Response to Reviewer 3

Comment 1:

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

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

Response 1:

Thank you for your comments. We will revise the manuscript to address the matters raised. We will shorten the section explaining the virtual framework. While the conceptual basis for using a virtual streamflow framework to benchmark SRG's is generic, we agree there are some aspects of the implementation which require future extensions to generalise the framework for a wider range of case studies (other stochastic rainfall generators, hydrological models and locations). In the revised paper discussion we will outline the aspects of the implementation that need further work. Naturally, some of the outcomes are tied to the specific case study we have used to demonstrate the framework (and the two tests). In the revised paper we will clearly outline the outcomes that are generic and the outcomes that are specific to the case study, to better guide the reader. Greater detail is given in response to specific comments below.

Comment 2:

Abstract section.

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

Response 2:

We will revise the abstract to provide a clearer explanation of the framework, including:

- Page 1| lines 12-15 – more detail will be provided on the method[1]; and
- Page 1| lines 18-20 – a more general description will be given, since reference to a 'wetting-up' period is not a clearly defined term based on the abstract.[2]

---

[1] The evaluation methods developed in this paper apply the observed and simulated rainfall to a hydrological model so that the impact of rainfall on a given catchment can be assessed in terms of its hydrological significance. The hydrological response of simulated rainfall is evaluated relative to the hydrological response of observed rainfall so that complications with observed streamflow are sidestepped. In this virtual framework it is further possible to create spliced records of simulated rainfall embedded in the observed record so the influence of a given month can be isolated, which is useful because catchment storages can aggregate rainfall over multiple months.

[2] The framework was able to identify that deficiencies in rainfall for some key months (e.g. May, June) had persistent impacts on subsequent higher flow months (Jul, Aug, Sep) that significantly affected the estimate of total annual flow.

Comment 3:

Section 2.

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

Response 3:

This comment aligns with feedback from other reviewers. We will shorten Section 2 and merge Figures 1 and 2 as suggested. A flow chart will be added to Section 2.4 to make it easier to follow the method and results.

Comment 4:

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

Response 4:

Thank you. Other reviewers have made similar comments (reviewer #1 comments 14-17, reviewer # 2 comments 7 and 11). We will address all the issues raised by the reviewer in Section 3 of the revised paper.[3]

This insight would not have been identified using existing methods and highlights the important of the virtual hydrological evaluation framework for stochastic rainfall model evaluation.

[3] Missing details to be added to the manuscript in Section 3 include:

- 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.

Comment 5:

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

Response 5:

We will provide a summary of the calibration approach for the rainfall model in the revised Section 3 so that it is easier for the reader to understand the model without needing to also read Bennett et al (2018)

Comment 6.1:

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

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

Response 6.1:

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

The selection of a rainfall model and its calibration approach, are independent of the generic evaluation framework. We agree with the reviewer, that if a different calibration approach (i.e. different rainfall statistics) were used, then the results may change, and the streamflow errors may reduce. Indeed, evaluating how different calibration approaches can influence the streamflow would be an excellent future use of this evaluation framework. As the current paper is already of considerable size (7 Figures and ~8,500 words) and will increase substantially to address the issues raised by the reviewers, examining this issue is outside of scope. We will include this idea in the revised discussion section (Section 5) as a future research application of the framework.

Comment 6.2:

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

Response 6.2:

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

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

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

Comment 7:

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

Response 7:

Thank you for pointing this out. The sentence will be revised to remove the text in brackets, so that the reader need not refer back to Bennett et al. (2018) at this point.

Comment 8:

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

Response 8:

Thank you. We will rephrase the sentence.

Comment 9:

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

Response 9:

Yes, the 13 errors compared arise from unit tests conducted for each of the 12 influencing months as well as an integrated test. We will revise the text to clarify this point.

Comment 10:

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

Response 10:

Thank you for pointing this out. We will revise the description of the integrated test in Section 2.3.1 to make clear that it is evaluated at the monthly scale (to allow comparison with the unit test).

Comment 11:

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

Response 11:

Thank you, the streamflow gauge position will be added to the revised figure.

Comment 12:

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

Response 12:

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

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

We agree with the reviewer that this is an important period of simulation to get right, and as a result of this analysis, we will examine hydrological models that have the potential to better simulate this period.[4] The evaluation using different models is unlikely to change the conclusion that the rainfall in the wetting-up months are important, as the GR4J model already performs reasonably well. However, it may change the magnitude of the rainfall's impact on the hydrological performance – especially for future comparisons to observed streamflow. Further research will investigate this issue, and will be added to the discussion section.

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.

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

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

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

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.