Environmental Modelling & Software 72 (2015) 33–43

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Toward creating simpler hydrological models: A LASSO subset selection approach

W.E. Bardsley [a, *], V. Vetrova [a], S. Liu [b]

[a] Faculty of Science and Engineering, University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand
[b] Key Lab of Water Cycle & Related Land Surface Processes, Institute of Geographic Sciences & Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

ARTICLE INFO

Article history:
Received 21 May 2015
Received in revised form
21 June 2015
Accepted 24 June 2015
Available online 12 July 2015

Keywords:
Model simplification
Linear basis functions
Linear LASSO
Subset selection
Finite mixture distributions
Linear programming
Runoff model
Groundwater model

ABSTRACT

A formalised means of simplifying hydrological models concurrent with calibration is proposed for use when nonlinear models can be initially formulated as over-parameterised constrained absolute deviation regressions of nonlinear expressions. This provides a flexible modelling framework for approximation of nonlinear situations, while allowing the models to be amenable to algorithmic simplification. The degree of simplification is controlled by a user-specified forcing parameter λ. That is, an original over-parameterised linear model is reduced to a simpler working model which is no more complex than required for a given application. The degree of simplification is a compromise between two factors. With weak simplification most parameters will remain, risking calibration overfitting. On the other hand, a high degree of simplification generates inflexible models. The linear LASSO (Least Absolute Shrinkage and Selection Operator) is utilised for the simplification process because of its ability to deal with linear constraints in the over-parameterised initial model.

1. Introduction

There tends to be a preference in hydrological modelling toward larger multi-purpose models in the various subject areas, as noted by Fenicia et al. (2011) in the context of rainfall-runoff modelling. While using a "big model" approach has understandable attraction, there have been concerns over whether models may be overly complex in practical applications. See, for example, Perrin et al. (2001) and Jakeman and Hornberger (1993). Beven (2006) identified finding a means of reduction of model dimensionality as one of a number of important research topics in runoff modelling.

There have been many proposed qualitative and quantitative approaches to model simplification in the hydrological literature. A review is beyond the scope of this paper but selected works include Dooge (1997), Schoups et al. (2008), Sivapalan et al. (2003), Sivakumar (2008), Fenicia et al. (2008), Hill (2006), Tonkin and Doherty (2005), Hunt et al. (2007), Arkesteijn and Pande (2013), and Diodato et al. (2014).

One generic approach toward simpler models is data-based mechanistic modelling (DBM) were options are restricted to simpler but physically plausible models consistent with data (Young, 2003, 2006; Young and Garnier, 2006; Young et al., 1996). Some combinations of DBM with other approaches are presented by Young and Ratto (2009, 2011) and Young (2013).

In the spirit seeking model simplicity, the present paper outlines the potential for algorithmic hydrological model simplification through use of the LASSO (Least Absolute Shrinkage and Selection Operator). The requirement here is that both the model and model fitting function are first set up as linear expressions. That is, the initial model as applied to a calibration data set is expressed as a sequence of linear equality and inequality expressions. The initial model is then formally reduced to a simpler model no more complex than required for application to a given data set.

There is of course an apparent contradiction in carrying out linear modelling of nonlinear hydrological processes. However, we make a case that constrained linear modelling of the type considered here can be formulated to be as "nonlinear" as necessary, through the use of linear combinations of nonlinear basis functions. A basis function can be defined as an element of a particular basis for a function space. For example, a quadratic polynomial comprises

* Corresponding author.
E-mail addresses: web@waikato.ac.nz (W.E. Bardsley), liusx@igsnrr.ac.cn (S. Liu).

http://dx.doi.org/10.1016/j.envsoft.2015.06.008
1364-8152/

the basis functions 1, x, and x^2 and the expression $a1 + bx + cx^2$ is a linear combination of basis functions.

The method will be illustrated with respect to an example rainfall-runoff model but the intention is for general application. However, this simplification approach is very much in its infancy and we use "evaluation" throughout the paper in preference to "validation" to avoid any implications of confirmation at this time.

With respect to the paper organisation, Section 2 gives an illustration of how a general nonlinear time series model can be reformulated in a linear way for later simplification. Section 3 gives a brief introduction to the LASSO method. Section 4 defines the linear LASSO simplification algorithm utilised here. Section 5 is an illustrative simplification of an over-paramerised basic rainfall-runoff model. Section 6 is a brief comment on possible application to groundwater modelling. Section 7 discusses issues of modelling philosophy which arise when models are derived from an automated simplification process. Possible further developments are considered in Section 8. Concluding comments are given in Section 9.

2. Linear models as nonlinear approximations

A requirement before the simplification process is the creation of an initial linear model for the nonlinear situation under study. That is, the entire model must be specified as a sequence of linear constraints, with fitting to data being a linear programming (LP) minimisation of absolute deviations.

Achieving accurate linear approximation to a nonlinear reality is not necessarily a trivial task and a full review of all mechanisms by which it might be achieved is beyond the scope of this paper. However, a sense of the type of formulation required is illustrated in this section with respect to creating a linear approximation for a general nonlinear time series model.

The nonlinear conceptual time series model is familiar and not necessarily specific to hydrology: Events occur at points in time and each event marks the initiation of a continuous non-negative nonlinear response which may be of arbitrary form but eventually declines to zero with increasing time. The responses may change from one event to the next depending on event magnitude and the current state of the system. The event responses sum together, producing the model time series output for some recording point. Variations of this approach have long been used in the context of rainfall-runoff modelling where the current state of a catchment influences the nature of runoff responses from rainfall events, with the individual event responses summing to give model discharge, possibly superimposed on a constant baseflow.

A conversion of this conceptual model to a linear approximation model is demonstrated by first considering a single event and its response. Defining this event to occur at time $t = \tau$, the magnitude of the response at any subsequent time t is expressed as a weighted finite mixture of L pre-selected non-negative nonlinear functions $g(t)$, all with origin at time τ:

$$f(t, Z_\tau) = \sum_{i=1}^{L} \omega_i(Z_\tau) g_i(t) \quad t \geq \tau \tag{1}$$

The $\omega_i(Z_\tau)$ terms in Eq. (1) are non-negative weighting expressions which give greater or lesser emphasis to individual $g_i(t)$ functions. The particular set of $g(t)$ functions chosen by the modeller would be representative of a range of possible event responses for the physical process under consideration. For example, in the case of a rainfall-runoff model this could be a number of different hydrograph forms characteristic of the catchment type and size. The choice of $g(t)$ functions will inevitably include some which will not in fact be helpful for a given application to data. However, these redundant functions will be eliminated later in the simplification process. A greater number of $g(t)$ functions would be chosen for a model intended for more general use. This will result in a greater number of $g(t)$ eliminations during the subsequent simplification when applied to data.

The $\omega_i(Z_\tau)$ terms in Eq. (1) are linear combinations of a set of M independent variables Z_τ whose magnitude may have influence on the system at event time τ:

$$\omega_i(Z_\tau) = \sum_{j=1}^{M} a_{ij} Z_{\tau j} \quad a_{ij} \geq 0 \quad Z_{\tau j} \geq 0 \tag{2}$$

The avoidance of negative $\omega_i(Z_\tau)$ terms ensures Eq. (1) cannot yield a negative prediction for the positive-valued response process concerned.

Eq. (2) defines the ith of the L weighting expressions in Eq. (1) and is a linear combination of the same set of independent variables Z. However, the weighting coefficients a_{ij} differ in value from one weighting expression to the next. The initial choice of the M independent causal variables represents a physical working hypothesis and it may happen that most are later eliminated in the linear LASSO simplification process, described in the next section.

In summary, the nonlinear response following a single event is modelled as a positive-valued weighted mixture of pre-chosen nonlinear $g(t)$ functions, with their associated weights being linear combinations of M independent causal variables. As noted earlier, pre-chosen functions like $g(t)$ are referred to as basis functions (Bishop, 2006, p.138) and when used as weighted mixtures can approximate many different nonlinear functions when L is sufficiently large.

With respect now to multiple events, the events are defined to occur at respective times τ[1], τ[2], τ[3] …, with the same set of $g(t)$ functions and Z variables operative for each τ[i]. However, the respective weights $\omega(Z)$ for each of the $g(t)$ will differ from one event to the next because the magnitudes of the Z variables change with time.

The model time series output at any given time t is the sum of the responses from all previous events up to that time. Defining $t = 0$ as the start of the time series, at some subsequent time t there will have been $K(t)$ previous events which occurred at times τ[1], τ[2] … τ[K]. Therefore, at time t the model-generated value $h(t,Z)$ can be written:

$$h(t, Z) = \omega_0 + \sum_{n=1}^{K(t)} f\left(t, Z_{\tau[n]}\right) \tag{3}$$

where $Z_{\tau[n]}$ denotes the magnitudes of the independent variables at the time of the nth event. The constant ω_0 may be set to zero depending on the context. For example, a non-zero value might represent some constant baseflow in a rainfall-runoff model.

As an aside, if an event can be thought of as the input of a set of particles into a store and the response is the time-varying rate of exit of those particles from the store, then at any time t the model-defined mean residence time $T(t,Z)$ of particles (derived from all the prior events) exiting the store is given by the weighted average of the previous event times:

$$T(t, Z) = \sum_{n=1}^{K(t)} \tau[n] f\left(t, Z_{\tau[n]}\right) \Big/ \sum_{n=1}^{K(t)} f\left(t, Z_{\tau[n]}\right) \tag{4}$$

Eq. (4) could have application, for example, in considering the age of water exiting from a catchment.

Having expressed the conceptual model as a linear

approximation, it remains to set out the coefficients as an LP minimisation matrix. The coefficients are all constrained to be non-negative to avoid negative $h(t,Z)$ values and the minimisation is with respect to least absolute deviations.

Following from Eq. (3), the matrix will have U rows (with each row corresponding here to a unit of time) and $L \times M + 2U$ columns, with one additional column with all values set to 1.0 if ω_0 is permitted to have an unknown nonzero value. The $2U$ columns here are the utility fitting variables required for least absolute deviations regression, two per data observation, and are not part of the model (Bloomfield and Steiger, 1983; ch. 6). As far as the model parameters are concerned, ω_0 would be an unknown to be solved for, along with the $L \times M$ unknown a coefficients. All the unknowns are constrained to be non-negative in the LP solution, as required by the specification of Eq. (1) and Eq. (2).

The number of matrix rows U will generally be less than the original number of rows in the time series concerned. This is because the user must define the first row in the matrix to be corresponding to a t large enough to avoid any response effects which may be still present from events prior to the start of the time series at $t = 0$.

The LP matrix does not define a linear model with M independent Z variables corresponding to the independent X variables of a standard linear regression. The Z variables do not influence the model in a direct linear way, but indirectly through the intermediary of determining the weight magnitudes via the time-varying weighted linear combinations of the L different nonlinear $g(t)$ functions. It may happen that the Z variables are themselves outputs from pre-chosen nonlinear expressions. The $L \times M$ variables here might be better termed pseudo variables because they combine the effect of different $g(t)$ functions as opposed to physical variables.

This type of initial model with numerous $g(t)$ functions will inevitably result in many superfluous parameters and there is no suggestion that such models should be applied directly in practice. Instead, they are only a means to an end and serve as the necessary preliminary stage before initiating the subsequent linear LASSO simplification to produce models for application. The following section gives a brief general description of the LASSO concept before considering the simplification algorithm.

3. The LASSO and linear LASSO

The LASSO (Least Absolute Shrinkage and Selection Operator) was introduced by Tibshirani (1996) as a means of eliminating less informative variables in least squares multiple linear regression. It has been applied in many fields but has only relatively recently been introduced into the hydroclimatic literature (Hammami et al., 2012). To our knowledge, the present paper is the first application of the LASSO in the more general context of model simplification rather than simply selection of a subset of informative independent variables in linear regression.

Briefly, the LASSO concept maintains the linear regression approach of seeking to match a linear function to a data set, but with the additional aspect of some degree of forcing of the parameters toward zero. Scaling is required prior to avoid preferential elimination of parameters because of units of measurement. A user-specified positive parameter λ defines the relative partitioning between optimising the parameter values toward data matching or forcing the parameters to zero. A large value of λ will cause all parameters to be set to zero while a small λ will not have any simplifying effect. See also Hammami et al. (2012), Wheeler (2009), and Tibshirani (2011).

The deleted variables will be those whose elimination has least effect on fitting the linear regression model to data while all parameters are being forced toward zero. Deletion might arise, for example, if a variable has weak explanatory power. Alternatively, some variables may be highly correlated so that when one is eliminated to zero another can take its place.

The nonzero parameters remaining after a LASSO process will have values biased toward zero. This can be offset with a subsequent standard linear regression with the independent variables now being just that surviving subset. The parameter values from the second regression will usually have larger absolute values and the model will better fit the data because the biasing effect will be at least partly removed.

However, the least-squares LASSO has a disadvantage for model simplification purposes. Specifically, linear constraints cannot be included without transforming the fit procedure into a quadratic optimisation exercise, which may be slow to run for large problems and will not necessarily yield a global minimum.

There is particular advantage in being able to incorporate linear constraints into models, both as part of the model description and because the constraints may result in many parameters never becoming part of the model. Such model improvement from inclusion of constraints has previously was noted, for example, by Gharari et al. (2014).

The constraints here might be as simple as avoiding negative discharge in a hydrological model or could be a more complex constraint set to approximate some physical process.

We therefore utilise here the linear LASSO (Wang et al., 2006) as the LASSO version most suited to model simplification. This permits linear constraints while still giving a single optimal global solution for calibration fitting with specified λ. In addition, the linear LASSO can be applied when there are more parameters than data points.

Given N data values of some dependent variable Y, and J independent variables in the absolute deviations regression matrix, the linear LASSO can be written as the LP minimisation:

$$\text{Minimise} \quad \sum_{t=1}^{N} \left| Y_t - \left(a_0 + \sum_{i=1}^{J} X_{it} a_i \right) \right| + \lambda \sum_{i=0}^{J} |a_i| \tag{5}$$

where J is not necessarily less than N.

Eq. (5) is a penalised regression operation that is a compromise between minimising absolute deviations from data and forcing the absolute values of the a coefficients toward zero. The minimisation here is with respect to finding the set of a coefficients which minimise Eq. (5) conditional on a specified value of λ. Increasing λ has the effect of forcing more a coefficients to zero. As with the least-squares LASSO, the X values are first standardised to dimensionless values (making λ dimensionless) so that no one a_i is preferentially forced to zero because of units of measurement. As before, when λ is sufficiently large the a coefficients will be forced to zero to give the limit case of maximum simplification but no predictive ability.

4. Model simplification with the linear LASSO

It is assumed that a model has already been formulated as an LP optimisation matrix so that the Eq. (5) minimisation can apply. The required data here is a calibration set and associated independent variables, together with a second data set for model evaluation purposes.

Our approach to linear LASSO model simplification concurrent with calibration is summarised in the sequential minimisation schematic:

For $\lambda = 0, D,$ **Step** $\Delta\lambda$

$$\text{Minimise} \quad \sum_{t=1}^{N}\left|Y_t - \left(a_0 + \sum_{i=1}^{J} X_{it}a_i\right)\right| + \lambda\sum_{i=0}^{J}|a_i| \qquad (6)$$

$$\text{Minimise} \quad \sum_{t=1}^{N}\left|Y_t - \left(a_0 + \sum_{i=1}^{J(\lambda)} X_{it}a_i\right)\right|$$

End

where the a values may be subject to linear constraints as part of the model specification. J and $J(\lambda)$ are respectively the number of model parameters before and after the linear LASSO.

In words, λ is incremented from zero to some value D and for given λ there are two minimisations carried out in sequence, adjusting the a_j values each time. The first minimisation is the linear LASSO expression of Eq. (5) which serves to force some a values to zero. The second minimisation for bias correction is ordinary calibration matching to data using an LP routine to minimise least absolute deviations. This second minimisation uses only the $J(\lambda)$ number of X variables which survived the linear LASSO process. These variables are likely to gain new a coefficients, some of which may in fact be zero if the coefficients have been constrained to be non-negative.

Depending on the model, the X variables in Eq. (6) may be independent variables in the usual regression sense, but they might also be pseudo-variables as noted in the previous section.

As λ is incremented from zero there will be a general decline in the number of surviving independent variables. At the same time, there will tend to be an improvement in the fit of the model to the evaluation data set because overfitting effects are progressively reduced. However, at some point any further increase in λ causes worse fits to the evaluation data because the model has become overly simplified and inflexible. Between these end points there will be some optimal zone which may include a number of moderately simplified models which might all fare reasonably in both calibration and evaluation. In this case a subjective choice of appropriate model must be made, which might favour the simplest model within the acceptable range. The evaluation data set here plays a role in the choice of λ and ideally further sets of independent evaluation data should be utilised as a true test of the resulting simplified model.

5. Example application

The method will be illustrated with respect to simplifying a basic rainfall-runoff model constructed like the time series model described in Section 2. It is not our intention to establish the viability of the simplification approach by creating a fully-developed new simplified runoff model which is demonstrably at least as functional as some already in use. Our purpose is only to give an example demonstration of the simplification process mechanics for one rather basic model. However, we hope that sufficient interest may be generated for specialists to consider evaluating more developed linear LASSO simplified models for possible application in a range of different fields.

The example data comprises 600 observations of hourly rainfall and discharge data from a small 14 km^2 sub-catchment within the Mahurangi River drainage basin in the Northland region of New Zealand. Visual inspection of the hydrograph indicated rapid discharge responses and quite short recessions typical of a small catchment.

The $g(t)$ functions chosen comprise nine inverse Gaussian probability density functions with parameterisation given by:

$$g(t) = \left[\frac{\mu\phi}{2\pi(t-\tau)^3}\right]^{1/2} e^{\phi}\exp\left\{-\frac{1}{2}\phi\left(\frac{t-\tau}{\mu}+\frac{\mu}{t-\tau}\right)\right\} \quad t>\tau \qquad (7)$$

where ϕ is a dimensionless shape parameter and $\tau + \mu$ is the distribution mean. The parameter μ can also be thought of as the distribution mean as measured from origin at τ. Table 1 lists the nine μ, ϕ pairings which define the nine specific $g_i(t)$ functions chosen for the model, selected here to have distribution mode values soon after a rain event (Figs. 1 and 2).

The inverse Gaussian distributions have some linkage to hydrology in that they are derived as arrival time distributions which have analogies to the water arrival times of a hydrograph (Bardsley, 1983). The family of gamma distributions are another option for $g(t)$ functions with a hydrological linkage, in this case through a linear reservoir cascade model (Nash, 1957).

The choice of inverse Gaussian distribution parameters and the number of distributions selected was subjective and based only on having a sufficient range of $g(t)$ functions which were deemed typical of the hydrographs. A greater number of $g(t)$ functions is likely to be required for more complex hydrographs. A similar set of different inverse Gaussian distributions would have served equally well, as could other distributions discussed by various authors (Nadarajah, 2007; Pramanik et al., 2010; Muneepeerakul et al., 2010).

The subjectivity aspect is not a critical factor here because the hydrographs are being modelled as finite mixture distributions. For sufficiently large L such distributions have ability to mimic each other by adjustment of the weights of their respective component distributions.

The development of the specific runoff prediction model is now considered, followed by its simplification when applied to the example data set.

With respect to constructing the initial linear model, each hour of non-zero rainfall is an event which initiates a 9-component hydrograph as a weighted mixture of the nine $g(t)$ functions plotted in Figs. 1 and 2. Eighteen Z variables are defined as rainfall, the square of rainfall, and the cube of rainfall, with respect to both the current rainfall initiating the hydrograph and for each of the rainfalls in the five previous hours. That is, for a given event the weight assigned to each of the nine $g(t)$ functions is an 18-term cubic polynomial function of rainfall in the current hour and in each of the previous 5 h. All the polynomial coefficients are constrained to be non-negative, which in this case avoids negative discharge.

Because only rainfalls are being referenced as causal variables we symbolise here using R rather than Z. Eq. (3) can be written in this case to give the ith of the 9 weights which apply to a 9-component hydrograph originating from rain R_τ at time τ:

$$\omega_{\tau i}(R) = \sum_{t=\tau-5}^{\tau} a_{i1t}R_t + a_{i2t}R_t^2 + a_{i3t}R_t^3 \qquad (8)$$

Table 1
Parameter values of the nine inverse Gaussian distribution distributions utilised as $g(t)$ expressions. Both μ and distribution modes are measured as hours after time τ. Plots are displayed in Figs. 1 and 2.

Distribution	1	2	3	4	5	6	7	8	9
μ	3.00	5.00	5.00	7.00	10.00	20.00	20.00	50.00	150
ϕ	3.60	2.81	1.43	4.38	0.99	0.80	0.30	0.18	0.10
Mode	2	3	2	5	3	5	2	3	5

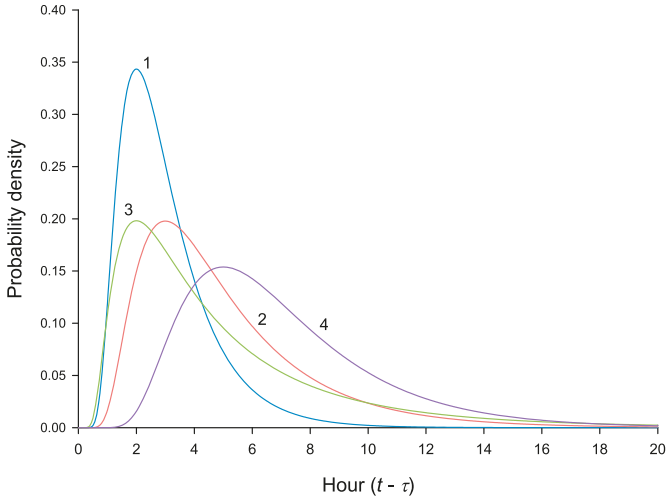

Fig. 1. Plot of chosen $g(t)$ functions: inverse Gaussian distributions 1–4 as listed in Table 1.

R_τ is always greater than zero but some or all of the rainfall for the previous 5 h may be zero, depending on previous rainfall frequency.

The weight specification in Eq. (8) does not mean that discharge is being modelled as a cubic polynomial function of rainfall. The polynomials relate to the individual weights, which are likely to influence different parts of the hydrograph. Also, many of the a coefficients will be removed during simplification so it is possible that some weights in the final simplified model will be reduced to linear functions of rainfall while others might be cubic or quadratic. For example, this allows for discovery of hydrograph peaks increasing nonlinearly with rainfall while recession discharge increases linearly with rainfall amount.

There is no specific hydrological reason for selecting the polynomial weight functions used here, other than giving a flexible range of possible linkages between rainfall and the respective weights. Implementing flexible empirical expressions as approximations to complex physical situations is not unknown in hydrology. For example, Young (2003, Eq. (8c)) transforms recorded rainfall to effective rainfall by way of a cubic polynomial function of an estimate of current soil water storage.

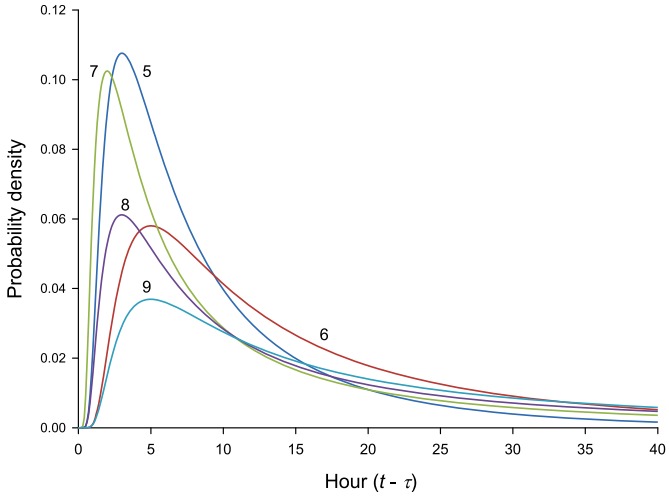

Fig. 2. Plot of chosen $g(t)$ functions: inverse Gaussian distributions 5–9 as listed in Table 1.

The non-negative constraint on the a coefficients in Eq. (8) might result in some of the power terms causing parts of an event hydrograph to increase rapidly with increasing rainfall, as might occur in reality as the extent of catchment saturation increases. Flexible functions other than polynomials could serve equally well here as long they are constrained to be positive. As will be seen, just the non-negative constraint may itself result in a dramatic reduction in the number of a parameters even before the linear LASSO is applied.

The use of previous rainfalls as causal variables could have some hydrological justification in terms of acting as a proxy for current catchment wetness at time of rainfall, but are used here mainly for the convenience of illustration. It is also noted that easily measured proxy variables for catchment wetness are a component in DBM modelling (Young, 2003). Ideally, a rainfall-runoff model should incorporate where possible direct measurements of soil moisture and other catchment variables, which would also be defined over space in the case of a distributed model.

From Eq. (1) and Eq. (3) the pre-simplification model $Q(t,R)$ for discharge at time t can be written:

$$Q(t,R) = \omega_0 + \sum_{n=1}^{K(t)} f\left(t, \tau[n], R_{\tau[n]}\right)$$
$$f\left(t, \tau[n], R_{\tau[n]}\right) = \sum_{i=1}^{9} \omega_{\tau[n]i}\left(R_{\tau[n]}\right) g_i(t) \quad (9)$$

where ω_0 is baseflow, $\omega_{\tau[n]i}(R_{\tau[n]})$ is from Eq. (8), and $g_i(t)$ is defined by Eq. (7) and Table 1.

A rainfall-runoff model of the same general form as Eq. (9) was developed by Bardsley and Liu (2003), although simplification possibilities were not considered in that paper.

The baseflow constant, the nine $g(t)$ expressions, and the 18-term weighting expressions collectively define an LP matrix which includes 163 model variables with coefficients to be either solved for or eliminated during the calibration/simplification process.

The data of the first 400 h in Figs. 3 and 4 are utilised for calibration/simplification, with hours 400–600 used for model evaluation. The evaluation data set was deliberately selected to include a peak discharge considerably exceeding the largest events in the calibration set.

Table 2 summarises the results of the application of the process of Eq. (6) to the 400 calibration data points, for selected values of the simplification parameter λ. For $\lambda > 0$, the simplification process is equivalent to an optimal allocation of weights between the nine $g(t)$ expressions, taking into account the twin aims of data fitting and model simplification. For $\lambda = 0$ the model seeks only the best fit to the calibration data with no simplification forcing. Various descriptive indices in Table 2 are listed for comparison of model values and data, for both the calibration and evaluation data sets.

The utilised goodness of fit index V in Table 2 is a measure which varies between 0 for worst fit and 1 for a perfect fit (Bardsley, 2013):

$$V = r^2 \Big/ (2 - E) \quad 0 \le V \le 1 \quad (10)$$

$$E = 1 - \frac{\sum (O_i - P_i)^2}{\sum \left(O_i - \overline{O}\right)^2} \quad -\infty < E \le 1$$

where E is the Nash-Sutcliffe efficiency (Nash and Sutcliffe, 1970) and r^2 is the coefficient of determination between the discharge observations O_i and model predictions P_i. Any V exceeding $0.5r^2$

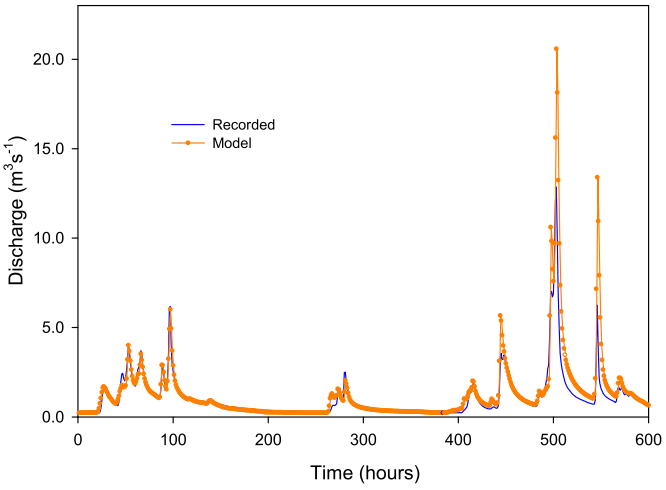

Fig. 3. Model calibration fitting ($\lambda = 0$, 13 nonzero parameters) prior to simplification. The calibration data is for the first 400 h, with the last 200 h being used for evaluation.

reflects that the model gives a better fit than the mean of the observed data. Therefore any V exceeding 0.5 ensures the model must match the data better than the data mean, with the degree of matching being quantified by comparison of squared deviations. A fit measure based on squared deviations instead of absolute deviations was used here to reduce the possibility of creating overly-favourable fit values after minimising least absolute deviations with the linear LASSO calibration.

It is evident from Table 2 that just the constraint against negative values is sufficient for initial elimination of a large number of model parameters, with the original 163 parameters being reduced to 13 for $\lambda = 0$. That is, any value other than zero for 150 parameters would give a worse calibration fit. Thirteen linear parameters is still a relatively large number for a runoff model and overfitting can be seen from the calibration and evaluation fit values being 0.92 and 0.52, respectively.

As λ increases there is a rapid improvement in the progressively simplified model's ability to match the evaluation data up to a maximum Ve of 0.87, corresponding to six model parameters and three $g(t)$ functions. When simplification forcing increases further there is a decline in both calibration and evaluation fits because of reduced model flexibility.

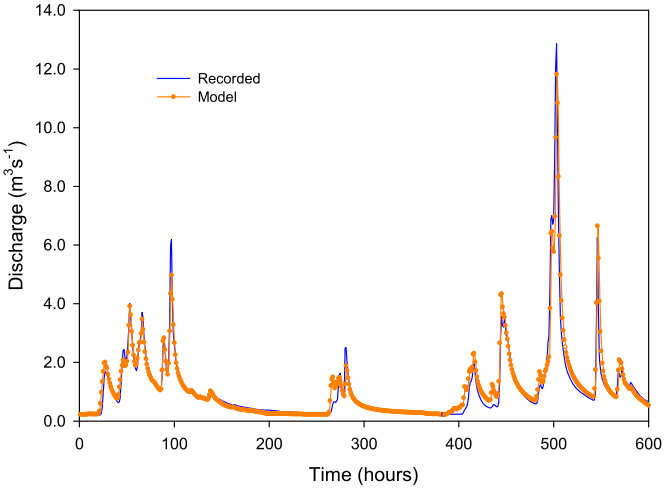

Fig. 4. Model fitting (6 parameters) to calibration data after previous linear LASSO simplification with ($0.9 \leq \lambda \leq 2.0$), giving the best match to evaluation data.

The linear LASSO induces abrupt model changes as λ increases past threshold values. For example, λ increasing over the small interval from 2.0 to 2.1 results in worse fits because of induced changes in the weighting function coefficients, although the total number of parameters and $g(t)$ functions remain unchanged. In contrast, the simplified model remains unchanged over the range $0.9 \leq \lambda \leq 2.0$, which also happens to give the best matching to the evaluation data.

The model's selection of $g(t)$ functions was quite stable with $L(\lambda) = 3$ for all nonzero values of λ. These groups consistently comprise the three distributions 1, 2, and 9 for λ in the range $0.4 \leq \lambda \leq 2.5$. Also, distributions 1 and 3 are present in all the groups of three for the whole range of λ investigated. The simplification process here evidently has the effect of progressively reducing the number of terms in the weighting expressions, but always keeping three $g(t)$ functions in the simplified models.

The reduced number of terms in the weighting expressions during simplification has the net effect of reducing the model discharge peak magnitudes in the evaluation period. This shows clearly in the listed d values of Table 2 where the difference between actual and model values for the highest flood peak shifts from over-prediction to under-prediction as λ increases. A similar pattern is evident for mean of the model discharges for the evaluation period. As might be expected, the mean model discharges for the calibration period are close to the true mean value of 0.79 $m^3\ s^{-1}$.

Fig. 3 shows a calibration/evaluation time series plot for the 13-parameter model for $\lambda = 0$, with poor matching to the evaluation data generally and there is considerable error in flood peaks.

Fig. 4 gives the corresponding time series for the best evaluation data match ($0.9 \leq \lambda \leq 2.0$). It is of interest that this level of simplification apparently gives a reasonable estimate of the 12.8 $m^3\ s^{-1}$ peak discharge around hour 500, despite the peak discharge of the calibration data being much lower (6 $m^3\ s^{-1}$). This simplified model comprises three $g(t)$ functions and six parameters, with one of the six being a non-zero baseflow constant. However, there is still probably a residual effect of calibration overfitting here. This can be seen by writing out the expression for the model discharge from a single rain event at time τ. Following from Eq. (9), this gives:

$$Q(t,R) = \omega_0 + \alpha_1 R_{\tau-1}^2 g_1(t) + \alpha_2 R_\tau^2 g_2(t) + (\alpha_3 R_\tau + \alpha_4 R_{\tau-1} + \alpha_5 R_{\tau-5}) g_9(t) \tag{11}$$

where $t > \tau$, ω_0 is a positive baseflow constant and the α terms are non-zero numerical values corresponding to the surviving subset of the a coefficients defined in Eq. (8).

The suggestion of some overfitting effect here derives from the model's $R_{\tau-5}$ term in Eq. (11), because there seems no hydrological reason why a rainfall 5 h previously should specifically influence the weighting associated with distribution 9.

This influence of overfitting notwithstanding, the simplified model of Eq. (11) has an interesting aspect in that the selected inverse Gaussian distributions 1 and 2 have weight terms increasing with the square of rainfall. The only other surviving $g(t)$ function is distribution 9, which is weighted by a linear function of current and previous hourly rainfalls. From Table 1, distributions 1 and 2 are peaked $g(t)$ functions with relatively small mean values while distribution 9 is strongly skewed with a large mean, providing much of the hydrograph tail component. That is, the hydrograph peaks will tend to increase nonlinearly with rainfall while the hydrograph recession discharges will increase linearly with rainfall. Fig. 5 illustrates the simplified model's predicted nonlinear increase

Table 2
Results from the calibration/simplification process as a function of the forcing parameter λ. *Vc* = calibration goodness of fit; *Ve* = evaluation goodness of fit; $J(\lambda)$ = number of parameters remaining after simplification; $L(\lambda)$ = number of $g(t)$ functions remaining after simplification; *d* = model minus observed discharge for the peak flood discharge near hour 500 in the evaluation set; *Qc* = model mean discharge for calibration data; *Qe* = model mean discharge for evaluation data. Actual mean discharges for the calibration and evaluation data are 0.79 and 1.62 $m^3\ s^{-1}$, respectively. A constant non-zero baseflow parameter is present in all parameter sets.

λ	0.0	0.4	0.5	0.6	0.9	1.5	2.0	2.1	2.5	3.3	3.5	10.0
Vc	0.92	0.92	0.90	0.90	0.89	0.89	0.89	0.83	0.83	0.80	0.75	0.68
Ve	0.52	0.57	0.78	0.81	0.87	0.87	0.87	0.80	0.80	0.71	0.61	0.52
$J(\lambda)$	13	9	8	7	6	6	6	6	5	5	4	4
$L(\lambda)$	5	3	3	3	3	3	3	3	3	3	3	3
d ($m^3\ s^{-1}$)	7.70	7.20	1.80	1.53	−1.04	−1.04	−1.04	−3.90	−3.80	−5.50	−7.10	−8.10
Qc ($m^3\ s^{-1}$)	0.78	0.78	0.79	0.79	0.79	0.79	0.79	0.78	0.78	0.78	0.77	0.73
Qe ($m^3\ s^{-1}$)	2.32	2.18	2.01	1.95	1.82	1.82	1.82	1.72	1.72	1.65	1.54	1.43

in peak discharge, using hydrographs from progressively increasing simulated rainfalls. One implication of the influence of R^2 terms in the model is that catchment peak runoff response will be more difficult to predict when there is spatial variability of rainfall.

A differentiation between a nonlinear peak response and linear recession response is not unusual in hydrology, but it is of interest to check how well this effect is established by the model. In this regard, one useful approach that LASSO simplification permits is a form of sensitivity analysis by noting which model parameters are most resistant to elimination by further increasing λ.

This effect on model parameter configuration with increasing λ is shown in Table 3. The somewhat unlikely nature of the $R_{\tau-5}$ parameter for distribution 9 tends to be reinforced by its disappearance after $\lambda = 2.5$. A consistent feature, however, is the continued association of nonlinear weighting terms with the most rapid-response $g(t)$ distributions (distributions 5 and below). At the same time, there is continued association of linear weighting with the longer-tailed distributions 8 and 9. There is some degree of swapping among similar distributions with increasing λ, but the stability of the respective linear and nonlinear $g(t)$ weight associations suggests detecting a real hydrological effect from this limited data set.

This purely illustrative model is deficient in many respects and as noted earlier there is no suggestion that a new functional rainfall-runoff model has now been created by the simplification process. There is no water balancing or residence time analysis, evaporation is included only implicitly, and the calibration period is too small to establish seasonality effects. Also, the evaluation data was used in deciding the appropriate λ. For proper evaluation of any LASSO-simplified model it would be necessary to consider fully independent evaluation data sets subsequent to the selection of λ. It would also be desirable to use multiple evaluation criteria and not just a single matching index (Bennett et al., 2013).

Furthermore, despite its large number of parameters the initial linear model incorporates no more hydrological structure than a basic unit hydrograph scheme with time-varying parameters. It could be desirable, for example, to include threshold effects so that small amounts of rainfall might not necessarily generate a hydrograph response. At the expense of an increased number of initial parameters, this could be achieved using higher order polynomials to allow an effectively zero $g(t)$ weighting to small rain events. Finally, no explicit consideration has been given to any error structure, which should be part of any viable predictive model.

6. Possibilities for the linear LASSO with groundwater models

Any model which is already expressed in linear form will be amenable to LASSO simplification. The obvious example of

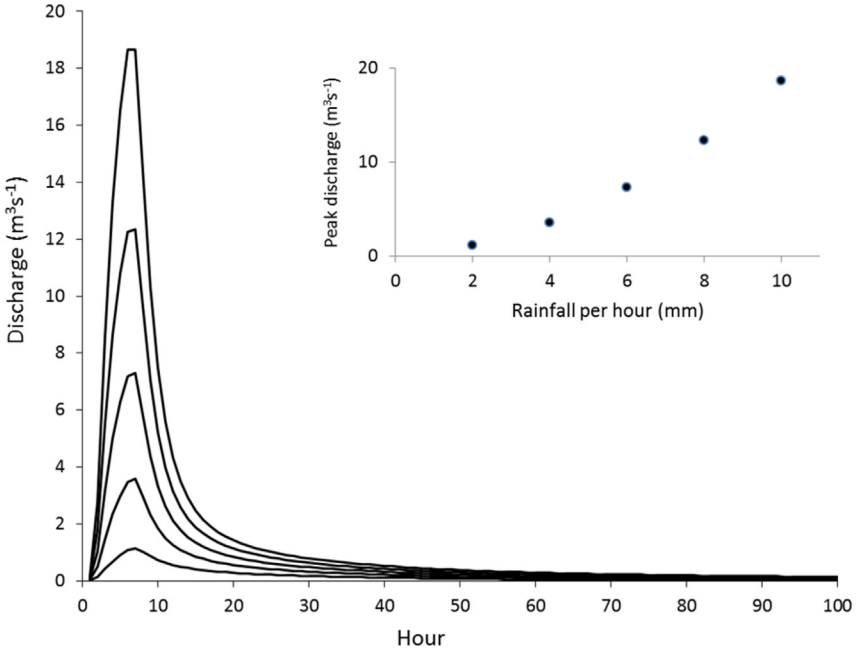

Fig. 5. Simulated hydrographs from the simplified model of Eq. (11), for 6 h of continuous rain at a constant rate, starting at zero time. Hydrographs from smallest to largest correspond, respectively, to rain rates of 2, 4, 6, 8, and 10 mm/h. Inset shows the nonlinear increase of peak discharge as a function of simulated rainfall rate.

Table 3
Parameters associated with surviving $g(t)$ functions for the larger values of λ. The $\lambda = 2$ column corresponds to the simplified model given by Eq. (11). All models also include a non-zero baseflow parameter ω_0.

g(t)	R	λ:	2	2.1	2.5	3.3	3.5	10
1	$R_{\tau-2}$			●	●	●	●	
1	$R_{\tau-3}$							●
1	$R^2_{\tau-1}$		●					
2	R^2_{τ}		●	●	●			
5	R^2_{τ}					●	●	
8	$R_{\tau-2}$							●
9	R_{τ}		●	●	●	●	●	●
9	$R_{\tau-1}$		●	●	●	●	●	
9	$R_{\tau-5}$		●	●				

numerical groundwater models is considered briefly here. Creating groundwater models no more complex than needed was considered by Hill (2006) by way of building up a groundwater model incrementally. Other studies discussing groundwater model simplification include Fienen et al. (2009) and Blakers et al. (2011).

For a runoff model the limit physical situation for total simplification is when all parameters are forced to zero so the model predicts zero discharge. For a groundwater model the equivalent situation is having all water fluxes forced to zero at all locations so there is no longer a groundwater flow field.

Considering the specific case of finite difference groundwater models, one approach to LASSO simplification would be to seek to match some calibration data set while at the same time trying to force all head gradients to zero. The whole groundwater model here is set up and solved as an LP minimisation, where some of the finite difference equations may not be solved exactly because of the simplification effect. All boundary conditions here would have to be specified as linear constraints. In the ordinary groundwater calibration following the LASSO step the head gradients which had been forced to zero would be now fixed to be zero head gradients as an additional set of boundary conditions.

The motivation for LASSO simplification of a groundwater model is likely to be different from, say, simplification of a runoff model. Having a groundwater model with a considerable extent of local zero head gradients may well be far removed from known hydrogeological reality. However, the simplified models in this case could play a role as groundwater emulation models. That is, the simplified models could still answer the same groundwater questions as the original model but possibly at less cost because there is no need for detailed geological information from the new zero-flow regions of the emulation model. The emulation models would also have a further advantage in that they are likely to run faster.

7. Discussion

This section anticipates some of the issues and concerns likely to arise from the simplification modelling approach considered in this paper. The LASSO simplification approach is essentially opposite to usual modelling in hydrology whereby a model structure is first selected and then effort expended seeking to optimise a relatively small number of nonlinear parameters. In contrast, the simplification process starts with a large number of user-specified fixed-value nonlinear parameters and then seeks optimal combinations of linear mixtures of functions of the fixed parameters. Unhelpful functions and parameters are eliminated at the same time, with the model structure remaining unknown until completion of the process.

Although the nonlinear parameters are fixed and not optimised the procedure actually gives (i) greater flexibility and (ii) has more calibration utility. The first holds because finite mixtures of many nonlinear functions can give greater model flexibility than by optimising a few nonlinear parameters in a few functions. The second applies because the calibration process can be reduced to an LP optimisation problem, giving fast solutions and globally optimal parameters for a given level of simplification.

Creating simplified models from a linear LASSO process might be loosely described as nonparametric modelling because individual parameters will often have minimal identifiable connection to physical reality. This has implications for parameter interpretation, identifiability of parameters and models, and formulation of model error structure. It also raises the deeper issue of the nature of hydrological knowledge and how this should be incorporated into models.

All these topics are currently active research themes in the hydrological literature. For example, see Yen et al. (2014) and cited references for error analysis in a watershed modelling context. Beven and Young (2013) include a detailed description of various error types. Shin et al. (2015) discuss a range of methods for assessing identifiability of rainfall-runoff models.

Identifiability issues arise from LASSO-simplified models because a relatively large number of linear parameters proxy for a small number of nonlinear parameters. Unlike the case with most hydrological models, the linear parameters do not attempt to represent a physical reality. Individual parameters can be expected to appear and disappear with minor changes in calibration data. This is analogous to small changes in magnitude of a nonlinear parameter. Linear parameter instability is are also likely to occur from using different fitting weights in weighted calibration, as suggested by (Liang and Bárdossy, 2012) to emphasise some data points which are seen as being more informative. The simplified model parameter values will also change with different $g(t)$ choices, different levels of simplification forcing, and different methods of pre-LASSO standardisation of coefficient values. Similar issues will arise with any system of models with structural adaption to the situation of application – see, for example, Gray and Wotherspoon (2012).

Linear LASSO simplified models and their associated parameters are therefore poorly defined in the sense that a hypothetical large extension of calibration data is not guaranteed to give convergence to a particular set of final parameters. This differs from the limit behaviour of an ideal model as envisaged by Andréassian et al. (2012). In fact, it is not even helpful to refer to a simplified linear model as “the” model because there will be different models created from one calibration to the next in the same time series, even when using long calibration periods.

All this may create a sense of unease relative to the certainty of using established hydrological models with a fixed nonlinear structure and a relatively small fixed number of parameters with hydrological names. However, we suggest that unpredictable algorithmic model simplification is actually advantageous because it forces attention toward model output and away from a focus on model parameters with their associated issues of identifiability, estimation accuracy, and sometimes dubious physical linkages.

It is worth noting in this context the comments by M. Sivapalan (cited by Beven, 2008) that seeking detailed error analysis structures could be something of a distraction from the core scientific task of developing better hydrological modelling approaches and measurement techniques. Similarly, a focus on model-related identifiability will not necessarily lead to a useful conclusion in the absence of ability to clearly confirm that some particular feature has been uniquely “identified”. There are distant echoes here of a rather sterile period in engineering hydrology seeking probability

distributions and estimation methods which were best suited to histograms of extreme event data (Bardsley, 1994).

There is of course an essential need for error frameworks in any modelling endeavour but with simplification-driven variable model structure the main error focus is likely to be on predictive uncertainty. Fortunately, predictive uncertainty is mostly encapsulated by the differences in observed and model-predicted values (Mantovan et al., 2007), largely avoiding issues of error associated with model structural uncertainty. In addition some quantification of model uncertainty might be possible using linear programming tolerance analysis (Arsham, 2007).

One criticism that is likely to be raised of the linear LASSO simplification process is that the resulting models are not physically based because they are linear and hydrological processes may be strongly nonlinear. As discussed in Section 2, the reply here is that a linear model using pre-selected $g(t)$ or other nonlinear expressions is simply a different way of representing nonlinearity and is no less nonlinear because it does not explicitly incorporate nonlinear parameters. The standard example is that a dependent variable Y may be modelled as a nonlinear polynomial function of some independent variable X but the model is still linear because the powers of X have been calculated prior.

There is no reason therefore why the simplified linear models should not extrapolate as well as their nonlinear equivalents, providing the nonlinearities have been sufficiently captured in calibration. Like any other system of hydrological models, the calibration/simplification approach will fail at evaluation if the calibration data is not sufficiently extensive to include the full range of natural variation, or if spurious relations remain after simplification.

Concerns might also be raised over the use of minimising absolute deviations for model calibration and simplification, as opposed to least squares. This was discussed in earlier hydrological literature when Natale and Todini (1976a, b) demonstrated some improvement in unit hydrograph estimation by constrained least squares compared to a previous use of minimum absolute deviations (Eagleson et al., 1966). In practice, when dealing with constrained models with large numbers of observations it seems unlikely that least squares results would be very different, although there may be numerical problems with the resulting quadratic programming exercise, as noted in Section 3. More recently, Arkesteijn and Pande (2013) found it helpful to use a least absolute deviations fit measure in a formal analysis of complexity in hydrological models.

On a more technical matter, it is noted that parameters derived from LASSO simplification are not optimal in any formal sense (Fan and Li, 2001). Our use of the linear LASSO is purely pragmatic in that it quickly enables elimination of large numbers of irrelevant parameters. However, some potentially useful parameters might be eliminated and some poor ones selected. In specific instances involving smaller numbers of linear parameters it may still be best to use more classical methods. For example, Bardsley and Manly (1985) used all subsets regression in a simple estimation model as a means of identifying less informative rain gauges for removal.

Finally, a brief concluding comment is offered here with respect to hydrological knowledge and hydrological models generally. It could happen, for example, that a more fully developed version of the inverse Gaussian runoff model of Section 5 is found to give good results over multiple calibration and evaluation data sets. However, this would not imply that the small-scale processes of hydrograph generation within a catchment are well described by the directional random walk basis of inverse Gaussian distributions. Similarly, obtaining a good result if the model incorporated gamma distributions would not indicate that catchment water transfer is by routing through a series of linear reservoirs via the cascade model of Nash (1957). A more correct statement is that event hydrographs are positively skewed by delayed flow so mixtures of positively-skewed distributions are geometrically well suited to mimic hydrographs.

Indeed, it would be equally justifiable to replace the $g(t)$ functions with a set of subjectively-defined histograms without any reference to mathematical expressions. Going further, we would argue that at most spatial scales hydrological knowledge is essentially qualitative. However, local information may still be usefully incorporated into models (Hughs, 2010). This does not deny the existence of well-defined physics at the scale of the very small and very uniform, but simply that the heterogeneity of nature makes upscaling mathematically unjustifiable. Beven (2014) provides useful discussion on this topic.

Qualitative hydrological knowledge does not lend itself to explicit mathematical expression, but it should still be possible to express that knowledge as model inequality statements (Gharari et al., 2014). Further, if the entire model can be formulated in the first instance as a large number of linear inequalities then the linear simplification framework considered here is a natural means of creating hydrological models specific to a given application.

8. Further development

If the linear simplification approach outlined here proves useful in practice then it could be helpful to accumulate and share hydrological knowledge in the form of selected $g(t)$ functions or other nonlinear expressions relevant to the various fields of research. In the context of catchment modelling such collections would be analogous to the runoff parameter libraries suggested by Perrin et al. (2008).

It would also be of interest to apply the method to distributed catchment models (after an initial linear formulation). This should lead to identification, for example, of the localities within a catchment which are of most importance to predicting discharge characteristics. The resulting reduced model is also likely to show improvement in computational performance.

A further extension of the method could be to have the equivalent of a multi-objective calibration, which can be a useful means of incorporating further information into a model (Efstratiadis and Koutsoyiannis, 2010). This could be achieved in the framework considered here by still keeping a single linear objective function but with grouped terms weighted by their importance. Each group here would correspond to a different objective. In such situations it could happen that some of these embedded objective functions are eliminated in the simplification process.

The linear LASSO was discussed in this paper as the method of choice for simplifying hydrological models formulated as constrained linear systems. The method is well established, easy to understand, and has a large literature in both theoretical and applied fields. However, other selection methods are available as well and there is active research in the field. By way of some selected references, Li (1991) gives an inverse method of dimension reduction in general regression models and Kohn et al. (2001) describe a Bayesian methodology for linear basis function model simplification. Various aspects of subset selection techniques are reviewed by Johnstone and Titterington (2009), Miller (2002), and Galelli et al. (2014).

9. Conclusion

A method was presented which uses the linear LASSO as a means of formalising the simplification of a class of linear hydrological models. This is a departure from the traditional use of the LASSO for selecting a subset of predictor variables in multiple linear

regression.

The simplification method has attraction in its generality, being in principle applicable over many fields of nonlinear hydrology and in other subject areas of environmental science as well. However, the value of the approach is certainly dependent on meeting the challenge of formulating hydrological models using a sufficiently extensive set of linear equalities and inequalities so as to capture the main features of the nonlinear environmental processes. The way is then open for use of the simplification process to create fully functional reduced models no more complex than required for application to a given data set. Whether this can be achieved remains an open question for now. We hope that sufficient interest might be created to encourage in-depth studies of models derived from the linear LASSO or other linear simplification procedures.

Acknowledgements

We are grateful for helpful comments from various reviewers. A reviewer of an earlier version of the paper made the useful observation that some of our analysis was equivalent to using linear combinations of basis functions. Thanks also go to colleagues for their useful feedback on parts of this work presented at the 2012 Asia Oceania Geosciences Society Conference in Singapore and the 2013 MODSIM conference in Adelaide. Hydrology data from the Mahurangi catchment were provided by Ross Woods (NIWA), funded by FRST contract "Land Use Intensification: Sustainable Management of Water Quantity and Quality (C01X0304)". Much of the material of this paper was prepared by the senior author during a period of study leave in Beijing, funded by the Chinese Academy of Sciences. The project number is 2010T2Z33. The second author was supported by a PhD scholarship funded by the New Zealand Ministry of Business, Innovation, and Employment.

Appendix 1. Setting up the LP matrix for the time series model of Section 2.

Each row in the LP matrix is taken as corresponding to a particular time point t. For a given t the first element in that row in the LP matrix is given by:

$$\sum_{n=1}^{K(t)} g_1(t) Z_{\tau[n],1} \tag{A1}$$

where $Z_{\tau[n],1}$ denotes the magnitude of the first causal variable at the time of the n th rain event.

Similarly, the second element in the same row is given by:

$$\sum_{n=1}^{K(t)} g_1(t) Z_{\tau[n],2} \tag{A2}$$

where $Z_{\tau[n],2}$ denotes the magnitude of the second causal variable at the time of the n th rain event.

This process continues through to the M th element of the row:

$$\sum_{n=1}^{K(t)} g_1(t) Z_{\tau[n],M} \tag{A3}$$

Element $M+1$ in the row is the start of the terms for the second $g(t)$ function:

$$\sum_{n=1}^{K(t)} g_2(t) Z_{\tau[n],1} \tag{A4}$$

This process continues until the $M \times L$ elements of the row have been completed, and then all rows completed. The $M \times L + 1$ column in the LP matrix is set to 1.0 if ω_0 is permitted to be a positive constant. The matrix columns will then be rescaled, perhaps to a constant mean value, so that no one column is preferentially excluded in the linear LASSO minimisation simplification. Additional columns are then added corresponding to the LP utility variables used in the minimisation operation, in the usual setup for minimising least absolute deviations (Bloomfield and Steiger, 1983; ch. 6). Finally, the recorded time series is placed in the last column. With reference to the specifications of Section 2, all the unknown coefficients will be constrained to be non-negative in the LP solution process.

Once set up, the LP minimisation can be carried out with any suitable LP package. We used the open source software *lpsolve* (Berkelaar et al., 2004). The linear LASSO algorithm described by Wang et al. (2006) will be more computationally efficient, which could be a factor for the large LP minimisations created from extended calibration data sets and more complex initial linear models.

References

Andréassian, V., Le Moine, N., Perrin, C., Ramos, M.-H., Oudin, L., Mathevet, T., Lerat, J., Berthet, L., 2012. All that glitters is not gold: the case of calibrating hydrological models. Hydrol. Process. 26, 2206–2210.

Arkesteijn, L., Pande, S., 2013. On hydrological model complexity, its geometrical interpretations and prediction uncertainty. Water Resour. Res. 49, 7048–7063. http://dx.doi.org/10.1002/wrcr.20529.

Arsham, H., 2007. Construction of the largest sensitivity region for general linear programs. Appl. Math. Comput. 189, 1435–1447.

Bardsley, W.E., 1983. An alternative distribution for describing the instantaneous unit hydrograph. J. Hydrol. 62, 375–378.

Bardsley, W.E., 1994. Against objective statistical analysis of hydrological extremes. J. Hydrol. 162, 429–431.

Bardsley, W.E., 2013. A goodness of fit measure related to r^2 for model performance assessment. Hydrol. Process. 27, 2851–2856.

Bardsley, W.E., Liu, S., 2003. An Approach to Creating Lumped-parameter Rainfall-runoff Models for Drainage Basins Experiencing Environmental Change. IAHS Publication 281, pp. 67–74.

Bardsley, W.E., Manly, B.F.J., 1985. Note on selecting an optimum raingauge subset. J. Hydrol. 76, 197–201.

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andréassian, V., 2013. Characterising performance of environmental models. Environ. Model. Softw. 40, 1–20.

Berkelaar, M., Eikland, K., Notebaert, P., 2004. Lpsolve: Open Source (Mixed-integer) Linear Programming System. Eindhoven University of Technology. http://lpsolve.sourceforge.net/5.5/.

Beven, K., 2006. A manifesto for the equifinality thesis. J. Hydrol. 320, 18–36.

Beven, K., 2008. On doing better hydrological science. Hydrol. Process. 22, 3549–3553.

Beven, K., 2014. 'Here we have a system in which liquid water is moving; let's just get at the physics of it' (Penman 1965). Hydrol. Res. 45, 727–736.

Beven, K., Young, P., 2013. A guide to good practice in modeling semantics for authors and referees. Water Resour. Res. 49, 5092–5098.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer, Singapore.

Blakers, R.S., Croke, B.F.W., Jakeman, A.J., 2011. The influence of model simplicity on uncertainty in the context of surface – groundwater modelling and integrated assessment. In: 19th International Congress on Modelling and Simulation, Perth, Australia, 12–16 December. http://www.mssanz.org.au/modsim2011/I9/blakers.pdf.

Bloomfield, P., Steiger, W., 1983. Least Absolute Deviations: Theory, Applications and Algorithms. Birkhäuser.

Diodato, N., Brocca, L., Bellocchi, G., Fiorillo, F., Guadagno, M., 2014. Complexity-reduction modelling for assessing the macro-scale patterns of historical soil moisture in the Euro-Mediterranean region. Hydrol. Process. 28, 3752–3760.

Dooge, J.C.I., 1997. Searching for simplicity in hydrology. Surv. Geophys. 5, 511–534.

Eagleson, P.S., Mejia-R, R., March, F., 1966. Computation of optimum realizable unit hydrographs. Water Resour. Res. 2, 755–764.

Efstratiadis, A., Koutsoyiannis, D., 2010. One decade of multi-objective calibration approaches in hydrological modelling: a review. Hydrol. Sci. J. 55, 58–78.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. 96, 1348–1360.

Fenicia, F., Savenije, H.H.G., Matgen, P., Pfister, L., 2008. Understanding catchment behavior through stepwise model concept improvement. Water Resour. Res. 44, W01402. http://dx.doi.org/10.1029/2006WR005563.

Fenicia, F., Kavetski, D., Savenije, H.H.G., 2011. Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. Water Resour. Res. 47, W11510. http://dx.doi.org/10.1029/2010WR010174.

Fienen, M., Hunt, R., Krabbenhoft, D., Clemo, T., 2009. Obtaining parsimonious hydraulic conductivity fields using head and transport observations: a Bayesian geostatistical parameter estimation approach. Water Resour. Res. 45, W08405. http://dx.doi.org/10.1029/2008WR007431.

Galelli, S., Humphrey, G.B., Maier, H.R., Castelletti, A., Dandy, G.C., Gibbs, M.S., 2014. An evaluation framework for input variable selection algorithms for environmental data-driven models. Environ. Model. Softw. 62, 33–51.

Gharari, S., Hrachowitz, M., Fenicia, F., Gao, H., Savenije, H.H.G., 2014. Using expert knowledge to increase realism in environmental system models can dramatically reduce the need for calibration. Hydrol. Earth Syst. Sci. 18, 4839–4859.

Gray, R., Wotherspoon, S., 2012. Increasing model efficiency by dynamically changing model representations. Environ. Model. Softw. 30, 115–122.

Hammami, D., Lee, T.S., Ouarda, T.B.M.J., Lee, J., 2012. Predictor selection for downscaling GCM data with LASSO. J. Geophys. Res. 117 http://dx.doi.org/10.1029/2012JD017864. D17116.

Hill, M.C., 2006. The practical use of simplicity in developing ground water models. Ground Water 44, 775–781.

Hughs, D.A., 2010. Hydrological models: mathematics or science? Hydrol. Process. 24, 2199–2201.

Hunt, R.J., Doherty, J., Tonkin, M.J., 2007. Are models too simple? Arguments for increased parameterization. Ground Water 45, 254–262.

Jakeman, A.J., Hornberger, G.M., 1993. How much complexity is warranted in a rainfall–runoff model? Water Resour. Res. 29, 2637–2649.

Johnstone, I.M., Titterington, D.M., 2009. Statistical challenges of high-dimensional data. Philosophical Trans. R. Soc. A – Math. Phys. Eng. Sci. 367, 4237–4253.

Kohn, R., Smith, M., Chan, D., 2001. Nonparametric regression using linear combinations of basis functions. Statistics Comput. 11, 313–322.

Li, K., 1991. Sliced inverse regression for dimension reduction. J. Am. Stat. Assoc. 86, 316–327.

Liang, J., Bárdossy, A., 2012. Improving the calibration strategy of the physically-based model WaSiM-ETH using critical events. Hydrol. Sci. J. 57, 1487–1505.

Mantovan, P., Todini, E., Martina, M.L.V., 2007. Reply to comment by Keith Beven, Paul Smith and Jim Freer on "Hydrological forecasting uncertainty assessment: incoherence of the GLUE methodology". J. Hydrol. 338, 319–324.

Miller, A., 2002. Subset Selection in Regression In: Monographs on Statistics and Applied Probability, second ed., vol. 95. Chapman & Hall/CRC London.

Muneepeerakul, R.S., Azaele, S., Botter, G., Rinaldo, A., Rodriguez-Iturbe, I., 2010. Daily streamflow analysis based on a two-scaled gamma pulse model. Water Resour. Res. 46 http://dx.doi.org/10.1029/2010WR009286. W11546.

Nadarajah, S., 2007. Probability models for unit hydrograph derivation. J. Hydrol. 344, 185–189.

Nash, J.E., 1957. The Form of Instantaneous Unit Hydrograph. IAHS Publication 45, pp. 114–121.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models. Part I: a discussion of principles. J. Hydrol. 10, 282–290.

Natale, L., Todini, E., 1976a. A stable estimator for linear models 1. Theoretical development and Monte Carlo experiments. Water Resour. Res. 12, 667–671.

Natale, L., Todini, E., 1976b. A stable estimator for linear models 2. Real world hydrologic applications. Water Resour. Res. 12, 672–676.

Perrin, C., Andréassian, V., Rojas Serna, C., Mathevet, T., Le Moine, N., 2008. Discrete parameterization of hydrological models: evaluating the use of parameter sets libraries over 900 catchments. Water Resour. Res. 44, W08447. http://dx.doi.org/10.1029/2007WR006579.

Perrin, C., Michel, C., Andréassian, V., 2001. Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. J. Hydrol. 242, 275–301.

Pramanik, N., Panda, R.K., Sen, D., 2010. Development of design flood hydrographs using probability density functions. Hydrol. Process. 24, 415–428.

Schoups, G.N., van de Giesen, C., Savenije, H.H.G., 2008. Model complexity control for hydrologic prediction. Water Resour. Res. 44, W00B03. http://dx.doi.org/10.1029/2008WR006836.

Shin, M.-J., Guillaume, J.H.A., Croke, B.F.W., Jakeman, A.J., 2015. A review of foundational methods for checking the structural identifiability of models: results for rainfall-runoff. J. Hydrol. 520, 1–16.

Sivakumar, B., 2008. Dominant processes concept, model simplification and classification framework in catchment hydrology. Stoch. Environ. Res. Risk Assess. 22, 737–748.

Sivapalan, M., Blöschl, G., Zhang, L., Vertessy, R., 2003. Downward approach to hydrological prediction. Hydrol. Process. 17, 1201–2111.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. 58, 267–288.

Tibshirani, R., 2011. Regression shrinkage and selection via the lasso: a retrospective. J. R. Stat. Soc. 73, 273–282.

Tonkin, M.J., Doherty, J., 2005. A hybrid regularized inversion methodology for highly parameterized environmental models. Water Resour. Res. 41 http://dx.doi.org/10.1029/2005WR003995. W10412.

Wang, L., Gordon, M.D., Zhu, J., 2006. Regularized least absolute deviations regression and an efficient algorithm for parameter tuning. In: IEEE Sixth International Conference on Data Mining, Proceedings, pp. 690–700.

Wheeler, D.C., 2009. Simultaneous coefficient penalization and model selection in geographically weighted regression: the geographically weighted lasso. Environ. Plan. A 41, 722–742.

Yen, H., Wang, X., Fontane, D.G., Harmel, R.D., Arabi, M., 2014. A framework for propagation of uncertainty contributed by parameterization, input data, model structure, and calibration/validation data in watershed modeling. Environ. Model. Softw. 54, 211–221.

Young, P.C., 2003. Top-down and data-based mechanistic modelling of rainfall–flow dynamics at the catchment scale. Hydrol. Process. 17, 2195–2217.

Young, P.C., 2006. The data-based mechanistic approach to the modelling, forecasting and control of environmental systems. Annu. Rev. Control 30, 169–182.

Young, P.C., 2013. Hypothetico-inductive data-based mechanistic modeling of hydrological systems. Water Resour. Res. 49 http://dx.doi.org/10.1002/wrcr.20068.

Young, P.C., Garnier, H., 2006. Identification and estimation of continuous-time, data-based mechanistic (DBM) models for environmental systems. Environ. Model. Softw. 21, 1055–1072.

Young, P.C., Ratto, M., 2009. A unified approach to environmental systems modelling. Stoch. Environ. Res. Risk Assess. 23, 1037–1057.

Young, P.C., Ratto, M., 2011. Statistical emulation of large linear dynamic models. Technometrics 53, 29–43.

Young, P., Parkinson, S., Lees, M., 1996. Simplicity out of complexity in environmental modelling: Occam's razor revisited. J. Appl. Statistics 23, 165–210.