Hydrol. Earth Syst. Sci., 11(1), 647–662, 2007
www.hydrol-earth-syst-sci.net/11/647/2007


Hydrology & Earth System Sciences

# Issues of diffuse pollution model complexity arising from performance benchmarking

M.G. Hutchins[1], C. Dilks[2], H.N. Davies[1] and A. Deflandre[1]

[1]CEH Wallingford, Maclean Building, Crowmarsh Gifford, Wallingford, Oxfordshire, OX10 8BB, UK
[2]The Macaulay Institute, Craigiebuckler, Aberdeen, AB15 8QH, UK

Email for corresponding author: mihu@ceh.ac.uk

## Abstract

Flow and nitrate dynamics were simulated in two catchments, the River Aire in northern England and the River Ythan in north-east Scotland. In the case of the Aire, a diffuse pollution model was coupled with a river quality model (CASCADE-QUESTOR); in the study of the Ythan, an integrated model (SWAT) was used. In each study, model performance was evaluated for differing levels of spatial representation in input data sets (rainfall, soils and land use). In respect of nitrate concentrations, the performance of the models was compared with that of a regression model based on proportions of land cover. The overall objective was to assess the merits of spatially distributed input data sets. In both catchments, specific measures of quantitative performance showed that models using the most detailed available input data contributed, at best, only a marginal improvement over simpler implementations. Hence, the level of complexity used in input data sets has to be determined, not only on multiple criteria of quantitative performance but also on qualitative assessments, reflecting the specific context of the model application and the current and likely future needs of end-users.

**Keywords:** flow, nitrate, catchment, benchmark, model, CASCADE, QUESTOR, SWAT, Aire, Ythan

## Introduction

Assessing the performance of a particular model is fundamental to benchmarking, as developed, e.g. in the EC 5th Framework Programme Benchmark Models for the Water Framework Directive (BMW) project (Kämäri *et al.*, 2006; www.environment.fi/syke/bmw). Initially, on the basis of its suitability for evaluating management strategies, a model capable of satisfying an application by an end-user has to be identified. The performance assessment in BMW requires the output variables considered relevant to be specified at the outset of the modelling study, together with an objective measure or measures of their goodness-of-fit, derived by comparing observed and simulated data.

Goodness-of-fit statistics for calibration, sensitivity analysis and comparison of model performance for diffuse pollution models have been introduced previously (Hutchins *et al.*, 2006). Perrin *et al*. (2006a) details various options for defining quantitative measures for model evaluation and comparison in rainfall–runoff and water quality domains; they stress the need to define an appropriate benchmark to standardise assessments. For benchmarking purposes, goodness-of-fit statistics are potentially powerful in identifying the level of complexity which is appropriate. However, an increase in model complexity generally increases the number of parameters, unconstrained in their values, which have to be calibrated. This, including all available process knowledge, may lead to over-parameterisation (Addiscott *et al.*, 1995). In optimising free parameters, a good performance, obtained for undesirable reasons (Kirchner *et al.*, 1996), may mask deficiencies in the structure of the model and limit the power of the model evaluation tests. Using a range of criteria, Perrin *et al.* (2003) assess the trade-off between improved model performance and increased complexity as increasing reliance on model calibration. In catchment-scale applications, models developed at plot or field scale may become over-parameterised as the feasibility of measuring parameters declines (Addiscott *et al.*, 1995; Wheater and Beck, 1995).

Striving to balance input data quality with model complexity is likely to pinpoint optimal model structures for predictive purposes. (Van Rompaey and Govers, 2002).

In this study, simple regression-based methods have been implemented in an initial benchmark simulation of water quality status. Subsequently, two cases for which modelling studies using different process model codes already reported (Hutchins *et al.*, 2006; Dilks *et al*., 2004), have been developed further by considering that benchmark levels of performance can be defined by the simplest implementation of a process model code considered (by modelling specialists and end-users) suitable to answer the agreed management question. Results from more complex applications of the same code can then be evaluated against the benchmark. With recent GIS-based modelling approaches, the spatial scale at which models may be assumed accurate must be determined, with consequent implications for appropriate resolution of input data (Wagenet and Hutson, 1996). Here, therefore, specific consideration is given to the effect on model performance of increased spatial detail in input data; various model implementations can be evaluated against the initial benchmark provided by the regression model. Model performance is sensitive to the calibration process. When increasing the spatial detail in input data, any change in the sensitivity of performance to this process can be illustrated.

## Background to case studies

### MODELLING OBJECTIVES

Before beginning modelling applications, the objectives must be specified and associated performance criteria determined. Here, N concentration, in the form of nitrate as opposed to ammonium or total-N (i.e. nitrate-N), is to be simulated at a catchment outlet as a basis for evaluating the impact of N on water bodies downstream. In this paper nitrate-N is simply referred to as N, as opposed to ammonium or total-N for example. The modelling approaches chosen allow exploration of management options for bringing about changes in the N regime.

Data availability constrained the analysis to national data sets and those collected from routine monitoring. The time step chosen (1 day) for the simulation depends on the data, the dynamics of the modelled system, and the residence times in the water bodies downstream. For model evaluation, a measure of efficiency adapted for benchmark purposes as proposed by Seibert (2001), was applied to flow and N concentration. This benchmark criterion ($E_{bench}$) (Seibert, 2001) is a generalised form of the $R^2$ measure of efficiency (Nash and Sutcliffe, 1970) (Eqn.1):

$$E_{bench} = 1 - \left\{ \frac{\sum_{i-1}^{n} [O_i - S_i]^2}{\sum_{i-1}^{n} [O_i - B_i]^2} \right\} \quad (1)$$

where, $n$ represents the number of paired observed ($O_i$) and simulated ($S_i$) data in the time series (at a time-step $i$) and $B_i$ the benchmark series. The criterion allows the fit of modelled time-series to be evaluated in the context of either a measured standard (e.g. the mean of observations), a simulation of this standard, or a benchmark time-series. In the former two cases, all the elements in series $B_i$ take the same value, the standard. Simple regression models can provide simulation of a standard, such as an estimate of the mean observed N concentration. A benchmark time-series can be provided by a simple implementation of a process model. More complex implementations of the process model, for example using more detailed spatial representation of input data, can then be tested against this benchmark. A positive efficiency coefficient ($0 < E_{bench} \leq 1$) indicates that the model output fits the observations better than the benchmark model, while a negative efficiency coefficient ($E_{bench} < 0$) indicates that the model application is performing less well than the benchmark.

Error indices in the flow duration curve are additional quantitative performance indicators. The index of cumulative frequency error (Refsgaard and Knudsen, 1996) calculates errors between observed and simulated data at each percentile value from 1 to 99, where a maximum value of 1 is a perfect fit. The mean absolute percentage error (MAPE) across each percentile value was also calculated. Similar indices were derived for N.

### STUDY CATCHMENTS

The catchments studied were the 865 km$^2$ Aire at Lemonroyd (UK National Grid Reference (NGR) SE 381282) (Hutchins *et al.* 2006) and the 540 km$^2$ Ythan at Ellon (UK NGR NJ 947303) (Dilks *et al*. 2004) (Fig. 1). Flow and N data are available at both locations. A rural sub-catchment of the Aire at Kildwick (282 km$^2$) (UK NGR SE 013457) was also monitored and used in calibrating the Aire Model.

Almost 50% of the Aire catchment at Lemonroyd is non-agricultural, with urban/suburban land-cover accounting for 21% (occurring almost exclusively downstream and to the south east of Kildwick), and 29% was woodland and upland moor (Fig. 2a). The agricultural area is 80% grassland and 20% arable. The annual rainfall averages 998 mm and ranges from less than 650 mm in the urban centres of Leeds-Bradford in the lowland south-east, to over 1250 mm in the hills of the Yorkshire Dales to the north-west. The low

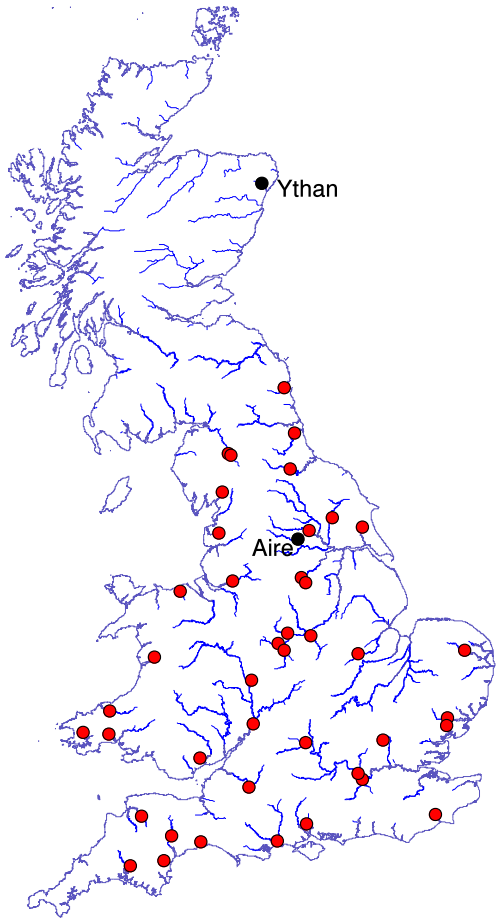


Fig. 1. *A map of Great Britain showing location of the two study catchments and the location of the 43 additional catchments used to formulate the landcover-based multiple regression model of mean N concentration*

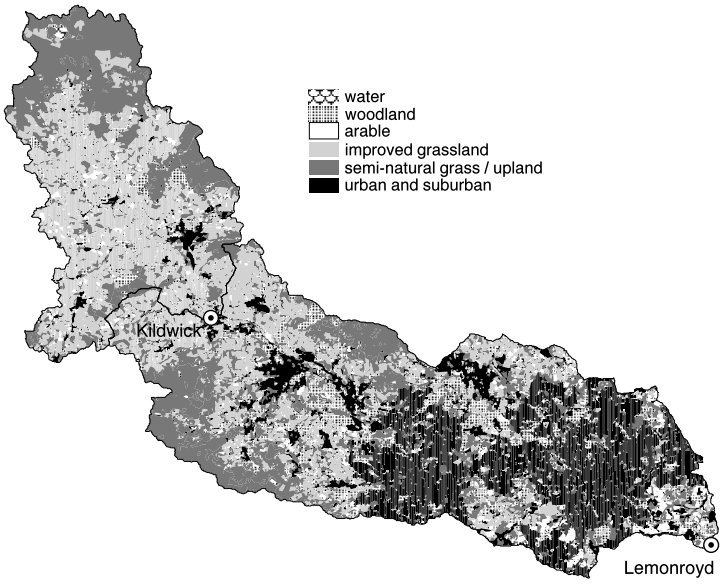


Fig. 2(a). *Map showing land use in the Aire catchment and the location of gauging stations at Lemonroyd and in a sub-catchment at Kildwick. Land use percentages used for regression modelling: arable = 4%, upland and semi natural grass (upland) = 32%, urban and suburban (urban) = 24%, Given the near absence of point source inputs, at Kildwick, the CASCADE model alone was applied. At Lemonroyd a linked model (CASCADE-QUESTOR) was applied.*

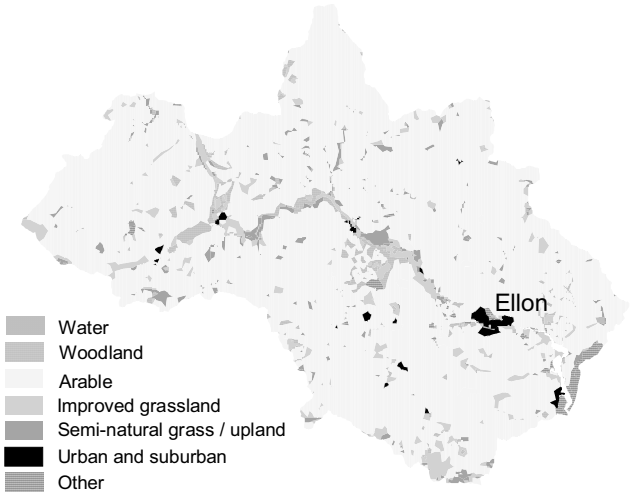


Fig. 2(b). *Map showing land use in the Ythan catchment and location of the gauging station at Ellon. Land use percentages used for regression modelling: arable = 64%, upland and semi natural grass (upland) = 7%, urban and suburban (urban) = 1%.*

permeability of most of the catchment soils means they are prone to seasonal waterlogging but more freely-draining soils are prevalent in the southern part of the catchment near Bradford.

Land use in the Ythan catchment is predominantly agricultural (95%), 64% is cropped arable land and the rest is grazed and mown grassland. Urban areas are less than 1% with the main population centre, Ellon, located at the outlet of the modelled area. The remaining 4% of the catchment is woodland and moorland (Fig. 2b). Soils, a mixture of humus iron-podzols, brown forest soils and non-calcareous gleys, are generally free-draining although less so towards the east of the region. Mean annual rainfall and water yield (at Ellon) are 815 mm and 450 mm, respectively. Annual rainfall over the region varies by less than 7% (Dunn *et al.*, 1998). Flow is dominated by slow sub-surface response, with baseflow making up 75 to 80% of the total annual discharge.

## Modelling approaches

### REGRESSION MODEL FOR N CONCENTRATION

Davies and Neal (2004) have shown that mean N concentration at a given location can be simulated using information on the proportions of constituent land cover

classes within the catchment. Therefore, a model was set up, based on data throughout England and Wales, against which data from the case study catchments could be tested.

Land cover composition was estimated on a catchment basis using the CEH Land Cover Map 2000 (LCM2000) (Fuller *et al.*, 2002). Classes in LCM2000 were aggregated as described by Davies and Neal (2004) into three land cover classes, arable (A), urban (U) and upland (UP); these were the only classes that showed a relationship between their percentage cover in a catchment and N concentration.

For calibration, N concentrations from rivers were taken from the EA Harmonised Monitoring Network (HMN) dataset, choosing sites throughout England and Wales so that differences in geology, rainfall, terrain and size were included in a comprehensive spatial distribution. Catchment sizes ranged from 157 km$^2$ to 9950 km$^2$. For the period of calibration chosen, 1988–1992, more than 40 readings were available for each site. In the calibration, 43 sites were used and nested catchments were generally avoided to reduce possible interdependencies (Fig. 1). Sites were chosen near flow gauging stations to allow flow weighted concentrations to be estimated if required. Given the uncertainties associated with determining accurate proportions of specific land uses, particularly in the smaller catchments, land-cover figures in the three aggregated classes were rounded to the nearest whole percent. Nevertheless, model predictions are very sensitive to small changes in the land-cover proportions. The land- cover proportions in the catchments used for the calibration range from 3–76% for A, 3–61% for UP and 1–54% for U. The calibrated multiple regression ($R = 0.919$, $p<0.001$) is:

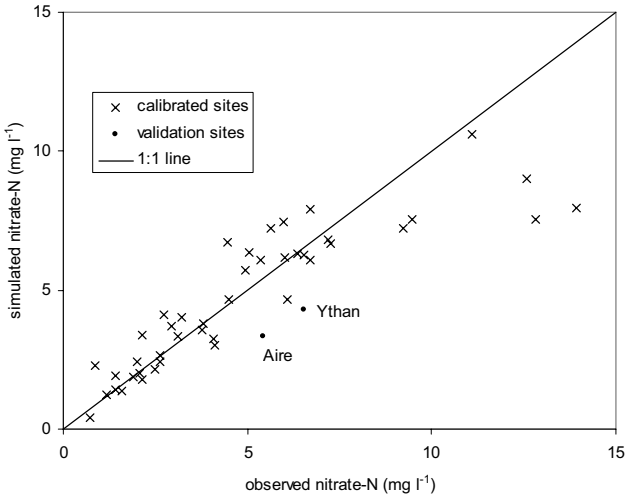


Fig. 3. *Calibration of the landcover regression model showing model performance for the Aire (Lemonroyd) and Ythan catchments*

$$\ln(N) = [\ln(A) \times (0.269 \pm 0.066)] + [\ln(U) \times (0.299 \pm 0.044)] - [\ln(UP) \times (0.314 \pm 0.080)] + [0.947 \pm 0.373] \quad (2)$$

where nitrate concentration ($N$) is in mg N l$^{-1}$ and +/– indicates twice the standard error.

Data from the Kildwick and Lemonroyd catchments of the Aire were excluded from the calibration and included in a dataset for model validation. The regression model for England and Wales was also applied in the Ythan catchment to determine average N concentrations in the River Ythan. Land cover proportions were estimated using the LCS88 land-cover data set (MLURI, 1993). The fit of the model calibration and the performance of the model for the Aire and the Ythan are shown (Fig. 3).

## AIRE CATCHMENT: THE CASCADE-QUESTOR PROCESS MODEL

A linked modelling approach was chosen. The CASCADE (Catchment Scale Delivery) model, (Cooper and Naden, 1998), was used to quantify delivery to watercourses of diffuse water flows and associated pollutants (in this case N). The outputs from CASCADE, as daily time-series and flow from component sub-catchments, were input to QUESTOR (Boorman, 2003), which represented the river channel and combines the inputs from diffuse sources (CASCADE) with the effects of point source inputs and abstractions. In this linked approach, model performance is influenced by processes of parameter optimisation. For CASCADE, data are available to constrain parameterisation of profile-scale soil hydrology. However, six additional parameters, representing catchment-scale hydrological response, have to be optimised against daily flow observations. Here, the near absence of point source inputs above Kildwick led to the CASCADE model alone being optimised against data from the gauge at Kildwick and the resultant parameters were then applied to the rest of the catchment (Fig. 2a). Point source inputs are considerable downstream of Kildwick so CASCADE-QUESTOR was applied. In the QUESTOR model, two parameters, controlling the process of nitrification and denitrification, were optimised against data at Lemonroyd. Hutchins *et al.* (2006) detail the two models and their calibration, including issues specific to the Aire catchment regarding the nature of the linkage. Also, for the case study, a qualitative judgement was made of the most sensitive parameters which must be calibrated. (Table 1).

The CASCADE model was applied using either lumped or distributed input data (representing each of rainfall, soil type and land-use). A previous implementation of the model to the Aire catchment used the distributed data (Hutchins *et*

*Table 1*. Parameters identified as being particularly sensitive in the context of the specific case-study applications

| CASCADE-QUESTOR | SWAT |
|---|---|
| ● percolation from upper soil store: magnitude and shape parameters | ● soil parameters relating to available water capacity, hydraulic conductivity and depth |
| ● lateral flow from upper soil store: magnitude and shape parameters | ● groundwater delay, affecting the timing of groundwater inputs to the river· groundwater nitrate concentration |
| ● in-river denitrification parameter | ● lag coefficient representing fraction of total surface runoff entering a reach on any one time-step |
| ● in-river nitrification parameter | |

*al*., 2006). A DTM was used to define 151 sub-catchments (Hydrological Response Units: HRUs) (Cooper and Naden, 1998), for each of which rainfall, soils and land-use information were defined. For rainfall input, one of six available daily time-series was selected and used in the model along with the percentage abundance of soil types, as classified by the HOST system (Boorman *et al*., 1995), and land-use classes (derived from a land cover map of Great Britain (Fuller *et al*., 1994) and agricultural census statistics). Only those categories of soil and land use comprising more than 1% of the total catchment area were considered. Soil classification defined soil hydrological parameterisation, whereas land-use classification drove monthly N inputs.

In addition to the distributed rainfall, soils and land-use data, a lumped representation of each was defined and used throughout in each component HRU. For rainfall, a catchment average time-series was calculated from a 1 km$^2$ grid of data calculated from daily Met Office raingauges using the triangle method (Jones, 1983). Soil parameters were derived from statistics of the National Soil Resources Institute SEISMIC database (Hallett *et al*., 1993) summarising soils in HOST class 24, the dominant soil type in the catchment. For land use, N accumulation rates for the catchment were calculated monthly using area-weightings of the land-use-specific values presented by Hutchins *et al*. (2006).

As Table 2 shows, eight implementations of the CASCADE model were now available, each driven by different combinations of input data complexity. The process of parameter optimisation was the same as that described by Hutchins *et al*. (2006). Results from these model applications were compared with observations from Kildwick.

For modelling at Lemonroyd, the linked CASCADE-QUESTOR model was used. The CASCADE calibration ensured total flows at Lemonroyd were not overestimated and contributions from predominantly urban HRUs were omitted. The rationale behind these modifications, which are necessary in catchments with significant urbanisation, is described in detail in Hutchins *et al.* (2006). To test the effect of input data complexity on model performance, two implementations of the linked model were set up:

(i) CASCADE-QUESTOR Aire Model A which used HRU-based data generated from CASCADE Aire Model 1 (least complex);
(ii) CASCADE-QUESTOR Aire Model B which used HRU-based data from CASCADE Aire Model 8 (most complex).

## YTHAN CATCHMENT: THE SWAT PROCESS MODEL

SWAT (Arnold *et al*, 1998) was applied in the Ythan catchment as described in Dilks *et al.* (2004). A DEM was used to divide the catchment, above Ellon (the tidal limit), into 32 sub-catchments, which were parameterised using a series of response units (RU). The RUs, unique combinations of soil and land-use data, were characterised in terms of soil properties and land management, reflecting differences in processes such as soil and crop evapotranspiration and surface runoff. The RUs were defined using distributed coverages of input data comprising of 37 soil series and 12 land-use classes. Land-use information was derived from a combination of the LCS88 spatial land cover (MLURI, 1993) and parish census data. Soil and land-use categories making up more than 5% of a sub-catchment, by area, were included in the RU generation, resulting in the formation of 695 RUs, which were used in the SWAT applications.

In total, 12 model applications, where the RUs were parameterised using varying degrees of spatially representative input data, were defined for the Ythan catchment. The applications comprised all possible combinations of two rainfall, three soil, and two land-use coverages, representing distributed and lumped inputs. For rainfall, measurements from eight precipitation stations within and around the Ythan catchment were used to generate an area-weighted average precipitation for the catchment. Two distributed soil data sets were used, soil

*Table 2.* Details of model implementations and performance assessment statistics

| *Implementation* | *Site* | *Model* | *Rainfall* | *Soil* | *Landuse* | $E_{bench}$ *flow* | $E_{bench}$ *N* | *EI flow* | *EI N* | *MAPE flow* | *MAPE N* | *Validation* $E_{bench}$ *flow* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Aire Model 1 | Kildwick | CASCADE | L | L | L | 0.70 | -5.01 | * | * | * | * | 0.71 |
| Aire Model 2 | Kildwick | CASCADE | L | L | D | 0.70 | -8.31 | * | * | * | * | * |
| Aire Model 3 | Kildwick | CASCADE | D | L | L | 0.78 | 0.13 | * | * | * | * | * |
| Aire Model 4 | Kildwick | CASCADE | D | L | D | 0.78 | -0.25 | * | * | * | * | * |
| Aire Model 5 | Kildwick | CASCADE | L | D | L | 0.75 | -0.97 | * | * | * | * | * |
| Aire Model 6 | Kildwick | CASCADE | L | D | D | 0.75 | -2.15 | * | * | * | * | * |
| Aire Model 7 | Kildwick | CASCADE | D | D | L | 0.79 | -0.05 | * | * | * | * | * |
| Aire Model 8 | Kildwick | CASCADE | D | D | D | 0.79 | -0.55 | * | * | * | * | 0.72 |
| Aire Model A | Lemonroyd | CASCADE- QUESTOR | L | L | L | 0.63 | 0.51 | 0.89 | 0.93 | 13.3 | 6.2 | * |
| Aire Model B | Lemonroyd | CASCADE- QUESTOR | D | D | D | 0.44 | 0.49 | 0.89 | 0.93 | 9.7 | 7.3 | * |
| Ythan Model 1 | Ellon | SWAT | L | L | L | 0.49 | -1.88 | 0.93 | 0.82 | 8.4 | 16.7 | 0.46 |
| Ythan Model 2 | Ellon | SWAT | L | L | D | 0.61 | -0.65 | * | * | * | * | * |
| Ythan Model 3 | Ellon | SWAT | D | L | L | 0.69 | -1.84 | * | * | * | * | * |
| Ythan Model 4 | Ellon | SWAT | D | L | D | 0.74 | 0.12 | * | * | * | * | * |
| Ythan Model 5 | Ellon | SWAT | L | HOST | L | 0.65 | 0.32 | * | * | * | * | * |
| Ythan Model 6 | Ellon | SWAT | L | HOST | D | 0.66 | 0.56 | * | * | * | * | * |
| Ythan Model 7 | Ellon | SWAT | D | HOST | L | 0.74 | 0.34 | * | * | * | * | * |
| Ythan Model 8 | Ellon | SWAT | D | HOST | D | 0.77 | 0.50 | * | * | * | * | * |
| Ythan Model 9 | Ellon | SWAT | L | Series | L | 0.61 | 0.38 | * | * | * | * | * |
| Ythan Model 10 | Ellon | SWAT | L | Series | D | 0.69 | 0.55 | * | * | * | * | * |
| Ythan Model 11 | Ellon | SWAT | D | Series | L | 0.73 | 0.37 | * | * | * | * | * |
| Ythan Model 12 | Ellon | SWAT | D | Series | D | 0.77 | 0.54 | 0.90 | 0.92 | 10.2 | 7.5 | 0.71 |

L=lumped, D=Distributed

$E_{bench}$ = benchmark efficiency

EI = cumulative frequency error index

MAPE = mean absolute percentage error

* = statistics not determined

Validation periods were: 1993-1999 (Ythan) and February-December 1987 (Aire)

All other columns of statistics refer to calibration periods: 1984-1992 (Ythan) and 1988-1

series and HOST class, comprising 37 and 17 categories, respectively. The HOST classification provided an intermediate level of soil spatial detail. The dominant soil series, determined by area, was used as the lumped coverage. The distributed land use contained 12 general classes and the dominant class was used for the lumped implementations. Table 2 describes the model implementations driven by different combinations of input data complexity. These model implementations cover a wide range of possible spatial representations.

With the exception of a few changes to nitrogen parameters, Ythan Model 12 (most complex) is the same as the previous application of SWAT in the Ythan (Dilks *et al*., 2004)

For each application the parameters were calibrated manually, adjusted one at a time within specified ranges. In general, they were treated as lumped basin-wide variables during calibration because there was insufficient information about the parameters at the RU level. Discharge calibration focused on curve numbers (SWAT estimates runoff volume using the modified SCS curve number method (Soil Conservation Society, 1972)), influencing the split between surface and groundwater contributions to discharge; an evaporation compensation factor, affecting the depth profile from which the soil's evaporative demand can be met; soil available water capacity, influencing soil water storage; lag times for surface runoff and groundwater contributions, affecting the timing of water inputs to the channel; and groundwater recession, affecting the shape of the baseflow recession curve. Given the focus on N concentration rather than N load in this current application compared to that of Dilks *et al*. (2004), slight modifications were made to some nitrogen parameters to improve predictions of model N concentrations. Groundwater nitrate concentration was set at 6.6 mg $l^{-1}$ N. Denitrification was increased by reducing the soil water content at which SWAT initiates the denitrification process compared to the application of Dilks *et al*. (2004). This change increased sensitivity to soil wetness rather than the rate at which denitrification occurred. Increasing denitrification reduces the large peaks in N concentration predicted in autumn and winter. Specific to the case study, a qualitative judgement was made of the most sensitive parameters that require calibration. These are listed in Table 1.

## Results

The effect of input data complexity on benchmark efficiency in the modelled catchments is illustrated in Table 2. In evaluating the performance of a model with respect to mean daily flow, mean observed flow from the same period was used as a benchmark but, for periodic N measurements, a benchmark was derived using the simulated mean concentration as estimated using the land-cover regression model. At Kildwick and Lemonroyd, in the Aire catchment, the regression estimated means of 1.63 and 3.33 mg N $l^{-1}$ were somewhat lower than the mean long-term observations of 2.31 and 5.43 mg N $l^{-1}$, respectively. Likewise, in the Ythan the mean N concentration estimated was 4.28 mg N $l^{-1}$, as opposed to the observed mean of 6.54 mg $l^{-1}$. Consequently, the N benchmarks defined using simulated means are considerably less demanding than those that would result from applying observed means. This difference is considered in 'choice of model benchmark', below, together with a discussion concerning possible reasons for the underestimations. Model performance was also evaluated with respect to the time-series generated by the least complex applications (Aire Model 1 and Ythan Model 1). Results from this assessment generally failed to yield further insights and as such have not been included in Table 2. Where use of this second benchmark suggested any differences in model performance, these have been highlighted in the text. It is argued that if the ultimate purposes of the modelling approach are: (i) to give estimates of N dynamics in basins that are not part of chemical monitoring programmes and (ii) to predict the impact of future management scenarios on N losses, then, for these case-studies, use of a simulated benchmark represents a fairer test of model performance than a benchmark defined by the mean of long-term observations.

### AIRE AT KILDWICK

The results from CASCADE implementations (Table 2) show that in terms of flow, inclusion of a distributed coverage of rainfall yields greatest improvements in performance. The nature of the soils data used appears to have little impact. Mechanistically, CASCADE simulated flow is insensitive to land-use, as reflected by the $E_{bench}$ (flow) model performance statistics (Hutchins *et al*., 2006). The performance of the flow model appears robust during a period of validation (Feb–Dec 1987, a short period constrained by data availability). During validation the benchmark efficiency of Aire Model 1 improved slightly (0.71) with respect to the calibration period, whereas there is only slight deterioration in the case of Aire Model 8 (0.72). Figure 4 shows the comparison between observed and simulated flows and N on a time-series basis. Only Aire Models 1 and 8 are displayed. Of the two models, Aire Model 8 gives better fit at high flows but overestimates at low and moderate flows.

For N, although the benefits of distributed rainfall are

again seen in terms of $E_{bench}$ (N) values, it appears that increasing the spatial accuracy of other input data sources is detrimental to model performance. Indeed, apart from Aire Model 3, the statistics suggest that the initial benchmark (the simple regression model) performs better. There are large overestimates of N in autumn (Fig. 4). This overestimation is far greater for Aire Model 1 than Aire Model 8 and is least severe for Aire Model 3. These models simulated mean N concentrations of 3.84, 2.47 and 2.83 mg $l^{-1}$ respectively, all considerably higher than the mean of observations. The performance in terms of N could be improved by incorporating in-stream processes such as denitrification, as represented by QUESTOR. Therefore, at this stage, the statistics of N performance have limited meaning in terms of evaluating model suitability.

### AIRE AT LEMONROYD

CASCADE was linked to QUESTOR to simulate flow and N in the Lemonroyd catchment. When including the net effect of point source inputs and discharges, the errors in 5-year total flow (1988–92) were +6% and +10% for Aire Models A and B, respectively. When partitioning the total modelled flows into sources, the importance of point sources (*c*. 25%) and shallow sub-surface flows (c. 60%) are depicted by both models.

Time series of observed and simulated flow and N concentration are shown in Fig. 5. In terms of N, point sources contribute 46% to total modelled inputs to the river system. Following calibration of nitrification and denitrification rates, there is a net retention of modelled N. Optimised values are comfortably within the range of the calibrated values specified by Eatherall *et al.* (1998) working on a more detailed reach-by-reach calibration of the nearby River Wharfe. Delivery of N is estimated as being 85% and 83% for Aire Models A and B respectively. Figure 5 illustrates the ability of the model to capture seasonal variability in N concentration.

Plots of observed and simulated flow duration curves and N cumulative frequency curves are shown in Fig. 6 (a) and (b) respectively. Quantitatively, in terms of cumulative frequency error index, performance is identical although Aire Model B performs better at low flows and worse at high flows. The models overestimate the very highest flows. During 1992 there was consistent underestimation of N by both models (Fig. 5), which may be important in accounting for the overall underestimation of percentile values.

### YTHAN AT ELLON

Model performance statistics indicate that flow simulations were acceptable for all 12 Ythan Model applications. The statistics were in line with those achieved by Perrin *et al*. (2006b) when applying the GR4J rainfall-runoff model (Perrin *et al*., 2003) in the same catchment (calibration: 0.80; validation: 0.60). In the present application, the greatest increases in model performance were achieved using distributed rainfall. Incorporating more detailed land-use information also resulted in slight improvements. The effect of including a spatial representation of soils data was less clear. Model performance increased when using HOST compared to the lumped coverage but no additional improvement was obtained by increasing the detail further and using a soil series input coverage. Performance statistics calculated using Ythan Model 1 as the benchmark displayed an identical pattern to those using the mean observed value.

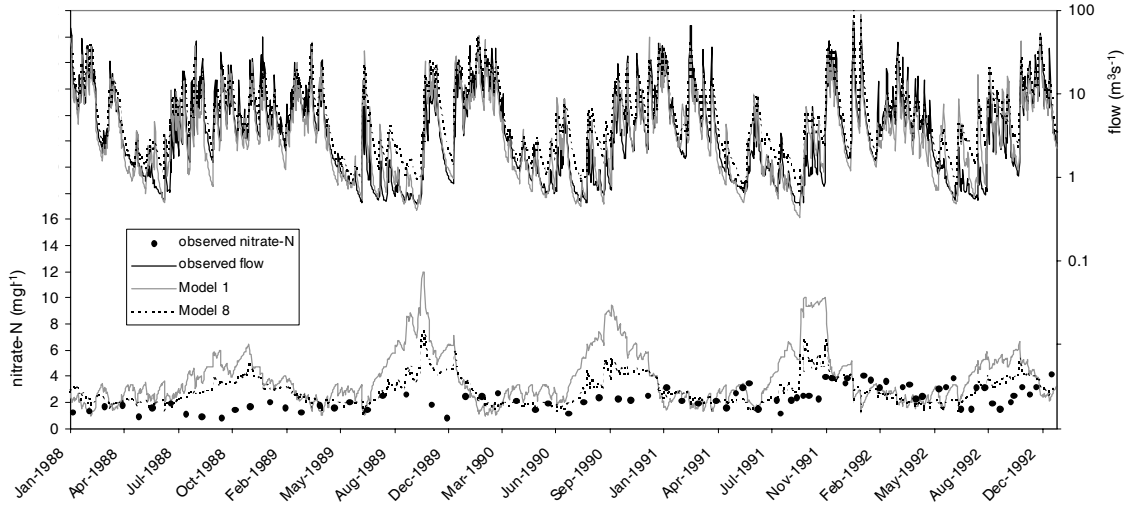


Fig. 4(a). *Observed and simulated flow and nitrate-N concentration at Kildwick for the calibration period (1988-92). Simulations generated by CASCADE. (The right hand y-axis is logarithmic)*

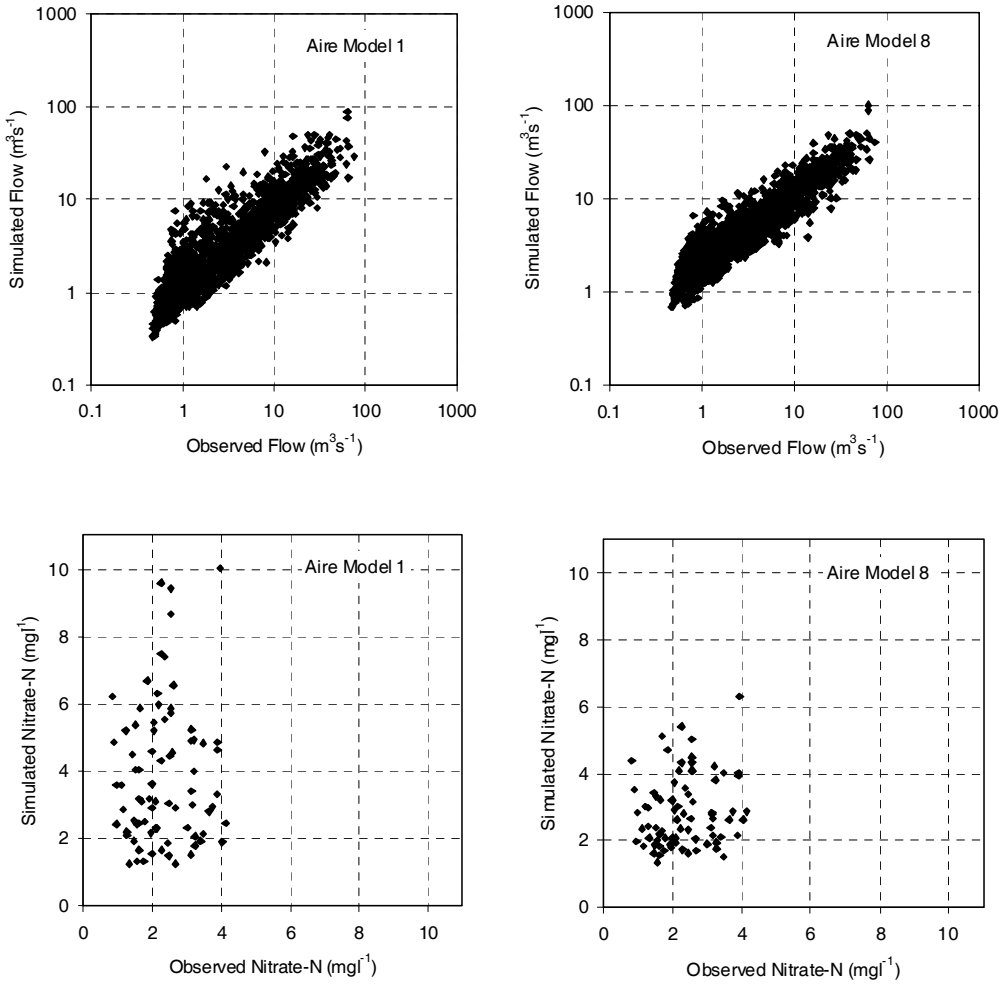


Fig. 4(b). *Scatter plots of model performance*

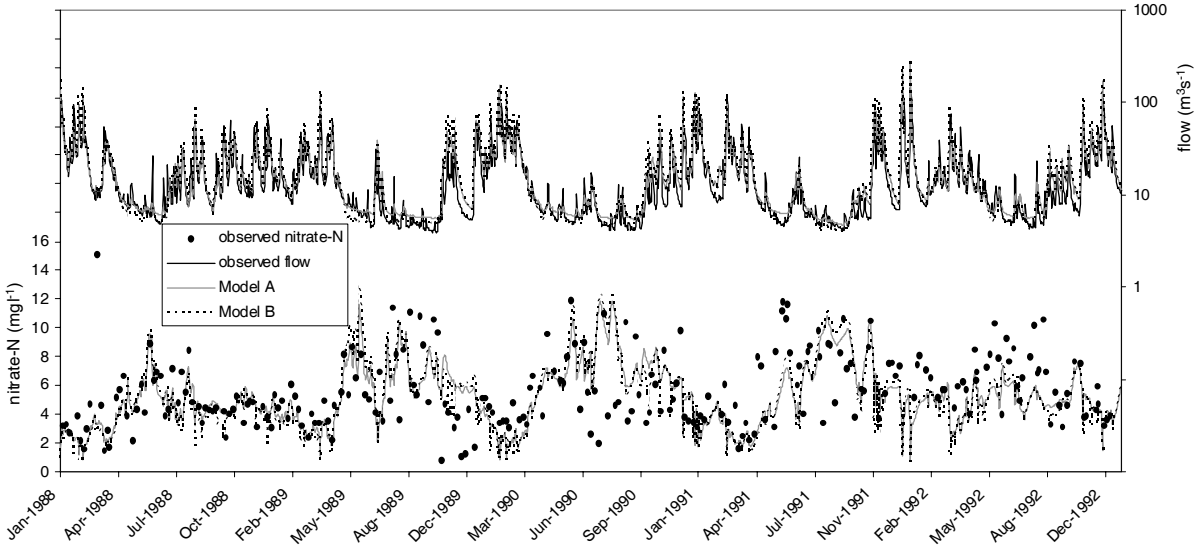


Fig. 5(a). *Observed and simulated flow and nitrate-N concentration at Lemonroyd for the calibration period (1988-92). Simulations generated by CASCADE-QUESTOR linked model (the right-hand axis is logarithmic).*

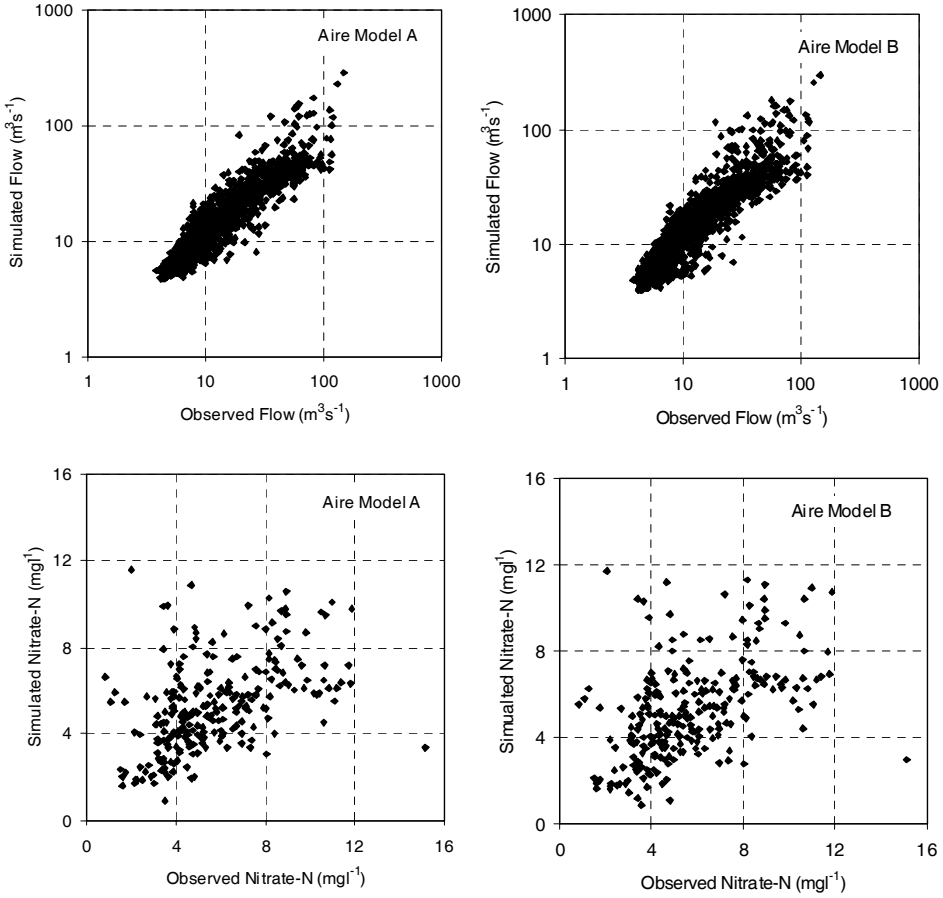


Fig. 5(b). *Scatter plots of model performance*

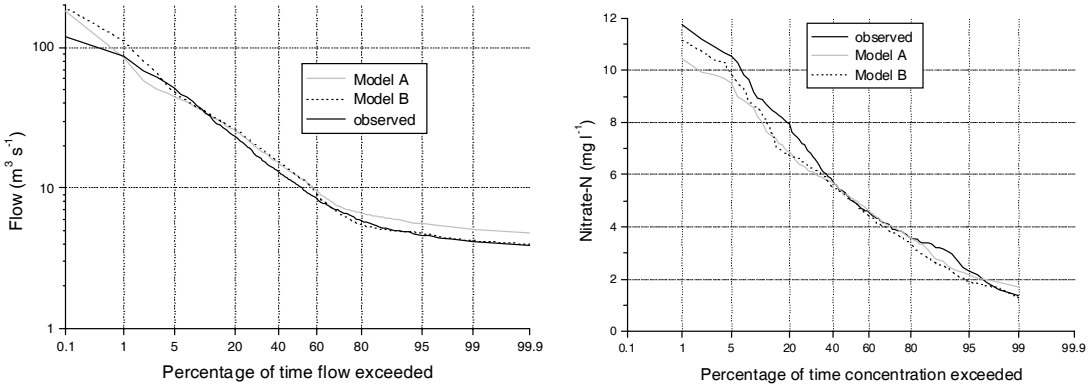


Fig. 6. *Observed and simulated (CASCADE-QUESTOR) cumulative frequency for the calibration period (1988–92) of (a) flow and (b) nitrate-N at Lemonroyd*

Positive values of these $E_{bench}$ statistics indicated that Ythan Models 2 to 12 performed better than Ythan Model 1. The performance of the flow models appeared robust during a validation period (1993–1999) with only slight deterioration in the $E_{bench}$ statistics, that varied between 0.46 (Ythan Model 1) and 0.71 (Ythan Model 12).

In terms of N, performance statistics indicate that increasing model complexity, either through use of distributed soils (HOST) information or by including spatial variability of land use, improved model performance, whilst including spatially distributed rainfall made little difference. Given that the modelling objective was to simulate N at the

catchment outlet, Ythan Model 6 (lumped rainfall, HOST soil, distributed land use), which achieved the highest performance, was considered the best of the 12 SWAT applications. Average daily N simulations from Ythan Models 5 to 12 were the closest to observed mean values, with mean simulations notably higher for Ythan Models 1 to 4, which used the lumped soil input coverage.

Figure 7 presents timeseries of observed and simulated flow and N for Ythan Models 1 (least complex) and 12 (most complex). Both models represented the observed flow successfully although some mismatches are apparent at both the highest and lowest flows, a result also highlighted by the flow duration curves presented in Fig. 8. The flow performance statistic for Ythan Model 1 was heavily impacted by a small number of over-estimated events. The N time series, on the other hand, exhibited a generally poor

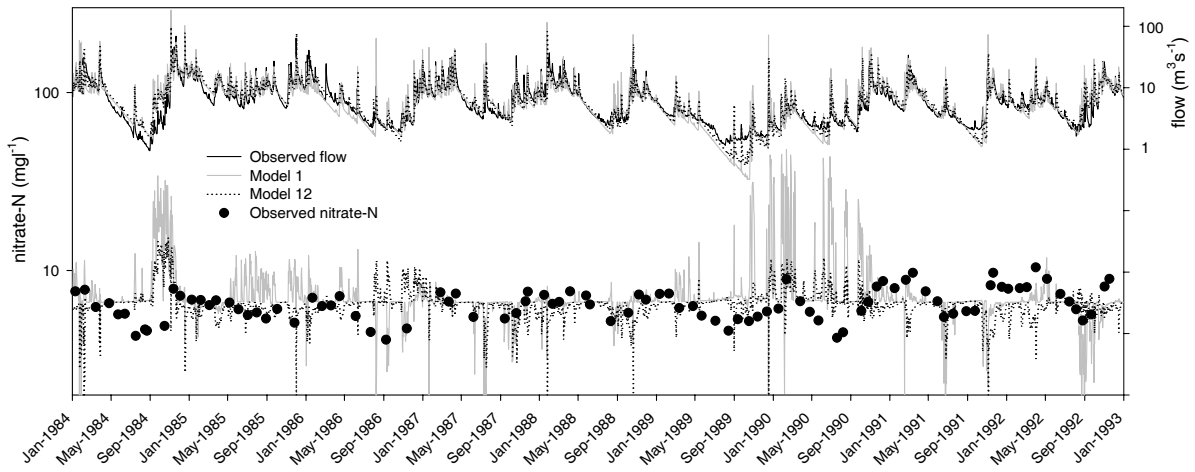


Fig. 7(a). *Observed and simulated flow and nitrate-N concentration at Ellon. Simulations generated using SWAT model (y-axes are logarithmic)*

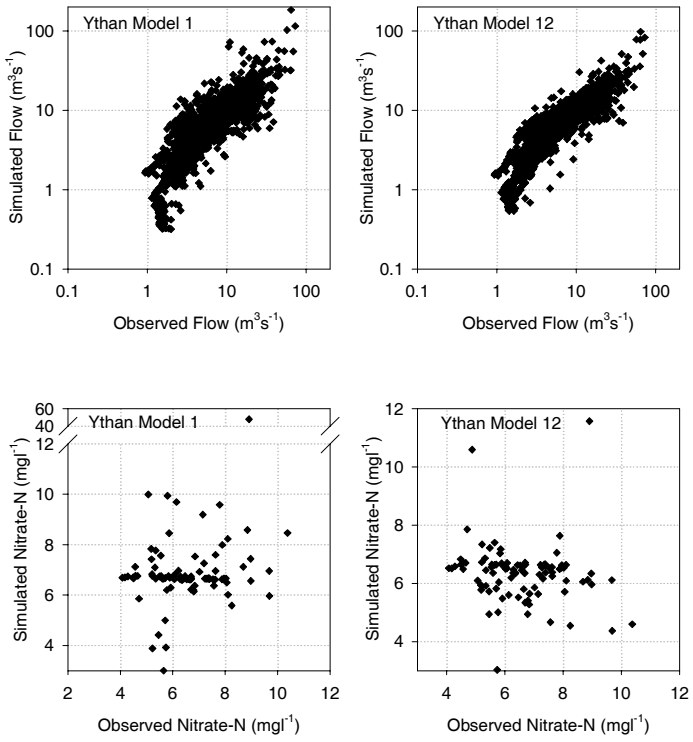


Fig. 7(b). *Scatter plots of model performance*

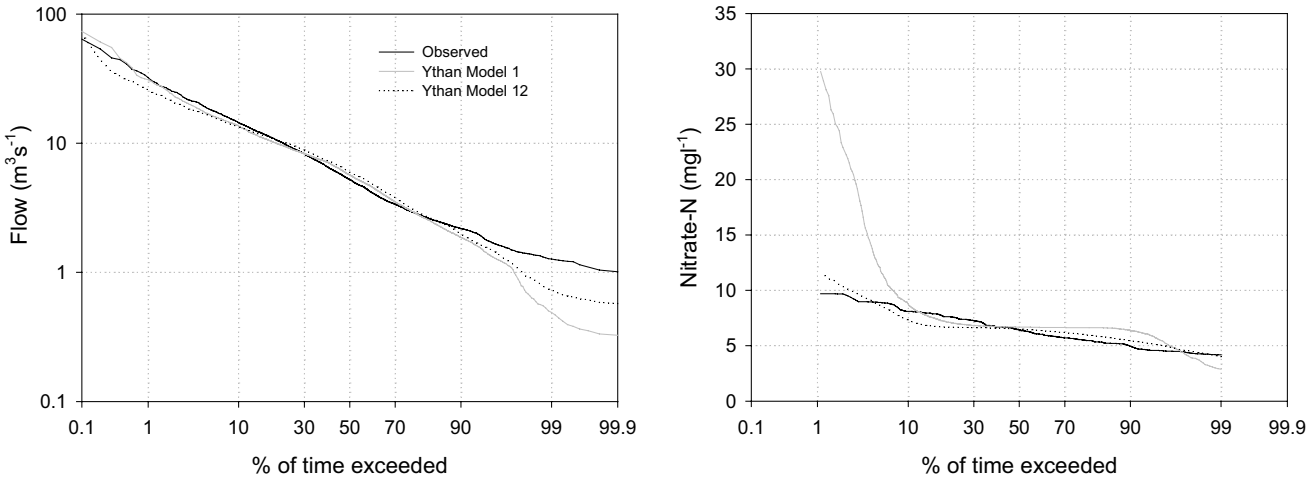


Fig. 8. *Observed and simulated (SWAT) cumulative frequency for the period 1984 – 1992 of (a) flow and (b) nitrate-N at Ellon*

fit to observed data (Figs. 7 and 8). Notably higher peaks of N concentration were simulated by Ythan Model 1 compared to Ythan Model 12 for a number of periods. Other periods, where the simulated data series of Ythan Models 1 and 12 exhibited conflicting peaks and troughs, were also seen. The simulation of large N peaks, considerably greater than observed data, by Ythan Model 1 is clear, as is the dominance of groundwater N contributions. The combination of clustering of model simulations around the observed mean (Figs. 7 and 8) and a poor benchmark results in falsely high $E_{bench}$ (N) statistics in the Ythan.

## Discussion

Two process-based models have been applied in two UK catchments, CASCADE-QUESTOR in the Aire and SWAT in the Ythan. Several implementations of each model, covering a wide range of possible spatial representations of rainfall, soil and land-use data, have been examined to assess the effect of input data complexity on model performance. Results indicate that model performance may be improved by increasing the spatial representation of some but not all of the tested input coverages; different coverages influenced different model outputs.

### PERFORMANCE ASSESSMENT: MODELLING FLOW

The partitioning of flow, as estimated by the model implementations at Lemonroyd and Ellon, is consistent with conceptual models of soil hydrological response represented in the HOST classification (Boorman *et al.*, 1995). The Aire is dominated by soils from HOST class 24, soils exhibiting gleyed horizons close to the surface underlain by strata of low-permeability, prone to seasonal waterlogging, reflecting the importance of shallow sub-surface flows (approx. 30%) in the region. The Ythan, on the other hand, is characterised by freely draining soils developed on loamy drifts underlain by either porous or hard coherent rocks at depths of greater than one metre, classified as HOST classes 6 and 17, respectively (Boorman *et al*., 1995). These soils reflect the importance of the slow hydrograph response in this baseflow-dominated catchment.

Difficulties were encountered in both case-studies when modelling the extremes of the flow distribution. In the Aire, CASCADE-QUESTOR overestimated the highest flows, whereas in the Ythan the lowest flows are underestimated by SWAT. Particular difficulties simulating low flows in the Ythan during the second half of 1989, a year with low precipitation, suggest that the model, as it is currently set up, may not reproduce observed flows well under unusually dry conditions. Representing the extremes of distributions is a common difficulty in model applications. The differing problems encountered by these two case-studies may reflect the different choice of model or the contrasting nature of soil hydrology in the regions. Additionally, in the Aire case-study, flows in the middle of the distribution were slightly over-estimated, even by Aire Model B (most complex). A more detailed representation of the spatial variability in rainfall may help alleviate this problem.

### PERFORMANCE ASSESSMENT: MODELLING NITRATE CONCENTRATION

Diffuse sources of N predominate in the Aire Kildwick catchment and the Ythan catchment where agriculture is the dominant land use. Typically, observed data from such

agricultural catchments will broadly reflect seasonal patterns of N accumulation, with the greatest N concentrations occurring in autumn, when peaks result from the leaching of N from crop residues following harvest. Such patterns were seen in observations from Ellon (Ythan) but not from Kildwick (Aire). Conversely, modelled time series for Kildwick indicated the occurrence of autumn maxima whereas the SWAT output for the Ythan failed to capture this seasonality. As a result of the differences between observed and simulated data, the visual and statistical fits derived from CASCADE at Kildwick were poor. However, there was scope to improve N performance by including in-stream processes represented in QUESTOR (nitrification and denitrification). Aire Model 3 produced the best $E_{bench}$ (N) value and in terms of flow was as good as the most complex implementation (Aire Model 8). Ythan simulations, on the other hand, consistently varied around 6.6 mg $l^{-1}$, with periods of peaks and troughs associated with rainfall events. Differences in the responses to these rainfall events reflect changes in the relative proportions of water from different sources within the catchment. The modelled time series shows stable concentrations at baseflow conditions. This reflects the dominance of groundwater in the Ythan region and the simple representation of groundwater processes by the SWAT model, which assumes a constant groundwater N concentration. In-stream processes have not been included in the current SWAT applications due to problems associated with their calibration. However, their inclusion could potentially improve N simulations in the Ythan. Similar problems calibrating in-stream processes have also been experienced by other users (Arnold and Fohrer, 2005).

The Aire Lemonroyd catchment contains many point sources in addition to diffuse sources of N. Here, the highest concentrations were observed during summer low flows when the relative contribution of high-concentration point sources is at its greatest. The lowest concentrations were observed during winter and early spring when diffuse sources dominate volumetrically and the autumn flushes of N mobilised from crop residues have started to subside. Visually, these seasonal trends were captured adequately by the linked CASCADE-QUESTOR model as illustrated in the time series plot (Fig. 5) and values of $E_{bench}$ (N).

The CASCADE-QUESTOR implementations, Aire Models A and B, produced similar $E_{bench}$ (N) values (approx. 0.5). For the SWAT Ythan models, Ythan Model 6 achieved the best performance for N prediction although a drop in flow performance was seen in comparison to Ythan Model 12 (most complex).

## IMPACT OF MODEL INPUT DATA COMPLEXITY

For the two applications, certain similarities were seen when assessing the impact of spatial input data complexity on model performance. Using a distributed rainfall coverage clearly improved flow simulation, even in the Ythan where spatial variability of annual mean precipitation across the catchment is small. The benefit of increasing land-use information was minimal in both catchments. The use of distributed N accumulations in the Aire catchment yielded, in general, better performance at low flows, when high concentrations of N are likely to be particularly significant ecologically. Similar improvements were not seen in the Ythan where at low flow groundwater is the main source of N.

Increasing the spatial representation of soils benefited the Ythan application but provided minimal utility in the Aire in terms of either flow or N prediction. In the Ythan the change from lumped to distributed, using the HOST classification, exerted a notable change in simulated mean annual N concentration and on qualitative performance criteria. Little was gained by further increasing the distribution using soil series. Under the lumped soil coverage (Ythan Models 1 to 4) considerable peaks in N concentration, not seen in observed data, were associated with both the spring and autumn seasons (see Fig. 7). These differences in the extremes of simulated N concentration, apparent for the various Ythan implementations, are thought to be associated with differences in the level of denitrification represented in the model applications. The denitrification process within SWAT is triggered once the soil water content exceeds a specified multiple of field capacity. The impact of this is that less denitrification will be modelled in the more freely-draining soils compared to the less freely-draining soils which spend a greater proportion of time above the threshold soil moisture level. The lumped soil coverage (Ythan Models 1 to 4) was represented by the dominant soil series in the Ythan catchment. This soil is more freely draining than several other soils in the region; consequently the total amount of denitrification simulated in the catchment is likely to be less than when a distributed soil coverage is used. A better representation of the lumped soil may be an average soil with parameter values calculated as an area-weighted average of the soil series or soil HOST class values. The use of catchment-specific soil parameterisation was not tested in the Aire application. The Ythan results suggest there would be little benefit to be gained with their inclusion. Cotter *et al*. (2003), also working with SWAT, suggest there is unlikely to be significant benefit in using increasingly detailed soils data at spatial resolutions finer than 1 $km^2$.

The performance of Aire Models A (least complex) and B (most complex) appeared very similar in terms of N. In this

respect, the influence of point source discharges, abstractions and in-stream processes dominates over the effects of the strikingly different levels of input data complexity used, as corroborated by Deflandre *et al*. (2006). Although in terms of $E_{bench}$ (flow), Aire Model A outperformed Aire Model B, the latter model performed better at low flows (see Fig. 6a) as reflected in the MAPE flow statistics (Table 2). Consequently, the high N concentrations occurring at these low flows were better simulated (Fig. 6b). The models underestimated throughout much of the final year (1992). Decadal monitoring data from the Aire at Lemonroyd suggest an increase in mean N concentration of approximately 0.3 mg $l^{-1}$ per annum, between 1985 and 1995. The reasons are not obvious, and their exploration is outside the scope of this paper.

Overall, the study determined that unless the ultimate purpose of the model application was to inform specific spatially-targeted mitigation options, there was little benefit to model performance by distributing all three (rainfall, soil, land-use) input coverages. Importantly, however, for the modelling approaches chosen in both case studies, any improvement in model performance gained through the use of distributed rather than lumped inputs is not offset by the consequences of over-parameterisation as there is no increase in the number of parameters requiring optimisation. Nevertheless, equifinality, or non-uniqueness of best-fit, where an optimal performance statistic may be obtained with more than one set of calibrated parameter values, reflects over-parameterisation. This remains an issue to be considered when evaluating the utility of the modelling approaches in these case studies.

Separate and ultimately more fundamental issues that arise from the performance assessment are the reasons as to why, in both case-studies, the improvements gained from increasing the spatial resolution of input data are only marginal. This may reflect inadequacies in the model structures (shortcomings in process representation) manifested as an inability to take advantage of the greater accuracy of the input data. On the other hand, parish-level agricultural census data used to specify the land-use distribution may be spatially insufficiently precise to support modelling at the HRU/RU resolutions adopted. Furthermore, catchments act as filters, dampening the effects of inputs in time and space and lessening the effect of improved inputs. The influence of input resolution may also be impaired by the pragmatic decision to use a model time-step of 1 day which may be close to the hydrological response times of the catchments.

### CHOICE OF MODEL BENCHMARK

Quantitative measures of performance suggest that both in terms of time-series N concentration responses ($E_{bench}$) and of quantitative measures of the simulated N distribution (e.g. MAPE), the Aire simulations were more satisfactory than the Ythan when using lumped representations of inputs. For fully distributed inputs there is little difference in performance between the two case studies. Moreover, the values of $E_{bench}$ (N) are heavily influenced by the standard used (from the regression model). In both the Aire and the Ythan, the regression model underestimated mean N concentration. If, instead, the observed mean N concentration was used, an appreciable reduction in the values of $E_{bench}$ (N) would have occurred; to values of 0.08 and 0.05 respectively for Aire Models A and B and to negative values in the case of all the Ythan models. Therefore, performance benchmarking reveals evidence that the model applications have predictive power only when evaluation is made against a simulated rather than an observed mean N concentration. Variability in the performance of a land-cover-based regression model for N concentration may be due to a variety of causes. The density of grazing livestock, or the N input from atmospheric sources, are examples of variables (i) to which the regression model is insensitive, so affecting its performance, and (ii) that, if not included in process models, compromise their ability to generate good simulations. Hence, use of the regression model in the $E_{bench}$ (N) statistic, as illustrated here, serves to provide an initial benchmark of the level of performance that might be expected of a more complex process model.

## Conclusions

Two case studies have been undertaken on different catchments, assessing the effect of data complexity on model performance using contrasting model codes. Despite the differences, and the restrictions imposed by assessing just two studies, some general conclusions arise for consideration when undertaking such catchment-scale modelling of flow and N concentration. Aggregations of each of the two sets of model results reveal similar features. Both case studies indicate the level of benefit to be gained in using distributed input data. The importance of distributed rainfall data, especially for flow simulation, is notable as is the adequacy in the use of the HOST classification rather than series-specific information when deriving soil parameters. The two modelling studies also showed that the sensitivity of individual parameters representing diffuse pollution processes may be greatly diminished by consequences of in-river processes and mixing of sources, especially point

discharges, prevalent in catchments of significant size.

The minimum amount of input data required by a model is determined by the nature of the application. Implementations of models using this minimum input requirement can be considered as providing benchmark simulations. The use of input data to a level above and beyond the minimum may or may not yield significant performance benefits. Specific applications are not prescribed here and, hence, benchmark performance is not fixed but results generated by two models have been assessed to illustrate the likely benefits of additional data, and how these benefits may differ in their nature between model codes and catchments.

As a general recommendation, judgement on an appropriate level of input data complexity for a given model application should be based on a combination of multiple quantitative performance criteria and qualitative assessment. Any benefits gained, or not, from using distributed input data, may reflect both on the quality of the input data itself and on the quality of the model representation. It is important that these issues are considered in a model application. The robustness of model performance during validation periods should also be considered. Whether or not a specific model implementation is deemed 'good enough' should be informed partly through value judgements as well as by a range of quantitative criteria. All too often model applications are justified purely on the basis of a single performance criterion; value judgments should be set in the context of the application itself and broader management requirements of end-users.

## Acknowledgements

The authors acknowledge financial support for the project *Benchmark Models for the Water Framework Directive* from the European Commission (EK1-CT2001-00093), the Natural Environment Research Council (NERC), and the Scottish Executive Environment and Rural Affairs Department (SEERAD). David M. Cooper (CEH, Wallingford) provided a version of the CASCADE model code which was adapted for use in this case study. The authors thank the Environment Agency (EA) and the Scottish Environmental Protection Agency (SEPA) for providing monitoring data. The analysis undertaken in this paper was also stimulated by a new initiative, the Catchment Hydrology, Resources, Economics and Management (ChREAM) project, funded under the joint ESRC, BBSRC and NERC Rural Economy and Land Use (RELU) programme.

## References

Addiscott, T., Smith, J. and Bradbury, N., 1995. Critical evaluation of models and their parameters. *J. Environ. Qual.*, **24,** 803–807.

Arnold, J.G. and Fohrer, N., 2005. SWAT2000: current capabilities and research opportunities in applied watershed modeling. *Hydrol. Process.,* **19,** 563–572.

Arnold, J.G., Srinivasan, R., Muttiah, R.S. and Williams, J.R., 1998. Large area hydrological modelling and assessment. Part I: Model Development. *J. Amer. Water Resour. Assoc.*, **34,** 73–89.

Boorman, D.B., 2003. LOIS in-stream water quality modelling. Part 1: Catchments and methods. *Sci. Total Envir.*, **314/316,** 379–395.

Boorman, D.B., Hollis, J.M. and Lilly, A., 1995. *Hydrology of soil types: a hydrologically-based classification of the soils of the United Kingdom*. Report No. 126, Institute of Hydrology, Wallingford, UK.

Cooper D.M. and Naden, P.S., 1998. Approaches to delivery modeling in LOIS. *Sci. Total Envir.*, **210/211,** 483–498.

Cotter, A.S., Chaubey, I., Costello, T.A., Soerens, T.S. and Nelson, M.A., 2003. Water quality model output uncertainty as affected by spatial resolution of input data. *J. Amer. Water Resour. Assoc.*, **39,** 977–986.

Davies, H.N. and Neal, C., 2004. GIS-based methodologies for assessing nitrate, nitrite and ammonium distributions across a major UK basin, the Humber. *Hydrol. Earth Syst. Sci.*, **8.** 823–833.

Deflandre, A., Williams, R.J., Elorza, F.J., Mira, J. and Boorman, D.B., 2006. Analysis of the QUESTOR water quality model using a Fourier Amplitude Sensitivity Test (FAST) for two UK Rivers. *Sci. Total Envir.,* (in press).

Dilks, C.F., Dunn, S.M. and Ferrier, R.C., 2004. Evaluation of a model benchmarking procedure through application of the Soil Water Assessment Tool in the Ythan catchment, UK. In: *Science and Practice for the 21st Century,* Proc. BHS Int. Conf., London, July 2004. **I**, 260–267.

Dunn, S.M., Mc Alister, E. and Ferrier, R.C., 1998. Development and application of a distibuted catchment-scale hydrological model for the river Ythan, NE Scotland. *Hydrol. Process.,* **12,** 401–416.

Eatherall, A., Boorman, D.B., Williams, R.J. and Kowe, R., 1998. Modelling in-stream water quality in LOIS. *Sci. Total Envir.,* **210/211,** 499–517.

Fuller, R.M., Groom, G.B. and Jones, A.R., 1994. The Land Cover Map of Great Britain: an automated classification of Landsat Thematic Mapper data. *Photogram. Eng. Remote Sens.*, **60,** 553–562.

Fuller, R.M., Smith, G.M., Sanderson, J.M., Hill, R.A. and Thompson, A.G., 2002. The UK Land Cover Map 2000: Construction of a parcel-based vector map from satellite images. *Cartographic J.*, **39,** 15–25.

Hallett, S.H., Jones, R.J.A. and Keay, C.A., 1993. SEISMIC: a spatial environmental information system for modelling the impact of chemicals. In: *Environmental Modelling: The Next 10 Years,* A.R.D. Stebbing, K. Travis and P. Matthieson, (Eds.), Report of a Symposium at The Society of Chemical Industry, London, 16 Dec 1992. 40–49.

Hutchins, M.G., Deflandre, A. and Boorman, D.B., 2006. Performance benchmarking linked diffuse pollution and in-stream water quality models. *Arch. Hydrobiol. - Large Rivers.* (in press)

Jones, S.B., 1983. *The estimation of catchment average point rainfall.* Report No. 87, Institute of Hydrology, Wallingford, UK.

Kämäri, J., Boorman, D., Icke, J., Perrin, C., Candela, L., Elorza, F., Ferrier, R., Bakken, T. and Hutchins, M., 2006. Process for benchmarking models: dialogue between water managers and modellers. *Arch. Hydrobiol. - Large Rivers*. (in press)

Kirchner, J.W., Hooper, R.P., Kendall, C., Neal, C. and Leavesley, G., 1996. Testing and validating environmental models. *Sci. Total Envir.,* **183,** 33–47.

MLURI, 1993. *The Land Cover of Scotland*. Final report on the Land Cover of Scotland project. Macaulay Land Use Research Institute, Aberdeen, Scotland.

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

Perrin, C., Michel, C. and Andreassian, V., 2003. Improvement of a parsimonious model for streamflow simulation. *J. Hydrol.*, **279,** 275–289.

Perrin, C., Andreassian, V. and Michel, C., 2006a. Simple benchmark models as a basis for criteria of model efficiency. *Arch. Hydrobiol. - Large Rivers*. (in press)

Perrin, C., Dilks, C., Barlund, I., Payan, J.L. and Andreassian, V., 2006b. Use of simple rainfall-runoff models as a baseline for the benchmarking of the hydrological component of complex catchment models. *Arch. Hydrobiol. - Large Rivers*. (in press)

Refsgaard, J.C. and Knudsen, J., 1996. Operational validation and intercomparison of different types of hydrological models. *Water Resour. Res.*, **32,** 2189–2202.

Seibert, J., 2001. On the need for benchmarks in hydrological modelling. *Hydrol. Process.*, **15,** 1063–1064.

Soil Conservation Society, 1972. Section 4: Hydrology, In: *National Engineering Handbook*, SCS, Washington DC, USA.

Van Rompaey, A. and Govers, G., 2002. Data quality and model complexity for regional scale soil erosion prediction. *Int. J. Geogr. Inf. Sci.*, **16,** 663–680.

Wagenet, R.J. And Hutson, J.L., 1996. Scale-dependency of solute transport modelling/GIS applications. *J. Environ. Qual.*, **25,** 499–510.

Wheater, H.S. and Beck, M.B., 1995. Modelling upland stream water quality: process identification and prediction uncertainty. In: *Solute Modelling in Catchment Systems*, S. Trudgill (Ed.) Wiley, Chichester, UK. 305–324.