The replies to referees are indicated in red. We thank the reviewers for pointing out unclear points in the manuscript.

Referee#2

In this manuscript the authors develop an ensemble forecasting system using an analog-based weather generator. They test this ensemble forecasting system for NAO and the temperature at several weather stations. They focus on the forecasts of temporal averages from 5 to 80 days. The forecast is made for each averaging period at the first corresponding lead time. The skill of these forecasts are evaluated through skill scores (the correlation and the continuous rank probability score, CRPS, the latter being well adapted to ensemble forecasts). The authors claim that there is some skill of the temperature and NAO up to seasonal time scales. I am not convinced by the system they propose, nor by the skill they found, for three important reasons:

(i) The system they propose suffers from a very important drawback, which is the progressive convergence toward the climatological mean as illustrated on the right column of Figure 3. There is only little variability of the forecasts for long time averages, indicating that the ensemble forecast is unreliable. This makes of this system a very poor probabilistic forecasting system, since the forecasts do not span the set of possible values of the observed variable. Reliability is one essential ingredient of ensemble forecasts that can also be easily checked with the decomposition of the CRPS in reliability and resolution. I therefore do not consider this ensemble system appropriate.
We never hide the fact that there is a convergence towards climatology (this is mentioned in the text). But long term forecasts with full scale climate models yield the same feature (as outlined by Hersbach (2000) and others). Our claim is that this system does a better job than usual references (Climatology or Persistence) or AR1 models (see response to referee #1). The ease of use of this system makes it possible at low cost to investigate the limit for large lead times. We consider that the fact that the scores are positive for shorter lead times (20 days ahead) is interesting. We now mention (and use) the CRPS decomposition of Hersbach (2000) in terms of reliability and potential CRPS.

(ii) It is not clear at all to me why the authors are looking at the first lead time of the 5, … 80 days averages. Using this approach, one can certainly expect that if one start from an initial state close to the reality, the forecast of the averages will always be better than the climatological average (provided we have access to an infinite sample). In other words some positive correlation will always be present, even if it is very small. This skill is artificial (due to averaging from the initial state) and I am wondering why the authors did not have looked at the skill of the daily values of NAO or temperatures. My guess is that there is no skill beyond a month or so.
We never claim the contrary and discussed it in the text. Starting from an observed state, daily trajectories tend to diverge from each other. The computing T-averages for various lead times allows accessing to the limit of predictability of our system.
But if an autoregressive model (mAR1, see response to referee#1) is initialized from observations, there is NO skill (correlation or CRPS). As stated in the text, we do not consider that the system has any skill beyond a month.

(iii) The analysis of the skill of ensemble forecasts should be done with appropriate tools. The CRPSS is one of them, but it is much more important to look at its decomposition in reliability, resolution and uncertainty. These are standard tools that can be found in classical books or papers (e.g. H. Hersbach, 2000, Weather and Forecasting, 15, 559-570).

Thank you for this suggestion. We add a discussion on the decomposition of CRPS (citing the paper of Hersbach 2000) in terms of reliability and potential CRPS. The problem with reliability is that its magnitude depends on the unit of the variable to be predicted (as discussed by Hersbach 2000). The results reported by Hersbach give very small values of reliability (for precipitation forecast) when the ECMWF analysis is used. But those numbers are small because the variable values to predicted are small.
We used the R package "verification" (by E. Gilleland) to compute this decomposition. The relative of variations of reliability that we obtain for temperature or NAO forecasts is in the same range of what is reported in the paper of Hersbach (2000) for lead times of 5 to 10 days. We now discuss the values of reliability, which appears in Figures 4-5. The reliability values for NAO are small ($\approx$ 8 10^{-3}), and the ratio to the CRPS value is in the same range of what is reported in Hersbach's paper.

Some additional (less important) points

1. The algorithm of page 4 (section 3.2) is far from clear. It would be nice to visualize the algorithm, together with the relations that are used for evaluating the weights.
OK. A graphical illustration is added (see below) to visualize the iteration procedure and the choice of weights to sample analogues.

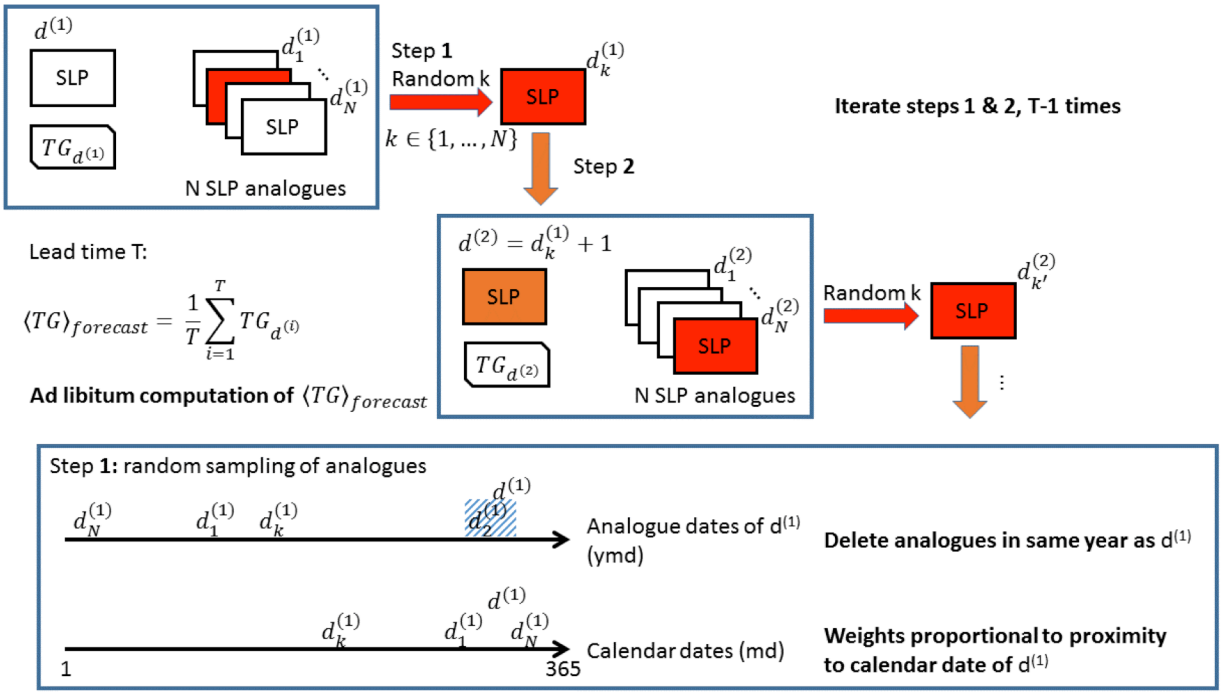

Illustration 1 : Schematic of the stochastic analogue weather generator. ymd indicates when absolute time is considered. md indicates when calendar time (i.e. time in the year) is used.

2. Page 6, line 5. Is S=N? This is not clear to me.
We compute the N=20 best analogues for each day. At each time increment, we simulate S=100 trajectories, sampled from those 20 best daily analogues. For a lead time of, say, 10 days, there are 20^{10} possible trajectories, which is far larger than S. This will be emphasized in the text.

3. An additional concern I have is the comparison with the persistence in Figs 4 and 5. It seems to me that the observables based on persistence display a higher variability than the forecasts constructed here (that are converging to the climatology). I therefore suspect

that the reliability of the persistent forecast is better than the one of the stochastic forecasts (the reliability term in the CRPS decomposition should be smaller for the persistence case), which is not reflected here in the analysis of the CRPSS. I would be very useful to evaluate the different terms of the CRPS to clarify the difference between the two systems. This will allow in particular to clarify why one gets 0.45
for all averages for NAO and why the skill increases for temperature.
A discussion on the CRPS decomposition for the different forecasts is added. The reliability value of CRPS for the persistence or the climatology give higher values (roughly twice larger) than for our model.