Response to Reviewer 2

Major Comments

Comment 1:

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

Response 1:

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

Comment 2:

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

Response 2:

Thank you, we will shorten Section 2 while retaining the key points.

Comment 3:

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

Response 3:

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

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

1. To include the details of a second SRG and/or a second hydrological model, as well as providing a complete explanation of the details of the framework, observed-rainfall evaluation, virtual observed streamflow evaluation, two different tests, the integrated and unit tests, would make the paper and/or analysis overly long. Reviewer #1 has asked for extra details and additional figures/tables to explain a wide range of details – including the hydrological model calibration (see reviewer #1 comment 15), extra streamflow analysis (see reviewer #1 comment 3)[1]. This is over and above the seven figures already included. If we included another SRG and hydrological model, the number of the figures could

[1] At least two additional figures will be incorporated to provide information on the seasonality of the catchment (rainfall and runoff). Table 2 will also be extended to characterise the rainfall time series at each site in more detail and supplementary material provided to give further detail again. Additionally, we will be examining the flow duration curves to see if they provide additional insight on the deficiencies of the rainfall model, over and above what is already presented. Where appropriate, we will add them to the manuscript and/or supplementary material with additional discussion.

increase dramatically making the paper overly long and lose focus on the presentation of the framework.

2. The framework has not been presented before, in particular the unit test. Therefore, in the manuscript we present the evaluation of a single stochastic rainfall model to demonstrate the framework. This application to a single model has demonstrated some new insights; that the errors in streamflow for a particular month can be affected by errors in the in rainfall from the previous 2 to 3 months. This innovation is something that has not been previously identified in the literature.

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

Comment 4:

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

Response 4:

Thank you.

We will tone done the discussion of the framework's usefulness for flood risk applications as this is not demonstrated in the manuscript. For example, on page 9 line 17 we will use an example not related to flood risk in explaining considerations relevant for choosing an appropriate hydrological model.

Additionally, reviewer 1 (see comment 3) has suggested that we examine additional streamflow characteristics, in particular flow duration curves. We plan to do this and include the flow duration curves, where they provide insight. This will provide a broader demonstration of the framework through an application that considers statistics at the finer (daily) scale.

Comment 5:

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

Response 5:

Thank you for pointing this out. We will revise the example in Section 2.3.2 which explains how the unit test figure (those like Figure 3b) are constructed and used as a diagnostic. The Figure caption will also be improved. The revised text will clarify that each boxplot is a summary of the error (the difference between the simulated and virtual-observed performance statistic, Eq. 4) for the 10,000 replicates of the simulated 73 year time series. The "13 errors" relates to the number of boxplots displayed in figures of this type (the unit test for 12 influencing months and 1 integrated test). As you point out, this description of "13 errors" on page 8 line 24 is confusing. It will be revised to make clearer that there are 13 error summary boxplots, and also set out how the boxplots are constructed (a related response on Figure 3 is given to reviewer 1, comment 11).

The integrated test is designed to be a compact evaluation that includes multiple sites and statistics. Once the integrated test is completed and problems identified in the simulated streamflow, the more detailed unit test is applied to sites of interest (Step 6 of the framework). Figure 3 describes the unit tests, which are not designed to be undertaken on a large number of sites – they are designed to be more probing and are only undertaken on certain sites with problems, such as those identified by the integrated test.

The reviewer is right to point out that the example on page 3 does make it look like the proliferation of error metrics is identified as a limitation. We can see that the accompanying example emphasises the large number of statistics rather than our intended key point: that there are difficulties in assessing trade-offs or the relative importance of statistics. The example will be revised to clarify that the key issue for observed-rainfall evaluation is the difficulty in understanding the relative importance of rainfall features in terms of streamflow generation and what to do when performance is 'mixed'.

Comment 6:

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

Response 6:

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

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

Comment 7:

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

Response 7:

The reviewer is correct, as the focus is on evaluating stochastic rainfall generators, the other forcings are supplied independently. In our case study, the potential evapotranspiration (PET) time series (our only other meteorological forcing) is unchanged from the observed values in all hydrological simulations (i.e. the same PET time series is used in the simulation of the virtual-observed streamflow, integrated tests and unit tests). This is important as the hydrological evaluation is a relative comparison of the observed and simulated rainfall, hence all other time series and parameters relating to the hydrological model are kept the same in all instances. 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. Figure 1 shows the rainfall-PET correlations for a drier summer month (January) and a wet winter month (June).

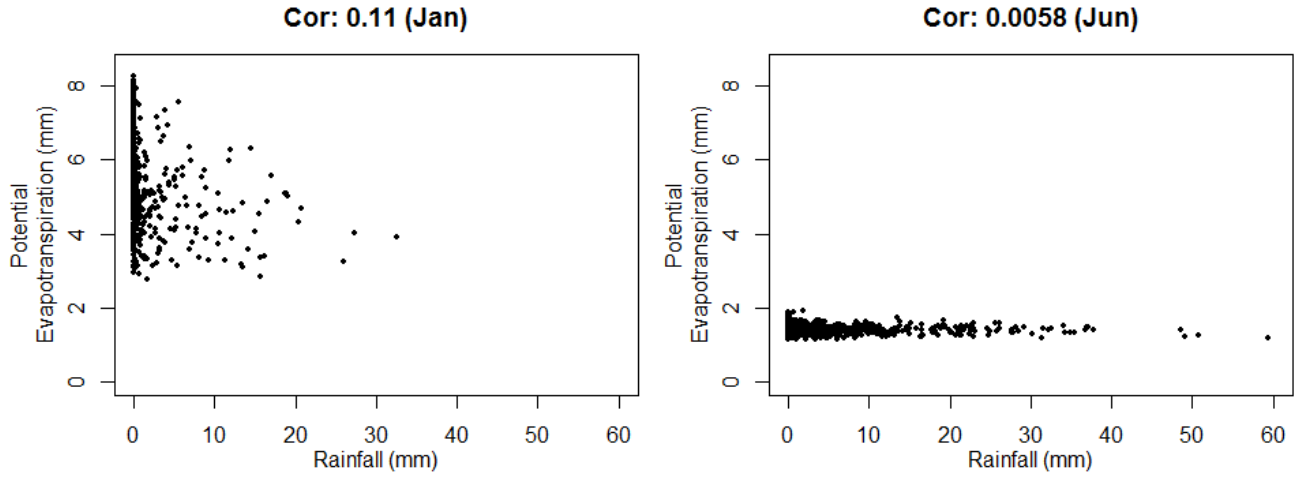

Figure 1: Rainfall-PET correlation (left) January and (right) June.

While there is some non-zero relationship, we do not consider it to undermine the case study (since all other statistics of PET are reproduced and the relationship is mild). However, this may not be the case for other locations where the model is applied. We will therefore discuss the matter in the paper, and suggest a method to identify whether it is significant. The method is to perform a bootstrap of the observed rainfall, for example, shuffle the order of years of rainfall but keep the same sequence of PET–thus breaking the rain-PET correlation–and test for differences in terms of virtual-observed streamflow compared to the observed time series of rainfall and PET.

We will include the reviewer's recommendation to apply the framework more generally to stochastic weather generators. The application would require care to ensure that the PET generator does not introduce other deficiencies.

Comment 8:

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

Response 8:

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

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

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

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

2. "*I would suggest that a method that "tracks" the propagation of rainfall through the model might be more effective … most clear way of doing this is to track how different rainfall statistical moments translate to different statistical moments in the streamflow, using both historical and synthetic rainfall*" – This idea offers scope to extend the framework, and is something to consider in the future. In this paper, we describe the integrated test and then introduce the unit test in terms of 'splicing' monthly blocks of rainfall. However, the reviewer is right that the approach could be formulated differently to use different 'splicing' approaches. For example, to examine the percentage changes in resultant streamflow as a function of a particular change in the inputted rainfall. We will clarify that the framework in its currently presented form is not the only way to undertake this type of investigation.

3. "*I recommend that the authors delete the statement that this framework "should be an essential step in the development and application of stochastic rainfall models*"" – We will soften the wording of this statement.

4. "*On a related note, the authors should comment on how this technique would apply to distributed (i.e. high-resolution gridded) SRGs and hydrologic models*." – This is an important topic. We will comment on how this technique could apply to distributed rainfall and hydrologic models in Sections 5.2 and 5.3.

Comment 9:

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

Response 9:

This is a valid and interesting point. Bennett et al. (2018) demonstrated that the monthly autocorrelations are small for the Onkaparinga catchment (from -0.2 in drier summer months to 0.3 in the wetter winter months), and as a result this issue was not considered in the presentation of the framework.[2] We can appreciate that monthly/seasonal autocorrelation is a significant feature of other locations and that could be a limitation when applying this method. We briefly suggested in Section 5.3 that the issue of monthly autocorrelation could be explored as an extension of the model – we will provide further details to explain how this might be achieved. For example, rather than naïvely splicing rainfall it might be possible to conditionally sample the simulated rainfall in a manner that preserves monthly correlations. The efficacy of this technique would require some exploration since there may be limitations arising from the conditional sampling.

Comment 10:

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

Response 10:

The reference to 90% limits indicates that the boxplot whiskers extend to from the 5^{th} to 95^{th} percentile values of the metrics based on the 10,000 replicates. The initial description of these figures (Section 2.3.2) and all figure captions will be revised to indicate that the "(90% limits shown)" indicates that the boxplot whiskers extend to the 90% limits of the 10,000 simulations for the presented statistic. Please also see the response to comment 5.

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

Comment 11:

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

Response 11:

Thank you, the basin size will be included in the revised manuscript (323 km^2).

All the sites identified in Figure 2 were used to estimate the catchment average rainfall (in the revised Section 3 of the manuscript we will better explain this) for the rainfall-runoff modelling calibration. When estimating catchment average rainfall it is fairly common to use sites outside the catchment, to better represent the spatial variability and to avoid boundary effects. It is therefore important that a stochastic rainfall model is able to reproduce the rainfall statistics at all of the sites outlined in Figure 2. This is why we evaluated the stochastic rainfall model at all the sites indicated in Figure 2.

There are further reasons why this is valid for a virtual approach. Most notably, because there is no comparison made with observed streamflow. The virtual hydrological evaluation uses the calibrated hydrological model as a tool to process the observed and simulated rainfall for comparison. The virtual-observed streamflow can be thought of as a virtual stream flow gauge. The virtual stream gauges have no physical location that they are trying to replicate. Instead the virtual stream gauges enable a synthetic test of the simulated rainfall.

Virtual hydrological evaluation of a single rainfall site is analogous to treating the information at the selected rainfall gauge (observed and simulated) as being representative of the catchment rainfall. This 'catchment rainfall' is then routed through the chosen hydrological model to produce simulated and virtual-observed streamflow at the 'virtual catchment outlet'.

This type of virtual approach was used in a different context (the development of new techniques for flood frequency estimation) in which a calibrated hydrological model was 'moved' all over Australia (Li et al., 2016).

Comment 12:

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

Response 12:

Thank you, we will comment on this issue in Section 5.3.

The reviewer is right to point out the complicated nature of comparing stochastically generated rainfall against a nonstationary continuous simulation. In this paper we took steps to minimise this impact by careful selection of the observed rainfall period.[3] Evaluating non-stationarity is considered an extension to this framework and therefore left for future research. It is conceivable that the same general framework can be applied for a selected non-stationary period, but 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.

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

Minor Comments

Comment 13:

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

Response 13:

Thank you for pointing this out. We will change 'months' to 'seasons'.

Comment 14:

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

Response 14:

Thank you, we will introduce and acronym for stochastic rainfall models and use it throughout the manuscript.

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

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

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

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

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

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

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

Comment 22: ***Page 1 line 24-25: I recommend deleting "risks" after "floods" and "droughts" and changing it to "hazards" after "hydrologic"***

Comment 23: ***Page 3 line 16: add comma after "poor"***

Comment 24: ***Page 5 line 10: The goal is not to match streamflow observations. It is to match the statistics of streamflows***

Comment 25: ***Page 3 line 12: add "model" before parameters***

Comment 26: ***Page 3 line 13: Why would you call ET "extraneous"? It is generally very very important.***

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

Comment 28: ***Page 11 line 16: grammar problem "was fit good"***

Comment 29: ***Page 19 line 3-4: streamflow arises from more than just rainfall integration over a catchment area-what about ET, etc.?***

Comment 30: ***Page 20 line 13: good place to mention that multiple SRGs could be used too, not just multiple hydrologic models.***

Comment 31: ***Page 20 line 18: put "memory" in quotations.***

Comment 32: ***Page 14 line 14: This sentence is a bit awkward. It isn't perhaps so "common and obvious" to the reader.***

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

Response 15-33:

We thank the reviewer for pointing out the above editorial corrections (comments 15 - 33) and for their thorough consideration. We will incorporate these corrections in the revised manuscript.

References

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

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

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

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

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

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

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

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

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

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