Overall Response to Editor

This document describes our response to the reviewers' comments and our revision of the manuscript entitled 'A virtual hydrological framework for evaluation of stochastic rainfall models' (HESS 2018-489). We agreed with the majority of the issues raised by the reviewers and editor. The revised paper is a significant improvement on the original submission. The key modifications we have made in the revised paper include:

- Clarifying and emphasising the novelty of the virtual hydrological evaluation framework more strongly (see Section 1: Introduction). In particular, we have provided better referencing and discussion on existing uses of virtual-observed streamflow evaluation and have emphasised the significance of formalising virtual-observed streamflow evaluation using a comprehensive and systematic evaluation (CASE) framework approach (originally developed in Bennett et al. 2018), as well as the novelty of the new unit test.
- Restructuring and shortening the motivation for the virtual hydrological evaluation approach as well as descriptions of the theoretical elements (see Sections 1 and 2).
- Extending the presented analysis to include additional runoff characteristics (i.e. flow duration curves) in the diagnosis of the sources of deficiencies in streamflow (see Section 4.3). This evaluation of flow duration curves has been incorporated in to the virtual hydrological evaluation framework as part of Step 3 (see Section 2.4.3.).
- Streamlining the virtual evaluation framework steps (from seven steps down to three), summarising this procedure graphically using a flow chart (see Figure 2) and restructuring Section 2 to better integrate the framework steps and the specific virtual tests (i.e. integrated and unit tests). Section 4 (Results) has also been re-structured to follow the streamlined framework procedure.
- Improving the explanations of the stochastic rainfall model and the hydrological model used in the case study, including the providing descriptions of the calibration approaches (see Section 3).

Please find below our response to specific comments.

Response to Editor

Thank you for posting your responses to the three referees' reports. The referees raised some critical comments and suggestions that I urge you to consider. I feel your work, if thoroughly revised, could ultimately be accepted for publication in HESS. I invite you to upload a revised manuscript, incorporating the proposed changes and additions, and making any other modifications where you see fit. Alongside a thoroughly revised version of your manuscript, please provide a detailed response ("item-by-item") to each of the referees' remarks.

In addition to the comments from the referees, I kindly ask you to add "Author contribution" and "Data availability" sections to the manuscript, as indicated in the guidance for authors (see https://www.hydrology-and-earth-system-sciences.net/for_authors/manuscript_preparation.html).

Response:

We have thoroughly revised the manuscript in light of the reviewer comments. 'Author contribution' and 'Data availability' sections have been added to the revised manuscript.

Our item-by-item responses appear below.

Response to Reviewer 1

Comment 1:

The manuscript introduces a new framework for the evaluation of generated rainfall time series in terms of their ability to reproduce runoff time series characteristics. This is done by two tests, an integrated test and a unit test. This topic is of broad interest for the hydrological scientific community and suitable for a publication in HESS.

However, I consider the integrated test not as a novelty, since it has been applied before in different studies, but the unit test is useful for rainfall model evaluation. Hence, I suggest to move the focus to the unit test and extend the validation by other runoff characteristics. Also, the theoretical elements of the paper are very long, the application and validation of the test should be extended and there is a lack of some crucial information regarding the applied r-r model and its calibration procedure (for details please see my specific comments). Due to the resulting workload I suggest a major revision of the manuscript.

Response 1: We are pleased that the reviewer found our manuscript suitable for publication and we are grateful for the insightful and constructive comments. They have been very helpful. The suggestions provided show careful consideration and have led to an improved revision of the manuscript. Regarding the specific matters raised:

- Novelty of the integrated test – we have reworded objective 2[1] and provided better referencing and discussion on existing uses of virtual-observed streamflow evaluation (calibration, validation, model selection and diagnosis) in the Introduction. The revised text is:

  Page 4, Line 21- 25: *"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) including the identification of rainfall features of interest in terms of hydrological behaviour (Sikorska et al., 2018), (ii) for developing new techniques for flood frequency analysis (Li et al., 2014, 2016), and (iii) to support SRM selection (Müller and Haberlandt, 2018) as well as calibration and validation (Kim and Olivera, 2011) through a comparison to virtual streamflow."*

  We feel that there are some important aspects of the implementation of our framework that distinguish it from existing presentations of virtual-observed streamflow evaluations. In particular, the presentation of the integrated test in this paper is the first time, a virtual-observed streamflow evaluation has been formalised used using a comprehensive and systematic evaluation (CASE) framework approach (Bennett et al. 2018). This has been clarified in the text:

  Page 4, Line 26 – Page 5, Line 3: *"The framework presented in this paper is significant advance from previously reported virtual experiments because it presents a formal framework to identify key deficiencies in the SRM by utilising (1) A comprehensive and systematic evaluation (CASE) framework (developed by Bennett et al., 2018 and used by Evin et al., 2018, Khedhaouiria et al., 2018) that systematically categorises performance at multiple spatial and temporal scales using quantitative criteria for each statistic, and (2) two types of virtual experiments that are able to identify the source of key deficiencies in SRM at specific locations and time periods."*

[1] Page 5, Line 6, *Objective 2: "*To present two different tests which are part of the framework: the integrated test as well as introduce a new test, the unit test. Combined use of these tests allows streamflow discrepancies to be attributed to their original source in the SRM according to site and season."

This distinguishing feature and others are further discussed in response to comment 2. We have taken the reviewer's advice and emphasized the novelty of the unit test and its diagnostic ability in the revised manuscript.

- *Evaluation using other runoff characteristics* – we have included an examination of flow duration curves to provide additional insight into the deficiencies of the rainfall model. This evaluation has been incorporated in to the virtual hydrological evaluation framework as part of Step 3 (see Section 2.4.3):

  > Page 12, Lines 1-8: *"Following a side-by-side comparison of integrated test and unit test results in terms of the relative errors the sources of poor performance should be classified in terms of in which influencing months streamflow deficiencies originate (e.g. poor streamflow arises from rainfall deficiencies mostly in the same month, a prior month or a contiguous block of months) … To complement this analysis a comparison of the virtual-observed flow duration curve for the evaluated time period with the flow duration curves resulting from unit tests for key influencing months is also recommended."*

- *Length of theoretical elements* – the motivation and description of the theoretical elements (now Sections 1 – 2.1) have been shortened by 40 lines (from 167 lines down to 127 lines).
- *Information on the r-r model* – we have improved the explanation of the rainfall-runoff model, including adding references of the calibration and validation procedure (see Section 3). The revised text is:

  > Page 13, Lines 1-12: "*The hydrological model GR4J (Perrin et al., 2003) was used to simulate virtual-observed streamflow at a daily time step. GR4J is a daily lumped hydrological model that simulates daily streamflow in a parsimonious manner using four parameters. The GR4J model was calibrated according to the procedure set out in Westra et al. (2014b) for the stationary version of the GR4J hydrological model. The details are provided in (Westra et al., 2014a) and a short summary is provided here. The multi-site rainfall gauges were Thiessen weighted to calculate the catchment average rainfall. The model was calibrated to the streamflow data at Houlgrave Weir (see Figure 4) using model calibration period of 15 years (1985-1999). The parameters were estimated using maximum likelihood estimation procedure with a weighted least squares likelihood function. The model parameters that maximised the likelihood function were found using a multi-start quasi-Newton optimisation procedure with 100 random starts. Overall, the GR4J model was a good fit to the observed streamflow, with a Nash-Sutcliffe efficiency of 0.8. A similar type of hydrological model and calibration approach has been used for other virtual evaluation studies (Li et al. 2014; 2016). The same set of hydrological model parameters are used for both the unit and integrated tests so that the same transformation of rainfall to flow is used."*

We further elaborate on these items in response to subsequent comments made.

Comment 2:

P2l20-23: The so-called "virtual-observed streamflow"-approach and the integrated test is not new and a widely used evaluation method, especially in data-sparse regions or research fields. For example in urban hydrology, where measured runoff characteristics are not often available, the simulation of a reference streamflow is very common (e.g. Müller and Haberlandt, 2018). The authors even mention other studies using the integrated test (Li et al., 2014, 2016). However, the unit test is interesting and indeed provides useful insights into the rainfall-runoff (r-r) transformation process.

Response 2:

Thank you for supplying the references: Müller and Haberlandt, 2018; Sikorska et al. 2018. We have included them in the revised introduction text along with additional references (e.g. Kim and Olivera, 2012).

We agree that the concept of a virtual-observed streamflow evaluation is not new and we have revised the introduction, including the presentation of objective 2, to make this clear and discuss that the approach has been used in a variety of contexts (e.g. calibration, validation, model selection and as a diagnostic tool). The revised text is:

> Page 4, Line 21- 25: "*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) including the identification of rainfall features of interest in terms of hydrological behaviour (Sikorska et al., 2018), (ii) for developing new techniques for flood frequency analysis (Li et al., 2014, 2016), and (iii) to support SRM selection (Müller and Haberlandt, 2018) as well as calibration and validation (Kim and Olivera, 2011) through a comparison to virtual streamflow.*"

However, there are some important aspects of our framework that distinguish it from existing presentations of virtual-observed streamflow evaluation, as outlined below.

1. This is the first time the virtual-observed streamflow evaluation approach has been formalised using a Comprehensive and Systematic Evaluation (CASE) framework (pioneered by Bennett et al., 2018 and used by Evin et al. 2018, Khedhaouiria et al. 2018) to evaluate stochastic rainfall models in terms of the ability to produce key runoff statistics of interest. The integrated tests presented in this paper follow the CASE approach because they (i) present a comprehensive range of key statistics of interest, (ii) systematically categorise performance at specific spatial and temporal scales using quantitative criteria for each statistic, and (iii) systematically categorise aggregate performance over multiple spatial and/or temporal scales. This novelty has been clarified in the revised manuscript:

   > Page 4, Line 26 – Page 5, Line 3: *"The framework presented in this paper is significant advance from previously reported virtual experiments because it presents a formal framework to identify key deficiencies in the SRM by utilising (1) A comprehensive and systematic evaluation (CASE) framework (developed by Bennett et al., 2018 and used by Evin et al., 2018, Khedhaouiria et al., 2018) that systematically categorises performance at multiple spatial and temporal scales using quantitative criteria for each statistic, and (2) two types of virtual experiments that are able to identify the source of key deficiencies in SRM at specific locations and time periods."*

   In contrast, previous papers (Müller and Haberlandt, 2018; Sikorska et al. 2018, Kim and Olivera, 2012) have used a virtual-observed streamflow evaluation approach, but have not used a CASE framework to evaluate the performance of stochastic rainfall model at multiple rainfall sites in terms of its ability capture key streamflow statistics of interest. For example, Müller and Haberlandt (2018) established the need for spatial consistency of rainfall generation in modelling sewer networks by comparing rainfall disaggregation approaches with or without spatial consistency. This virtual-observed streamflow evaluation is performed for identified extreme rainfall events only and therefore does not use a CASE approach that considers multiple temporal scales and the longer term effects of the applied rainfall on the translation of subsequent rainfall to streamflow. Sikorska et al. (2018) focused on identifying rainfall features of interest in terms of resultant hydrological behaviour for the purposes of determining the effective daily precipitation duration with a view to selecting a suitable rainfall disaggregation scheme. Their evaluations determined that detailed temporal rainfall time series were not needed to reproduce annual or seasonal peaks in their modelled catchments. Although the evaluations presented are comprehensive the motivation of the Sikorska et al. (2018) paper is different and does not provide a general formalised framework for systematically categorising stochastic rainfall model performance at specific and aggregate temporal and spatial scales. Kim & Olivera (2012) used virtual-observed streamflow evaluation as part of a larger calibration and validation approach in which various weights were trialled to reflect the importance of various rainfall statistics within a modified Bartlett-Lewis rectangular pulse (MBLRP) model. However, the focus was on the improvement and validation of the MBLRP model rather than the presentation of separate framework for model evaluation. Finally, Li et al. (2014, 2016) used a virtual-observed streamflow evaluation approach, to evaluate the ability of range of techniques to estimate the derived annual flood frequency distribution - they did not use a CASE approach to evaluate stochastic rainfall models.

2. The formalisation of the integrated test using a comprehensive and systematic evaluation (CASE) approach, forms a baseline for subsequent application of the unit test which has greater ability to pinpoint issues with respect to the source of the rainfall error on a monthly basis. For example, in Step 3 of the framework the following comparison is made:

> Page 11, Lines 4-6: "*Side-by-side comparison of the results of the integrated test and unit tests are 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*."

3. As identified by the reviewer, we introduce an innovative unit test, which has never been used before in a virtual-observed streamflow evaluation approach. The key advantage of this unit test is that by splicing together the observed and simulated rainfall in a systematic manner, it is able to develop new insights on which months have deficiencies in simulated rainfall that produce poor performance in streamflow. We have put greater emphasis on this new innovative unit test in the revised manuscript including extending the unit test to evaluate streamflow deficiencies via an examination of daily flow duration curves.

Comment 3:

It would be useful to move the focus on this test and proof it with additional runoff characteristics, e.g. flow duration curves, not using only the monthly runoff amount. Therefore, no new simulations are necessary, only additional analyses of the existing r-r simulation results.

Response 3:

We have included an examination of flow duration curves to provide additional insight into the deficiencies of the rainfall model in the revised manuscript (see Section 4.3). An evaluation of flow duration curves has been incorporated in to the virtual hydrological evaluation framework as part of Step 3 (see Section 2.4.3) to augment the unit test evaluations. An example of the examination and discussion of the flow duration curves and a flow duration curve figure (Fig. 9) are reproduced below:

> Page 18, Lines 9-15: "*This need to improve September in preference to preceding months is also illustrated via Fig. 9 (c) where the September daily flow duration curves are shown for the cases where August (orange shading) and September (blue shading) are the influencing months compared against the virtual-observed September flow duration curve (purple dots). Where August is selected as the influencing month, the virtual-observed flow duration curve largely sits inside the 90% limits of the flow duration curves resulting from the unit testing procedure. Whereas, the virtual-observed flow duration curve sits outside the 90% limits of the unit test flow duration curve when September is taken as the influencing month. Thereby providing further evidence that to improve September flows that September rainfall should be improved in preference to other months.*"

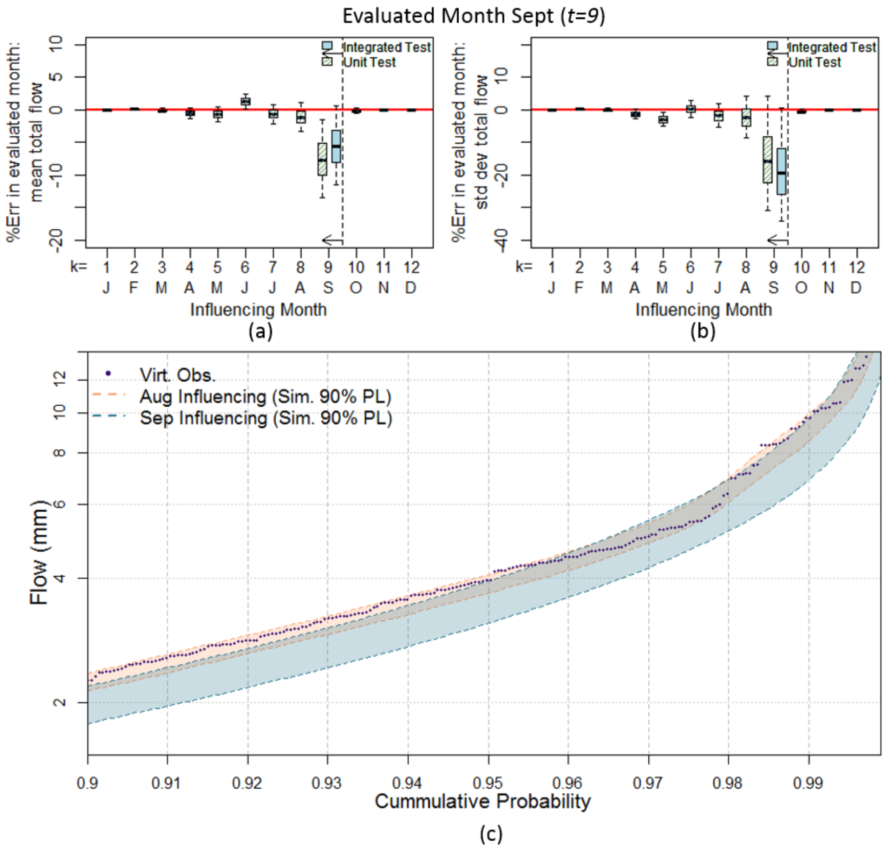


Fig. 1 Lobethal, Site 12 (90% limits shown) (a) unit test error in mean monthly flow (September), (b) unit test error in standard deviation of monthly flow (September), (c) Unit test September flow duration curve when August and September are selected as influencing months (top 10% of flow days shown).

Comment 4:

P2l23-25 The sentence is not clear without the explanations given in section 2. Either here more information are provided or the sentence is left out.

Response 4:

Thank you. The sentence has been left out in the revised manuscript.

Comment 5:

P3l9-14 The idea behind the example provided by the authors is clear. Nevertheless, some of the rainfall characteristics mentioned are not clear and, since it is only an example, can be left out or can be replaced by other rainfall characteristics:
- rainfall on wet days - What does this characteristic represent (the daily total rainfall itself is mentioned later)?;
- Extreme value analysis on a monthly basis and autocorrelation on an annual basis are from my understanding rather uncommon rainfall characteristics for the evaluation of rainfall time series

Response 5:

Due to a restructure of the paper the example does not appear in the revised manuscript.

Comment 6: P3l20-21 The details provided in brackets can be left out, since without reading the reference there are no additional information for the reader.

Response 6: Due to the shortening of the introduction and motivation of the new evaluation framework (see comment 7 below) the sentence no longer appears in the revised manuscript.

Comment 7:

P3l7-P4l20 The motivation for the introduction of the new evaluation strategy is quite long and can be shortened by the half. I think the majority of the community is quite aware of the issue with overlapping errors. Also Fig. 1 and Fig. 2 are quite clear from the text and could be left out. If kept, a box with "True rainfall" should be added in Fig.1a) to be consistent with Fig. 1b ("True streamflow")

Response 7:

The motivation for the new evaluation strategy has been restructured and shortened by 50 lines (from 167 lines to 127 lines) while maintaining the key points of the introduction and section 2.1 of the original manuscript. We have retained Fig. 1 and 2 but merged them (now Fig. 1 in the revised manuscript) and made the amendments suggested above. Based on our experiences explaining this work these figures are helpful to avoid misconceptions. For convenience the merged figure is reproduced below.

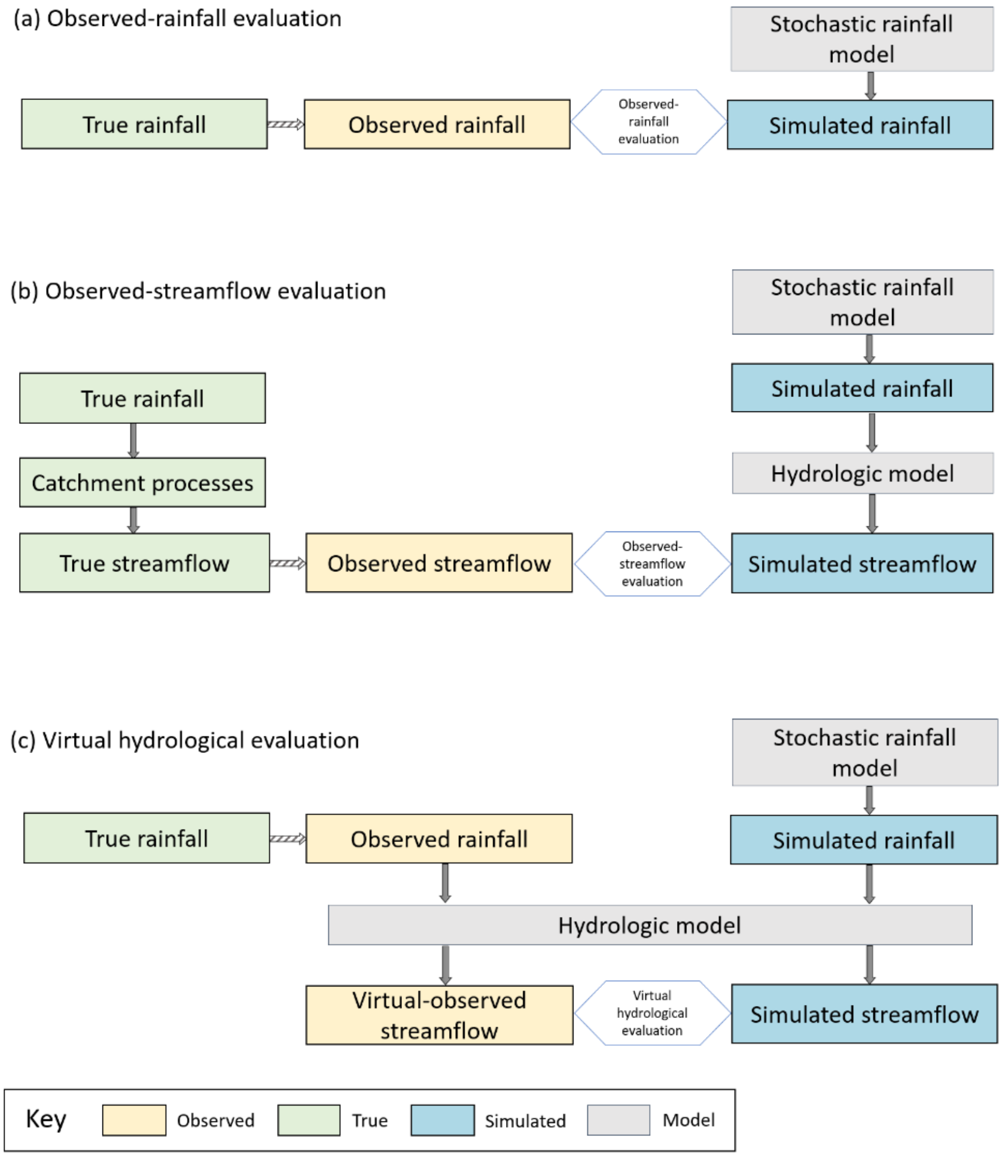

(a) Observed-rainfall evaluation
Stochastic rainfall model
True rainfall
Observed rainfall
Observed-rainfall evaluation
Simulated rainfall
(b) Observed-streamflow evaluation
Stochastic rainfall model
Simulated rainfall
True rainfall
Catchment processes
Hydrologic model
True streamflow
Observed streamflow
Observed-streamflow evaluation
Simulated streamflow
(c) Virtual hydrological evaluation
Stochastic rainfall model
True rainfall
Observed rainfall
Simulated rainfall
Hydrologic model
Virtual-observed streamflow
Virtual hydrological evaluation
Simulated streamflow
Key
Observed
True
Simulated
Model

Fig. 2 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 (c) virtual hydrological evaluation framework where simulated streamflow is compared against virtual-observed streamflow.

Comment 8:

P5l10 "to match streamflow observations" -> "to match streamflow observations or statistics"

Response 8: The sentence has been modified as suggested.

Comment 9:
P6Table1 The authors should include a definition of the applied symbols in the caption, since the difference between "x" and "-" is not too intuitive (from only the table). Is in the last line, first column something missing (virtual hydrological…)?

Response 9:

To avoid ambiguity Table 1 has been revised so that text (i.e. 'Yes' and 'No') is used in place of the original symbols and an additional column is now included that indicates whether the analysis is streamflow-based. The last line, first column has been amended to read 'virtual hydrological evaluation'. Table 1 is reproduced below for convenience.

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

	Source of error			Streamflow-based evaluation
	Stochastic rainfall model	Hydrological model	Observed streamflow	
Observed-rainfall evaluation	Yes	No	No	No
Observed-streamflow evaluation	Yes	Yes	Yes	Yes
Virtual hydrological evaluation	Yes	No	No	Yes

Comment 10:

P6Table 1 From my opinion the results from the virtual-observed streamflow approach can still be biased by the applied r-r model. For example, rainfall is generated in space and two rainfall generation methods show differences in terms of rainfall characteristics, but not in the simulated streamflow. After what I've read in the introduction and methods section, the conclusion is that the compared rainfall characteristics are then not practicable ("no impact") und useless (for the study region). But this also depends on i) the model choice (including e.g. spatial resolution, model type (fully / semi-distributed), several model approaches) and ii) the parameter identification. In a semi-distributed model differences in spatial rainfall could be dampened, while they are (maybe) not dampened in a fully-distributed model. The parameters have to be chosen a priori – a calibration on one of the rainfall data sets is not possible to avoid biases. Will the parameters be calibrated by an additional rainfall data set (the observed data) and if so, how can be avoided that this calibration introduces a bias (e.g. maybe the observed rainfall data is more similar to rainfall data set A under investigation than to B)? So all of the results depend on the chosen setup for the r-r simulations and drawn conclusions are only valid in context with the model setup and parameter set. This is of course always the case in hydrology, but it becomes more important if a virtual runoff time series is applied, since the "relation" between the model output and reality gets lost. However, the authors point these issues out later in their investigation (p20), but it should be communicated earlier to the reader.

Response 10:

The reviewer has raised some excellent discussion points. Although we have discussed some of these points later in the investigation (Section 5), we now raise these key issues earlier with the reader (Section 2.1).

We provide specific responses below to the discussion points raised.

1. "*Virtual-observed streamflow approach can still be biased by the applied r-r model*" – Yes, we agree. This is a very important matter to consider. As a result we acknowledge that virtual hydrological evaluation should use an appropriate hydrological model in Section 2 :

    Page 5, Lines 29-30: "*It is also important that the selected hydrological model is fit for purpose so that it can simulate the streamflow characteristics of interest.*"

    The need for an appropriate hydrological model is further discussed later in Section 5.3:

    Page 27, Lines 25-30: "*Hydrological structural errors may potentially skew interpretation of the SRM evaluation if the hydrological model poorly represents the catchment processes. To reduce these impacts the steps taken in this study included (a) using a well-tested hydrological model that has demonstrated good performance on a wide range of catchments (e.g. the GR4J model has been widely tested , see (Perrin et al., 2003, Coron et al., 2012); (b) calibrating and*

> *evaluating the hydrological on a catchment close to the observed rainfall sites to ensure it provided sufficiently good performance (e.g. GR4J was calibrated to the Onkaparinga catchment - see (Westra et al., 2014a, Westra et al., 2014b).”*

Future research on the application of the VHE will consider multiple hydrological models to avoid the reliance on single rainfall-runoff models and/or single calibration schemes. This discussion has been added to Section 5.3:

> Page 27. Lines 30-32: “*Future research will use multiple, well-tested hydrological models with sufficiently good performance to reduce the reliance on a single hydrological model and ensure the identification of SRM deficiencies is not dependent on a single hydrological model.”*

2. “*Differences in terms of rainfall characteristics, but not in the simulated streamflow*” – Yes, there is the potential that a chosen rainfall-runoff model is insensitive to certain important differences in modelled rainfall. While we agree with the reviewer that spatial patterns are important, temporal differences in rainfall patterns, which are capture by a lumped rainfall-runoff model are potentially equally as important.

3. “*But this also depends on i) the model choice … and ii) the parameter identification*” – Yes, as with the comments in (1), all elements of the modelling method can potentially introduce bias. The end-user’s impact of interest and associated modelling process can influence an outcome. These observations reinforce the need for care when applying the framework. This includes ensuring that (i) the rainfall-runoff model is ‘fit for purpose’,

> Page 5, Lines 29-30: “*It is also important that the selected hydrological model is fit for purpose so that it can simulate the streamflow characteristics of interest.”*

(ii) that this rainfall model is well-tested,

> Page 27, Lines 25-30: “*Hydrological structural errors may potentially skew interpretation of the SRM evaluation if the hydrological model poorly represents the catchment processes. To reduce these impacts the steps taken in this study included (a) using a well-tested hydrological model that has demonstrated good performance on a wide range of catchments (e.g. the GR4J model has been widely tested , see (Perrin et al., 2003, Coron et al., 2012)”*

and (iii) well calibrated.

> Page 27, Lines 25-30: “*(b) calibrating and evaluating the hydrological on a catchment close to the observed rainfall sites to ensure it provided sufficiently good performance (e.g. GR4J was calibrated to the Onkaparinga catchment - see (Westra et al., 2014a, Westra et al., 2014b).”*

A description of the hydrological model calibration approach is reproduced below:

> Page 14, Lines 1-17: “*The hydrological model GR4J (Perrin et al., 2003) was used to simulate virtual-observed streamflow at a daily time step. GR4J is a daily lumped hydrological model that simulates daily streamflow in a parsimonious manner using four parameters. The GR4J model was calibrated according to the procedure set out in Westra et al. (2014b) for the stationary version of the GR4J hydrological model. The details are provided in (Westra et al., 2014a) and a short summary is provided here. The multi-site rainfall gauges were Thiessen weighted to calculate the catchment average rainfall. The model was calibrated to the streamflow data at Houlgrave Weir (see Figure 4) using model calibration period of 15 years (1985-1999). The parameters were estimated using maximum likelihood estimation procedure with a weighted least squares likelihood function. The model parameters that maximised the likelihood function were found using a multi-start quasi-Newton optimisation procedure with 100 random starts. Overall, the GR4J model was a good fit to the observed streamflow, with a Nash-Sutcliffe efficiency of 0.8. A similar type of hydrological model and calibration approach has been used for other virtual evaluation studies (Li et al. 2014; 2016).*”

4. "*If so, how can be avoided that this calibration introduces a bias*" – This is an excellent question, which we will address with further discussion. While best-practice models and methods are important, this does not necessarily guard against the possibility that a model poorly represents key processes of

interest. A remedy for this limitation would be to use multiple rainfall-runoff models and this is now clearly mentioned as the first step of Future research in new Section 5.3, Limitations and future research.

> Page 27, Line 22-32: "*The formalisation of the virtual hydrological framework for SRM evaluation provides the opportunity for further improvements in the future, including:*
>
> *(i) Using multiple, well-tested hydrological models - a potential limitation of the virtual hydrologic evaluation framework is that it is reliant on the use of a hydrological model. Hydrological structural errors may potentially skew interpretation of the SRM evaluation if the hydrological model poorly represents the catchment processes. To reduce these impacts the steps taken in this study included (a) using a well-tested hydrological model that has demonstrated good performance on a wide range of catchments (e.g. the GR4J model has been widely tested , see (Perrin et al., 2003, Coron et al., 2012); (b) calibrating and evaluating the hydrological on a catchment close to the observed rainfall sites to ensure it provided sufficiently good performance (e.g. GR4J was calibrated to the Onkaparinga catchment - see (Westra et al., 2014a, Westra et al., 2014b). Future research will use multiple, well-tested hydrological models with sufficiently good performance to reduce the reliance on a single hydrological model and ensure the identification of SRM deficiencies is not dependent on a single hydrological model.*"

Comment 11:

P9Fig3b Maybe the authors can spend a more detailed explanation of the two different indices k and t. For me the difference was not quite clear at the beginning. Also, it is clear that rainfall in June can affect the runoff in July (or from April by filling storages and hence affecting runoff in July). But how can rainfall in July affect runoff in June, although the months August to January obviously don't? Is the rainfall information transformed into runoff over such a long period in the model? Since there is no rain in the summer half year, shouldn't the storages run empty?

Response 11:

We have more clearly indicated the meaning of indices *k* and *t* in the descriptive text (see Sections 2.4.1 and 2.4.2), Figure 3b (Figure 4b in the revised manuscript) and its caption to aid the reader.

> Section 2.4.1, Pages 9 - 11: "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(a), 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, k, and has the same length as the corresponding observed R^{obs} and simulated R^{sim} time series.
>
> $$R_{(k)}^{spl} = \bigcup_{m=1}^{12} \begin{cases} R_m^{sim}; m = k \\ R_m^{obs}; m \neq k \end{cases} \qquad (3)$$
>
> 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:
>
> $$R_{(6)}^{spl} = \{R_1^{obs}, \ldots, R_5^{obs}, R_6^{sim}, R_7^{obs}, \ldots, R_{12}^{obs}\} \qquad (4)$$
>
> The ensemble of $k = 1,\ldots,12$ spliced rainfall time series $R_{(k)}^{spl}$ for all influencing months and additional inputs (e.g. potential evapotranspiration) indicated by '…' are transformed according to a hydrological

model $g[\,]$ to produce an ensemble of simulated streamflow, $Q_{(k)}^{spl}$. This procedure is repeated for all simulated rainfall replicates.

$$Q_{(k)}^{spl} = g\left[R_{(k)}^{spl}, \dots\right] \tag{5}$$

By construction, the 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.

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 results of the unit test and the integrated test (Steps 1-2) are then investigated and compared selecting each month as the evaluated time period in turn as well as other key time periods (e.g. annual). ”

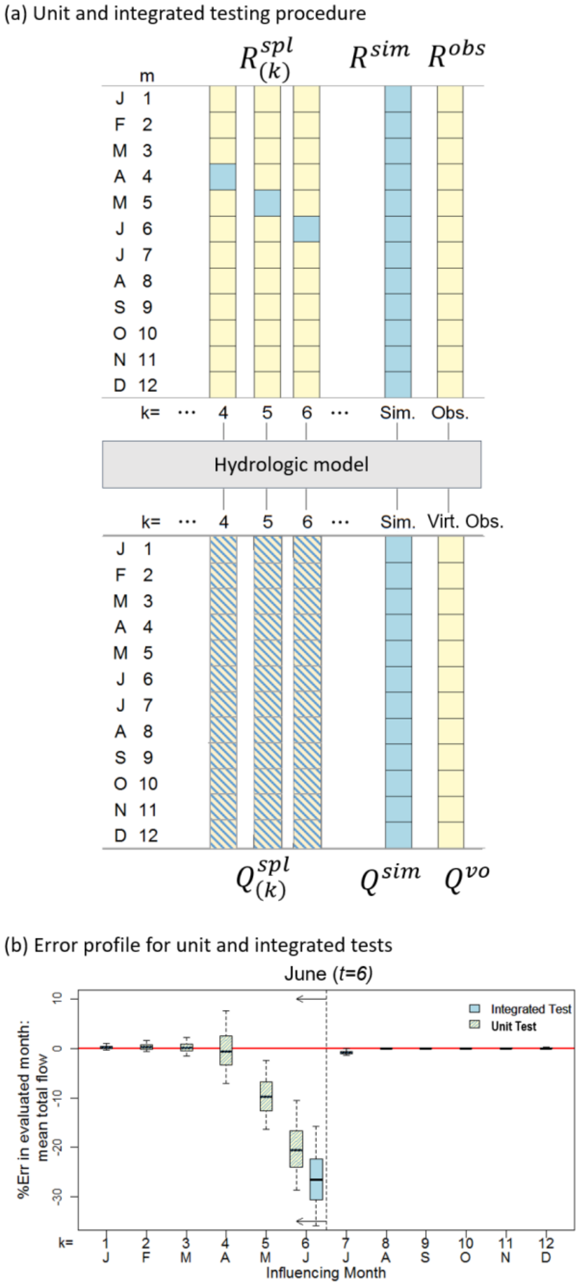


Fig. 3 Schematic of (a) the method of constructing a unit test by embedding simulated months in an observation time series, and (b) the error profile produced when using the integrated and unit tests for the evaluated time period of June (*t*=6) (box plot whiskers

indicate the 90% limits of the simulated streamflow replicates). For the unit test the errors in the evaluated period (t) are calculated as the difference between $Q_{(k)}^{spl}$ and $Q_{(t)}^{vo}$. For the integrated test the errors are calculated as difference between Q^{sim} and $Q_{(t)}^{vo}$.

> Section 2.4.2, Page 11, Lines 11-20: "Using the function $h[\,]$ to denote a calculated statistic of interest (e.g. mean or standard deviation), the relative error in an evaluated time period t (e.g. annual or particular month) is given by
>
> $$\%Err_{(t)} = \frac{h\left[Q_{(t)}^{eval}\right]-h\left[Q_{(t)}^{vo}\right]}{h\left[Q_{(t)}^{vo}\right]} \times 100 \qquad (6)$$
>
> where $Q_{(t)}^{vo}$ is the virtual-observed streamflow and $Q_{(t)}^{eval}$ is the simulated streamflow from the selected virtual hydrologic test (i.e. Q^{sim} if integrated test or $Q_{(k)}^{spl}$ if unit test selected) in the evaluated time period t. This procedure is repeated for all replicates of the simulated streamflow such that a range of errors is reported for each test for the target time period.
>
> Following the calculation of this error metric for all replicates of the integrated test and ensemble of unit tests ($k = 1, \dots, 12$) it is possible to investigate deficiencies in the simulated streamflow in terms of which influencing month(s) contribute more to the deficiencies in streamflow for the target time period based on that statistic of interest. Thus, for each site, statistic and evaluated time period there are 13 sets of errors to compare."

A unit test is undertaken by evaluating the ensemble of simulated streamflows from transforming the spliced rainfall from the 12 potential influencing months for an evaluated month, *t*. We believe it is necessary to evaluate all 12 potential influencing months because *a priori* the impacts of 'poor' rainfall can have long-term impacts on streamflow statistics due to catchment storage in the rainfall-runoff model. Some catchment models have short-term stores to represent features such as depressions, basins, and channels, but other catchments models have long-term stores to represent the long-term memory in subsurface catchment storages) that can have memory over multiple months. For the case study catchment the storages do not run empty each year in summer, so there is potential for persistence at longer timescales due to this 'memory' in the catchment. Therefore, it is plausible that rainfall from 12 months prior can influence the current state of a catchment (especially if that month/season was anomalously wet or dry). An additional figure of monthly rainfall and streamflow boxplots is now provided to illustrate the highly seasonal nature of the case study catchment in Section 3 (also see response to comment 14). The new figure (Fig. 6) is reproduced below for convenience.

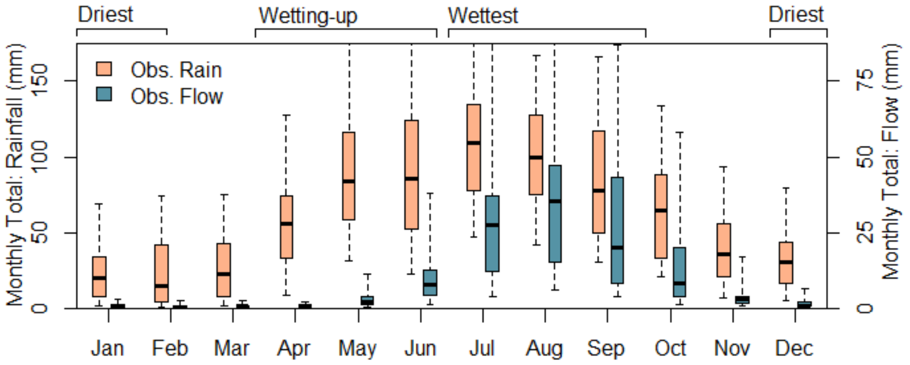


Fig. 6 Seasonal variation of catchment average rainfall and flow at Houlgrave Weir. Boxplots show the variation across years. Wettest indicates (high rainfall, high flow), direst indicates (low rainfall, low flow) and wetting-up indicates (high rainfall, low flow).

Comment 12:

P9Section 2.4 It would be useful for the reader to illustrate the implementation of the framework with a flow chart, since the authors use step 1, step 6 and so on throughout the section (and the manuscript).

Response 12:

We have incorporated a flow chart to illustrate the framework procedure in Section 2.1 (see Fig. 2 in the revised manuscript). For convenience the flow chart is reproduced below.

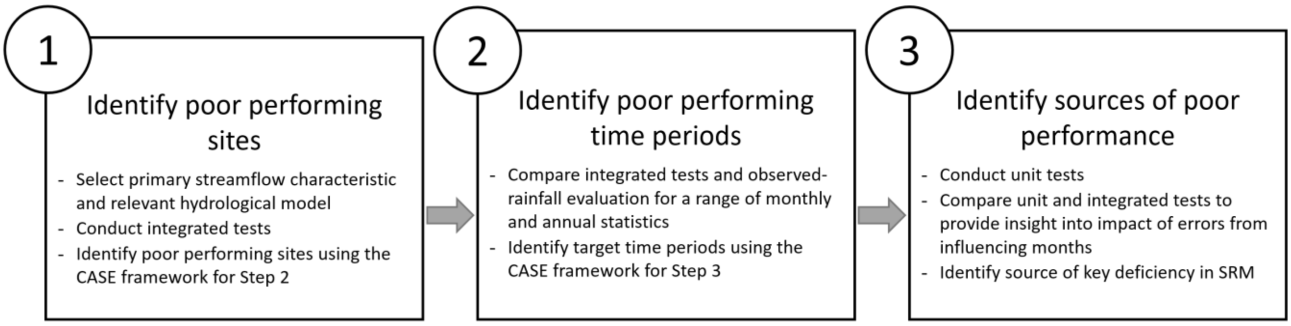

1
Identify poor performing sites
- Select primary streamflow characteristic and relevant hydrological model
- Conduct integrated tests
- Identify poor performing sites using the CASE framework for Step 2
2
Identify poor performing time periods
- Compare integrated tests and observed-rainfall evaluation for a range of monthly and annual statistics
- Identify target time periods using the CASE framework for Step 3
3
Identify sources of poor performance
- Conduct unit tests
- Compare unit and integrated tests to provide insight into impact of errors from influencing months
- Identify source of key deficiency in SRM

Fig. 2 - Virtual hydrological evaluation procedure.

In the revised manuscript the framework steps have been streamlined (from seven steps down to three) to further improve the presentation of the framework and Section 2 has been restructured to better integrate the framework steps and the specific virtual tests (i.e. integrated and unit tests). The structure of Section 2 is now as follows:

2. Virtual hydrological framework
2.1 Overview
2.2 Step 1 – Identify poor performing sites
2.2.1 Selection of primary streamflow characteristic and relevant hydrological model
2.2.2 Integrated test procedure
2.2.3 Identify poor performing sites using CASE framework
2.3 Step 2 – Identify poor performing time periods
2.4 Step 3 – Identify sources of poor performance
2.4.1 Unit test procedure
2.4.2 Compare unit tests and integrated tests
2.4.3 Identify types of key deficiencies

Section 4 (Results) has also been restructured to follow the streamlined framework procedure. The structure of Section 4 is now as follows:

4. Results
4.1 Step 1 – Identify poor performing sites
4.2 Step 2 – Identify poor performing time periods
4.3 Step 3 – Identify sources of poor performance
4.3.1 Streamflow errors mostly originate from rainfall model deficiencies in the evaluated month
4.3.2 Streamflow errors originate from rainfall model deficiencies over a contiguous block of months
4.3.3 Streamflow errors originate from rainfall model deficiencies in a preceding month more so than evaluated month
4.3.4 Influence of monthly rainfall on annual flow volumes

Comment 13:
P10l14-17 What is the 90 % limit of the simulated statistic? If m=10 mm, everything between 1 mm and 19 mm is considered as good? Here an additional explanation is required.

Response 13:
Thank you. We have added additional information to explain the 90% limit test as requested both in text and graphically (see Section 2.2.3, Table 2 and Fig. 3). The text and figures are reproduced below for convinience:

> Page 7, Line 15-22: “ *The integrated test results aim to identify the sites that are poor performing for the primary streamflow characteristic. Model performance is categorised using a CASE framework approach as ‘good’, ‘fair’ or ‘poor’ following Bennett et al. (2018). The quantitative tests for each performance category are provided in Table 2 alongside an illustration of each in Fig. 4. The quantitative tests proceed by comparing the statistics of the virtual-observed streamflow against those calculated from replicates of the simulated streamflow. Performance was categorised as ‘good’ if the selected statistic for the virtual-observed streamflow fell within the 90% limits of the statistic calculated from the simulated streamflow replicates (Fig. 4, case i), as ‘fair’ if the virtual-observed statistic fell outside the 90% limits of the simulated streamflow replicates but within the 99.7% limits (Fig. 4, case ii) and otherwise as ‘poor’ (Fig. 4, case iii).”*

Table 2 CASE performance classification criteria. Adapted from Bennett et al. (2018).

Performance Classification	Test	Key
‘good’	Observation lies within the 90% limits (case i)	🟩
‘fair’	Observation lies outside the 90% limits but within the 99.7% limits (case ii)	🟨
‘poor’	Otherwise (case iii)	🟥

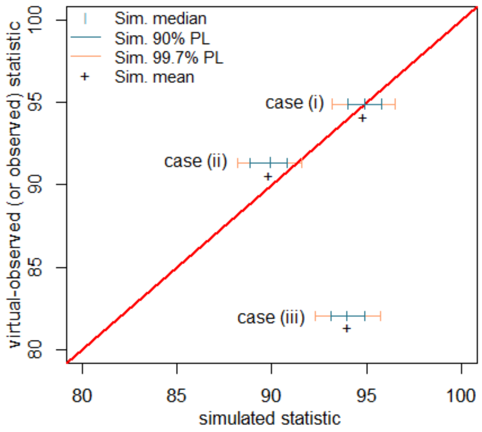


Fig. 4 Illustration of performance classification, case (i) shows ‘good’ performance, case (ii) shows ‘fair’ performance and case (iii) shows ‘poor’ performance. Adapted from Bennett et al. (2018).

Comment 14:
P11l9-10 From Table 2 it cannot be seen, how long the time series used for the calibration of the rainfall generator are. It would be useful to the reader to characterize the time series more in detail (wet spell durations and amount, dry spell durations and maybe even on a monthly basis, since further investigations are carried out on a monthly basis). At least a hint to Fig. 6 and Fig. 7, which include some monthly observations, would be useful.

Response 14:
We agree with the reviewer and recognise that the high-level summaries need more tangible details on the rainfall and streamflow statistics on a monthly basis to help the reader understand the seasonal behaviour of the case study catchment. We have revised Table 2 (now Table 3) to characterise the rainfall time series in more detail including the addition of columns that present rainfall statistics (total rainfall, no. of wet days,

average daily rainfall, average wet day length, average dry spell durations) in different seasons – for brevity in this table we show two months: January to represent the dry summer and July to represent the wet winter. Further detail on these statistics for all months at each site is also provided as supplementary material. Table 3 is reproduced below:

Table 3 Site names, locations and seasonal rainfall characteristic summary. Sites ordered from lowest to highest elevation.

				January					July				
Site No	Site Name	Elev (m)	Ann. Av. Rain (mm)	Total (mm)	No. Wet (days)	Daily Av. (mm)	Wet-spell (days)	Dry-spell (days)	Total (mm)	No. Wet (days)	Daily Av. (mm)	Wet-spell (days)	Dry-spell (days)
19	Old Noarlunga	7	520	20	4.1	0.6	1.6	9.6	72	17	2.3	3.1	2.8
16	Morphett Vale	90	560	20	4.1	0.6	1.5	8.9	76	17	2.4	3.3	2.8
10	Happy Valley	148	640	22	4.8	0.7	1.7	8.2	88	18	2.8	3.6	2.6
21	Willunga	158	640	23	4	0.7	1.6	10	95	17	3	3.2	2.7
5	Clarendon	223	820	25	4.7	0.8	1.7	8.9	114	17	3.7	3.4	2.8
6	Coromandel	234	710	24	4.8	0.8	1.8	9.2	102	18	3.3	3.6	2.8
13	Macclesfield	302	730	28	5.3	0.9	1.8	7.9	99	17	3.2	3	2.7
15	Cudlee Creek	311	830	29	5	0.9	1.8	8.4	123	18	3.9	3.8	2.7
11	Harrogate	335	550	23	3.5	0.7	1.6	12	75	12	2.4	2.2	3.8
4	Cherry gardens	345	920	30	5.4	1	1.8	7.7	134	18	4.3	3.8	2.6
8	Gumeracha	346	790	27	5.3	0.9	1.8	7.8	108	18	3.5	3.5	2.6
9	Hahndorf	347	850	29	5.4	0.9	1.9	8.1	123	18	4	3.4	2.7
17	Mount Barker	349	770	28	5.9	0.9	1.9	7.2	104	18	3.3	3.3	2.6
7	Echunga	375	805	28	5	0.9	1.8	8.7	110	17	3.5	3.3	2.6
3	Bridgewater	376	1050	32	5.2	1	1.9	8.9	154	18	4.9	3.6	2.7
14	Meadows	384	870	30	4.8	1	1.7	8.5	122	17	3.9	3.2	2.7
2	Birdwood	385	720	25	4.4	0.8	1.8	9.6	104	17	3.4	3.4	2.8
1	Belair	386	790	28	4.6	0.9	1.8	9.8	111	16	3.6	3.2	3
22	Woodside	387	800	27	4.3	0.9	1.6	8.3	121	16	3.9	2.9	2.7
18	Nairne	403	680	28	4.7	0.9	1.6	8	93	16	3	2.8	2.8
12	Lobethal	470	880	28	4.9	0.9	1.8	8.4	133	18	4.3	3.5	2.6
20	Uraidla	499	1090	35	4.7	1.1	1.8	9	161	17	5.2	3.4	2.7

In addition, a new figure that shows seasonal variation of catchment average rainfall and streamflow on a monthly basis has been added to Section 3 (Case Study) of the main paper.

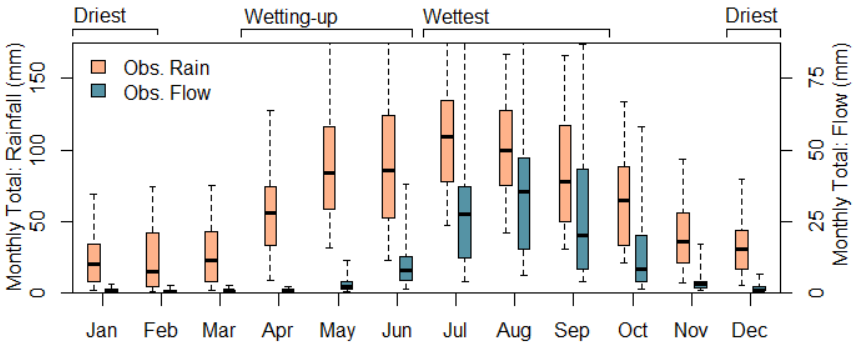


Fig. 6 - Seasonal variation of catchment average rainfall and flow at Houlgrave Weir. Boxplots show the variation across years. Wettest indicates (high rainfall, high flow), direst indicates (low rainfall, low flow) and wetting-up indicates (high rainfall, low flow).

Comment 15:

P11l16 For the calibration of the model the reader is referred to Westra et al. (2014), which is a non-reviewed technical report with 100+ pages, as far as I can see. In context with my former specific comment it is necessary to provide information in the actual manuscript, how the model has been calibrated. Which rainfall data was used for the calibration? If all 22 stations have been applied, how was the areal rainfall estimated as input for the lumped r-r model?

Response 15:

The calibration of the hydrological model was actually based on the approach used in a water resources research paper (Westra et al., 2014a) is now be cited alongside the report. The paper provides a compact peer-reviewed summary of the model and its calibration – for a neighbouring catchment (Scott Creek). This paper was acknowledged with a Research Spotlight Award from American Geophysical Union (top 5% of papers in

AGU). The reference to the report, Westra et al. (2014b), is also retained since it gives details specific to the Onkaparinga catchment used in this paper and because it is comprehensive. The Scott Creek and Onkaparinga catchments were calibrated as part of the same project using consistent models and techniques.
The relevant aspects of the calibration and model-selection have been added to the manuscript.
Page 13, Lines 1-12:

> *"The hydrological model GR4J (Perrin et al., 2003) was used to simulate virtual-observed streamflow at a daily time step. GR4J is a daily lumped hydrological model that simulates daily streamflow in a parsimonious manner using four parameters. The GR4J model was calibrated according to the procedure set out in Westra et al. (2014b) for the stationary version of the GR4J hydrological model. The details are provided in (Westra et al., 2014a) and a short summary is provided here. The multi-site rainfall gauges were Thiessen weighted to calculate the catchment average rainfall. The model was calibrated to the streamflow data at Houlgrave Weir (see Figure 4) using model calibration period of 15 years (1985-1999). The parameters were estimated using maximum likelihood estimation procedure with a weighted least squares likelihood function. The model parameters that maximised the likelihood function were found using a multi-start quasi-Newton optimisation procedure with 100 random starts. Overall, the GR4J model was a good fit to the observed streamflow, with a Nash-Sutcliffe efficiency of 0.8. A similar type of hydrological model and calibration approach has been used for other virtual evaluation studies (Li et al. 2014; 2016). The same set of hydrological model parameters are used for both the unit and integrated tests so that the same transformation of rainfall to flow is used."*

Comment 16:
P11Section 3 Although the observed discharge time series is not used in the investigation, it would be useful for the reader to provide some runoff characteristics (e.g. mean discharge) to get a feeling for the catchment.

Response 16:
Details of the catchment's runoff characteristics at the annual and seasonal level have been added to Table 2 (now Table 3 in the revised manuscript) and a new figure (Fig. 6) showing the catchment's seasonal runoff and rainfall characteristics has also been added (see response to comment 13). Additional descriptions of key portions of the catchment cycle (i.e. "wettest", "driest", "wetting-up") have also been added and used throughout the manuscript. The new figure is reproduced below for convenience.

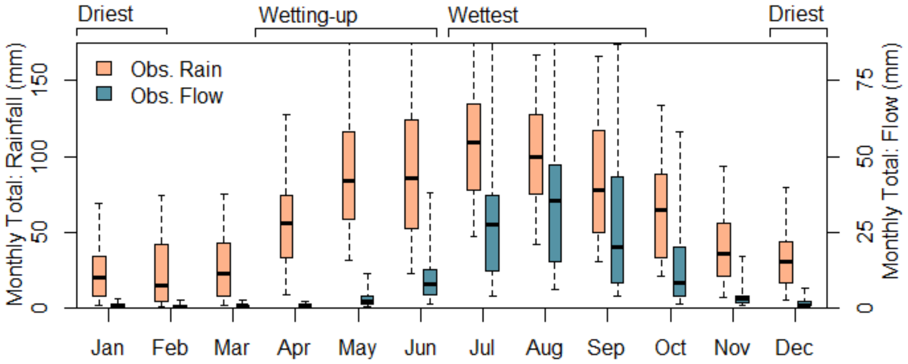


Fig. 6 - Seasonal variation of catchment average rainfall and flow at Houlgrave Weir. Boxplots show the variation across years. Wettest indicates (high rainfall, high flow), direst indicates (low rainfall, low flow) and wetting-up indicates (high rainfall, low flow).

Comment 17:
P11l14-18 On p9l15-17 you mention "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." Is the lumped model able to simulate the physical processes of a catchment with a few 100 km2 catchment area (I could not find the catchment area in the manuscript).

Response 17:
Thank you for pointing this out. The catchment area (323 km^2) has been included in Section 3 of the revised manuscript.

It is important that the chosen hydrological model is fit for purpose (see also discussion of comment 10). The GR4J model used in this paper is for catchment inflows to the Mount Bold reservoir and is appropriate for analysis of catchment yield (i.e. focussed on means and variances of annual inflow)[2].

However, if we were examining impacts on instantaneous peak flows impacts, this model would not be suitable and if we wanted to look at impacts of distributed rainfall, we would need a distributed rainfall-runoff model. However, for the purpose of this paper, which is to demonstrate the virtual hydrological framework (including the unit test) for evaluating the ability of a stochastic rainfall model to estimate catchment yield, the model is deemed sufficient.

Comment 18:
P12l8-10 Which result is analyzed? Integrated test or unit test?

Response 18:

We agree this is not clear, thank you for pointing it out. The result is from the integrated test (step 4). This is clarified in the revised text.

Comment 19:

P19l19 In Fig. 5 the results for rainfall are worse than for runoff (for mean values).

Response 19:
There are many interesting features of Fig 5 (now Figure 7) like this. The fact that there is not a direct correspondence between 'good' rainfall and 'good' runoff, and/or 'poor' rainfall and 'poor' runoff is one of the motivations for the virtual hydrological evaluation framework in addition to observed rainfall-based evaluation. Figure 7 shows that it is possible for seemingly 'poor' rainfall to yield 'good' runoff (as also 'good' rainfall can yield 'poor' runoff). We note that the discrepancy in Figure 7 is not in terms of mean values, but for the standard deviation of monthly aggregates (see Figure 5, *sd*(total) for rainfall and runoff in Jan, Mar, May, Jun, Oct, Nov, Dec). For drier months (Nov-Mar) the lack of correspondence (i.e. 'poor' rainfall producing 'good' runoff) is due to the low amount of runoff. While in wetter months (May-Oct) the relationship is more complicated as shown in the unit test demonstrations (Section 4.3).
The description of this Figure has been revised to better explain these features in Section 4.2 of the revised manuscript and is reproduced below.

> Page 15, Line 11 – Page 16, line 12: "*The poor performing sites identified in Step 1 were then compared in terms of both an observed-rainfall evaluation and virtual hydrological evaluation via an integrated test. Fig. 7 graphically summarises this comparison, with each row presenting monthly or annual performance of the following statistics:*
>
> - *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*

[2] The GR4J model has a calibrated Nash Sutcliffe of 0.8 (reported in the original manuscript). This model (and its non-stationarity variants) were used to project climate change impact on the Onkaparinga catchment (Westra et al. 2014b).

- *aggregate streamflow statistics (mean and standard deviation of total flow).*

The first to fourth columns of Fig. 7 summarise the observed-rainfall evaluation and the fifth and sixth of Fig. 7 summarise the virtual hydrological evaluation. The first column of Fig. 7 indicates that of the poor performing sites the SRM exhibited 'good' performance in simulating daily rainfall means and standard deviations as well as the mean number of wet days for all sites and months and at an annual level according to the observed-rainfall evaluation. Whereas the second column indicates that there is mixed performance across sites and months in simulating the variability in the number of wet days (sd(nwet)). Likewise, the third and fourth columns indicate overall 'good' performance in simulating mean monthly totals and mixed performance in simulating the monthly or annual total standard deviations (sd(total)). Whereas the virtual hydrological evaluation (fifth and sixth) columns show mostly 'good' performance in all months other than those in the 'wettest' or 'wetting-up' periods.

A clear trend, from Fig. 7 is the contrast in performance between the observed-rainfall evaluation and the virtual hydrological evaluation. One constrast is that, in the driest months (Dec, Jan, Feb) 'poor' performance in simulating rainfall (based on observed-rainfall evaluation) did not necessarily translate to 'poor' performance in simulating streamflow (based on virtual hydrological evaluation). For example, examining the first row of Fig. 7, the observed-rainfall evaluation shows that in January the SRM's ability to simulate variability in the number of wet days, sd(nwet), was 'poor' for all sites. However, in contrast the virtual hydrological evaluation shows that most sites had 'good' performance in simulating the January distribution of monthly total flow (i.e. m(total) and sd(total)).

A second contrast is that 'good' performance in the observed-rainfall evaluation does not necessarily translate to 'good' performance for the virtual hydrological evaluation, particularly for months in the 'wettest' and 'wetting-up' periods. For example, in Fig. 7 the rows summarising June and August show large percentages of 'poor' sites in the virtual hydrological evaluation of 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. 7 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."

Comment 20:
P19 Results-section Before it was mentioned that also the influence of spatial rainfall patterns can be evaluated. Since this is not done in the manuscript, it can be moved to the outlooks of the manuscript. Otherwise a spatial analyses can be implemented in the manuscript (what I would recommend), to show further advantages of the unit test.

Response 20:
This concept has not been demonstrated in the main paper, therefore it has been deferred to the discussion on outlooks (Section 5.3). We believe the unit test has sufficient novelty to represent a substantial contribution, hence we will the leave spatial rainfall evaluation for future developments.

Comment 21:

P20l20-21 With the introduced framework it is still not possible to identify, which rainfall characteristics are important for streamflow prediction. Based on the high non-linearity of the rainfall-runoff transformation process, a single rainfall characteristic cannot be sufficient to draw conclusions about the impact on the resulting runoff. If this would be the case, r-r models wouldn't have to be used anymore. However, could the authors identify, based on their analysis, which rainfall characteristics are important for the resulting runoff behaviour? (of course, the results depend on the study site, model choice and so on, but nevertheless…)

Response 21:

The comment is correct that the proposed method does not identify the impact of specific singular rainfall characteristics on the resulting runoff. However, the framework does provide a clear approach to isolate which set of components of the rainfall model require further attention. The integrated test focuses attention on hydrological properties, and the unit test can isolate deficiencies in rainfall by month. When applied to the case study in our paper, the limitations of the model are in the variability of the rainfall and not in the rainfall mean. The initial motivation for the approach can be seen in Figure 7 (formerly Figure 5) where the mean of the annual rainfall is 'good', but the mean of the annual runoff is 'poor'. Figure 7 also shows that this is mostly attributed to 'poor' mean runoff in June, July and August. Unit tests were then used to show that the rainfall in the catchment 'wetting-up' period (May-June) is of key importance. This is greater insight than could have been achieved with observed-rainfall evaluation and is greater insight than could be gathered from other virtual-observed streamflow approaches.

However, we agree with the reviewer that the framework cannot currently distinguish between particular features of the rainfall (e.g. "Is it rainfall correlation, magnitude, or intermittency that causes a low standard deviation in monthly streamflow?"). Nonetheless, the framework has significant potential to be extended to diagnose which are rainfall characteristics should be improved. This can be done by comparing multiple rainfall model variants (parametrically, or via bootstrap techniques) which are designed to have contrasting features of a key characteristic (e.g. intermittency, rainfall correlation). Such an approach was undertaken by Evin et al. (2018) using an observed-rainfall evaluation approach to compare model variants. In the revised manuscript, this is identified as a limitation and this extension is highlighted for further research (Section 5.3).

> Page 27, Line 22 - Page 28, Line 7: *"The formalisation of the virtual hydrological framework for SRM evaluation provides the opportunity for further improvements in the future, including: …*
>
> *(ii) Comparison of SRMs – this framework can be extended to provide more direct guidance on which rainfall features (in terms of components of the SRM) should be modified to improve streamflow performance. This can be done by comparing multiple rainfall model variants (parametrically, or via bootstrap techniques) which are designed to have contrasting features of a key characteristic (e.g. intermittency, rainfall correlation). Such an approach was undertaken by Evin et al. (2018) using an observed-rainfall evaluation approach. If the SRMs have monthly/seasonal autocorrelation (these were not significant for the rainfall in the Onkaparinga catchment) the unit testing approach would need to be extended by conditionally sample the simulated rainfall in a manner that preserves monthly correlations."*

Comment 22:
P21l32-p22l1 This example is hard to follow, maybe the authors can extend it. From my understanding it depends on the calibration of the storage coefficients. If storage coefficients are small, the results from the monthly rainfall will be transferred to runoff immediately. This would be possible with the "traditional“ approach.

Response 22:
Thank you for pointing out that the example is hard to follow. The example and paragraph framing the example have been revised to better step through the example (i.e. what elements come from the integrated test or unit test). We have also added additional information on catchment seasonality in Section 3 to better explain the importance of the 'wetting-up' months and storage in the catchment (also see responses to comments 14 and 16).

Comment 23:
P23 There is a reference of Li et al. (2015b), but no Li et al. (2015a). Also Li et al. (2016) is mentioned before Li et al. (2015b)

Response 23:
Thank you, this has been corrected.

Response to Reviewer 2

Major Comments

Comment 1:

The authors present a streamflow-based evaluation framework to assess the adequacy of hydrologic predictions from stochastic rainfall generators (SRGs). This is a "virtual framework" in that it benchmarks these predictions against streamflows produced by historical continuous simulations, rather than observed streamflow timeseries. The authors point out that this avoids the complicating issues of model structural errors. This is a useful approach to benchmarking SRGs, and could perhaps be applied to other fluxes of interest (not just streamflow) for which long-term observation records aren't just available.

Response 1:

We are pleased that the reviewer found the approach useful for evaluating stochastic rainfall models and appreciated the wider potential application of the approach for other fluxes of interest.

Comment 2:

I agree with the first reviewer, who stated that "the theoretical elements of the paper are very long". There seems to be a fair bit of repetition, or at least over-explanation, of the motivation, and I strongly recommend that the authors look closely at how Section 2 can be shortened.

Response 2:

We have restructured the Introduction and Section 2 to shorten the presentation (and remove repetition) of the theoretical elements by 50 lines while maintaining the key points. Please also see the response to Reviewer 1, comment 7.

Comment 3:

Generally, I think that the demonstration would be more illuminating if the authors used it to compare two or more SRGs and/or hydrologic models.

Response 3:

We support this idea. In the future, the framework can be used to compare two or more SRG's for particular hydrological applications. Furthermore, by utilising two or more hydrological models in the virtual evaluation framework, it would reduce the dependence on the choice of hydrological model (which was raised by reviewer #1 see comment 10), because one could look for patterns of errors for a single SRG across two hydrological models.

However, our preference is not to include this in this paper, for the following reasons:

1. To include the details of a second SRG and/or a second hydrological model, as well as providing a complete explanation of the details of the framework, observed-rainfall evaluation, virtual observed streamflow evaluation, two different tests, the integrated and unit tests, would make the paper and/or analysis overly long. Reviewer #1 has asked for extra details and additional figures/tables to explain a wide range of details – including the hydrological model calibration (see reviewer #1 comment 15), extra streamflow analysis (see reviewer #1 comment 3). This is over and above the seven figures already included. If we included another SRG and hydrological model, the number of the figures could increase dramatically making the paper overly long and lose focus on the presentation of the framework.
2. The framework has not been presented before, in particular the unit test. Therefore, in the manuscript we present the evaluation of a single stochastic rainfall model to demonstrate the framework. This application to a single model has demonstrated some new insights; that the errors in streamflow for a

> particular month can be affected by errors in the in rainfall from the previous 2 to 3 months. This innovation is something that has not been previously identified in the literature.

Once the framework has been established and explained in this paper, future work will undertake multiple SRG and/or multi-hydrological model comparisons, as suggested by the reviewer. A comment on this has been incorporated into the discussion (Section 5).

> Page 27, Line 22 - Page 29, Line 7: *"The formalisation of the virtual hydrological framework for SRM evaluation provides the opportunity for further improvements in the future, including:*
>
> *(i) Using multiple, well-tested hydrological models - a potential limitation of the virtual hydrologic evaluation framework is that it is reliant on the use of a hydrological model. Hydrological structural errors may potentially skew interpretation of the SRM evaluation if the hydrological model poorly represents the catchment processes. To reduce these impacts the steps taken in this study included (a) using a well-tested hydrological model that has demonstrated good performance on a wide range of catchments (e.g. the GR4J model has been widely tested , see (Perrin et al., 2003, Coron et al., 2012); (b) calibrating and evaluating the hydrological on a catchment close to the observed rainfall sites to ensure it provided sufficiently good performance (e.g. GR4J was calibrated to the Onkaparinga catchment - see (Westra et al., 2014a, Westra et al., 2014b). Future research will use multiple, well-tested hydrological models with sufficiently good performance to reduce the reliance on a single hydrological model and ensure the identification of SRM deficiencies is not dependent on a single hydrological model.*
>
> *(ii) Comparison of SRMs – this framework can be extended to provide more direct guidance on which rainfall features (in terms of components of the SRM) should be modified to improve streamflow performance. This can be done by comparing multiple rainfall model variants (parametrically, or via bootstrap techniques) which are designed to have contrasting features of a key characteristic (e.g. intermittency, rainfall correlation). Such an approach was undertaken by Evin et al. (2018) using an observed-rainfall evaluation approach. If the SRMs have monthly/seasonal autocorrelation (these were not significant for the rainfall in the Onkaparinga catchment) the unit testing approach would need to be extended by conditionally sample the simulated rainfall in a manner that preserves monthly correlations."*

Comment 4:

The demonstration of these methods is provided at the monthly timescale. While this timescale might be useful for applications of water supply, it is not meaningful for flood processes in all but the very largest watersheds. It is easy to picture a hydrologic model that produces adequate performance in terms of monthly flows, but not daily or subdaily extremes, while the opposite is also possible. Similarly, it is also probably an easier task to create a stochastic rainfall generator that works well for producing monthly means and associated variability than fine-scale extremes. Thus, the virtual framework in this manuscript may not be as broadly useful for extremes as the authors claim (at least flood extremes, droughts might be a different story). I thus recommend that the authors acknowledge this shortcoming, and "tone down" the framework's purported usefulness for flood risk (e.g. page 9 lines 17), since this remains unproven.

Response 4:

Thank you. We have toned done the discussion of the framework's usefulness for flood risk applications as this is not demonstrated in the manuscript. For example, we have removed the example that referred to flood risk which appeared on page 9 line 17 of the original manuscript so the sentence now reads:

> "*For example, the distribution of annual total flow would be a suitable characteristic when investigating yield"*

Reviewer 1 (see comment 3) has suggested that we examine additional streamflow characteristics, in particular flow duration curves. We have done so and included the flow duration curves to provide supporting evidence of the identified deficiencies in the simulated rainfall as part of the framework. This provides a broader demonstration of the framework through an application that considers statistics at the finer (daily) scale.

Comment 5:

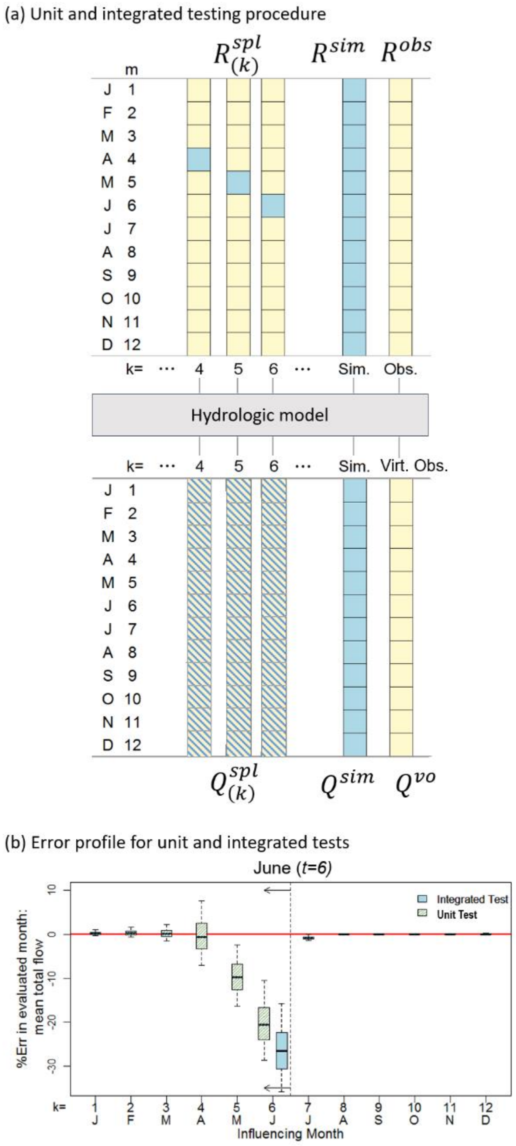
It isn't clear how the boxplots (e.g. figure 3) are constructed. Is it the "13 errors" mentioned in page 8 line 24? Or is it somehow derived from the 10,000 synthetic rainfall years? Or the 73 years of observed data with synthetic rainfall "spliced in"? Either way, it isn't clear that the authors have avoided the proliferation of error metrics that they identify as a limitation of previous measures on page 3. If this method is applied to a large number of sites, it still seems like a not-entirely compact evaluation scheme. Perhaps the authors could clarify how this compares to other methods in this respect.

Response 5:

Thank you for pointing this out. A related response is given to reviewer 1, comment 11. We have revised the example which explains how the unit test figure (those like Figure 4b in the revised manuscript) are constructed and used as a diagnostic (see Section 2.4.2 and 2.4.3). The companion figure (4b) and the caption has also been improved to step the reader through the figure.

The revised text clarifies that each boxplot is a summary of the error (the difference between the simulated and virtual-observed performance statistic, Eq. 6) across all replicates of the simulated time series. The "13 errors" relates to the number of boxplots displayed in figures of this type (the unit test for 12 influencing months and 1 integrated test). The description is reproduced below.

> Section 2.4.1, Pages 9 - 11: "Following Fig. 3(a), 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 $m \in \{1, \dots, 12 | \, m \neq k\}$. The resulting spliced rainfall time series $R_{(k)}^{spl}$ is denoted with respect to the influencing month, k, and has the same length as the corresponding observed R^{obs} and simulated R^{sim} time series.
>
> $$R_{(k)}^{spl} = \bigcup_{m=1}^{12} \begin{cases} R_m^{sim}; m = k \\ R_m^{obs}; m \neq k \end{cases} \qquad (3)$$
>
> 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:
>
> $$R_{(6)}^{spl} = \{R_1^{obs}, \dots, R_5^{obs}, R_6^{sim}, R_7^{obs}, \dots, R_{12}^{obs}\} \qquad (4)$$
>
> The ensemble of $k = 1, \dots, 12$ spliced rainfall time series $R_{(k)}^{spl}$ for all influencing months and additional inputs (e.g. potential evapotranspiration) indicated by '…' are transformed according to a hydrological model $g[\,]$ to produce an ensemble of simulated streamflow, $Q_{(k)}^{spl}$. This procedure is repeated for all simulated rainfall replicates.
>
> $$Q_{(k)}^{spl} = g\left[R_{(k)}^{spl}, \dots\right] \qquad (5)$$
>
> By construction, the 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.
>
> 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 results of the unit test and the integrated test (Steps 1-2) are then investigated and compared selecting each month as the evaluated time period in turn as well as other key time periods (e.g. annual)."



(a) Unit and integrated testing procedure

(b) Error profile for unit and integrated tests

Fig. 5 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 time period of June (*t*=6) (box plot whiskers indicate the 90% limits of the simulated streamflow replicates). For the unit test the errors in the evaluated period (*t*) are calculated as the difference between $Q^{spl}_{(k)}$ and $Q^{vo}_{(t)}$. For the integrated test the errors are calculated as difference between Q^{sim} and $Q^{vo}_{(t)}$.

Section 2.4.2, Page 11, Lines 11-20: "Using the function $h[\,]$ to denote a calculated statistic of interest (e.g. mean or standard deviation), the relative error in an evaluated time period t (e.g. annual or particular month) is given by

$$\%Err_{(t)} = \frac{h\left[Q^{eval}_{(t)}\right] - h\left[Q^{vo}_{(t)}\right]}{h\left[Q^{vo}_{(t)}\right]} \times 100 \quad (6)$$

where $Q^{vo}_{(t)}$ is the virtual-observed streamflow and $Q^{eval}_{(t)}$ is the simulated streamflow from the selected virtual hydrologic test (i.e. Q^{sim} if integrated test or $Q^{spl}_{(k)}$ if unit test selected) in the evaluated time period t. This procedure is repeated for all replicates of the simulated streamflow such that a range of errors is reported for each test for the target time period.

Following the calculation of this error metric for all replicates of the integrated test and ensemble of unit tests ($k = 1, \dots, 12$) it is possible to investigate deficiencies in the simulated streamflow in terms of which influencing month(s) contribute more to the deficiencies in streamflow for the target time

> period based on that statistic of interest. Thus, for each site, statistic and evaluated time period there are 13 sets of errors to compare."

The integrated test is designed to be a compact evaluation that includes multiple sites and statistics. Once the integrated test is completed and problems identified in the simulated streamflow, the more detailed unit test is applied to sites of interest (Step 3 of the framework in the revised manuscript). Figure 4 (formerly Figure 3) describes the unit tests, which are not designed to be undertaken on a large number of sites – they are designed to be more probing and are only undertaken on certain sites with problems, such as those identified by the integrated test.

The reviewer is right to point out that the example does make it look like the proliferation of error metrics is identified as a limitation. We can see that the accompanying example emphasises the large number of statistics rather than our intended key point: that there are difficulties in assessing trade-offs or the relative importance of statistics. The introduction has been revised to clarify that the key issue for observed-rainfall evaluation is the difficulty in understanding the relative importance of rainfall features in terms of streamflow generation and what to do when performance is 'mixed'.

> Page 2, Lines 3-9: "*Observed-rainfall evaluation is the most common method for SRM evaluation (Rasmussen, 2013, Wilks, 2008, Baxevani and Lennartsson, 2015, Srikanthan and Pegram, 2009, Evin et al., 2018, Bennett et al., 2018). As shown in Fig. 1(a) it involves comparisons between observed and simulated rainfall typically using a large number of evaluation statistics. Often, this method shows 'mixed' performance where many statistics are reproduced well, but some are poor. While these assessments are useful, a drawback is that 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. This means 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 (Bennett et al., 2018, Evin et al., 2018).*"

Comment 6:

Relatedly, in Figure 6 and 7, shouldn't the "obs rain" and "virtual obs. Flow" be a range, rather than a single value? There are 73 years of monthly rainfall and simulated flows… this variability would be valuable context for evaluating the variability of the stochastic realizations.

Response 6:

Yes, the variability of the rainfall and simulated flows is valuable context.

In Figures 6 and 7 (now Figures 8 and 11) we present a higher-level summary to evaluate the model performance considering both mean conditions and their variability (i.e. standard deviations). The statistics presented in these figures are the observed rain and the virtual-observed flow means (left column) and standard deviations (right column) over the full 73 years, calculated for the 12 months respectively. There are 12 monthly means and 12 monthly standard deviations per realisation. We are not calculating a separate statistic for each year of the timeseries. The boxplots show the range of these monthly statistics for the 10,000 stochastic rainfall model replicates. This convention is common to other papers in the field (e.g. Bennett et al. 2018, Khedhaouiria et al. 2018, Evin et al. 2018, Frost et al. 2011, Frost et al. 2004, Srikanthan et al. 2004, etc.)

Comment 7:

It is unclear how other meteorological forcings (temperature, etc.) are handled in this framework. The authors focus on stochastic rainfall generators, as opposed to stochastic weather generators, meaning that the other forcings must be supplied independently of the rainfall. I would imagine that this could create some serious issues in some cases if synthetic rainfall is spliced together with inappropriate series of temperature or other forcings; one can imagine getting strange results in terms of precipitation vs. ET balances, with unclear consequences for the evaluation results.

Response 7:

The reviewer is correct, as the focus is on evaluating stochastic rainfall generators, the other forcings are supplied independently. In our case study, the potential evapotranspiration (PET) time series (our only other meteorological forcing) is unchanged from the observed values in all hydrological simulations (i.e. the same PET time series is used in the simulation of the virtual-observed streamflow, integrated tests and unit tests). This is important as the hydrological evaluation is a relative comparison of the observed and simulated rainfall, hence all other time series and parameters relating to the hydrological model are kept the same in all tests. This approach was also taken in Sikorska et al. (2018), where the impact of using different rainfall disaggregation schemes on resultant flow was tested using a hydrological model. For all these tests the historical observed temperature time series was used to enable a comparison between the rainfall elements only.

To assess this assumption for the Onkaparinga case study we have evaluated the rainfall-PET correlation in all months. There is a negative relationship, which accounts for a small portion of the variance, up to R^2= 0.11 in drier summer months. shows the rainfall-PET correlations for a drier summer month (January) and a wet winter month (June).

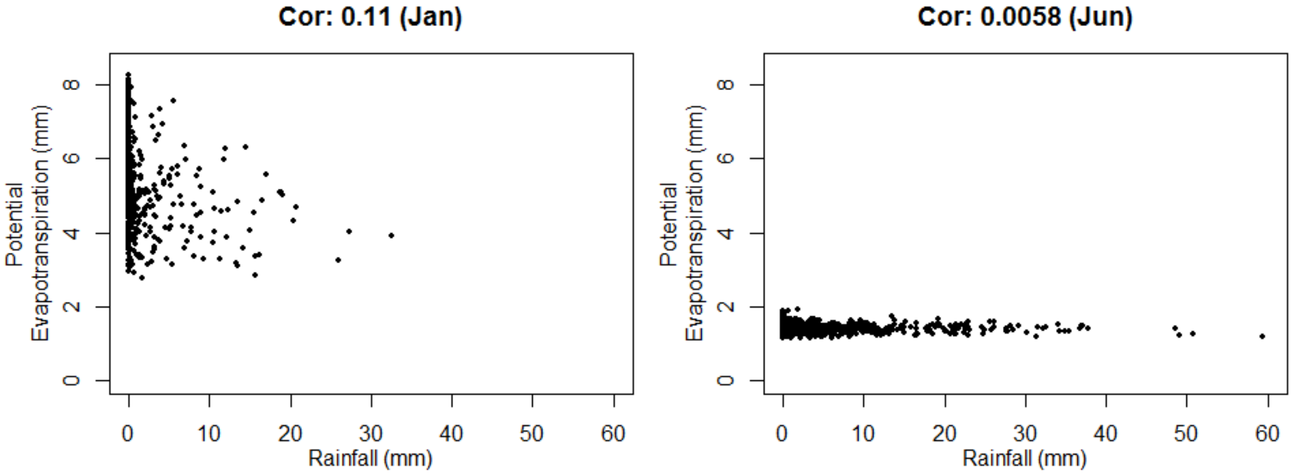


Figure 1: Rainfall-PET correlation (left) January and (right) June.

While there is some non-zero relationship, we do not consider it to undermine the case study (since all other statistics of PET are reproduced and the relationship is mild). However, this may not be the case for other locations where the model is applied.

We have included the reviewer's recommendation to apply the framework more generally to stochastic weather generators in Section 5.3. The application would require care to ensure that the PET (or other weather variable) generator does not introduce other deficiencies.

> Page 27, Line 22 - Page 28, Line 19: *"The formalisation of the virtual hydrological framework for SRM evaluation provides the opportunity for further improvements in the future, including: …*
>
> *(iv) Evaluation of spatial performance – there are multiple opportunities to develop tests for spatial performance including (a) repeating the integrated test for all sites and for catchment average rainfall means it would be possible to diagnose whether specific locations or the spatial dependence causes*

poor reproduction of streamflow statistics, (b) developing a spatial unit test (which is analogous to the temporal unit test but extended to space) where different combinations of sites are 'spliced' in the construction of catchment average rainfall – to evaluate the impact of 'mixed' performance in the SRMs between sites on the catchment average rainfall, and (c) these spatial unit tests could be used to evaluate stochastic weather generators (SWG) more generally as well as spatially distributed SRGs – though these would require a spatially distributed hydrological model."

Comment 8:

I have never developed my own SRG, but I imagine that it might be hard to know exactly how to use the results from this analysis to refine that generator, despite the authors' claim that this is a valuable use of the framework. It identifies performance by month, rather than by "rainfall characteristics" (pg. 21 line 9). It is useful to know whether the SRG performs well for some months than others, but what next? If the authors plan to continue research on this topic, I would suggest that a method that "tracks" the propagation of rainfall through the model might be more effective. To me, the most clear way of doing this is to track how different rainfall statistical moments translate to different statistical moments in the streamflow, using both historical and synthetic rainfall. Such an approach would be amenable to changing the evaluation timescale. For these reasons, I recommend that the authors delete the statement that this framework "should be an essential step in the development and application of stochastic rainfall models" (page 21 line 22-23). On a related note, the authors should comment on how this technique would apply to distributed (i.e. high-resolution gridded) SRGs and hydrologic models.

Response 8:

The reviewer has raised some excellent discussion points. We provide specific responses below to the discussion points raised.

1. "*It is useful to know whether the SRG performs well for some months than others, but what next?*" – We agree that it is a useful feature and it provides much more information than observed-rainfall evaluation alone. This is a key innovation of the paper. Following an observed-rainfall evaluation the focus would have been on the months Jan, Feb, Nov, Dec, May and June (Bennett et al., 2018). However, based on the results of the virtual hydrologic framework, we now know that May-July are the key months when considering the hydrology and that the problems with modelled rainfall in Jan-Feb, Nov-Dec are less important. Also, we now know that rainfall in preceding months is important and not just the month in which the flow is evaluated, which is more information than before.

   We agree that it does not tell us exactly which rainfall characteristics to focus on. However, it is unlikely to be that simple – a single rainfall statistic is unlikely to translate into a single runoff statistic because streamflow integrates a range of rainfall processes (see also reviewer #1, comment 21).

   Now that we know in which months deficiencies originate, we can focus on those months and trial various alternatives to the rainfall model to address the problem. This is left for future research, as mentioned in Section 5.3.

   > Page 27, Line 22 - Page 28, Line 7: *"The formalisation of the virtual hydrological framework for SRM evaluation provides the opportunity for further improvements in the future, including: …*
   >
   > *(ii) Comparison of SRMs – this framework can be extended to provide more direct guidance on which rainfall features (in terms of components of the SRM) should be modified to improve streamflow performance. This can be done by comparing multiple rainfall model variants (parametrically, or via bootstrap techniques) which are designed to have contrasting features of a key characteristic (e.g. intermittency, rainfall correlation). Such an approach was undertaken by Evin et al. (2018) using an observed-rainfall evaluation approach. If the SRMs have monthly/seasonal autocorrelation (these were not significant for the rainfall in the Onkaparinga catchment) the unit testing approach would need to be extended by conditionally sample the simulated rainfall in a manner that preserves monthly correlations."*

2. "*I would suggest that a method that "tracks" the propagation of rainfall through the model might be more effective … most clear way of doing this is to track how different rainfall statistical moments translate to different statistical moments in the streamflow, using both historical and synthetic rainfall*" – This idea offers scope to extend the framework, and is something to consider in the future. In this paper, we describe the integrated test and then introduce the unit test in terms of 'splicing' monthly blocks of rainfall. However, the reviewer is right that the approach could be formulated differently to use different 'splicing' approaches. For example, to examine the percentage changes in resultant streamflow as a function of a particular change in the inputted rainfall. To allow for alternative approaches that 'track' the propagation of error we have softened our concluding remarks.
3. "*I recommend that the authors delete the statement that this framework "should be an essential step in the development and application of stochastic rainfall models*"" – We have softened the wording of this statement. It now reads:

   Page 29, Lines 2-4: "*The virtual hydrologic evaluation framework provides insights not available through conventional approaches and provides useful diagnostic ability for the development and application of SRMs."*
4. "*On a related note, the authors should comment on how this technique would apply to distributed (i.e. high-resolution gridded) SRGs and hydrologic models*." – This is an important topic. We have commented that this technique could be extended to distributed rainfall and hydrologic models in Section 5.3.

Comment 9:

I wonder if this framework should consider the autocorrelation in monthly rainfalls when doing this splicing. I don't know too much about the climate of South Australia, but I can imagine that autocorrelation at least in dry periods can be quite important, and this is likely not preserved during the splicing. It's not clear what the implications would be for the resulting evaluation.

Response 9:

This is a valid and interesting point. Bennett et al. (2018) demonstrated that the monthly autocorrelations are small for the Onkaparinga catchment (from -0.2 in drier summer months to 0.3 in the wetter winter months), and as a result this issue was not considered in the presentation of the framework.[3] We can appreciate that monthly/seasonal autocorrelation is a significant feature of other locations and that could be a limitation when applying this method. We briefly suggested in Section 5.3 that the issue of monthly autocorrelation could be explored as an extension of the model and have provided further details to explain how this might be achieved in the revised manuscript. For example, rather than naïvely splicing rainfall it might be possible to conditionally sample the simulated rainfall in a manner that preserves monthly correlations. The efficacy of this technique would require some exploration since there may be limitations arising from the conditional sampling.

Comment 10:

Figure 3 and elsewhere: I don't understand what "(90% limits shown)" means.

Response 10:

The reference to 90% limits indicates that the boxplot whiskers extend to from the 5th to 95th percentile values of the metrics based on the 10,000 replicates. The initial description of these figures (see Section 2.4.2 in the revised manuscript) and all figure captions has been revised to clarify that the "(90% limits shown)" indicates that the boxplot whiskers extend to the 90% limits of the 10,000 simulations for the presented statistic.

[3] Bennett at al. (2018) also demonstrated that the model sufficiently reproduced these small monthly autocorrelations.

Comment 11:

Section 3: Mention basin size. Also, why are stations outside the watershed used?

Response 11:

The basin size is included in the revised manuscript (323 km^2). All the sites identified in Figure 2 (now Figure 5) were used to estimate the catchment average rainfall for the rainfall-runoff modelling calibration. This is stated in the revised Section 3 of the manuscript.

> Page 13, Lines 1-12: *"The hydrological model GR4J (Perrin et al., 2003) was used to simulate virtual-observed streamflow at a daily time step. GR4J is a daily lumped hydrological model that simulates daily streamflow in a parsimonious manner using four parameters. The GR4J model was calibrated according to the procedure set out in Westra et al. (2014b) for the stationary version of the GR4J hydrological model. The details are provided in (Westra et al., 2014a) and a short summary is provided here. The multi-site rainfall gauges were Thiessen weighted to calculate the catchment average rainfall. The model was calibrated to the streamflow data at Houlgrave Weir (see Figure 4) using model calibration period of 15 years (1985-1999). The parameters were estimated using maximum likelihood estimation procedure with a weighted least squares likelihood function. The model parameters that maximised the likelihood function were found using a multi-start quasi-Newton optimisation procedure with 100 random starts. Overall, the GR4J model was a good fit to the observed streamflow, with a Nash-Sutcliffe efficiency of 0.8. A similar type of hydrological model and calibration approach has been used for other virtual evaluation studies (Li et al. 2014; 2016). The same set of hydrological model parameters are used for both the unit and integrated tests so that the same transformation of rainfall to flow is used."*

When estimating catchment average rainfall it is fairly common to use sites outside the catchment, to better represent the spatial variability and to avoid boundary effects. It is therefore important that a stochastic rainfall model is able to reproduce the rainfall statistics at all of the sites outlined in Figure 5. This is why we evaluated the stochastic rainfall model at all the sites indicated in Figure 5. There are further reasons why this is valid for a virtual approach. Most notably, because there is no comparison made with observed streamflow. The virtual hydrological evaluation uses the calibrated hydrological model as a tool to process the observed and simulated rainfall for comparison. The virtual-observed streamflow can be thought of as a virtual stream flow gauge. The virtual stream gauges have no physical location that they are trying to replicate. Instead the virtual stream gauges enable a synthetic test of the simulated rainfall.

Virtual hydrological evaluation of a single rainfall site is analogous to treating the information at the selected rainfall gauge (observed and simulated) as being representative of the catchment rainfall. This 'catchment rainfall' is then routed through the chosen hydrological model to produce simulated and virtual-observed streamflow at the 'virtual catchment outlet'. This type of virtual approach was used in a different context (the development of new techniques for flood frequency estimation) in which a calibrated hydrological model was 'moved' all over Australia (Li et al., 2016).

Comment 12:

More importantly, is the rainfall hydroclimate stationary? If not, then it seems as though this whole issue of stochastic generation and comparison of resulting streamflows against a nonstationary continuous simulation would be more complicated. Please comment on this.

Response 12:

We have now commented on this issue in Section 5.3.

> Page 27, Line 22 - Page 28, Line 12: *"The formalisation of the virtual hydrological framework for SRM evaluation provides the opportunity for further improvements in the future, including: …*
>
> *(iii) Evaluation of temporal non-stationarity – this framework can be extended to evaluate the impact of non-stationarity on SRM model performance by applying it on a selected non-stationary period. Care would be needed in the selection of statistics to identify model performance (since the performance in different sub-periods could be masked when evaluating an overall period). A related issue is that the hydrological model should provide adequate performance across the range of non-stationary climate forcings to which it is subjected."*

The reviewer is right to point out the complicated nature of comparing stochastically generated rainfall against a nonstationary continuous simulation. In this paper we took steps to minimise this impact by careful selection of the observed rainfall period.[4]

Minor Comments

Comment 13:

Page 1 line 15: change "months" to "seasons"- that is a more broadly relevant term. Hydrology varies seasonally, months are an arbitrary construct (this comment applies elsewhere in the paper, such as page 2 line 31)

Response 13:

Thank you for pointing this out. Where possible the terms 'seasons' and 'time periods' are used in preferred to 'months'. Where 'months' form part of the specific analysis (i.e. influencing months) the term has been retained.

Comment 14:

Throughout paper: I recommend introducing an acronym for stochastic rainfall models and using it throughout.

Response 14:

Thank you, we have introduced an acronym for stochastic rainfall models (SRMs) and used it throughout the revised manuscript.

[4] The catchment experiences a significant rainfall decline in the early 2000's (see Westra et al 2014a and 2014b) due to the 'millennium drought'. This is why we choose an earlier rainfall period that finishes in 1986. Although this does not mean we have eliminated the impact of non-stationarity it has been reduced by taking this step.

Comment 15: ***Page 1 line 9: change "is" to "has been"***

Comment 16: ***Page 1 line 10: change "is given" to "has been paid"***

Comment 17: ***Page 1 line 12: delete "whenever the simulated rainfall are poor"***

Comment 18: ***Page 1 line 19: change "catchment cycle" to "annual hydrologic cycle"***

Comment 19: ***Page 1 line 28: delete comma after "targeted"***

Comment 20: ***Page 2 line 10: "and/or" is not appropriate in technical writing. Use "or"***

Comment 21: ***Page 2 line 12: put "virtual experiments" in quotations when mentioned for the first time, for emphasis***

Comment 22: ***Page 3 line 16: add comma after "poor"***

Comment 23: ***Page 5 line 10: The goal is not to match streamflow observations. It is to match the statistics of streamflows***

Comment 24: ***Page 3 line 12: add "model" before parameters***

Comment 25: ***Page 11 line 16: grammar problem "was fit good"***

Comment 26: ***Page 20 line 18: put "memory" in quotations.***

Comment 27: ***Page 10 line 15: I think that "observed/virtual" is a strange term. Observations have very little usage in this study…***

Comment 28: ***Page 14 line 14: This sentence is a bit awkward. It isn't perhaps so "common and obvious" to the reader.***

Comment 29: ***Page 20 line 13: good place to mention that multiple SRGs could be used too, not just multiple hydrologic models.***

Response 15 – 29:

Fixed.

Comment 30: ***Page 1 line 24-25: I recommend deleting "risks" after "floods" and "droughts" and changing it to "hazards" after "hydrologic"***

Response 30:

'Risks' has been deleted after 'floods' and 'droughts'. 'Hydrologic risks' has been retained in preference to 'hydrologic hazards'.

Comment 31: ***Page 3 line 13: Why would you call ET "extraneous"? It is generally very very important.***

Response 31:

We agree ET is important it was referred to as extraneous only to evaluations of rainfall.

Comment 32: ***Page 19 line 3-4: streamflow arises from more than just rainfall integration over a catchment area-what about ET, etc.?***

Response 32:

Thank you, we agree. The impact of ET and other catchment properties related to the production of streamflow are mentioned later in Section 5.1.

Comment 33: ***Page 20 line 11-13: I don't understand this sentence. Certainly model performance depends on the chosen model.***

Response 33:

Due to a revision of the text the sentence no longer appears.

Response to Reviewer 3

Comment 1:

The manuscript proposes a "virtual hydrological" framework useful for the performance evaluation of stochastic rainfall generators (SRGs). Differently from other studies involved on this topic, this work proposes 1) to evaluate the rainfall performances directly in terms of discharge by considering as benchmark the "virtual observed streamflow", i.e. the streamflow obtained by running the observed rainfall into the hydrological model, 2) to use two different tests to highlight discrepancies between observed and simulated rainfall for a specific site or month.

Although the topic is surely of interest for the readership of HESS, a major revision is required before to consider the manuscript suitable for the publication. Indeed, throughout the manuscript some important information are missing (for details see specific comments below) whereas the section 2 and details about the virtual hydrological framework should be shortened. Moreover, the outcomes of this study seem linked to the specific case study and the authors should discuss how the results could be generalized for different SRGs and hydrological models.

Response 1:

Thank you for your comments. We have revised the manuscript to address the matters raised in particular we have:

- Shortened the Section 2 and other sections describing the virtual hydrological evaluation framework.
- Provided further details on the stochastic rainfall and hydrological models in Section 3.
- Outlined and discussed aspects of the virtual hydrological framework implementation that need further work (Section 5.3).
- Added better signposting in text to make clear the outcomes that are generic and those that are specific to the case study.

Greater detail is given in response to specific comments below.

Comment 2:

Abstract section.

This section should be made clearer concerning both the explanation of the virtual hydrological framework features and the results obtained in the work. Specifically, lines 12-15 and 18-20 in page 1 are not clear without reading the paper.

Response 2:

The abstract has been revised to provide more detail on the framework as well as improving signposting (within the text to better distinguish the framework outcomes and case study specific outcomes.

The revised abstract text is reproduced below:

> Page 1, Lines 8-24: "*Stochastic rainfall modelling is a commonly used technique for evaluating the impact of flooding, drought or climate change in a catchment. While considerable attention has been given to the development of stochastic rainfall models (SRMs), significantly less attention has been paid to performance evaluation methods. Typical evaluation methods employ a wide range of rainfall statistics. However, they give limited understanding about which rainfall statistical characteristics are most important for reliable streamflow prediction. To address this issue a formal evaluation framework 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 seasons. The*

virtual hydrologic evaluation framework uses two types of tests, integrated tests and unit tests, to attribute deficiencies that impact on streamflow to their original source in the SRM according to site and season. The framework is applied to a case study of 22 sites in South Australia with a strong seasonal cycle. In this case study, the framework demonstrated the surprising result that apparently 'good' modelled rainfall can produce 'poor' streamflow predictions, whilst 'poor' modelled rainfall may lead to 'good' streamflow predictions. This is due to the representation of highly seasonally catchment processes within the hydrological model that can dampen or amplify rainfall errors when converted to streamflow. The framework identified the importance of rainfall in the 'wetting-up' months (months where the rainfall is higher but streamflow lower) of the annual hydrologic 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 SRM evaluation."

Comment 3:

Section 2.

This section should be shortened deleting multiple repetitions about the framework description in sub-sections 2.2 and 2.3. Moreover, Figure 1 and 2 could be merged into one figure. Conversely, the section 2.4 should be improved (also adding a flowchart) to allow the readers to easily follow the section "results".

Response 3:

This comment aligns with feedback from other reviewers. We have shortened the introduction and Section 2 (from 167 lines down to 127 lines) and merged Figures 1 and 2 as suggested.

Please also see the response to Reviewer 1 comment 12. We have significantly improved the explanation of the virtual hydrological framework procedure. The procedure has been simplified from seven steps down to three steps and the presentation of the framework in Section 2 has been restructured to better integrate the steps and specific virtual tests (i.e. integrated and unit tests). The structure of Section 2 is now as follows:

2. Virtual hydrological framework
2.1 Overview
2.2 Step 1 – Identify poor performing sites
2.2.1 Selection of primary streamflow characteristic and relevant hydrological model
2.2.2 Integrated test procedure
2.2.3 Identify poor performing sites using CASE framework
2.3 Step 2 – Identify poor performing time periods
2.4 Step 3 – Identify sources of poor performance
2.4.1 Unit test procedure
2.4.2 Compare unit tests and integrated tests
2.4.3 Identify types of key deficiencies

A flow chart (figure 2) has been added to graphically summarise the process to make it easier to follow in terms of the method and results. The figure is reproduced below.

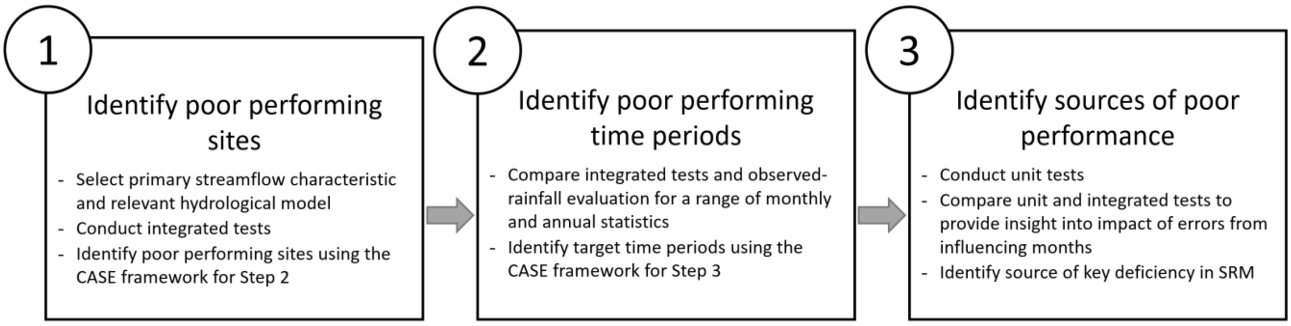

1
Identify poor performing sites
- Select primary streamflow characteristic and relevant hydrological model
- Conduct integrated tests
- Identify poor performing sites using the CASE framework for Step 2
2
Identify poor performing time periods
- Compare integrated tests and observed-rainfall evaluation for a range of monthly and annual statistics
- Identify target time periods using the CASE framework for Step 3
3
Identify sources of poor performance
- Conduct unit tests
- Compare unit and integrated tests to provide insight into impact of errors from influencing months
- Identify source of key deficiency in SRM

Fig. 2 - Virtual hydrological evaluation procedure.

Comment 4:

Section 3. Some important details are missing in this section. In addition to the area of the catchment and the temporal resolution of the simulated rainfall, the authors should specify how the GR4J model is forced by observed rainfall and how it is calibrated. Is the observed catchment average rainfall used to force and calibrate the hydrological model? How many years of observed discharge data are used for calibration? Is this set of parameters used to simulate streamflow within the unit and integrated test?

Response 4:

Other reviewers have made similar comments (reviewer #1 comments 14-17, reviewer # 2 comments 7 and 11). The following details have been included in the revised Section 3:

- Catchment area (323 km^2)
- Rainfall model resolution (daily)
- Hydrological model resolution (daily)
- Hydrological model calibration details such as the number of observed years (model calibration and selection: 1985-1999, model evaluation: 2000-2009), Thiessen weighting of rainfall gauges was used to calculate catchment average rainfall, and the impact of rainfall errors was considered in detail (see Westra et al. 2014a, Westra et al. 2014b).
- The same set of hydrological model parameters are used for the unit and integrated tests so that the same transformation of rainfall to flow is used.

Section 3 is reproduced below for convenience.

> Page 12, Line 10 – Page 13, Line 12: "*The Onkaparinga catchment in South Australia is used as a case study (Fig. 6). The 323 km^2 catchment lies 25 km south of the Adelaide metropolitan area and contains the largest reservoir in the Adelaide Hills supplying the region (Mount Bold Reservoir). The catchment has a strong seasonal cycle (shown in Fig. 6) where the driest months (December, January and February) exhibit low rainfall and low streamflow, the wettest months (July, August and September) have high rainfall and high streamflow and the 'wetting-up' period (April, May and June) has high rainfall and lower streamflow.*
>
> *There is a strong rainfall gradient (Table 3), 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). A breakdown of the rainfall characteristics (annual total, number of wet days, daily average amounts, wet-spell and dry spell durations) at each site on a monthly basis is provided in Supplementary Material A.*
>
> *The simulated daily rainfall was determined from the latent variable autoregressive daily rainfall model of Bennett et al. (2018) using at-site calibrated parameters. This rainfall model uses a latent variable concept, which relies on sampling from a normally distributed 'hidden' variable. The latent variable can then be transformed to a rainfall amount by truncating values below zero and by rescaling values above zero to match the observed rainfall's distribution. Here, the rainfall is rescaled using a power transformation.*
>
> *To calibrate the model the rainfall data at a given site is partitioned on a monthly basis and separate parameters are fit for each month. The mean and standard deviation of rainfall amounts, as well as the proportion of dry days is calculated. These statistics are matched to the corresponding properties of the truncated power transformed normal distribution. The at-site lag-1 temporal correlation is then calculated based on the observed wet day periods for a given month. This statistic is transformed to the equivalent correlation of the underlying latent variable by accounting for the effects of truncation to determine the autocorrelation parameter. Full details of the calibration procedure are provided in Bennett et al. (2018).*

> *In this study the daily rainfall model was calibrated and simulated at 22 locations throughout the catchment that have long, high-quality records (Table 3). 10,000 replicates of simulated rainfall covering a 73 year period (1914-1986) were used.*
>
> *The hydrological model GR4J (Perrin et al., 2003) was used to simulate virtual-observed streamflow at a daily time step. GR4J is a daily lumped hydrological model that simulates daily streamflow in a parsimonious manner using four parameters. The GR4J model was calibrated according to the procedure set out in Westra et al. (2014b) for the stationary version of the GR4J hydrological model. The details are provided in (Westra et al., 2014a) and a short summary is provided here. The multi-site rainfall gauges were Thiessen weighted to calculate the catchment average rainfall. The model was calibrated to the streamflow data at Houlgrave Weir (see Figure 4) using model calibration period of 15 years (1985-1999). The parameters were estimated using maximum likelihood estimation procedure with a weighted least squares likelihood function. The model parameters that maximised the likelihood function were found using a multi-start quasi-Newton optimisation procedure with 100 random starts. Overall, the GR4J model was a good fit to the observed streamflow, with a Nash-Sutcliffe efficiency of 0.8. A similar type of hydrological model and calibration approach has been used for other virtual evaluation studies (Li et al. 2014; 2016). The same set of hydrological model parameters are used for both the unit and integrated tests so that the same transformation of rainfall to flow is used."*

Comment 5:

Finally, major details should be added to this section about the rainfall statistics used for the calibration of the SRG of Bennett et. al. (2018).

Response 5:

In the revised manuscript a summary is provided of the calibration approach for the rainfall model (see Section 3) so that it is easier for the reader to understand the model without needing to also read Bennett et al (2018).

> Page 12, Lines 19-32: "*The simulated daily rainfall was determined from the latent variable autoregressive daily rainfall model of Bennett et al. (2018) using at-site calibrated parameters. This rainfall model uses a latent variable concept, which relies on sampling from a normally distributed 'hidden' variable. The latent variable can then be transformed to a rainfall amount by truncating values below zero and by rescaling values above zero to match the observed rainfall's distribution. Here, the rainfall is rescaled using a power transformation.*
>
> *To calibrate the model the rainfall data at a given site is partitioned on a monthly basis and separate parameters are fit for each month. The mean and standard deviation of rainfall amounts, as well as the proportion of dry days is calculated. These statistics are matched to the corresponding properties of the truncated power transformed normal distribution. The at-site lag-1 temporal correlation is then calculated based on the observed wet day periods for a given month. This statistic is transformed to the equivalent correlation of the underlying latent variable by accounting for the effects of truncation to determine the autocorrelation parameter. Full details of the calibration procedure are provided in Bennett et al. (2018).*
>
> *In this study the daily rainfall model was calibrated and simulated at 22 locations throughout the catchment that have long, high-quality records (Table 3). 10,000 replicates of simulated rainfall covering a 73 year period (1914-1986) were used."*

Comment 6.1:

Section 4. In this section the authors should address the following points:

1) as the authors know, the rainfall simulated by the SRG are function of the rainfall statistic properties used to estimate the model parameters. According to the authors, in which way the rainfall statistical properties and the results obtained by the unit and integrated test are linked? If different rainfall statistical properties are used for the SRG calibration, are the results different? For instance, is the identification of the 10 "poor" sites sensitive a variation of rainfall statistics? If different statistics are used for SRG calibration, is it possible to reduce the streamflow errors? These aspects should be demonstrated/discussed by the authors in order to provide to the readers a general framework not tailored for a specific case study.

Response 6.1:

It is important to clarify, that the goal of this paper is introduce a generic virtual-observed streamflow framework, and two tests (integrated and unit) that provide greater insight than traditional observed-rainfall evaluation approaches. This evaluation framework was demonstrated using a case study that included the rainfall model of Bennett et al. (2018) and a hydrological model for the Onkaparinga catchment. We demonstrated new insights/outcomes that are necessarily case study specific – but, they could not be derived, using traditional observed-rainfall evaluation techniques, and hence are a demonstration of the framework. As requested by the reviewer, we have more clearly signposted the generic components of the framework and the case study specific outcomes in the revised paper.

The selection of a rainfall model and its calibration approach, are independent of the generic evaluation framework. We agree with the reviewer, that if a different calibration approach (i.e. different rainfall statistics) were used, then the results may change, and the streamflow errors may reduce. Indeed, evaluating how different calibration approaches can influence the streamflow would be an excellent future use of this evaluation framework. As the paper has increased substantially to address the issues raised by the reviewers (13 figures, ~10,000 words), examining this issue is outside of scope. We have briefly touched on this idea in the revised discussion section (Section 5.3) as a future research application of the framework.

Comment 6.2:

2) as the streamflow generation is a results of the mean areal (rather than single-site) rainfall over the basin, before to apply the integrated test to identify sites for which the rainfall simulation is not good, it could be interesting to estimate the streamflow errors coming from the mean areal rainfall, evaluated as average over the 22 sites. How good are these streamflow time series with respect to the "virtual-observed" ones (obtained by the rainfall observed over the 22 sites and averaged over the catchment)? More interesting, the authors should highlight the benefits deriving from the use of integrated test. For instance, they should show what are the streamflow errors if only the 12 "good" sites are retained to evaluate the mean areal rainfall. Is it better than using all 22 sites?

Response 6.2:

The reviewer is right – in a 'real-world' catchment, streamflow is the result of rainfall over a basin rather than a single site. Our framework does enable spatial characteristics to be tested and in this respect we agree with the reviewer about the potential utility of the proposed test. By using the framework it would be possible to undertake evaluations with catchment average rainfall.

In this paper we have not evaluated how the observed and simulated catchment average rainfall compare in terms of the resultant streamflow. This is because, as a matter of first priority, our approach focuses on identifying issues with rainfall at each site and getting this right before moving on to assess deficiencies in spatial properties. We therefore prefer to assess the at-site performance prior to the catchment average performance. Future work will demonstrate and apply the framework to catchment average rainfall.

The reviewer makes an excellent suggestion regarding future investigations into the impact of 'mixed' performance in the rainfall model between sites. We have added a discussion on how this could be explored in future work and indicated that the proposed investigation of resultant streamflow (where different combinations of sites are 'spliced' in the construction of catchment average rainfall) is analogous to our temporal unit test but extended to space.

> Page 27, Line 22 – Page 28, Line 19: *"The formalisation of the virtual hydrological framework for SRM evaluation provides the opportunity for further improvements in the future, including: …*
>
> *(iv) Evaluation of spatial performance – there are multiple opportunities to develop tests for spatial performance including (a) repeating the integrated test for all sites and for catchment average rainfall means it would be possible to diagnose whether specific locations or the spatial dependence causes poor reproduction of streamflow statistics, (b) developing a spatial unit test (which is analogous to the temporal unit test but extended to space) where different combinations of sites are 'spliced' in the construction of catchment average rainfall – to evaluate the impact of 'mixed' performance in the SRMs between sites on the catchment average rainfall, and (c) these spatial unit tests could be used to evaluate stochastic weather generators (SWG) more generally as well as spatially distributed SRGs – though these would require a spatially distributed hydrological model."*

Comment 7:

Page 2, lines 19-21: the example of Bennet et al. (2018) is not clear without reading the paper.

Response 7:

Thank you for pointing this out. Due to a restructure of the introduction this example no longer appears.

Comment 8:

Page 7, lines 14-15: This sentence is not clear. Please rephrase it.

Response 8:

Due to a restructure the sentence no longer appears.

Comment 9:

Page 8, line 24: Why the authors write "13 errors to compare"? Are the authors considering also the integrated test? It is not clear.

Response 9:

Yes, the 13 errors compared arise from unit tests conducted for each of the 12 influencing months as well as an integrated test. We have restructured the presentation of methodology and revised the text explaining the calculation and comparison or errors to clarify this point. Please also see the response to reviewer 2, comment 5.

Comment 10:

Page 8, lines 27-28. This example related to the integrated test is difficult to understand in this section. The reason is that in the previous section, where the authors describe the test there is no mention to the fact that the evaluation of the integrated test is carried out also at the monthly scale.

Response 10:

We have improved the explanation of the framework to make clear that at Step 2 of the framework the integrated test results are evaluated at the monthly scale to allow a comparison with observed-rainfall evaluation and the unit test (see Section 2.3).

Comment 11:

Figure 4: the position of the streamflow gauging station should be added in the figure.

Response 11:

The streamflow gauge position has been added to the revised figure. It is indicated by the purple square.

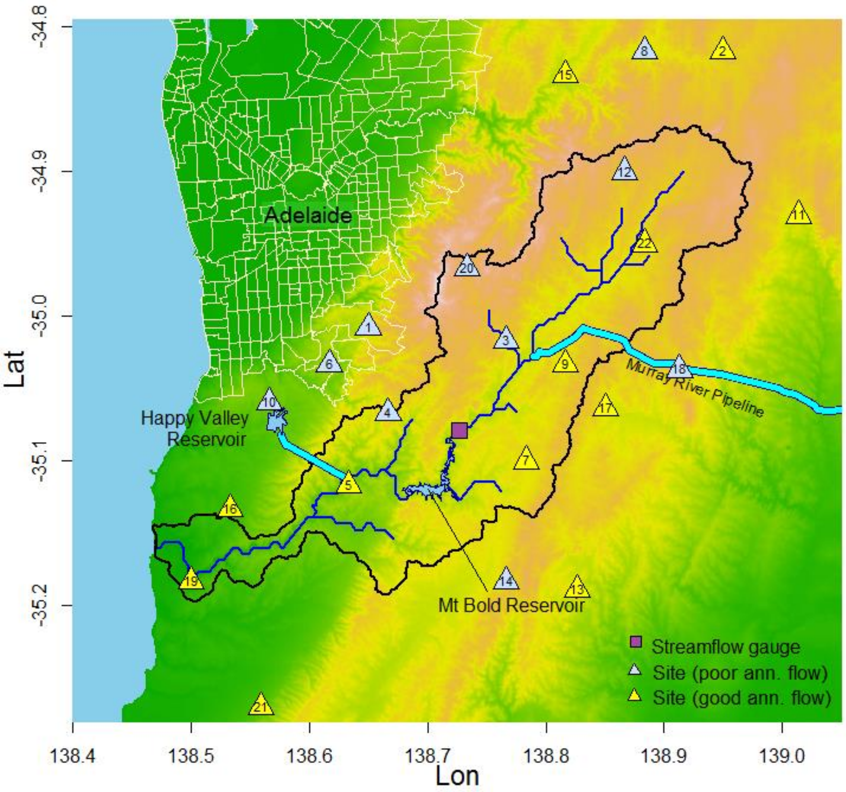


Fig. 6 Onkaparinga catchment, South Australia. Sites indicated by blue triangles are explored in greater detail in this paper due to the relatively poorer ability of simulated rainfall to reproduce annual streamflow totals at these sites.

Comment 12:

Page 15, lines 19-20: the conclusion about the transitional months should be drawn carefully. Indeed, moving from dry to wet conditions the process of formation of flood is very sensitive to the antecedent soil moisture conditions. Is the hydrological model able to reproduce observed streamflow in the transition period?

Response 12:

Thank you, this is a fair point. Our application has focused on streamflow characteristics relevant to yield. In the framework we stress the importance of choosing a 'fit for purpose' hydrological model in terms of reproducing the streamflow characteristic of interest (Step 2). In our investigation we have chosen a widely applied model (GR4J) that has been calibrated using a rigorous approach (see Westra et al 2014a and 2014b).

The focus on rainfall in the wetting-up period was identified after we applied the new virtual framework. This key period would not have been identified using an observed-rainfall evaluation approach. Therefore, it was not possible to set hydrological model performance throughout this period as a criterion for hydrological model selection at the outset.

We agree with the reviewer that this is an important period of simulation to get right, and as a result of this analysis, we will examine hydrological models that have the potential to better simulate this period.[5] The

[5] Such models may include non-stationary variants of GR4J (Westra et al. 2014a and 2014b) or SUPERFLEX (Fenicia et al. 2011, Kavetski and Fenicia, 2011).

evaluation using different models is unlikely to change the conclusion that the rainfall in the wetting-up months are important, as the GR4J model already performs reasonably well. However, it may change the magnitude of the rainfall's impact on the hydrological performance – especially for future comparisons to observed streamflow. Further research will investigate this issue and has been added to the discussion Section.

> Page 28, Lines 22-302: "*The formalisation of the virtual hydrological framework for SRM evaluation provides the opportunity for further improvements in the future, including:*
>
> *(i) Using multiple, well-tested hydrological models - a potential limitation of the virtual hydrologic evaluation framework is that it is reliant on the use of a hydrological model. Hydrological structural errors may potentially skew interpretation of the SRM evaluation if the hydrological model poorly represents the catchment processes. To reduce these impacts the steps taken in this study included (a) using a well-tested hydrological model that has demonstrated good performance on a wide range of catchments (e.g. the GR4J model has been widely tested , see (Perrin et al., 2003, Coron et al., 2012); (b) calibrating and evaluating the hydrological on a catchment close to the observed rainfall sites to ensure it provided sufficiently good performance (e.g. GR4J was calibrated to the Onkaparinga catchment - see (Westra et al., 2014a, Westra et al., 2014b). Future research will use multiple, well-tested hydrological models with sufficiently good performance to reduce the reliance on a single hydrological model and ensure the identification of SRM deficiencies is not dependent on a single hydrological model*."

References

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

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

Fenicia, F., D. Kavetski, and H. H. G. Savenije (2011). Elements of a flexible approach for conceptual hydrological modelling: 1. Motivation and theoretical development, Water Resour. Res., 47, W11510, doi:10.1029/2010WR010174

Frost, A.J., Cowpertwait, P. and Srikanthan, R. (2004). Stochastic generation of point rainfall data at subdaily timescales: a comparison of DRIP and NSRP. CRC for Catchment Hydrology.

Frost, A.J., Charles, S.P., Timbal, B., Chiew, F.H., Mehrotra, R., Nguyen, K.C., Chandler, R.E., McGregor, J.L., Fu, G., Kirono, D.G. and Fernandez, E. (2011). A comparison of multi-site daily rainfall downscaling techniques under Australian conditions, Journal of Hydrology, 408(1-2), pp.1-18.

Kavetski, D., and F. Fenicia (2011), Elements of a flexible approach for conceptual hydrological modelling: 2. Application and experimental insights, Water Resour. Res., 47, W11511, doi:10.1029/2011WR010748.

Khedhaouiria, D., Mailhot, A. and Favre, A.C. (2018). Daily Precipitation Fields Modeling across the Great Lakes Region (Canada) by Using the CFSR Reanalysis. Journal of Applied Meteorology and Climatology, 57(10), pp.2419-2438.

Kim, D. and F. Olivera (2012). "Relative Importance of the Different Rainfall Statistics in the Calibration of Stochastic Rainfall Generation Models." Journal of Hydrologic Engineering 17(3): 368-376.

Leonard, M., Metcalfe, A. and Lambert, M. (2008) Frequency analysis of rainfall and streamflow extremes accounting for seasonal and climatic partitions. Journal of hydrology, 348(1-2), pp.135-147.

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

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

Sikorska, A. E., Vivrioli, D., Seibert, J. (2018): Effective precipitation duration for runoff peaks based on catchment modelling, J. Hydrol., 556, 510–522

Srikanthan, R., Chiew, F. and Frost, A. (2004). Stochastic Climate Library, User Guide.

Thyer, M. and Kuczera, G. (2000). Modeling long-term persistence in hydroclimatic time series using a hidden state Markov Model. Water resources research, 36(11), pp.3301-3310.

Westra, S., Thyer, M. Leonard, M., Kavetski, D. and Lambert, M. (2014a). A strategy for diagnosing and interpreting hydrological model nonstationarity, Water Resour. Res., 50, 5090–5113, doi: 10.1002/2013WR014719.

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

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~~has been given to the development of stochastic rainfall models~~,~~ (SRMs), significantly less attention ~~is given~~has been paid to performance evaluation methods. Typical evaluation methods employ a ~~variety~~wide range of rainfall statistics. However, they give limited understanding about which rainfall statistical characteristics are most important for reliable streamflow prediction~~ whenever the simulated rainfall are poor.~~. To address this issue a ~~new~~formal evaluation ~~method for rainfall models~~framework 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~~seasons. The virtual hydrologic evaluation framework uses two types of tests, integrated tests and unit tests, to attribute deficiencies that impact on streamflow to their original source in the SRM according to site and season. The framework is applied to a case study of 22 sites in South Australia~~.~~ with a strong seasonal cycle. In this case study, the framework demonstrated the surprising result that apparently 'good' modelled rainfall can produce 'poor' streamflow predictions, whilst 'poor' modelled rainfall may lead to 'good' streamflow predictions~~, as~~. This is due to the representation of highly seasonally catchment processes within the hydrological model that 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~~(months where the rainfall is higher but streamflow lower) of the annual hydrologic 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~~SRM evaluation.

1 Introduction

Stochastic rainfall model (SRM) 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~~floods (e.g. Camici et al., 2011, Li et al., 2016) or ~~drought risks~~droughts (e.g. Henley et al., 2013, Paton et al., 2013, Mortazavi-Naeini et al., 2015). When evaluating the efficacy of ~~stochastic rainfall models~~SRM's, current approaches~~, which~~ that 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~~SRM'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~~SRMs.

Observed-rainfall evaluation is the most common method ~~of~~for SRM evaluation (Rasmussen, 2013, Wilks, 2008, Baxevani and Lennartsson, 2015, Srikanthan and Pegram, 2009, Evin et al., 2018, Bennett et al., 2018). As shown in Fig. 1~~and relies on~~(a) it involves comparisons between observed and simulated rainfall typically using a large number of evaluation statistics~~.~~ ~~Typically~~. Often, 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~~. While these ~~limitations, an alternative is to evaluate the rainfall model's performance in terms of streamflow, hereafter referred to as 'observed-streamflow evaluation' . A challenge with observed-streamflow evaluation~~assessments are useful, a drawback 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 , and/or model structural errors .~~

~~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 , (ii) evaluating how well simpler conceptual hydrological models characterise complex processes , and (iii) for developing new techniques for flood frequency analysis . 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 with the procedures for the integrated test and unit test outlined in Section . 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).~~

~~21 Virtual hydrological evaluation framework~~

~~2.1 Limitations of existing evaluation techniques~~

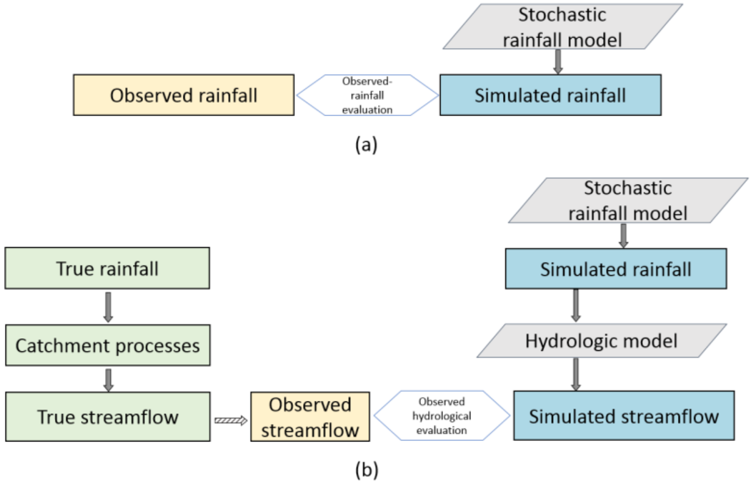
~~Observed-rainfall evaluation involves comparisons between observed and simulated rainfall statistics, as shown in (a). This method involves a large number of evaluation statistics and it is often unclear which are the most important and how trade-offs in respective performance between them might be 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,~~This means 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 (Bennett et al., 2018, Evin et al., 2018)~~. For example, 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, 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~~the conventional alternative is to evaluate the rainfall model's performance in terms of streamflow (e.g. Camici et al., 2011, Blazkova and Beven, 2002, 2009, McMillan and Brasington, 2008)~~. This method is indicated in (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 . For example, observed-streamflow evaluation has been used to predict the annual flood frequency distribution .~~

~~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~~SRM 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. 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 SRM. Hydrological model predictive performance can vary substantially from catchment to catchment due to data errors (rainfall or streamflow) and model structural errors (Evin et al., 2014, Coxon et al., 2015, Andreassian et al., 2001, Kuczera and Williams, 1992, Renard et al., 2011, McInerney et al., 2017)~~For example, 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, 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 . In some ephemeral catchments, the predictive uncertainty has a precision (standard deviation) of 40% of mean observed streamflow . These predictive errors can be substantially worse where there are significant rainfall errors .~~ which makes it difficult to evaluate the performance of the SRM and identify opportunities for improvement.

~~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.~~


Stochastic rainfall model
Observed rainfall
Observed-rainfall evaluation
Simulated rainfall
(a)
Stochastic rainfall model
True rainfall
Simulated rainfall
Catchment processes
Hydrologic model
True streamflow
Observed streamflow
Observed hydrological evaluation
Simulated streamflow
(b)

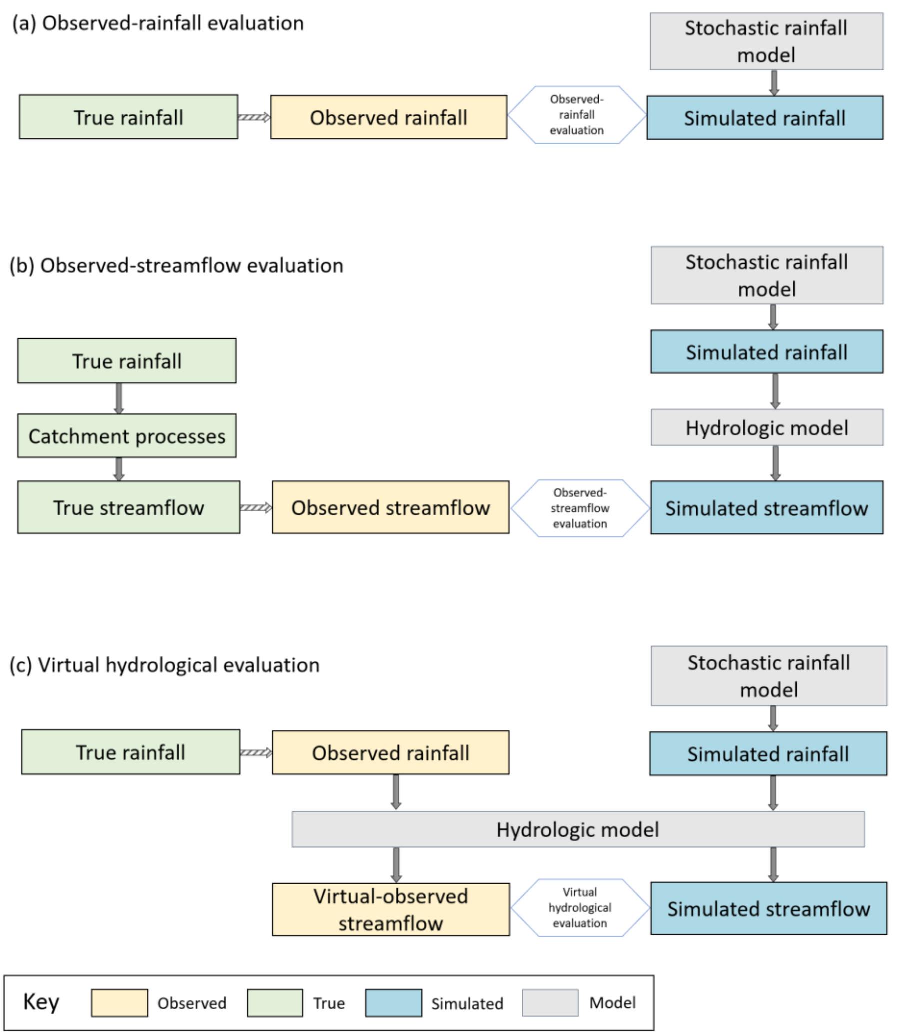

(a) Observed-rainfall evaluation
Stochastic rainfall model
True rainfall
Observed rainfall
Observed-rainfall evaluation
Simulated rainfall
(b) Observed-streamflow evaluation
Stochastic rainfall model
True rainfall
Simulated rainfall
Catchment processes
Hydrologic model
True streamflow
Observed streamflow
Observed-streamflow evaluation
Simulated streamflow
(c) Virtual hydrological evaluation
Stochastic rainfall model
True rainfall
Observed rainfall
Simulated rainfall
Hydrologic model
Virtual-observed streamflow
Virtual hydrological evaluation
Simulated streamflow
Key
Observed
True
Simulated
Model

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~~.~~ (c) virtual hydrological evaluation framework where simulated streamflow is compared against virtual-observed streamflow.

~~2.2~~ The focus~~Overview~~ of ~~the~~this paper is the development and application of a virtual hydrological evaluation framework for streamflow-based evaluation of SRMs. As shown in Fig. 1

~~The~~ (c), a virtual hydrological evaluation involves the comparison of simulated streamflow statistics (produced by the hydrological model by inputting simulated rainfall from a SRM) against virtual-observed streamflow statistics (produced by the hydrological model by inputting observed rainfall). This framework is designed to focus on streamflow predictions of interest, similar to observed-streamflow evaluation, but to reduce the sources of error to only those introduced by the ~~stochastic rainfall model.~~SRM. To illustrate this, Table 1 gives an overview of the sources of error for the three evaluation frameworks~~.~~ and indicates whether the evaluations are streamflow-based. The observed-rainfall evaluation framework is used to identify errors in the ~~rainfall model~~SRM, 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~~.~~ or statistics. However, with this approach it is not possible to readily identify whether discrepancies in the simulated streamflow are attributed to the ~~rainfall model~~SRM, the streamflow observations or poor process representation within the hydrological model. In contrast, the virtual hydrological evaluation 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.~~ SRM.

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

	Source of error			
~~**Evaluation framework**~~	**Stochastic rainfall model**	**Hydrological model**	**Observed streamflow**	**Streamflow-based evaluation**
Observed-rainfall evaluation	~~✓~~Yes	~~✗~~No	~~✗~~No	No
Observed-streamflow evaluation	~~✓~~Yes	~~✓~~Yes	~~✓~~Yes	Yes
Virtual hydrological evaluation	~~✓~~Yes	~~-~~No	~~-~~No	Yes

Merged Cells

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) including the identification of rainfall features of interest in terms of hydrological behaviour (Sikorska et al., 2018), (ii) for developing new techniques for flood frequency analysis (Li et al., 2014, 2016), and (iii) to support SRM selection (Müller and Haberlandt, 2018) as well as calibration and validation (Kim and Olivera, 2011) through a comparison to virtual streamflow.

The framework presented in this paper is significant advance from previously reported virtual experiments because it presents a formal framework to identify key deficiencies in the SRM by utilising (1) A comprehensive and systematic evaluation (CASE) framework (developed by Bennett et al., 2018 and used by Evin et al., 2018, Khedhaouiria et al., 2018) that systematically categorises performance at multiple spatial and temporal scales using quantitative criteria for each statistic, and (2) two types of virtual experiments that are able to identify the source of key deficiencies in SRM at specific locations and time periods.

The key objectives of this paper are:

1. To introduce a formalised framework for the virtual hydrological evaluation of SRMs.
2. To present two different tests which are part of the framework: the integrated test as well as introduce a new test, the unit test. Combined use of these tests allows streamflow discrepancies to be attributed to their original source in the SRM according to site and season.
3. To demonstrate the framework evaluation on a SRM and contrast the outcomes with conventional evaluation methods.

The virtual hydrologic evaluation framework is explained in Section 2 with the procedures for the integrated test and unit test outlined in Sections 2.2.2 and 2.4.1. SRMs 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

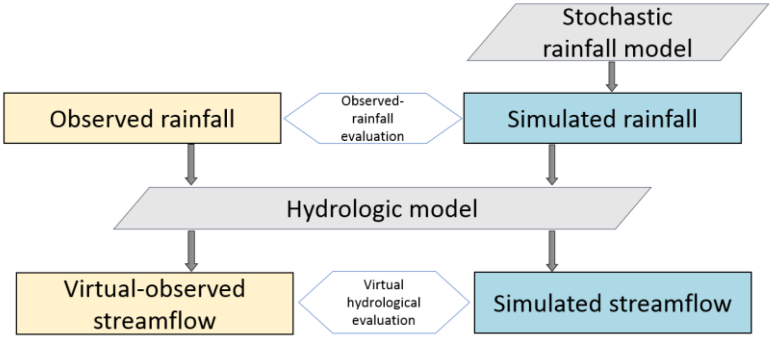
~~2.3~~2.1 Overview

~~As shown in , ~~A virtual hydrological evaluation involves the comparison of simulated streamflow statistics to virtual-observed streamflow statistics (Fig. 1~~,~~(c)), 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.~~SRM. Because the hydrological evaluation is a relative comparison of the observed and simulated rainfall, it is important that all other model parameters and extraneous variables (e.g. potential evapotranspiration) relating to the hydrological model are kept the same ~~in all instances. This is because ~~for the ~~ultimate goal~~simulation of ~~the modelling process remains the same: to match ~~both virtual-observed and simulated streamflow~~ for a catchment~~. It is also important that the selected hydrological model is fit for purpose so that it can simulate the streamflow characteristics of interest.

~~Therefore, ~~The virtual hydrological 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~~SRM 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~~SRM. This is because the ultimate goal of the SRM modelling process remains the same: to match observed streamflow for a catchment of interest.



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

The formal implementation~~Fig. Schematic~~ of the virtual ~~hydrologic~~hydrological evaluation framework ~~where simulated streamflow~~ is summarised in Fig. 2~~compared against virtual ~~. It uses a series of steps to identify poor performing sites, then poor performing time periods and then the key deficiencies in the SRM for those sites and time periods. It compares both observed ~~streamflow.~~

~~2.4 Diagnostic tests enabled by~~rainfall-evaluation and virtual hydrological evaluation. The virtual hydrological ~~framework~~

~~The virtual framework enables several techniques for isolating the deficiencies in simulated rainfall on the key streamflow predictions of interests, including~~evaluation includes two different types of tests, an ‘integrated test’ that isolates issues for a given site, and ‘unit tests’ that isolate issues for specific time periods. This enables the diagnosis of the key deficiencies in the simulated rainfall. The following sections explain the three steps in turn.

~~2.4.1 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 ().~~

~~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.~~

~~2.4.2 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 , consider the time series of observed, $R_{\_}^{obs}$, and simulated, $R_{\_}^{sim}$, daily rainfall for each year (and replicate) at a given site. (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.~~

$$R_{(k)}^{spl} = \bigcup_{m=1}^{12} \begin{cases} R_m^{sim} ; m = k \\ R_m^{obs} ; m \neq k \end{cases}$$ ~~(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:~~

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

~~The ensemble of $k = 1, \ldots, 12$ spliced rainfall time series $R^{spl}_{(k)}$~~

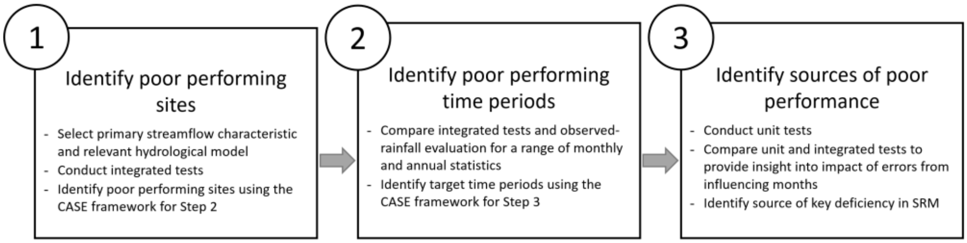

1
Identify poor performing sites
- Select primary streamflow characteristic and relevant hydrological model
- Conduct integrated tests
- Identify poor performing sites using the CASE framework for Step 2
2
Identify poor performing time periods
- Compare integrated tests and observed-rainfall evaluation for a range of monthly and annual statistics
- Identify target time periods using the CASE framework for Step 3
3
Identify sources of poor performance
- Conduct unit tests
- Compare unit and integrated tests to provide insight into impact of errors from influencing months
- Identify source of key deficiency in SRM

Fig. 2~~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 $Q^{sim}_{(t,k)}$ relating to the evaluated month, t,~~

~~$Q^{sim}_{(t,k)} \underset{t}{\in} g\left[R^{spl}_{(k)}\right]$ (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.~~ Virtual hydrological evaluation procedure.

2.2 Step 1 – Identify poor performing sites

The first step focuses on using integrated tests to identify poor performing sites for further evaluation. Following the selection of a primary streamflow characteristic of interest and a fit for purpose hydrological model, integrated tests are conducted for each rainfall site. The results of the integrated tests are then used to identify sites that are poor performing, according to the CASE framework, for the primary streamflow characteristic.

2.2.1 Selection of primary streamflow characteristic and relevant hydrological model

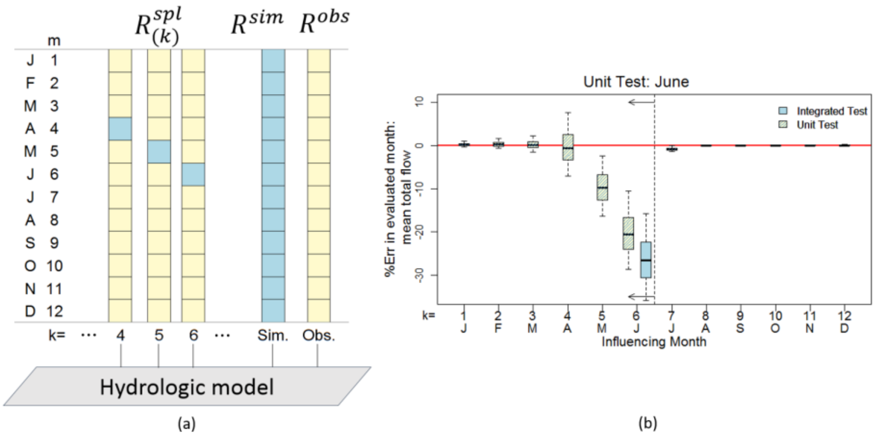
In order to undertake an integrated test ~~Using the function $h[\,]$ to denote a calculated statistic of interest, the relative error in an evaluated month is given by~~

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

~~where $Q^{vo}_{(t)}$ is the virtual-observed streamflow and $Q^{sim}_{(t)}$ 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. (b) illustrates for the evaluated month of June ($t = 6$), a typical error profile from a unit test of the mean monthly flow. (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.~~



~~**Fig. 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).**~~

~~2.5 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 is required 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.~~ flow volumes is important for yield. A streamflow characteristic of interest, herein termed the ‘primary streamflow characteristic’, is then selected to enable a method for filtering sites and concentrating the investigation of the rainfall model on sites that perform poorly in terms of its intended application. For example, the distribution of annual total flow would be a suitable characteristic when investigating yield. Following the identification of the hydrological model and primary streamflow characteristic an integrated test is conducted for each rainfall site, which serves as an overall test of the SRM's performance.

2.2.2 Integrated test procedure

The ~~third step is to conduct an~~ integrated test proceeds for ~~each rainfall~~a single site~~, which serves as an overall test of~~ by transforming the ~~rainfall model's performance. The test uses the~~ time series of observed and simulated rainfall, via the hydrological model (Fig. 1(c)). Consider the time series of observed, R^{obs} daily rainfall for each year at a given site. This rainfall time series ~~and~~ is transformed according to a hydrological model $g[\,]$ to produce the virtual-observed streamflow, denoted as Q^{vo} and '...' are additional inputs (e.g. potential evapotranspiration).

$$Q^{vo} = g[R^{obs}, \dots] \qquad (1)$$

Likewise, all replicates of the simulated, R^{sim} daily rainfall ~~time series as inputs to~~ for each year at a given site are transformed according the hydrological model ~~().~~ $g[\,]$ to produce simulated streamflow replicates, Q^{sim}.

~~The fourth step~~$$Q^{sim} = g[R^{sim}, \dots] \qquad (2)$$

If there is ~~to reduce the set of investigated~~a discrepancy between the simulated streamflow, Q^{sim}, and the virtual-observed streamflow, Q^{vo}, distributions, this indicates that there is a deficiency in the simulated rainfall for that site.

2.2.3 Identify poor performing sites using CASE framework

The integrated test results aim to ~~a smaller subset that poorly simulated~~identify the sites that are poor performing for the primary streamflow characteristic. Model performance is categorised using a CASE framework approach as 'good', 'fair' or 'poor' following Bennett et al. (2018). The quantitative tests for each performance category are provided in Table 2 alongside an illustration of each in Fig. 3. The quantitative tests proceed by comparing the statistics of the virtual-observed streamflow against those calculated from replicates of the simulated streamflow. Performance was categorised as 'good' if the selected statistic for the virtual-observed streamflow fell within the 90% limits of the statistic calculated from the simulated streamflow replicates (Fig. 3, case i), as 'fair' if the virtual-observed statistic fell outside the 90% limits of the simulated streamflow replicates but within the 99.7% limits (Fig. 3, case ii) and otherwise as 'poor' (Fig. 3 ~~during the integrated test in step three. This step assumes a multi-site evaluation and is optional if only a single site is evaluated~~, case iii).

Table 2 CASE performance classification criteria. Adapted from Bennett et al. (2018).

Performance Classification	Test	Key
'good'	Observation lies within the 90% limits (case i)	🟩
'fair'	Observation lies outside the 90% limits but within the 99.7% limits (case ii)	🟨
'poor'	Otherwise (case iii)	🟥

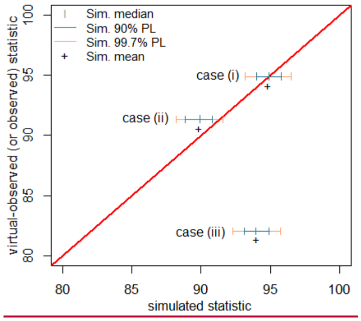


Fig. 3 Illustration of performance classification, case (i) shows 'good' performance, case (ii) shows 'fair' performance and case (iii) shows 'poor' performance. Adapted from Bennett et al. (2018).

2.3 Step 2 – Identify poor performing time periods

The ~~fifth~~second framework step is to ~~conduct~~identify poor performing time periods by conducting a detailed analysis of the integrated test results and comparing these results with an observed-rainfall evaluation 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). For each of the poor performing sites, each of these statistics for each month are categorized into 'good', 'fair' and 'poor' using the CASE framework. See Section 2.2.3~~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) for the subset of sites showing:~~ for further explanation of the categorization procedure. This enables the identification of poor-performing time periods from the perspective of the virtual hydrological evaluation.

Poor performance in reproducing virtual-observed streamflow is then contrasted against an observed-rainfall evaluation so that specific poor-performing time periods can be identified for further investigation in Step 3. By contrasting CASE performance categories ('good', 'fair' and 'poor') for observed-rainfall evaluation against virtual-observation streamflow evaluation, poor performing time periods from both rainfall and streamflow perspective can be identified. This comparison between the observed-rainfall evaluation and the virtual hydrological evaluation (integrated test) can be summarised graphically (e.g. see Fig. 7, Section 4.2).

2.4 Step 3 – Identify sources of poor performance

The third step of the framework is to identify sources of poor performance in streamflow according to deficiencies in the simulated rainfall. Step 2 identifies the poor performing time periods from a streamflow perspective. However, due to catchment 'memory', the poor performance in streamflow could be due to deficiencies in the simulated rainfall from a range of potential influencing months during or prior to the poor performing time period. For example poor streamflow performance in an evaluated month maybe due to the influence of: (i) rainfall deficiencies mostly in the same month (i.e. concurrent influencing months), (ii) rainfall deficiencies over a contiguous block of months including and preceding the evaluated month (i.e. prior and concurrent influencing months), or (iii) rainfall model deficiencies in a preceding month more so than in the evaluated month (i.e. prior influencing months). The integrated test cannot isolate which influencing months produce these deficiencies. Therefore, the unit test is designed to enable the identification of sources of poor performance in streamflow. The sources of poor performance are described in terms of which influencing months exhibit key deficiencies in simulated rainfall and therefore which SRM components should be improved.

2.4.1 Unit test procedure

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. 4(a), consider the time series of observed, R^{obs}, and simulated, R^{sim}, daily rainfall for each year (and replicate) at a given site. Fig. 4(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, k, and has the same length as the corresponding observed R^{obs} and simulated R^{sim} time series.

$$R_{(k)}^{spl} = \bigcup_{m=1}^{12} \begin{cases} R_m^{sim}\, ; m = k \\ R_m^{obs}\, ; m \neq k \end{cases} \qquad (3)$$

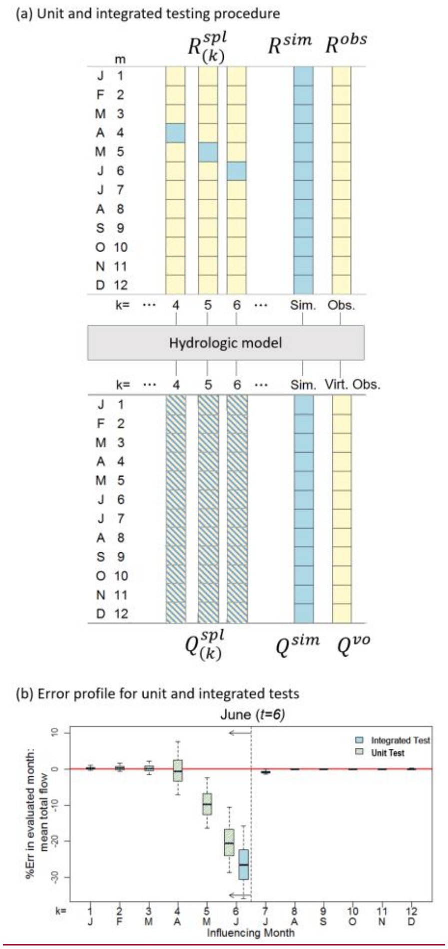
For example, if June ($k = 6$) is selected as the influencing month, each year of the spliced time series, $R_{(6)}^{spl}$, would be composed as follows:

$$R_{(6)}^{spl} = \{R_1^{obs}, \ldots, R_5^{obs}, R_6^{sim}, R_7^{obs}, \ldots, R_{12}^{obs}\} \qquad (4)$$

The ensemble of $k = 1, \ldots, 12$ spliced rainfall time series $R_{(k)}^{spl}$ for all influencing months and additional inputs (e.g. potential evapotranspiration) indicated by '…' are transformed according to a hydrological model $g[\,]$ to produce an ensemble of simulated streamflow, $Q_{(k)}^{spl}$. This procedure is repeated for all simulated rainfall replicates.

$$Q_{(k)}^{spl} = g[R_{(k)}^{spl}, \ldots] \qquad (5)$$

By construction, the 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.


(a) Unit and integrated testing procedure

(b) Error profile for unit and integrated tests

Fig. 4 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 time period of June (*t*=6) (box plot whiskers indicate the 90% limits of the simulated streamflow replicates). For the unit test the errors in the evaluated period (*t*) are calculated as the difference between $Q^{spl}_{(k)}$ and $Q^{vo}_{(t)}$. For the integrated test the errors are calculated as difference between Q^{sim} and $Q^{vo}_{(t)}$.

- ~~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 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 results of the unit test and the integrated test (Steps 1-2) are then investigated and compared selecting each month as the evaluated time period in turn as well as other key time periods (e.g. annual).

2.4.2 ~~The seventh step is to investigate and compare the results of the~~Compare unit ~~test (Step 6)~~tests and ~~the~~ integrated ~~test (Step 5) selecting each month as the evaluated month in turn.~~ tests

Side-by-side comparison of the results of the integrated test and ~~the~~ unit ~~test is~~tests are given in terms of the errors for selected monthly and annual statistics (see illustration in Fig. 4(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~~A relative error metric, $\%Err$, is used enable a comparison between the virtual-observed streamflow and the evaluated streamflow replicates from the integrated or unit tests (Q^{sim} or $Q_{(k)}^{spl}$) for time periods of interest (e.g. annual level or particular season or month). In this way the targeted hydrological evaluation centres on a specific subset of streamflows relating to the evaluated time period, t. In this paper, examples are provided for evaluated time periods at the monthly and annual scales.

Using the function $h[\ ]$ to denote a calculated statistic of interest (e.g. mean or standard deviation), the relative error in an evaluated time period t (e.g. annual or particular month) is given by

$$\%Err_{(t)} = \frac{h\left[Q_{(t)}^{eval}\right] - h\left[Q_{(t)}^{vo}\right]}{h\left[Q_{(t)}^{vo}\right]} \times 100 \qquad (6)$$

where $Q_{(t)}^{vo}$ is the virtual-observed streamflow and $Q_{(t)}^{eval}$ is the simulated streamflow from the selected virtual hydrologic test ~~across~~(i.e. Q^{sim} if integrated test or $Q_{(k)}^{spl}$ if unit test selected) in the evaluated time period t. This procedure is repeated for all replicates of the simulated streamflow such that a range of errors is reported for each test for the target time period.

Following the calculation of this error metric for all replicates of the integrated test and ensemble of unit tests ($k = 1, \dots, 12$) it is possible to investigate deficiencies in the simulated streamflow in terms of which influencing month(s) contribute more to the deficiencies in streamflow for the target time period based on that statistic of interest. Thus, for each site, statistic and evaluated time period there are 13 sets of errors to compare.

A typical error profile from integrated and unit tests is shown in Fig. 4(b) where mean monthly flow is selected as the statistic of interest for the evaluated time period of June ($t = 6$). In this figure the sets of errors from the integrated and unit tests are summarised as boxplots with the boxplot whiskers indicating the 90% limits of the errors from the evaluated streamflow replicates. Fig. 4~~12~~(b) shows the integrated test produced a median error of 27% (blue shaded boxplot) from all simulated rainfall replicates indicating a deficiency in the simulated streamflow for June. Examination of the unit tests (yellow and blue striped boxplots) for the target time period (June) shows that the median error is 20% when the influencing month is June ($k = 6$), the median error is 10% when the influencing month is May ($k = 5$) 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 SRM deficiencies in June and May respectively.

2.4.3 Identify types of key deficiencies

Following a side-by-side comparison of integrated test and unit test results in terms of the relative errors the sources of poor performance should be classified in terms of in which influencing months ~~with the integrated test can lead to several possible interpretations of where streamflow errors in the evaluation month~~streamflow deficiencies originate~~.~~ (e.g. poor streamflow arises from rainfall deficiencies mostly in the same month, a prior month or a contiguous block of months). Differentiating ~~the~~between cases allows for ~~rainfall model~~SRM 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.~~To complement this analysis a comparison of the virtual-observed flow duration curve for the evaluated time period with the flow duration curves resulting from unit tests for key influencing months is also recommended. 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. 5). The 323 km^2 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 (shown in Fig. 6~~with~~) where the ~~majority of rainfall occurring in winter and spring (June to November) and with a predominantly dry summer season~~driest months (December~~ to ~~, January and February~~).~~) exhibit low rainfall and low streamflow, the wettest months (July, August and September) have high rainfall and high streamflow and the ‘wetting-up’ period (April, May and June) has high rainfall and lower streamflow.

There is a strong rainfall gradient (Table 3), 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). A breakdown of the rainfall characteristics (annual total, number of wet days, daily average amounts, wet-spell and dry spell durations) at each site on a monthly basis is provided in Supplementary Material A.

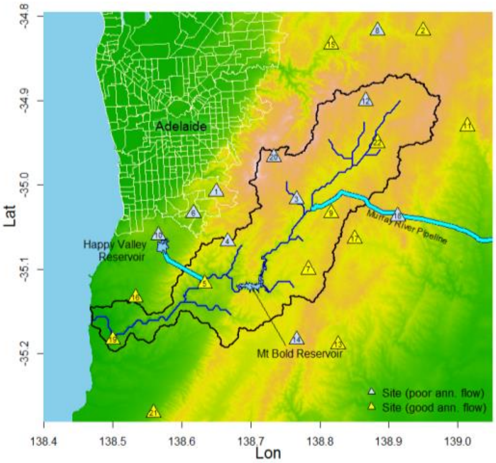
The simulated daily rainfall was determined from the latent variable autoregressive daily rainfall model of Bennett et al. (2018) using at-site calibrated parameters. This rainfall model uses a latent variable concept, which relies on sampling from a normally distributed ‘hidden’ variable. The latent variable can then be transformed to a rainfall amount by truncating values below zero and by rescaling values above zero to match the observed rainfall’s distribution. Here, the rainfall is rescaled using a power transformation.

To calibrate the model the rainfall data at a given site is partitioned on a monthly basis and separate parameters are fit for each month. The mean and standard deviation of rainfall amounts, as well as the proportion of dry days is calculated. These statistics are matched to the corresponding properties of the truncated power transformed normal distribution. The at-site lag-1 temporal correlation is then calculated based on the observed wet day periods for a given month. This statistic is transformed to the equivalent correlation of the underlying latent variable by accounting for the effects of truncation to determine the autocorrelation parameter. Full details of the calibration procedure are provided in Bennett et al. (2018)~~The~~.

In this study the daily rainfall model was calibrated and simulated at 22 locations throughout the catchment that have long, high-quality records (Table 3). 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).~~at a daily time step. GR4J is a daily lumped hydrological model that simulates daily streamflow in a parsimonious manner using four parameters. The GR4J model was calibrated according to the procedure set out in Westra et al. (2014b)~~to simulate streamflow for the Onkaparinga catchment~~ for the stationary version of the GR4J hydrological model. The details are provided in (Westra et al., 2014a) and ~~was~~a short summary is provided here. The multi-site rainfall gauges were Thiessen weighted to calculate the catchment average rainfall. The model was calibrated to the streamflow data at Houlgrave Weir (see Figure 4) using model calibration period of 15 years (1985-1999). The parameters were estimated using maximum likelihood estimation procedure with a weighted least squares likelihood function. The model parameters that maximised the likelihood function were found using a multi-start quasi-Newton optimisation procedure with 100 random starts. Overall, the GR4J model was a good fit to the observed streamflow, with a Nash-Sutcliffe efficiency of 0.8. ~~The~~A similar type of hydrological model and calibration approach has~~ also~~ been used for other virtual evaluation studies (Li et al. 2014; 2016). The same set of hydrological model parameters are used for both the unit and integrated tests so that the same transformation of rainfall to flow is used.


-34.8
-34.9
-35.0
-35.1
-35.2
Lat
138.4
138.5
138.6
138.7
138.8
138.9
139.0
Lon
Adelaide
Happy Valley
Reservoir
Mt Bold Reservoir
Murray River Pipeline
Site (poor ann. flow)
Site (good ann. flow)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

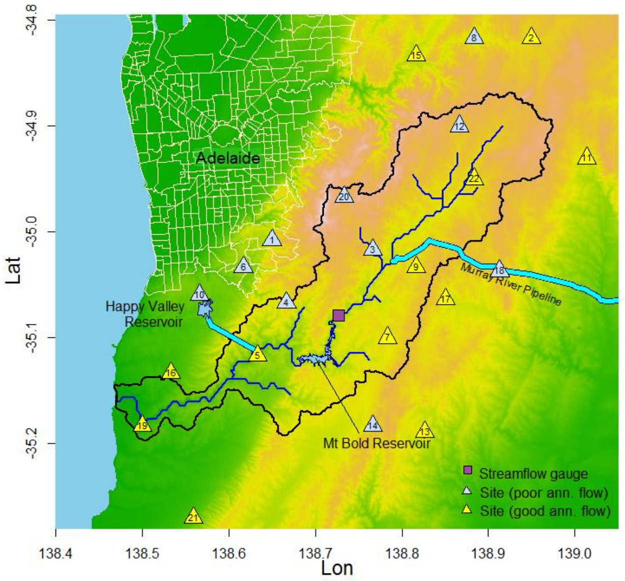


Fig. 5 Onkaparinga catchment, South Australia. Sites indicated by blue triangles are explored in greater detail in this paper due to the relatively poorer ability of simulated rainfall to reproduce annual streamflow totals at these sites.

Table 3 Site names ~~and~~, locations and seasonal rainfall characteristic summary. Sites ordered from lowest to highest elevation.

~~8~~ ~~Gumeracha~~ ~~346~~ ~~790~~ ~~19~~ ~~Old Noarlunga~~ ~~7~~ ~~520~~

						Site Name		Ann	~~Site No~~		January			~~Site Name~~		July		
		Site No					Elev (m)	Av. Rai	Total (mm)	No. Wet (days)	Daily Av. (mm)	Wet-spell (days)	Dry-spell (days)	Total (mm)	No. Wet (days)	Daily Av. (mm)	Wet-spell (days)	Dry-spell (days)
3	Bridgewater		376	1050	32	5.2	1	. 1.9	8.9 ~~Hah~~	~~347~~	~~1850~~	18		4.9		203.6	2.7	~~Uraid~~ 4
~~10~~14	Meadows~~Happ~~		~~148~~3	~~640~~87	30	~~21~~4.8	~~Willung~~	~~458~~	1.7	~~640~~		122	17	3.9		3.2		2.7
2	Birdwood		385	720	25	4.4	0.8		1.8		9.6	104	17	3.4		3.4		2.8
1	Belair		386	790	28	4.6	0.9		1.8		9.8	111	16	3.6		3.2		3
22	Woodside	19	387	800	27	~~Old~~ 4.3	0.9	520	1.8 20	4.1 3.3	0.6 ~~B~~	~~136~~1	~~700~~9	72.9	~~12~~1	~~Lobe~~ 2.8	~~4703~~	2.8 ~~802.~~
18	Nairne	16	403	680	28	Morphett 4.7	90 0.9	560	1.6 20	4.1 8	0.6	93 1.5	16 8.9	76 3	17	2.4 2.8	3.3	2.8 2.8
12	Lobethal	10	470	880	28	Happy 4.9	148 0.9	640	1.8 22	4.8 8.4	0.7	13 1.7	18 8.2	88 4.3	18	2.8 3.5	3.6	2.6 2.6
20	Uraidla	21	499	1090	35	Willunga 4.7	158 1.1	640	1.8 23	4 9	0.7	16 1.6	17 10	95 5.2	17	3 3.4	3.2	2.2 2.7
		5				Clarendon	223	820	25	4.7	0.8	1.7	8.9	114	17	3.7	3.4	2.8
		6				Coroman	234	710	24	4.8	0.8	1.8	9.2	102	18	3.3	3.6	2.8
~~2~~	~~Birdwood~~		~~385~~	~~720~~	13	Macclesfi	302	730	28	5.3	0.9	1.8	7.9	99	17	3.2	3	2.7
		~~3~~15				Cudlee	~~376~~	830	29	~~14~~5	~~Mead~~	~~384~~1	~~870~~8	123	18	3.9	3.8	2.7
		11				Harrogate	335	550	23	3.5	0.7	1.6	12	75	12	2.4	2.2	3.8
		4				Cherry	345	920	30	~~155.~~	~~Cudl~~	~~311~~1	~~830~~7	134	18	4.3	3.8	2.6
		8				Gumerach	346	790	27	5.3	0.9~~Cl~~	~~223~~1	~~820~~7	108	~~16~~1	~~Morp~~	~~90~~3.	~~560~~2.
		9				Hahndorf	347	850	29	5.4	0.9	1.9	8.1	123	18	4	3.4	2.7
~~6~~	~~Coromandel~~		~~234~~	~~710~~	17	Mount	349	770	28	5.9	0.9	1.9	7.2	104	18	3.3	3.3	2.6
		7				Echunga	375	805	28	~~18~~5	0.9~~N~~	~~403~~1	~~680~~8	110	17	3.5	3.3	2.6

Merged Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Deleted Cells
Deleted Cells
Deleted Cells
Deleted Cells
Deleted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells
Inserted Cells

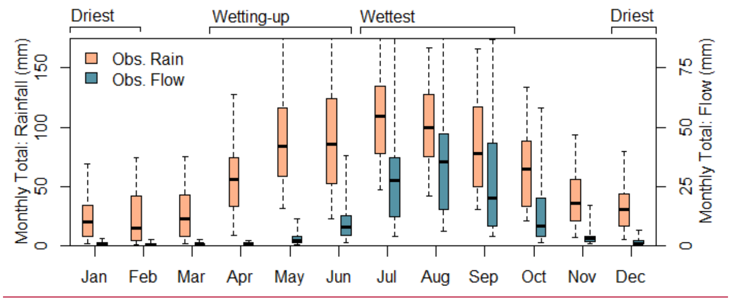


Fig. 6 Seasonal variation of catchment average rainfall and flow at Houlgrave Weir. Boxplots show the variation across years. Wettest indicates (high rainfall, high flow), direst indicates (low rainfall, low flow) and wetting-up indicates (high rainfall, low flow).

4 Results

4.1 ~~Following the selection of~~ Step 1 – Identify poor performing sites

To undertake Step1, annual total flow volumes were designated as the primary streamflow characteristic ~~(Step 1)~~to narrow the number of sites investigated. Following the selection of the primary streamflow characteristic and selection of the hydrological model ~~(Step 2), Step 3 was~~, GR4J, integrated tests were 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~~10 of the 22 sites as 'poor' and 12 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 2.2.3 ~~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 .~~ for category definitions).

~~4.1 Integrated test~~

~~compares the rainfall model's performance for the~~ The 10 sites categorised ~~'poor'~~ as 'poor' are the focus of subsequent hydrologic evaluation framework steps. These 'poor' performing sites are indicated by the blue triangles in Fig. 5.

4.2 Step ~~4~~2 – Identify poor performing time periods

The poor performing sites identified in Step 1 were then compared in terms of both an observed-rainfall evaluation and virtual hydrological evaluation via an integrated test. Fig. 7 ~~(Step 5),~~ graphically summarises this comparison, with each row ~~summarising~~presenting monthly or annual performance~~.~~

~~Although~~ of the following statistics:

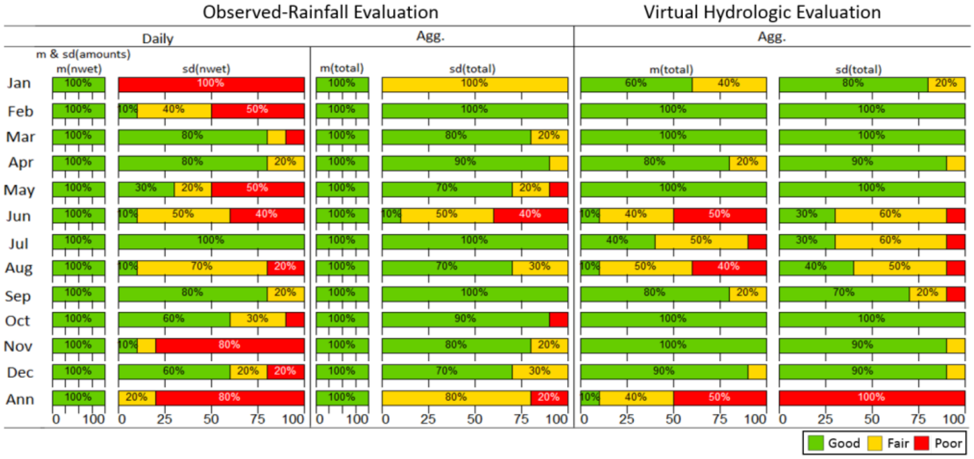
- simulated daily rainfall statistics (mean (*m*) daily amounts, standard deviation (*sd*) of daily amounts, mean number of wet days (*nwet*) and the standard deviation of the number of wet days);
- aggregate rainfall statistics (mean and standard deviation of total rainfall); and
- aggregate streamflow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

The first to fourth columns of Fig. 7 summarise the observed-rainfall evaluation and the fifth and sixth of Fig. 7 summarise the virtual hydrological evaluation. The first column of Fig. 7~~rainfall model performed well in simulating the total~~ indicates that of the poor performing sites the SRM exhibited 'good' performance in simulating daily rainfall means and standard deviations ~~for the majority of~~as well as the mean number of wet days for all sites and months and ~~sites, it~~at an annual level according to the observed-rainfall evaluation. Whereas the second column indicates that there is ~~clear from that 'poor'~~ mixed performance across sites and months in simulating the variability in the number of wet days (*sd(nwet)*). Likewise, the third and fourth columns indicate overall 'good' performance in simulating mean monthly totals and mixed performance in simulating the monthly or annual total standard deviations (*sd(total)*). Whereas the virtual hydrological evaluation (fifth and sixth) columns show mostly 'good' performance in all months other than those in the 'wettest' or 'wetting-up' periods.

A clear trend, from Fig. 7~~in rainfall~~ is the contrast in performance between the observed-rainfall evaluation and the virtual hydrological evaluation. One constrast is that, in the driest months (Dec, Jan, Feb) 'poor' performance in simulating rainfall (based on observed-rainfall evaluation) did not necessarily translate to 'poor' performance ~~is~~in simulating streamflow~~.~~ (based on virtual hydrological evaluation). For example, examining the first row of Fig. 7~~, it can be seen that~~ , the observed-rainfall evaluation shows that in January the ~~rainfall model's~~SRM's ability to simulate variability in the number of wet days, *sd(nwet)*, was 'poor' for all sites. However, in contrast the virtual hydrological evaluation 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,~~A second contrast is that 'good' performance in the observed-rainfall evaluation does not necessarily translate to 'good' performance for the virtual hydrological evaluation, particularly for months in the ~~simulation of monthly total flows.~~'wettest' and 'wetting-up' periods. For example, in Fig. 7 the rows summarising June and August ~~() have~~show large percentages of 'poor' sites ~~for simulating~~in the virtual hydrological evaluation of 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. 7~~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.



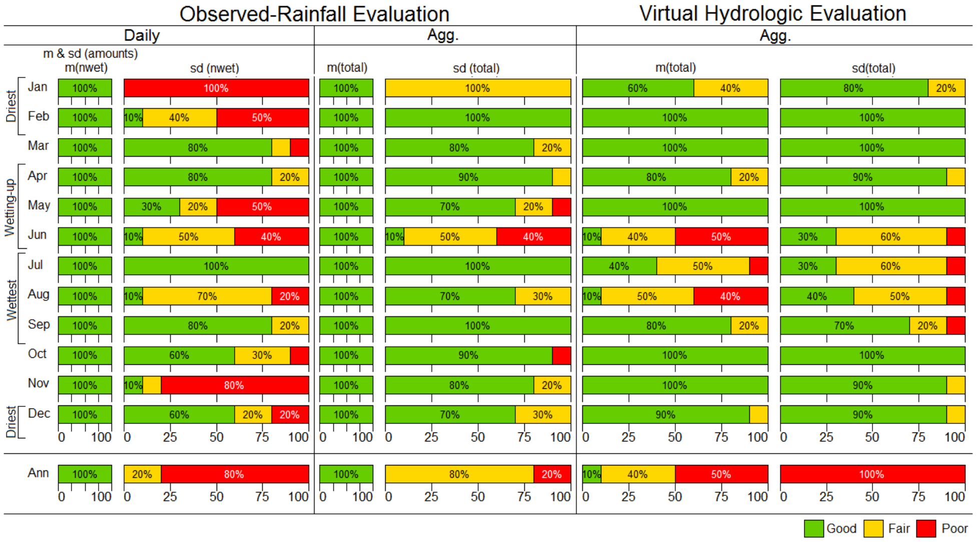


Fig. 7 Integrated test, comparing observed-rainfall evaluation (left) with the virtual hydrologic evaluation (right). Comparison of daily and aggregate ('Agg.') rainfall statistics against aggregate flow statistics for individual months and years. Daily rainfall statistics presented include the means 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~~

4.3 ~~The unit test was run for the subset of 10 sites (~~Step ~~6) and the results of the integrated and~~ 3 – Identify sources of poor performance

To undertake Step 3, unit tests were run to evaluate the source of deficiencies in poor performing time periods. The results of these test were compared ~~(Step 7). The unit test results and~~ against integrated tests in terms of their relative errors. From this comparison ~~with integrated test results~~the source and type of key deficiency in the simulated rainfall that lead to poor performance in simulated streamflow were identified. A comparison of the virtual-observed flow duration curve for the poor performing time periods and the flow duration curves resulting from unit tests for key influencing months was also undertaken to illustrate the impact of these key deficiencies on the daily flow duration curve.

Here, four examples of the different types of key deficiencies are illustrated ~~for~~using two locations, Site 12 and Site 10 (see Fig. 8 ~~and~~ to Fig. 13). For completeness these results are presented ~~alongside~~together with the results of the observed-rainfall evaluation (panels (a) and (b) of Fig. 8 and Fig. 11~~).~~).

~~4.2.1~~4.3.1 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. 8 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 in Fig. 9 since it shows biased flow.

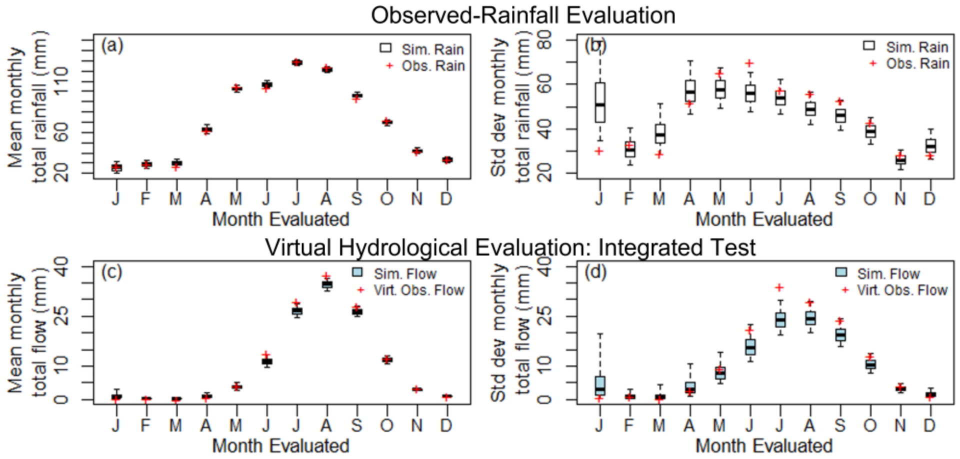


Fig. 8 Lobethal, Site 12 (a) observed-rainfall evaluation mean monthly total rainfall, (b) observed-rainfall evaluation standard deviation of monthly total rainfall, (c) virtual hydrological evaluation (integrated test) mean monthly total streamflow, (d) virtual hydrological evaluation (integrated test) standard deviation of monthly total streamflow. Boxplot whiskers indicate the 90% limits of the simulated streamflow or rainfall replicates.

Taking September as the evaluated month~~,~~ (*t*=9), Fig. 9~~ (e~~(a) and (~~f~~b) 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, *k*=9) the resultant error is greatest and closest to the error for the integrated test for both the mean monthly total flow (Fig. 9~~(e~~ (a)) and standard deviation of monthly total flow (Fig. 9~~(f~~ (c)). 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. 9~~(f~~ (b)). Therefore, to improve September flows, September rainfall should be improved in preference to all other months.

This need to improve September in preference to preceding months is also illustrated via Fig. 9 (c) where the September daily flow duration curves are shown for the cases where August (orange shading) and September (blue shading) are the influencing months compared against the virtual-observed September flow duration curve (purple dots). Where August is selected as the influencing month, the virtual-observed flow duration curve largely sits inside the 90% limits of the flow duration curves resulting from the unit testing procedure. Whereas, the virtual-observed flow duration curve sits outside the 90% limits of the unit test flow duration curve when September is taken as the influencing month. Thereby providing further evidence that to improve September flows that September rainfall should be improved in preference to 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.~~

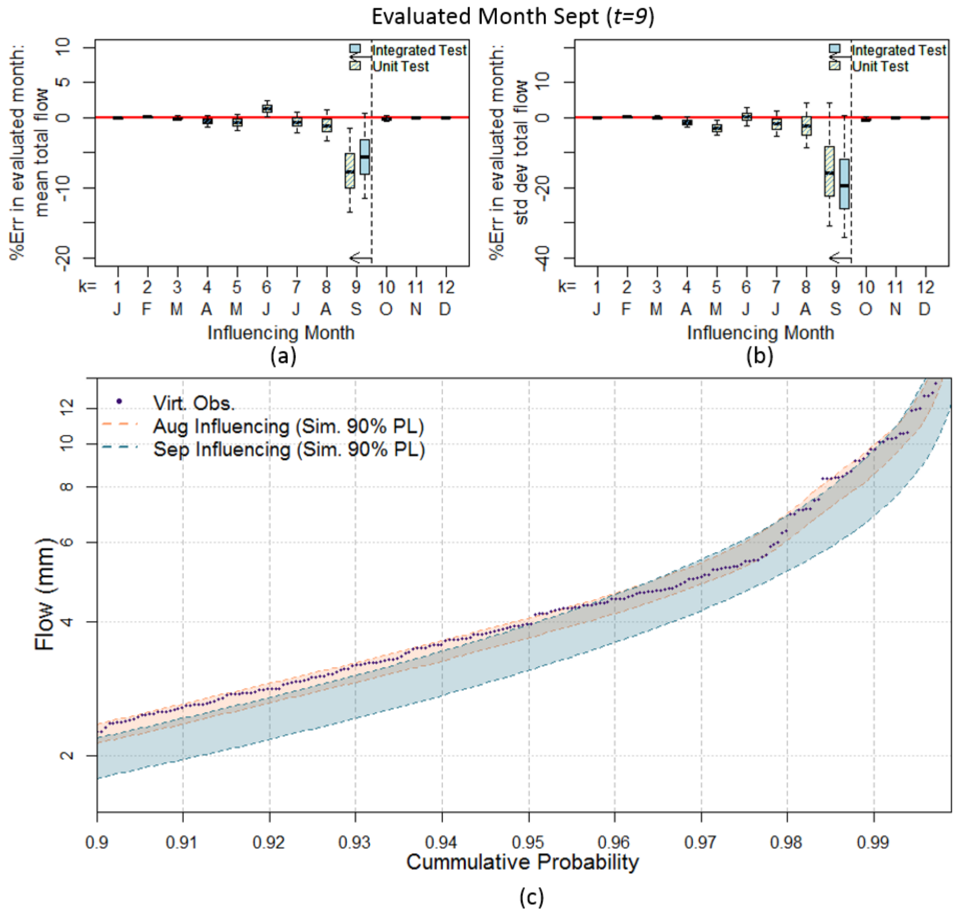


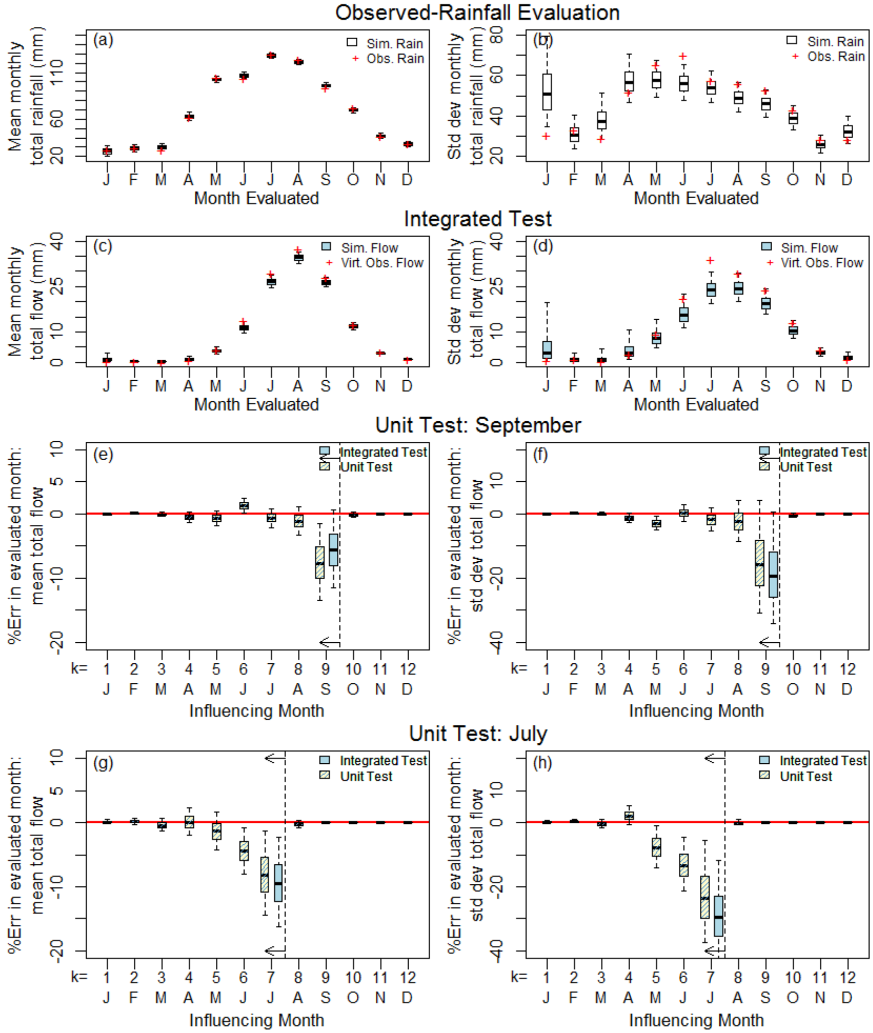
Fig. 9 Lobethal, Site 12 (a) unit test error in mean monthly flow (September), (b) unit test error in standard deviation of monthly flow (September), (c) Unit test September flow duration curve when August and September are selected as influencing months (top 10% of flow days shown). Boxplot whiskers indicate the 90% limits of the simulated streamflow replicates.

~~4.2.2~~4.3.2 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. 10~~ (g~~ (a) and (~~h~~b)) demonstrates that the errors in July streamflow do not originate in the July rainfall alone (unlike the case for September – see Section 4.3.1). 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.

This need to improve July and preceding months is also illustrated via Fig. 10~~Typically, autumn (transitional) and winter months (May to August)~~(c) where the July daily flow duration curves are shown for the cases where June (orange shading) and July (blue shading) are the influencing months compared against the virtual-observed July flow duration curve (purple dots). For both cases the virtual-observed flow duration curve sits outside the 90% limits of the flow duration curves resulting from the unit testing procedure.

Typically, ‘wetting-up’ and ‘wettest’ months 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.


Observed-Rainfall Evaluation
(a)
Sim. Rain
Obs. Rain
Mean monthly total rainfall (mm)
20
60
110
J F M A M J J A S O N D
Month Evaluated
(b)
Std dev monthly total rainfall (mm)
20
30
40
60
80
Integrated Test
(c)
Sim. Flow
Virt. Obs. Flow
Mean monthly total flow (mm)
0
10
25
40
(d)
Std dev monthly total flow (mm)
Unit Test: September
(e)
Integrated Test
Unit Test
%Err in evaluated month: mean total flow
10
5
0
-10
-20
k= 1 2 3 4 5 6 7 8 9 10 11 12
Influencing Month
(f)
%Err in evaluated month: std dev total flow
10
0
-20
-40
Unit Test: July
(g)
(h)

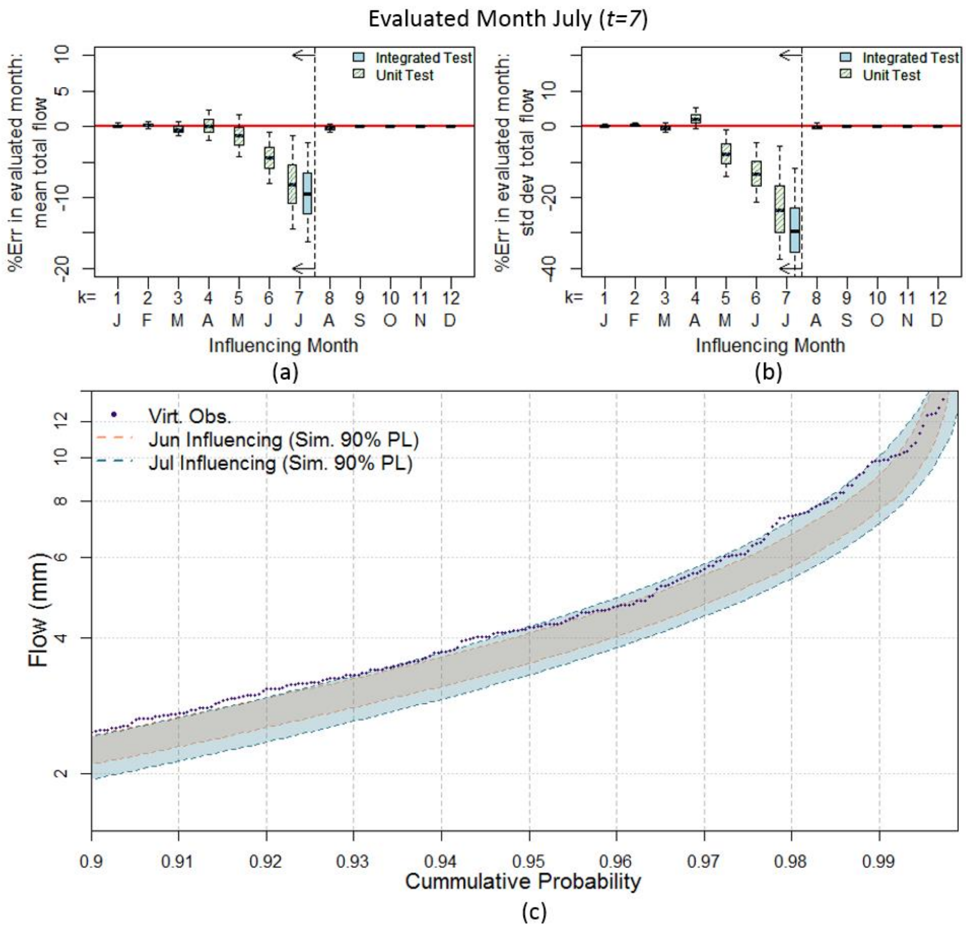


Fig. 10 Lobethal, Site 12 (~~90% limits shown) (a) observed-rainfall evaluation mean monthly total rainfall, (b) observed-rainfall evaluation standard deviation of monthly total rainfall, (c) integrated test mean monthly total rainfall, (d) integrated test standard deviation of monthly total rainfall (e) unit test error in mean monthly flow (September), (f) unit test error in standard deviation of monthly flow (September), (g~~a) unit test error in mean monthly total flow (July), and (~~h~~b) unit test error in standard deviation of monthly total flow (July~~).~~), (c) July flow duration curve when June and July are selected as influencing months in unit test (top 10% of flow days shown). Boxplot whiskers indicate the 90% limits of the simulated streamflow replicates.

~~4.2.3~~4.3.3 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. 11 ~~(~~(c) and (d)), but did not show any bias in the simulated rainfall (see Fig. 11 ~~(~~(a) and (b)).

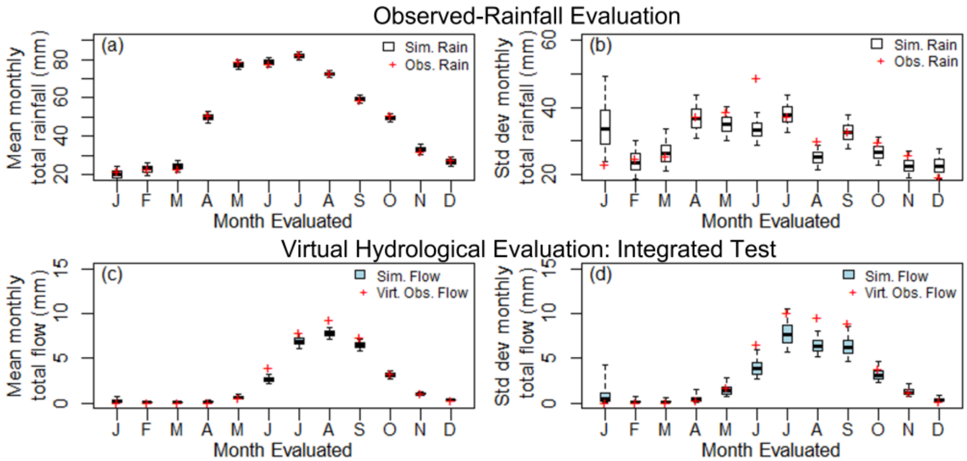


Fig. 11 Happy Valley (Site 10) (a) observed-rainfall evaluation mean monthly total rainfall, (b) observed-rainfall evaluation standard deviation of monthly total rainfall, (c) virtual hydrological evaluation (integrated test) mean monthly total streamflow (d) virtual hydrological evaluation (integrated test) standard deviation of monthly total streamflow. Boxplot whiskers indicate the 90% limits of the simulated streamflow or rainfall replicates.

The largest contributor to error in July flow is not July rainfall but June rainfall (Fig. 12 ~~(e~~(a) and (~~f~~b)). 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. This deficiency in June rainfall can also be seen in an examination of the July flow duration curves (Fig. 12(c)) where the virtual-observed streamflow sits within the 90% limits of the simulated flow duration curve where July is designated as the influencing month, whereas when June is designated as the influencing month the virtual-observed streamflow sits outside the 90% limits for a number of the higher flow days.

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.

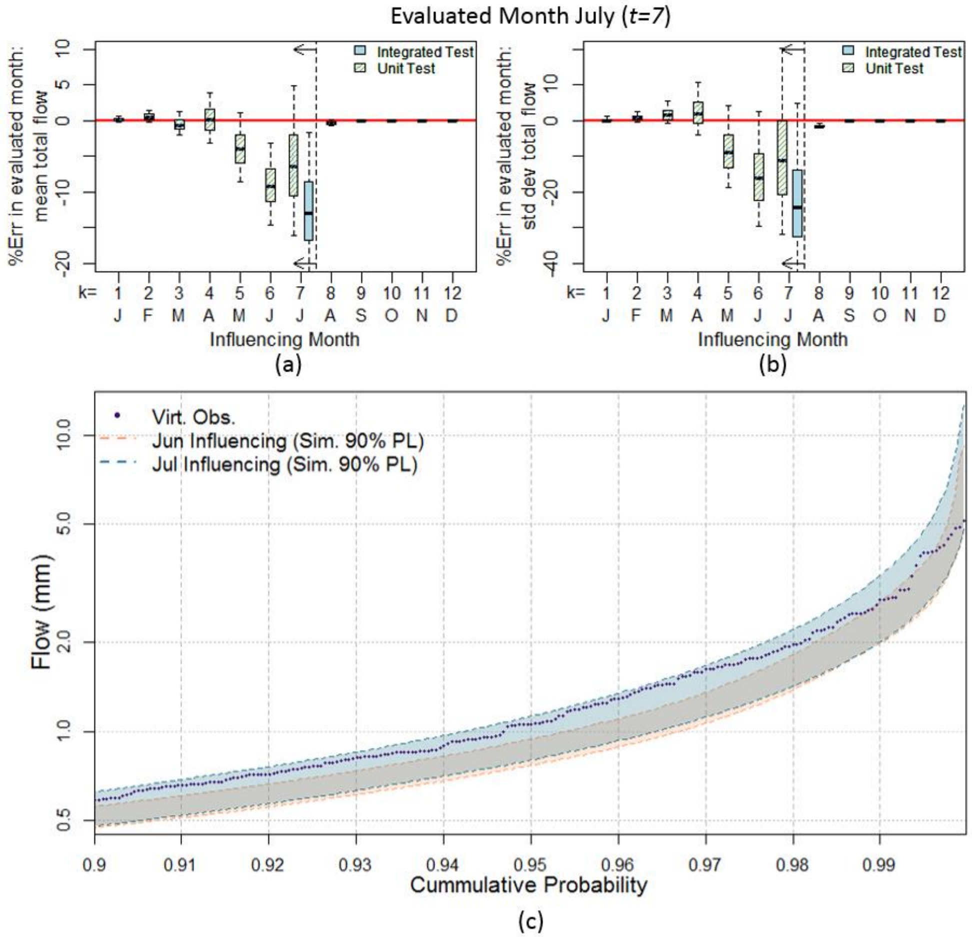
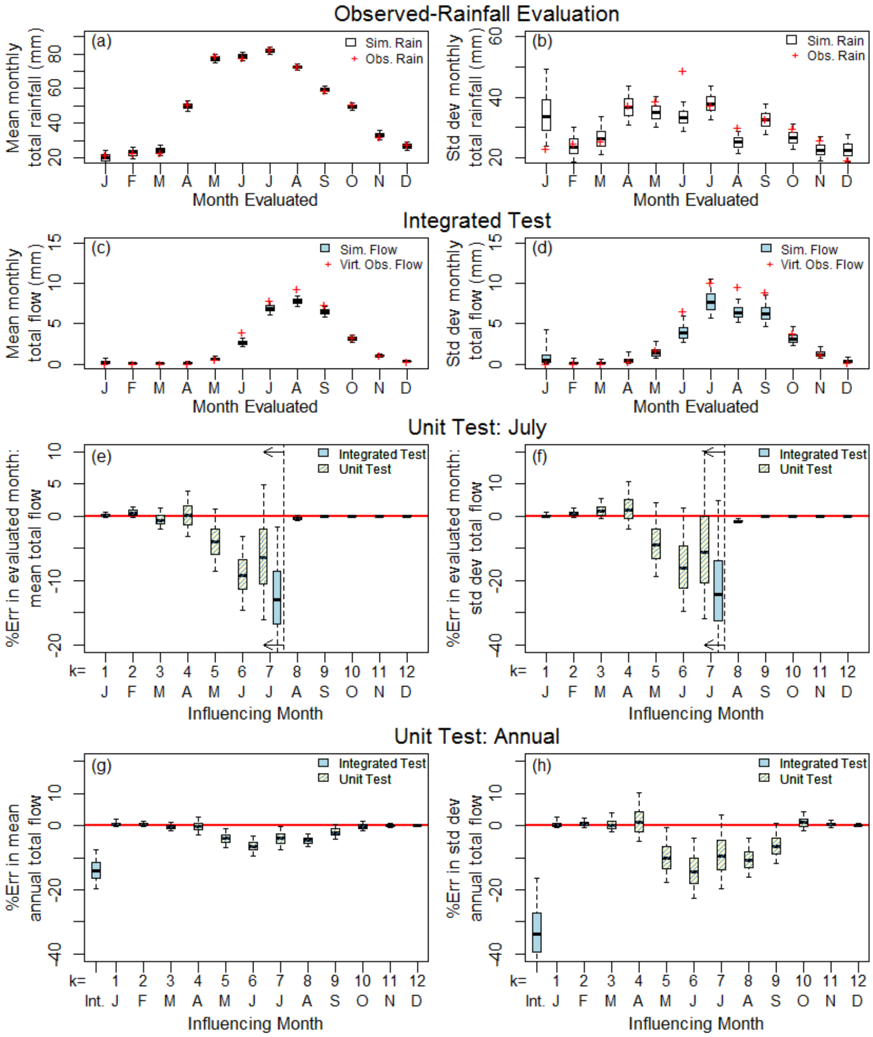


Fig. 12 Happy Valley (Site 10) (a) unit test error in mean monthly flow (July), (b) unit test error in standard deviation of monthly flow, and (c) July flow duration curve when June and July are selected as influencing months in unit test (top 10% of flow days shown). Boxplot whiskers indicate the 90% limits of the simulated streamflow replicates.

~~4.2.4~~4.3.4 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. 13~~ (g~~(a) and (~~h~~b) 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. This deficiency can also be seen via an examination of the flow duration curve (Fig. 13(c)) in which the virtual-observed flow duration curve sits outside portions of the simulated flow duration curves where May or June are designated as the influencing month. Improvements to the simulation of annual total flow will therefore come from improving the rainfall model in the ‘wetting-up’ and wettest months of the seasonal catchment cycle (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 5.2).

Field Code Changed


Observed-Rainfall Evaluation
(a)
Mean monthly total rainfall (mm)
20 40 60 80
□ Sim. Rain
+ Obs. Rain
J F M A M J J A S O N D
Month Evaluated
(b)
Std dev monthly total rainfall (mm)
20 40 60
□ Sim. Rain
+ Obs. Rain
J F M A M J J A S O N D
Month Evaluated
Integrated Test
(c)
Mean monthly total flow (mm)
0 5 10 15
Sim. Flow
+ Virt. Obs. Flow
J F M A M J J A S O N D
Month Evaluated
(d)
Std dev monthly total flow (mm)
0 5 10 15
Sim. Flow
+ Virt. Obs. Flow
J F M A M J J A S O N D
Month Evaluated
Unit Test: July
(e)
%Err in evaluated month: mean total flow
-20 -10 -5 0 5 10
Integrated Test
Unit Test
k= 1 2 3 4 5 6 7 8 9 10 11 12
J F M A M J J A S O N D
Influencing Month
(f)
%Err in evaluated month: std dev total flow
-40 -20 0 10
Integrated Test
Unit Test
k= 1 2 3 4 5 6 7 8 9 10 11 12
J F M A M J J A S O N D
Influencing Month
Unit Test: Annual
(g)
%Err in mean annual total flow
-40 -20 0 10
Integrated Test
Unit Test
k= 1 2 3 4 5 6 7 8 9 10 11 12
Int. J F M A M J J A S O N D
Influencing Month
(h)
%Err in std dev annual total flow
-40 -20 0 10
Integrated Test
Unit Test
k= 1 2 3 4 5 6 7 8 9 10 11 12
Int. J F M A M J J A S O N D
Influencing Month

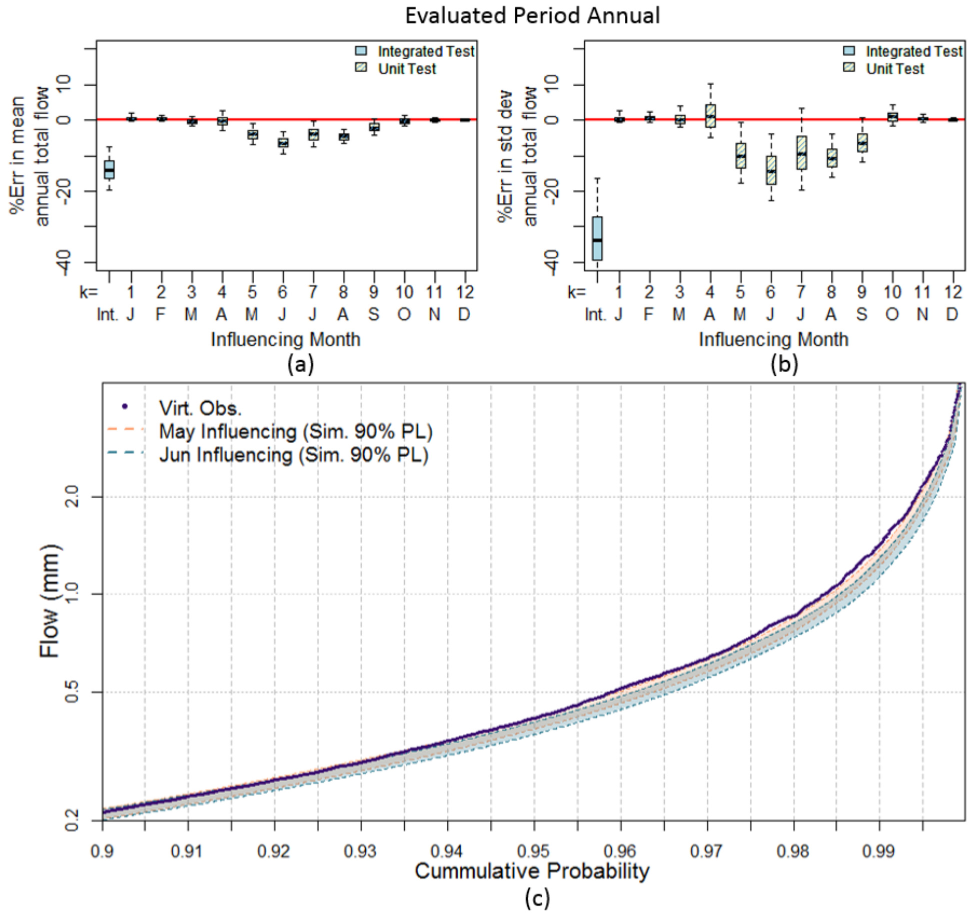


Fig. 13 Happy Valley (Site 10) (~~90% limits shown) (a) observed-rainfall evaluation mean monthly total rainfall, (b) observed-rainfall evaluation standard deviation of monthly total rainfall, (c) integrated test mean monthly total rainfall, (d) integrated test standard deviation of monthly total rainfall (e) unit test error in mean monthly flow (July), (f) unit test error in standard deviation of monthly flow, (g~~a) unit test error in mean annual total flow, and (~~h~~b) unit test error in standard deviation of annual total flow~~.~~, and (c) annual flow duration curve when May and June are selected as influencing months in unit test (top 10% of flow days shown). Boxplot whiskers indicate the 90% limits of the simulated streamflow replicates.

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. These behaviours were clearly identified and demonstrated in Step 2 of the virtual hydrological evaluation framework which compares observed-rainfall evaluation and virtual hydrological evaluation (see Section 4.2).

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. 7 shows that July rainfall statistics were classified as ‘good’, yet despite this, the streamflow response was ‘poor’ (see Section 4.2 for further discussion). It could be argued that the rainfall results presented in Fig. 7 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. 7 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~~benefits of the virtual evaluation framework

~~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 (). While there are several streamflow gauges in the Onkaparinga catchment, their records are shorter and complicated by inter-catchment pipeline transfers () 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 . 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 – see . 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 ((g) and (h)).~~A benefit of virtual hydrological evaluation is that it is a relative measure of performance, where the hydrological model is a common factor in the construction of virtual-observed and simulated streamflow. This enables discrepancies in the streamflow to be identified in terms of SRM features. In contrast, observed-streamflow evaluation is typically hampered by difficulties in separating the impact of data errors, hydrological model predictive performance from the errors in the SRM. A further benefit is the ability to undertake streamflow-based evaluation at any site where rainfall is observed and simulated. This enables insights into the SRM performance for simulating streamflow on a site-by-site basis.

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~~SRM. This paper introduces an innovative 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~~This unit test~~,~~ provides added value over and above the integrated test because it identifying which are the influencing months which have deficiencies in the modelled rainfall that produce poor streamflow predictions. For example, Section 4.3.2 illustrated that while the integrated test identified that were was poor streamflow in July for Site 12, the unit test was able to identify that the simulated rainfall in the prior influencing months of both May and June (Fig. 10) made significant contributions (10-15%) errors to July's poor streamflow. A second example is shown in the influence of monthly rainfall on the errors in annual flow volumes in Section 4.3.4~~ 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~~The unit in Section 4.3.4 show, May and June combined ~~contribute~~make up 13% ~~in~~of the total annual ~~total~~ flow volume (Fig. 11~~ (~~(c~~)), compared~~)). However, they contribute to 11% of the error in the mean annual total flow (Fig. 13~~ (g~~(a)) and 24% error in the standard deviation (Fig. 13~~ (h~~(b)). 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~~SRM improvement than a high-flow month such as September.

5.3 Limitations and future research

The formalisation of the virtual hydrological framework for SRM evaluation provides the opportunity for further improvements in the future, including:

(i) Using multiple, well-tested hydrological models - a potential limitation of the virtual hydrologic evaluation framework is that it is reliant on the use of a hydrological model. Hydrological structural errors may potentially skew interpretation of the SRM evaluation if the hydrological model poorly represents the catchment processes. To reduce these impacts the steps taken in this study included (a) using a well-tested hydrological model that has demonstrated good performance on a wide range of catchments (e.g. the GR4J model has been widely tested , see (Perrin et al., 2003, Coron et al., 2012); (b) calibrating and evaluating the hydrological on a catchment close to the observed rainfall sites to ensure it provided sufficiently good performance (e.g. GR4J was calibrated to the Onkaparinga catchment - see (Westra et al., 2014a, Westra et al., 2014b)~~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.~~

. Future research will use multiple, well-tested hydrological models with sufficiently good performance to reduce the reliance on a single hydrological model and ensure the identification of SRM deficiencies is not dependent on a single hydrological model.

(ii) Comparison of SRMs – this framework can be extended to provide more direct guidance on which rainfall features (in terms of components of the SRM) should be modified to improve streamflow performance. This can be done by comparing multiple rainfall model variants (parametrically, or via bootstrap techniques) which are designed to have contrasting features of a key characteristic (e.g. intermittency, rainfall correlation). Such an approach was undertaken by Evin et al. (2018) using an observed-rainfall evaluation approach. If the SRMs have monthly/seasonal autocorrelation (these were not significant for the rainfall in the Onkaparinga catchment) the unit testing approach would need to be extended by conditionally sample the simulated rainfall in a manner that preserves monthly correlations.

(iii) Evaluation of temporal non-stationarity – this framework can be extended to evaluate the impact of non-stationarity on SRM model performance by applying it on a selected non-stationary period. Care would be needed in the selection of statistics to identify model performance (since the performance in different sub-periods could be masked when evaluating an overall period). A related issue is that the hydrological model should provide adequate performance across the range of non-stationary climate forcings to which it is subjected.

(iv) Evaluation of spatial performance – there are multiple opportunities to develop tests for spatial performance including (a) repeating the integrated test for all sites and for catchment average rainfall means it would be possible to diagnose whether specific locations or the spatial dependence causes poor reproduction of streamflow statistics, (b) developing a spatial unit test (which is analogous to the temporal unit test but extended to space) where different combinations of sites are ‘spliced’ in the construction of catchment average rainfall – to evaluate the impact of ‘mixed’ performance in the SRMs between sites on the catchment average rainfall, and (c) these spatial unit tests could be used to evaluate stochastic weather generators (SWG) more generally as well as spatially distributed SRGs – though these would require a spatially distributed hydrological model.

6 Conclusions

This paper has introduced ~~the~~a formal 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~~formalises virtual streamflow investigations by (1) using a comprehensive and systematic evaluation (CASE) approach to evaluate performance (2) introducing 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~~‘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 framework provides insights not available through conventional approaches and ~~should be an essential step in~~provides useful diagnostic ability for the development and application of ~~stochastic rainfall models~~SRMs.

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. We thank the anonymous reviewers for their constructive comments and feedback that helped substantially improved the manuscript.

8 Author contributions

BSB conceived and executed the analysis, with input from MT, ML, MFL and BCB. MT, ML, MFL and BCB have contributed to assisting with method development, interpretation and analysis. BSB, MT and ML mostly wrote, reviewed and revised the paper.

9 Data availability

All the data used in this study can be requested by contacting the corresponding author Bree Bennett at bree.bennett@adelaide.edu.au.

~~8~~10 References

ANDREASSIAN, V., PERRIN, C., MICHEL, C., USART-SANCHEZ, I. & LAVABRE, J. 2001. Impact of imperfect rainfall knowledge on the efficiency and the parameters of watershed models. *Journal of Hydrology*, 250**,** 206-223.

ANG, A. & TANG, W. 2007. *Probability Concepts in Engineering: Emphasis on Applications to Civil and Environmental Engineering (v. 1)*.

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

BAXEVANI, A. & LENNARTSSON, J. 2015. A spatiotemporal precipitation generator based on a censored latent Gaussian field. *Water Resources Research*.

BENNETT, B., THYER, M., LEONARD, M., LAMBERT, M. & BATES, B. 2018. A comprehensive and systematic evaluation framework for a parsimonious daily rainfall field model. *Journal of Hydrology,* 556**,** 1123-1138.

BLAZKOVA, S. & BEVEN, K. 2002. Flood frequency estimation by continuous simulation for a catchment treated as ungauged (with uncertainty). *Water Resources Research,* 38**,** 14-1-14-14.

BLAZKOVA, S. & BEVEN, K. 2009. 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.

CAMICI, S., TARPANELLI, A., BROCCA, L., MELONE, F. & MORAMARCO, T. 2011. Design soil moisture estimation by comparing continuous and storm-based rainfall-runoff modeling. *Water Resour. Res.,* 47**,** W05527.

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

CORON, L., ANDREASSIAN, V., PERRIN, C., LERAT, J., VAZE, J., BOURQUI, M. & HENDRICKX, F. 2012. Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments. *Water Resources Research,* 48.

COXON, G., FREER, J., WESTERBERG, I., WAGENER, T., WOODS, R. & SMITH, P. 2015. A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations. *Water resources research,* 51**,** 5531-5546.

EVIN, G., FAVRE, A.-C. & HINGRAY, B. 2018. Stochastic generation of multi-site daily precipitation focusing on extreme events. *Hydrology and Earth System Sciences,* 22**,** 655-672.

EVIN, G., THYER, M., KAVETSKI, D., MCINERNEY, D. & KUCZERA, G. 2014. Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity. *Water Resources Research,* 50**,** 2350-2375.

HENLEY, B. J., THYER, M. A. & KUCZERA, G. 2013. Climate driver informed short- term drought risk evaluation. *Water Resources Research,* 49**,** 2317-2326.

KHEDHAOUIRIA, D., MAILHOT, A. & FAVRE, A.-C. 2018. Daily Precipitation Fields Modeling across the Great Lakes Region (Canada) by Using the CFSR Reanalysis. *Journal of Applied Meteorology and Climatology,* 57**,** 2419-2438.

KIM, D. & OLIVERA, F. 2011. Relative importance of the different rainfall statistics in the calibration of stochastic rainfall generation models. *Journal of Hydrologic Engineering,* 17**,** 368-376.

KUCZERA, G. & WILLIAMS, B. J. 1992. Effect of rainfall errors on accuracy of design flood estimates. *Water Resources Research,,* 28**,** 1145-1154.

LI, J., THYER, M., LAMBERT, M., KUCZERA, G. & METCALFE, A. 2014. An efficient causative event-based approach for deriving the annual flood frequency distribution. *Journal of Hydrology,* 510**,** 412-423.

LI, J., THYER, M., LAMBERT, M., KUZERA, G. & METCALFE, A. 2016. Incorporating seasonality into event-based joint probability methods for predicting flood frequency: A hybrid causative event approach. *Journal of Hydrology,* 533**,** 40-52.

MCINERNEY, D., THYER, M., KAVETSKI, D., LERAT, J. & KUCZERA, G. 2017. Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors. *Water Resources Research,* 53**,** 2199-2239.

MCMILLAN, H. K. & BRASINGTON, J. 2008. End-to-end flood risk assessment: A coupled model cascade with uncertainty estimation. *Water Resources Research,* 44**,** W03419.

MORTAZAVI-NAEINI, M., KUCZERA, G., KIEM, A. S., CUI, L., HENLEY, B., BERGHOUT, B. & TURNER, E. 2015. Robust optimization to secure urban bulk water supply against extreme drought and uncertain climate change. *Environmental Modelling & Software,* 69**,** 437-451.

MÜLLER, H. & HABERLANDT, U. 2018. Temporal rainfall disaggregation using a multiplicative cascade model for spatial application in urban hydrology. *Journal of Hydrology,* 556**,** 847-864.

NICÓTINA, L., ALESSI CELEGON, E., RINALDO, A. & MARANI, M. 2008. On the impact of rainfall patterns on the hydrologic response. *Water Resources Research,* 44.

PASCHALIS, A., MOLNAR, P., FATICHI, S. & BURLANDO, P. 2013. A stochastic model for high‐resolution space‐time precipitation simulation. *Water Resources Research,* 49**,** 8400-8417.

PATON, F. L., MAIER, H. R. & DANDY, G. C. 2013. 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.

PERRIN, C., MICHEL, C. & ANDRÉASSIAN, V. 2003. Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology,* 279**,** 275-289.

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

RENARD, B., KAVETSKI, D., LEBLOIS, E., THYER, M., KUCZERA, G. & FRANKS, S. W. 2011. Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation. *Water Resour. Res.,* 47**,** W11516.

SHAH, S. M. S., O'CONNELL, P. E. & HOSKING, J. R. M. 1996. Modelling the effects of spatial variability in rainfall on catchment response. 2. Experiments with distributed and lumped models. *Journal of Hydrology,* 175**,** 89-111.

SIKORSKA, A. E., VIVIROLI, D. & SEIBERT, J. 2018. Effective precipitation duration for runoff peaks based on catchment modelling. *Journal of Hydrology,* 556**,** 510-522.

SRIKANTHAN, R. & PEGRAM, G. G. S. 2009. A nested multisite daily rainfall stochastic generation model. *Journal of Hydrology,* 371**,** 142-153.

WESTRA, S., THYER, M., LEONARD, M., KAVETSKI, D. & LAMBERT, M. 2014a. Impacts of climate change on surface water in the Onkaparinga catchment-Final report volume 1: hydrological model development and sources of uncertainty.

WESTRA, S., THYER, M., LEONARD, M., KAVETSKI, D. & LAMBERT, M. 2014b. A strategy for diagnosing and interpreting hydrological model nonstationarity. *Water Resources Research,* 50**,** 5090-5113.

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

WILSON, C. B., VALDES, J. B. & RODRIGUEZ-ITURBE, I. 1979. On the influence of the spatial distribution of rainfall on storm runoff. *Water Resources Research,* 15**,** 321-328.

Commented [BSB1]: References added\ include:

Khedhaouria et al. 2018
Kim & Olivera (2011)
Muller & Haberlandt (2018)
Sikorska et al. (2018)
Westra et al (2014b)