Response to Editor

I have now received the reports from the referees. While the first reviewer mentioned several minor issues to be corrected, the second referee has more critical comments and suggestions that I would like you to consider. I kindly ask you to revise the manuscript once again. The revised manuscript will then be re-evaluated.

Looking forward to reading the revised text and best regards.

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 two reviewers and editor. The key modifications we have made in the revised paper include:

- Improved the description of 'virtual experiments' in Section 1 (see Response 5).
- Revised Figure 1 to indicate the 'special' status of virtual-observed streamflow within the figure (see Response 4).
- Added clarification surrounding the differences between the CASE framework presented in Bennett et al. (2018) and the virtual hydrological evaluation framework presented in this manuscript (see responses to comments 15, 20, 22 and 45).
- Rephrased the first paper objective to clarify that the virtual hydrological evaluation framework is a novelty but draws on the systematic evaluation and tests introduced by the CASE framework.
- Incorporated additional details and clarified phrasings where suggested to aid reader understanding (see responses to comments 2-3, 4-8, 10, 15, 18-20, 22-23, 30-33, 36, 38, 41-44, 46-52).
- Revised Figures 3, 5, 9-10 and 12-13 as well as Table 3 to improve reader comprehension.

In particular we would like to thank Reviewer #2 for their thoroughness in raising over 40 issues and suggestions in this second revision of the manuscript. Their attention to detail has greatly improved this manuscript. We agree and addressed 90% of this reviewer's comments. The issues we have not addressed were either (i) issues previously raised in the 1st round of reviews and responded to in the 1st response (we have included the relevant text in this 2nd response for completeness); (ii) Deemed beyond the scope of the current paper, which already quite lengthy (revised paper ~10,300 words, 13 Figures, 3 Tables). Responding to these issues thoroughly would require substantial analysis better suited to a dedicated separate future paper to ensure they are addressed and presented in a meaningful way.

Our item-by-item responses to the specific comments appear below.

Response to Reviewer 1

Comment 1:

Second Review of HESS-2018-489: "A virtual hydrological framework for evaluation of stochastic rainfall models" by Bennett et al.

The authors have by and large addressed my concerns, and I think the manuscript is almost ready for publication. I have a few minor recommendations:

Response 1:

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

- Improved the description of 'virtual experiments' in Section 1 (see Response 5).
- Revised Figure 1 to indicate the 'special' status of virtual-observed streamflow within the figure (see Response 4).
- Incorporated the recommended phrasing changes to improve the clarity of explanations (see Responses 2-3, 6-8).

Greater detail is given in response to specific comments below.

Comment 2:

Pg 1 Line 10: I recommend changing "performance evaluation methods" to "developing methods to evaluate their performance". It isn't actually clear what you're referring to as it is written now.

Response 2:

Done. The sentence has be re-written as suggested.

Comment 3:

Pg 1 Line 22: change "higher" and "lower" to "high" and "low", since the current wording is ambiguous- higher or lower than what?

Response 3:

Fixed. The sentence has be re-written as suggested.

Comment 4:

Figure 1: The "virtual observed streamflow" box should definitely not be yellow, since it isn't observed. It should either be blue, or, I would suggest a diagonal line running through the box, with half being blue and half being yellow to indicate its "special" status.

Response 4:

Thank you for pointing this out. Figure 1 has been amended to that the 'special' status of virtual-observed streamflow is clearly indicated. Virtual-observed streamflow is denoted via the use of a diagonal line running through the box with half shaded blue and half shaded yellow. Figure 1(c) and the key are reproduced below:

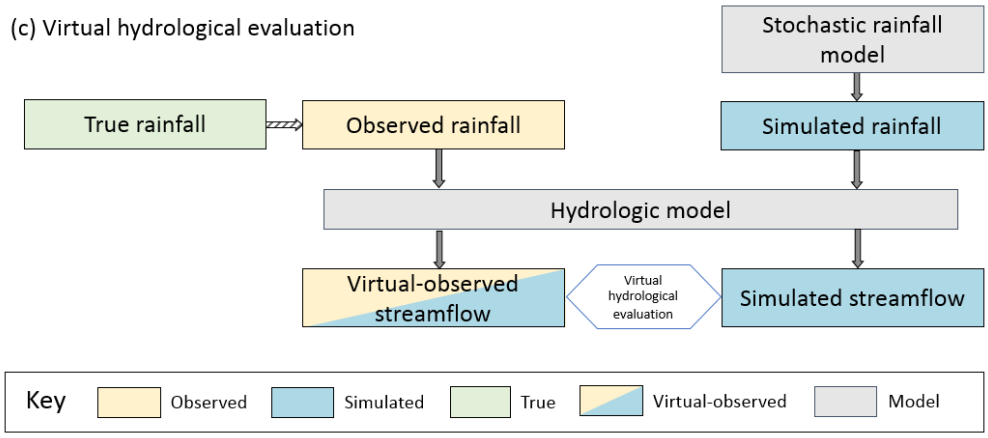


Figure 1 - Excerpt of revised Figure 1(c) in manuscript now indicating 'special' status of virtual-observed streamflow.

Comment 5:

Page 4 line 21: carefully define what is meant by virtual experiment.

Response 5:

The paragraph has been updated to define 'virtual experiments' as experiments that focus on comparisons between streamflow simulated under different conditions or inputs without relying on comparisons to observed streamflow. The paragraph is reproduced below:

> Page 4, Line 21: '*To date, 'virtual experiments', that is, experiments that focus on comparisons between streamflow simulated under different conditions or inputs (i.e. virtual streamflow) without relying on comparisons to observed streamflow, have been used in a variety of contexts. Examples include (i) the evaluation of hydrological model sensitivity (e.g. 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 (e.g. Sikorska et al., 2018), (ii) the development of new techniques for flood frequency analysis (e.g. Li et al., 2014, 2016), and (iii) the calibration, validation and selection of SRMs (e.g. Müller and Haberlandt, 2018, Kim and Olivera, 2011).'*

Comment 6:

Page 4 line 21: Also, in this paragraph, references should be of the style "e.g." since these are likely just a few examples. The general idea behind the integrated test gets used over and over in hydrologic subdisciplines, so the referencing in this paragraph is in no way comprehensive.

Response 6:

Fixed. All referenced examples are prefaced with 'e.g.' to clarify that the lists are not exhaustive.

Comment 7:

Pg 6 line 15: replace "fit for purpose" with "suitable"

Response 7:

Done. 'Suitable' has been inserted in place of 'fit for purpose'.

Comment 8:

Section 5.3: There are a number of small issues with parentheses in the wrong place on references. Also pg 27 line 29 appears to be missing the word "model" and pg 28 line 6 should say "sampling" rather than "sample"

Response 8:

Fixed.

Response to Reviewer 2

Comment 9:

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.

I'm involved as reviewer for the second time. The revised manuscript has been significantly improved in response to the review comments and most of the issues raised in the first round of reviews have been addressed in a satisfactory way. The presentation and discussion of the method and the results is now clearer and more robust with the FDCs. However, there is still the need of improving some presentational aspects, especially in terms of method description. I have some follow-up questions on the responses and changes, plus some new suggestions, which are necessary to consider for the readability of the manuscript. Thus I would recommend moderate revisions for the manuscript.

Response 9:

Thank you. We are pleased that the reviewer notes the significant improvement as a result of the first round of reviewer comments and that the topic is broad interest for the hydrological scientific community. We thank the reviewer for this thorough approach and attend to detail. Over 90% of the 44 comments raised by the reviewer have been addressed.

The suggestions and follow-up questions provided by the reviewer in this second round of comments have again led to an improved version of the manuscript. Regarding the specific matters raised:

- Added clarification surrounding the differences between the CASE framework presented in Bennett et al. (2018) and the virtual hydrological evaluation framework presented in this manuscript (see responses to comments 15, 20, 22 and 45). We also rephrased the first paper objective to clarify that the virtual hydrological evaluation framework is a novelty but draws on the systematic evaluation and tests introduced by the CASE framework.
- Incorporated additional details and clarified phrasings where suggested to aid reader understanding (see comments 10, 15, 18-20, 22-23, 30-33, 36, 38, 41-44, 46-52).
- Revised Figures 3, 5, 9-10 and 12-13 as well as Table 3 to improve reader comprehension.

Greater detail is given in response to specific comments below.

Comment 10:

P1l29 „…evaluating…the efficacy of SRM's…comparisons to observed rainfall or streamflow are limited." Maybe the authors want to replace the „or" with „and" to emphasize the streamflow comparison. Comparisons of generated rainfall time series with observed ones are state of the art and applied in the majority of rainfall generation manuscripts (as the authors point out later), so I would not consider the body of literature as „limited".

Response 10:

Thank you for pointing out that the sentence is ambiguous. The intention of this sentence was to point of that current approaches, while state-of-the-art, have limitations in their ability to evaluate the efficacy of SRM's in terms of the resulting streamflow. The sentence has been reworded as follows to avoid ambiguity:

> Page 1, Line 28: '*When evaluating the efficacy of SRM's, current approaches that make comparisons to observed rainfall or streamflow have limited diagnostic ability. They are unable to make a targeted evaluation of the SRM's ability to reproduce streamflow characteristics of practical interest.'*

Comment 11:

P2l16 „poor predictive performance" In general, the rainfall-runoff (r-r) models are calibrated in before with observed time series. This pre-calibration and a subsequent comparison of simulated and observed runoff characteristics enable conclusions, if the hydrological model can be used for the estimation of the runoff characteristics. Hence, for the following „observed-streamflow evaluation" it is from my point of view not challenging to ascertain the „poor performance" as the authors point out later (p2l17). Maybe the authors can implement a discussion on that issue in the manuscript. In general, for the calibration of the r-r models several possibilities exist (using observed time series with a shorter time length than analysed later in the comparison or with lower network density or with coarser temporal resolution or from different data sources (satellite or radar data instead of station-data),….) and can be discussed in the manuscript.

Comment 12:

P2l10 Data errors do not necessarily occur only for single catchments and not for others, they can also appear in one catchment, but for a limited time period only. In combination with my previous comment, data errors should be identified before or during the calibration process. In general, I would exclude „data errors" and "r-r model structural errors" from the motivation for the introduced framework.

Responses 11–12:

The reviewer raises issues around the difficulty of managing 'r-r model structural errors' and the presence of 'data errors'.

In the authors' experience the 'r-r model structural errors' are ubiquitous. This is often because simplified models are used to represent complex catchment processes. Therefore, these structural errors can have the error of masking the impact of errors in the SRM.

We agree with reviewer's comment that data errors can occur at any time. We also agree that ideally data errors would be identified before and during calibration. However, this is not routinely done, because typically it is quite hard to do and requires a great deal of both effort and information that is not routinely available. For example, data errors include runoff errors, which can be estimated if the rating curve (and any changes) is available, using Bayesian rating curve analysis (e.g. BaRatin; Le Coz et al. (2014)). Data errors also include errors in estimating catchment average rainfall from a limited number of rainfall gauges. These are quite challenging to estimate because the 'true' catchment rainfall is typically unknown (Renard et al., 2010), has high uncertainty often with standard errors exceeding 20%-30%, especially if the gauge network is sparse (Thyer et al., 2009, Linsley et al., 1982), and requires advanced hierarchical Bayesian techniques to integrate as part of the hydrological model calibration process (Renard et al., 2011). Typically, modellers take the best available runoff data, and catchment average rainfall data based on available data (with associated errors). Hence, for the majority of practical catchments, data errors are common, and cannot be excluded from model calibration process.

Both of these sources of error contribute to errors in the observed streamflow evaluation and mask the impact of the errors in the SRM. Based on this, we are of the opinion that they should be included as motivation for the framework.

Comment 13:

P4l15-17 Observed runoff is not required for the virtual hydrological evaluation, but it is for the calibration of the model in before.

Response 13:

Yes, in the calibration of the hydrological model observed streamflow would be used. However, it is not the basis of the streamflow evaluation for the virtual hydrological framework. The virtual hydrological evaluation framework is based on a relative comparison between the virtual-observed streamflow and the simulated streamflow–as both simulated and observed rainfall undergo transformation by the same process representation (i.e. the hydrological model). Provided that the hydrological model is calibrated such that the

necessary catchment streamflow features can be simulated (perhaps even via calibration to a neighbouring catchment with similar catchment properties) insight can be gained from this relative comparison.

Comment 14:

P5l1 The authors state that the framework "categorises performance at multiple spatial and temporal scales". The case study includes one catchment, simulated with a lumped r-r model lumped and with daily resolution. What are the multiple spatial and temporal scales?

Response 14:

The virtual hydrological evaluation framework is designed to evaluate the performance of SRMs (and not hydrological models). Hence the reference to multiple temporal scales refers to the evaluation of the SRM at the daily, monthly and annual scale in this manuscript and the ability to test the simulated rainfall at multiple sites or in aggregate across the catchment.

Comment 15:

P5l5 The authors state that one key objective of the manuscript is the "introduction of a formalised framework for the virtual hydrological evaluation of SRMs", but the authors state before (p4l28) that this framework was developed before by Bennett et al. (2018, reference is missing in the reference list). What is the difference between both frameworks? Is the "introduction of the framework" still a novelty or is it the application of the existing framework introduced by Bennett et al. (2018)? Throughout the manuscript Bennett et al. (2018) is referenced. It is necessary to explain the differences between both studies and to enable a full understanding of the applied method/framework without reading the reference of Bennett et al. (2018).

Response 15:

The Comprehensive And Systematic Evaluation (CASE) framework presented in Bennett et al. (2018) is an observed-rainfall evaluation framework. The CASE framework enables the systematic comparison of SRM performance through the use of a quantitative classification system that is applied at multiple spatial and temporal scales. The CASE framework, as presented in Bennett et al. (2018), is an observed-rainfall evaluation framework and therefore it does not consider the simulation of streamflow by the SRM as part of its evaluation.

In this manuscript the underlying principles presented from the CASE framework (based on observed-rainfall evaluation) are adopted and as extended to enable a systematic and comprehensive evaluation within the virtual hydrological evaluation framework (based on virtual-streamflow evaluation). The virtual hydrological evaluation framework also introduces two types of virtual experiments. None of these advances were part of the Bennett et al. (2018) paper. Hence, we believe this study represents a substantial advance on Bennett et al. (2018).

To help the reader understand this, clarification of the differences between the two frameworks has been added to the revised manuscript:

> Page 5, Line 1: '*The framework presented in this paper is a significant advance from previously reported virtual experiments because it presents a formal framework to identify key deficiencies in the SRM by (1) extending the comprehensive and systematic evaluation (CASE) framework (developed by Bennett et al., 2018 for observed-rainfall evaluation 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 for use in virtual hydrological evaluations, and (2) utilising two types of virtual experiments that are able to identify the source of key deficiencies in SRM at specific locations and time periods.*'

Additionally, the first paper objective has been re-worded to clarify that the virtual hydrological evaluation framework is a novelty but draws on the systematic evaluation and tests introduced by the CASE framework.

> Page 5, Line 8: '*1. To introduce a formalised framework for the virtual hydrological evaluation of SRMs: the new framework is a stepwise procedure that enables the identification poor performing sites, then*

poor performing time periods and then the key deficiencies in the SRM for those sites and time periods by drawing on the systematic application of quantitative performance criteria.'

We do not believe the reference to Bennett et al. (2018) was missing and is given on Page 29, Line 27 of the manuscript and is also reproduced below. This reference was a part of the original submission to HESS and has been included in all revisions.

> 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.

Comment 16:

P5l23-26 For the comparison of different rainfall data sets (observed vs. generated time series) the extraneous variables are kept. This introduces a new bias, since extraneous variables from a rainy day can occur with a dry day from the generated rainfall time series (and vice versa), while for the observed time series always a "perfect match" occurs. Depending on the used equation for the calculation of the potential evapotranspiration (not mentioned so far) for example very sunny days with high radiation can occur simultaneously with rainfall the whole day. This bias has to be quantified by a sensitivity study, because it is not related to the SRM generation itself. A possible solution would be to use a weather generator (to have consistent weather data set as input for the r-r model) instead of only a rainfall generator (but of course, that introduces biases as well from the other climate time series), but this is maybe beyond the scope of the study. A discussion on the new introduced bias and its quantification would solve this issue.

Response 16:

This issue was already raised in the previous round of reviewer comments and has been addressed through additions in the discussion section. We have copied the text from the previous response below for convenience that provided additional explanation and outlines the changes made. Further, as the reviewer concedes we consider a more extended discussion to be out of scope for this study.

Excerpt from Response to Reviewer # 2 Comment 7, Revision Round 1:

> **[***'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).*

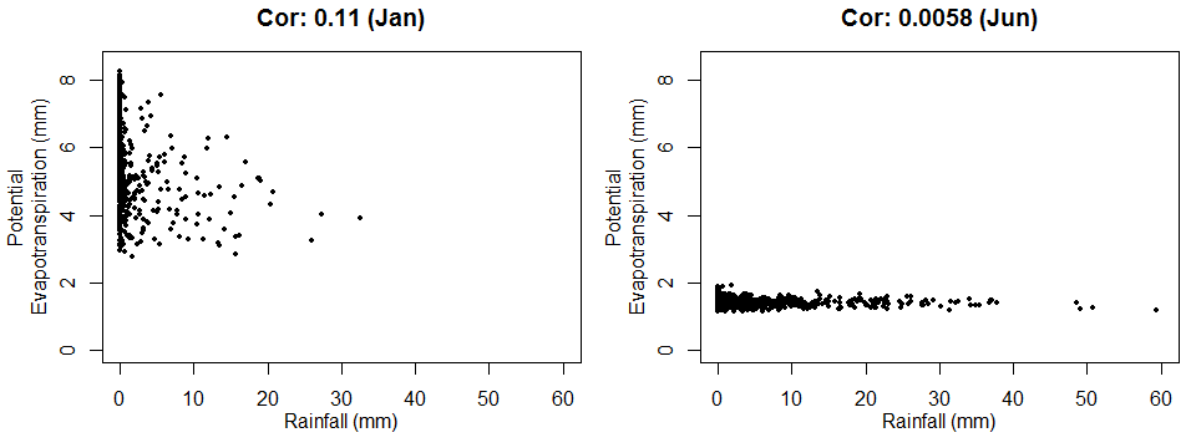


Figure 2- Rainfall-PET correlation (left) January and (right) June.

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."* **] end of excerpt.**

Comment 17:

P5l23-26 Also, one aim of the rainfall generation is to provide longer input time series than the existing observed ones. How can this be handled by the framework? For example, would a generated 600 year time series split into "30 realisations" because only 20 years of observations exist? How are completely unobserved catchments (no climate data) validated with this method. Some information on these issues should be provided in the manuscript.

Response 17:

In this paper 10,000 replicates of 73 year timeseries generated by an SRM are evaluated. In this case the generated SRM timeseries is of equal length to the observed rainfall time series. As the comparisons made between the virtual-observed (or observed) timeseries are performed through the comparison of summary statistics it is possible to compare generated timeseries that are longer than the virtual-observed (or observed) in terms of how well the reproduce the statistics of interest. The evaluation of a longer generated timeseries against a shorter observed period is also reported in observed-rainfall evaluation approaches (Rasmussen, 2013).

The virtual hydrological evaluation framework is based on a relative comparison between the virtual-observed streamflow and the simulated streamflow–as both simulated and observed rainfall undergo transformation by the same process representation (i.e. the hydrological model). Therefore, in instances where the streamflow of a catchment is unobserved the comparison between virtual-observed and simulated streamflow is still possible for the purpose of identifying deficiencies in the SRM. In such cases it is recommended that the hydrological model is calibrated such that the necessary catchment streamflow features can be simulated (e.g. via calibration to a neighbouring catchment with similar catchment properties).

Comment 18:

P6l7 Maybe the authors consider to replace "compare" by "combine"? If a comparison is the intention behind Fig. 2, I cannot understand how the comparison is carried out.

Response 18:

Done. The sentence has been rephrased as: '*The formal implementation of the virtual hydrological evaluation framework is summarised in Fig.2 … It combines both observed rainfall-evaluation and virtual hydrological evaluation*.'

Comment 19:

P6l14 The authors mention integrated tests here, but introduce them a few sections later. This is a bit confusing for the reader. Either the reader is referred to the subsection where the test is introduced or (what I would prefer) the test is introduced before mentioning it.

Response 19:

A reference to the subsequent subsection where the test is introduced has been inserted (as this section simply provides a brief overview of Step 1).

> Page 6, Line 20: '*Following the selection of a primary streamflow characteristic of interest and a suitable hydrological model, integrated tests are conducted for each rainfall site (described below in Section 2.2.2).*'

Comment 20:

P6l17 & Fig. 2 The CASE framework was mentioned only in the introduction (p4l28), the provided reference is not included in the reference list. (How) Does the CASE framework differ from the framework applied in this study? If the CASE framework is important for the reader to follow the investigation, it should be (briefly) explained in the current manuscript (the reader should not be forced to read other manuscripts to be able to follow the current investigation).

Response 20:

Please also see the response to comment 15. At this point in the manuscript (introductory text setting out the elements of Step 1) the important factor is that the systematic application of quantitative performance criteria is used to identify poor performing sites. The text has been revised to read:

> Page 6, Line 20: '*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 suitable hydrological model, integrated tests are conducted for each rainfall site (described below in Section 2.2.2). The results of the integrated tests are then used to identify sites that are poor performing, according to the systematic application of quantitative performance criteria (see Section 2.2.3), for the primary streamflow characteristic*.'

We do not believe the reference to Bennett et al. (2018) was missing. The reference appears on Page 29, Line 27 of the manuscript and is also reproduced below:

> 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.

Comment 21:

P6l22 Is it also possible to take into account more than one primary streamflow characteristic?

Response 21:

Yes, where a combination of streamflow features are important the 'primary streamflow characteristic' may take into account multiple factors. The purpose of the primary streamflow characteristic is to focus the investigation of SRM deficiencies on to those that impact on streamflow features that are important for the studied catchment.

Comment 22:

P7l16 Again, the CASE framework should be explained before.

Response 22:

Thank you. The CASE framework is now described earlier in the manuscript. Please also see responses to comments 15, 20 and 45.

Comment 23:

Fig. 3 The location of the elements case (i-iii) is hard to identify since it remains unclear, which element (the mean or the range) determines the position on the y-axis. I suggest to put both elements on the same level in direction of y-axis instead of having two different elements. Also, does the range result from different months or from the 10,000 realisations of the SRM or from both? Would it be worth to distinguish between both ranges?

Response 23: We have amended the Figure so that the mean and range are aligned. Thank you for pointing this out. The mean (indicated by the '+' symbol) should be on the same level as the y-axis (it was originally separated out so that they element could be seen better, but we agree this is confusing).

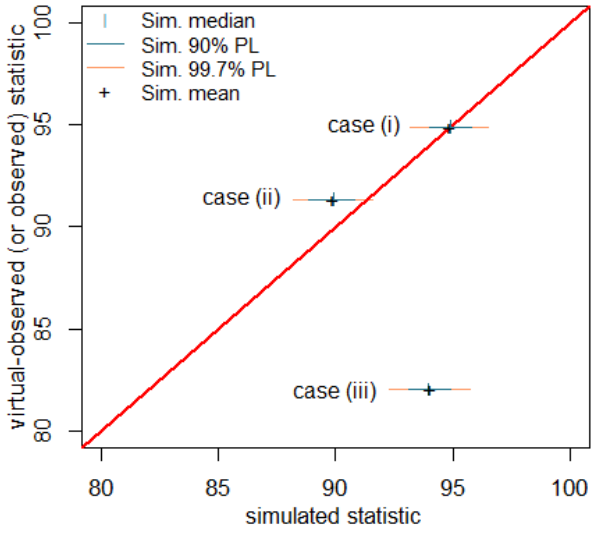


Figure 3 - Illustration of performance classification, case (i) shows 'good' performance, case (ii) shows 'fair' performance and case (iii) shows 'poor' performance. Adapted from Bennett et al. (2018).

Yes, the range is the result of the 10,000 realisations of the SRM. As the figure is statistic based, the statistics for the individual months can be separated out. For example, the performance of the SRM's simulation of mean flow for three separate months would look like Figure 3 where 3 ranges (made up of 10,000) are shown.

Comment 24:

P7l20-22 How have the threshold of 90 % and 99.7 % been determined? Especially the latter one seems to have a certain origin and was not been chosen arbitrarily…The threshold should depend on the criterion used for the validation, right? So the threshold would be different for a derived flood with a 100 year return period in comparison to the mean discharge or some drought indices.

Response 24:

In this manuscript the thresholds of 90% and 99.7% applied are taken from Bennett et al. (2018) with the 99.7% threshold being used to apply a threshold that is 3 standard deviations from the mean.

Yes, the thresholds should depend on the criterion used for the validation and can be adjusted as needed to suit the criterion. For example, a relative difference metric that provides an indication of the practical tolerance for the particular streamflow or rainfall feature could be adopted instead of a metric that is a function of the simulated range of statistics.

Comment 25:

P8l5-17 How is snowfall handled in this context? If snow falls and accumulates over the winter period, it depends on the temperature time series when it starts to melt and to contribute to the total runoff. Again, in this context, how is the bias quantified? I can imagine that the differences are quite high depending on the interplay between rainfall and temperature.

For example, if the SRM generates "precipitation" and due to the temperature time series it falls as snow and accumulates over days/weeks until it contributes to runoff, it causes a high difference to the observed time series, which had no precipitation in that cold period. This is from my understanding not covered by the virtual hydrological framework, and has to be quantified as the bias mentioned before.

Response 25:

The case study catchment (Onkaparinga catchment, South Australia) has a Mediterranean climate and an ephemeral streamflow response. It does not experience snowfall or any build-up of snow over the winter period. This a common catchment type for Australia (and in similar climates internationally) and represents different challenges to ones where there is snowfall. The reviewer raises a valid point regarding the impact of snow, but this is not applicable to this catchment. We look forward to addressing this issue in future research when we apply this framework to a catchment with snow.

Comment 26:

P12l19-29 The description of the rainfall generator is not sufficient to understand it. For further details the reader is referred to a reference which is not provided in the reference list. This section has to be clear to the reader, since it is an essential basic for the evaluation of the introduced framework.

Response 26:

Please see the two responses from the first round of reviewer comments below in which additional details on the rainfall generator have already been provided. We also do not believe that the reference was missing.

The focus of this manuscript is the virtual hydrological evaluation framework, which is independent of the rainfall generator tested (i.e. essentially any rainfall generator could be tested in its place).

Note that the rainfall generator is clearly explained in Bennett et al. (2018) which was provided in the reference list and applied to the same catchment. We do not think it is necessary to add extra length (to an already long paper) explaining a rainfall generator that is not part of the key innovations and is explained in an easily accessible paper.

Excerpt from Response to Reviewer #3 Comment 5, Revision Round 1:

> **['***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."] **End of excerpt.**

Excerpt from Response to Reviewer #3 Comment 4, Revision Round 1:

['*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. 5). 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..."' **] End of excerpt.**

Comment 27:

P12l30-31 Regarding the parameter estimation: Parameters are estimated from 1914-1986 and have almost no overlay with the calibration period of the hydrological model. Which time series period of the other climate variables is used for the simulations, the period of the calibration period of the r-r model or the period which was also used for the estimation of the SRM parameters?

Response 27:

Yes, the SRM model parameters are estimated on the 1914-1986 which does not have a large overlay with the hydrological model calibration period and associated climate variable time series (in this case PET). While these differences would present an impediment for observed streamflow evaluation they are not critical for the use of a virtual hydrological evaluation which uses a relative comparison of simulated and virtual-observed streamflow to interrogate the simulated rainfall. In a virtual hydrological evaluation the same hydrological model parameters are used in the production of the virtual-observed streamflow and simulated streamflow. Likewise, the same related climate variables (in this case PET) are used in all simulations. Therefore, the relative comparison between the virtual-observed and simulated streamflow are not impacted by the different calibration period.

Comment 28:

P13l2-4 Was the calibration carried out for the same period and with the same station density as in this study? If not, does it has an impact on the simulated runoff (e.g. if one station was missing in the calibration period, which has a high influence (weight) on the areal rainfall, but the rainfall generation for that station is different due to its altitude or snow/rainfall occurrences)?

Response 28:

Similarly to the response to comment 27, while these differences are not as important when using a virtual hydrological evaluation approach which uses a relative comparison of the simulated and virtual-observed streamflow to interrogate the simulated rainfall, as the same hydrological model parameters are used in the production of both the virtual-observed streamflow and simulated streamflow.

In terms of the issue surrounding station density, in this paper we have not evaluated how simulated catchment average rainfall performs 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.

Comment 29:

Fig. 5: 7 out of 22 rain gauges have no influence on areal rainfall using Thiessen polygons: station 21, 13, 19, 8, 2, 10, 8. In total, 3 out of them are explored in detail due to "the relatively poorer ability of simulated rainfall to reproduce annual streamflow totals at these sites". How can these points have a negative impact on the simulated runoff if they do not contribute at all to the areal rainfall (if Thiessen polygons are applied)? Or did I miss a certain point in the description of the areal rainfall determination? Is the areal rainfall for the framework analysis estimated only by one (or less than all 22 stations)?

Response 29:

Thiessen polygons are not applied for the evaluation represented in Fig. 5, which summarises the outcome of Step 1 of the virtual hydrological framework procedure. In Step 1 of the proposed framework, rainfall from each site is applied independently and repeated for each site, so that any discrepancy between virtual-observed streamflow and simulated streamflow can be attributed to a deficiency in rainfall parameters at that site. Also see the response to comment 34.

Comment 30:

P13l5-6 To which streamflow data were the model parameters calibrated? Mean flow (on a seasonal basis), annual extreme flows or something different?

Response 30:

The rainfall-runoff model was calibration to daily streamflow time series. The text has been amended as follows:

> Page 13, Line 4: '*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 hydrological model was calibrated to the daily streamflow data at Houlgrave Weir (see Fig. 5) using model calibration period of 15 years (1985-1999). The model parameters were estimated using maximum likelihood estimation procedure with a weighted least squares likelihood function. The set of hydrological 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 able to simulate streamflow with a good fit to the observed daily streamflow, with a Nash-Sutcliffe efficiency of 0.8*.'

Comment 31:

P13l11 "The same set" – For all further analysis all 100 parameter sets are kept or only the best one?

Response 31:

The hydrological model parameters (i.e. the set of parameters) that maximised the likelihood function were used for both the unit and integrated test. This has been clarified in the manuscript:

> Page 13, Line 8: '*The model parameters were estimated using maximum likelihood estimation procedure with a weighted least squares likelihood function. The set of hydrological model parameters that maximised the likelihood function were found using a multi-start quasi-Newton optimisation procedure with 100 random starts ... 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 32:

P13l Providing a NSC-value is only useful, if the authors state for which variable it was calculated. I assume for simulated daily discharge values?

Response 32:

Yes, the NSE value was calculated based on comparing observed and simulated daily streamflow. This detail has been added to the paper.

> Page 13, Line 11: '*Overall, the GR4J model was able to simulate streamflow with a good fit to the observed daily streamflow, with a Nash-Sutcliffe efficiency of 0.8*.'

Comment 33:

Table 3 The abbreviations used in the table header include event-based characteristics. How do the authors define an event (when does it start and end)? This information is important for the reader to understand the characteristics.

Response 33:

A note has been added to Table 3 to indicate that a wet day is defined as a day with rainfall above 0.1 mm such that the number of days in a row above this threshold are deemed a wet-spell (and vice versa for dry-spells). The note is produced below:

> Page 14, Bottom of Table 3: '*Note: Wet days are defined as days where the rainfall exceeded a 0.1 mm threshold with wet-spells defined as the number of days in a row above the threshold (and vice versa for dry-spells).*'

Comment 34:

P15l6 I can't follow how 10 out of 22 rain gauges are categorized as "poor", if the areal mean is used as input for r-r modelling. Or are the r-r models driven by single-station input only?

Response 34:

Yes, the hydrological model is driven by single-station input only. Also see response to comment 29.

Comment 35:

P15l9-10 I would prefer to see the results for all 22 sites to test the framework more in detail.

Response 35:

The full rainfall-based evaluation is provided in Bennett et al. (2018). The focus of the virtual hydrological evaluation framework (and in particular Steps 2-3) is to identify sources of error in streamflow-based evaluation. Therefore it logically follows that we would choose to apply this to sites where there are streamflow errors. This is why the analysis focuses on the 10 sites with poor streamflow performance. If we had undertaken the subsequent analysis on the other 12 sites, it is unlikely to result in additional knowledge or the identification of actionable changes to the SRM because it performs well in terms of both rainfall and resultant streamflow. This analysis would have taken additional space in an already long manuscript (13 Figures, 3 Tables, ~10,300 words).

Comment 36:

P15l19-27 I struggle with the description of Fig. 7. The first column includes simulated rainfall statistics: mean of daily rainfall amounts and mean number of wet days, while in the second one the standard deviation for both is shown. How is in the first column for one sites decided, if it is a good, fair or poor quality, when two criteria are taken into account? Do both criteria for one site have to be "good", to result in a final "good?" Or is the mean of both relative errors chosen for the final decision? A short explanation would be helpful.

Response 36:

In Figure 7 each quantitative criteria is applied separately such that a site may be deemed 'good' in terms of the simulation of daily rainfall means but 'poor' in terms of the standard deviation. However, in the first column it shows the performance for the sites for three separate statistics (mean daily rainfall amounts, the standard deviation of daily rainfall amounts and the mean number of wet day) as all of these statistics are classified as 'good' for 100% of sites for all the months (and at the annual level). This was done to avoid the repetition of three columns of 100% 'good' bar-charts.

Descriptive text explaining this has been added to the manuscript to avoid confusion:

> Page 15, Line 19: '*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. Each of the three statistics presented in the first column are assessed separately but are presented together to avoid repetition*.'

Comment 37:

P16l1-3 The small deviation from the observed runoff in January result from the very low runoff generated in January (see Fig. 6), that the r-r model is trained to simulate-the rainfall has no effect in that months. It would be useful if the authors provide an example from the wettest periods (as done later).

Response 37:

An example of the type requested (i.e. discussion of wetter months) appears in the subsequent paragraph.

Comment 38:

Fig. 7 caption: Since the rainfall and runoff characteristics are mentioned before (p15l15-18), they can be removed from the caption.

Response 38:

Fixed. The previously described rainfall and runoff characteristics have been removed from the caption.

Comment 39:

Fig. 8 Why is for the standard deviation (b) and d)) no range? There should be a range from the number of months used for the simulation (15 years) and from the 100 parameter sets mentioned before.

Comment 40:

Fig. 9 Again, shouldn't there be a range for the virtual observations as well?

Responses 39–40:

As mentioned in the response to comment 31, only the 'best' set of hydrological model parameters (i.e. those that maximise the likelihood function) are used in the simulation of virtual-observed and simulated streamflow.

The red crosses in Figure 8 (b) and (d) represent the standard deviations of the observed rainfall and the virtual-observed flow for each month calculated from the 73 years. The standard deviation represents the variability. They are single numbers. If we were presenting all the monthly values, then they would be a range, but we are presenting statistics, the mean (left-hand side) and standard deviation (right-hand side) for each month of the

year. The boxplots show the range of these monthly statistics for the 10,000 stochastic rainfall model replicates. This is a standard way of evaluating stochastic models (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.).

This issue regarding the range has also been dealt with in the first round of revisions. See text below.

Excerpt from Response to Reviewer #2 Comment 6, Revision Round 1:

> ***[****'… 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.)*'**] End of excerpt.**

Comment 41:

P22l20 Do the 10 % refer to the 10 bad out of 22 stations or to all 22 stations?

Response 41:

Thank you for pointing this out. It refers to the set of 'poor' sites/months (identified in Step 2) in terms of the attributed cause. Clarification has been added.

> Page 22, Line 19: '*The category where streamflow errors originate from rainfall model deficiencies in a preceding month represents about 10% of the evaluated site/month combinations (i.e. those identified in Step 2).*'

Comment 42:

P26l11 I'm not familiar with the reference of Ang & Tang (2007) and the information provided in the reference list are insufficient to find it (is it a book? v1 sounds like a handbook or a model description). However, the r-r transformation is strongly non-linear, hence an error propagation as mentioned in the text should not be carried out (even not as an example).

Response 42:

Yes, Ang & Tang (2007) is a textbook commonly used as reference text to teach statistics to engineering students. The reference list has been updated to make this book easier to identify and find. The reference supplied is as follows:

> ANG, A. & TANG, W. 2007. Probability Concepts in Engineering: Emphasis on Applications to Civil and Environmental Engineering (2nd Edition), Hoboken, New Jersey, USA, John Wiley & Sons Inc., hardback ISBN-10 0-471-72064-X.

While we agree with the reviewer, that the rainfall-runoff transformation is non-linear, this does not invalidate this discussion point. The example provided is used to demonstrate the potential amplification of the errors and is likely to be a lower bound, because it assumes linearity in the transformation to perform the error propagation. The revised text now highlights this point. This discussion point simply serves as an illustration of the advantages of the using streamflow-based evaluation approaches. The analysis of error propagation is not a key point of the paper.

> Page 26, Line 9: '*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 (Chiew, 2006), using the principles of error propagation (Ang and Tang, 2007), assuming linearity it follows that a 10% error in mean/standard deviation of rainfall could potentially be amplified to 20-35% error in the mean/standard of streamflow. This estimate represents a lower-bound of the potential amplification,*

since the non-linear nature of the rainfall-runoff transformation will likely produce a larger potential amplification of errors. 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.'

Comment 43:

P28l21-22 This sentence sounds as if it was possible to determine, which rainfall characteristics are most important for the r-r modelling. Indeed, this was not done in the study. A few rainfall characteristics have been analysed, but there was no detailed and quantitative investigation which one has higher impacts on the simulated runoff.

Response 43:

Fixed. The sentence has been reworded to remove the implication that it determines which rainfall characteristics should be modified.

Page 28, Line 24: '*This paper has introduced a virtual hydrologic evaluation framework that enables targeted hydrological evaluation of SRMs*.'

Comment 44:

P29l1 It would be useful if the authors can provide references for conventional approaches.

Response 44:

We agree, conventional approaches were not well defined and have removed the term. The wording of this part of the conclusion, has been modified to be more direct and consistent with the discussion in Section 5.2.

Page 29, Line 2: '*The unit test identified the importance of the simulated rainfall in the transition months of May and June (late autumn/early winter) during the 'wetting-up' phase of the catchment cycle for producing low errors in subsequent high streamflow months (July/August/September) and the annual streamflow distribution. The virtual hydrologic evaluation framework provides valuable additional diagnostic ability for the development and application of SRMs, not available by using rainfall-based evaluation techniques alone.'*

Technical corrections:

Comment 45:

P2l9 The reference of Bennett et al. (2018) is missing and could not be proven regarding its content and relevance for the current manuscript.

Response 45:

We do not believe that the reference was missing, it could be found on page 40, line 1 of the marked-up version of the manuscript provided as well as page 29, line 24 of the un-annotated version as part of the first round of manuscript revisions. Furthermore, this reference was a part of the original submission to HESS and has been consistently included in all revisions.

The reference Bennett et al. (2018) is provided on Page 29, Line 27 of the revised manuscript and is also reproduced below for convenience:

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.

Comment 46:

Fig. 9, 10, 13 "cummulative" -> "cumulative"

Response 46:

Fixed. The x-axis labels in panel (c) of Figures 9, 10, 12 and 13 have been amended.

Comment 47:

P13l6 Fig. 5 instead of Fig. 4? But I can't see the Houlgrave weir in Fig. 5 anyway.

Response 47:

The figure cross-reference has been fixed. Figure 5 has been amended so that Houlgrave Weir is explicitly labelled to make it easier to locate on the map. The streamflow gauge position is indicated by the purple square. The figure is reproduced below:

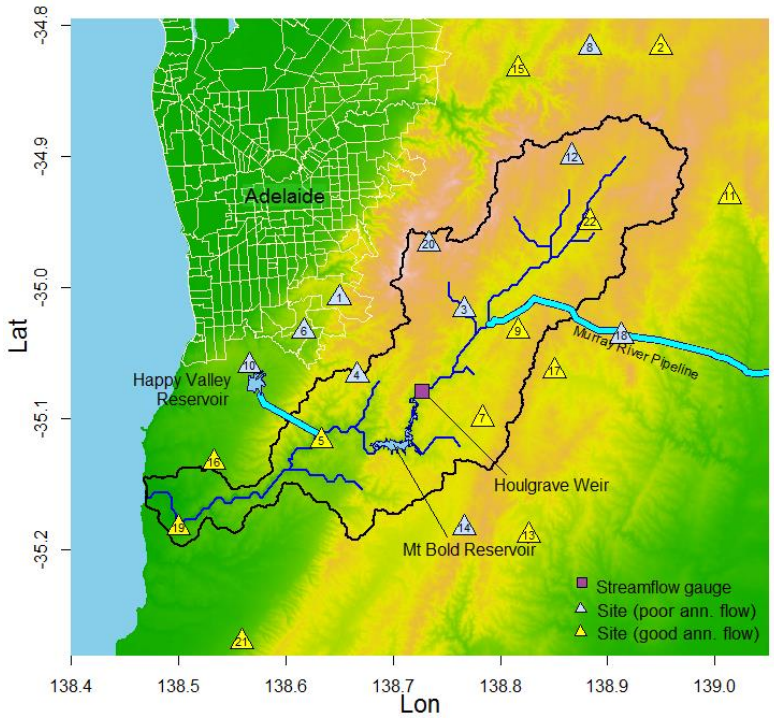


Figure 4 - Onkaparinga catchment, South Australia. Sites indicated by blue triangles are explored in greater detail in this paper due to the relatively poorer ability of simulated rainfall to reproduce annual streamflow totals at these sites.

Comment 48:

P24l6 "sits" is an inappropriate verb in this context

Response 48:

The verb 'sits' has been removed and replaced with 'is located'.

Comment 49:

P27l8 "identifying" -> "identifies"

Response 49:

Fixed.

Comment 50:

P27l13 Remove the comma.

Response 50:

Fixed.

Comment 51:

P27l25 "a hydrological model" -> "a single hydrological model" Maybe this is what the authors want to say?

Response 51:

Done – 'a single hydrological model' has been inserted in place of 'a hydrological model'.

Comment 52:

Comments from the first review:

Reply to comment 2: "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"

From the manuscript it remains unclear what the CASE framework is exactly (is it the applied framework illustrated in Fig. 2?), the reference for explanations (Bennett et al., 2018) is missing.

„(iii) systematically categorise aggregate performance over multiple spatial and/or temporal scales"

-> Only one spatial and temporal scale are analyzed.

Response 52:

Clarification surrounding the CASE framework as it pertains to the virtual hydrological evaluation framework has been added to the manuscript (see Sections 1 and 2.2; also see the responses to comments 15, 20, 22 and 45) as well as clarification around the spatial and temporal scales addressed (see response to comment 13).

References

ANG, A. & TANG, W. 2007. *Probability Concepts in Engineering: Emphasis on Applications to Civil and Environmental Engineering (2nd Edition),* Hoboken, New Jersey, USA, John Wiley & Sons Inc, hardback ISBN-10 0-471-72064-X.

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

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.

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

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.

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.

LE COZ, J., RENARD, B., BONNIFAIT, L., BRANGER, F. & LE BOURSICAUD, R. 2014. Combining hydraulic knowledge and uncertain gaugings in the estimation of hydrometric rating curves: A Bayesian approach. *Journal of Hydrology,* 509**,** 573-587.

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.

LINSLEY, R., KOHLER, M. A. & PAULHUS, J. L. 1982. Hydrology for Engineers. MacGraw-Hill. *Inc., New York, USA*.

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.

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

RENARD, B., KAVETSKI, D., KUCZERA, G., THYER, M. & FRANKS, S. W. 2010. Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resour. Res.,* 46**,** W05521.

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.

THYER, M., RENARD, B., KAVETSKI, D., KUCZERA, G., FRANKS, S. W. & SRIKANTHAN, S. 2009. Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis. *Water Resources Research,* 45.

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.

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.

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 has been given to the development of stochastic rainfall models (SRMs), significantly less attention has been paid to ~~performance evaluation methods~~developing methods to evaluate their performance. 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 high~~er~~ but streamflow low~~er~~) 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.

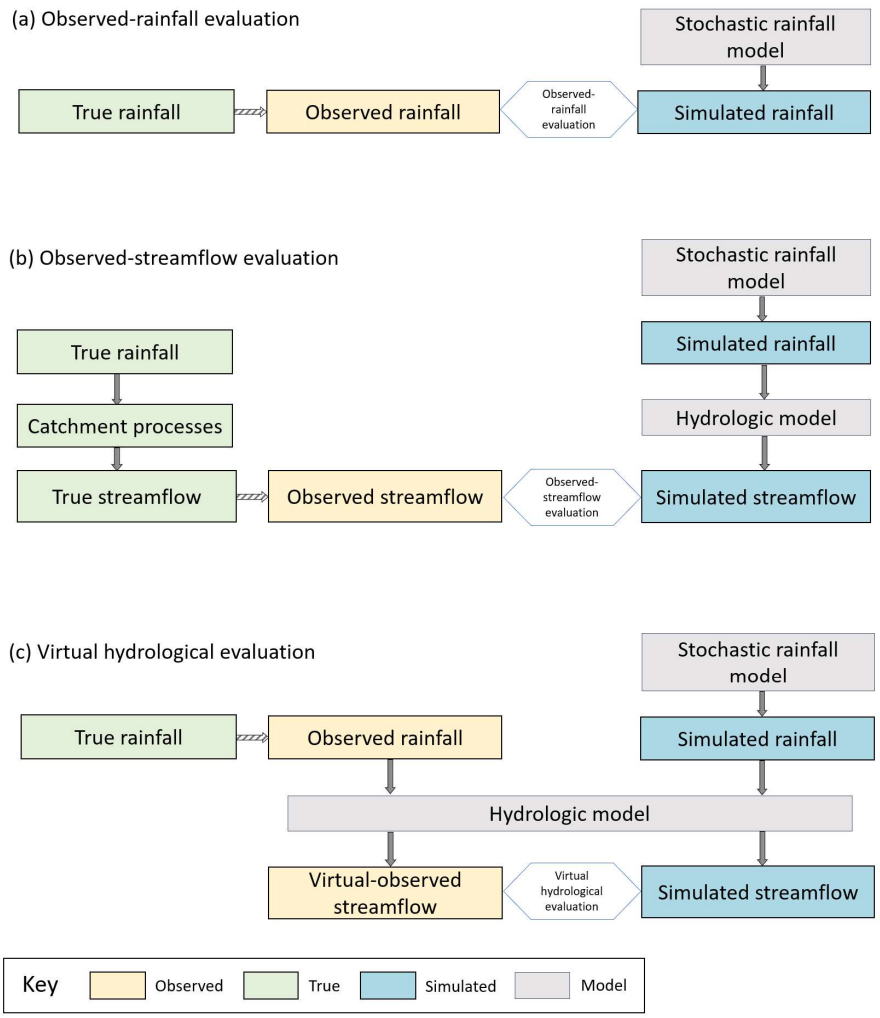
1 Introduction

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 floods (e.g. Camici et al., 2011, Li et al., 2016) or droughts (e.g. Henley et al., 2013, Paton et al., 2013, Mortazavi-Naeini et al., 2015). When evaluating the efficacy of SRM's, current approaches that make comparisons to observed rainfall or streamflow have~~are~~ limited diagnostic

ability. They are unable to make a targeted evaluation of the SRM's ability to reproduce streamflow characteristics of practical interest. This paper introduces a new virtual framework that enables targeted hydrological evaluation of SRMs.

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).

To overcome limitations in observed-rainfall evaluation methods, 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) and is referred to as 'observed-streamflow evaluation'. From Fig. 1(b), observed-streamflow evaluation typically involves (1) a 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) which makes it difficult to evaluate the performance of the SRM and identify opportunities for improvement.


(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

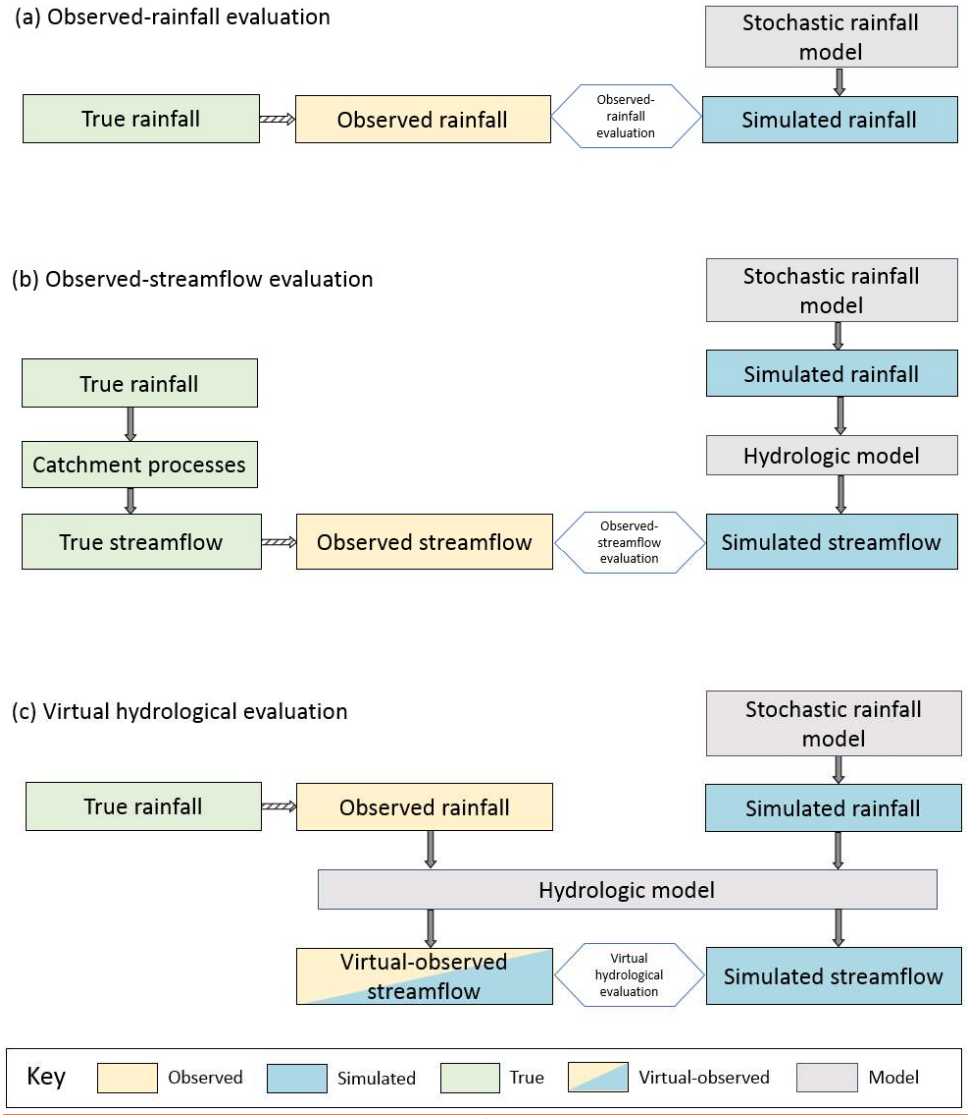

(a) Observed-rainfall evaluation
Stochastic rainfall model
True rainfall
Observed rainfall
Observed-rainfall evaluation
Simulated rainfall
(b) Observed-streamflow evaluation
Stochastic rainfall model
True rainfall
Simulated rainfall
Catchment processes
Hydrologic model
True streamflow
Observed streamflow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

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

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			
	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

To date, 'virtual experiments', that is, experiments that focus on comparisons between streamflow simulated under different conditions or inputs (i.e. virtual streamflow) without relying on comparisons to observed streamflow, have been used in a variety of contexts. Examples include, including (i) the evaluation of hydrological model sensitivity (e.g. 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 (e.g. Sikorska et al., 2018), (ii) for the development ofing new techniques for flood frequency analysis (e.g. Li et al., 2014, 2016), and (iii) the calibration, validation and selection ofto support SRMs selection

(e.g. Müller and Haberlandt, 2018, Kim and Olivera, 2011) ~~as well as calibration and validation (e.g. Kim and Olivera, 2011) ,through a comparison to virtual streamflow~~.

The framework presented in this paper is a significant advance from previously reported virtual experiments because it presents a formal framework to identify key deficiencies in the SRM by ~~utilising~~ (1) Aextending thea comprehensive and systematic evaluation (CASE) framework (developed by Bennett et al., 2018 for observed-rainfall evaluation 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 for use in virtual hydrological evaluations, and (2) utilising 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: the new framework is a stepwise procedure that enables the identification poor performing sites, then poor performing time periods and then the key deficiencies in the SRM for those sites and time periods by drawing on the systematic application of quantitative performance criteria.
2. To present two different tests which are part of the framework: the integrated test ~~as well as introduce~~and a new type of 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).

Formatted: List Paragraph, Numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0.91 cm + Indent at: 1.55 cm

2 Virtual hydrological evaluation framework

2.1 Overview

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 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 for the simulation of both virtual-observed and simulated streamflow. It is also important that the selected hydrological model is fit for purpose so that it can simulate the streamflow characteristics of interest.

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 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 SRM. This is because the ultimate goal of the SRM modelling process remains the same: to match observed streamflow for a catchment of interest.

The formal implementation of the virtual hydrological evaluation framework is summarised in Fig. 2. 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~~ combines both observed rainfall-evaluation and virtual hydrological evaluation. The virtual hydrological 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.

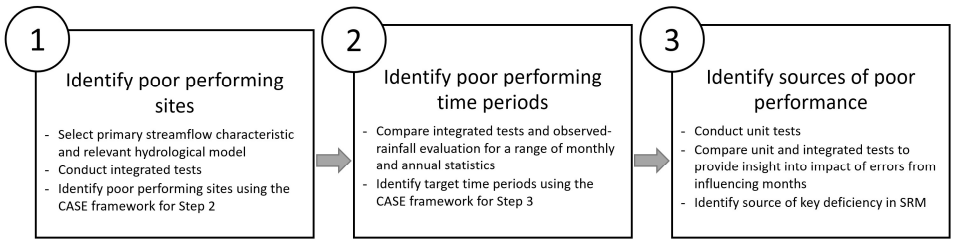


Fig. 2 Virtual hydrological evaluation procedure.

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~~suitable hydrological model, integrated tests are conducted for each rainfall site (described below in Section 2.2.2). The results of the integrated tests are then used to identify

sites that are poor performing, according to the systematic application of quantitative performance criteria (see Section 2.2.3)~~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 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 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 integrated test proceeds for a single site by transforming 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 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).

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

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

$$Q^{sim} = g[R^{sim}, \dots] \quad (2)$$

If there is 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 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, 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)	

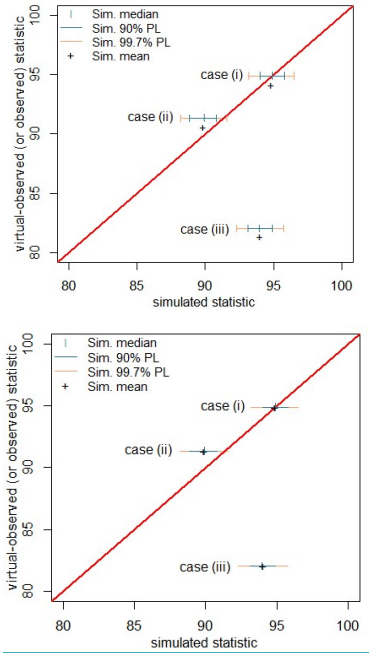


Fig. 3 Illustration of performance classification, case (i) shows 'good' performance, case (ii) shows 'fair' performance and case (iii) shows 'poor' performance. Adapted from Bennett et al. (2018).

2.3 Step 2 – Identify poor performing time periods

The second framework step is to 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.

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 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 perspectives 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 $m \in \{1, \dots, 12 | \, m \neq k\}$. 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.

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

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:

$$R_{(6)}^{spl} = \left\{R_1^{obs}, \dots, R_5^{obs}, R_6^{sim}, R_7^{obs}, \dots, R_{12}^{obs}\right\} \quad (4)$$

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.

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

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.

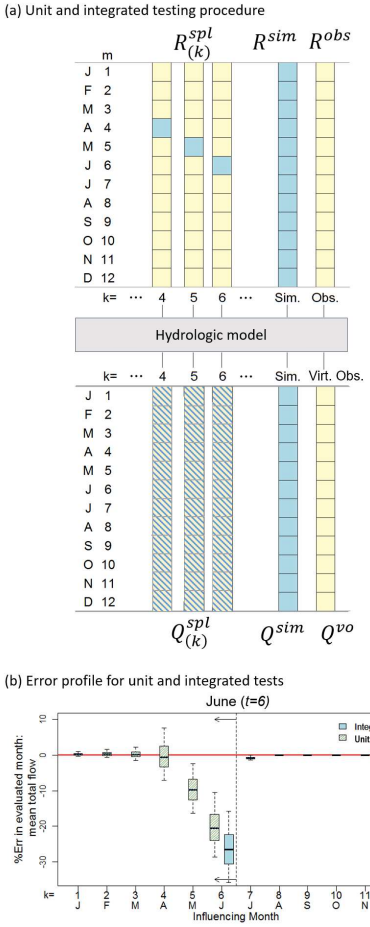


Fig. 4 Schematic of (a) the method of constructing a unit test by embedding simulated months in an observation time series, and (b) the error profile produced when using the integrated and unit tests for the evaluated time period of June (*t*=6) (box plot whiskers indicate the 90% limits of the simulated 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 $Q^{vo}_{(t)}$.

The full set of spliced rainfall (e.g. spliced rainfall for each month designated as the influencing month $R^{spl}_{(k)}; k = 1, \dots, 12$) 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 Compare unit tests and integrated tests

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. 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.

A relative error metric, $\%Err$, is used to enable a comparison between the virtual-observed streamflow and the evaluated streamflow replicates from the integrated or unit tests (Q^{sim} or $Q^{spl}_{(k)}$) 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

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

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 ($k = 1, \ldots, 12$) 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(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 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 between cases allows for SRM improvements to be targeted in terms of their ultimate impact on streamflow statistics. 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 1.

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 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.

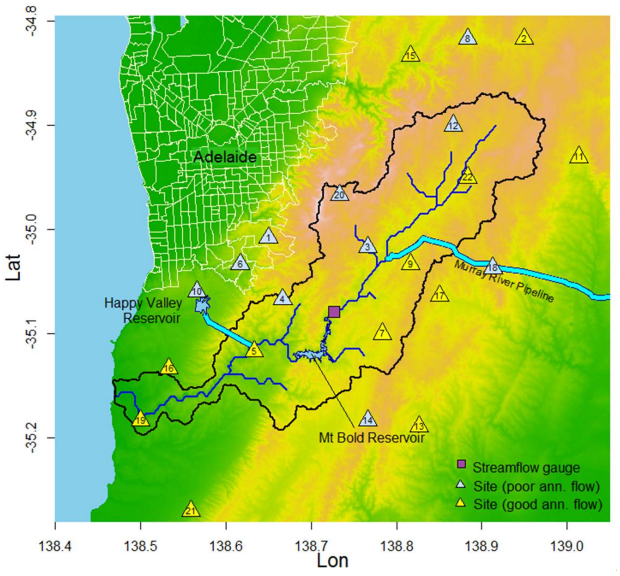
There is a strong rainfall gradient (Table 3), with average annual rainfall ranging from approximately 500 mm on the coast (Site No. 19) to over 1000 mm in the region of highest elevations (Site No. 20). A breakdown of the rainfall characteristics (annual total, number of wet days, daily average amounts, 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 hydrological model was calibrated to the daily streamflow data at Houlgrave Weir (see ~~Fig.~~ 5~~Figure 4~~) using model calibration period of 15 years (1985-1999). The model parameters were estimated using maximum likelihood estimation procedure with a weighted least squares likelihood function. The set of hydrological 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 able to simulate streamflow with a good fit to the observed daily 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.


Adelaide
Happy Valley Reservoir
Mt Bold Reservoir
Murray River Pipeline
Streamflow gauge
Site (poor ann. flow)
Site (good ann. flow)
Lat
Lon
-34.8
-34.9
-35.0
-35.1
-35.2
138.4
138.5
138.6
138.7
138.8
138.9
139.0

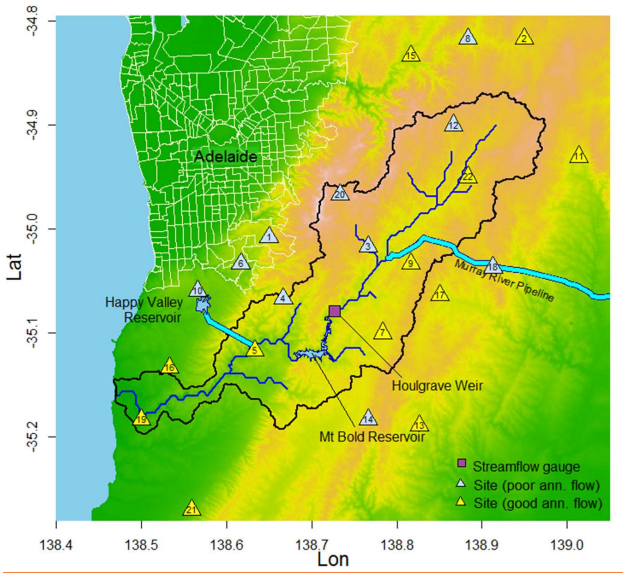


Fig. 5 Onkaparinga catchment, South Australia. Sites indicated by blue triangles are explored in greater detail in this paper due to the relatively poorer ability of simulated rainfall to reproduce annual streamflow totals at these sites.

Table 3 Site names, 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

Note: Wet days are defined as days where the rainfall exceeded a 0.1 mm threshold with wet-spells defined as the number of days in a row above the threshold (and vice versa for dry-spells).

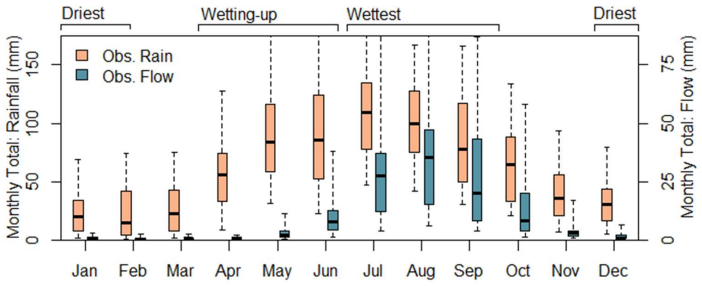


Fig. 6 Seasonal variation of catchment average rainfall and flow at Houlgrave Weir. Boxplots show the variation across years. Wettest indicates (high rainfall, high flow), direst indicates (low rainfall, low flow) and wetting-up indicates (high rainfall, low flow).

Formatted Table

Formatted: Left

4 Results

4.1 Step 1 – Identify poor performing sites

To undertake Step1, annual total flow volumes were designated as the primary streamflow characteristic to narrow the number of sites investigated. Following the selection of the primary streamflow characteristic and selection of the hydrological model, 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. This step categorised 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).

The 10 sites categorised 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 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 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
- 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. Each of the three statistics presented in the first column are assessed separately but are presented together to avoid repetition. 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.

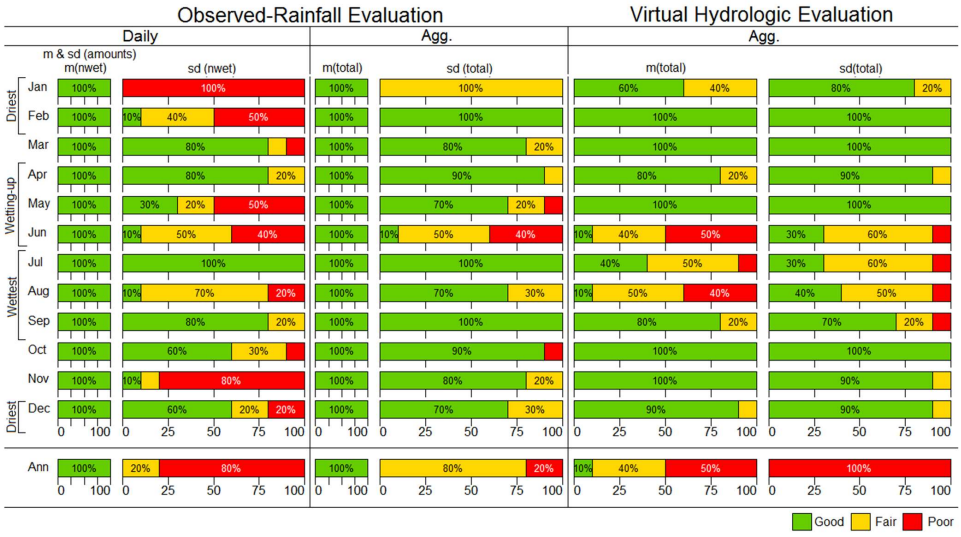


Fig. 7 Integrated test, comparing observed-rainfall evaluation (left) with the virtual hydrologic evaluation (right). Comparison of daily and aggregate ('Agg.') rainfall statistics against aggregate flow statistics for individual months and years. ~~Daily rainfall statistics presented include the means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

4.3 Step 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 against integrated tests in terms of their relative errors. From this comparison the source and type of key deficiencies~~y~~ 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 using two locations, Site 12 and Site 10 (see Fig. 8 to Fig. 13). For completeness these results are presented together with the results of the observed-rainfall evaluation (panels (a) and (b) of Fig. 8 and Fig. 11).

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

A common 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.

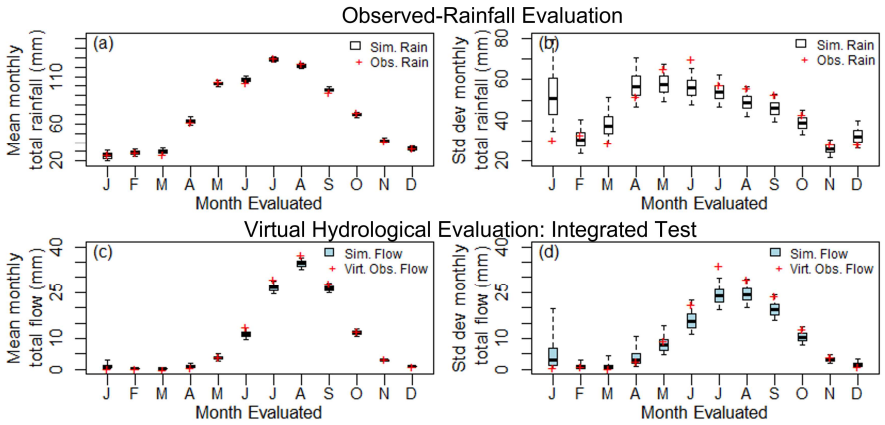
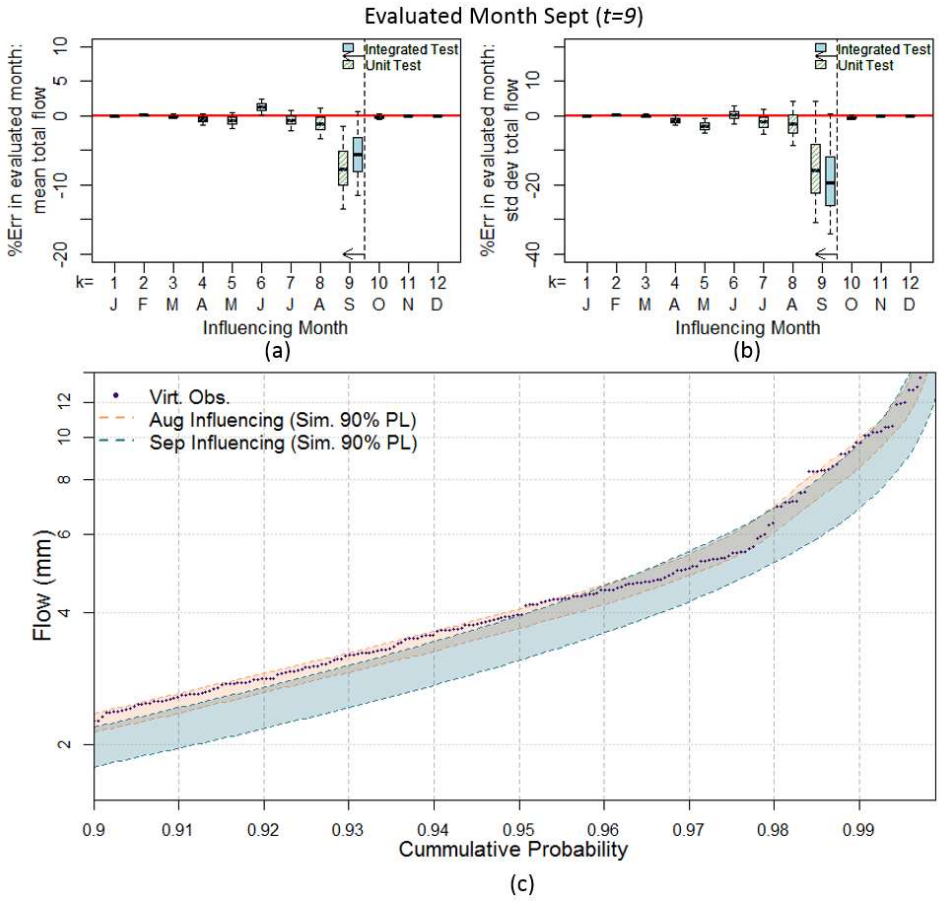


Fig. 8 Lobethal, Site 12 (a) observed-rainfall evaluation mean monthly total rainfall, (b) observed-rainfall evaluation standard deviation of monthly total rainfall, (c) virtual hydrological evaluation (integrated test) mean monthly total streamflow, (d) virtual hydrological evaluation (integrated test) 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 (a) and (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 (a)) and standard deviation of monthly total flow (Fig. 9 (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 (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~~ is located 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.


Evaluated Month Sept (t=9)
%Err in evaluated month: mean total flow
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
(a)
%Err in evaluated month: std dev total flow
Integrated Test
Unit Test
Influencing Month
(b)
Virt. Obs.
Aug Influencing (Sim. 90% PL)
Sep Influencing (Sim. 90% PL)
Flow (mm)
Cummulative Probability
(c)

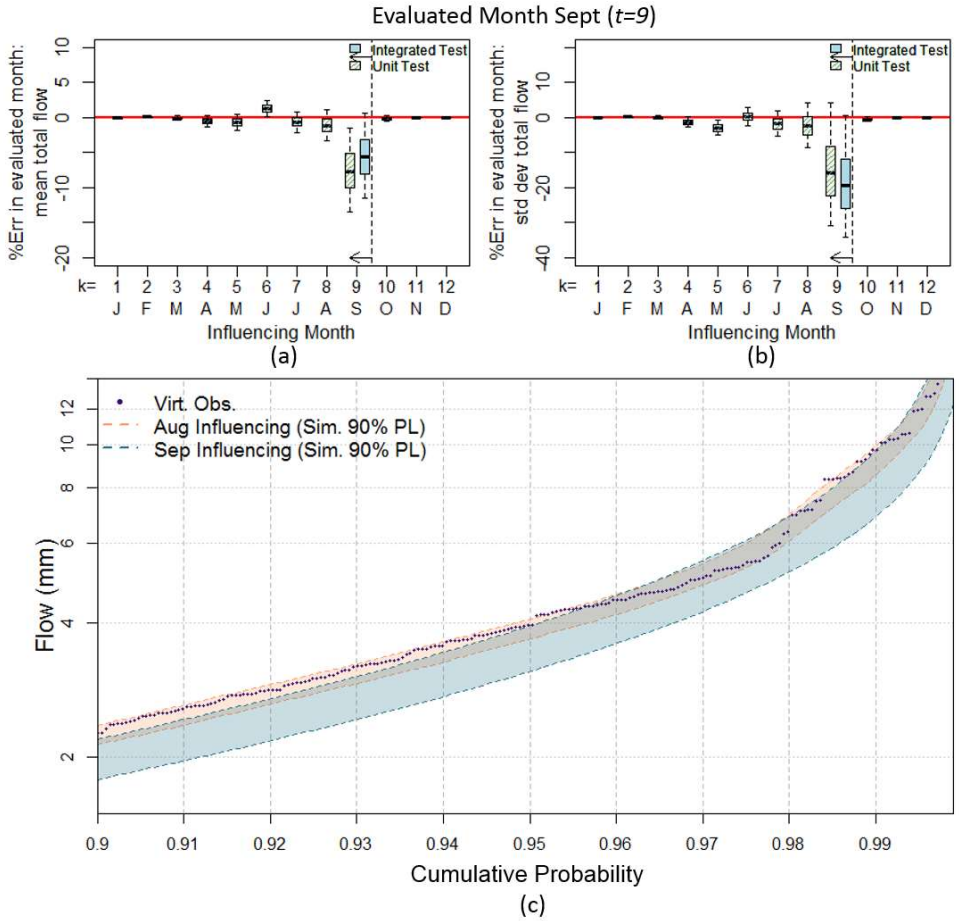


Fig. 9 Lobethal, Site 12 (a) unit test error in mean monthly flow (September), (b) unit test error in standard deviation of monthly flow (September), (c) Unit test September flow duration curve when August and September are selected as influencing months (top 10% of flow days shown). Boxplot whiskers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

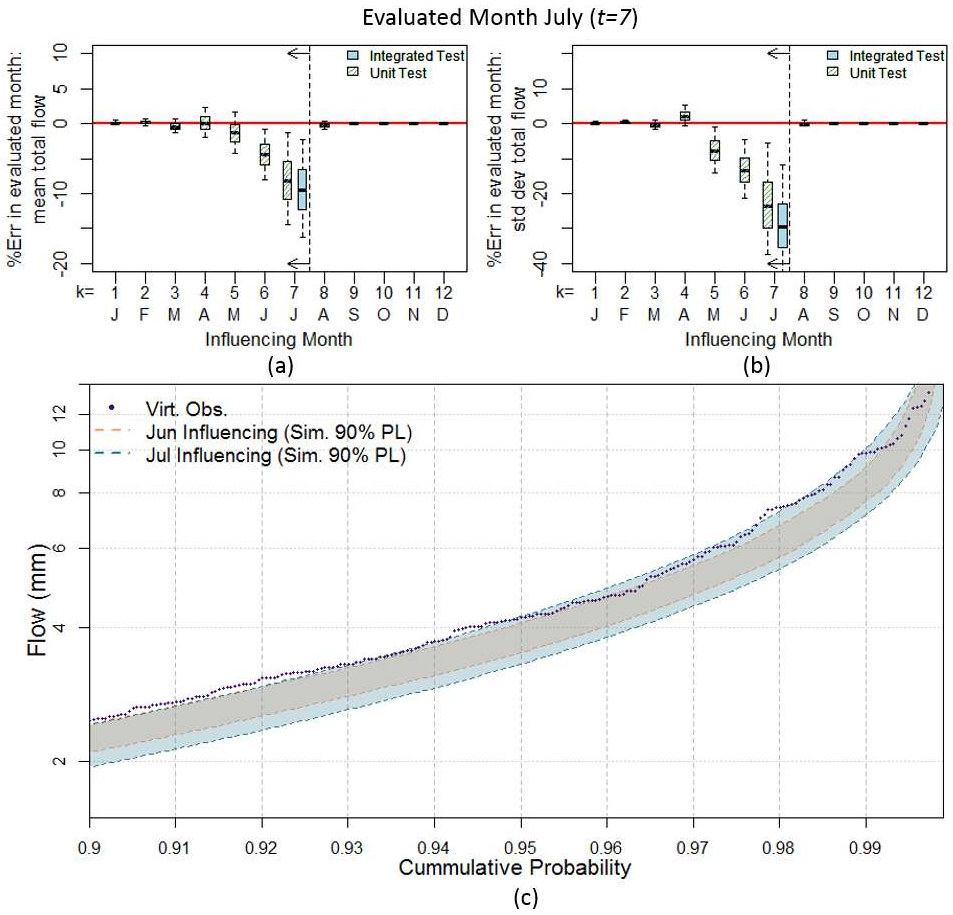
4.3.2 Streamflow errors originate from rainfall model deficiencies over a contiguous block of months

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 (a) and (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 (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~~ is located 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.


Evaluated Month July (t=7)
%Err in evaluated month: mean total flow
Integrated Test
Unit Test
k=
Influencing Month
(a)
%Err in evaluated month: std dev total flow
(b)
Virt. Obs.
Jun Influencing (Sim. 90% PL)
Jul Influencing (Sim. 90% PL)
Flow (mm)
Cummulative Probability
(c)

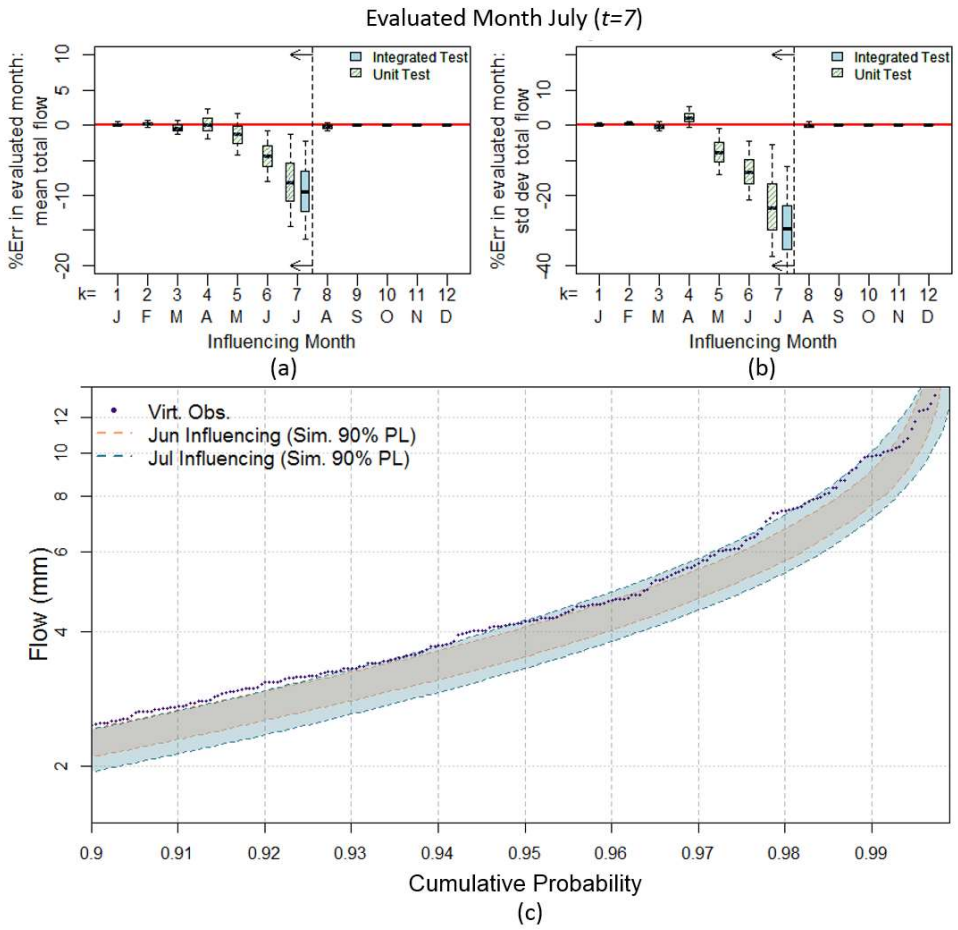


Fig. 10 Lobethal, Site 12 (a) unit test error in mean monthly total flow (July), and (b) unit test error in standard deviation of monthly total flow (July), (c) July flow duration curve when June and July are selected as influencing months in unit test (top 10% of flow days shown). Boxplot whiskers indicate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

4.3.3 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)).

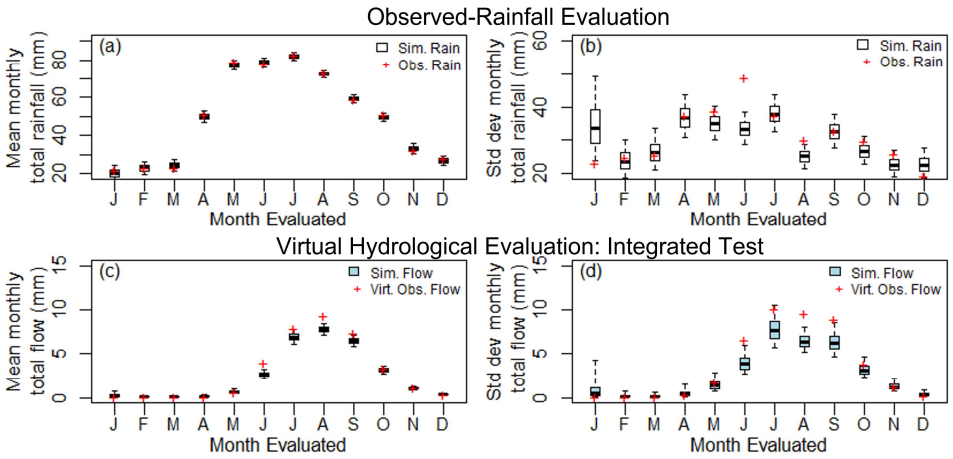
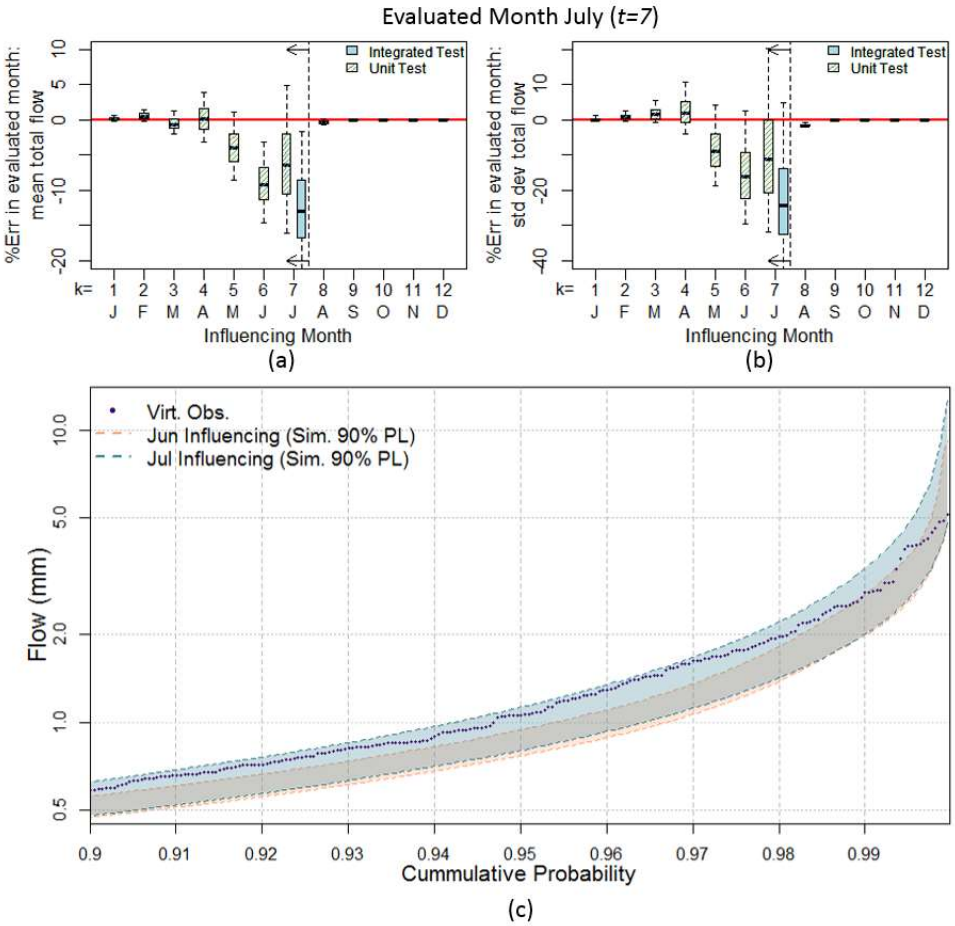


Fig. 11 Happy Valley (Site 10) (a) observed-rainfall evaluation mean monthly total rainfall, (b) observed-rainfall evaluation standard deviation of monthly total rainfall, (c) virtual hydrological evaluation (integrated test) mean monthly total streamflow (d) virtual hydrological evaluation (integrated 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 (a) and (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 evaluated site/month combinations (i.e. those identified in Step 2) ~~for this case study~~.


Evaluated Month July (t=7)
%Err in evaluated month: mean total flow
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
(a)
%Err in evaluated month: std dev total flow
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
(b)
Virt. Obs.
Jun Influencing (Sim. 90% PL)
Jul Influencing (Sim. 90% PL)
Flow (mm)
Cummulative Probability
(c)

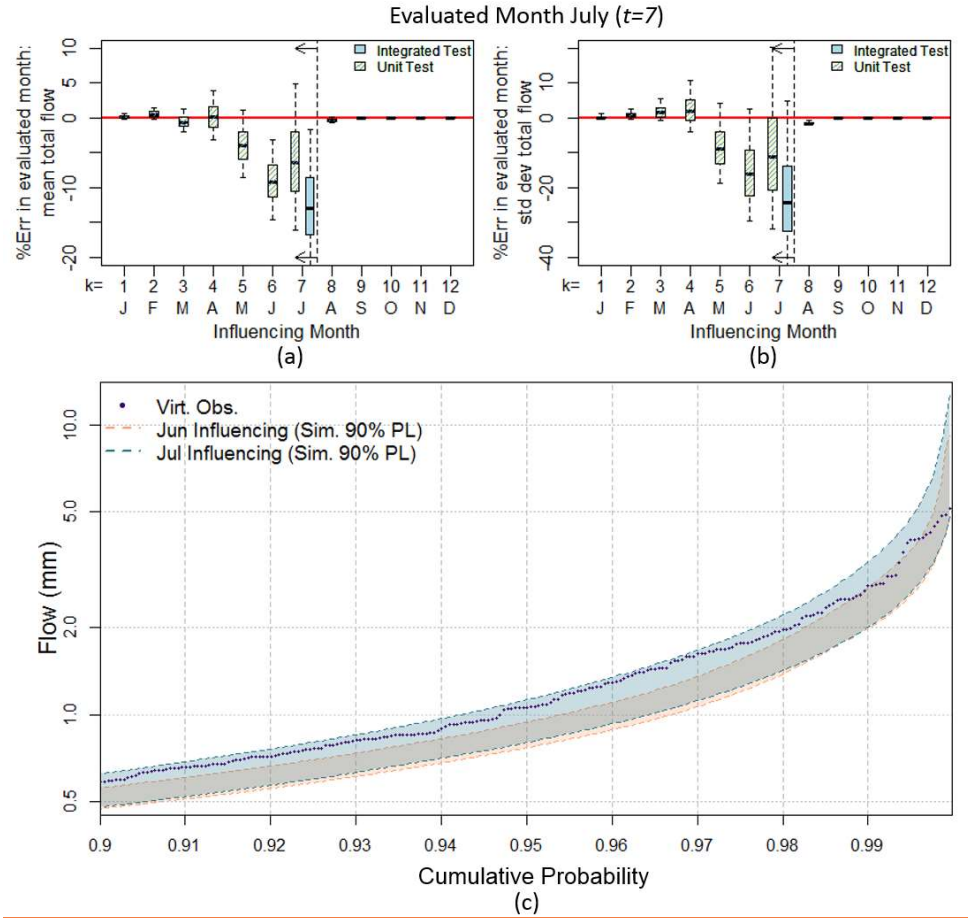
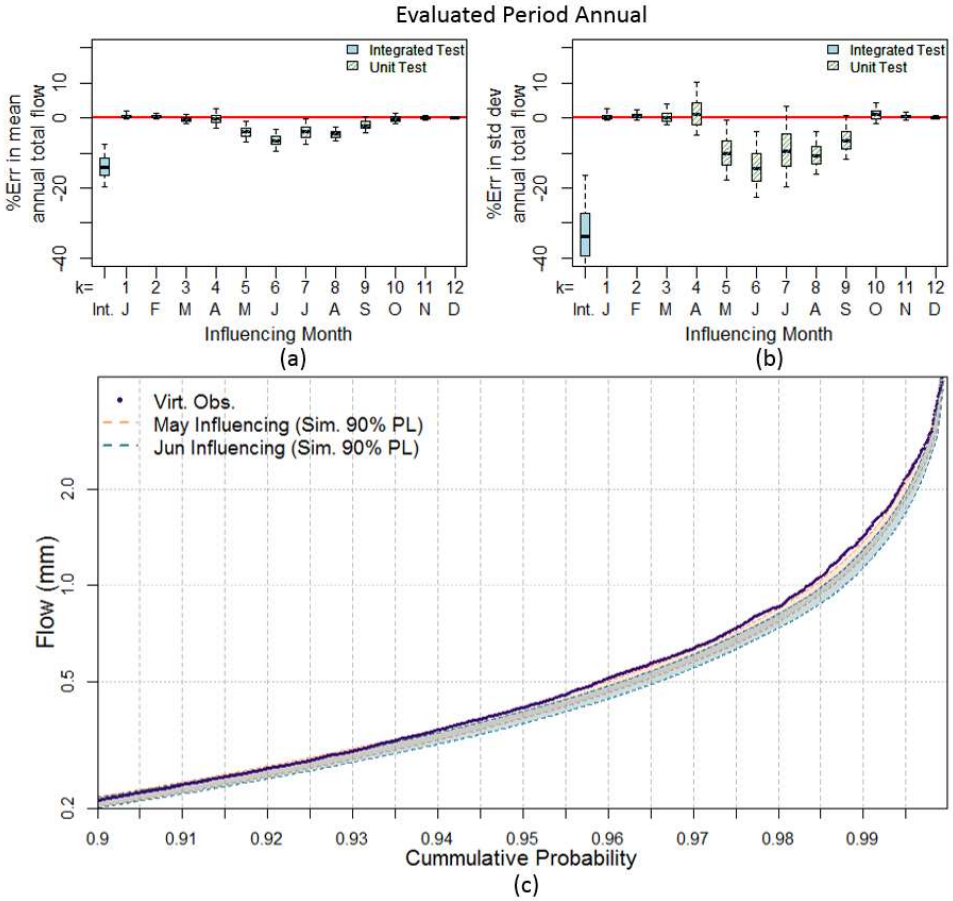


Fig. 12 Happy Valley (Site 10) (a) unit test error in mean monthly flow (July), (b) unit test error in standard deviation of monthly flow, and (c) July flow duration curve when June and July are selected as influencing months in unit test (top 10% of flow days shown). Boxplot whiskers indicate the 90%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

4.3.4 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 (a) and (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 lies~~is located 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). This insight from the use of unit testing would be difficult to obtain using other evaluation strategies (further discussed in Section 5.2).


Evaluated Period Annual
Integrated Test
Unit Test
%Err in mean annual total flow
%Err in std dev annual total flow
k=
Int. J F M A M J J A S O N D
Influencing Month
(a)
(b)
Virt. Obs.
May Influencing (Sim. 90% PL)
Jun Influencing (Sim. 90% PL)
Flow (mm)
Cummulative Probability
(c)

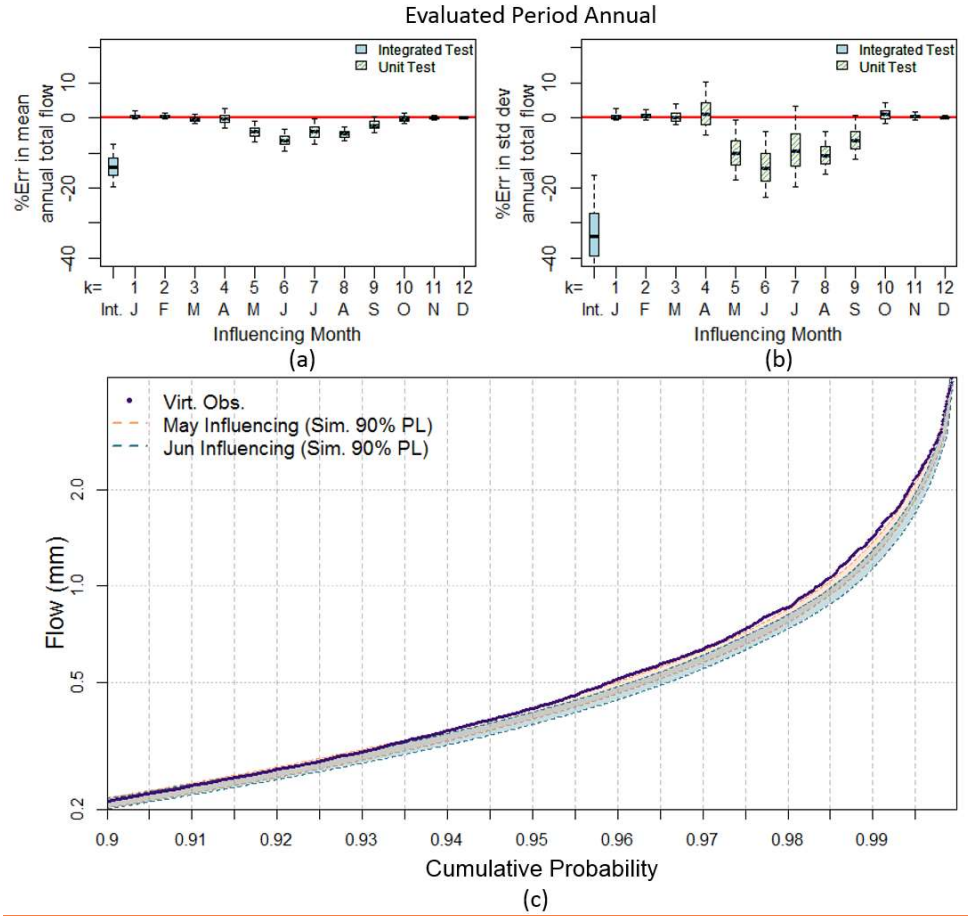


Fig. 13 Happy Valley (Site 10) (a) unit test error in mean annual total flow, and (b) unit test error in standard deviation of annual total flow, and (c) annual flow duration curve when May and June are selected as influencing months in unit test (top 10% of flow days shown). Boxplot whiskers indicate 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 (Chiew, 2006), using the principles of error propagation ~~of linear functions~~ (Ang and Tang, 2007), assuming linearity it follows that a 10% error in mean/standard deviation of rainfall could potentially be amplified to 20-35% error in the mean/standard of streamflow. This estimate represents a lower-bound of the potential amplification, since the non-linear nature of the rainfall-runoff transformation will likely produce a larger potential amplification of errors. 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.

5.2 The benefits of the virtual evaluation framework

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 SRM. This paper introduces an innovative unit test that was used 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.

This unit test provides added value over and above the integrated test because it identifiesying which are the influencing months which that 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 there 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. 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). The unit tests in Section 4.3.4 show that, May and June combined make up 13% of the total annual flow volume (Fig. 11 (c)). However, they contribute to 11% of the error in the mean annual total flow (Fig. 13 (a)) and 24% error in the standard deviation (Fig. 13 (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 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 single 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) and Coron et al. (2012)); (Perrin et al., 2003, Coron et al., 2012); (b) calibrating and evaluating the hydrological model 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. (2014b) and Westra et al. (2014a) (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 samplinge 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~~ rainfall generators – though these would require a spatially distributed hydrological model.

6 Conclusions

This paper has introduced a virtual hydrologic evaluation framework that enables ~~the~~ targeted hydrological evaluation of SRMs. ~~This paper has introduced 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 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. The integrated test compares simulated streamflow and virtual-observed streamflow to detect overall deficiencies in the ability of at-site stochastic rainfall to reproduce streamflow statistics. The unit test enables the attribution of detected streamflow errors to specific months of stochastic rainfall. The integrated and unit tests enabled different conclusions to be reached in terms of priorities for improving the rainfall model. These conclusions would not otherwise have been possible with conventional evaluation methods

that focus either on rainfall statistics, or on high streamflow months. The integrated test demonstrated that while large discrepancies were identified in low rainfall months these did not translate to deficiencies in streamflow due to the dry state of the catchment. The test also indicated instances where modelled rainfall categorised as 'good' translated to 'poor' flow due to the influence of catchment 'memory' and rainfall from prior months. The unit test identified the importance of the simulated rainfall in ~~transition~~the transition months of May and June (late autumn/early winter) during ~~in~~ the 'wetting-up' phase of the catchment cycle for producing low errors in subsequent high streamflow months (July/August/September) and the annual streamflow distribution. ~~. 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~~ provides valuable additional ~~useful~~ diagnostic ability for the development and application of SRMs, not available by using rainfall-based evaluation techniques alone.

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.

10 References

Commented [BSB1]: Note that the Ang & Tang (2007) reference has been updated to provide additional details.

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 (2nd Edition),* Hoboken, New Jersey, USA, John Wiley & Sons Inc, hardback ISBN-10 0-471-72064-X.

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.