This paper presents analyses the performance of three stochastic weather generators based on circulation analogs to simulate daily temperature and precipitation over the Aare river basin (Switzerland). The paper is overall interesting (comparison of three models of increasing complexity) and clearly written.

We thank the reviewer for this positive feedback.

Yet, I think that the experimental set up could be improved and some discussions do not seem to be supported by the results or the figures. Therefore, I feel that there is ample room for improvement of the manuscript to optimize its impact.

We appreciate these comments which will help us to improve the manuscript. We hope that the proposed modifications will answer your questions.

C1: I do not think that stochastic weather generators (especially those based on analogs) are efficient or even useful to simulate long (i.e. multi-annual) sequences of climate variables, because they cannot take low-frequency variability (due to the ocean or global warming) into account. Instead, they can be very useful to simulate very large ensembles of short sequences in a stationary climate. The manuscript never compares long term variability of model simulations and observations, but focuses on seasonal probability distributions. Therefore, the introduction and interpretation should focus on the challenge of reproducing the probability distribution of climate variables, rather than a centennial reconstruction that is not even analyzed. This would also be more relevant for potential users (as claimed in the abstract and introduction), and would make room for comparisons of probability distributions (past vs. present vs. future).

Thank you for those comments. We partly agree with the reviewer's statement. Yes, the majority of stochastic weather generators (WGEN) are not able to simulate the low – frequency variability of the weather. A number of papers have shown that they cannot simulate a relevant interannual variability of precipitation. The case of WGENs based on analogs is different. By construction, they are conditioned by a sequence of large-scale circulation patterns which presents variability at multiple scales, from daily to interannual and even multi-decadal scales. Thus, WGENs based on analogs are also able to generate long sequences (multiple decades) of weather with relevant multi-annual variability as it derives from the one contained in the large-scale forcing data available for the time period considered for the generation. This is obviously a strength of such WGENs and we will correct our manuscript to better emphasize this point. In the SCAMP+ version, note that the additional step which generates non-observed sequences of large-scale circulation, also generates other realizations of low-frequency variations**.**

In the current manuscript, we only present the mean annual cycle of precipitation and the seasonal probability distributions (for observations and simulations). We agree that this does not allow appreciating the ability of ANALOG and SCAMP generators to produce a centennial reconstruction of regional weather (the reconstruction that can be achieved when the generators are forced by reanalyzes). In addition, the reader cannot appreciate the ability of the models to generate relevant multi-annual variability (for the reconstructed periods for ANALOG and SCAMP, or for long periods obtained from resampled large-scale centennial circulation sequences for SCAMP+). This will be clarified in the revised manuscript. We will add figures that show the time series of observed and generated annual variables (e.g. annual mean, seasonal precipitation) over the last century and over different 100-year resampled large-scale circulation sequences. We will also present how the additional SCAMP+ generation step influences the low -frequency variability of the generated scenarios.

C2: Does the comparison of seasonal precipitations (Fig. 6) depend on choices of predictors to compute analogues, or even how the seasonality is taken into account?

In a preliminary work (not shown in the manuscript), we have considered many different versions of each WGEN based on different sets of predictors respectively. The set of predictors considered in our manuscript has been selected so as to maximize the skill of the WGEN for the prediction of the daily precipitation and temperature observations over the whole simulation period (Chardon et al. 2016, Raynaud et al. 2018). The skill is estimated with the Continuous Ranked Probability Skill Score (CRPSS), a probabilistic evaluation score typically used for the verification of probabilistic or ensemble weather forecasts. This will be clarified in the revised manuscript.

These preliminary analyses showed that the results of the generations depend on the choice of predictors used for the analog selection. This could have been presented in our manuscript. However, results obtained with other predictor sets are not really relevant to consider. The lower skill of these other sets for the prediction of daily variables directly translates to a lower skill for the reproduction of observed seasonal probability distributions. A comment will be added in the revised manuscript.

For the second part of the question, the seasonality is accounted for in different ways:

1. As indicated in Section 3.3.1of the current manuscript, the large-scale predictors are likely to differ from one season to the other. In our work, the first level analogy variables used to identify the candidate analog days are the same but the second level analogy variables differ according to the season. From September to May they are the vertical velocities at 600 hPa and the large scale temperature at 2 meters. In summer, the vertical velocities but also other predictors such as the Convective Available Potential Energy (CAPE) led to a rather poor prediction of precipitation due to the coarse resolution of the atmospheric reanalysis that prevent it from providing an accurate simulation of convective processes. Consequently, large scale precipitation from the reanalysis has been used instead, resulting in predictive skills similar to the ones obtained for the rest of the year.

2. The large-scale / small scale downscaling relationship is likely to differ from one season to the other. To account for this, the candidate analogs are identified within a 2 months calendar moving window centered on the target day (day of simulation). For instance, when the current simulation day is a 6th june, all days between the 6th of May and the 6th of July of each year are considered as candidate analogs. This calendar constraint for the selection of candidates was not indicated in the manuscript and this point will be added in the revised manuscript.

3. A last calendar constraint is used for the first step SCAMP+ (generation of large scale circulation sequences). This constraint is given at l. 268-270 of the present manuscript version: “To insure that two consecutive days of the generated sequences belong to the appropriate season, the five 2-day analogue sequences are identified within a +/-15-day moving window centred on the calendar day of the target simulation day”.

C3: I am surprised that the discussion of the results is so qualitative: the authors show boxplots or return value plots that yield rather small changes, but never compute actual scores of performance that would quantify the performance of the simulations. Continuous Rank Probability Scores (CRPS) or Tallagrand diagrams (or just quantile plots) would be more useful than a subjective appreciation of Fig. 7.

We could present the CRPSS or Tallagrand diagrams obtained for the ANALOG and the SCAMP models. Both models are indeed expected to reproduce the time variations of observed precipitation. This is not the case for SCAMP+. SCAMP+ produces its own trajectories of large-scale

variables. These trajectories are by construction different from the observed one. As a result, the time series of weather variables generated with SCAMP+ are not expected to fit the observed ones. Note in addition that the main interest of SCAMP+ is that is allows to better explore the diversity of weather configurations and sequences. This is highlighted by the larger range obtained with SCAMP+ for different weather characteristics. A quantitative assessment of SCAMP+ via its ability to reproduce the observed sequence of some variables is thus not really relevant nor interesting.

C4: I see no discussion of uncertainties of the results (e.g. with respect to model parameters).

It is true that a discussion on this issue should be incorporated. Among the model parameters that can have an impact on the results, we can stress the importance of:
1. the set of predictors used in the selection of analogs.
2. the number of analogs selected as potential candidates (100 for the first level of analogy and 30 for the second level).
3.the transition probability p between large scale trajectory in the first-generation step of SCAMP (p=1/7 in the manuscript).

C5: My bet for the strange performance of SCAMP+ to simulate a reasonable range of summer temperatures is that summer temperature follow a distribution that depends on the mean state (e.g. Parey, S., Dacunha-Castelle, D., & Hoang, T. H. (2010). Mean and variance evolutions of the hot and cold temperatures in Europe. Climate dynamics, 34(2-3), 345-359.). Just perturbing with a Gaussian distribution with a fixed variance lowers the variance, with respect to the true temperature variance.

Thanks for this comment. As discussed at l. 420-428 of the current manuscript, the limitations of SCAMP+ concerning the generation of hot summers and cold winters are very likely related to the temperature increase experienced over the 20th century, which appears clearly when looking at the hottest summers and the coldest winters. Additional experiments will be performed in order to verify this assumption. In details, we will detrend observed temperatures using a regional linear long-term trend, as done in Evin et al. (2018). We will then redo all the analyses on these detrended temperature observations. We expect this pre-processing to solve this particular issue.

Evin, Guillaume, Anne-Catherine Favre, and Benoit Hingray. 2018. "Stochastic Generators of Multi-Site Daily Temperature: Comparison of Performances in Various Applications." *Theoretical and Applied Climatology*, February, 1–14. https://doi.org/10.1007/s00704-018-2404-x.

C6: My notions of Alpine geography are rather limited. Indications of longitude and latitude in Fig. 1 would be useful.

Longitudes and latitudes will be added in Fig. 1.

C7: Using geopotential heights for analogs is certainly a good idea, but the authors should be aware of long term trends (due to temperature increase), which induce biases in analog computations, especially in ERA20C. The authors could consider removing such a trend.

This issue is indeed potentially critical. We use geopotential heights for the first analogy level in the analog selection. The Teweles–Wobus score (TWS) proposed by Teweles and Wobus (1954) is used there. This score has been found to lead to higher performances than a more classical Euclidian or Malahanobis distance (Kendall et al. 1983; Guilbault et Obled, 1998; Wetterhall et al., 2005). It quantifies the similarity between two geopotential fields comparing their spatial gradients. It allows

selecting dates that have the most similar spatial patterns in terms of atmospheric circulation at a given (or several) geopotential level(s). As a consequence, it does not compare the absolute values of the geopotential fields between 2 days. We are aware that the mean value of geopotential fields is expected to change with regional warming. The Teweless-Wobus has the great advantage to remove the influence of this long term trend, and should therefore avoid biases for the analog identification. A comment will be added on this point.

Kendall, M., Stuart, A., Ord, J.K., 1983. The Advanced Theory of Statistics. Design and Analysis, and Time-series, vol. 3. Oxford Univ Press, New York. 780 p.

Teweles J, Wobus H. 1954. Verification of prognosis charts. Bulletin of the American Meteorological Society 35: 2599–2617.

Guilbaud S, Obled C. 1998. Prévision quantitative des précipitations journalières par une technique de recherche de journées antérieures analogues: optimisation du critère d'analogie (Daily quantitative precipitation forecast by an analogue technique: optimisation of the analogy criterion). Comptes Rendus de l'Académie des Sciences – Series IIA, Earth and Planetary Science Letters 327: 181–188. doi:10.1016/S1251-8050(98)80006-2.

Wetterhall F, Halldin S, Xu CY. 2005. Statistical precipitation downscaling in central Sweden with the analogue method. Journal of Hydrology 306: 174–190. doi:10.1016/j.jhydrol.2004.09.008.

C8: The authors compare (with two different visualizations) 1 day, 3 days, 5 days (Fig. 8) and 92 days (Fig. 7a) precipitation values. What is the cut-off duration for which the three weather generators give similar results (Fig. 7a)? If a generalized Pareto distribution was fitted to precipitation, would the ANALOGUE or SCAMP weather generators be within confidence intervals?

We thank the reviewer for this comment. There are two different aspects. First, in Fig.7, we assess some features of the **climatology**, i.e. the mean of the precipitation amounts at the seasonal scale. We also present the seasonality and precipitation values at the monthly scale in Fig. 6. We could also present the same results at the weekly or daily scale, but it does not present so much interest since it will be similar scaled results (i.e. the monthly mean is equal to the daily mean times the number of days in the month). Second, in Fig. 8, we have a look at the features of **extreme values** for different durations. It does make so much sense to assess annual maxima at the monthly scale (i.e. the maxima of 12 values) since they are not "extremes" in the sense of the extreme value theory (i.e. the maxima of samples of infinite size). We will however consider the interest of including results for 10days maxima which are also of relevance for hydrometeorological extremes in the considered catchment. For the last question, we do not know if the confidence interval obtained from a GPD fitted to precipitation observations would match the intervals obtained from the ANALOGUE and SCAMP simulations, but it can be investigated and discussed.