002b). In addition, the Ouse has been used in studies in Climate Hydrochemistry and Economics of Surface-water Systems (CHESS), funded by the European Commission and NERC (Boorman, 2003), and in a modelling study of the fate of agricultural pollutants, funded by the UK Department of Environment, Food and Rural Affairs (Ministry of Agriculture Fisheries and Food, 2002).

Ten years of data were available for the period 1986 to 1995, allowing three separate GQA assessments each for three-year periods: Period I, 1987–1989; Period, II 1990–1992; and Period III, 1993–1995. Period II corresponds exactly with the period considered in NRA (1994).

On the Ouse, eight sites with monitoring data allow the GQA class to be evaluated for all three periods. On the Aire there are 50 such sites. On both rivers, but especially on the Aire, there are concerns about the independence of the data from all sites, since many are close together. Some of these sites are on river stretches not included in the modelled network, but for all sites a GQA assessment was generated during each study period.

On the Aire the GQA class improved slightly over time (Table 2). In Period I, 18% of monitoring sites were classified as F ‘bad’ and 8% as A or B ‘good’ but, by Period III, these figures had changed to 0% and 26% respectively. However, in all three periods the modal class is E ‘poor’. Between periods, there is also a change in the parameter in

Table 2. GQA class derived from observed data at in the Aire (50 sites) and Ouse (8 sites).

Class	Aire			Ouse		
	Period			*Period*		
	I	II	III	I	II	III
A Good	0	0	3	0	0	3
B Good	4	9	10	6	5	3
C Fair	7	7	6	0	1	1
D Fair	10	13	10	0	0	0
E Poor	20	17	21	2	2	1
F Bad	9	4	0	0	0	0

the GQA that limits its class (i.e. the individual parameter falling in the lowest class). In Period I this is equally likely to be BOD or DO%, and less likely to be NH_4, but in Period III it is most frequently BOD that limits the classification, although all three parameters play a more balanced role in the assessment.

For the Aire, the GQA assessment shows that quality decreases downstream. Many head-water reaches are in Class B but the lower reaches of the Aire (and of the Calder, also in the Aire catchment) are classified E. Superposed on this general trend, a few 'hot-spots' of local downgrading in class are followed by downstream recovery. The GQA assessments corresponding to Period II as derived for this study are shown in Fig. 2.

The GQA classes for the monitoring sites on the Ouse show it to be somewhat cleaner than the Aire (Table 2); again there is some improvement in quality in the final period. Within the classification, the limiting parameter is most frequently BOD. Again, the map in NRA (1994) shows that the quality decreases downstream, in this instance from Class A in the headwaters to Class C at the tidal limit and, subsequently, to Class F in the tidal river. As on the Aire, some individual reaches are downgraded by one class but these are generally short reaches and the quality improves again downstream. The GQA assessments corresponding to Period II are shown in Fig. 2.

While improving the quality of rivers is one of the objectives of catchment management, without further investigation, the improvement over time cannot be said to be the result of better management, rather than a reflection

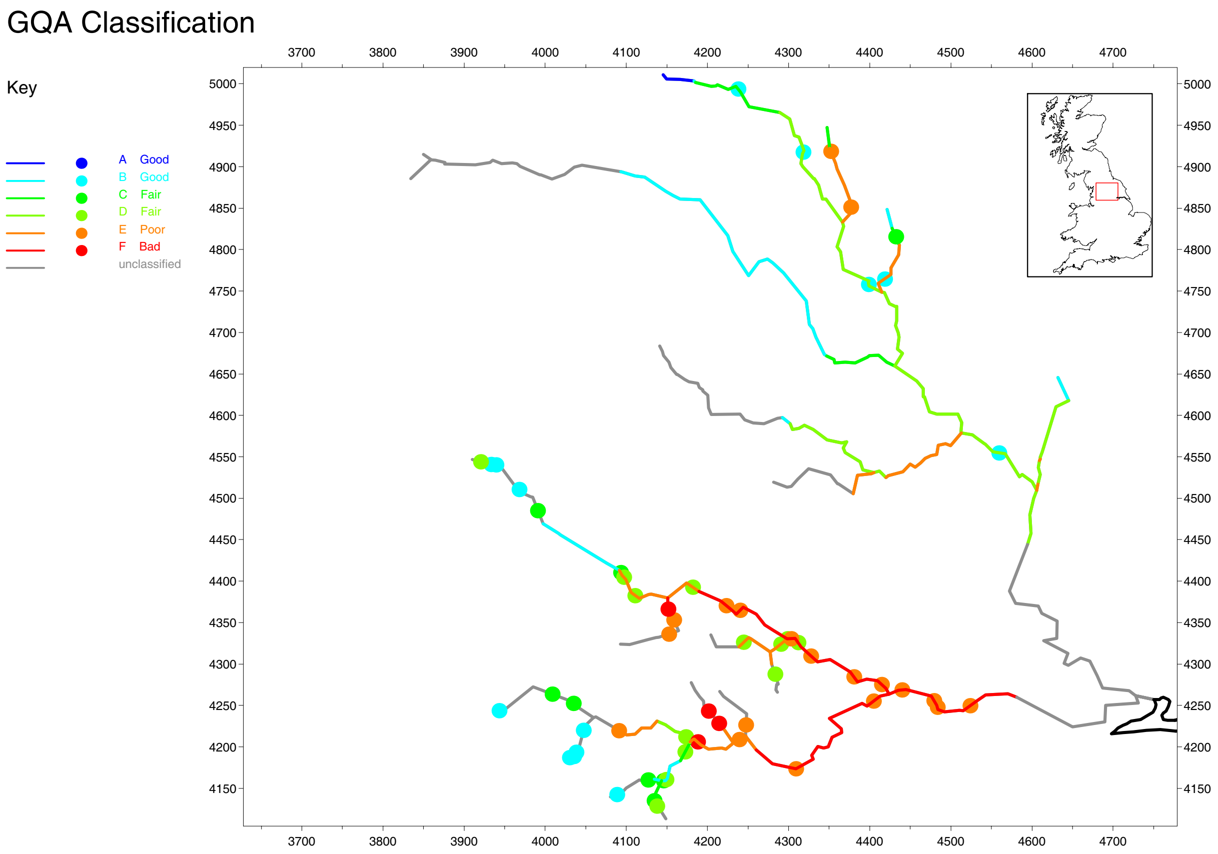

Fig. 2. *The modelled networks of the Rivers Ouse (top) and Aire (bottom) showing the GQA classification as derived by the benchmark model (lines) and from observed data (dots) for Period II.*

of other factors, such as differences in weather between the periods.

THE TEST MODEL

The test model was the Quality Evaluation and Simulation Tool for River-systems (QUESTOR) (Eatherall *et al*., 1998; Naden *et al*., 2001; Boorman, 2003); this model had previously been applied to the catchments studied and represented the items required for the GQA assessment. The determinands modelled were flow, temperature, DO, BOD, NH_4 and nitrate (NO_3). The relevant processes within the model representation were aeration (at the surface and at weirs), benthic oxygen demand, DO uptake against BOD decay, BOD sedimentation, nitrification and denitrification. Temperature is modelled conservatively.

The modelled networks represent a length of 86.3 km on the Aire and 282 km on the Ouse. Inputs to the river modelled river network come from tributaries and discharges, and in both cases were derived where possible from observations. Where this was impossible, data were generated from analogous sites, or set to default values typical of industry types (e.g. food processing, paper industries, sewage treatment). Similarly, time series were constructed for abstractions from the rivers. All time series were daily to correspond with the daily time step used in the model (although a finer time step can be used where this is required by the numerical method used in the model algorithms). Further details of the model and of the input data are given in Eatherall *et al*. (1998), Naden *et al*. (2001) and Boorman (2002a).

The process representation within the model is simple and requires some calibration. The first four years (i.e. until the end of the first GQA assessment period) were used for calibration and the last six years for independent evaluation. The calibration was a manual process using an interactive graphics interface that allows time series and distributions to be displayed alongside observations as the model is running. The interface allows model parameters to be adjusted on a trial-and-error basis, without formal optimisation, but it does report root mean square and efficiency measures.

Above it was noted that, over the whole period of the investigation, the quality as indicated by GQA has improved in the basins studied. Also, it is known at the outset that conditions in the calibration and validation periods differ, so that the models may be required to extrapolate (slightly) to represent the later periods. Whether this extrapolation requires a variation in model processes or if it depends on input data will be investigated later.

THE ASSESSMENT PROTOCOL

The objective stated above is to estimate the GQA class in the two rivers. Neither the benchmark nor the test model does this directly. What the benchmark and test model applications do is to generate time series of a number of modelled quantities, from some of which distributions are derived, key percentiles are abstracted and these are used to define the GQA class.

The question therefore arises whether the benchmarking assessment should just be against GQA, or whether it should consider steps in the modelling process leading to GQA estimation. The former approach addresses the stated objective of the modelling directly. It does not, however, ignore the fact that those other stages exist; they may be used in calibrating the test model, and in fact have been in this instance. The second approach requires the intermediate stages to be considered in both calibration and validation data sets. While it is hard to imagine that a test model will perform worse than a benchmark model when judged against the ultimate criterion, it may perform less well against one of the intermediate criteria. In such a situation it would be necessary to decide if this results in the model failing to achieve the benchmark standard. Regardless of this specific question relating to the benchmarking procedure, these intermediate comparisons are likely to be informative and should be undertaken as part of good modelling practice.

Therefore, as part of the evaluation process, it was decided to adopt both assessment schemes, i.e. directly against GQA, and stepwise including DO, BOD, NH_4, time series and percentiles in deriving GQA.

For any of the modelled quantities, including GQA class, an assessment can be made both subjectively and based on a number of statistics. The argument in favour of statistics is their objectivity, but the act of selecting any one statistic is necessarily subjective. This becomes especially difficult when a number of variables is being considered at the same time (as in this case), and where the assessments can be made at several sites (also as in this case). Such issues have been widely addressed in the literature (e.g. Nash and Sutcliffe, 1970; Aitken, 1973; Loague and Green, 1991). In this study, comparisons made will be made subjectively using tabulated and graphical results, and objectively using two indicative statistics, RMSE and bias defined as

$$RMSE = \sqrt{\frac{\sum (X_i - O_i)^2}{n}}$$

$$bias = \frac{\sum (X_i - O_i)}{n}$$

where X_i and O_i are the i^th^ simulated and observed values from a set of *n* values.

The splitting of data into calibration and validation sets is used in many hydrological and hydrochemical modelling studies. During calibration the modeller endeavours to maximise the goodness-of-fit by adjusting model parameters, constrained either by physical interpretations of the parameters, or by previous experience. Validation checks that the goodness-of-fit is similar when the model is applied to a second data set from the same case study, and is intended to reveal whether the calibration has general applicability to the case study or is tailored overmuch to the particular characteristics of the data set used for calibration. In the context of the above discussion on the setting of a benchmark, calibration is setting a level of performance that is used to assess model performance during validation. It is generally recognised that model performance from the validation data set will be lower than for calibration, but deciding how much lower is still acceptable is another subjective element brought to a supposedly objective process.

Introducing the idea of a benchmark model helps with these comparisons since, in both calibration and validation, the performance of the test model can be compared with the performance of a benchmark model

This comparison between benchmark and test model will be made for both the calibration and validation periods using the subjective and objective assessments described above.

Direct comparison of observed and simulated GQA class

Comparisons below consider the two rivers separately and refer to results presented in Tables 3-6. For Period II the results from both rivers are presented graphically, in Fig. 2 for the benchmark model, and in Fig. 3 for the test model.

THE RIVER AIRE

On the Aire catchment 26 of the 50 monitoring sites are on the modelled river network and can be used to assess the performance of the two models. Using the benchmark model, during the first period 11 sites are assessed correctly, with the remaining 15 being within one class of the assessment based on monitored data (Table 3). Of the 15, the model assesses the quality as worse than observed in 11 cases. Through time, the most notable difference between the assessment made by the benchmark model and that derived from the data is an increase in the spread of the differences. The test model correctly estimates a greater number of GQA classes in the first period, but thereafter there is increasing divergence from the observed classification in the same way as for the benchmark model. Overall both models give an unbiased estimation of class in all three periods.

Table 3. Comparison of GQA classes from the benchmark and test models for each of the 3 sub-periods on the Aire catchment. (Note: –ve class difference means that the model class is worse than the class derived from monitored data.)

MODEL	BENCHMARK			TEST		
Class	*Period*			*Period*		
difference	I	II	III	I	II	III
-3			1			1
-2		3	0		1	0
-1	11	17	7	1	6	4
Correct	11	2	16	17	15	15
+1	4	3	0	8	3	4
+2		1	2		1	2
+3						

Table 4. Comparison of the GQA class derived from the benchmark and test models for all reaches in the Aire catchment.

MODEL	BENCHMARK			TEST		
Class	*Period*			*Period*		
	I	II	III	I	II	III
A Good	0	0	3	0	0	2
B Good	2	9	7	2	9	8
C Fair	9	2	1	9	2	3
D Fair	4	8	5	14	12	19
E Poor	50	32	87	72	73	71
F Bad	38	52	0	6	7	0

An assessment of changes through time, as simulated by the models can be made using all 103 of the modelled reaches (Table 4). From these, a very slight improvement in quality is seen, generally by no more than one GQA. Most notable are the 38 and 58 reaches that are classified as class F (‘bad’) using the benchmark model in the first and second periods, respectively; in Period III, these improve to class E (‘poor’). The test model classifies fewer reaches as class F in the first two periods but, as with the benchmark model, no reaches are classed as F in Period III. With the benchmark model, in all cases either BOD or NH_4 was the limiting parameter in making the GQA assessment. It was on the basis of high BOD values that all assessments as class F were made. Using the test model, all three GQA parameters act in limiting the allocated class.

THE RIVER OUSE

All eight of the monitoring sites in the Ouse catchment that

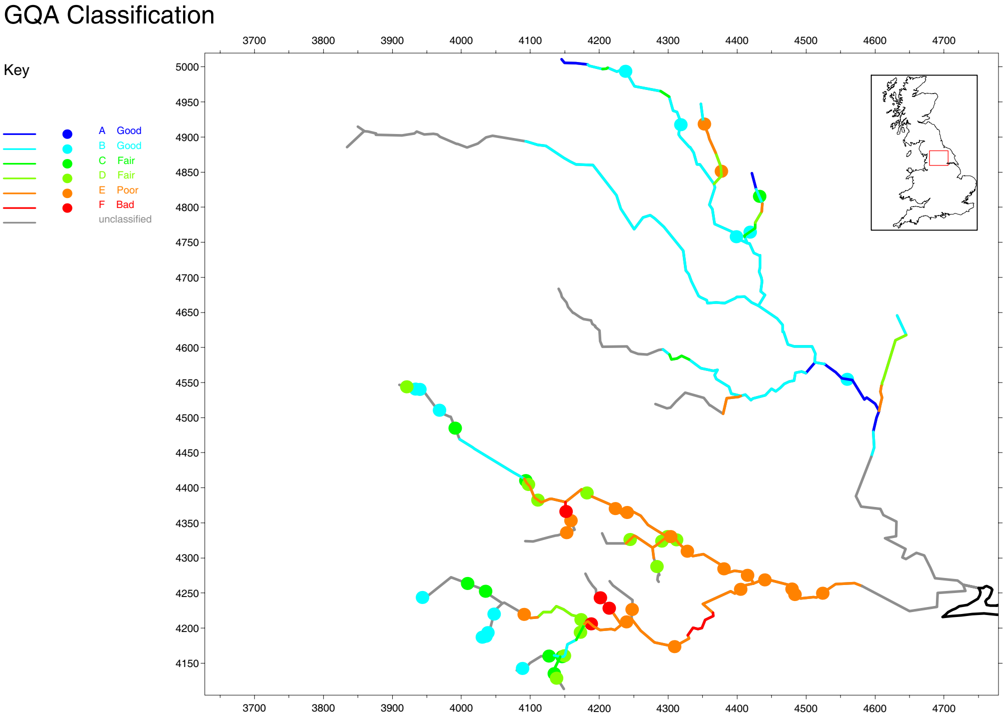

Fig. 3. *The modelled networks of the Rivers Ouse (top) and Aire (bottom) showing the GQA classification as derived by the test model (lines) and from observed data (dots) for Period II.*

can be used to derive GQA classes are on the modelled network, and can therefore be used to compare assessments based on data and those derived from the models. The benchmark model always assesses the class as the same, or worse than, that derived from observed data, i.e. never better (Table 5). Although based on just eight sites, it appears that misclassification is greatest in better quality waters (i.e. it seems likely that a 'good' quality as assessed from monitored data, may be downgraded to 'fair' or even 'poor' by the benchmark model). The test model provided a good central estimate of GQA class, and was always within one class of the value derived from monitored data (Table 5). In the last period, there appeared to be a bias towards underestimation of quality but only by one class and this was based on a small number of comparisons.

Table 5. Comparison of GQA classes from the benchmark and test models for each of the 3 sub-periods on the Ouse catchment. (Note: –ve class difference means that the model class is worse than the class derived from monitored data.)

MODEL	BENCHMARK			TEST		
Class	*Period*			*Period*		
difference	I	II	III	I	II	III
-3	1	1	3			
-2	3	3	3			
-1	2	1	1	3	2	5
Correct	2	3	1	4	4	1
+1				1	2	2
+2						
+3						

Changes in water quality between the three periods can be reviewed based on the classification of all reaches within the modelled network, i.e. 155 reaches (Table 6). Using the benchmark model it suggests that water quality decreased slightly between the first and second periods, but then improved to be slightly better than originally in the third

Table 6. Comparison of the GQA class derived from the benchmark and test models for all reaches in the Ouse catchment

MODEL	BENCHMARK			TEST		
Class	*Period*			*Period*		
	I	II	III	I	II	III
A Good	7	3	12	20	16	53
B Good	21	25	22	101	107	76
C Fair	33	19	28	13	10	10
D Fair	76	77	78	9	9	11
E Poor	18	31	15	12	13	5
F Bad	0	0	0	0	0	0

period. In almost all cases, NH_4 was the limiting parameter within the GQA classification. The simulated changes through time in the GQA classification using the test model show a distinct improvement in the Period III, with many reaches, 53, being placed in the highest class compared with 20 and 16 in the earlier periods (Table 6). DO is rarely the limiting parameter within the GQA assessment, while BOD and NH_4 values limit the classification approximately equally.

DISCUSSION

On the Ouse, the test model assessment of GQA class is more accurate than that of the benchmark model, whereas on the Aire the test model offers relatively little improvement in accuracy over the benchmark model.

For the Ouse, the relative performance of the models can be explained by the absence of processes from the benchmark model. In the benchmark model relatively small amounts of BOD and NH_4 entering the river are transported conservatively, and in the lower reaches of the river network result in a classification as Class D, when the observed class is B. The inclusion of processes within the test model ensures a better estimate of the observed class, although some uncertainty in the model gives some misclassification when compared with the observed class. There is a suggestion that model performance is slightly poorer in the third than in the earlier two periods. Perhaps some changes in actual discharges are not reflected in model input data, or the model calibration is biased towards the earlier period.

For the Aire, two other factors are important. Firstly, the explanation for the benchmark model indicating a better quality than that observed in the river can be attributed to the model's boundary conditions. If these have been set slightly too 'high' then, in the first reaches of the river, they will remain high. With the large number of monitoring sites on the Aire, such over-estimation of quality gets included in the assessment, whereas this is not the case on the Ouse. This over-estimation of quality is seen in both the benchmark and test models, although in the latter case it could be caused by the process representation. Secondly, much of the lower part of the Aire is classified by observations as Class E. If, as on the Ouse, the quality is underestimated by the benchmark model, the maximum error is just one class since Class F is the lowest possible state in the GQA system.

Table 7. RMSE for DO (not DO%), BOD and NH_4 at three sites (upper, middle and lower) on the River Aire (Note the bold values indicate where the benchmark model gives a lower RMSE value than the test model.)

Model	Benchmark			Test		
Variable	*Period*			*Period*		
	I	II	III	I	II	III
DO						
Upper	2.0	2.9	2.1	1.9	2.7	1.8
Middle	4.8	3.9	**2.7**	2.0	3.3	**3.1**
Lower	3.3	3.1	3.0	2.0	2.9	2.7
BOD						
Upper	2.7	3.9	3.7	2.2	3.1	2.9
Middle	7.3	6.1	3.8	4.3	2.9	2.9
Lower	6.1	6.8	5.1	2.0	2.9	2.2
NH_4						
Upper	0.7	0.7	0.3	0.3	0.5	0.2
Middle	1.2	0.7	0.4	1.2	0.2	0.4
Lower	1.3	1.7	1.1	1.0	0.3	0.5

Stepwise assessment of benchmark and test model performance

THE RIVER AIRE

The test and benchmark models simulate time series of DO, BOD and NH_4 that can be compared with the observed time series of these variables. For the Aire, this assessment can be made at the 26 sites with data in each of the three periods. Three sites have been chosen to represent the upper, middle and lower reaches of the Aire, and for each site, and all periods, the root mean square error (RMSE) has been derived (Table 7). A summary of these is:

(1) RMSE values from the test model are in all but one instance less than or equal to those from the benchmark model,
(2) the largest RMSE values occur in the middle and lower reaches, especially in the benchmark model, and
(3) that in the benchmark model RMSE decreases with time.

The RMSE values provide a figure of the goodness-of-fit relevant to continuous simulation of each of the variables at each site. The next step towards deriving a GQA class from the simulated results is to fit distributions, as described above in the context of the monitored data. A next stage in comparing the observed and simulated values would be to generate a RMSE value from an ordered set of values as this would assess the goodness-of-fit across the whole distribution of observed and simulated values. A RMSE derived in this way from ordered data will always give a lower RMSE than using unordered data. While this would generally be a sensible assessment, in this case the classification depends on a specific value in the distribution, *Table 8*. RMSE and bias (in brackets) derived from 25 sites on the River Aire. (Note the bold values indicate where the benchmark model gives a lower RMSE and bias value than the test model.)

Model	Benchmark			Test		
Variable	*Period*			*Period*		
	I	II	III	I	II	III
10%ile DO%	59.5 (56.4)	41.7(40.7)	42.9 (38.6)	17.8 (9.8)	15.4 (-7.4)	23.9 (-3.4)
90%ile BOD	5.5 (3.2)	5.6 (4.1)	4.0 (2.6)	3.0 (-0.9)	2.4 (-0.4)	2.4 (-0.9)
90%ile NH_4	**1.1 (-0.1)**	1.1 (0.8)	0.6 (0.4)	**1.7 (-1.0)**	0.8 (-0.1)	0.4 (-0.1)

and so the required percentiles have been abstracted from the observed and simulated data. RMSE and bias based on all 25 sites have been derived and are presented in Table 8. The test model clearly performs better in terms of RMSE than the benchmark model for both DO% and BOD. For NH_4, the benchmark model performs better than the test model in the calibration period, but slightly less well in the two validation periods. This result will be discussed later.

Comparing the RMSE values from all data at selected sites (Table 7), with RMSE values for the target percentiles at all sites (Table 8) indicates that for both BOD and NH_4 the errors are broadly similar, i.e. estimation of an extreme value is no less accurate than the estimation of the complete data set. Although the figures in Tables 7 and 8 are not directly comparable, since the former has DO and the later DO%, in both models extremely low values of DO are less well estimated than the data set as a whole. This is particularly true for the benchmark model in which the in-stream processes that contribute to the low DO values are omitted.

The bias values for the benchmark model indicate that the model almost always overestimates the actual value (Table 8). In the case of DO% this represents a better environmental value than is observed; for BOD and NH_4 the opposite is the case. Since for the benchmark model the bias in many cases is similar to the RMSE, then almost the whole of the error is the result of model bias. This is an expected outcome since in the benchmark model BOD and NH_4 are determined by dilution alone, i.e. likely sinks of both are not represented. In contrast, for the test model the bias is always considerably less than the RMSE. The bias is in all but one case towards underestimation of observed values. This suggests a small bias in the calibration process, using Period I, which is perpetuated in Periods II and III.

THE RIVER OUSE

Three sites representing the upper and lower reaches of the Ouse, and a small tributary have been selected and RMSE values derived (Table 9). The Ouse is a cleaner river than the Aire with lower values of BOD and NH_4 and there are, therefore, smaller errors in estimating these variables than on the Aire. However, the values of DO are also better estimated for the Ouse than for the Aire, which implies that the benchmark model (i.e. saturated DO) is a better starting point on the Ouse than it was on the Aire. This is confirmed by the relatively modest improvement in performance by the test model over the benchmark model.

For both BOD and NH_4, the RMSE from the test model is less than the corresponding value from the benchmark model. However, although the errors are small, the values from the test model are roughly half of the corresponding values from the benchmark model and, given the non-linear nature of the GQA scheme, there is likely to be a miss-classification on a clean river. The RMSE values show no great difference between sites or between assessment periods.

Table 9 RMSE for DO (not DO%), BOD and NH_4 at three example sites on the River Ouse.

Model	Benchmark			Test		
Variable	*Period*			*Period*		
	I	II	III	I	II	III
DO						
Upper	1.2	0.7	0.8	1.1	0.6	0.7
Tributary	3.1	3.5	3.5	2.4	2.3	2.8
Lower	1.7	1.4	1.5	1.3	1.1	1.4
BOD						
Upper	2.0	1.9	2.2	1.0	0.9	1.0
Tributary	3.6	3.9	3.6	1.5	1.5	1.2
Lower	1.8	2.0	1.8	0.9	0.8	0.7
NH_4						
Upper	1.2	1.2	1.3	0.1	0.2	0.1
Tributary	2.4	2.3	2.3	1.	0.9	0.9
Lower	1.0	1.0	1.0	0.1	0.1	0.1

Table 10. RMSE and bias (in brackets) derived from eight sites on the River Ouse.

Model *Variable*	Benchmark *Period* I	 II	 III	Test *Period* I	 II	 III
10%ile DO%	26.2 (21.4)	26.4 (21.7)	19.9 (17.4)	9.6 (2.3)	7.1 (1.8)	7.0 (-0.5)
90%ile BOD	2.8 (1.7)	3.4 (1.8)	2.4 (1.9)	1.0 (-0.2)	1.3 (-0.5)	0.8 (0.1)
90%ile NH_4	1.9 (1.0)	2.0 (1.4)	2.0 (1.0)	1.1 (-0.3)	0.7 (0.3)	1.2 (-0.2)

As on the Aire, the RMSE values from the selected sites on the Ouse are similar to the values for the 90 percentile values of BOD and NH_4, whereas the error in estimation of the extreme low values of DO distribution is greater than the mean error.

RMSE and bias statistics based on the eight monitoring sites on the Ouse (Table 10) indicate that the test model performs better than the benchmark model for all variables and periods. As on the Aire the benchmark model is seen to be biased towards overestimation of all three variables, whereas the test model appears relatively unbiased.

Discussion

Examining the simulation of the benchmark and test models in terms of statistics based on time series at selected sites and distribution percentiles of the variables at all sites, has given a clearer idea of how the models perform. The benchmark model gives results that are greatly biased with respect to the observed data, and the removal of this bias is certainly one positive aspect of the performance of the test model.

However, the insights into model performance come at the expense of having to generate and examine a great many intermediate data sets and statistics. Within those statistics presented above were two instances where the benchmark model performed better than the test model. The question of whether this constitutes a failure of the benchmarking process should not be addressed only after the results have been obtained. At the outset, the objective was stated as simulating GQA class, rather than the simulation of the time series of each variable, and therefore in this instance this should not be seen as failure.

There is of course no reason why it cannot be stated *a priori* that the benchmark comprises multiple criteria, and that to achieve the benchmark each of these must be satisfied in turn. Care would need to be taken, however, in defining exactly how to apply such criteria, since the rigorous application of intermediate criteria may lead either to unnecessary additional effort, or to the rejection of a model that actually performs well against the ultimate objective.

It is the difficulty of addressing such multiple criteria that leads to the test model performing less well than the benchmark model in two of the above comparisons. The calibration of the test model was a subjective process that sought to balance the fit between DO, BOD and NH_4 at individual sites in terms of both time series and distributions. In adjusting model parameters, there were instances where achieving a good fit against one variable necessitated compromising the fit of one of the other variables. In addition, maximising the fit at an upstream site could compromise the fit downstream. It was, therefore, quite likely that, given the degree of compromise that had to be adopted during calibration and the many comparisons made, there would be a few instances in which the benchmark model would perform better than the test model.

Two issues that emerge from this are the parameterisation of the model and the subjectivity of the calibration process. To a large extent both of these should be addressed by the model developer, i.e. the developer should provide advice on, or tools to explore, parameter sensitivity and, similarly, advice, or tools, to assist model calibration. However, providing such advice, or tools, is far from straightforward for the model developer since a general purpose model is being used to fulfill a specific requirement (i.e. estimating a quality class based on three variables, for many reaches). Further understanding of these issues is an obvious objective of further research.

Conclusion

A particular concept of a benchmark against which model performance can be assessed has been presented; it takes the form of a benchmark model against which a test model can be compared. Because of the definition adopted, the two models are in fact different versions of the same model. Introducing such a benchmark model was, therefore, a straightforward process using existing modelling software, and requires no more data than the test model.

Using the benchmark model, by definition, enables the performance of the test model to be compared with another modelling approach, albeit one that is almost certainly biased in its results. Any alternative form of model comparison is likely to involve a second model, which may require additional expertise and data. However, inevitably, introducing any second model into a modelling study introduces the possibility of very many more comparisons of model performance (i.e. not just model with data, but also model with model).

From the application described above, it is clear that to adopt the benchmarking approach requires more than defining a benchmark model. It is equally necessary to have an *a priori* description of the objectives and how they will be implemented. Given these two components, it seems that the benchmarking process can make a positive contribution to a modelling exercise. In this example, a great deal about the rivers, classification and model processes could be gained without delving into the intermediate simulations required to produce the GQA assessment. It is, however, only by exploring these intermediate simulations that it is possible to see if a correct GQA assessment is in fact being produced by a poor model.

From the two applications it is obvious that the benchmarking process is catchment- and data-dependent. The conclusion from this is that the results of benchmarking on one study may have little or no relevance to another application. However, by its very nature, it is easy to repeat the benchmarking for a new application. This is in contrast to conventional model comparisons, where the greater effort involved may encourage the wider extrapolation of the results.

The benchmark model, even within a single domain such as in-stream modelling, need not be the same in all situations. By basing the benchmark model on the test model, it seems equally valid to allow, for example, the DO simulation to be set as conservative or as saturated. Clearly some further applications of the benchmarking process, involving other models on the same case studies and other case study rivers, are needed to define better how the concept of a benchmark model can be applied more generally.

A final conclusion is that the benchmark approach is probably suited only to those applications with easily defined objectives. Where the purpose of modelling is, for example, to gain a better understanding of how a river system functions, it may be appropriate to adopt some other form of good modelling practice that considers in greater detail the processes and intermediate stages within the model. However, the idea underpinning the benchmarking approach of beginning with a simple model before introducing complexity is widely applicable and relevant to the exploration of catchment management.

While the benchmarking concept as presented in this paper is easily adopted for use in other model applications, some next steps in developing this approach to benchmarking are obvious. They are: to use other models within the same domain (i.e. rivers); to transfer the concept to other domains (e.g. lakes, diffuse runoff); to explore the relationship between calibration and the objective; and to set the comparison with a benchmark model against model sensitivity and uncertainty analyses.

Acknowledgements

This study has been undertaken as part of the project *Benchmark Models for the Water Framework Directive* (BMW) and the author acknowledges the financial support from the UK Natural Environment Research Council and the EC Energy, Environment and Sustainable Development Programme (contract EVK1-CT-2001-00093).

References

Aitken, A.P., 1973. Assessing systematic errors in rainfall-runoff models. *J Hydrol.*, **20,** 121–136.

Boorman, D.B., 2003a. LOIS in-stream water quality modelling. Part 1: Catchments and methods. *Sci. Total Envir.,* **314-316**, 379–396.

Boorman, D.B., 2003b. LOIS in-stream water quality modelling. Part 2: Results and scenarios. *Sci. Total Envir.,* **314-316**, 397–410.

Boorman, D.B., 2003. Climate, Hydrochemistry and Economics of Surface-water Systems (CHESS): adding a European dimension to the catchment modelling experience developed under LOIS. *Sci. Total Envir*., **314-316**, 411–438.

Crabtree, R.W., Cluckie, I.D., Forster, C.F. and Crockett, C.P., 1986. A comparison of two river quality models. *Water Res.,* **20**, 53–61.

Eatherall, A., Boorman, D.B., Williams, R.J., and Kowe, R., 1998. Modelling in-stream water quality in LOIS. *Sci. Total Envir.,* **210/211**, 499–518.

Loague, K. and Green, R.E., 1991. Statistical and graphical methods for evaluating solute transport models: Overview and application. *J. Contam. Hydrol.*, **7**, 51–73.

Ministry of Agriculture Fisheries and Food, 2002. *An integrated approach to modelling the fate of agricultural pollutants at national scale*. 1-21.

Naden, P.S., Cooper, D.M. and Boorman, D.B., 2001. Modelling large-scale river basins. In: H*A Land Ocean Interaction Study: measuring and modelling fluxes from rivers to the coastal ocean.* D. Huntley, G.J.L. Leeks and D.E. Walling (Eds.). IAWQ, London.

Nash, J.E. and Sutcliffe, J.V., 1970. River flow forecasting through conceptual models: 1. A discussion of principles. *J. Hydrol.,* **10**, 282–290.

National Rivers Authority, 1991. *The quality of rivers, canals and estuaries in England and Wales.* NRA, Bristol, UK. 63pp.

Proctor, R., Holt, J.T., Harris, J., Tappin, A.D. and Boorman, D.B., 1999. Modelling the Humber estuary catchment and coastal zone. *Proc. 6th Int. Conf. on Estuarine and Coastal Modelling American Society of Civil Engineers*, New Orleans, USA.

Tappin, A.D., Harris, J.R.W., Uncles, R.J. and Boorman, D.B., 2002. Potential modification of the fluxes of nitrogen from the Humber estuary catchment (UK) to the North Sea in response to changing agricultural inputs and climate patterns. *Hydrobiologica*, **475-476**, 65–77.