I am grateful for the feedback from the second reviewer. Please see below for specific points.

First, the evaluation should be strengthened beyond visual comparison. Most results are shown as QQ plots and time-series comparisons, but the manuscript would benefit from quantitative metrics for wet-day frequency, wet-day mean precipitation, annual totals, extremes, and wet/dry spell durations. These metrics would help readers assess performance more objectively.

Thanks for this feedback. These will be explained more carefully in the revised paper. The WG for precipitation has been designed to reproduce prescribed wet-day frequency and wet-day mean precipitation either aggregated over the entire period, or on an annual basis if prescribed for individual years. Hence, the purpose of comparing those statistics with observations in this manuscript is a "sanity test" to see that the routine works as it should.

The WG produces annual statistics[1] with a random sequence unless it's told to reproduce the prescribed annual statistics, but still with an overall agreement with the prescribed wet-day frequency and wet-day mean precipitation.

The annual total is approximately the product of the wet-day frequency and wet-day mean precipitation, and follows the two former annual statistics.

An evaluation of extremes and spell durations, on the other hand, provide a more critical test of the WG, and I will elaborate on this in the revised version. The sequencing of the daily amounts is random in the WG simulations, as opposed to the mentioned marginal statistics, and since the number of spells are fewer than the length of the time series, we must expect a degree of random statistical fluctuations in the spell statistics. For all intents and purposes, the spell durations of the observations may also be regarded as random if the observation can be considered as one realisation of many possible. Since the WG involves random number generations, a proper evaluation needs to involve a large number of runs in order to provide robust statistics and meaningful quantitative metrics.

Second, the underestimation of wet-spell and dry-spell duration is an important limitation. The author notes that the WG tends to generate too many short wet and dry spells, but this issue should be discussed more centrally. Spell persistence strongly affects hydrological model outputs, especially runoff generation, antecedent moisture, flood peaks, and drought duration. The manuscript should either improve the method to better preserve spell statistics or clearly state that the current WG is less suitable for applications that depend strongly on precipitation persistence.

---

[1] **by 'statistics', I refer to both wet-day frequency and wet-day mean precipitation.**

This is a good point, and the spell duration statistics may be improved through subsequent swapping of wet and dry days. The need to do so appears to depend on the chosen location and the wet-day frequency climatology (whether there are rainy seasons or not). This can be controlled in the model through an argument to its R call, as it also could be useful to have the very simple type for when the spell durations don't matter. It's important to take into account that the timing of wet and dry days has up to this point been based on the wet-day frequency climatology, since a desire has been to reproduce the rainy seasons, so a limited shuffling of wet and dry days needs to be carried out in a fashion that doesn't severely corrupt the rainy seasons for the sake of getting good spell duration statistics.

Finally, the temperature generator is presented rather briefly. The precipitation WG is the main focus, but the manuscript also reports reasonable skill for daily temperature. The temperature section should either be expanded with more complete evaluation across sites and seasons or shortened and framed as a preliminary extension.

Thanks. This is a good point, and the revised paper will provide more results from evaluations of the temperature-based WG.

Specific comments:

1. Please consider evaluating the WG at more sites or across a broader range of climate regimes to better demonstrate transferability. **The WG for precipitation was evaluated at a larger number of sites in Norway with different climate types for reviewer #1. The Norwegian rain gauge stations were used in this case because they are open access data from thredds which makes it easy for others to reproduce the same analyses, if they can use R and have the most recent esd-package installed. One point I want to make is that this WG is easy and quick to test with any daily rain gauge station data, and users should also do so for the location of their interest - it only takes a minute or so.**

2. Please add quantitative performance metrics to complement the figures. **Quantitative performance metrics will be included in the revised paper based on standard statistical tests to measure whether the observations belong to the same statistical population as the WG simulation. An ensemble of WG may be needed to to properly test the method due to its stochastic nature, e.g. for the spell statistics (the figures will also be revised to show the results for an ensemble). The statistical distribution of daily amounts, simulated aggregated precipitation statistics, as well as spell durations will be compared with that of the observations using the Kolmogorov-Smirnov (K-S) Two-Sample Test. The revised paper will include a table with such metrics.**

3. Please discuss the lack of spatial dependence among multiple stations. This may limit applications to basin-scale hydrological modeling. **This is a good point, and the revised paper will include a more extensive discussion on spatial dependencies. There are two aspects to this, as the WG can incorporate spatial dependencies for *annual* scales, where the covariance for annual wet-day frequency and wet-day mean precipitation is accounted for in WG simulations. On the *daily* time scale, on the other hand, the WG has not been designed to depend on the daily precipitation of an adjacent site.**

4. Please discuss the lack of dependence between precipitation and temperature. This may matter for snow and evapotranspiration applications. **This point is appreciated, and is somewhat related to the point above. A more comprehensive discussion in the revised paper will elaborate on this issue, and it may also be connected to the issue of how to adjust spell durations in order to get a better match with that seen in the observations.**

5. Some figures are difficult to read because labels and legends are small. Please improve figure readability. **Thanks. This point was also made by reviewer #1, and the revised paper will have improved labels and legends.**

6. The conclusion should include a concise limitations paragraph, such as spell-duration bias, stationary seasonality, single-site design, lack of spatial correlation, and limited temperature evaluation. **Thanks for this comment. The conclusion will be expanded in the revised paper to include the limitations of the WGs.**