The need for operational reasoning in data-driven rating curve prediction of suspended sediment.

Journal:	*Hydrological Processes*
Manuscript ID:	HYP-11-0353.R1
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Mount, Nick; University of Nottingham, Geography Abrahart, Robert; University of Nottingham, School of Geography Dawson, Christian; Loughborough University, Computer Science Ab Ghani, Ngahzaifa; University of Nottingham, School of Geography
Keywords:	suspended sediment, data-driven, rating curve, modeling, operational validity

The need for operational reasoning in data-driven rating curve prediction of suspended sediment.

Nick J. Mount [a,*], Robert J. Abrahart [a], Christian W. Dawson[b] and Ngahzaifa Ab Ghani [a,c]

[a] School of Geography, University of Nottingham, Nottingham, NG7 2RD, UK

[b] Department of Computer Science, Loughborough University, Loughborough, LE11 3TU, UK

[c] Faculty of Civil Engineering and Earth Resources, Universiti Malaysia Pahang, 26300 Kuantan, Pahang Darul Makmur, Malaysia

* corresponding author. Tel: +44 115 951 5438; Fax: +44 115 951 5249; E-mail address: nick.mount@nottingham.ac.uk

Abstract

The use of data-driven modelling techniques to deliver improved suspended sediment rating curves has received considerable interest in recent years. Studies indicate an increased level of performance over traditional approaches when such techniques are adopted. However, closer scrutiny reveals that, unlike their traditional counterparts, data-driven solutions commonly include lagged sediment data as model inputs and this seriously limits their operational application. In this paper we argue the need for a greater degree of operational reasoning underpinning data-driven rating curve solutions and demonstrate how incorrect conclusions about the performance of a data-driven modelling technique can be reached when the model solution is based upon operationally-invalid input combinations. We exemplify the problem through the re-analysis and augmentation of a recent and typical published study which uses gene expression programming to model the rating curve. We compare and contrast the previously-published, solutions, whose inputs negate their operational application, with a range of newly developed and directly comparable traditional and data-driven solutions which do have operational value. Results clearly demonstrate that the performance benefits of the published gene expression programming solutions are dependent on the inclusion of operationally-limiting, lagged data inputs. Indeed, when operationally-inapplicable input combinations are discounted from the models, and the analysis is repeated, gene expression programming fails to perform as well as many simpler, more standard multiple linear regression, piecewise linear regression and neural network counterparts. The potential for overstatement of the benefits of the data-driven paradigm in rating curve studies is thus highlighted.

Keywords

Suspended sediment; data-driven; rating curve; modeling; operational validity

1. INTRODUCTION

In this paper, we highlight arguments surrounding the operational applicability of data-driven discharge / suspended sediment rating curve solutions that use lagged data as inputs. This is an issue that has received no comprehensive attention despite more than 10-years of data-driven, suspended sediment modelling research activities. Moreover, it is a highly significant issue as the development of any model that cannot actually be used inevitably raises questions about the value of the modelling that is being done and the approaches that are being used to do it. We define the operational applicability of a data-driven solution simply as the extent to which its input and parameter requirements limit its application with regard to common operational tasks; a definition which is distinct from a solution's functional performance (i.e. a poorly performing solution may be operationally-applicable, whereas a better performing solution is not). Indeed, this definition conforms to others which *emphasise the need for solutions rather than the search for hydrological knowledge per se* (Wilby and Davies, 1997; p. 195). In this study we repeat and augment the analysis of a typical, recently-published data-driven rating curve problem and, in so doing, exemplify how solutions that have no operationally application, may nonetheless be evaluated as offering better functional performance than their operationally-applicable counterparts. This can result in an over-statement of the data-driven paradigm's ability to provide better solutions to end-users. To this end we replicate the basic modelling methodology of Aytek and Kisi (2008), who develop a set of data-driven rating curve solutions for the Tongue River, USA and augment this with a range of new, operationally-applicable solutions for comparison. By repeating previously published methods we include certain methodological weaknesses that we identify in the original work, and which are highlighted at appropriate points in the text. However, these inclusions are justified by the comparative nature of the study which requires the exact re-application of the original methodological approach if the new, additional solutions are to be adequately compared to the originals. The aim of this study is, therefore, not to develop a comprehensive critique or reanalysis of the original modelling methodology, but to re-apply it so that an illustrative case study emerges, which demonstrates the importance of including both functional performance assessment and a higher-level rationale of the operational application of data-driven rating curves solutions. This paper is the third in a series by the authors that critically evaluates different aspects of data-driven methodologies used in suspended sediment modelling. In Mount and Abrahart (2011a), we examine how data-series should be delivered to the modelling process (i.e. as concentration or load / logged or

unlogged) and the impact that decision has on the form and consistency of outputs. In Abrahart et al. (2011) we examine how contextual hydrological and dataset knowledge can be incorporated into the process of developing and selecting data-driven suspended sediment models and call for an end to blind model justification on the sole basis of goodness-of-fit metrics.

In recent years, practical opportunities for the implementation of data-driven modelling mechanisms has caught the attention of the hydrological modelling community and this has included efforts to develop superior discharge / sediment rating curve solutions. The use of a data-driven model (DDM) in this way has been described as advanced curve fitting, in which the form of the response function is unrestricted by the *a priori* constraints imposed in conventional, empirical models (Mount and Abrahart, 2011a). Satisfactory combinations of simplicity and predictive power in both single-input single-output (SISO) and multi input-single output (MISO) versions have been reported (c.f. Kisi, 2005; Zhu et al. 2007). The value of such models in suspended sediment forecasting and hindcasting has long been recognised as most catchments lack the comprehensive suspended sediment monitoring equipment from which to quantify suspended sediment directly. Consequently discharge-driven rating curves are often used as operational tools for estimating short and long-term suspended sediment yield in catchments and for interpolating missing values in suspended sediment records (e.g. Stott and Mount, 2007). However, the risks associated with their use, especially for estimating high-magnitude / low frequency data, should not be underestimated (Walling, 1977; Walling and Webb, 1988).

In the data-driven paradigm, one or more input data sets that are presumed useful predictors of suspended sediment, are interrogated using machine learning and artificial intelligence algorithms with the optimal form of the response function being learned directly from patterns and structures that are present the data sets. A comprehensive review of data driven approaches in hydrology can be found in Solomatine et al., (2008). The standard conceptual and practical modelling processes used by data-driven modellers is shown in Figure 1, with an ordered set of more abstract, reasoning elements informing the practical workflow of the modeller. Conceptual reasoning is undertaken to formulate the overall concept and goal of the modelling activity (e.g. process knowledge discovery versus curve fitting) and to determine the appropriateness of a data-driven versus knowledge-driven approach (Dubois et al., 2000). It also determines the extent to which evidence of causal hydrological relationships should be

captured in the suspended sediment model's response function and predictor set selections and, therefore, the conceptual set of predictive drivers for which data will be required.

Procedural reasoning informs the decisions governing data set acquisition and any processing procedures applied to optimise the model's predictive power and robustness. In a practical sense it guides how each predictor is objectified as a concrete data set that can be delivered to the model as an input. This includes any raw data manipulations and transformations applied to improve the model's predictive power or robustness, such as temporal lagging (Ciğizoğlu (2002); Ciğizoğlu and Kisi (2006); Aytek and Kisi (2008); Partal and Ciğizoğlu (2008); Cobaner et al. (2009); and Kisi (2009)), log-transformation (Mount and Abrahart, 2011a; Alp and Ciğizoğlu, 2007) and unit conversion to remove spurious correlation (McBean and Al Nasri, 1988; Annadale, 1990; Nordin, 1990; Wahl, 1990; Milhous, 1990; McBean and Al Nasri, 1990 a,b;). Functional reasoning directs the modeller's identification of an optimal data-driven modelling algorithm or technique, and/or the identification and selection of the optimum combination of input data sets delivered to it. In practice the approaches tend to be rather ad hoc and partial in scope. The array of possible input combinations are seldom modelled exhaustively, and preferred algorithms and/or input combinations are usually identified through reference to one or more goodness-of-fit metrics.

Parallels can be drawn between the reasoning elements of data-driven suspended sediment models detailed in Figure 1, and many of those underpinning physically-based hydrological models. However, at present the reasoning used to inform the development of DDMs is too heavily biased towards functional elements. Indeed, papers assessing the relative performance of different data-driven algorithms, techniques and model configurations under different modelling scenarios are far more numerous than those which examine the influence that the modeller's reasoning processes may be having on the usefulness or correctness of their assessments, or the physical interpretation of their results – issues that have received far more attention from physically-based modellers (c.f. Beven, 1989). As a consequence, a number of key criticisms of the data-driven paradigm applied to suspended sediment modelling remain largely unaddressed including a general:

1. lack of justification for the model configuration and response function structures used (Minns and Hall, 1996; Babovic, 2005; Solomatine et al., 2008);

2. lack of established procedures to ensure robust model configurations and outputs (Mount and Abrahart, 2011a; Abrahart et al., 2008a);

3. lack of justification for using more complex, data-driven response functions over simpler empirical counterparts (Mount and Abrahart, 2011b)

4. over-reliance on simplistic, goodness-of-fit metrics in the identification of preferred models (Legates and McCabe, 1999).

5. lack of operational applications for the model *i.e.* can the resultant model actually be applied for an operational purpose? (Abrahart et al., 2008b; Abrahart and Mount, 2011).

The last of these is particularly important as, without a justification of its operational applicability, the practical value of a model becomes unclear and the conceptual reasoning underpinning the modelling goal becomes open to question. The problem is well exemplified by the publication of numerous suspended sediment studies in which the 'best' model input configuration incorporates current discharge together with some mix of lagged discharge and lagged sediment records e.g. Ciğizoğlu (2002); Ciğizoğlu and Kisi (2006); Aytek and Kisi (2008); Partal and Ciğizoğlu (2008); Cobaner et al. (2009); and Kisi (2009a).

The widespread use of lagged sediment as a predictor has, to date, been supported by functional reasoning, which asserts that DDMs produce better goodness-of-fit metrics if such inputs are used (e.g. Kisi, 2005; Aytek and Kisi, 2008). However, it is important to recognise that the use of past sediment records as an input to the modelling process not only reduces the problem to something approaching a near-linear modelling operation (in which case the need for using a DDM becomes questionable anyway), but it also makes little operational sense. This is because a full set of observed measurements of suspended sediment is needed as an input to the model for the period being modelled minus the lag. Consequently, there would be no need to model the suspended sediment record as it would already, by definition, be known at least up until the period of the lag. Similarly, extrapolation of suspended sediment outputs beyond the period for which suspended sediment values are known is not supported. Indeed, problems even remain if the prediction of sediment response is made only within the period of the lag. This is because the standard inclusion of current discharge input means that the related sediment response would have happened before it could be modelled – negating the value of the prediction. Finally, model solutions cannot be transferred to similar rivers, or to

different periods, if no observed suspended sediment records are available for use as inputs to the sediment prediction model. The supposed advantages of both implicit and explicit DDMs are thus negated on several counts since the resultant models and/or equations are restricted to the original stations for which sediment data must be collected and transferring unique solutions to different reaches or catchments makes neither conceptual nor operational sense.

Given the importance of such issues, the explicit inclusion of an operational reasoning element to inform the modeller about how a particular data-driven model configuration may restrict its operational value would appear essential. To this end, Figure 1 can be amended so that operational reasoning is explicitly included (Figure 2), and model input configurations which result in models of little operational value can be identified and rejected prior to any functional efforts to identify 'best' or 'preferred' models.

In this paper, we demonstrate how incorrect conclusions about the performance of a data-driven modelling technique can be reached when the model solution is based upon operationally-invalid input combinations. This is achieved through a re-appraisal of two gene expression programming (GEP) models developed for the Tongue River, USA, originally published by Aytek and Kisi (2008). Each model, developed on operationally-invalid input combinations, is compared to a range of new GEP solutions, developed in an identical manner to the original, but with only operationally-valid inputs included. In addition we develop a range of new, operationally-valid linear and non-linear model counterparts against which the performance of both GEP solutions can be assessed. Consequently, using the context of the Tongue River, this paper examines:

1. the impact of removing operationally-invalid input combinations on the explicit outputs and goodness-of-fit metrics obtained by the GEP technique and;

2. the extent to which the GEP maintains its functional advantages over other, increasingly sophisticated modelling approaches when operationally-invalid model configurations are excluded from the analysis.

The paper progresses with a review of the Tongue River data sets, a re-examination of the input combinations used by Aytek and Kisi (2008) and identification of those combinations that lack operational application. We then repeat Aytek and Kisi's original analysis using only the input combinations deemed to be of operational value. The results from the original work and those obtained from the re-analysis are then compared and the impact of rejecting

the operationally-invalid input combinations is highlighted. Finally, we compare the results obtained from the re-analysed GEP approach and those obtained from sediment rating curve, linear regression and non-linear neural network counterparts.

2. THE TONGUE RIVER: STUDY AREA AND DATA

2.1. Overview

The Tongue River data set has been a focus of several recent papers examining data-driven modelling approaches in hydrology (Kisi, 2004; Aytek and Kisi, 2008; Guven and Kisi, 2010). Of particular importance to this study is the paper by Aytek and Kisi (2008) in which an explicit formulation of the sediment-discharge relationship in the Tongue River is developed using GEP. The Tongue River watershed encompasses some 13,983 km^2 of Wyoming and Montana (Figure 3). This river is a tributary of the Yellowstone River, it rises in the Big Horn Mountains of Wyoming, flows through northern Wyoming and southeastern Montana, and then empties into the Yellowstone River at Miles City, Montana. The river is 396 km long and the watershed is for the most part rural. From elevations of 2,400–3,000 m, it drops to low, rugged mountains and badlands. Below the mountains the stream runs through a long, narrow valley confined by high bluffs and terraces. HydroSolutions Inc. (2008) report that streamflow is driven by precipitation, although the relationship is complex. Variations in the pattern and timing of precipitation over the basin, and lag time between snowfall and snowmelt, are some of the complicating factors. The river is fed by winter snow pack from the higher elevations of the Big Horn Mountains, by early snow runoff in the lower elevations of the drainage basin, and by ground water springs. The river rises in March and April due to snowmelt in the lower elevations, and again in June as summer weather melts the higher elevation snow pack. In the plains region, with elevations from 900 m to 1800 m above mean sea level, annual average precipitation ranges from 250 to 350 mm, and rainfall is the more dominant form. Average monthly precipitation is greatest from April through September, and maximum temperatures occur in July, while minimum values occur in January. About 75 percent of the annual precipitation falls as rain during the April-September growing season. May and June are usually the wettest months of the year.

In accordance with standard rating curve approaches, two datasets were used in Aytek and Kisi (2008). They comprise daily time series of streamflow (Q in m^3s^{-1}) and suspended

sediment (S in ton day^{-1}) records for an upstream station (UPST: Station No: 6307830, situated below Brandenberg Bridge near Ashland) and a downstream station (DNST: Station No: 6308500 at Miles City) on the Tongue River, in Montana, USA. The reported upstream drainage area for UPST is 10,521 km^2; for DNST it is 13,932 km^2.

2.2. Data subsets for model training and testing

In the subsequent analyses we replicate the data acquisition, processing and sub-setting procedures outlined in Aytek and Kisi (2008). Every available record for the two test stations was downloaded from the US Geological Survey (USGS) web server (http://webserver.cr.usgs.gov/sediment) on 01 April 2011. The full set of downloaded material comprised daily records for 1 October 1975 - 30 September 1981 at UPST and for 31 August 1977 - 4 December 1985 at DNST. Modelling was performed on a common overlapping subset of the full available record: a 3-water-year period that spanned 1st October 1977 – 30th September 1980 (SET A) that was used for model development / training purposes; and a 1-water-year model testing period that spanned 1st October 1980 – 30th September 1981 (SET B) that was used from model testing. The record for 31 August 1977 was also used in the modelling processes to fill a missing record at t-1 for 1 September 1977.

Time series plots and scatterplots for the UPST and DNST split sample datasets are provided in Figure 4.

The statistical parameters of daily stream flow and daily sediment for both stations are given in Table 1. The discharge and suspended sediment load dataset for both stations, particularly for the training datasets (SET B), is observed to be highly variable, highly peaked and highly skewed. Large magnitudinal differences in sediment, but not in discharge, are observed to exist between the two gauging stations. Such differences reflect their relative positions in the catchment with regard to contributing sources. Large magnitudinal differences also exist between SET A (training data) and SET B (testing data) at each station such that final model assessment is performed on what equates to a minor, limited and somewhat unrepresentative fraction of a larger set of processes appertaining to that region. The largest events are part of the development period, thus sidestepping the need to perform difficult extrapolation tasks

for peak sediment loads during testing, but also increasing the probability of overestimating low(er) sediment values. Similarly, the smallest magnitude events are unevenly distributed amongst the development and testing datasets. The minimum values for SET A at DNST are higher than those of its corresponding testing dataset, which could cause downward extrapolation difficulties in estimating lower sediment values.

Figure 4 identifies a number of important outliers in the dataset, arising from some common set of non-linear processes that a DDM might be expected to encapsulate, and that should not simply be interpreted as measurement or sampling oddities.. These extreme values have their origins in snowmelt processes that are unlikely to be captured by a data-driven model trained on relatively short data records and which do not include local climatic and/or meteorological predictors. Indeed, Figure 4 shows low flow conditions, interrupted in March and April by lower elevation snowmelt, and again in June by higher elevation snowmelt. This produces high discharges, sediment flushing and probable large clockwise hysteresis loops. The result is that at DNST, two observed suspended sediment load values exceed 80,000 ton day^{-1}, whilst the remaining sediment values are below 50,000 ton day^{-1}. Similarly, at UPST, one observed suspended sediment load value exceeds 27,000 ton day^{-1}, whilst the other values are below 20,000 ton day^{-1}.

Table 2, which provides a cross correlation matrix for SET A and SET B, also reveals potential problems. Higher correlation coefficients for all variable pairs are reported in SET B, in comparison to SET A, and this is likely to lead to artificially high testing scores. Moreover, the highest correlation against S is St-1; and the correlation between S and St-1 is much higher for SET B, compared to SET A. Consequently there is a danger that autocorrelation in the predicted dataset, is confused with causation, resulting in poorly constructed arguments for the inclusion of past sediment inputs as drivers.

3. MODEL INPUT CONFIGURATION SELECTIONS

In Aytek and Kisi's original paper, models are developed on seven input combinations comprising Q_t, Q_t-1, S_t-1 and S_t-2 (Table 3), although the remaining 8 input combinations that are theoretically possible are not included in the analysis and a justification for their lack of inclusion is not provided. Thus conclusions drawn in the study should not be generalised but considered as observations, specific only to, and valid in, the particular cases which are

listed (Aksoy et al., 2007). When operational reasoning is applied, and the use of lagged suspended sediment inputs is thus discounted, the candidate input set is reduced to Q_t and Q_t-1, resulting in only three possible input combinations: [Q_t]; [Q_t-1]; [Q_t, Q_t-1]. The first combination equates to concurrent input-output modelling in a similar manner to that of a traditional sediment rating curve approach. The second is, in effect, a lagged version of the traditional rating curve and, therefore, one would not expect its performance to exceed that of Qt. Consequently, we have discounted its use in this study; again on grounds of operational reasoning (the operational value of an inferior model configuration is difficult to justify). The third combination was not considered in the original paper despite its clear potential for improved modelling where complex hysteresis may exist in the record. Consequently, we do include it in our subsequent analyses.

4. TONGUE RIVER RE-ANALYSIS: MODELLING APPROACHES

In their original paper, Aytek and Kisi (2008) conclude that *"the results obtained with GEP models are better than those obtained using the conventional rating curve and multiple linear regression techniques"*. They also assert that *"the results suggest that GEP may provide a superior alternative to the sediment rating curve and multiple linear regression techniques"* (p. 297). Similar conclusions remain in more recent studies (c.f. Guven and Kisi, 2010). However, these conclusions are based on comparisons between rating curve and regression solutions and the best performing GEP solution, which is developed on operationally-inapplicable input combinations (i.e. [Q_t, Q_{t-1}, S_{t-1}]). Consequently, there is a need to assess the relative performance of GEP solutions developed on operationally-applicable input combinations with i) Aytek and Kisi's best performing GEP solution and ii) other operationally-applicable linear and non-linear solutions.

To that end, Aytek and Kisi's preferred GEP solutions (Figures 3 and 5 and Equations 7 and 14 from the original paper), which were developed using inputs [Qt, Qt-1, St-1] were;

1. re-modelled with two new GEP solutions developed on inputs [Qt] and [Qt, Qt-1], using the same basic methodology and settings used by Aytek and Kisi (2008);

3. re-modelled with three additional statistical and data-driven approaches that provide a range of linear and non-linear counterparts against which the relative performance of the GEP solutions can be assessed.

The modelling approaches are detailed below. Models developed on one input are labelled 1 [Qt]; models developed on two inputs are labelled 2 [Qt, Qt-1].

3.1. Symbolic regression gene expression programming approach

GEP solutions for UPST and DNST were developed in GeneXproTools 4.0 (Ferreira, 2001; 2006; http://www.gepsoft.com/). For each station two different models were produced, providing St output on inputs [Qt] and [Qt, Qt-1]. Wherever possible parameter settings used by Aytek and Kisi (2008) were replicated and, unless their methods indicated otherwise, default settings were adopted. However, some apparent inconsistencies in the methodological descriptions in their paper meant that some minor assumptions had to be made. Specifically, the original paper (p.290) reported that the function set used contained four basic arithmetic operators (+, -, *, /) and six basic mathematical functions (√, ln(x), log(x), e^x, 10^x, power) — yet their final models, inter alia, contained abs(), x^2 and $\sqrt[3]{}$ functions. Similarly, the original paper (p.290) reported that the fitness function used was 'Absolute Error With Selection Range' (AESR) — yet mean absolute error is subsequently reported as the fitness function on p. 291. Since neither issue could be resolved from the original paper, our method follows what is documented in the methodology on p.290 of Aytek and Kisi's paper. It should be noted that Aytek and Kisi's choice of the fitness function settings (SR =100, p =0.1) is of importance because major events that deliver large errors on specific models could fall outside the selection range (SR) and would therefore be treated as outliers and excluded from the fitness landscape that is used to evolve the solutions. Consequently, snow melt instances are not likely to be properly accommodated during the modelling process. Finally, no 'stopping condition' was reported in the original paper, so we opted to stop at 10,000 generations. Our settings are listed in Table 4.

3.2. Sediment rating curve counterpart

Aytek and Kisi (2008) provide a bias-corrected (Ferguson, 1986) single-input single-output sediment rating curve solution (e.g. Asselman, 2000) as a benchmark comparison for their transparent GEP solutions. We include this counterpart in our analysis, accepting that it represents an important baseline comparator against which the potential benefits of more complex, data-driven models should be assessed. The rating curve comprises a bias corrected,

linear least squares regression model of log S_t on log Q_t, depicted in the manner of a power function. The equation is provided in (1) below:

$$SRC = a \,.\, CF \,.\, Q^b \quad (1)$$

For UPST a = 0.4296, b = 2.1022, CF = 1.389

For DNST a = 0.7066, b = 2.0589, CF = 1.496

3.3. Linear regression counterparts

Linear models (Raghuwanshi et al., 2006) should be used as benchmarks against which non-linear data-driven modelling methodologies are tested in order to establish the extent to which the numerical relationship that is presented for modelling is linear or near-linear. Without such comparators it is impossible to properly justify the need for the application of more complex modeling methodologies, irrespective of theoretical arguments about the established nature of the underlying scientific process that is of interest (Abrahart and See, 2007a; 2007b). Therefore, linear modelling solutions should always be used to (i) test the numerical strength of empirical relationships in each dataset and (ii) identify the nature and extent of residual non-linearities that necessitate the adoption of a different sort of additional model to represent such factors (Curry and Morgan, 2003).

For each station, as a baseline linear comparator, ordinary least squares linear regression (OLS) models were developed. Two different models were produced, providing St output on inputs [Qt] and [Qt, Qt-1], named OLS1 and OLS2 respectively. Equation parameters for the OLS regression counterparts are provided in Table 5.

In the present study the 'Waikato Environment for Knowledge Analysis' Data Mining Toolkit v3.6 (WEKA: Witten and Frank, 2005) was used to derive M5 Model Tree (M5MT) piecewise linear regression predictions of St on inputs [Qt] and [Qt, Qt-1], termed M5MT1 and M5MT2. The M5MT algorithm Quinlan (1992) splits the input data into non-intersecting regions and thereafter fits a linear regression model to each of the data subsets. The size of the regions progressively narrows, producing increasingly complex piecewise models.

Exhaustive search is used to examine all possible splits, to select the one that delivers maximum reduction in the standard deviation of the resulting models, and the splitting process terminates when no significant variation in the result is achieved via further splitting. M5MT has been used to predict harbour basin sedimentation in the Port of Rotterdam (Bhattacharya and Solomatine, 2006); to model bed-load and total-load sediment transport in alluvial channels (Bhattacharya et al., 2007); and for modelling suspended sediment load in rivers (Janga Reddy and Ghimire, 2009). Upstream M5MT solution parameters are presented in Table 6, with downstream solution parameters given in Table 7.

3.4. Non-linear neural network counterparts

Numerous different types of neural network (NN) exist with feedforward solutions, that are trained using the back propagation of error algorithm, providing one of the most widely used data-driven approaches in hydrological modelling. These include applications for sediment yield estimation (Abrahart and White, 2001). They therefore represent an established non-linear modelling benchmark against which more novel approaches can be compared.

NN models producing St output on inputs [Qt] and [Qt, Qt-1], named NN1 and NN2, were developed for each station using an in-house program, written in Pascal, that has delivered sound performance on a number of previous occasions using similar settings e.g. Dawson et al. (2002, 2006). Ten different architectures were used to model each station: each setup comprising 1:{1,5}:1 and 2:{1,5}:1 configurations, in which the number of hidden units {1,5} used in a particular model is indicated by means of brackets e.g. NN1(1). In all cases, prior to modelling, each dataset was standardised in a linear manner to a common range {0.1-0.9}. Each network configuration was thereafter trained using the traditional 'back-propagation of error with momentum' algorithm. The algorithm parameters (objective function = sum squared error; learning rate = 0.1; momentum factor = 0.9; number of epochs = 20,000) are commensurate with the relatively simple NN architecture being used in this paper. Each model was assessed on root mean squared error for the training and testing datasets at 1000 epoch intervals such that a preferred solution could be selected. These configurations mirror those adopted in other hydrological modelling tasks of similar scope and scale.

5. EVALUATION METRICS

It is standard practice to assess the relative performance of a given model via reference to one or more quantitative evaluation metrics. However, the selection and use of these statistics is often problematic and it is sometimes difficult for readers or users to interpret how well a particular model reproduces the observed dataset or how well a model compares with other models (American Society of Civil Engineers, 1993; Legates and McCabe, 1999). To provide a transparent comparison with past modelling efforts, this paper includes identical metrics to Aytek and Kisi (2008) – namely root mean square error (RMSE) and R-squared (RSqr). These two metrics are strongly influenced by a model's replication of lower frequency, medium and high magnitude events in the data. Although these comprise a relatively small proportion of the observations in the data set, they are arguably some of the most important events in an operational context as they may be responsible for delivering much of the sediment yield of a catchment. All evaluation statistics were generated using HydroTest (http://www.hydrotest.org.uk): a standardised, open access web site that performs the required numerical calculations (Dawson et al., 2007; 2010). This service supports a broad spectrum of quantitative tests and provides descriptive statistics for the comparison of observed and predicted datasets. It also documents the underlying equations upon which the calculations are based.

6. COMPARISON OF GEP SOLUTIONS

One of the key benefits of GEP is the production of an explicit set of tree diagrams which provide an explicit documentation of the model solution. These can then be used to re-apply the solution to other data, or can be used as the basis for comparison to other GEP solutions. Gene expression tree diagrams for Aytek and Kisi's (2008) UPST solution, together with those for UPST-GEP1 and UPST-GEP2 are presented in Figure 5. Observed versus predicted plots for training and testing periods for each sub expression, together with the combined response function, are presented in Figures 6 and 7. By plotting individual sub expressions, it becomes possible to identify the importance of the contribution that each makes to the combined response function. Moreover, as not all inputs are necessarily utilised in each sub expression (for example see Aytek and Kisi Sub-ET 1 where lagged sediment is not included) it becomes possible to determine whether all inputs are implicated as main drivers of the combined response function.

The patterns presented in Figures 6 and 7 demonstrate that for all GEP solutions one dominant sub expression comprises the majority of the combined response function, with the other two sub functions providing small, local adjustments. In the case of Aytek and Kisi's solution, Sub-ET 3 is the dominant sub function and models the global trend in the data, particularly at the lower data ranges. The non-dominant expression tree Sub-ET 2 can be seen to have little influence on lower-range data, but particular influence on the upper-range data where it acts to increase their values. Sub-ET 1 displays a similar pattern to Sub-ET 2, but in a much subdued manner. Importantly, both Sub-ET 3 and Sub ET 2 include St-1 as an input and, in so doing, lagged sediment is implicated as an important driver of their solution; particularly for modelling upper-range data. In the case of GEP1 and GEP2 Sub-ET 1 acts as the dominant sub function with the other two providing negligible input – suggesting that three sub expression trees in solutions based solely on Qt and Qt-1 is unnecessarily complex. The inclusion of Qt-1 in Sub-ET 1 of GEP2 indicates that, when made available, lagged discharge is selected as an important driver of the model.

The solutions for the DNST gauge (Figures 8-10) follow a similar basic pattern to those for UPST, with a single sub expression capturing much of the global trend in the data and, in so doing, comprising the majority of the combined response function in all solutions. Again, lagged sediment is included as an input to the dominant sub expression of Aytek and Kisi's solution and to Sub-ET 2, both of which act to increase the two highest data points. Sub-ET 1 contributes little to Aytek and Kisi's combined response function; again indicating that the use of three sub expressions is more complex than the modelling problem demands. GEP1 and GEP2 are seen to perform comparatively poorly, especially in their ability to model the two high magnitude data points. As in the UPST models, a single sub-expression tree is responsible for almost all of the response function.

Evaluation metrics for the models are presented in Table 8. The explicit solutions published by Aytek and Kisi (2008) make it possible to compute the full range of evaluation metrics, for both SET A and SET B data, despite the fact that only SET B results are reported in the original paper. In all cases, Aytek and Kisi's solution outperforms GEP1 and GEP2. The inclusion of Qt-1 in GEP2 is seen to enhance its performance compared to GEP1, but without

the inclusion of a lagged sediment input, the evaluation metrics for GEP1 or GEP2 fail to equate to those of Aytek and Kisi's solution by a substantial margin.

The expression tree plots highlight the critical importance of lagged sediment in adequately modelling high magnitude instantaneous suspended sediment values. This is reflected strongly in the evaluation metrics which are particularly influenced by the effectiveness with which upper-range data are modeled. Both RMSE and RSqr values indicate Aytek and Kisi's lagged suspended sediment model as performing consistently, and considerably better, especially in upper-range data. However, when operational reasoning is incorporated, and lagged suspended sediment is rejected as an input, the ability of GEP to model suspended sediment is seen to be significantly limited, with RMSE and RSqr statistics reduced. This therefore raises the important question of whether GEP maintains its performance advantage under such a scenario; a question which is addressed through the subsequent comparison of GEP1 and GEP2 solutions against a range of linear and non-linear counterparts.

7. COMPARISON OF GEP AND COUNTERPART SOLUTIONS

Comparison of GEP1 and GEP2 solutions to: i) SRC, ii) linear counterparts, and iii) non-linear counterparts are presented in Tables 9, 10 and 11-12 respectively. Comparisons between GEP1, GEP2 and SRC indicate mixed results (Table 9). Metrics for the UPST gauge indicate GEP2 offers a generally better solution than either GEP1 or SRC, with it resulting in the best metric scores for all but SET B RMSE. Scores from the more complex SET A data highlight significant improvement in RMSE when GEP solutions are applied, irrespective of whether Qt-1 is included as an input; whilst RSqr scores are roughly comparable across the models. This improvement in RMSE is, however, not observed in the SET B data and there is relatively little difference between GEP1, GEP2 and SRC scores. This is probably a result of SET B lacking the outliers evident in SET A, which are likely to be influencing the RMSE values. At the DNST gauge GEP performs slightly less well than SRC for both SET A and SET B data, with the inclusion of lagged discharge in GEP2 worsening the metric scores. Given the simplicity of an SRC model over GEP, these results indicate that for operationally-applicable models, the functional performance benefits of GEP over SRC solutions are difficult to argue.

When operationally-applicable GEP solutions are compared to their linear counterparts (Table 10), the evaluation metrics provide a clear picture. GEP performs similarly to its OLS counterparts with metrics for GEP2 and OLS2 (both including Qt-1 as an input) indicating that simple multiple linear regression provides a better model than the more complex GEP solution. This finding is in clear contradiction to Aytek and Kisi's (2008) paper, which identified the multiple linear regression solution as having inferior metric scores compared to its GEP counterpart. Whilst the advantages of a piecewise linear solution based solely on [Qt] (M5MT1) is minimal, the piecewise solution based on [Qt, Qt-1] (M5MT2) has the best metric scores for both UPST and DNST gauges and for both SET A and SET B data. These results therefore indicate that either a multiple linear regression or a piecewise linear solution incorporating lagged discharge is preferable to its counterpart GEP solution.

The message emerging from our NN non-linear findings (Tables 11 and 12) is not as well defined as that delivered by the linear modelling solutions (Tables 9 and 10). For models based solely on [Qt] (Table 11) a general pattern of marginal improvement in metric scores is evident in the NN solutions with those of NN1(3) suggesting it to be the best overall model for SET A data at the UPST and DNST gauge. For set B data the picture is less clear with GEP1 providing the best RSqr scores and NN1(2) providing the best RMSE scores at both UPST and DNST gauges. For solutions based on Qt and Qt-1 (Table 12), NN solutions result in better metric scores than their GEP counterparts, although it is clear that different NN configurations perform differently at UPST and DNST gauges, and for SET A and SET B data. Those NN solutions with high numbers of hidden units (NN2(4) and NN2(5)) have the best metric scores on SET A data, and this reflects the relative complexity of that data set, which is best modeled by a more complex NN configuration. The more simple SET B data are well modeled by NNs with lower numbers of hidden units, particularly at the DNST gauge where the relationship between discharge and suspended sediment is least complex. It is important to note that the higher the number of hidden units in an NN, the greater the chance of over-fitting. Normally this is detected by ensuring a comparable degree of fit for the solution to training and testing data. However, this relies on the training and testing data sets containing mutually representative statistical and temporal patterns. The data split used by Aytek and Kisi (2008), and replicated here, has resulted in data sets that are not particularly representative of one another (Table 1 and Figure 4) and it is, therefore, difficult to discount over-fitting in those NN solutions with higher numbers of hidden units. However,

the fact that even the NN solutions with low numbers of hidden units (i.e. those less likely to be overfitted) result in better metric scores than their GEP counterparts is strong evidence of the preferential solutions delivered by NN approaches. In the context of the operationally-applicable models for the Tongue River, NN solutions are therefore preferred over their GEP counterparts. It is possible that different objective functions could deliver contradictory results, but a comprehensive exploration of such matters is beyond the scope of our current paper.

8. SUMMARY

This re-analysis of suspended sediment modelling for the Tongue River dataset provides an important example that demonstrates how the inclusion of lagged suspended sediment as an input can be an essential factor in data-driven modelling techniques offering better functional performance than their linear and non-linear counterparts. If lagged sediment is rejected as a GEP model input, the performance of Aytek and Kisi's original solution can not be matched; either by other GEP solutions or by the range of counterpart models. Indeed, the operationally-applicable GEP solutions are seen to be some of the worst performers. However, several operationally-applicable counterparts are able to achieve evaluation metric scores that are close the Aytek and Kisi's original GEP solution (Table 13), and in so doing, offer an operationally useful alternative with only minimal reduction in evaluation statistics. Of particular note in this regard are NN and piecewise linear regression solutions based on [Qt, Qt-1]. Whilst NN solutions lack the explicit outputs of their GEP counterparts and may be prone to overfitting, M5MTs have the advantage of being both explicit and robust.

9. CONCLUSIONS

This study highlights the fact that input configurations which include lagged sediment, and which are commonly used in data-driven suspended sediment rating curve models, lack operational justification and have little operational value. Their inclusion in future studies should, therefore, either be explicitly justified with respect to the conceptual purpose of the modeling exercise, or rejected. It also highlights how sensitive the performance of GEP modeling approaches are to their input configurations in the case of this example data set. These two factors support the authors' call for the inclusion of a proper appraisal of the operational-applicablity of different input configurations to a data-driven modeller's

workflow before any functional evaluations based on evaluation metrics are made – an element that is currently missing.

The original input configurations used by Aytek and Kisi (2008) in their analysis of the Tongue River lacked operational application due to their inclusion of lagged sediment as an input. Their GEP models did indeed produce improved evaluation metric scores. Therefore, on the basis of the purely functional reasoning employed in their study, their claims that GEP "*may provide a superior alternative to the sediment rating curve and multiple linear regression techniques"* are supported. However, the re-analysis presented here clearly demonstrates that, once operational reasoning is applied and operationally-inapplicable input combinations are discounted, GEP fails to perform as well as many simpler, more standard multiple linear regression, piecewise linear regression and NN counterparts. Indeed, the reported superiority of GEP in the Tongue River is shown to be dependent on the inclusion of an operationally-invalid lagged sediment input.

The conclusion of the study is, therefore clear. Operational reasoning should be an essential element in all data-driven suspended sediment modeling workflows to ensure that:

1. the resultant models can actually be used for operational purposes and;
2. the performance of complex data-driven solutions are not overstated in comparison to their simpler model counterparts.

REFERENCES

Abrahart RJ, Mount, NJ. 2011. Discussion of "Neuro-fuzzy models employing wavelet analysis for suspended sediment concentration in rivers". *Hydrological Sciences Journal* **56**: 1-5.

Abrahart RJ, Mount NJ, AbGhani N, Clifford NJ, Dawson CW. 2011. DAMP: a protocol for contextualizing goodness-of-fit statistics in sediment-discharge data-driven modeling. *Journal of Hydrology* **409**: 596-611.

Abrahart RJ, See LM, Dawson CW. 2008a. Neural network hydroinformatics: maintaining scientific rigour. In: *Practical Hydroinformatics*, Abrahart RJ, See LM, Solomatine DP (Eds.), 33-48, Springer-Verlag, Berlin, Heidleberg.

Abrahart RJ, See LM, Heppenstall AJ, White SM. 2008b. Neural network estimation of suspended sediment: potential pitfalls and future directions. In: *Practical Hydroinformatics*, Abrahart RJ, See LM, Solomatine DP (Eds.), 139-164, Springer-Verlag, Berlin, Heidleberg.

Abrahart RJ, See LM. 2007a. Neural network emulation of a rainfall-runoff model. *Hydrology and Earth System Sciences Discussions* **4**: 287-326.

Abrahart RJ, See LM. 2007b. Neural network modelling of non-linear hydrological relationships. *Hydrology and Earth System Sciences* **11**: 1563-1579.

Abrahart RJ, White SM. 2001. Modelling Sediment Transfer in Malawi: Comparing Backpropagation Neural Network Solutions Against a Multiple Linear Regression Benchmark Using Small Data Sets. *Physics and Chemistry of the Earth, Part B* **2**: 19-24.

Aksoy H, Guven A, Aytek A, Yuce MI, Unal NE. 2007. Discussion of "Generalized regression neural networks for evapotranspiration modelling", *Hydrological Sciences Journal* **52**(4): 825-828.

Alp M, Ciğizoğlu HK. 2007. Suspended sediment load simulation by two artificial neural network methods using hydrometeorological data. *Environmental Modelling & Software* **22**: 2-13.

American Society of Civil Engineers. 1993. Criteria for evaluation of watershed models. *Journal of Irrigation and Drainage Engineering* **119**: 429-442.

Annandale GW. 1990. Uncertainty in suspended sediment transport curves. Discussion. *Journal of Hydraulic Engineering* **116**: 140-141.

Asselman NEM. 2000. Fitting and interpretation of sediment rating curves. *Journal of Hydrology* **234**: 228-248.

Aytek A, Kisi, Ö. 2008. A genetic programming approach to suspended sediment modelling. *Journal of Hydrology* **351**: 288-298.

Bhattacharya B, Price RK, Solomatine DP, 2007. Machine Learning Approach to Modeling Sediment Transport. *Journal of Hydraulic Engineering* **133**: 440-450.

Bhattacharya B, Solomatine DP, 2006. Machine learning in sedimentation modelling, *Neural Networks* **19**: 208-214.

Babovic V. 2005. Data mining in hydrology. *Hydrological Processes* **19:** 1511-1515.

Beven K. 1989. Changing ideas in hydrology – the case of physically-based models. *Journal of Hydrology* **105**: 157-172.

Ciğizoğlu HK. 2002. Suspended Sediment Estimation for Rivers using Artificial Neural Networks and Sediment Rating Curves. *Turkish Journal of Engineering and Environmental Sciences* **26**: 27-36.

Ciğizoğlu HK, Kisi Ö. 2006. Methods to improve the neural network performance in suspended sediment estimation. *Journal of Hydrology* **317**: 221-238.

Cobaner M, Unal B, Kisi Ö. 2009. Suspended sediment concentration estimation by an adaptive neuro-fuzzy and neural network approaches using hydro-meteorological data. *Journal of Hydrology* **367**: 52-61.

Curry B, Morgan PH, 2003. Neural networks, linear functions and neglected non-linearity. *Computational Management Science* **1**: 15–29.

Dawson CW, Abrahart RJ, See LM. 2010. 'HydroTest: further development of a web resource for the standardised assessment of hydrological models'. *Environmental Modelling & Software* **25**:1481-1482.

Dawson CW, Abrahart RJ, Shamseldin AY, Wilby RL. 2006. Flood estimation at ungauged sites using artificial neural networks. *Journal of Hydrology* **319**: 391–409.

Dawson CW, Abrahart RJ, See LM. 2007. HydroTest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. *Environmental Modelling & Software* **22**: 1034-1052.

Dawson CW, Harpham, C, Wilby RL, Chen Y. 2002. An Evaluation of Artificial Neural Network Techniques for Flow Forecasting in the River Yangtze, China. *Hydrology and Earth System Sciences* **:**, 619-626.

Dubois D, Hajek P, Prade H. 2000. Knowledge-driven versus data-driven logics. *Journal of Logic, Language and Information* **9**: 65-89.

Ferguson RI. 1986. River Loads Underestimated by Rating Curves. *Water Resources Research* **22**: 74-76.

Ferreira C. 2001. Gene Expression Programming: A New Adaptive Algorithm for Solving Problems. *Complex Systems* **13**: 87-129.

Ferreira C. 2006. *Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence*. 2nd Ed. Berlin: Springer-Verlag.

Guven A, Kisi O. 2010. Estimation of suspended sediment yield in natural rivers using machine-coded linear genetic programming. *Water Resources Management* **25**: 691-704.

HydroSolutions Inc. 2008. *Tongue River Hydrology Report: Tongue River Information Program*.http://www.hydrosi.com/MBOGC04_TongueR_HydroRpt_Final_05162008.pdf

Janga Reddy M, Ghimire, BNS. 2009. Use of Model Tree and Gene Expression Programming to Predict the Suspended Sediment Load in Rivers. *Journal of Intelligent Systems* **1**: 211–228.

Kisi, Ö. 2004. Daily suspended sediment modelling using a fuzzy differential evolution approach. *Hydrological Sciences Journal* **49**: 183-197.

Kisi, Ö. 2005. Suspended sediment estimation using neuro-fuzzy and neural network approaches. *Hydrological Sciences Journal* **50:** 683-696.

Kisi, Ö. 2009. Evolutionary fuzzy models for river suspended sediment concentration estimation. *Journal of Hydrology* **372**: 68-79.

Legates D, and McCabe G. 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research* **35**: 233-241.

Lee H-Y, Lin Y-T, Chiu Y-J. 2006. Quantitative Estimation of Reservoir Sedimentation from Three Typhoon Events. *Journal of Hydrologic Engineering* **1:**, 362-370.

McBean EA, Al-Nassri S. 1988. Uncertainty in suspended sediment transport curves. *Journal of Hydraulic Engineering* **114**: 63-74.

McBean EA, Al-Nassri S. 1990a. Closure to "Uncertainty in Suspended Sediment Transport Curves" *Journal of Hydraulic Engineering* **116**: 150-151.

McBean EA, Al-Massri S. 1990b. Closure to "Uncertainty in Suspended Sediment Transport Curves" *Journal of Hydraulic Engineering* **116**: 732.

Milhous RT. 1990. Uncertainty in suspended sediment transport curves. Discussion. *Journal of Hydraulic Engineering* **11**: 730-732.

Minns AW, Hall MJ. 1996. Artifical neural networks as rainfall-runoff models. *Hydrological Sciences Journal* **41**: 399-417.

Mount NJ, Abrahart RJ. 2011a. Load or concentration, logged or unlogged? Addressing ten years of uncertainty in neural network suspended sediment prediction. *Hydrological Processes*, DOI:10.1002/hyp.8033.

Mount NJ, Abrahart RJ. 2011b. Discussion on 'River flow estimation from upstream flow records by artificial intelligence methods' by M.E. Turan and M.A. Yurdusev [J. Hydrol. 369 (2009) 71-77]. *Journal of Hydrology* **396**: 193-196.

Nordin CG. 1990. Uncertainty in suspended sediment transport curves. Discussion. *Journal of Hydraulic Engineering* **116**: 145-148.

Partal T, Ciğizoğlu HK. 2008. Estimation and forecasting of daily suspended sediment data using wavelet-neural networks. *Journal of Hydrology* **358**: 317-331.

Quinlan JR. 1992. Learning with continuous classes. In: *Proceedings of the Fifth Australian Joint Conference on Artificial Intelligence*, Adams N, Sterling L. (eds.) Hobart, Tasmania, Australia, 343-348. Singapore: World Scientific.

Raghuwanshi NS, Singh R, Reddy LS. 2006. Runoff and sediment yield modeling using artificial neural networks: Upper Siwane River, India. *Journal of Hydrologic Engineering* **11**: 71-79.

Singh P, Deo MC. 2007. Suitability of different neural networks in daily flow forecasting. *Applied Soft Computing* **7**: 968-978.

Solomatine DP, See LM, Abrahart RJ. 2008. Data-driven modelling: concepts, approaches and experiences. In: *Practical Hydroinformatics: Computational Intelligence and Technological Developments in Water Applications*, Abrahart RJ, See LM, Solomatine DP. (Eds.) Water Science and Technology Library Volume **68**, Springer Berlin Heidelberg, 17-30.

Stott TA, Mount NJ. 2007. Alpine proglacial suspended sediment dynamics in warm and cool ablation seasons: implications for global warming? Journal of Hydrology, **332**(3-4): 259-270.

Wahl KL. 1990. Uncertainty in suspended sediment transport curves. Discussion. *Journal of Hydraulic Engineering* **116**: 148-150.

Walling D. 1977. Assessing the accuracy of suspended sediment rating curves for a small basin. *Water Resources Research* **13**: 531-538.

Walling DE, Webb BW. 1988. The reliability of rating curve estimates of suspended sediment yield: some further comments. In: *Sediment Budgets (Proceedings of the Porto Alegre Symposium, December 1988)*, International Association of Hydrological Sciences Publication no. 174. pp. 337-350.

Wilby RL, Davies G. 1997. Operational hydrology. In: *Contemporary Hydrology,* Wilby RL. (Ed.), 195–240. John Wiley & Sons Ltd, Chichester, West Sussex, UK.

Witten, I.H., Frank, E. 2005. *Data Mining: Practical MachineLearning Tools and Techniques*. 2nd Ed. San Francisco, CA, USA: Morgan Kaufmann Publishers.

Zhu YM, Lu XX, Zhou Y. 2007. Suspended sediment flux modelling with artificial neural network: an example of the Longchuanjiang River in Upper Yangtze Catchment, China. *Geomorphology* **84:** 111-125.

Table 1. Statistical descriptors for UPST and DNST training (SET A) and testing (SET B) datasets.

	UPST		DNST	
Set A	*Q (m^3s^{-1})*	*S (ton day^{-1})*	*Q (m^3s^{-1})*	*S (ton day^{-1})*
Mean	15.07	443.90	15.65	996.35
Median	10.00	52.00	7.67	47.00
Std. Dev.	20.58	1725.76	24.87	4859.58
Variance	423.61	2978241.03	618.48	23615510.07
Kurtosis	23.22	95.10	19.38	166.58
Skewness	4.40	8.48	4.16	11.50
Range	213.16	27198.50	216.10	84396.60
Minimum	1.84	1.50	1.90	3.40
Maximum	215.00	27200.00	218.00	84400.00
Count	1096	1096	1096	1096

	UPST		DNST	
Set B	*Q (m^3s^{-1})*	*S (ton day^{-1})*	*Q (m^3s^{-1})*	*S (ton day^{-1})*
Mean	10.28	234.88	9.00	369.01
Median	6.65	26.00	5.47	23.00
Std. Dev.	12.54	834.75	12.88	1235.10
Variance	157.19	696803.56	166.00	1525483.09
Kurtosis	9.12	34.61	9.19	15.53
Skewness	3.09	5.43	3.12	4.04
Range	60.92	7997.00	62.24	7399.87
Minimum	1.98	3.00	0.06	0.13
Maximum	62.90	8000.00	62.30	7400.00
Count	365	365	365	365

Table 2. Product moment correlation matrix for UPST and DNST datasets.

	UPST	*Q*	*Qt-1*	*Qt-2*	*S*	*St-1*	*St-2*
Set A	*Q*	1.00					
	Qt-1	0.98	1.00				
	Qt-2	0.95	0.98	1.00			
	S	0.80	0.74	0.68	1.00		
	St-1	0.82	0.80	0.74	**0.87**	1.00	
	St-2	0.82	0.82	0.80	0.74	0.87	1.00
Set B	*Q*	1.00					
	Qt-1	0.99	1.00				
	Qt-2	0.97	0.99	1.00			
	S	0.85	0.80	0.74	1.00		
	St-1	0.86	0.85	0.80	**0.91**	1.00	
	St-2	0.87	0.86	0.85	0.84	0.91	1.00
	DNST	*Q*	*Qt-1*	*Qt-2*	*S*	*St-1*	*St-2*
Set A	Q	1.00					
	Qt-1	0.97	1.00				
	Qt-2	0.92	0.97	1.00			
	S	0.74	0.63	0.52	1.00		
	St-1	0.74	0.74	0.63	**0.80**	1.00	
	St-2	0.66	0.74	0.74	0.53	0.80	1.00
Set B	Q	1.00					
	Qt-1	0.99	1.00				
	Qt-2	0.97	0.99	1.00			
	S	0.91	0.87	0.82	1.00		
	St-1	0.93	0.91	0.87	**0.97**	1.00	
	St-2	0.93	0.93	0.91	0.92	0.97	1.00

*Strongest relationship between predictor and predictand in bold

Table 3. Reasons for the inclusion / rejection of input combinations in Tongue River re-analysis.

Input combinations	***Included in original analysis?***	***Included in re-analyses?***	***Reason for inclusion / rejection in re-analyses***
Q_t	Yes	Yes	Equates to standard, concurrent input-output modelling used in standard rating curves. Operationally applicable and provides useful benchmark.
S_{t-1}	Yes	No	Operationally-inapplicable: utilises lagged suspended sediment
Q_t and S_{t-1}	Yes	No	Operationally-inapplicable: utilises lagged suspended sediment
S_{t-1} and S_{t-2}	Yes	No	Operationally-inapplicable: utilises lagged suspended sediment
Q_t, S_{t-1} and S_{t-2}	Yes	No	Operationally-inapplicable: utilises lagged suspended sediment
Q_t, Q_{t-1}, and S_{t-1}	Yes	No	Operationally-inapplicable: utilises lagged suspended sediment
Q_t, Q_{t-1}, S_{t-1} and S_{t-2}	Yes	No	Operationally-inapplicable: utilises lagged suspended sediment
Q_t and Q_{t-1}	No	Yes	Operationally applicable and offers potential for modelling hysteresis which is not captured in Q_t alone.

Table 4. GeneXproTools 4.0 used in the Tongue River re-analysis

Function Set	***Symbol***	***Weight***	***Arity***
Addition	+	4	2
Subtraction	-	4	2
Multiplication	*	4	2
Division	/	1	2
Power	Pow	1	2
Square Root	Sqrt	1	1
Exponential	Exp	1	1
10^x	Pow10	1	1
Natural Logarithm	Ln	1	1
Logarithm of base 10	Log	1	1
General			
Chromosomes		30	
Genes		3	
Head Size		7	
Tail Size		8	
Dc Size		8	
Gene Size		23	
Linking function		Addition	
Fitness function			
Error Type		Absolute with SR	
Precision		0.01	
Selection Range		100	
Genetic Operators			
Mutation Rate		0.044	
Inversion Rate		0.1	
IS Transposition Rate		0.1	
RIS Transposition Rate		0.1	
One-point Recombination Rate		0.3	
Two-point Recombination Rate		0.3	
Gene Recombination Rate		0.1	
Gene Transposition Rate		0.1	
Numerical Constants			
Constants per Gene		2	
Data Type		Floating-Point	
Lower Bound		-10	
Upper Bound		10	
RNC Mutation		0.01	
Dc Mutation		0.044	
Dc Inversion		0.1	
Dc IS Transposition		0.1	

Table 5: Ordinary least squares regression solution parameters for UPST and DNST.

	UPST		DNST	
Coefficient	OLS1	OLS2	OLS1	OLS2
Intercept	-572.575	-542.403	-1275.807	-1130.763
Qt	67.444	179.244	145.171	440.156
Qt-1	-	-113.814	-	-304.242

Table 6. M5MT linear piecewise solution parameters for UPST.

M5 pruned model tree for M5MT1

```
Qt0 <= 15.25 :
|   Qt0 <= 9.7 : LM1 (529)
|   Qt0 >  9.7 : LM2 (339)
Qt0 >  15.25 : LM3 (228)

LM num: 1
Lt0 =
      1.5168 * Qt0
      + 11.2219

LM num: 2
Lt0 =
      19.5906 * Qt0
      - 123.7161

LM num: 3
Lt0 =
      74.2324 * Qt0
      - 1037.8266
```

M5 pruned model tree for M5MT2

```
Qt0 <= 15.25 :
|   Qt0 <= 9.7 : LM1 (529)
|   Qt0 >  9.7 : LM2 (339)
Qt0 >  15.25 :
|   Qt0 <= 50.55 : LM3 (182)
|   Qt0 >  50.55 :
|   |   Qt0 <= 110.5 : LM4 (36)
|   |   Qt0 >  110.5 : LM5 (10)

LM num: 1
Lt0 =
      0.3711 * Qt0
      + 1.0575 * Qt-1
      + 12.5534

LM num: 2
Lt0 =
      18.4449 * Qt0
      + 1.0575 * Qt-1
      - 122.3847

LM num: 3
Lt0 =
      79.2117 * Qt0
      - 34.6599 * Qt-1
      - 280.8026

LM num: 4
Lt0 =
      211.9839 * Qt0
      - 179.3727 * Qt-1
      + 2398.7186

LM num: 5
Lt0 =
      119.1819 * Qt0
      - 30.6555 * Qt-1
      - 2612.4667
```

Table 7. M5MT linear piecewise solution parameters for UPST.

M5 pruned model tree for M5MT1	M5 pruned model tree for M5MT2
Qt0 <= 10.95 : LM1 (723) Qt0 > 10.95 : | Qt0 <= 38.4 : LM2 (298) | Qt0 > 38.4 : LM3 (75) LM num: 1 Lt0 = 2.9506 * Qt0 + 25.1496 LM num: 2 Lt0 = 81.8322 * Qt0 - 767.3094 LM num: 3 Lt0 = 221.7397 * Qt0 - 9500.7118	Qt0 <= 10.95 : LM1 (723) Qt0 > 10.95 : | Qt0 <= 38.4 : LM2 (298) | Qt0 > 38.4 : LM3 (75) LM num: 1 Lt0 = 2.4904 * Qt-1 + 32.3512 LM num: 2 Lt0 = 219.5664 * Qt0 - 136.6173 * Qt-1 - 763.5686 LM num: 3 Lt0 = 390.512 * Qt0 - 204.2036 * Qt-1 - 6669.4627

Table 8. Evaluation metrics for GEP solutions. * indicates a value that has been taken from Aytek and Kisi's original paper and † indicates a value computed using the explicit equations defined by their and our gene expression trees. The value of the best performing solution for each metric in each training and testing set is highlighted in bold.

UPST		A + K (Qt, Qt-1, St-1)	GEP1	GEP2
A	RMSE:	**662.05†**	1018.40†	1010.05†
	RSqr:	**0.85†**	0.67†	0.68†
B	RMSE:	**231.00***	496.56†	440.73†
	RSqr:	**0.94***	0.78†	0.82†
DNST		A + K (Qt, Qt-1, St-1)	GEP1	GEP2
A	RMSE:	**3250.15†**	3415.13†	3930.32†
	RSqr:	**0.93†**	0.68†	0.54†
B	RMSE:	**331.00***	918.70†	1013.07†
	RSqr:	**0.93***	0.88†	0.84†

Table 9. Comparison of GEP1, GEP2 (shaded columns) and SRC metrics for training and testing data sets in both upstream and downstream gauges. The value of the best performing solution for each metric in each training and testing set is highlighted in bold.

UPST		GEP1	GEP2	SRC
A	RMSE:	1018.40	**1010.05**	1620.44
	RSqr:	0.67	**0.68**	0.66
B	RMSE:	496.56	440.73	**395.90**
	RSqr:	0.78	**0.82**	0.78
DNST		GEP1	GEP2	SRC
A	RMSE:	3415.13	3930.32	**2597.97**
	RSqr:	**0.68**	0.54	**0.68**
B	RMSE:	918.70	1013.07	**448.74**
	RSqr:	**0.88**	0.84	**0.88**

Table 10. Comparison of GEP1, GEP2 (shaded columns), OLS1, OLS2, M5MT1 and M5MT2 metrics for training and testing data sets in both upstream and downstream gauges. The value of the best performing solution for each metric in each training and testing set is highlighted in bold.

UPST		GEP1	GEP2	OLS1	OLS2	M5MT1	M5MT2
A	RMSE:	1018.40	1010.05	1024.91	926.22	1000.26	**815.04**
	RSqr:	0.67	0.68	0.65	0.71	0.66	**0.78**
B	RMSE:	496.56	440.73	478.69	391.70	402.88	**359.22**
	RSqr:	0.78	0.82	0.72	0.81	0.77	**0.89**
DNST		GEP1	GEP1	OLS1	OLS2	M5MT1	M5MT2
A	RMSE:	3415.13	3930.32	3251.39	2672.99	3025.40	**2571.12**
	RSqr:	0.68	0.54	0.55	0.70	0.61	**0.72**
B	RMSE:	918.70	1013.07	971.85	876.71	591.41	**410.70**
	RSqr:	0.88	0.84	0.82	0.87	0.83	**0.90**

Table11. Comparison of GEP1 (shaded column) and NN1(*n*) metrics for training and testing data sets in both upstream and downstream gauges. The value of the best performing solution for each metric in each training and testing set is highlighted in bold.

UPST		GEP1	NN1(1)	NN1(2)	NN1(3)	NN1(4)	NN1(5)
A	RMSE:	1018.40	982.39	997.62	**971.81**	978.98	973.24
	RSqr:	0.67	**0.68**	0.67	**0.68**	**0.68**	**0.68**
B	RMSE:	496.56	480.21	**440.96**	449.49	481.92	451.59
	RSqr:	**0.78**	0.76	0.74	0.75	0.74	0.74
DNST		GEP1	NN1(1)	NN1(2)	NN1(3)	NN1(4)	NN1(5)
A	RMSE:	3415.13	2612.97	2832.49	**2212.43**	2219.07	2217.85
	RSqr:	0.68	0.71	0.67	**0.80**	0.79	0.79
B	RMSE:	918.70	719.12	**486.07**	520.98	493.46	595.51
	RSqr:	**0.88**	0.87	0.86	0.83	0.84	0.80

Table 12. Comparison of GEP2 (shaded column) and NN2(*n*) metrics for training and testing data sets in both upstream and downstream gauges. The value of the best performing solution for each metric in each training and testing set is highlighted in bold.

UPST		GEP2	NN2(1)	NN2(2)	NN2(3)	NN2(4)	NN2(5)
A	RMSE:	1010.05	855.41	867.17	634.82	608.29	**605.94**
	RSqr:	0.68	0.76	0.75	0.87	**0.88**	**0.88**
B	RMSE:	440.73	345.19	336.24	343.92	**295.38**	304.33
	RSqr:	0.82	**0.88**	0.87	0.84	**0.88**	0.87
DNST		GEP2	NN2(1)	NN2(2)	NN2(3)	NN2(4)	NN2(5)
A	RMSE:	3930.32	1920.13	2008.12	1879.27	**1515.80**	2033.11
	RSqr:	0.54	0.85	0.83	0.86	**0.90**	0.84
B	RMSE:	1013.07	528.95	**427.99**	468.61	583.44	453.56
	RSqr:	0.84	**0.90**	0.89	0.89	0.84	0.87

Table 13. Comparison of the best performing, operationally-valid models with Aytek and Kisi's (2008) original GEP solution.

UPST		BEST PERFORMING M5MT SOLUTION		BEST PERFORMING NN SOLUTION		AYTEK AND KISI (2008) GEP SOLUTION
A	RMSE:	M5MT2	815.04	NN2(5)	605.94	662.05
	RSqr:	M5MT2	0.78	NN2(5)	0.88	0.85
B	RMSE:	M5MT2	359.22	NN2(4)	295.38	231.00
	RSqr:	M5MT2	0.89	M5MT2	0.89	0.94
DNST						
A	RMSE:	M5MT2	2571.12	NN2(4)	1515.8	3250.15
	RSqr:	M5MT2	0.72	NN2(4)	0.90	0.93
B	RMSE:	M5MT2	410.70	NN2(2)	427.99	331.00
	RSqr:	M5MT2	0.90	NN2(1)	0.90	0.93

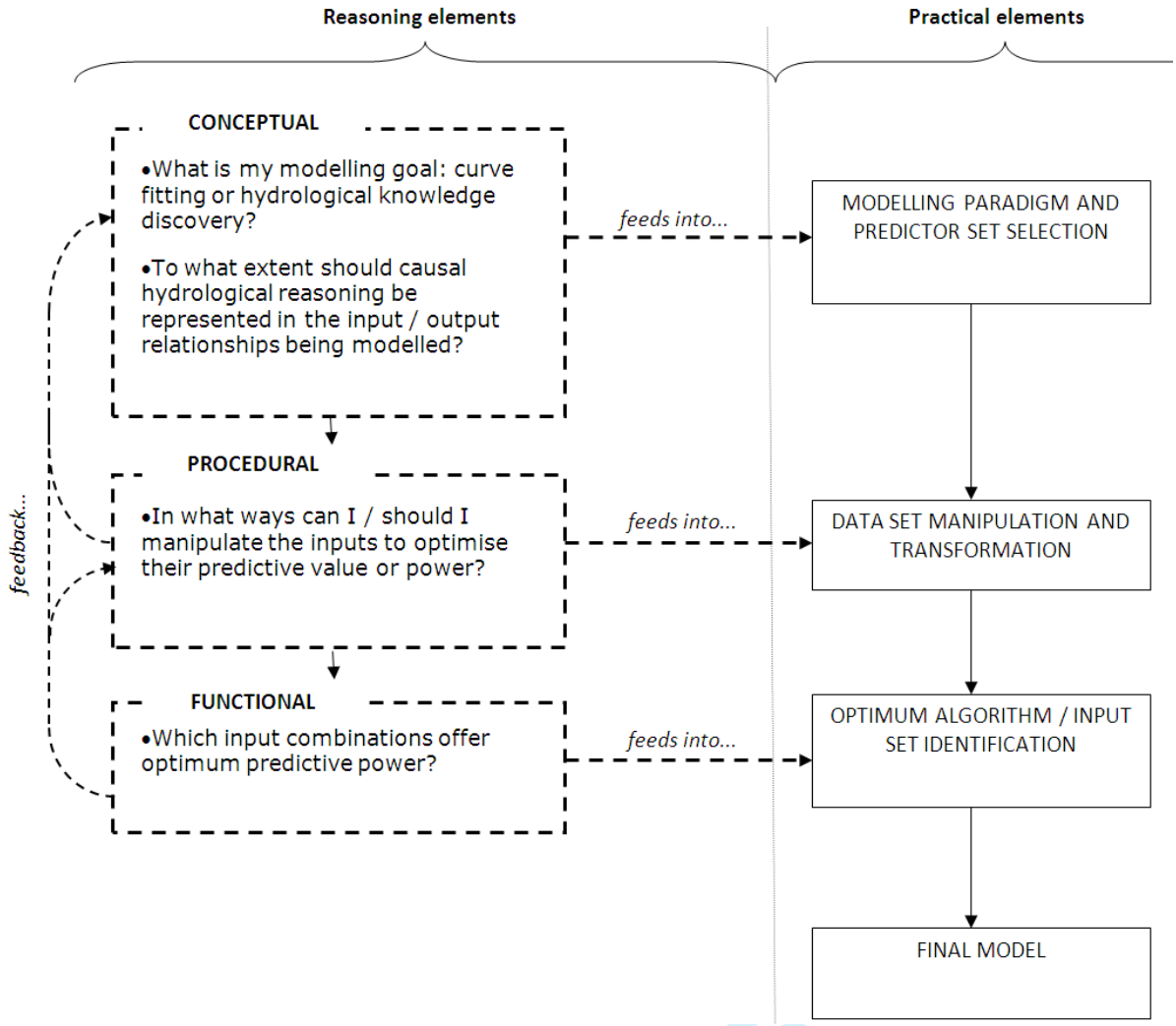


Figure 1. The reasoning and practical modelling processes most commonly utilised in data-driven suspended sediment studies.

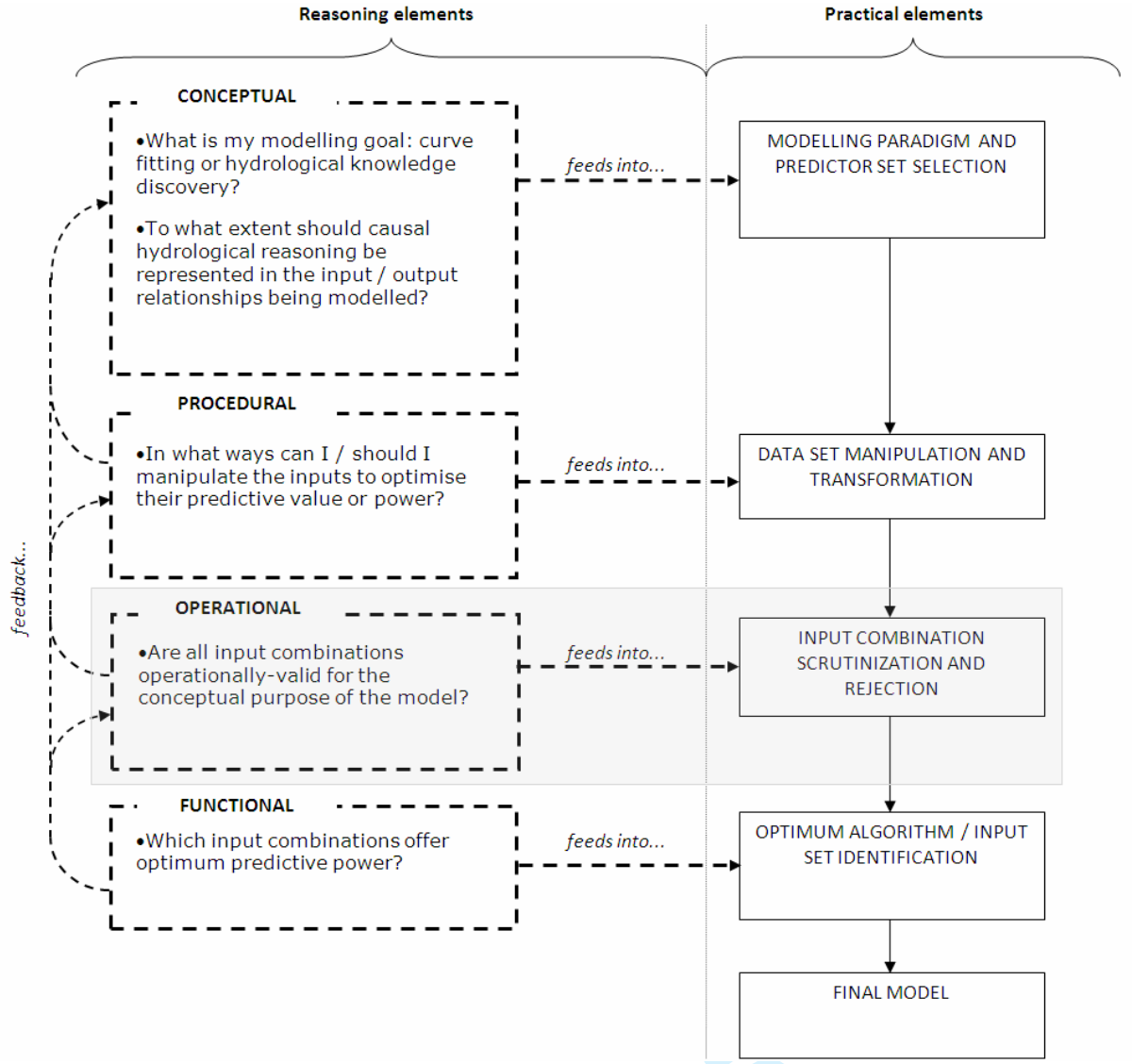


Figure 2. The reasoning and practical modelling processes most commonly utilised utilised in data-driven suspended sediment studies, augmented to incorporate operational reasoning and input combination scrutinisation.

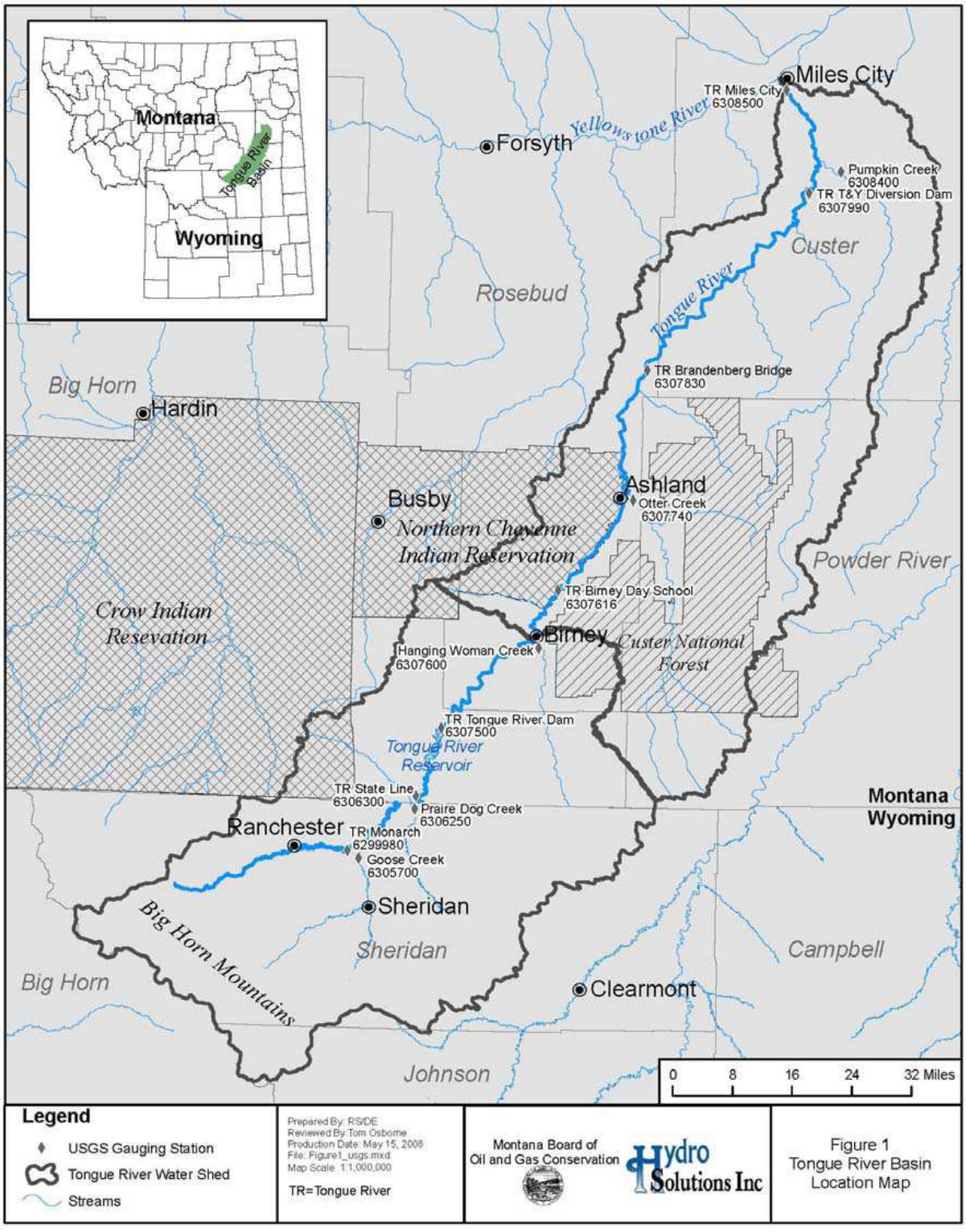


Figure 3. Tongue River Basin (courtesy of HydroSolutions Inc., 2008).

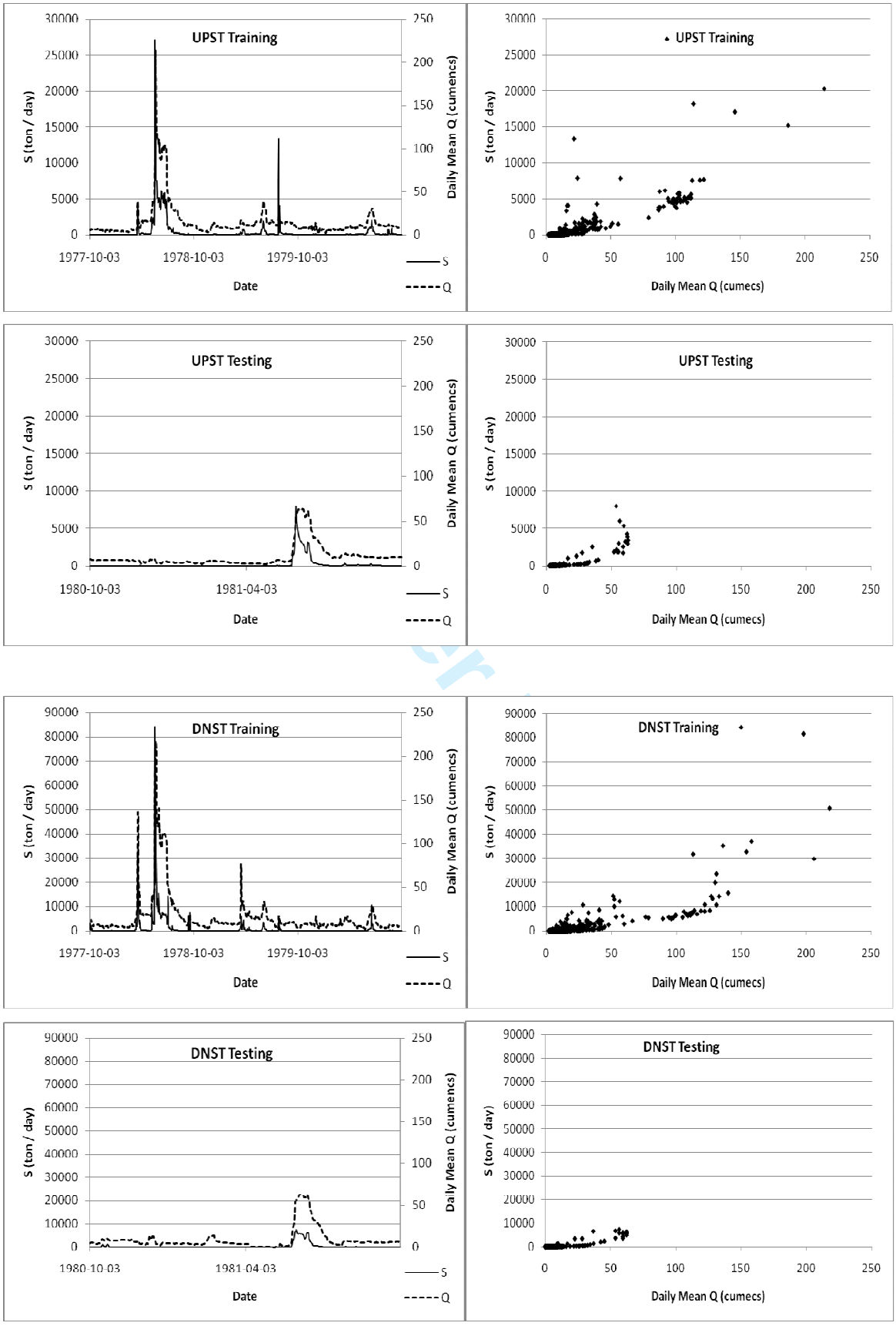
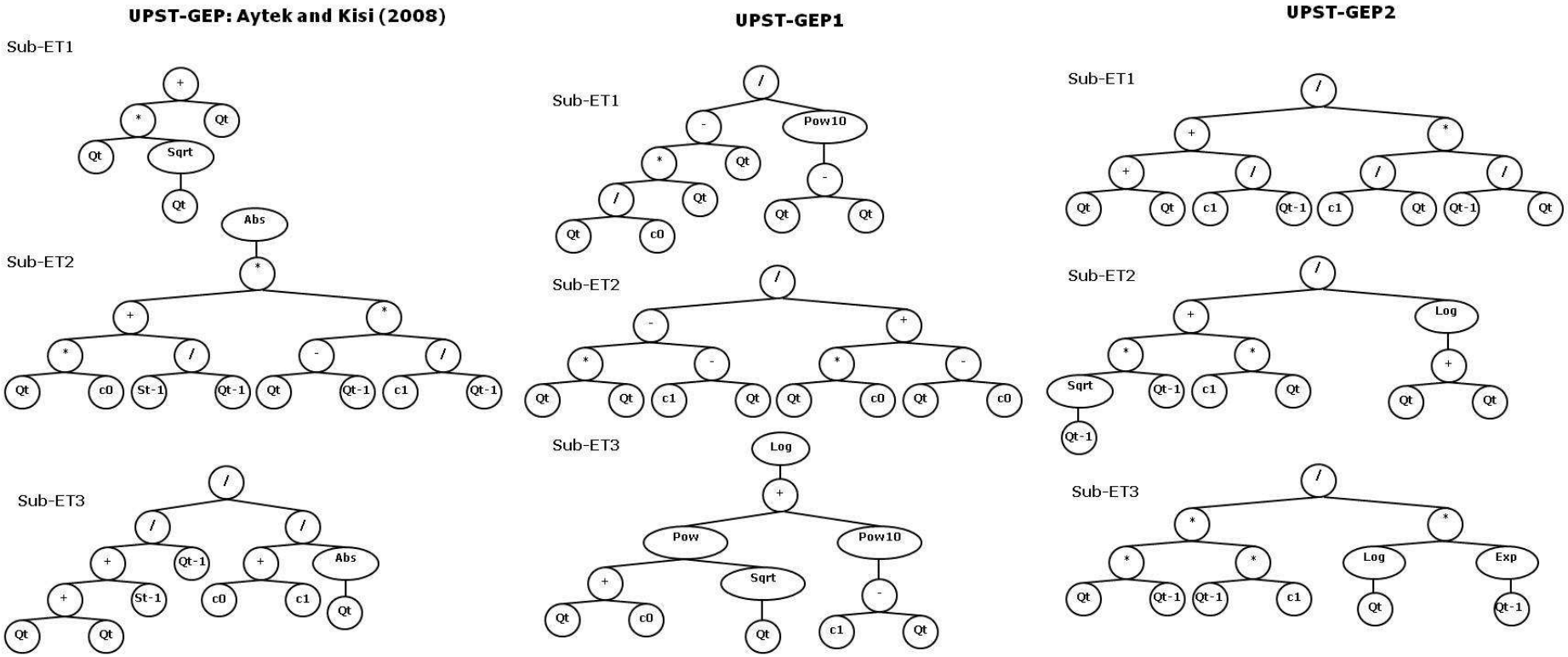


Figure 4. UPST and DNST data sets for training (SET A) and testing (SET B) periods.



	UPST GEP: Aytek and Kisi (2008)	UPST-GEP1	UPST-GEP2
SubET1 c0	-6.058	1.665	2.497
SubET1 c1	-9.641	-4.497	3.997
SubET2 c0	5.135	-6.057	-8.433
SubET2 c1	-9.641	9.583	-3.168
SubET3 c0	1.636	-1.564	9.715
SubET3 c1	0.021	8.792	2.637

Figure 5. Upstream gene expression trees and associated constant values for Aytek and Kisi's (2008) solution, GEP1 and GEP2.

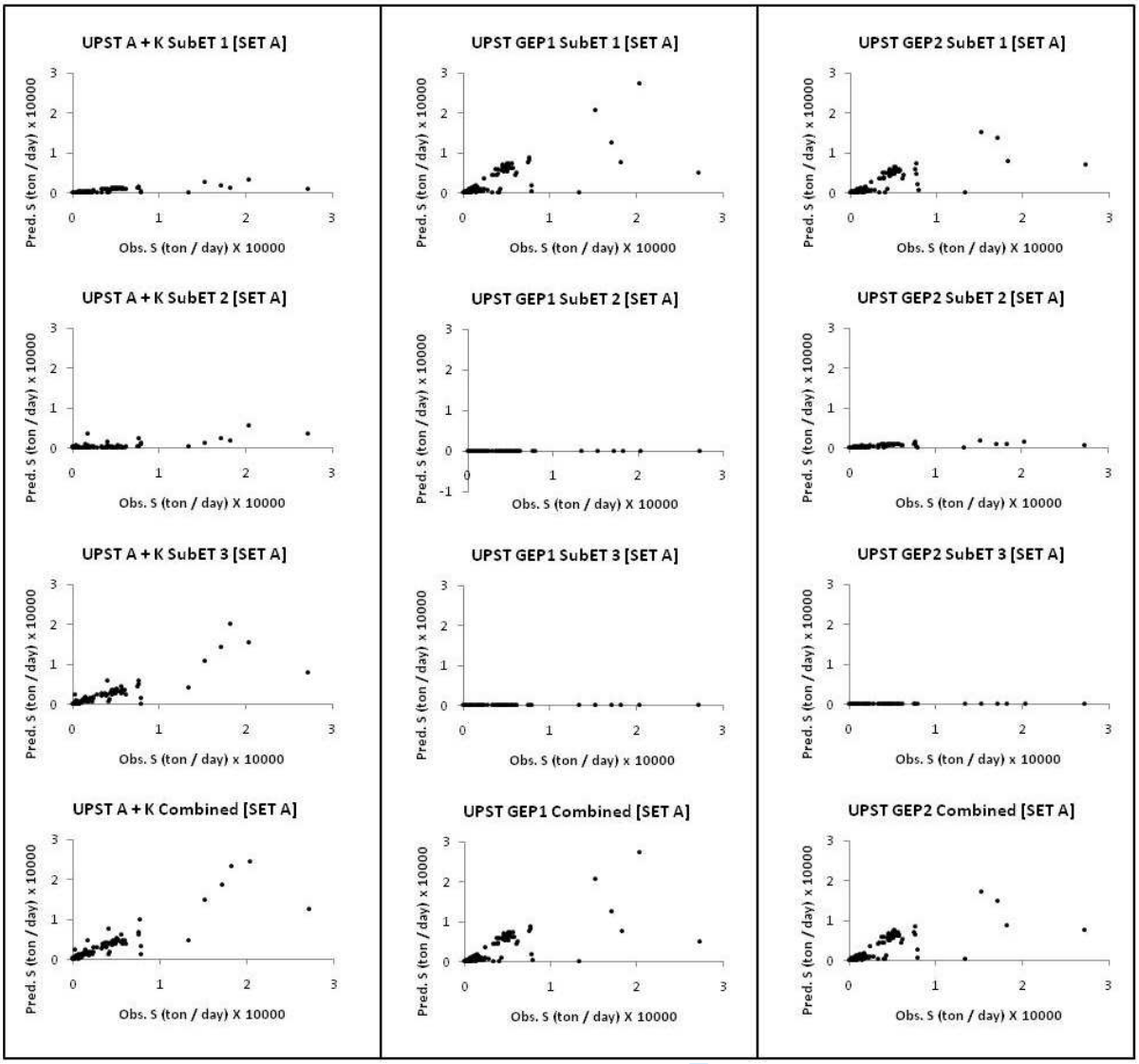


Figure 6. Observed versus predicted plots of GEP sub expressions and the combined function for UPST training data set (SET A): Aytek and Kisi's (2008) solution, GEP1 and GEP2.

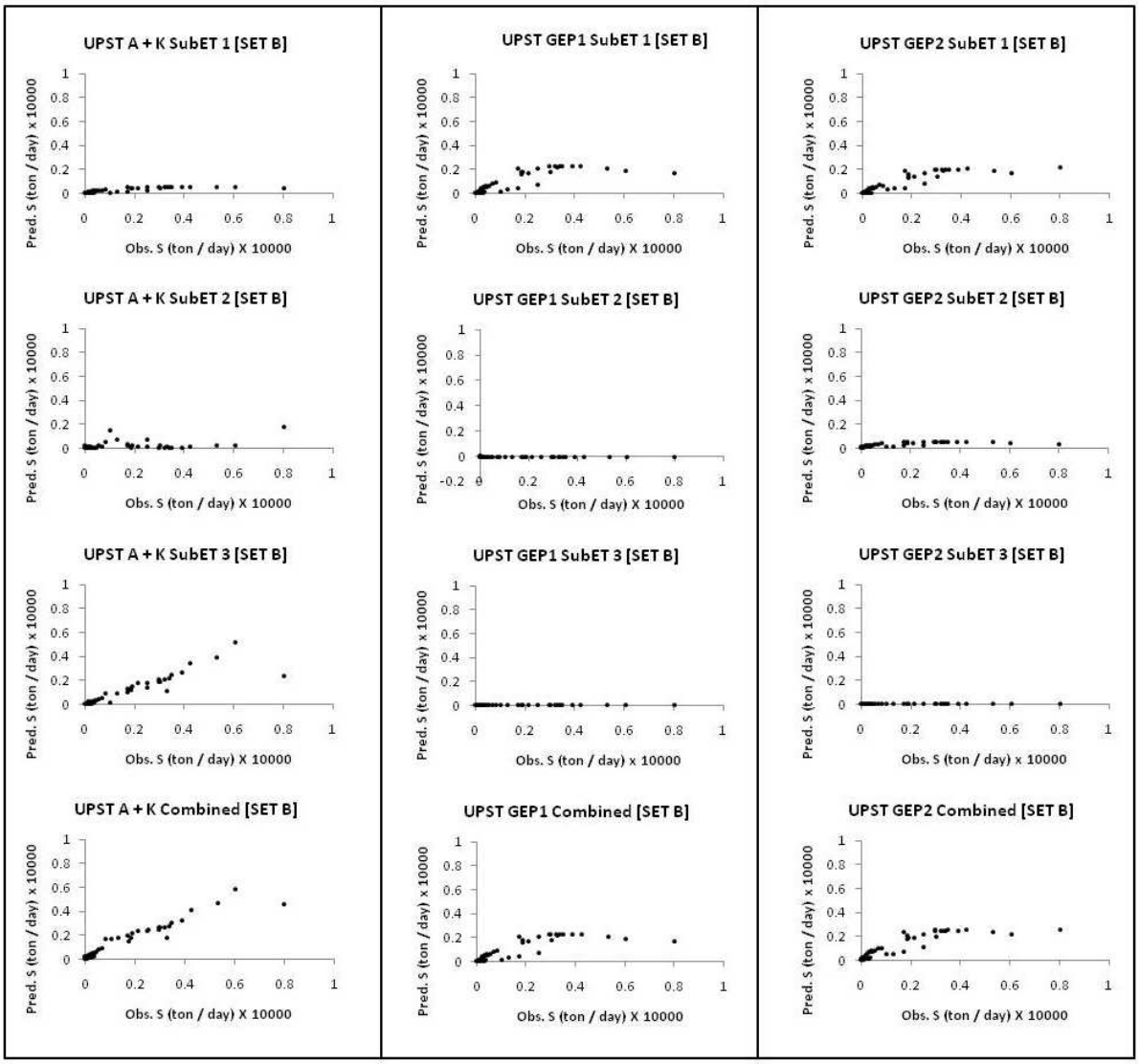
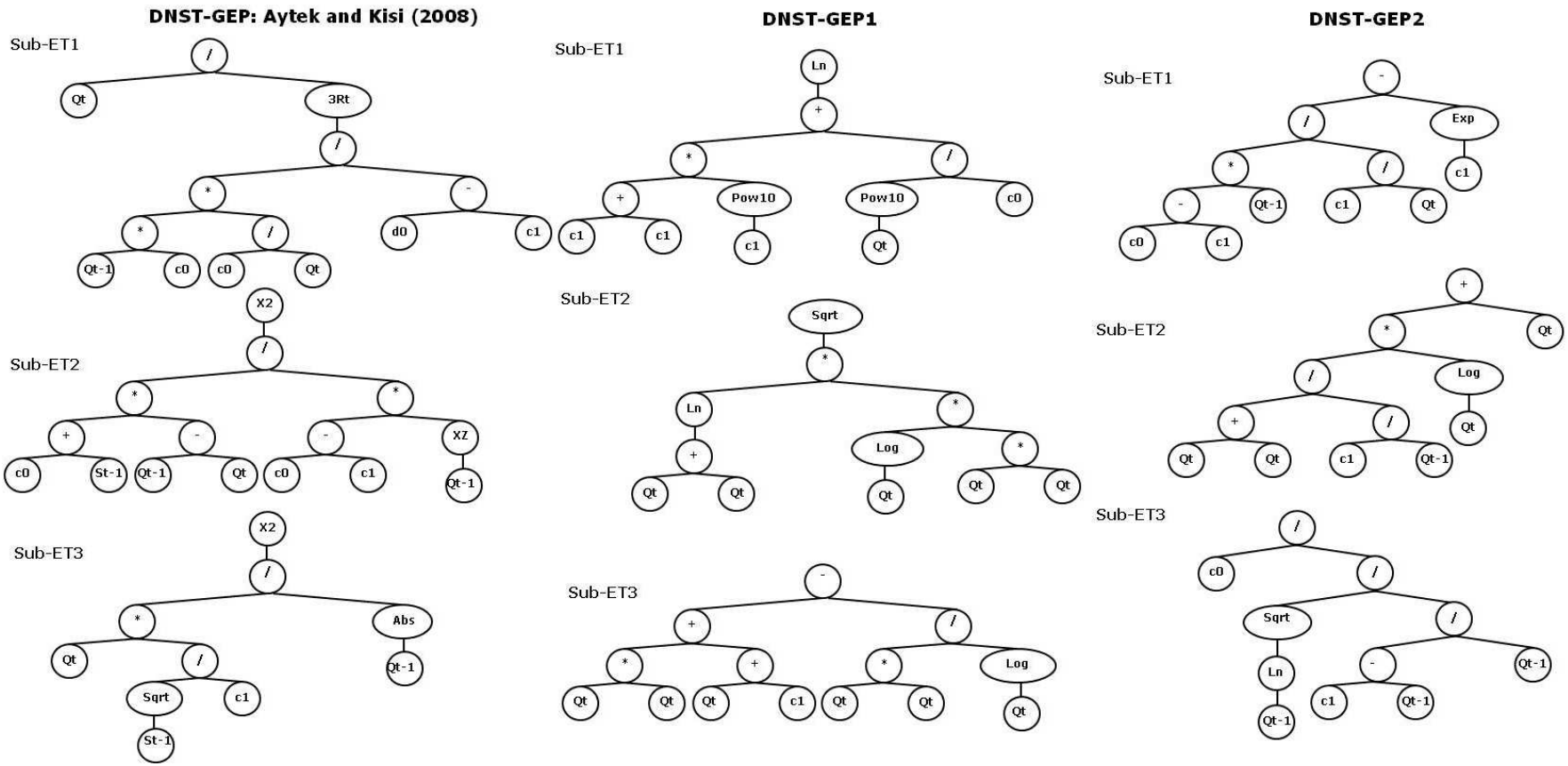


Figure 7. Observed versus predicted plots of GEP sub expressions and the combined function for UPST testing data set (SET B): Aytek and Kisi's (2008) solution, GEP1 and GEP2.



	DNST GEP: Aytek and Kisi (2008)	DNST-GEP1	DNST-GEP2
SubET1 c0	0.055	2.851	-4.845
SubET1 c1	-7.995	4.947	-3.730
SubET2 c0	-1.415	4.600	6.344
SubET2 c1	-4.465	-0.291	9.798
SubET3 c0	-2.526	8.425	7.395
SubET3 c1	1.379	3.437	4.836

Figure 8. Downstream gene expression trees and associated constant values for Aytek and Kisi's (2008) solution, GEP1 and GEP2.

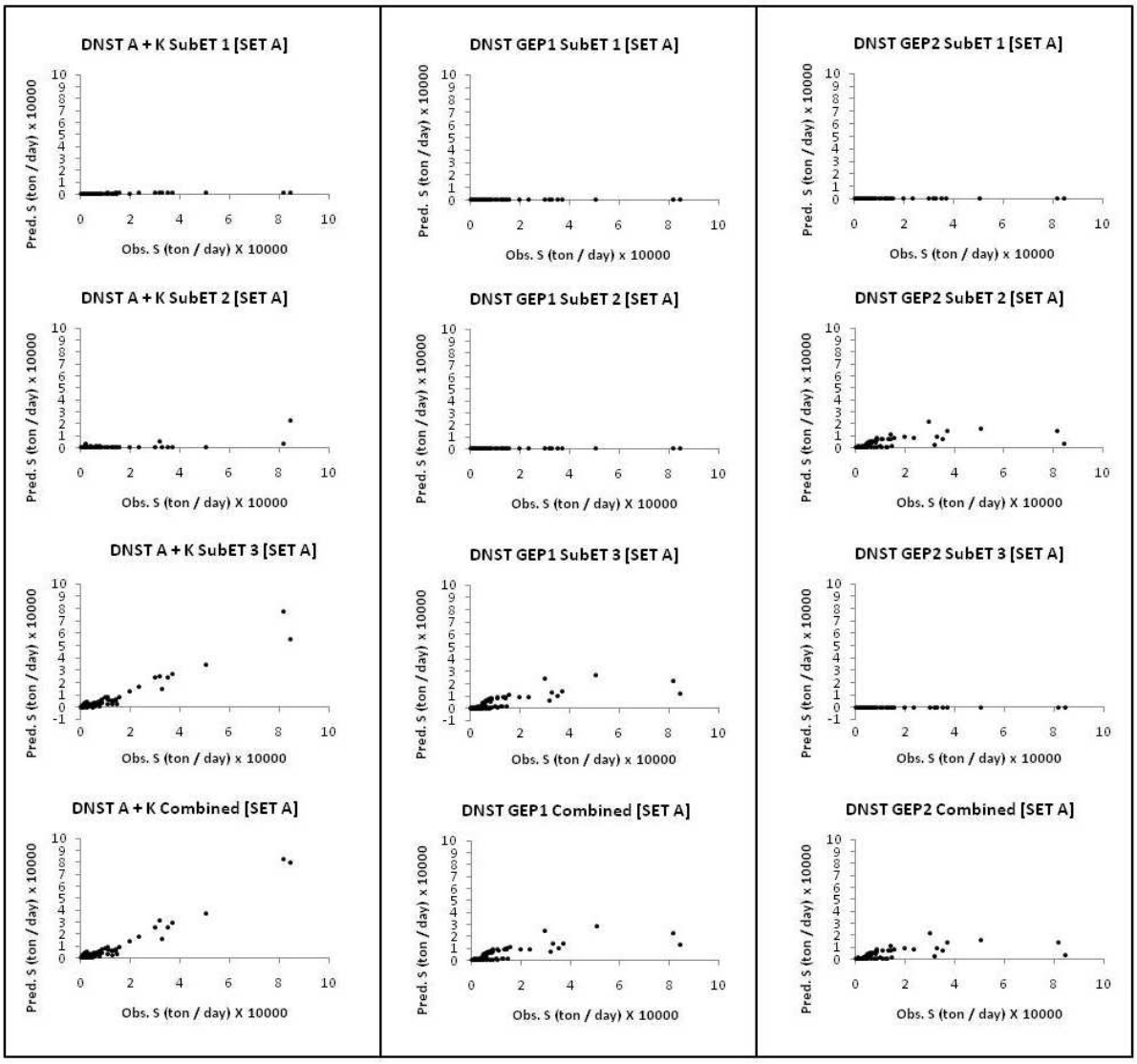


Figure 9. Observed versus predicted plots of GEP sub expressions and the combined function for DNST training data set (SET A): Aytek and Kisi's (2008) solution, GEP1 and GEP2.

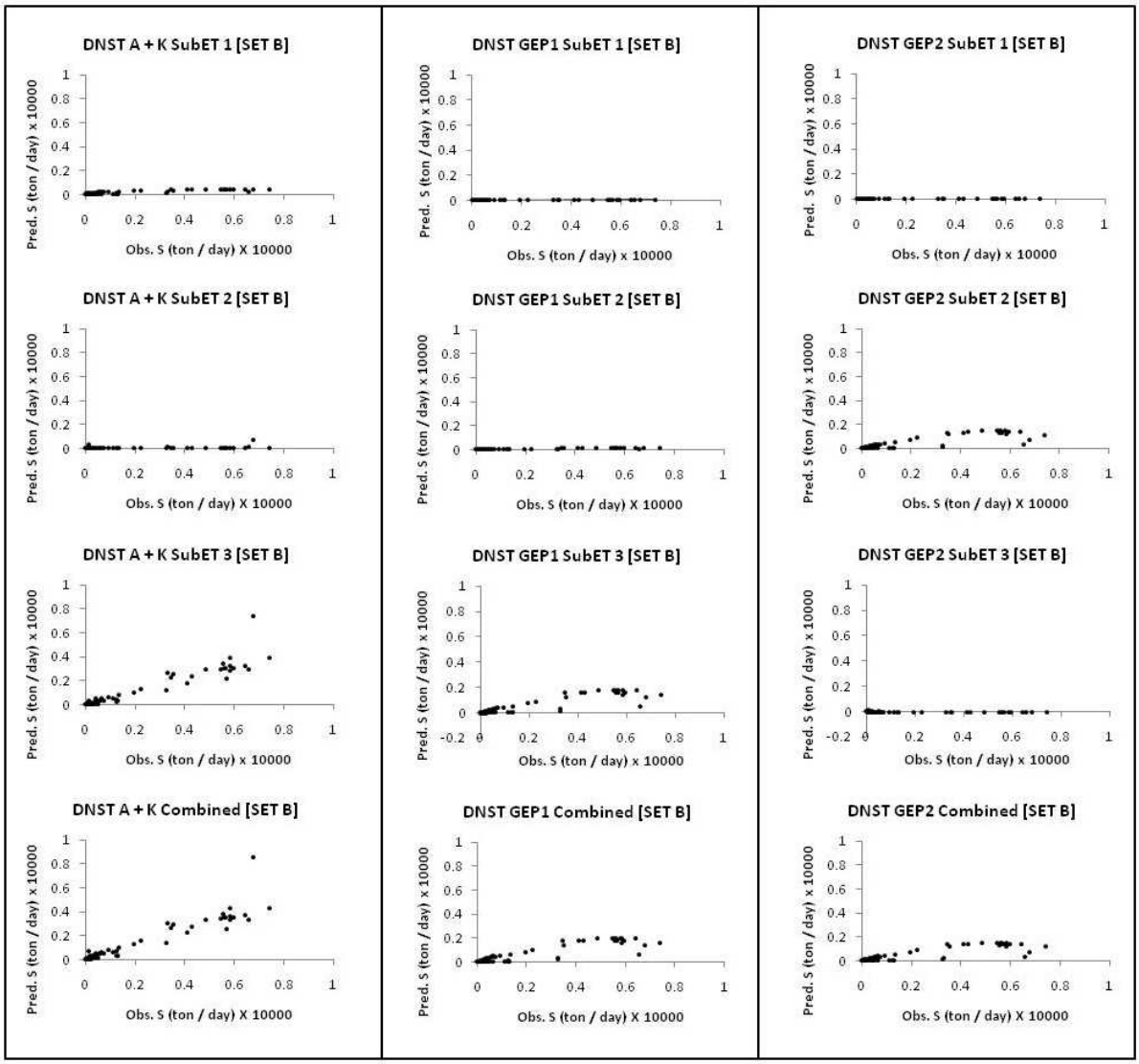


Figure 10. Observed versus predicted plots of GEP sub expressions and the combined function for DNST testing data set (SET B): Aytek and Kisi's (2008) solution, GEP1 and GEP2.