Response to Interactive comment by Anonymous Referee #1

We thank the referee for this thorough review and for the numerous constructive suggestions. Please find below, our answers to these suggestions.

2. I missed some validation or references to validation of the GLS data. As mentioned by the authors, Safran has a number of biases. Crocus might be based on assumptions which are not always fulfilled and so the end product, GLS, might also suffer from a number of shortcomings.
We will add several references emphasizing that SAFRAN and Crocus have been well validated.

3. In addition, there is no validation of the GEV models, just the final selection among the models in Table 2. These are based on AIC and likelihood ratios. So the best model is selected. But do they fit well ? What if none of the models were really adequate (even the best one among them) ? Maybe some qq-plots analyses should be included.
Quantile-Quantile (Q-Q) analysis is performed for all selected models. To apply this analysis to both stationary and non-stationary model, we rely on [1] that suggests 1) to transform the data to stationary Gumbel 2) to use a Q-Q plot analysis on the transformed data w.r.t. to a Gumbel distribution. Q-Q plots reveal that transformed data is well fitted by a stationary Gumbel distribution, hence that data is well fitted by the selected models.
Moreover, according to the comparative study [2], the most powerful Goodness of Fit test for the Gumbel distribution is a combination of the Anderson-Darling test and the Maximum Likelihood Estimator. We apply this test on the transformed data, and found using [3] that we cannot reject the null hypothesis (samples generated from the Gumbel model) at the 5% significance level for 98% of the time series, justifying the good fit of our selected models.
We will add these test results at the beginning of the Result section. In an Appendix section, we will detail an explanation on the methodology of [1], and display Q-Q plots for the time series presented in section 2.

4. Given the amount of literature, I found it a bit disappointing that no attempt was made to rely on models that make use of more data, not only annual maxima as mentioned in the discussion. For instance, the tail index is taken to be constant in view of the difficulty to estimate it. There are many ways around this, one of which is the so-called regional analysis.
SAFRAN reanalysis are the result of a postprocessing of the meteorological observations at the massif scale and, as such, already represents "regionalized" data. In this context, it does not seem clear how a regional analysis could be performed.

5 The authors argue that the number of years of the GSL reanalysis is too short to attempt to use anything else than linear relationships in the non-stationary models. Nevertheless, they recognize that other extreme value approaches, such as peaks-overthreshold, can be apply to exploit more data (more than a maxima per year). This seems a bit contradictory. If the authors could show that the GEV models with

linear non-stationarities fit well the data without too much uncertainty in the estimates, then it would alleviate this issue

Our goal is to implement a clear comparison with French standards. For this reason,,thus we prefer to rely on the Gumbel distribution & extensions of this distribution, which explains our choice to use Gumbel and GEV distributions.

Furthermore, the impact of the uncertainty in the estimates is already shown on our main figures (black bars on Figure 9). Despite that these uncertainty interval can sometimes be large, it does impact the main conclusions of this article. For instance, we would still have between 40 and 80% of massifs whose return levels in 2019 exceed French standards.

6. Although the paper is generally well written, I think it can be improved on a number of aspects

We will correct expression/syntax mistakes that are mentioned. Also in the modified manuscript, we will clarify the following points:

6.1. I found that the abstract was not conveying too well the main analyses and conclusions of the paper.

We will work on improving the abstract once the content of the modified manuscript is finalized.

6.2. P.3 What is the spatial resolution of Safran ? Does GSL has the same spatial resolution ?

As explained on l.59, SAFRAN does not provide gridded data, it gives massif-level data. More precisely, as detailed in [6]: " The principle of SAFRAN is to perform a spatialization of the available weather data in mountain ranges with so-called "massifs" of about 1000 km2 where meteorological conditions are assumed to depend only on altitude." In the Data section, we will add a sentence to make that point clear. Otherwise, Crocus snowpack model takes SAFRAN data as inputs to produce SWE (which we use to compute GSL), therefore yes, GSL data has the same spatial resolution as Safran.

6.2. P.11 last sentence : "... often above effective return levels " effective in what ways ?not sure what it means here

While classical stationary return levels do not depend on time, return levels are denoted as effective when they depend on time [4, 5]. To quote [4]: "[Effective design value] has an interpretation similar to that for an ordinary design value (i.e., the quantile corresponding to a specified return period), except that it varies depending on the time of year.".

6.3. -P.13 L.245-250 : " ... start the non-stationarity after the most likely year ", what is meant by most likely year ?

For each model with a linearity in some parameters of the distribution we could choose to start the linearity only after some starting year.

The most likely starting year is the year that gives the maximum likelihood for this linear model [6]. However, in the end we decided not to use this approach. Therefore we propose to remove this sentence altogether from the discussion Section to avoid confusing the reader with unnecessary details.

[1] Richard W. Katz. (2012). Statistical methods for nonstationary extremes. In Extremes in a Changing Climate - Detection, Analysis & Uncertainty (pp. 15–38). Springer Science & Business Media.
[2] Abidin, N. Z., Adam, M. B., & Midi, H. (2012). The Goodness-of-fit Test for Gumbel Distribution: A Comparative Study. Matematika, 28(1), 35–48. Retrieved from http://www.matematika.utm.my/index.php/matematika/article/view/313
[3] Ali Saeb (2018). gnFit R package https://www.rdocumentation.org/packages/gnFit
[4] Katz, R. W., Parlange, M. B., & Naveau, P. (2002). Statistics of extremes in hydrology. Advances in Water Resources, 25(8–12), 1287–1304. https://doi.org/10.1016/S0309-1708(02)00056-8
[5] Mondal, A., & Daniel, D. (2019). Return Levels under Nonstationarity: The Need to Update Infrastructure Design Strategies. Journal of Hydrologic Engineering, 24(1), 04018060. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001738
[6] Nousu, J.-P., Lafaysse, M., Vernay, M., Bellier, J., Evin, G., & Joly, B. (2019). Statistical post-processing of ensemble forecasts of the height of new snow. Nonlinear Processes in Geophysics, 1–32. https://doi.org/10.5194/npg-2019-27
[7] Blanchet, J., Molinié, G., & Touati, J. (2016). Spatial analysis of trend in extreme daily rainfall in southern France. Climate Dynamics, 51(3), 799–812. https://doi.org/10.1007/s00382-016-3122-7