Review (1) of « Assessment of stochastic weather forecast of precipitation near European cities, based on analogs of circulation » by Krouma et al.

We thank the reviewer for the positive and constructive comments, which we considered in the new version of our manuscript.

General comments :
The article assesses the skill of a stochastic weather generator to forecast precipitation in 4 cities of Western Europe. The SWG is based on random sampling of analogs of geopotential height. It was developed in another article by Yiou and Déandréis 2019, where it was applied to temperature. This study complements the latter for precipitation. As a refinement, a time embedding of 4 days is considered in the distance for the search of analogs (however the considered distance is not a mathematical distance anymore - see below). Skill scores are evaluated for lead times of 5 to 20 days. Results show positive skills up to 10 days. A comparison to ECMWF forecasts is provided but I have some concerns about this part (see below). The study is interesting, clear and well written. Precipitation forecasting is an important subject of research and I support the idea basing SWG on analogs. However I have several main concerns :
- The results are mainly shown for NCEP which used to cover a longer period than ERA5. However ERA5 is now available since 1950. Given its much better resolution, I recommend considering ERA5 for all the results.
⇒ We extended the search of analogs to 1950 using the ERA5 database. We showed in section 4.1 that there is no difference between NCEP and ERA5 from 1950. And we decided to show results with NCEP in the paper.
- I'm surprised that all the applied tests (Table 1 and p 14) have pvalues equal to 2.2^{-16} (I guess you mean 2.2 10^{-16}?). Isn't that strange ? More importantly, I doubt that Kolmogorov Smirnov gives such low pvalues given the differences in the CDF of Figure 6.
⇒ For all the correlation tests we are showing in table 1,2,3 and 4 we computed the confidence interval. For the Kolmogorov Smirnov (K-S) test, the p-values were given by the K-S test. In fact the K-S test with low p-values meant the rejection of the null hypothesis that the two distributions are equal. Hence, the Figure 6 and the conclusion of the KS test are consistent and we explicitly said that.

- I'm concerned about the comparison with ECMWF forecasts since ECMWF are gridded data, whereas SWG is based on point data (ECAD). Have you considered comparing ECMWF forecasts with SWG based on E-OBS, since both have an horizontal resolution of 0.25°x0.25° ?
⇒ We used E-obs for clarifications and we found that they are giving the same results as ECA&D as explained before.
- The evaluation of CRPSS conditional on weather regimes is interesting but I wonder whereas considering the weather regime of the last day of the sequence (t0+T) is representative for the weather regime of the whole sequence.
⇒ We took into consideration this suggestion and we considered the most frequent weather regime for each simulation (rather than at the last day) as explained in

subsection 3.4 and we verified the relation between CRPS and weather regimes. We represent results in subsection 4.4

More minor, some references are missing (see below). There are issues in the units of CRPS. There are several equation issues.

Detailed comments :
- l 65 : Reference to Klein Tank is missing
- l 67 : please specify that ECAD provides point (station) data
- l 72-81 : actually ERA5 is now available since 1950.
- l 84 : Reference to Herschbach is missing. By the way, is it the right reference ?
- l 87 : so ECMWF forecasts have the same resolution as EOBS.

⇒ **We added the missing references.**

- l 104 eq (1) : This is a good idea to account for several days in the distance, however D in (1) is not anymore a mathematical distance. Of course this is not mandatory for analog search, however why not using [sum_x sum_i {Z500(x,t+i)-Z500(x,t'+i)²}], which is a mathematical distance? By the way, could you pelase provide a comparison of the results with the Euclidean distance (based on 1 day) vs. the D distance (based on 4 days)? And why 4 days ?

⇒ **We added the results of the simulation of the precipitation with analogs computed based on 1 day and 4 days embedding in the subsection 4.1. The comparison showed that the skill of the forecast with 4 days embedding is better than 1 day.**

- l 111-118 : explanations are quite confusing. I had to read Yiou and Déandréis to understand Please consider rewritting the method.

⇒ **We clarified the subsection 3.3.**

- Figure 1 : a) please consider placing the red rectangle somewhere else within the 30 days for clarity since its date is not necessarily the same as the target day. b) the largest window doesn't match the coordinates given l 80. I would be happy to see some results on the other windows of analogy. Otherwise I think it's not worth showing them. Also there are several syntax issues in the caption

⇒ **We modified the figure 1 a and b and made it clearer. We explained in subsection 4.1 the choice of the windows of analogy. In fact there is no difference between the small blue boxes and the red rectangle. However for the big blue rectangle we provided a table 1, where we are showing the difference between the two domains.**

- l 122 : please specify that persistence is computed over year k (unlike the climatology which is computed over all years)

⇒ **done**

- l 124 : « control forecast » I dont't understand

⇒ **We changed this.**

- l 137 : I guess that averaging the 100 trajectories smooths out the predictions. So at the end, is there a real gain (in terms of CRPSS) compared to considering only one analog ? (maybe that's already studied in another article, I haven't checked)
⇒ The CRPS/CRPSS are based on the 100 trajectories, not by considering only the mean. The "averaged" trajectory is just for illustration purposes, and does not influence the CRPS score computations. The correlations are computed over the averaged trajectories.

- l 150 : P(x) should be P(x,t) for day t. Please also rephrase the sentence
⇒ We corrected the equation.

- l 153 eq (2) : the equation is confusing. Should be CRPS(P,t) and t should be in the right side as well. The inferiori limit is 0 for precipitation.
⇒ We corrected this equation as well.

- l 159 : seasonality → climatology
⇒ done.

- l 162 eq (3) : Equation issues. there should a sum over the days (or mean) in the numerator and denominator
⇒ We corrected it.

- Table 1 : is it Pearson correlation ? I'm surprised that all pvalues equal 2.2^{-16} (I guess you mean 2.2 10^{-16}?).
⇒ We are using a Spearman (rank) correlation. We used confidence intervals instead of p-values.

- l 198-201 : this paragraph should go after l 204. Please refer to Fig 3.
⇒ We corrected this.

- l 212 : syntax issue
⇒ done
- Table 2 : I guess this is for NCEP ?
⇒ done.
- l 228 : remove the brackets
⇒ done.

- l 230 : so you obtain 100 classifications. How do you deal with that ?
⇒ We clarified this part.
- l 241 : is the weather regime at time t0+T representative of the sequence from t0 to t0+T. Why don't you consider the most frequent WR within t0 to t0+T ?
⇒ We decided to consider the most frequent weather regime in each simulation and we get results that we discuss and present on subsection 4.4

- Fig 3 : please consider plotting both reanalyses on the same plot for ease of comparison. Actually for persistence ERA5 seems to give larger CRPSS. Caption : persistence in lowercase letter. Please use either « reeference » or « baseline » along the article. For the boxplots, why don't you consider correlation with the mean instead of the median (as the predictions of SWG) ? (but it should not change much)

⇒ We added a table 2 where we are shown the comparison between the simulations from both reanalyses.

- l 243 : don't you mean below the 25th quantile ? Where is this used ?

⇒ We considered values below the 25th quantile as indicator of good forecast quality, and quantile beyond 75th quantile for poor forecast quality

- Figure 4 caption : blank space after (BLO)

⇒ done.

- l 251 and followings : I see there are differences depending on the WR but it seems to depend on the city. Can we have explanations why, e.g. CRPS for BLO is better in Orly ?

⇒ We related this to the difference in the local weather.

- Figure 5 : Units of CRPS are not mm. What is the lead time here ? Given l 243-244, I would have expected here to see boxplots for the two classes of predictability.

⇒ That was clarified and the units of CRPS was corrected in the subsection 4.5.

- l 257 : what is the reference for CRPSS ? (I guess climatology)

⇒ yes climatology. This will be clarified.

- l 261 : do I understand correctly that in ECMWF forecasts, CRPSS is given for the whole of Europe whereas CRPS are avalaible at every grid point ? As said above, I find difficult comparing the skills of ECMWF vs SWG given that the horizontal resulution is different (0.25°x0.25° vs point data). Comparison of ECMWF with EOBS at the same resolution may be easier .

⇒ Yes ECMWF forecasts are given for the whole of Europe, we extracted forecasts in single points which have the same coordinates then the studied stations, then we did the comparison. ECAD data are provided at station level (This is what interests us in this study). Moreover, the E-OBS are made from the ECAD data. By comparing the E-OBS and ECAD there is a strong correlation between data.

- Table 3 : please specify the reference. You may want to add here the CRPSS of Europe with ECMWF.

⇒ climatology. This was clarified.

- l 264 : CPRSS are actually hard to compare since they are not based on the same data (different resolution)

⇒ We cannot find literature that explains how CRPS/CRPSS values should depend on data spatial resolution. The main difference stems from the ensemble size. We did the simulations with E-OBS data (that yields the same horizontal resolution as the ECMWF forecast) and we found the same results.

- l 266 « We found... » I don't understand the sentence (syntax issues). Anyway according to the CDF of CRPS in Fig 6, ECMWF seems significantly better (a much larger proportion of low values)
⇒ We corrected this.

- l 270 again 2.2^{-16} ? Anyway, I think something's wrong here because the CDFs in Fig 6 do seem different. A difference of 0.2 between CDFs is large actually.
⇒ small p-values of Kolmogorov Smirnov indicate that the null hypothesis is rejected, and it is in agreement with the D (difference between CDFs) we explained this in subsection 4.5.

- l 276 and Fig 7 : I think something's wrong because ECMWF shows a much larger proportion of small CRPS for Toulouse and Madrid (see Fig 6). The difference in CRPSS for Orly between 5 and 10 days is very surprising.
⇒ We will compare what we find for precipitation with temperature in order to better understand this relation; for Toulouse and Madrid, we will verify this.

- l 300 : designed
⇒ done.

- Some references are missing. There is no year for Cassou.
⇒ We added all the references.

Review (2) of the manuscript "Assessment of stochastic weather forecast of precipitation near European cities, based on analogs of circulation" by M. Krouma et al.

We thank the reviewer for the positive and constructive comments, which we considered in the new version of our manuscript.

This is a very interesting manuscript, owning a good potential to become a high impact paper with positive repercussions on different societal sectors. A stochastic rain generator is produced exploiting the relationship between Z500 and precipitation in different European cities. The work is worth publication, but it needs a substantial revision about three distinct points:
1 - an improved description of the methodology is needed, in order to better understand the workflow and some of the choices that have been employed.
⇒ We rewrote the methodology, we clarified our choices for the analog search and the configuration of the SWG. (subsection 3.1 & 3.2)

2 - the use of the prolonged ERA5 dataset (since 1950) is urged, in order to understand whether the differences in skill with NCEP are actually due to the length of the analog database, or to the database itself.
⇒ We extended the analog search in ERA5 to 1950. And we did a comparison with NCEP. We represented the results on a table 2 subsection 4.1.

3 - a thorough and comprehensive revision of the English language is needed. Many subject-predicate inconsistencies, missing s', wrong sentence structures make some parts of the manuscript very hard to read.
⇒ We took care of the English language in this new version of the paper.

Specific comments are stored in the attached files. In bold font, those that pertain to the above-mentioned major observations.
Please also note the supplement to this comment: https://gmd.copernicus.org/preprints/gmd-2021-36/gmd-2021-36-RC2-supplement.pdf

Lines 79-80 and figure 1. Why exactly that region? Is there a process-knowledge approach behind this choice, a literature review or a regression/correlation between Z500 and rain over each station was applied? Whatever, the choice, it should be justified by means of references or, in the last case, with a graph/map that certifies the link between Z500 and precipitation.
⇒ We justified our choice with references in the introduction, we mentioned recent studies where the relation of Z500 and precipitation was explained. Regarding the geographical region, we explained further in subsection 4.1, that we explored the relationship between different regions. However, we found that the region with the correlations 30°W-20°E; 40°-60°N is the optimal as it allows to calculate analogs for the different regions and also at a lower cost.

Line 98. Why did the author choose exactly 20 analogues? Would not be better to base the choice on a maximum Euclidean distance?
⇒ We explained that the choice of 20 analogs was based on experimental experience. We explained that we do not find changes on the Euclidean distance when we exceed some number of analogs. We justified our choice using a recent study of Platzer et al. (2020, https://arxiv.org/pdf/2101.10640) that shows that for complex systems the use of a large number of analogs (exceeding 30 analogues) doesn't make a big change for a forecast with analogs.

Line 100. The 4-day time embedding is not clear to me. Why is it necessary? Why does it preserve the temporal derivative of the atmospheric field? Please explain.
⇒ We explained that 4 day embedding enhances a better simulated persistence and yields better skill scores for the forecast. We added the results of that on subsection 4.1.

Lines 113-118. Despite the mechanism is quite clear to me, the sentences "In order to go […] precipitation between t0 and t0 + T" are not well formulated. Since this is a crucial part for the understanding of the method, I would rephrase and expand this part.
⇒ We rephrased the subsection 3.3.

Line 120. "of the properties" is redundant.
⇒ We corrected this.

Line 123. More than the average value, the persistence consists in the anomaly between t0 - T and t0. Also, the climatological forecast takes.
Lines 128-134. Again, it is not clear upon which basis those domains (in Fig. 1b) have been chosen and the final domain selected among the four attempts. Also, the entire paragraph needs a language revision.
Line 150. A verb is missing (meet?).
⇒ We rephrased and we corrected the mistakes.

Figure 2. A visual legend is needed. Also, what's in the y-axis? Are the 5th and 95th percentiles calculated over the 1948-2019 time series?
⇒ We added a visual legend to make the figure clearer. And we clarified the caption of figure 3.

Line 184: six??
⇒ We meant four (a mistake).

Lines 201-205. It seems to me that part of the methodology is described here, where a description of results is expected.
⇒ We added this result on a new subsection 4.1 "parameter optimization", where we showed the results of the different experiments that we did (databases, geographical domain, embedding..)

Line 206 and following. Very little description is given for figure 3. First of all, I think it is useful to illustrate all the stations in the main text, instead of showing only Orly while relegating the others to the Appendix. This is one of the main results of the study and deserves a better stage (instead, I would recommend to place Fig. 4 the Appendix, since the WRs are not a result of this study). Besides, fig. 3 and fig. A1 show a very interesting characteristic that should be discussed: most of the times, in fact, the summer SWG forecast vs. persistence improves with lead times, which is somewhat unexpected. Any thoughts about it?

⇒ **We illustrated the results of the different studied stations. We chose to show results with NCEP to be more precise.**

Line 217. De Bilt?
Line 225. They help describe.
Line 232. Missing year for the Cassou reference.
⇒ **we corrected the sentence and added the missing reference.**

Lines 246-255. Fig. 5 (B1) should be described with more care. What do "Good forecasts (low quantiles of CRPS)" and "The low quality forecasts (high quantiles of CRPS)" mean? The caption for figure 5 is totally unclear and does not describe the plot. On the top of the blocking bar (panel b) a group of dots appear.

⇒ **We clarified the caption in figure 6, we explained further the plot, and how we sampled the weather regimes dates. We explained on the subsection 3.4 how we determined the relation with the CRPS, our definition of good/low forecast quality.**

Lines 257-276. This paragraph is hard to read, there are many inconsistencies between subjects and predicates and other grammar errors. Also, the first paragraph is not clear: what is it meant to demonstrate? Maybe the lower skill of the ECMWF forecast? The latter is calculated over the entire European domain, how can it be compared with a forecast over single stations? A table with both ECMWF and SWG CRPSS would be more informative than a few words, if the authors find the way to make a fair comparison between the two.

⇒**We rephrase the paragraph. And we explained that we extracted the ECMWF forecasts at single points that have the same coordinates as the studied stations and we made the comparison.**

Line 279. The input of our model was analogs of geopotential heights at 500 hPa (Z500). This sentence should be rephrased.
⇒ **We corrected the sentence.**

Line 283. I cannot accept this conclusion. The only way to test it is to compute the analogs with the 70-year ERA5 dataset, now available since 1950. This is a very important test that should be included in this study, because it clarifies the role of the different reanalyses as well as the role of database length.

⇒ **We computed analogs from the ERA5 dataset including data from 1950 to 1978. The results with ERA5 (1950-2019) and NCEP (1948-2019) are very similar, we are showing this in the subsection 4.1.**