EUROHARP 8-2004

Modelling approaches: Model parameterisation, calibration and performance assessment methods in the EUROHARP project

Editors
Martyn Silgram, ADAS, United Kingdom
Oscar F. Schoumans, ALTERRA, the Netherlands

EUROHARP Report no 8
Modelling approaches: Model parameterisation, calibration and performance assessment methods in the EUROHARP project

Silgram, M., and Schoumans, O.F. (eds.), 2004. “Modelling approaches: Model parameterisation, calibration and performance assessment methods in the EUROHARP project”. EUROHARP report 8-2004, NIVA report SNO 4740-2003, Oslo, 18 pp.

A publication developed within Work Packages 3 and 4 of the EUROHARP project: ″Application and evaluation of phosphorus and nitrogen loss quantification tools”.

Published by: Norwegian Institute for Water Research (NIVA), Norway

Editors: Oscar Schoumans, ALTERRA, The Netherlands - Martyn Silgram, ADAS, United Kingdom

Authors: Hans Estrup Andersen, NERI, Denmark - Steven Anthony, ADAS, United Kingdom - Berit Arheimer, SMHI, Sweden - Alan Barr, KMM, Ireland - Horst Behrendt, FV-IGB, Germany - Fayçal Bouraoui, EC-JRC, Italy - Helene Ejhed, IVL, Sweden - Piet Groenendijk, ALTERRA, The Netherlands - Michel Jeuken TERRA, The Netherlands- Holger Johnsson, SLU, Sweden - Brian Kronvang, NERI, Denmark- Guillaume Le Gall, BETURE-CEREC, France - Alison Murdock, KMM, Ireland - Antonio Lo Porto, IRSA-CNR, Italy - Laura Fawcett, ADAS, United Kingdom - Oscar Schoumans, ALTERRA, The Netherlands - Martyn Silgram, ADAS, United Kingdom - Robert Smit, ALTERRA, The Netherlands - Ekaterini Varanou, NTUA, Greece - Ulrike Zweynert, FV-IGB, Germany.

Number Printed: 60

ISBN: 82-577-4491-3

Pages:

Date of printing: February 2004

The Project: The EC funded EUROHARP project encompasses 22 research institutes from 17 European countries (2002-2005). The overall objective of the EUROHARP work is to provide end-users with guidance for an appropriate choice of quantification tools to satisfy existing European requirements on harmonisation and transparency for quantifying diffuse nutrient losses, e.g. to facilitate the implementation of the Water Framework Directive and the Nitrates Directive. The project includes both the assessment of the performance of individual models and the applicability of the same models in catchments with different data availability and environmental condition throughout Europe. The basis for the performance and applicability studies is the compilation of a harmonised GIS/database for all catchment data and the analysis of these data (trends, watercourse retention).

Key words: Model assessment. Nitrogen. Phosphorus. Diffuse losses. Water Framework Directive. Nitrates Directive.

The report may also be downloaded from the EUROHARP web site: www.euroharp.org.
Editors, authors and NIVA assume no liability for any loss resulting from the use of this report.

Contents

1. Introduction

Due to the wide range of different modelling tools used in the project and presence of significant earlier work focusing exclusively on modelling approaches, it was decided not to attempt to be unrealistically prescriptive in the detailed elements of individual modelling approaches. However, there is still a need to document our common understanding regarding the basic approaches to be adopted in the EUROHARP project.

As a result, this report is intended as a collation of key text developed during the course of the first two years of the EUROHARP project, together with a reiteration of the agreements reached at several project meetings (Meze, Berlin, Wolverhampton, York, Wageningen, Oslo) during 2002 and 2003. The purpose is simply to bring together all this relevant information in one place to serve as a reference document during the modelling exercise, to accompany the model review document circulated separately. This document may be subsequently revised and updated, access the EUROHARP website (www.euroharp.org) for the latest version.

This report is separated into several sections covering agreements reached on different aspects of the catchment modelling strategy (Section 2), edited versions of the validation methods discussed at project meetings (Section 3 – a specific EUROHARP deliverable to the EC), and an abridged version of the HARP guideline on Flow Normalisation developed within the HARP framework (Section 4). A useful document on "Good Modelling Practice" developed in the Netherlands is also noted and is available separately as an Adobe pdf file. Any updates to these files will be posted on the appropriate part of the project website http://www.euroharp.org.

2. Main Agreements

2.1 Catchment-model selection

EUROHARP requires all nine quantification tools to be applied to the three core catchments: the Vansjo-Hobol (Norway), the Yorkshire Ouse (England), and the Enza (Italy).

Additional catchment to be modelled in WP5 were also selected for each of the nine quantification tools. These catchments are identified here for completeness:
ANIMO: 1. Denmark; 2. Czech Republic; 3. Germany/Netherlands
N-LESS: 1. Finland; 2. Luxembourg; 3. Spain
TRK: 1. Germany / Netherlands; 2. Hungary; 3. France
EVENFLOW: 1. Germany; 2. Czech Republic; 3. Greece
REALTA: 1. Germany; 2. Lithuania; 3. France
MONERIS: 1. Lithuania; 2. Ireland; 3. Greece
SWAT: 1. Sweden; 2. Austria; 3. Spain
NOPOLU: all 17 catchments
Source Appointment: all 17 catchments
MONERIS will also be applied to catchments in Austria, Hungary, Luxembourg, Germany, Germany/Netherlands, and Czech Republic (in other projects)

In the case that these additional catchments have insufficient data of an appropriate type, or cannot be modelled in this project for other reasons, three reserve catchments were also selected for each of the models:
ANIMO: 4. Sweden; 5. Austria; 6. Luxembourg
N-LESS: 4. Ireland; 5. Lithuania; 6. Sweden
TRK: 4. Austria; 5. Denmark; 6. Finland
EVENFLOW: 4. Finland; 5. Lithuania; 6. Ireland
REALTA: 4. Hungary; 5. Germany / Netherlands; 6. Denmark
MONERIS: 4. Denmark; 5. Sweden; 6. Spain
SWAT: 4. Luxembourg; 5. Denmark; 6. Hungary
NOPOLU: Not necessary as all 17 will be modelled

2.2 Model Review

- To aid transparency in the modelling process, a short model description, extended model description, and accompanying documentation concerning the theory, key equations, and application of each of the nine quantification tools (models) have been developed and are available for download from the EUROHARP website. A separate model review document is available for this purpose.

- A number of new criteria for evaluation of the candidate models were proposed, requiring consideration of additional factors when assessing model characteristics:
  - Intended purpose/status and history of model application (maturity)
  - Dependencies on previous models (scientific evolution)
  - Operational experience and skills requirement of users
  - Participation in previous model comparison studies
  - Existing sensitivity analyses
  - Historic data requirements for initialisation
  - Sub-modules that can be independently checked

2.3 Good Modelling Practice

- Alterra has provided a substantial "Good Modelling Practice" document developed in the Netherlands which covers many of the issues faced by catchment modellers in the EUROHARP project in considerable detail. The document is available from the EUROHARP website. All modellers are urged to consult this document in support of their activities. The document is also available at the following addresses:
  http://www.info.wau/nl/research projets/pub-pdf/gmp.pdf
  and from
  http://www.harmoniqua.org
- Linking with the above document, all modellers agreed to document the sources of all parameters used in modelling work, to provide a transparent audit trail to facilitate subsequent intercomparisons of model performance and the writing of scientific papers.

2.4 Model Parameterisation Issues

- Some models include their own rainfall interpolation routines which may be integrated into the model structure such that it is not possible to bypass them and use alternative externally generated datasets. This means that the preferred option of all models being applied using the same interpolated rainfall input data for a particular catchment is not possible. As a result, modellers agreed to retain the option of using rainfall data on a point basis or on an interpolated grid basis.
- Modellers agreed to conceptually divide parameters into the following groups:
  - Parameters based on readily available field and catchment information (measured parameters)
  - Parameters based on relevant *published* literature sources or appropriate default values
  - Parameters based on *transfer functions* from available data
  - Parameters *fitted* during the calibration process

The extent to which models rely on the last category (fitted parameters) will be taken into account as part of the assessment of model performance. The model parameters in each of the above groups, the values used, and the source used for identifying the value used, should all be documented (see Section 2.3).

2.5 Calibration issues

- Given the need to assess performance of both sub-annual and annual timestep models, modellers agreed the calibration process should be:
  - split in time by dividing each subcatchment timeseries in half with the first half for calibration and second half for validation, and also
  - split in space by using 3-5 subcatchment monitoring stations in addition to the main catchment outlet.

The precise subcatchment boundaries used in the calibration in space approach will be dependent on the availability of subcatchment monitoring data and have been decided on a catchment basis.
All modelling institutes have agreed on the same subcatchment gauging stations and associated land boundaries to use (see later section).

- Three options were identified for estimating riverine load for comparison against monitoring data:
  - According to the OSPAR HARP guideline on flow normalisation (see later section)
  - Linear interpolation method
  - Catchment owners' own method (documentation needed)

2.6 Modelling approach

- Model applications will typically follow the order Norwegian catchment, English catchment, Italian catchment (except for NL-CAT and SWAT, which will be applied in parallel).
- Both the prediction of concentrations (Water Framework Directive) and loads (OSPAR) were identified as issues of interest to end-users of model results from this project.
- Institutes agreed to focus on annual loads (all models) supplemented by subannual load assessments e.g. weekly, monthly (some models).
- **Three stages have been agreed for testing the models: (i) blind test, (ii) calibration, (iii) validation → all statistics.** The Blind test is not a requirement of the proposal – it is an additional exercise suggested during project meetings. The blind test involves model predictions without detailed calibration timeseries, and allow us to explore the extent to which some models are capable of producing reasonable predictions without recourse to detailed (and costly) site-specific calibration. Gauged river monitoring datasets will be split in time and space for performance assessment purposes i.e. modelling subcatchment outlets as well as the main catchment outlet, and using (usually the first) half of the timeseries for calibration and half for validation. The agreed procedure is as follows:

  1. Blind test: Work Package 2 (WP2) leader releases one year of data now (to allow modellers to check structure). On 15 March 03, WP2 leader will provide all INPUT data (climate, soils, geology, agriculture, etc) for each of the catchments and subcatchments
  * EveNFlow, SWAT, NL-CAT, TRK. Deadline for model results end May
  Submit blind test results to WP2 leader (who forwards them to WP3/4/5 leaders);
  WP2 leader will send half of the river data for calibration (stage 2 below)
  * MONERIS, NLESS, NOPOLU need flow data as input → no blind test possible
  * REALTA needs loads for subcatchments (for each risk class) → no blind test possible

  2. Calibration: The WP2 leader will provide half of the output (river flow and concentration) data for all catchments and subcatchments to modellers on receipt of the blind test results (see above).

  Norway
  Calibration 1990-1994. Gauging stations at Mosselva and Hobolelva/Hoifoss will be used for assessing model performance.
  UK
  Data are available for nitrate and (unfiltered) orthophosphate for three identified subcatchments.

  Phosphorus – split in space
  TP data is only available for several years for five sites, so for P, the catchment can be split in space and some subcatchments are used for calibration and others for validation. ADAS suggested (and it was agreed) that the Swale (Station id: S1) and the Nidd (Station id: N4) are used for calibration for TP. The upper Swale is largely upland; the Nidd contains upland and some lowland areas. Validation can be done using station 'O6' at the base of the whole catchment which include upland and lowland areas and the whole of the Swale and Nidd basins.

  For modellers prepared to use orthophosphate data, then data are provided for the full 10 year period and for the same locations described below for nitrogen.

Nitrogen
For modelling N, three subcatchments will be used (i.e. the Swale at Catterick Bridge, the River Wiske at Kirkby Wiske and River Nidd at Skip Bridge) together with the main outlet at the base of the entire catchment (the Ouse at Nether Poppleton). Calibration should be conducted on the 1990-1994 data and validation on the 1995-2000 data.

Italy
Six gauging stations were agreed for modelling flow, loads and concentrations in the Enza. They are: Vetto, Buvolo, Cerezzola, Traversetolo, S_Ilario, and Coenzo. These should be split in time -> 1992-96 for calibration and data continuing on to 2002 for validation (the previous most downstream station is now not used due to its confluence with the Po).

3. Validation: The remaining half of the output (river flow and concentration) data for all (sub)catchments will be provided the to modellers from the EUROHARP database on receipt of calibration results (see above).

For all the above stages (1), (2) and (3):

- Modellers should record all key parameters and data input files used (and version of the model) → to allow model run to be reproduced/verified etc. at a later date
- Retention → For models with their own retention routine, modellers need to provide output both with and without their retention estimate (so we can compare their independent retention estimate with Brian's from WP5)
- Modellers agree to submit all model results (flows, concentrations, and loads) to the Work Package leaders. “Model results” should include flow (units m3/s), concentrations (units mg/l), and loads (units kg/ha); and be reported for all chemical species for which the model is capable of providing output and the most detailed time interval possible (e.g. day, month or year). Reported modelled concentrations and loads must specify whether units are elemental or refer to the anion (i.e. N or NO_3). A standard reporting format for model results has been provided for this purpose (available form the website).

- In its simplest form, the agreed procedure for assessing performance is:

Model (i) flow and (ii) N and/or P loads and concentrations on annual
(and where possible, sub-annual) basis for subcatchments and main catchment

↓

Allow for net effect of Point Source contributions

↓

Deduct retention estimates (see below)
(using alternative WP5 estimates and, where possible, model’s own estimate)

↓

Assess performance using simple statistical tools
(for different time/space/retention method combinations)

- The agreed performance tests for all model outputs are presented in Section 3 and are:
    - Relative volume error
    - Annual timestep output: mean deviation; mean absolute deviation; standard deviation
    - Sub-annual model output: RMSE; Relative error; coefficient of determination

Additional tests of model performance against measured data (e.g. Nash-Sutcliffe) may be calculated by agreement, depending on the nature of model outputs. Details of some of these approaches are included in Section 3.3.2.

2.7 Retention (input from WP5)

- Estimates developed within WP5 will relate to the main catchment outlet, and to specific subcatchments previously identified by catchment data contacts. Additional work will include the provision of seasonal (summer, winter) estimates of retention for the above catchment and subcatchment areas.
- The retention group will also apply the two methods documented in the OSPAR HARP guidelines for estimation of mean annual retention for each of the 17 catchments.
- The SWAT, TRK, MONERIS, EVENFLOW and ANIMO/NL-CAT models which include routines for estimating retention in the river system should produce two sets of model output – with and without the retention processes active – to aid comparison with the expert retention group estimates.

3. Model Performance Assessment (Validation Protocol)

3.1 Introduction

Performance assessment in the Euroharp project should clarify the applicability of quantification tools for different circumstances concerning climate, landscape, land use, etc. within the EC. Validation is one of the elements contributing to the process of judging model applicability. The type of validation to be performed is related to the aim of the modelling effort.

Aspects relevant to the type and method of validation to be performed

Aim of the modelling effort	Fresh water ecology within the catchment (definition and restoration of reference situation)
	Drinking water supply, safeguarding water quality
	Fresh water ecology downstream of the catchment
	Evaluation of impacts of fertilisation reduction on loads
	Evaluation of eutrophication reduction measures at catchment scale
	Evaluation of impacts of global change, water management, land use change, etc. on water quality
	Evaluation of eutrophication reduction measures at national scale (reporting obligation)
Output items	Flows
	N & P loads (specified for N and P species)
	N & P concentrations (specified for N and P species)
	Others, e.g. Retention and balances
Spatial scale of output required	Distribution within the catchment
	Some nested sub-watershed(s)
	Catchment at the outlet
Temporal aspects	Daily or monthly timestep of output required
	Annual timestep of output required
	Extreme values within annual cycle or long term extremes
	Extreme annual averages, return periods
	Trends within the annual cycle or long term trends
	Sudden changes in time series
	Dynamics within the annual cycle
Features of calibrated model parameters	Robust or sensitive
	Exact or uncertain

In the Euroharp project the validation focuses mainly on the evaluation of water quality at the outlet of the catchment or a sub-watershed on annual base and the evaluation of eutrophication reduction measures. Flows, loads and concentration timeseries are important parameters to be tested.

3.2 General approach

- The fact that half the models provide only annual timestep of outputs requires the use of (a) longer time series (up to 10 years for calibration and for validation) than that originally proposed, and (b) the use of a nested subcatchment approach (as documented in the Description of Work (DoW) document. It was agreed that model estimates would focus on a small number (typically 2-4) subcatchments in each main catchment, in addition to the main catchment outlet. These subcatchments should preferably be of different size (first and third order) and agri-environment character to enable validation of model leaching and retention estimates. Furthermore, the modelled subcatchments considered together should capture as much as possible of the dominant soil/land use/hydrology/climate typologies found within the main catchment itself.

- Model performance will be assessed (a) based on best available knowledge and (b) after formal calibration using measured river data (flow, load, concentration).
- Modellers will report predicted flow, N and P load (including speciation if possible), and flow-weighted mean concentration for each gauging site and time period which has been modelled.
- Modellers will use simple statistics to characterise model performance by subcatchment, catchment, and across all core catchments considered together.
  These simple statistics include:
  1. Mean deviation
  2. Mean absolute deviation
  3. Standard deviation
  Additional performance assessment criteria may also be used.
- The statistical results will be collated and tabulated to aid comparison between the same model on different core catchments, and between different models on the same core catchment.
- In addition there will be (where possible) an intercomparison regarding intermediate (internal) values associated with specific pathways (e.g. runoff, root zone losses, groundwater/baseflow component) to help identify the main sources of differences between the predictions from the various quantification tools. All institutes agreed that this “internal” checking of model performance should be undertaken where possible (e.g. checking model predictions of crop yield, loss from the soil root zone or tile drains). Although formal validation is usually not possible in these cases, such intermediate model outputs can be checked against plausible values in consultation with the catchment data owners, and thus identify potential sources of error or parametric uncertainty.

3.3 Model performance assessment

The scientific component of model evaluation is described as the assessment of consistency between model-predicted results and prevailing scientific theory, which for the scope of the EUROHARP project will be largely covered in the EUROHARP Model Review. An assessment of accuracy and precision represent the operational components of the model evaluation process (Willmot et al., 1985). Loague and Green (1991) define accuracy as the extent to which the model predicted values approach a corresponding set of measured observations. Precision is the degree to which model-predicted values approach a linear function of measured observations.

The model validation criteria presented below are concerned with the model output at a daily, monthly or annual time-step and include both statistical criteria and suggested graphical displays. This combination of assessments are proposed measures for use in the EUROHARP project to evaluate elements of the performance of the range of competing models in terms of accuracy and precision e.g. the expected magnitude of errors or the tendency for systematic bias in the model errors.

In assessing a model’s performance, qualitative measures are also valuable such as the simplicity of the model and ease of model use: these qualitative factors are addressed in the EUROHARP Model Review document.

Aggregation procedure for different frequencies of model output and data timeseries

Daily output timestep	Annual output timestep
Apply model to (1) Blind test, (2) Calibration timeseries and (3) Validation timeseries for those models capable of producing output at daily timesteps (i.e. NL-CAT, TRK, SWAT, EveNFlow)	Apply model to (1) Blind test, (2) Calibration timeseries and (3) Validation timeseries for those models capable of producing output at an annual timestep only (i.e. (MONERIS, NLES, NOPOLU, REALTA, SA)
If the validation dataset comprise time series on daily timestep, then compare output with measured data using methods listed below	
If the validation dataset comprise time series on annual timestep, then aggregate and calculate annual flows/loads and flow weighted mean concentrations →	Compare output with measured data using methods listed below

Depending on the availability of flow data, loads can be estimated by processing the flow data and concentration time series (see HARP flow normalisation method later). If the flow data are not reliable, the only remaining option for validation is the comparison of simulated and measured concentration time series. **The minimum option should therefore be the comparison of total flow, total load, and flow weighted mean concentrations for at least five years both at the main catchment outlet and previously identified subcatchment gauging stations**. In the case of annual timestep models, the small number of data pairs will mean this approach has only a limited ability to characterise performance – hence the need for separate assessments of performance for different subcatchments within each main catchment (see earlier).

Comparison with surface water concentrations implies the availability of retention estimates within the surface water system. When the model only generates output concerning flows, loads and concentrations at the soil root zone, then estimates for net retention between root zone and surface water body are also required.

3.3.1 Methods to assess model performance: Annual timesteps

In the case of annual timestep models, the small number of modelled and measured data pairs means that only a few criteria can be applied to test the model performance.

Graphical evaluation

- 1:1 plots (simulated against measured)
- Plotting of frequency distributions

Statistical evaluation for time-series

- Absolute error (observation minus prediction)
- Residual error (e.g. discernible systematic bias; correlation coefficient)

In addition: statistical evaluation for long time-series

- Type of distribution (normal, uniform, skewed)
- Distribution of extreme values (Gumbel)
- Kendall tau test for trends.

Assessment of model performance for future trend predictions

- Equifinality (i.e. does a unique set of parameters yield a specific result?)
- Uncertainty of parameters (how many parameters have been identified from measurements, or from literature, or from expert knowledge, or from transfer functions, or fitted during calibration?)

- Uncertainty in the prediction of future trends (e.g. Generalised Likelihood Uncertainty Estimation - GLUE)

Overall assessment of Euroharp model suite

- Ranking the models to the performance regarding a certain statistical criterion (e.g. Absolute error, RMSE)
- Rejection or acceptance on the basis of exceeding or non-exceeding a certain tolerance limit. Such a tolerance limit depends on the aim of the modelling effort, and would require criteria to be developed to judge whether a model application is valid or non-valid for a particular site or scenario.

3.3.2 Methods to Assess Model Performance: Sub-annual timesteps

Scope

Internal system checking

Many of the models for quantifying diffuse N and P losses from agricultural land, selected for the EUROHARP project are able to predict a variety of variables other than N and P concentrations or loads in the river. For example, many models estimate the variables of flow, soil moisture, leached nitrate at the base of the soil profile. Although some internal intercomparisons between model predictions will be undertaken (e.g. at the root zone), due to data availability the EUROHARP model validation exercise is restricted to objective assessment of the ability of a model to predict the N or P concentration in the river relative to other models and measured data. Further since only some models can produce daily outputs, subannual validation and statistical tests used are at the discretion of the modelling institute.

Statistical tests

Statistical validation measures can be applied to both the estimation period and the validation period for comparison, and can include:

i) Concentration time-series,
ii) load time-series,
iii) sorted and unsorted data.

Graphical displays can be useful for showing trends, types of errors and distribution patterns (Loague and Green, 1991). If a model is capable of providing output at a subannual temporal resolution, then modelling institutes have agreed to provide subannual timestep output for each subcatchment with monitoring data which has been identified and defined by the catchment data institute. Suitable performance assessment criteria are described in the following section.

Criteria for Measuring the Performance of Models

Included below are some suggested statistical tests to evaluate the performance of the models applied in EUROHARP which are capable of generating data on a daily or monthly time steps. Loague and Green (1991) comment that, independently, many of these measures of model performance are limited by assumptions of normality, equality of variance and independence. If these assumptions are violated the derived statistic could be potentially unreliable. For this reason model outputs are tested against a range of statistical methods of varying complexity.

The statistical tests fall into two categories (Beck et al., 1994):

i) unpaired tests: individual values of (O_i) and (P_i) are not matched with one another.
ii) paired tests: individual values are matched.

Unpaired tests

The unpaired tests examine the following properties of the observed and modelled data:

- A comparison of cumulative distribution functions: i.e. the frequency with which certain values of a variable are found to occur. These can be compared between models and also for the same model to examine spatial variations i.e. distributions for different subcatchments;
- Mean and variance of these distribution functions.

Tests based on unpaired sets of data such as the mean and variance of cumulative distribution functions for (O_i) and (P_i), do not concern themselves with the contemporaneous variations in these quantities (Beck et al., 1994). These tests are included as part of the EUROHARP performance assessment tests because the development of some models may lend greater importance to accuracy in modelling the distributions of diffuse N loss (which may be especially important for a model developed for policy support purposes) rather than the precise timing of individual events on specific days (e.g. the frequency of exceedance of the 50 mg NO3/l limit stipulated in the Nitrates Directive).

Cumulative Frequency Distributions

The Kolmogorov-Smirnov statistic is a non-parametric goodness of fit test, that can be applied to determine whether the cumulative frequency distributions of the observed and modelled flow and concentration or load series are 'significantly different' from one another. The null hypothesis for the Kolomogorov-Smirnov goodness of fit test is that the sample data which produce the observed cumulative frequency curve have been drawn from a population that possess the specified theoretical distribution, in this case the distribution of the measured data. If the maximum difference between the two distributions (D) does not exceed the critical value at the specified level of significance then the two distributions are said to be 'not significantly different', (Spear, 1970; Parkinson and Young, 1998; Ebdon, 1985).

Paired tests

The following tests are included to test the model match between the predicted and observed values at the same points in time and space.

The paired tests may be further categorised into tests concerned with:

i) Measures of bias (measures of difference)
ii) Measures of statistical association
iii) Residual diagnostics

These tests are parametric and therefore assume that the data are sampled from a Gaussian distribution where the model residuals (errors) are independent and normally distributed with mean zero and constant variance.

Measures of bias / Measures of difference

The Root Mean Square Error ($RMSE$), Mean Percentage Error (MPE), and Relative Error (RE) are measures which provide a quantitative estimate of the size of differences between models. There is no absolute value for a 'good' $RMSE$ or MPE, instead these diagnostics will be used for relative comparison of the output between models.

Root Mean Squared Error (RMSE):

$$RMSE = \frac{100}{\overline{O}}\sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}}$$

where n is the number of data points, O_i are the ith observed (measured) data points and P_i are the ith modelled data points; $\overline{O}$ is the mean of the observed (measured) data (Smith et al., 1997). The lower limit for RMSE is zero.

Relative Error

The bias in the total difference between simulations and measurements is determined by calculating the relative error (RE) (Smith et al., 1997).

$$RE = \frac{100}{n}\sum_{i=1}^{n}(O_i - P_i) / O_i$$

where n is the number of data points, O_i are the ith observed (measured) data points and P_i are the ith modelled data points.

<u>Measures of statistical association</u>

The precision of the model can be checked using tests which measure how well the modelled time series fits the data.

Coefficient of Determination

The coefficient of determination (R^2) is a measure of the proportion of the total variance of the observed data explained by the predicted data. There are several definitions of the coefficient of determination (Aitken, 1973; Nash and Suttcliffe, 1970), in this instance the Nash-Sutcliffe coefficient of determination will be applied (Nash and Suttcliffe, 1970):

$$Efficiency\ \ factor = \frac{\sum_{i=1}^{n}(O_i - \overline{O})^2 - \sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2}$$

where O_i are the ith observed (measured) data points and P_i are the ith modelled data points; $\overline{O}$ is the mean of the observed (measured) data; n is the number of data points.

The upper-bound for R^2 is unity and it can assume negative values, which implies that the model introduces more ambiguity than is introduced by simply using the mean value for the observation as an estimator.

4. Quantification of the Monitored Riverine Load of Nitrogen and Phosphorus, including Water Flow Normalisation Procedures

(Selected excerpt from OSPAR HARP guideline 7[1])

4.1 Objectives

To describe procedures for the quantification of the total riverine load of nitrogen and phosphorus, including methods for the normalisation of riverine loads.

This Guideline describes procedures for:

- The quantification and reporting of the total riverine load of nitrogen and phosphorus and
- The normalisation of riverine loads (c.f. section 4).

4.2 Quantification of the total riverine load of nitrogen and phosphorus

4.2.1 Monthly data resolution

Where appropriate and practicable, the following riverine time-series data, covering the period from 1985 onwards, should be calculated on the basis of:
Time-series of river flow (flow data on a monthly basis, preferably based on daily values); and
Time-series, with calculated riverine loads of nitrogen (NO_3-N and total-N) and phosphorous (PO_4-P and total-P)- dissolved and particulate- the data resolution should at least be on a monthly basis.

4.3 Normalisation of riverine load data

4.3.1 General

The following two major approaches for flow-normalisation are described:
a. Empirical hydrological normalisation (further referred to as category 1); and
b. Model-based hydrological normalisation (further referred to as category 2).

Six different empirical hydrological normalisation methods are described below (further referred to as 1A1, 1A2, 1A3, 1B1, 1B2 and 1C). They should be used for the reporting of annual riverine or stream loads.

Generally, methods 1A1, 1A2 and 1A3 are most suitable when trends in the riverine loads are small, or when the relationship between load and flow, or concentration and flow do not change over time.
4.3.1.4 Method 1A1 is less suitable when the concentration/flow-relationships are strong (c.f.: example in the Annex). Methods 1B1 and 1B2 are particularly useful in situations where the transport/flow relationship is gradually changing over time (i.e. when diffuse or point sources are increasing or decreasing over time).

[1] Full text available from http://www.ospar.org and www.euroharp.org

4.3.2 Empirical hydrological normalisation (category 1)

General

Category 1 methods concern:
1A: Methods that can be applied to systems with random variation around a fairly constant long-term mean;
1B: Methods that can be applied to systems with trends; and
1C: Methods that can be applied to systems where the flow may be divided between various pathways.

1A. Methods that can be applied to systems with random variation around a fairly constant long-term mean

The three formulas given may be used when the trends in the riverine loads are small. The first method (1A1) represents the easiest approach, where annual normalised loads are estimated by:

$$\boxed{\tilde{L}_i = L_i \frac{\bar{q}}{q_i}} \quad \text{(1A1)}$$

Where

L_i denotes the mean annual load the ith year;
q_i is the mean annual flow in the ith year; and
$\bar{q}$ the long-term mean annual flow (calculated over the time period from 1985 onwards).

The disadvantage with this method is the rather inefficient use of the statistical information in the concentration and flow data. This is particularly true in situations with dependency between concentration and water discharge (c.f.: example in Annex). It is therefore recommended to use this method only if the other proposed methods are considered to be inadequate.

Method 1A2 uses the normally good relationship that exists between riverine loads and flow (i.e. water discharge). The relationship may be modelled by a simple regression equation of the following form:

$$\boxed{L_{ij} = \alpha \;\; + \beta \;\; q_{ij} + \varepsilon_{ij}, \quad i = 1,2,\ldots,n, \quad j = 1,2,\ldots,m,} \quad \text{(1A2)}$$

Where

L_{ij} denotes the load during the jth season (normally monthly or fortnightly point samples) of the ith year;
q_{ij} is the flow during the same period; and
ε_{ij} is a random error term: α (intercept) and β (slope) are model parameters.

For the sake of simplicity, this approach is exemplified with a linear model. Any model-function (not necessarily linear) is, however, possible. With this model-structure, flow-normalised seasonal values may be calculated according to the equation:

$$\boxed{\tilde{L}_{ij} = L_{ij} - (q_{ij} - \bar{q}_{..})\hat{\beta}}$$

Where
$\hat{\beta}$ is the estimated slope parameter; and
$\bar{q}_{..}$ the average flow for a reference period.

In order to reduce the risk of obtaining negative loads, one can also apply a flow-normalisation, according to the equation:

$$\boxed{\tilde{L}_{ij} = L_{ij} \cdot \frac{\hat{\alpha} + \hat{\beta}\bar{q}_{..}}{\hat{\alpha} + \hat{\beta}q_{ij}}},$$

Where
$\hat{\alpha}$ is the estimated intercept parameter.

Annual flow-normalised values are also obtained by simple aggregation of the seasonal values according to the equation:

$$\boxed{\tilde{L}_i = \sum_j \tilde{L}_{ij}}$$

If the relationship between nitrogen and phosphorus load and flow shows seasonality, the regression model 1A2 can be extended to the equation:

$$\boxed{L_{ij} = \alpha_j + \beta_j q_{ij} + \varepsilon_{ij}, \quad i = 1,2,\ldots,n, \quad j = 1,2,\ldots,m,} \quad \text{(1A3)}$$

Where
L_{ij} denotes the load during the jth season (month) of the ith year;
q_{ij} the flow during the same period; and
ε_{ij} is a random error term: α_j and β_j are model parameters.

In such cases, flow-normalised values can be calculated according to the equation:

$$\boxed{\tilde{\hat{L}}_{ij} = L_{ij} - (q_{ij} - \bar{q}_{.j})\hat{\beta}_j}$$

Where
$\hat{\beta}_j$ denotes the estimated slope parameter for the jth season; and
$\bar{q}_{.j}$ is the average flow during the jth season.

Annual flow-normalised values are obtained in a similar way as for method 1A2.

1B. Methods that can be applied to systems with trends

The 1A methods described above are relevant for situations whereby the momentary concentration or riverine load is a time-independent function of the simultaneous flow or of time-lagged runoff values. However, concentration-flow and load-flow relationships may change gradually over time. Two flow-normalisation methods, which can accommodate gradual changes in transport-flow relationships, are described below.

Method 1B1 represents basically an extension of methods 1A2 and 1A3. The time series are divided into separate time periods (1985-1989 and 1990-1994) and then analysed separately according to methods 1A1, 1A2 or 1A3.

Method 1B2 accomplishes gradual and smooth changes in relationships between load and runoff. More precisely, it describes a semi-parametric regression model on the following form

$$\boxed{L_{ij} = \alpha_j + \beta_{ij} q_{ij} + \varepsilon_{ij}, \quad i = 1,2,\ldots,n, \quad j = 1,2,\ldots,m,} \quad \text{(1B2)}$$

in which the variation of slope parameters β_{ij} from season to season and year to year is only restricted by non-parametric constraints.

The model parameters are estimated by minimising an expression of the form:

$$\boxed{S(\alpha,\beta) = \sum_{i,j} (L_{ij} - \alpha_j - \beta_{ij} q_{ij})^2 + \lambda_1 \sum_{i,j} (\beta_{ij} - \frac{\beta_{i+1,j} + \beta_{i-1,j}}{2})^2 + \lambda_2 \sum_{i,j} (\beta_{ij} - \frac{\beta_{i,j+1} + \beta_{i,j-1}}{2})^2,}$$

Two penalty factors λ_1 and λ_2, are used to define a desired compromise between overfitting and specification errors. This semi-parametric regression approach is also referred to as a roughness penalty technique. Suitable levels of the penalty factors λ_1 and λ_2 can be established by undertaking a cross-validation study of relationships between L_{ij} and q_{ij}. One may also apply further restrictions: the generalised degrees of freedom of the model could be a constant or the ratio $\frac{\lambda_1}{\lambda_2}$ of the penalty factors could be a constant.

Seasonal flow-normalisation could be accomplished in an additive way by employing the formula:

$$\boxed{\hat{L}_{ij} = L_{ij} - (q_{ij} - \bar{q}_{.j})\,\hat{\beta}_{ij},}$$

or by multiplication by employing the formula:

$$\boxed{\tilde{L}_{ij} = L_{ij} \cdot \frac{\hat{\alpha}_j + \hat{\beta}_{ij}\bar{q}_{.j}}{\hat{\alpha}_j + \hat{\beta}_{ij} q_{ij}}}$$

Where
$\hat{\beta}_{ij}$ and $\hat{\alpha}_j$ depict parameter estimates obtained by employing the roughness penalty approach described above.

Annual flow-normalised values are obtained in a similar way as for method 1A2. Method 1B2 cannot be run automatically in standard software packages.

Method 1B2 can be extended with regard to:
The parameterisation of the intercept parameter α, which may vary from year to year; and
Further normalisation variables, e.g. the temperature. This requires an extension of the penalty expression above and appropriate restrictions to the penalty factors.

5. References

Behrendt, H. 1997. ‘Detection of anthropogenic trends in time series of riverine load using windows of discharge and long- term means’, ICES-Report cm1997/env: 11 of the ICES/OSPAR workshop on the identification of statistical methods for temporal trends, Annex 5, 20-29, 1997.

OSPAR, 1996. ‘Principles of the Comprehensive Study on Riverine Inputs and Direct Discharges (RID)

Stålnacke, P. and Grimvall, A. 1997. Semi-parametric approaches to flow-normalisation and source apportionment of substance transport in rivers. *Envirometrics*.

Aitken, A. P. (1973) Assessing systematic errors in rainfall-runoff models. *Journal of Hydrology*, 20(2), 121-136.

Beck, M. B., Mulkey, L A., Barnell, T. O. (1994) Model Validation for Predictive Exposure Assessments

Ebdon, D. 1985 Statistics in Geography, Blackwell Publishers, Oxford.

UK government (1999) White paper on the Nature and Scope of Issues on Adoption of Model Use Acceptability Guidance, the Science Policy Council Model Acceptance criteria and peer review White paper Working group.

Finney, D. J. (1980) Statistics for Biologists. Science Paperbacks, Bristol, Great Britain.

Loague, K. and Green, R.E (1991) Statistical and graphical methods for evaluating solute transport models: Overview and application, *Journal of Contaminant Hydrology*, 7, 51-73.

Nash, J. and Sutcliffe, J. (1970) River flow forecasting through conceptual models part 1 – a discussion of principles. Journal of Hydrology, 10, 282-290.

Parkinson S. and Young, P.C. (1998) Uncertainty and sensitivity in global carbon cycle modelling. Climate Research 142 pp

Spear, R. C. (1970) The application of Kolmogorov-Renyi statistics to problems of parameter uncertainly in systems design. International Journal of Control. 11(5), 771-778.

Willmot, C. J, Ackleson, S. G., Davis, R. E., Feddema, J. J. Klink, K, M. Legates, D. R., O’Donnell, J., Rowe, C. M. (1985) Statistics for the evaluation and comparison of models. *Journal of Geophysical Research*, 90 (5), 8895-9005.