OFFICIAL

Australia's National Science Agency

MD-WERP Deliverable 2.8a.5

# Continued investigation of skill of streamflow forecasts to assist management of the Narran Lakes (Dharriwaa)

James C Bennett, David E Robertson, Zeqing Huang

Australian Government

Murray–Darling Water & Environment Research

## Citation

Bennett JC, Robertson DE, Huang Z (2024) Continued investigation of skill of streamflow forecasts to assist management of the Narran Lakes (Dharriwaa). MD-WERP Deliverable T.2.8a.5. CSIRO, Australia.

## Copyright



## Important disclaimer

# Contents

# Figures



# Tables

# Acknowledgments

This work was conducted on the traditional lands of the Boonwurrung and Djarra peoples of the Kulin Nation. We acknowledge their continuing custodianship of these lands and the rivers that flow through them, and pay our respects to their elders, past and present. This work was funded through the Murray-Darling Basin Authority's Water Environment Research Program. We also acknowledge the traditional custodians of the catchments and rivers used in this study, including the Yuwaalaraay/Euahlayi people. We thank Christopher Pickett-Heaps, Ulrike Bende-Michl and Wendy Sharples (Bureau of Meteorology) for advice on the availability of Australian Water Outlooks retrospective forecasts. Thanks to Michael Peat, Philippa Russell Brown and Peter Brownhalls for advice on management of the Narran Lakes. The retrospective forecasts generated for this study were run on CSIRO's Petrichor and Bowen Cloud high-performance computing infrastructure.

# Executive summary


This study continues the assessment of potential operational forecasting products for the purpose of assisting management of water bird breeding events in the Narran Lakes (Dharriwaa). This study assessed two products:

1. A modified version of the Bureau of Meteorology's seven-day ensemble streamflow (SDES) forecasting system, termed SWIFT forecasts in this study
2. Runoff outlooks from the Bureau of Meteorology's Australian Water Outlooks (AWO) statistically processed with the Bayesian Joint Probability model

Both systems were configured to produce daily streamflow forecasts at key points in the Balonne River system, in particular for flows at the Balonne River at St George gauge. Neither system takes account of irrigation extractions or the effects of Beardmore Dam, which is upstream of the St George Gauge.

In a previous study we trialled modifications to the SDES system, finding streamflow forecast skill only to 7-8 days. With the additional modifications trialled in this study: to rainfall observations, rainfall-runoff objective function, error model bias-correction and rainfall forecast forcing, we show that SWIFT forecasts can be skilful to 15+ days, and potentially beyond 30 days. A key source of additional skill was the inclusion of statistically processed Numerical Weather Predictions of precipitation within the modelling chain. The SWIFT forecasts were substantially more skilful than the statistically processed AWO forecasts at the Balonne River at St George gauge.

Skill in SWIFT forecasts was also evident for forecasts of large flows, up to 500 GL averaged over 30 days. These are strongly skilful forecasts, which if operational could benefit the management of the Narran Lakes. Our major findings and recommendations are:

1. Modified SDES forecasting methods, termed 'SWIFT forecasts' in this report, produce forecasts that promise enough skill to inform more effective management of large inflow events into the Narran Lakes. A crucial aspect of this modification was jointly estimating transformation and hydrological model parameters with a likelihood.

2. Skilful rainfall forecasts are a key component of the modelling chain, and should be included in any attempts to operationalise forecasts

3. The forecasts trialled here are not in operation. All methods used in this report are encoded in operations grade software, allowing expedited operationalisation.

# 1 Introduction

The Narran Lakes (Dharriwaa) are a system of wetlands in Northern NSW of international significance as a breeding site for water birds. Dharriwaa supports several threatened species and is listed under the Ramsar Convention (1971). Water bird breeding events require a level of at least 1.08 m on the back lake gauge to meet water bird breeding conditions. Dharriwaa receives inflows from the Narran River, a distributary of the Balonne River (Figure 1). There are no water storages upstream of Dharriwaa that are large enough to buffer inflows to the lake. To keep water levels above 1.08 m, inflows can be managed by restricting the diversion of water for irrigation by purchasing irrigation entitlements. Modelling inflow to the Narran Lakes is difficult due to the complexity of hydraulic and hydrological processes in this region – notably the anabranching of the Balonne River into the Culgoa, Bokhara and Narran Rivers. For management purposes, water levels are approximated by volumes of flow events that pass the Balonne River at St George gauge (Figure 1). Note that the duration of these events is unspecified – they could occur over anything from 1-30 days. For water-bird breeding, events between 250 and 500 GL can be augmented with water purchased from irrigators to prolong ideal breeding conditions. We summarise key cumulative inflow volume thresholds for this study in Table 1.

This study continues to investigate if currently operational water forecasting systems may be helpful in helping manage the Narran Lakes. It builds on work by Bennett and Robertson (2023), who investigated hydrological forecasting methods used in the Bureau of Meteorology's 7-day ensemble forecasting system to predict flows at key gauges in the Balonne and Maranoa Rivers. In their initial setup, Bennett and Robertson (2023) did not use informative rainfall forecasts, so the forecasts relied purely on catchment memory for skill. Bennett and Robertson (2023) found that forecast skill was only available with these methods for lead times of ~7 days, and this skill varied considerably with season. This level of skill is unlikely to be sufficient to improve the management of the Narran Lakes.

In this study, we continue to assess the skill of 7-day forecasting methods under different configurations, and in addition to assessing another forecasting product, the Australian Water Outlook. The additional experiments we conduct are as follows:

1. **7-day forecasting methods**
    a. We assess different rainfall information for the 7-day forecasting methods
    b. We experiment with different error model configurations within the 7-day forecasting methods
    c. We incorporate skilful rainfall forecasts into the 7-day forecasting methods
2. **Australian Water Outlook**
    a. We assess retrospective forecasts from the Australian water outlooks
    b. We investigate whether these forecasts can be improved through post-processing

Table 1 Magnitude of flow events relevant to management of Narran Lakes. The duration of flow events is variable.

| VOLUME OF EVENT AT BALONNE R AT ST GEORGE GAUGE (GL) | ENVIRONMENTAL BENEFIT | FURTHER BENEFIT POSSIBLE BY PURCHASING WATER FROM IRRIGATORS |
|---|---|---|
| **>154** | Minimum event size to ensure Narran Lake levels support vegetation regeneration. | Yes |
| **250-499** | 250 GL is the minimum event size to ensure Narran Lake level of 1.08 m needed to support water bird breeding. | Yes |
| **500** | Event size that guarantees Narran Lake level exceeds the 1.08 m level needed to support water bird breeding. | No |

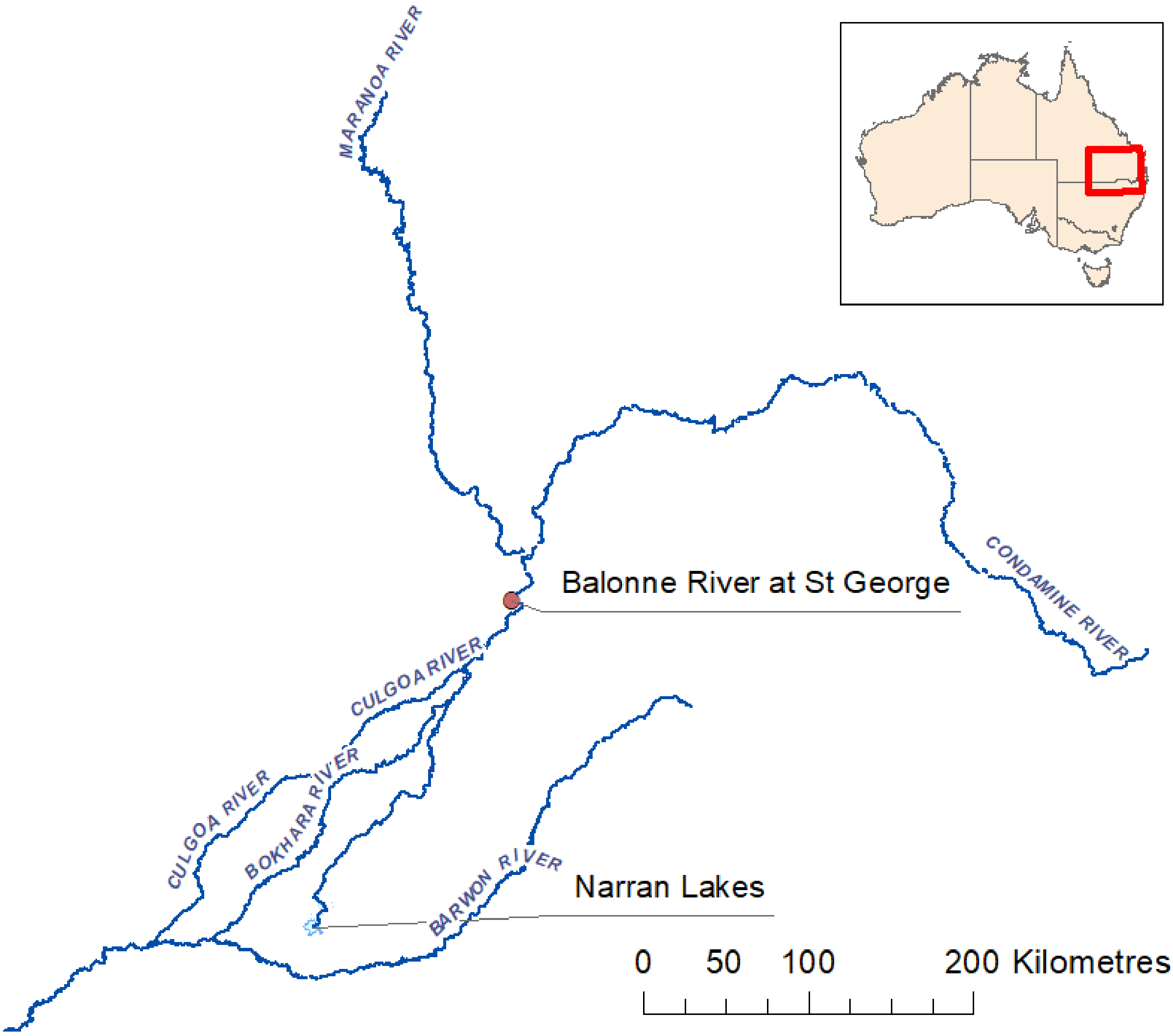


Figure 1 Location of Narran Lakes and the Balonne River at St George gauge.

# 2 Methods

## 2.1 SWIFT forecasts: SDES forecasting system

A schematic describing the modelling workflow to generate the SDES forecasts is shown in Figure 2. The method combines: (1) Numerical Weather Prediction (NWP) forecasts, (2) statistical procesing, (3) conceptual rainfall-runoff modelling and (4) hydrological error modelling. Forecast skill – that is, information in the forecast that improves upon a simple climatology prediction such as average monthly streamflow – comes from:

1) Atmospheric initial conditions, as estimated in the NWP (1)
2) Hydrological initial conditions, estimated by the conceptual rainfall-runoff model GR4J (3)
3) Memory of hydrological errors, estimated by the ERRIS (error reduction and representation in stages) error model (4)

NWP rainfall forecasts are developed from atmosphere-only models, which tend to run out of skill at lead times of 5-6 days. The information from hydrological errors often persists for an even shorter time horizon. Information from hydrological initial conditions can persist much longer – in some cases for weeks or even months - depending on the catchment.

At present the system runs at the hourly time step to produce forecast to 7 days. To meet the requirements of this study, the system can run at the daily timestep, but must run to 30 days. We will refer to forecasts generated with modified SDES methods as 'SWIFT' forecasts, after the Short-term Water and Information Forecasting Tools (SWIFT, Perraud et al., 2015)) software package that is used to generate them.

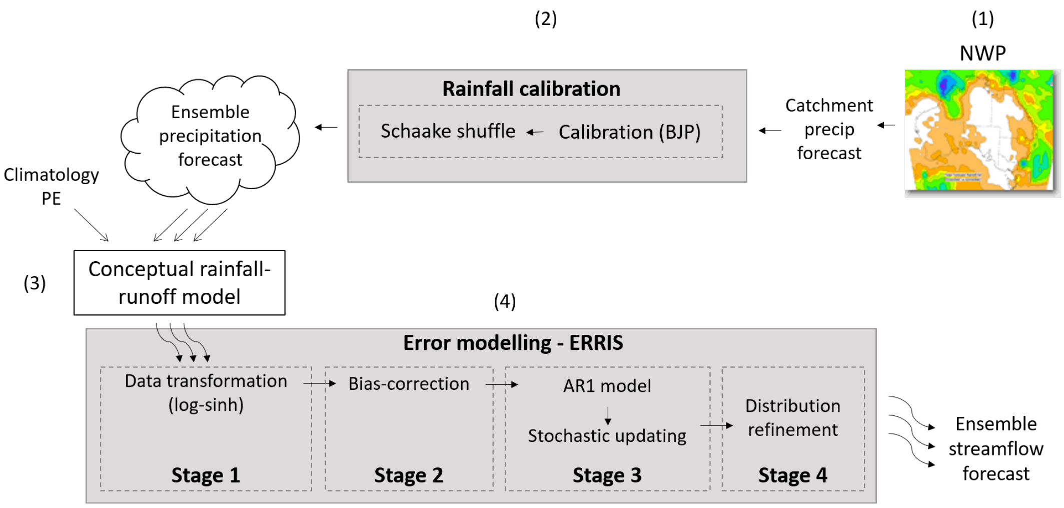


Figure 2 Modelling workflow used to generate SDES forecasts.

### 2.1.1 Adapting the SDES forecasting system to run at the daily time step and to longer lead times in SWIFT

#### Adaptations to hydrological and error models

The GR4J hydrological model was originally designed to run at the daily time step (Perrin et al., 2003), so no changes are required to this component. The ERRIS error model was also originally designed to run at the daily time step (Li et al., 2016), and requires no changes to run at the daily time step. We note, however, that we updated ERRIS to improve its function in ephemeral catchments (Bennett et al., 2021). This update has not been included in the current SDES system but is slated for future inclusion. The Maranoa River – one of the key tributaries of the Balonne River, and one that we calibrate the SDES models to explicitly – regularly ceases to flow. For this reason, we use the newer version of ERRIS. The newer version of ERRIS uses a revised restriction on updating and a dynamic bias-correction, which corrects biases measured over a preceding time window, and the duration of this window must be specified. For hourly timesteps, Bennett et al. (2021) recommend a window of 10 days, and this was used by Bennett and Robertson (2023).

In this study, we trial the following variations on hydrological and error models:

1. We trial a different calibration objective for GR4J: employing maximum likelihood and fitting the hydrological and transformation parameters jointly.
2. We use a 30-day bias-correction window for ERRIS.

#### Adaptations to forecast rainfall

Changing of the rainfall components is less straight forward. Climate forecasting systems, generated by coupled ocean-atmosphere-land-surface models, can take advantage of more persistent elements of the climate system – e.g., the land surface, sea-surface temperatures – to produce skilful forecasts of rainfall for the coming month (and sometimes more). They can, however, be negatively skilful (that is, worse than a seasonal climatology), and generally require post-processing. These methods exist; however, even when implemented, skill in seasonal precipitation forecasts often only contribute a small fraction of skill to streamflow forecasts (Bennett et al., 2017). In this study, therefore, we initially explore what skill is available purely from initial hydrological conditions and hydrological model errors. To do this, we generate so-called 'extended streamflow prediction' (ESP) forecasts.

ESP forecasts have been used for decades (Day, 1985), and work by forcing an initialised hydrological model with an ensemble of rainfall sequences sampled from the historical record. For this study we sample rainfall sequences from 1970-2023, using the same buffered leave-one-out cross-validation strategy described in Section 2.1.4. This results in a 50-member ensemble of 'forecast' rainfall. The rainfall forecasts do not have any information but are unbiased and correctly represent the uncertainty of historical rainfall. Streamflow forecast skill accordingly derives solely from initial hydrological conditions.

In the study by Bennett and Robertson (2023), only ESP rainfall forecasts and 'perfect' rainfall forecasts (i.e. observed rainfalls) were used. This is because even with the use of perfect rainfall forecasts, skill only extended to 7-8 days, indicating that forecast performance is constrained largely by the hydrological model. As we were able to improve the hydrological model performance substantially in this study, we have trialled the use of Numerical Weather Prediction (NWP) forecasts in the forecasting chain.

NWP precipitation forecasts were sourced from the European Centre for Medium Range Weather Forecasting (ECMWF) ensemble system (ECMWF-ens), retrieved from ECMWF's MARS operational archive. Operational ECMWF-ens forecasts are made up of 51 ensemble members – 1 control member, and 50 perturbed ensemble members. The number of parallel data retrieval requests from the MARS archive is limited (currently to 1), and as retrieving the control ensemble member requires a separate request to the perturbed ensemble members, we retrieved only the 50 perturbed ensemble members. We retrieved forecasts issued at 12:00 UTC every day for the period 1 Jan 2019 to 31 Dec 2022 (4 years). ECMWF-ens is widely recognised as the best performed medium range ensemble NWP, both in Australia and more widely.

ECMWF-ens forecasts are available at a 1 h time step for the first 90 h of the forecast, a 3 h timestep from 93-144 h, and a 6 h timestep from 150-360 h. To match this with observed precipitation (see Section 2.1.2) we first disaggregate forecasts for all lead times to hourly, disaggregating by simple division (i.e. 3 h accumulations were divided by 3, 6 h accumulations by 6). Both SILO and AGCD daily observed rainfalls are accumulated to 9:00 am (23:00 UTC the previous day). We therefore discard the first 11 lead times of each forecast, and aggregate the remaining hourly forecasts to daily. This leaves a 14-day daily precipitation forecast issued at 23:00 UTC/9:00 AEST, aligning with both SILO and AGCD.

NWP predictions usually suffer from bias and are often too narrow in ensemble spread, and thus require statistical processing. We pursue a simple post-processing strategy for this study, processing only the ensemble mean of ECMWF-ens. To produce 30-day forecasts, we augment the 14-day NWP forecasts with the ensemble mean of ESP precipitation forecasts (described above) for lead days 15-30. These predictions will not be skilful, but will reflect mean seasonal variations in climate. We correct bias and ensemble spread by processing the forecasts with the Bayesian Joint Probability (BJP) model (Wang & Robertson, 2011). The BJP has been widely used to produce seasonal streamflow (www.bom.gov.au/water/ssf) and precipitation forecasts (Schepen et al., 2020) and so we do not offer a detailed description here, but refer the reader to the source papers. BJP is highly effective at correcting bias and ensemble spread, and returns a climatology when skill is absent; the latter two properties are significant advantages over simpler bias-correction methods (Zhao et al., 2017).

We apply the BJP at each lead time and subarea under the cross-validation scheme described in Section 2.1.4. This means the BJP produces statistically reliable predictive distributions at each lead time and location, but these are not linked in time or space. We imbue the BJP-processed forecasts with realistic temporal and spatial correlations with the Schaake shuffle (Clark et al., 2004). We generate an ensemble of 200 members with the BJP.

Table 2 Streamflow gauge summary information

| RIVER | GAUGE NAME | GAUGE ID | CATCHMENT AREA (KM$^2$) | RECORD AVAILABLE |
|---|---|---|---|---|
| **Maranoa** | Maranoa River at Cashmere | 422404A | 18,778 | 1969-Present |
| **Balonne** | Balonne River at Weribone | 422213A | 51,495 | 1969-Present |
| **Balonne** | Balonne River at St George | 422201E/422201F | 75,446 | 1971-Present |

### 2.1.2 Hydrological modelling and data

#### Data

The rainfall runoff model used in this study requires rainfall and potential evaporation forcings and must be calibrated to streamflow observations. In the previous study by Bennett and Robertson (2023), daily rainfall data was retrieved from the Bureau of Meteorology's Australian Gridded Climate Data (AGCD; Evans et al., 2020). AGCD is interpolated from high quality gauges in the Bureau's network, but these do not always offer good coverage, which may increase the uncertainty of rainfall estimates over catchments. In this study we compare AGCD rainfall to another gridded rainfall product, SILO (Jeffrey et al., 2001). SILO was developed over Queensland, and appears to be interpolated from a denser gauge network than AGCD for the Balonne River catchment (Figure 3).

Potential evaporation data was taken from the Bureau of Meteorology's gridded Australian Water Resources Assessment Landscape model (AWRA-L; Frost et al., 2018). Streamflow data for three sites (Table 2, Figure 4) were extracted from the Bureau's Water Data Online database

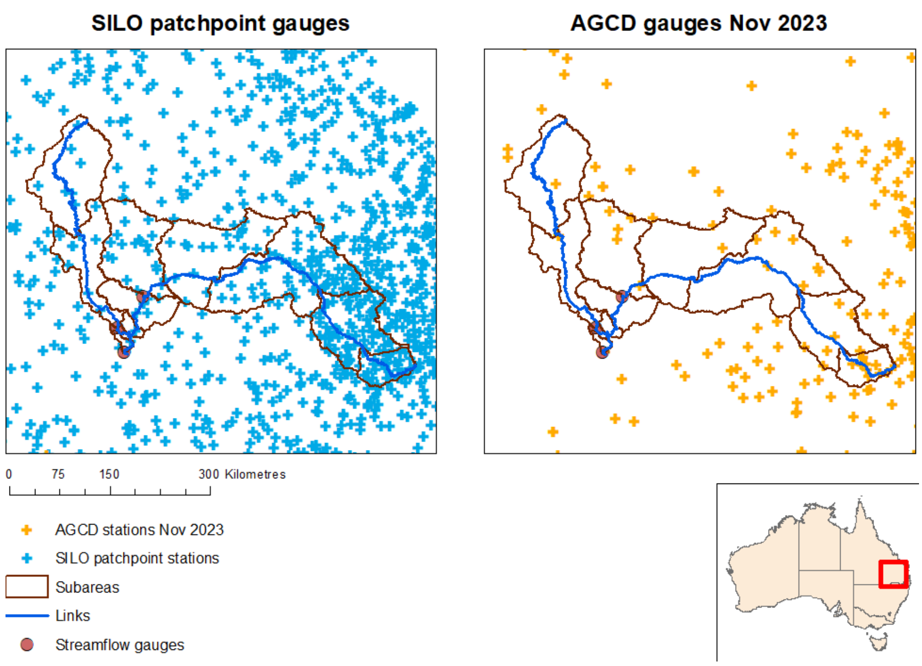


Figure 3 Comparison of gauge coverage for the SILO (left) and AGCD (right) interpolated datasets. Note that the AGCD gauge network may be denser than displayed: these gauges were used in Nov 2023, but a greater number is likely to have been used in previous years. SILO shown include all gauges used throughout the period, which likely overstates the number of gauges used at any particular time.



OFFICIAL

(http://www.bom.gov.au/waterdata/). While the Balonne River at St George site is the main site of interest for this study, we focus in this study on the two other sites for several reasons:

1) They are not directly impacted by a storage (in the St George's case, Beardmore Dam), which can have serious repercussions for error modelling
2) They have long, high quality records.
3) They provide a reasonable basis to prove forecasting concepts

#### Catchment delineation

We use the same delineation as Bennett and Robertson (2023) to model streamflow at Balonne River at St George. The catchment is divided into 13 subareas (Figure 4) to capture the east-to-west rainfall gradient and broken also at key gauges (Section 2.1.2).

The Condamine River is the major tributary of the Balonne in the eastern reaches of the catchment. The Condamine is heavily exploited for irrigation. The rainfall-runoff model only attempts to represent natural aspects of the catchment rather than irrigation extractions. In this initial study, we ignore the effects of irrigation extractions. This will likely lead to lower hydrological model performance than might be attained with better accounting of extractions. For this study this is justifiable because:

1) The rainfall-runoff model should still be able to model hydrological memory in the entire system
2) The extractions occur mainly in the upper reaches of the Balonne catchment; applying an error model at a lower gauge (at Balonne River at Weribone) corrects biases from headwater catchments

Similarly, we do not account for the function of Beardmore dam, which impounds the small storage of Lake Kajarabi just above the Balonne River at St George gauge. The dam can only store a small amount of water in comparison to the magnitude of flow in the catchment – its ~80 GL capacity can be exceeded by a single day's flow – which means it effectively functions as a weir. This means that it does not have a major impact on streamflow, particularly when we are interested in accumulations of large streamflow over several days, as in this application. While Lake Kajarabi's influence on streamflow is not negligible, for this preliminary study we assume it will not have a major impact on the predictability of multi-day large streamflow.

### 2.1.3 Generating hydrological simulations and forecasts with SWIFT

We retain the hydrological modelling setup from the SDES system: a semi-distributed model where runoff at each subarea is generated by the GR4J model and routed downstream using Lag-and-Route routing. The SDES uses a hybrid objective used in the SDES: Nash-Sutcliffe efficiency (NSE), bias and NSE of log-transformed flows (Hapuarachchi et al., 2022). However, we found in testing (not shown for brevity) that this objective did not perform as well as the more complex likelihood estimation method developed by Wang et al. (2020).

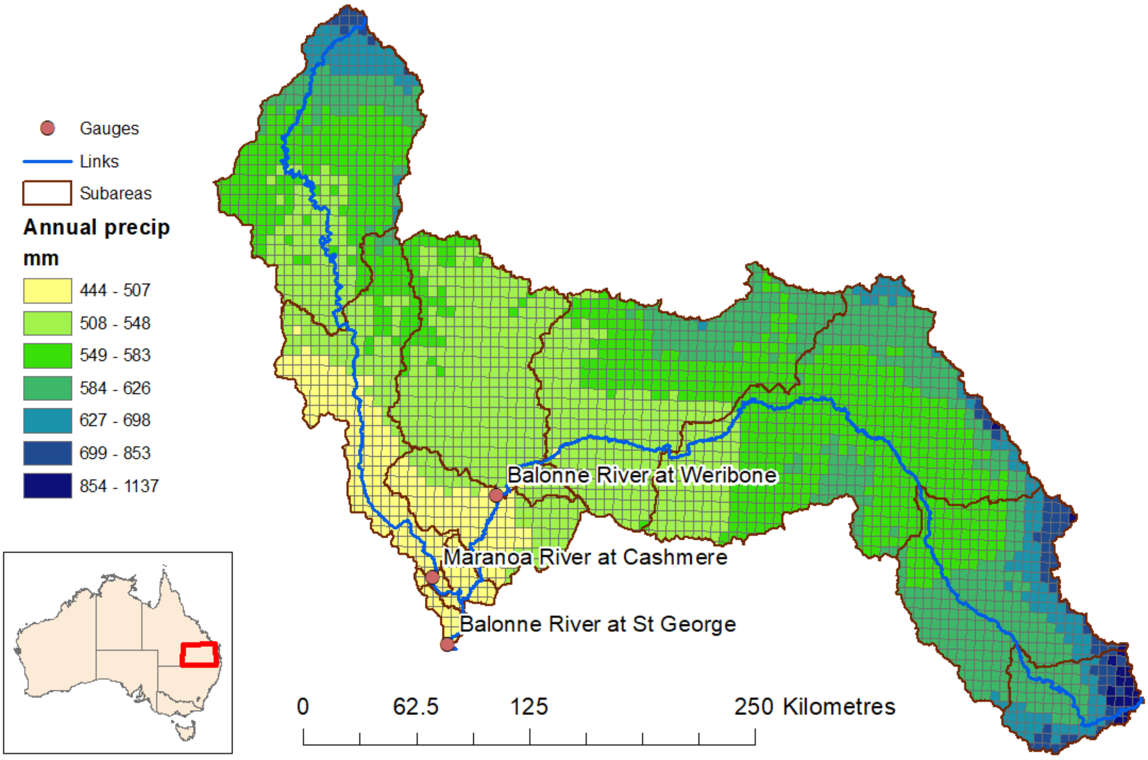


**Figure 4 Hydrological model catchment delineation. Annual mean precipitation is taken from the AGCD dataset.**

We assess both hydrological simulations, generated by forcing hydrological models with observed rainfall and potential evaporation, and forecasts, generated by forcing initialised hydrological models with 'forecast' rainfall and climatology. As noted above (Section 2.1.1), this study has 3 types of forecast rainfall:

1. **ESP forecasts:** An ensemble of historical rainfall and PE sequences (ESP forcings)
2. **Perfect forecasts:** To give an ideal counterpoint, we also generate streamflow forecasts with 'perfect' (observed) rainfall forecasts, allowing us to understand what could happen if perfectly accurate rainfall forecasts are available
3. **BJP-processed NWP:** Rainfall forecasts from the ECMWF-ens NWP, processed with the BJP.

To compare the modified SDES forecasts to other forecasting systems we generated retrospective forecasts for two periods:

1. 1982-2017, issued at the beginning of each calendar month. This allows direct comparison to the AWO forecasts
2. 2019-2022, issued every day. This allows assessment of the value of numerical weather prediction forecasts in the forecasting chain.

Note that these periods differ from Bennett and Robertson (2023), who used 1990-2020.

### 2.1.4 Cross-validation

To assess forecast and simulation performance rigorously, parameter estimation is carried out under strict cross-validation. For hydrological model parameters, the cross-validation scheme is a buffered leave-one-year out scheme with a buffer of one year. It is most easily illustrated by an example. When we wish to assess performance for 2010, we omit streamflow data from 2010 from the calibration, but also 2011. The trailing year (2011) year is also omitted because states from 2010 can influence flows in 2011. Therefore, if we include data from 2011, the calibration could be 'learning' from forcing data in 2011. This cross-validation is repeated for every year. Performance of both simulations and forecasts is assessed under this cross-validation.

For BJP parameter estimation, we use a simpler leave-one-year out cross-validation, as precipitation has much lower autocorrelation and memory than streamflow.

## 2.2 Australian Water Outlooks

The Australian Water Outlooks (AWO) are a gridded forecast product produced by the Bureau of Meteorology. AWO produces forecasts of several variables; in this study we focus only on runoff. The forecasts are produced using the following modelling chain:

1) Daily climate forecasts from the ACCESS-S climate forecasting system (Hudson et al., 2017)
2) Downscaling and bias-correction climate forecasts with quantile mapping
3) Land surface and Hydrological modelling with AWRA-L

AWO forecasts are produced to match the AGCD grid (~5 km). Retrospective forecasts are available from Australia's National Computing Infrastructure from 1981-2017. For this study, we make use of forecasts from 1990 onwards. To produce forecasts at a gauge, runoff from all grid cells upstream of a gauge are aggregated through simple averaging.

The AWO ensemble is constructed from the ACCESS-S ensemble: ACCESS-S generates a 3-member burst ensemble every day. For the AWO retrospective forecasts, 9 forecasts issued near the beginning of each month are aggregated, to create a time-lagged ensemble of 27 members. Forecasts are issued for 3-months, of which we use the first 30 days.

### 2.2.1 Statistical processing of AWO with the Bayesian Joint Probability model

AWO outputs are not designed to give accurate predictions at gauges; the Bureau uses it to predict deviations from internal climatology. While AWRA-L is calibrated to gauges, it features only one parameter set for all of Australia. This means we expect strong biases at individual gauges from AWRA-L. Further, AWRA-L does not attempt to route flows (other than implicitly, through lags in the AWRA-L rainfall-runoff algorithms), and thus we do not expect AWRA-L simulations or forecasts to be highly accurate.

To address these issues we process the AWO ensemble mean with the BJP. The method is very similar to that described for processing rainfall forecasts: we apply the BJP to each lead time and each location (gauge) independently. Streamflow tends to be conditionally autocorrelated – i.e., it is more autocorrelated on the falling limb of hydrographs than on the rising limb, and for this reason the Schaake Shuffle (which does not account for conditional autocorrelation) is unsuitable. Instead, to reassemble the temporal features of hydrographs, we take rank patterns from the AWO forecasts, following Bennett et al. (2022). This is called ensemble copula coupling (ECC), after Schefzik (2017).

## 2.3 Verification

For brevity, we focus on ensemble forecast verification of ensemble forecasts. Ensemble forecasts require specialised measures of performance scores, summarised in Table 3. The scores are calculated independently at each lead time, to illustrate how forecast performance varies with lead time. In addition to these scores, we assess performance with scatter plots comparing forecast quantiles with observations, as well as conventional visual inspection of forecast hydrographs.

In addition, we compare forecast performance to a simple climatology alternative. The climatology forecast is produced by sampling historical streamflow from matching ordinal dates from the period 1970-2023. We use the same buffered leave-one-year out cross-validation (Section 2.1.4) to generate these climatologies, resulting in a 50-member ensemble constructed from historical flows.

**Table 3. Performance scores used to assess hydrological and rainfall forecasts. Equations are described for hydrological forecasts.**

| Score | Description | Equation | Range |
|---|---|---|---|
| Continuous ranked probability score (CRPS) (Hersbach, 2000) | Measures forecast errors. Reduces to mean absolute error for deterministic forecasts. Units are the units of measurement (e.g., ML/day). | $CRPS = \frac{1}{T}\sum_{t=1}^{T}\int_{-\infty}^{\infty}\left(C(t,x) - H(q_o(t) \leq x)\right)^2 dx$<br>where:<br>$t = \{1,2,\ldots,T\}$ is time<br>$C(t,[\ \ ])$ $is$ the empirical cumulative distribution function (CDF) of the ensemble forecast at $t$<br>$H$ is the Heaviside step function | $\infty\ (worst)$ $to\ 0\ (best)$ |
| Threshold-weighted CRPS (Gneiting & Ranjan, 2011) | Like CRPS, but uses a weighting function to focus in on the area of the predictive CDF of interest. In our case, we focus on regions above high flow thresholds. | $twCRPS =$ $\frac{1}{T}\sum_{t=1}^{T}\int_{-\infty}^{\infty} w(x)\left(C(t,x) - H(q_o(t) \leq x)\right)^2 dx$<br>where:<br>$w(x)$ is a weighting function<br>We use the simple weighting function<br>$w(x) = \mathbb{1}(x > r)$ for a threshold $r$ | $\infty\ (worst)$ $to\ 0\ (best)$ |
| Probability integral transform (PIT) (Laio & Tamea, 2007) | Measures statistical reliability of ensemble spread. Forecasts are reliable when PIT values are uniformly distributed | $p(t) = \begin{cases} C\left(t, q_o(t)\right) & q_o(t) > 0 \\ U(0,1) \times C(t,0) & q_o(t) = 0 \end{cases}$<br>where:<br>$U$ is a random uniform number | $0\ to\ 1$ |
| $\boldsymbol{\alpha}$-index (Renard et al., 2010) | Summary statistic assessing uniformity of PIT values. | $\alpha = 1 - \frac{2}{T}\sum_{t=1}^{T}\lvert p(t) - p_U(t)\rvert$<br>where:<br>$p_U(t)$ is the theoretical value of $p(t)$ drawn from a uniform distribution | $1\ (best)$ $to\ 0\ (worst)$ |



OFFICIAL

# 3 Results

## 3.1 Sensitivity of forecast accuracy to different rainfall data sources

We show the accuracy of SWIFT streamflow forecasts using ESP rainfall forcing and 'perfect' (observed) rainfall forecasts from AGCD and SILO in Figure 5. In each case, the entire SDES modelling chain, including streamflow and error model parameter estimation, is carried out using each rainfall data source throughout. Forecasts are issued at the start of every month for the period 1982-2017.

For ESP rainfall forecasts – which, to reiterate, have no skill – the streamflow forecast accuracy is generally insensitive to the source of rainfall observations. This is not the case with 'perfect' rainfall forecasts: forecasts for the St George and especially Weribone gauges are more accurate with SILO. At St George, SILO improves forecast accuracy above AGCD for lead day 7 to lead day 12. We can draw two conclusions from this finding:

1. The spatial representation of rainfall in the wetter eastern hills in the SILO dataset materially improves forecasts
2. Improvements in forecasts due to rainfall representation only manifest to the extent that rainfall forecasts are accurate.

We note that for both rainfall datasets, forecasts are more skilful at St George than reported by Bennett and Robertson (2023). ESP Forecasts at St George are more accurate than climatology to >10 lead days, while Bennett and Robertson (2023) reported no skill beyond 7-8 days. This is explainable mainly by the change in the period over which forecasts have been assessed. In catchment with such variable flow regimes, forecast skill can be somewhat volatile. The longer

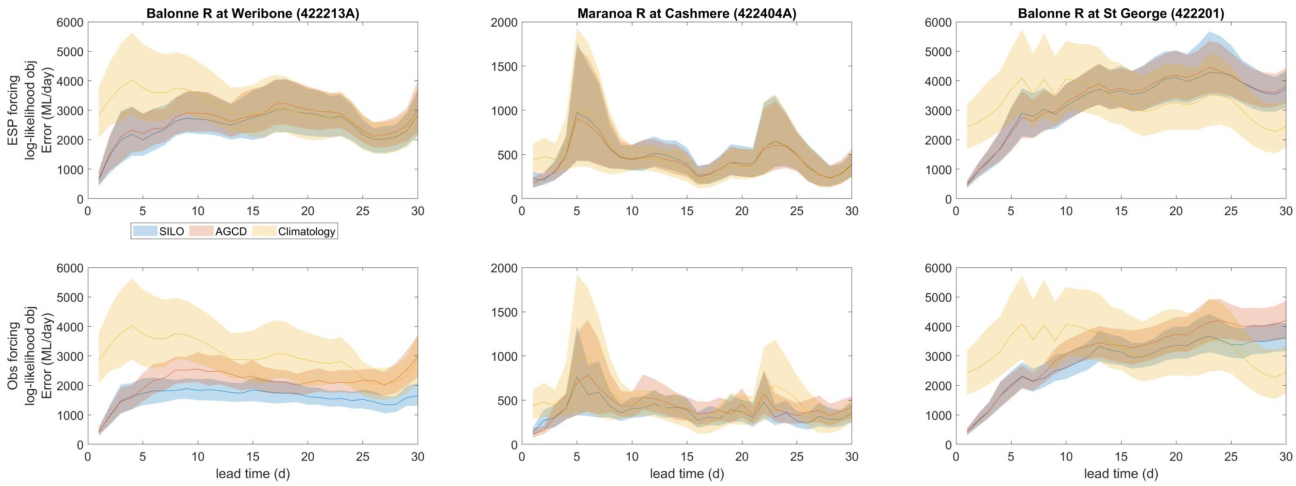


**Figure 5 Influence of rainfall dataset on forecast error (CRPS) for the period 1982-2017.**
**Columns show gauges, top row shows ESP forecasts, bottom row shows forecasts generated with 'perfect' rainfall predictions. Confidence intervals are generated by bootstrapping mean CRPS calculations with 200 repeats. SDES model parameters are estimated by minimising a log-likelihood.**

assessment period used in this study is likely to give a more robust estimate of forecast accuracy. We note that even with SILO and 'perfect' rainfall forecasts, the forecasting system is less accurate than a climatology prediction at the St. George gauge at lead times beyond ~15 days. This is still problematic; an ideal forecast system should at worst be similarly accurate to climatology. We now explore possible improvements to the SDES parameter estimation.

## 3.2 Improvements from hydrological model parameter estimation

We show the improvements to forecast accuracy that can be achieved with different objectives for calibrating the hydrological model. We compare three different objectives: two are variations of a log-likelihood, both described by Wang et al. (2020). The log-likelihood methods rely on the log-sinh transformation (Wang et al., 2012) to normalise data before hydrological model parameters are estimated. The methods differ in how the parameters of the log-sinh transformation are estimated, as follows:

1. **Log-likelihood**. Log-sinh parameters are first fitted to observed data. These parameters are fixed, and then hydrological model parameters are estimated.
2. **Log-likelihood joint**. Hydrological model parameters are estimated jointly with transformation parameters. Joint estimation theoretically allows the model more flexibility to fit a range of data, but can result in over-fitting (i.e., poorer performance outside the calibration period).

We also include the hybrid deterministic objective used by the Bureau in the SDES, which combines the Nash-Sutcliffe Efficiency (NSE), bias, and the NSE of log-transformed flow (Hapuarachchi et al., 2022) for comparison, which we term the *7-day* objective.

In the first instance, we present the effect of the rainfall-runoff model objective on forecasts generated with SILO rainfalls (Figure 6). For ESP forecasts, the performance of all objectives at the Weribone and Cashmere gauges is broadly similar. Consistent with Bennett and Robertson (2023), the log-likelihood objective outperforms the 7-day objective. This is most noticeable at lead times >7 days at the St George gauge. The log-likelihood joint objective produces clearly the most accurate forecasts at the St George gauge: notably, forecasts are always at least as accurate as climatology predictions. Interestingly, using perfect rainfall forecasts causes some issues with the log-likelihood joint objective at the Cashmere gauge: forecast errors become very large between lead days 15 and 30. Further investigation showed that this was due to a very large over-prediction in January 2008 (not shown for brevity). This shows the necessity of bootstrapping CRPS. The poor performance at the Maranoa gauge of the log-likelihood joint objective combined with perfect forcings propagates to the downstream St George gauge. Interestingly, this issue does not occur with the AGCD rainfalls, where the log-likelihood joint objective produces forecasts that are consistently as accurate as, or more accurate than, the other objectives (Figure 7). While the log-likelihood joint objective overall produces the most promising forecasts, the poor performance with perfect SILO rainfall forecasts points to the danger of overfitting with this objective.

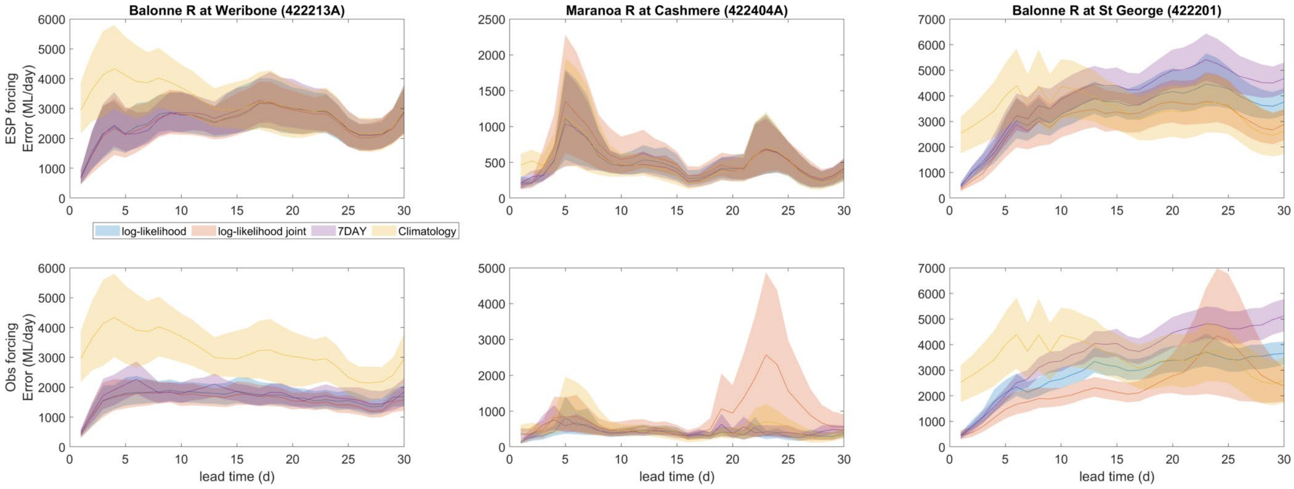


**Figure 6 Influence of rainfall-runoff model calibration objective function on forecast error (CRPS) with SILO forcing for the period 1982-2017.**
**Columns show gauges, top row shows ESP forecasts, bottom row shows forecasts generated with 'perfect' rainfall predictions. Confidence intervals are generated by bootstrapping mean CRPS calculations with 200 repeats.**

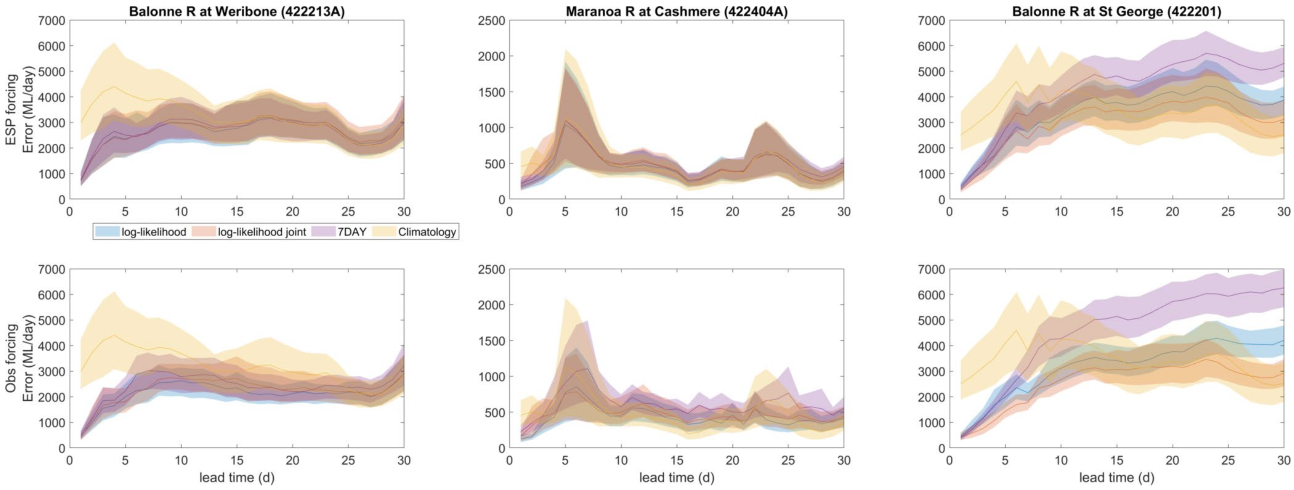


**Figure 7 Influence of rainfall-runoff model calibration objective function on forecast error (CRPS) with AGCD forcing for the period 1982-2017.**
**Columns show gauges, top row shows ESP forecasts, bottom row shows forecasts generated with 'perfect' rainfall predictions. Confidence intervals are generated by bootstrapping mean CRPS calculations with 200 repeats.**

## 3.3 Comparison of modified SWIFT forecasts to BJP-processed AWO

The BJP is extremely effective at correcting ensemble spread, and the statistically processed AWO forecasts are generally more reliable than ESP SWIFT forecasts (Figure 8). While SWIFT forecasts are highly reliable until lead day 5, reliability declines thereafter. This highlights the difficulty of propagating uncertainty through a forecast. Nonetheless, SWIFT forecasts are reasonably reliable even at lead day 30, with PIT-alpha values > 0.75. We note both the BJP-processed AWO forecasts and SWIFT predictions have the same number of ensemble members – PIT-alpha calculations are sensitive to differences in ensemble size.

However, SWIFT forecasts are markedly more accurate than BJP-processed AWO forecasts at shorter lead times at the Weribone and St George gauges (Figure 9). This finding holds also for

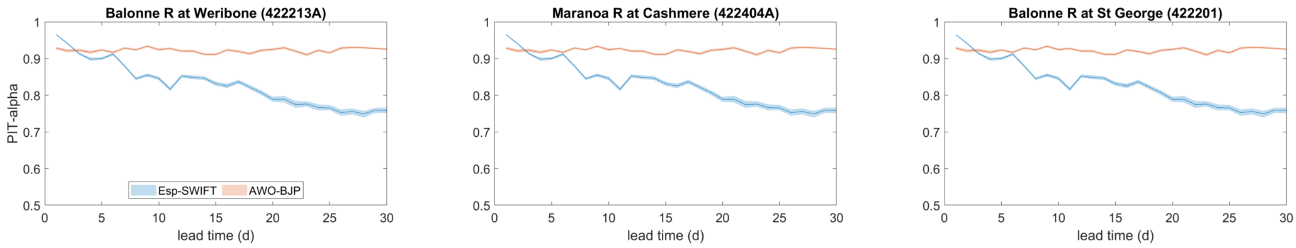


**Figure 8 Reliability of BJP-processed AWO and SWIFT ESP forecasts for the period 1982-2017. Columns show gauges. SWIFT is configured with the log-likelihood joint objective forced with SILO ESP forecasts.**

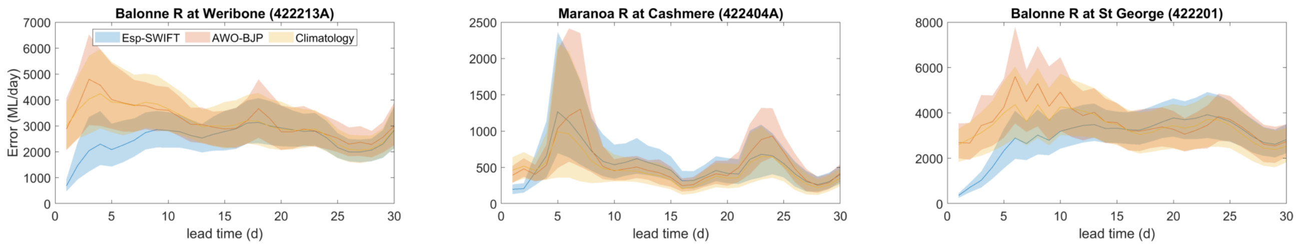


**Figure 9 Error (CRPS) of BJP-procssed AWO and SWIFT ESP forecasts (1982-2017). Columns show gauges. SWIFT is configured with the log-likelihood joint objective forced with SILO ESP forecasts.**

accumulated flow volumes (Figure 10). This demonstrates the advantage of a hydrological model that is tuned specifically to the catchment of interest, and particularly the advantages of error modelling over BJP-processing for streamflow predictions. The memory in streamflow allows error models to be highly effective at correcting forecasts at shorter lead times. Processing with the BJP does not take recent errors into account. We note that statistically processing AWO at the monthly time step is likely to produce more satisfactory results, as timing errors tend to cancel and it is easier for methods like the BJP to identify signal from noise. However, because the daily time step is crucial for this application, we conclude that SWIFT forecasts are far more likely to be beneficial for managing the Narran Lakes than daily BJP-processed AWO forecasts.

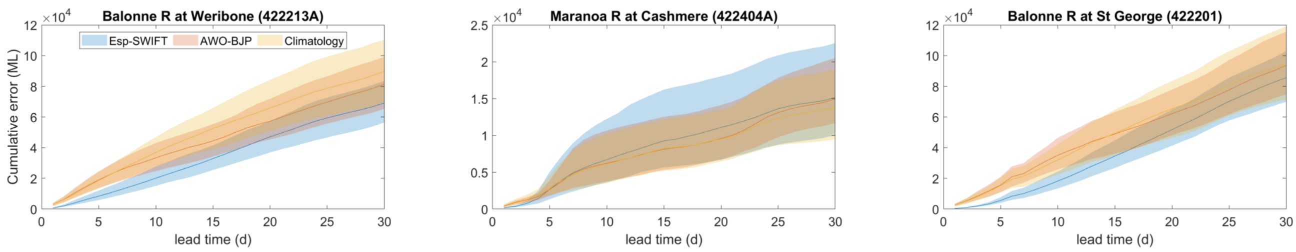


**Figure 10 Error (CRPS) of accumulated BJP-processed AWO and SWIFT ESP forecasts (1982-2017). Columns show gauges. SWIFT is configured with the log-likelihood joint objective forced with SILO ESP forecasts.**

## 3.4 Benefit of NWP forecasts in the SWIFT modelling chain

In the period for which we were able to assess retrospective forecasts (2019-2022), using BJP-processed ECMWF NWP forecasts in the modelling chain substantially increased the skill of forecasts compared with climatology (Figure 11). We note that SWIFT ESP forecasts are generally more skilful during this period than the 1982-2017 period used to compare SWIFT forecasts to BJP-processed AWO forecasts, again highlighting the volatility of forecast scores in these highly variable catchments. ESP forecasts at the St George gauge are skilful to lead day ~20, while for the 1982-2017 skill was positive to lead day ~15. Notwithstanding this volatility, including NWP forecasts in the modelling chain greatly improves forecast skill, notably from lead day 5 onwards. This is although NWP rainfall forecasts can only possibly be informative in the first 14 lead times (and are typically informative only in the first 7 days or so). It is possible to have skilful streamflow forecasts for much longer lead times because of the memory catchment that is represented in the SWIFT modelling system. Similarly, accumulated SWIFT forecasts generated with BJP-processed NWP forcings also tend to have much lower errors than ESP forecasts (as well as climatology predictions), as shown in Figure 12.

These findings hold true for forecasts of higher flow thresholds. We show forecast skill computed with the threshold weighed CRPS skill score for flow cumulative flow thresholds of 153, 250 and 500 GL, which we assume are averaged over the 30 days of the forecast period (Figure 13). These are the key accumulated volumes used as rules-of-thumb to predict sufficient inflows to the Narran Lakes for water bird breeding events and other environmental outcomes. Forecast skill declines with the magnitude of flow thresholds, particularly at the Cashmere gauge. However, forecast skill remains positive at the St George gauge for all flow thresholds when BJP-processed NWP forecasts are used in the SWIFT modelling chain.

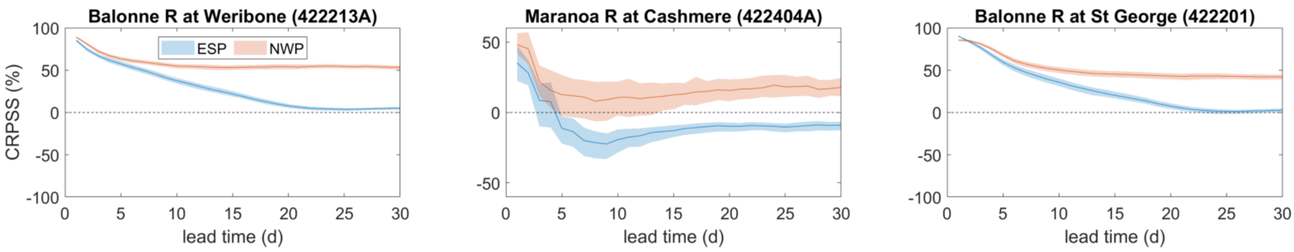


**Figure 11 Forecast skill compared to climatology of SWIFT forecasts forced with ESP (blue) and BJP-processed NWP (red) rainfalls assessed for the period 2019-2022.**
**Columns show gauges. SWIFT is configured with the log-likelihood joint objective with SILO rainfall observations.**

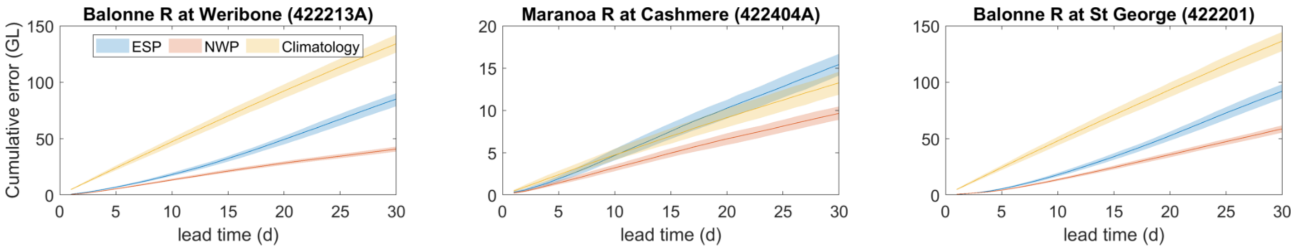


**Figure 12 Error (CRPS) of accumulated SWIFT forecasts forced with ESP (blue) and NWP (red) forecasts compared to climatology (yellow) for the period 2019-2022.**
**Columns show gauges. SWIFT is configured with the log-likelihood joint objective with SILO rainfall observations.**

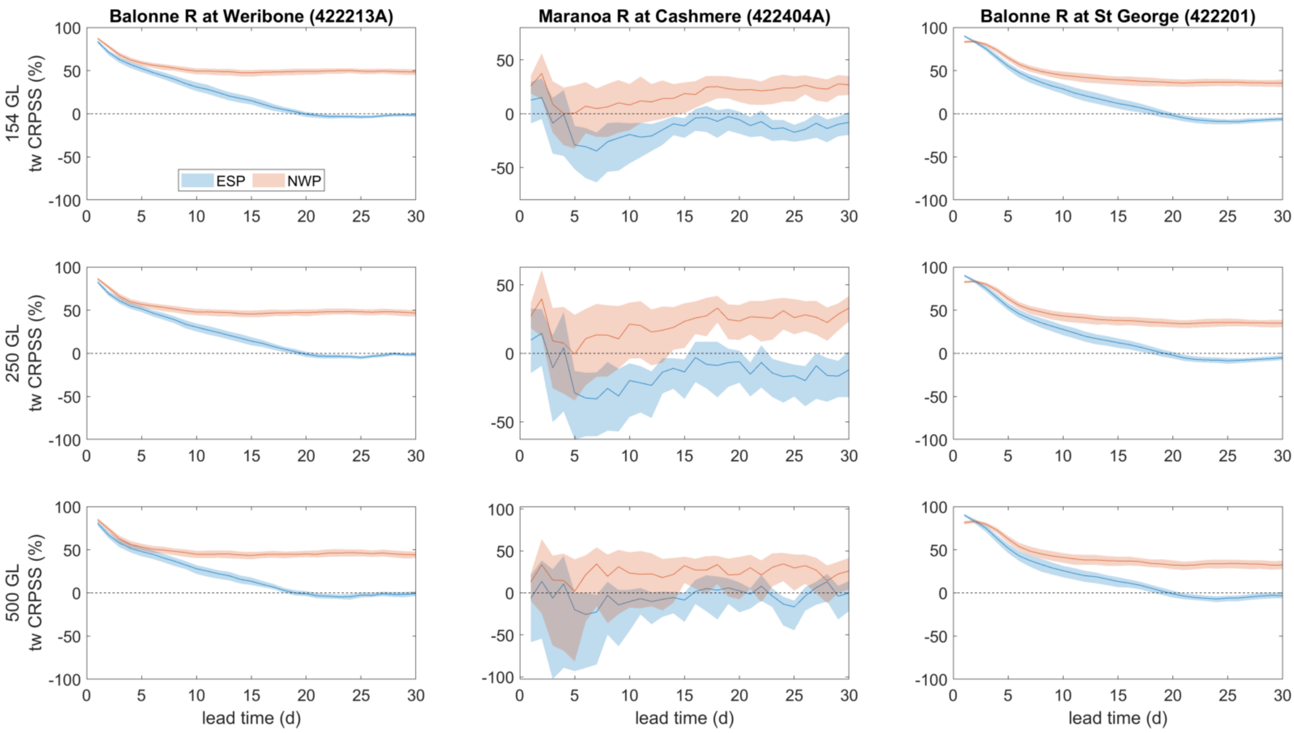


**Figure 13 Forecast skill computed with the threshold-weighted CRPS skill score compared to climatology of SWIFT forecasts forced with ESP (blue) and BJP-processed NWP (red) rainfalls assessed for the period 2019-2022.**
**Columns show gauges. Rows show cumulative thresholds of 154, 250 and 500 GL, averaged over 30 days. SWIFT is configured with the log-likelihood joint objective with SILO rainfall observations.**

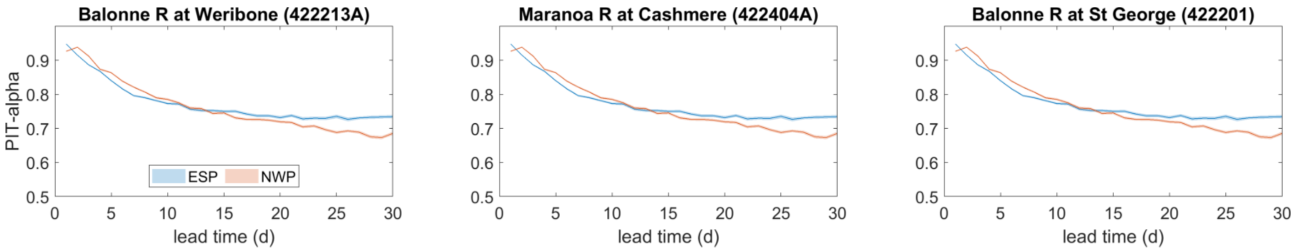


**Figure 14 Reliability of SWIFT forecasts forced with ESP (blue) and BJP-processed NWP (rainfall) forcings for the period 2019-2022.**
**Columns show gauges. SWIFT is configured with the log-likelihood joint objective forced with SILO rainfall observations.**

The only detraction for using BJP-processed NWP forecasts in the modelling chain is a very slight decline in reliability at longer lead times (Figure 14). However, this is only a minor trade-off for the substantial increase in forecast skill.

# 4 Discussion and summary

## 4.1 Discussion

We have demonstrated that the SWIFT forecasts – a modified version of the Bureau's SDES system – can produce skilful forecasts for Balonne River at St George gauge. Perhaps the most important modification was the way forecast rainfall has been treated in the forecast chain: the use of BJP-processed daily rainfall forecasts added substantial skill to the forecasts. We have chosen to statistically processed ensemble NWP forecasts from the skilful ECMWF-ens system for lead days 1-14, and augment these with reliable, but non-informative, climatology forecasts to extend lead times to 30 days. Because of the memory in the hydrological models (both the rainfall-runoff model and the error model), this results in streamflow forecasts that are skilful to long lead times. We note, however, that instead of these forcings, we could have used predictions from a coupled climate forecasting system like the Bureau's ACCESS-S2 or ECMWF's SYS5. Coupled models realise longer-range skill by explicitly predicting slower-varying aspects of the weather/climate system, notably the ocean and land-surface. While they are typically less accurate than NWP forecasts at shorter lead times, they may be more accurate than ESP forecasts at lead times 15-30 days. It is therefore possible that additional forecast skill can be gleaned by including seasonal climate prediction models for lead times 15-30 days (or beyond, if necessary).

We note that the strong skills shown for both ESP and NWP-forced streamflow forecasts during 2019-2022 may not be robust, as the climatology predictions may not have represented this period as well as the longer period (1982-2017) used to compare the BJP-processed AWO forecasts to SWIFT. What our results unequivocally demonstrated was the value of including informative precipitation forecasts in the modelling chain.

For daily streamflow forecasts, the SWIFT system outperformed BJP-processed AWO forecasts, even with ESP forcings. We reiterate that this finding may not hold at other time steps (e.g. monthly streamflow) where statistical processing with the BJP may better be able to isolate signal from noise. As daily forecasts are crucial to this application, our results point to the SWIFT approach being a better candidate for operationalisation.

We note that the SWIFT forecast system tested here is not operational. Both the BJP used to process NWP forecasts and the SWIFT software itself have been deployed operationally by the Bureau and elsewhere. The relatively simple methods used to forecast streamflow in this report – where no account is taken of extractions from the system – e.g. by irrigators in the Condamine catchment – nor of operation of Beardmore dam – are very amenable for operationalisation. Only real-time data feeds of observed precipitation and streamflow are required to run these forecasts in an operational setting. We note also that while SILO rainfalls resulted in more accurate forecasts than AGCD rainfalls, most likely because of improved representation of the eastern portion of the catchments due to a denser rain gauge network, a smaller number of rain gauges may be available to generate rainfall surfaces in an operational setting, reducing forecast performance.

Should such a system be operationalised, improved representation of water extractions and the operation of Beardmore dam may further improve the quality of predictions.

## 4.2 Summary and recommendations

We have shown that with sufficient modification to the SDES system, SWIFT forecasts are skilful to at least 15 days, and may be skilful beyond 30 days. Skill is evident generally, and also for forecasts of high flows. The SWIFT forecasts outperformed the BJP-processed AWO forecasts at the daily time step. Accordingly, our main findings are as follows:

1. Modified SDES forecasting methods, termed 'SWIFT forecasts' in this report, produce forecasts that promise enough skill to inform more effective management of large inflow events into the Narran Lakes. A crucial aspect of this modification was jointly estimating transformation and hydrological model parameters with a likelihood.
2. Skilful rainfall forecasts are a key component of the modelling chain, and should be included in any attempts to operationalise forecasts
3. The forecasts trialled here are not in operation. All methods used in this report are encoded in operations grade software, allowing expedited operationalisation.

# References


Bennett, J. C., Robertson, D. E., & CSIRO. (2023). *Investigating skill of streamflow forecasts to assist management of the Narran Lakes (Dharriwaa)*. CSIRO Environment.

Bennett, J. C., Robertson, D. E., Wang, Q. J., Li, M., & Perraud, J.-M. (2021). Propagating reliable estimates of hydrological forecast uncertainty to many lead times. *Journal of Hydrology*, 126798. https://doi.org/10.1016/j.jhydrol.2021.126798

Bennett, J. C., Schepen, A., Robertson, D. E., & CSIRO. (2022). *SLEIP streamflow outlooks*. CSIRO.

Bennett, J. C., Wang, Q. J., Robertson, D. E., Schepen, A., Li, M., & Michael, K. (2017). Assessment of an ensemble seasonal streamflow forecasting system for Australia. *Hydrol. Earth Syst. Sci.*, *21*(12), 6007–6030. https://doi.org/10.5194/hess-21-6007-2017

Clark, M. P., Gangopadhyay, S., Hay, L., Rajagopalan, B., & Wilby, R. (2004). The Schaake shuffle: a method for reconstructing space–time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, *5*, 243–262. https://doi.org/10.1175/1525-7541(2004)005<0243:TSSAMF>2.0.CO;2

Day, G. N. (1985). Extended streamflow forecasting using NWSRFS. *Journal of Water Resources Planning and Management*, *111*(2), 157–170. https://doi.org/10.1061/(ASCE)0733-9496(1985)111:2(157)

Evans, A., Jones, D., Smalley, R., Lellyett, S., & Meteorology, B. of. (2020). An enhanced gridded rainfall analysis scheme for Australia. In *Bureau Research Report* (Issue BRR041). Bureau of Meteorology. http://www.bom.gov.au/research/publications/researchreports/BRR-041.pdf

Frost, A. J., Ramchurn, A., & Smith, A. (2018). The Australian Landscape Water Balance model (AWRA-L v6). Technical Description of the Australian Water Resources Assessment Landscape model version 6. In *Bureau of Meteorology Technical Report*. Bureau of Meteorology. http://www.bom.gov.au/water/landscape/assets/static/publications/AWRALv6_Model_Description_Report.pdf

Gneiting, T., & Ranjan, R. (2011). Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules. *Journal of Business & Economic Statistics*, *29*(3), 411–422. https://doi.org/10.1198/jbes.2010.08110

Hapuarachchi, H. A. P., Bari, M. A., Kabir, A., Hasan, M. M., Woldemeskel, F. M., Gamage, N., Sunter, P. D., Zhang, X. S., Robertson, D. E., Bennett, J. C., & Feikema, P. M. (2022). Development of a national 7-day ensemble streamflow forecasting service for Australia. *Hydrol. Earth Syst. Sci.*, *26*(18), 4801–4821. https://doi.org/10.5194/hess-26-4801-2022

Hersbach, H. (2000). Decomposition of the Continuous Ranked Probability Score for ensemble prediction systems. *Weather and Forecasting*, *15*(5), 559–570. https://doi.org/10.1175/1520-0434(2000)015<0559:dotcrp>2.0.co;2

Hudson, D., Alves, O., Hendon, H. H., Lim, E.-P., Liu, G., Luo, J.-J., MacLachlan, C., Marshall, A. G., Shi, L., Wang, G., Wedd, R., Young, G., Zhao, M., & Zhou, X. (2017). ACCESS-S1: The new Bureau of Meteorology multi-week to seasonal prediction system. *Journal of Southern Hemisphere Earth Systems Science*, *67*(3), 132–159. https://doi.org/10.22499/3.6703.001

Jeffrey, S. J., Carter, J. O., Moodie, K. M., & Beswick, A. R. (2001). Using spatial interpolation to construct a comprehensive archive of Australian climate data. *Environmental Modelling & Software*, *16*, 309–330. https://doi.org/10.1016/S1364-8152(01)00008-1

Laio, F., & Tamea, S. (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences*, *11*(4), 1267–1277. https://doi.org/10.5194/hess-11-1267-2007

Li, M., Wang, Q. J., Bennett, J. C., & Robertson, D. E. (2016). Error reduction and representation in stages (ERRIS) in hydrological modelling for ensemble streamflow forecasting. *Hydrology and Earth System Sciences*, *20*(9), 3561–3579. https://doi.org/10.5194/hess-20-3561-2016

Perraud, J.-M., Bridgart, R., Bennett, J. C., & Robertson, D. E. (2015). SWIFT2: High performance software for short-medium term ensemble streamflow forecasting research and operations. In T. Weber, M. J. McPhee, & R. S. Anderssen (Eds.), *MODSIM2015, 21st International Congress on Modelling and Simulation* (pp. 2458–2464). Modelling and Simulation Society of Australia and New Zealand. www.mssanz.org.au/modsim2015/L15/perraud.pdf

Perrin, C., Michel, C., & Andréassian, V. (2003). Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology*, *279*, 275–289. https://doi.org/10.1016/S0022-1694(03)00225-7

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., & Franks, S. W. (2010). Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research*, *46*(5), W05521. https://doi.org/10.1029/2009wr008328

Schefzik, R. (2017). Ensemble calibration with preserved correlations: unifying and comparing ensemble copula coupling and member-by-member postprocessing. *Quarterly Journal of the Royal Meteorological Society*, *143*(703), 999–1008. https://doi.org/10.1002/qj.2984

Schepen, A., Everingham, Y., & Wang, Q. J. (2020). On the Joint Calibration of Multivariate Seasonal Climate Forecasts from GCMs. *Monthly Weather Review*, *148*(1), 437–456. https://doi.org/10.1175/mwr-d-19-0046.1

Wang, Q. J., Bennett, J. C., Robertson, D. E., & Li, M. (2020). A Data Censoring Approach for Predictive Error Modeling of Flow in Ephemeral Rivers. *Water Resources Research*, *56*(1). https://doi.org/10.1029/2019WR026128

Wang, Q. J., & Robertson, D. E. (2011). Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences. *Water Resources Research*, *47*, W02546. https://doi.org/10.1029/2010WR009333

Wang, Q. J., Shrestha, D. L., Robertson, D. E., & Pokhrel, P. (2012). A log-sinh transformation for data normalization and variance stabilization. *Water Resources Research*, *48*, W05514. https://doi.org/10.1029/2011WR010973.

Zhao, T., Bennett, J. C., Wang, Q. J., Schepen, A., Wood, A. W., Robertson, D. E., & Ramos, M.-H. (2017). How suitable is quantile mapping for post-processing GCM precipitation forecasts? *Journal of Climate*, *30*(9), 3185–3196. https://doi.org/10.1175/jcli-d-16-0652.1

**Contact us**

1300 363 400
+61 3 9545 2176
csiro.au/contact
csiro.au

**For further information**

Environment
Dr James Bennett
+61 3 9545 2462
james.bennett@csiro.au
csiro.au/en/about/people/business-units/Environment