

Response to Reviewer 2

Manuscript ID: egusphere-2026-2325

“unseen-awg v1.0: spatio-temporal weather generation using analogs and unseen data”

Jonathan Wider, Jakob Zscheischler

October 5, 2026

Below, we respond to the reviewer’s assessment of our manuscript. Their comments are shown in *gray italics*, our replies in regular text.

General comments

This manuscript introduces unseen-awg, an analog weather generator that uses ECMWF re-forecasts and block sampling to generate long daily weather series over Europe. The approach is innovative, relevant for climate-risk applications, and supported by openly available code and data.

The manuscript is generally well written and structured. The method is promising and could make a valuable contribution. However, the interpretation of unseen extremes, parameter choices, and validation procedures should be strengthened.

Response: We thank the reviewer for their positive assessment of our study, their thorough review of our manuscript, and their critical yet constructive feedback. Their review will help us improve our study considerably. Below, we reply to each point individually.

Major comments

1. The daily extremes generated by unseen-awg are unprecedented relative to ERA5 for 2003–2023, but they already exist in the reforecast archive. The authors should clarify this distinction and avoid implying that the weather generator creates entirely new daily states. They should also examine whether the most extreme values are associated with lead times or ensemble members and assess whether they could reflect forecast artifacts.

Response: We agree with the reviewer that, while the daily weather states used by our weather generator are “unseen” and potentially record-breaking with respect to observational data, they are simply resampled from an existing reforecast dataset. We will make this clearer in the revised manuscript.

We will conduct additional tests to assess whether the distribution of lead times during extreme events differs from that of non-extreme states included in the generated time series for the types of extremes we study (see response to comment 4 of reviewer 1). While this is not sufficient to exclude “forecast artifacts”, it can at least serve as a partial diagnostic.

Two effects make the lead-time distribution of the generated time series non-uniform:

1. Through sampling in blocks with $\tau = 30$ days, more states with medium lead times are included than states with very short or very long lead times, because block starts can vary but each block must be $\tau = 30$ days long.
2. For small t_{lead} , the forecast ensemble members are still statistically dependent, and their effective sample size is lower than for long lead times. Therefore, even in a situation without any biases, the chance that the nearest neighbor for a state at the end of a block (with large t_{lead}) will have small t_{lead} is reduced.

We do not consider either of these effects to be problematic per se, but acknowledge that, in combination with lead-time-dependent biases, they may skew the distribution of generated weather data. To make this aspect more transparent, we will include a histogram of lead times in the simulated time series in the Appendix of the revised manuscript and reference it in the main text.

A full test for forecast artifacts would require a comprehensive evaluation of the IFS extended-range system and is beyond the scope of our study. We will highlight the limitations that arise due to the underlying dataset more clearly in the manuscript.

2. The reforecast dataset is described as containing approximately 3,500 years of weather states, but these states are not independent because of overlapping initializations, lead times, and ensemble members. This may affect analog selection, extreme-event sampling, and return-level interpretation. The effective sample size should be investigated through sensitivity tests using subsets of initializations, lead times, or ensemble members.

Response: Effective sample size depends on predictability: it is smaller for more predictable variables and types of events (e.g. Kelder et al., 2020). If events are predictable, members from the same ensemble forecast will be more dependent; the same is true for forecasts from nearby initialization times. Therefore we deem a comprehensive analysis of effective sample size infeasible.

We already highlight the limitations due to these dependencies and the related risks in several places in the manuscript:

Original text (l. 267–268):

“Caution is advised when interpreting return levels [...]”

Original text (l. 83–84):

“While dependencies, especially for small forecast lead times and between consecutive initializations, clearly decrease the effective sample size, the dataset remains extensive.”

In the revised manuscript, we will expand on this by adding the following aspects:

- Small lead times should be excluded when trying to estimate return periods.
- Return levels for very long return periods may be unreliable. At one point, generating more and more data will not yield ever more extreme daily records, because the resampled dataset imposes a finite upper limit.

3. Please clarify whether the unweighted Euclidean is normalized, whether anomalies or raw $Z_{500\text{hPa}}$ fields are used, the choice not to use area weighting should also be justified because it gives greater influence to high-latitude grid cells.

Response: Strictly speaking we used (unweighted) root mean squared error (RMSE), not Euclidean distance. This was described incorrectly in the original manuscript and will be corrected in the revised manuscript. These metrics differ only by a constant factor that does not affect analog selection as long as σ is adjusted accordingly (e.g., by tuning it).

No normalization was used when computing distances between analogs and we work with absolute values rather than anomalies. The differences from working with anomalies should be limited: the windowed selection ensures that candidate states are from days of the year similar to that of the base state; therefore, the climatological contribution approximately cancels when computing the pattern distance. Generally, using absolute values rather than anomalies should lead to the

selection of analogs from more similar days of the year because climatological differences, which grow with calendar distance, add to the distance.

Although we previously used unweighted RMSE for simplicity, we will use area weights in the revised version.

4. The selection of $\tau=30$ days is based mainly on a qualitative compromise between autocorrelation and flexibility. Figure 3b only evaluates summer temperature at one grid cell. The authors should provide broader sensitivity analyses across variables, seasons, and regions.

Response: Mechanistically, we do not see any reason why the finding would not generalize: the main reason for underestimated autocorrelation lies in block transitions, where 1) analogs are similar but not identical to the true next state, and 2) similarity in circulation only partially enforces similarity in impact variables. If the autocorrelation in the reforecast data is unbiased in this respect, fewer block transitions should therefore always improve autocorrelation.

This is consistent with Figure 6 of the original manuscript, which shows that correlations between seasonal aggregates improve with increasing τ .

As a simple spatially resolved test, we will compare maps of autocorrelation at a fixed lag between ERA5_{2003–2023} and weather generator simulations with different values of τ . We will include plots for JJA and DJF.

5. The parameter σ is optimized for $\tau = 1$ but then applied to simulations using $\tau = 30$. The assumption that the optimal value transfers between these configurations should be tested. In addition, the forecast experiment benefits from a one-day “look-ahead” and uses a lead-zero reforecast state as ground truth. Its interpretation as an actual forecast experiment should therefore be clarified.

Response: Thank you for highlighting this important point, which was not sufficiently addressed in our original manuscript. We have already conducted some preliminary analyses that led to improvements to our weather generator.

We agree that transferring the value of σ between the two settings was not fully tested in the original manuscript. We also agree that these settings differ in several respects:

1. The initialization states in the forecast experiments come from small lead times, while for $\tau = 30$ d, the lead time of the base state during analog selection (at the end of a block) will always be large.
2. The candidate states are restricted to small lead times when using $\tau = 30$ d because they need to sit at the start of a valid 30-day block.
3. For the forecasts, we exclude candidates with t_{valid} later than the forecast initialization time.
4. For the generated time series, it is impossible to sample the true next state (because it would never be a valid start for a new block), while this is possible for the tuning analog forecasts.

We cannot easily test the full $\tau = 30$ d setting because the predictability of daily geopotential height vanishes with increasing lead time.

We will test three strategies:

1. We will try to make the tuning analog forecast setting as close as possible to the actual time series generation. This could include:
 - Excluding long lead times from the set of allowed candidates.

- Excluding only states valid between the initialization and target time of the analog forecast (plus a generous buffer), rather than all states valid after the initialization.
- Excluding transitions to “true” next states, as these are impossible in the $\tau = 30$ d setting.

Unlike in the full $\tau = 30$ d setting, we will not be able to start analog forecasts at long lead times because we need $t_{\text{lead}} = 0$ d at the start of each forecast.

2. The weather generator assigns a probability distribution over the candidate analogs whenever an analog is sampled (equation 2 in the original manuscript). As a measure of how uniform or concentrated this distribution is, we will compute its entropy. We will compare the entropies of these distributions in the tuning forecasts with those in our default $\tau = 30$ d. If the analog-forecast setting is representative of the long simulations, we would expect entropies during transitions to be similar as well.
3. We will use a more extensive set of combinations of τ and t_{forecast} for ranges where there is still analog forecast skill on daily scales for geopotential height.

While doing some exploratory analyses on this new entropy metric, we noticed a bug in our weather generator that caused the set of candidate states in autumn to be much smaller than expected. This was a consequence of an error in reorganizing the data to allow similarity computations to be vectorized, which excluded some data points in this season from the reorganized dataset. We will fix this issue in the revised version and rerun all analyses with the corrected weather generator.

We do not consider the “look-ahead” an issue that would preclude using tuning forecasts to determine parameters of the weather generator for long simulations. Of course, this look-ahead needs to be taken into account if one explicitly wants to use the generator as a forecast model and compare it with other forecast models. This is already mentioned in Appendix A3 of the original manuscript.

We are also considering including a configuration in which σ can vary with the day of the year.

6. The method is presented as generating stationary present-day climate, although the 2003–2023 period contains temperature trends. The authors should clarify how “present-day climate” is defined and discuss whether trends affect the generated extremes and dependencies.

Response: This is a valid point and a complication that is common to analog-based weather generators like ours and that of [Yiou \(2014\)](#). Compared to [Yiou \(2014\)](#), we have the advantage of covering only 21 years of climate rather than 65 years of NCEP reanalysis data.

Combining data from multiple years without accounting for trends could inflate the estimated variability and, in turn, the magnitude of anomalies.

Our analyses should be less susceptible to this: the effect on quantile exceedances is more indirect. The fact that we do regional rather than global analyses and that internal variability therefore plays a larger role also reduces the effect of “mixing” years with different global mean temperatures.

We will discuss this point more clearly as a limitation of our approach and adapt our phrasing from “present-day climate” to “2003–2023 climate”.

Minor comments

- | Line 39: “As a consequence, also local weather” → “As a consequence, local weather is also”.
- | Line 50: “An alternative are reforecasts” → “An alternative is the use of reforecasts”.
- | Line 79: Replace “0th to 45th forecast day” with “forecast days 0–45”.
- | Line 237: “coniguous” → “contiguous”.
- | Line 246: “climate-sensitive system” → “climate-sensitive systems”.
- | Line 253: “with dimensionality” → “with the dimensionality”.
- | Line 270: “may be never be generated” → “may never be generated”.
- | Line 279: “size of reforecast dataset” → “size of the reforecast dataset”.
- | Line 304: “climate impacts models” → “climate-impact models”.
- | Line 452: “ECMWFs” → “ECMWF’s”.
- | Line 468: “AI assisted” → “AI-assisted”.

Response: Thank you for pointing out these errors and inconsistencies. We will fix them in the revised manuscript.

References

- Kelder, T., Müller, M., Slater, L. J., Marjoribanks, T. I., Wilby, R. L., Prudhomme, C., Bohlinger, P., Ferranti, L., and Nipen, T.: Using UNSEEN Trends to Detect Decadal Changes in 100-Year Precipitation Extremes, *npj Climate and Atmospheric Science*, 3, 47, <https://doi.org/10.1038/s41612-020-00149-4>, 2020.
- Yiou, P.: AnaWEGE: A Weather Generator Based on Analogues of Atmospheric Circulation, *Geoscientific Model Development*, 7, 531–543, <https://doi.org/10.5194/gmd-7-531-2014>, 2014.