Response to Reviewer 1

Comment 1:

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

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

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

Novelty of the integrated test – we will provide better referencing and discussion on existing uses of virtual-observed streamflow evaluation (calibration, validation, model selection and diagnosis) in the introduction, including the rewording presentation of objective 2. However, we feel that there are some important aspects of our implementation of our framework that distinguish it from existing presentations of virtual-observed streamflow evaluations. In particular, the presentation of the integrated test in this paper is the first time, a virtual-observed streamflow evaluation has been formalised used using a comprehensive and systematic evaluation (CASE) framework approach (Bennett et al. 2018). This distinguishing feature and others are further discussed in response to comment 2. We will take the reviewer's advice and emphasize the novelty of the unit test and its diagnostic ability in the revised manuscript.

Evaluation using other runoff characteristics – we will examine if additional runoff characteristics such as flow duration curves provide additional insight on the deficiencies of the rainfall model, over and above what is already presented and incorporate discussion of these insights where appropriate.

Length of theoretical elements – we will reduce the length of the relevant sections.

Information on the r-r model – we will provide better explanation, including references of the calibration and validation procedure of this rainfall-runoff model.

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

Comment 2:

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

Response 2:

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

We agree that the concept of a virtual-observed streamflow evaluation is not new and we will revise the introduction, including the presentation of objective 2, to make this clear and discuss that the approach has been used in a variety of contexts (e.g. calibration, validation, model selection and as a diagnostic tool).[1] However, there are some important aspects of our framework that distinguish it from existing presentations of virtual-observed streamflow evaluation, as outlined below.

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

Previous papers (Müller and Haberlandt, 2018; Sikorska et al. 2018, Kim and Olivera, 2012) have used a virtual-observed streamflow evaluation approach, but have not used a CASE framework to evaluate the performance of stochastic rainfall model at multiple rainfall sites in terms of its ability capture key streamflow statistics of interest. For example, Müller and Haberlandt (2018) established the need for spatial consistency of rainfall generation in modelling sewer networks by comparing rainfall disaggregation approaches with or without spatial consistency. This virtual-observed streamflow evaluation is performed for identified extreme rainfall events only and therefore does not use a CASE approach that considers multiple temporal scales and the longer term effects of the applied rainfall on the translation of subsequent rainfall to streamflow. Sikorska et al. (2018)

[1] We will rewrite the literature review to point out that this test is not new and that it has been employed in a variety of contexts, including:

> *Calibration* – Using virtual streamflow to directly improve the calibration of a rainfall model. For example, Kim & Olivera (2012) derived weights to reflect the importance of various rainfall statistics in terms of streamflow. As another example, Li et al. (2014, 2016) used catchment simulations to estimate soil moisture distributions as part of a new technique to derive flood frequencies.
>
> *Validation* – to establish a model as fit-for-purpose together with other validation tests (Kim & Olivera, 2012).
>
> *Model selection* – to identify key rainfall features of multiple competing models or model options in terms of hydrological behaviour. For example, Müller and Haberlandt (2018) established the need for spatial consistency of rainfall generation in modelling sewer networks.
>
> *Diagnostic* – to identify rainfall features of interest in a given rainfall model in terms of hydrological behaviour. For example, Sikorska et al. (2018) found that detailed rainfall time series were not needed to reproduce peaks in the modelled catchments and that simple rainfall disaggregation approaches were sufficient.

focused on identifying rainfall features of interest in terms of resultant hydrological behaviour for the purposes of determining the effective daily precipitation duration with a view to selecting a suitable rainfall disaggregation scheme. Their evaluations determined that detailed temporal rainfall time series were not needed to reproduce annual or seasonal peaks in their modelled catchments. Although the evaluations presented are comprehensive the motivation of the Sikorska et al. (2018) paper is different and does not provide a general formalised framework for systematically categorising stochastic rainfall model performance at specific and aggregate temporal and spatial scales. Kim & Olivera (2012) used virtual-observed streamflow evaluation as part of a larger calibration and validation approach in which various weights were trialled to reflect the importance of various rainfall statistics within a modified Bartlett-Lewis rectangular pulse (MBLRP) model. However the focus was on the improvement and validation of the MBLRP model rather than the presentation of separate framework for model evaluation. Finally, Li et al. (2014, 2016) used a virtual-observed streamflow evaluation approach, to evaluate the ability of range of techniques to estimate the derived annual flood frequency distribution - they did not use a CASE approach to evaluate stochastic rainfall models. In the revised paper we will improve the presentation of the approach, highlighting the key points above to more clearly demonstrate the novelty.

Additionally, the formalisation of virtual-observed streamflow evaluation using a comprehensive and systematic evaluation (CASE) approach, the integrated test, forms a baseline for subsequent application of the unit test which has greater ability to pinpoint issues with respect to the source of the rainfall error on a monthly basis.

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

Comment 3:

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

Response 3:

Good idea, we will examine if additional runoff characteristics such as flow duration curves 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.

Comment 4:

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

Response 4:

Thank you. We will leave out the sentence.

Comment 5:

P3l9-14 The idea behind the example provided by the authors is clear. Nevertheless, some of the rainfall characteristics mentioned are not clear and, since it is only an example, can be left out or can be replaced by other rainfall characteristics:

- rainfall on wet days - What does this characteristic represent (the daily total rainfall itself is mentioned later)?;

- Extreme value analysis on a monthly basis and autocorrelation on an annual basis are from my understanding rather uncommon rainfall characteristics for the evaluation of rainfall time series

Response 5:

Thank you. We will modify the example to be clearer, by only using rainfall statistics that are well-known and require no additional explanation.

To address the specific questions regarding our original choice of statistics, our interest in some of these rainfall statistics arises from our context. For example, (i) rainfall on wet days is important because the calibration should match the moments of the truncated and power-transformed Gaussian upper tail; (ii) extreme value analysis of months is of interest to strongly seasonal locations (e.g. Leonard et al., 2008); and (iii) auto-correlation of annual totals is of interest due to teleconnections in the rainfall signal (Thyer and Kuczera, 2000).

Comment 5:

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

Response 5:

We agree that the details in brackets can be left out and will do so in the revised manuscript.

Comment 7:
P3l7-P4l20 The motivation for the introduction of the new evaluation strategy is quite long and can be shortened by the half. I think the majority of the community is quite aware of the issue with overlapping errors. Also Fig. 1 and Fig. 2 are quite clear from the text and could be left out. If kept, a box with "True rainfall“ should be added in Fig.1a) to be consistent with Fig. 1b ("True streamflow“)

Response 7:

We will shorten the explanation while maintaining the key points of the introduction. Based on our experiences explaining this work, we feel the figures in section 2 are helpful to avoid misconceptions. We prefer to retain them and will amended them as suggested.

Comment 8:

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

Response 8:

Thank you. We will modify the sentence as suggested.

Comment 9:

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

Response 9:

Thank you for this suggestion. Table 1 will be revised so that text is used in place of the original symbols ('Yes', 'No', 'Not Applicable'). The Table caption will also be amended to clarify this also (i.e. 'Yes' indicates that a source of error in included in the evaluation, 'Not Applicable' indicates that a source of error is not relevant to the evaluation and 'No' indicates a source of error is not included in the evaluation). The last line, first column will be amended to read 'virtual hydrological evaluation'.

Comment 10:

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

Response 10:

The reviewer has raised some excellent discussion points. The immediate response is that while we have discussed some of these points later in the investigation (Section 5.2), we will communicate the key issues earlier to the reader. We appreciate the centrality of the issue raised and will highlight it in Section 2.2. We provide specific responses below to the discussion points raised.

1. "*Virtual-observed streamflow approach can still be biased by the applied r-r model*" – Yes, we agree. This is a very important matter to consider.
2. "*Differences in terms of rainfall characteristics, but not in the simulated streamflow*" – Yes, there is the potential that a chosen rainfall-runoff model is insensitive to certain important differences in modelled rainfall. Further to this, it is important to mention that a unit test can still get insights using a lumped rainfall-runoff model. This initial test is a necessary, but not sufficient condition for spatial rainfall models. If a rainfall model, cannot get the virtual-observed streamflow statistics from a well calibrated, well-known lumped rainfall-runoff model right, there is limited value in examining the spatial statistics.
3. "*But this also depends on i) the model choice … and ii) the parameter identification*" – Yes, as with the comments in (2), all elements of the modelling method can potentially introduce bias. The end-user's impact of interest and associated modelling process can influence an outcome. These observations reinforce the need for care when applying the framework. We have chosen a widely applied model, GR4J. We also adopted rigorous calibration which is presented compactly in a journal paper (Westra

et al. 2014a), but with a detailed report also available (Westra et al. 2014b). The calibration of the hydrological model is further discussed in the response to comment 15.

4. "*If so, how can be avoided that this calibration introduces a bias*" – This is an excellent question, which we will address with further discussion. While best-practice models and methods are important, this does not necessarily guard against the possibility that a model poorly represents key processes of interest. One remedy for this limitation would be to use multiple rainfall-runoff models and this is discussed in Section 5.2.

Comment 11:

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

Response 11:

We will more clearly indicate the meaning of indices k and t in the descriptive text, Figure 3b and its caption to aid the reader.

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

Comment 12:

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

Response 12:

We agree, good idea. We will incorporate a flow chart to illustrate the implementation of the framework in Section 2.4.

Comment 13:

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

Response 13:

Thank you. We will add additional information to explain the 90% limit test as requested both in text and graphically (potentially as supplementary material). The relevant information is available in Bennett et al. (2018), but to be more accessible this information will be reproduced in the current paper.

Comment 14:

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

Response 14:

We agree with the reviewer and recognise that the high-level summaries need more tangible details on the rainfall and streamflow statistics on a monthly basis to help the reader understand the seasonal behaviour of the case study catchment. We will revise Table 2 to characterise the rainfall time series in more detail including the addition of columns that present rainfall statistics (total rainfall, no. of wet days, average daily rainfall, average wet day length, average dry spell durations) in different seasons – for brevity in this table we will show two months: January to represent the dry summer and July to represent the wet winter. Further detail on these statistics for all months at each site will be also provided as supplementary material. In addition, a new figure that shows seasonal variation of catchment average rainfall and streamflow on a monthly basis will be added to Section 3 (Case Study) of the main paper to address the suggestion of the reviewer.

Comment 15:

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

Response 15:

A paper (Westra et al., 2014a) will now be cited alongside the report. The paper provides a compact peer-reviewed summary of the model and its calibration – for a neighbouring catchment (Scott Creek). This paper was acknowledged with a Research Spotlight Award from American Geophysical Union (top 5% of papers in AGU). The reference to the report, Westra et al. (2014b), is also retained since it gives details specific to the Onkaparinga catchment used in this paper and because it is comprehensive. The Scott Creek and Onkaparinga catchments were calibrated as part of the same project using consistent models and techniques.

The model development and calibration was comprehensive and considered a range of aspects including multiple sources of uncertainty (input, output, parameters, etc.)[2] and it is beyond of the scope of this paper to include all the details. Instead relevant aspects of the calibration and model-selection will be added to the manuscript. The relevant features to be included will be: calibration approach, including parameter optimisation method and objective used, calibration and validation results (NSE etc.), and an explanation of the rainfall and runoff data used for calibration and validation.

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

Response 16:
Thanks. Details of the catchment's runoff characteristics at the annual and seasonal level will be added to the revised Table 2 and the new figure (see response to comment 13).

Comment 17:
P11l14-18 On p9l15-17 you mention "The hydrological model should be selected on the basis that it is capable of simulating streamflow for the timescales, magnitudes and physical processes of interest to the intended application." Is the lumped model able to simulate the physical processes of a catchment with a few 100 km2 catchment area (I could not find the catchment area in the manuscript).

Response 17:
The catchment area is 323 km^2 and will be mentioned in the revision (Section 3).

It is important that the chosen hydrological model is fit for purpose (see also discussion of comment 10). The GR4J model used in this paper is for catchment inflows to the Mount Bold reservoir and is appropriate for analysis of catchment yield (i.e. focussed on means and variances of inflow)[3]. However, if we were examining impacts on instantaneous peak flows impacts, this model would not be suitable and if we wanted to look at impacts of distributed rainfall, we would need a distributed rainfall-runoff model. However, for the purpose of this paper, which is to demonstrate the virtual hydrological framework (including the unit test) for evaluating the ability of a stochastic rainfall model to estimate catchment yield, the model is deemed sufficient.

---

[2] The hydrological model calibration considered 24 model variants combined with likelihood estimation of a heteroskedastic error model. The calibration separated out multiple sources of uncertainty (input uncertainty from gauges and radar, output uncertainty associated with streamflow gauges, and model uncertainty). All 22 rainfall stations were used to estimate areal rainfall. The areal rainfall was interpolated using kriging with external drift on a daily basis using a similar latent-variable Gaussian model as the stochastic model from Bennett et al. (2018). The areal rainfall estimation was performed using Thiessen weights for comparison. The number of gauges is relatively dense and the uncertainty due to rainfall inputs was also assessed relative to other sources of uncertainty (Westra et al. 2014b; pg. 7).

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

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

Response 18:

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

Comment 19:

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

Response 19:
There are many interesting features of Fig 5 like this. The fact that there is not a direct correspondence between 'good' rainfall and 'good' runoff, and/or 'poor' rainfall and 'poor' runoff is one of the motivations for the virtual hydrological evaluation framework in addition to observed rainfall-based evaluation. Figure 5 shows that it is possible for seemingly 'poor' rainfall to yield 'good' runoff (as also 'good' rainfall can yield 'poor' runoff). We note that the discrepancy in Figure 5 is not in terms of mean values, but for the standard deviation of monthly aggregates (see Figure 5, *sd*(total) for rainfall and runoff in Jan, Mar, May, Jun, Oct, Nov, Dec). For drier months (Nov-Mar) the lack of correspondence (i.e. 'poor' rainfall producing 'good' runoff) is due to the low amount of runoff. While in wetter months (May-Oct) the relationship is more complicated as shown in the unit test demonstrations (Section 4.2). This will be fully explained in the revised Section 4.1 of the paper.

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

Response 20:
This concept has not been demonstrated in the main paper, therefore it will be deferred to the discussion on outlooks (Section 5.2). We believe the unit test has sufficient novelty to represent a substantial contribution, hence we will the leave spatial rainfall evaluation for future developments.

Comment 21:

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

Response 21:
The comment is correct that the proposed method does not identify the impact of specific singular rainfall characteristics on the resulting runoff. However, the framework does provide a clear approach to isolate which set of components of the rainfall model require further attention. The integrated test focuses attention on hydrological properties, and the unit test can isolate deficiencies in rainfall by month. When applied to the case study in our paper, the limitations of the model are in the variability of the rainfall and not in the rainfall mean. The initial motivation for the approach can be seen in Figure 5 where the mean of the annual rainfall is 'good', but the mean of the annual runoff is 'poor'. Figure 5 also shows that this is mostly attributed to 'poor' mean runoff in June, July and August. Unit tests were then used to show that the rainfall in the catchment 'wetting-up' period (May-June) is of key importance. This is greater insight than could have been achieved with observed-rainfall evaluation and is greater insight than could be gathered from other virtual-observed streamflow approaches.
However, we agree with the reviewer that the framework cannot currently distinguish between particular features of the rainfall (e.g. "Is it rainfall correlation, magnitude, or intermittency that causes a low standard deviation in monthly streamflow?"). Nonetheless, the framework has significant potential to be extended to diagnose which are rainfall characteristics are. This can be done by comparing multiple rainfall model variants (parametrically, or via bootstrap techniques) which are designed to have contrasting features of a key characteristic (e.g. intermittency, rainfall correlation). Such an approach was undertaken by Evin et al. (2018) using an observed-rainfall evaluation approach to compare model variants. In the revised paper, this will now be identified as a limitation and this extension will be highlighted for further research.

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

Response 22:
Thank you for pointing out that the example is hard to follow. We will add additional information on catchment seasonality in the case study description to better explain the importance of the 'wetting-up' months and storage in the catchment (also see responses to comments 14 and 16). We will then revisit this example to explain the concept more concretely.

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

Response 23:
Thank you, this will be corrected.

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.

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

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

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

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

Li, J., Thyer, M., Lambert, M., Kuzera, G., Metcalfe, A., 2016. Incorporating seasonality into event-basedjoint probability methods for predicting flood frequency: A hybrid causative event approach. J. Hydrol., 533: 40-52.

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

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

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

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