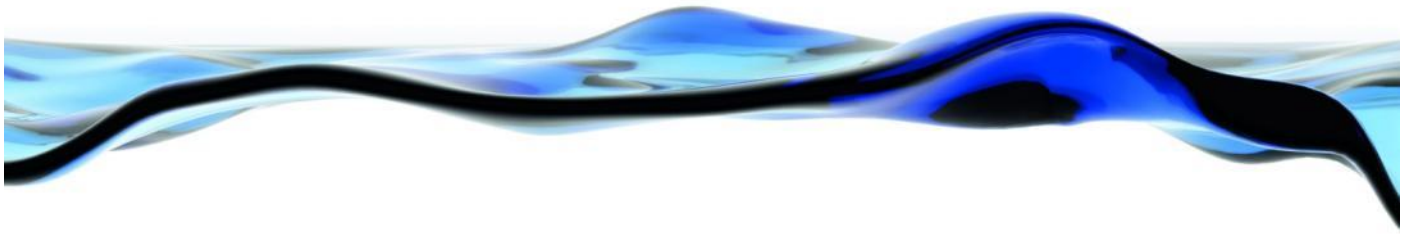
Impacts of Climate Change on Surface Water in the Onkaparinga Catchment

Final Report Volume 1: Hydrological Model Development and Sources of Uncertainty

Westra, S., Thyer, M., Leonard, M., Kavetski, D. & Lambert, M.



Goyder Institute for Water Research

Technical Report Series No. 14/22

www.goyderinstitute.org

Goyder Institute for Water Research Technical Report Series ISSN: 1839-2725

The Goyder Institute for Water Research is a partnership between the South Australian Government through the Department of Environment, Water and Natural Resources, CSIRO, Flinders University, the University of Adelaide and the University of South Australia. The Institute will enhance the South Australian Government's capacity to develop and deliver science-based policy solutions in water management. It brings together the best scientists and researchers across Australia to provide expert and independent scientific advice to inform good government water policy and identify future threats and opportunities to water security.

The following Associate organisations contributed to this report:

Enquires should be addressed to: Goyder Institute for Water Research
Level 1, Torrens Building
220 Victoria Square, Adelaide, SA, 5000
tel: 08-8303 8952
e-mail: enquiries@goyderinstitute.org

Citation

Westra, S., Thyer, M., Leonard, M., Kavetski, D. & Lambert, M., 2014, *Impacts of Climate Change on Surface Water in the Onkaparinga Catchment – Volume 1: Hydrological Model Development and Sources of Uncertainty,* Goyder Institute for Water Research Technical Report Series No. 14/22, Adelaide, South Australia.

Copyright

Disclaimer

The Participants advise that the information contained in this publication comprises general statements based on scientific research and does not warrant or represent the completeness of any information or material in this publication.

Table of Contents

List of Figures

Executive Summary

This is the first of three final volumes for the University of Adelaide component of *Task 4: Application Test Bed.* The application test bed is the fourth and final task in the Goyder Climate Change project, with the overall project aim being to develop a benchmark suite of downscaled climate projections and climate variable time series for South Australia. The specific contribution of Task 4 is to apply downscaled data generated in Task 3 in a series of hydrological test cases, and provide regular feedback on the downscaling activity throughout the project lifecycle.

This first report examines the principal sources of streamflow predictive uncertainty in the Onkaparinga catchment, including the relative contribution of errors in the model inputs, outputs and structure. The information is used to improve the model structure, with 22 alternative hydrological model structures identified and compared. Four of these models are selected to form the basis of the climate change projections in the Onkaparinga catchment.

Streamflow predictive uncertainty analysis

Streamflow projections under a warmer future climate typically are highly uncertain. This uncertainty arises due to a complex modelling chain that commences with scenarios of global future greenhouse gas emissions, which are then modelled using global climate models to produce estimates of future global and regional climate variables. These variables are then downscaled using one or more downscaling models to produce estimates of hydrometeorological variables (e.g. precipitation, temperature, humidity, wind) at the catchment scale, and finally the hydrometeorological variables are converted to streamflow using one or more rainfall-runoff models. This uncertainty is often conceptualised as a 'cascade', with each source of uncertainty influencing subsequent steps of the modelling chain.

The first volume of this report focuses on uncertainty due to the rainfall-runoff transformation for the Onkaparinga catchment, with uncertainty due to the emission scenarios, global climate models and the downscaling algorithm addressed in the third volume of this series. The purpose of the analysis of rainfall-runoff model uncertainty is to improve the reliability of hydrological projections under a warmer future climate. Outcomes were as follows:

- Input uncertainty, particularly the uncertainty associated with deriving spatial rainfall estimates based on gauges and radar, was found to be an important source of uncertainty for medium and high flows, but less so for low flows. The magnitude of input uncertainty was similar to the magnitude of uncertainty captured in the residual error model.

- Output uncertainty was moderate based on a comprehensive rating curve analysis. The likely reasons are the relatively stable rating curves and high number of streamflow gaugings for each of the streamflow sites. Timing issues when removing Murray pipeline flows from the recorded flows at Houlgrave Weir were addressed by adopting a censoring approach during model calibration, to ensure that the calibration procedure adopted to estimate the final parameter sets was not affected by timing errors.

- Parameter uncertainty was small based on a simulation of the joint posterior distribution of the parameters, indicating that the record length is sufficient relative to the model complexity to enable precise estimation of model parameters.

- Structural uncertainty (the uncertainty due to the hydrological model structure not being able to reflect true flow behaviour) was identified as an important source of total predictive uncertainty. This was based on a detailed analysis of model diagnostics including flow duration curves, the rising and falling limb of the hydrograph, and information-theoretic measures that assess the non-stationarity of hydrological model parameters.

Future climate change projections are usually developed relative to a climatological 'baseline' derived from historical simulations of rainfall and PET. Therefore, in evaluating the role of the above sources of uncertainty in developing future climate change projections, it should be noted that the main purposes of observational (instrumental) data are to (1) select one or several hydrological models that faithfully represent the historical rainfall-runoff relationship, and (2) estimate the hydrological model parameters for each hydrological model. As a result, investigations into the suitability of the hydrological model structure are critical, particularly when extrapolating outside the domain over which the model has been calibrated. Therefore the emphasis of the remainder of the report is on understanding model structural errors that are likely to affect future climate change predictions.

Improving hydrological model structure

GR4J was selected as the hydrological model for this analysis, based on its parsimony and the extensive testing of this model over a wide range of catchments. Investigations of the capacity of GR4J to simulate flows found deficiencies particularly in the simulation of hydrograph recessions, as well as the identified non-stationary of parameter θ_1. Therefore, in addition to the standard GR4J, 21 alternative model structures were developed to address some of the limitations of the standard model. Model structures included various combinations of the following:

- Sinusoidal variation in θ_1 with a period of one year;
- Allowing θ_1 to vary as a function of the previous 365-day rainfall and PET;
- Allowing θ_1 to vary as a function of a linear trend;
- Inclusion of an additional parameter that controls the proportion of net rainfall that enters the production store; and
- A modification to the way that actual evapotranspiration is estimated in the model.

The models were evaluated using the Akaike Information Criterion (AIC) based on a Gaussian heteroscedastic likelihood function with a low flow threshold to censor flows. This likelihood function performed well in capturing the distribution aspects of the flows, and thus the estimated residual error model parameters obtained from this function could thus be used to estimate predictive uncertainty limits for streamflow projections.

In addition to the AIC, model diagnostics used included the Nash-Sutcliffe coefficient of efficiency (NSE), the annual flow volume, and the various quantiles of the flow duration curve. The standard GR4J model and 21 alternative models were tested over an exploratory period (from 1977, 1985 and 1993 in Houlgrave Weir, Scott Creek and Echunga Creek, respectively, up to 1999 for all sub-catchments) and a drier confirmatory period (from 2000 to 2009 for all sub-catchments). The modified models all showed improvements over the standard GR4J model, with the most notable improvements being attributable

to simulating the seasonal cycle for θ_1 and the inclusion of an additional parameter that controls the proportion of the net rainfall that enters the production store.

Compared to the standard GR4J model that overestimated flows in the confirmatory period by 17%, the 'AIC-best' model ($g_{3.11}$) underestimated flows by only 2.6%, representing a significant improvement in the simulation of the catchment water balance. Furthermore, models that accounted for a sinusoidal variation in θ_1 were much better in representing the flow duration curve in autumn (typically referred to as the 'wetting' season when the commencement of the rainfall season fills catchment stores without leading to a significant increase in runoff), whereas the inclusion of the additional parameter that controls the proportion of net rainfall that enters the production store also led to a significant improvement in the model's representation of the observed flow duration curve.

Hydrological model selection

The 22 potential model structures were divided into three sets: (1) a set of models ($g_{1.1}$,…, $g_{1.8}$) that included different combinations of time-varying covariates to simulate GR4J parameter θ_1; (2) a set of models ($g_{2.1}$,…,$g_{2.4}$) that included different modifications to the GR4J model structure; and (3) a set of models ($g_{3.1}$,…,$g_{3.12}$) that included different combinations of models from the first two sets. An ensemble of four hydrological models was then selected for the development of future climate change projections for the Onkaparinga. These models comprised the standard GR4J model as well as the best model in each of the three sets, and were selected to obtain a diversity of model structures to simulate hydrological model uncertainty. The specific models are:

- Model $g_{1.1}$ (the standard GR4J model) as a benchmark against which other models can be evaluated;
- Model $g_{1.8}$, which was the best model in the first set, and accounts for non-stationarity due to seasonal variability, the effect of the previous 365-day rainfall and PET as well as a linear trend in the capacity of the production store;
- Model $g_{2.2}$, which was the best model in the second set, and includes an additional parameter that controls the portion of rainfall that enters the production store; and
- Model $g_{3.11}$, which was the best model in the third set, and accounts for non-stationarity due to seasonal variability, the effect of the previous 365-day rainfall and PET, a linear trend in the capacity of the production store, as well as an additional parameter to control the portion of rainfall that enters the production store.

Models $g_{1.8}$ and $g_{3.11}$ incorporate the effects of a linear trend, and rather than extrapolate this linear trend into the future, the contribution of this predictor at the end of the calibration period (31 December 1999) is held constant for future simulations. Although it is not possible to attribute the trend to a particular feature of catchment change, it is likely that at least part of the trend is attributable to an increase in on-farm dams. Given this, the decision to fix the trend parameter at the 1999 value is appropriate because changes to the regulation of on-farm dams have limited the increase in total storage since the early 2000s.

The four hydrological models will form the basis of future streamflow projections in the Onkaparinga catchment. The due hydrological model uncertainty will be combined with GCM and the representative concentration pathway (RCP) uncertainty, to provide a thorough exploration of the uncertainty associated with climate change projections. This is discussed in the third volume of this report.

1 Introduction

This is the first of three final reports for the University of Adelaide component of *Task 4: Application Test Bed* for the Goyder Climate Change project. The overall Goyder Climate Change project aims to develop a benchmark suite of downscaled climate projections and climate variable time series for South Australia. The contribution of Task 4 is to apply the downscaled data in a series of hydrology test cases to provide iterative feedback on the overall downscaling activity throughout the project lifecycle.

The Onkaparinga catchment has been identified as the primary case study location for this project. The catchment was selected because of the availability and quality of observational data and its importance as a water supply catchment for the Adelaide region. The University of Adelaide component of Task 4 involves applying the rainfall-runoff model 'GR4J' [*Perrin et al.*, 2003] to three sub-catchments in the Onkaparinga: Houlgrave Weir, Echunga Creek and Scott Creek. Each of these sub-catchments has long records of historical daily flows, and collectively they represent the majority of the flow volume in the Onkaparinga upstream of the Happy Valley diversion. This enables identification of one or more suitable hydrological models for the region, as well as an assessment of the magnitude of uncertainties associated with the transformation from rainfall to runoff.

The selected rainfall-runoff models can be used to test the downscaled hydrometeorological forcing variables (rainfall, temperature, radiation, humidity and pressure) in terms of their capacity to simulate flows derived using the instrumental record of rainfall and potential evapotranspiration (PET). The implications of future climate change on flows in the three sub-catchments can then be evaluated.

The work has been divided into the following three reports:

> **Report 1 (this report):** *Hydrological Model Development and Sources of Uncertainty*. This report focuses on assessing the relative contribution of the principal sources of hydrological model uncertainty: input errors, output errors and model structural errors. The Bayesian Total Error Analysis methodology is used as the basis of the analysis. Findings are used to improve the model structure, and develop a set of models that can be used to produce the climate projections.
>
> **Report 2:** *Hydrological evaluation of Non-homogenous Hidden Markov Model (NHMM) projections*. This report describes the comparison of historical flows in three sub-catchments of the Onkaparinga. Estimated flows are obtained by passing the NHMM projections of rainfall and other meteorological variables through a calibrated hydrological model. A total of five General Circulation Models (GCMs) from the Coupled Model Intercomparison Project Phase 3 (CMIP3) archive, 15 GCMs from the CMIP5 archive and a reanalysis model run are evaluated.
>
> **Report 3:** *Impact of climate change on flows in the Onkaparinga catchment.* This report outlines projections for future flows in the Onkaparinga catchment, for 30-year future time slices centred on 2030, 2050, 2070 and 2085. Attributes of future flows include aggregate annual and seasonal flow patterns, low flows and peak high flows.

2 Overview of this Report

This report describes the basis for the hydrological modelling used in this study. The report draws largely from the first [*Leonard et al.*, 2011] and second [*Westra et al.*, 2012] milestone reports, with additional updates to account for modifications to the processing of instrumental data used as the hydrological model inputs and outputs. A large portion of this report is based on a manuscript that is currently in press in Water Resources Research [*Westra et al.*, 2014a], and this manuscript has been included as Appendix 1 of this report.

In Section 3, a brief overview is provided of the Onkaparinga catchment. This is followed in Section 4 by a summary of the rainfall, PET and runoff data used to support the hydrological models. In Section 5 we describe the principal sources of uncertainty associated with the hydrological modelling, concluding with recommendations for model improvement. Section 6 briefly describes a set of non-stationary versions of GR4J that have been developed for use under potentially changed future climate conditions, and summarises the ensemble of models that were ultimately selected for use in the climate change assessment (report volume 3 of this series). Finally, conclusions are given in Section 7. A manuscript describing a general strategy for diagnosing and interpreting hydrological non-stationarity is provided as Appendix 1 of this report, and this manuscript is currently undergoing peer review in Water Resources Research.

3 Background to the Onkaparinga Catchment

The Onkaparinga catchment is situated to the south-east of Adelaide in the southern portions of the Mount Lofty Ranges, and exits to the ocean near Old Noarlunga. It is a significant source of municipal water for metropolitan Adelaide, and also provides water to farm dams and the environment.

The catchment has a total area of 553 km^2, with a significant elevation gradient ranging from low-lying coastal plains near the mouth at Port Noarlunga to elevations of 700 m in the upper reaches within the Mount Lofty Ranges. Partly due to these orographic features, the catchment has a substantial rainfall gradient, with a median catchment-average rainfall of about 780 mm but ranging from approximately 500 mm along the coast to 1100 mm at higher elevations. The areal potential evapotranspiration calculated using Morton's APET formulation [*McMahon et al.*, 2013] is approximately 1300 mm per year, while the pan evaporation recorded at Mt Bold reservoir is 1560 mm per year [*Teoh*, 2002].

Mount Bold Reservoir has been operational since 1938, and this reservoir diverts water to supply Happy Valley Reservoir. Catchment inflows to the reservoir are supplemented with water pumped from the Murray River, which enters the system near Hahndorf. Houlgrave Weir is situated just upstream of Mount Bold Reservoir, and the flows past this weir are mostly derived from the upstream catchment during the winter months, and from the Murray River during the summer months. In the upper reaches of the catchment there are a number of towns, including Aldgate, Bridgewater, Balhannah, Lobethal, Hahndorf, Stirling, Summertown, Uraidla and Woodside.

A study by Teoh [2002] gives a detailed account of the hydrology in the catchment, focussing on the role of farm dams. Using aerial photography, the report identified 2700 farm dams in the catchment in 1999. Based on an empirical relationship between the dam surface area and volume, they estimated that the dams represent a total storage capacity of 8.5 GL. The report concluded that these dams harvested approximately 4.3 GL of the water entering Mount Bold Reservoir, representing 8% of median annual adjusted flow.

Over the 12 year period from the first aerial photograph in 1987, the farm dam storage volume in the Onkaparinga catchment increased by about 11%, although the rate of increase in some sub-catchments was much greater. For example, in Scott Creek no farm dams could be observed in 1987, yet in 1999 the total volume of dams was 148 ML. Because of these changes, some shifts in the rainfall-runoff relationship in these catchments might be expected. Furthermore, approximately 9.3% of the catchment is irrigated, and the water is partly obtained from groundwater extractions from individual groundwater bores; this is also likely to contribute to the total catchment water balance.

A more recent study by Heneker and Cresswell [2010] covers the issue of climate change impact on the Mount Lofty Ranges. The purpose of the analysis was to determine likely changes to water storage reservoir inflows. Their assessment was based on a statistical downscaling approach (the non-homogenous Markov model; NHMM) in conjunction with existing hydrological models. They used one simulation from a single GCM for each of two emission scenarios, A2 and B2 [*IPCC*, 2000], and focused on the time slice 2035-2065. They concluded that climate change represents a significant risk to Adelaide's water supply and that changing weather patterns could potentially

reduce annual rainfall by 13%, translating to a potential reduction in annual runoff from the Mount Lofty Ranges water supply catchments of more than 30%. The largest changes are expected to be during the autumn and early winter months, with projected rainfall reduced by as much as 25% during this period. Changes in evaporation rates were not modelled as part of their downscaling process.

There is some evidence of a net groundwater export from some of the sub-catchments in the Onkaparinga. For example, estimates of groundwater export provided by the Department for Environment, Water and Natural Resources (DEWNR; pers. comm. Graham Green, 24/09/2012) were 993 ML/year and 1257 ML/year for Scott Creek and Echunga Creek, respectively; this corresponds to 34 mm/year and 32 mm/year catchment-average depth. These numbers are highly approximate, however, and thus are difficult to incorporate directly in water balance computations.

The Onkaparinga catchment is shown in Figure 1. The portion of the catchment of interest for this study – namely the portion that supplies water to Adelaide – is located upstream of the Happy Valley reservoir diversion, and is largely represented by three sub-catchments: Scott Creek, Houlgrave Weir and Echunga Creek. The choice of these three catchments is based on the presence of high-quality streamflow gauges at each of the catchment outlets, and the selection of these gauges is discussed in the following section.

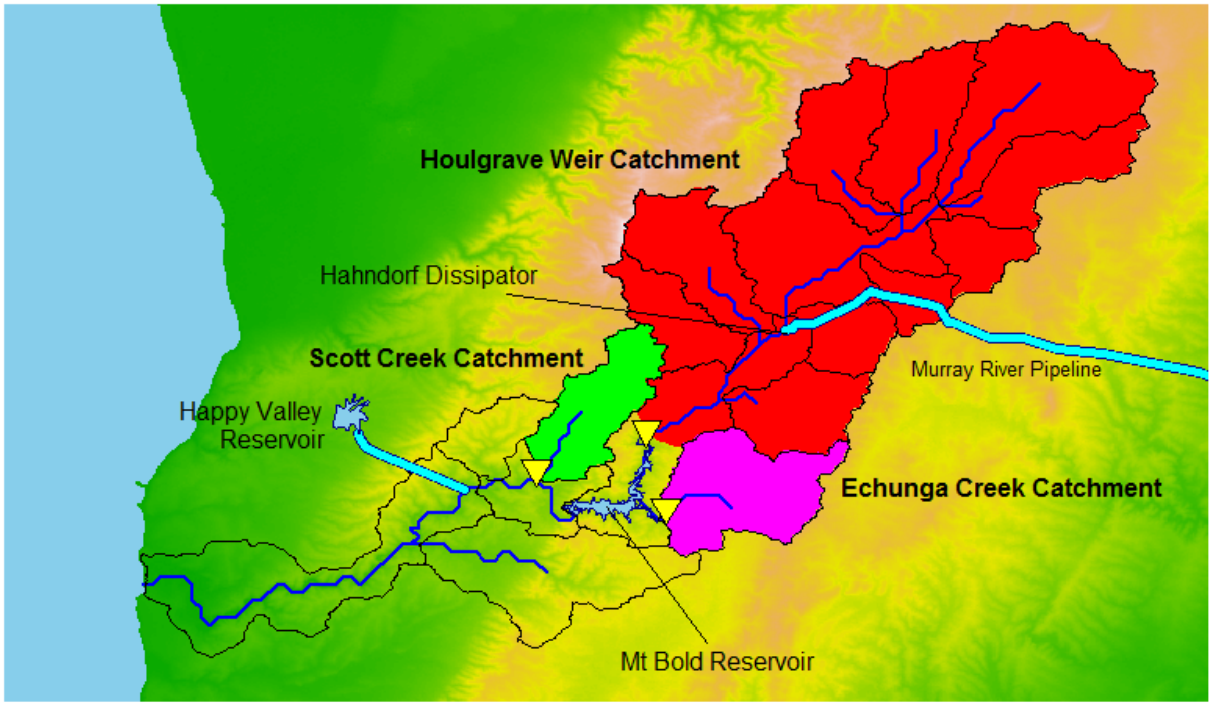


Figure 1: Catchments used in the analysis

4 Data to Support the Hydrological Modelling

The ultimate objective of this study is to assess the implications of anthropogenic climate change on hydrological response in the Onkaparinga catchment. This will be achieved using downscaled sequences based on the non-homogenous hidden Markov model (NHMM) provided in Task 3 of this project, to simulate hydrological response under both historical and future climates.

In this context, observational (instrumental) data is used to calibrate the hydrological model, assess its performance over an independent confirmatory period, and quantify predictive uncertainty. The observational data comprises streamflow gaugings, gauged rainfall and a set of meteorological variables used to compute potential evapotranspiration. These are discussed in turn below.

4.1 Runoff

A total of 24 continuous streamflow gauges were identified within the Onkaparinga Catchment, available from the Surface Water Archive:https://www.waterconnect.sa.gov.au/Systems/SWD/SitePages/Home.aspx.These sites are listed in Table 1 and shown in Figure 2.

Table 1: Summary of streamflow data

ID	Station Location	Start	End	Latitude	Longitude
Q_5030500	Clarendon Weir	20/09/1937	21/02/2011	-35.111	138.635
Q_5030502	Scott Creek	28/03/1969	20/12/2010	-35.101	138.673
Q_5030503	Bakers Gully	12/04/1969	21/12/2010	-35.139	138.607
Q_5030504	Houlgrave	18/04/1973	21/02/2011	-35.082	138.725
Q_5030505	Snow Hill	6/11/1972	19/01/1983	-35.15	138.724
Q_5030506	Echunga	23/03/1973	26/01/2011	-35.127	138.728
Q_5030507	Lenswood	19/05/1972	6/03/2011	-34.937	138.822
Q_5030508	Inverbrackie	18/05/1972	26/04/2010	-34.947	138.926
Q_5030509	Aldgate	14/07/1972	28/04/2011	-35.016	138.731
Q_5030521	Verdun	30/06/1977	4/11/1982	-34.997	138.795
Q_5030522	Noarlunga	28/06/1973	14/02/1988	-35.171	138.52
Q_5030524	Piccadilly Valley - Vince Creek	8/06/1982	1/04/1987	-34.961	138.724
Q_5030525	Piccadilly Valley - Sutton Creek	23/07/1982	4/07/1988	-34.969	138.741
Q_5030526	Cox Creek	24/06/1976	13/04/2011	-34.975	138.734
Q_5030528	DS Mount Bold Reservoir	4/08/1977	7/02/1989	-35.123	138.674
Q_5030529	Burnt Out	13/01/1978	19/12/2010	-35.128	138.704
Q_5030530	Kerber	31/07/1987	7/11/1989	-34.955	138.896
Q_5030531	near Charleston	10/08/1987	1/11/1989	-34.911	138.901
Q_5030537	Hahndorf Ck DS STW	25/03/1993	3/08/2011	-35.02	138.793
Q_5030545	US Scott Creek	9/02/2001	1/06/2009	-35.095	138.681
Q_5031001	Onkaparinga R US Dissipater	22/06/2002	3/08/2011	-35.022	138.791
Q_5031004	DS Clarendon Weir	20/05/2006	13/02/2011	-35.113	138.631
Q_5031005	Old Noarlunga - estuary ford	26/05/2006	30/05/2011	-35.176	138.513
Q_5031006	Woodhouse Wetland Outlet	12/07/2006	3/03/2011	-34.986	138.738

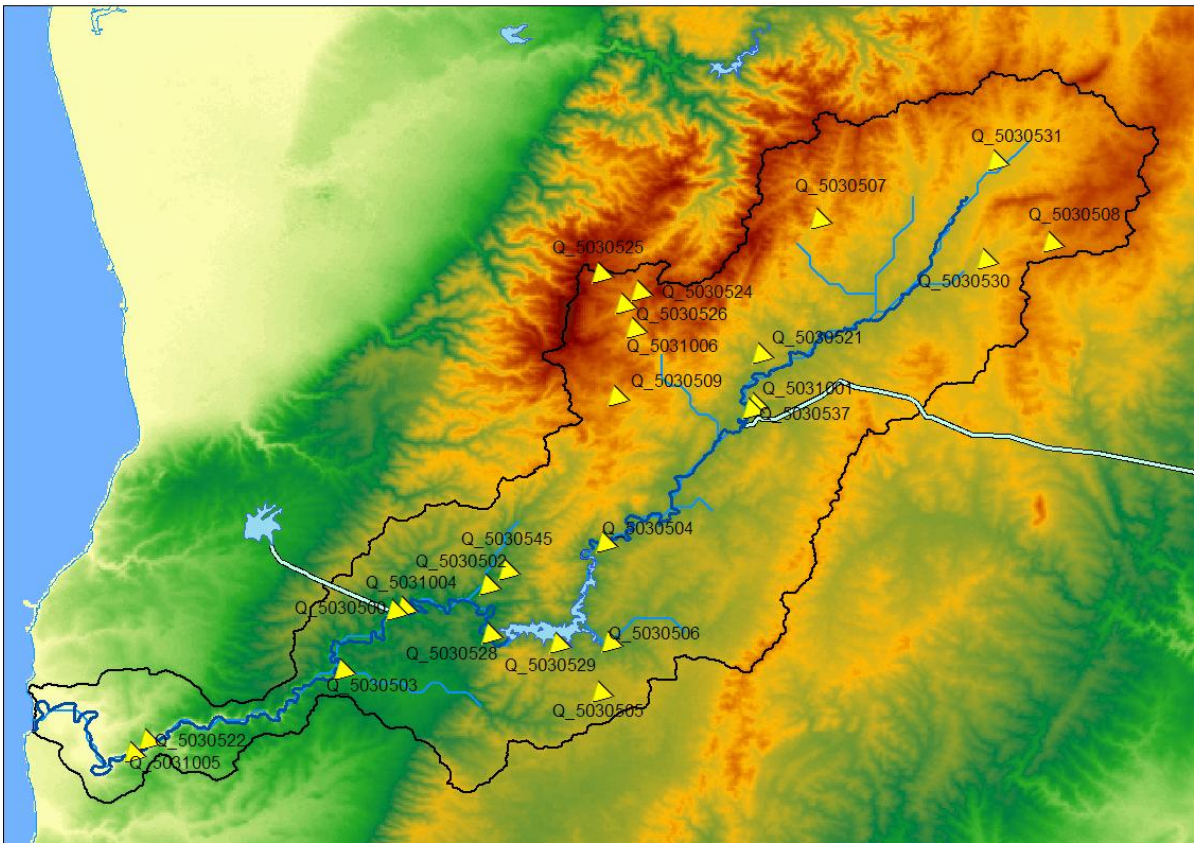


Figure 2: Location of stream gauge sites

Figure 3 illustrates the length and quality of the records across the catchment, and shows that 10 of the gauges have records that are less than 10 years long. In terms of modelling the catchment flows that are available for diversion to Happy Valley Reservoir, the most significant gauges due to their record length and location are Clarendon Weir (5030500), Scott Creek (5030502), Houlgrave Weir (5030504) and Echunga Creek (5030506). Even though Scott Creek is only a sub-catchment of the total area contributing to Clarendon Weir, it is notable for its continuous unbroken record, whereas Clarendon Weir has 2119 missing days over the entire record with 162 missing days since 1985 (the analysis period used in this report). Aside from the missing data, this record would also require corrections to account for the volumes of releases from Mount Bold as well as the diversions to Happy Valley Reservoir. Until recent years of the record, the operational releases were regularly faxed but not archived and therefore cannot be used (pers. Comm. Rob Daly 2011). Moreover, there has been instrumentation issues affecting the accuracy of the flume used to measure the diversion flows to Happy Valley. We therefore use the data from Scott Creek as being representative of flows from the Clarendon Weir sub-catchment. Estimates for the Clarendon Weir catchment could nevertheless be obtained by using the parameter estimates obtained for Scott Creek catchment, but with the Clarendon Weir catchment-average rainfall time series as inputs. This approach is supported by the finding that regionalisation approaches are based on using the parameter estimates from geographically proximate catchments are often competitive with more sophisticated regionalisation approaches that take other catchment characteristics into account [e.g. *Merz and Bloschl*, 2004].

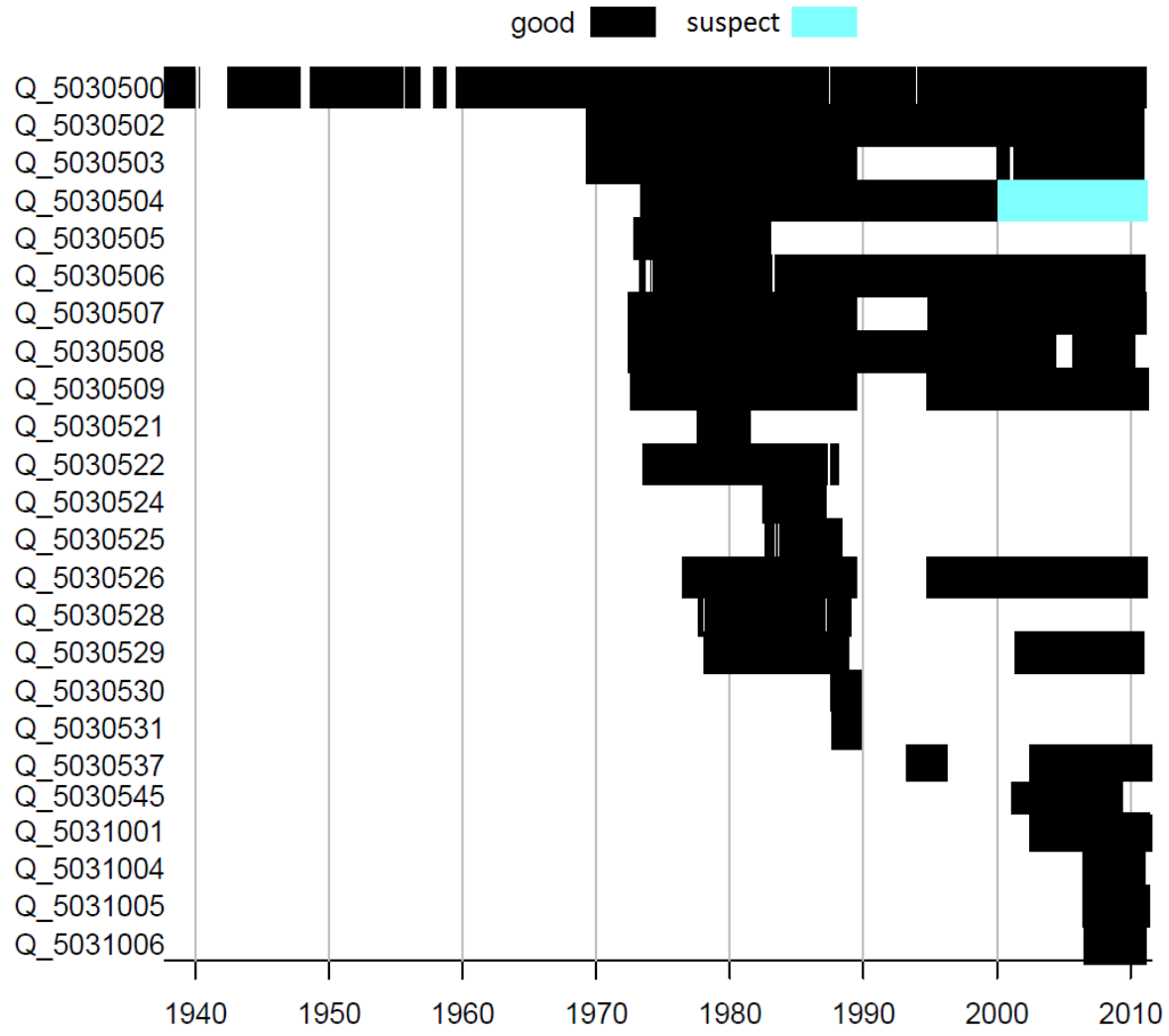


Figure 3: Data quality at streamflow gauges

The runoff estimates from the Surface Water Archive are available as daily totals, defined from 00:00 to 24:00 each day. To be comparable with the daily rainfall data, daily runoff was required from 09:00 to 09:00 each day. Sub-daily runoff measurements in increments of five minutes were therefore obtained for the three selected sub-catchments, and this data was aggregated to daily runoff corresponding to the daily rainfall measurements. The record lengths of the gauges, as well as the contributing area upstream of the gauges, are given in Table 2.

Table 2: Summary of sub-daily streamflow data

ID	Station Location	Sub-catchment Area (km^2)	Start	End
5030502	Scott Creek	29	29/03/1969	02/11/2011
5030504	Houlgrave Weir	323	18/04/1973	03/11/2011
5030506	Echunga Creek	39	02/08/1989	13/06/2011

A timing issue exists at Houlgrave Weir, as flows past this point include both the natural catchment flows plus flows from the Murray pipeline. Hourly data from 1/05/2003 to 29/05/2013 at Hahndorf Dissipator was obtained from SA Water. Using this data, it was found that there was a lag of about 5-7 hours between water leaving the Hahndorf Dissipator (the point at which Murray pipeline flows are measured) and the water arriving at Houlgrave Weir gauge (the point at which the combined flows from the natural catchment and the Murray pipeline are measured).

This lag means that directly subtracting daily flows at the Dissipator from daily flows at Hahndorf Weir will create a timing error. The pre-2003 record of Murray pipeline flows at Hahndorf Dissipator is only available at a daily resolution, and therefore it is not possible to correct this timing error. The timing does not affect aggregate flow volumes at Houlgrave Weir (as an overestimate of the flow rate for one day would be compensated by an underestimate of the flow rate for the following day), but it will affect the individual daily runoff amounts, and this in turn can influence the hydrological model calibration. Nevertheless, it is possible to use information on changes in the flow rate at Hahndorf Dissipator from one day to the next to identify when a timing error is likely to occur. Therefore calibration was performed using the flows at Houlgrave Weir after subtracting Dissipator flows, and censoring days from the calibration that experienced a large change to the flow rate at Hahndorf Dissipator. This censoring was conducted when flows at Hahndorf Dissipator changed by more than a threshold of 0.2 mm in a given day (units of catchment-average runoff depth)

4.2 Rainfall

Teoh [2002] identified 93 rainfall records from the Bureau of Meteorology within and surrounding the Onkaparinga catchment. They then selected a subset of 23 gauges that have long records and are evenly spaced over the region. This set of gauges (Figure 4) was chosen for further investigation. The daily rainfall at these locations was obtained from the SILO Patched Point Dataset (PPD; http://www.longpaddock.qld.gov.au/silo/), which has infilled values for missing and/or accumulated observations [*Jeffrey et al.*, 2001]. The timespan of each gauge is shown in Table 3 with the majority of gauges covering the entire record from 01/1889 to 06/2011. Days are defined as the 24 hours prior to 9 am.

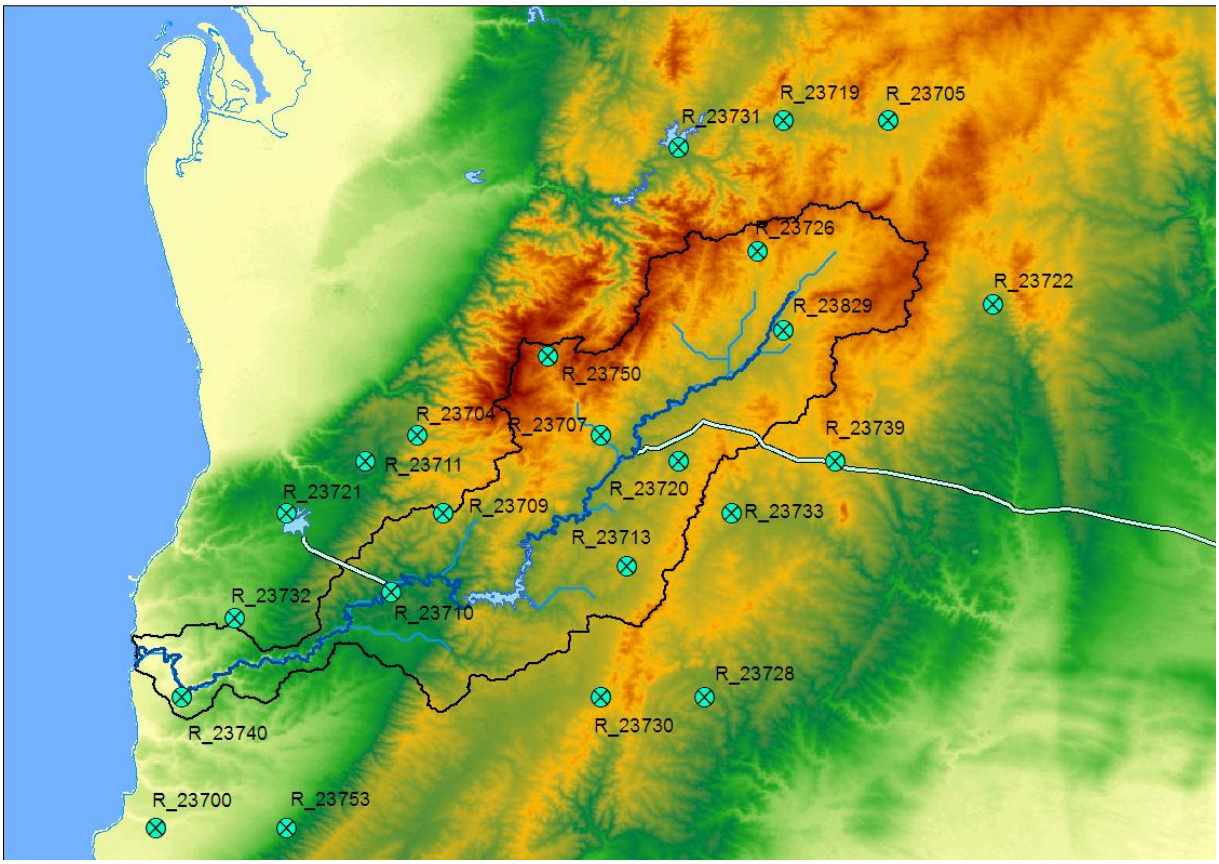


Figure 4: Rain gauge locations in the Onkaparinga catchment

A detailed temporal homogeneity analysis was undertaken to test for:

- Changes in the location of the measuring gauge;
- Changes in the observer;
- Changes in watering pattern in the vicinity of the gauge; or
- Growth or trees, crops or buildings near the gauge.

The tests were performed following the method of Allen *et al*. [1998a], with detailed results presented in the first milestone report [*Leonard et al.*, 2011]. The Happy Valley Reservoir record (23721) was used as the reference record since it has previously been verified as a high-quality station. The outcome of the analysis was that five sites had values outside the 95% confidence interval, with Bridgewater (site 23707) having the largest excursions. Coromandel Valley, Gumeracha, Hahndorf and Morphett Vale also show some level of inhomogeneity.

Because Bridgewater has a central location and high rainfall, it is expected to have high influence on the interpolation. An additional analysis was conducted using three shorter records closer to Bridgewater that were not in the original 23 sites, and this comparison did not show similar excursions to the Bridgewater site. Thus, Bridgewater was excluded from the study. The disadvantage of this approach is that as a high-elevation site and therefore has relatively high annual rainfall (1046 mm), however as the interpolation technique includes elevation as a covariate, and

because there are other high-elevation sites (e.g. Urailda, 23750 and Lobethal, 23726) in the catchment, the elevation effect is already incorporated into the interpolated rainfall. With the exception of Bridgewater, the remaining sites were retained because their detected inhomogeneities were not as large and because the additional sites, especially towards the outer extent of the region, increase the level of spatial information available to capture the variability in daily rainfall. The site details for the selected 22 sites are summarised in Table 3, and form the basis for most of the remaining analyses in this report.

Table 3. Summary of rainfall data within Onkaparinga catchment.

SiteID	SiteName	Lat.	Lon.	AnnAve (mm)	Elev. (m)	Start	End
23700	ALDINGA POST OFFICE	-35.16	138.29	508	32	1893	1992
23704	BELAIR (STATE FLORA NURSERY)	-35.01	138.39	786	386	1889	2011
23705	BIRDWOOD	-34.49	138.57	723	385	1889	2011
23709	CHERRY GARDENS	-35.04	138.4	924	345	1899	2011
23710	CLARENDON	-35.07	138.38	818	223	1889	2011
23711	COROMANDEL VALLEY (BRANDEN)	-35.02	138.37	714	234	1890	1986
23713	ECHUNGA GOLF COURSE	-35.06	138.47	805	375	1889	2011
23719	GUMERACHA	-34.49	138.53	793	346	1889	2011
23720	HAHNDORF	-35.02	138.49	845	347	1889	2011
23721	HAPPY VALLEY RESERVOIR	-35.04	138.34	638	148	1891	2011
23722	HARROGATE	-34.56	139.01	552	335	1896	2011
23726	LOBETHAL	-34.54	138.52	882	470	1889	2011
23728	MACCLESFIELD	-35.11	138.5	730	302	1889	2011
23730	MEADOWS	-35.11	138.46	869	384	1889	2011
23731	CUDLEE CREEK (MILLBROOK)	-34.5	138.49	831	311	1914	2011
23732	MORPHETT VALE	-35.08	138.32	562	90	1889	2011
23733	MOUNT BARKER	-35.04	138.51	766	349	1889	2011
23739	NAIRNE	-35.02	138.55	678	403	1889	2011
23740	OLD NOARLUNGA POST OFFICE	-35.11	138.3	522	7	1889	1998
23750	URAIDLA	-34.58	138.44	1088	499	1890	2011
23753	WILLUNGA	-35.16	138.34	642	158	1889	2011
23829	WOODSIDE	-34.57	138.53	801	387	1889	2011

The average rainfall over each of the three sub-catchments was estimated using weights obtained through a kriging procedure, in which rainfall totals are first regressed against elevation, followed by the interpolation of the residual. This ensures that the spatial interpolation accounts for the strong rainfall gradient in the catchment, which is likely to be associated with orographic effects (and hence elevation). The kriging method, and a comparison of the interpolated surface with the Australian Water Availability Project (AWAP) product (http://www.bom.gov.au/jsp/awap/) is presented in the first milestone report [*Leonard et al.*, 2011]. The outcome of this comparison is that the interpolated surface captures the spatial patterns of the rainfall observed in the AWAP product, including the ridge of higher rainfall along the western edge of the Onkaparinga and the lower rainfall estimates towards the catchment outlet. At the annual scale, the krigged data is on average within 4mm/year of the AWAP data, with the maximum discrepancy in a single year of 20 mm. It is noted, however,

that since AWAP is a gridded product, it could not be used directly as the basis for deriving catchment rainfall estimates because a procedure is required that can also be applied to interpolate outputs from the multi-site NHMM downscaling model.

The SILO database was used to obtain rainfall over a common period for constructing catchment averages. From the first milestone report [*Leonard et al.*, 2011] it was highlighted that the SILO database used interpolations to construct the uninterrupted record. The catchment averages constructed by means of kriging are therefore to some extent doubly interpolated. Interpolated values are always smoother (less variable or extreme) than original values, so there is the potential that this approach over-smooths the data, which is of particular concern for high rainfall events (i.e. the interpolation could potentially underestimate them). It is important to emphasize that the procedure used to construct catchment averages could be performed directly from daily rainfall observations rather than SILO and thus avoid this issue, but SILO was used here to retain simplicity in the overall approach.

The main concern regarding over-smoothing is whether there is a relationship between the missing observations and the rainfall amount (e.g. is it possible that larger rainfall events had more missing observations and thus become biased). This relationship was investigated in a number of ways using scatterplots, regressions and summary statistics for different sets of rainfall. Figure 5 gives a typical result, which shows a scatterplot of Houlgrave Weir catchment average rainfall and the percent missing sites on any given day. “Missing” values are those values used by SILO which were interpolated and this incorporates instances of accumulated observations in addition to truly missing or corrupt observations. From this plot, there is only a very weak relationship between the average rainfall and the percent missing sites, which can be seen from the shift in the shaded density and from the regression line (R^2=0.033). Additional checks were performed using averages for the other catchments and for averages that are strictly made from non-infilled SILO observations and these checks returned similarly weak relationships.

Based on the regression line, each rainfall day has approximately 20-30% of sites missing across the catchment. However, because the gauging density is relatively high (with a total of 22 gauges in or nearby the catchment), this means that is still a relatively large number of sites with data in most cases to develop the spatially averaged estimates of rainfall. Nevertheless, there are a small number of high rainfall days that have a high percentage of missing sites, and thus are likely to be uncertain. These high rainfall days could potentially have a significant influence on estimated streamflow, but techniques are currently unavailable to evaluate the influence of these points on the overall predictions. It is recommended that estimating the influence of high rainfall points, and developing time-varying rainfall uncertainty estimates at the daily timescale, be subject to future research.

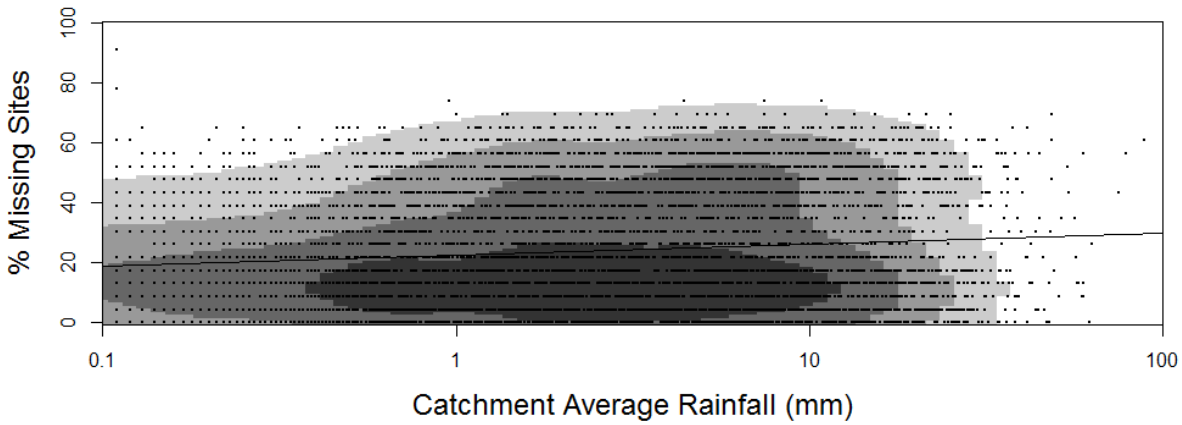


Figure 5: Scatterplot of the Houlgrave Weir catchment average rainfall with the percentage of missing sites on any given day in the period 1970 onwards. The underlying density (bandwidth=0.5) indicates that there is not a strong relationship between the two variables. "Missing" values are those that did not have an exact total recorded on that day and thus required interpolation by SILO (i.e. accumulations, actual missing or poor quality data).

4.3 Potential Evapotranspiration

Potential Evapotranspiration (PET) was estimated using Morton's method [*McMahon et al.*, 2013] but with Penman's approach [*Allen et al.*, 1998b] adopted for the radiation calculations. This approach is consistent with other similar studies in Australia [e.g. *Li et al.*, 2009], and uses daily minimum and maximum temperature, incoming solar radiation and vapour pressure deficit as input variables. These variables are also produced by the non-homogenous Markov model (NHMM) downscaling technique in Task 3 (discussed further in the second report of this series), and therefore enables a consistent basis for future climate change assessments.

PET was calculated using two alternative data sources. The first data source was the observed instrumental data from the Kent Town high-quality weather station, and then modified to produce PET estimates over the Onkaparinga using conversion factors. These factors were calculated using the AWAP gridded product (http://www.eoc.csiro.au/awap/) by comparing the annual AWAP PET at each of the Onkaparinga rain gauges with the annual AWAP PET at Kent Town. The second data source was daily minimum and maximum temperature, incoming solar radiation and vapour pressure deficit obtained from the SILO PPD. The two data sources produced estimates that were within 4% of each other, and the SILO PPD was selected as the data source for the remainder of the work to ensure consistency with the NHMM simulations, which were also based on the SILO PPD records.

The catchment average wet-environment areal PET, calculated using Morton's method using the SILO PPD, was 1300 mm per year. In contrast, observed pan evaporation at Mount Bold Reservoir was 1560 mm per year. These numbers would not be expected to be equal, as Morton's APET represents the potential evapotranspiration that would occur under steady state meteorological conditions in which the soil/plant surfaces are saturated and there is an abundant water supply [*McMahon et al.*, 2013], while pan evaporation tends to overestimate evaporative demand due to the incidence of solar radiation on the top and sides of the evaporation pan. Morton's wet-

environment APET was selected in favour of the Mount Bold Reservoir pan evaporation series because:

(1) The pan evaporation data does not allow for an assessment of the impacts of climate change in the future, since it is necessary to develop relationships between potential evaporation and meteorological variables such as temperature and vapour pressure deficit; and

(2) A strong and unexplainable negative trend was detected at Mount Bold Reservoir from about 1970 to 1990, with this trend reversing from the mid 1990's onward. This is shown in Figure 6 for Mount Bold and a nearby gauge at McLaren Vale. A review of trends in pan evaporation data at other locations showed that some locations exhibited increasing trends while other locations (often in close proximity) exhibited decreasing trends. Therefore these trends do not appear to be region-wide phenomena, and thus may be at least partially attributed to measurement issues at individual gauges.

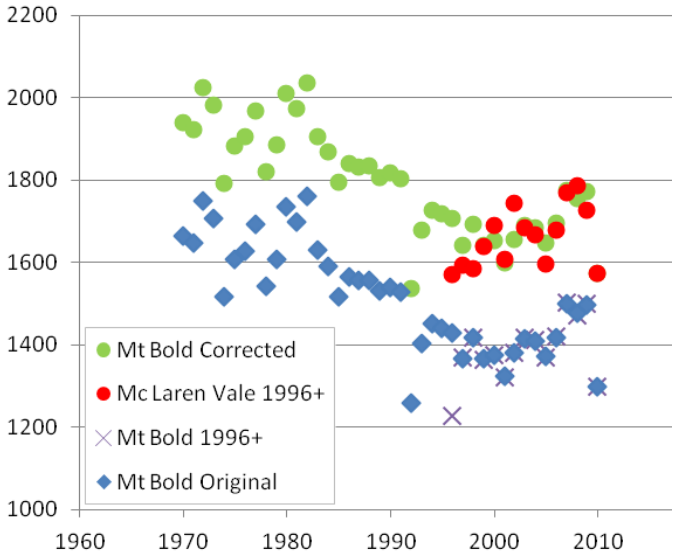


Figure 6: Time series of annual total pan evaporation. The 'corrected' version of the Mount Bold Reservoir data was based on the recommendation in Teoh [2002] that due to the proximity of the station to a water body and a pine forest in the surrounding area, it was necessary to adjust the records upwards.

4.4 Modifications to the data used since second milestone report

Since the second milestone report, a number of modifications to the observational datasets were made, which have led to improvements in overall model performance. These included:

- The timing issue of daily flows (originally defined from 00:00 to 24:00) being inconsistent with daily rainfall (from 09:00 to 09:00) has been rectified by re-deriving daily flows from a sub-daily record in each catchment. This has also led to a shortening of the period of calibration compared to the period used in Westra et al. [2012], since for example sub-daily data at Echunga Creek was not available prior to 1993.

- Timing errors associated with the release of flows from Hahndorf Dissipator could not be corrected due to the absence of a sufficiently long record of sub-daily flow data at this location. Therefore a new approach of censoring the flows during calibration was adopted, to ensure that such timing issues do not affect the calibrated parameter set.

- The potential evapotranspiration calculations were changed from a Penman-Monteith formulation to Morton's areal potential evapotranspiration. Furthermore, rather than using the observed meteorological data at Kent Town, the SILO PPD data was used from the 22 gauges in and surrounding the Onkaparinga catchment. This modification was undertaken to ensure that the PET estimates used for calibrating the hydrological model parameters are consistent with those used by the NHMM downscaling algorithm.

4.5 Summary of data used for hydrological model calibration and validation

In summary, the hydrologic data used for this analysis are as follows:

- **Streamflow**: Streamflow from three gauges – Scott Creek, Echunga Creek and Houlgrave Weir – were used. These gauges describe the major sub-catchments in the Onkaparinga catchment upstream of the Happy Valley Reservoir diversion.
- **Rainfall**: Catchment average rainfall for each of the three sub-catchments was obtained by the 22 gauges given in Table 3, using a kriging technique to develop the catchment averages.
- **Potential Evapotranspriation (PET)**: Aerial PET was calculated using Morton's method based on daily minimum and maximum temperature, incoming solar radiation and vapour pressure deficit. These variables were obtained from the SILO PPD. A single daily time series was used for all sub-catchments, as the instrumental basis of the SILO PPD (comprising only two gauges located within the Onkaparinga catchment – E23734 and E23801 – see Section 6.3.4 in Leonard et al, 2011) was not of sufficient resolution to support higher spatial resolution estimates.

Based the availability of high-quality observational data, the records were separated into an exploratory (calibration) period used for parameter estimation, and a confirmatory period used for model evaluation (usually this is referred to as a 'validation' period but we prefer to use the term 'confirmatory' given that it is not possible to validate a model's future performance using only historical data; see [*Oreskes et al.*, 1994]). The dates of each period are summarised in Table 4. The start dates for the exploratory period were selected based on a rating curve analysis (discussed further in Section 5.3.2), with the objective being to maximise the period of record available for

model calibration subject to the quality of the data being deemed sufficient for use. The confirmatory period was selected to be the dry decade from 2000 to 2009; this provides an important test for a hydrological model to evaluate its capacity to simulate hydrological response to changed climate forcings. Further detail on the basis for selecting the exploratory and confirmatory periods is provided in *Westra et al.* [2014a].

Table 4: The exploratory and confirmatory periods used in the analysis

	Exploratory		Confirmatory	
Site Name	**Start**	**End**	**Start**	**End**
Houlgrave Weir	01/01/1977	31/12/1999	01/01/2000	31/12/2009
Scott Creek	01/01/1985	31/12/1999	01/01/2000	31/12/2009
Echunga Creek	01/01/1993	31/12/1999	01/01/2000	31/12/2009

5 Quantification of Uncertainty

5.1 Overview

Quantifying and (where possible) reducing uncertainty remains an on-going challenge for climate impact assessments. Henderson-Sellers [1993] developed the concept of a 'cascade of uncertainty' in which uncertainty is introduced at each step of the modelling chain from large-scale climatic processes to local impacts. To assess the uncertainty in runoff projections under a future climate for the Onkaparinga catchment, sources of uncertainty and approaches used to quantify their relative importance include:

- **Future greenhouse gas emissions scenarios.** Simulations are provided for two representative concentration pathways (RCPs), to simulate two plausible scenarios for future greenhouse gas concentrations.

- **Global climate models.** GCM-based simulations are provided from the World Climate Research Program Coupled Model Intercomparison Project Phase 5 (CMIP5) archive.

- **The downscaling method.** Here a single downscaling method – the non-homogenous hidden Markov model – is used, but 100 replicates are provided for each RCP and GCM to represent the uncertainty in downscaling from the GCM scale to local gauged scale, conditional on the chosen downscaling method.

- **The hydrological model** that translates projections of rainfall and PET into runoff.

This volume focuses on the fourth source of uncertainty: that associated with the hydrological model. The remaining sources of uncertainty are considered in the third volume of this series. Hydrological model uncertainty may be because of:

- Biases or random errors in the data used to calibrate the hydrological model, such as instrumentation errors or errors associated with translating point data into catchment-averaged data;

- Use of a finite calibration record to estimate the hydrological model parameters; and

- Model structural deficiencies, since hydrological models are simplified representations of the complex processes involved in translating rainfall and PET into runoff.

The following sections address each of these sources of hydrological model uncertainty, focusing separately on observational data errors and model structural errors. The uncertainty associated with each of these error sources has been quantified using the BATEA methodology, as described briefly below. The final component of this analysis (section 5.5) is a comparison of the contribution of the various sources of uncertainty to the hydrological model predictions.

5.2 Bayesian Total Error Analysis (BATEA)

Bayesian Total Error Analysis (BATEA) is a model calibration and prediction framework introduced in Kavetski et al. [2002] and generalized in subsequent publications [*Kavetski et al.*, 2006; *Kuczera et*

al., 2006; *Renard et al.*, 2011]. The main objectives of BATEA are to improve the reliability and precision of parameter estimates and model predictions and, where possible, gain insights into predictive uncertainty by decomposing it into its multiple contributing sources (Figure 7).

The core ideas and steps within BATEA are as follows:

1. Specify explicit probability models for each source of uncertainty (input, output and model structural errors);

2. Where necessary, use hierarchical techniques to implement these probability models within a Bayesian inference framework;

3. Where available, include a priori information about the catchment behaviour and data uncertainty;

4. Jointly infer the parameters of the hydrological model and the error models; and

5. Examine posterior diagnostics to check the assumptions of the error models made in step 1.

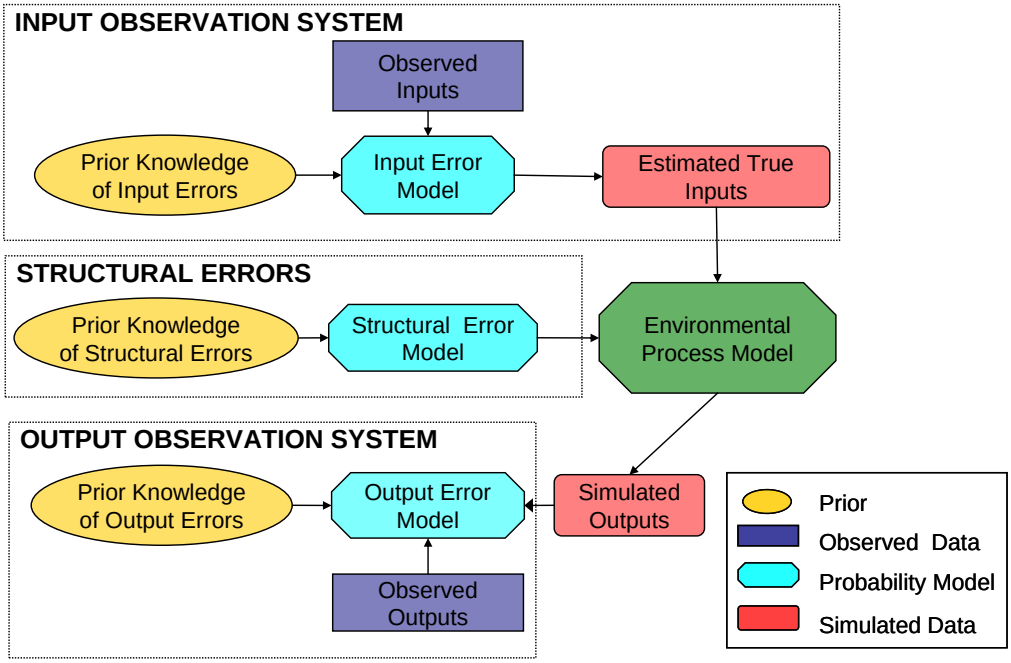


Figure 7: Schematic of BATEA

The BATEA methodology is implemented within the 'BATEAU' software platform, which is a generic toolkit for model calibration, prediction and uncertainty analysis. BATEAU has following capabilities:

1. A selection of optimization algorithms including quasi Newton (QN), shuffled complex evolution (SCE), dynamic dimensioned search (DDS).

2. Ability to undertake parameter and predictive uncertainty estimation and analysis using Bayesian techniques and Markov chain Monte Carlo (MCMC) analysis.

3. Ability to link to models through a variety of interfaces, including via direct linking using models coded in Fortran, via Dynamic Linked Libraries (e.g. models coded in C#), and linking to program executables.

4. Wide range of calibration schemes, including the commonly used standard least squares (SLS) approach and weighted least squares (WLS).

BATEAU has a comprehensive set of diagnostics to analyse model assumptions and predictive performance. BATEAU also has an interface with the Bayesian Analysis Diagnostics (BAD) package written in the R statistical computing language, which enables post-processing of the outputs from BATEAU to produce a large number of diagnostic plots and statistics to aid in the analysis and interpretation of results. The BAD package consists of a number of functions, and a beta version is available from http://code.google.com/p/bad-/.

The BATEA method, in combination with the BATEAU and BAD software packages, provide a unified framework to analyse and interpret hydrological model outputs, while providing detailed information on parameter and predictive uncertainty. The method uses information on expected input errors (e.g. from spatially averaging point-based rainfall data) and outputs errors (e.g. from rating curve uncertainty) to determine the relative contribution of each source of uncertainty. The remaining uncertainty is then attributed to deficiencies in the model structure in capturing the complex processes that cause rainfall to be converted to runoff. The following section describes the results of the data-related uncertainty analysis, to be followed by a description of model-related uncertainty in Section 5.4.

5.3 Data Errors

Projections of hydrological response under a future climate will be obtained by comparing the characteristics of runoff derived from NHMM simulations of rainfall and PET in a future climate with those derived from NHMM simulations of rainfall and PET under historical climate forcings. Thus, the instrumental data will not be used in this analysis except through the process of hydrological model calibration and validation. With this in mind, the main emphasis of the assessment of data errors is to assess the impacts on the calibrated model parameters.

5.3.1 Input Errors

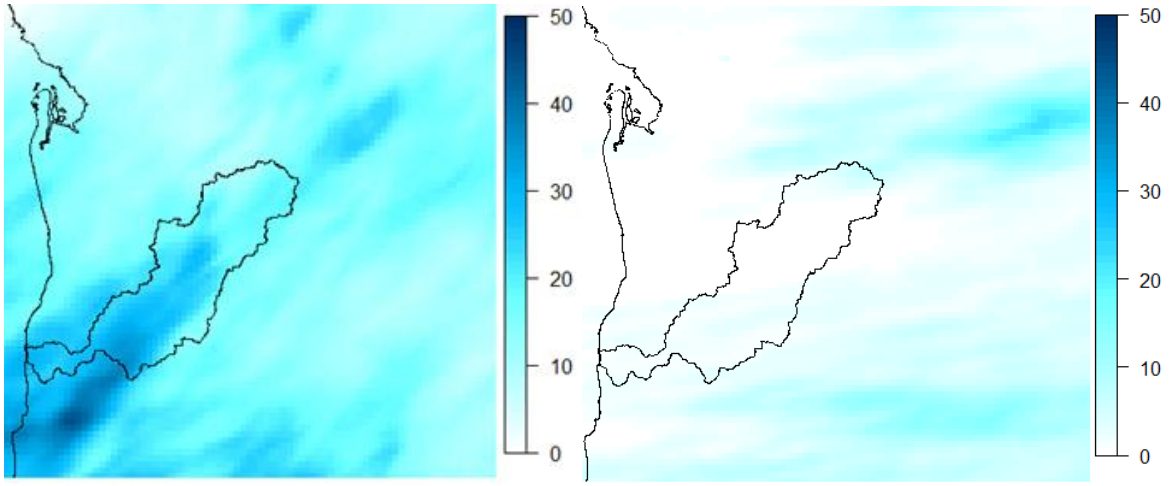
Input errors are errors associated with the rainfall and PET data used as inputs to the rainfall-runoff model. These errors can affect the model through its parameters, confidence or credibility intervals, and predictions. This section focuses on quantifying the magnitude of rainfall errors, for the following reasons:

(1) A combination of gauge-based and radar-derived rainfall products are available for the Onkaparinga catchment, which can support the quantification of rainfall uncertainty. In contrast, there is limited observational data for the variables that drive potential evapotranspiration, with the nearest high-quality weather station located at Kent Town. The difference between scaled Morton's APET using the Kent Town data and the estimates using the SILO PPD was only about 4%.

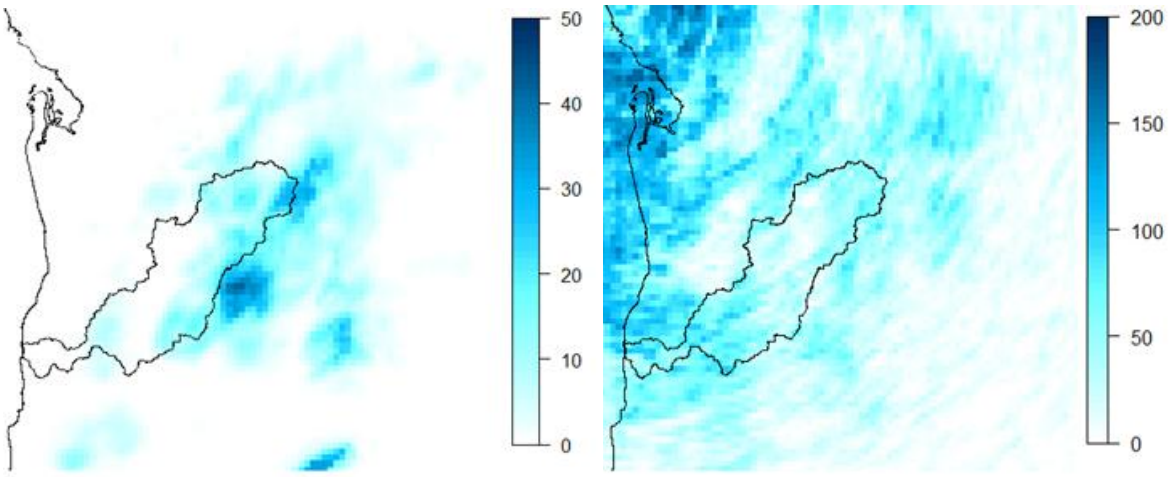
(2) Hydrological models are typically much more sensitive to changes in rainfall than they are to changes in PET. For example [*Jones et al.*, 2006] showed that runoff is about 3 to 5 times more sensitive to changes in rainfall than changes in PET.

Radar data is used to determine the variability induced in catchment-average rainfall estimates for the Scott Creek, Echunga Creek and Houlgrave Weir sub-catchments. Weather radars measure reflectivity from electromagnetic pulses as they scan the sky, and these pulses are then converted to rainfall estimates according to the relationship $Z = a\,R^b$, where a and b are climatological parameters determined for that region. There are many complications in the measurement of reflectivity and subsequent conversion from reflectivity to rainfall estimates, so that radar estimates are often globally (i.e., over the whole domain of the radar) and locally biased. Nevertheless, radar imagery is invaluable since it provides detailed information on the spatial structure of rainfall over large regions.

Figure 8 provides examples of radar images on selected days based on the Buckland Park radar outputs. The images are at a 24 hour scale and were aggregated from images at a 10 minute interval. The radar covers an area of 256 km x 256 km, although Figure 8 shows only a quarter of its domain.



(a) South-Westerly Rainfall 25/06/2009 (b) Westerly Rainfall 17/12/2008



(c) Localised Convective Rainfall 01/02/2009 (d) Radial Artefact 08/02/2009

Figure 8: A sample of radar images covering the Onkaparinga Catchment showing different types of rainfall activity

Panel (a) of Figure 8 shows a frontal southwesterly storm – common for the Adelaide region – along with more westerly storms, as shown in panel (b). Panel (c) shows some localised convective activity. The two are different in that frontal events have a longer correlation length scale in the direction of the storm whereas convective events do not show the same spatial dependence. Finally, panel (d) shows an image with radial artefacts in the estimates.

The rainfall conversion for Buckland Park is based on a global bias correction so that the overall regional average corresponding to rain gauges is preserved. The annual average total of the record is shown in Figure 9 and it suggests that the highest rainfalls were north of the Onkaparinga catchment and in the lower catchment near the outlet. This does not compare to typical isohyets for this region, where Mount Lofty, bordering the North Western side of the catchment, experiences the highest rainfall (~1200 mm on average) and where there is a strong gradient toward the south eastern extent which has an annual average of approximately 800 mm. Also the region near the catchment outlet has approximately 600 mm on average.

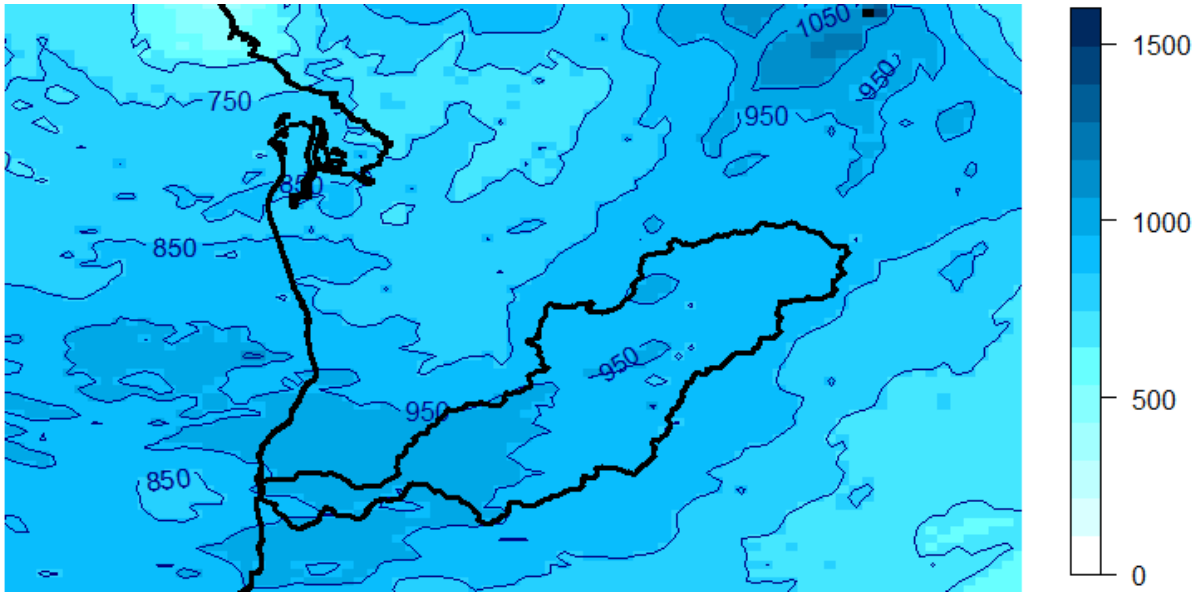


Figure 9: Annual average rainfall observed by Buckland Park radar

It is possible that these discrepancies were due to sampling variability, so that the short radar record may have been higher near the outlet and now experiences a strong gradient in that period. Figure 10 gives the gauge-based rainfall averages for the same period and shows a different pattern; namely, that there is a stronger gradient and that the rainfall amount in the western portion of the catchment is considerably less than the radar estimate. This demonstrates the difference between global and local bias correction. Although it is possible that artefacts in the radar imagery will have some impact, the high density of gauges used in the following method suggests that it will not be large.

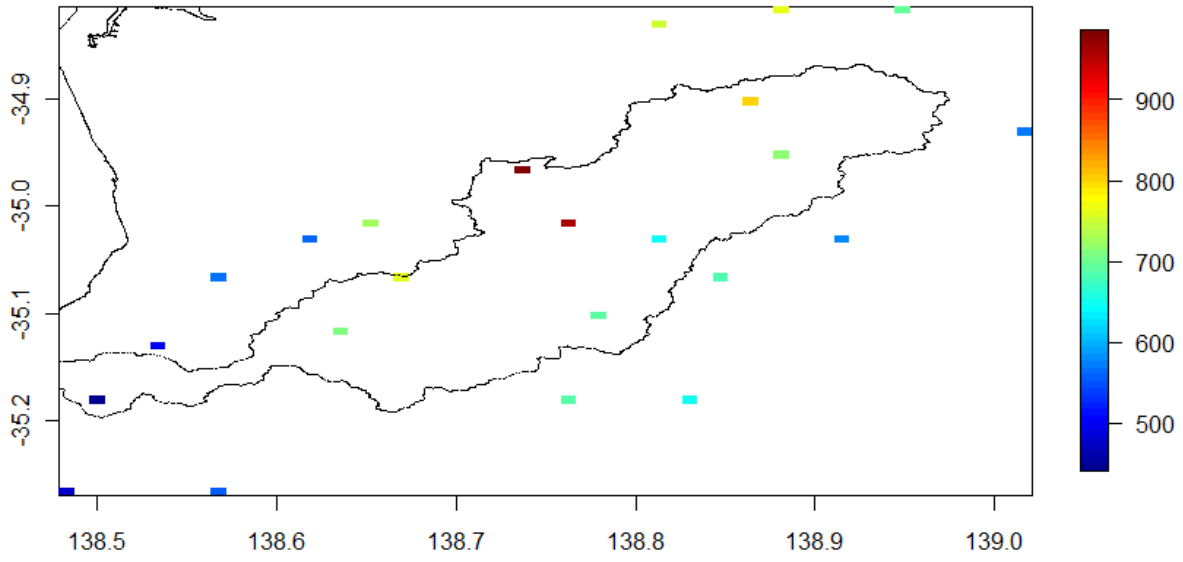


Figure 10: Annual average rainfall observed for the period matching the radar.

When a network of rain gauges observes a rainfall event, they obtain only sparse and noisy information. While rain gauges are useful for determining temporal properties, they are limited in the spatial domain. The spatial imagery of radars can be exploited to determine the error induced by the sparse sampling of the gauge network. The method follows Villarini and Krajewski [2008] who define a multiplicative error term so that the true rainfall over the catchment is obtained from the observed network along with a distribution of errors. These multiplicative errors are used to assess the conditional mean error with respect to increasing observed rainfall, where the condition μ=1 represents unbiasedness. The conditional standard deviation is also calculated as above a given threshold. The multiplicative error is defined as:

$$\varphi = \frac{r_{OBS}}{r_{TRUE}} \tag{3}$$

where r_{OBS} represents the set of locations that observe the rainfall event and r_{TRUE} is the true average rainfall over the catchment. Panels (a) to (c) in Figure 11 show the masks that were applied to the radar imagery in order to estimate the catchment average r_{TRUE} for the three catchments. Panel (d) shows the locations of the 23 rain gauges and the corresponding radar pixels adopted to determine the value of r_{OBS}. In each case an arithmetic average is used to determine the average daily rainfall, although careful analysis of the observed weights with respect to the radar imagery may lead to more accurate assessments. The same 23 gauges are used for all three catchment estimates which is because they are all within the 24 hour correlation length scale. While there were 415 wet days in the record, not all of these can be used since if the r_{OBS} is zero, an infinite value is generated. Also, the multipliers are bounded at zero from below, yet are unbounded above one so that large multipliers can be generated (especially when r_{OBS} is close to zero).

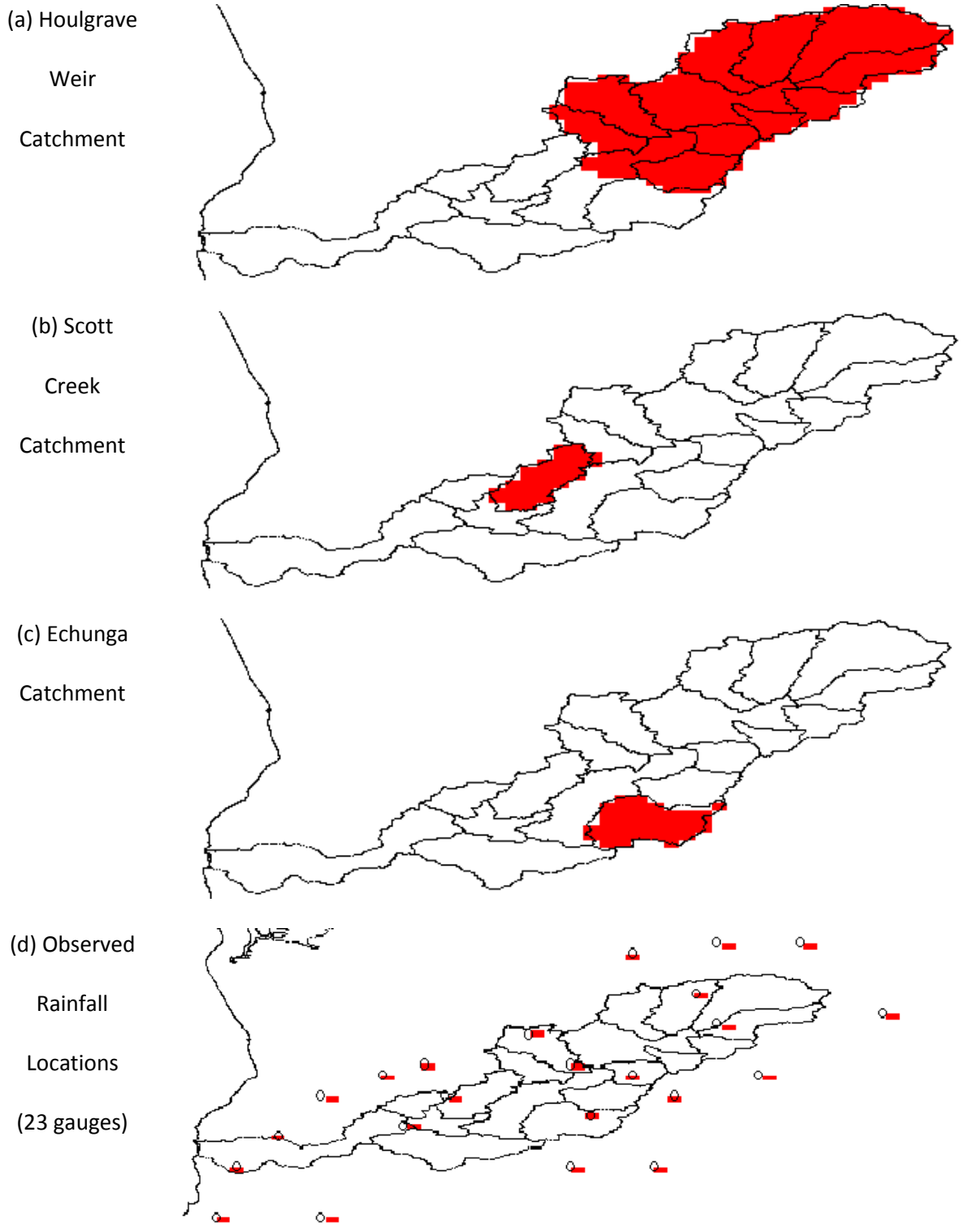
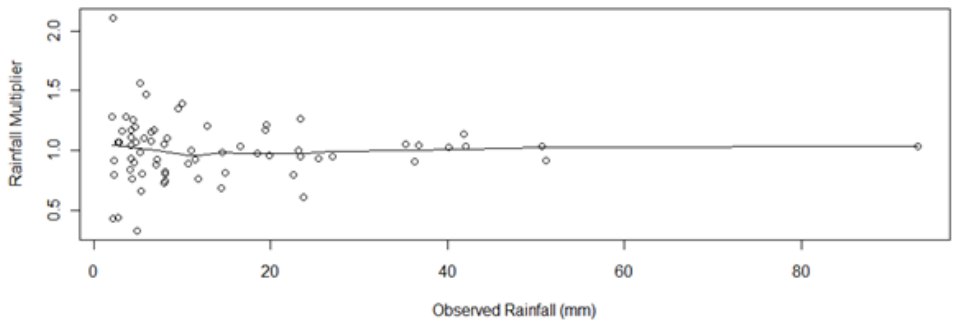


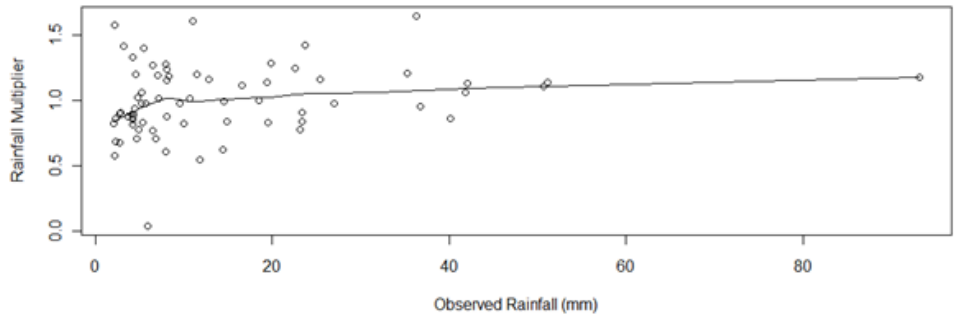
Figure 11: Masked images to obtain true rainfall estimate for relevant subcatchments of the Onkaparinga (Houlgrave Weir Catchment, Scott Creek Catchment, Echunga Catchment). The rainfall gauge locations provide an estimate of the observed rainfall.

The multipliers obtained for each of the three catchments are shown in Figure 12 along with a lowess curve to visualise any significant departure from the mean. Only those observed rainfalls greater than 0.1 mm are plotted. It does not appear that there is any significant bias in this record. In all three cases the variability of the errors decreases with increasing rainfall amount.

(a) Houlgrave Weir Catchment



(b) Scott Creek Catchment



(c) Echunga Catchment

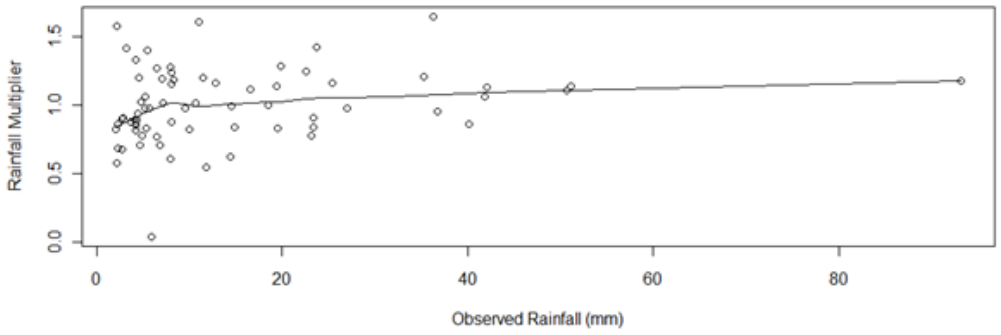


Figure 12: Multipliers for separate catchments plotted against observed rainfall.

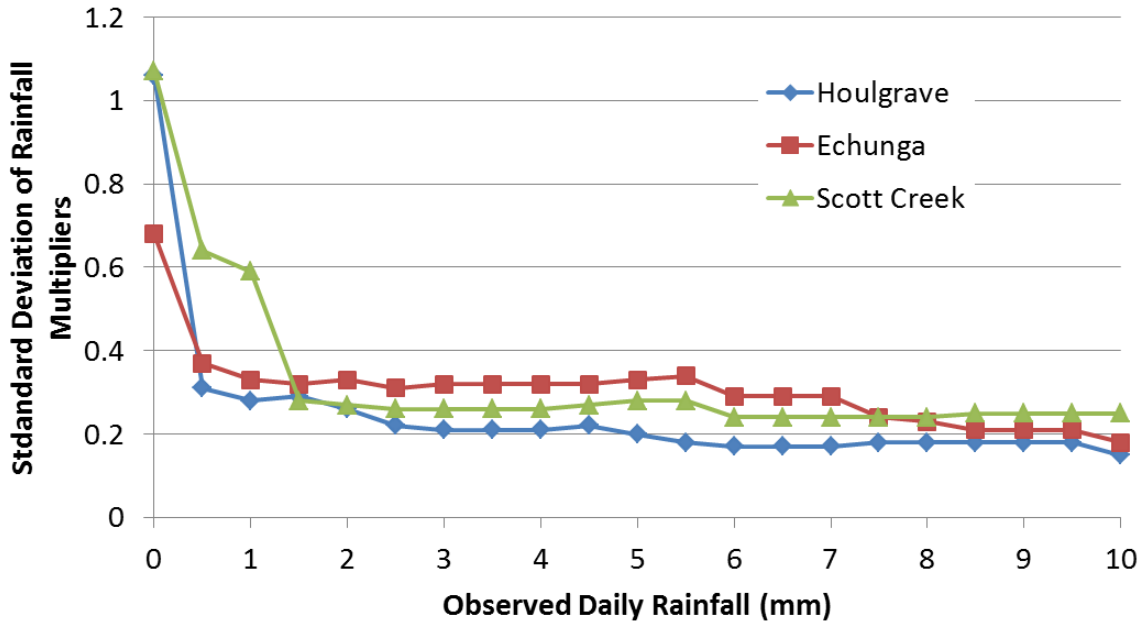


Figure 13: Standard deviation of rainfall multipliers for plotted against observed rainfall for each catchment.

Figure 13 shows the standard deviation of the rainfall multipliers, and indicates a decreasing power law with respect to the observed daily rainfall in all three cases, plateauing at approximately 2-2.5 mm. Houlgrave Weir is the least variable at large rainfalls, while Scott Creek is the most variable. This is understandable given the smaller catchment area of Scott Creek. Above 5mm the standard deviation in Echunga Creek appears to reduce, but this may be due to sampling variability.

Based on Figure 13, the highest standard deviations occur for the lowest rainfall events, and then decrease with increasing rainfall intensity. As the majority of flow volume comes from moderate to large rainfall events (e.g. the top 10 flow days in a year accounts for approximately half the total annual runoff in all three sub-catchments), it is important to specify a standard deviation for the multipliers that correspond to those events. Therefore we estimated the standard deviation of the rainfall multipliers based on all rainfall days greater than 10mm, which resulted in standard deviations of 0.25, 0.18 and 0.15 for Scott Creek, Echunga Creek and Houlgrave Weir catchment, respectively. This means that the 95 percent prediction interval for the rainfall multiplier for moderate to large events would be between 0.5 and 1.5 (Scott Creek), 0.64 and 1.36 (Echunga Creek) 0.7 and 1.3 (Houlgrave Weir).

More complex rainfall error models, such as those that model the standard deviation as a function of rainfall intensity, may more accurately represent the decreasing rainfall multiplier variability with increasing rainfall depth in a more physically realistic manner, but are not considered this case due to limited amount of radar data available for the analysis.

5.3.2 Output Errors

Output errors are mostly due to errors in the rating curve which transforms river height into streamflow. A detailed analysis of the Scott Creek, Echunga Creek and Houlgrave Weir rating curves was conducted and presented in the second Milestone report [*Westra et al.*, 2012]. The adopted rating curve error model is based on the difference between the river gauging $\left(Q_{gauge}\right)$ and the rating curve predicted (RCP) runoff $\left(Q_{rc}\right)$:

$$\varepsilon_Q = Q_{rc} - Q_{gauge} \tag{1}$$

Previous studies [*Thyer et al.*, 2009] have found that this runoff error increases as the RCP runoff increases. This provided the motivation to develop a heteroscedastic error model with the standard deviation increasing as a function of the RCP runoff. This is similar to runoff error models developed in previous studies [*Thyer et al.*, 2009], and is summarised as follows:

$$\begin{aligned} &\varepsilon_Q \sim N(\mu_Q, \sigma_Q) \\ &\sigma_Q = a_\sigma + b_\sigma * \left(Q_{rc}\right)^{c_\sigma} \\ &\mu_Q = a_\mu + b_\mu * Q_{rc} \end{aligned} \tag{2}$$

The parameters a_μ, b_μ of the output error model quantify the evidence of bias in the RCP runoff. The parameters $a_\sigma, b_\sigma, c_\sigma$ quantify the evidence of heteroscedasticity in the runoff errors. The model parameters were fitted to the runoff error data using the WINBUGS software [*Spieghalter et al.*, 2003] to evaluate the posterior distribution. Each parameter was included in the final model if the posterior probability of the parameter having a value of zero was negligible. It should be noted that the rating curve analysis described here is based on instantaneous flow data, and thus are likely to overestimate the total uncertainty in the daily streamflow time series.

As discussed in further detail in Westra et al. [2012], the analysis showed that there was that significant extrapolation to the rating curve at Scott Creek and Echunga Creek for flows greater than 10 mm (approximately a 1 in 6 month flow), whereas the rating curve at Houlgrave Weir was supported by measurements up to 20 mm (approximately a 1 in 20 year flow). No major changes in the rating curve could be detected in the Echunga Creek and Houlgrave Weir catchments, although there was evidence of a significant bias prior to 1984 at Scott Creek [*Westra et al.*, 2012] due to change in the rating curve. For Echunga there was some evidence of small relative bias of 3%. To reduce the likelihood that systematic runoff measurement biases will impact on model calibration performance, all subsequent analyses focus on Scott Creek data from 1985. The parameters from the output error model for each sub-catchment are provided in Westra et al. [2012].

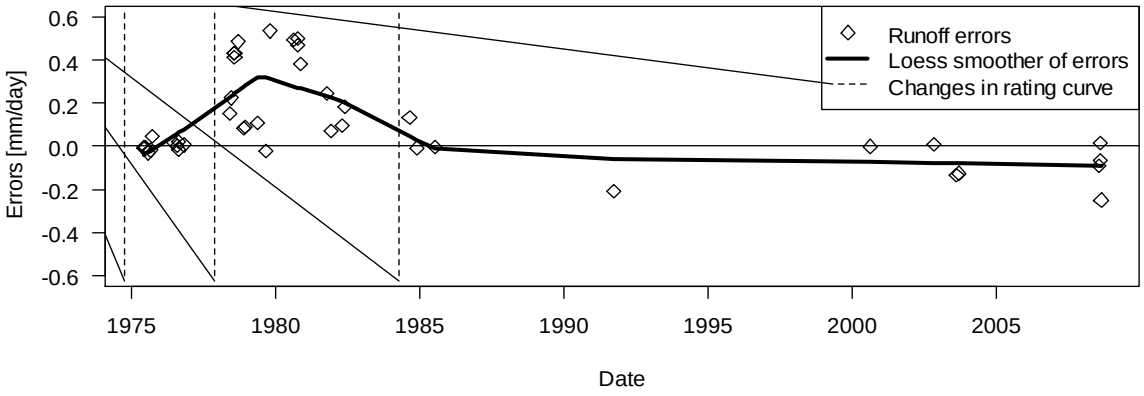


Figure 14: Runoff error time series at Scott Creek. Runoff errors = streamflow predicted by rating curve – streamflow gauging.

Table 5. Summary of Runoff Error Analysis

Site	Degree of Extrapolation	Frequency rating curve changes	Runoff Error Model Parameters				
			a_μ	b_μ	a_σ	b_σ	c_σ
Scotts Creek	Above 1 in 6 month/year flow	One major change in 1984	0.0	0.0	0.034	0.082	0.57
Echunga Creek	Above 1 in 3 year flow	None	0.0	0.03	0.006	0.06	0.74
Catchment runoff at Houlgraves Weir (includes MBO pipeline errors)	Above 1 in 20 year flow	None	0.0	0.016	0.031	0.05	1.0

A further issue relates to timing errors due to the release of flows from Hahndorf Dissipator, which needs to be accounted for when estimating the natural catchment flows at Houlgrave Weir. This issue was discussed in Section 4.1 and the outcome was to censor the streamflow on days which are likely to have significant timing error.

5.4 Hydrological Model Errors

Errors associated with the hydrological model can be due to parameter uncertainty, and the issue that the hydrological model does not represent the 'true' model of the catchment. These issues are discussed in turn below.

5.4.1 Overview of hydrological model GR4J

All hydrological models considered in this work are derived from the lumped conceptual rainfall-runoff model GR4J [*Perrin et al.*, 2003]. The published version of GR4J has four calibration parameters, namely the production store capacity (θ_1, units of mm), the groundwater exchange

coefficient (θ_2, units of mm), the one day-ahead maximum capacity of the routing store (θ_3, units of mm), and the time base of the unit hydrograph (θ_4, units of days). The model structure is depicted in Figure 15.

GR4J belongs to the family of soil moisture accounting models and was developed to provide, on average, good performance across a wide range of catchment conditions [*Renard et al.*, 2011]. Detailed testing of the model in simulating runoff for the Onkaparinga catchment is described in the second milestone report [Westra et al, 2012]. The simplicity of the model makes GR4J particularly suitable as a starting point for model modifications and refinements, including the versions constructed in this work as part of detecting and quantifying hydrological non-stationarity as discussed in Appendix 1 of this report.

In addition to the four hydrological model parameters that need to be estimated, there are two heteroscedastic residual error model parameters (a_ε and b_ε) that also need to be estimated during the inference. The heteroscedastic error model allows errors to be normally distributed $\varepsilon \sim N(0, \sigma_\varepsilon)$ but with the standard deviation increasing linearly as a function of simulated flow (i.e. $\sigma_{\varepsilon(t)} = a_\varepsilon + b_\varepsilon y_t$).

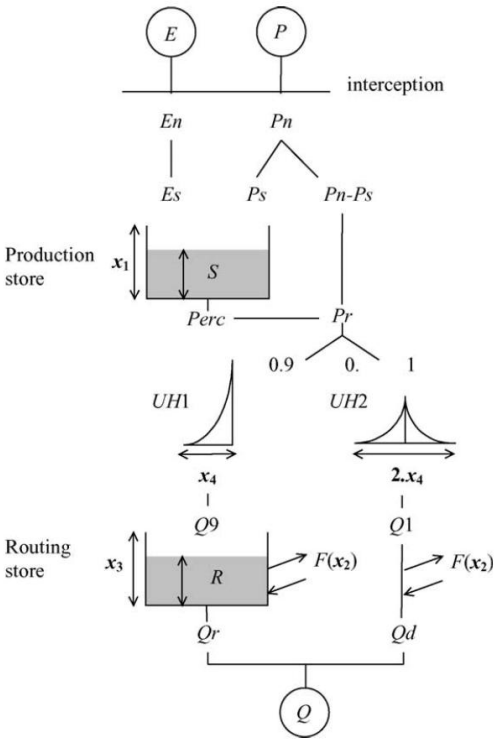


Figure 15: Structure of GR4J model. Extracted from [*Perrin et al.*, 2003].

5.4.2 Parameter Uncertainty

Model calibration involves finding a set of parameters that maximise the similarity between the simulated and observed streamflow, with the similarity measured using an objective function such as simple least squares or weighted least squares. The BATEAU software allows uncertainty bounds

to be estimated for each of the GR4J parameters, and these results were presented in the second milestone report [*Westra et al.*, 2012]. The 2.5 and 97.5 percentiles of each parameter were within 15% of the maximum likelihood parameter estimates in all cases except for parameter θ_4 at Scott Creek. This parameter relates to the timing of the hydrograph, and given that the model is run at a daily time step and the response time in Scott Creek is much shorter than a day, the uncertainty of this parameter is because this parameter is very small (close to zero) and thus appears to vary substantially as a percentage but does not vary substantially in absolute terms.

The conclusion is that parametric uncertainty is a relatively minor source of uncertainty, particularly as the model being used (GR4) is relatively parsimonious and the observational dataset available for model calibration is relatively long. This is discussed in more detail in Section 5.5 where the role of different sources of uncertainty on model predictions is evaluated.

5.4.3 Structural Uncertainty

The preceding sections showed that uncertainty associated with input and output data, together with uncertainty in the estimation of the GR4J model parameters, collectively are unlikely to explain the full hydrological model uncertainty. The remaining uncertainty is attributable to model structure, since the GR4J model is unable to completely represent the complex nature of the transformation from rainfall to runoff.

In the second milestone report, a number of diagnostics were adopted to identify systematic areas of model bias. Diagnostics included traditional performance measures such as the Nash-Sutcliffe efficiency (NSE), annual flow volumes, the quantile distribution of annual flows, monthly flow totals, and various versions of the flow duration curve. The results are presented in detail in *Westra et al*. [2012] and also covered for Scott Creek in Appendix 1. They are summarised briefly here:

- The NSE ranged from 0.599 to 0.782 during the exploratory analysis (calibration) and from 0.476 to 0.803 during the confirmatory analysis. The highest (best) values were for Houlgrave Weir.

- GR4J generally overestimated the annual flow during both the exploratory and confirmation periods, with the overestimation being larger during the confirmatory period. No systematic biases could be detected for the simulation of low flow years relative to high flow years, suggesting that any difficulties in reproducing annual flow volumes were consistent across all flow years.

- Flows were consistently overestimated during the spring drying period, suggesting the presence of biases in how the model represents the hydrograph recession.

- GR4J substantially overestimated the duration of the hydrograph recessions, which is consistent with the finding that the model overestimates spring runoff. The weakness of GR4J in simulating hydrograph recessions may be attributable to the inability of the model to simulate cease-to-flow conditions.

- GR4J performs well on the rising limb of the hydrograph.

- In addition to the above diagnostics, each of which are commonly used in hydrological model evaluation, an additional diagnostic was proposed in which the hydrological model parameters were simulated as time-varying functions of a set of covariates, including a sinusoidal function with a period of one year, the 365-day antecedent rainfall and PET, and a linear trend. This was conducted to test whether the hydrological model exhibited non-stationary behaviour, in which the model parameters were found to be inconsistent from one period to the next. The conclusion of this analysis is that the hydrological model was non-stationary, with significant evidence of GR4J parameter θ_1 varying seasonally as well as increasing systematically over the calibration period.

The overall conclusion of the second milestone report was that there were a number of structural deficiencies associated with the GR4J model, particularly because of the identified non-stationarity of parameter θ_1, which need to be accounted for when developing projections of future hydrological response. Furthermore, the consistent biases in the simulation of the recession limb of the hydrograph suggest that additional flexibility is required in simulating hydrograph recessions. These changes have led to a new class of non-stationary hydrological model, described briefly in Section 6 and extensively in Appendix 1.

5.5 Impact of hydrological model errors on predictions

The previous sections described the primary sources of error associated with modelling the rainfall-runoff transformation. In this section we assess the implications of the errors on model predictions.

5.5.1 Evaluating the role of input error on the model parameters

The contribution of the input error to the total error is examined for Houlgrave Weir, as this catchment represents the largest fraction (83%) of the combined catchment area. It should be noted, however, that as discussed in Section 5.3.1, the input errors are slightly lower in Houlgrave Weir compared to the other catchments, and therefore the input error bands for the Scott Creek and Echunga Creek sub-catchments are likely to be slightly wider than for Houlgrave Weir.

The effect of input errors on the overall predictive errors is shown in Figure 16. Here, the partial predictive distribution of the input error is shown as blue shading, while the total predictive error (comprising the combination of input, output and model structure error) is given as red shading. As can be seen, the input error represents a significant contribution of total error for medium and high flows, but a lower contribution of the low flows.

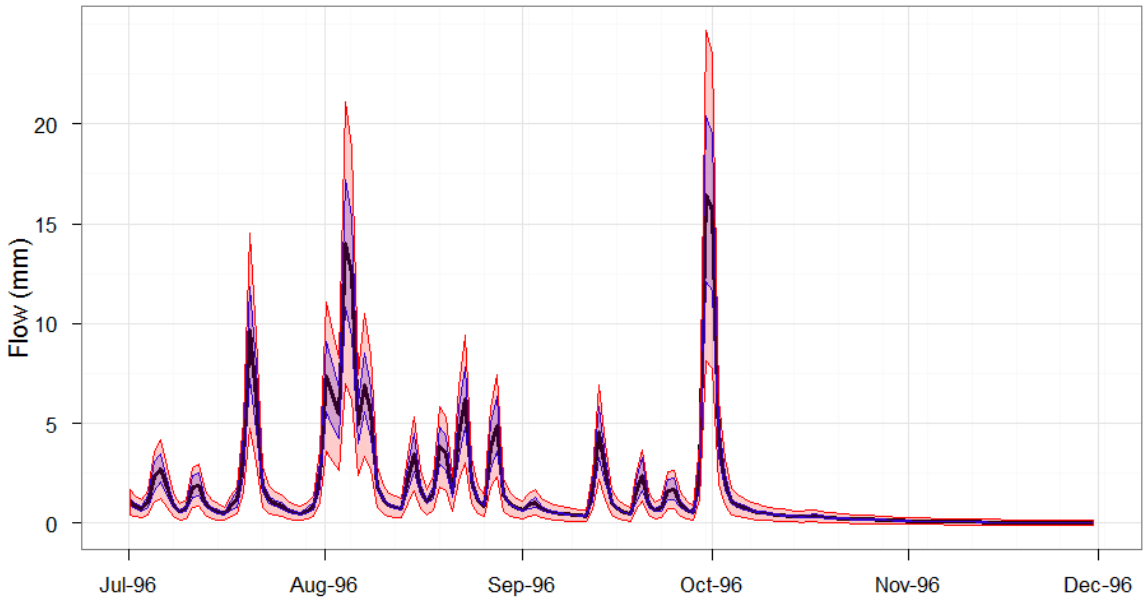


Figure 16: Implications of input error on total error for Houlgrave Weir a five-month period in 1996. Blue shading indicates rainfall error, while red shading indicates the residual error model.

The implications of the input errors on the maximum likelihood of the hydrological model GR4J's parameter estimates are also examined. As shown in Table 6, the change in GR4J parameters θ_1 and θ_4 are relatively small (with the largest parameter difference being -6% for θ_1), with larger differences for θ_2, which controls the groundwater exchange, and θ_3, which governs the one day-ahead maximum capacity of the routing store. The input error also reduces the residual error model parameters (a_ε and b_ε), which represents the portion of the residual error that is not attributed to errors in the model inputs.

Ultimately, the most important measure of the effect of incorporating input error into the hydrological modelling is to examine the model predictions, which is the subject of the following section.

Table 6: Implication of accounting for input error on the maximum likelihood estimate of the model parameters

Parameter	No Input Error	Input Error	% change
Houlgrave Weir			
θ_1	358	339	-6%
θ_2	-0.520	-0.812	36%
θ_3	9.30	11.6	20%
θ_4	1.31	1.31	0.6%
a_ε	0.07	0.0652	-8%
b_ε	0.302	0.240	-26%
Echunga Creek			
θ_1	367	353	-4%
θ_2	-2.20	-2.95	26%
θ_3	8.85	11.4	22%
θ_4	1.21	1.21	-0.2%
a_ε	0.0312	0.0243	-29%
b_ε	0.419	0.352	-19%
Scott Creek			
θ_1	388	401	3%
θ_2	-1.11	-1.65	33%
θ_3	11.5	15.6	27%
θ_4	1.18	1.17	-0.6%
a_ε	0.0597	0.0589	-1%
b_ε	0.408	0.281	-45%

5.5.2 Comparing the effect of different sources of uncertainty on the hydrological predictions

To assess the total effect of input, parameter, output and residual model errors on predictions, we present uncertainty bounds for Houlgrave Weir over the exploratory period (1977-1999). The uncertainty bounds are calculated using each type of error for the flow duration curves (Figure 17 for all flows above 0.01 mm, and Figure 18 for all flows above 1 mm), and annual average flows together with the 95 and 99 percentile of the flow duration curves (Table 7).

The uncertainty bounds were calculated based on 1000 stochastic replicates, obtained as follows:

- The input error distribution described in Section 5.3.1 was applied to the observed time series of rainfall, which was simulated through GR4J using the maximum likelihood-derived input error model parameters given in Table 6.
- The residual error distribution was simulated using the heteroscedastic residual error model using the calibrated parameters a_ε and b_ε, using the maximum likelihood estimates of parameters given in Table 6 without accounting for input error.
- The parameter error distribution was simulated using the posterior distribution of the GR4J parameters, calculated using the BATEA software.
- The output error was simulated using the output error model parameters given in Westra et al. [2012].

The results of the analysis show that the input error model and the residual error model produce uncertainty bounds of similar width, whereas the parameter error distribution and the output error distribution produce narrower uncertainty bounds. This is particularly evident in the zoomed-in portion of the flow duration curve (Figure 18) and in Table 7.

There remains a bias between the simulated flows and the observed flows, which is not fully accounted for by any of the error models. There are a number of possible reasons for this:

- The error models do not account for autocorrelation, and thus will lead to narrower bounds for specific quantiles and for aggregated variables such as annual average flows. This is discussed at length in Appendix 1.
- High very flow days (~2% of flows) are being under predicted, and the linear heteroscedastic model does not account for this.
- There are also structural model biases for low flows, due to the lack of a cease-to-flow mechanism in GR4J.

The issue of structural model errors are the basis for recommending multiple hydrological models for developing climate change projections, which is discussed more fully in Section 6.

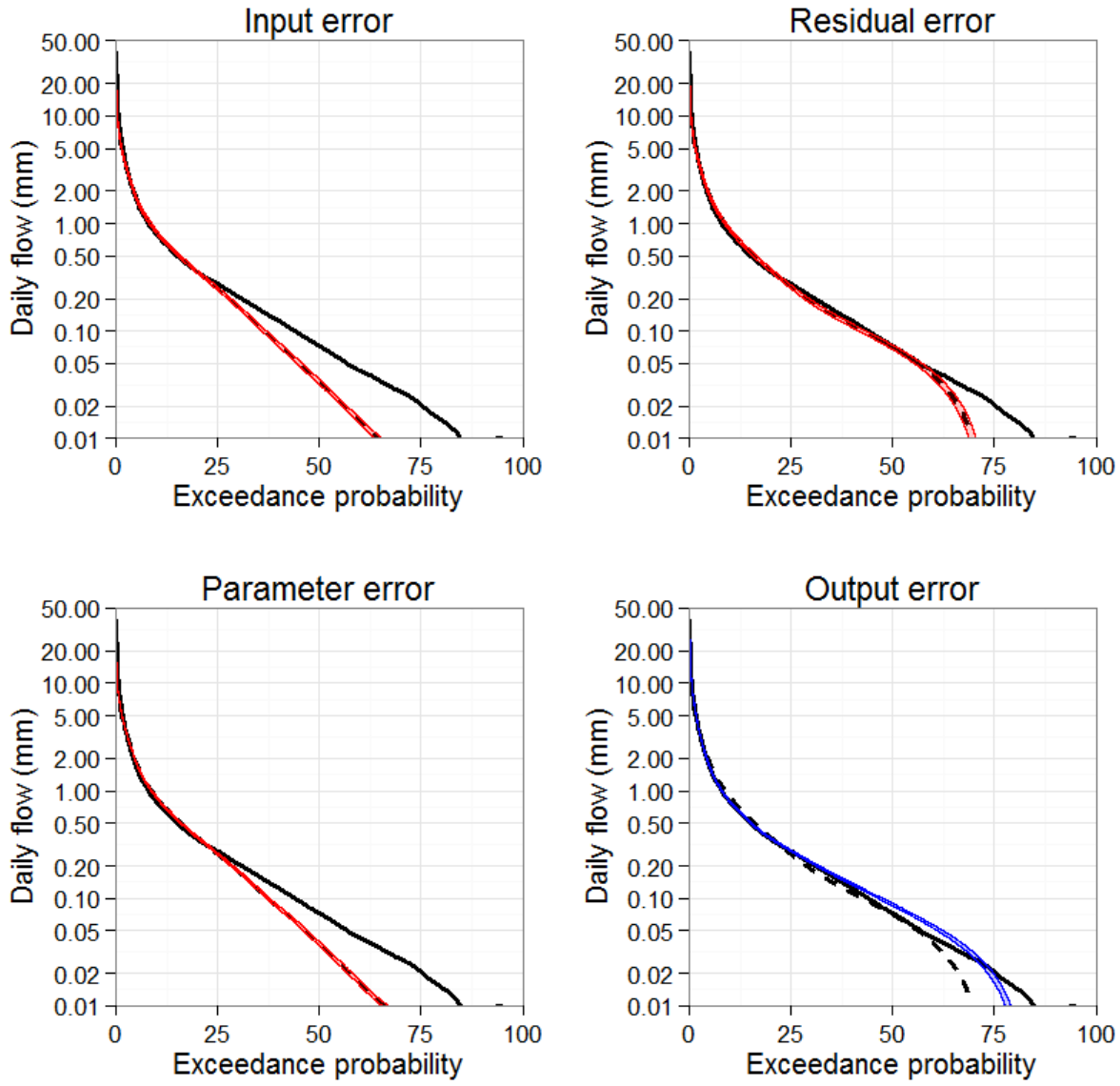


Figure 17: Uncertainty intervals (2.5% and 97.5%) for alternative error models, for flows greater than 0.01mm. Red shading indicates that the errors are calculated relative to simulated flows, and blue shading indicates that errors are calculated relative to observed flows.

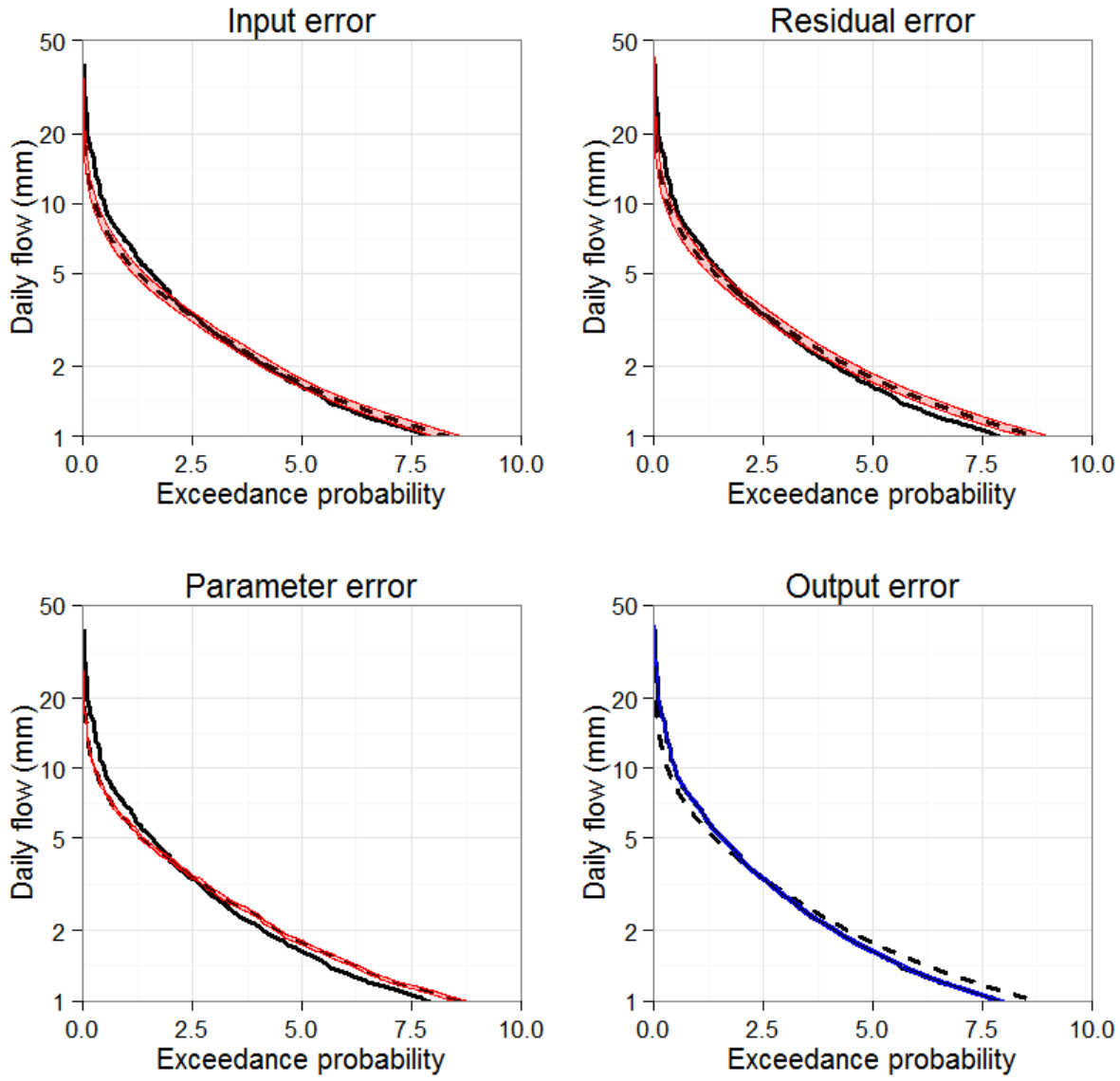


Figure 18: Uncertainty intervals (2.5% and 97.5%) for alternative error models, for flows greater than 1mm. Red shading indicates that the errors are calculated relative to simulated flows, and blue shading indicates that errors are calculated relative to observed flows.

Table 7: Observed flow and uncertainty ranges for different types of hydrological model errors.

	Observed	Input error only	Residual error model	Parameter error only	Output error only*
Annual mean	*152.1*	*127.5-143.0*	*130.6-150.1*	*137.2-143.0*	*151.5-159.9*
95 percentile flow	*1.63*	*1.61-1.76*	*1.70-1.86*	*1.74-1.82*	*1.60-1.67*
99 percentile flow	*6.90*	*5.36-6.09*	*5.65-6.42*	*5.88-6.09*	*6.71-7.06*

* Output error is calculated relative to the observed flow, whereas the other error models are calculated relative to the simulated flow series.

5.6 Summary of uncertainty modelling

A detailed analysis of hydrological model uncertainty was conducted in the second milestone using the Bayesian Total Error Analysis (BATEA) methodology. This analysis was documented in *Westra et al*. [2012], and summarised briefly in the sections above. The conclusions are that:

- Input uncertainty, particularly the uncertainty associated with deriving spatial rainfall estimates based on gauges and radar, was found to be an important source of uncertainty for medium and high flows, but less so for low flows. The magnitude of input uncertainty was similar to the magnitude of uncertainty captured in the residual error model.

- Output uncertainty was moderate based on a comprehensive rating curve analysis. The likely reasons are the relatively stable rating curves and high number of streamflow gaugings for each of the streamflow sites. Timing issues when removing Murray pipeline flows from the recorded flows at Houlgrave Weir were addressed by adopting a censoring approach during model calibration, to ensure that the calibration procedure adopted to estimate the final parameter sets was not affected by timing errors.

- Parameter uncertainty was small based on a simulation of the joint posterior distribution of the parameters to obtain model predictions, indicating that the record length is sufficient relative to the model complexity to enable precise estimation of model parameters.

- Structural uncertainty (the uncertainty due to the hydrological model structure not being able to reflect true flow behaviour) was identified as an important source of total predictive uncertainty. This was based on a detailed analysis of model diagnostics including flow duration curves, the rising and falling limb of the hydrograph, and information-theoretic measures that assess the non-stationarity of hydrological model parameters (Section 6).

These conclusions need to be considered in the context of the intended purpose of the hydrological modelling, which is to assess the impact of future changes to rainfall and PET on runoff. This analysis is to be undertaken relative to historical GCM-based simulations of rainfall and PET, and thus observational data is only used to select the hydrological model and obtain its parameters. This issue is discussed more fully in the second volume of this series.

The difficulty in simulating hydrograph recessions, as well as the identified non-stationarity of parameter θ_1, has led to the development of multiple alternative hydrological model structures to be used as the basis for predicting hydrological response under future climate forcings in the Onkaparinga. This is described further in the following section.

6 Non-Stationary Model Development

6.1.1 Addressing structural model uncertainty

In the previous section, structural model uncertainty was identified as the principal source of hydrological uncertainty for the Onkaparinga catchment. A set of new 'GR4J-like' models have therefore been developed, which have all been derived from the standard GR4J model of Perrin et al. [2003], but modified to address a number of the structural model deficiencies.

The structural deficiencies were first described in *Westra et al*. [2012], and include a systematic overestimation of the duration of the hydrograph recessions, together hydrological model parameter non-stationarity, in which the hydrological model parameters vary in time, and thus depend on the period of record used for their estimation.

Hydrological model non-stationarity was expressed in terms of variability of the maximum capacity of the production store as a function of time and other covariates such as the seasonal cycle and the previous 365 day's rainfall and potential evapotranspiration. This non-stationarity represents a major limitation when applying the model to predict runoff under future climate change, since the hydrometeorological forcing conditions (i.e. the rainfall and PET) will likely be very different to those experienced in the historical record.

A total of 22 versions of GR4J were developed to address these structural deficiencies, and were designed to achieve the following aims:

1. to simulate non-stationary model behaviour by allowing the primary parameter governing the maximum capacity of the production store (θ_1) to change as a function of time-varying parameters, including (i) an annual sinusoid; (ii) the previous 365-day rainfall and potential evapotranspiration; and (iii) a linear trend; and

2. to compare the performance of several alternative model structures in their ability to reduce parameter non-stationarity.

These modifications are discussed at length in Appendix 1, and form part of a manuscript that has been published in Water Resources Research [*Westra et al.*, 2014a]. A range of hydrologically oriented diagnostics were used to evaluate each of the 22 models, including the Nash-Sutcliffe coefficient of efficiency (NSE), annual flow volumes, flow duration curves as well as an information theoretic measure (the Akaike Information Criterion). Based on a detailed analysis of these diagnostics, it was shown that for the Scott Creek catchment used for model development, most of the modified versions of GR4J performed significantly better than the standard version. The assessment was based both on an analysis over an exploratory period used for parameter estimation, as well as over an independent (and much drier) confirmatory period. This latter conclusion is particularly important given the expectation (described in the third volume of this report series) that the future climate will be much drier than the historical climate. The manuscript in Appendix 1 presents the results for the Scott Creek sub-catchment.

6.1.2 Selecting models for use in the climate change assessment

Many of the modified GR4J models outperform the standard GR4J model, however based on the non-stationarity analysis described in Appendix 1, it is difficult to identify a single 'best' model to form the basis of future climate change projections. Given that it is unlikely that any of the models will perfectly represent the rainfall-runoff transformation, an ensemble of four models was ultimately selected for use. These include:

1. Model $g_{1.1}$ (the standard GR4J model) as a benchmark against which other models can be evaluated;

2. Model $g_{1.8}$, which accounts for non-stationarity due to seasonal variability, the effect of the previous 365-day rainfall and PET as well as a linear trend in the capacity of the production store;

3. Model $g_{2.2}$, which incorporates an additional parameter to control the portion of rainfall that enters the production store;

4. Model $g_{3.11}$, which accounts for non-stationarity due to seasonal variability, the effect of the previous 365-day rainfall and PET as well as a linear trend in the capacity of the production store, as well as an additional parameter to control the portion of rainfall that enters the production store.

The choice of these four models was based on including the standard GR4J model for comparison purposes and selecting a subset of the 22 models trialled in Westra et al. [2014a] that performed well in the exploratory and confirmatory analyses, while also capturing a range of model structures.

Models $g_{1.8}$ and $g_{3.11}$ both include a linear trend in the maximum capacity of the production store (θ_1). Rather than extrapolate this linear trend into the future, which is likely to be unreliable for long future time horizons, the contribution of this predictor at the end of the calibration period (31/12/1999) is held at the same value for future simulations. Although further research is required before it is possible to attribute the trend to a particular feature of catchment change, it is likely that at least part of the trend is attributable to an increase in on-farm dams, and the development of on-farm dams has slowed considerably since the 1990s due to increased regulation [*Teoh*, 2002]. Therefore, the assumption of fixing the 'trend' component to its value at 31/12/1999 is likely to be more physically realistic compared to the alternative of allowing this trend to continue linearly until 2100. These models are denoted as $g_{1.8}^*$ and $g_{3.11}^*$ – they are equivalent to models $g_{1.8}$ and $g_{3.11}$ except that it is assumed that the linear trend does not continue past the year 2000. The final calibrated model parameters for each of the four models and three sub-catchments, as well as the residual error model parameters (a_ε and b_ε), are presented in Table 8.

The performance of the four selected models in the exploratory period is presented in Table 9, and the performance in an independent confirmatory period is presented in Table 10. Performance metrics include annual average flows, various quantiles of the flow duration curve, and the NSE. As can be seen from both periods, the model performance is quite variable between the different models, although the NSE values were generally quite high, being above 0.7 in all cases and with

approximately half the models having NSE values above 0.8. Over the exploratory period, the models $g_{1.1}$ and $g_{2.2}$ perform reasonably well in estimating the average annual flow, except for Houlgrave Weir where the flows are underestimated by 9.6% and 7.8%, respectively.

Interestingly, models $g_{1.8}^{*}$ and $g_{3.11}^{*}$ both underestimate total annual flow for Scott Creek and Houlgrave Weir and overestimate flow at Echunga Creek. This issue was much less apparent for models $g_{1.8}$ and $g_{3.11}$ (performance metrics shown in parentheses below the $g_{1.8}^{*}$ and $g_{3.11}^{*}$ in Table 9), with both models simulating an increasing trend in θ_1 for Scott Creek and Houlgrave Weir and decreasing trend for Echunga Creek. Thus, it is likely that the biases in models $g_{1.8}^{*}$ and $g_{3.11}^{*}$ are due to the fixing of the trend parameter at 31/12/1999 levels, whereas in reality the catchment stores have changed over the exploratory period.

Considering other flow metrics, it can be seen that none of the models are clearly superior in representing all the flow percentiles. For example, considering model $g_{3.11*}$ over the confirmatory period, it can be seen that this model performed well in simulating average annual flows for all catchments, but performed poorly for low flows. This model was also the best model out of the four selected models for high flows (95 and 99 percentiles) at Houlgrave Weir, second best at Echunga Creek, and had relatively good performance at Scott Creek for the 95 percentile but relatively poor performance at Scott Creek for the 99 percentile.

As a result of this analysis, there does not appear to be a single model that clearly outperforms the remaining models across all flow metrics, highlighting the importance of an ensemble of models to produce climate change projections, rather than using a single model that is deemed 'best' according to one or more performance criteria.

Table 8: Calibrated model parameters for the models used for the climate impact assessment, presented to two significant figures. Models $g_{1.8}^{*}$ and $g_{3.11}^{*}$ are equivalent to $g_{1.8}$ and $g_{3.11}$, except that trend term is calculated at 31/12/1999 and incorporated into parameter p_1.

	Parameter	Scott Creek	Echunga Creek	Houlgrave Weir
Model $g_{1.1}$	λ_1	390	370	360
	θ_2	-1.1	-2.2	-0.52
	θ_3	11	8.8	9.3
	θ_4	1.2	1.2	1.3
	a_ε	0.060	0.031	0.070
	b_ε	0.41	0.42	0.30
Model $g_{1.8}$	λ_1	2400	-1200	-440
	λ_2	0.055	-0.010	0.0083
	λ_3	170	100	110
	λ_4	160	110	250
	λ_5	-0.16	-0.018	0.042
	λ_6	-1.4	1.1	0.39
	θ_2	-0.33	-8.5	-2.0
	θ_3	9.0	25	16
	θ_4	1.2	1.2	1.3
	a_ε	0.021	0.010	0.071
	b_ε	0.44	0.43	0.28
Model $g_{1.8^*}$	λ_1	3000	-1300	-360
	λ_3	170	100	110
	λ_4	160	110	250
	λ_5	-0.16	-0.018	0.042
	λ_6	-1.4	1.1	0.39
	θ_2	-0.33	-8.5	-2.0
	θ_3	9.0	25	16
	θ_4	1.2	1.2	1.3
	a_ε	0.021	0.010	0.071
	b_ε	0.44	0.43	0.28

Model $g_{2.2}$	λ_1	480	380	380
	θ_2	-2.2	-3.8	-0.92
	θ_3	25	16	15
	θ_4	1.2	1.2	1.3
	θ_5	1.3	1.5	1.5
	a_ε	0.033	0.017	0.057
	b_ε	0.42	0.44	0.31
Model $g_{3.11}$	λ_1	2500	-780	7.4
	λ_2	0.068	-0.018	0.021
	λ_3	180	10	96
	λ_4	120	120	14
	λ_5	-0.41	0.025	-0.18
	λ_6	-1.4	0.76	0.39
	θ_2	-1.5	-10	-1.4
	θ_3	20	27	18
	θ_4	1.1	1.2	1.3
	θ_5	1.3	1.7	1.3
	a_ε	0.020	0.0086	0.056
	b_ε	0.2	0.43	0.30
Model $g_{3.11}^*$	λ_1	3200	-850	200
	λ_3	180	10	96
	λ_4	120	120	14
	λ_5	-0.41	0.025	-0.18
	λ_6	-1.4	0.76	0.39
	θ_2	-1.5	-10	-1.4
	θ_3	20	27	18
	θ_4	1.1	1.2	1.3
	θ_5	1.3	1.7	1.3
	a_ε	0.020	0.0086	0.056
	b_ε	0.2	0.43	0.30

Table 9: Performance of the four selected models described in Table 8 relative to observed flow over the exploratory period. Results in parentheses represent simulation results from $g_{1.8}$ and $g_{3.11}$ (i.e. the models including linear trend).

	Observed flow	$g_{1.1}$	$g_{1.8*}$ ($g_{1.8}$)	$g_{2.2}$	$g_{3.11*}$ ($g_{3.11}$)
Scott Creek					
Annual average	138.6	139.2	116.0 (128.2)	129.9	116.9 (131.3)
10 percentile	0.00596	0.00203	0.00809 (0.00649)	0.00571	0.0085 (0.00721)
50 percentile	0.0878	0.0452	0.0880 (0.0793)	0.0797	0.0914 (0.0823)
95 percentile	1.55	1.78	1.37 (1.51)	1.59	1.370 (1.52)
99 percentile	5.25	4.98	3.53 (4.06)	4.03	3.34 (3.93)
NSE		0.806	0.701 (0.746)	0.805	0.729 (0.776)
Echunga Creek					
Annual average	62.2	62.5	67.2 (62.7)	59.4	70.6 (62.0)
10 percentile	0.00222	0.000389	0.000830 (0.000655)	0.000811	0.000949 (0.000618)
50 percentile	0.00712	0.00870	0.0116 (0.00892)	0.0129	0.0136 (0.00921)
95 percentile	0.529	0.697	0.721 (0.669)	0.660	0.755 (0.674)
99 percentile	2.89	3.14	2.69 (2.59)	2.66	2.83 (2.52)
NSE		0.826	0.774 (0.811)	0.792	0.724 (0.798)
Houlgrave Weir					
Annual average	152.1	140.3	137.4 (152.4)	137.5	124.6 (139.7)
10 percentile	0.00248	0.00265	0.00212 (0.00232)	0.00412	0.00478 (0.00445)
50 percentile	0.0724	0.0374	0.0323 (0.0389)	0.0516	0.0542 (0.0503)
95 percentile	1.63	1.79	1.74 (0.192)	1.70	1.52 (1.71)
99 percentile	6.90	6.02	5.76 (6.55)	5.48	4.75 (5.64)
NSE		0.811	0.822 (0.814)	0.800	0.760 (0.805)

Table 10: Performance of the four selected models described in Table 8 relative to observed flow over the confirmatory period.

	Observed flow	$g_{1.1}$	$g_{1.8*}$	$g_{2.2}$	$g_{3.11*}$
Scott Creek					
Annual average	107	125	106	118	108
10 percentile	0.00298	0.00193	0.00753	0.00571	0.00844
50 percentile	0.0699	0.0468	0.0837	0.0826	0.0895
95 percentile	1.03	1.48	1.16	1.34	1.20
99 percentile	4.20	4.92	3.31	3.82	3.32
NSE		0.773	0.747	0.804	0.779
Echunga Creek					
Annual average	64.2	61.8	57.5	58.2	60.3
10 percentile	0.00222	0.000367	0.000581	0.000731	0.000743
50 percentile	0.00665	0.00976	0.0105	0.0155	0.0139
95 percentile	0.697	0.720	0.675	0.699	0.721
99 percentile	3.03	2.88	2.58	2.55	2.68
NSE		0.774	0.802	0.756	0.808
Houlgrave Weir					
Annual average	118	136	133	134	120
10 percentile	0.00243	0.00261	0.00205	0.00408	0.00502
50 percentile	0.0680	0.0426	0.0366	0.0589	0.0630
95 percentile	1.30	1.80	1.70	1.70	1.49
99 percentile	4.47	5.02	4.95	4.86	4.10
NSE		0.825	0.838	0.836	0.845

7 Summary and Conclusions

This is the first of three final reports for the University of Adelaide component of *Task 4: Application Test Bed* for the Goyder Climate Change project. The focus of this report is to identify the principal sources of hydrological uncertainty, including the relative contributions of model input, output and structural errors. The BATEA methodology is used as the basis of the analysis. Findings are then used to improve the model structure, leading to a set of models that significantly improve model predictions. These models will be used to produce climate projections (to be covered in the second [*Westra et al.*, 2014b] and third [*Westra et al.*, 2014c] volumes of this report series).

The outcomes of the uncertainty analysis are as follows:

- Input uncertainty, particularly the uncertainty associated with deriving spatial rainfall estimates based on gauges and radar, was found to be an important source of uncertainty for medium and

high flows, but less so for low flows. The magnitude of input uncertainty was similar to the magnitude of uncertainty captured in the residual error model.

- Output uncertainty was moderate based on a comprehensive rating curve analysis. The likely reasons are the relatively stable rating curves and high number of streamflow gaugings for each of the streamflow sites. Timing issues when removing Murray pipeline flows from the recorded flows at Houlgrave Weir were addressed by adopting a censoring approach during model calibration, to ensure that the calibration procedure adopted to estimate the final parameter sets was not affected by timing errors.

- Parameter uncertainty was small based on a simulation of the joint posterior distribution of the parameters to obtain model predictions, indicating that the record length is sufficient relative to the model complexity to enable precise estimation of model parameters.

- Structural uncertainty (the uncertainty due to the hydrological model structure not being able to reflect true flow behaviour) was found to be the dominant source of total predictive uncertainty. This was based on a detailed analysis of model diagnostics including flow duration curves, the rising and falling limb of the hydrograph, and information-theoretic measures that assess the non-stationarity of hydrological model parameters.

The purpose of this uncertainty analysis described in this report was to identify the largest source(s) of uncertainty associated with the rainfall-runoff modelling transformation, and hence improve the reliability of the hydrological predictions that will be used to assess the impact of future GCM-derived projections of rainfall and PET on runoff in the Onkaparinga catchment. The future climate projections will be based on changes in flow relative to a 'baseline' climate obtained from GCM-derived historical simulations of rainfall and PET. Therefore in this analysis, observational (instrumental) data is not used directly in the development of future climate projections. Rather, the role of the observational data is to estimate the hydrological model parameters (including the residual error model parameters that provide estimates of hydrological uncertainty), and to apply calibration-based diagnostics as the basis for model selection.

To address the limitations of the standard GR4J model in simulating hydrograph recessions, as well as an identified non-stationary of parameter θ_1, a total of 21 alternative model structures were developed. These model structures included various combinations of the following:

- Sinusoidal variation in θ_1 with a period of one year;
- Allowing θ_1 to vary as a function of the previous 365-day rainfall and PET;
- Allowing θ_1 to vary as a function of a linear trend;
- Inclusion of an additional parameter that controls the proportion of net rainfall that enters the production store; and
- A modification to the way that actual evapotranspiration is estimated in the model.

The models were evaluated using the Akaike Information Criterion (AIC), as well as other model diagnostics including the Nash-Sutcliffe coefficient of efficiency, the annual flow volumes, and the flow duration curve. The standard GR4J model and the 21 alternative models were tested using these models over both the calibration period and a drier confirmatory (or validation) period from 2000 to 2009.

The modified models all showed improvements over the standard GR4J model, with the most notable improvements being due to the sinusoidal term for θ_1, and the inclusion of an additional parameter that controls the proportion of the net rainfall that enters the production store. Compared to the standard GR4J model that overestimated flows in the confirmatory period by 17%, the 'AIC-best' model ($g_{3.11}$) underestimated flows by only 2.6%, representing a significant improvement in the reliability of the hydrological predictions.

Based on these results, an ensemble of four hydrological models was selected to develop the future climate change projections for the Onkaparinga. These are:

- Model $g_{1.1}$ (the standard GR4J model) as a benchmark against which other models can be evaluated;
- Model $g_{1.8}$, which accounts for non-stationarity due to seasonal variability, the effect of the previous 365-day rainfall and PET as well as a linear trend in the capacity of the production store;
- Model $g_{2.2}$, which incorporates an additional parameter to control the portion of rainfall that enters the production store;
- Model $g_{3.11}$, which accounts for all the forms of non-stationarity described in Equation 1 of Appendix 1 as well as an additional parameter to control the portion of rainfall that enters the production store.

Models $g_{1.8}$ and $g_{3.11}$ incorporates the effects of a linear trend, and rather than extrapolate this linear trend into the future, the contribution of this predictor at the end of the calibration period (31 December 1999) is held at the same value for future simulations. Although further research is required before it is possible to attribute the trend to a particular feature of catchment change, it is likely that at least part of the trend is attributable to an increase in on-farm dams. Given this, the decision to fix the trend parameter at the 1999 value is because the development of further on-farm dams has slowed due to increased regulation of the construction of new dams. It is recommended that future investigation be conducted into the farm-dam storage capacity over the exploratory and confirmatory periods, as this could assist in trend attribution. For example, the storage capacity could be used as a covariate for the non-stationary model for θ_1, and this could be achieved in a more spatially distributed fashion using a larger number of sub-catchments to capture the different rates of farm dam construction in different parts of the Onkaparinga.

These four hydrological models will be used as the basis of projection of future hydrological response in the Onkaparinga as a result of anthropogenic climate change. The hydrological model

uncertainty will be combined with uncertainty due to the GCM and the representative concentration pathway (RCP), to provide a thorough exploration of the uncertainty associated with climate change projections.

Before developing the climate change projections in Volume 3 of this series, we turn to evaluating the performance of historical simulations of runoff derived from NHMM simulations of rainfall and PET. This will be the focus of volume 2 of this series.

References

Allen, R. G., L. S. Pereira, D. Raes, and M. Smith (1998a), Statistical anlaysis of weather datasets, in *Crop Evapotranspiration - Guidelines for Computing Crop Water Requirements*, edited, FAO - Food and Agriculture Organisation of the United Nations.

Allen, R. G., L. S. Periera, D. Raes, and M. Smith (1998b), Crop Evapotranspiration: Guidelines for Computing Crop Requirements, Irrigation and Drainage Paper No. 56*Rep.*, 300 pp, FAO, Rome, Italy.

Henderson-Sellers, A. (1993), An antipodean climate of uncertainty? , *Climatic Change*, *25*(3-4), 203-224.

Heneker, T. M., and D. Cresswell (2010), Potential Impact on Water Resource Availability in the Mount Lofty Ranges due to Climate Change*Rep.*, Government of South Australia, through Department for Water, Adelaide.

IPCC (2000), Special Report on Emission Scenarios *Rep.*, WMO and UNEP.

Jeffrey, S. J., J. O. Carter, K. B. Moodie, and A. R. Beswick (2001), Using spatial interpolation to construct a comprehensive archive of Australian climate data, *Environmental Modelling & Software*, *16*(4), 309-330.

Jones, R. N., F. H. S. Chiew, W. C. Boughton, and L. Zhang (2006), Estimating the sensitivity of mean annual runoff to climate change using selected hydrological models, *Advances in Water Resources*, *29*(10).

Kavetski, D., S. W. Franks, and G. Kuczera (2002), Confronting input uncertainty in environmental modeling, in *Calibration of Watershed Models*, edited by Q. Y. Duan, H. V. Gupta and S. Sorooshian, pp. 49-68, AGU series, Washington.

Kavetski, D., G. Kuczera, and S. W. Franks (2006), Bayesian analysis of input uncertainty in hydrological modeling: 2. Application, *Water Resources Research*, *42*(3).

Kuczera, G., D. Kavetski, S. W. Franks, and M. Thyer (2006), Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterizing model error using storm-dependent parameters, *Journal of Hydrology 331*(1-2), 161-177.

Leonard, M., M. Thyer, M. F. Lambert, H. R. Maier, and G. Dandy (2011), Task 4, Application Test Bed, Onkaparinga Catchment Case Study: Surface Water Hydrological Modelling*Rep.*, University of Adelaide, Adelaide.

Li, L., R. J. Donohue, T. R. McVicar, T. G. van Niel, J. Teng, N. J. Potter, I. N. Smith, D. G. C. Kirono, J. M. Bathols, W. Cai, S. P. Marvanek, S. N. Gallant, F. H. S. Chiew, and A. J. Frost (2009), Climate data and their characterisation for hydrological scenario modelling across northern Australia*Rep.*, CSIRO.

McMahon, T. A., M. C. Peel, L. Lowe, R. Srikanthan, and T. R. McVicar (2013), Estimating actual, potential, reference crop and pan evaporation using standard meteorological data: a pragmatic synthesis, *Hydrological Earth Systems Science*, *17*, 1331-1363.

Merz, R., and G. Bloschl (2004), Regionalisation of catchment model parameters, *Journal of Hydrology 287*(1-4).

Oreskes, N., K. Shrader-Frechette, and K. Belitz (1994), Verification, Validation and Confirmation of Numerical Models in the Earth Sciences, *Science*, *263*(5147), 641-646.

Perrin, C., C. Michel, and V. Andreassian (2003), Improvement of a parsimonious model for streamflow simulation, *Journal of Hydrology*, *279*, 275-289.

Renard, B., D. Kavetski, E. Leblois, M. Thyer, G. Kuczera, and S. W. Franks (2011), Toward a reliable decomposition of predictive uncertainty in hydrological modelling: Characterizing rainfall errors using conditional simulation, *Water Resources Research*, *47*(W11516).

Spieghalter, D. J., A. Thomas, and N. Best (2003), WinBugs, Version 1.4, User Manual*Rep.*, MRC and Imperial College of Science, Technology and Medicine.

Teoh, K. S. (2002), Estimating the Impact of Current Farm Dams Development on the Surface Water Resources of the Onkaparinga River Catchment*Rep.*, Department of Water, Land and Biodiversity Conservation.

Thyer, M., B. Renard, D. Kavetski, G. Kuczera, S. W. Franks, and R. Srikanthan (2009), Critical evaluation of parameter consistency and predictive uncertainty in hydrological modelling: A case study using Bayesian total error analysis, *Water Resources Research*, *45*(W00B14).
Villarini, G., and W. F. Krajewski (2008), Empirically-based modeling of spatial sampling uncertainties associated with rainfall measurements by rain gauges, *Advances in Water Resources*, *31*(7), 1015-1023.
Westra, S., M. Leonard, M. Thyer, D. Kavetski, and M. Lambert (2012), Task 4, Application Test Bed, Onkaparinga Catchment Case Study: Surface Water Hydrological Model. Milestone 2 Report*Rep.*
Westra, S., M. Thyer, M. Leonard, D. Kavetski, and M. Lambert (2014a), A strategy for diagnosing and interpreting hydrologic non-stationarity, *Water Resources Research*, *50*, 5090-5113.
Westra, S., M. Thyer, M. Leonard, and M. Lambert (2014b), Impacts of Climate Change on Surface Water in the Onkaparinga Catchment - Volume 2: Hydrological Evaluation of the CMIP3 and CMIP5 GCMs and the Non-homogenous Hidden Markov Model (NHMM)*Rep.*, Adelaide, South Australia.
Westra, S., M. Thyer, M. Leonard, and M. Lambert (2014c), Impacts of Climate Change on Surface Water in the Onkaparinga Catchment - Volume 3: Impacts of Climate Change on Runoff *Rep.*

Appendix 1: "A Strategy for diagnosing and interpreting hydrological non-stationarity" – manuscript published in Water Resources Research

A strategy for diagnosing and interpreting hydrological model non-stationarity

Seth Westra, Mark Thyer, Michael Leonard, Dmitri Kavetski and Martin Lambert

School of Civil, Environmental and Mining Engineering,

University of Adelaide, Adelaide, Australia 5005

E-mail: seth.westra@adelaide.edu.au; mark.thyer@adelaide.edu.au; michael.leonard@adelaide.edu.au;

dmitri.kavetski@adelaide.edu.au; martin.lambert@adelaide.edu.au

Keywords: hydrological modeling, non-stationarity, model selection, model diagnostics, AIC, GR4J

Abstract

This paper presents a strategy for diagnosing and interpreting hydrological non-stationarity, with the aim of improving hydrological models and their predictive ability under changing hydroclimatic conditions. The strategy consists of four elements: (i) detecting potential systematic errors in the calibration data; (ii) hypothesizing a set of non-stationary parameterisations of existing hydrological model structures, where one or more parameters vary in time as functions of selected covariates; (iii) trialing alternative stationary model structures to assess whether non-stationarity can be reduced by modifying the model structure; and (iv) selecting one or more models for prediction. The Scott Creek catchment in South Australia and the hydrological model GR4J are used to illustrate the strategy. Streamflow predictions improve significantly when the GR4J parameter describing the maximum capacity of the production store is allowed to vary in time as a combined function of: (i) an annual sinusoid; (ii) the previous 365-day rainfall and potential evapotranspiration; and (iii) a linear trend. This improvement provides strong evidence of model non-stationarity. Based on a range of hydrologically-oriented diagnostics such as flow-duration curves, the GR4J model structure was modified by introducing an additional calibration parameter that controls recession behaviour and by making actual evapotranspiration dependent only on catchment storage. Using information theoretic measures (the Akaike Information Criterion) for model selection, together with several hydrologically oriented diagnostics, it was shown that these modifications clearly improve predictive performance in the Scott Creek catchment. Based on a comparison of 22 versions of GR4J with different representations of non-stationarity and other modifications, the model selection approach applied in the exploratory period (used for parameter estimation) correctly identifies models that perform well in a much drier independent confirmatory period.

1. Introduction

The development of hydrological models that produce credible predictions under a changing climate is one of the most challenging aspects of hydrological modelling [Klemes, 1986]. This challenge is particularly pertinent when models are extrapolated outside the range of observed data used for parameter estimation, which is often necessary when looking at long lead times or high warming scenarios [Milly et al., 2008]. Moreover, under such conditions model evaluation and selection requires methods that make the best use of available historical data to assess the model's extrapolative ability [Anderson and Woessner, 1992; Oreskes et al., 1994].

One of the most stringent tests of hydrological model credibility is 'differential split sample testing' [Klemes, 1986]. In these tests, the performance of a calibrated model is evaluated on one or more periods that are climatologically different from the period used for parameter estimation; for example a model calibrated under "wet" conditions can be tested on a "dry" period, and vice versa. For a model capable of such extrapolation, parameter estimates and predictive performance should remain similar across the two periods. However, numerous studies concluded that parameter estimates depended on the calibration period [Gan and Burges, 1990; Wagener et al., 2003; Choi and Beven, 2007; Le Lay et al., 2007; Marshall et al., 2007; Wu and Johnston, 2007; Vaze et al., 2010; Merz et al., 2011; Zhang et al., 2011; Coron et al., 2012; Seiller et al., 2012]. Furthermore, seasonal variations of hydrological parameters have been reported by Ye et al. [1997] and Paik et al. [2005].

We define the term "hydrological model non-stationarity" as the situation where hydrological model parameters vary in time, and thus depend on the period of record used for their estimation. Such non-stationarity can lead to poor predictions, especially when the model is applied to a climatologically different period [Gharari et al., 2013]. For example, Coron *et al*. [2012] found that models calibrated to a period with a wetter climate overestimated the mean annual runoff when applied to a drier period, and vice versa. The severity of the non-stationarity problem and its implications on model prediction depend on multiple factors, including: (i) the length and variability of the historical record; (ii) the magnitude of future climate change; and (iii) the hydrological model [e.g. Brigode et al., 2012].

There are many possible reasons for hydrological model non-stationary, including systematic data errors, weaknesses in calibration procedures, numerical artefacts, model structural deficiencies and others [Beven and Binley, 1992; Wagener et al., 2003; Clark et al., 2011; Kavetski et al., 2011]. For example, streamflow records can become biased due to siltation of weirs and changes in the channel flow geometry [Guerrero et al., 2012]; rainfall records can be affected by changes in the location and quality of rain gauges [Molini et al., 2005], and so forth. Similarly, poor choice of objective function can cause non-stationarity in the calibrated model parameters. For example, Thyer *et al.* [2009] showed that calibration to different time periods using a standard least squares objective function produced distinctly different estimates of hydrological parameters; these discrepancies were substantially reduced when a weighed least squares objective function was used.

A fundamental concern with hydrological non-stationarity is the possible implication that one or more important physical processes are not adequately represented [Lin and Beck, 2007; de Vos et al., 2010], or that changes in the catchment (e.g. land use changes) are occurring but are not

explicitly represented by the model. We therefore argue that, provided that robust data, numerical methods and calibration procedures are used, hydrological model non-stationarity must be caused by the approximate nature of hydrological models relative to the physical system under investigation [Anderson and Woessner, 1992]. From this perspective, models with time-invariant parameters are more likely to be reliably representing the key physical processes. This is particularly important when predicting catchment response to future climatic forcings, as accurate process representation is critical when extrapolating a model outside of its calibrated range. Stationarity of model parameters can therefore be viewed as a necessary condition for the hydrological model to provide credible projections under extrapolation, and tests for stationarity can be useful as part of model selection for climate impact studies [Seiller et al., 2012].

A pragmatic approach to detect and mitigate non-stationarity is to calibrate the model to one or more historical periods that are analogous to the expected future hydroclimatic conditions [e.g. Vaze et al., 2010]. Provided such historical analogues are available, this approach reduces the extent of model extrapolation, and thus may be adequate for short future time horizons and small levels of climate change. An obvious limitation is that there may not be any historical periods that are sufficiently representative of the projected future conditions. This limitation can be particularly significant when it is recognised that hydroclimatic changes are expressed not only in terms of changes in annual average precipitation and potential evapotranspiration, but, just as importantly, in terms of the seasonality, intermittency and intensity of future precipitation events [Bates et al., 2008; Westra et al., 2013]. Furthermore, by maximising the 'similarity' of the historical climate sequences to the projected future climate, it becomes necessary to use only relatively short portions of the historical record for model calibration, so that potentially valuable information on catchment behaviour is ignored during parameter estimation. This is a type of bias-variance trade-off: to maximise the similarity between the calibration period and expected future climate (and hence reduce parameter bias), we need to use shorter periods of the historical record as the basis for calibration (which will usually increase parameter variance) [Brigode et al., 2012]. Furthermore, this approach does not characterise and/or resolve the cause of this non-stationarity.

This paper develops a strategy to diagnose non-stationarity in hydrological model parameters and identify possible causes that require further investigation. The major distinct elements of this strategy are the characterization of parameter non-stationarity by representing hydrological model parameter(s) as a function of a set of time-varying covariates, the trialling of alternative model structures, and the assessment of empirical support for each proposed description of non-stationarity and/or alternative model structures using multiple model selection criteria including information-theoretic metrics [Burnham and Anderson, 2010] and hydrological diagnostics [Gupta et al., 2008]. Compared to the existing approach of separately calibrating the hydrological model to different historical periods, the proposed approach has the following advantages:

1. A larger portion of the historical record is used for parameter estimation. This avoids the potential loss of information when discarding large portions of observed data.

2. By representing selected hydrological model parameters as continuous functions of selected covariates, it becomes possible to at least tentatively extrapolate these parameters to different hydroclimatic regimes (note that the difficulties of model evaluation under extrapolation

described by Klemes [1986] still apply). Such extrapolation is not possible when the parameters are kept constant at values calibrated to a subset of the historical record.

3. The use of model selection techniques such as split-sample testing and/or information-theoretic approaches allows an assessment of whether the additional model complexity associated with the description of parameter non-stationarity produces a significant improvement in the model's predictive ability. In contrast, it is not clear how to evaluate the trade-off between model fit, complexity and length of record when calibrating parameters to different historical periods.

4. Additional insights are provided on the nature of possible deficiencies in the model structure [de Vos et al., 2010]. As parameter non-stationarity can be symptomatic of poor representation of important hydrological processes, it can serve as a valuable diagnostic of the suitability of the existing model for extrapolation. The nature of suggested non-stationarity can help guide model improvement, especially when the non-stationarity can be attributed to a specific cause, such as a particular poorly represented process in the model or a major change in catchment conditions.

The paper is structured as follows. The key elements of the proposed strategy for diagnosing and interpreting hydrological non-stationarity are presented in Section 2, followed by a description of the case study catchment in Section 3. Section 4 provides a detailed investigation of data quality, including the analysis of possible systematic changes in the quality of rainfall, evapotranspiration or streamflow data. Section 5 describes a set of 22 candidate hydrological models with different combinations of non-stationarity parameters to be evaluated, and Section 6 describes the approach to parameter estimation. Section 7 details an AIC-based approach for model selection and diagnosis of hydrological non-stationarity. Results are presented in Section 8, followed by discussion in Section 9 and conclusions in Section 10.

2. Overview of the strategy for diagnosing and interpreting hydrological non-stationarity

Our strategy for developing hydrological models for predicting catchment runoff under changing hydroclimatic conditions follows the philosophical approach of 'multiple working hypotheses', described originally by Chamberlain [1890] and more recently in the hydrological context by Clark et al. [2011]. In this approach, a set of candidate models ('hypotheses') is constructed and evaluated, with each model providing an alternative representation of catchment behaviour, including any non-stationarities. The models are calibrated to observed data in an exploratory period, and an information-theoretic measure (the AIC) is used to evaluate the level of support from the data for each model. A selected subset of models is then tested on an independent confirmatory period that is climatologically different from the period used for parameter estimation, thus representing a differential split sample test [Klemes, 1986]. The four elements of the strategy are outlined next.

2.1 Detecting systematic errors in the calibration data

Biases and systematic changes in the measurement of hydrological data can significantly affect parameter estimation and can also lead to non-stationarity in hydrological model parameters. In situations where biases and/or changes in data quality cannot be excluded *a priori*, they must be

retained among the 'working hypotheses' to be evaluated *a posteriori* as part of model calibration and analysis. In this study, we use a set of standard diagnostics to assess the quality of the rainfall, potential evapotranspiration and runoff data (Section 4).

2.2 Modelling one or more parameters as functions of time-varying covariates

As we define hydrological model non-stationarity as the case where hydrological model parameters change in time, a practical strategy for detecting non-stationarity is to allow the model parameters to vary in time as functions of selected covariates and examine the resulting impact on model performance. In this study, covariates are selected to represent the major timescales of hydrological variability. For example, we use a sinusoidal function to represent seasonal changes in the catchment storage capacity. The covariates are discussed further in Section 5.1, and resemble some of the timescales of variability used in the 'unobserved components' within the data-based mechanistic modelling (DBM) philosophy [Young and Beven, 1994; Young, 1998]. In DBM, however, time-varying covariates are used as part of model identification and development using transfer functions, whereas in our case the purpose is largely as a diagnostic for structural errors in conceptual hydrological models.

2.3 Comparison of alternative model structures

One of the possible causes of non-stationarity in hydrological model parameters is poor process representation within the model. This step therefore aims to identify missing or poorly represented processes; this information can be used either to improve the hydrological model, or to better characterise its predictive uncertainty. In this paper we use various hydrological diagnostics, such as flow duration curves stratified by season and by the phase of the hydrograph (rising and falling limbs), to isolate possible weaknesses in the conceptual model GR4J when simulating runoff in Scott Creek catchment. Based on this assessment, we make two modifications to the standard GR4J model; these are discussed in Section 5.2.

Alternative approaches for model development and comparison include flexible model frameworks such as FUSE [Clark et al., 2008] and SUPERFLEX [Fenicia et al., 2011]. These frameworks can be used to analyse larger and more diverse sets of model structures. However, incorporating flexible model structures into the second step of the non-stationarity analysis strategy (Section 2.2) requires further work to support non–nested structures with distinctly different conceptualisations and parameterisations. For example, in non-nested models it may not be possible to apply non-stationary covariates to the same parameter, making it difficult to consistently compare the degree of parameter non-stationarity across all models under consideration.

2.4 Model selection and evaluation

The final step is to evaluate the empirical support for the model structures hypothesized and calibrated in Steps 2 and 3. Many model selection approaches can be used, including:

- cross-validation based methods [e.g., Schoups et al., 2008; Hastie et al., 2009], including split sample testing, in which one or more models are fitted using a portion of the historical record (usually referred to as the 'calibration' period) and tested on the remainder of the

record (usually referred to as the 'validation' or 'verification' period). This has been the preferred approach in the hydrological literature for estimating 'out of sample' model error [e.g. Hastie et al., 2009];

- information theory [e.g., Burnham and Anderson, 2010], which is receiving increased interest in the hydrological literature [e.g., Gupta et al., 2008; Weijs et al., 2010]. The information-theoretic framework aims to estimate the 'in-sample' prediction error from the likelihood (objective) function calculated during model calibration, while also attempting to account for the expected model 'optimism' arising from the assessment of model performance over the calibration period itself [Hastie et al., 2009]. The Akaike Information Criterion (AIC) [Akaike, 1974] and its small sample approximation (AICc) [Sugiura, 1978] are widely used model selection criteria derived using information theory.

- Bayesian approaches, such as the Bayesian Information Criterion (BIC) [Schwarz, 1978; Marshall et al., 2005; Martinez and Gupta, 2011], Kashyap's Information Criteria (KIC) [Kashyap, 1982; Martinez and Gupta, 2011] and Bayesian model averaging [Hoeting et al., 1999; Claeskens and Hjort, 2008].

There are ongoing debates in the hydrological and broader communities on the advantages, limitations and interpretations of different model selection criteria [e.g. Ye et al., 2008; Burnham and Anderson, 2010]. An increasing number of studies compare multiple model selection approaches, often with contradictory results that appear to depend on specific features of the data and models being investigated [Schoups et al., 2008; Ye et al., 2008; Burnham and Anderson, 2010; Dai et al., 2012; Engelhardt et al., 2013]. In this study we adopt the AIC for selecting between multiple model hypotheses because it is a simple yet widely used criterion that seeks to maintain parsimony while selecting the model with the greatest predictive ability [McQuarrie and Tsai, 2007; Burnham and Anderson, 2010]. The key properties of this criterion are given in Section 7.

Note that in this paper we use the term 'exploratory period' to refer to the period used for parameter estimation ('calibration'), model comparison and selection. Furthermore, the term 'confirmatory period' refers to the period used for independent model evaluation. The confirmatory period is commonly referred to as the 'validation' or 'verification' period in the hydrological literature, however the term 'confirmatory' is intended to emphasize that future model performance cannot be 'validated' or 'verified' from past performance alone [Oreskes et al., 1994].

3. Case study catchment

The four steps of the strategy for analysing non-stationarity of hydrological model parameters are illustrated using the Scott Creek catchment in South Australia. This catchment has an area of 29 km^2 and forms a part of the larger Onkaparinga catchment – Adelaide's primary surface water source (Figure 1). The median annual rainfall (P) in Scott Creek is 905 mm, and the median annual potential evapotranspiration (PET) is 1600 mm. The long-term average runoff is 123 mm, giving a runoff coefficient of 0.14.

The Scott Creek catchment is classified as semi-arid and has a winter-dominated rainfall regime. February is the driest month (monthly average of 20 mm), while July is the wettest (monthly average of 130 mm). In contrast, monthly PET varies from 50 mm in July to 250 mm in January. Therefore, in summer the catchment is water-limited (P<<PET), whereas in winter it is energy-limited (P>>PET). The combined effect of seasonality in P and PET is that, in an average year, the runoff is highly seasonal, with over 75 % occurring in the three-month period from July to September. The seasonality of the catchment suggests that different physical mechanisms may be governing the rainfall-runoff relationships in summer and winter.

In addition to seasonal variations, the runoff characteristics of the Scott Creek catchment also vary inter-annually. At the aggregated annual scale, the relationship between catchment-average rainfall and runoff is approximately linear (with a Pearson correlation R^2 of 0.80), and a 1 % change in annual rainfall yields an approximately 3 % change in runoff. This catchment sensitivity is within the typical range for semi-arid catchments in southeast Australia [Chiew, 2006]. The runoff coefficient, when calculated for each calendar year, varies from 0.06 in the driest year (2006) to 0.22 in the wettest year (1986).

The streamflow varies over four orders of magnitude, with approximately 21% of days over the exploratory and confirmatory periods having flows below 0.01 mm /day, and with only 22 days having flows above 10 mm /day. This corresponds to approximately 30% of the flow volume occurring in the top 1% of flow days, and 68% of flows occurring in the top 10% of flow days.

The 1985-1999 period is used for the exploratory analysis (parameter estimation and model selection), and the 2000-2009 period is used for the confirmatory analysis (model evaluation). Prior to both periods, a four-year spin-up period is used to reduce the impact of unknown initial conditions. The confirmatory period is much drier than the exploratory period, with 19 % less runoff on average, and therefore provides a stringent differential split sample test.

4. Identifying systematic errors in calibration data

The first element of the strategy is to identify systematic errors in the observed data. In this study, we examine the quality of observed streamflow, potential evapotranspiration and rainfall.

Streamflow estimates for Scott Creek were obtained from a rectangular stepped weir near the catchment outlet, which has operated continuously since 1969. Analysis of the differences between streamflow gaugings and streamflow estimates from the rating curve suggests a significant increase in rating curve errors during 1980-1984, with some evidence of systematic bias (Figure 2). Furthermore, the gauging station metadata indicates that a major rating curve change occurred in 1984. Hence, to avoid the impact of potentially biased streamflow data on the inference of non-stationarity, our analysis is based exclusively on post-1984 data. The drawback of selecting this time period is that it has a smaller number of rating curve measurements, so that all flows greater than 10 mm (1-in-6-month flow) are extrapolated.

Catchment-average PET was estimated using Morton's areal potential evapotranspiration (APET) method [Morton, 1983; McMahon et al., 2013], which is based on temperature, vapour pressure and

incoming solar radiation data from the Australian SILO 0.05° latitude/longitude gridded dataset [Jeffrey et al., 2001]. The time series of annual APET have a slight upward trend from 1985 to 2009. A similar trend is present in Morton's APET estimated at the high-quality Kent Town weather station (the nearest high-quality weather recording station), indicating that this trend is unlikely to be caused by measurement errors.

Three rainfall gauges are located within or very close to Scott Creek catchment. Continuous rainfall data for these gauges were obtained from the SILO patched point database, and these data are occasionally infilled using interpolated data when observed data is missing or suspect [Jeffrey et al., 2001]. Therefore, to detect potential systematic errors, a homogeneity analysis [Allen et al., 1998] was performed by comparing the rainfall time series at each gauge in Scott Creek catchment to time series from the rain gauge at Happy Valley, which is part of Australia's high quality gauge network [Lavery et al., 1992]. No statistically significant evidence of inhomogeneity was found. The catchment-average rainfall for Scott Creek was obtained by kriging the three gauges, and is dominated by a single gauge at Cherry Gardens (see Figure 1), which has a weight of 0.9.

Based on the analysis of streamflow, PET and rainfall data in Scott Creek catchment, we conclude that this data is of relatively high quality from 1985 onwards, and we therefore use only post-1984 data for model development and evaluation. A negative consequence of using stringent criteria for data selection is that potentially long portions of the historical record might be discarded from the analysis. For the present case study, the record retained is sufficient for the intended analysis, and reduces the potential contribution of poor data quality to model non-stationarity.

An alternative way of addressing data quality is to develop more comprehensive data error models. For example, rainfall error models could be based on detailed geostatistical analysis [Renard et al., 2011]. However this requires considerable additional information and was not pursued in this study.

5. Candidate hydrological models

All hydrological models considered in this work are derived from the lumped conceptual rainfall-runoff model GR4J [Renard et al., 2011]. The published version of GR4J has four calibration parameters, namely the production store capacity (θ_1, units of mm), the groundwater exchange coefficient (θ_2, units of mm), the one day-ahead maximum capacity of the routing store (θ_3, units of mm), and the time base of the unit hydrograph (θ_4, units of days).

GR4J was developed to provide, on average, good performance across a wide range of catchment conditions [Renard et al., 2011]. This makes GR4J particularly suitable as a starting point for model modifications and refinements, including the versions constructed in this work as part of detecting and quantifying hydrological non-stationarity. The GR4J modifications are described next.

5.1 Simulating hydrological model non-stationarity

Parameter θ_1 is allowed to vary in time to represent several potential time scales of non-stationarity. We focus on θ_1 because it represents the primary storage of water in the catchment.

Previous studies [Kuczera et al., 2006; Renard et al., 2011] have indeed suggested that θ_1 is the most sensitive GR4J parameter, with Renard *et al*. [2011] showing through a sensitivity analysis that stochastic variations of θ_1 have the largest impact on model predictions. By treating θ_1 as a function of multiple covariates representing seasonal, annual and longer-term variability, we attempt to characterize the major potential time scales of non-stationarity in the catchment, as follows:

(1) Seasonal-scale variability in catchment characteristics are represented by conditioning θ_1 on a sine function with a yearly period, parameterized by its amplitude and phase. In the Scott Creek catchment, a major source of seasonality might be the switch from water limitations in summer to energy limitations in winter (Section 3).

(2) Annual-scale variability due to hydrometeorological changes is represented by conditioning θ_1 on the 365-day antecedent daily rainfall and potential evapotranspiration. This conditioning aims to account for non-stationarity in the predictive errors, such as when a hydrological model systematically overestimates flows during dry years and underestimates flows during wet years [e.g. Coron et al., 2012; Pathiraja et al., 2012].

(3) Long-term changes in catchment response are represented using a linear trend in θ_1.

The full non-stationary model for parameter θ_1 is:

$$\theta_1(t \mid \boldsymbol{\lambda}) = \lambda_1 + \underbrace{\lambda_2 t}_{\text{linear trend}} + \underbrace{\lambda_3 \sin\left(2\pi \frac{t+\lambda_4}{365}\right)}_{\text{seasonal variability}} + \underbrace{\lambda_5 P_{365} + \lambda_6 PET_{365}}_{\text{annual variability}} \quad (1)$$

where *t* is the number of days since the start of simulation and $\lambda_1, \ldots, \lambda_6$ are six "non-stationarity" parameters. Parameter λ_1 is a constant term, λ_2 represents the linear trend, $\{\lambda_3, \lambda_4\}$ represent the amplitude and phase of the sine term, and $\{\lambda_5, \lambda_6\}$ represent the influence of previous 365-day rainfall (P_{365}) and potential evapotranspiration (PET_{365}). Note that parameters λ_1 and λ_4 depend on the starting date of the simulation (here selected as 1 January in both the exploratory and confirmation periods).

As discussed in Section 9.3, there may be physically interpretable reasons for temporal changes in catchment storage capacity. For example, an increase in on-farm dams [e.g. Coron et al., 2012; Pathiraja et al., 2012] in Scott Creek catchment may lead to an increase in the total available storage volume, and thus to a larger value of the storage parameter θ_1. Other forms of non-stationarity might be less physically interpretable. For example, in Scott Creek the total volume of available storage in the soil matrix is unlikely to change regularly each season, so that the presence of a sinusoidal pattern in θ_1 does not immediately indicate a seasonal change in actual catchment storage capacity. Therefore we view the primary purpose of the covariates described in this section as diagnostic: by representing the main time scales of likely variation in model parameters, it

becomes possible to identify deficiencies in the model structure, which in turn can be used to identify areas for model improvement.

5.2 Modifying the structure of GR4J

Non-stationarity in hydrological model parameters can indicate that a hydrological process is either absent or incorrectly represented in the hydrological model. We test this proposition by making several modifications to GR4J, based on the results of model diagnostics (discussed further in Section 7).

5.2.1 Representation of recession dynamics

Inspection of hydrographs predicted using the standard GR4J model indicated systematic deficiencies in the representation of the falling limb (Section 8). To improve the representation of recession behaviour, an additional parameter θ_5 is introduced to provide greater flexibility in the GR4J equation that controls the partitioning of net rainfall between the production and routing stores:

$$P_s = \frac{\theta_1\left(1-\left(\frac{S}{\theta_1}\right)^{\theta_5}\right)\tanh\left(\frac{P_n}{\theta_1}\right)}{1+\frac{S}{\theta_1}\tanh\left(\frac{P_n}{\theta_1}\right)} \qquad (2)$$

where P_s is the portion of net rainfall P_n that enters the production store and S is the water content in the production store [compare with Equation 3 in Perrin et al., 2003].

5.2.2 Representation of evapotranspiration dynamics

The low runoff coefficient in the Scott Creek catchment and the general aridity of its regional environment indicate a large contribution of evapotranspiration to the overall water balance. Analysis of the GR4J simulations found that almost 30 % of the rainfall was being converted to actual evapotranspiration on rainy days, before the rainfall entered the production store. In the original version of GR4J, actual evapotranspiration (AET) is determined from two different model processes. The first process occurs on all rainy days when P>PET; here the net rainfall is calculated $P_n = P - PET$, and AET occurs at the potential rate. The second process occurs on days when P<PET, and the AET is calculated as a function of the water level in the production store. In the modified GR4J, an alternative formulation is considered, in which $P_n = P$ (i.e. removing the first process), and AET is only a function of the volume of water in the production store. This representation is common in hydrological models, including HBV [Bergstrom, 1995], TOPMODEL [Beven et al., 1995] and others.

5.3 Model structure groupings

To assist in the systematic comparison of predictive performance across a large number of candidate models, we define the following three model structure groupings, as described in Table 1:

1. A set of eight model structures, labelled $g_{1.1},\ldots,g_{1.8}$, are used to cover all possible combinations of the three non-stationary components developed in Section 5.1. Note that the individual terms within each distinct non-stationarity component in Equation (1) are always considered jointly (e.g., we do not split the annual variability representation into individual P and PET terms).

2. A set of four model structures, labelled $g_{2.1},\ldots,g_{2.4}$, are used to examine the impact of GR4J structural modifications presented in Section 5.2 for improving the representation of recession and evapotranspiration dynamics. Note that the original GR4J model ($g_{1.1}$) is included in this grouping as model $g_{2.1}$, and is used as a 'reference' against which this set of model modifications are compared.

3. A set of 12 model structures, labelled $g_{3.1},\ldots,g_{3.12}$, are given by different combinations of non-stationarity models and GR4J structural modifications. In this grouping, the 'reference' model ($g_{3.1}$) is selected to be model $g_{2.2}$, as model $g_{2.2}$ was found to be the best model in the model grouping $g_{2.x}$ (Section 8.2). Note that the model grouping $g_{3.x}$ does not include all possible combinations of covariates for non-stationary θ_1 and other model modifications, as this would have led to an excessively large number of candidate models. Rather, important groups of parameters were identified based on the analysis of the first two model groupings ($g_{1.x}$ and $g_{2.x}$); this is discussed further in Section 8.

6. Parameter estimation

This section describes the method of maximum likelihood used in this study to estimate the model parameters. This method requires the construction of a likelihood function, followed by parameter optimization through likelihood maximisation.

6.1 Specification of the likelihood function

The likelihood function $\mathrm{L}()$ is defined as the joint probability of the observed streamflow given the observed forcings and the parameters $\boldsymbol{\theta}$ of a predictive model, i.e., $\mathrm{L}(\boldsymbol{\theta}) = p(\tilde{y}_1,\tilde{y}_2,\ldots,\tilde{y}_n \mid \tilde{x}_1,\tilde{x}_2,\ldots,\tilde{x}_n,\boldsymbol{\theta}) = p(\tilde{\mathbf{y}} \mid \tilde{\mathbf{x}},\boldsymbol{\theta})$. The predictive model is constructed by combining a hydrological model with a description of predictive uncertainty, as detailed next.

Consider a deterministic hydrological model *h*(), such as GR4J. At time step *t*, the model predictions of streamflow y_t are:

$$y_t = h(\tilde{\mathbf{x}}_{1:t};\boldsymbol{\theta}_h) \tag{3}$$

where $\tilde{\mathbf{x}}_{1:t}$ is the time series ($t = 1\ldots,n$) of observed hydrological inputs (here, daily rainfall and PET) and $\boldsymbol{\theta}_h$ is the vector of hydrological model parameters (here, $\boldsymbol{\theta}_h = \{\theta_1,\ldots,\theta_5\}$).

Next, consider an additive residual error model, defined as $\varepsilon_t = \tilde{y}_t - y_t$, where $\tilde{y}_t$ is the observed streamflow at time step *t*. We assume that the residuals are independent in time and follow a

Gaussian distribution with zero mean and standard deviation σ_ε, i.e., $\varepsilon \sim N(0, \sigma_\varepsilon)$. As hydrological model residuals are typically heteroscedastic [Sorooshian, 1981; Schoups and Vrugt, 2010], we allow σ_ε to vary in time as a linear function of predicted streamflow, i.e., $\sigma_{\varepsilon(t)} = a_\varepsilon + b_\varepsilon y_t$. The error model parameters $\boldsymbol{\theta}_\varepsilon = \{a_\varepsilon, b_\varepsilon\}$ are unknown and are therefore estimated as part of the inference.

Under the residual error assumptions listed above, the following log likelihood is obtained:

$$\log \mathrm{L}(\boldsymbol{\theta}) = \log \mathrm{L}(\boldsymbol{\theta}_h, \boldsymbol{\theta}_\varepsilon) = \sum_{t=1}^{n} \log f\left[\varepsilon_t(\boldsymbol{\theta}_h); 0, \sigma_{\varepsilon(t)}(\boldsymbol{\theta}_h, \boldsymbol{\theta}_\varepsilon)\right] \tag{4}$$

where $f(z; \mu, \sigma)$ is the Gaussian probability density function with mean μ and standard deviation σ, evaluated at point *z*. Note that the residuals depend solely on hydrological parameters, whereas the standard deviations of the residuals depend on both hydrological and error model parameters.

6.2 Extension to models with non-stationary parameters

As detailed in Section 2.2, hydrological non-stationarity can be investigated by allowing one or more hydrological model parameters θ_h to vary in time as functions of selected covariates. For example, θ_1 is modelled as a function of covariates as described in Equation (1). This can be accommodated within the likelihood function by no longer calibrating θ_1 and instead calibrating $\lambda_1, \ldots, \lambda_6$.

The remainder of the paper uses the short-hand notation '*g*' to represent the combined hydrological and error models,

$$\begin{aligned} g &= g(\tilde{\mathbf{x}}_{1:t}; \boldsymbol{\theta}) \\ &= h(\tilde{\mathbf{x}}_{1:t}; \boldsymbol{\theta}_h, \boldsymbol{\lambda}) + \varepsilon \end{aligned} \tag{5}$$

As discussed in Section 5.3, we compare the performance of 22 alternative models listed in Table 1. The individual models are identified by an index on *g*. Note that each of the models have different numbers of calibrated parameters, e.g., model $g_{1.2}$ can be written as $g_{1.2} = h(\tilde{\mathbf{x}}_{1:t}; \theta_1, \theta_2, \theta_3, \theta_4, \lambda_1, \lambda_2) + \varepsilon$.

6.3 Mitigating deficiencies in the assumed likelihood function

The assumption of independent residual errors in Equation (4) is poor in most hydrological applications [Sorooshian and Dracup, 1980; Evin et al., 2013]. Moreover, near-zero flows exert a strong influence on the inference when using a likelihood that accounts for heteroscedasticity. Therefore, two changes are made to the likelihood function, as detailed below.

6.3.1 Handling low (close to zero) flows in the likelihood function

The Scott Creek catchment is highly seasonal, typically with very little runoff during summer. The handling of low flows in the likelihood function is the subject of ongoing research [e.g. Smith et al.,

2010]. To avoid this issue negatively impacting on the analysis, observed daily flows below a threshold of 0.09 mm are censored from the likelihood function. The resulting streamflow data set is referred to as $\tilde{\mathbf{y}}^{(>0.09)}$. This censoring threshold corresponds to the streamflow value for which, based on the rating curve analysis, there is a 95 % probability that the streamflow predicted by the rating curve is greater than zero. Over the exploratory period, 55 % of days have flows below 0.09 mm, yet these censored days contribute less than 5 % of the total catchment flow volume.

Residual error diagnostics were checked in all cases, and are presented here for the simplest model ($g_{1.1}$) and for one of the most complex models ($g_{3.11}$). Density plots of the standardized residuals in Figure 3 show a good match between empirical and theoretical density functions. Furthermore, the reliability of the total predictive uncertainty was assessed using a predictive quantile-quantile (PQQ) plot [Thyer et al., 2009] (not shown). The observed *p*-values were very close to the 1:1 line suggesting that the error model provides a reasonably reliable approximation of the probability distribution of the residuals.

6.3.2 Handling autocorrelation in the residuals

Autocorrelation of residual errors can significantly influence model inference and selection, yet is omitted in Equation (4). In this case study, statistically significant error autocorrelation was found for all models. For the most complex model ($g_{3.11}$), the lag-1 autocorrelation coefficient for the residuals (after the low flow threshold is applied) is 0.32, which, although relatively low in the context of rainfall-runoff applications, is statistically significant at the 5 % level. To reduce the impact of ignoring autocorrelation in the likelihood function, all hydrological models were re-calibrated to a "thinned" streamflow set comprising every k^{th} day of record. We trialed several values of *k*, and identified the minimal value of *k* for which the lag-1 autocorrelation coefficient was no longer significant at the 5 % level. For almost all the models, this led to a six-day sampling interval (*k*=6).

The thinning is incorporated into the likelihood function in Equation (4) by only including the model residuals from every sixth day of record, while still censoring days with observed flows below the 0.09 mm threshold, i.e., $\varepsilon_t : t \in \{1,7,13,\ldots\} \cap \tilde{y}_t > 0.09$. The corresponding streamflow set is referred to as $\tilde{\mathbf{y}}^{(>0.09)}_{t:=1+6j}$. The sensitivity of the results to the particular choice of thinned period is investigated by calibrating (separately) to six non-overlapping sets of thinned residuals, defined as $\varepsilon_t : t \in \{2,8,14,\ldots\} \cap \tilde{y}_t > 0.09$, $\varepsilon_t : t \in \{3,9,15,\ldots\} \cap \tilde{y}_t > 0.09$, and so on. The corresponding streamflow sets are referred to as $\tilde{\mathbf{y}}^{(>0.09)}_{t:=2+6j}$, $\tilde{\mathbf{y}}^{(>0.09)}_{t:=3+6j}$, and so on.

More complex residual error models, such as those including specialized treatment of low flows [Smith et al., 2010] and direct treatment of error autocorrelation [Evin et al., 2013], are clearly of interest to improve the specification of the likelihood function. However, practical difficulties have been encountered when jointly inferring error autocorrelation and heteroscedasticity, including strong interactions of the error autocorrelation parameter with the GR4J mass balance parameter θ_2 [Evin et al., 2013]. Moreover, combined treatment of error autocorrelation and low flows requires separate theoretical development. Hence, censoring of low flows and calibrating to thinned streamflow sets was used as a pragmatic approach to reduce the violations of the likelihood assumptions.

6.4 Parameter optimization

The parameter values that maximize the likelihood function in Equation (4) were estimated using a quasi-Newton optimization method. Optimization was repeated with 100 random starting points to reduce the probability of being trapped in local optima.

7. Model evaluation and selection

We use an information-theoretic approach in combination with multiple hydrologically-oriented diagnostics to evaluate the performance of the hydrological models. This section details the specific metrics used.

7.1 The Akaike Information Criterion

Information-theoretic techniques use the Kullback-Leibler information to compare an approximate probability model $p(\tilde{\mathbf{y}} \mid \boldsymbol{\theta})$ (note the conditioning on $\tilde{\mathbf{x}}$ used previously has been removed for notational convenience) against the (unknown) 'true' probability density function $p_{true}(\tilde{\mathbf{y}})$ describing the system of interest [Burnham and Anderson, 2010]:

$$\begin{aligned} I_{KL}\left(p_{true}(\tilde{\mathbf{y}}) \,\square\, p(\tilde{\mathbf{y}} \mid \boldsymbol{\theta})\right) &= \int \log \frac{p_{true}(\tilde{\mathbf{y}})}{p(\tilde{\mathbf{y}} \mid \boldsymbol{\theta})}\, p_{true}(\tilde{\mathbf{y}})\, \mathrm{d}\,\tilde{\mathbf{y}} \\ &= \int \log\left(p_{true}(\tilde{\mathbf{y}})\right) p_{true}(\tilde{\mathbf{y}})\, \mathrm{d}\,\tilde{\mathbf{y}} - \int \log\left(p(\tilde{\mathbf{y}} \mid \boldsymbol{\theta})\right) p_{true}(\tilde{\mathbf{y}})\, \mathrm{d}\,\tilde{\mathbf{y}} \end{aligned} \tag{6}$$

The Kullback-Leibler information $I_{KL}\left(p_{true}(\tilde{\mathbf{y}}) \,\square\, p(\tilde{\mathbf{y}} \mid \boldsymbol{\theta})\right)$, often referred to as the "Kullback-Leibler divergence of $p(\tilde{\mathbf{y}} \mid \boldsymbol{\theta})$ from $p_{true}(\tilde{\mathbf{y}})$", can be interpreted as the information lost when an approximate likelihood $p(\tilde{\mathbf{y}} \mid \boldsymbol{\theta})$ is used to represent the "true" likelihood $p_{true}(\tilde{\mathbf{y}})$. Since $p_{true}(\tilde{\mathbf{y}})$ represents the 'truth', it does not vary as a function of the parameters, whereas $p(\tilde{\mathbf{y}} \mid \boldsymbol{\theta})$ varies over the parameter space $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. We stress that $p(\tilde{\mathbf{y}} \mid \boldsymbol{\theta})$ refers to the complete probability model of the data, which here is constructed by combining a deterministic component (i.e., the hydrological model) and a stochastic component (i.e., the error model).

In real environmental systems $p_{true}(\tilde{\mathbf{y}})$ is unknown and therefore the Kullback-Leibler information cannot be calculated. However, since the term $\int \log\left(p_{true}(\tilde{\mathbf{y}})\right) p_{true}(\tilde{\mathbf{y}})\, \mathrm{d}\,\tilde{\mathbf{y}}$ in Equation (6) is a constant that depends only on the (unknown) 'truth', it is possible to calculate the difference in Kullback-Leibler information between any two models $p_1(\tilde{\mathbf{y}} \mid \boldsymbol{\theta})$ and $p_2(\tilde{\mathbf{y}} \mid \boldsymbol{\theta})$. This difference can be treated as a measure of relative empirical support in favour of one of the models.

Under a set of assumptions discussed below, choosing the model that maximises the AIC yields the smallest Kullback-Leibler divergence from the true model p_{true} [Akaike, 1974]. The derivation of the AIC, A, from the Kullback-Leibler information is described in Burnham and Anderson [2010], and requires the use of the maximum likelihood estimate of the parameter vector, $\hat{\boldsymbol{\theta}}$:

$$\mathrm{A} = \log \mathrm{L}(\hat{\boldsymbol{\theta}}) - K \tag{7}$$

The term *K* denotes the number of calibrated parameters in the model, and is often described as a 'complexity penalty' that accounts for the fact that the model parameters $\hat{\boldsymbol{\theta}}$ are being calibrated to the (finite) observed data.

The AIC differences, denoted by $\Delta\mathrm{A}$, and can be interpreted as the loss of information when model *i* is used instead of the AIC-best model in a set of models being compared:

$$\Delta\mathrm{A}_i = \mathrm{A}_i - \mathrm{A}_{\min} \tag{8}$$

This metric can be evaluated for each model *i* = 1,…,*M* in the set of *M* models being compared, with $\mathrm{A}_{\min}$ being the lowest (best) AIC value produced by the models in the set.

Akaike 'weights', $w^{(\mathrm{A})}$, defined for model *i* from the set of *M* models as:

$$w_{i|M}^{(\mathrm{A})} = \frac{\exp\left(-\frac{1}{2}\Delta\mathrm{A}_i\right)}{\sum_{j=1}^{M}\exp\left(-\frac{1}{2}\Delta\mathrm{A}_j\right)} \tag{9}$$

can then be interpreted as the "weight of evidence in favour of model *i*", i.e., the probability that, given the set of *M* models, model *i* will obtain the highest likelihood value when predicting new data arising from the same system.

The Akaike weights facilitate a probabilistic interpretation of AIC differences. Values of $\Delta\mathrm{A}_i$ less than 2 are usually interpreted as indicating "substantial" support for model *i*, whereas values greater than 10 indicate that there is "virtually no support" for that model [Burnham and Anderson, 2010].

Two major assumptions underlying the AIC should be considered. Firstly, the term *K* in Equation (7) is derived under the assumption that the sample size is 'large'. A 'large' sample is usually defined when $n/K > 40$ [Burnham and Anderson, 2010], and in this study the criterion is met in all cases. Secondly, the AIC is derived under the assumption that the likelihood function provides a 'good' approximation to the actual system. This assumption is questionable in this study, in particular because the error model used to derive the likelihood in Equation (4) assumes the residuals are independent (Section 6.3.2). Since neglecting the serial dependence of the errors results in an over-estimation of the information content of the data and may affect the AIC assumptions, this paper does not use the full interpretation of the AIC weights described in the preceding paragraph. However, despite these limitations, we proceed on the assumption that AIC-based rankings and the relative magnitudes of AIC difference and the AIC weights can still help guide model selection (see Section 8).

7.2 Hydrologically-oriented model diagnostics

Since the AIC comparison only considers statistical aspects of model performance, additional metrics with hydrological interpretation are used for a more thorough model comparison:

- The Nash-Sutcliffe coefficient of efficiency (NSE), which is widely used in the hydrological literature and therefore enables direct comparison with other studies;
- The differences between modeled and observed annual total flow volume, which is a measure of the catchment water balance;
- Daily-scale flow duration curves, which allow comparing the probability distributions of observed and modeled flows and can provide a visual indication of potential biases (e.g. compensating behavior with overestimation of low flows and underestimation of high flows). We consider stratified flow duration curves for: (i) all flows throughout the year; (ii) flows in individual seasons; and (iii) flows in the rising and falling limb of the hydrographs.

This list is not intended as a comprehensive set of diagnostics for hydrological model evaluation. In addition to general metrics, the diagnostics should reflect the modelling goals. We refer the reader to Martinez and Gupta [2011] for further details.

8 Results

This section examines the performance of the 22 hydrological models (Table 1) over the exploratory and confirmatory periods (Section 3). For convenience, the comparison makes use of the model structure groupings described in Section 5.3. The impact of thinning the streamflow set used in the calibration (Section 6.3.2) is also investigated.

Figure 4 shows the AIC differences, the residual error parameters (a_ε and b_ε), the NSE and the magnitude of the groundwater flux calculated over the exploratory period using streamflow set $\tilde{\mathbf{y}}^{(>0.09)}$. The AIC differences when estimating parameters using streamflow set $\tilde{\mathbf{y}}^{(>0.09)}_{t:=1+6j}$ and the AIC values for the AIC-best model in each model structure grouping are also shown.

8.1 Model grouping $g_{1.x}$: Non-stationary GR4J

The results for the model grouping $g_{1.x}$ are presented as red bars in Figure 4. When calibrating to streamflow set $\tilde{\mathbf{y}}^{(>0.09)}$, the best AIC value is achieved by model $g_{1.8}$, which is the most complex model in the comparison and includes all forms of non-stationarity. In contrast, when calibrating to streamflow set $\tilde{\mathbf{y}}^{(>0.09)}_{t:=1+6j}$, model $g_{1.6}$ is the AIC-best model, with model $g_{1.8}$, very closely behind. The only difference between these two models is that $g_{1,8}$ has the linear trend in parameter θ_1.

8.1.1 Interpretation of the AIC weights

The AIC-best model estimated using the streamflow set $\tilde{\mathbf{y}}^{(>0.09)}$ has an AIC weight of 0.994, while the second best model has a weight of 0.006. In contrast, the AIC-best and AIC-second best models estimated using streamflow set $\tilde{\mathbf{y}}^{(>0.09)}_{t:=1+6j}$ have near-equal weights of 50.1 and 49.9, respectively, with

almost no weight for the remaining models. This would indicate that the remaining models have almost no probability of being selected as AIC-best under an independent confirmatory period, but it is unlikely that this interpretation holds in this case. This is because the assumption of independence in the model residuals is unlikely to hold exactly, even when sampling every sixth day. Other deficiencies in the likelihood function, including the GR4J hydrological model, the Gaussian distribution and linear heteroscedasticity of the residual errors, may also affect the 'good model' assumption underlying the AIC and reduce its interpretability. Despite these limitations, the relative magnitude of AIC differences in Figure 4 is instructive, and suggests which model modifications are responsible for the greatest improvements in model performance. For example, comparison of models $g_{1.2}$, $g_{1.3}$ and $g_{1.4}$ shows that the sinusoid representation of the seasonal-scale non-stationarity in θ_1 delivers by far the greatest improvement in predictive ability.

8.1.2 Increasing trend in parameter θ_1

Figure 5 shows the time variation of parameter θ_1 (i.e., the catchment storage capacity) and the actual storage in the production store for the two AIC-best models, $g_{1.6}$ and $g_{1.8}$, over the exploratory period. The sinusoidal variation is prominent for both models. There is also an apparent trend, with higher values of the production store observed in the second half of the record. The magnitude of this trend is similar regardless of whether a linear trend is included ($g_{1.8}$) or not included ($g_{1.6}$) as one of the covariates. It is likely that covariation exists between parameters λ_2 (representing the linear trend) and λ_5 (representing the previous 365-day PET) as a trend was found in PET (Section 3), and this could explain the similarity in performance between the two models. In both models, the increase in θ_1 over the exploratory period means that the responsiveness of the catchment to rainfall is decreasing through time (as a larger storage capacity provides a stronger damping of the effects of rainfall variability on the streamflow).

The actual water level in the production store is highly seasonal, with the store reaching a maximum value in late winter / early spring, and a minimum value in summer. This is not surprising given the seasonal nature of rainfall and PET in this catchment. More interesting is the timing of the sinusoid function for θ_1, with a maximum value occurring at the beginning of the year and a minimum value occurring in the middle of the year. As the production store affects the catchment responsiveness and the partitioning of rainfall between actual evapotranspiration and runoff/groundwater recharge, this result suggests that, in summer, the model without a sinusoidal term in θ_1 is overestimating the runoff responses and/or underestimating the actual evapotranspiration flux. The opposite effect is present in the winter predictions.

8.1.3 Other measures of model performance

The residual error model parameters a_ε and b_ε can serve as additional measures of hydrological model performance and are shown in Figure 4. These parameters need to be interpreted jointly, as a_ε describes the standard deviation of the residual error model at low flows, while b_ε describes the rate of increase in the standard deviation of the residual error model with predicted flow. Figure 4

shows that, as we consider models with lower AIC, a_ε decreases faster than b_ε. In fact, b_ε for the AIC-best model (b_ε = 0.40) is only slightly larger than b_ε for the AIC-worst model (b_ε = 0.36). This indicates that the GR4J modifications provide the greatest improvements when predicting low flows. The value of θ_1 for model $g_{1.1}$ (original GR4J model, where θ_1 is constant in time) is approximately 400 mm, which is closer to the winter minimum value of θ_1 θ_1 when the parameter is allowed to vary sinusoidally. This suggests that the non-stationarity model in Equation (1) provides the largest improvements during periods of low flow, particularly in summer and autumn. This is apparent when examining the autumn flow duration curves for models $g_{1.1}$, $g_{1.2}$, $g_{1.3}$, and $g_{1.4}$ in Figure 6: a significant improvement is provided by model $g_{1.3}$, in which θ_1 varies sinusoidally over the year, whereas the improvements are much more limited for the other models.

In contrast to the AIC-based rankings, the NSE yields a very different ranking of models, with model $g_{1.1}$ selected as the 'best' model, both with streamflow set $\tilde{\mathbf{y}}^{(>0.09)}$ and with streamflow set $\tilde{\mathbf{y}}^{(>0.09)}_{t:=1+6j}$. The NSE ranking is more consistent with the annually-aggregated flow error metric: the models with the lowest error in total annual flow also have the highest NSE values. This is probably due to the highly-skewed nature of flows in the catchment, with the majority of flow volume occurring in a relatively small number of wet days, and with the NSE reflecting the performance of the model in simulating those high-flow days. In contrast, the AIC is informed by the heteroscedastic likelihood model, which allows for a greater contribution from low flows whose total flow volume is small.

8.2 Model grouping $g_{2.x}$: Improved the process representation

The models in grouping $g_{2.x}$ represent modifications to the recession and ET equations in the GR4J model (except for $g_{2.1}$ which represents the original GR4J). Figure 4 shows that these modifications yield substantial improvements to model performance. Regardless of whether the model was calibrated using streamflow set $\tilde{\mathbf{y}}^{(>0.09)}$ or $\tilde{\mathbf{y}}^{(>0.09)}_{t:=1+6j}$, model $g_{2.2}$ was selected as the AIC-best model, followed by model $g_{2.4}$. Both models contain the additional parameter θ_5, and model $g_{2.4}$ also includes the modified representation of actual evapotranspiration.

In contrast to the model selection results based on the AIC or the inferred parameters of the residual error model, model $g_{2.1}$ was best in reproducing annual average flows. The most notable difference was the groundwater export volume, with model $g_{2.4}$ having either 2.5 or 1.9 times the groundwater flux compared to model $g_{2.1}$, depending on whether streamflow set $\tilde{\mathbf{y}}^{(>0.09)}$ or $\tilde{\mathbf{y}}^{(>0.09)}_{t:=1+6j}$ were used, respectively. For models $g_{2.3}$ and $g_{2.4}$ the total groundwater export is of a similar magnitude to the streamflow, thereby representing a major component of the catchment water balance.

Figure 7 shows the flow duration curves for simulated runoff from models $g_{1.1}$ and $g_{2.2}$. The model predictions are adequate for flows greater than the 30 % exceedance probability. However, model $g_{2.2}$ clearly outperforms $g_{1.1}$ for lower flows, supporting the earlier conclusion that the largest improvements occur for low flows. To further investigate the models' predictive performance, flow duration curves were plotted separately for the rising and falling limbs of the hydrographs, as shown inFigure 7. The most significant improvements occur in the falling limb. This is not surprising, as θ_5

was specifically introduced to improve the shape of hydrograph recessions (by modifying the partitioning of net rainfall into the production and routing store).

8.3 Model grouping $g_{3.x}$: Combining the non-stationary GR4J with improved process representation

The models in grouping $g_{3.x}$ combine the non-stationarity characterization of parameter θ_1 with the recession and ET modifications to the standard GR4J model. As noted in Section 5.3, the AIC-best model from grouping $g_{2.x}$, i.e., model $g_{2.2}$, is included in model grouping $g_{3.x}$ as model $g_{3.1}$.

The AIC-based model rankings are almost identical whether streamflow set $\tilde{\mathbf{y}}^{(>0.09)}$ or $\tilde{\mathbf{y}}^{(>0.09)}_{t:=1+6j}$ is used. The AIC-best and second-best models ($g_{3.11}$ and $g_{3.12}$, respectively) are equivalent except for the evapotranspiration term. The third-best model is $g_{3.4}$, which does not use the previous 365 day rainfall and PET as covariates.

Models with the sinusoidal term ($g_{3.3}$, $g_{3.4}$, $g_{3.6}$, $g_{3.7}$, $g_{3.9}$, $g_{3.10}$, $g_{3.11}$ and $g_{3.12}$) generally rank much higher than the models without this term: all six top-ranked models include this term. This suggests that the model modifications described in Section 8.2 were not able to eliminate the sinusoidal variation in θ_1. Models with parameter θ_5 perform better than models without this parameter, which is consistent with the results in Section 8.2. In contrast, Figure 4 shows that the inclusion of the linear trend in the non-stationarity model of θ_1 has a much smaller effect on the model rankings (i.e., compare models $g_{3.3}$ versus $g_{3.4}$, $g_{3.6}$ versus $g_{3.7}$, and $g_{3.9}$ versus $g_{3.10}$).

Similar to the case for model groupings $g_{1.x}$ and $g_{2.x}$, there is no close relationship between the AIC value of models within the groupings $g_{3.x}$ and their errors in total annual flow volume. This implies that the AIC and likelihood values are not good predictors of annual flow error over the exploratory period. This is not surprising, as the likelihood used in this study is based on a heteroscedastic error model that provides a balanced fit to low and high flows. Consequently, the likelihood function is not overly sensitive to errors in the total flow volume.

8.4 Sensitivity of AIC-based model rankings to the choice of thinned dataset

This section reports the sensitivity of the AIC-based model selection to the choice of thinned data set (see Section 6.3.2). Figure 9 shows the model ranks for each of the three sets of models (i.e. $g_{1.x}$, $g_{2.x}$ and $g_{3.x}$) using all six distinct thinned datasets. The colours are used to distinguish between three groupings of model structure that were found to perform similarly (Sections 8.1-8.3): (i) models with a sinusoid parameterization of θ_1, (ii) models with the additional parameter θ_5, and (iii) models with both a sinusoid parameterization of θ_1 and the recession parameter θ_5.

The findings show reasonable consistency in the rankings between models, particularly when accounting for the major model structural groupings. In set $g_{1.x}$, models with the sinusoid function consistently outperform those without this function. Model $g_{1.6}$ is the AIC-best for four of the thinned datasets, whereas $g_{1.5}$ and $g_{1.8}$ are each the AIC-best for one thinned dataset. Similarly, in set $g_{2.x}$, models $g_{2.2}$ and $g_{2.4}$ consistently outperform the remaining models, except for the sixth thinned dataset, in which model $g_{2.3}$ is ranked AIC-second best. Finally, in set $g_{3.x}$, the six models with both

the sinusoid function and the recession parameter θ_5 are consistently ranked amongst the top four models, although there is some variation in their individual rankings. Thus, the conclusions in Section 8.3 regarding the covariates with the most dominant influence on model performance appear to be reasonably robust with respect to the choice of thinned dataset used in the exploratory period.

8.5 Model evaluation over an independent confirmatory period

This section reports the performance of the models over the confirmatory period. The focus of this evaluation is to establish whether the AIC-best models over the exploratory period also perform well over the independent confirmatory period (where, as discussed in Section 3, the annual flows are on average 19 % lower than in the exploratory period).

Figure 10 shows the observed and simulated hydrographs for representative half-year sub-periods of the exploratory period (upper panel) and confirmatory period (lower panel). The confidence intervals are calculated using the estimated residual model parameters a_ε and b_ε. The cooler half-year (May-November) is shown as the majority of annual flow occurs during this period. The figure compares the predictions of the simplest model ($g_{1.1}$) and the AIC-best model ($g_{3.11}$). The models' ability to capture the observed hydrographs is difficult to determine by visual inspection alone, with both models underestimating some days and overestimating other days. Flow duration curves (Figure 6-Figure 8) are arguably better diagnostics for assessing the predictive performance at individual streamflow quantiles. However, it can be seen from Figure 10 that, in the confirmatory period, the simplest model ($g_{1.1}$) significantly overpredicts the observed flows for the majority of flow events, while the AIC-best model ($g_{3.11}$) matches the observations much better. Another noticeable feature is the narrower confidence interval for model $g_{3.11}$, particularly for low flows. This highlights that the non-stationary model yields a significant improvement in predictive precision, while maintaining a good description of residual errors (Figure 3).

Figure 11 presents the performance metrics, namely likelihood, NSE and the annual average flow volume error, for all models. Note that the AIC is not included in this comparison because the AIC is based on the maximum-likelihood parameter values estimated over the exploratory period, and it cannot be assumed that those parameters will also be the maximum likelihood estimates over the confirmatory period. Based on likelihood values, model $g_{1.1}$ is the worst performing model in the confirmatory period, and this also has the highest flow error of 18 %. In contrast the best model in the confirmatory period is model $g_{3.11}$, , and this model underestimates the average flow rate by only 2.6 % – a significant improvement on the original GR4J model $g_{1.1}$. Figure 11 also shows that including a linear trend in θ_1 leads to an underestimation of flows in the confirmatory period (by 6.7% on average), while the absence of this term leads to an overestimation by a similar magnitude (7.7% on average). Potential reasons for this finding are discussed in Section 9.3.

Figure 12 shows the AIC calculated over the exploratory period against the likelihood calculated over the confirmatory period, for both the full and thinned data sets. This plot examines the ability of the AIC to predict model performance in the confirmatory period. It can be seen that lower (better) AIC values over the exploratory period are associated with higher (better) log likelihood values in the confirmatory period. This association is statistically significant, with correlation coefficients of -0.66 and -0.44 for the full and thinned data sets, respectively. Therefore, even though the AIC tended to

favour more complex models in the exploratory period, this complexity appears to be justified by the data: these more complex models also have the highest likelihood values in the confirmatory period.

9. Discussion: Model selection for future climate predictions

This section discusses three alternative perspectives for selecting one or more models to be used for prediction. The discussion is not intended as exhaustive (e.g., Section 2 lists further approaches).

9.1 An information-theoretic approach to model selection

Section 6.1 discussed the use of the Kullback-Leibler information as a measure of the information lost when representing environmental processes using a (necessarily approximate) model. As noted by Burnham and Anderson [2010], a key theoretical appeal of the AIC is that, given a set of models, it identifies the model that approximately minimises the Kullback-Leibler information. However, as emphasized in Section 6.1, this appealing feature can be undermined if the assumptions in the likelihood function (and thus the AIC) are strongly violated. Note that this includes assumptions in both the deterministic and error models, i.e., AIC-based conclusions may be sensitive to deficiencies in either/both physical process representation and the statistical description of uncertainty.

Of the 22 models, whether calibrating to data set $\tilde{\mathbf{y}}^{(>0.09)}$ or $\tilde{\mathbf{y}}^{(>0.09)}_{t:=1+6j}$, model $g_{3.11}$ gives the lowest (best) AIC value, followed by model $g_{3.12}$ (models $g_{3.3}$ and $g_{3.4}$ also perform well in some of the thinned datasets, see Figure 9). Models $g_{3.11}$ and $g_{3.12}$ are the most complex models considered in this work, incorporating all the covariates and differing only in the calculation AET. Similar results have been found in other studies [e.g. Engelhardt et al., 2013], where the AIC tended to favour very complex models when compared to Bayesian selection criteria such as the BIC or KIC. Nevertheless, model $g_{3.11}$ is also found to maximise the likelihood in the confirmatory period, although significant scatter is observed (Figure 12).

The problems with applying the AIC weights in cases where the hydrological and error model assumptions are not met are demonstrated by comparing the weights of the candidate models. In the case where streamflow set $\tilde{\mathbf{y}}^{(>0.09)}$ is used for parameter estimation, the model with highest AIC rank has an Akaike weight of close to one, while all other models have weights close to zero. This may be due to the omission of error autocorrelation from the likelihood function, which results in an inference that over-estimates the information content of the data. If streamflow set $\tilde{\mathbf{y}}^{(>0.09)}_{t:=1+6j}$ is used, the residual error autocorrelation is no longer statistically significant at the 5% significance level, yet the AIC weight of the preferred model decreases only slightly, to 0.98. Since the AIC is derived under the assumption that the entire predictive model (here, GR4J and the WLS error model) is a sufficiently "good" approximation of the real system, it may be that the AIC is affected more by deficiencies in the hydrological model (i.e., in GR4J and its variants) than by deficiencies in the error model. Hence, further research is required to improve the specification of likelihood functions in hydrological modelling and understand the sensitivity of AIC weights to violations in the deterministic and stochastic components of the likelihood function. The use of 'hydrologically meaningful' measures of model performance is hence of clear importance, as described next [e.g. see Martinez and Gupta, 2011].

9.2 A multiple diagnostics approach to model selection

The limitations of single-metric approaches can be avoided by using multiple 'hydrologically meaningful' diagnostics [e.g. see Legates and McCabe, 1999; Martinez and Gupta, 2011]. These diagnostics can be constructed to scrutinize the ability of the model to reproduce specific hydrological features of interest. For example, in this study, seasonal flow duration curves were used to establish that model $g_{1.3}$ (for which parameter θ_1 is allowed to vary sinusoidally over the year), outperforms models based on other representations of non-stationarity (inter-annual, etc). From a physical perspective, this can be attributed to the seasonality of the catchment, with summer being water-limited and winter being energy-limited. Flow duration curves also helped to establish that, of all GR4J modifications considered in this study, the introduction of parameter θ_5 to control the portion of net rainfall directed to the production store yields the largest improvement in the simulation of hydrograph recessions.

The annual flow error (or bias) is another useful diagnostic, given its obvious relevance for studies such as reservoir yield analyses. However, in this study the predictive power of this statistic appears limited, with no statistically significant correlation between the flow errors in the confirmatory versus exploratory periods. These results indicate that a 'good' model in terms of overall mass balance over a calibration period may not be a 'good' model when applied in prediction.

In contrast to the flow error, the NSE performed better as a diagnostic tool, with statistically significant associations between the NSE in the exploratory and confirmatory periods. In contrast to the AIC, the NSE generally favoured simpler models, such as model $g_{2.2}$ (followed closely by $g_{2.3}$ and $g_{2.1}$) when calibrating to the streamflow set $\tilde{\mathbf{y}}^{(>0.09)}$, and $g_{1.1}$ when calibrating to the streamflow set $\tilde{\mathbf{y}}^{(>0.09)}_{t:=1+6j}$. As a result, in this study the models favored by the NSE are very different to the models favored by the AIC. This is likely to be due to the different weighting of low and high flows in the WLS-based likelihood (which attempts to balance the fitting of low and high flows) vs the NSE metric (which is generally insensitive to low flows).

Given that different models are favoured by different metrics, it is unclear how to best use multiple diagnostics for model selection and for constructing multi-model ensemble. Which models should be included in the ensemble, and how should they be weighted?

9.3 Use of independent information to assist in model selection

In many cases, information on a particular catchment may be difficult to include directly into a hydrological modeling framework, but may nevertheless enable the physical realism of the model predictions to be assessed against the empirical evidence. This is referred to as the 'principle of hydrological consistency' in Martinez and Gupta [2011].

In this study, the observed trend in parameter θ_1 might at least partially be explained by independent evidence suggesting an increase in farm dams in the catchment. In particular, the report by Teoh *et al*. [2002] shows that no farm dams were present in the catchment in 1987, increasing to 140 farm dams with a total storage volume of 118 ML in 1996, and to 161 farm dams with a total storage volume of 148 ML in 1999. The 1999 volume equates to a catchment-averaged

depth of 5.1 mm, and represents 4 % of the annual average catchment discharge over the exploratory period. Controls on the development of new farm dams have been instigated in the early 2000s [Teoh, 2002], and it is therefore likely that the total storage volume of farm dams would not have increased substantially since that time. Interestingly, during the confirmatory period all the models without a trend overestimated total annual flows, whereas all the models with a trend underestimated total annual flows (see Section 8.5). This is consistent with the independent evidence on trends in farm dams, however other changes (e.g. groundwater extraction due to agricultural activities) may also have occurred over this time, and cannot be ruled out as alternative potential physical causes of the non-stationarity in θ_1.

Catchment groundwater flux is an alternative source of information that can be used to evaluate hydrological consistency. In most models calibrated in this study, groundwater represents an important component of the water balance, although the total groundwater flux estimates varies substantially between models, ranging from 0.064 to 0.411 mm/day (Figure 4). The best available estimate of groundwater export (calculated as net recharge minus baseflow) was approximately 995 ML/year (0.094 mm/day) when averaged over the 30 year period from 1975-2004, although the estimates are very approximate and confidence intervals are not available [Adelaide and Mount Lofty Ranges Natural Resources Board, 2013]. Therefore, available evidence on the groundwater flux is consistent with the modeling results presented here, in that all evidence points to a groundwater export. However more detailed estimates of groundwater fluxes are needed before individual models can be more confidently excluded from the analysis.

10. Conclusions

This paper proposes and illustrates a strategy for diagnosing and interpreting hydrological non-stationarity. The major aim is to improve the ability of a hydrological model to provide extrapolative predictions under changing hydroclimatic conditions, since future hydroclimatic conditions may be outside of the domain of the data used for model selection and parameter estimation.

The strategy consists of four elements: (1) detecting, and where possible, eliminating, systematic errors in data; (2) allowing one or more hydrological model parameters to vary over time as functions of covariates intended to capture the relevant time scales of hydrological model non-stationarity (e.g., seasonal, annual and interannual); (3) trialing alternative model structures, with the aim of reducing hydrological model non-stationarity; and (4) model selection and evaluation including the combined use of information-theoretic metrics (such as the AIC) and hydrologically oriented diagnostics (such as flow duration curves).

The strategy is illustrated for a small catchment in South Australia, using the GR4J hydrological model as the initial hypothesis. A weighted least squares likelihood is applied to a thresholded and thinned data set to reduce the impact of low flows and residual error autocorrelation, respectively. An exploratory period is used for model calibration and selection, and a confirmatory period that is much drier than the exploratory period is used to test whether the models are robust under extrapolation.

The key conclusions of implementing the non-stationarity analysis strategy in the case study are:

1. Improved model predictions are obtained when the GR4J storage capacity parameter (θ_1) is made dependent on covariates describing seasonality, annual variability and longer term trends. No systematic errors were found in the calibration data itself, suggesting that the non-stationarity model of θ_1 is compensating for structural errors how the model represents changes in the hydrological dynamics of the catchment.

2. The model selection analysis highlights the impact of the choice of model evaluation metrics and methodology. The AIC approach often reports a strong difference between models, compared to the NSE metric which has a much lower discriminatory power. Hydrological models with low AIC values in the exploratory period also perform well in terms of the AIC in the confirmatory period. In contrast, models selected using the NSE performed poorly over the confirmatory period.

3. Hydrologically oriented model diagnostics, such as the flow duration curves (stratified by season, rising and falling hydrograph limbs, etc), are useful for detecting model weaknesses. For example, they can help detect systematic biases in predictions of low and high flows, motivate and guide changes in the model representation of recessions and actual evapotranspiration, and so on.

4. Overall, reasonable improvements in predictive performance are achieved: whereas the original GR4J model overestimates annual average flows in the confirmatory period by 18 %, the best-performing modified models underestimate the flows by only 3-7 %.

When using the inferred non-stationary models for developing streamflow projections for a future climate, scientific judgement is still required to estimate how the identified parameter trends might continue over time. For example, in this study, the identified trend of increasing model storage capacity could be tentatively explained by an increase in farm dams within the catchment, although other hypotheses such as changes in vegetation dynamics or groundwater extractions cannot be excluded. Given this uncertainty, projections should be made available using an ensemble of possible models, encompassing a range of possible future changes to catchment stores. This offers the best chance to adequately capture the uncertainty in future catchment behaviour.

Future research is recommended on: (1) extending the non-stationarity approach to multiple model parameters, to detect and quantify non-stationarity across non-nested models (e.g. models that do not share common parameters); (2) further exploring the AIC-based model selection methodology and comparing its results to other selection approaches such those identified in Section 2; and (3) applying the non-stationary approaches and model selection strategy to flexible model structures such as FUSE [Clark et al., 2008] and SUPERFLEX [Fenicia et al., 2011; Kavetski and Fenicia, 2011], with the aim of finding model structures that minimise parameter non-stationarity.

Acknowledgements

This research was funded by the Goyder Institute for Water Research as part of the project: *C.1.1 Development of an agreed set of climate projections for South Australia*, and their support is gratefully acknowledged. We also wish to acknowledge Associate Editor Jasper Vrugt, Hoshin Gupta and an anonymous reviewer, for constructive feedback that has helped improve the manuscript.

References

Adelaide and Mount Lofty Ranges Natural Resources Board (2013), Water Allocation Plan - Western Mount Lofty Ranges, edited.

Akaike, H. (1974), A new look at the statistical model identification, IEEE Transactions on Automatic Control, 19(6), 716-723.

Allen, R. G., L. S. Pereira, D. Raes, and M. Smith (1998), Statistical anlaysis of weather datasets, in Crop Evapotranspiration - Guidelines for Computing Crop Water Requirements, edited, FAO - Food and Agriculture Organisation of the United Nations.

Anderson, M. P., and W. W. Woessner (1992), The role of the postaudit in model validation, Advances in Water Resources, 15(1992), 167-173.

Bates, B. C., Z. W. Kundzewicz, S. Wu, and J. P. Palutikof (2008), Climate Change and Water. Technical Paper of the Intergovernmental Panel on Climate Change, edited, p. 210, IPCC Secretariat, Geneva.

Bergstrom, S. (1995), The HBV model, in Computer Models of Watershed Hydrology, edited by V. P. Singh, pp. 443-476, Highlands Ranch, CO.

Beven, K., and A. M. Binley (1992), The future of distributed hydrological models: Model calibration and uncertainty prediction, Hydrological Processes, 6, 279-298.

Beven, K., R. Lamb, P. F. Quinn, R. Romanowicz, and J. Freer (1995), TOPMODEL, in Computer Models of Watershed Hydrology, edited by V. P. Singh, Highlands Ranch, Colorado.

Brigode, P., L. Oudin, and C. Perrin (2012), Hydrological model parameter instability: an additional uncertainty in estimating the hydrological impacts of climate change?, Journal of Hydrology

Burnham, K. P., and D. R. Anderson (2010), Model Selection and Multimodel Inference, Springer, New York.

Chamberlain, T. C. (1890), The method of multiple working hypotheses, Science (reprinted in 1965), 15(92).

Chiew, F. H. S. (2006), An Overview of Methods for Estimating Climate Change Impact on Runoff, in 30th Hydrology and Water Resources Symposium, edited, Launceston.

Choi, H. T., and K. Beven (2007), Multi-period and multi-criteria model conditioning to reduce prediction uncertainty in an application of TOPMODEL within the GLUE framework, Journal of Hydrology 332, 316-336.

Claeskens, G., and N. L. Hjort (2008), Model selection and model averaging, Cambridge University Press, Cambridge.

Clark, M. P., A. G. Slater, D. E. Rupp, R. A. Woods, J. A. Vrugt, H. V. Gupta, T. Wagener, and L. E. Hay (2008), Framework for understanding structural errors (FUSE): A modular framework to diagnose differences between hydrological models, Water Resources Research, 44.

Clark, M. P., D. Kavetski, and F. Fenicia (2011), Pursuing the method of multiple working hypotheses for hydrological modelling, Water Resources Research, 47(W09301).

Coron, L., V. Andreassian, C. Perrin, J. Lerat, J. Vaze, M. Bourqui, and F. Hendrickx (2012), Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments, Water Resources Research, 48(W05552).

Dai, Z., A. Wolfsberg, P. Reimus, H. Deng, E. Kwicklis, M. Ding, D. Ware, and M. Ye (2012), Identification of sorption processes and parameters for radionuclide transport in fractured rock, Journal of Hydrology, 414-415, 516-526.

de Vos, N. J., T. H. M. Rientjes, and H. V. Gupta (2010), Diagnostic evaluation of conceptual rainfall-runoff models using temporal clustering, Hydrological Processes, 24, 2840-2850.

Engelhardt, I., J. G. De Aguinaga, H. Mikat, C. Schuth, and R. Liedl (2013), Complexity vs Simplicity: Groundwater Model Ranking using Information Criteria, Groundwater.

Evin, G., D. Kavetski, M. Thyer, and G. Kuczera (2013), Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration, Water Resources Research, 49(7), 4518-4524.

Fenicia, F., D. Kavetski, and H. H. G. Savenije (2011), Elements of a flexible approach for conceptual hydrological modelling: 1. Motivation and theoretical development, Water Resources Research, 47(W11510).

Gan, T. Y., and S. J. Burges (1990), An assessment of a conceptual rainfall-runoff model's ability to represent the dynamics of small hypothetical catchments: 2. Hydrologic responses for normal and extreme rainfall, Water Resources Research, 26(7), 1605-1619.

Gharari, S., M. Hrachowitz, F. Fenicia, and H. H. G. Savenije (2013), An approach to identify time consistent model parameters: sub-period calibration, Hydrological Earth Systems Science, 17, 149-161.

Guerrero, J.-L., I. K. Westerberg, S. Halldin, C.-Y. Xu, and L.-C. Lundin (2012), Temporal variability in stage-discharge relationships, Journal of Hydrology 446, 90-102.

Gupta, H. V., T. Wagener, and Y. Liu (2008), Reconciling theory with observations: elements of a diagnostic approach to model evaluation, Hydrological Processes, 22, 3802-3813.

Hastie, T., R. Tibshirani, and J. Friedman (2009), The Elements of Statistical Learning: Data Mining, Inference and Prediction.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999), Bayesian Model Averaging: A Tutorial, Statistical Science, 14(4), 382-417.

Jeffrey, S. J., J. O. Carter, K. B. Moodie, and A. R. Beswick (2001), Using spatial interpolation to construct a comprehensive archive of Australian climate data, Environmental Modelling & Software, 16(4), 309-330.

Kashyap, R. L. (1982), Optimal choice of AR and MA part sin autoregressive moving average models, IEEE Trans. Pattern Anal. Machine Intell., 4(2), 99-104.

Kavetski, D., and F. Fenicia (2011), Elements of a flexible approach for conceptual hydrological modelling: 2. Application and experimental insights, Water Resources Research, 47(W11511).

Kavetski, D., F. Fenicia, and M. P. Clark (2011), Impact of temporal data resolution on parameter inference and model identification in conceptual hydrological models: Insights from an experimental catchment, Water Resources Research, 47(W05501).

Klemes, V. (1986), Operational testing of hydrological simulation models, Hydrological Sciences Journal, 31(1), 13-24.

Kuczera, G., D. Kavetski, S. W. Franks, and M. Thyer (2006), Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterizing model error using storm-dependent parameters, Journal of Hydrology 331(1-2), 161-177.

Lavery, B., A. Kariko, and N. Nicholls (1992), A historical rainfall dataset for Australia, Australian Meteorological Magazine, 40, 33-39.

Le Lay, M., S. Galle, G. M. Saulnier, and I. Braud (2007), Exploring the relationship between hydroclimatic stationarity and rainfall-runoff model parameter stability: A case study in West Africa, Water Resources Research, 43(W07420).

Legates, D. R., and G. J. McCabe (1999), Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation, Water Resources Research, 35(1), 233-241.

Lin, Z., and M. B. Beck (2007), On the identification of model structure in hydrological and environmental systems, Water Resources Research, 43(W02402).

Marshall, L., D. Nott, and A. Sharma (2005), Hydrological model selection: a Bayesian alternative, Water Resources Research, 41(W10422).

Marshall, L., D. Nott, and A. Sharma (2007), Towards dynamic catchment modelling: a Bayesian hierarchical mixtures of experts framework

Martinez, G. F., and H. V. Gupta (2011), Hydrologic consistency as a basis for assessing complexity of monthly water balance models for the continental United States, Water Resources Research, 47(W12540).

McMahon, T. A., M. C. Peel, L. Lowe, R. Srikanthan, and T. R. McVicar (2013), Estimating actual, potential, reference crop and pan evaporation using standard meteorological data: a pragmatic synthesis, Hydrological Earth Systems Science, 17, 1331-1363.

McQuarrie, A. D. R., and C.-L. Tsai (2007), Regression and Time Series Model Selection, Singapore.

Merz, R., J. Parajka, and G. Bloschl (2011), Time stability of catchment model parameters: implications for climate impact analyses, Water Resources Research, 47(W02531).

Milly, P. C. D., J. Betancourt, M. Falkenmark, R. M. Hirsch, W. Zbigniew, Z. W. Kundzewicz, D. P. Lettenmaier, and R. J. Stouffer (2008), Stationarity is Dead: Whither Water Management?, Science, 319, 573-574.

Molini, A., L. G. Lanza, and P. La Barbera (2005), The impact of tipping bucket rain gauge measurement errors on design rainfall for urban-scale applications, Hydrological Processes, 19(5), 1073-1088.

Morton, F. I. (1983), Operational estimates of areal evapotranspiration and their significance to the science and practice of hydrology, Journal of Hydrology, 66, 1-76.

Oreskes, N., K. Shrader-Frechette, and K. Belitz (1994), Verification, Validation and Confirmation of Numerical Models in the Earth Sciences, Science, 263(5147), 641-646.

Paik, K., J. H. Kim, H. S. Kim, and D. R. Lee (2005), A conceptual rainfall-runoff model considering seasonal variation, Hydrological Processes, 19, 3837-3850.

Pathiraja, S., S. Westra, and A. Sharma (2012), Why continuous simulation? The role of antecedent moisture in design flood estimation, Water Resources Research, 48(W06534).

Perrin, C., C. Michel, and V. Andreassian (2003), Improvement of a parsimonious model for streamflow simulation, Journal of Hydrology, 279, 275-289.

Renard, B., D. Kavetski, E. Leblois, M. Thyer, G. Kuczera, and S. W. Franks (2011), Toward a reliable decomposition of predictive uncertainty in hydrological modelling: Characterizing rainfall errors using conditional simulation, Water Resources Research, 47(W11516).

Schoups, G., N. van de Giesen, and H. H. G. Savenije (2008), Model complexity control for hydrological prediction, Water Resources Research, 44(W00B03).

Schoups, G., and J. A. Vrugt (2010), A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroskedastic and non-Gaussian errors, Water Resources Research, 46(10).

Schwarz, G. (1978), Estimating the dimension of a model, Annals of Statistics, 6(2), 461-464.

Seiller, G., F. Anctil, and C. Perrin (2012), Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions, Hydrological Earth Systems Science, 16, 1171-1189.

Smith, T. J., A. Sharma, L. Marshall, R. Mehrotra, and S. A. Sisson (2010), Development of a formal likelihood function for improved Bayesian inference of ephemeral catchments, Water Resources Research, 46(W12551).

Sorooshian, S., and J. A. Dracup (1980), Stochastic parameter estimation procedures for hydrological rainfall-runoff models: correlated and heteroscedastic error cases, Water Resources Research, 16(2), 430-442.

Sorooshian, S. (1981), Parameter estimation of rainfall-runoff models with heteroskedastic straemflow errors - The noninformative data case, Journal of Hydrology 52(1-2), 127-138.

Sugiura, N. (1978), Further analysis of the data by Akaike's information criterion and the finite corrections, Communications in Statistics, Theory and Methods, A7, 13-26.

Teoh, K. S. (2002), Estimating the Impact of Current Farm Dams Development on the Surface Water Resources of the Onkaparinga River CatchmentRep., Department of Water, Land and Biodiversity Conservation.

Thyer, M., B. Renard, D. Kavetski, G. Kuczera, S. W. Franks, and R. Srikanthan (2009), Critical evaluation of parameter consistency and predictive uncertainty in hydrological modelling: A case study using Bayesian total error analysis, Water Resources Research, 45(W00B14).

Vaze, J., D. A. Post, F. H. S. Chiew, J.-M. Perraud, N. R. Viney, and J. Teng (2010), Climate non-stationarity - Validity of calibrated rainfall-runoff models for use in climate change studies, Journal of Hydrology 394, 447-457.

Wagener, T., N. R. McIntyre, M. J. Lees, H. S. Wheater, and H. V. Gupta (2003), Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiability analysis, Hydrological Processes, 17, 455-476.

Weijs, S., G. Schoups, and N. de Giesen (2010), Why hydrological predictions should be evaluated using information theory, Hydrological Earth Systems Science, 14(12), 2545-2558.

Westra, S., J. P. Evans, R. Mehrotra, and A. Sharma (2013), A conditional disaggregation algorithm for generating fine time-scale rainfall data in a warmer climate, Journal of Hydrology, 479, 86-99.

Wu, K., and C. A. Johnston (2007), Hydrologic response to climate variability in a Great Lakes Watershed: A case study with the SWAT model, Journal of Hydrology, 337, 187-199.

Ye, M., P. D. Meyer, and S. P. Neuman (2008), On model selection criteria in multimodel analysis, Water Resources Research, 44(W03428).

Ye, W., B. C. Bates, N. R. Viney, M. Sivapalan, and A. J. Jakeman (1997), Performance of conceptual rainfall-runoff models in low-yielding ephemeral catchments, Water Resources Research, 33(1), 153-166.

Young, P., and K. Beven (1994), Data-based mechanistic modelling and the rainfall-flow non-linearity, Environmetrics, 5, 335-363.

Young, P. (1998), Data-based mechanistic modelling of environmental, ecological, economic and engineering systems, Environmental Modelling & Software, 13, 105-122.

Zhang, H., G. H. Huang, D. Wang, and X. Zhang (2011), Multi-period calibration of a semi-distributed hydrological model based on hydroclimatic clustering, Advances in Water Resources, 34, 1292-1303.

List of Figure Captions

Tables

Table 1: Modified GR4J models used in this paper. The parameters for the non-stationary model of θ_1 are represented by $\lambda_1 ... \lambda_6$ as described in Equation (1). Parameter θ_5 is described in Equation (2). The last column describes the approach used to calculate net rainfall (P_n).

New process	**-**	***Trend***	***Seasonality***		***Antecedent***		**-**	**-**	**-**	**-**	***Net precip***
					Rain	***PET***					
Model	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	θ_2	θ_3	θ_4	θ_5	P_n
$g_{1.1}$	✓						✓	✓	✓		$P - E$
$g_{1.2}$	✓	✓					✓	✓	✓		$P - E$
$g_{1.3}$	✓		✓	✓			✓	✓	✓		$P - E$
$g_{1.4}$	✓				✓	✓	✓	✓	✓		$P - E$
$g_{1.5}$	✓	✓	✓	✓			✓	✓	✓		$P - E$
$g_{1.6}$	✓		✓	✓	✓	✓	✓	✓	✓		$P - E$
$g_{1.7}$	✓	✓			✓	✓	✓	✓	✓		$P - E$
$g_{1.8}$	✓	✓	✓	✓	✓	✓	✓	✓	✓		$P - E$
$g_{2.1} = g_{1.1}$	✓						✓	✓	✓		$P - E$
$g_{2.2}$	✓						✓	✓	✓	✓	$P - E$
$g_{2.3}$	✓						✓	✓	✓		P
$g_{2.4}$	✓						✓	✓	✓	✓	P
$g_{3.1} = g_{2.2}$	✓						✓	✓	✓	✓	$P - E$
$g_{3.2}$	✓	✓					✓	✓	✓	✓	$P - E$
$g_{3.3}$	✓		✓	✓			✓	✓	✓	✓	$P - E$
$g_{3.4}$	✓	✓	✓	✓			✓	✓	✓	✓	$P - E$
$g_{3.5}$	✓	✓					✓	✓	✓		P
$g_{3.6}$	✓		✓	✓			✓	✓	✓		P
$g_{3.7}$	✓	✓	✓	✓			✓	✓	✓		P
$g_{3.8}$	✓	✓					✓	✓	✓	✓	P
$g_{3.9}$	✓		✓	✓			✓	✓	✓	✓	P
$g_{3.10}$	✓	✓	✓	✓			✓	✓	✓	✓	P
$g_{3.11}$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	$P - E$
$g_{3.12}$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	P

Figures

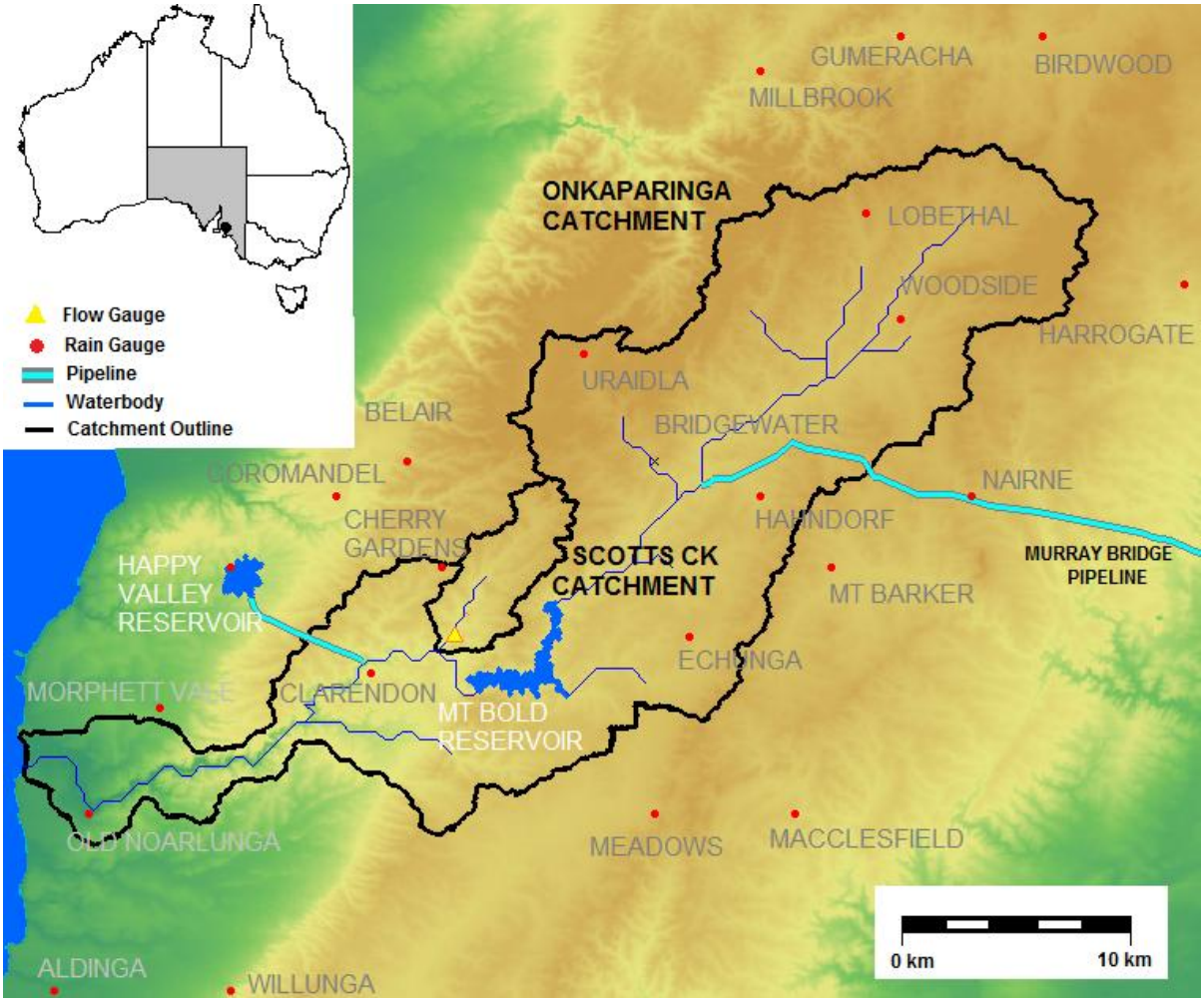


Figure 1. Map of the Onkaparinga catchment.

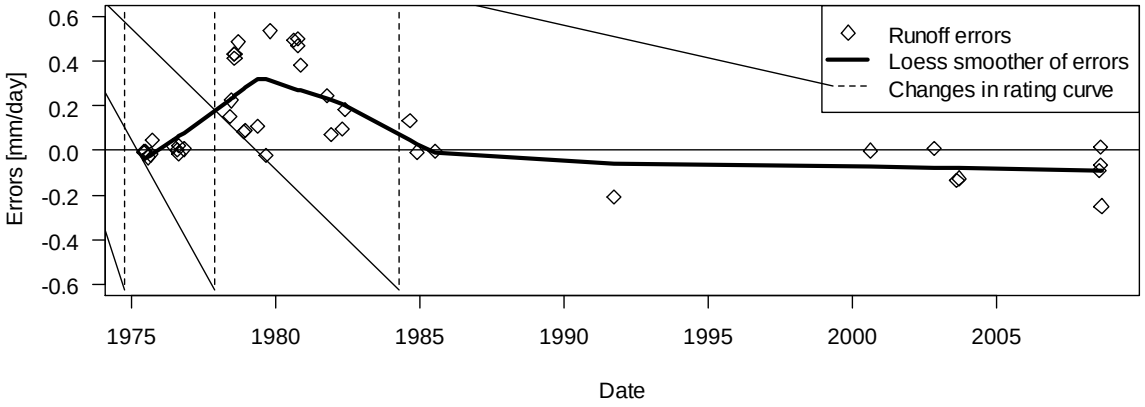


Figure 2. Runoff error time series at Scott Creek over the exploratory (1985-1999) and confirmatory (2000-2009) periods. Runoff errors are defined as the differences between the streamflows predicted by rating curve and the actual streamflow gauging. The loess smoother [Hastie et al., 2009] of errors shows a clear overprediction of streamflow prior to last rating curve change in 1984.

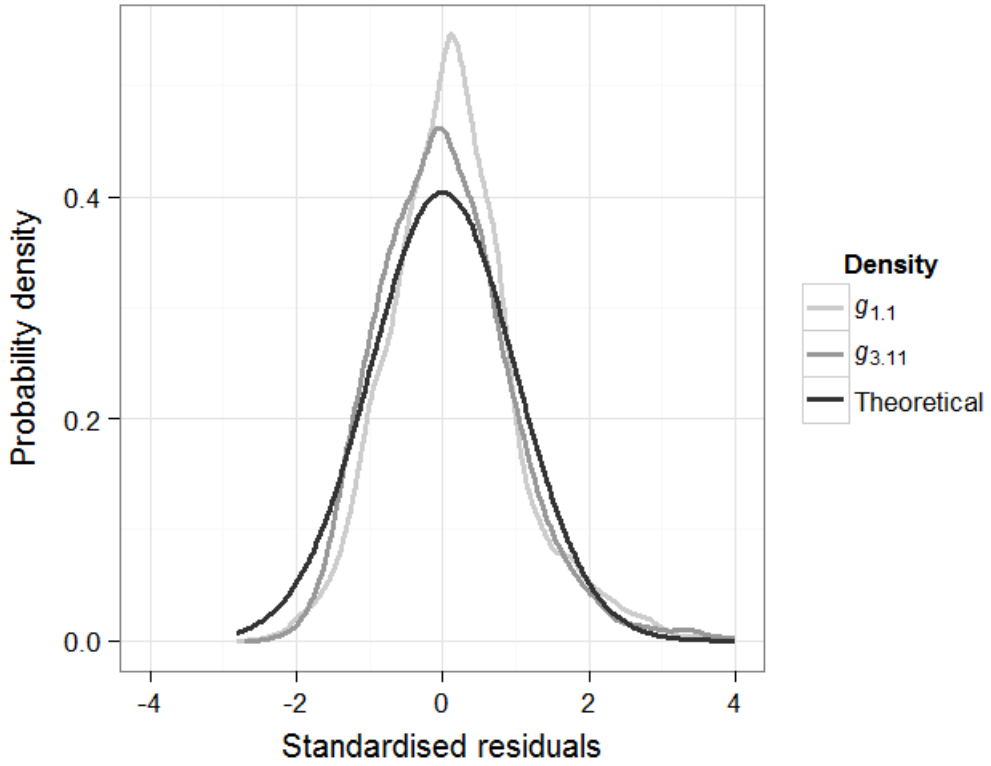


Figure 3. Density plot of the standardised residuals in the exploratory period for models $g_{1.1}$and $g_{3.11}$. The standard Gaussian distribution is shown for reference.

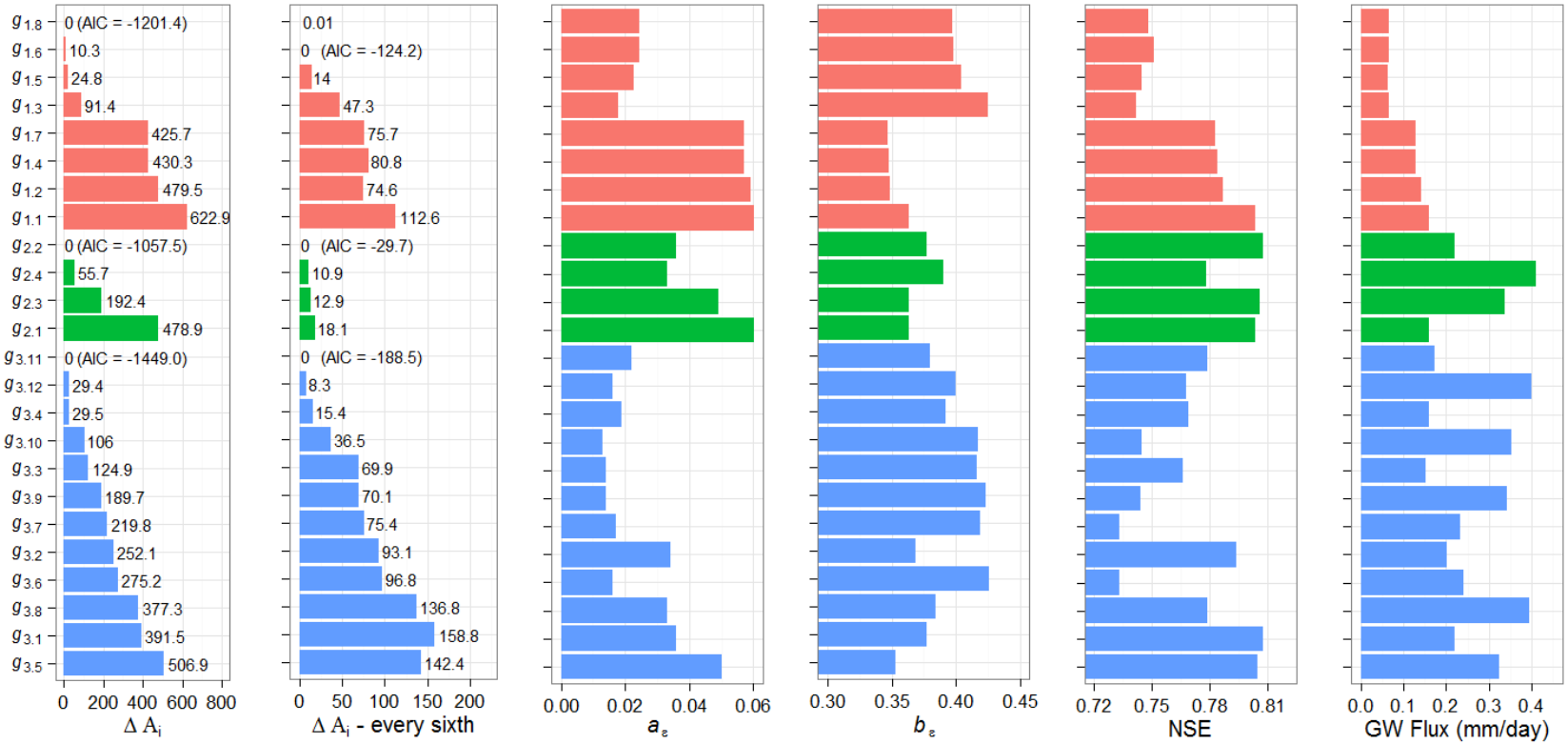


Figure 4. Model comparison in the exploratory period: Akaike differences (Δ_i) when using every day and every sixth day in the likelihood function, residual error model parameters (a_ε and b_ε), Nash Sutcliffe Efficiency (NSE) and groundwater flux for all models. The red, green and blue colours indicate the model structure groupings $g_{1.x}$, $g_{2.x}$ and $g_{3.x}$ respectively. Within each grouping, the models are ordered from best to worst performance, as given by the AIC differences.

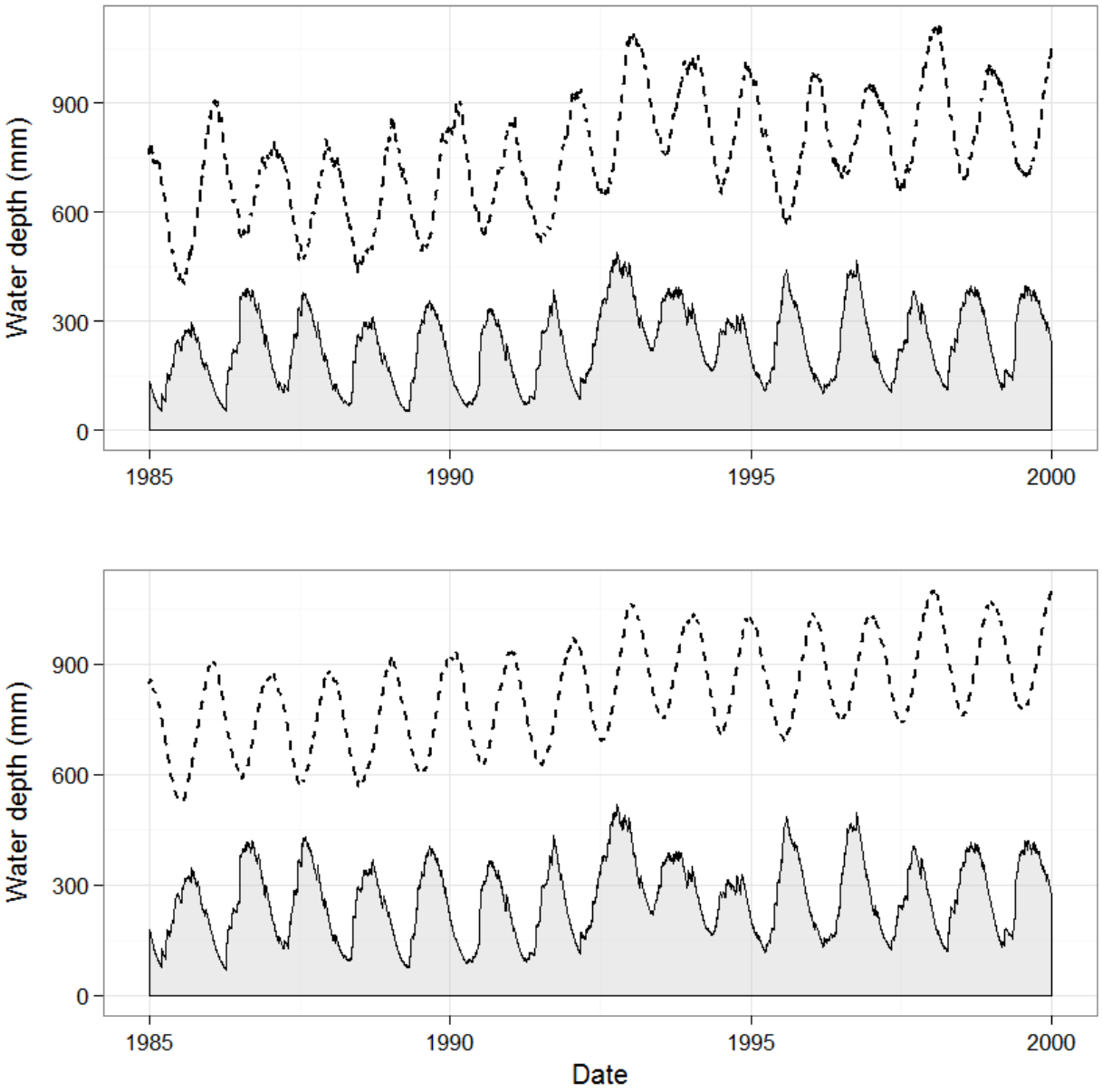


Figure 5. Time series of the production store capacity parameter ϑ_1 (dotted line) and the actual storage S in the production store (solid grey curves). The top panel shows the results for the AIC-best model ($g_{1.8}$) obtained when calibrating to streamflow set $\tilde{\mathbf{y}}^{(>0.09)}$ whereas the bottom panel shows the results for the AIC-best model ($g_{1.6}$) obtained when calibrating to streamflow set $\tilde{\mathbf{y}}^{(>0.09)}_{t:=1+6j}$;. See Section 6.3 for a description of the streamflow sets.

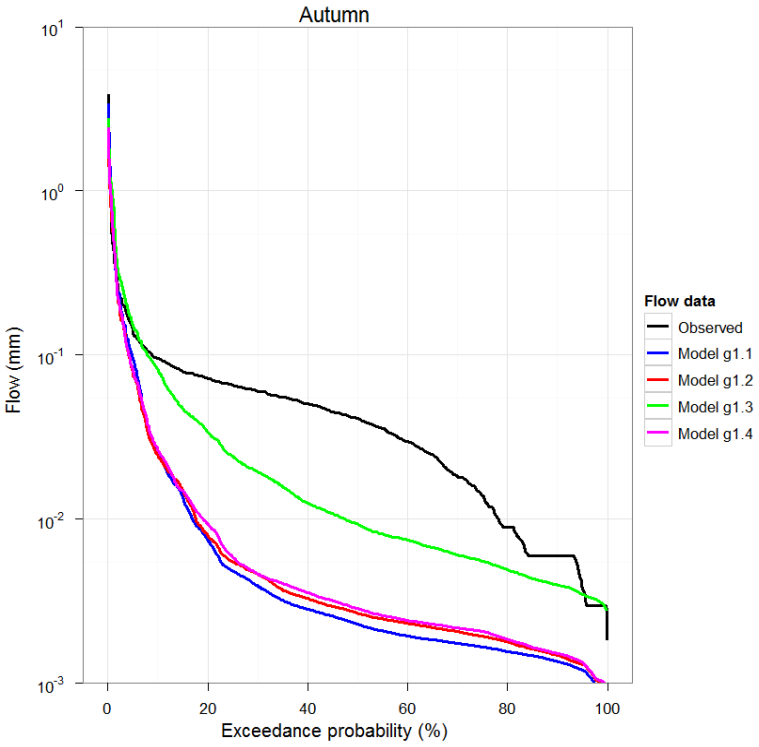


Figure 6. Flow duration curve during autumn. Observed data (black line) and models $g_{1.1}$ (blue line), $g_{1.2}$ (red line), $g_{1.3}$ (green line), $g_{1.4}$ (magenta line) are shown over the exploratory period (1985-1999).

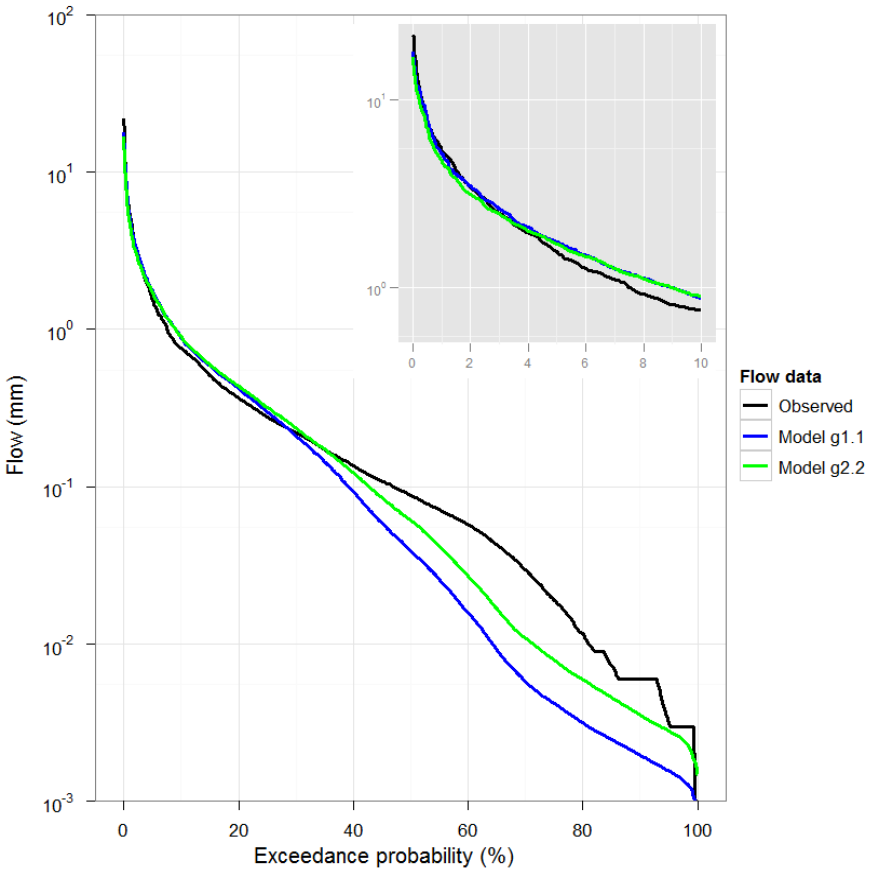


Figure 7. Observed and simulated flow duration curves over the exploratory period. The inset zooms in on the highest 10 % of flow days.

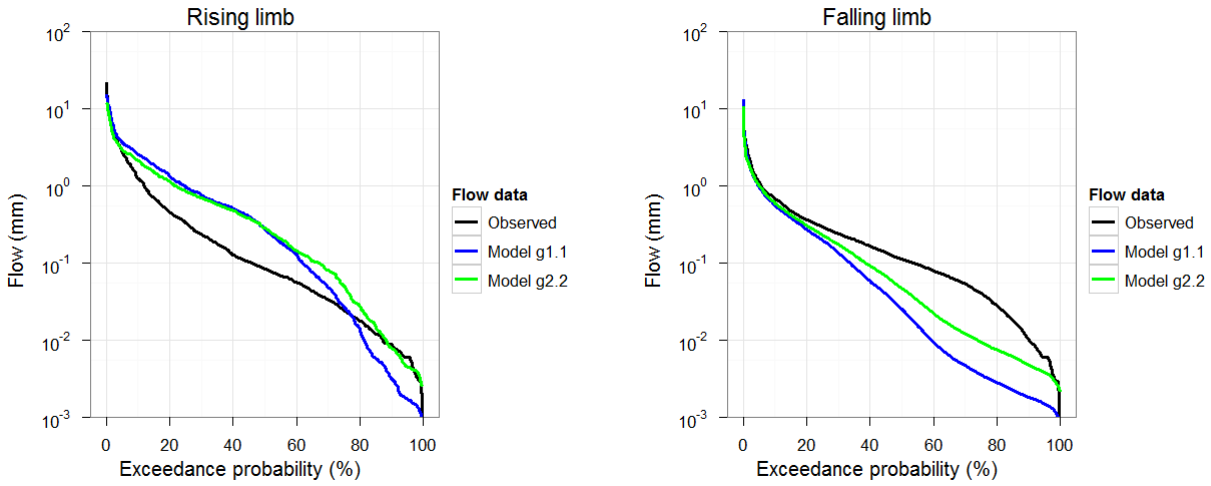


Figure 8. Observed and simulated flow duration curves for the rising limb (left panel) and falling limb (right panel) of hydrographs in the exploratory period.

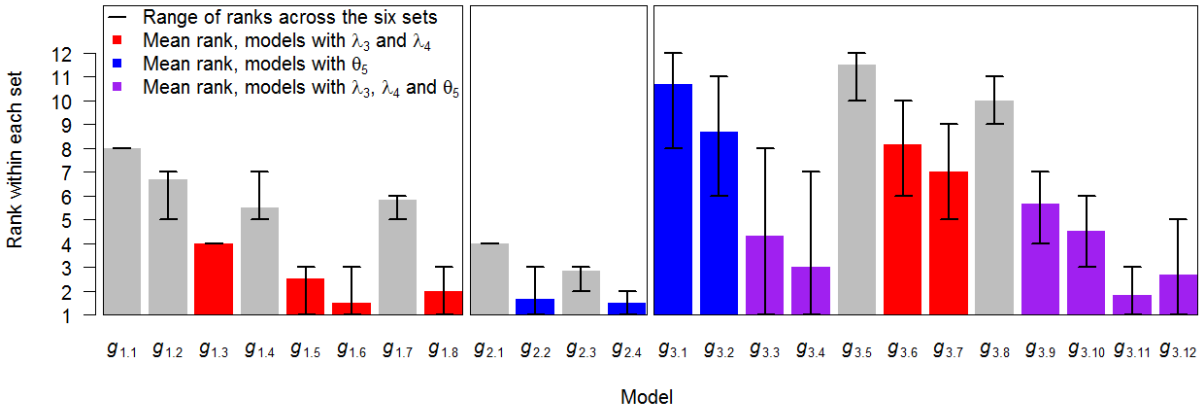


Figure 9: AIC ranking when models are calibrated separately to the six distinct thinned datasets (Section 6.3.2). The model sets $g_{1.X}$, $g_{2.X}$ and $g_{3.X}$ are shown in separate panels. Within each panel, the mean rank and the range for each of the six sets is presented. The colours denote an alternative way of grouping the models, based on the presence of particular calibrated parameters (as indicated in the legend).

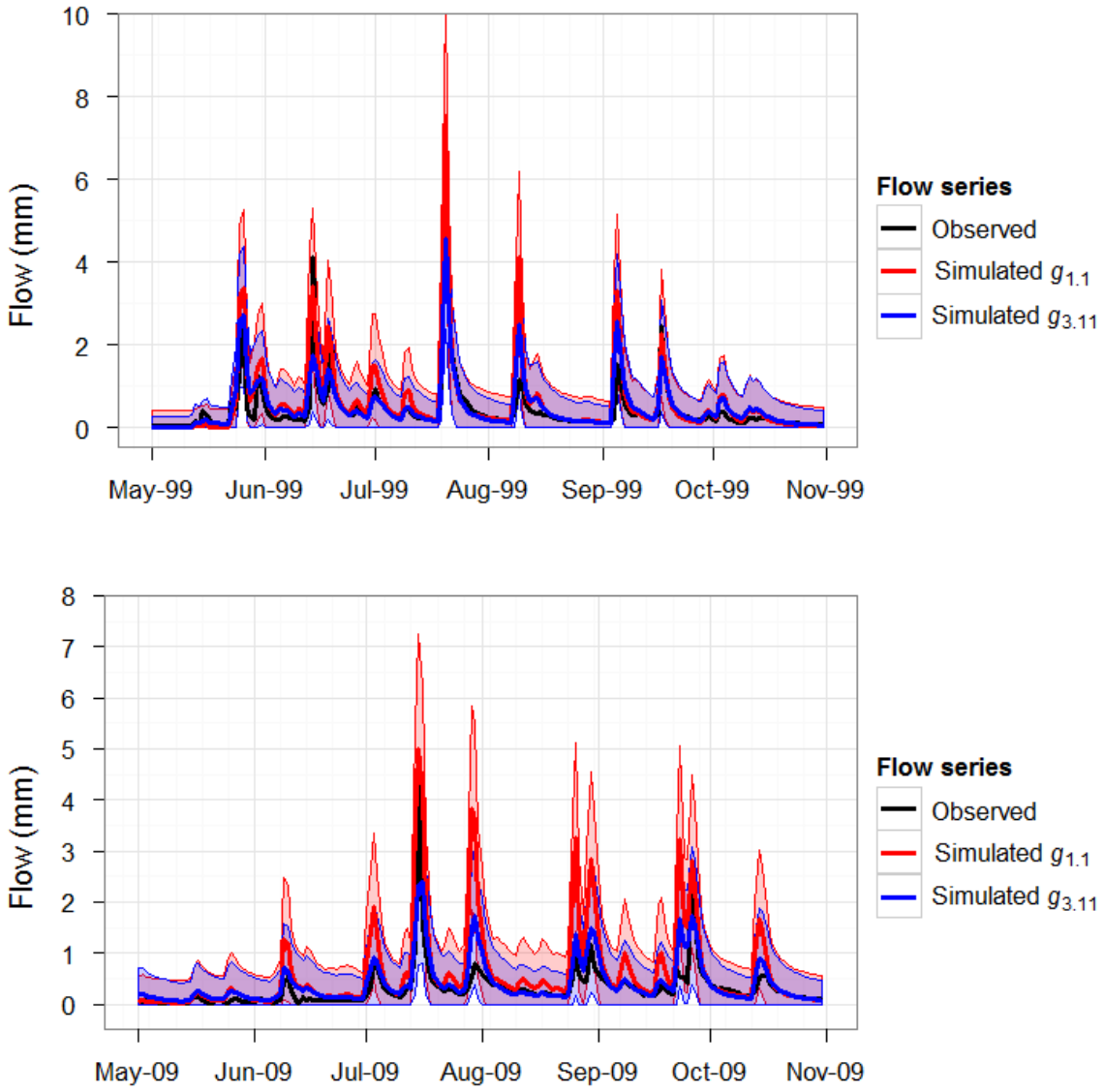


Figure 10: Observed and simulated flow series for a representative year from the exploratory period (upper panel) and an independent drier confirmatory period (lower panel), for the standard GR4J model (model $g_{1.1}$, red lines) and the AIC-best model (model $g_{3.11}$, blue lines). The 90 % confidence intervals are shown using shading.

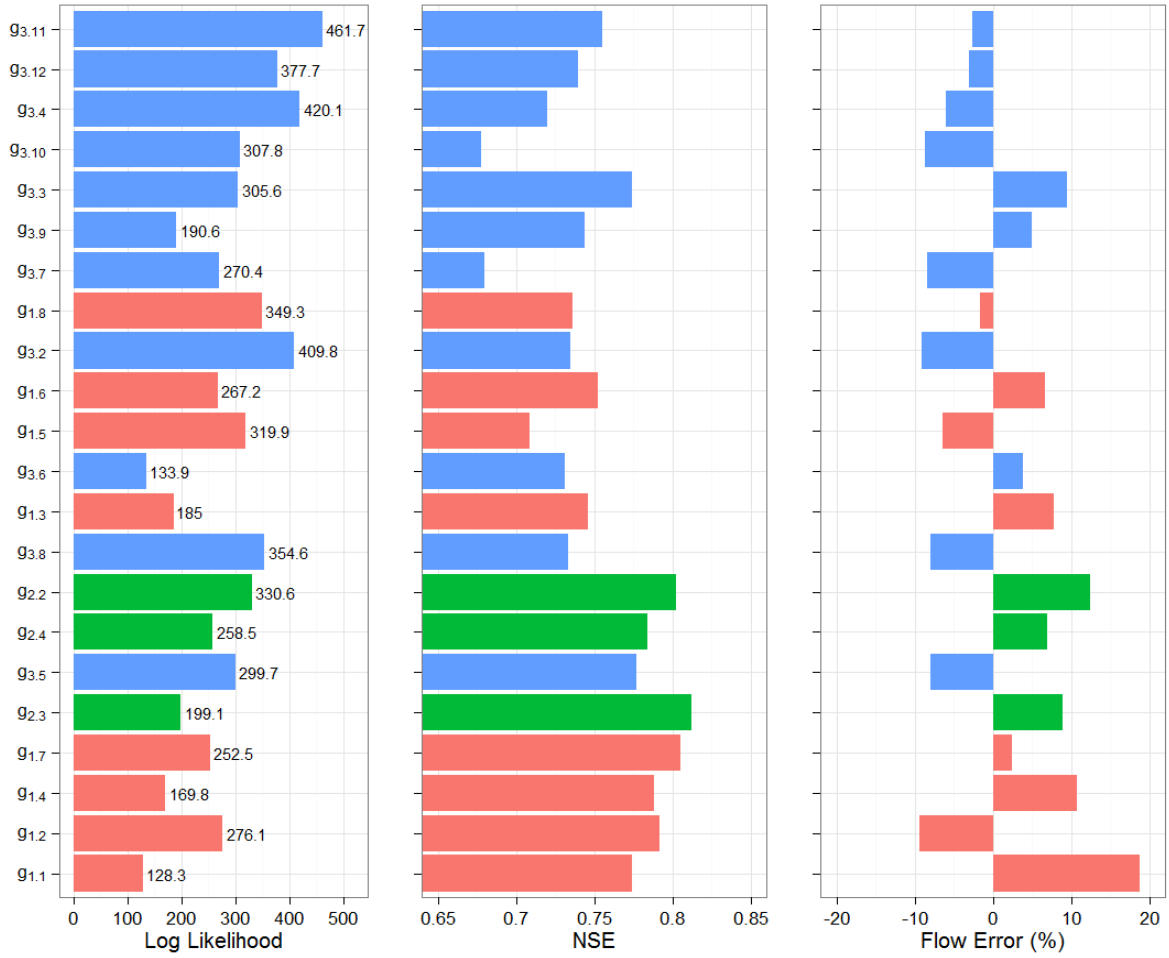


Figure 11: Model likelihood, NSE and flow error (as a percentage of total annual flow) evaluated during the confirmatory period. Red, green and blue colours indicate the model structure groupings $g_{1.X}$, $g_{2.X}$ and $g_{3.X}$ respectively. All models are ordered from best to worst performance, as given by the AIC differences over the exploratory period (see Figure 4).

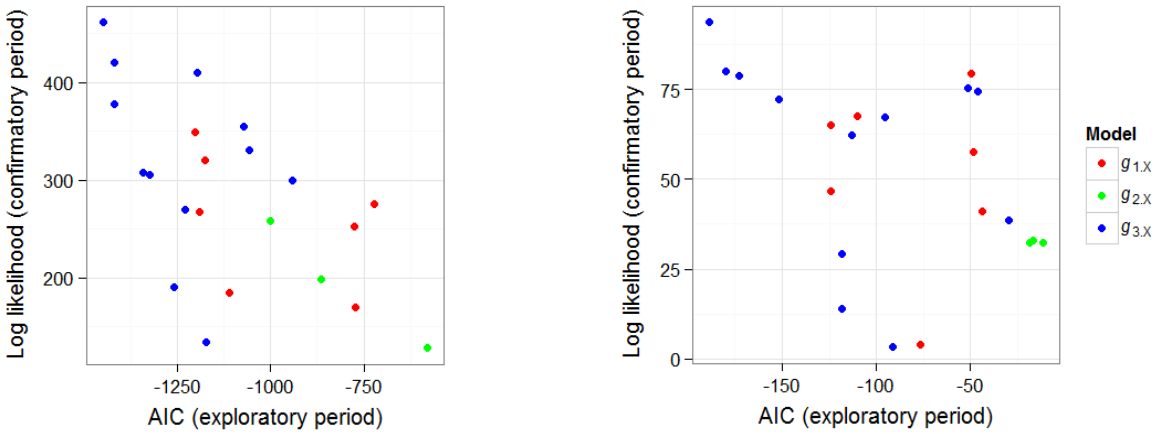
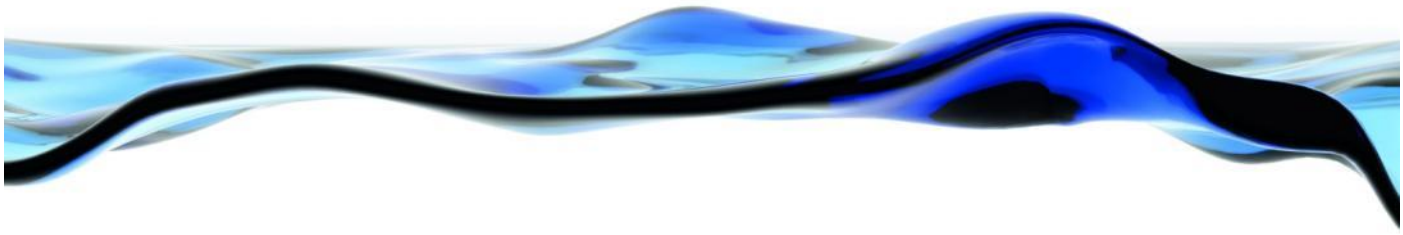


Figure 12. Likelihood function values computed over the confirmatory period (2000-2009) plotted against AIC values computed over the exploratory period (1988-1999). Left panel shows results for the full data set; right panel shows results for the thinned dataset. Correlation coefficients are -0.66 and -0.44, respectively, which are statistically significant at the 5 % level.

GOYDER INSTITUTE FOR WATER RESEARCH



Government of South Australia
Department of Environment, Water and Natural Resources

CSIRO

THE UNIVERSITY of ADELAIDE

University of South Australia

Flinders UNIVERSITY ADELAIDE • AUSTRALIA

The Goyder Institute for Water Research is a partnership between the South Australian Government through the Department of Environment, Water and Natural Resources, CSIRO, Flinders University, the University of Adelaide and the University of South Australia.