A virtual hydrological framework for evaluation of stochastic rainfall models

Bree Bennett[1], Mark Thyer[1], Michael Leonard[1], Martin Lambert[1], Bryson Bates[2]

[1]School of Civil, Environmental and Mining Engineering, University of Adelaide, North Terrace Campus, 5005, South Australia
[2] School of Agriculture and Environment, The University of Western Australia, Crawley, 6009, Western Australia

Correspondence to: Bree Bennett (bree.bennett@adelaide.edu.au)

Abstract. Stochastic rainfall modelling is a commonly used technique for evaluating the impact of flooding, drought or climate change in a catchment. While considerable attention is given to the development of stochastic rainfall models, significantly less attention is given to performance evaluation methods. Typical evaluation methods employ a variety of rainfall statistics. However, they give limited understanding about which rainfall characteristics are most important for reliable streamflow prediction whenever the simulated rainfall are poor. To address this issue a new evaluation method for rainfall models is introduced, with three key features: (i) streamflow-based — to give a direct evaluation of modelled streamflow performance, (ii) virtual — to avoid the issue of confounding errors in hydrological models or data, and (iii) targeted — to isolate the source of errors according to specific sites and months. The virtual hydrologic evaluation framework is applied to a case study of 22 sites in South Australia. The framework demonstrated that apparently 'good' modelled rainfall can produce 'poor' streamflow predictions, whilst 'poor' modelled rainfall may lead to 'good' streamflow predictions, as catchment processes can dampen or amplify rainfall errors when converted to streamflow. The framework identified the importance of rainfall in the 'wetting-up' months of the catchment cycle (May and June in this case study) for providing reliable predictions of streamflow over the entire year despite their low monthly flow volume. This insight would not have been found using existing methods and highlights the importance of the virtual hydrological evaluation framework for stochastic rainfall model evaluation.

1 Introduction

Stochastic rainfall model simulations are primarily used as inputs to a hydrological model, for simulating realisations of streamflow. Streamflow simulations are then used to assess hydrological risks, such as flood risks (Camici et al., 2011; Li et al., 2016) or drought risks (Henley et al., 2013; Mortazavi-Naeini et al., 2015; Paton et al., 2013). When evaluating the efficacy of stochastic rainfall models, current approaches, which make comparisons to observed rainfall and/or streamflow, are limited. This is because they are unable to make a targeted evaluation of the stochastic rainfall model's ability to reproduce streamflow characteristics of practical interest. This paper introduces a new virtual framework that enables targeted, hydrological evaluation of stochastic rainfall models.

Observed-rainfall evaluation is the most common method of evaluation and relies on comparisons between observed and simulated rainfall statistics (Baxevani and Lennartsson, 2015; Bennett et al., 2018; Evin et al., 2018; Rasmussen, 2013; Srikanthan and Pegram, 2009; Wilks, 2008). Typically, this method shows 'mixed' performance where many statistics are reproduced well, but some are poor and the implications for improving streamflow simulations are not clear. To overcome these limitations, an alternative is to evaluate the rainfall model's performance in terms of streamflow, hereafter referred to as 'observed-streamflow evaluation' (Blazkova and Beven, 2002, 2009; Camici et al., 2011; McMillan and Brasington, 2008). A challenge with observed-streamflow evaluation is that when there is poor predictive performance, it is difficult to ascertain if the poor performance was caused by the hydrological model or the stochastic rainfall model. This is because hydrological model performance can vary substantially, with the possibility of significant data errors (Andreassian et al., 2001; Kuczera and Williams, 1992; McMillan et al., 2010), and/or model structural errors (Clark et al., 2008; Renard et al., 2010; Renard et al., 2011; Smith et al., 2008).

There is the possibility of a different approach to evaluation, involving the use of virtual experiments, which has not previously been developed for the evaluation of stochastic rainfall models. To date, virtual experiments have been used in a variety of contexts, including (i) the evaluation of hydrological model sensitivity (Ball, 1994; Nicótina et al., 2008; Paschalis et al., 2013; Shah et al., 1996; Wilson et al., 1979), (ii) evaluating how well simpler conceptual hydrological models characterise complex processes (Li et al., 2015b), and (iii) for developing new techniques for flood frequency analysis (Li et al., 2014; 2016). Stochastic rainfall models can be a significant source of error in streamflow simulation, yet virtual experiments have not been designed to identify these deficiencies. An evaluation technique is needed that can identify deficiencies in simulated rainfall while retaining a focus on streamflow.

This paper introduces a virtual hydrological framework for evaluating stochastic rainfall models. The key difference between this framework and conventional frameworks is the introduction of a new baseline. Rather than use observed streamflow, the hydrological model can be used to translate observed rainfall to corresponding streamflow (conditioned on that hydrological model), referred to here as 'virtual-observed streamflow'. The framework overcomes the limitations of observed-rainfall evaluation because it evaluates the performance of the stochastic rainfall model in terms of the key streamflow characteristics of interest. Likewise, the framework overcomes the limitations of observed-streamflow evaluation because it can separate out the sources of error in the comparison of streamflows.

The key objectives of this paper are:

1. To introduce a framework for virtual hydrological evaluation of stochastic rainfall models.
2. To introduce two different tests which are part of the framework: an integrated test and a unit test. Combined use of these tests allows streamflow discrepancies to be attributed to their original source in the rainfall model according to site and month.
3. To demonstrate the framework evaluation on a stochastic rainfall model and contrast the outcomes with conventional evaluation methods.

The framework of virtual hydrologic evaluation is explained in Section 2.2 with the procedures for the integrated test and unit test outlined in Section 2.3. Stochastic rainfall models have been developed for 22 sites in the Onkaparinga catchment, South Australia (Section 3) and are used to illustrate the procedure (Section 4). Discussion and conclusions emphasize the features of the framework and the different recommendations it can identify for improving the rainfall model (Sections 5 and 6).

2 Virtual hydrological evaluation framework

2.1 Limitations of existing evaluation techniques

Observed-rainfall evaluation involves comparisons between observed and simulated rainfall statistics, as shown in Fig. 1(a). This method involves a large number of evaluation statistics and it is often unclear which are the most important and how trade-offs in respective performance between them might be evaluated. For the example of a daily stochastic rainfall model, these statistics might include the mean and standard deviation of: rainfall on wet days, wet-spell and dry-spell durations, wet-day proportions, extreme values, and daily total rainfall. This basic list would involve 12 daily-scale statistics per site per month. Aggregated statistics might include the mean, standard deviation and autocorrelation at aggregated monthly and annual time scales, producing another 39 statistics per site. For a multi-site rainfall model having 10 gauges, the evaluation would involve 1830 temporal statistics before considering the statistics of spatial aggregates.

Typically, observed-rainfall evaluation shows ‘mixed’ performance where many statistics are reproduced well but some are poor which presents a number of challenges for evaluating predictive performance. Firstly, it is difficult to ascertain if the rainfall model’s performance is sufficient in terms of predictions of practical interest, which are typically streamflow-based. Secondly, it is unclear if it is necessary to invest time and effort to address instances of poor performance, when the majority of statistics are well reproduced. For example, Bennett et al. (2018) found that the standard deviation of the number of wet days varied monthly and by site (monthly performance varied between January approximately 100% ‘poor’ and July 100% ‘good’). As another example, Evin et al. (2018) showed good reproduction of a wide range of daily scale statistics but that model variants differed in terms of their ability to reproduce 3-day extremal totals. While these assessments are useful, they are nonetheless a step removed from their ability to reproduce streamflow characteristics.

To overcome limitations in observed-rainfall evaluation methods, a conventional alternative is to evaluate the rainfall model’s performance in terms of streamflow. This method is indicated in Fig. 1 (b) and is referred to as ‘observed-streamflow evaluation’. Rainfall modelling is not an end in itself, therefore hydrological statistics of interest (such as properties of the flow duration curve as well as annual and monthly total flows), can give better perspective on the suitability of the rainfall model for its intended application (see Cowpertwait, 2006). For example, observed-streamflow evaluation has been used to predict the annual flood frequency distribution (Blazkova and Beven, 2002, 2009; Camici et al., 2011; McMillan and Brasington, 2008).

© Author(s) 2018. CC BY 4.0 License.

A challenge with observed-streamflow evaluation is that when there is poor predictive performance (i.e. a significant discrepancy between the observed and predicted streamflow) it is difficult to ascertain if the poor performance was caused by the hydrological model or the stochastic rainfall model. From Fig. 1 (b), observed-streamflow evaluation typically involves (1) a stochastic model that produces simulations of rainfall, that are (2) input to a hydrological model to produce simulated streamflow, which is (3) converted to the predictions of interest (e.g. the flood frequency distribution), and (4) compared against the observed streamflow predictions of interest. For example, Camici et al. (2011) used this technique to show that observed discharge peaks were matched by synthetic design storm simulations. However, hydrological model performance can vary substantially. One reason is due to data errors, for example, in their analysis of streamflow data errors, Coxon et al. (2015) found that for mean and high flows, over 80% of gauging stations had a 95% uncertainty interval of 40% of the flow. Another reason is model structural errors, as even hydrological models with ‘good’ performance can have predictive errors with a standard deviation at 25% of the predicted streamflow (Evin et al., 2014). In some ephemeral catchments, the predictive uncertainty has a precision (standard deviation) of 40% of mean observed streamflow (McInerney et al., 2017). These predictive errors can be substantially worse where there are significant rainfall errors (Andreassian et al., 2001; Kuczera and Williams, 1992; Renard et al., 2011).

With predictive errors of this magnitude it is difficult to evaluate the ability of a stochastic rainfall model to produce the key streamflow predictions of interest using observed-streamflow evaluation. This is because when the simulated streamflow produced by the combination of a stochastic rainfall model and a hydrological model is in error, it is difficult to determine if the error is due to a deficiency in the stochastic rainfall model or predictive errors within the hydrological model. This means it is unclear whether or not the stochastic rainfall model should be improved, or how to do so to improve key streamflow predictions of interest.

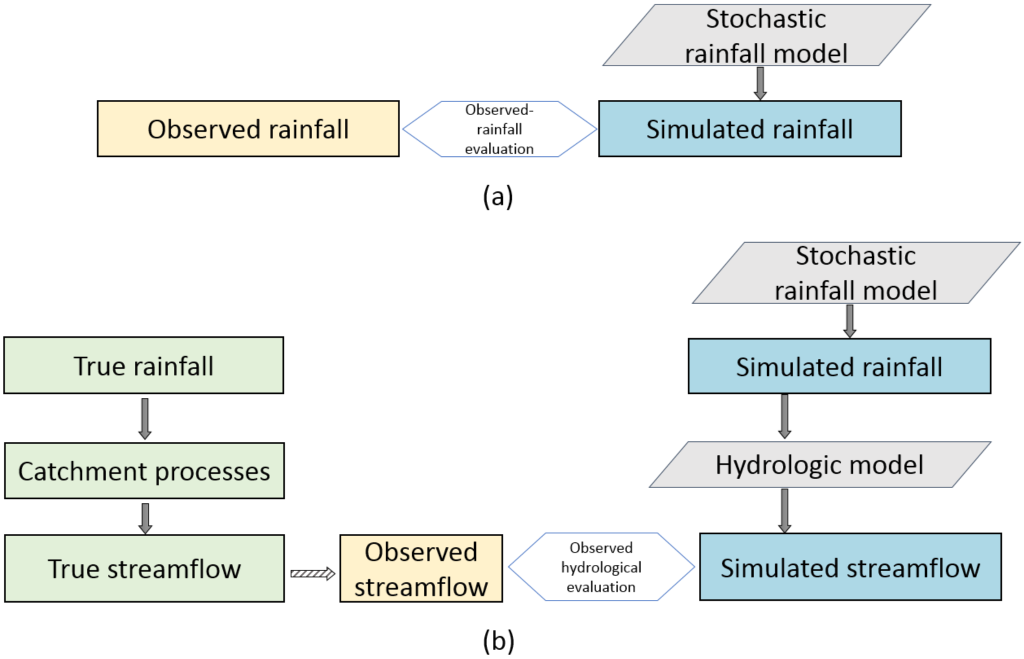


Fig. 1 Schematic of (a) observed-rainfall evaluation where simulated rainfall is compared against observed rainfall and (b) observed-streamflow evaluation where simulated streamflow is compared against observed streamflow.

2.2 Overview of the virtual hydrological evaluation framework

The virtual hydrological evaluation framework is designed to focus on predictions of interest, similar to observed-streamflow evaluation, but to reduce the sources of error to only those introduced by the stochastic rainfall model. Table 1 gives an overview of the sources of error for the three evaluation frameworks. The observed-rainfall evaluation framework is used to identify errors in the rainfall model, but is not able to determine their implications for streamflow. The observed-streamflow framework provides an absolute measure of performance, since ultimately the goal is to match streamflow observations. However, with this approach it is not possible to readily identify whether discrepancies in the simulated streamflow are attributed to the rainfall model, the streamflow observations or poor process representation within the hydrological model. In contrast, the virtual framework is a relative measure of performance, where the hydrological model is a common factor in the production of simulated streamflow and virtual-observed streamflow (elaborated on below) that is used as a baseline for comparison. By using a virtual baseline, observed streamflow is not directly required in the evaluation as both simulated and observed rainfall undergo transformation by the same process representation (i.e. the hydrological model). This enables discrepancies in the streamflow to be identified in terms of features of the stochastic rainfall model.

Table 1 Comparison of the sources of error for observed-rainfall, observed-streamflow and virtual hydrological based evaluation frameworks.

Evaluation framework	Stochastic rainfall	Hydrological model	Observed streamflow
Observed-rainfall	✓	×	×
Observed-streamflow	✓	✓	✓
Virtual hydrological	✓	-	-

As shown in Fig. 2, a virtual hydrological evaluation involves the comparison of simulated streamflow statistics to virtual-observed streamflow statistics, defined as:

- Simulated streamflow — is streamflow produced by the hydrological model by inputting simulated rainfall at a given site.
- Virtual-observed streamflow — is streamflow produced by the hydrological model by inputting observed rainfall at the same given site.

The virtual framework undertakes a relative assessment of the simulated and observed rainfall after its transformation by the same hydrological model to provide insight into the performance of the stochastic rainfall model. Because the hydrological evaluation is a relative comparison of the observed and simulated rainfall, it is important that all other parameters and extraneous variables (e.g. potential evapotranspiration) relating to the hydrological model are kept the same in all instances. This is because the ultimate goal of the modelling process remains the same: to match observed streamflow for a catchment of interest. Therefore, the virtual evaluation framework is best used to augment and complement existing evaluation methods, rather than act as a replacement. The three evaluation frameworks could work together as follows, where: (i) observed-rainfall evaluation identifies any deficiencies in the stochastic rainfall model prior to any hydrological considerations; (ii) the virtual hydrological framework identifies which of these rainfall deficiencies impact on the key predictions of interest, that is, simulated streamflow; and (iii) observed-streamflow evaluation provides a final validation. Therefore, together they enable a more focused approach to identify opportunities for improvement of a stochastic rainfall model.

© Author(s) 2018. CC BY 4.0 License.

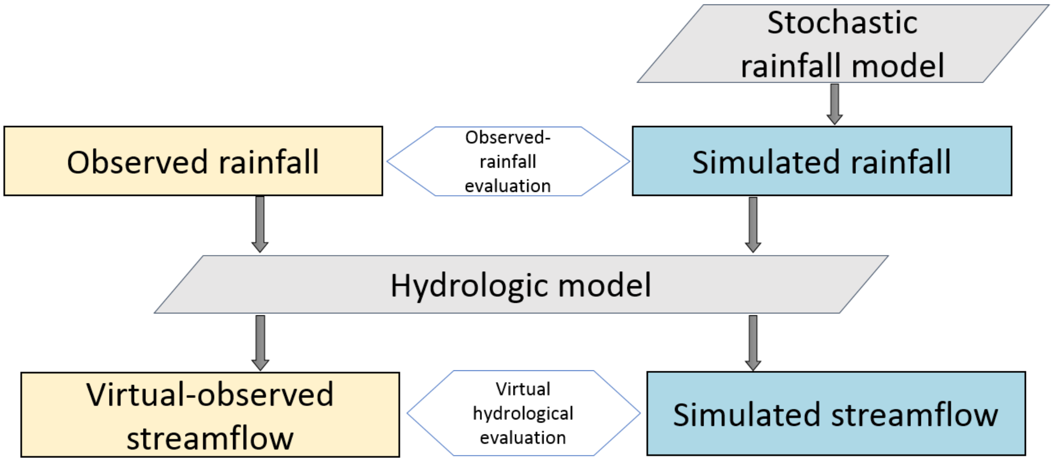

Stochastic rainfall model
Observed rainfall
Observed-rainfall evaluation
Simulated rainfall
Hydrologic model
Virtual-observed streamflow
Virtual hydrological evaluation
Simulated streamflow

Fig. 2 Schematic of the virtual hydrologic evaluation framework where simulated streamflow is compared against virtual-observed streamflow.

2.3 Diagnostic tests enabled by virtual hydrological framework

The virtual framework enables several techniques for isolating the deficiencies in simulated rainfall on the key streamflow predictions of interests, including an 'integrated test' that isolates issues for a given site, and 'unit tests' that isolate issues for specific time periods.

2.3.1 Integrated tests

The integrated test acts as an overall test of the stochastic rainfall model's performance. The integrated test proceeds for a single location of interest by transforming the time series of observed and simulated rainfall, via the hydrological model (Fig. 2). The hydrological model evaluates the effects of rainfall over longer timescales due to the integrating property of its conceptual catchment stores. If there is a discrepancy in the distribution of the simulated streamflow and the virtual-observed streamflow, this indicates that the simulated rainfall at that site is deficient.

Due to the virtual nature of this evaluation, it is not bound by a comparison to observed streamflow sites and avoids the requirement to construct estimates of catchment rainfall. Without a virtual approach, the framework would be unable to evaluate each rainfall site separately, as necessary for isolating rainfall model deficiencies. The application of an integrated test at individual sites is the conceptual equivalent of saying that single site rainfall represents the average rainfall for that catchment, which is a situation not unlike many hydrological studies that rely on one or two rain gauges.

As a further integrated test, it is possible to use the catchment average rainfall. If all individual sites show minimal discrepancies, but the catchment average test shows a discrepancy, this indicates an issue in the spatial dependence structure of the rainfall model rather than the at-site statistics. Repeating an integrated test for all sites and for catchment average(s), it is possible to diagnose whether specific locations or the spatial dependence causes poor reproduction of streamflow statistics.

2.3.2 Unit tests

While an integrated test is able to isolate deficiencies in stochastic rainfall for an individual site, it is unable to identify which specific time period(s) the issues are related to, due to storage properties of the hydrological model. For this reason, an additional test is introduced to isolate time periods and is referred to as a 'unit test'. The unit test investigates the impact of simulated rainfall in a given influencing month on the production of streamflow in an evaluated month of interest. This is achieved by splicing observed and simulated rainfall into a single time series which is used to produce simulated streamflow.

Following Fig. 3, consider the time series of observed, R^{obs}, and simulated, R^{sim} , daily rainfall for each year (and replicate) at a given site. Fig. 3(a) illustrates the embedding of simulated rainfall R_k^{sim} in an influencing month, k, within observed rainfall R_m^{obs} for all other months $m \in \{1, \ldots, 12 |\, m \neq k\}$. The resulting spliced rainfall time series $R_{(k)}^{spl}$ is denoted with respect to the influencing month, and has the same length as the corresponding observed R^{obs} and simulated R^{sim} time series.

$$R_{(k)}^{spl} = \cup_{m=1}^{12} \begin{cases} R_m^{sim}\,; m = k \\ R_m^{obs}\,; m \neq k \end{cases} \qquad (1)$$

For example, if June ($k = 6$) is selected as the influencing month, each year of the spliced time series, $R_{(6)}^{spl}$, would be composed as follows:

$$R_{(6)}^{spl} = \{R_1^{obs}, \ldots, R_5^{obs}, R_6^{sim}, R_7^{obs}, \ldots, R_{12}^{obs}\} \qquad (2)$$

The ensemble of $k = 1, \ldots, 12$ spliced rainfall time series $R_{(k)}^{spl}$ for all influencing months is transformed according to a hydrological model $g[\,]$ to produce an ensemble of simulated streamflows. However, the hydrological evaluation centres on a specific subset of flows $Q_{(t,k)}^{sim}$ relating to the evaluated month, t,

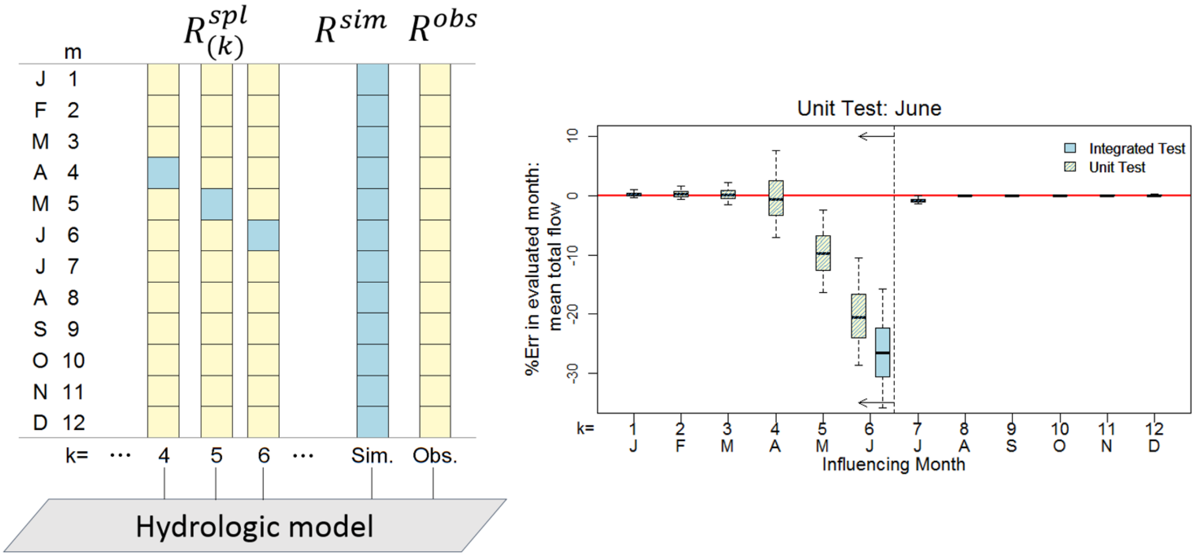
$$Q_{(t,k)}^{sim} \underset{\mathrm{t}}{\subset} g\left[R_{(k)}^{spl}\right] \qquad (3)$$

The ensemble of simulated streamflows corresponding to influencing month k and evaluated month t are used to determine errors with respect to the virtual-observed streamflow in the evaluated month. Using the function $h[\,]$ to denote a calculated statistic of interest, the relative error in an evaluated month is given by

$$\%Err_{(t)} = \frac{h\left[Q_{(t)}^{sim}\right] - h\left[Q_{(t)}^{vo}\right]}{h\left[Q_{(t)}^{vo}\right]} \times 100 \qquad (4)$$

where $Q_{(t)}^{vo}$ is the virtual-observed streamflow and $Q_{(t)}^{sim}$ is the simulated streamflow in the evaluated month (whether from the integrated test or any of the 12 unit tests). Thus, for each site, statistic and evaluated month there are 13 errors to compare.

By construction, spliced rainfall is identical to the observed rainfall for all months other than the influencing month, so any errors in streamflow statistics can be attributed to the influencing month free from other factors. Fig. 3(b) illustrates for the evaluated month of June ($t = 6$), a typical error profile from a unit test of the mean monthly flow. Fig. 3(b) shows the integrated test produced a median error of 27% (blue shaded boxplot), when the influencing month is June ($k = 6$) the median error is 20%, when the influencing month is May ($k = 5$) the median error is 10%, and when the influencing month is April ($k = 4$) the median error is negligible. Therefore, the bias in mean June streamflow is primarily due to rainfall model deficiencies in June and May respectively.



(a)

(b)

Fig. 3 Schematic of (a) the method of constructing a unit test by embedding simulated months in an observation time series, and (b) the error profile produced when using the integrated and unit tests for the evaluated month of June (90% limits shown).

2.4 Implementation of Framework

The procedure for virtual hydrologic evaluation is implemented in a stepwise manner. The first step is to select a streamflow characteristic of interest, herein termed the ‘primary streamflow characteristic’. The reason for specifying a primary streamflow characteristic is to enable a method (carried out in Step 4) for filtering sites and concentrating the investigation of the rainfall model on sites that perform poorly in terms of its intended application. For example, a flood frequency distribution would be a suitable characteristic for a flood risk study, or the distribution of annual total flow when investigating yield.

The second step is to select an appropriate hydrological model to simulate the streamflow. The hydrological model should be selected on the basis that it is capable of simulating streamflow for the timescales, magnitudes and physical processes of interest to the intended application. For example, a capability for simulating peak flows is important for flood risk studies.

The third step is to conduct an integrated test for each rainfall site, which serves as an overall test of the rainfall model’s performance. The test uses the observed rainfall time series and all replicates of the simulated rainfall time series as inputs to the hydrological model (Fig. 2).

The fourth step is to reduce the set of investigated sites to a smaller subset that poorly simulated the primary streamflow characteristic during the integrated test in step three. This step assumes a multi-site evaluation and is optional if only a single site is evaluated

The fifth step is to conduct a detailed analysis of the integrated test results at the monthly scale for the remaining subset of sites. Evaluating monthly total flows is a valuable test of rainfall model performance as the production of monthly total flow volumes relies on the integration of many daily rainfall characteristics (amount, duration, persistence). Errors in reproducing virtual-observed streamflow are contrasted against the observed-rainfall evaluation so that specific sites and months can be identified for further investigation in steps six and seven. The comparison between the observed-rainfall evaluation and the integrated test can be summarised graphically (e.g. see Fig. 5, Section 4.1) for the subset of sites showing:

- simulated daily rainfall statistics (mean (*m*) daily amounts, standard deviation (*sd*) of daily amounts, mean number of wet days (*nwet*) and the standard deviation of the number of wet days);
- aggregate rainfall statistics (mean and standard deviation of total rainfall); and
- aggregate streamflow statistics (mean and standard deviation of total flow).

For convenience, model performance is categorised following Bennett et al. (2018) as ‘good’, ‘fair’ or ‘poor’. Performance was categorised as ‘good’ if the observed/virtual statistic fell within the 90% limits of the simulated statistic, as ‘fair’ if the observed/virtual statistics fell outside the 90% limits of the simulated statistic but within the 99.7% limits and otherwise as ‘poor’.

The sixth step is to implement the unit test for each site of interest. The full set of spliced rainfall (e.g. spliced rainfall for each month designated as the influencing month $R_{(k)}^{spl}; k = 1, \dots, 12$) is input to the hydrological model. This step is repeated for all available replicates of the spliced time series.

The seventh step is to investigate and compare the results of the unit test (Step 6) and the integrated test (Step 5) selecting each month as the evaluated month in turn. Side-by-side comparison of the results of the integrated test and the unit test is given in terms of the errors for selected monthly and annual statistics (see illustration in Fig. 3(b)). The comparison of errors from the unit test forms the basis of interpretation for hydrological insights and their relationship to the rainfall model.

Comparing the unit test across all 12 influencing months with the integrated test can lead to several possible interpretations of where streamflow errors in the evaluation month originate. Differentiating the cases allows for rainfall model improvements to be targeted in terms of their ultimate impact on streamflow statistics. For example, streamflow errors might originate from (i) rainfall model deficiencies mostly in the evaluated month, (ii) rainfall model deficiencies over a contiguous block of months including and preceding the evaluated month, or (iii) rainfall model deficiencies in a preceding month more so than in the evaluated month. Examples for each of case are presented in Section 4.

3 Case Study

The Onkaparinga catchment in South Australia is used as a case study (Fig. 4). The catchment lies 25 km to the south of the Adelaide metropolitan area and contains the largest reservoir in the Adelaide Hills supplying the region (Mount Bold Reservoir). The catchment rainfall has a strong seasonal cycle with the majority of rainfall occurring in winter and spring (June to November) and with a predominantly dry summer season (December to February). There is a strong rainfall gradient (Table 2), with average annual rainfall ranging from approximately 500 mm on the coast (Site No. 19) to over 1000 mm in the region of highest elevations (Site No. 20).

The simulated rainfall was determined from the latent variable autoregressive daily rainfall model of Bennett et al. (2018) using at-site calibrated parameters. The rainfall model was calibrated and simulated at 22 locations throughout the catchment that have long, high-quality records (Table 2). 10,000 replicates of simulated rainfall covering a 73 year period (1914-1986) were used.

In this paper annual total flow volumes were designated as the primary streamflow characteristic to narrow the number of sites investigated (Step 1).

The hydrological model, GR4J (Perrin et al., 2003) was used to simulate virtual-observed streamflow (Step 2). GR4J is a daily lumped hydrological model that simulates daily streamflow in a parsimonious manner using four parameters. The GR4J model was calibrated to simulate streamflow for the Onkaparinga catchment (Westra et al., 2014) and was good fit to the observed streamflow, with a Nash-Sutcliffe efficiency of 0.8. The model has also been used for other virtual evaluation studies (Li et al. 2014; 2016).

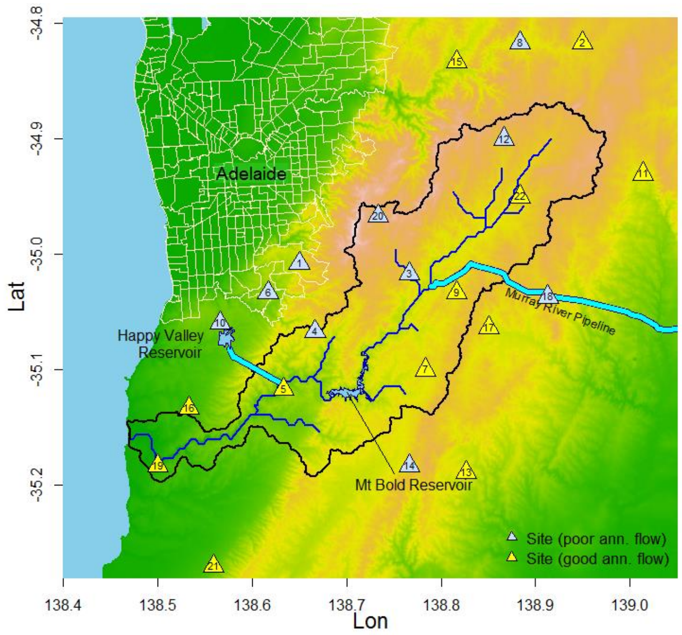


Fig. 4 Onkaparinga catchment, South Australia. Sites indicated by blue triangles are explored in greater detail in this paper due to the relatively poorer ability of simulated rainfall to reproduce annual streamflow totals at these sites.

Table 2 Site names and locations.

Site No	Site Name	Elev (m)	Ann. Ave. Rain (mm)	Site No	Site Name	Elev (m)	Ann. Ave. Rain (mm)
1	Belair	386	790	12	Lobethal	470	880
2	Birdwood	385	720	13	Macclesfield	302	730
3	Bridgewater	376	1050	14	Meadows	384	870
4	Cherry gardens	345	920	15	Cudlee Creek	311	830
5	Clarendon	223	820	16	Morphett Vale	90	560
6	Coromandel Valley	234	710	17	Mount Barker	349	770
7	Echunga	375	805	18	Nairne	403	680
8	Gumeracha	346	790	19	Old Noarlunga	7	520
9	Hahndorf	347	850	20	Uraidla	499	1090
10	Happy Valley	148	640	21	Willunga	158	640
11	Harrogate	335	550	22	Woodside	387	800

4 Results

Following the selection of annual total flow as the primary streamflow characteristic (Step 1) and selection of the hydrological model (Step 2), Step 3 was undertaken to evaluate the simulated rainfall at the 22 sites.

The annual total flow distribution was used to give a broad indication of performance (see Step 4). This step categorised 12 of the 22 sites as ‘good’, which is in strong contrast to earlier evaluation efforts using observed-rainfall evaluation (Bennett et al., 2018) that categorised the majority of sites and statistics as ‘good’ (see Section 2.4 for category definitions). Following Step 4, the 10 sites categorised ‘poor’ are the focus of subsequent steps in the hydrologic evaluation framework. The chosen subset of ‘poor’ sites is indicated by the blue triangles in Fig. 4.

4.1 Integrated test

Fig. 5 compares the rainfall model's performance for the 10 sites categorised 'poor' in Step 4 in terms of both observed-rainfall evaluation and virtual hydrological evaluation via an integrated test (Step 5), with each row summarising monthly or annual performance.

Although the rainfall model performed well in simulating the total rainfall means and standard deviations for the majority of months and sites, it is clear from Fig. 5 that 'poor' performance in rainfall did not necessarily translate to 'poor' performance is simulating streamflow. For example, examining the first row of Fig. 5, it can be seen that in January the rainfall model's ability to simulate variability in the number of wet days, *sd(nwet),* was 'poor' for all sites. However, Fig. 5 shows that despite this, most sites had 'good' performance in simulating the January distribution of monthly total flow (i.e. *m(total)* and *sd(total)*). Similar results were seen in months February, May and November.

Conversely, 'good' performance in the observed-rainfall evaluation does not necessarily translate to 'good' performance in the simulation of monthly total flows. For example, June and August (Fig. 5) have large percentages of 'poor' sites for simulating monthly total flow. This deficiency would have been difficult to infer using the observed-rainfall evaluation due to the 100% 'good' performance of *m(total)* rainfall and 'good/fair' performance of *sd(total)* rainfall in these months. Likewise, by examining the bottom row of Fig. 5 that summarises annual performance, it can be seen that the observed-rainfall evaluation shows unbiased mean annual total, *m(total)*, rainfall (100% 'good') and yet the mean annual total flows showed only 10% of sites as 'good'. Discussion of the unit tests in the following section will investigate reasons why apparently 'good' rainfall can yield 'poor' flow.

© Author(s) 2018. CC BY 4.0 License.

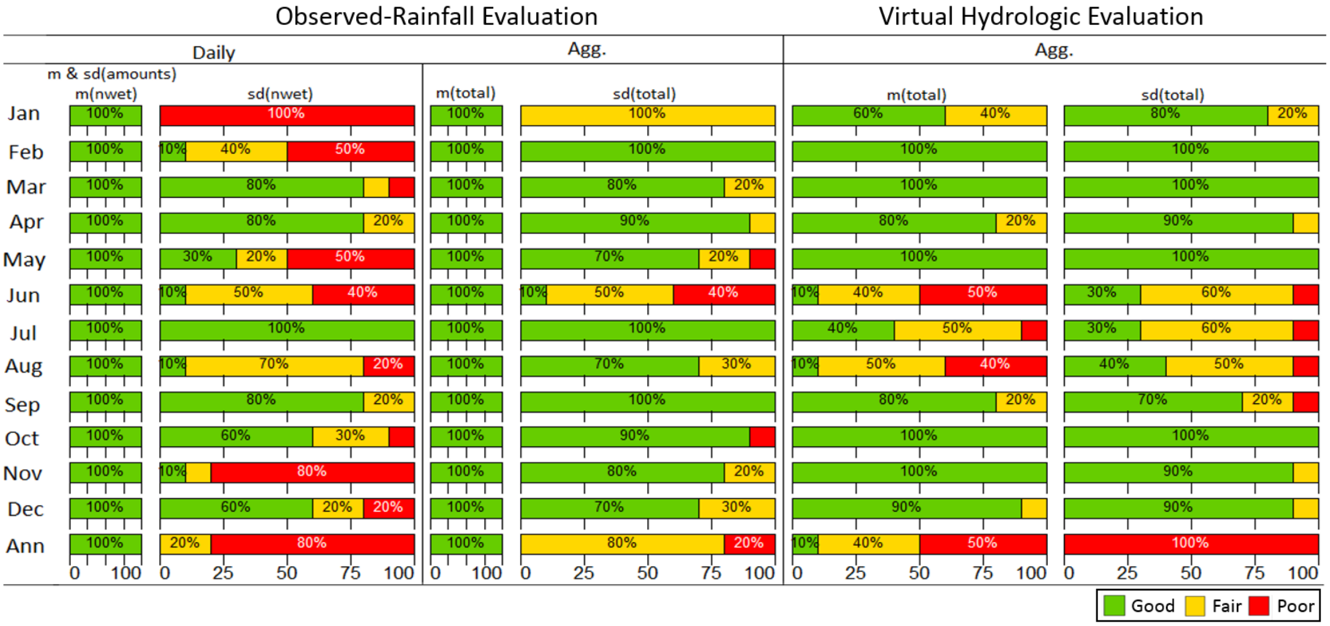


Fig. 5 Integrated test, comparing observed-rainfall evaluation (left) with the virtual hydrologic evaluation (right). Comparison of daily and aggregate ('Agg.') rainfall statistics against aggregate flow statistics for individual months and years. Daily rainfall statistics presented include the means and standard deviations of amounts (*m(amounts), sd(amounts)*), means and standard deviations of the number of wet days (*m(nwet), sd(nwet)*). The aggregate rainfall statistics presented include the means and standard deviations of total rainfall volumes (*m(total), sd(total)*). The aggregate streamflow statistics presented are the means and standard deviations of total flow volumes (*m(total), sd(total)*).

4.2 Unit test – influence of individual monthly rainfall isolated

The unit test was run for the subset of 10 sites (Step 6) and the results of the integrated and unit tests were compared (Step 7). The unit test results and their comparison with integrated test results are illustrated for two locations, Site 12 and Site 10 (see Fig. 6 and Fig. 7). For completeness these results are presented alongside the results of the observed-rainfall evaluation (panels (a) and (b) of Fig. 6 and Fig. 7).

4.2.1 Streamflow errors mostly originate from rainfall model deficiencies in the evaluated month

A common and obvious case for streamflow errors is that they originate from rainfall in the same month. This case can be illustrated using Site 12 in Fig. 6 where left-side panels show results for the mean and right-side panels show the standard deviation and where panels (a) and (b) summarise the observed-rainfall evaluation, (c) and (d) summarise the integrated test. From panels (a) and (b), the simulated monthly rainfall is generally unbiased, but from (c) and (d) the mean and standard deviation of the simulated streamflow is lower than the virtual-observed flow from June to September. Here, September is selected as an illustrative case for an application of the unit test since it shows biased flow.

Taking September as the evaluated month, Fig. 6 (e) and (f) compare the unit tests for all 12 influencing months (yellow and blue striped boxplots) with the integrated test (blue shaded boxplot) in terms of the error in the simulated flow. When the influencing month is September (i.e. the September rainfall is 'spliced' into the observed record) the resultant error is greatest and closest to the error for the integrated test for both the mean monthly total flow (Fig. 6(e)) and standard deviation of monthly total flow (Fig. 6(f)). For the example of the standard deviation, when the influencing month is July (i.e. July rainfall is spliced into the observed record) the median error is less than 2%, whereas when September is taken as the influencing month the median error is approximately 16% (Fig. 6(f)). Therefore, to improve September flows, September rainfall should be improved in preference to all other months.

Analysing other sites and months suggests that over 50% of the evaluations correspond to this case, and they typically occur in spring and summer months when the catchment is drying out. Examination of the simulated flow duration curves instead of monthly totals (not shown) yields a similar conclusion.

4.2.2 Streamflow errors originate from rainfall model deficiencies over a contiguous block of months

An illustration of the case where streamflow errors originate from rainfall model deficiencies over a contiguous block of months is provided by Site 12, where July is selected as the evaluated month. Comparison of the July performance in the integrated and unit tests (Fig. 6 (g) and (h)) demonstrates that the errors in July streamflow do not originate in the July rainfall alone (unlike the case for September – see Section 4.2.1). Although the largest percentage error in flow is attributable to July (a median error of 8% in mean monthly total flow and 25% in the standard deviation of monthly total flow when the influencing month is July) a significant proportion of the error for July streamflow originates in prior months. June and May rainfall have a significant influence on the July flow with percentage errors of up to 15% in July flow when June or May are the influencing month. Therefore, to improve July flows, it is not just the July rainfall that should be improved but also the preceding two months. Typically, autumn (transitional) and winter months (May to August) fall in this case where streamflow errors originate from rainfall model deficiencies over a contiguous block of months, approximately 40% of the site/month combinations.

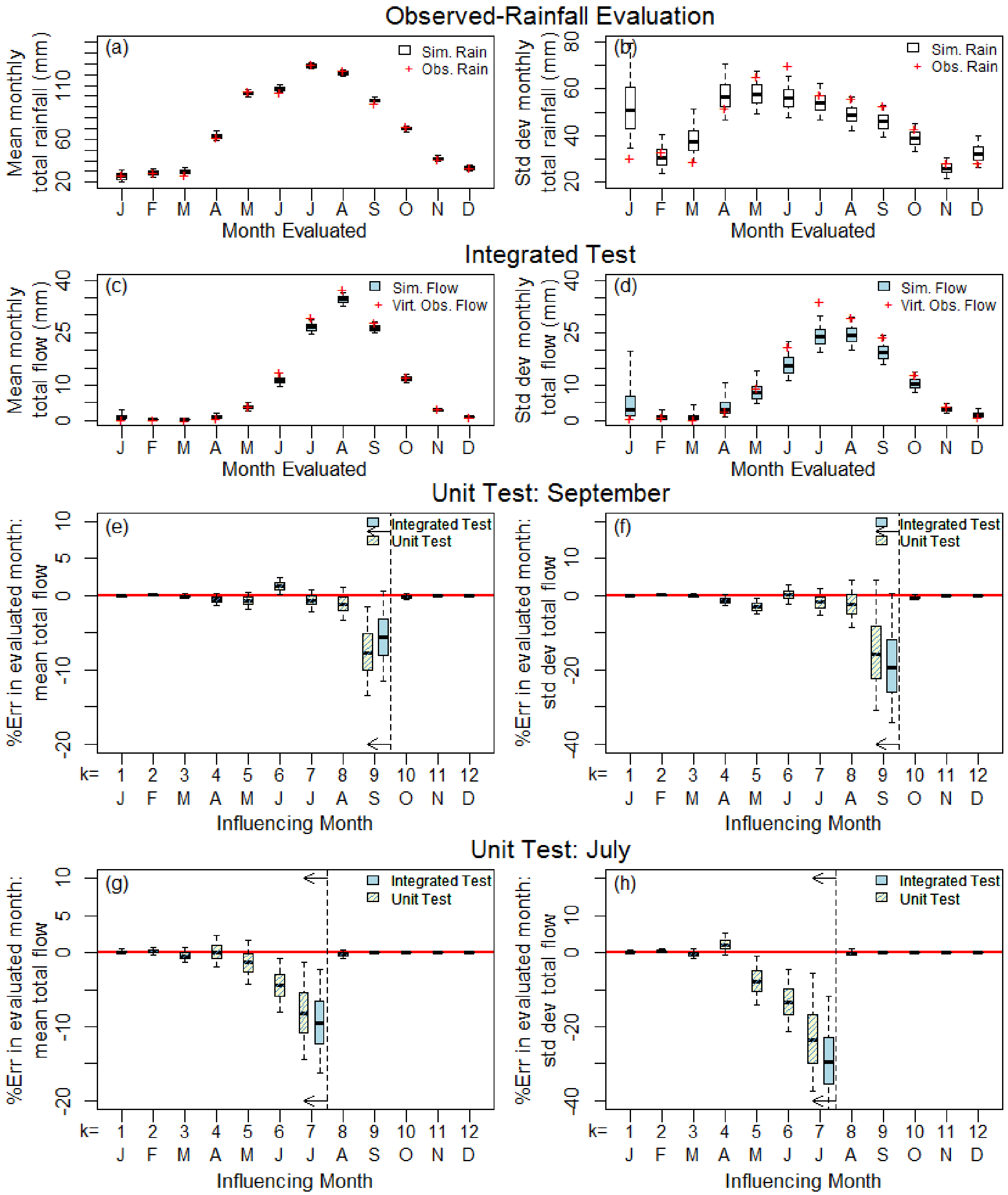


Fig. 6 Lobethal, Site 12 (90% limits shown) (a) observed-rainfall evaluation mean monthly total rainfall, (b) observed-rainfall evaluation standard deviation of monthly total rainfall, (c) integrated test mean monthly total rainfall, (d) integrated test standard deviation of monthly total rainfall (e) unit test error in mean monthly flow (September), (f) unit test error in standard deviation of monthly flow (September), (g) unit test error in mean monthly total flow (July), and (h) unit test error in standard deviation of monthly total flow (July).

4.2.3 Streamflow errors originate from rainfall model deficiencies in a preceding month more so than evaluated month

An example of the case where the largest contribution to streamflow errors arises from rainfall deficiencies in a preceding month is provided by Site 10, where July is selected as the evaluated month. July is selected as an illustrative case for application of the unit test since it shows biased flow (See Fig. 7 (c) and (d)), but did not show any bias in the simulated rainfall (See Fig. 7 (a) and (b)). The largest contributor to error in July flow is not July rainfall but June rainfall (Fig. 7 (e) and (f)). That is, the largest errors occur when there is observed rainfall for July spliced with simulated rainfall for June. In contrast, simulated July rainfall spliced with observed rainfall in other months, yields a smaller median error.

While improving the July rainfall will improve the simulation of July flow, a more significant improvement will be obtained by focusing on improving the June rainfall. The category where streamflow errors originate from rainfall model deficiencies in a preceding month represents about 10% of the site/month combinations for this case study.

4.2.4 Influence of monthly rainfall on annual flow volumes

While annual simulated rainfall was unbiased, annual simulated streamflow was biased. An illustration of how errors in annual total streamflow arise from rainfall is shown for Site 10. Fig. 7 (g) and (h) show that when the months of May to August are assessed as the influencing month they produce the largest errors in distribution of annual total flow for Site 10. Splices of other months do not significantly degrade the simulation of total annual flow. Improvements to the simulation of annual total flow will therefore come from improving the rainfall model in the 'wetting-up' months of May to August in late-autumn and winter. This insight from the use of unit testing would be difficult to obtain using other evaluation strategies (further discussed in Section 5.3).

© Author(s) 2018. CC BY 4.0 License.

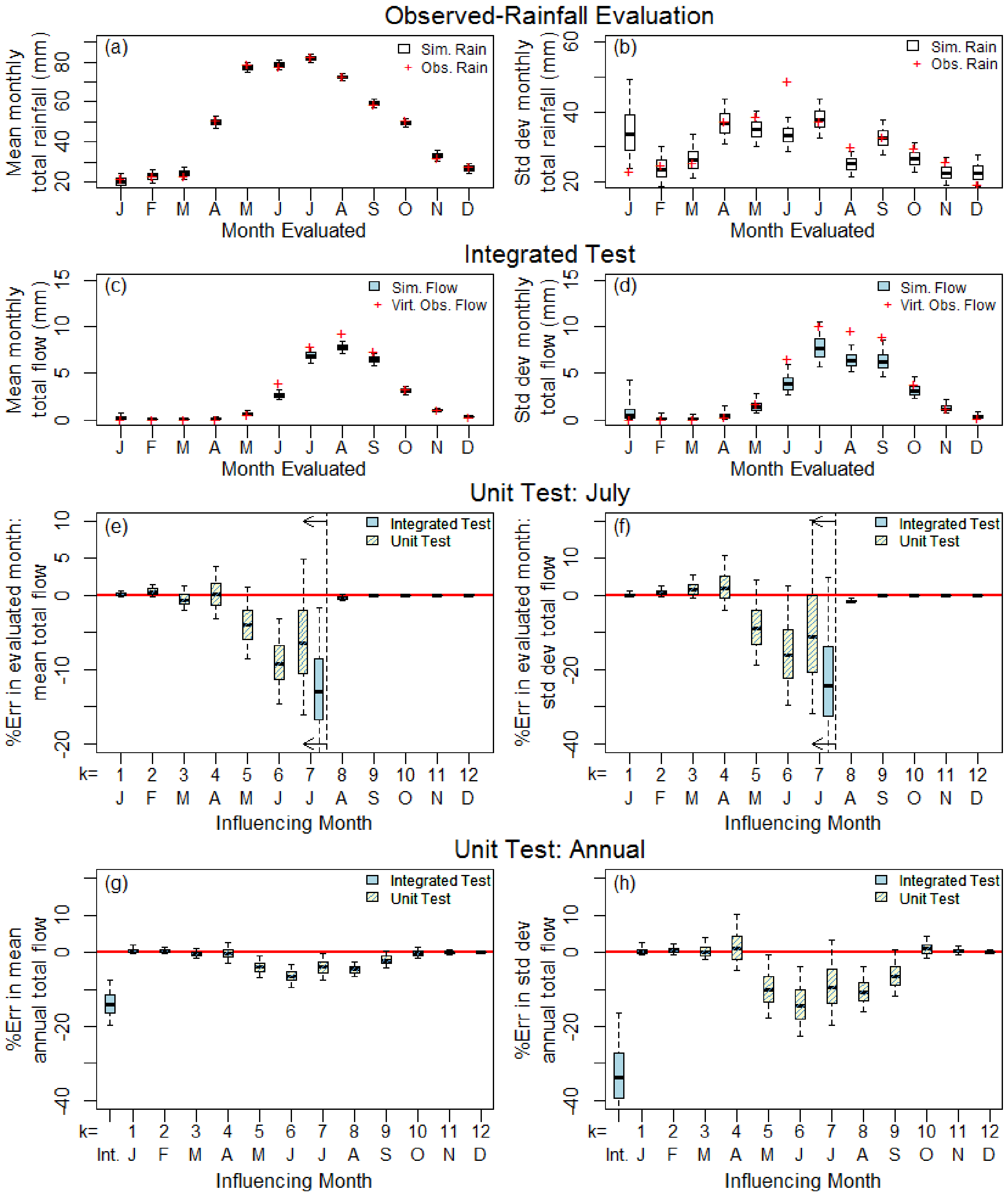


Fig. 7 Happy Valley (Site 10) (90% limits shown) (a) observed-rainfall evaluation mean monthly total rainfall, (b) observed-rainfall evaluation standard deviation of monthly total rainfall, (c) integrated test mean monthly total rainfall, (d) integrated test standard deviation of monthly total rainfall (e) unit test error in mean monthly flow (July), (f) unit test error in standard deviation of monthly flow, (g) unit test error in mean annual total flow, and (h) unit test error in standard deviation of annual total flow.

5 Discussion

5.1 The importance of streamflow-based evaluation

Streamflow arises from the integration of rainfall processes (e.g. rainfall amounts, occurrences and wet-dry patterns) over a catchment. Features of the catchment, such as catchment storage, thresholds and nonlinearities in the rainfall-streamflow response function, can either act to amplify or dampen the rainfall errors at different times of the year.

In terms of amplification, the elasticity of the rainfall-streamflow response (Chiew, 2006) suggests that catchments can have strong sensitivities to discrepancies in rainfall. Given that the rainfall elasticity of streamflow to rainfall is a factor of 2 to 3.5, using the principles of error propagation of linear functions (Ang and Tang, 2007) it follows that a 10% error in mean/standard deviation of rainfall could be amplified to 20-35% error in the mean/standard of streamflow. This indicates that streamflow-based evaluation of rainfall models provides a stronger test than observed-rainfall evaluation in terms of the sensitivity of the statistics. For example, Fig. 5 shows that July rainfall statistics were classified as ‘good’, yet despite this, the streamflow response was ‘poor’ (see Section 5.3 for further discussion). It could be argued that the rainfall results presented in Fig. 5 were classified as ‘good’ because the observed-rainfall evaluation was limited, but the evaluation was methodical and used a comprehensive range of daily and monthly statistics (Bennett et al., 2018). While many rainfall statistics were preserved (means, standard deviation, extremes, marginal distributions of daily rainfall) the rainfall-streamflow response of the catchment exposes that there are deficiencies in the rainfall model not clearly identified by the observed-rainfall evaluation (Bennett et al., 2018).

In terms of dampened influence, catchment storages and high evapotranspiration can also act to suppress errors in the rainfall simulations. For example, Fig. 5 showed that the variability in the number of wet days, *sd*(*nwet*), was ‘poor’ for all sites in January, yet this did not result in ‘poor’ streamflow. The high potential evapotranspiration in January indicates that the majority of rainfall in January is converted into actual evapotranspiration yielding little streamflow. Hence, any errors in rainfall do not noticeably impact on January streamflow.

It is clear that streamflow-based evaluation is beneficial in addition to conventional observed-rainfall evaluation. However, a hydrological framework is not without its own challenges (see Sections 5.2 and 5.3 below).

5.2 The importance of virtual evaluation

A benefit of virtual hydrologic evaluation is the ability to undertake streamflow-based evaluation at any site where rainfall is observed and simulated. In contrast, observed-streamflow evaluation relies on having measurements of streamflow – which is only undertaken at limited number of locations (e.g. across Australia there are over 8, 000 observed daily rainfall sites, typically greater than 50 years length, compared to approximately 3,300 observed streamflow sites, typically 20 years length). Virtual evaluation can be undertaken at a much wider number of locations. For example, annual total flow volume was compared at 22 virtual-observed streamflow sites and identified that only 10 sites showed ‘poor’ performance (Fig. 4). While there are several streamflow gauges in the Onkaparinga catchment, their records are shorter and complicated by inter-catchment pipeline transfers (Fig. 4) and nonstationarities (Westra et al, 2014) making observed-streamflow evaluation difficult. Furthermore, the process of comparing to observed streamflow requires estimates of the catchment rainfall, which is constructed from multiple rainfall sites and obscures the ability to evaluate rainfall at individual sites.

A potential limitation of the virtual hydrologic evaluation framework is that it is reliant on the use of a hydrological model – which is common to both the observed and simulated rainfall. There is the potential for hydrological structural errors to skew interpretation of the rainfall model evaluation if the hydrological model poorly represents the catchment processes. To reduce these impacts it is recommended to use a well-tested hydrological model that has demonstrated good performance on a wide range of catchments. In this study GR4J was used, which has been tested on hundreds of catchment in Europe and Australia (Coron et al., 2012; Perrin et al., 2003). It is further recommended to use a hydrological model that has been calibrated and evaluated using observed data located close to the observed rainfall sites. For example, in this study the GR4J model was calibrated to the Onkaparinga catchment (Westra et al., 2014) – see Fig. 4. Even where a well-tested hydrological model is calibrated, there is the remaining question of whether or not the conclusion of a virtual evaluation is dependent on the choice of the hydrological model. To address this limitation, multiple hydrological models could be used as part of the virtual framework. As this is the first proof-of-concept study that has introduced the virtual hydrologic evaluation framework, the use of multiple hydrological models is outside the scope of this paper and is left for future research.

5.3 Virtual hydrological evaluation and unit testing – a unique opportunity for targeted evaluation of rainfall model performance

Previous discussion has highlighted the benefits of the streamflow-based and virtual elements of an evaluation framework of rainfall models when compared to existing evaluation methods (observed-rainfall evaluation and observed-streamflow evaluation). However, these innovations alone lack the ability to target which periods of the rainfall model produce poor streamflow simulations. For example, it would be possible to identify a discrepancy in July for Site 12, but it would be impossible to determine whether this was due to a deficiency in rainfall in July, or a prior month such as May or June (Fig. 6(g) and (h)).

The use of a virtual hydrological framework for evaluation provides the unique opportunity to develop innovative tests that can target specific aspects of the stochastic rainfall model. A unit test was introduced as a method for isolating the influence of rainfall in a month (i.e. the influencing month) on streamflow in an evaluated month while excluding the possibility of deficiencies from other rainfall months. The test enables a procedure for targeting months that are influential in terms of streamflow production rather than interpret model performance based on blunt evaluation of rainfall or streamflow.

Without the unit test, identifying which months have deficiencies in the modelled rainfall that produce poor streamflow predictions would not have been possible. If the modeller, had focussed on improving the rainfall model by focusing on months with the highest contribution to annual total flow, July to September would have been identified as important, whereas the unit test identifies a different focus (May-August). For example, May and June combined contribute 13% in the annual total flow (Fig. 7 (c)), compared to 11% error in the mean (Fig. 7 (g)) and 24% error in the standard deviation (Fig. 7 (h)). By contrast, September is a high flow month contributing 21% of the annual total flow, but only 2% error in the mean and 6% error in the standard deviation. Without the unit test, it would have been less clear that the 'wetting-up' months such as May and June were a more important focus for rainfall model improvement than a high-flow month such as September.

Although it is possible to isolate which months were the main contributors to errors in the annual total flow volume, questions remain surrounding the influence of longer term wet-dry patterns, month-to-month correlations, and inter-annual correlations on annual flow. It is possible to extend the framework to address these questions.

6 Conclusions

This paper has introduced the virtual hydrologic evaluation framework to overcome the problems of existing evaluation methods which are not able to identify which rainfall characteristics are important for streamflow prediction. The framework enables two key innovations, an integrated test and a unit test, by using a hydrological model to convert rainfall observations to corresponding 'virtual-observed' streamflow. The integrated test compares simulated streamflow and virtual-observed streamflow to detect overall deficiencies in the ability of at-site stochastic rainfall to reproduce streamflow statistics. The unit test enables the attribution of detected streamflow errors to specific months of stochastic rainfall. The integrated and unit tests enabled different conclusions to be reached in terms of priorities for improving the rainfall model. These conclusions would not otherwise have been possible with conventional evaluation methods that focus either on rainfall statistics, or on high streamflow months. The integrated test demonstrated that while large discrepancies were identified in low rainfall months these did not translate to deficiencies in streamflow due to the dry state of the catchment. The test also indicated instances where modelled rainfall categorised as 'good' translated to 'poor' flow due to the influence of catchment memory and rainfall from prior months. The unit test identified the importance of transition months May and June (late autumn/early winter) in the 'wetting-up' phase of the catchment cycle. By contrast, a conventional approach focusing on high flows would have emphasised the importance of correctly simulating rainfall in high flow months. The virtual hydrologic evaluation provides insights not available through conventional approaches and should be an essential step in the development and application of stochastic rainfall models.

7 Acknowledgements

This work was supported by an Australian Research Council Discovery grant: A new flood design methodology for a variable and changing climate DP1094796. Additional support was provided by the CSIRO Climate Adaptation Flagship.

© Author(s) 2018. CC BY 4.0 License.

8 References

Andreassian, V., Perrin, C., Michel, C., Usart-Sanchez, I., and Lavabre, J.: Impact of imperfect rainfall knowledge on the efficiency and the parameters of watershed models, Journal of Hydrology, 250, 206-223, 2001.

Ang, A. and Tang, W.: Probability Concepts in Engineering: Emphasis on Applications to Civil and Environmental Engineering (v. 1), 2007.

Ball, J. E.: The influence of storm temporal patterns on catchment response, Journal of Hydrology, 158, 285-303, 1994.

Baxevani, A. and Lennartsson, J.: A spatiotemporal precipitation generator based on a censored latent Gaussian field, Water Resources Research, 2015. 2015.

Bennett, B., Thyer, M., Leonard, M., Lambert, M., and Bates, B.: A comprehensive and systematic evaluation framework for a parsimonious daily rainfall field model, Journal of Hydrology, 556, 1123-1138, 2018.

Blazkova, S. and Beven, K.: Flood frequency estimation by continuous simulation for a catchment treated as ungauged (with uncertainty), Water Resources Research, 38, 14-11-14-14, 2002.

Blazkova, S. and Beven, K.: A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic, Water Resources Research, 45, W00B16, 2009.

Camici, S., Tarpanelli, A., Brocca, L., Melone, F., and Moramarco, T.: Design soil moisture estimation by comparing continuous and storm-based rainfall-runoff modeling, Water Resour. Res., 47, W05527, 2011.

Chiew, F. H.: Estimation of rainfall elasticity of streamflow in Australia, Hydrological Sciences Journal, 51, 613-625, 2006.

Clark, M., Slater, A., Rupp, D., Woods, R., Vrugt, J., Gupta, H., Wagener, T., and Hay, L.: Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, Water Resour. Res., 44, 2008.

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

Cowpertwait, P. S. P.: A spatial-temporal point process model of rainfall for the Thames catchment, UK, Journal of Hydrology, 330, 586-595, 2006.

Coxon, G., Freer, J., Westerberg, I., Wagener, T., Woods, R., and Smith, P.: A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations, Water resources research, 51, 5531-5546, 2015.

Evin, G., Favre, A.-C., and Hingray, B.: Stochastic generation of multi-site daily precipitation focusing on extreme events, Hydrology and Earth System Sciences, 22, 655-672, 2018.

Evin, G., Thyer, M., Kavetski, D., McInerney, D., and Kuczera, G.: Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity, Water Resources Research, 50, 2350-2375, 2014.

Henley, B. J., Thyer, M. A., and Kuczera, G.: Climate driver informed short- term drought risk evaluation, Water Resources Research, 49, 2317-2326, 2013.

Kuczera, G. and Williams, B. J.: Effect of rainfall errors on accuracy of design flood estimates, Water Resources Research,, 28, 1145-1154, 1992.

Li, J., Thyer, M., Lambert, M., Kuczera, G., and Metcalfe, A.: An efficient causative event-based approach for deriving the annual flood frequency distribution, J Hydrol, 510, 412-423, 2014.

Li, J., Thyer, M., Lambert, M., Kuzera, G., and Metcalfe, A.: Incorporating seasonality into event-based joint probability methods for predicting flood frequency: A hybrid causative event approach, Journal of Hydrology, 533, 40-52, 2016.

Li, Z., Lü, Z., Li, J., and Shi, X.: Links between the spatial structure of weather generator and hydrological modeling, Theoretical and Applied Climatology, doi: 10.1007/s00704-015-1691-8, 2015b. 1-9, 2015b.

McInerney, D., Thyer, M., Kavetski, D., Lerat, J., and Kuczera, G.: Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors, Water Resources Research, 53, 2199-2239, 2017.

McMillan, H., Freer, J., Pappenberger, F., Krueger, T., and Clark, M.: Impacts of uncertain river flow data on rainfall- runoff model calibration and discharge predictions, Hydrological Processes: An International Journal, 24, 1270-1284, 2010.

McMillan, H. K. and Brasington, J.: End-to-end flood risk assessment: A coupled model cascade with uncertainty estimation, Water Resources Research, 44, W03419, 2008.

Mortazavi-Naeini, M., Kuczera, G., Kiem, A. S., Cui, L., Henley, B., Berghout, B., and Turner, E.: Robust optimization to secure urban bulk water supply against extreme drought and uncertain climate change, Environmental Modelling & Software, 69, 437-451, 2015.

Nicótina, L., Alessi Celegon, E., Rinaldo, A., and Marani, M.: On the impact of rainfall patterns on the hydrologic response, Water Resources Research, 44, 2008.

Paschalis, A., Molnar, P., Fatichi, S., and Burlando, P.: A stochastic model for high- resolution space- time precipitation simulation, Water Resources Research, 49, 8400-8417, 2013.

Paton, F. L., Maier, H. R., and Dandy, G. C.: Relative magnitudes of sources of uncertainty in assessing climate change impacts on water supply security for the southern Adelaide water supply system, Water Resources Research, 49, 1643-1667, 2013.

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

Rasmussen, P.: Multisite precipitation generation using a latent autoregressive model, Water Resources Research, 49, 1845-1857, 2013.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W.: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, Water Resour. Res., 46, W05521, 2010.

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

Shah, S. M. S., O'Connell, P. E., and Hosking, J. R. M.: Modelling the effects of spatial variability in rainfall on catchment response. 2. Experiments with distributed and lumped models, Journal of Hydrology, 175, 89-111, 1996.

Smith, P. J., Beven, K. J., and Tawn, J. A.: Detection of structural inadequacy in process-based hydrological models: A particle-filtering approach, Water Resources Research, 44, W09403, 2008.

Srikanthan, R. and Pegram, G. G. S.: A nested multisite daily rainfall stochastic generation model, Journal of Hydrology, 371, 142-153, 2009.

Westra, S., Thyer, M., Leonard, M., Kavetski, D., and Lambert, M.: Impacts of climate change on surface water in the Onkaparinga catchment-Final report volume 1: hydrological model development and sources of uncertainty, 1839-2725, 2014.

Wilks, D. S.: High-resolution spatial interpolation of weather generator parameters using local weighted regressions, Agricultural and Forest Meteorology, 148, 111-120, 2008.

Wilson, C. B., Valdes, J. B., and Rodriguez-Iturbe, I.: On the influence of the spatial distribution of rainfall on storm runoff, Water Resources Research, 15, 321-328, 1979.