Continuous rainfall-runoff model comparison and short-term daily streamflow forecast skill evaluation
Thomas Pagano, Prasanta Hapuarachchi, and Q.J. Wang

June 2010

Report to Bureau of Meteorology

A water information R & D alliance between the Bureau of Meteorology and CSIRO’s Water for a Healthy Country Flagship
Australia is founding its future on science and innovation. Its national science agency, CSIRO, is a powerhouse of ideas, technologies and skills.

CSIRO initiated the National Research Flagships to address Australia’s major research challenges and opportunities. They apply large scale, long term, multidisciplinary science and aim for widespread adoption of solutions. The Flagship Collaboration Fund supports the best and brightest researchers to address these complex challenges through partnerships between CSIRO, universities, research agencies and industry.

The Water for a Healthy Country Flagship aims to provide Australia with solutions for water resource management, creating economic gains of $3 billion per annum by 2030, while protecting or restoring our major water ecosystems.

The work contained in this report is collaboration between CSIRO’s Water for a Healthy Country Flagship and the Bureau of Meteorology’s Water Division under the Water Information Research and Development Alliance (WIRADA).

WIRADA is a five-year, $50 million partnership that will use CSIRO’s R&D expertise to help the Bureau of Meteorology report on availability, condition and use of Australia's water resources.

For more information about the partnership visit www.csiro.au/partnerships/WIRADA.html
For more information about Water for a Healthy Country Flagship or the National Research Flagship Initiative visit www.csiro.au/org/HealthyCountry.html

Citation: Pagano, T, P. Hapuarachchi, and Q.J. Wang, 2010. Continuous rainfall-runoff model comparison and short-term daily streamflow forecast skill evaluation. CSIRO: Water for a Healthy Country National Research Flagship

Copyright and Disclaimer
© 2010 CSIRO To the extent permitted by law, all rights are reserved and no part of this publication covered by copyright may be reproduced or copied in any form or by any means except with the written permission of CSIRO.

Important Disclaimer:
CSIRO advises that the information contained in this publication comprises general statements based on scientific research. The reader is advised and needs to be aware that such information may be incomplete or unable to be used in any specific situation. No reliance or actions must therefore be made on that information without seeking prior expert professional, scientific and technical advice. To the extent permitted by law, CSIRO (including its employees and consultants) excludes all liability to any person for any consequences, including but not limited to all losses, damages, costs, expenses and any other compensation, arising directly or indirectly from using this publication (in part or in whole) and any information or material contained in it.

Cover Photograph:
From http://www.abc.net.au/news/stories/2010/01/02/2784276.htm
Description: Castlereagh River, northeast New South Wales “Floodwaters cover a vast expanse of land near Coonamble” (02/01/2010)
Photographer: Amy Griffiths (Daily Liberal)
© 2010 Approval pending from Rural Press
# CONTENTS

Acknowledgments .................................................................................................................. 5  
Executive Summary ................................................................................................................ 6  

1. **INTRODUCTION** ........................................................................................................... 7  
   1.1. Motivation .................................................................................................................. 7  
   1.2. Structure of this document ....................................................................................... 8  

2. **RAINFALL-RUNOFF MODELS** .................................................................................. 9  
   2.1. Recovering Initial Loss Model (RILM) ..................................................................... 9  
   2.2. GR4J .......................................................................................................................... 10  
   2.3. Summary and discussion ......................................................................................... 11  

3. **MODEL CALIBRATION AND EVALUATION** .............................................................. 13  
   3.1. Parameter estimation and objective functions ......................................................... 13  
   3.2. Calibration and validation methods ......................................................................... 14  
   3.3. Summary and Discussion ...................................................................................... 19  

4. **MODELLING DATASETS** ........................................................................................... 20  
   4.1. Benchmark Catchments ......................................................................................... 20  
   4.2. Forcing and observed data ...................................................................................... 20  

5. **MODEL INTER-COMPARISONS** ............................................................................... 22  
   5.1. Base models (Rfact = off, Routing = lincatch) ....................................................... 22  
   5.2. Models without additional routing (Routing = none) ............................................. 25  
   5.3. Use of a rainfall multiplier (Rfact = on) .................................................................. 29  
   5.4. Which configurations are best for each model? .................................................... 32  
   5.5. Benefits of combined models ............................................................................... 35  
   5.6. Summary and discussion ...................................................................................... 39  

6. **ERROR CORRECTION** ............................................................................................... 41  
   6.1. Introduction ............................................................................................................. 41  
   6.2. Dual pass error correction ...................................................................................... 41  
   6.3. Results ..................................................................................................................... 42  
   6.4. Summary and discussion ...................................................................................... 47  

7. **ENSEMBLE STREAMFLOW PREDICTION** ................................................................. 49  
   7.1. Introduction ............................................................................................................. 49  
   7.2. Retrospective forecasts .......................................................................................... 49  
   7.3. Results ..................................................................................................................... 51  
   7.4. Summary and discussion ...................................................................................... 62  

8. **FINAL SUMMARY AND RECOMMENDATIONS** ....................................................... 64  

REFERENCES .................................................................................................................... 66
LIST OF FIGURES

Figure 2.1 The RILM model structure and equations .............................................................. 9
Figure 2.2 The GR4J model structure ..................................................................................... 11
Figure 3.1 Diagram of double split-sample validation procedure ............................................ 15
Figure 3.2 Validation performance of models for stitching (solid) and averaging (dashed) ..... 16
Figure 3.3 As figure 3.2 but for additional diagnostic skill scores ......................................... 17
Figure 3.4 As figure 3.1 but for the difference between averaging and stitching skill scores. 18
Figure 3.5 As figure 3.4 but for additional diagnostic skill scores ......................................... 18
Figure 4.1 240 Benchmark catchments in Southeast Australia (in red) ............................... 20
Figure 4.2 Aridity and runoff production of study catchments .............................................. 21
Figure 5.1 Calibration (dashed) and validation (solid) performance of models using linear reservoir catchment routing and no rainfall multiplier .............................................. 23
Figure 5.2 As figure 5.1 but for additional diagnostic skill scores .......................................... 24
Figure 5.3 As figure 5.1 but for the difference between validation and calibration ............... 25
Figure 5.4 As figure 5.3 except for the ancillary skill measures ............................................. 25
Figure 5.5 Validation performance of models using linear reservoir catchment routing (solid lines) versus no additional catchment routing (dashed lines) ........................................ 26
Figure 5.6 As figure 5.5 but for additional diagnostic skill scores .......................................... 27
Figure 5.7 As figure 5.5 except for the difference in skill when using catchment routing versus not .................................................................................................................. 28
Figure 5.8 As figure 5.7 except for ancillary skill measures .................................................. 29
Figure 5.9 As figure 5.5 except with the use of a rainfall multiplier (Rfact, dashed) versus non-use of a rainfall multiplier (NoRfact, solid) ......................................................... 30
Figure 5.10 As figure 5.9 except for the ancillary skill measures ............................................ 31
Figure 5.11 As figure 5.9 except for the difference in skill between using and not using rainfall multipliers .............................................................................................................. 31
Figure 5.12 As figure 5.11 except for the ancillary skill measures .......................................... 32
Figure 5.13 Map of validation performance of GR4J without catchment routing or a rainfall multiplier ................................................................. 34
Figure 5.14 Validation performance of individual models (solid lines) and a multi-model combination (dashed line) .......................................................... 36
Figure 5.15 As figure 5.14 except for the additional diagnostic scores .................................. 37
Figure 5.16 As figure 5.14 except for the difference in skill between individual models and the multi-model combination ................................................................. 38
Figure 5.17 As figure 5.16 except for the ancillary skill measures .......................................... 38
Figure 6.1 Model performance for raw model (dashed) and with error correction (solid) .... 45
Figure 6.2 As figure 6.1 but for additional diagnostic skill scores ......................................... 46
Figure 6.3 As figure 6.1 but for the difference between no error correction and no error correction .............................................................................................................. 46
Figure 6.4 As figure 6.3 but for additional diagnostic skill scores ......................................... 47
Figure 7.1 Location of retrospective ESP forecasts.......................................................... 50
Figure 7.2 Skill of retrospective ensemble forecasts by lead-time (x-axis)......................... 52
Figure 7.3 Performance of retrospective ESP forecasts................................................... 53
Figure 7.4 As figure 7.3 except for the additional diagnostic scores............................... 53
Figure 7.5 Seasonal climatology (1974-2006) of Oxley at Eungella (top) and Pirron Yallock (bottom). .................................................................................................................. 55
Figure 7.6 Oxley forecast skill by month......................................................................... 56
Figure 7.7 Pirron Yallock forecast skill by month......................................................... 56
Figure 7.8 Simulation performance of GR4J. ................................................................. 58
Figure 7.9 Performance of ESP forecasts with 1 day lead-time. .................................... 59
Figure 7.10 Performance of ESP forecasts with 3 day lead-time..................................... 60
Figure 7.11 Performance of ESP forecasts with 7 day lead-time.................................... 61
LIST OF TABLES

Table 2.1 Parameters of the RILM model................................................................. 10
Table 2.2 Parameters of the GR4J model............................................................... 11
Table 2.3 Summary of daily time-step rainfall-runoff models................................. 12
Table 3.1 Forecast evaluation measures.............................................................. 19
Table 4.1 Characteristics of 240 study catchments............................................. 21
Table 5.1 Performance of configurations (rows) by rainfall runoff model (columns).... 33
Table 6.1 Optimal error correction parameters for GR4J (top), AWBM (middle) and RILMCL (bottom). ........................................................................................................... 43
ACKNOWLEDGMENTS

This work is part of the water information research and development alliance between CSIRO’s Water for a Healthy Country Flagship and the Bureau of Meteorology.

This work would not have been possible without the assistance and guidance of hydrologists at the Bureau, in particular Jim Elliott, Soori Sooriyakumaran, Chris Leahy, Justin Robinson, Adam Smith, Peter Baddiley, and Gnanathikkam Amirthanathan. However, the views expressed herein are solely those of the authors and do not necessarily represent those of the Bureau or its employees.

The authors are indebted to CSIRO colleagues that provided feedback, datasets and model code, such as Jai Vaze, Neil Viney and Julien Lerat.
EXECUTIVE SUMMARY

This research was done to support the development and improvement of an operational short-term (i.e. <14 days into the future) streamflow forecasting service at the Australian Bureau of Meteorology.

Part of this research is to compare various available rainfall-runoff models and determine which model would be suitable for the task. Four rainfall runoff models (AWBM, GR4J, PDM, and SimHyd) were run on 240 Southeast Australian catchments from 1974-2006 and the results compared to two naïve baseline models (rainfall multiplied by a ratio and a simple loss model that has been used experimentally in the Bureau operational environment).

For practically every measure, GR4J outperformed every other model considered. GR4J rose to the task of performing well in the measures it was asked to perform well in during model setup. Part of the data was withheld during model setup and GR4J had the least decline in performance of any of the models when asked to simulate streamflow from this independent period. Most remarkably, it also performs among the best of all the models in measures that were not part of the model setup (related to the shape and spikiness of the hydrograph). This suggests that GR4J is capturing realistic hydrologic behaviour and is not simply fitting the data. This is all the more remarkable because GR4J has 4 tunable parameters, fewer than most rainfall-runoff models.

The models were tried in a variety of configurations and tested on the 240 catchments to determine the optimal setup. A tunable parameter to rescale rainfall inputs did not reliably improve the four rainfall runoff models. Additional catchment routing routine marginally benefitted SimHyd (which already had ways of delaying and attenuating surface runoff), but it did not improve GR4J (which uses a unit hydrograph). A skill-weighted average of the hydrographs from the four models was not significantly better than GR4J on its own.

Error correction uses the difference between recent observed and simulated flow to improve the simulations. An error correction routine that corrects for quickly (1-day) and slowly (1-year) varying biases was proposed and applied. This is a novel approach and showed the most improvement during low flow conditions. The naïve models benefitted the most, although the final performance of uncorrected rainfall-runoff models was still better than the corrected naïve model.

The Ensemble Streamflow Prediction method was used to generate retrospective forecasts (as opposed to simulations where future rainfall is known with certainty). Streamflow forecast skill diminished with lead-time, sometimes rapidly (i.e. +1 day ahead). A fair amount of the apparent forecasting skill was due to the model’s ability to reproduce the seasonal cycle of streamflow. When compared to a more challenging baseline that considered the seasonal cycle, many catchments did not have forecast skill beyond +3 days ahead. To the authors’ knowledge, this is the first Australian study to measure short-term streamflow forecasting skill at more than 125 catchments.
1. INTRODUCTION

1.1. Motivation

Reliable water forecasts are critical for the management of extreme events and for optimising river and water resources management. These forecasts need lead-times ranging from hours to weeks to be of value. The Australian Bureau of Meteorology (the Bureau) is seeking to expand and improve its short-term streamflow forecasting services. This is the fourth in a series of technical reports on the research being done to support the science behind these services. This research is being done as part of the Water Information Research and Development Alliance (WIRADA), one component of which is project “4.1 Water Forecasting and Prediction – Short Term.” WIRADA brings together CSIRO’s R&D expertise in water and information sciences and the Bureau of Meteorology’s operational role in hydrological analysis and prediction to transform the way Australia manages water resources.

Project 4.1 aims to (1) extend the Bureau’s current event based flow modelling for flood prediction to provide continuous flow forecasting through the use of soil moisture accounting models; (2) improve the accuracy of flow and flood prediction at both catchment outlet and internal (ungauged) points through using spatially variable rainfall input; (3) improve forecast accuracy through statistical post-processing and model updating; (4) improve forecast lead-time through using Numerical Weather Prediction products; (5) improve model initializations through the use of blended multi-sensor precipitation estimates; (6) develop practical methods for quantifying forecast uncertainty; (7) develop a hydrological modelling system for providing flash flood guidance; (8) evaluate potential modelling frameworks for upgrading the Bureau’s current flood forecasting system. To achieve some of the above goals, WIRADA has developed a modelling application called Short-term Water Information Forecasting Tools (SWIFT).

Demonstration of the improvement of alternative approaches over existing practices is an important ingredient in the successful adoption of new technologies. This demonstration must be under operationally realistic conditions to be relevant and credible. In particular, the testing methodology should allow the models to make best use of information that would be available to forecasters in real-time and disallow the use of non-operationally available information or the use of methods that would not be practical within operational constraints (of computing power, timeliness of forecasts and so on). The current operational forecasts are likely to benefit from forecaster expertise, experience and intuition, and so the challenge to the researcher is to construct a baseline that resembles existing practice and yet can be objectified.

This report details the latest progress towards developing SWIFT, its rainfall runoff models, its parameter calibration routines and performance evaluation measures. It also evaluates the ability of the models to simulate historical flow and to use recent simulation errors to keep the simulations on track (i.e. error correction). The project has also taken an important modelling step over the threshold into forecasting, as opposed to running simulations where rainfall is known with certainty. Although the forecasting method is fairly basic, the results may indicate the expected performance of these models when used to make real-time forecasts. Furthermore, the results may provide insight into and scientific understanding of key hydrologic processes in Southeast Australia, at lease from the perspectives of the models.
1.2. Structure of this document

This report is organised as follows:

- Chapter 2 introduces two rainfall-runoff models that have been integrated into the streamflow forecasting application. This adds to the many models already in the application. These other models were described in previous reports.
- Chapter 3 describes the calibration procedures and objective functions that have been used. The model cross validation procedure is described.
- Chapter 4 describes the model forcing data and a set of benchmark study catchments.
- Chapter 5 documents the results of preliminary experiments using the above mentioned rainfall-runoff models, calibration procedures, and datasets. The models are configured in a variety of ways to find the optimal model setup. The performance of a multi-model combination is also considered.
- Chapter 6 describes and evaluates an error correction routine that accounts for both slowly and quickly varying simulation errors.
- Chapter 7 runs the best performing rainfall-runoff model in retrospective ensemble prediction mode for a subset of catchments. The skill of these forecasts are analysed by location and by season.
- Chapter 8 concludes with a discussion of the above results and makes recommendations for future directions.
2. RAINFALL-RUNOFF MODELS

As part of WIRADA research activities a streamflow forecasting application has been developed, the Short-term Water Information Tools (SWIFT) program. This application contains a variety of rainfall-runoff models which can be combined with a variety of catchment and channel routing models. SWIFT’s routing algorithms are detailed in Pagano et al. [2009b] and the majority of SWIFT’s rainfall-runoff models are described in Pagano et al. [2009a]. This section will introduce two rainfall-runoff models that were not included in those reports but are used in this study. The first is a Recovering Initial Loss Model (RILM) that will serve as a baseline for comparison to the other models. Next is modèle du Génie Rural à 4 paramètres Journalier (Model of Agricultural Engineering 4 parameters Daily, GR4J), a conceptual rainfall runoff model. The following sections describe the models, their major processes, and their parameters (default values, reasonable ranges, and controls on hydrograph behaviour).

2.1. Recovering Initial Loss Model (RILM)

Carroll [2007] describes the event-based rainfall-runoff routing model (Unified River Basin Simulator, URBS) that is currently used operationally by the Bureau of Meteorology. Event-based models are significantly different from continuous models (all the other models in this study) in that soil moisture states are not simulated but rather estimated by the operator and these estimates vary from event to event. Typically, an initial loss is specified and there is no runoff until cumulative rainfall satisfies this loss. Once the initial loss is satisfied, there is a continuing loss (constant or proportional) that is deducted from runoff. This model is not useful for long runs (i.e. weeks) because it does not allow for the recovery of the initial loss during interstorm periods.

Carroll [2007] proposed an extended model (figure 2.1, table 1) that allows the recovery of the initial loss; when rainfall is less than the continuing loss (CL), initial loss is increased by a fraction (frac) of the difference between continuing loss and rainfall. Initial loss cannot be greater than a maximum initial loss (MaxIL). The constant (as opposed to proportional) continuing loss model is used here. Carroll [2007] recommends a value of frac between 0.1 and 0.5. Throughout this report RILM and RILMCL are used interchangeably.

Elsewhere in this report the results of RILM and the “coefficient” model are compared. The main differences between these two models are that the coefficient model’s losses are fixed through time (they do not recover) and are proportional (as opposed to RILM’s continuing loss).

![Figure 2.1 The RILM model structure and equations. In the diagram on the left, parameters are shown in parentheses.](image)

If Rain > IL then
Runoff = Rain – IL
IL = 0
Else
Runoff = 0
IL = IL – Rain
End if

If Runoff > CL then
Runoff = Runoff – CL
Else
Runoff = 0
End if

If rain > CL ; IL = IL + frac * (CL – Rain)
IL = Min(MaxIL, IL)
### Table 2.1 Parameters of the RILM model

<table>
<thead>
<tr>
<th>Name</th>
<th>Default</th>
<th>Min</th>
<th>Max</th>
<th>Units</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>MaxIL</td>
<td>2</td>
<td>0</td>
<td>400</td>
<td>mm</td>
<td>Higher values mean increased losses, later rising hydrographs</td>
</tr>
<tr>
<td>CL</td>
<td>0.25</td>
<td>0</td>
<td>20</td>
<td>mm/hr</td>
<td>Higher values mean increased losses after initial loss is satisfied</td>
</tr>
<tr>
<td>Frac</td>
<td>0.3</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td>Higher values mean increased losses, less runoff between events, later rising hydrographs. A value of 0 reduces the model to an event-based model without recovery.</td>
</tr>
</tbody>
</table>

### 2.2. GR4J

Perrin et al. [2003] developed the modèle du Génie Rural (Agricultural Engineering Model) à 4 paramètres Journalier (4 parameters Daily, GR4J, figure 2.2, table 2). This model was created using the “Top Down” approach; the developers proposed a structure that was general enough to encompass the components of existing efficient models and systematically reduced complexity while preserving adequate performance. The developers used pieces of 19 well-known models from the literature to construct 235 different model structures and the results were evaluated on 429 basins across four continents including Australia. The model with the least number of parameters and best performance during independent validation was retained.

One of GR4J’s unique aspects is the use of a “groundwater exchange coefficient” (X2 in the diagram below). When this parameter is away from 0, the model relaxes the requirement that runoff balances precipitation and evapotranspiration; the authors defend this based on the possibility of transport across the lateral boundary of the subsurface of catchments. Although the parameter is conceptually linked to groundwater, Le Moine et al. [2007] showed that it is also effective at reducing model bias, more so than simple scaling parameters for rainfall, potential evapotranspiration or runoff. When it is unclear what process a parameter is controlling (Is it describing groundwater flows? Or compensating for non-representativeness of station forcing data?), it is more difficult to conduct certain experiments with the model. For example, what effect will land use change have on streamflow? Of course, the user of any conceptual hydrologic model faces the same difficulties.

Mathevet et al. [2004] tested 140 possible structural modifications to GR4J to develop a model that could be run on an hourly (as opposed to daily) time-step. Mathevet called this model GR5H. When tested on 145 catchments, GR5H performed significantly better in validation than four other well-known models. SWIFT has implemented a 4-parameter variant of GR5H suggested by Lerat (J. Lerat, CSIRO Land and Water, personal communication 2010). The difference between GR4J and the hourly model is primarily a rescaling of some of the model parameters, constants and exponents. All of the simulations here use the daily model.
**Figure 2.2** The GR4J model structure. This figure originally from Perrin et al. [2003]. Refer to the original source for model equations.

Table 2.2 Parameters of the GR4J model.

<table>
<thead>
<tr>
<th>Name</th>
<th>Default</th>
<th>Min</th>
<th>Max</th>
<th>Units</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>350</td>
<td>1</td>
<td>3000</td>
<td>mm</td>
<td>Higher values increase soil moisture, giving hydrograph longer memory.</td>
</tr>
<tr>
<td>X2</td>
<td>0</td>
<td>-27</td>
<td>27</td>
<td>mm</td>
<td>Higher values increase streamflow, violating water balance.</td>
</tr>
<tr>
<td>X3</td>
<td>150</td>
<td>1</td>
<td>660</td>
<td>mm</td>
<td>Controls the discharge of baseflow</td>
</tr>
<tr>
<td>X4</td>
<td>40</td>
<td>1</td>
<td>240</td>
<td>hr</td>
<td>Higher values delay, attenuate the hydrograph</td>
</tr>
</tbody>
</table>

**2.3. Summary and discussion**

This chapter has documented the various rainfall-runoff models newly implemented into the application being developed for this research. Table 3 summarizes these models as well as other relevant models in the SWIFT system (previously described in Pagano et al. [2009a; 2009b]).

Although not discussed further in this report, the Sacramento Soil Moisture Accounting (SAC-SMA) model [Burnash, 1995] is a well-known streamflow forecasting model that is used operationally in many countries. The model has a very flexible structure with many
parameters that are commonly manually calibrated. It would be a candidate for future evaluation, primarily because of its ubiquitous international use. While SAC-SMA is included in SWIFT, it was not included in this research because its implementation is still in developmental mode.

The structures of the various models are similar in that most contain a fixed capacity soil moisture store that efficiently produces runoff when filled. This store also generates recharge that fills a bottomless groundwater reservoir which is a pathway for the production of (slow) baseflow.

Table 2.3 Summary of daily time-step rainfall-runoff models. The second column provides the number of calibrated parameters, whereas the number in parenthesis is the total available parameters including those that the model authors recommend remain fixed.

<table>
<thead>
<tr>
<th>Name</th>
<th>Calibrated Parameters (Total)</th>
<th>Number of storages</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>AWBM</td>
<td>3 (7)</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>SimHyd</td>
<td>7 (7)</td>
<td>2</td>
<td>Has a 3rd storage for interception although its contents are evaporated at each time-step. Has its own routing.</td>
</tr>
<tr>
<td>PDM</td>
<td>4 (6)</td>
<td>2</td>
<td>PDM’s soil moisture storage is tracked as an average storage across a continuous distribution of many individual stores</td>
</tr>
<tr>
<td>RILM</td>
<td>3</td>
<td>1</td>
<td>Storage is the initial loss</td>
</tr>
<tr>
<td>COEFF</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>GR4J</td>
<td>4</td>
<td>2</td>
<td>Violates water balance. Has its own routing.</td>
</tr>
</tbody>
</table>

Other relevant models

<table>
<thead>
<tr>
<th>Name</th>
<th>Calibrated Parameters (Total)</th>
<th>Number of storages</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAC-SMA</td>
<td>13 (17)</td>
<td>6</td>
<td>Still testing implementation</td>
</tr>
</tbody>
</table>

Later sections will describe the skill improvement associated with surface water routing. The simulations done here are spatially lumped and therefore no channel routing is used. However, some models (i.e. SimHyd, PDM and GR4J) have their own routing methodology. For example, GR4J uses a 1 parameter Unit Hydrograph. PDM uses a 2 parameter cascade of two linear reservoirs. For all the various models, in experiments labelled “lincatch”, surface runoff (but not baseflow) has been routed through PDM’s routing scheme, in addition to the models existing routing schemes. In experiments labelled “nocatch”, either no routing is done or the model’s (i.e. SimHyd and GR4J) original routing is used.
3. MODEL CALIBRATION AND EVALUATION

3.1. Parameter estimation and objective functions

All of the above mentioned rainfall-runoff models have parameters that must be estimated using, for example, expert knowledge of catchment behaviour, an objective parameter optimization routine, or a combination of both. Beven [2001] provides a good overview of the parameter estimation problem and the various approaches in use. SWIFT’s automatic calibration routines were previously described in Pagano et al. [2009a; 2009b]. This section will briefly recapture the salient points.

The application uses the following objective functions (as used in Wang et al. [2010], modified from Zhang et al. [2008]):

\[ F_{ns} = 1 - \frac{\sum_{i=1}^{n} (Q_{sim,i} - Q_{obs,i})^2}{\sum_{i=1}^{n} (Q_{sim,i} - \bar{Q}_{obs})^2} \]

\[ F_{log,ns} = 1 - \frac{\sum_{i=1}^{n} [\ln(Q_{sim,i} + \epsilon) - \ln(Q_{obs,i} + \epsilon)]^2}{\sum_{i=1}^{n} [\ln(Q_{sim,i} + \epsilon) - \ln(\bar{Q}_{obs} + \epsilon)]^2} \]

\[ F_{corr} = \frac{\sum_{i=1}^{n} (Q_{sim,i} - \bar{Q}_{sim})(Q_{obs,i} - \bar{Q}_{obs})}{\sqrt{\sum_{i=1}^{n} (Q_{sim,i} - \bar{Q}_{sim})^2} \sum_{i=1}^{n} (Q_{obs,i} - \bar{Q}_{obs})^2} \]

\[ F_{bias} = 1 - \left[ \max\left(\frac{\bar{Q}_{sim}}{\bar{Q}_{obs}}, \frac{\bar{Q}_{obs}}{\bar{Q}_{sim}}\right) - 1 \right]^2 \]

where \( Q_{sim,i} \) and \( Q_{obs,i} \) are the simulated and observed streamflows, respectively, at time step \( i \), \( n \) is the number of days in the calibration, and \( \epsilon \) is a small value (equal to the smallest non-zero observed flow value of the entire period of record). \( F_{ns} \) is the Nash-Sutcliffe model efficiency coefficient (NS, [Nash and Sutcliffe, 1970]), an observed-variance-normalized mean squared error that gives highest weight to large errors (which often, but not always, happen during periods of high flow). \( F(\log) \) is the NS of logarithmically transformed flows with greater weight given to errors during low flows. \( F(\text{corr}) \) is the Pearson correlation coefficient, which measures the covariability of the simulated and observed without penalizing for bias. \( F_{bias} \) is a symmetric measure of the match between the average simulation and average observation (i.e. a Bias Score).

Other objectives that have been implemented in the application include the NS of Box-Cox transformed flows

\[ F_{box} = 1 - \frac{\sum_{i=1}^{n} (Q'_{sim,i} - Q'_{obs,i})^2}{\sum_{i=1}^{n} (Q'_{sim,i} - \bar{Q}'_{obs})^2} \]

where streamflow is transformed by

\[ Q'_{sim,i} = \frac{(Q_{sim,i} + 1)^{0.3} - 1}{0.3} \]

following the recommendations of Misirli et al. [2002]. Do note that the Box-Cox transformation used here, with its exponent of 0.3, is a specific instance of a more general transformation. With this exponent value, the score is a good measure of intermediate flows.
Mathematically, when the exponent equals 1.0, \( F_{\text{boxcox}} \) equals \( F_{ns} \). When the exponent tends to 0.0, the transformation approaches the natural logarithm, although the score will not exactly equal \( F_{\log ns} \) due to the choice of an offset of 1 versus \( \varepsilon \).

Viney et al. [2009] proposed the following

\[
F_{\text{viney}} = 1 - \frac{\sum_{i=1}^{n}(Q_{\text{sim},i} - Q_{\text{obs},i})^2}{\sum_{i=1}^{n}(Q_{\text{sim},i} - \bar{Q}_{\text{obs}})^2} - 5 \ln(1 + \frac{\sum_{i=1}^{n}(Q_{\text{sim},i} - Q_{\text{obs},i})}{\sum_{i=1}^{n}Q_{\text{obs},i}})^{2.5}
\]

which combines the Nash Sutcliffe with a log-bias constraint. A perfect score is 1, lack of skill lies below 0.

The SWIFT program also has the ability to calculate a skill score based on indices of hydrograph shape. Specifically, the rising/falling limb density (RLD, FLD) is the number of peaks divided by the cumulative time of rising/falling limbs. Shamir et al. [2005] applied these measures during the calibration of hydrologic model parameters. Within SWIFT, the measures have been fashioned into a skill score.

\[
F_{\text{RLD}} = 1 - \left( \frac{\text{RLD}_{\text{sim}}}{\text{RLD}_{\text{obs}}} - 1 \right)^2
\]

\[
F_{\text{FLD}} = 1 - \left( \frac{\text{FLD}_{\text{sim}}}{\text{FLD}_{\text{obs}}} - 1 \right)^2
\]

Each day of a rising/falling limb must have an absolute rate of change at least twice the smallest observed rate of change on record, or the event is discarded. This makes the procedure more robust and less sensitive to sensor “flutter” during relatively flat baseflow periods.

Pagano et al. [2009a; 2009b] described how a combination of objective functions could be used for parameter calibration. In this study, an unweighted average of \( F_{\log ns} \), \( F_{ns} \), \( F_{\text{corr}} \), \( F_{\text{bias}} \) will be used for parameter calibration, following Wang et al. [2010]. The results will be evaluated using all of the above mentioned scores, however.

### 3.2. Calibration and validation methods

As before, Pagano et al. [2009a; 2009b] describes SWIFT’s automatic calibration routines and therefore only a brief treatment is given here. During the calibration period, the Shuffled Complex Evolution (SCE, [Duan et al., 1994]) technique is used to calibrate model parameters. This scheme evaluates many potential combinations of parameter values to find the best fit between simulated and observed streamflow. Fit is evaluated by maximizing the objective function described in section 3.1. The use of multiple objectives during calibration is believed to give more stable estimates of model parameters and prevents overfitting.

The historical record of observed flow data was divided into two periods with equal amounts of non-missing data. The parameters were calibrated using data from the first part of the record (“SCE-Straight”) and the resulting optimal parameters were used to simulate the second part of the record (“RUN-Straight”). Next, the periods are reversed and calibration occurs in the second half of the record (“SCE-Swapped”) and validated on the first half of the record (“RUN-Swapped”). “RUN-Swapped” and “RUN-Straight” can be spliced together to form a time series of flows that spans the entire record and is of equal length to a spliced “SCE-Straight”/“SCE-Swapped”. Furthermore, the difference in skill between calibration and validation will not be due to differing climate regimes between the two periods. In all experiments, the period 1/1/1974-1/1/1979 is reserved for model warmup.
Figure 3.1 Diagram of double split-sample validation procedure. In the top diagram, the model is calibrated (SCE) on the first half of the period and validated (RUN) on the second half. In the 2nd diagram, the model is calibrated on the second half and validated on the first. In the bottom diagrams, the results from individual periods are combined. Note that no additional calibration occurs when simulations are stitched together.
The question arises if the skill score statistics should be calculated for the entire stitched time series ("stitched"), or if they should be calculated on individual halves of the record and averaged together ("avg")? It will be shown here that the former may mask biases and the latter should be used instead. Consider the case of a catchment that has a large step change in bias near the middle of the record. The model may calibrate the parameters to minimize bias in the first half of the record and may over-simulate in the second half in validation. Next, the model may minimize bias in the second half of the record and under-simulate in the first half. Under both averaging and stitching, biases will appear small in calibration. However, in validation, stitching will allow the biases from the first half of the record to cancel out the biases from the second half and the bias would appear small. In contrast, if the skill scores from the individual poorly performing periods were averaged together, the result would be more informative and more representative of the expected bias that would be experienced during validation.

Figure 3.2 Validation performance of models for stitching (solid) and averaging (dashed). The skill scores are described in section 3.1. Generally higher values mean better performance. The lower right plot is the average of 4 scores and the lower right plot is the difference between averaging and stitching.
Figures 3.2 and 3.3 show the validation period performance of 240 catchments (discussed in chapter 4) for the various skill measures calculated in the two ways. The interpretation of this kind of graph is discussed in more detail in chapter 5; the main purpose of showing these graphs here is to merely illustrate that the method of calculating the skill score has important consequences.

In these graphs higher values mean seemingly better performance. The model configurations shown are the non-error corrected RILM and AWBM models with catchment routing and a rainfall multiplier (discussed in later sections) and GR4J without catchment routing or a rainfall multiplier. These are the same collection of model configurations as is used in chapter 6. Stitching is shown as the solid line and averaging is shown as the dashed line.

Figures 3.4 and 3.5 and the lower right panel of figure 3.2 show the difference between averaging and stitching; negative values mean that averaging results in lower skill scores than stitching. In this context, negative is preferable because it means that averaging provides more information about model performance than stitching.

Measures that include correspondence (e.g. $F_{boxcox}$, $F_{ns}$, $F_{logns}$) appear slightly worse in averaging than stitching and the naïve models appear more greatly affected than the full rainfall-runoff models. However, as expected, averaging gives much lower $F_{bias}$ than stitching (lower right figure 3.4), especially for naïve models. The rising/falling limb density scores (figure 3.5) are also affected; these scores are calculated by counting the number of peaks relative to the total time spent rising or falling and therefore it might be possible to overestimate the time spent rising in one half of the record and underestimate in another and on the whole achieve a high score under stitching (but not averaging).
Figure 3.4 As figure 3.1 but for the difference between averaging and stitching skill scores. Negative means averaging gives lower scores than stitching.

Figure 3.5 As figure 3.4 but for additional diagnostic skill scores.
3.3. Summary and Discussion

A summary of skill measures is presented in Table 4.

Table 3.1 Forecast evaluation measures. No-skill threshold for all scores is zero.

<table>
<thead>
<tr>
<th>Name</th>
<th>Range (worst to best)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_{ns}$</td>
<td>$-\infty$ to 1</td>
<td>Sensitive to large errors, peak flows</td>
</tr>
<tr>
<td>$F_{log ns}$</td>
<td>$-\infty$ to 1</td>
<td>Sensitive to errors during low flow</td>
</tr>
<tr>
<td>$F_{corr}$</td>
<td>0 to 1</td>
<td>Like $F_{ns}$ but does not penalize for bias</td>
</tr>
<tr>
<td>$F_{bias}$</td>
<td>$-\infty$ to 1</td>
<td>Does not measure correspondence. Is presented as a skill score and therefore does not provide information about bias direction.</td>
</tr>
<tr>
<td>$F_{boxcox}$</td>
<td>$-\infty$ to 1</td>
<td>Sensitive to errors during mid-range flows</td>
</tr>
<tr>
<td>$F_{vines}$</td>
<td>$-\infty$ to 1</td>
<td>Combines $F_{ns}$ and $F_{bias}$</td>
</tr>
<tr>
<td>$F_{RLD}$</td>
<td>$-\infty$ to 1</td>
<td>Sensitive to speed of rising limb. Does not measure correspondence.</td>
</tr>
<tr>
<td>$F_{FLD}$</td>
<td>$-\infty$ to 1</td>
<td>Sensitive to speed of falling limb. Does not measure correspondence.</td>
</tr>
</tbody>
</table>

A cross validation procedure is used in which the model is calibrated on half of the record and validated on the other. This procedure is repeated, with the calibration and validation periods swapped. The skill scores are calculated on the two individual halves of the record and are averaged together; this was shown to be more informative than calculating a single skill score for the entire record.

WIRADA is still in pursuit of objective functions that match forecasters’ subjective impressions of simulation quality. The United States National Weather Service has an objective procedure for emulating manual calibration; The objective function is a linearly weighted combination of mean error, mean square error, error in the histogram, and maximum errors of under- and over-estimation of hourly, daily, 10-daily and monthly flows [NWS, 2009]. Pokhrel et al. [in press] also have an interesting set of similar “diagnostic signature measures” designed for rainfall-runoff models. Ongoing communication between CSIRO and the Bureau of Meteorology will help foster objectives that are relevant to Australian modelling.

Later in this report retrospective ensemble forecasts will be evaluated. The greatest benefit of ensemble forecasting is that it provides a measure of forecast uncertainty. Deterministic measures, like the ones mentioned above, do not measure the probabilistic reliability of the forecasts. If the forecasts are reliable (this is an assumption) it is likely that a deterministic distillation of the ensemble forecast (e.g. the ensemble median or mean) could be effectively evaluated using deterministic measures. Barnston [1992] found monotonic relationships between (deterministic) correlation and (probabilistic) Heidke and Linear Error in Probability Space (LEPS) scores. Therefore, if forecasts are reliable, the relative ranking of the skill of various ensemble forecasts would be very similar for probabilistic and deterministic measures. Nonetheless, it will be essential to measure the probabilistic reliability of the ensembles in future work.
4. MODELLING DATASETS

4.1. Benchmark Catchments

A primary objective of the WIRADA research program is to evaluate the performance of streamflow forecasts on sub-daily timescales. High quality continuous sub-daily timeseries data are less available than daily data and therefore must be cultivated for this project. The original intent of this report was to compare the performance of models, in forecasting mode, for a set of 15-20 Bureau-selected catchments with sub-daily data. To date, only sub-daily data for the Ovens catchment has been made available (see Pagano et al. [2009a; 2009b]).

Previously, WIRADA research focused on a set of 320 unregulated catchments distributed around Australia (from Peel et al. [2000]). Subsequent research found that some of those catchments were indeed regulated or had significant land use change. Perraud et al. [2009] developed a new set of 240 unregulated and data rich catchments in southeast Australia (figure 4.1). The hydrologic modelling community has been encouraged to use this set of catchments in order to facilitate the comparison of results across studies. Daily streamflow values were provided by other CSIRO researchers (J Vaze, CSIRO Land and Water, personal communication, December 2009).

![Image of 240 Benchmark catchments in Southeast Australia](image-url)

Figure 4.1 240 Benchmark catchments in Southeast Australia (in red). Surface water management areas shown in gray and significant rivers in blue. The background is a Landsat7 visible image.

4.2. Forcing and observed data

Daily rainfall and potential evaporation data were obtained from the SILO system [Jeffery, 2006; Jeffrey et al., 2001] These data are spatial interpolations of gauge measurements to a 5-km grid across the entire Australian mainland and Tasmania. The models are run in
spatially lumped mode, using areal average forcings. The forcing data are serially complete from 1/1/1974 to 31/12/2006. SILO is updated through real-time, although the limiting factor was the lack of streamflow data in recent years.

Table 4.1 Characteristics of 240 study catchments. Reference period is 1974-2006.

<table>
<thead>
<tr>
<th></th>
<th>Area (km²)</th>
<th>Precipitation (mm/year)</th>
<th>Runoff (mm/year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum (0%)</td>
<td>50</td>
<td>448</td>
<td>9</td>
</tr>
<tr>
<td>25th percentile</td>
<td>160</td>
<td>715</td>
<td>59</td>
</tr>
<tr>
<td>Median (50%)</td>
<td>320</td>
<td>867</td>
<td>115</td>
</tr>
<tr>
<td>75th percentile</td>
<td>630</td>
<td>1,102</td>
<td>261</td>
</tr>
<tr>
<td>Maximum (100%)</td>
<td>2,000</td>
<td>2,113</td>
<td>1,170</td>
</tr>
</tbody>
</table>

Most catchments (figure 4.2) are moisture limited. Table 5 provides information about the characteristics of the study catchments. For reference, Australia’s area-average long-term annual rainfall is approximately 500 mm/year [Lavery et al., 1997], meaning that these study catchments over-represent Australia’s relatively wet regions. Similarly, the wettest areas (as wet as 4,400 mm/year in northern Queensland) are not represented.

240 Southeast Australia Benchmark Catchments

![Figure 4.2 Aridity and runoff production of study catchments. The x-axis, Runoff/Precipitation is a measure of the relative efficiency of runoff production. On the y-axis, catchments near the top have an abundance of rainfall and are energy limited whereas lower catchments have strong evaporative demand and are moisture limited. Values are for the period of record 1974-2006.](image-url)
5. MODEL INTER-COMPARISONS

This section compares the performance of the various daily simulation models in a variety of configurations. Section 5.1 is for the base models in calibration and validation. Section 5.2 compares the effect on validation of not adding catchment routing. Section 5.3 measures the validation benefit of including a rainfall multiplier. Section 5.4 studies the performance of a simple multi-model average and Section 5.5 provides discussion and conclusions.

5.1. Base models ($\text{Rfact} = \text{off}$, $\text{Routing} = \text{lincatch}$)

The remainder of this chapter describes the performance of the daily simulation models (AWBM, PDM, SimHyd, and GR4J). All models have their surface water component catchment-routed through a cascade of two linear reservoirs (a structure originally from the PDM model). Baseflow is not catchment routed.

The results are compared to two naïve baselines, the coefficient model (COE, i.e. Runoff = a * Precipitation, a = [0-1]) and the Recovering Initial Loss Model (RILM, denoted RILMCL here to emphasize the use of continuing loss, see section 2.1) with PDM-style catchment routing. Both naïve baselines account for the seasonality of rainfall. The coefficient model with catchment routing can smooth and delay the flow and is somewhat like what one would expect from an Event-based proportional-loss model when the initial loss is satisfied. RILMCL is somewhat more sophisticated but, like COE, does not account for the seasonality in potential evaporation.

Figure 5.1 shows the performance of the calibration and validation results of the six models. While AWBM, SIMHYD and PDM display comparable performance, GR4J consistently performs better than the others. All models except the naïve baselines seem to perform comparably at minimizing bias (figure 5.1, centre right). Somewhat surprisingly the coefficient (COE) model seems to outperform the recovering loss (RILMCL) model for the majority of catchments, although roughly $\frac{1}{3}$ of the highest skill RILMCL models outperform the COE models. The reader interested in the geographic patterns of model skill can skip to figure 5.13 in section 5.4.

The lower right panel of figure 5.1 (and figure 5.3) show the difference in skill between validation and calibration. Negative values indicate a loss of skill during validation compared to calibration. The 4-measure average score typically loses 0.05 points of skill. The amount of skill lost is not evenly distributed among the individual components of the 4-measure average. For example, $F_{ns}$ suffers the most among the correspondence scores, although bias is responsible for loss of major loss of skill in a select minority of catchments.

Figures 5.2 and 5.4 show the performance of the remainder of the scores. The performance of $F_{boxcox}$ (mid-range flows) is similar to the other measures. $F_{viny}$ summarizes a combination of $F_{ns}$ and $F_{bias}$. GR4J is a strong performer for the Rising Limb Density ($F_{RLD}$), although surprisingly, COE is the best model. This is suggestive that Australian catchments (at least of this size) are extremely fast responding and most of the signal on the rising limb is due to coincident rainfall. On sub-daily timescales this result may change. On the Falling Limb Density, GR4J again is a strong performer although SIMHYD is the most skilful for many catchments. The Limb Density results for the naïve baselines should be interpreted in the light that the baseline models are still using catchment routing (compare with section 5.2).
Figure 5.1 Calibration (dashed) and validation (solid) performance of models using linear reservoir catchment routing and no rainfall multiplier. The skill scores are described in section 3. Generally higher values mean better performance. The lower right figure shows the difference in skill between validation and calibration. Negative values mean that calibration is better than validation.
Figure 5.2 As figure 5.1 but for additional diagnostic skill scores.
5.2. Models without additional routing (Routing = none)

As mentioned in section 2, all of the models had their surface runoff component routed through a cascade of two linear reservoirs, a structure originally used in the PDM model. This routing model has two parameters, one to control the length of the delay and another to control the linearity of the recession. In the extreme case of the second parameter equalling zero, the routing model reduces to a single linear reservoir. This reservoir responds to an impulse of rainfall (a given amount of rainfall on a single timestep) with a simple recession (falling limb only). As the second parameter approaches one, an impulse of rainfall produces a hydrograph with both a rising and falling limb. Conceptually, this routing model is representing the “fast” component of flow and therefore baseflow is not routed.

Two of the models (SIMHYD and GR4J) already contain surface runoff delaying components and therefore coupling these models to PDM’s routing may be redundant. In other words “Is runoff being double-routed?” Indeed, by adding more parameters, the chance of overfitting the model increases, potentially leading to a loss of skill in validation. This section addresses the impact, positive or negative, of coupling PDM’s routing model to other rainfall routing models. AWBM and PDM are not discussed in this section because these two models do not contain surface runoff delaying components that would be redundant.

Figure 5.5 shows the impact of removing catchment routing. If catchment routing improved the simulations, one would expect solid lines closer to the upper left corner than dashed lines of the same color. GR4J’s performance is largely unchanged, except that its performance is consistently slightly deteriorated at low flows when PDM-style catchment routing is used (upper left figure 5.7).
The lower right hand plot of figure 5.5 (note rescaling from previous plots) shows that a small minority of catchments are improved, balanced out by a small number that are degraded. In comparison, the naïve baselines suffer massive loss of skill without catchment routing (many models have negative skill scores). Without catchment routing, the naïve baselines have no recession and respond only to coincident rainfall. In this case, when routing is present, the 4-measure average objective function is likely dominated by measures with correspondence (i.e. $F_{\log_{10}}$, $F_{n}$, $F_{corr}$). When routing is absent, the automatic calibration gives up on its ability to achieve correspondence and can only hope to minimize bias.

![Graphs showing validation performance](image)

Figure 5.5 Validation performance of models using linear reservoir catchment routing (solid lines) versus no additional catchment routing (dashed lines). See also figure 5.1. These models did not use a rainfall multiplier. In the lower right figure, the difference in skill is shown. Positive means that catchment routing improves performance.

This result is further illustrated in figure 5.6. Notice in particular that in the absence of catchment routing practically none of the naïve baseline models can have positive skill for the Falling Limb Density ($F_{FLD}$, lower right corner). Despite this, the Rising Limb Density
(\( F_{RLD} \), lower left corner) performance is not entirely negative, which is likely due to many events going from trough to peak in a single day; again, this result may not hold for sub-daily simulations and for larger catchments.

Figure 5.6 As figure 5.5 but for additional diagnostic skill scores.
Figure 5.7 As figure 5.5 except for the difference in skill when using catchment routing versus not. Positive means that catchment routing improves skill.
5.3. Use of a rainfall multiplier (Rfact = on)

Many conceptual rainfall-runoff models accept precipitation and potential evaporation as forcings, calculate losses to actual evaporation, assume no aquifer recharge, and produce runoff. However, it is difficult to measure rainfall and runoff, particularly during large events. Furthermore, the forcing data (likely interpolated from limited station measurements) may not be representative of true catchment averages. A catchment may be affected by human influences, such as consumptive uses.

Boughton [2006] used the AWBM model to simulate Australian streamflow and found that two-thirds of poorly performing model simulations could be rehabilitated by using a multiplier on the rainfall and/or potential evaporation data. In 80% of cases, potential evaporation was scaled down an average of 14% to give the best fit. Optimal rainfall multipliers were fairly evenly distributed positive and negative.

As mentioned earlier, GR4J stands apart from the other models in that it has a groundwater exchange coefficient to represent non-conservative losses or gains. Le Moine et al. [2007] tested a variety of rescaling methods (adjusting rainfall, potential evaporation, catchment area) and found that GR4J’s exchange coefficient was the most effective. Le Moine et al simulated many French catchments and was able to relate the parameter’s value to catchment lithology.

Here, the models are modified to allow for a rainfall multiplier (Rfact) that is calibrated. The coefficient model (COE) is itself simple model that says that effective precipitation is a fraction of actual precipitation. Therefore, the coupling of the COE model to a rainfall multiplier is redundant.

Figure 5.9 shows the validation performance for models including a rainfall multiplier (dashed lines) versus those that do not (solid). If a multiplier improved performance, the dashed line would be closer than the solid line to the upper left corner of each plot. For the standard skill measures, a rainfall multiplier has no net benefit for most of the models; some catchments are improved, some degraded (figure 5.11 and lower right corner of figure 5.9). As discussed above, the COE model is unchanged, however the use of a multiplier has a significant positive impact on the RILM model. This result is put to a finer point in figure 5.10, which shows the validation performance of the ancillary skill scores. In particular, rising and falling limb density skill scores are greatly improved for RILM with the rainfall multiplier (figure 5.12). Of all the models, GR4J shows the least change, likely because it already has a (more
efficient) mechanism for solving problems with the water balance.

Figure 5.9 As figure 5.5 except with the use of a rainfall multiplier (Rfact, dashed) versus non-use of a rainfall multiplier (NoRfact, solid). These models use catchment routing. In the lower right figure, the difference in skill is shown. Negative means rainfall multipliers improve performance.
Figure 5.10 As figure 5.9 except for the ancillary skill measures.

Figure 5.11 As figure 5.9 except for the difference in skill between using and not using rainfall multipliers. Negative values mean that rainfall multipliers improve performance.
5.4. Which configurations are best for each model?

The combination of results from the prior two sections are summarized in table 5.1. The entries in the table give, for each rainfall-runoff model, the percentage of catchments that had the highest performance measure for a given configuration. Performance here is an unweighted average of all 8 skill scores, including the ancillary skill measures. The top half of the table is for calibration, the bottom half is validation. Deeper colors indicate that this combination was frequently optimal. Each column sums to 100%. Gray cells are for configurations that were not attempted. For example, the upper left most entry (80%) means that 192 (80% of 240) catchments preferred AWBM with catchment routing to have a rainfall multiplier whereas 49 (20%) performed better in calibration without a rainfall multiplier. One must recognize that the calibration objective was 4 skill scores whereas this table shows the results for 8 skill scores.

This table shows that the optimal combination for most models is to use both catchment routing and a rainfall multiplier. However, for the non-naive models, the balance shifts in validation so that the models show a nearly equal preference for rainfall multipliers or not using multipliers. This suggests that perhaps if the hydrologist had evidence that a rainfall multiplier is needed (based on knowledge of the catchment or the data), its use could help, but that it should not be used simply by default.

GR4J stands out among the other models as being the only one that prefers not to include catchment routing. Similarly, the shift in preference about rainfall multipliers from calibration to validation suggests that the original structure of the model performs fine without these accoutrements. Of course these results may be specific to the 240 catchments tested.
Table 5.1 Performance of configurations (rows) by rainfall runoff model (columns). Deeper colours indicate a better performing configuration for each rainfall runoff model. See text for more discussion.

### Calibration

<table>
<thead>
<tr>
<th>Configuration</th>
<th>AWBM</th>
<th>SIMHYD</th>
<th>GR4J</th>
<th>FDM</th>
<th>RILMCL</th>
<th>COE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Route On, Rfact On</td>
<td>30%</td>
<td>36%</td>
<td>19%</td>
<td>75%</td>
<td>84%</td>
<td>49%</td>
</tr>
<tr>
<td>Route On, Rfact Off</td>
<td>20%</td>
<td>22%</td>
<td>5%</td>
<td>25%</td>
<td>5%</td>
<td>46%</td>
</tr>
<tr>
<td>Route Off, Rfact On</td>
<td>28%</td>
<td>64%</td>
<td>10%</td>
<td>1%</td>
<td>0%</td>
<td>4%</td>
</tr>
<tr>
<td>Route Off, Rfact Off</td>
<td>14%</td>
<td>13%</td>
<td>0%</td>
<td>4%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Validation

<table>
<thead>
<tr>
<th>Configuration</th>
<th>AWBM</th>
<th>SIMHYD</th>
<th>GR4J</th>
<th>FDM</th>
<th>RILMCL</th>
<th>COE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Route On, Rfact On</td>
<td>55%</td>
<td>27%</td>
<td>10%</td>
<td>54%</td>
<td>60%</td>
<td>43%</td>
</tr>
<tr>
<td>Route On, Rfact Off</td>
<td>45%</td>
<td>36%</td>
<td>15%</td>
<td>46%</td>
<td>7%</td>
<td>53%</td>
</tr>
<tr>
<td>Route Off, Rfact On</td>
<td>19%</td>
<td>42%</td>
<td>13%</td>
<td>1%</td>
<td>0%</td>
<td>3%</td>
</tr>
<tr>
<td>Route Off, Rfact Off</td>
<td>18%</td>
<td>33%</td>
<td>0%</td>
<td>3%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Across all the model configurations, GR4J without catchment routing and with a rainfall multiplier performed the best in terms of the performance of the average of 8-measures of skill; the geographic median score of 240 catchments was 0.828, marginally outperforming (by 0.3%) the same configuration but with no rainfall multiplier (0.825). Although the improvement from a rainfall multiplier holds up in validation, one must question if the gains are worth having to increase the number of parameters from 4 to 5.

Therefore, for having the best performance for the fewest number of parameters, GR4J without routing or multipliers will be the focus for closer analyses. Figure 5.13 shows a map of validation performance skill (the original 4-measure score) for this model. This corresponds to the data for the dashed blue line in the lower left plot of figure 5.5. Skill is highest in central Victoria, and coastal northeast New South Wales. Skill is worst in isolated catchments near the New South Wales/Victoria border and in interior northeast New South Wales. More geographic patterns of skill are presented in chapter 7.
Figure 5.13 Map of validation performance of GR4J without catchment routing or a rainfall multiplier. Blue indicates better performance, red indicates worse. The background is a Landsat7 visible image. See text for discussion. Note change of scale increments above 0.90.
5.5. Benefits of combined models

Inevitably, all models are imperfect representations of reality; each is a different perspective on a system. Operational hydrology often focuses on the use of a single tool or a single model in developing forecast guidance. In many natural science and economic settings, research consistently reveals that a consensus forecast based on the output of many tools almost always outperforms the best individual tool within the ensemble [Armstrong, 2001].

The approach of creating forecasts based on an ensemble of tools (e.g. “Super-ensembles”) has gained acceptance in the operational meteorological and climatological communities, and the evolution of hydrologic practice along these lines would be logical and would benefit users. Previously, the resources required to maintain many different modelling systems made such an enterprise prohibitively expensive, especially if the incremental improvement in the forecasts was small compared to the cost of maintaining many different systems. Operational meteorologists and climatologists rely heavily on automation and leverage partnerships with outside research groups (e.g. universities) running their own models; there are no obvious reasons why the same approach could not be used by hydrologists.

A simple method of multi-model combination is to take a weighted average of several models. For each of the 4 models (m, i.e. AWBM, SIMHYD, PDM and GR4J) and for each half of the record (p, i.e. first and second), calculate the combined objective function during model calibration period. If the objective function is negative, the value is set to zero.

\[ F_{avg}(m, p) = \max[0, F_{\log}(m, p) + F_{nu}(m, p) + F_{corr}(m, p) + F_{bias}(m, p)] \]

The multi-model combined streamflow is calculated for the opposite period by the following weighted average.

\[ Q(combo, p') = \frac{\sum_{m=1}^{4} F_{avg}(m, p)Q(m, p')}{\sum_{m=1}^{4} F_{avg}(m, p)} \]

For example, if the model was calibrated on the first half of the period (p), the calibration performance is used to combine the streamflows from the validation (p’) period, the second half of the record. The process is repeated for taking the calibration results from the second half of the record to determine how validation flows from the first half are weighted. The skill scores are calculated for the individual periods and averaged together, using the same method as earlier in this report.

Here, AWBM, SIMHYD, PDM and GR4J are included in this average and the performance of the resulting hydrograph (“combo”) is shown as the dashed black line in figures 5.14 and 5.15. The difference of the skill of the individual models (e.g. AWBM) from the multi-model combination is shown in figures 5.16 and 5.17 and the lower right panel of 5.14. All models except GR4J use PDM-style catchment routing and use the rainfall multiplier (in previous sections it was shown that these two add-ons benefitted those models). GR4J is run without routing or a rainfall multiplier as this was determined in the previous section to be the best performing model with the fewest number of parameters.

Figure 5.14 shows that the performance of GR4J and the multi-model combination are practically indistinguishable except for \( F_{\log} \), low flows, where the combination is better than any individual model. Figure 5.15 suggests that the falling limb density is substantially improved by multi-model combination, a result that is not readily explainable.

Figure 5.16 and 5.17 largely confirm these results. The multi-model combination is better than SIMHYD, PDM or AWBM in nearly 95% of catchments for nearly every correspondence measure. In non-correspondence measures (e.g. \( F_{bias} \), \( F_{RLD} \), \( F_{FLD} \)), the multi-model
combination improves a smaller fraction (but still a majority) of catchments. Except for $F_{\text{log ns}}$, GR4J is relatively evenly split between being better or worse than the multi-model combination.

In this case, results from a very good model (GR4J) were combined with the results from three fair models. There may be a high level of collinearity among the model results and therefore the averaging method may dilute the results of the best model by outnumbering it with fair models. Other methods may be able to prevent this diluting, such as using principal components analysis.

Figure 5.14 Validation performance of individual models (solid lines) and a multi-model combination (dashed line). In the lower right corner, the performance of the combination is subtracted from the performance of the individual models; negative means that the model combination performs better than the individual models.
Figure 5.15 As figure 5.15 except for the additional diagnostic scores. Dashed line is a multi-model combination.
Figure 5.16 As figure 5.14 except for the difference in skill between individual models and the multi-model combination. Negative values mean that multi-model combination improves performance.

Figure 5.17 As figure 5.16 except for the ancillary skill measures. Negative values mean that multi-model combination improves performance.
5.6. Summary and discussion

In this chapter, daily simulations from four rainfall-runoff models were compared to two naïve baselines on 240 catchments across southeast Australia. The models were compared in a double split sample evaluation. In this process, the model is calibrated on the first half of the period and then run on the second half. Next, the model is calibrated on the second half and run on the first half. Skill scores are calculated on each time series and the scores from the two validation periods are averaged together. The same is done for the calibration period. This results in measures of skill for calibration and validation that come from time series with identical climatic characteristics.

Not surprisingly, validation performance was slightly worst than calibration performance for all models. GR4J suffered the least loss of performance, however, suggesting a possible robustness to its model structure and a low chance of model overfitting during calibration. One would be tempted to attribute this to GR4J parsimony and relatively low number free parameters but the RILMCL model, which has one fewer parameter than GR4J, had the greatest loss of skill in validation of any of the models considered.

All of the models were compared on nine measures of performance, one of which considers bias alone ($F_{bias}$), one for correspondence alone ($F_{corr}$), three for bias and correspondence together ($F_{ns}$, $F_{logns}$, $F_{boxcox}$) and two for hydrograph shape characteristics ($F_{RLD}$, $F_{FLD}$). Two of the scores ($F_{viney}$ and $(F_{ns} + F_{logns} + F_{bias} + F_{corr})/4$) were weighted averages of other scores. The last of these measures, the 4-score average, was used as the objective function during calibration.

For practically every measure, the cumulative distribution of skill of GR4J is greater than every other model. GR4J rises to the task of performing well in the measures it was asked to perform well in during calibration, it sustains this performance better than any other model into validation, but most remarkably, it also loses the least skill (from calibration to validation) of all the models at the ancillary skill scores that were not used during calibration. One is left with the impression that GR4J is capturing the *Gestalt* (essence of an entity's complete form) of the hydrology and hydrograph. It would be interesting to test, for individual catchments with comparable standard skill scores (e.g. Nash is the same for GR4J and AWBM), if GR4J’s simulations would rate higher than other models in a visual/subjective evaluation.

This chapter also evaluated the benefit/detriment of using PDM’s cascade of two linear reservoirs for catchment routing of (quick) surface runoff. The catchment routing largely had minor impact on lumped models that already had mechanisms for delaying of surface runoff (i.e. SIMHYD). However, catchment routing was an essential element of good performance for the naïve baseline models, and therefore it is no surprise that the operational models (whose rainfall-runoff components are quite simple) have a heavy emphasis on routing. PDM’s catchment routing method is also spatial scale independent, as opposed to GR4J and SIMHYD; this issue was not explored in these spatially lumped simulations, but may be important in future semi-distributed simulations.

A rainfall multiplier was used as a parameter to help close the water balance. Again, this multiplier helped some of the naïve models but had a neutral impact on the more complex rainfall-runoff models. The effect was least on GR4J, which already has a parameter to control non-conservative losses/gains. The selection of test catchments may affect the results. These are largely long-record high-quality regulation-free catchments. They may not be representative of what is typically encountered in operations, e.g. catchments with consumptive uses and sparse data networks. The rainfall multiplier may be useful in later experiments when inhomogeneity in the forcing data is expected (e.g. the use of Numerical Weather Prediction models).

Past studies have suggested that a multi-model combination will likely outperform any individual model. Although common in other fields, such as climatology and meteorology, the use of multiple hydrologic models in a single operational centre is very rare and new. Part of the issue is associated with the practical operational overhead of calibrating/maintaining
multiple models and the level of human intervention typically involved in keeping simulations on track in real-time.

Nonetheless, this study showed that a multi-model combination typically performed no worse than the best single model. It did not show improvement over the best model, likely due to the simplicity of the model combination scheme. In particular, a second tier of three fair performing models outnumbered the single best model (GR4J). Ideally, the combination method should consider the amount of intercorrelation of models in the pool as well as the relative performance of each. Fully formalized (e.g. Bayesian) combination methods exist (e.g. Ajami et al. [2006]).

One of the objectives of WIRADA’s research program is to compare the performance of new methods to existing operational practice. Two baseline models were included, a simple coefficient on rainfall, and a recovering initial loss model, the latter being close to the model being used operationally. When used in isolation, both models were extremely poor performers. When coupled with catchment routing routines, their performance was passable but still much worse than the continuous simulation models, such as GR4J.

One is strongly and vigorously cautioned against concluding from this information that the performance of the naïve baseline models is about the same as what is seen in operational practice. In these simulations, the catchments are lumped and the model parameters are fixed. In operations, the models are semi-distributed and humans adjust the parameters throughout flood events. A fairer comparison would involve time-varying error correction and state/parameter updating.
6. ERROR CORRECTION

6.1. Introduction

A significant element of hydrologic forecasting is to keep model simulations consistent with recent observations of non-forcing variables (e.g. soil moisture, streamflow). Differences between simulations and observations can persist over time and adjustment of the simulations to compensate for errors can lead to more skilful predictions. There are many kinds of such data assimilation techniques, often divided into several classes (from Sene [2010]):

- Input updating – adjustment of the input data or forecasts to the model
- State updating – adjustment of the initial model states
- Parameter updating – adjustment of the parameters of the model
- Output updating – adjustments of the outputs of the model

This chapter focuses on output updating, whose alternative terms include error correction, error prediction, output correction, among others. Autoregressive moving average (ARMA) models are ubiquitously used for error correction (e.g. Broersen and Weerts [2005]). One limitation of such models is their inability to distinguish relatively quickly varying errors from slow drift. Such slow drift may be due to emerging biases such as undocumented diversions, land use change, sensor calibration issues and so on and, in theory, should change more slowly and therefore persist longer into the future than transient errors. Section 6.2 describes a method that may overcome this limitation. Section 6.3 describes the results of applying this method to the study catchments mentioned above. Section 6.4 concludes with a discussion.

6.2. Dual pass error correction

Define

\[
Q_L(t) = Q_{raw}(t) \left( \frac{\sum_{j=1}^{365} Q_{obs}(t-j)}{\sum_{j=1}^{365} Q_{raw}(t-j)} \right)^{\alpha}
\]

Where the subscripts obs, raw and L mean observed streamflow, raw (unadjusted) model simulation, and model simulation adjusted for bias over the prior year, respectively. One year was selected for the analysis period so that seasonality would not have an effect on the results. An adjustment factor for the long-term correction is \(\alpha\), which varies between 0 (no adjustment) and 1 (full adjustment). If any days over the past year are missing for the simulated or observed, only concurrent records are used. If all of the data from one or all variables are entirely missing over the last year, then

\[
Q_L(t) = Q_{raw}(t)
\]

Next, define a “second pass”;

\[
Q_{LS}(t) = \max[0, Q_L(t) + \beta(Q_{obs}(t-1) - Q_L(t-1))]
\]

Where the subscript LS means model simulation adjusted for both long-term and short-term bias. A second correction factor, \(\beta\), adjusts the strength of the adjustment for short term biases (again, 0 means no adjustment, 1 means full adjustment). Originally a multiplicative model was used for the previous equation, but trials on 240 catchments (not shown) suggested the additive model was superior. In classical ARMA modelling, if \(\beta\) were less than 1, one would need to include a constant term that accounts for mean differences.
between simulated and observed. This term is not included because it is believed that the “first pass” of the error correction will remove these systematic biases. Also, bias during the calibration period was commonly absent.

If all of the data from one or all variables is missing over the prior time-step, then

\[ Q_{LS}(t) = Q_L(t) \]

This method has two parameters to be calibrated for simulation mode. When it comes to forecasting, a third parameter may be used to control the shape of an exponential decay blending the adjustment back to the raw simulation at long lead-times. Specifically, forecasts should be heavily weighted towards \( Q_{LS}(t) \) at short lead-times and at long lead-times it should revert to \( Q_L(t) \). In this study, only one step ahead error correction is considered and therefore \( Q_{LS}(t) \) has an effective weighting of 1.0 and this parameter is unnecessary.

### 6.3. Results

For the 240 catchments mentioned in prior sections, simulations from three models were used to study the benefits of error correction. This includes one strong performer, GR4J without catchment routing and no rainfall multiplier (a good performer), the AWBM model with catchment routing and a rainfall multiplier (a fair performer), and a naïve baseline RILMCL with catchment routing and a rainfall multiplier.

Each model’s results from its two calibration periods (i.e. Sce-straight and Sce-swapped, see section 3.2) were used to form \( Q_{raw}(t) \). In 6 increments of 0.2 from 0 to 1 (i.e. 0, 0.2, 0.4, 0.6, 0.8, 1.0), 36 combinations of \( \alpha \) and \( \beta \) were tried. The period of record was divided into a “straight” and “swapped” period (again, see section 3.2) and the combination of \( \alpha \) and \( \beta \) that gave the highest \( F_{viny} \) (section 3.3) for \( Q_{LS}(t) \) vs \( Q_{obs}(t) \) for each period was found. \( F_{viny} \) was selected because of its ease of implementation in this exploratory work and its ability to quantify bias and correspondence simultaneously. Future research should use the full 4-measure objective function described in earlier sections although the results will likely be similar.

Next, these optimal parameters were applied to the model’s validation time series. In this sense, both model hydrologic parameters and error correction parameters are being validated on an independent dataset. The result is time series of \( Q_{LS}(t) \) for two non-overlapping periods. As above, a warmup period of 1974-1979 was discarded.

Table 6 shows the optimal error correction parameters for the three models (GR4J- top, AWBM- middle, RILMCL- bottom). “Alpha (long)” is the gain factor for the bias correction based on the 365 day period whereas “Beta (short)” is the gain factor for the bias correction based on the error from the previous single day. Values of zero mean zero adjustment, values of one mean full adjustment. “Earlier” is the optimal parameter calibrated on the first half of the record, “Later” is the optimal parameter for the second half. The red numbers are the joint frequency that a given 2 parameter combination was determined as optimal whereas the blue numbers are the marginal distribution of individual parameters. For example, in the upper left most entry of the table 6.1D, 25% (60 of 240) catchments had optimal long-term gain factors of zero for both the first half and second half of the record for the GR4J model. This means that the model largely passes up the opportunity to correct flows based on long-term biases.
Table 6.1 Optimal error correction parameters for GR4J (top), AWBM (middle) and RILMCL (bottom). Alpha is the long-term correction factor whereas beta is the short-term correction factor. Earlier and later refer to the first and second half of the period of record. Deeper colours indicate more frequent occurrences. Where a combination of parameters never occurred, the cell is blank. See text for further explanation.

<table>
<thead>
<tr>
<th></th>
<th>GR4J</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>beta (short)</td>
<td>(earlier)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>10%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>8%</td>
<td>2%</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>2%</td>
<td>4%</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0%</td>
<td>2%</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>1%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Sum</td>
<td>20%</td>
<td>21%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>AWBM</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>beta (short)</td>
<td>(earlier)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>8%</td>
<td>5%</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>5%</td>
<td>1%</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0%</td>
<td>1%</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0%</td>
<td>1%</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0%</td>
<td>1%</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0%</td>
<td>1%</td>
</tr>
<tr>
<td></td>
<td>Sum</td>
<td>19%</td>
<td>31%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>RILMCL</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>beta (short)</td>
<td>(later)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>2%</td>
<td>4%</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0%</td>
<td>3%</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0%</td>
<td>1%</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0%</td>
<td>2%</td>
</tr>
<tr>
<td></td>
<td>Sum</td>
<td>4%</td>
<td>12%</td>
</tr>
</tbody>
</table>
For sub-tables 6.1-3 A and D, the parameters would be considered stable if most of the numbers ran along the diagonal. If the numbers were evenly distributed across the table, then the parameter is poorly identified. The short-term error correction factor ($\beta$) appears relatively stable (table A), although the long-term error correction factor ($\alpha$) is often zero (table D).

For the rainfall-runoff models (GR4J and AWBM) the error correction tends to gravitate towards $\alpha$ values less than 0.4 and $\beta$ values of 0.2 or 0.4. For the na"ive baseline model (RILMCL), the model relies heavily on short-term error correction, with $\beta$ more in the range of 0.4 to 0.8, although $\alpha$ still prefers to be less than 0.4.

These results are somewhat expected. All of the models considered have a parameter that can help in removing static long-term biases (e.g. the rainfall multiplier, GR4J’s subsurface exchange parameter) during calibration. The error correction was optimized for the calibration period and therefore it would only rely heavily on the long-term error correction if the biases persisted for a year or more but varied within the calibration period (14 years).

Figure 6.1 shows the improvement of including error correction in the validation period. Here, the average of the two periods scores are shown, as per the discussion at the start of this section. The solid line is for the error corrected results and the dashed line is the raw model output.

Figures 6.1 and 6.3 show that error correction has many benefits. $F_{\log ns}$ is improved, in particular for the RILM model. This is reassuring given the potential for an additive error correction model (the “second pass”) to produce unreasonably low low-flows. $F_{ns}$ scores are improved, again substantially so for the RILM model. This too is a positive outcome because the multiplicative correction (the “first pass”) might have produced unreasonably high high-flows.

$F_{bias}$ is improved, so much so for the na"ive baseline that it outperforms the uncorrected results from the full rainfall-runoff models (the solid black line is higher than the blue and yellow dashed lines in middle right figure 6.1). In aggregate though, the 4-measure score is consistently improved although the results are not enough to change the relative rankings of the models (i.e. GR4J outperforms AWBM outperforms RILM if error correction is applied to all models or none).

The lower right figure of 6.1 shows the change in skill of including error correction (positive means error correction is beneficial) for the calibration (solid line) and validation (dashed line) periods. This shows that the benefits of error correction are greater in the validation period than in the calibration period. This result is expected given the relative lack of bias in the calibration period. However, the result is also encouraging in that the optimization of $\alpha$ and $\beta$ on the calibration dataset did not result in overfitting.

For ancillary measures (Figure 6.2, 6.4), $F_{boxcox}$ and $F_{viney}$ results largely confirm the above. Interestingly, the hydrograph shape skill scores are changed. For the rising limb density ($F_{RLD}$), RILMCL is substantially improved with error correction and the others are unaffected. For falling limb density, RILM is very much improved, AWBM stays the same and GR4J is deteriorated with error correction. Closer inspection of the GR4J results (figure 6.4) shows that the median change in skill was zero with nearly as many catchments improved as deteriorated with error correction. Although the average change was positive (error correction improves) there were more catchments with moderate decreases in skill ($\sim$0.15) than moderate increases). Visual inspection of hydrographs showed an increase in small flutters in the error corrected time series during periods of low flow, meaning that error correction produces time series (for GR4J) that are spikier than the observed.
Figure 6.1 Model performance for raw model (dashed) and with error correction (solid). In subplots 1-5, the validation results are shown, coloured by model. The lower right figure shows the difference between using and not using error correction during calibration (solid) and validation (dashed). More positive values means error correction has a beneficial effect. See text for discussion and interpretation.
Figure 6.2 As figure 6.1 but for additional diagnostic skill scores.

Figure 6.3 As figure 6.1 but for the difference between no error correction and no error correction. Negative means error correction improves performance.
6.4. Summary and discussion

“Dual pass” error correction has been shown to be an effective way of removing short and long-term biases in simulations. A simpler model (RILMCL) showed much more improvement than more complex models (AWBM, GR4J), although the difference was not enough to make the simple model with error correction better than the complex model without error correction. This experiment compared good, fair and poor models; future research should test the other models such as SIMHYD and PDM.

The error correction model parameters seem mostly stable throughout time for short-term error correction. In comparison, the long-term error correction parameter tended to be small, likely the result of absence of long-term bias in calibration. The long-term parameter may be ill defined because it is insensitive in calibration. Future research could easily test this by fixing $\beta$ to zero. Such an experiment would quantify the relative benefit from short and long-term error correction in isolation.

In this study, optimal parameters were found through a coarse grid search of the parameter space. It would be possible to include the error correction parameters as part of the Shuffled Complex Evolution optimization routine. However, the above-mentioned difficulties in identifying the long-term error correction parameter may mean that the high precision result of such an optimization may not be that useful. Alternatively, a constant value of the parameter (e.g. 0.6) could be used for all catchments, determining the need for calibration of this parameter. This value may be found by finding the optimal error-correction parameters from the validation period results.

Another future task for this work would be to estimate the decay timescale for the error correction. The autocorrelation of simulation errors may inform the selection of this parameter, although it could also be optimized, as above.
It should be recognized that the error correction routine tested here is only one among many. In particular, it would be useful to test the performance of this routine against other methods, such as ARMA or methods that consider the timing/phase of recent errors. Furthermore, this study assumed that long-term errors were multiplicative and short-term errors were additive; other combinations may be more appropriate.
7. ENSEMBLE STREAMFLOW PREDICTION

7.1. Introduction

The greatest source of uncertainty in many streamflow forecasting applications is unknown future precipitation (e.g. Pagano et al. [2004]; Welles et al. [2007]). This is particularly true in catchments that lack seasonal snowcover and their hydrology is instead rainfall-driven. It is common for hydrologic modelling studies (as was done in chapters 5-6) to assume that future rainfall is perfectly known. However, these results will provide only an upper limit to possible skill in forecasting situations; actual forecasting skill will almost always be lower, sometimes significantly so.

When future precipitation is not known, one has several alternatives. First, one may assume a naive scenario, such as zero precipitation. One may survey the range of possibilities by testing various volumes of precipitation (e.g. assume constant intensity rainfall of 1, 2, 5, and 10 mm/hr over 24 hours). Numerical Weather Prediction models also provide forecasts of precipitation, either as a single deterministic forecast or as an ensemble. Hydrologists may also consult a professional meteorologist to determine a plausible scenario. The latter is commonly used at the Bureau of Meteorology. Unfortunately, these precipitation forecasts would be difficult to reproduce or generate retrospectively.

The United States National Weather Service addressed this issue by using the Ensemble Streamflow Prediction (ESP) method for close to three decades [Schaake, 1978]. According to Day [1985], “The ESP procedure assumes that meteorological events that occurred in the past are representative of events that may occur in the future. Each year of historical meteorological data is assumed to be a possible representation of the future and is used to simulate a streamflow trace.”

The ESP procedure has been widely used for seasonal forecasting although it can be applied to models that run on any length timestep. It has been particularly successful in springtime in snowmelt dominated regions where the initial conditions (snowpack and soil moisture) largely determine the expected volume of flow and the climate scenarios (e.g. temperature) provide a range of possibilities for timing and rate of flow. ESP has been applied in non-snowmelt dominated areas, mostly in combination with seasonal climate (e.g. rainfall) forecasts to determine the likelihood of each ensemble trace [Croley and Hartmann, 1987]. In these non-snowmelt dominated areas, ESP has been used for short-range (i.e. < 15 day leadtime) operational forecasting. For example, four of the nine National Weather Service’s non-mountainous River Forecast Centres use short-range ESP, sometimes conditioning the ensembles on weather forecasts. Regrettably, this use is poorly documented in the research literature (Kevin Werner, National Weather Service, personal communication 3 July 2010).

The remainder of this chapter quantifies skill of the ESP method on selected Australian study catchments.

7.2. Retrospective forecasts

Ensemble forecasts are generated retrospectively using the ESP method. First, the model is initialized to the “current” date. For example, if “now” is 15 June 1983, the model will forced with observed precipitation and evaporation from 1 January 1974-14 June 1983. The model state is saved on 15 June 1983. The model is the forced with historical precipitation and evaporation data from 15 June 1974-30 June 1974 (15 days). This will generate one ensemble member of the forecast. Next, 15 June 1983’s model state is paired with historical precipitation and evaporation data from 15 June 1975-30 June 1975, 1976, 1977 and so on until 2006. In total 33 ensemble members will be generated, each forecasting 15 days into the future.
This process is repeated for many initialization points in the historical record. Generating these retrospective forecasts can be extremely computationally expensive. Therefore, only a subset of 16 catchments was used to generate a serially complete set of retrospective forecasts (starting every day from 1979-2006, 10,197 forecasts per catchment, each with 33 members). These catchments were selected through a subjective process that favoured a diversity of climates and a range of model skills. An additional 112 catchments were randomly selected from the remaining 224 to have start dates staggered every 5 days (e.g. 1 Jan 1979, 6 Jan 1979, 11 Jan 1979…, resulting in 2,043 forecasts per catchment, each with 33 members).

In this study, the daily mean of the ensemble was calculated and used as a deterministic forecast, to facilitate easier comparison with the results from previous sections. However, if so desired, one could easily create a probabilistic forecast from the ensemble and this would contain more information than the deterministic forecast. The skill scores from section 3.1 were used.

Given the computational expensiveness of retrospective forecasts, only one model and parameter set was run. This was the GR4J model without PDM-style catchment routing or a rainfall multiplier (the justification for this configuration is provided in section 5.4). The model used was calibrated for the second half of the period of record. This will result in a slight inflation of skill during part of the record. Also, the ensemble contains the “simulation” trace, the trace of observed rainfall during the forecast year. Although the result is averaged in with 32 other ensemble members and will contribute to less than 3.1% of the signal, it too will lead to a slight inflation of skill [Franz et al., 2003]. Although the setup is less than ideal, the results are expected to be at least qualitatively informative about relative spatial patterns of skill. Note that only one parameter set was used and therefore earlier discussions about stitching and averaging of calibration and validation are not relevant here.
Forecast users often are vulnerable to errors in terms of real-world units (e.g. the difference in flow between the forecast and observed). However, the scientific perspective is often more one of wanting to know the information content of the forecast above and beyond a naïve baseline. For example, a scientist may be unimpressed that a hydrologic model was able to reproduce the seasonal cycle of streamflow for catchments with a strong seasonal cycle to rainfall. It is more challenging to adequately capture the interannual variability of flows due to soil moisture variability.

The ESP procedure provides the model with an ensemble of future rainfall traces that properly represent the seasonal cycle of rainfall. Therefore, it would also be interesting to evaluate the characteristic memory timescales of the model that are associated with soil moisture. In addition to calculating skill scores of forecasts, segregated by forecast lead-time, the next section will also subset the forecasts by the month of the year of the target period (e.g. all forecasts for January will be pooled together). In the latter case, the model is not rewarded for simply distinguishing the wet season from the dry season, but rather it is measured by its ability to forecast the relative anomaly from the seasonal cycle.

7.3. Results

Figure 7.2 shows the skill of a small selection of catchments versus lead-time, in the context of the skill of the broader collection of 112 catchments mentioned in section 7.2. As expected, forecast performance of the ensemble mean is lower than simulation performance and forecast performance diminishes with lead-time.

The exception is Dalrymple Creek whose 1-day ahead skill is lower than the 3-day ahead skill, particularly for the $F_{bias}$. One would normally expect bias to be independent of forecast lead-time. It may be relevant that Dalrymple is a highly variable ephemeral stream (zero flow on 44% of days, coefficient of variation 9.66). For the correlation score, Dalrymple has the expected smoothly diminishing decline of skill versus lead-time. For all catchments the skill versus lead-time relationship is largely flat beyond 7 days.

What is interesting to note is that the relative ranking of forecast skill at 3 days is fairly different from the relative ranking of simulation skill. For example, Mount Pleasant Creek (bright pink line) is a strong performer at simulation skill, but it has very low skill at even a 1-day ahead forecast. In comparison, Morass creek (red line) is below median for simulation skill, but has relatively high skill at longer lead-times. There is not an obvious relationship between catchment area and forecast skill at lead-time among the selected subset or the broader collection of 112 catchments. For example, the catchment with the longest skill “tail” (Pirron Yallock) is also the smallest in terms of area. This result may not necessarily hold for very large catchments where the time-to-concentration is longer than the model time-step.
Figure 7.2 Skill of retrospective ensemble forecasts by lead-time (x-axis). “Sim” is the performance of the simulation in which rainfall is perfectly known. Skill is defined as the 4-skill score average, as used in chapter 5. The gray background (black dashed line) is the 25-75% exceedence range (median) of the forecast skill for the entire collection of 128 catchments.

Figure 7.3 shows the performance of the forecasts at various lead-times for 128 catchments (16 serially complete, 112 issued every 5 days, see section 7.2, figure 7.1). This model was calibrated on the second half of the period and a consistent parameter set used throughout and therefore the simulation skill (black line) in figure 7.3 combines both calibration and validation and therefore is not directly comparable to any other line on plots in chapters 5-6 although it is most similar to the blue line on figure 5.1.

The distance between the red and black line of the upper right of figure 7.3 shows that skill drops very low for even the shortest lead-time forecast. This result makes sense in that the largest peaks are driven by large coincident rainfall and the ESP forecasts would generally be conservatively low in those situations. $F_{ns}$ skill has mostly stabilized by day 3 lead-time. Clearly rainfall forecasts would be necessary for skilful short-term (i.e. lead-time 2-7 day) flood forecasts for catchments of this size.

In comparison, $F_{\log ns}$ (upper left figure 7.3) and $F_{boxcox}$ (figure 7.4) skill drops very slowly with lead-time and continues to drop beyond day 15. This may be due to the persistent nature of baseflow conditions. Interestingly, the drop in $F_{bias}$ (middle right figure 7.3) at the shortest lead-time (red line) and eventual recovery at longer lead-times suggests that the earlier results at Dalrymple are not necessarily an anomaly.
Figure 7.3 Performance of retrospective ESP forecasts. The performance is shown by lead-time ("Sim" assumes perfect knowledge of rainfall). In the lower right corner, the performance of the simulation is subtracted from the performance at various lead-times (negative means lower skill at longer lead-time).

Figure 7.4 As figure 7.3 except for the additional diagnostic scores.
The relative lack of change in forecast skill between lead-times of 3 and 15 days suggests that the forecasts appear skilful simply because they are reproducing the seasonal cycle of streamflow (perhaps by reflecting the seasonal cycle of rainfall). One would not need a rainfall-runoff model to make such forecasts as they could instead be derived from historical streamflow averages. However, these forecasts would not necessarily be informative and could be considered naïve. The remainder of this section will test the forecasts against more challenging baselines that account for streamflow seasonality.

Figure 7.5 shows the seasonal cycle of the water balance for a catchment at the far northeast of the study area (Oxley) in Southeast Queensland and southwestern Victoria (Pirron Yallock) near Lake Corangamite. Both of these are among the 16 catchments with serially complete retrospective forecasts.

Oxley is a wet catchment with autumn rains much in excess of potential evaporation. The catchment is drier in spring, with rainfall (and the ratio of rain to evaporation) reaching a minimum in September. The catchment is efficient at producing runoff, with a maximum in March and a minimum in October. Although October-January precipitation is high, some of this is likely lost to soil moisture recharge and evaporation.

Pirron Yallock is a drier catchment with precipitation reaching a seasonal maximum in late winter and a minimum in late summer. The catchment is inefficient at producing runoff, with a precipitation to runoff ratio less than 20%. Late summer average runoff is close to zero and nearly every year the river goes dry between mid-February and late April. Similar to Oxley, the first few months of the rainy season (in Pirron Yallock’s case, April-July) are not as efficient at producing runoff as the months near the end of the rainy season (August-October).

Figure 7.6 shows the skill scores for forecasts pooled by month of year of the forecast target for Oxley. As earlier, forecast skill diminishes with lead-time, with $F_{ns}$ diminishing very quickly and $F_{log ns}$ diminishing slowly. The correlation between forecast and observed (lower right plot) is highest at long lead-times in August when the catchment is near the end of the dry season. The correlation between simulated and observed is still high during the late summer when runoff is high, but the skill of forecasts is low, suggesting runoff driven by coincident rainfall and highly dynamic catchment conditions. Oxley has no issues with forecast bias.
Figure 7.5 Seasonal climatology (1974-2006) of Oxley at Eungella (top) and Pirron Yallock (bottom). Shown are monthly average precipitation (green), potential evaporation (red) and runoff (blue). Precipitation and potential evaporation are serially complete whereas runoff is for the available non-missing months of data.

Figure 7.7 shows the skill of the Pirron Yallock by month. In figure 7.3 the Pirron Yallock had both very high simulation and forecasting skill, especially at long lead-times. Figure 7.7 suggests that this was largely due to the strong seasonal cycle of runoff. Once that factor has been removed, skill is lower and is negative in some months. By some measures it is comparable or worse than Oxley at long lead-times. In particular, this catchment suffers from bias issues during the dry season. This is in part because GR4J will always produce at least some small non-zero flow (like all the other rainfall runoff models considered here) and in part because $F_{bias}$ becomes relatively unstable when observed flow approaches zero.
Figure 7.6 Oxley forecast skill by month. See text for details.

Figure 7.7 Pirron Yallock forecast skill by month. Compare with figure 7.6.
Figures 7.8-11 show the geographic patterns of model performance for simulation and at various forecast lead-times. The performance metric is the 4-measure skill score average, as used in figure 7.2. In all graphs, blue colors indicate good performance of the model (scores closer to 1) and deep reds indicate poor performance (scores closer to 0). The metric is a skill score, the improvement of the forecasts over a naïve baseline. In this case, the baseline is the average observed of the collection of pooled forecasts.

Results that include the entire collection of forecasts (year-round) are shown in the top subplots. Only forecasts with a target period in March are included in the middle plots and only forecasts for September are shown on the bottom. One should not misinterpret “forecasts for September” as meaning that these are seasonal (long-range) forecasts with a month-long target period. The model time-step is always daily and the performance is calculated on a daily interval. For example, a forecast issued at the start of August 28 with a lead-time of 7 days has a target date of September 3. When pooling September forecasts together, this forecast would be included with other forecasts whose target falls in September. Forecasts issued, for example, September 10 for September 6 would be included whereas forecasts issued September 28 for October 4 are not included in the mix (because the target is in October).

March is the end of the summer in Australia and it follows a period of high evaporation. It is a period of high rainfall along the east coast, north of Sydney. September is the end of winter, when evaporation is low. It is relatively dry for most of the basin but it is the wettest time of the year for western Victoria. The area north and east of Melbourne in eastern Victoria has a fairly flat seasonal cycle to rainfall.

Figure 7.8 shows the performance of the models in simulation. When all the months are pooled together (top of figure 7.8) nearly all the catchments perform very well. When the seasonal cycle is removed, simulations in March perform mostly fair, with strong performance in Northeast New South Wales. In September, the simulation performance is uniformly good, perhaps relatively strongest in central Victoria.

The picture for 1-day ahead forecasts (figure 7.9) is significantly different. All-year forecast performance is typically about 0.2 worse than simulation, which is about a 25% drop in skill. Forecasts are best in Victoria and less skilful in New South Wales. When the seasonal cycle is removed, however, March forecasts in western Victoria and Queensland have negative skill. September forecasts are better in New South Wales.

For 3-day ahead forecasts (figure 7.10) March forecasts have negative skill in many areas and only perform well in eastern Victoria and southern New South Wales. September 3-day ahead forecasts have uniformly mild skill. For 7-day ahead forecasts (figure 7.11), the skill of the forecasts pooled together annual is primarily due to the seasonal cycle of flow. There is practically no skill in predicting March flow 7 days head although there is some skill in September, particularly in New South Wales.
Figure 7.8 Simulation performance of GR4J. See text for discussion.
Figure 7.9 Performance of ESP forecasts with 1 day lead-time. See text for discussion.
Figure 7.10 Performance of ESP forecasts with 3 day lead-time. See text for discussion.
Figure 7.11 Performance of ESP forecasts with 7 day lead-time. See text for discussion.
7.4. Summary and discussion

Hydrologic models are commonly evaluated on their ability to simulate historical flow. This evaluation assumes that rainfall is known with complete confidence. In real-world forecasting situations, future rainfall will be uncertain and actual forecasting skill may be much less than simulation skill. This chapter used a standard operational technique (the ESP method) to assume that future rainfall is uncertain but is likely to be similar to what was seen in past years. Retrospective forecasts from 1979-2006 were generated, each with 33 ensemble members. This was done to test and validate real-time forecast accuracy. Generating these retrospective forecasts is computationally expensive, so only a subset of 16 catchments had forecasts issued every day while half the remaining (112) catchments had forecasts issued every 5 days.

When compared to a baseline that does not account for the seasonal cycle of streamflow, the forecasts appear very skilful, even at long lead-times. Some catchments did see a significant loss of skill going from simulation to 1-day ahead forecasting, indicating that the runoff responds to rainfall on the same day the rainfall occurs (adjusted for the daily timestamp shift between the rainfall and runoff data, which is typically 15 hours).

When the results were pooled by month, the simulation performance was not stable across the year. Furthermore, it appears much more difficult to beat baselines that account for the seasonal cycle of streamflow. In particular, +3 day ahead forecasts in March had negative skill in many locations.

This initial foray into forecasting has many limitations, as follows;

- The ensemble forecasts were distilled down to the ensemble mean and the forecasts were evaluated deterministically. Therefore, one could not make any statements about the probabilistic reliability of the forecasts. Future studies should use probabilistic evaluation measures, such as the ranked probability skill score, reliability diagrams, and rank histograms.
- The ESP method assumes near total uncertainty about future rainfall. In real-time operations, forecasters would have quantitative and qualitative guidance from meteorologists and Numerical Weather Prediction models. Assuming that the weather forecasts are skilful and free from systematic biases, this study represents a lower limit of possible streamflow forecasting skill.
- The collection of study catchments was limited. To be representative of operational conditions, one would need to include more catchments that have poor/short data records, are in different climates/states and are of different sizes. In particular, larger catchments are likely to have greater skill at longer lead-times than small catchments. Semi-distributed (as opposed to spatially lumped) models should be used for these large catchments.
- No data assimilation or error correction was used. Chapter 6 showed that simple error correction can have cross-validated benefits in simulation skill, especially during low flow conditions and for relatively simple models. If error correction was applied to the forecasts, their skill may improve. It is unknown how far into the forecast horizon these benefits will be felt, however. It is also unknown if a data assimilation/error correction method that was calibrated on historical simulations would perform well in forecasting model. Consider, for example, figure 7.3 and how forecast bias depended on forecast leadtime.

At the very least, this study demonstrated the benefit of evaluating the models from a forecasting perspective. Furthermore, the results confirm the findings of Hashino et al. [2006] who wrote:
“Clearly, the speculation that months with high relative accuracy or low biases in the historical simulation will have similar characteristics in their probabilistic forecasts is not true. Indeed, due to biases, some months have very little probabilistic forecast skill, despite having good skill in simulation mode.”
8. FINAL SUMMARY AND RECOMMENDATIONS

This research evaluated the performance of a variety of rainfall runoff models in a number of configurations. These experiments were done using double split sample evaluation because this method removes the effects of climate non-stationarity, leading to a more reliable interpretation of the results.

For practically every measure, the cumulative distribution of skill of GR4J was greater than every other model considered. GR4J rises to the task of performing well in the measures it was asked to perform well in during calibration, it sustains this performance better than any other model into validation, but most remarkably, it also performs among the best of all the models at the ancillary skill scores that were not used during calibration. Catchment routing had minor impact on lumped models that already had mechanisms for delaying of surface runoff, although it provided a significant boost in performance for the naïve baseline models. Similarly, the use of a multiplier to rescale precipitation had little benefit for complex rainfall-runoff models. However, the multiplier was useful for the naïve baseline model. Despite these last two findings, catchment routing and rainfall multipliers may still be useful for semi-distributed simulations.

A relatively simple weighted average was used to combine the results from multiple models. The weighting was based on the performance of the individual models in the calibration period. The combined result was better than the individual models that performed fair, but was not substantially better than the best individual model.

A (“dual pass”) error correction method that considers quickly and slowly varying errors was effective at improving simulation performance. This is a novel approach and showed the most improvement during low flow conditions. The method’s quickly-varying bias (“second pass”) correction factor was more reliably estimated than the slowly-varying bias (“first pass”) factor. The improvement to the complex model was much less than the improvement to the naïve baseline, although the final performance of uncorrected complex model was still better than the corrected naïve baseline.

The Ensemble Streamflow Prediction method was used to generate retrospective forecasts. Forecast skill diminished with lead-time, sometimes rapidly. A fair amount of the apparent forecasting skill was due to the model’s ability to reproduce the seasonal cycle of streamflow. When compared to a more challenging baseline that considered the seasonal cycle, many catchments did not have forecast skill beyond +3 days ahead.

These findings lead to the following recommendations;

**Recommendation 1:** Due to its parsimony and strong performance, GR4J should be the model of choice for future modelling experiments. For daily time-step lumped simulations, GR4J without catchment routing or a rainfall multiplier is sufficient.

**Recommendation 2:** Calculating skill scores on the joined (“stitched”) time series from the double cross validation was found to mask biases. Averaging the skill scores from the individual periods overcame this artefact. The latter method should be used in all future studies.

**Recommendation 3:** Continue to improve the realism of the naïve baseline models. An operationally realistic baseline model would include state updating and possibly error correction. Like the other models, the baseline performance must be evaluated in cross validation.

**Recommendation 4:** Continue to develop the “dual pass” error correction method. Compare “dual pass” error correction method to other methods such as ARMA or the Mike-11 phase-correction method.

**Recommendation 5:** Retrospective forecasting is a very useful tool for understanding forecasts and such experiments should be strongly encouraged and supported. Future studies should evaluate the probabilistic reliability of the ensemble forecasts. The benefit of
using weather forecasts should be compared against the ESP method. Error correction/state updating should be done in conjunction with retrospective forecasting while being careful to adequately validate the results.

**Recommendation 6:** SWIFT should implement additional performance evaluation measures that do not reward for simply reproducing the seasonal cycle of streamflow. These may include persistence indices or measures whose baseline varies by the day of the year.

**Recommendation 7:** It would be interesting to know if skill-lead-time relationships are related to catchment characteristics, such as catchment size or seasonality of rainfall and potential evaporation. Large-sample retrospective forecasting experiments could be used to conduct “top down” studies of the correlation between skill and catchment characteristics. “Bottom up” studies could also be done, for example, studying the structure of models, deriving relationships between model parameters and expected predictability. Toy models may be useful for the bottom up studies. To the authors’ knowledge, such research has not been done before in Australia or overseas.
REFERENCES


