## Comments from Referee #1

The paper raises a very important issue: in order to obtain robust estimates, especially in a nonstationary setting, we need to use a multitude of available information sources. However, it is a challenge to connect properly these different information sources and requires a lot of scientific effort. I consider this paper to be one of the initial steps in the right direction regarding the quantification of changing extreme events due to global warming. The paper has a clear structure, the language is understandable despite the many technical details and steps, the figures are illustrative. I appreciate a lot that the authors discuss critically the drawbacks of their method in Sec. 5.2, and that they provide a clear and concise overview of related studies and how this work is embedded in the current scientific literature. Another strength of this work is that a range of possible parametrisations is explored and the optimal one is chosen in an objective way, based on a mean logarithmic score. I tend to suggest the manuscript to be accepted for publication, but only after a few very important issues are clarified.

We thank the reviewer #1 for this positive feedback and these useful comments on our manuscript. Please find below a detailed response to individual comments and questions.

### Main comments:

RC1#1. My main criticism is that the authors use a non-stationary GEV setting without showing that the annual maxima of snow load are properly estimated by a GEV distribution. The authors write that “...due to these theoretical justifications, the GEV distribution enables a robust estimation of return levels”. Yes, the GEV distribution enables (under certain conditions) a robust estimation, but only if:

1) the chosen block size □ is large enough, as the theory applies for □→∞. The block size of one year is not automatically large enough (one might need 2 years or 10 years or even longer)–it has to be shown that the annual maxima can be reliably estimated based on a GEV distribution (see convergence plots and diagnostic plots in Coles, 2001).
2) the auto-correlation is weak enough -the stronger the auto-correlation, the larger the smallest block size for which the GEV limit is valid. These convergence issues are discussed extensively in Galfi et al.(2017, Complexity).

I ask the authors to show that the convergence to the GEV distribution is good enough for the chosen block size, i.e. that the annual maxima are properly modelled by a GEV distribution. This should be done before building up the non-stationary framework. If this cannot be shown, I’m afraid that the whole experimental setup described in the paper is useless, as it is a necessary condition. In the theoretical description of the GEV distribution in Section 3.1 more emphasize is needed to underline the asymptotic nature of the theory.

We agree with the reviewer that we make the strong assumption that annual maxima follows a non-stationary GEV distribution whereas our block size is limited to one year. As done in many studies on climate extremes, the choice of annual blocks represents a compromise between having sufficiently large blocks in order to obtain maxima on a large number of values (365 days), and not too large in order to have enough blocks to avoid very uncertainty GEV parameter estimates (which would be the case with a 10-year block size, leading to very small samples). The revised version of the manuscript will clarify this pragmatic choice, and recognize that we may still  not totally fulfill the convergence conditions, but that according to the data our model choice appears to be sensible.

In detail we will further present a quantitative evaluation of the goodness of fit. We will rely on the Anderson–Darling statistical test, which is the most powerful test for the Gumbel distribution (Abidin et al., 2012), similarly to what was proposed in Le Roux et al. (2021). This test assesses if the residuals follow a standard Gumbel distribution (see the Appendix A of our article for a definition of these residuals). For every selected model, the p-value of this test was computed for each GCM-RCM pair separately (and the observations) for each massif. In the Figure below, we observe that the test is rejected for 20% of the 483 cases (23 massifs x 21 time series). This relatively high number seems to be mainly explained by the small values reached at the end of the century for many GCM-RCM runs. Indeed, the same tests applied at an elevation of 2700 m (see Figure 2 below) show a much smaller percentage of rejections (7%) and larger p-values. The inadequacy of the selected GEV model to represent these zero values will be discussed in the revised version of the manuscript.

The fact that snow loads maxima reach a lower bound at the end of the century is clearly a major challenge for this application, and is related to the difficulties in adjusting the shape parameter (see Section 4.2). However, we emphasize the fact that this application can be considered as a non-trivial example of the proposed methodological approach, which also illustrates its limitations. An application to unbounded variables (e.g. annual maxima of temperatures) would be easier to present but would be less innovative in terms of specific application, and would fail to illustrate these challenges. All in all, we firmly believe that these results are sufficient to support our modeling choices with regards to our application, and to illustrate the potential of the framework we propose for many applications.

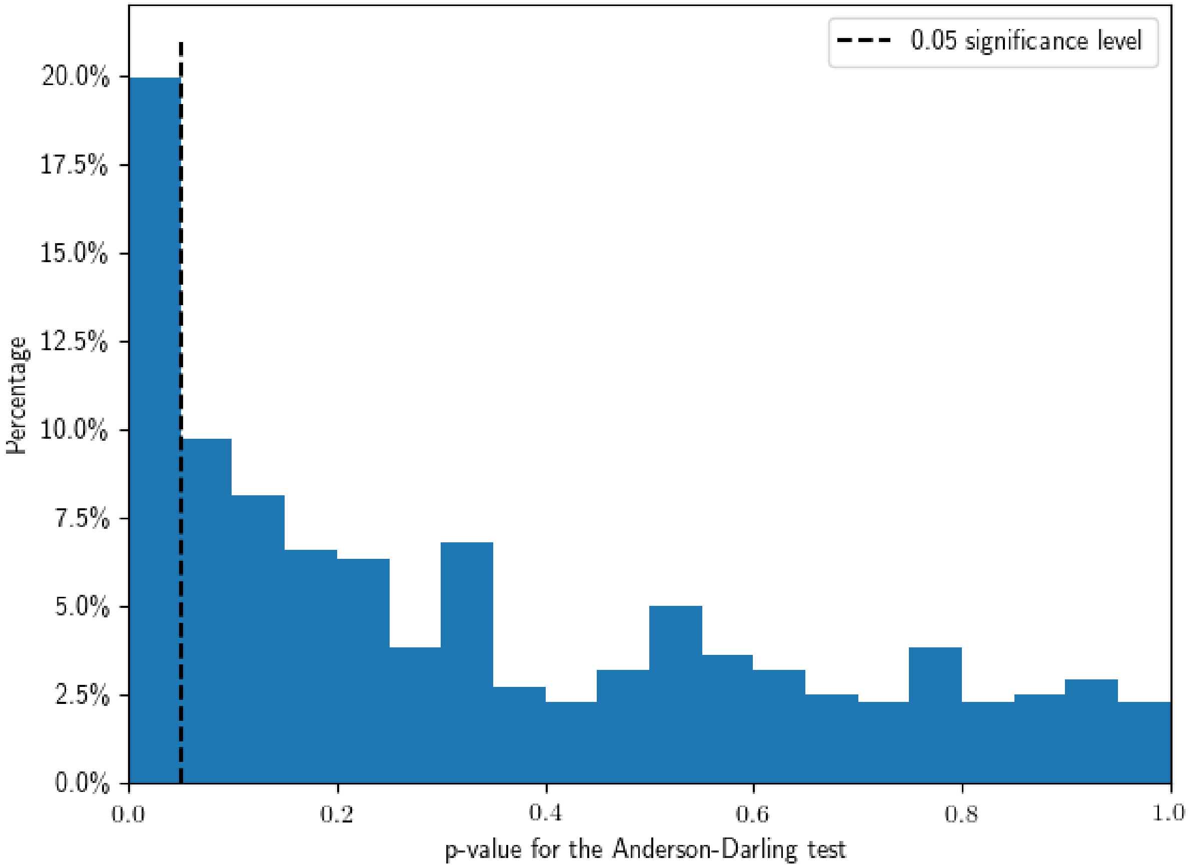

Figure 1: Distribution of p-values for the Anderson-Darling test for the elevation 1500 m. For every selected model, a p-value was computed for each GCM-RCM pair separately.

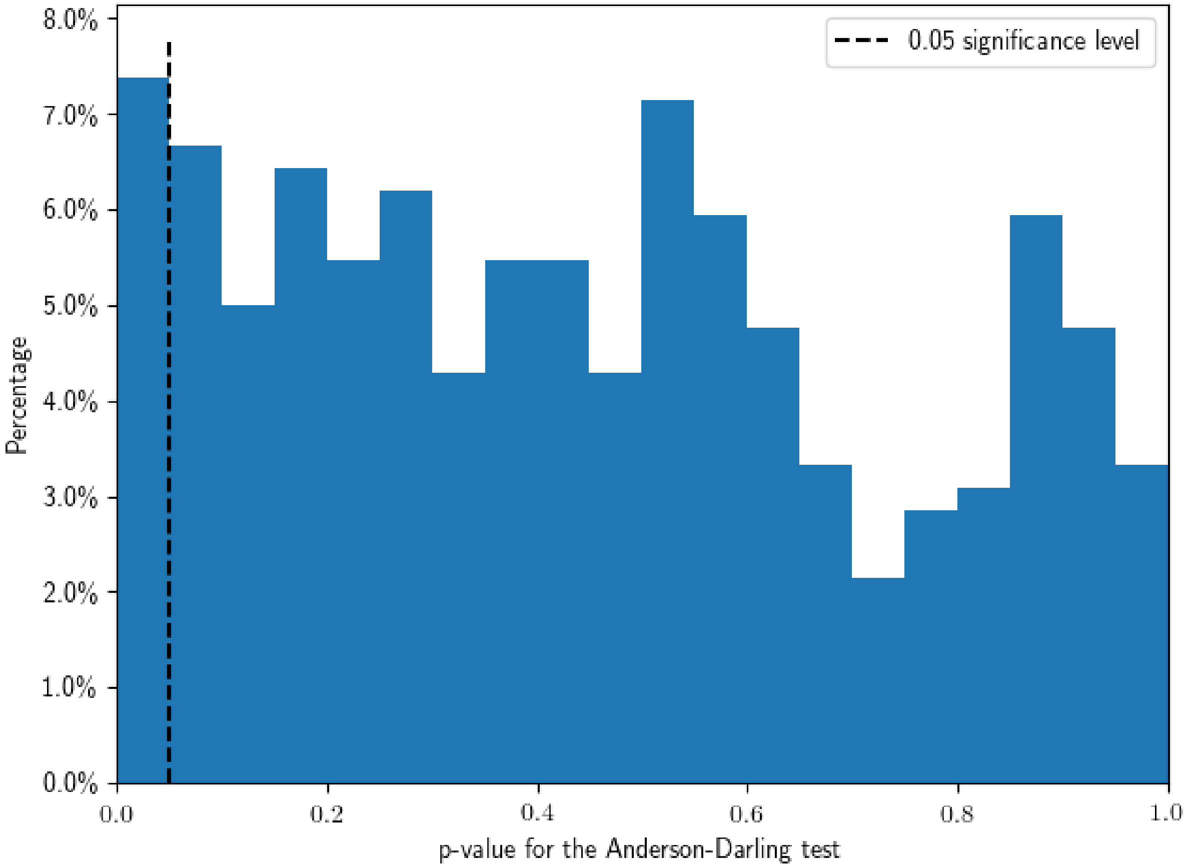

Figure 2: Distribution of p-values for the Anderson-Darling test for the elevation 2700 m. For every selected model, a p-value was computed for each GCM-RCM pair separately.

RC1#2. It is a bit disappointing that even after performing the adjustments and applying the complex methodology, the results still do not follow the observations, as shown by Fig. 4 and discussed shortly in the text. One reason for this could be that, as the authors explain, the evaluation set for adjustment parametrization contains only a few (24, 17,10) maxima, thus the selection of the optimal parametrisation might be misleading. I believe a more thorough discussion is needed here. It is also not clear for me how to overcome the issue regarding the estimation of the shape parameter: in case of no adjustment, the model results do not follow the observations, in case it is adjusted it can lead to prediction failures. The authors do not suggest any solution for this, although it would be crucial for the applicability of the method in future studies.

It is true that this issue is particularly challenging and open to discussion. On the one hand, we can imagine some applications where it seems reasonable that the model follows the observation betters, for example if a long series of past observations is available or if it is assumed that only past observations provide a relevant information about the tail of the distributions (i.e. if it is assumed that climate simulations are not able to simulate reliable climate extremes). A simple solution for doing this could be to put more weight on the observations in the likelihood (Eq. 4) in comparison to the climate simulations. However, this approach has not been tested thoroughly.

RC1#3. It is not totally clear for me why is it important to know return levels of a variable whose extremes are expected to decrease with global warming (assuming a simple direct relationship between them), thus becoming less extreme? I think the authors should put more emphasize to show the relevance of this subject.

Thanks for this question. Beyond the methodological contribution of our submission, our second objective is to check whether return levels are expected to increase or decrease. Indeed, the literature points to a decrease of mean winter SWE (IPCC 2019), i.e. to a decrease of mean winter snow load. However, to the best of our knowledge, projected trends in extreme snow loads have never been studied before.

Since annual means of snow loads are expected to decrease, our first hypothesis was that extreme snow load would also decrease. However, in cold regions (high elevation regions for instance) we expect extreme snowfall to increase with climate change (O'Gorman 2014), thus our second hypothesis was that this increase of extreme snowfall can lead to an increase of extreme snow load. According to our results, the former hypothesis is the most likely hypothesis in the French Alps.

Even if extremes are expected to decrease, a quantification of these decreases are of prime interest:

- to study compounds extremes, e.g. extreme snow load combined with extreme wind,
- to adapt structures standards, e.g. to decrease the constraints used to design new structures, which may reduce the construction cost.

These motivations will be added to the manuscript.

Specific comments:

RC1#4. L93 Are the time periods "historical" 1951-2005 and "future" 2006-2100 correct?

Yes, these are the standard historical and future time periods used in the EUROCORDEX experiment obtained from a CMIP5 ensemble, the RCP scenarios prescribing greenhouse gas concentration trajectories from 2006.

RC1#5. Eq.4 There should be two equations here: □̂=argmax(□(□|□))and □(□|□)=∏...

Thank you for this remark, this was a mistake. We will modify this equation.

RC1#6. L177-L178 It is not totally clear what "RL50" stands for because it is written that "is computed without adjustment coefficients", but later in Figure 4 it is used also for the case with adjustment coefficients.

Thanks for this remark. Whether or not the model has adjustment coefficients, RL50 corresponds to the 50-year return level computed without adding the adjustment coefficients.

To clarify that, this paragraph "*In other words, if the selected parameterization has adjustment coefficients, RL50 is computed without these adjustment coefficients since using* *these coefficients would provide the 50-year return level of the GCM-RCM pairs*.” will be replaced by “*In other words, if the selected parameterization has adjustment coefficients, we do not add these coefficients to compute the RL50*.”

RC1#7. Figure 4: I think that in case of the line with the “warm” colours the legend should be different for the 4 subplots –the legend is the same for each subplot, although the text and the figure caption suggest the opposite.

The legend is not exactly the same for each subplot. Indeed, at the end of the line with the “warm” colors it is either written ‘all GCM-RCM pairs’, ‘each GCM’, ‘each RCM’, or ‘each GCM-RCM pair’.

RC1#8. Figure 5: To which year or period do these return levels refer?

As indicated in the legend of Figure 5, these return levels correspond respectively to the 50-year return levels for T=+1, T=+2, T=+3 and T=+4 degrees of global warming.

## References

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 https://doi.org/10.11113/matematika.v28.n.313.

Le Roux, E., G. Evin, N. Eckert, J. Blanchet, & S. Morin. (2021) Elevation-Dependent Trends in Extreme Snowfall in the French Alps from 1959 to 2019. The Cryosphere 15, n$^{o}$ 9: 4335- 56. https://doi.org/10.5194/tc-15-4335-2021.