Reply to referee #1

A) GENERAL COMMENTS

The manuscript represents a weather-generator-type algorithm for simulating extreme events like heatwaves. Using so-called empirical importance sampling in generating the daily sequences of weather, the algorithm can be tailored to produce extremely warm summers much (presumably, several orders of magnitude) more commonly than they would occur in the real world or without the importance sampling. The algorithm is generic and should therefore also be applicable to other types of “long” extreme events such as extended periods of heavy precipitation.
Importance sampling has previously been applied to simulations by dynamical atmospheric models (e.g., Ragone et al. (2017) cited in the manuscript). Compared with this, the use of a stochastic weather generator requires much less computing time, and therefore potentially allows a much larger number of simulations. However, I have a major scientific concern about the fidelity of the method. If the motivation is anything else than to find the extremes of the summer (June-July-August) mean temperature and the associated summer mean atmospheric circulation, the realism of the daily time series also matters. However, as shown by Fig. 5, the warmest simulated summers have no seasonality at all, with equally high temperatures occurring from the beginning of June to the end of August. This appears quite unrealistic, since very high temperatures are much less probable (if possible at all) in the beginning and the end of the summer than in July or early August. The problem likely results from the fact that the method has no strict constraint on the time of year of the circulation analogies, and the warmest trajectories therefore sample days from the height of the summer even in the beginning and the end of the summer. This could be mitigated, though at the cost of reduced sampling space, by only accepting analogue days from a moving window of (e.g.) +/- 15 calendar days around the target day. Alternatively, it might be possible to work with anomalies (removing the mean seasonal cycle before applying the algorithm) or with normalized anomalies of Z500 and T (removing the mean seasonal cycle and dividing by standard deviation), although this is uncharted terrain that might create its own problems.

Referee#2 had a similar comment. We thank both of them for their careful reviews. The weight that is devoted to the calendar day was indeed too weak. We were deceived by the experiments of Ragone et al. (2017) of perpetual summer. We now use a “stiffer nudging” for calendar days, and obtain more realistic summers (i.e., summers that eventually end into a Fall season). The procedure is explained below.

B) SPECIFIC COMMENTS

1. P5L1. Does this mean the best 20 analogues regardless of the time of the year (cf. general comments)?
Yes, indeed. Then, they are weighed by their calendar day.

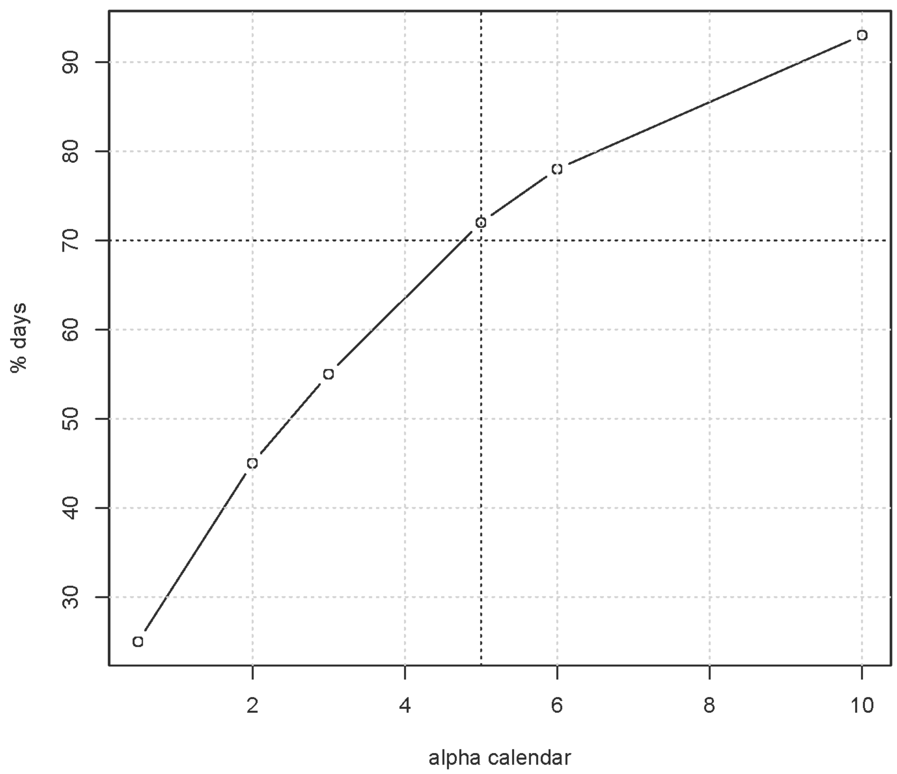
2. P5L18-19 and 26-27. I suppose the weight is directly, not inversely

proportional to the correlation, i.e., days with higher correlation are more likely chosen as analogues.
This is true. Please note that this condition is not used in the importance sampling SWG. This is clarified in the revised text.

3. P5L28-30. Does this procedure lead to a realistic autocorrelation of the daily mean temperatures?
This was checked in the original paper of Yiou (GMD, 2014, Figure 4). The analog stochastic weather generators generally underestimate temporal auto-correlation. “Real” temperature is not as random as a stochastic weather generator.

4. P7L10. I assume that this is an addition to the factors 1-2 on P5L24-27, not a replacement.
Thank you for this remark. This rule actually replaces the weight on the spatial correlation, and the weights on the calendar days are maintained. This is clarified in the revised manuscript.

5. P7L16-18. Is physical relevance really ensured? See the general comment about the lack of seasonality.
This is indeed a crucial point (see reply to general comment). This problem was treated by increasing the weight on the calendar days. This “calendar nudging” parameter was estimated by trial and error, by taking the smallest value for which most (e.g. more than 70%) of the dynamic simulations end with dates after the second half of August. In our case (summer temperature simulations), a parameter value of 5 is deemed reasonable for summer temperature simulations. All the figures are changed accordingly. This will be explained in the revised text. Note that if this criterion is used (more than 70% of simulations end towards the end of the season), then the value of this weight parameter might be different for another season. This will be explained in the revised manuscript, with an additional figure (below) showing the dependence to the calendar weight of the percentage of simulations for which the last day is within the last two weeks of the summer.



6. P9L26-27. The only way in which this paper currently answers this question ("how likely is the occurrence of an event") is by fitting a normal distribution to the observed JJA mean temperatures (right scale in Figure 4). Would it be possible to refine these estimates based on the SWG approach (cf. Figure 4 in Ragone et al. (2017), cited in the manuscript)?

This is a good question on the theory. One can give a heuristic formula for the probability of trajectories from the analogues. For example, let A be the smallest number (of analogues) for which $\frac{1}{S}\sum_{k=1}^{A}\exp(-\alpha_T k)>1-\epsilon$, where ϵ is a small positive number (for example 1/Number of simulations that are needed to observe an event) and S is the sum of all exponential weights.Then the probability of trajectories is close to $\left(\frac{A}{K}\right)^M$, where M is the number of independent days in the season (M=18 for a season of 90 days) and K is the number of analogues (K=20). In this case, we verified that such an estimate is close to the Gaussian quantiles, for values of α_T that are not too large (< 1), if K=20 analogues are considered. Such a formulation needs several parameter adjustments (e.g., ϵ and M). A rigorous formulation would be a statistics paper in itself. In the case of seasonal temperature averages, it is probably wiser to stick to the Gaussian approximation, which does not require too many trials. This will be mentioned in the revised manuscript.

7. P12L22-24. This interpretation "small perturbations of the atmospheric Z500 structures can add≈4K..." is dubious, because there is no one-to-one physical relationship between Z500 and surface temperature. Much more likely, the 4 K addition in temperature comes from the tendency of the algorithm to select the warmest days among days with similar Z500 fields. The slight changes in the Z500 anomalies are a side effect of this, but they are not large enough to "cause" the change in surface temperature. Note that, for the whole

atmospheric column to be 4 K warmer, the layer between 500 and 1000 hPa should become 80 meters thicker.
Point taken. What we meant was that a small modification of Z500 patterns (a few meters at most) can be associated to large surface temperature changes (4K, i.e. larger changes than the expected isostatic change). This is larger than the temperature trend (~0.1K/decade). This is clarified in the revised manuscript.

8. P12L27-29. How long does the simulation remember its initial conditions? Would there still be a difference in the Z500 fields if they were only averaged over July-August?
The “regular” stochastic weather generator remembers the initial conditions (for temperature) up to one month ahead (Yiou and Déandréis, GMD, 2019). Starting all simulations in June, the SLP/Z500 fields averaged over July-August (rather than JJA) do quantitatively change, although the patterns are still anticyclonic over Europe (see figures below).

C) TECHNICAL COMMENTS AND CORRECTIONS

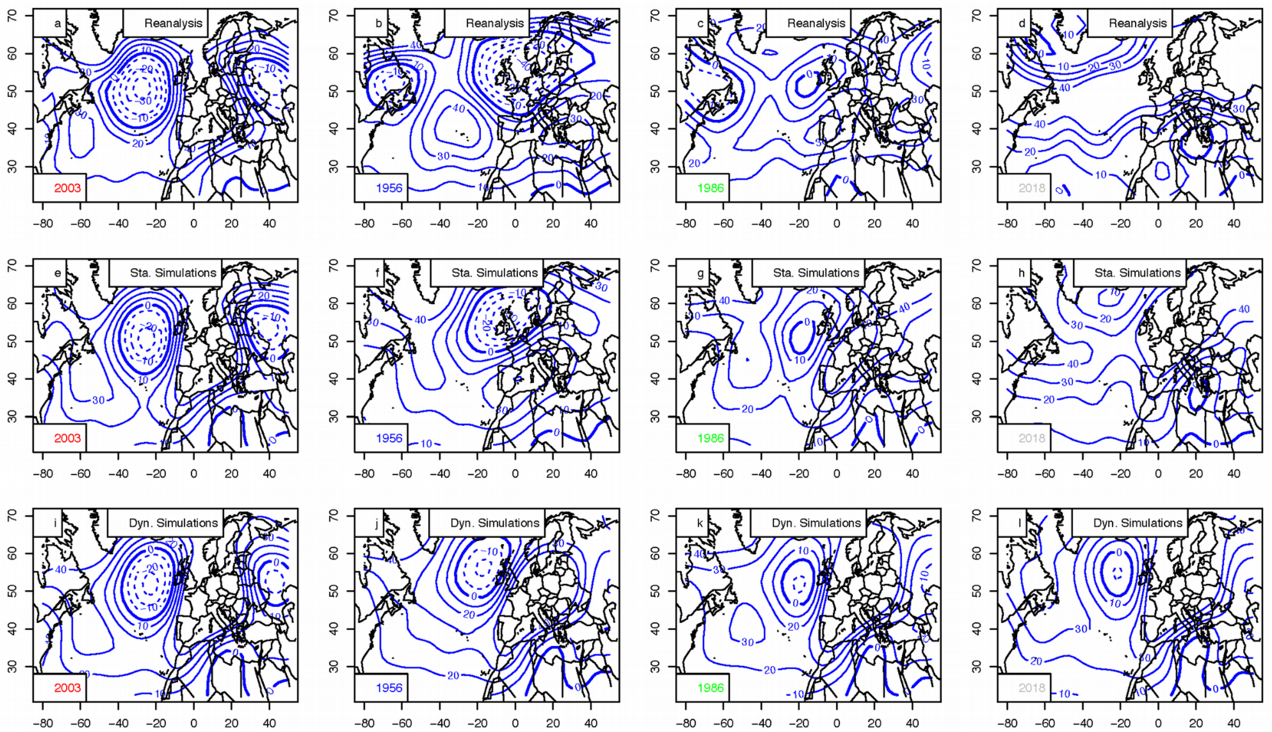
1. P1L19: ensembles in plural
OK.

2. P2L26: linked to
OK.

3. P3L1: recalls
OK.

4. Figure 3 could be improved by including the values of the alpha parameter in the figure panels. In addition, it would be useful to describe the interpretation of the box plots in the caption (there are several versions around, although some are more common than others).
The caption for boxplot figures will be completed in the revision (see comment of referee#2).

5. Figures 6, 7, A2, A4 and A6. These maps could be improved by using different colors for positive and negative anomalies. It would also be better to use the same contour interval in all panels of each figure.
The contour lines use the same interval increments. The lines will be continuous for positive anomalies, dashed for negative anomalies, and thick for 0. This facilitates reading the figure on a B&W printout. An example of maps is given below (Z500 anomalies):


a
Reanalysis
2003
b
Reanalysis
1956
c
Reanalysis
1986
d
Reanalysis
2018
e
Sta. Simulations
2003
f
Sta. Simulations
1956
g
Sta. Simulations
1986
h
Sta. Simulations
2018
i
Dyn. Simulations
2003
j
Dyn. Simulations
1956
k
Dyn. Simulations
1986
l
Dyn. Simulations
2018