Regression models for within-year capacity adjustment in reservoir planning

A. J. ADELOYE
School of the Built Environment, Heriot-Watt University, Edinburgh EH14 4AS, UK
a.j.adeloye@hw.ac.uk

F. LALLEMAND
22 Allée des Jonquilles, F-83320 Carqueiranne, France
frlallemand@yahoo.com

T. A. McMAHON
Department of Civil and Environmental Engineering, University of Melbourne, Victoria 3010, Australia
t.mcmahon@civag.unimelb.edu.au

Abstract Streamflow data from 12 international rivers were used to develop predictive relationships for total (i.e. within-year plus over-year) reservoir capacity as well as the within-year capacity adjustment for use during reservoir planning. The models were then validated using data from three other international rivers. In general, it was found that the performance, during both calibration and validation, of the combined capacity model was better than that of the within-year capacity adjustment model. The two models are applicable only if an estimate of the over-year capacity is available. Methods of independently estimating the over-year capacity, as well as the other derived factors used as independent variables in the regression models, are described for gauged and ungauged sites.

Key words reservoir planning; sequent peak algorithm (SPA); over-year capacity; within-year capacity; storage–yield performance; storage–yield relationships; Gould-Gamma model; regression models; gauged sites; ungauged sites

INTRODUCTION

Situations often arise during water resources planning in which the development of reservoir storage–yield performance (SYP) characteristics through analysis of at-site data is considered to be either too detailed for the purpose, or infeasible. An example
of the former situation concerns the initial screening of potential reservoir sites, where a rapid analysis to show the relative advantages of the respective sites in terms of their storage–yield function is all that is required. Once the screening has narrowed down the options to a smaller number of sites, these sites can then be investigated using more detailed analysis techniques such as simulation. Obtaining SYP information at ungauged sites is a clear example of the latter situation.

The above needs for rapid SYP methodologies have led to the development of generalized storage–yield performance relationships. By being generalized, these relationships are meant to produce answers to the storage–yield problem using input factors such as the coefficient of variation ($C_v$) of streamflow, the demand, system reliability and perhaps a knowledge of the probability distribution of the inflows at the site. An example of such generalized relationships is the Gould-Gamma model (see McMahon, 1993; Vogel & McMahon, 1996):

$$K_A = \frac{z_g^2}{4(1-D)}(C_v)^2$$

$$z_g = \frac{2}{\gamma}\left[1 + \frac{\gamma}{6}\left(z_f - \frac{\gamma}{6}\right)^3\right] - 1$$

$$z_f = \left(1 - f\right)^{0.135}/0.1975$$

where $K_A$ is the over-year capacity as ratio of mean annual flow; $D$ is the demand as ratio of mean annual flow; $C_v$ is the coefficient of variation of annual flow; $z_f$ is the standardized normal variate of $f$ probability (decimal) of non-exceedence; $(1 - f)$ is the annual time-based reliability; $z_g$ is the equivalent standardized gamma variate; and $\gamma$ is the skew coefficient of annual flow.

However, the Gould-Gamma model (equation (1)) and all the other generalized relationships developed so far have been limited to systems characterized by over-year behaviours; where a reservoir has significant within-year effects, the currently available generalized SYP relationships will underestimate the capacity requirements. Generalizing the over-year SYP relationship has been relatively straightforward because of the few variables involved. For example, with respect to the probability distribution of the flows, only three parameters—the mean, $C_v$ and skew coefficient—are often sufficient to describe the annual flows. Besides, these three parameters can be selected arbitrarily to cover most of the world’s streamflow regimes; hence there is no need to base the analysis on any particular river basin. Indeed, all the attempts that have tackled the over-year SYP generalization problem have relied on such arbitrary, albeit plausible, parameter space to drive the analysis. However, attempting to include within-year consideration in such a framework will be fraught with enormous difficulties because of the much larger number of variables involved. One way which obviates this potential difficulty is to adjust the over-year capacity for within-year effects. This has been attempted by Hardison (1965) (see also McMahon & Mein, 1986), but because these studies were limited to USA streams, the applicability of the approach is limited. Other within-year adjustment approaches applicable over a much wider range of flow characteristics are therefore required.
Consequently, this study has developed predictive regression equations for within-year capacity adjustment by carrying out SYP analysis of historic monthly and annual flow data from 12 international catchments located in four continents: North America (USA); Oceania (Australia); Europe (United Kingdom) and Africa (South Africa). For each catchment, reservoir capacity for meeting specified demands and time-based reliability targets over the respective historic records was determined using the modified sequent peak algorithm, SPA (Adeloye et al., 2001). The analysis used first the annual data and then the monthly data, thus enabling the within-year effects to be determined. The capacity estimates were then regressed on flow and reservoir characteristics to develop the predictive models; the models were later validated at three independent sites.

**METHODOLOGY**

**Flows records**

The study used monthly runoff data records from 15 international rivers summarized in Table 1. The first 12 sites were used for model calibration while the remaining three were used for validation. In general, the data records are at least 30 years long with the catchments ranging in size from 101 to 19,654 km². The 15 streams represent a wide range of annual Cv (0.19–1.47), which covers most conditions in the world (McMahon et al., 1992). Furthermore, most of the 15 monthly flow regimes of the world as defined by Haines et al. (1988) are represented by the data set. Thus the data set offers a unique opportunity for studying how the combined within-year and over-year storage

<table>
<thead>
<tr>
<th>Site</th>
<th>River</th>
<th>Gauging station</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Country</th>
<th>Catchment area (km²)</th>
<th>Mean annual flow ((x 10^6 \text{ m}^3))</th>
<th>Mean annual runoff (mm)</th>
<th>Cv</th>
<th>Years (and span) of data record</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Renoster</td>
<td>Koppies Dam</td>
<td>−27.25</td>
<td>27.68</td>
<td>South Africa</td>
<td>2196</td>
<td>112.36</td>
<td>51.17</td>
<td>0.991</td>
<td>40 (1920–1959)</td>
</tr>
<tr>
<td>3</td>
<td>Vis</td>
<td>Harderug Dam</td>
<td>−31.82</td>
<td>20.37</td>
<td>South Africa</td>
<td>1463</td>
<td>18.52</td>
<td>12.66</td>
<td>1.004</td>
<td>33 (1927–1959)</td>
</tr>
<tr>
<td>4</td>
<td>Brak</td>
<td>Bellair Dam</td>
<td>−33.72</td>
<td>20.6</td>
<td>South Africa</td>
<td>546</td>
<td>2.28</td>
<td>4.18</td>
<td>1.072</td>
<td>40 (1920–1959)</td>
</tr>
<tr>
<td>5</td>
<td>Homochitto</td>
<td>Eddiceton Dam</td>
<td>31.50</td>
<td>−90.78</td>
<td>USA</td>
<td>466.2</td>
<td>238.17</td>
<td>510.88</td>
<td>0.395</td>
<td>46 (1938–1983)</td>
</tr>
<tr>
<td>6</td>
<td>Dee</td>
<td>Erbiston Rectory</td>
<td>52.92</td>
<td>−2.97</td>
<td>UK</td>
<td>1040</td>
<td>1000.26</td>
<td>961.79</td>
<td>0.201</td>
<td>32 (1938–1969)</td>
</tr>
<tr>
<td>7</td>
<td>Ongaparinga</td>
<td>Clarendon Weir</td>
<td>−35.12</td>
<td>138.63</td>
<td>Australia</td>
<td>445</td>
<td>81.47</td>
<td>183.08</td>
<td>0.683</td>
<td>69 (1869–1936)</td>
</tr>
<tr>
<td>8</td>
<td>Werribee</td>
<td>Ballan Weir</td>
<td>−37.60</td>
<td>144.23</td>
<td>Australia</td>
<td>101</td>
<td>21.49</td>
<td>212.77</td>
<td>0.707</td>
<td>30 (1944–1973)</td>
</tr>
<tr>
<td>9</td>
<td>Earn</td>
<td>Kinkell Bridge</td>
<td>56.33</td>
<td>−3.67</td>
<td>UK</td>
<td>590.5</td>
<td>648.29</td>
<td>1097.87</td>
<td>0.189</td>
<td>34 (1949–1982)</td>
</tr>
<tr>
<td>10</td>
<td>Hatchie</td>
<td>Bolivar Dam</td>
<td>35.28</td>
<td>−88.98</td>
<td>USA</td>
<td>3833.2</td>
<td>2178.40</td>
<td>568.30</td>
<td>0.363</td>
<td>55 (1930–1984)</td>
</tr>
<tr>
<td>11</td>
<td>Richmond</td>
<td>Casino Dam</td>
<td>−28.87</td>
<td>153.03</td>
<td>Australia</td>
<td>1790</td>
<td>699.46</td>
<td>390.76</td>
<td>0.653</td>
<td>41 (1944–1984)</td>
</tr>
<tr>
<td>12</td>
<td>Paria</td>
<td>Lees Ferry</td>
<td>36.87</td>
<td>−111.59</td>
<td>USA</td>
<td>3651.9</td>
<td>26.76</td>
<td>7.33</td>
<td>0.404</td>
<td>61 (1923–1983)</td>
</tr>
<tr>
<td>13</td>
<td>Prins</td>
<td>Prins River Dam</td>
<td>−33.52</td>
<td>20.75</td>
<td>South Africa</td>
<td>761</td>
<td>3.85</td>
<td>5.06</td>
<td>1.47</td>
<td>43 (1917–1959)</td>
</tr>
<tr>
<td>14</td>
<td>Mulgrave</td>
<td>Gordonvale</td>
<td>−17.10</td>
<td>145.8</td>
<td>Australia</td>
<td>554</td>
<td>886.34</td>
<td>1599.89</td>
<td>0.51</td>
<td>36 (1918–1953)</td>
</tr>
<tr>
<td>15</td>
<td>Coruh</td>
<td>Karsikoy</td>
<td>41.45</td>
<td>41.73</td>
<td>Turkey</td>
<td>1965.4</td>
<td>5922.41</td>
<td>301.33</td>
<td>0.26</td>
<td>39 (1943–1981)</td>
</tr>
</tbody>
</table>
capacities vary in general with annual and seasonal streamflow variability. Vogel et al. (1999) carried out a similar activity for streams in the USA.

**Storage–yield performance analysis**

The modified SPA storage–yield performance simulations were implemented with MS Excel. The performance index considered was the time-based reliability over the historic record periods. However, since this index differs depending on whether annual or monthly data are used (Nawaz & Adeloye, 1999), it was important to ensure that, before comparing them, capacities derived using monthly data and annual data relate to the same time-based reliability.

Time-based reliability computed on the basis of annual simulation is different from that computed from monthly data simulation, because, in the annual analysis, each failure month is not recognized in its own right. So, whether one, two, or all 12 months in a given year failed in a monthly simulation, they all constitute a single failure when computing the annual time-based reliability. To illustrate this, assume in an annual simulation that there is only one failure year in a $T$-year record, because the flow in that year is extremely low. The annual time-based reliability is $(T - 1)/T$. Assuming that a monthly simulation is then carried out, resulting in $N$ failed months, then the corresponding monthly time-based reliability will be $(12T - N)/12T$, which can be simplified into $(T - N/12)/T$. Obviously, except when $N$ is twelve, $N/12$ is less than unity and the monthly time-based reliability will be higher than the annual time-based reliability.

The way this problem was resolved in this work was to restrict consideration to the annual reliability and to convert the monthly reliability to the equivalent annual reliability before comparing the capacity estimates. Thus, when a monthly SPA simulation is completed, the Excel spreadsheet is examined to see how the failure months are distributed among the years in the record; this distribution is then used to determine the equivalent annual reliability. For example, assume the design is for four failure months. Then if all four failure months occur in the same year, that is taken to be equivalent to one failure year and the estimated capacity from the monthly analysis is compared with the capacity based on annual simulation with one failure year. Similarly, if the failure months are contained within two years, then the monthly storage estimate is compared with the annual estimate for two failure years, and so on.

For the analysis, demand ratios from 0.1 to 0.9 with a step of 0.1 were considered. For the monthly analysis, the total annual demand was assumed equally distributed among the 12 months. The analysis considered annual time-based reliabilities of 90–100%. When using the modified SPA to design for failures, the volumetric shortfall during the failures can also be pre-specified. This was fixed at 25% of the demand throughout.

**RESULTS**

**SYP functions**

For the no-failure situations, the storage–yield functions based on annual and monthly simulations are shown in Fig. 1 for all the 12 rivers. Lallemand (2001) has developed similar functions for various failure targets, but these have not been presented here for lack of space. For each of the rivers in Fig. 1, both the over-year (based on annual
Regression models for within-year capacity adjustment in reservoir planning
simulation) and the total (i.e. within-year plus over-year; based on monthly simulation) capacities have been plotted on the same graph. This helps to see how the annual $C_v$ influences the within-year component of the total capacity estimate for a reservoir. In general, it is clear from the plots in Fig. 1 that storage–yield functions based on monthly data analyses are always to the right of those based on annual data analyses. In other words, for a given demand ratio, the storage capacity obtained using monthly analyses exceeds that obtained using the annual data. This is to be expected because monthly analysis produces the combined within-year and over-year storage capacity, or total capacity, whereas the annual analysis merely produces the over-year capacity.

The difference between the total capacity and the over-year capacity is the within-year capacity. Although this may not be immediately obvious in Fig. 1, because the scales of the capacity axes are not similar, this difference appears to increase as the $C_v$ of annual flows reduces. Systems with low $C_v$s tend to be dominated by within-year behaviour, particularly at low demand ratios, whereas systems with high $C_{vs}$ tend to be dominated by over-year behaviour. Indeed, for most of the low variability streams, e.g. $C_v \leq 0.4$, there is no over-year storage requirement below a demand ratio, $D$, of 0.4, whereas these same rivers exhibit significant within-year requirement below this demand ratio threshold. However, as the demand increases for these low variability streams, the total and over-year capacity functions gradually converge, implying that the influence of within-year on total capacity reduces as the demand ratio exceeds 0.4.

For the high variability streams, i.e. $C_v > 0.6$, however, there is a non-zero over-year capacity even for a demand ratio as low as 0.1. Indeed at this low demand ratio, the total and over-year capacity curves appear almost indistinguishable, implying that within-year effects are minimal for $D \leq 0.1$. As the demand ratio increases, however, there is a discernible within-year contribution, but this is still small (relative to the over-year capacity) and remains constant irrespective of $D$. As a consequence, both the total capacity and over-year capacity curves tend to run parallel to each other as $D$ increases. Therefore, in cases of high variability streams it would appear that the demand level plays no significant role in influencing the behaviour of the reservoir, which is dominated throughout by over-year effects.

Thus a tentative conclusion arising from this study is that a reservoir will behave essentially as within-year if the annual $C_v < 0.4$ and $D < 0.4$. Conversely, a reservoir will be dominated by over-year behaviour if $C_v < 0.4$ and $D > 0.6$, or if $C_v > 0.6$ whatever the demand ratio. Situations outside this $C_v$–$D$ space will be expected to exhibit a uniform mix of the two behaviour modes. Montaseri & Adeloye (1999) noted that knowing whether a reservoir will behave as within-year or over-year prior to analysis is advantageous because it aids in the selection of an appropriate temporal scale for the analysis. For example, if a reservoir is demonstrably over-year, then analysis can be based on annual data series whereas shorter-term data must be used if the reservoir is within-year.

To reinforce further the significance of within-year at various $C_v$ and $D$ values, Table 2 has been prepared, which shows the $K_T/K_A$ ratio, where $K_T$ is the total capacity and $K_A$ is the over-year capacity. A ratio $\gg 1$ implies significant within-year whereas a ratio close to unity implies significant over-year. As seen in Table 2, the $K_T/K_A$ ratio is close to unity for the high variability streams irrespective of the demand ratio. However, for the low variability streams the ratio is very much above unity at the demand ratio of 0.4 but very much less so as the demand ratio exceeds 0.4.
Table 2 $K_T/K_d$ ratios for demand ratios $D = 0.4, 0.6$ and $0.8$.  

<table>
<thead>
<tr>
<th>$Cv$</th>
<th>Yield ratio, $D$</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.072</td>
<td>1.14</td>
<td>1.14</td>
<td>1.08</td>
<td></td>
</tr>
<tr>
<td>1.012</td>
<td>1.82</td>
<td>1.47</td>
<td>1.19</td>
<td></td>
</tr>
<tr>
<td>1.004</td>
<td>1.42</td>
<td>1.28</td>
<td>1.25</td>
<td></td>
</tr>
<tr>
<td>0.991</td>
<td>1.17</td>
<td>1.09</td>
<td>1.07</td>
<td></td>
</tr>
<tr>
<td>0.707</td>
<td>1.22</td>
<td>1.30</td>
<td>1.17</td>
<td></td>
</tr>
<tr>
<td>0.694</td>
<td>1.34</td>
<td>1.35</td>
<td>1.24</td>
<td></td>
</tr>
<tr>
<td>0.653</td>
<td>1.88</td>
<td>1.27</td>
<td>1.27</td>
<td></td>
</tr>
<tr>
<td>0.404</td>
<td>6.51</td>
<td>1.50</td>
<td>1.16</td>
<td></td>
</tr>
<tr>
<td>0.395</td>
<td>2.62</td>
<td>1.74</td>
<td>1.21</td>
<td></td>
</tr>
<tr>
<td>0.363</td>
<td>69.83</td>
<td>1.53</td>
<td>1.26</td>
<td></td>
</tr>
<tr>
<td>0.201</td>
<td>$\sim\infty$</td>
<td>2.55</td>
<td>1.32</td>
<td></td>
</tr>
<tr>
<td>0.189</td>
<td>$\sim\infty$</td>
<td>$\sim\infty$</td>
<td>2.80</td>
<td></td>
</tr>
</tbody>
</table>

Development of regression equations for within-year capacity adjustment

Correlation analysis To establish the likely significant variables to be used as independent variables, a correlation analysis was carried out using the variables that have been shown by various studies to influence reservoir capacity. These include the time-based reliability, the coefficient of variation of annual flow, the demand and the length of data record (see Montaseri & Adeloye, 1999), and their logarithmic transformations. Where it is known that a given variable might be zero, one is added to that variable to eliminate numerical overflow problems when logarithms are taken. A summary of the variables and their meanings are shown in Table 3.

The SPA simulations for the various demands and time-based reliability targets considered resulted in 432 cases, all of which were used in the correlation analysis. The resulting correlation matrix is shown in Table 4. As expected, some of the correlation coefficients are very low, e.g. between $Cv$ and $D$, for the simple reason that there

Table 3 Summary of the variables and their meanings.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Cv$</td>
<td>Coefficient of variation of annual flow</td>
</tr>
<tr>
<td>$D$</td>
<td>Demand (ratio of mean annual flow)</td>
</tr>
<tr>
<td>$R$</td>
<td>Annual time-based reliability (%)</td>
</tr>
<tr>
<td>$K_d$</td>
<td>Over-year capacity (ratio of mean annual flow)</td>
</tr>
<tr>
<td>$K_T$</td>
<td>Total capacity (ratio of mean annual flow)</td>
</tr>
<tr>
<td>$D$</td>
<td>Within-year capacity ($10^6 m^3$)</td>
</tr>
<tr>
<td>$d/M$</td>
<td>Within-year capacity (ratio of mean annual flow)</td>
</tr>
<tr>
<td>$M$</td>
<td>Standardized net inflow parameter ($= (1 – D)/Cv$)</td>
</tr>
<tr>
<td>$L$</td>
<td>Length of record (years)</td>
</tr>
<tr>
<td>$\log(Cv)$</td>
<td>$\log(D)$</td>
</tr>
<tr>
<td>$\log(D)$</td>
<td>$\log((1 + d)/Cv)$</td>
</tr>
<tr>
<td>$\log((1 + d)/M)$</td>
<td>$\log((1 + d/M)$</td>
</tr>
<tr>
<td>$\log((1 + m)$</td>
<td>$\log((1 + K_d)$</td>
</tr>
<tr>
<td>$\log((1 + K_T)$</td>
<td>$\log((1 + K_T)$</td>
</tr>
<tr>
<td>$\log(L)$</td>
<td>$\log(L)$</td>
</tr>
</tbody>
</table>
Table 4 Correlation matrix for variables investigated.

<table>
<thead>
<tr>
<th></th>
<th>$C_V$</th>
<th>$D$</th>
<th>$R$</th>
<th>$d$</th>
<th>$d/M$</th>
<th>$m$</th>
<th>$L$</th>
<th>$\log C_V$</th>
<th>$\log D$</th>
<th>$\log 1p d$</th>
<th>$\log 1p d/M$</th>
<th>$\log 1p m$</th>
<th>$\log 1p L$</th>
<th>$K_A$</th>
<th>$K_T$</th>
<th>$\log 1p K_A$</th>
<th>$\log 1p K_T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_V$</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D$</td>
<td></td>
<td>2.39E-18</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R$</td>
<td>-0.037</td>
<td>-1.7E-17</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$d$</td>
<td>-0.301</td>
<td>0.292</td>
<td>0.026</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$d/M$</td>
<td>0.451</td>
<td>0.700</td>
<td>-0.124</td>
<td>0.174</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m$</td>
<td>-0.650</td>
<td>-0.571</td>
<td>-0.035</td>
<td>-0.087</td>
<td>-0.641</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$L$</td>
<td>-0.167</td>
<td>7.4E-18</td>
<td>0.285</td>
<td>0.157</td>
<td>-0.084</td>
<td>-0.094</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\log C_V$</td>
<td>0.973</td>
<td>0</td>
<td>0.002</td>
<td>-0.259</td>
<td>0.448</td>
<td>-0.708</td>
<td>-0.0271</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\log D$</td>
<td>-3.5E-18</td>
<td>0.955</td>
<td>-4.7E-18</td>
<td>0.275</td>
<td>0.669</td>
<td>-0.545</td>
<td>3.80E-18</td>
<td>3.4E-18</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\log 1p d$</td>
<td>-0.484</td>
<td>0.411</td>
<td>0.008</td>
<td>0.684</td>
<td>0.175</td>
<td>-0.014</td>
<td>0.095</td>
<td>-0.446</td>
<td>0.440</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\log 1p d/M$</td>
<td>0.453</td>
<td>0.728</td>
<td>-0.119</td>
<td>0.192</td>
<td>0.995</td>
<td>-0.677</td>
<td>-0.074</td>
<td>0.454</td>
<td>0.706</td>
<td>0.201</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\log 1p m$</td>
<td>-0.676</td>
<td>-0.654</td>
<td>-0.017</td>
<td>-0.071</td>
<td>-0.735</td>
<td>0.972</td>
<td>-0.033</td>
<td>-0.714</td>
<td>-0.603</td>
<td>0.011</td>
<td>-0.765</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\log 1p L$</td>
<td>-0.152</td>
<td>-1.3E-18</td>
<td>0.290</td>
<td>0.183</td>
<td>-0.087</td>
<td>-0.106</td>
<td>0.994</td>
<td>-0.013</td>
<td>-5.9E-18</td>
<td>0.111</td>
<td>-0.076</td>
<td>-0.043</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$K_A$</td>
<td>0.510</td>
<td>0.656</td>
<td>0.027</td>
<td>-0.040</td>
<td>0.751</td>
<td>-0.551</td>
<td>-0.124</td>
<td>0.484</td>
<td>0.561</td>
<td>-0.069</td>
<td>0.741</td>
<td>-0.692</td>
<td>-0.115</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$K_T$</td>
<td>0.522</td>
<td>0.690</td>
<td>-0.023</td>
<td>-0.008</td>
<td>0.826</td>
<td>-0.590</td>
<td>-0.121</td>
<td>0.500</td>
<td>0.602</td>
<td>-0.032</td>
<td>0.816</td>
<td>-0.729</td>
<td>-0.114</td>
<td>0.991</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\log 1p K_A$</td>
<td>0.554</td>
<td>0.711</td>
<td>0.035</td>
<td>0.000</td>
<td>0.809</td>
<td>-0.649</td>
<td>-0.105</td>
<td>0.542</td>
<td>0.632</td>
<td>-0.039</td>
<td>0.811</td>
<td>-0.787</td>
<td>-0.099</td>
<td>0.968</td>
<td>0.977</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$\log 1p K_T$</td>
<td>0.555</td>
<td>0.752</td>
<td>-0.016</td>
<td>0.049</td>
<td>0.871</td>
<td>-0.699</td>
<td>-0.096</td>
<td>0.548</td>
<td>0.689</td>
<td>0.020</td>
<td>0.878</td>
<td>-0.827</td>
<td>-0.092</td>
<td>0.940</td>
<td>0.966</td>
<td>0.989</td>
<td>1</td>
</tr>
</tbody>
</table>
is no causal relationship between the two variables concerned. For the purpose of predicting the within-year capacity, the most relevant variables to use as dependent variables are \( d/M \) (or its logarithmic transformation \( \log(1+pdM) \)) and \( K_T \) (or \( \log(1+K_T) \)). From Table 4, \( d/M \) is strongly correlated with \( K_A \), \( D \), \( m \), with respective correlation coefficients of 0.751, 0.700 and −0.641, and less so with the \( Cv \), the reliability \( (R) \) and record length \( (L) \) with correlation coefficients of 0.451, −0.124 and −0.084, respectively. The low correlation between \( d/M \) and \( Cv \) should not pose any problem if \( m \) is used as an independent variable (see definition of \( m \) in Table 3); otherwise both the \( Cv \) and \( D \) will have to be used as independent variables.

As expected, the total capacity ratio \( K_T \) is highly correlated with \( K_A \) (0.991), \( D \) (0.690) and \( Cv \) (0.522). This shows that \( K_A \) will be a very good predictor for \( K_T \). Finally, using logarithmic transformations increased all the correlation coefficients slightly. For example, the correlation coefficients between \( \log(1+pdM) \) and \( \log(1+K_A) \), between \( \log(1+pdM) \) and \( \log(1+m) \) and, between \( \log(1+pdM) \) and \( \log(D) \) are respectively 0.811, −0.765 and 0.706.

**Model 1**: \( d/M = fn(Cv, D, R, K_A) \)

Given the above correlations, model 1 used \( Cv \), \( D \), \( R \) and \( K_A \) as independent variables. Since the logarithmic transformations produced only marginal improvements in the correlation coefficients, which is unlikely to translate to better predictions or correlation when the variables are anti-logged, as will be noted later (see also McCuen et al., 1990), a linear model of the form:

\[
d/M = a + bCv + cD + hR + gK_A + \varepsilon
\]  

(2)

was hypothesized, where \( a \), \( b \), \( c \), \( h \) and \( g \) are the model parameters, \( \varepsilon \) is the error term and all the other variables are as defined in Table 3. Estimates of the model parameters and their statistics as obtained using least squares regression are shown in Table 5. These results give a particular form of model 1 as:

\[
d/M = 0.677 + 0.231Cv + 0.464D - 0.009R + 0.040K_A \quad (R^2 = 0.719)
\]  

(3)

As seen in Table 5, the \( t \) statistics associated with the parameter estimates are all significant at the 5% level. In addition, various diagnostic checks, involving probability and scatter plots, carried out by Lallemand (2001) on the error term, confirmed that the error term possesses the right attributes of independence, normality and homoscedasticity. Figure 2(a) compares the observed \( d/M \) and those predicted by equation (3). The performance of the model is quite good, although the model appears to predict negative \( d/M \) on occasions. The negative correlation between \( d/M \) and \( R \) (and hence the negative multiplier of \( R \) in equation (3)) may have contributed to the

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard error</th>
<th>( t ) statistic</th>
<th>Sig. at 5% (Y/N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A )</td>
<td>0.677</td>
<td>0.183</td>
<td>3.689</td>
<td>Y</td>
</tr>
<tr>
<td>( B )</td>
<td>0.231</td>
<td>0.024</td>
<td>9.591</td>
<td>Y</td>
</tr>
<tr>
<td>( C )</td>
<td>0.464</td>
<td>0.033</td>
<td>13.989</td>
<td>Y</td>
</tr>
<tr>
<td>( H )</td>
<td>−0.009</td>
<td>0.002</td>
<td>−4.560</td>
<td>Y</td>
</tr>
<tr>
<td>( G )</td>
<td>0.040</td>
<td>0.008</td>
<td>4.749</td>
<td>Y</td>
</tr>
</tbody>
</table>
incidences of negative predictions. However, the fact that other trial regression analyses (unreported herein) not including $R$ as an explanatory variable also produced negative $d/M$ predictions is proof that this is unlikely to be the sole cause. A straightforward explanation for the negative values is that some of the $d/M$ values used as dependent variables were close to zero. The negative predictions are concentrated in this region of low $d/M$ values.

Model 2: $K_T = fn(Cv, D, K_A)$

This model is an attempt to predict the combined (i.e. within-year and over-year) rather than the within-year adjustment and thereby possibly reduce or eliminate the negative values predicted by Model 1. Based on the correlation matrix in Table 1, a regression model of the form in equation (4) was set up:

$$K_T = \kappa + jCv + nD + sK_A + \varepsilon$$

where $\kappa$, $j$, $n$ and $s$ are the model parameters. The time-based reliability has not been included as an explanatory variable in Model 2 because of its very low correlation with $K_T$ (see Table 4).

The model parameter estimates and their associated statistics are shown in Table 6. The $t$ statistics in Table 6 are generally high, implying that the parameters are well estimated and are statistically different from zero. Similar diagnostic checks carried out on the residuals in Model 1 were also carried out for Model 2 and these were also found to be adequate. The resulting model therefore becomes:

$$K_T = -0.222 + 0.322Cv + 0.6D + 1.025K_A$$

($R^2 = 0.988$)  

Thus overall, Model 2 is a much better model than Model 1, which is further confirmed in Fig. 2(b) where the predicted and observed $K_T$ values are compared.

**Validation of regression equations**

To validate the models, they were applied to the three validation sites for the no-failure situation. The results are shown in Table 7. Also shown in Table 7 are the results of
### Table 6 Summary statistics for regression Model 2.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard error</th>
<th>t statistic</th>
<th>Sig. at 5% (Y/N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\kappa$</td>
<td>-0.222</td>
<td>0.030</td>
<td>-7.370</td>
<td>Y</td>
</tr>
<tr>
<td>$J$</td>
<td>0.322</td>
<td>0.031</td>
<td>10.217</td>
<td>Y</td>
</tr>
<tr>
<td>$N$</td>
<td>0.600</td>
<td>0.043</td>
<td>13.791</td>
<td>Y</td>
</tr>
<tr>
<td>$S$</td>
<td>1.025</td>
<td>0.011</td>
<td>93.238</td>
<td>Y</td>
</tr>
</tbody>
</table>

### Table 7 Validation of regression models.

<table>
<thead>
<tr>
<th>Site</th>
<th>Yield ratio, $D$:</th>
<th>0.4 SPA</th>
<th>Model</th>
<th>0.6 SPA</th>
<th>Model</th>
<th>0.8 SPA</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>$K_T$</td>
<td>1.16</td>
<td>1.5</td>
<td>2.81</td>
<td>2.67</td>
<td>5.74</td>
<td>5.45</td>
</tr>
<tr>
<td></td>
<td>$d/M$</td>
<td>0.18</td>
<td>0.34</td>
<td>0.8</td>
<td>0.48</td>
<td>1.14</td>
<td>0.67</td>
</tr>
<tr>
<td>14</td>
<td>$K_T$</td>
<td>0.24</td>
<td>0.24</td>
<td>0.72</td>
<td>0.70</td>
<td>1.28</td>
<td>1.30</td>
</tr>
<tr>
<td></td>
<td>$d/M$</td>
<td>0.18</td>
<td>0.08</td>
<td>0.33</td>
<td>0.19</td>
<td>0.42</td>
<td>0.30</td>
</tr>
<tr>
<td>15</td>
<td>$K_T$</td>
<td>0.07</td>
<td>0.10</td>
<td>0.21</td>
<td>0.30</td>
<td>0.56</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>$d/M$</td>
<td>0.07</td>
<td>0.02</td>
<td>0.13</td>
<td>0.12</td>
<td>0.22</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Applying the SPA to the three data records for no-failure. In general, the validation results are adequate and do confirm the relative superiority of Model 2 over Model 1. Of notable significance, however, is the good performance of Model 2 at site 13 whose $C_v (= 1.47)$ is outside the $C_v$ range employed for the model calibration. This is further proof of the efficacy of the model in providing reliable estimates of $K_T$ whatever the annual and monthly flow patterns being considered.

**DISCUSSION OF RESULTS**

The results presented above have shown that the two regression models are capable of providing reliable within-year adjustment for capacity estimates if the over-year capacity is known. Of the two models, Model 2 appears to perform better, not just in its prediction but also in its plausibility. For example, during the testing of Model 1, a relatively high number of negative values of $d/M$ were predicted, especially for low $d/M$ values. In general, low $d/M$ values will correspond to regions of low within-year effect, i.e. regions where over-year capacity will predominate and hence for which within-year adjustment will have little or no effect on the total capacity estimate. Attempts to rectify this problem using logarithmic transformation of $(1 + d/M)$, i.e. $\log_{10}(1 + d/M)$, did not succeed, because negative $d/M$ predictions still resulted after anti-logging and deducting unity (see Lallemand, 2001). In practical applications of the model, any negative value predicted must be set to zero.

The poor performance of Model 1 is not restricted to the low $d/M$ range alone. For example, as shown in Fig. 2(a), the predicted $d/M$ values also appear to deviate from the 1:1 line for large $d/M$. By contrast, Model 2 produced far fewer negative predictions and more importantly, its performance is uniformly good across the entire range of the reservoir capacity investigated. Model 2 also gives the total capacity directly, rather than the over-year adjustment provided by Model 1. Given all of this, it would seem that Model 2 is the preferred model to use; once $K_T$ is known, the $d/M$ component, if required, can be obtained readily by subtracting $K_A$ from $K_T$. 


The use of the models will require estimates of the $C_v$, $D$, $R$ and $K_A$. The time-based reliability, $R$, is normally specified as a planning parameter and is hence known. For a gauged site, the $C_v$ can easily be determined from the available data. The demand ratio, $D$, requires an estimate of the mean annual flow; this mean can also be obtained from the available data record. The over-year capacity ratio requires estimate of both the capacity (volumetric units) and the mean annual flow. The former can be obtained by analysing the annual data at the site using a suitable technique such as the modified SPA described by Adeloye et al. (2001). Alternatively, $K_A$ can be obtained using one of the available generalized SYP techniques such as the Gould-Gamma model given in equation (1). So the two models are readily applicable at gauged sites.

At ungauged sites, there are no flow data with which to estimate most of these variables and hence indirect approaches will have to be used. The Gould-Gamma method will estimate $K_A$ (and hence the over-year capacity ($m^3$)) if the mean, $C_v$ and skew coefficient of annual flows are known. A relatively accurate estimate of the mean annual runoff can be obtained using:

$$\bar{Q} = P - AE$$

where $\bar{Q}$ is the mean annual runoff (mm), $P$ is the mean annual rainfall (mm) and $AE$ is the mean annual actual evapotranspiration (mm). Often, only an estimate of the potential evapotranspiration ($PE$) is available which will have to be used to estimate the $AE$. For wet catchments, the $AE$ and the $PE$ are very close; however for drier catchments, the $PE$ often exceeds the $AE$. In the UK, a rainfall dependent correction factor is often used to obtain the $AE$ from the $PE$ such that (Gustard et al., 1992):

$$AE = wPE$$

$$w = \begin{cases} 0.00061P + 0.475 & P < 850 \\ 1.0 & \text{otherwise} \end{cases}$$

Once $\bar{Q}$ is estimated, the volumetric annual runoff can then be obtained by multiplying $\bar{Q}$ by the catchment area. Young et al. (1996) have shown that the combination of equations (6) and (7) performs very well in predicting the mean annual runoff for Scottish catchments.

In Australia, in addition to point and areal evapotranspiration maps, monthly and annual actual evapotranspiration maps are now available (Wang et al., 2001). As annual rainfall maps are also available, equation (6) can be readily solved.

A reliable estimate of the $C_v$ to use at ungauged catchments can be obtained by pooling together the at-site estimates of the $C_v$ at gauged sites in the region in which the ungauged site occurs (NERC, 1975):

$$\text{Pooled } C_v = \left( \frac{\sum_{i=1}^{ns} (T_i - 1)C_{vi}^2}{\sum_{i=1}^{ns} (T_i - 1)} \right)^{0.5}$$

where $ns$ is the number of gauged sites in the region, $T_i$ is the record length (years) and $C_{vi}$ is the observed $C_v$ at the $i$th gauged site. By the same token, a pooled estimate of the skew coefficient, $\gamma$, has also been recommended for use at ungauged sites as (NERC, 1975):
Regression models for within-year capacity adjustment in reservoir planning

\begin{equation}
\text{Pooled } \gamma = \frac{A_3}{(C_v)^3} \tag{(9a)}
\end{equation}

\begin{equation}
A_3 = \frac{\sum_{i=1}^{n_i} S_{3,i}}{\sum_{i=1}^{n_i} (T_i - 1)} \tag{(9b)}
\end{equation}

\begin{equation}
S_{3,i} = \sum_{i=1}^{T_i} \left( \frac{Q_{ji} - \bar{Q}_i}{\bar{Q}_i} \right)^3 \tag{(9c)}
\end{equation}

where \( C_v \) is the pooled \( C_v \) (see equation (8)), \( Q_{ji} \) is the annual flow in year \( j \) at the \( i \)th gauged site and \( \bar{Q}_i \) is the mean annual runoff at the \( i \)th gauged site. So the regression models are also applicable at ungauged sites.

Another issue worthy of a mention with regard to using the above models relates to steady state reliability vs no-failure reliability. The SYP implemented in this study with the SPA concerns the reliability over the respective records and would therefore be representative of the steady-state reliability if the record lengths were very long. This is the most commonly used form of reliability outside the USA. In the USA, where the no-failure probability is used, approximate procedure for converting from this form of reliability to the steady state reliability has been developed (see Vogel & McMahon, 1996). However, whatever the reliability measure being considered, the above models are still valid for two main reasons. First is that the reliability index is weakly correlated with either \( d/M \) or \( K_T \) and hence is not a strong explanatory variable for the capacity. This is why the reliability has not featured in Model 2. Secondly, both predictive models include \( K_A \), the over-year capacity estimate, as an explanatory variable. Thus, whatever reliability is incorporated in the estimate of the over-year capacity should be carried through in the estimate of \( d/M \) or indeed \( K_T \). As a consequence, the models presented in this work should equally apply in the context of both reliability definitions.

Finally, the analysis has only considered single realizations of the records at each of the sites. Consequently, it has not been possible to evaluate the stochastic properties of either \( d/M \) or \( K_T \). Such issues will be taken up in an extended study which will also use a much larger set of rivers and data.

**CONCLUSION**

Regression models have been developed to predict either the within-year capacity adjustment or the combined within-year plus over-year capacity estimates for use during reservoir planning. These regression models offer an interim solution of a generalized SYP relationship for systems with significant within-year effects. The basis of the data for calibrating the regression models was the SYP analysis of data from 12 international rivers using the modified SPA. The regression models were later validated using three other independent sites. The streams cover a wide range of annual flow variability as well as the majority of the unique monthly flow patterns of world streams; hence the regression models should be applicable in most regions of the world. The models are also applicable in both gauged and ungauged situations and procedures for obtaining the independent variables in either situation were presented. In general, the regression model for predicting the combined capacity was better.
during calibration and validation than the model for estimating the within-year capacity adjustment. Since both models use more or less the same independent variables, i.e. neither offers a relative advantage as regards obtaining the independent variables, the $K_T$ model with its better predictive capability should be preferred.

Acknowledgements The data used in this work were made available principally to T. A. McMahon through the assistance of a number of individuals including Alan Gustard (Centre for Ecology and Hydrology, Wallingford, UK), Jim Slack (US Geological Survey), Stefan van Biljon (Department of Water Affairs, Republic of South Africa) and Murray Peel (University of Melbourne, Australia). The authors would also like to thank the two anonymous reviewers for their helpful comments.

REFERENCES


Received 1 June 2002; accepted 1 April 2003