

Stochastic inflow modelling for hydropower scheduling problems

Geoffrey Pritchard*

3 December 2014

Abstract

We introduce a new stochastic model for inflow time series that is designed with the requirements of hydropower scheduling problems in mind. The model is an “iterated function system”: it models inflow as continuous, but the random innovation at each time step has a discrete distribution. With this inflow model, hydro-scheduling problems can be solved by the stochastic dual dynamic programming (SDDP) algorithm exactly as posed, without the additional sampling error introduced by sample average approximations. The model is fitted to univariate inflow time series by quantile regression. We consider various goodness-of-fit metrics for the new model and some alternatives to it, including performance in an actual hydro-scheduling problem. The numerical data used are for inflows to New Zealand hydropower reservoirs.

Key words and phrases: OR in energy, hydro-thermal scheduling, stochastic dual dynamic programming, time series, quantile regression.

1 Introduction

In a hydro- (or hydro-thermal) scheduling problem, one aims to optimally manage the operation of water reservoirs over a period of time. The objective may be to minimize the shortage that ensues when reservoirs run dry, or (more generally) the economic cost of meeting a given demand for electricity from hydro and other sources. Alternatively, it may be to maximize the revenues from selling water or hydro-electricity in external markets.

In all such problems, the central element is a stochastic model for natural inflows. This will typically be an autoregressive or Markov process which has been obtained by fitting to observed data.

The leading technique for the solution of hydro-scheduling problems is stochastic dual dynamic programming (SDDP). SDDP is a sampling-based version of the nested decomposition algorithm which dates back to [16]. Originally, it was designed to

*University of Auckland, New Zealand. g.pritchard@auckland.ac.nz

solve risk-neutral multistage stochastic linear programs. Recently, two extensions of the method to deal with risk-averse multistage stochastic programs have been proposed in [10, 11, 23] and have been applied to hydropower scheduling problems in [8, 19, 24]. The convergence of the method is proved in [14] for risk-neutral linear programs, in [5] for risk-neutral convex programs, and in [9] for risk-averse convex programs.

In problems amenable to SDDP, the flows and other quantities at each time step are described by a linear programming problem; thus they are continuous variables. But the SDDP algorithm also involves the computation of expectations, which (as a practical matter) requires a discrete probability space. For problems formulated with continuously-distributed random variables, the usual tactic is to replace each continuous distribution with a finite sample from it, giving a “sample average approximation” of the original problem.

Note that when we resort to sampling in this way, we incur both a statistical sampling error (when the inflow model is fitted to the original data) and a probabilistic, or simulation sampling error (when the fitted continuous inflow model is replaced by a discrete approximation obtained through sampling).

The motivating idea of the present paper is that errors of the second type should be avoidable by designing an inflow model in which the inflow variables themselves are continuous, but the random innovation at each time step is discrete. Such an inflow model renders the associated hydro-scheduling problem solvable by SDDP exactly as posed. The process of fitting the inflow model obtains, in one step, a good representation of the original data by a model of the final form required.

Another advantage of such a model is that it may actually be more general. By ignoring standard continuous model types (autoregressive, ARMA, etc.) and instead considering classes of discrete-innovation models that play well with SDDP, it may be possible to make better use of the limited flexibility that SDDP ultimately allows.

1.1 Hydro-scheduling and SDDP

The present paper will not attempt to give a complete description of the SDDP method as applied to hydro-scheduling problems. Instead, we refer the reader to [5, 9, 10, 11, 14, 15, 16, 23] for the full details, and focus here on those considerations relevant to stochastic modelling.

SDDP is an algorithm for solving multi-stage stochastic convex programming problems. In a practical implementation, the convex functions will be piecewise linear, which gives the problem the following general form. We wish to compute, or at least estimate, the functions

$$g_{t-1}(x) = E \left[\min_u \{c_t \cdot u + g_t(u) : A_t u = b_t + B_t x, u \geq 0\} \right] \quad (1)$$

for $t = 1, \dots, T$. Here c_t, b_t are random vectors and A_t, B_t are random matrices. The vector u may be thought of as representing the decisions taken during the t^{th} time period, or stage; we aim to make these decisions so as to minimize the sum of their immediate cost $c_t \cdot u$ and the expected cost $g_t(u)$ of their consequences in stages

$t + 1, \dots, T$. The possible decisions are constrained by the corresponding decisions x taken at the previous stage ($A_t u = b_t + B_t x$).

As a simple example, consider the classic hydro-thermal scheduling problem for a single reservoir, in which we must meet a demand δ_t for energy in stage t with some combination of thermal generation (which costs γ per unit) and hydropower (which costs nothing, though the water supply is limited). There is a random inflow I_t in stage t . The decision vector is $u = (u_\theta, u_H, u_S)$, with the components representing thermal generation, hydro generation, and reservoir storage at the end of the stage. Then the problem is

$$g_{t-1}(x) = E \left[\min_u \{ \gamma u_\theta + g_t(u) : u_\theta + u_H = \delta_t, u_H + u_S = I_t + x_S, u \geq 0 \} \right] \quad (2)$$

Note that the formulation (1) implies that the random elements at each stage (c_t, b_t, A_t, B_t) are independent of those at all previous stages (because $g_t(u)$ is non-random). For problems involving serially dependent processes, this limitation can be overcome by adding “decision” variables whose purpose is to transmit stochastic information to future stages. For example, suppose the inflows (I_t) above behave as an AR(1) process:

$$I_t = \alpha I_{t-1} + E_t,$$

where the E_t are independent. We can incorporate this behaviour into the problem by giving u an additional component u_I to represent the current stage inflow, and writing the constraint set in (2) as

$$u_\theta + u_H = \delta_t, \quad u_H + u_S - u_I = x_S, \quad u_I = \alpha x_I + E_t, \quad u \geq 0.$$

Now suppose that a problem of the form (1) is to be solved by stochastic dual dynamic programming (SDDP). In this method, the most demanding step (the “backward pass”) requires the computation of the expectation in (1). In problems where the random elements c_t, b_t, A_t, B_t have a discrete joint distribution, this can be done by considering all their possible values (and solving the optimization problem for each one). Otherwise, it is usually necessary to resort to sampling. Replacing the true probability distribution by a finite sample at each stage gives a “sample average approximation” of the problem which we can solve.

Note that there is no requirement in SDDP for the problem’s stages to resemble each other closely; it is enough that they all fit the general form (1). The implication for modelling is that random processes can be non-stationary – indeed, it is hardly necessary to think of them as processes at all. An inflow process, for example, could be modelled not only with different parameter values for winter and summer, but by entirely different statistical techniques for different times of year – a point exploited by [20]. The only requirement is that, whatever methods are used, the model must eventually be expressible in the form (1).

2 Inflow model forms

Consider the problem of developing a stochastic model of a (univariate or multivariate) inflow process, from raw data consisting of a time series of flow observations.

The model should be convenient for use within a hydro-scheduling problem of the general form (1).

It is well-known that stagewise-independent inflow processes are inadequate for this purpose. Inflows to real hydrological catchments usually have positive serial correlation; the effect of this is to increase the risk that reservoirs will run dry or overflow. To introduce appropriate serial dependence, our model must be expressible in the general explicit form

$$X_t = F_t(W_{t-1}). \quad (3)$$

That is, the inflow X_t to the catchment at stage t is expressible as some random function F_t of a state variable W_{t-1} . The latter is a vector whose entries comprise some of the previous stage subproblem's variables; it can be thought of as representing the "climatic state" at the previous stage. In the univariate (single catchment) case, a natural choice would be $W_{t-1} = X_{t-1}$ (or possibly $W_{t-1} = (X_{t-1}, X_{t-2}, \dots, X_{t-r})$, for a higher-order model).

The random functions $(F_t)_{t=1}^T$ must be independent, and concave with probability 1; in an actual implementation they will be expressed as piecewise linear functions:

$$X_t \leq \min_j F_t^j(W_{t-1}), \quad (4)$$

where the F_t^j are random affine functions. Although the formulation requires an inequality (\leq), the intent is that equality will prevail at optimality. This will generally be so in situations where water has positive marginal value (*i.e.* more is better). It may fail in situations where water has zero marginal value (*e.g.* spilling from a reservoir) or negative marginal value (*e.g.* a binding maximum flow constraint on a canal). To maintain the equality in these situations, it may be necessary to place a "big M " penalty on the amount of any inequality.

The present paper will confine itself to first-order models in which F_t itself is affine with probability 1. This turns out to be a quite adequate model in many instances, and avoids any restrictions on the sign of the marginal water value. That is, we have

$$X_t = D_t + M_t X_{t-1}, \quad (5)$$

where D_t is a random vector and M_t a random matrix. In the univariate case, D_t and M_t are random variables.

2.1 Autoregressive models

A commonly-used model is the autoregressive (AR(1)) process:

$$X_t = \alpha_t + \beta_t X_{t-1} + E_t,$$

where α_t , β_t are deterministic and the innovation process (E_t) is a sequence of independent random variables. This gives a special case of (5) in which only the intercept $D_t = \alpha_t + E_t$ is random, while the slope $M_t = \beta_t$ is deterministic.

Alternatively, the process may be applied at the logarithmic level:

$$\log(X_t) = \alpha_t + \beta_t \log(X_{t-1}) + E_t.$$

(It is common to transform flow data by taking logarithms, because they often have right-skewed distributions.) This gives

$$X_t = e^{\alpha_t} X_{t-1}^{\beta_t} e^{E_t}.$$

The function $x \mapsto x^{\beta_t}$ is concave (provided $0 < \beta_t < 1$, which we may expect), so we have a model of the required form. A piecewise linear approximation of the sort in (4) will be required; if this is simply an affine piece we obtain a model

$$X_t = (\theta_t + \eta_t X_{t-1}) R_t, \tag{6}$$

where θ_t, η_t are deterministic and $R_t = \exp(E_t)$ is random. See [24] for examples. This approach gives another special case of (5): both the intercept $D_t = \theta_t R_t$ and the slope $M_t = \eta_t R_t$ are random, but their randomness derives from the factor R_t , which they have in common. The only difference between “wet” and “dry” realizations of the random linear function F_t is a constant multiplicative factor.

Thus, neither variety of autoregressive model takes full advantage of the flexibility offered by the form (5).

2.2 Iterated function system models

The approach championed in the present paper will be to give F_t in (3) a finite discrete distribution. That is, there will be a finite collection of (affine) functions $f_t^1, \dots, f_t^{m_t}$ giving the possible values of F_t , with associated probabilities. The main motivation for this idea is the thought that we will eventually need a model of this form anyway, for (SDDP) computational purposes. Rather than approximating the data with a continuous model, and then approximating that model with a discrete one (a sample average approximation) by sampling, it might be more efficient to seek from the outset the best possible representation of the original data by a discrete model of the final form required.

Note that (3) with discretely distributed F_t implies that each X_t is also discretely distributed (provided the initial X_0 is discretely distributed). However, the number of possible values of X_t may be combinatorially large (as many as $\prod_{s=1}^t m_s$). It is thus useful in practice to continue to think of X_t as a continuous variable, even when it is not.

Stochastic processes of the general form (5) with discretely distributed random affine mappings have been studied for several decades under the name “iterated function systems”. See [2, 3] for early papers; also [1, 4, 13, 27]. The main problem in the field is the inverse problem: given a probability distribution μ on \mathbf{R}^d , find an iterated function system that reproduces μ . That is, we seek a finite collection f^1, \dots, f^m of affine maps on \mathbf{R}^d , with associated probabilities p_1, \dots, p_m , such that when the F_t are drawn independently from this discrete distribution, the Markov process defined by $X_t = F_t(X_{t-1})$ has limiting distribution (approximately) μ . The principal application to date has been to graphic image compression with $d = 2$.

In applying the iterated-function-system technique to inflow modelling, we face a couple of complications. One is that the targeted probability distribution for inflow

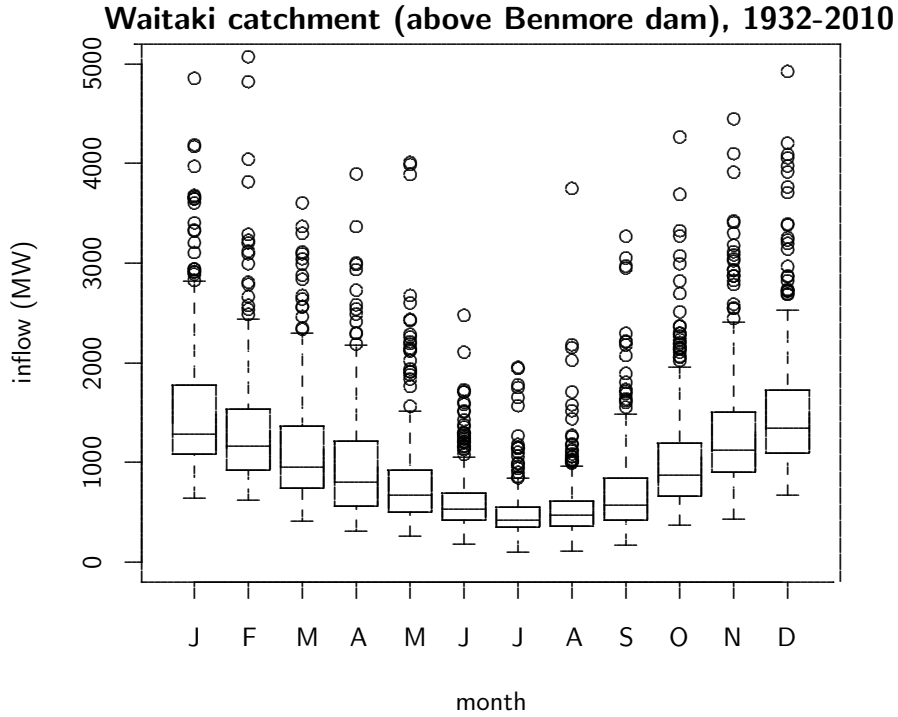


Figure 1: Raw inflow data.

varies seasonally. The Markov process (X_t) must therefore be time-inhomogeneous, with the distribution of F_t varying in a periodic (seasonal) fashion. Another is that we aim to reproduce not only the correct distribution for X_t , but also the serial dependence structure of the inflow process.

In the univariate case, both complications can be addressed by a technique of fitting the distribution of F_t via a quantile regression of X_t vs. X_{t-1} and seasonal regressors. The method will be described in detail in the next section.

3 Inflow model fitting

This section considers the inflow modelling approaches described in Section 2 in more detail, using a typical example univariate flow data set to illustrate.

The example data are weekly natural inflows to the Waitaki River system in New Zealand. This system includes eight hydroelectric generation sites, of which only the three furthest downstream have access to the whole flow. The water flows between dams are not heavily constrained, meaning that a hydropower scheduling model could reasonably treat the entire catchment as a single equivalent-energy reservoir. Accordingly, all inflow values have been converted to their electrical energy equivalent, giving a univariate time series for total inflow expressed in power units (megawatts).

The sources of the inflow include both summer melting of snowpack and glaciers

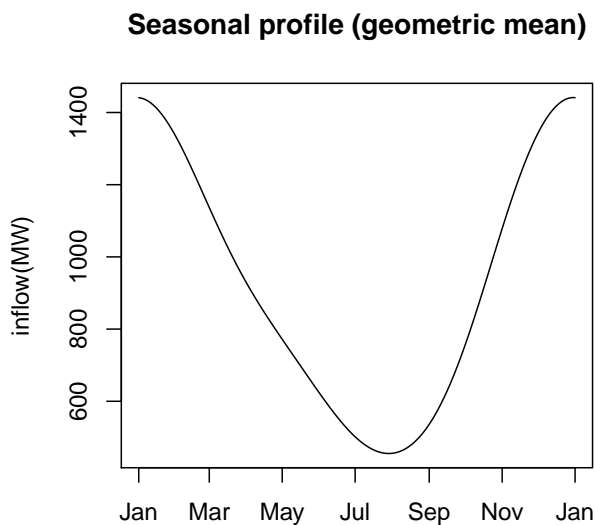


Figure 2: The fitted seasonal profile $g(t)$, a third-order trigonometric polynomial.

in high mountain areas and year-round rainfall in lower-lying parts of the catchment. This combination gives the raw inflow series a fairly strong seasonal dependence, with average summer peak flow around 3 times the average winter minimum.

As with many other hydrological series, the distribution of weekly inflow is strongly right-skewed, with prompt runoff from storm events producing observations exceeding 3 times the seasonal average at the rate of about one per year. However, there is no obvious concentration of these extremes at any particular time of year.

There is also a clear serial correlation, with a correlation length on the order of a few weeks. Closer examination reveals that this correlation structure also varies seasonally, with the strongest serial correlations in the late winter, and the weakest in late summer. This presumably reflects the different physical processes dominating at different times of year.

We begin by applying an initial scaling to the inflow series (X_t) :

$$Q_t = X_t/g(t),$$

where $g(t)$ is a seasonal periodic function. We work thereafter with (Q_t) . This scaling is not meant to imply any assumed stationarity of (Q_t) ; it merely simplifies the problem a little for later modelling efforts, by obviating the need to deal with large seasonal changes of magnitude.

The function $g(t)$ is constructed by Fourier regression at the logarithmic level:

$$\log(X_t) = f(t) + Z_t, \tag{7}$$

where $f(t)$ is a fitted trigonometric polynomial and (Z_t) are mean-zero errors. (The trigonometric polynomial form is used because the data are weekly, so observations

do not fall on the same dates in each year. For monthly data, it would be simpler to just estimate a seasonal adjustment factor for each of the 12 months.) We then take $Q_t = e^{Z_t}$ and $g(t) = e^{f(t)}$, making $g(t)$ an estimate of the conditional geometric mean inflow for the time of year of t .

Figure 2 shows the fitted scale function $g(t)$ for the example data set. The fitting, and all other statistical analysis in the present paper, were conducted using the statistical language R ([21]). Code and data are available from the author on request.

3.1 Autoregressive model

Following the approach outlined in Section 2.1, suppose we now postulate an AR(1) model for Z_t ($= \log Q_t$):

$$Z_t = \gamma(t)Z_{t-1} + \epsilon_t. \quad (8)$$

Here $\gamma(t)$ is a seasonally varying coefficient and (ϵ_t) a white noise process. There is no need for a constant term because (7) has removed it. This gives

$$Q_t = Q_{t-1}^{\gamma(t)} R_t,$$

where $R_t = \exp(\epsilon_t)$. Since $Q_t \approx 1$ is a typical value at any time of year, we approximate the function $q \mapsto q^{\gamma(t)}$ by its linearization about $q = 1$, *i.e.*

$$q^{\gamma(t)} \approx 1 - \gamma(t) + \gamma(t)q,$$

giving the model

$$Q_t = (1 - \gamma(t) + \gamma(t)Q_{t-1})R_t. \quad (9)$$

The estimation of $\gamma(t)$ requires some care. Perhaps the most obvious method is to fit the linear autoregression (8): this will be good as far as (8) goes, but it makes no allowance for the effect of the subsequent linear approximation step. Alternatively, one could reason (as in [24]) that (9) with $R_t = 1$ suggests the autoregression

$$Q_t - 1 = \gamma(t)(Q_{t-1} - 1) + \text{error}. \quad (10)$$

This approach is also questionable because it assumes the wrong error structure (additive, rather than multiplicative). For the purposes of the present paper, we fit (9) for the multiplicative error structure by maximum likelihood. Assuming normal errors ϵ_t , this is equivalent to minimizing the sum of squares

$$\sum_i \left(\log \left(\frac{q_i}{1 - \gamma(t) + \gamma(t)q_{i-1}} \right) \right)^2,$$

where the q_i are the observations of the process (Q_t). Differences among the fitting methods have a noticeable effect on the estimate obtained for $\gamma(t)$, as illustrated by Figure 3.

To completely characterize the stochastic process, it is also necessary to describe the distribution of the errors ϵ_t , or equivalently of R_t . The estimation of $\gamma(t)$ assumed

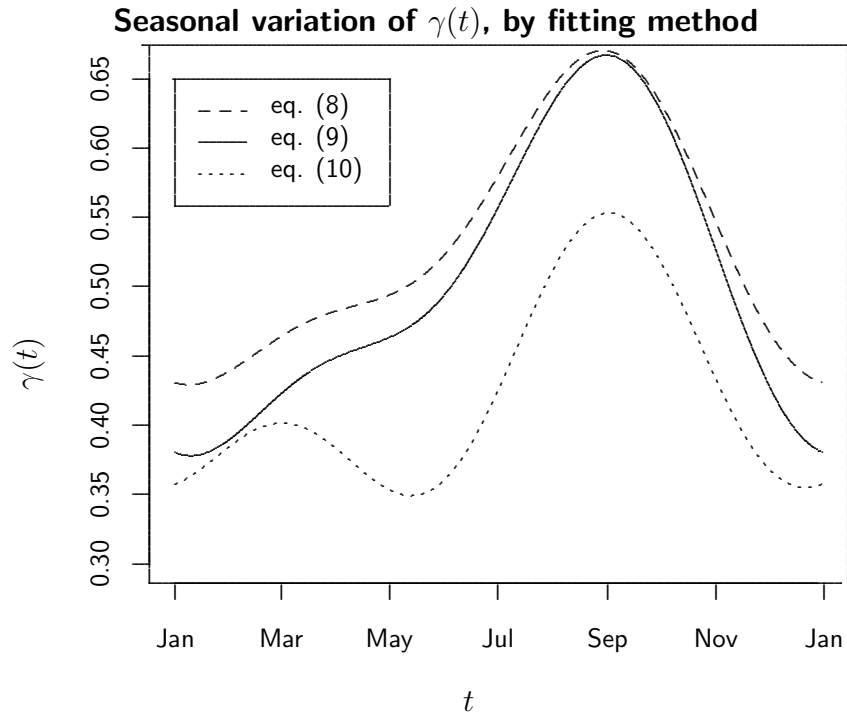


Figure 3: Seasonal autocorrelation structure. The graph shows second-order trigonometric polynomial models for $\gamma(t)$ fitted by three different methods.

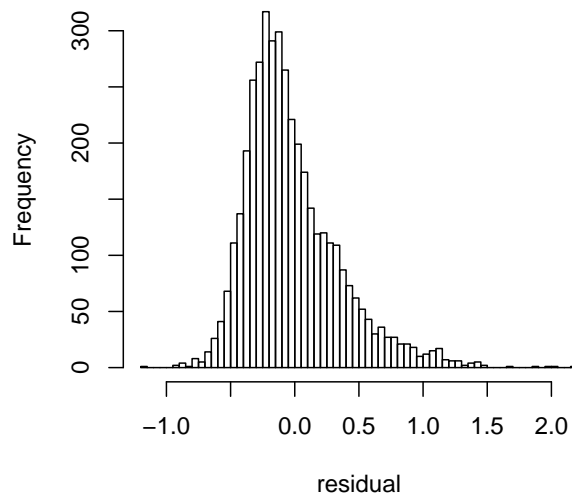


Figure 4: Regression residuals in (9), estimating the distribution of $\epsilon_t = \log R_t$.

a white noise process for (ϵ_t) ; the observed distribution of the regression residuals (Figure 4), together with their low serial correlations (-0.04 at lag 1, and no larger than 0.06 at any lag), reassure us that this is nearly enough true for accuracy in estimation. Nonetheless, there is appreciable non-normality in Figure 4, which we may wish to reflect in the stochastic process we construct.

We will take the errors ϵ_t to be independent random variables, with one of three distributions:

- (i) (AR-N) Normal, with mean zero and standard deviation (0.368) chosen to match that of the regression residuals.
- (ii) (AR-R) Re-sampled regression residuals, with all values depicted in Figure 4 being equally likely to be chosen. This gives better treatment of the non-normality.
- (iii) (AR-S) Regression residuals, re-sampled from those occurring at the same time of year as t . This allows for any non-normality to have a seasonally varying structure.

Other parametric families of distributions (e.g. the lognormal or hyperbolic) might also be considered, with parameters either constant or periodically varying to reflect seasonality. Also, the re-sampling methods could potentially benefit from the use of scenario selection to control the number of scenarios. As the goal of the present paper is to compare different inflow models, we do not attempt to use model selection procedures to determine which of these possibilities might be considered the best fit.

With the choice of distribution for ϵ_t , we have completed the description of a stochastic process modelling inflows. However, the resulting model is unlikely to be suitable as-is for use in optimization with SDDP. With distribution (i), we have a continuous model, which must be made discrete by some form of sampling. With distribution (ii), the model is already discrete, but the large number of scenarios (as many as the observations in the original data set: about 4000 for our example data) mean that some form of scenario selection will be needed to make the problem tractable. This may also be true with distribution (iii), although here there are fewer scenarios to begin with (as many as the *years* in the original data set: 79 for our example data).

3.2 Iterated function system model

The approach outlined in Section 2.2 suggests fitting models

$$Q_t = D_t + M_t Q_{t-1} \tag{11}$$

directly. The pair (D_t, M_t) should be drawn from a discrete probability distribution. That is, we need a collection of scenarios $(d_t^1, m_t^1), \dots, (d_t^k, m_t^k)$, with associated probabilities p_t^1, \dots, p_t^k , creating a discrete approximation to the conditional distribution

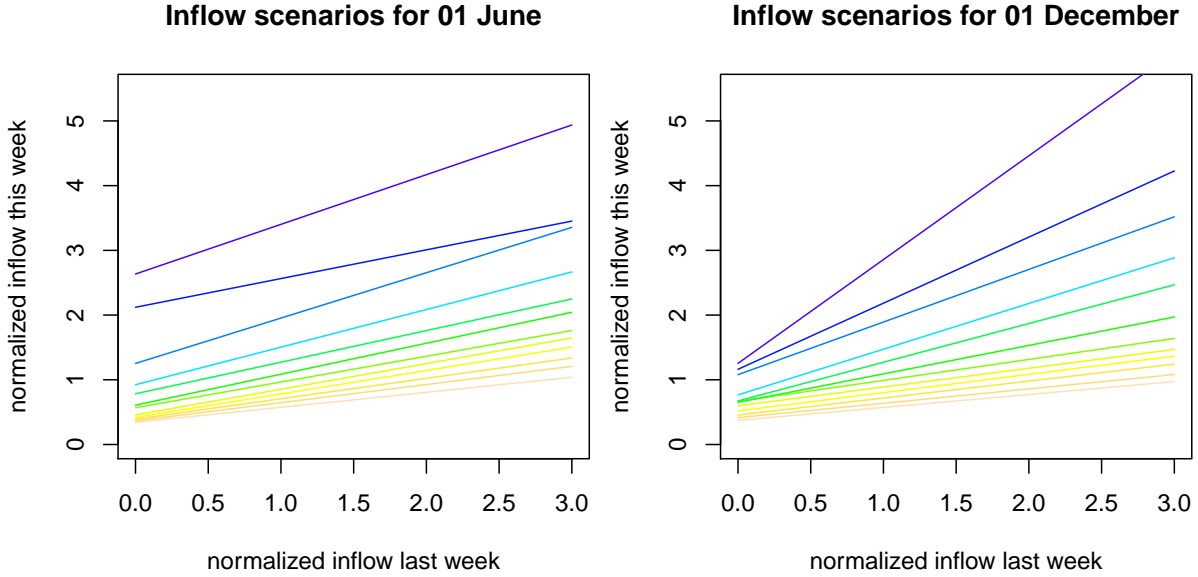


Figure 5: IFS inflow model, showing the linear dependence of the scenarios on the previous-stage inflow.

of Q_t given Q_{t-1} . The d_t^i , m_t^i , and p_t^i are all functions of time, giving a time-varying (*e.g.* seasonal) conditional dependence structure.

This can be achieved via quantile regression ([12]). Fitting (11) as a quantile regression, for a collection of k different quantiles r_1, \dots, r_k , gives us k different models $(d_t^1, m_t^1), \dots, (d_t^k, m_t^k)$. In these regressions, we take the d_t^i and m_t^i to have trigonometric polynomial form (in t), so that they (and thus the IFS model constructed from them) will have an annual seasonal variation. The regressors will thus be the lagged data (Q_{t-1}), a collection of trigonometric functions (sines and cosines) of time, and interaction terms (products) between the lagged data and the periodic functions.

Since scenario i represents outcomes falling near quantile r_i of the conditional distribution, a reasonable choice for the probabilities p_t^i is

$$p_t^i = |\{r \in [0, 1] : |r - r_i| \leq |r - r_j| \ \forall j\}|.$$

That is, $p_t^i = s_i - s_{i-1}$, where $s_0 = 0$, $s_k = 1$, and $s_i = (q_i + q_{i+1})/2$ for $i = 1, \dots, k-1$.

An appealing feature of this approach is that the stochastic model can be designed with as many or as few scenarios as desired, without resorting to any additional post-selection or sampling steps.

Figures 5 and 6 show twelve scenarios obtained by this approach, corresponding to the 2nd, 6th, 15th, 30th, 45th, 60th, 70th, 80th, 88th, 94th, 97th, and 99th percentiles of the conditional distribution of Q_t given Q_{t-1} . These quantiles were chosen to give good coverage of the distribution, with particular emphasis on the extremes. The upper tail is especially well-represented, as high-inflow events can be

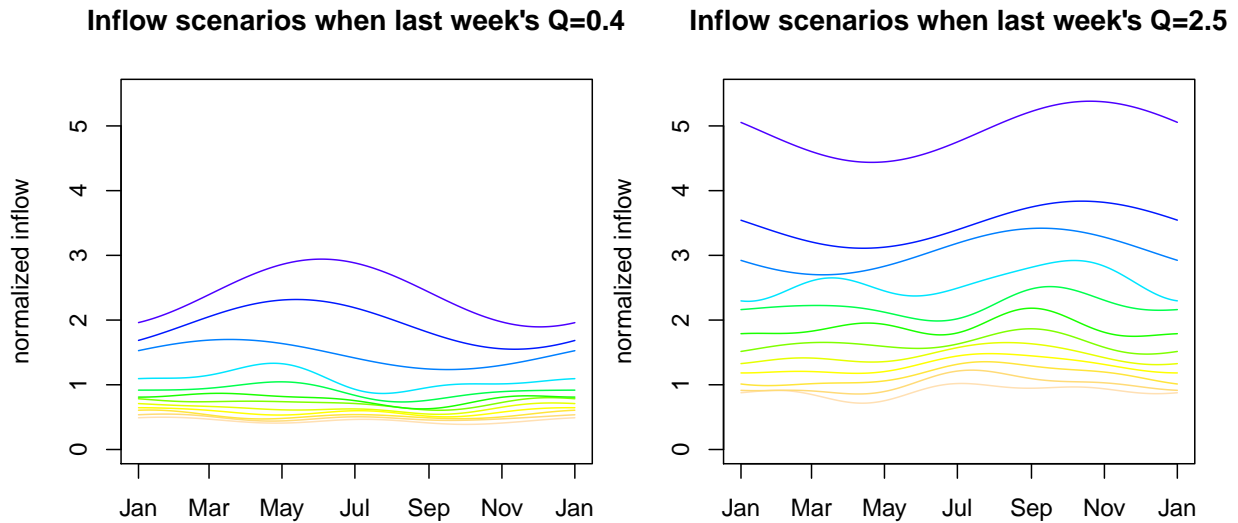


Figure 6: IFS inflow model, showing the seasonal dependence of the scenarios.

of great importance in reservoir optimization problems. Third-order trigonometric polynomials are used for the seasonal variation in the first nine scenarios. In the other three, a first-order model (*i.e.* a simple sinusoid) seems better attuned to the relative paucity of data at this upper end of the distribution.

4 Goodness of fit

The primary measure of merit in a stochastic model for inflow is whether it produces outputs which are statistically similar to the data on which it is based. In this section, we evaluate the performance of our models by simulating them and comparing the results to the original data.

There are many possible measures of statistical similarity, but the ones of most interest to us are those directly relevant to reservoir-management and hydro-scheduling optimization problems. We will consider two measures: the distribution of the total inflow over a time comparable to the characteristic time scale of the storage reservoir, and the distribution of extreme high inflows.

4.1 Monthly and quarterly flows

Consider the distribution of the total (or, equivalently, average) inflow over a time comparable to the characteristic time scale of the main storage reservoir involved. For the New Zealand hydrological system of our example data, this time scale is on the order of 1-3 months. Note that the distribution of inflow over such a period

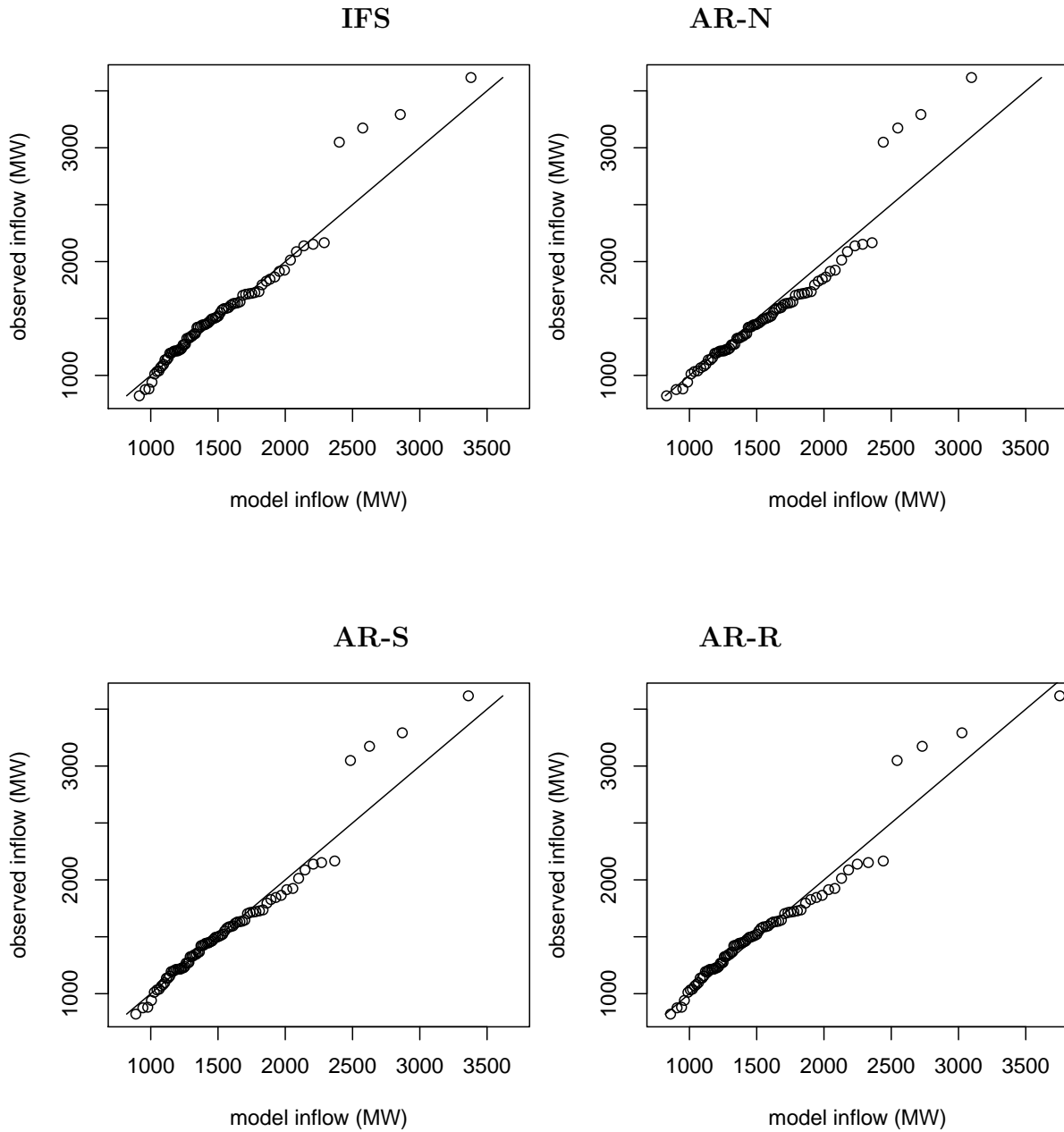


Figure 7: Quantile-quantile plot of inflow in the month of December, comparing original data to 8000-year simulations from each of four stochastic models. Clockwise from top left: IFS model, AR model with normal errors, AR model with errors re-sampled from regression residuals, AR model with errors re-sampled from seasonally appropriate regression residuals.

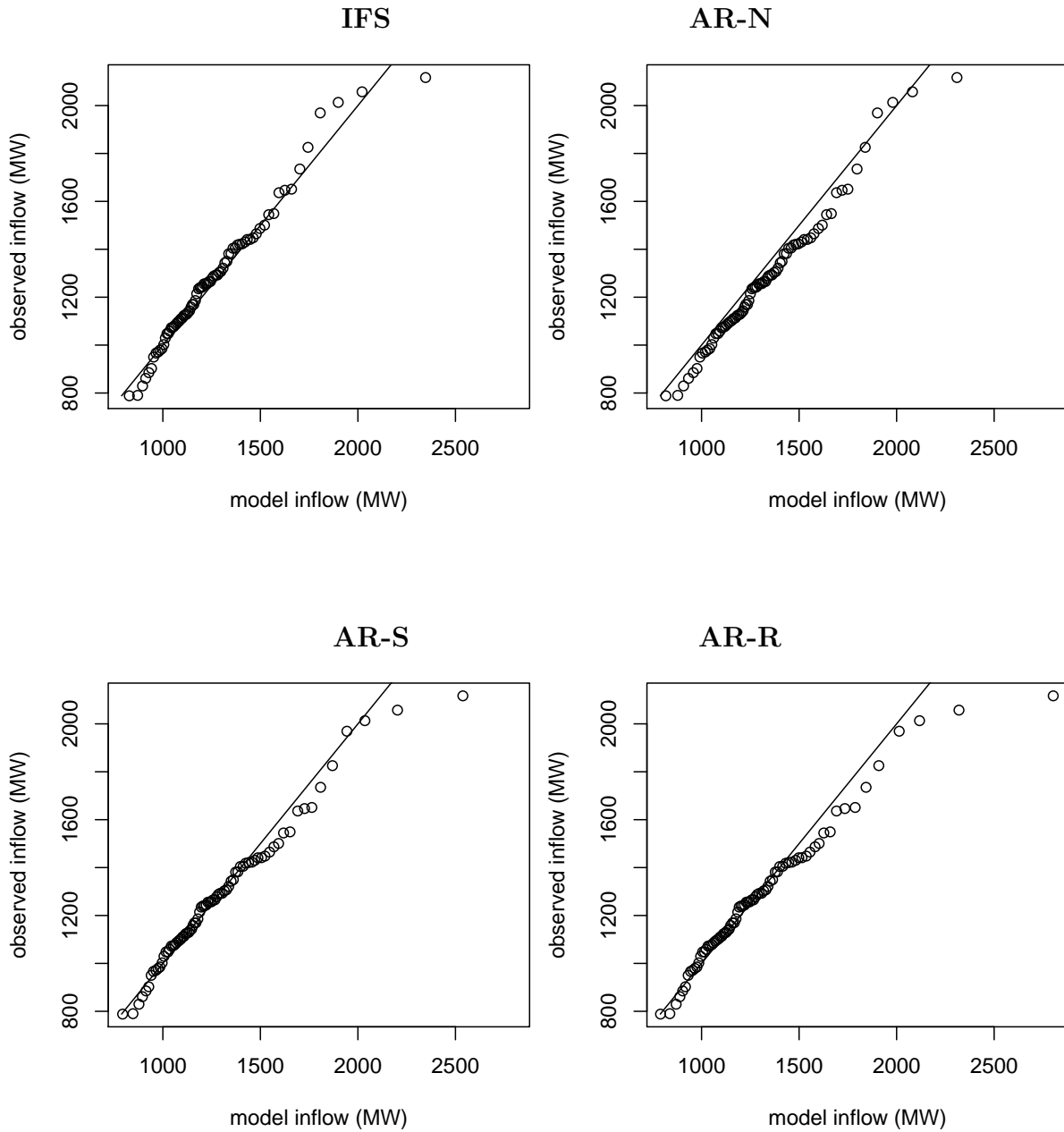


Figure 8: Quantile-quantile plot of inflow in the months of October–December, comparing original data to 8000-year simulations from each of four stochastic models. Clockwise from top left: IFS model, AR model with normal errors, AR model with errors re-sampled from regression residuals, AR model with errors re-sampled from seasonally appropriate regression residuals.

depends on both the distributions of the stage-wise (weekly) inflows and the degree of serial dependence in stage-wise inflows.

Consider, in particular, the inflow in a given calendar month or quarter. From the raw data, we have as many observations of this quantity as there are years in the data set (79 in the example), giving us a direct empirical estimate of the shape of its distribution. From a stochastic model for inflow, we can obtain an equivalent set of observations by Monte Carlo simulation, giving us an empirical estimate of the analogous distribution for the model. (This distribution would be difficult to derive in any other way). As the length of the simulation is limited only by our patience, this synthetic sample can be much larger than is available for the raw data. We can then compare the two empirical distributions via a quantile-quantile plot, with the direct observations (sorted into increasing order) plotted *vs.* the corresponding quantile in the synthetic sample.

Figures 7 and 8 show these results for our autoregressive and IFS models over a 1-month period (December) and a 3-month period (the fourth calendar quarter). In each case, the model inflows are taken from a simulation over 8000 years, or about 100 times as long as the original data series. All the models seem to do a reasonable job of reproducing the distributions of inflows over these time scales. In particular, the IFS model is not obviously handicapped by having only 12 scenarios at each stage, relative to the other models with 79, 4000, and infinitely many scenarios.

The 3-month results are particularly reassuring: having the right inflow distribution over this longer period suggests that the serial dependence of inflows has been well modelled.

4.2 Extreme high inflows

The greatest differences between the various models lie in the upper tail, *i.e.* in the way they extrapolate from the data to produce extreme high inflows. A common method for this kind of extrapolation is to fit a log-Pearson III distribution to the annual maximum flows ([6, 7, 18]). This has long been a standard technique in statistical hydrology, following its adoption by the US Water Resources Council in the 1960s ([25, 26]).

Figure 9 compares the annual maximum flows of our example data to the corresponding quantiles of a log-Pearson III distribution fitted by the mixed-moment method of Rao ([22]). The fit is quite good. The same log-Pearson III distribution is then used for each panel in Figure 10, where its quantiles are compared to the annual maxima of each of our 8000-year model simulations. As these simulated series are much longer than the original data, the highest inflows within them are more extreme than anything observed historically. It is apparent that the IFS model extrapolates to extreme high inflows in a way similar to the log-Pearson III method. The autoregressive model with normal errors produces an inflow distribution with a thinner upper tail – as we might have expected from observing the non-normality in Figure 4. Re-sampling the regression residuals, by contrast, produces relatively fat tails, suggesting perhaps that the low serial correlation among those residuals

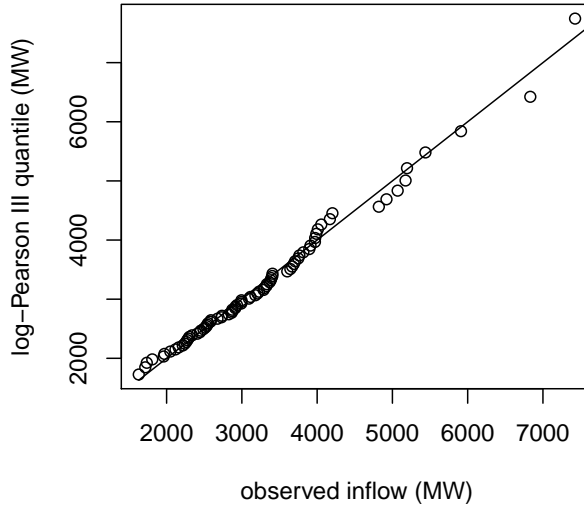


Figure 9: Annual maximum inflows for original data, plotted *vs.* quantiles of fitted log-Pearson III distribution.

belies the dependence at the extreme upper end of the distribution. The effect can be reduced, but not eliminated, by requiring the re-sampled residuals to come from the correct time of year.

5 Numerical example: hydro-scheduling

In this section we assess the behaviour of the inflow models in their intended application: a hydro-scheduling problem. We use a simple one-reservoir problem similar to that outlined in Section 1.1. The problem data are given in Table 1. The system described does not correspond to the real power generation and consumption infrastructure in the Waitaki Valley and New Zealand; all elements other than the inflow process are fictional.

The problem begins in the autumn when inflows are falling, and continues through the low-inflow winter period. Allowing the reservoir to run dry incurs a high penalty for the unserved demand, and there is a credit for stored water remaining after the last stage. The first-stage inflow is set to 500 MW (which is below average for that time of year, giving $Q_1 \approx 0.54$) and inflows evolve thereafter according to the chosen inflow model.

As noted in Section 3.1, the three versions of our autoregressive inflow model are unsuitable for immediate use in this context, as they have a large number (or in the case of normal errors, a continuous distribution) of scenarios at each stage. To put them on the same footing as our IFS model, some scenario selection is required. This was done by replacing the error distribution in each model by a set of 12 quantiles

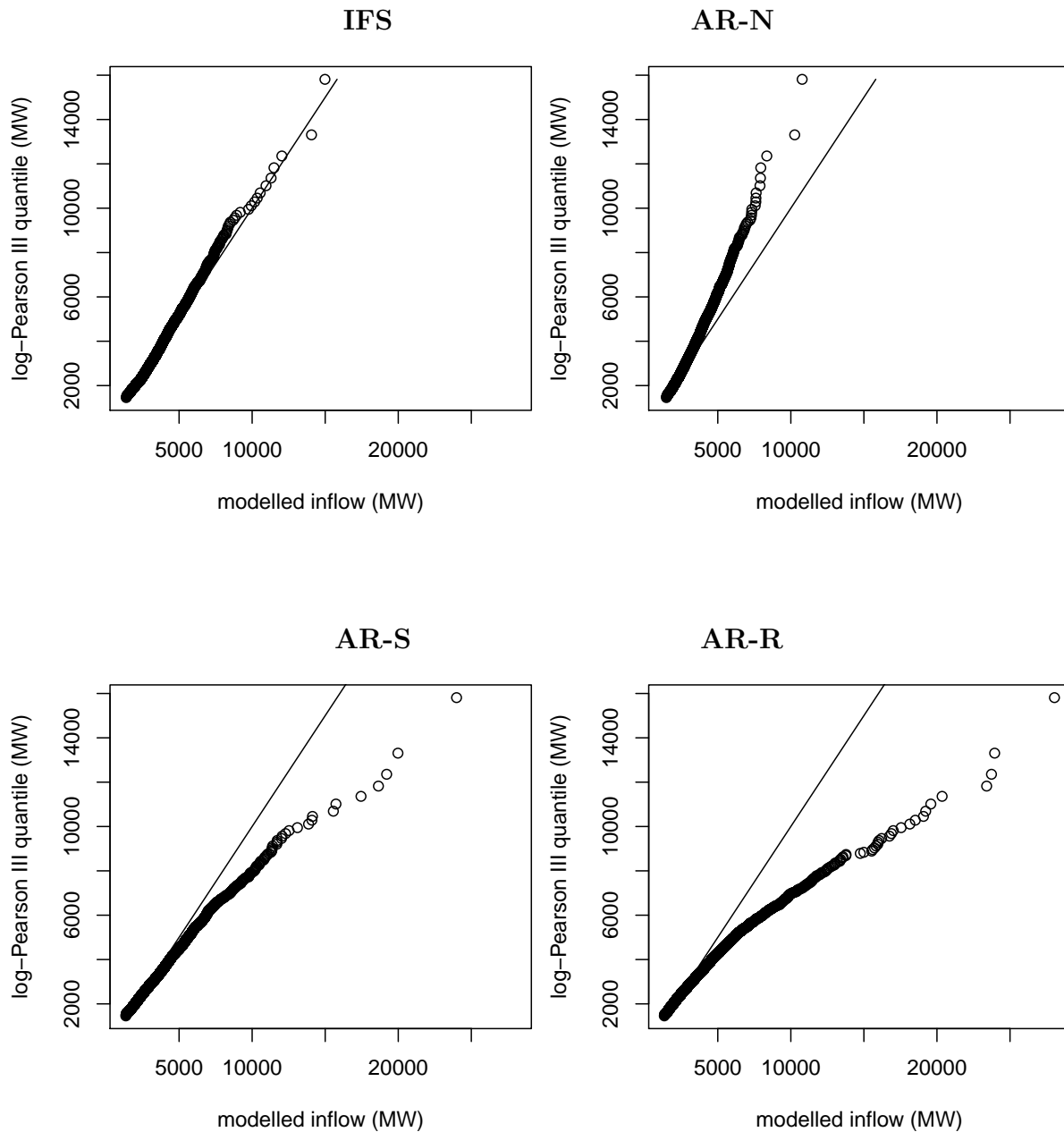


Figure 10: Distribution of annual maximum inflow for each of four stochastic models, plotted *vs.* quantiles of log-Pearson III distribution fitted to original data. Clockwise from top left: IFS model, AR model with normal errors, AR model with errors re-sampled from regression residuals, AR model with errors re-sampled from seasonally appropriate regression residuals.

Hydro generation capacity:	1500 MW
Hydro reservoir capacity:	800 GWh
Initial reservoir contents:	200 GWh
Initial (first stage) inflow:	500 MW
Thermal generation capacity:	900 MW
Thermal generation cost:	\$50/MWh
Demand:	1400 MW, constant
Cost of unserved demand:	\$1000/MWh
Stage structure:	30×7-day stages, April 1 – October 27
Residual water value after last stage:	\$150/MWh up to 84 GWh; then \$75/MWh up to 252 GWh; then \$60/MWh up to 420 GWh; then \$45/MWh up to 672 GWh; then \$25/MWh.

Table 1: Example hydro-scheduling problem data.

of that distribution: the same quantiles (and associated probabilities) as used for the IFS model in Section 3.2.

With the models in this common form, it is straight-forward to compute analytically the exact mean inflow delivered by each model over the 30 stages of the problem. The models can then be adjusted slightly so that this mean is the same for all of them. This is a small adjustment, but an important one to make when comparing inflow models of disparate kinds. Typical hydro-scheduling problems are most sensitive to the overall amount of water received, and we do not want any differences in the results to be due simply to one model running slightly wetter or drier than another. In this exercise, the inflows of the autoregressive models with normal, resampled, and seasonally-resampled errors were reduced by (respectively) 4.9%, 1.5%, and 0.5% (after the first stage) to bring them into line with the IFS model.

The problem was solved for each inflow model using Doasa 2.0, an SDDP code developed at the University of Auckland ([14]). In each case, 2500 forward and backward passes were computed. This is rather more than required to get a good solution; the primal variables have converged to within 5% after only a few hundred iterations. Running times were on the order of 10 minutes on a 32-bit x86 CPU. Each solution was then simulated over 40000 Monte Carlo sample paths to assess its performance. Common random numbers were used across inflow models, *i.e.* the initial random seed was the same for each model. Since all the models have the same number of inflow scenarios (twelve) at each stage, and the same probabilities for those scenarios, this means that corresponding scenarios were sampled.

A “wait-and-see” version of the problem, in which the inflows are visible in advance, was also solved; this is much less computationally demanding, as it is merely a linear program. The same common inflow scenarios as in the SDDP problem were

	IFS	AR-N	AR-R	AR-S
<i>Optimization:</i>				
Wait-&-see lower bound (\$m)	155.12 \pm 0.34	156.41 \pm 0.40	158.19 \pm 0.44	158.34 \pm 0.43
SDDP lower bound (\$m)	155.71	156.80	159.23	159.33
<i>Simulation:</i>				
Objective (\$m)	155.86 \pm 0.33	157.13 \pm 0.39	159.54 \pm 0.43	159.82 \pm 0.42
Mean shortage (MW)	0.276 \pm 0.025	0.530 \pm 0.040	0.705 \pm 0.044	0.697 \pm 0.043
P (shortage)	0.022 \pm 0.001	0.031 \pm 0.002	0.041 \pm 0.002	0.044 \pm 0.002
Mean spill (MW)	3.00 \pm 0.21	3.18 \pm 0.27	8.61 \pm 0.58	9.00 \pm 0.50
P (spill)	0.053 \pm 0.002	0.044 \pm 0.002	0.089 \pm 0.003	0.101 \pm 0.003
Mean energy price (\$/MWh)	65.42 \pm 0.45	71.68 \pm 0.52	81.55 \pm 0.65	82.10 \pm 0.70

Table 2: Example hydro-scheduling problem results, for four different inflow models. Objective values are in millions of dollars. Uncertainties shown are two times the standard error.

used. The expected optimal objective value for this problem serves as a lower bound for the objective value of the corresponding SDDP problem.

Table 2 shows the results. It is evident that in certain key respects the hydro-scheduling problem is very sensitive to small differences among inflow models; this is not atypical for such problems, and highlights the importance of detailed inflow modelling. The inflow models with resampling are particularly prone to spilling (reservoir overflow); this is consistent with the results on high inflow depicted in Figure 10. The risk of shortage is reflected in the energy prices, and these too vary substantially among the inflow models, even though there is little apparent difference among them at the low-inflow end of the distribution.

6 Conclusion

Inflow models of IFS type, which are highly compatible with the SDDP algorithm, appear to be capable of representing inflow processes in at least two important respects. Firstly, the distribution of inflow over problem-relevant time periods (several problem stages, or a time similar to the characteristic time scale of the reservoir) is reproduced with reasonable accuracy. This is also true over somewhat longer periods, suggesting that the all-important serial dependence of inflows has been well modelled. Secondly, extreme high inflows are generated in a way consistent with the standard log-Pearson III extrapolation technique.

The general form of SDDP-compatible linear inflow model (5) should be workable in principle for problems with multivariate inflow processes. However, the present paper has considered only univariate inflows. A suitable method for fitting an IFS-type model to multivariate inflow series has yet to be devised; this seems a promising area for future research.

7 Acknowledgements

The author acknowledges the contribution of the anonymous referees to this paper.

References

- [1] M. Barnsley, *Fractals Everywhere*, 2nd ed. (1989), Academic Press, New York.
- [2] M. Barnsley, S. Demko, Iterated function systems and the global construction of fractals, *Proceedings of the Royal Society of London* 399 (1985), 243–275.
- [3] P. Diaconis, D. Freedman, Iterated random functions, *SIAM Review* 41 (1999), 45–76.
- [4] B. Forte, E. R. Vrscay, Solving the inverse problem for measures using iterated function systems: a new approach, *Advances in Applied Probability* 27 (1995), 800–820.
- [5] P. Girardeau, V. Leclere, A.B. Philpott, On the convergence of decomposition methods for multistage stochastic convex programs, *Mathematics of Operations Research* 40 (2015), 130–145.
- [6] V. W. Griffis, J. R. Stedinger, Log-Pearson type 3 distribution and Its application in flood frequency analysis. I: Distribution characteristics, *J. Hydrologic Engineering* 12 (2007), 482–491.
- [7] V. W. Griffis, J. R. Stedinger, Log-Pearson type 3 distribution and its application in flood frequency analysis. II: Parameter estimation methods, *J. Hydrologic Engineering* 12 (2007), 492–500.
- [8] V. Guigues, SDDP for some interstage dependent risk-averse problems and application to hydro-thermal planning, *Computational Optimization and Applications* 57 (2014), 167–203.
- [9] V. Guigues, Convergence analysis of sampling-based decomposition methods for risk-averse multistage stochastic convex programs, *Optimization Online*.
- [10] V. Guigues, W. Römis, Sampling-based decomposition methods for multi-stage stochastic programs based on extended polyhedral risk measures, *SIAM J. Optimization* 22 (2012), 286–312.
- [11] V. Guigues, W. Römis, SDDP for multistage stochastic linear programs based on spectral risk measures. *Operations Research Letters* 40 (2012), 313–318.
- [12] R. Koenker, *Quantile regression*, (2005), Cambridge University Press, New York.

- [13] D. La Torre, E. R. Vrscay, Fractal-based measure approximation with entropy maximization and sparsity constraints, in *Bayesian inference and maximum entropy methods in science and engineering* (2012), AIP, 63–71.
- [14] A.B. Philpott, Z. Guan, On the convergence of stochastic dual dynamic programming and related methods, *Operations Research Letters* 36 (2008), 450–455.
- [15] M.V.F. Pereira, L.M.V.G. Pinto, Stochastic optimization of a multireservoir hydroelectric system – a decomposition approach, *Water Resources Research* 21 (1985), 779–792.
- [16] M.V.F. Pereira, L.M.V.G. Pinto, Multi-stage stochastic optimization applied to energy planning, *Mathematical Programming* 52 (1991), 359–375.
- [17] M. V. F. Pereira, G. C. Oliveira, C. C. G. Costa, J. Kelman, Stochastic stream-flow models for hydroelectric systems, *Water Resources Research* 20 (1984), 379–390.
- [18] H. N. Phien, T. J. Ajirajah, Applications of the log Pearson type-3 distribution in hydrology, *J. Hydrology* 73 (1984), 359–372.
- [19] A. Philpott, V. de Matos, Dynamic sampling algorithms for multi-stage stochastic programs with risk aversion, *European J. Operational Research* 218 (2012), 470–483.
- [20] J. M. Purdie, W. E. Bardsley, Seasonal prediction of lake inflows and rainfall in a hydro-electricity catchment, Waitaki River, New Zealand, *International J. Climatology* 30 (2010), 372–389.
- [21] R Development Core Team, *R: A Language and Environment for Statistical Computing* (2009), <http://www.R-project.org>.
- [22] D. V. Rao, Estimating log Pearson parameters by mixed moments, *J. Hydraulic Engineering - ASCE* 109 (1983), 1118–1132.
- [23] A. Shapiro, Analysis of Stochastic Dual Dynamic Programming method, *European J. Operational Research* 209 (2011), 63–72.
- [24] A. Shapiro, W. Tekaya, J. P. da Costa, M. P. Soares, Risk neutral and risk averse Stochastic Dual Dynamic Programming method, *European J. Operational Research* 224 (2013), 375–391.
- [25] United States Water Resources Council, *A uniform technique for determining flood flow frequencies*, Bulletin 15 (1967), U.S. Water Resources Council Hydrology Committee, Washington, D.C.
- [26] United States Water Resources Council, *Guidelines for determining flood flow frequency*, Bulletin 17B (1982), U.S. Water Resources Council Hydrology Committee, Washington, D.C.

- [27] J. Yang, Multiple description wavelet-based image coding using iterated function system, *Mathematical Problems in Engineering* 2013 (2013), article ID 924274.