Review of "EGGS-WG: An open source global gridded stochastic weather generator derived from ERA5-Land"

This manuscript presents EGGS-WG, an open-source stochastic weather generator parameterized from ERA5-Land. The model produces daily precipitation occurrence and amount, hourly 2 m 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.

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.

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 P_L 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.

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.

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.

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.

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.

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.

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.

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 $\bar{M}_H$, $\tilde{\mu}_H$, and $\tilde{\sigma}_H$ 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, dew-point-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.

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.

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.

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.

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.

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.

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".
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.
3. Line 40: "otherwise limited" is vague; specify the dimension, for example coverage, cost, variables, resolution, or transferability.
4. Lines 51–52: "benefits over traditional methods are relatively minor" – minor in what sense, and for which applications?
5. Line 54: "Stochatic" → "Stochastic".
6. Line 75: "with many of the older models simulation daily minimum and maximum temperatures" should be rephrased.
7. Eq. (1) and line 129: add a comma after the equation, and "Where" → "where".
8. Lines 144–145: "to ensure that the output the model produces is spatially coherent output" – rephrase.
9. Line 207: future versions might also consider surface heat fluxes alongside winds and radiation.
10. Line 288: "There are many reasons why a user would not want these adjustments" – give a couple of examples.
11. Line 293: "course scale" → "coarse scale".
12. Line 368: "This is deviation is much more pronounced" → "This deviation is much more pronounced".
13. Line 409: "sperate" → "separate".
14. Lines 473–490: recurring misspellings should be corrected, including "sperately", "equivilent", "occurence", "agregated", "indentified", and "emsemble". Also check "the the" at line 482.
15. Line 500: "unrelated o each other" → "unrelated to each other".
16. Line 545: "publically" → "publicly".
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.

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.

Recommendation

Major revisions. The model and its open release are a worthwhile contribution, and I expect the paper to be publishable once the authors sharpen the statement of novelty, address the New Guinea and reanalysis-validation issues, document the fitting, temperature-conditioning, and spatial-correlation steps to a reproducible standard, and clarify the role of the post-simulation adjustments.

References

Groisman, P. Y., et al. (1999). Changes in the probability of heavy precipitation: Important indicators of climatic change. *Climatic Change, 42*, 243–283.

Martinez-Villalobos, C., and Neelin, J. D. (2019). Why do precipitation intensities tend to follow Gamma distributions? *Journal of the Atmospheric Sciences, 76*, 3611–3631.

Martinez-Villalobos, C., and Neelin, J. D. (2023). Regionally high risk increase for precipitation extreme events under global warming. *Scientific Reports, 13*, 5579.