Reviewer 1:

This manuscript presents EGGS-WG, an open-source stochastic weather generator parameterized from ERA5-Land. The model produces daily precipitation occurrence and amount, hourly 2m air temperature and dew point temperature, and relative humidity over gridded land domains. The stated contribution is a portable, low-resource, globally applicable generator aimed at non-specialist users, with future climate applications in mind.

This is a useful contribution and, in my view, broadly within scope for Geoscientific Model Development. The open-source release with a versioned archive and download scripts is a real strength, the writing is mostly clear, and the validation indicates that EGGS-WG reproduces many ERA5-Land statistics well, particularly for air temperature. I would support publication after revision.

My main concerns are the following. First, the novelty relative to existing generators, and to Abbasnezhadi et al. (2019) in particular, is not stated sharply enough. Second, the justification for the precipitation amount distribution is weak, especially given the stated interest in future climate applications. Third, the New Guinea results point to a systematic problem in the one large tropical domain that is repeatedly noted and then mostly set aside. Fourth, the stationarity adjustment and the use of ERA5-Land as “observations” both sit underneath every validation comparison and need more scrutiny. Finally, several implementation choices are not documented precisely enough to reproduce. I expand on these below.

Thank you for the comments; we have tried to address all of them as best we can, and we believe that they have led to a substantially improved manuscript.

We have tried to clarify our views on novelty throughout the manuscript, although this can be a little difficult, as there are many weather generators that are used for different things, and ours is different from many of these in different ways. To summarise our views, the most important developments missing from other SWG (including Abhasnezhadi's) are the fact our model works globally, is publicly available and open source, and is straightforward to use. Other developments we have taken from Abhasnezhadi but missing from other models, include the regionally coherent simulation and hourly temperature simulation.

We agree with your assessment on the precipitation amounts distribution work. We think that the system in the model is more than serviceable; however, we will push towards implementing a gamma distribution approach in the future. Furthermore, we have also expanded the analysis of the New Guinea region to include regional subdivisions and include some additional regions. This provides much more detail about the nature of these biases. Content has been adjusted and text added to address this concern. Additionally, we have refined our discussion of ERA5 as observation comparisons and stationarity to make the limitations more clear. We have also added more detail for the purposes of reproducibility.

Major Comments

1. The Introduction (lines 40–75, and the recap at lines 505–512) lists global ERA5-Land parameterization, gridded multi-site simulation, hourly temperature, dew point and relative humidity, open code, and low resource use. All are worthwhile, but the reader is left to guess which is the primary contribution as opposed to a sensible engineering choice. Given that the design follows Abbasnezhadi et al. (2019) and the wider Richardson/Wilks family closely, please state plainly what is new. If the main advance is the automated global parameterization from a single reanalysis product, that should be the organizing claim, with the validation framed around demonstrating it. A short paragraph, or a table contrasting EGGS-WG with a few representative generators across these features, would make the contribution much easier to assess.

We have changed the text in the introduction and conclusions to try and make it more clear that our primary purpose in the paper is to present an open source model in the vein of Abbasnezhadi. We think part of the confusion here is that the abstract was already written to support this, but these other sections needed updating. Unlike the original work from abhasnezhadi, our model is designed with a focus on working for flexible simulation domains over the world. This has been reinforced throughout the text. We workshopped the suggested table but ultimately it was hard to get it to a form where it wasn't just more confusing than the corresponding text.

2. The justification of the precipitation amount distribution should be strengthened. At lines 120–127 the authors avoid empirical distributions partly because they could constrain future climate change applications, and then adopt a three-parameter mixed exponential without justifying that choice against the same criterion. This is where I think the reasoning is weakest. There is now a well-developed theory for why daily precipitation intensities follow a Gamma-like form, in which the cutoff or scale parameter PL has a direct physical interpretation in terms of moisture convergence fluctuations during precipitating events (Martinez-Villalobos and Neelin, 2019), and in which changes in that parameter provide an interpretable, theory-backed account of how the intensity distribution and its extremes shift under warming (Martinez-Villalobos and Neelin, 2023; see also Groisman et al., 1999 for the observational precedent). The mixed-exponential parameters do not carry a comparable interpretation or an established link to warming-driven change, so the stated future-climate motivation does not obviously support the chosen form. I am not asking the authors to switch distributions, but given that they invoke future climate applicability as part of the motivation, they should either justify the mixed exponential on that basis, or state explicitly how its parameters would be perturbed under climate change, and acknowledge the Gamma-like alternative.

Before this comment, the authors were unaware of the two Martinez-Villalobos papers referenced in this comment. We would like to thank the reviewer for bringing them to our attention, as they are very informative and show a clear pathway for inducing a future warming response that we had missed. If we had been aware of these papers at the beginning of the project, we would have opted for a gamma distribution for the model. We

still considered switching after this comment, but ultimately the development time (mostly associated with the testing of the parameter fitting process) would likely be unjustifiable for the potential improvements. I have started working with a student to build and test a parallel gamma model and investigate what the full set of impacts is, and when we release an updated version of the model, we will likely change the distribution to gamma. If the results from the student project show a significant improvement, we will likely just release a model update with a gamma distribution. We have completely rewritten the related section, making all of this clear.

3. The New Guinea behavior should be confronted directly rather than deferred. Across the main text and supplement, the one large tropical domain consistently underperforms: underestimated precipitation extremes, in contrast to the overestimate seen elsewhere (lines 369–372); reduced variance, especially in annual cumulative precipitation (lines 380–390); an emergent “regionwide” bias (lines 384–389); weaker dew-point spatial correlations (lines 421–423); and a much narrower ensemble spread than observed (lines 478–492). The domain is also much larger than the others (lines 315–316), and precipitation occurrence approaches 100% in some months (lines 322–324). The authors note these issues and then suggest that domain size may be involved, but this is not really tested. For a paper whose claim is global applicability, this is the case that most tests the claim. Is the problem domain size, the tropical/monsoonal regime, the near-100% occurrence rates, or their interaction with the spatial-correlation machinery? At minimum I would like a focused discussion; ideally one diagnostic experiment, for example re-running a New Guinea sub-domain comparable in size to the other regions, to separate the size effect from a genuine tropical-regime failure.

We have addressed this comment by running additional simulations across the New Guinea region by splitting it into four separate regions. This provides additional information about the biases in this region. We have also added a new section that discusses these biases directly as well as some additional regions that clarify the nature of these biases.

4. The stationarity adjustment (Sect. 3.2) should be justified and its effect on validation shown. The data are shifted to “2022 climate” by fitting a Fourier series, taking its linear terms to build a trend line, and inverting that trend (lines 218–224). Using the linear terms of a Fourier fit to extract a secular trend is unusual, and the manuscript does not explain why this is preferable to a direct linear trend, a seasonally varying trend, or a nonparametric trend. Because all validation is against these adjusted data, any artifact here propagates into every comparison in Sect. 4. Please justify the method and, if possible, show that the validation is not substantially changed by a reasonable alternative detrending.

We have added some text clarifying this issue in the manuscript. Because of the way the detrending is done, validating the detrended data against the stationary simulations is effectively the same as the original data vs our non-stationary simulations. Just to be explicit here we are fitting a Fourier series with a linear time dependant term and using this term. In the majority of cases this will be very similar to a direct linear trend. However, we wanted to be careful about any undue seasonal impact as this will be otherwise accounted for in the model.

5. Please be explicit about the limits of validating against reanalysis. ERA5-Land is used throughout as “observational” data, which the authors acknowledge is not strictly correct (line 208). This matters most for extremes: lines 529–530 concede that the model overproduces extremes while ERA5 underproduces them, “so the exact impact is difficult to determine.” This is a fair concession, but it also means the extreme-value comparisons are partly circular, since the target is biased in the same variable being evaluated. Please state clearly, where extremes are discussed, what can and cannot be concluded from agreement with a product known to underestimate them, and avoid language implying validation against ground truth in those sections.

We have added more discussion of this limitation throughout the manuscript and also included some guidance on model use which may help with these concerns.

6. The precipitation-parameter fitting needs to be documented well enough to reproduce. In lines 235–242 the mixed-exponential parameters are fit with SciPy, Nelder–Mead first and Powell on failure. Please state the objective function minimized, what constitutes “failure”, and what happens where both algorithms fail. The restriction of α to 0.5–0.9 (lines 241–242) is attributed to internal testing; please give the rationale and say whether it degrades the fit in particular regimes, especially very dry or very wet cells. Since the model is meant to run automatically anywhere on land, the robustness of this fit is central, and the current description would not let an independent implementer reproduce it.

The corresponding text has been updated to better describe the exact fitting approach used and to further discuss the alpha restriction used. We definitely agree with the statement about reproducibility and have tried to rectify this (although in a worst case scenario the code is open source, so a user could determine it themselves).

7. Please clarify how seasonality enters the precipitation model. Figure 2 shows monthly occurrence and amount are captured well, but the method sections (lines 109–119 for occurrence, 120–145 for amount) do not state whether the transition probabilities and amount-distribution parameters are estimated by month, season, day of year, or over the whole year. This is distinct from the stationarity discussion in Sect. 3.2 and should be stated explicitly for reproducibility.

This has been clarified throughout the manuscript. Particularly, text has been added to sections 2 and 3 to make it clear that these calculations are done on a monthly basis.

8. The spatial-correlation procedure needs more specifics. Sect. 3.4 states that the random numbers used for precipitation occurrence and amounts are calculated with the exact same process (line 244), but the description then focuses on occurrence and later states that “the precipitation occurrence simulation can be handled the same way” with the key difference that correlations can only be calculated when both points are wet (lines 266–267). I assume the latter sentence should refer to precipitation amounts rather than occurrence. Please clarify this. Line 257 also states that generated correlations are smaller than observed; a one-sentence explanation of why would help the non-specialist reader the paper targets. For reproducibility, please report the convergence factors η used (lines 263–266), the sensitivity to them, and the exact small positive value substituted for negative eigenvalues.

We have corrected the line identified that said occurrence instead of amount. We have added an explanation for the difference in correlation values. Likewise, we have also added several additional details for reproducibility, including eta values.

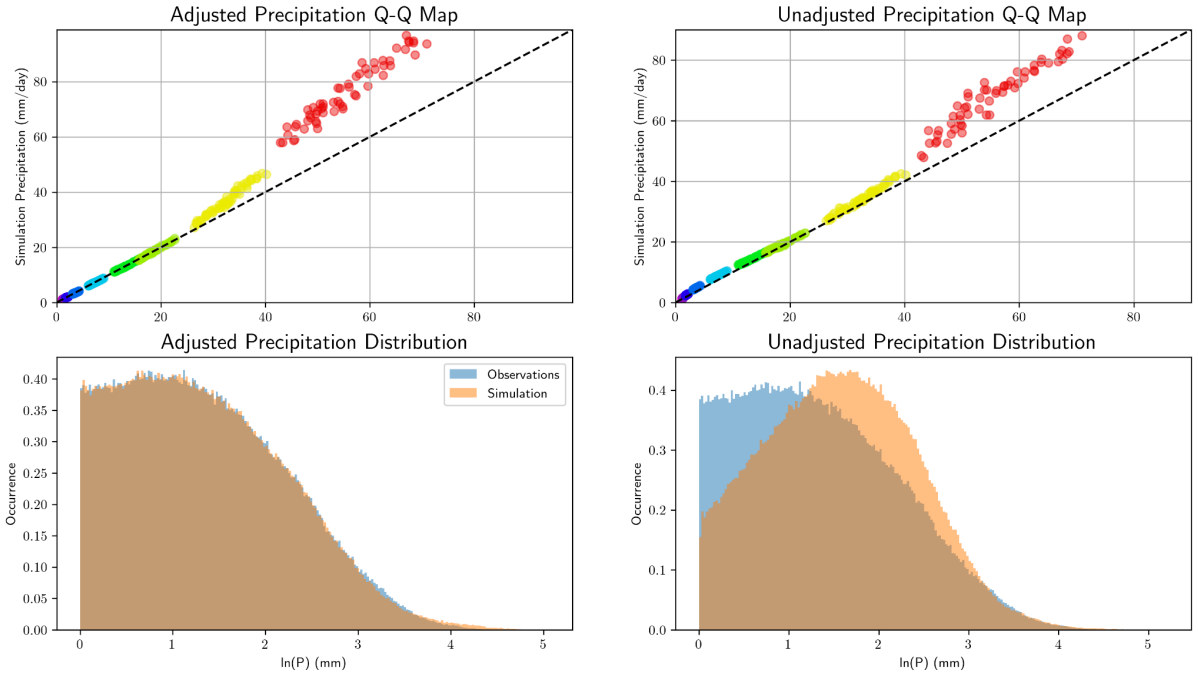
9. The temperature conditioning (Sect. 2.3) is the most distinctive component and the hardest to reproduce from the text. The wet/dry daily Fourier series and their use to condition the hourly series (eqs. 10–12, lines 182–194) are dense, and the notation shifts between ¯MH, ˜μH, and ˜σH without a fully worked statement of how the conditioned hourly mean and standard deviation are assembled from the unconditioned and conditioned series. Since this hourly, dewpoint-inclusive framework is central to the paper's claimed contribution, please add a clearer step-by-step description, or short pseudocode, sufficient to reproduce the temperature model and not only the precipitation model.

We have tidied up this section in general. This includes adding a sentence distinguishing between kinds of mean functions. We have also added more detail to the conditioning section to make it clearer how the conditioning takes place. Finally, we have added a step-by-step summary overview as requested.

10. The post-simulation adjustments (Sect. 3.7) and their contribution to skill should be clarified. The precipitation adjustment, based on a cubic spline fit to observed-minus-simulated differences at set percentiles (lines 287–298), reads as a coarse quantile-mapping or bias-correction step. If so, please say so directly. More importantly, are the Sect. 4 validation figures produced with these adjustments on or off? If on, please quantify how much of the agreement with ERA5-Land comes from the generator itself versus the correction. This is essential in the high-precipitation tail, where the largest errors occur and where a bias-correction step would do most of the work.

On this suggestion, we have now included some discussion and a figure in the supplementary (also below) which compares a simulation run with no adjustments to one with adjustments. Interestingly, the impact appears to much more notable on the smaller values of the precipitation occurrence than at the extreme end. In general, it is clear that the most notable impact of this correction happens below 8 mm/day. Although, the size of these changes is partially misleading as the log scale of the distribution plot means that smaller changes in magnitude will appear more significant if the existing magnitude was small.

We have also added text clarifying, as you suggested that we view this process as a coarse quantile mapping approach and clarified the role this bias correction approach has in general.



11. The validation regions should be justified, or the generality claims tempered. The model is described as able to simulate any terrestrial region (lines 52–75, 505–512), but the five test regions are three in New Zealand plus one each in Australia and New Guinea, i.e. concentrated in Australasia and the western Pacific. The wet/dry contrast is useful, but this set does not exercise continental midlatitudes, high-latitude or snow-dominated climates, or arid subtropics. Please either add a small number of contrasting regions or state clearly that the present validation is preliminary and regional, and soften the corresponding claims. Given comment 3, at least one additional tropical or monsoon region would be particularly valuable.

To address this issue we have expanded our analysis by including new simulation outputs from 4 sub-regions within the New Guinea region, 5 subregions from different climates over North America and 5 subregions over different climates within the Indian Subcontinent. This expands our analysis to a much wider range of climates and allows us to draw more defined conclusions from the results presented. In particular, we see some exceptional performance from the North American domains while mixed performance from the Indian subcontinent ones. When combined with the partitioned New Guinea data this provides much stronger evidence that the monsoon climate is the driving force behind the aforementioned issues. We have also added a new section to the paper which discusses these results and develops this interpretation.

12. The spell and extreme evaluations should be made quantitative. Figure 4 (spells) and the extreme percentiles in Figure 5 are discussed mainly by eye (lines 350–375). Since spell persistence is a direct product of the Markov chain, please report quantitative agreement, e.g. biases in mean and upper-percentile spell length, or a distance metric between observed and simulated spell-length distributions. For the extremes, the manuscript suggests a “more aggressive” adjustment could help but does not say what it would be. Please specify the intended approach and discuss the extremes-versus-overfitting trade-off concretely.

We have added some quantitative measures to figure 4 and altered the surrounding text to clarify the interpretations. In particular, we added some percentile values and the sum of absolute values. As suggested this clarifies our interpretation of these figures. We explored several additional metrics that could be used for figure 5 (and figure 9). However, we didn't find any gains in interpretation with the approaches we tried so have left this out of the paper update. We have also added some details around the more aggressive adjustment suggest, although to be clear, we think the downsides outweigh the upsides, so have no intention to implement.

13. The portability and low-resource claim needs substantiating. Portability and low resource use are presented as core contributions, but the evidence in Table S1 is a single laptop and a single cluster over the three New Zealand regions, with model-building already taking several hours on the laptop. Please indicate how runtime and memory scale with domain size, and give at least an indication of feasibility for a large domain, since that is the use case the "global" framing implies.

We have added detail throughout on portability and resources, particularly in the conclusions. This includes some discussions of domain size issues.

14. The limitations should be connected to concrete usage guidance. The limitations paragraph (lines 525–533) is useful. Please turn it into a short, practical statement of recommended and not-recommended uses: for instance, whether EGGS-WG is suitable for studying precipitation extremes at all, and for which variables or climates users should be cautious. Given the extreme precipitation and tropical results, this guidance would help potential users considerably.

We have added some details as suggested, and we have expanded our discussion on user guidance and usage issues.

Technical corrections and smaller comments

1. Lines 9–10: "Although, there are some clear issues. . . " is awkward; e.g. "However, some clear issues remain in the simulation of the most extreme precipitation values".

This has been updated as suggested

2. Lines 19–20: "fundamentally limited in their capabilities" – give one or two concrete examples, such as physical consistency, feedbacks, teleconnections, or response to nonstationary forcing.

This section has been reworked to include an example

3. Line 40: "otherwise limited" is vague; specify the dimension, for example coverage, cost, variables, resolution, or transferability.

I have added some text clarifying my intended meaning

4. Lines 51–52: “benefits over traditional methods are relatively minor” – minor in what sense, and for which applications?

We have refined our wording here. Hopefully the point is more clear now.

5. Line 54: “Stochatic” → “Stochastic”.

This typo has been fixed

6. Line 75: “with many of the older models simulation daily minimum and maximum temperatures” should be rephrased.

This has been reworked to make sense

7. Eq. (1) and line 129: add a comma after the equation, and “Where” → “where”.

The grammar for all equations has been checked and updated. Thank you for this comment, I was completely unaware this is the correct grammatical structure until this spurred some investigation

8. Lines 144–145: “to ensure that the output the model produces is spatially coherent output” – rephrase.

This has been reworked to make sense

9. Line 207: future versions might also consider surface heat fluxes alongside winds and radiation.

Great suggestion. I have added a sentence referencing this and other potential additions.

10. Line 288: “There are many reasons why a user would not want these adjustments” – give a couple of examples.

Additional detail has been added

11. Line 293: “course scale” → “coarse scale”.

This typo has been fixed

12. Line 368: “This is deviation is much more pronounced” → “This deviation is much more pronounced”.

This typo has been fixed

13. Line 409: “sperate” → “separate”.

This typo has been fixed

14. Lines 473–490: recurring misspellings should be corrected, including “sperately”, “equivilent”, “occurence”, “agregated”, “indentified”, and “emsemble”. Also check “the the” at line 482.

These have been fixed throughout. Thanks to this comment I realized my spell checker had broken....

15. Line 500: “unrelated o each other” → “unrelated to each other”.

This typo has been fixed

16. Line 545: “publically” → “publicly”.

This typo has been fixed

17. Several supplement captions refer to “all three simulation domains” when five are used, for example Figs. S2 and S5–S8. Please make the count consistent

This has been fixed

18. Some figure captions should be more self-contained: Figure 4 should state how frequencies are normalized, and Figure 5 should give precipitation units on the axes.

The text of several figure captions has been updated to add clarity on the data processing and make them interpretable. Figures have also been changed as suggested.

Reviewer 2:

Summary:

This study introduces EGGS-WG, a new open-source stochastic weather generator based on the ERA5-Land dataset. The model can simulate weather over gridded land regions worldwide, extending previous approaches that were mainly designed for individual sites. It also provides hourly temperature simulations and includes dew point temperature and relative humidity. Validation across five regions with different climates shows that the model generally reproduces regional averages and spatial patterns well. However, it has clear limitations, including systematic biases in extreme precipitation and weaker performance in representing tropical weather dynamics in New Guinea.

Overall, the study is innovative, useful, and clearly presented. I recommend minor revisions to address the following comments:

Thank you for the positive comments. We have tried to address your comments as thoroughly as possible. In conjunction with comments from the other reviewer, we have also substantially improved the discussion around existing limitations.

Introduction section:

1. Line 80: “The full code base for EGGS-WG are available on a….” - Please use “is” instead of “Are.”

We have updated the line as suggested

2. Line 70: The Introduction states, “if our model fails to accurately handle extremes,” suggesting that this is only a possibility. However, the Abstract and Results already show that the model does not accurately simulate the most extreme precipitation values and tends to overestimate them in most regions. Therefore, the conditional wording is misleading. This limitation should be presented as a known finding that the study aims to address, rather than as a hypothetical Issue.

Wording has been adjusted in the introduction to make it clear this issue was identified. We have also altered some of the conclusions to reflect this point.

3. The Introduction presents the model’s global land coverage as a major advantage, stating that the parameters are “always solvable.” However, this claim should be qualified because the model’s performance is not consistent across all regions. The Results section reports significant biases in the New Guinea test region, including reduced variability and poor representation of spatial dynamics. These findings are important limitations for a model presented as a global tool and should be acknowledged in the Introduction.

Our wording here has been changed to reflect our intention that this sentence was specifically about being able to calculate parameters for the precipitation distribution. This holds true due to the estimate approaches used and range of precipitation distribution

shapes. The issues with data simulation of the New Guinea region are unrelated to the fitting of these precipitation parameters, so we hope the change in text makes this clear. There is a new section discussing the New Guinea Region that further explores this. We have also updated our late discussion of the fitting process as discussed in other comments. We have also added more discussion of limitations to the end of the introduction.

4. The Introduction presents the addition of dew point temperature as an improvement. However, it does not acknowledge that the validation results show a consistent underestimation of the cross-correlation between air temperature and dew point temperature. This limitation should be mentioned in the Introduction to provide a more balanced description of the model's performance.

Along with the above comment we have added some discussion of limitations to the introduction.

Model design section:

5. Line 105: "a single order chain has been show to fail to produce" - Please use "shown" instead of "show."

We have updated the line as suggested

Implementation section:

6. Line 205: "Expansion to other ERA5-Land variables... are planned"- use "is planned" instead of "are planned"

We have updated the line as suggested

7. Line 220: "This trend lines can then be inverted..." - Use "These trend lines" for multiple trend lines, or "This trend line" for a single trend line.

We have updated the line as suggested

Line 240: "Restricting α to a value between 0.5 and 0.9 has also shown to improve the results in our internal testing."- The authors state that restricting α to values between 0.5 and 0.9 improved the results in their internal testing. Was this testing conducted globally or only in selected regions? Please clarify the geographical scope of the testing and indicate whether this parameter range performs consistently across different climatic regions.

We agree that this was a currently under-described in the text. The constraint here is less complex than implied. Alpha must always be between 0 and 1 by definition. Because the precipitation function is a mixed exponential, constraining alpha to 0.5 to 1 will have no implications for the solution but does appear to improve the convergence of the solver. Restricting the value as less than 0.9 mostly ensure that the solver will not resolve to a

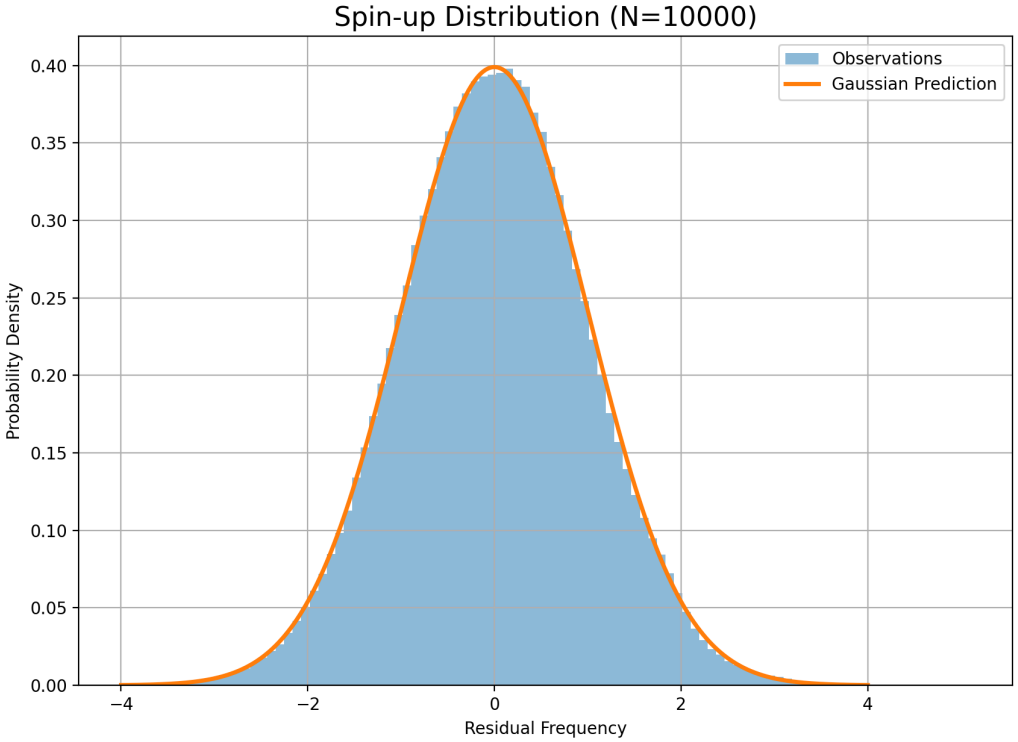
single exponential distribution. The corresponding text has been updated to clarify this as well as adding some details around the testing.

Line 295: “...adjustment is necessary for them to be adjusted”. - The phrase “for them to be adjusted” is redundant and should be removed or rephrased.

We have updated the line as suggested

Section 3.6: The model uses a fixed 1,000-day spin-up period to reach a random state, starting from completely dry conditions with temperature residuals set to zero. This simplified initialization may not represent the full range of realistic climatic starting conditions. The authors should justify this approach and discuss whether the results are sensitive to the chosen initial state and spin-up duration.

We agree that this is a sensible concern, so we have added the below figure to the supporting information and changed some corresponding text. This figure shows the distribution of the full set of temperature residuals from the end of the spin-up period from 10000 different spin-up simulations. What we see is that the starting residuals nearly perfectly follow a random Gaussian distribution. This means that from the perspective of the temperature fields, our starting data will be perfectly random. However, as the data has been run through the simulator, it should also have a sensible physical structure without unphysical spikes you might see with purely random data.



Results and discussion section:

Line 320: “...generally, the simulation appears to be able to capture to the mean regional properties...”. - Remove the extra “to” before “the mean regional properties.”

We have updated the line as suggested

Line 340: “The spatial dynamics are the very similar for the amount values...”. - Remove “the” before “very similar.”

We have updated the line as suggested

Line 355: “There appears to be some issues that with the New Guinea simulation for long wet streaks...”- Remove “that” and revise “appears” to “appear” to agree with the plural noun “issues.”

Text has been updated to fix the grammar

Line 460: “...maximum values t2m are well generally represented by the model...”. - Change “well generally represented” to “generally well represented.”

We have updated the line as suggested

Line 495: “...individual model runs successfully reproducing the variability within the observations...”. - Add “are” after “runs.”

We have updated the line as suggested

Line 500: “...individual ensemble runs are able to provide sufficiently variable simulations that are unrelated o each other” - Add the missing “t” in “to.”

Typo has been fixed

Line 345: The Canterbury results show that some grid-cell pairs have much lower simulated correlations than the observed values. The authors should clarify whether these outliers are geographically clustered, for example near mountainous areas, or whether they are likely due to random statistical variation.

This is a fantastic idea that the authors did not consider before receiving this comment. Some cursory exploration has not turned up anything obvious enough to be included in the paper at this stage. However, I am currently working with a student on this issue to see if anything extra can be determined using this approach. At this point, any useful result is likely to be such a significant deviation that it probably couldn't be included in this paper. However, if we find something useful, we will make the information publicly available (likely on the GitHub page for the project). This could also possible spur further coding improvements within the model.

Conclusions:

The authors describe the impact of the model's poor representation of extreme values as "minor" because such events are rare. This claim should be tempered. Stochastic weather generators are often used in hydrological risk assessment and flood modeling, where the accurate representation of extreme events is especially important. The systematic overestimation of extremes in most regions, and their underestimation in New Guinea, therefore represents a substantial limitation. The Conclusion should clearly advise users to exercise caution when applying the model to analyses involving extreme precipitation or related risks.

The conclusions have been reworked a bit in response to this comment and comments from the other reviewer. This includes tempering expectations where issues are identified and providing direct user guidance. This includes identifying guidance regarding extreme precipitation. Interestingly, simulations over some additional regions that have been included suggest that these biases are highly regional and not particularly consistent over different climates.