6th International Conference on Hydroinformatics - Liong, Phoon & Babovic (eds)

DEVELOPMENT OF A STRATEGIC DATA MANAGEMENT SYSTEM FOR A NATIONAL HYDROLOGICAL DATABASE, THE UK NATIONAL RIVER FLOW ARCHIVE

JAMIE HANNAFORD
Centre for Ecology and Hydrology, Maclean Building, Crowmarsh Gifford Wallingford, Oxfordshire OX10 8BB, United Kingdom

Many countries are experiencing increases in both the demand for river flow data and the financial pressures which constrain monitoring. This paper describes the implementation of data management system designed to maximise accessibility to user-focused river flow data of an appropriate quality for meeting strategic information needs. The system is based on a Service Level Agreement (SLA) which governs the transfer of data from regional data providers to the national database, and the subsequent quality control of data. The SLA employs scoring mechanisms to monitor performance of providers in relation to the completeness and quality of data along with the timeliness of provision. The SLA mechanisms and the associated data management infrastructure are discussed and results for the first year of implementation are presented. The potential of the system for improving the utility of flow data is discussed from a strategic perspective.

INTRODUCTION

Good quality, continuous time series of river flows are fundamental to the management of hydrological resources and risks at a national level. The quality of flow data is of paramount importance; models and decision support systems are only as reliable as the data used to develop and calibrate them, whilst the identification of non-stationarity or trend, which is crucial to the forecasting of future resource availability or risk magnitude, is dependent on the availability of long-term, continuous flow time series. The reliability and homogeneity of flow data assumes even greater importance at a time when climatic change is expected to modify hydrological regimes [1, 2]. However, funding and support for monitoring networks and data acquisition systems is under pressure in many parts of the world [3], adding to the challenges providing reliable data in a timely manner to researchers, water managers and policy makers. This paper discusses a data management initiative designed to improve the utility of the UK's nationally archived river flow datasets to allow national strategic requirements to be met within available resource limitations.

BACKGROUND TO THE NRFA

The National River Flow Archive (NRFA) is the UK's principal hydrometric database, containing daily mean river flow data from over 1300 monitoring sites. The NRFA is not responsible for collecting the river flow data – data are captured and processed primarily

by the government agencies responsible for environmental monitoring and protection in the UK: the Environment Agency (EA) in England and Wales; the Scottish Environment Protection Agency (SEPA); and the Rivers Agency of the Department of Agriculture and Rural Development in Northern Ireland (DARDNI). For routine archiving, these agencies submit data on an annual basis to the NRFA, for loading, inspection and quality control, and ultimately dissemination to the user community through a range of media.

DATA UTILITY AND THE SERVICE LEVEL AGREEMENT

Following a workshop convened to determine the strategic information needs of a range of stakeholders, a new set of core objectives for the NRFA was drawn up to meet changing national information needs [2]. These objectives include the need to establish a range of capabilities, including requirements to: assess national water resources; improve the ability to identify trends in runoff; meet national research objectives, including the development of hydrological models. Data archiving and management programmes such as the NRFA can play a key role in maximising the utility of flow data, particularly by improving the information content of datasets and adopting a user-focused approach. In order to maintain and improve the utility of nationally archived data, a Service Level Agreement (SLA) was introduced to govern the routine acquisition and validation of flow data. The SLA provides a framework to assess data utility using indices relating to three components: data completeness, data quality and timeliness of provision. The SLA framework is designed to facilitate monitoring of the performance of data providers, or of individual gauging stations or groups of gauging stations, according to these three core components. Performance is assessed using a scoring system which permits an objective, quantitative appraisal.

An enhanced data acquisition and validation system was developed to support the SLA framework. The NRFA data take-on model is illustrated in Fig.1. Data loading and validation are facilitated by a suite of PC front-end software applications, which populate a series of audit trails. These ensure that the process is managed efficiently (important given the distributed nature of the data providers and the volumes of data being exchanged), whilst also enabling the automatic calculation of SLA scores. Implementation of a ‘Batch’ database to hold new, provisional data prior to validation, offered considerable data security advantages, in keeping un-validated data separate from the main archive. More importantly, its functionality is essential to the effective management of the complex data transfer operation. Discrete ‘batches’ of data are linked by a unique primary key, which is used to track separate batches through the validation and query process. This approach also ensures that metadata (such as time of arrival, details of the data provider and the NRFA personnel involved in loading) is attributed to each batch, providing an efficient, auditable system capable of automatically yielding the information needed to calculate SLA scores.

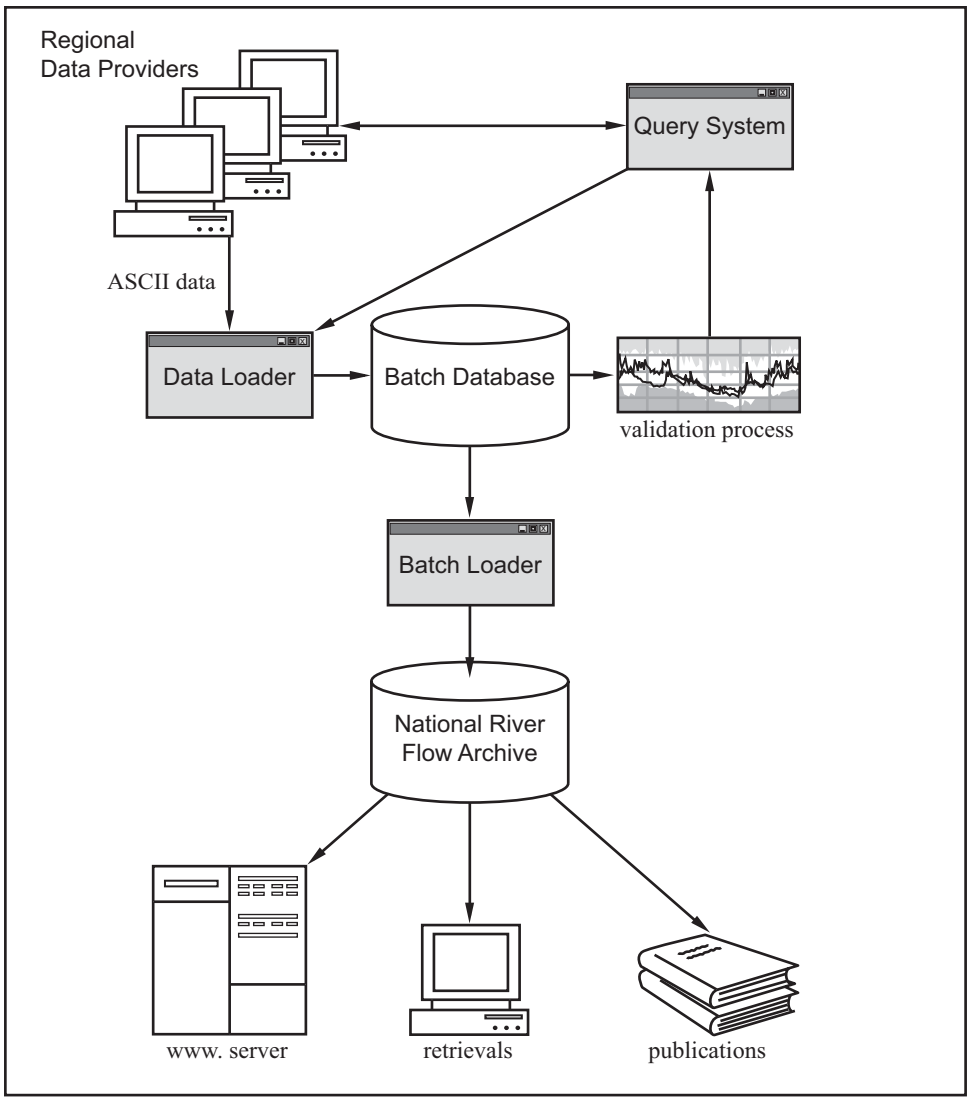

Figure 1. Schematic diagram of NRFA data acquisition and validation process.

The SLA was first implemented in 2002 in England and Wales. Table 1 illustrates the SLA scores for data suppliers from 8 regions. SLA scores were also computed for a number of sub-regional suppliers; for illustrative purposes, scores are shown for the sub-regional providers which make up region 3 (3a – 3c). The following sections discuss the various components of the SLA in more detail, with reference to the scores presented in this table. The scoring indices are referred to in the text through Roman Numerals as presented in the columns of Table 1.

Table 1. Service Level Agreement Scores for 8 regions and 3 sub-regions in 2002.

REGION	# Of SLA Sites	Time-liness (I)	Comp. values (II)	Comp. stations (%) (III)	Queried values (IV)	Queried stations (%) (V)	Query time-liness (VI)
1	62	100	9.16	87.1	9.64	93.6	15
2	70	100	9.64	95.7	9.70	92.9	100
3	64	81.2	7.95	52	9.82	93.7	100
3a	21	44.0	7.37	85	10	100	100
3b	20	100	7	0	9.90	89.5	100
3c	24	97.2	7	67	9.60	91.7	100
4	68	100	9.38	98.5	9.97	95.6	100
5	50	100	4.72	74	7.44	70	100
6	53	79.08	9.91	73.6	9.42	94.3	36.67
7	66	100	7.45	86.2	9.07	93.8	100
8	46	93.0	6.96	78	9.62	93.3	3.30

Timeliness and data loading

Timeliness of despatch has a direct impact on data utility in terms of the operational requirements of the NRFA - receipt of a data despatch determines the earliest time at which data can become available for processing and dissemination. A 31st March deadline was adopted for the annual transfer of the previous year's data. A maximum SLA timeliness score is awarded to data received before the deadline, whilst the score for data received late is decremented on a daily basis until it receives a zero score after fifty working days (I). In the first year operation, the majority of data providers achieved the deadline.

Completeness

The utility of a time-series dataset is strongly influenced by its continuity. Even minor gaps can preclude the calculation of a summary statistic (such as annual runoff), which may eventually lead to entire years being unsuitable for statistical analyses – a major constraint on the utility of a dataset. By virtue of the problems associated with data capture during very high or very low flows, gaps are found more frequently at the extremes of the flow range. In particular, misleading conclusions may be drawn from analyses applied to time-series which have missing data for those periods (e.g. the drought of 1976 [5] or the widespread flooding of autumn 2000 [6] in the UK) which define period-of-record minima and maxima. Clearly, any attempt to improve the utility of data must address this problem by monitoring data completeness and, through time, attempt to ameliorate the problem.

The first SLA completeness score (II) is based on the ratio of missing daily mean flows to the flows which are expected (normally 365 or 366, although the system is flexible to allow for unavoidable station downtime) for any station, and is scaled from 0 – 10. A second scoring mechanism (III) reports on the percentage of stations with complete data in any one region. Table 1 shows that, for most regions, there was a good degree of completeness when measured at the daily flow level – which indicates that overall, the amount of missing data is very low relative to the number of days present. The station completeness scores, however, reveal that missing data sequences affect a high proportion of stations. Together, these scores indicate that the major problem with completeness is not the volume of missing data in any region but the number of stations affected by gaps – which are often of short duration. The zero score for sub-region 3b, for example, is a result of a single daily mean flow being missing from every station; a contrast to the relatively good DMF completeness score.

The completeness component of the SLA was adapted in 2003, in an attempt to ameliorate the problem by placing greater emphasis on the infilling of gaps. In the UK, infilling can often be achieved during periods of stable flow (e.g. during recessions) by a modelled interpolation or data transfer from nearby or analogue sites. To encourage infilling/interpolation over the full span of the SLA cycle, the scoring algorithms are re-applied at the end of the year; the new system therefore monitors improvements in completeness seen over the SLA cycle.

Data Quality and the Query System

The river flow data are validated by a combination of mechanistic checks (based on the statistical characteristics of the time series) and visual appraisal undertaken by experienced operators familiar with the expected flow patterns at the target station. Dedicated time-series plotting software enables visual hydrograph appraisal (Fig 2). Hydrographs can be inspected against near-neighbour sites or analogue sites, and against hyetographs from raingauges. The software also displays a series of QA/QC flags which highlight, for example, rises in flow which rank in the top 5 rises in the whole record. In addition, the long-term minimum and maximum envelopes are featured, allowing operators to focus on anomalous sequences of greatest significance. This form of visual appraisal enables the user to identify periods of anomalous flow and investigate whether the anomaly reflects a possible error in data collection or processing. Equally important in guiding validation is the descriptive information which is accessible through the plotting software. This descriptive information includes details of the catchment characteristics, hydrometric performance of the station and artificial influences on the flow regime; the latter are particularly important in assessing the causes of anomalous flows. In the UK, the majority of rivers are affected by anthropogenic influences to some degree, so it is important to take account of these effects when validating data.

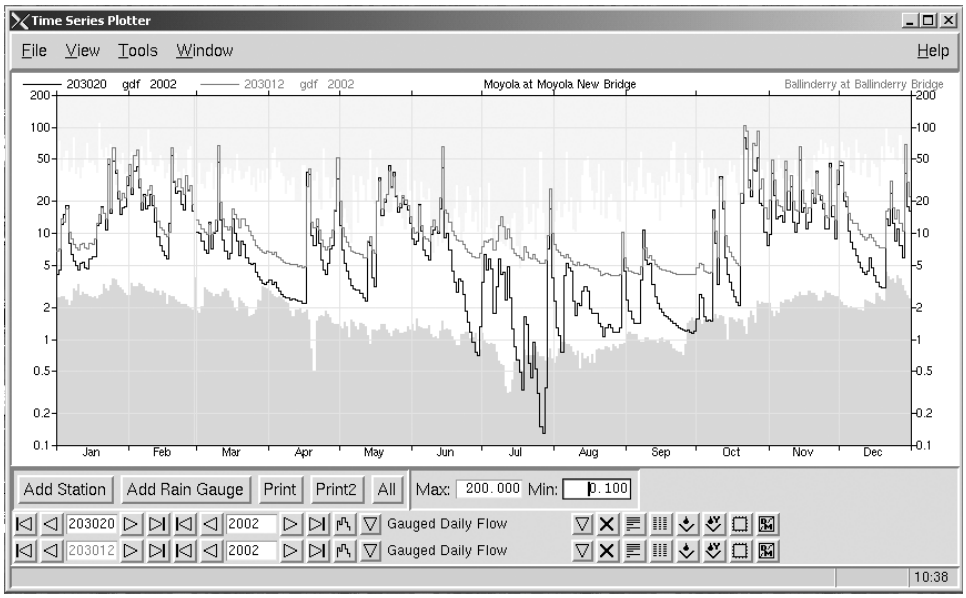

Figure 2: Screen shot of NRFA hydrograph plotting software. Dark trace shows flow from target site, light trace displays flows at a near-neighbour site. Maximum and minimum flow envelopes are shown above and below the line plots in light and dark grey respectively.

When anomalous flows are identified, a 'query' is logged using a front-end program (the Query System in Fig 1) connected to a validation audit-trail. Queries are then despatched to the measuring agencies so that anomalous values can be addressed at source. A standard 'query report' is exported from the query software, which automates Microsoft® Excel® to produce the pre-formatted report. The report lists the queried data along with comments describing the nature of the query, and descriptive information about the station. Hydrographs are exported from the validation software in Adobe® Acrobat® (.pdf) format, thus enabling a full numerical and graphical depiction of the query to be emailed to the measuring authorities. The query is normally resolved by the measuring authority sending the query report back with explanations for the observed flows, along with revised data if appropriate.

Under the SLA, data quality is indexed by reference to the number of queried values; the scoring algorithms are identical to the completeness scores, with a score based on queried days (IV) and another on the percentage of stations in a region with queries (V). A further score assesses the timeliness of response to queries (VI). The query scores for 2002 (Table 1) indicate that, whilst the total number of queried days is typically very low, typically between 5 and 10% of stations had some queried data which required resolution.

STRATEGIC IMPORTANCE OF THE SLA

The mechanistic algorithms employed to calculate SLA scores are designed primarily to monitor the performance of data providers. However, the programme has a wider scope - to make long-term improvements to the utility of data; the cornerstone of this objective is the user-focused approach of the SLA. This is emphasised in the completeness component, which seeks to promote the continuity of datasets by encouraging targeted and auditable gap-filling. The inclusion of suitably flagged estimates is normally preferable to gaps in the record. It is particularly important that this is addressed from a strategic perspective; gaps in a time-series may not be regarded as a problem in relation to local, operational needs – particularly if resources are stretched. Completeness may have a major impact on the suitability of the dataset for a range of applications, however; if a catchment is relatively undisturbed by anthropogenic impacts, for example, it will have particular relevance to the identification of climatic driven trends.

In terms of information delivery for strategic purposes, there is an important link between the SLA and the NRFA's ongoing national gauging station network review [4] – a project which aims to identify the more strategically valuable gauging stations in the UK. The network review categorises stations according to their utility for several strategic purposes, such as the 'Benchmark' catchments, which are relatively pristine, and hence most suitable for trend detection [7]. The SLA mechanisms, correspondingly, are prioritised towards the catchments with the greatest strategic importance.

One of the limitations of the SLA is that mechanistic scoring indicators do not give a fully characteristic account of the utility of a dataset for strategic purposes. Counting the number of queries allows quantitative comparison of data quality between regions, but the strategic impact of queries relating to an individual station cannot be captured in a simple index. The hydrograph in Fig. 2 illustrates a query at station 203020 – the low flows in July are the lowest on record and appear to be anomalous. If this query had not been resolved – the low flows were found to be the result of works on a downstream bridge – the original data would have erroneously extended the range of recorded variability and biased the statistical properties of the time-series for low flow analyses. The SLA scores alone do not give a complete account of the performance of a station, so it is vital that regular reviews are carried out to ensure that longer-term data quality issues are addressed. To this end, an annual hydrometric audit is carried out; this uses the SLA performance data along with information gathered during the validation phase, to inform a qualitative appraisal of station performance in relation to the strategic importance of the station. Over time, SLA performance scores combined with hydrometric audits, will inform decisions on the suitability of gauging stations for strategic purposes. Improvements in fitness-for-purpose will be achieved using longer term performance data to ensure that appropriate user-guidance material is disseminated with individual datasets; chronic hydrometric problems which become apparent - such as the impact of summer weed growth on the monitoring of stage – can be flagged up on the descriptive comments provided with data retrievals.

The SLA data management system aims to maximise the utility of hydrological data, whilst prioritising quality assurance activities to ensure that available resources are concentrated on data with the greatest strategic importance. Thus the overall objective is to secure significant improvements in the information content of datasets, with modest expenditure of effort relative to the resources devoted to data capture – particularly vital at a time when resources available for monitoring are stretched.

ACKNOWLEDGEMENTS

The author wishes to thank Terry Marsh for constructive comments on this paper, and Karin Cheetham for preparing the figures, in particular Fig 1. In addition, the assistance of NRFA staff in developing the data management structures is acknowledged, particularly the systems manager Oliver Swain, who programmed the QC software and much of the front-end PC applications.

REFERENCES

[1] IPCC, "*Climate Change 2001: The Scientific Basis. Contribution of working group 1 to the third assessment report of the Intergovernmental Panel on Climate Change*", Cambridge University Press, Cambridge and New York, (2001).

[2] Hulme, M., Jenkins, G.J., Lu, X., Turnpenny, J.R., Mitchell, T.D., Jones, R.G., Lowe, R., Murphy, J.M., Hassell, D., Boorman, P., McDonald, R. and Hill, S., "Climate Change Scenarios for the United Kingdom: the UKCIP02 Scientific Report", Tyndall Centre for Climate Change Research, School of Environmental Sciences, University of East Anglia, Norwich, (2002).

[3] Rodda, J.C., "Water Under Pressure", *Hydrological Sciences Journal*, Vol. 46, No. 6, (2001), pp 841 – 853

[4] Marsh, T.J., "Capitalising on river flow data to meet changing national needs – a UK perspective", *Flow Measurement and Instrumentation*, Vol. 13, (2001), pp 291-298

[5] Hamlin, M.J., and Wright, C.E., "The effects of drought on the River Systems", in "Scientific Aspects of the 1975-76 Drought in England and Wales", *Proceedings of the Royal Meteorological Society*, Series A, Vol. 363, (1978) pp 69-96

[6] Marsh, T.J., and Dale, M., "The UK Floods of 2000-2001: A Hydrometeorological appraisal", *J. Chartered Institute of Water and Environmental Management,* Vol. 16 (2002), pp. 180-188

[7] Bradford, R.B. and Marsh, T.J., "Defining a network of benchmark catchments for the UK", *Proceedngs of the Institution of Civil Engineers, Water and Maritime Engineering*, Vol. 156, (2003), pp. 109-116