Point by point reply to the Interactive comments of Anonymous Referees

Dear Editors,

Thank you for giving us the opportunity to submit a revised draft of our manuscript «Non-stationary Extreme Value Analysis of Ground Snow Loads in the French Alps: a Comparison with Building Standards».

We thank the referee for their thorough reviews and for the numerous suggestions that helped us greatly improve the manuscript.
Following their comments, the main modifications are that we:

- added references for the validation of our data.
- validated qualitatively & quantitatively the selected models.
- modified the Abstract to better emphasize the content of the article
- changed the title of the manuscript.
- improved Figure 9 (and Section 5.3) with an additional axis illustrating the mean relative difference between return levels and French standards.
- corrected expression/syntax mistakes/typos across the document.
- generated all Figures in the pdf format instead of the jpg format.

Please find below, point by point, our answers to all the other suggestions. Anonymous Referee #1 suggestions are in red, while suggestions from Anonymous Referee #2 are in blue.

We hope that our revised manuscript will be found suitable for publication in "Natural Hazards and Earth System Sciences"

Yours sincerely,

On behalf of the co-authors
Erwan Le Roux

Suggestions from Anonymous Referee #1

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 added a paragraph in Sect. 2 emphasizing that the SAFRAN-Crocus reanalysis 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 Coles (2001) 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 of Abidin et al., (2012), 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 Ali Saeb (2018) 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 added these test results at the beginning of the Result section. In an Appendix section, we detail an explanation on the methodology of Coles (2001) 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.

6.2. 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 Nousu et al, (2019): " 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 (Katz et al., 2002, Mondal et al., 2019) . To quote Katz et al., (2002): "[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 (Blanchet et al., 2016). However, in the end we decided not to use this approach. Therefore we remove this sentence altogether from the discussion Section to avoid confusing the reader with unnecessary details.

Suggestions from Anonymous Referee #2

Specific comments: The manuscript lacks a description of (i) error measures of GSL data used as basis of the extreme value statistics, and (ii) general remarks on the reanalysis used to provide that data. In particular it would be crucial to tell something about the BIAS or absolute errors of the yearly maximum GSL values. Otherwise provided uncertainty assessments are less valuable. Furthermore a general description of some aspects of the reanalysis is missing. How is GSL calculated for the massif scale? Is the 50-year GSL return level computed by your models valid for the whole massif just depending on altitude? The abstract of Vernay (2019) states also a dependency on aspect and slope. You should clarify if your results are valid for distinct elevations or elevation bands (as it is stated here and there). In the latter case you should explain, how GSL values are assigned to that band (see lines 59, 71 in your manuscript).

The SAFRAN-Crocus reanalysis has been evaluated against various observation datasets, as reported in previous publications (Lafaysse et al., 2013, Vionnet et al., 2016, Revuelto et al., 2018, Vionnet et al. 2019). In most cases, the evaluation is carried out against in-situ snow depth observations and remote sensing snow cover information. For example, Vionnet et al., (2016) evaluated SAFRAN-Crocus snow depth data against 79 observed snow depth data in the French Alps for the 2010-2014 time period, with mean bias and standard error values of 18 cm and 37 cm, respectively. This corresponds to typical values for snow modelling systems applied in various regions on Earth. Because of lower data availability, evaluations against observed SWE values are less frequent than against snow depth data, although we note that Crocus has been shown to perform extremely well compared to other snow cover models, in terms of SWE, across many observation sites worldwide (Krinner et al., 2018) and SAFRAN-Crocus exhibits satisfying performance in terms of snow depth and SWE in the Pyrenees (Quéno et al., 2016), providing confidence, with respect to other existing datasets, in using this model chain for ground snow load (GSL) values. Further model evaluations, using additional datasets, are required to continue assessing and improving the quality of the model chain.

Furthermore, we highlight that we only used SAFRAN-Crocus reanalysis values on flat field, and we did not used simulations on slopes, hence it is not relevant to discuss the impact of slope and aspect on the results of this study.

Technical corrections:

80: As maximum values are relevant in this study, the procedure of _removing the top annual maximum when considered exceptional_ should be shortly addressed. I can imagine that one can find information about that in the given reference, but this is in French…

The procedure is as follows: «If the ratio of the largest load value to the characteristic load determined without the inclusion of that value is greater than 1.5 then the largest load value shall be treated as an exceptional value» (Sanpaolesi et al., 1998). This will be added to the revised version of the manuscript.

84: What exactly do you mean with _relative change_? Relative to what? (see also line 48)

We meant "relative change of 50-year return levels of GSL between 1960 and 2010". We will clarify it when necessary, and maybe refer to formula 4 (detailed expression).

126: I wonder if these complex expressions are necessary to understand the content? If not you could remove them.

We do not believe that the expression of the AIC is particularly complex. Most importantly, we think that this expression is necessary to understand the model selection, since the penalization of the log-likelihood by the number of fitted parameters clearly appears.

219: Why of all things 1800 m? Is this because Vercors top heights are around 1800m?

This is because French standards for extreme snow loads are defined from 200 m to 2000 m (Section 2). As we consider available altitudes between 200 m and 2000 m, only results obtained with reanalysis from 300 m to 1800 m are shown. However, the SAFRAN-Crocus reanalysis can provide results at higher elevation for the mountain areas peaking above 1800 m elevation.

Figure 8. Top left panel: Do you have a clue, why the uncertainties at lower altitudes are larger than at higher altitudes? With respect to the smaller number of available reanalysis stations at higher altitudes, this should be inverted, as can be seen in all other panels.

The reviewer must refer to the top-right panel (Vercors massif & Selected model) which is certainly different from the other panel.

Indeed, uncertainties usually grow larger with the altitude. Looking at similar plots to Figure 8 for all other massifs (not shown), this pattern is always seen for the left panels, i.e. with the stationary Gumbel model. However, for the right panels, i.e. for the selected model, 6 massifs out of 23 (Vercors, Ubaye, Oisans, Mercantour, Maurienne, Haut Var Haut Verdon) do not present this pattern, i.e. have larger uncertainties at lower altitudes. Some of these uncertainties might be due to variance in the estimated parameters. In particular, the shape parameter of the GEV distribution is known to be difficult to estimate. As shown in Figure 4, at 900 m the Vercors massif (most western massif) is colored in brown, meaning than its shape parameter roughly equals 0.3. This might explain the high uncertainty at 900 m in the top-right panel, as small changes around 0.3 can have large effect in the 50-year return level.

257-258: You obtained the "same" results for time series with less than 10% of zero GSL values. Can you provide a similar number used by French standards for the decision to switch to a mixed discrete-continuous distribution?

In the French standards, the mixed discrete-continuous distribution was considered for all time series, those with less than 10%, of zero GSL values, as well as those with more than 10%.

299-300: This statement is unclear. I suggest to either remove it, or to provide more details. If you really would like to leave that here, you should provide at least a refer-ence for the European construction standards, and elaborate a little bit on those safety coefficients that might alter very widely according to country, professional, construction material, etc.

We agree with the reviewer that this paragraph should be more detailed, and we did that in the revised manuscript. Concerning European standards (Sanpaolesi et al. 1998, page 32, equation 8), the design value for the structure equals the sum of i) the characteristic value of permanent action, i.e. self-weight, multiplied by a safety coefficient equal to 1.35 and ii) the characteristic value of variable action, i.e. roof snow load, multiplied by a safety coefficient equal to 1.5.

References

Coles, S. G. (2001). An introduction to Statistical Modeling of Extreme Values. Springer Series in Statistics, 221. https://doi.org/10.1007/978-1-4471-3675-0

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

Saeb, A. (2018). gnFit R package. Retrieved from https://www.rdocumentation.org/packages/gnFit

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

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

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

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

Vionnet, V., Six, D., Auger, L., Dumont, M., Lafaysse, M., Quéno, L., Réveillet, M., Dombrowski-Etchevers I., Thibert, E. and Vincent, C.: Sub-kilometer precipitation datasets for snowpack and glacier modeling in alpine terrain, Front. Earth Sci., 7, 182, https://doi.org/10.3389/feart.2019.00182, 2019.

Vionnet V., Dombrowski-Etchevers I., Lafaysse M., Quéno L., Seity Y., and Bazile, E. : Numerical weather forecasts at kilometer scale in the French Alps : evaluation and applications for snowpack modelling, J. Hydrometeor., 17, 2591-2614, doi:10.1175/JHM-D-15-0241.1

Quéno, L., Vionnet, V., Dombrowski-Etchevers, I., Lafaysse, M., Dumont, M., and Karbou, F.: Snowpack modelling in the Pyrenees driven by kilometric-resolution meteorological forecasts, The Cryosphere, 10, 1571-1589, doi:10.5194/tc-10-1571-2016

Revuelto, J., Lecourt, G., Lafaysse, M., Zin, I., Charrois, L., Vionnet, V., Dumont, M., Rabatel, A., Six, D., Condom, T., Morin, S., Viani, A., and Sirguey, P. : Multi-criteria evaluation of snowpack simulations in complex alpine terrain using satellite and in situ observations, Remote Sensing, 10, 1171, doi:10.3390/rs10081171, 2018.

Krinner, G., Derksen, C., Essery, R., Flanner, M., Hagemann, S., Clark, M., Hall, A., Rott, H., Brutel-Vuilmet, C., Kim, H., Ménard, C. B., Mudryk, L., Thackeray, C., Wang, L., Arduini, G., Balsamo, G., Bartlett, P., Boike, J., Boone, A., Chéruy, F., Colin, J., Cuntz, M., Dai, Y., Decharme, B., Derry, J., Ducharne, A., Dutra, E., Fang, X., Fierz, C., Ghattas, J., Gusev, Y., Haverd, V., Kontu, A., Lafaysse, M., Law, R., Lawrence, D., Li, W., Marke, T., Marks, D., Nasonova, O., Nitta, T., Niwano, M., Pomeroy, J., Raleigh, M. S., Schaedler, G., Semenov, V., Smirnova, T., Stacke, T., Strasser, U., Svenson, S., Turkov, D., Wang, T., Wever, N., Yuan, H., and Zhou, W.: ESM-SnowMIP: Assessing snow models and quantifying snow-related climate feedbacks, Geosci. Model Dev., 11, 5027-5049, https://doi.org/10.5194/gmd-11-5027-2018, 2018.

Lafaysse, M.., S. Morin, C. Coléou, M. Vernay, D. Serça, F. Besson, J.-M. Willemet, G. Giraud and Y. Durand, 2013 : Towards a new chain of models for avalanche hazard forecasting in French mountain ranges, including low altitude mountains, Proceedings of the International Snow Science Workshop Grenoble - Chamonix Mont-Blanc - 2013, 7-11 October, Grenoble, France, 162-166.

Sanpaolesi, Luca and Currie, D and Sims, P and Sacre, C and Stiefel, U and Lozza, S and Eiselt, B and Peckham, R and Solomos, G and Holand, I. and others. (1998). Scientific support activity in the field of structural stability of civil engineering works: snow loads. Final Report Phase I. Brussels: Commission of the European Communities. DGIII-D3.

~~Accounting for Non-stationarity in~~ Non-stationary Extreme Value Analysis of Ground Snow Loads in the French Alps: a Comparison with Building Standards~~in the French Alps~~

Erwan Le Roux[1], Guillaume Evin[1], Nicolas Eckert[1], Juliette Blanchet[2], and Samuel Morin[3]

[1]Univ. Grenoble Alpes, INRAE, UR ETNA
[2]Univ. Grenoble Alpes, Grenoble INP, CNRS, IRD, IGE
[3]Univ. Grenoble Alpes, Univ. Toulouse, Météo-France, CNRS, CNRM, CEN Grenoble

Correspondence: Erwan Le Roux (erwan.le-roux@inrae.fr)

Abstract. In a context of climate change, trends in extreme snow loads need to be determined to minimize the risk of structure collapse. We study trends in ~~annual maxima~~ 50-year return levels of ground snow load (GSL) using non-stationary extreme value models. ~~Trends in return levels of GSL~~ These trends are assessed at a mountain massif scale from GSL data, provided for the French Alps from 1959 to 2019 by a meteorological reanalysis and a snowpack model. Our results indicate a temporal decrease in 50-year return levels from 900 m to 4200 m, significant in the Northwest of the French Alps until 2100 m. ~~Despite this decrease , in half of the massifs, the return level~~ We detect the most important decrease at 900 m with an average of -30% for return levels between 1960 and 2010. Despite these decreases, in 2019 ~~at 1800 m exceeds the return level~~ return levels still exceed return levels designed for French building standards under a stationary assumption. At worst, i.e. at 1800 m, return levels exceed standards by 15% on average, and half of the massifs exceeds standards. We believe that ~~this high number of exceedances is~~ these exceedances are due to questionable assumptions concerning the computation of ~~current~~ standards. For example, these were devised with GSL, estimated from snow depth maxima and constant snow density set to $150\ \mathrm{kg\ m^{-3}}$, which underestimate typical GSL values for the ~~full~~ snowpack.

1 Introduction

Extreme snow loads can generate economic damages and casualties. For instance, more than $200 million in roof damages occurred during the Great Blizzard of 1993 (O'Rourke and Auren, 1997). ~~Recently~~In 2006, at the Katowice International Fair~~in 2006~~, the roof of one of the buildings collapsed under a layer of snow, leading to 65 casualties and 140 injured (BBC News, 2006). In France, snow loads over Roussillon in 1986, caused both 17 million euros in damages and a major power outage due to overloading of electrical cables and pylons by sticking snow (Vigneau, 1987; Naaim-Bouvet et al., 2000).

Ground snow load (GSL) is defined as the pressure exerted by accumulated snow on the ground, which can be directly associated ~~to~~ with accumulated snow on structures, e.g. on roofs (Sanpaolesi et al., 1998). In details, the observed height of accumulated snow is called snow depth (in m). The density of this snow can vary widely between precipitation particles ($\rho_{\mathrm{SNOW}} \approx 100\ \mathrm{kg\ m^{-3}}$) and a ripe snowpack ($\rho_{\mathrm{SNOW}} \approx 500\ \mathrm{kg\ m^{-3}}$). Multiplying snow depth by snow density gives the sur-

facic mass of snow (in kg m^{-2}). Surfacic mass of snow corresponds to the snow water equivalent (SWE) which is the height of water (in mm) we could obtain if we melt all the snow in a 1 m^2 area. Indeed, since water density is $\rho_{WATER} = 1000$ kg m^{-3}, we have that 1 mm of water on 1 m^2 has a surfacic mass of 1 kg m^{-2}. Snow load is the pressure exerted by this surfacic mass of snow (in N m^{-2} or Pa) and equals the ~~snow water equivalent~~ SWE times the gravitational acceleration (g = 9.81 m s^{-2}).

Snowpack variables related to GSL (snow depth, ~~snow water equivalent~~SWE) evolve with climate change. As shown in Table 1, literature on past trends in snowpack variables for the Western Alps shows a decreasing trend. Literature on projected trends also points to a decrease (stronger for the second half of the 21st century under a high greenhouse gas emission scenario than with strong reductions in greenhouse gas emissions) for mean winter (December-May) ~~snow water equivalent~~ SWE in the European Alps (IPCC, 2019). However, anthropogenic climate change impacts climatic variables in their averages, but also in their extremes (Klein Tank and Können, 2003; IPCC, 2012). For instance, annual maxima of snow depth have decreased in Switzerland (Marty and Blanchet, 2012). Projected trends in extreme snowpack variables are prone to strong uncertainties (Strasser, 2008; Beniston et al., 2018) as both mean winter temperature (IPCC, 2019) and winter precipitation extremes (Rajczak and Schär, 2017) are projected to increase in the European Alps.

Variable	Indicator	Trend	Country	Time	Source
HS	Seasonal mean (Nov to Apr)	Decrease	CH	1931-1999	Laternser and Schneebeli (2003)
	Winter mean (Dec to Feb)	Decrease in the North	FR	1958-2007	Durand et al. (2009b)
	Mean annual maxima	Decrease	CH	1930-2010	Marty and Blanchet (2012)
	Seasonal mean (Nov to May)	Decrease	IT	1951-2010	Terzago et al. (2013)
	Seasonal mean (Nov to Apr)	Decrease in the South	CH	1961–2012	Schöner et al. (2019)
SWE	1st of April value	Decrease	IT	1965-2007	Bocchiola and Diolaiuti (2010)
	1st of April value	Decrease	FR, IT, CH	1968-2012	Marty et al. (2017)

Table 1. Past trends in snowpack variables, snow depth (HS) and snow water equivalent (SWE), according to existing studies in the Western Alps, i.e. in Italy (IT), France (FR), and Switzerland (CH). In the ~~trend~~ Trend column, "North" and "South" refer to the considered country.

The impact of climate change on GSL was not taken into account in current European standards for structural design, a.k.a Eurocodes (Sanpaolesi et al., 1998), which drive French standards (Biétry, 2005). These standards define that structures must withstand their own weight plus a pressure proportional to a characteristic value. The latter is the stationary 50-year return level of GSL, ~~i.e.~~ exceeded once every 50 years ~~in~~ on average. Thus, studying trends in 50-year return levels of GSL is needed for updating these standards (Croce et al., 2018). In the literature, past and projected trends in 50-year return levels of GSL have rarely been investigated with the exception of (Rózsás et al., 2016; Il Jeong and Sushama, 2018; Croce et al., 2018). In the French Alps, several studies focused on extreme snow variables (Biétry, 2005; Gaume et al., 2012, 2013) and their spatial dependence (Nicolet et al., 2015, 2016, 2017, 2018). However, trends in 50-year return levels of GSL remain unexplored.

We fill these gaps by studying annual maxima of GSL provided every 300 m of ~~elevation~~ altitude at a mountain massif scale for the 23 French Alps massifs. We rely on ~~a reanalysis~~ the SAFRAN-Crocus reanalysis (Vernay et al., 2019) produced by the SAFRAN-Crocus chain (Durand et al., 2009a; Vionnet et al., 2012) available for the period 1959-2019. The major

advantage of this reanalysis is to take benefit of an advanced snowpack model which provides daily estimates of ground snow load values, while previous studies relied on approximate values directly related to snow depth with a crude estimation of snow density (Biétry, 2005). ~~Our~~ Thus, our approach considers only natural snow processes, i.e. we do not account for snow removal throughout the year and consider all processes (accumulation, thaw/freeze, melt, compaction etc.) occurring during the winter season.

Our statistical methodology consists in applying stationary and non-stationary extreme value models to annual maxima time series. We select one model by massif and altitude with the AIC statistical criterion, validate the selected model with the Anderson-Darling test, and assess its significance with the likelihood ratio statistical test. Finally, for each massif and altitude, we compute the relative change ~~in~~ of 50-year return ~~level~~ levels of GSL between 1960 and 2010, and we compare the non-stationary return level in 2019 with the stationary return level designed for French building standards. ~~Our approach considers only natural snow processes, i.e. we do not account for snow removal throughout the year and consider all processes (accumulation, thaw/freeze, melt, compaction etc.) occurring during the winter season.~~

This paper is organized as follows. Section 2 presents our data. Section 3 describes standards for ground snow load. Then, section 4 explains our methodology. Results, discussion and conclusions are introduced in Sections 5, 6 and 7, respectively.

2 Ground snow load data

The study area covers the French Alps which are located between Lake Geneva to the north and the Mediterranean Sea to the south (Fig. 1). The climate is contrasted, colder and wetter in the northern Alps and much drier in the southern Alps (Durand et al., 2009a). This region is typically divided into 23 mountain massifs ~~. In this work, we rely on a~~ of about 1000 km^2. We rely on the SAFRAN-Crocus reanalysis (Vernay et al., 2019) from the SAFRAN-Crocus chain (Durand et al., 2009a; Vionnet et al., 2012) available from August 1958 to July 2019 at the scale of these massifs, ~~for~~ every 300 m ~~elevation bands~~ of altitude from 300 m to 4800 m. ~~However~~Contrary to gridded products, this reanalysis assumes for a given altitude the homogeneity of the different variables at the scale of the massif. Also, annual maxima are ~~only~~ available from 1959 to 2019. Indeed, ~~in this work, annual maxima denotes~~ annual maxima denote the maxima during a year centered on the winter season, e.g. annual maxima for ~~the year~~ 1959 correspond to the maxima from the 1st of August 1958 to the 31st of July 1959.

To sum up, GSL equals SWE from the SAFRAN-Crocus reanalysis times the gravitational acceleration. We study time series of annual maxima of GSL for each massif from 1959 to 2019 every 300 m of altitude from 300 m to 4800 m (Fig. 1).

The SAFRAN-Crocus reanalysis is produced by a chain of two models. First, SAFRAN meteorological reanalysis (Durand et al., 2009a) ~~provides consistent meteorological~~ performs a spatialization of the weather data (precipitation, temperature, humidity, radiation, wind speed) over the ~~considered mountain massifs and elevations. The~~ massifs and altitudes. Then, the Crocus snowpack model (Vionnet et al., 2012) infers snow depth and ~~snow water equivalent (SWE)~~ SWE based on SAFRAN time series. Crocus is a one-dimensional multilayer physical snow scheme, which simulates the snowpack evolution over time, by accounting for several processes such as thermal diffusion, phase changes and metamorphism. ~~A large intercomparison of snow models illustrates that Crocus is among the top models to simulate SWE (Krinner et al., 2018a). Furthermore, another~~

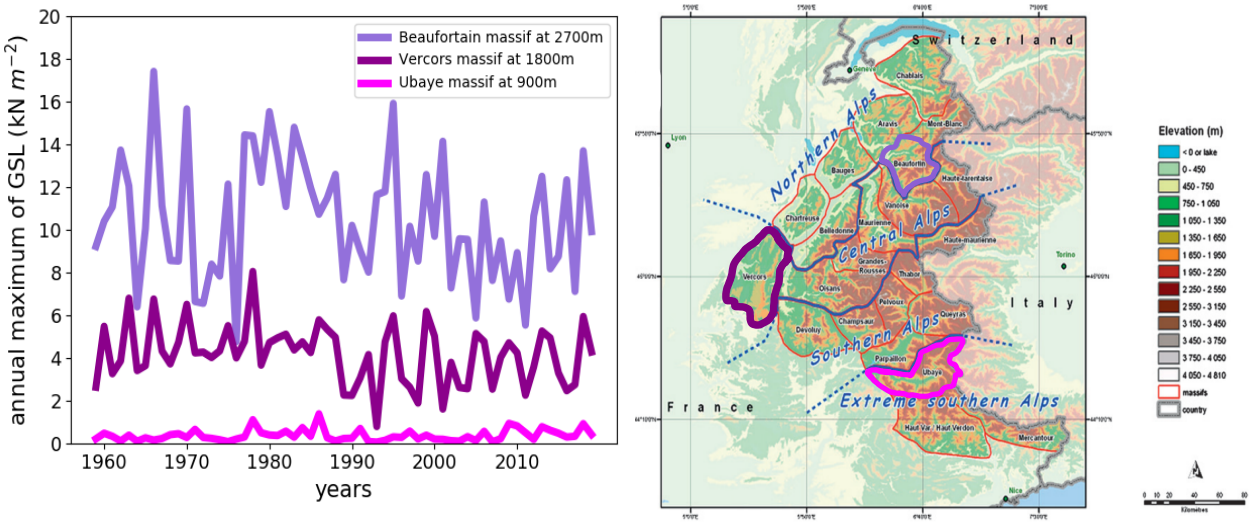

Figure 1. Left: Three time series of annual maxima ~~for~~ of ground snow load (GSL) from 1959 to 2019 for 3 massifs at low (900 m), mid (1800 m) or high (2700 m) altitude. Right: 23 mountains massifs of the French Alps and their orographic features (Durand et al., 2009a).

~~recent intercomparison of SWE products emphasizes Crocus usefulness as it concludes that ensembles containing Crocus (and/or Crocus driven by ERA-Interim reanalysis) perform better than those that do not (?).~~

~~In this paper, GSL equals the SWEfrom the~~ The SAFRAN-Crocus ~~chain times the gravitational acceleration. We study time series of annual maxima of GSL for each massif from 1959 to 2019 for 300 m elevation bands from 300 m to 4800 m (Fig. 1)~~reanalysis has been evaluated against various observation datasets, as reported in previous publications (Lafaysse et al., 2013; Vionnet et a. In most cases, the evaluation is carried out against in-situ snow depth observations and remote sensing snow cover information. For example, Vionnet et al. (2016) evaluated SAFRAN-Crocus snow depth data against 79 observed snow depth data in the French Alps for the 2010-2014 time period, with mean bias and standard error values of 18 cm and 37 cm, respectively. This corresponds to typical values for snow modelling systems applied in various regions on Earth. Because of lower data availability, evaluations against observed SWE values are less frequent than against snow depth data, although we note that Crocus has been shown to perform extremely well compared to other snow cover models, in terms of SWE, across many observation sites worldwide (Krinner et al., 2018b) and SAFRAN-Crocus exhibits satisfying performance in terms of snow depth and SWE in the Pyrenees (Quéno et al., 2016), providing confidence, with respect to other existing datasets, in using this model chain for GSL values. Further model evaluations, using additional datasets, are required to continue assessing and improving the quality of the model chain. Furthermore, we highlight that we only used SAFRAN-Crocus reanalysis values on flat field, and we did not used simulations on slopes, hence it is not relevant to discuss the impact of slope and aspect on the results of this study.

3 Standards for ground snow load in the French Alps

GSL French standards (Biétry, 2005) are based mostly on Eurocodes (Sanpaolesi et al., 1998) and on prior French standards (AFNOR, 1996). Each French department, and by extension each French Alps massif, is associated with a region (C or E) that sets characteristic 50-year return level values of GSL between 200 m and 2000 m of altitude (Fig. 2).

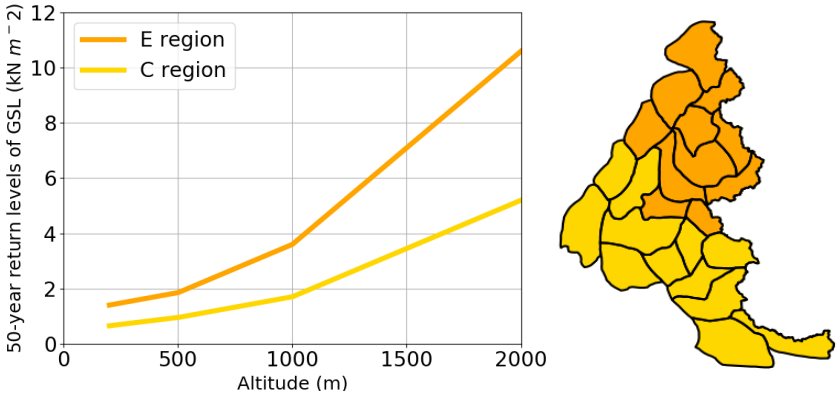

Figure 2. Left: French standards 50-year return levels of ground snow load (GSL) w.r.t~~the~~. altitude for regions C and E. Right: Regions map.

French standards were elaborated with annual maxima time series of snow depth on the ground measured ~~in~~ at stations from 1945 to 1992. GSL data were approximated from annual maxima of snow depth and by assuming that snow density equals 150 kg m^{-3}. Following Eurocodes, the characteristic value of GSL is defined as the 50-year return level of a Gumbel distribution (Sect. ~~4.1~~4). This distribution was fitted using the least squares method and by removing the top annual maximum when considered exceptional (Biétry, 2005) according to a criterion not explicitly mentioned in the French report cited as reference. However, in the Eurocodes, the standard method was to consider the top maximum as exceptional if it was 1.5 times larger than the second largest maximum (Sanpaolesi et al., 1998). In our methodology, we do not remove the top annual maximum.

4 Statistical Methodology

Following extreme value theory, we employ 2 stationary models and 6 non-stationary models for time series of annual maxima of GSL (Sect. 4.1). We select a single model for each time series (i.e. for each massif and altitude) with the AIC statistical criterion, validate this model with the Anderson-Darling test, and assess its significance with the likelihood ratio statistical test (Sect. **??**4.2). Finally, we compute the relative change ~~in~~ of 50-year return levels of GSL ~~, and quantify their uncertainty~~ between 1960 and 2010, quantify the uncertainty of return levels in 2019 to compare them with the stationary return levels designed for French standards (Sect. 4.3).

4.1 Stationary and non-stationary models based on extreme value distributions

Climate extremes are generally studied with statistics. As underlined in the IPCC special report on climate extremes, a large amount of statistical literature builds on extreme indices to examine moderate extremes (IPCC, 2012). However, since we focus on ~~rarer extremes~~ extremes that are more rare, it is recommended to rely on extreme value theory (EVT~~) (Coles, 2001)~~, Coles 2001). Such statistical models provide and hypothesize additional prior information in order to compensate the limited amount of empirical observations that commonly span only several decades. These models can be used to extrapolate beyond the empirical observations, and to estimate return levels (Sect. 4.3).

EVT offers a suitable framework to analyse extreme values, i.e. to model the form of the tail for almost any probability distribution. Asymptotically, as the central limit theorem motivates sample means modelling with the normal distribution, the Fisher–Tippett–Gnedenko theorem (Fisher and Tippett, 1928; Gnedenko, 1943) encourages sample maxima modelling with the GEV distribution. This theorem justifies that the maximum of finite-sized blocks with a large enough block size can be modeled with the GEV distribution. In practice, an annual maximum is thus usually considered as a realization of a GEV distribution. Three parameters define the GEV distribution: a location μ, a scale $\sigma > 0$ and a shape ζ (a.k.a extremal index or tail index). The GEV distribution includes three specific types of distributions: Weibull ($\zeta < 0$), Fréchet ($\zeta > 0$) and Gumbel ($\zeta = 0$). Thus, by definition, if ~~Y represents~~ Z represents an annual maximum of GSL, we can assume that ~~Y~~ Z follows a GEV distribution, i.e. ~~$Y \sim \mathrm{GEV}(\mu,\sigma,\zeta)$~~ $Z \sim \mathrm{GEV}(\mu,\sigma,\zeta)$, which implies that:

$$P(Z \leq z) = \begin{cases} \exp\left[-(1+\zeta\frac{z-\mu}{\sigma})_+^{-\frac{1}{\zeta}}\right] \text{ if } \zeta \neq 0 \text{ and where } u_+ \text{ denotes } \max(u,0), \\ \exp\left[-\exp\left(-\frac{z-\mu}{\sigma}\right)\right] \text{ if } \zeta = 0, \text{ in other words if } Z \sim \mathrm{Gumbel}(\mu,\sigma). \end{cases} \quad (1)$$

In a context of climate change, a large amount of hydrological literature builds on non-stationary modelling (Milly et al., 2008) to assess whether a time series is generated by a unique probability distribution (stationary model), or if the generating probability distribution is changing (non-stationary model). Non-stationary extremes are usually studied with both non-stationary modelling and EVT (Katz et al., 2002). Annual maxima are assumed independent but not necessarily identically distributed (Serinaldi and Kilsby, 2015). Such approaches combine a stationary random component (a fixed extreme value distribution) with non-stationary deterministic functions that map each temporal covariate t to the changing parameters of the distribution (Montanari and Koutsoyiannis, 2014). In a non-stationary context, Zhang et al. (2004) showed that tests based on this parametric approach have stronger power of detection when compared with non-parametric methods.

~~In this paper, we~~ We consider non-stationarity for both the Gumbel distribution and the more general GEV distribution, since they represent natural extensions of the Gumbel distribution which was used for French building standards (Sect. 3). For any model, we have ~~that $Y(t) \sim \mathrm{GEV}(\mu(t),\sigma(t),\zeta(t))$~~ $Z(t) \sim \mathrm{GEV}(\mu(t),\sigma(t),\zeta(t))$, as the Gumbel distribution ~~correspond~~ corresponds to $\zeta(t) = 0$. For a model $\mathcal{M}$, we denote as $\boldsymbol{\theta}_{\mathcal{M}}$ all parameters for its functions ($\mu(t), \sigma(t)$ and $\zeta(t)$). We focus on simple linear functions due to the limited length of time series (60 years). The linearity starts in 1959 which is the first winter with available data. As shown in Table 2, we consider only models with a constant shape parameter, but where the location and/or the scale parameter can vary linearly with years t.

Model type	Distribution	Model name	$\mu(t)$	$\sigma(t)$	$\zeta(t)$	$\boldsymbol{\theta}_{\mathcal{M}}$	$\#\,\boldsymbol{\theta}_{\mathcal{M}}$
Stationary	Gumbel	$\mathcal{M}_0$	μ_0	σ_0	0	(μ_0, σ_0)	2
	GEV	$\mathcal{M}_{\zeta_0}$			ζ_0	$(\mu_0, \sigma_0, \zeta_0)$	3
Non-stationary	Gumbel	$\mathcal{M}_{\mu_1}$	$\mu_0 + \mu_1 \times (t - 1959)$	σ_0	0	(μ_0, μ_1, σ_0)	3
	GEV	$\mathcal{M}_{\zeta_0,\mu_1}$			ζ_0	$(\mu_0, \mu_1, \sigma_0, \zeta_0)$	4
Non-stationary	Gumbel	$\mathcal{M}_{\sigma_1}$	μ_0	$\sigma_0 + \sigma_1 \times (t - 1959)$	0	$(\mu_0, \sigma_0, \sigma_1)$	3
	GEV	$\mathcal{M}_{\zeta_0,\sigma_1}$			ζ_0	$(\mu_0, \sigma_0, \sigma_1, \zeta_0)$	4
Non-stationary	Gumbel	$\mathcal{M}_{\mu_1,\sigma_1}$	$\mu_0 + \mu_1 \times (t - 1959)$	$\sigma_0 + \sigma_1 \times (t - 1959)$	0	$(\mu_0, \mu_1, \sigma_0, \sigma_1)$	4
	GEV	$\mathcal{M}_{\zeta_0,\mu_1,\sigma_1}$			ζ_0	$(\mu_0, \mu_1, \sigma_0, \sigma_1, \zeta_0)$	5

Table 2. Statistical models considered for annual maxima of GSL are based on the Gumbel or the GEV distribution, and are extensions of the stationary Gumbel model. For non-stationary models, the location and/or the scale vary linearly with years t after the starting year 1959.

4.2 Model selection, validation and significance

~~Let $\boldsymbol{y} = (y_{1959}, ..., y_{2019})$ represents~~ **Model selection**. Let $\boldsymbol{z} = (z_{1959}, ..., z_{2019})$ represent a time series of annual maxima of GSL, i.e. for a massif and an ~~elevation band~~ altitude (Sect. 2). First, models are fitted with the maximum likelihood method. For every model $\mathcal{M}$, we compute the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_{\mathcal{M}}$ which corresponds to the parameter $\boldsymbol{\theta}_{\mathcal{M}}$ that maximizes the likelihood:

$$\widehat{\boldsymbol{\theta}}_{\mathcal{M}} = \underset{\boldsymbol{\theta}_{\mathcal{M}}}{\operatorname{argmax}} \mathcal{L}(\boldsymbol{\theta}_{\mathcal{M}}; \boldsymbol{z}) \text{ where } \mathcal{L}(\boldsymbol{\theta}_{\mathcal{M}}; \boldsymbol{z}) = p(\boldsymbol{z}|\boldsymbol{\theta}_{\mathcal{M}}) = \prod_t p(y z_t|\mu(t), \sigma(t), \zeta(t)) = \prod_t \frac{\partial P(Y(t) \leq y_t)}{\partial y_t} \frac{\partial P(Z(t) \leq z_t)}{\partial z_t}. \tag{2}$$

Then, for each ~~$\boldsymbol{yz}$~~, i.e. for each massif and ~~elevation band~~altitude, we select the model $\mathcal{M}_N$ with the minimal AIC value (Akaike, 1974), as it is the best information criterion in a non-stationary context with small sample sizes (Kim et al., 2017). We define~~that~~:

$$\mathcal{M}_N = \underset{\mathcal{M} \in \text{Table2}}{\operatorname{argmin}} \text{AIC}(\mathcal{M}) \text{ where } \text{AIC}(\mathcal{M}) = 2 \times [\#\boldsymbol{\theta}_{\mathcal{M}} - \log \mathcal{L}(\widehat{\boldsymbol{\theta}}_{\mathcal{M}}; \boldsymbol{z})], \text{ where } \#\boldsymbol{\theta}_{\mathcal{M}} \text{ is the cardinality of } \boldsymbol{\theta}_{\mathcal{M}}. \tag{3}$$

The selected model $\mathcal{M}_N$ can be any model from Table 2, i.e. a stationary or a non-stationary model. The subscript N designates the number of additional parameters compared to the stationary Gumbel model $\mathcal{M}_0$, i.e. $N = \#\boldsymbol{\theta}_{\mathcal{M}_N} - \#\boldsymbol{\theta}_{\mathcal{M}_0}$.

~~If~~ **Model validation**. 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 Richard W. Katz (2012) that suggests (i) to transform the data to stationary Gumbel (ii) 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 (App. B). Moreover,

according to the comparative study of Abidin et al. (2012), 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 using Saeb (2018), and found that we cannot reject the null hypothesis (samples generated from the Gumbel model) at the 5% significance level for almost all our selected models (98%), justifying their good fit. We refer to App. B for more details.

Model significance. If the selected model $\mathcal{M}_N$ is not the model $\mathcal{M}_0$ then, since models are nested, we can compute the significance of $\mathcal{M}_N$ w.r.t. $\mathcal{M}_0$ with a likelihood ratio test (Coles, 2001). This test assess whether there is enough evidence to ~~move from~~ reject the stationary Gumbel model $\mathcal{M}_0$ ~~to~~ in favor of the selected model $\mathcal{M}_N$. The null hypothesis can be stated as: the N additional parameters of the model $\mathcal{M}_N$ can be set to zero. In other words, we want to check if setting to zero the N additional parameters of the model $\mathcal{M}_N$ is supported by the data ~~$\boldsymbol{y}$~~$\boldsymbol{z}$. Under the null hypothesis, the likelihood ratio test statistic (LR) has an asymptotic χ^2_N-distribution: ~~$\text{LR}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\mathcal{M}_N}}, \widehat{\boldsymbol{\theta}}_{\boldsymbol{\mathcal{M}_0}}, \boldsymbol{y}) = 2\log \frac{\mathcal{L}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\mathcal{M}_N}}; \boldsymbol{y})}{\mathcal{L}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\mathcal{M}_N}}; \boldsymbol{y})} \dot{\sim} \chi^2_N$~~ $\text{LR}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\mathcal{M}_N}}, \widehat{\boldsymbol{\theta}}_{\boldsymbol{\mathcal{M}_0}}, \boldsymbol{z}) = -2\log \frac{\mathcal{L}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\mathcal{M}_0}}; \boldsymbol{z})}{\mathcal{L}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\mathcal{M}_N}}; \boldsymbol{z})}$ where $\dot{\sim}$ means distributed under suitable regularity conditions.

In practice, the test works as follows. We first choose a 0.05 level of significance. Then, if LR is greater than $q_{\chi^2_N}$, the $1 - 0.05 = 0.95$ quantile of the χ^2_N distribution, ~~it means that~~ we reject the nested model $\mathcal{M}_0$ in favor of the selected model $\mathcal{M}_N$. ~~In this case, if~~ If the selected model $\mathcal{M}_N$ is non-stationary, ~~then~~ we consider the associated trend as significant.

4.3 Return ~~level~~levels

In a stationary context, the T-year return level, which corresponds to a return period of T years, is the classical metric to quantify hazards of extreme events (Cooley, 2012). For a stationary model, there is a one-to-one relationship between a return level (a quantile exceeded each year with probability p) and a return period (a duration exceeded every $T = \frac{1}{p}$ years ~~in~~ on average).

In a non-stationary context, return level and return period concepts (Cooley, 2012) become further ambiguous, prone to misconceptions and can lead to misleading conclusions (Serinaldi, 2015). We focus on the yearly level for a fixed probability of exceedance, a.k.a effective return level (Katz et al., 2002; Cheng et al., 2014), as it conveys best that hazard evolves with time.

For the stationary Gumbel model $\mathcal{M}_0$, the return level $z_p(\boldsymbol{\theta}_{\boldsymbol{\mathcal{M}_0}})$ is defined as the level exceeded each year with probability p. In other words, if ~~Y~~ Z denotes an annual maximum, then ~~$P(Y \leq z_p(\boldsymbol{\theta}_{\boldsymbol{\mathcal{M}_0}})) = 1 - p$~~ $P(Z \leq z_p(\boldsymbol{\theta}_{\boldsymbol{\mathcal{M}_0}})) = 1 - p$. This return level is constant through time and equals $z_p(\boldsymbol{\theta}_{\boldsymbol{\mathcal{M}_0}}) = \mu_0 - \sigma_0 \log(-\log(1-p))$. In this paper, we set $p = \frac{1}{50} = 0.02$ as it corresponds to the 50-year return period defined by French standards (based on European standards) for the design working life of buildings (Sect. 3).

For the selected model $\mathcal{M}_N$, ~~we define our return level~~ the return level is defined as the yearly level for a fixed probability of exceedance p. For any model considered in Table 2, we obtain ~~that~~ $z_p(\boldsymbol{\theta}_{\boldsymbol{\mathcal{M}_N}}, t) = \mu_0 + \mu_1 \times (t - 1959) - \frac{\sigma_0 + \sigma_1 \times (t-1959)}{\zeta_0}[1 - (-\log(1-p))^{-\zeta_0}]$, where we set μ_1, σ_1 or ζ_0 to 0 if they are not defined in the model $\mathcal{M}_N$. For example, for the Gumbel model $\mathcal{M}_0$, the return level is constant: for any year t, $z_p(\boldsymbol{\theta}_{\boldsymbol{\mathcal{M}_0}}, t) = \lim_{\zeta_0 \to 0}[\mu_0 + \frac{\sigma_0}{\zeta_0}(1 - (-\log(1-p))^{-\zeta_0})] = \mu_0 - \sigma_0 \log(-\log(1-p))$.

For any considered model, the time derivative of the return level is constant, as $\frac{\partial z_p(\boldsymbol{\theta}_{\boldsymbol{\mathcal{M}_N}},t)}{\partial t} = \mu_1 - \frac{\sigma_1}{\zeta_0}(1-(-\log(1-p))^{-\zeta_0})$. It quantifies the yearly change ~~in~~ of return level. Thus, the relative difference of return levels between year t_1 and year t_2 is:

$$\text{Relative change}(z_p(\boldsymbol{\theta}_{\boldsymbol{\mathcal{M}_N}},t_1), z_p(\boldsymbol{\theta}_{\boldsymbol{\mathcal{M}_N}},t_2)) = \frac{z_p(\boldsymbol{\theta}_{\boldsymbol{\mathcal{M}_N}},t_2) - z_p(\boldsymbol{\theta}_{\boldsymbol{\mathcal{M}_N}},t_1)}{z_p(\boldsymbol{\theta}_{\boldsymbol{\mathcal{M}_N}},t_1)} = \frac{t_2 - t_1}{z_p(\boldsymbol{\theta}_{\boldsymbol{\mathcal{M}_N}},t_1)} \times \frac{\partial z_p(\boldsymbol{\theta}_{\boldsymbol{\mathcal{M}_N}},t)}{\partial t}. \quad (4)$$

In the context of maximum likelihood estimation, uncertainty related to return levels can be derived by the delta method, which quickly provides confidence intervals both in the stationary and non-stationary case (Coles, 2001; Gilleland and Katz, 2016). First, the return level estimator associated ~~to~~ with the maximum likelihood estimator simply equals $z_p(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\mathcal{M}}})$. Then, due to the asymptotic normality of the maximum likelihood estimator (MLE), we can assume that, even with a finite number of data, the MLE is normally distributed. Therefore, ~~we have that~~ under regularity conditions, limits of the $1-\alpha = 95\%$ confidence interval are $\widehat{\boldsymbol{\theta}}_{\boldsymbol{\mathcal{M}}} \pm q_{\frac{\alpha}{2}} \times v_{z_p}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\mathcal{M}}})$ where $q_{\frac{\alpha}{2}}$ is the 1-$\frac{\alpha}{2}$ quantile of the standard normal distribution, and v_{z_p} is a function that associates to each parameter $\boldsymbol{\theta}_{\boldsymbol{\mathcal{M}}}$ the variance of the approximate normal distribution associated ~~to~~ with its return level $z_p(\boldsymbol{\theta}_{\boldsymbol{\mathcal{M}}})$. For a full expression of the function v_{z_p} and for details on the delta method, we refer to Theorem 2.4 of Coles (2001). In particular, this theorem explains that the delta method is valid for $\zeta_0 < 1$, which is respected in our case as $-0.5 \leq \zeta_0 \leq 0.5$ (Sect. ~~5~~4.4). Also, uncertainty of non-stationary return levels $z_p(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\mathcal{M}}}, t)$ can be obtained by incorporating the covariate t in the function z_p.

4.4 Application

First, we exclude 4 times series of annual maxima with more than 10% of zeros, i.e. years without GSL. Then, we fit models ~~on~~ to time series, and retain only those models with $-0.5 \leq \widehat{\zeta_0} \leq 0.5$. This impacts 3 time series. We ~~made~~ make this choice because $\widehat{\zeta_0} > 0.5$ designates distributions with an "exploding" tail which are known to be physically implausible (Martins and Stedinger, 2000). ~~We~~ Following Sect. 4.2, we select one model for each time series (i.e. for each massif and altitude) with the AIC statistical criterion~~, and assess its significance with the~~. Then, we exclude the 5 times series (2%) where the selected model do not pass the Anderson test. Finally, we assess if the selected model is significantly more appropriate than the stationary Gumbel model $\mathcal{M}_0$ with a likelihood ratio test~~(Sect. ??)~~.

5 ~~Result~~Results

5.1 Selected models

Figure 3 shows that a stationary model, i.e. models $\mathcal{M}_0$ and $\mathcal{M}_{\zeta_0}$, is selected for a majority (57%) of time series studied (Sect. 2). Models with a linearity in both the location and scale parameters are the most frequently selected non-stationary models (22%). For both stationary and non-stationary models, Gumbel models are always more often selected that their corresponding GEV models (Fig. 3, Fig. 4). All in all, we highlight that ~~40~~39% of selected models are significantly ~~different from~~ more appropriate than the stationary Gumbel model $\mathcal{M}_0$.

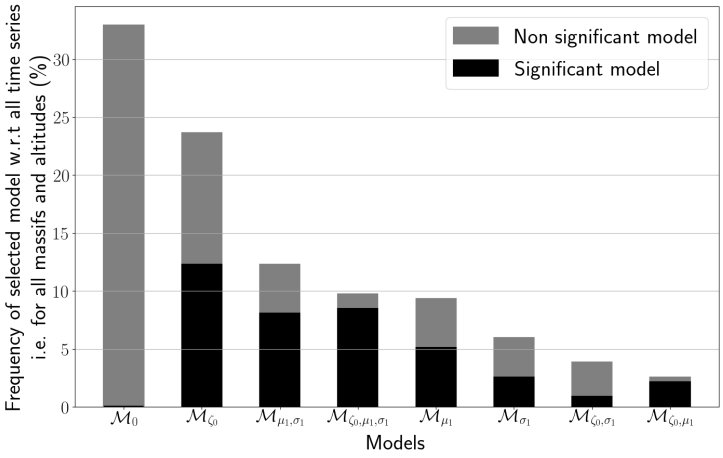

Figure 3. Distribution of selected models. Frequency of selected model (in %) w.r.t. all time series, i.e. for all massifs and altitudes. For the selection procedure and the definition of significance, we refer to Sect. ??4.2.

Figure 4 depicts shape parameter values for the selected ~~model~~ models at 900 m, 1800 m and 2700 m. We notice that a majority of massifs are ~~white-colored~~ white illustrating that a (stationary or non-stationary) Gumbel model (i.e. $\widehat{\zeta_0} = 0$) is selected (Sect. 5). ~~It~~ This emphasizes that a Gumbel distribution often explains more succinctly the data than a GEV distribution. Also, with the GEV distribution, the estimated most likely shape parameter $\widehat{\zeta_0}$ is often quite uncertain, i.e. confidence intervals are large, which is the main reason why French standards did not rely on it. This uncertainty in $\widehat{\zeta_0}$ likely comes from the limited length of time series~~, and longer time series would certainly reduce it, and thus estimate the most likely shape parameter more robustly~~. Therefore, additional data would enable to estimate $\widehat{\zeta_0}$ more robustly, and thus reduce uncertainty.

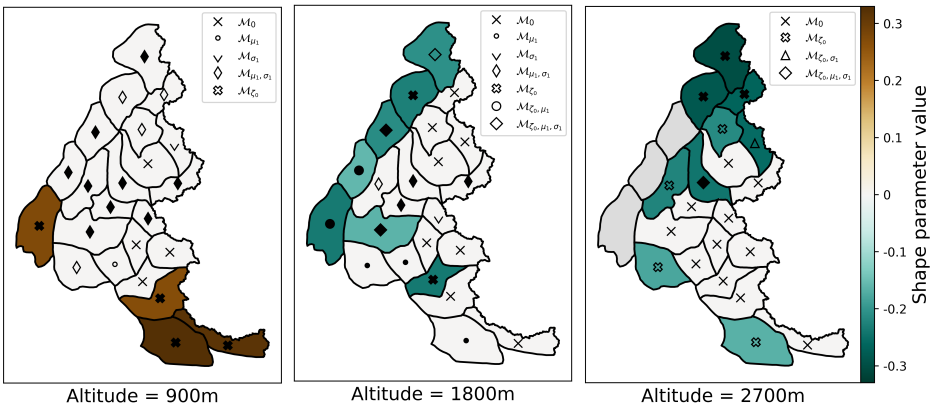

Figure 4. Shape parameter values for the selected models at low (900 m), mid (1800 m) or high (2700 m) altitude. Markers show selected model $\mathcal{M}_N$ while filled markers symbolize models that are significantly better than the Gumbel model $\mathcal{M}_0$ (Sect. 4.2). Grey areas denote either time series that were excluded (Sect. 4.4) or missing data, e.g. when the altitude considered is above the top altitude of the massif.

~~On~~ In Figure 4, we further observe that non-null shape parameters at low altitudes (~~≤~~900 m) are always positive (brown-colored massifs), i.e. a Fréchet distribution is preferred. On the other hand, for high altitudes (~~≥ 1500~~ 1800 m and 2700 m) non-null shape parameters are always negative (green-colored massifs), i.e. a Weibull distribution is favored. ~~We hypothesize that this might be due to~~ Similar results for the shape parameter have been observed for snow depth by Blanchet et al. (2009), Blanchet and Lehning (2010) and Schellander and Hell (2018). This reflects the different nature of annual maxima of GSL between low and high altitudes. At high altitudes, annual maxima are mainly due to snowpack accumulation during several months, while at low altitudes this accumulation is limited, and thus annual maxima roughly correspond to heavy ~~precipitations~~precipitation

~~Shape parameter values for the selected model at low (900 m), mid (1800 m) or high (2700 m) altitude. Markers show selected model $\mathcal{M}_N$. Filled markers symbolize significant trend. For the selection procedure and the definition of significance, we refer to Sect. ??.~~

5.2 Trends in return levels of ground snow load

Figure 5 maps the relative change ~~between 1960 and 2010 for~~ of 50-year return levels of GSL between 1960 and 2010 (Eq. 4) at 900 m, 1800 m and 2700 m (see App. A for maps at all altitudes). Quantitatively, for Northwest massifs, we observe that return levels have decreased by up to 60% at 900 m (dark blue), while at 1800 m this decrease is less marked (pale blue). Qualitatively, these decreasing trends are frequently due to significant changes both in the location and scale parameters of the Gumbel or GEV distribution (small and large diamond-shaped filled markers). At 2700 m, or in the South at 900 m and 1800 m, ~~no trends are found (whitecolor~~we often do not observe any trends (white), since stationary models are ~~often selected (~~selected (small and large cross-shaped markers).

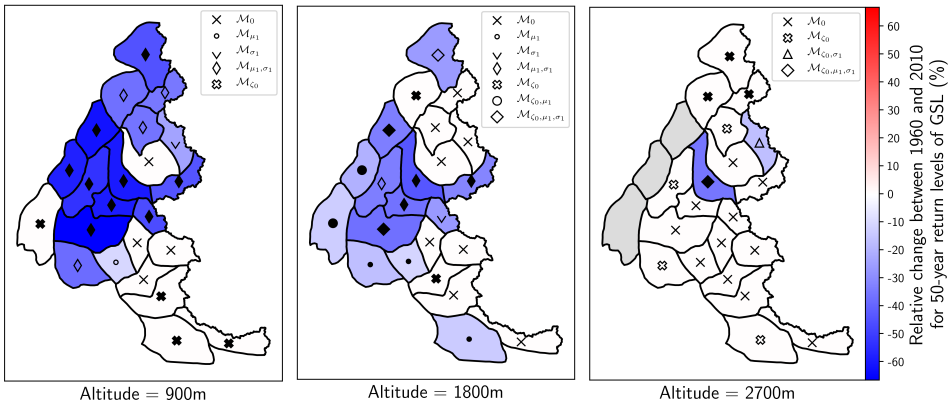

Figure 5. Trends in 50-year return levels of ground snow load (GSL) between 1960 and 2010 at low (900 m), mid (1800 m) or high (2700 m) altitude. Markers show selected model $\mathcal{M}_N$ ~~. Filled~~ while filled markers symbolize ~~significant trend~~models that are significantly better than the Gumbel model $\mathcal{M}_0$ (Sect. ~~Colors illustrate relative change for 50-year non-stationary return levels of GSL between 1960 and 2010. For~~ 4.2). Grey areas denote missing data, e.g. when the ~~selection procedure and~~ altitude considered is above the ~~definition~~ top altitude of ~~significance, we refer to Sect. ??~~the massif.

Figure 6 emphasizes the evolution of decreasing trends between 900 m and 4800 m of altitude. We observe that decreasing trends are significant for more than one-third of the massifs, located in the Northwest of the Alps (~~Sect~~App. A), until 2100 m (black bars). In half a century, return levels have dropped ~~in~~ on average by up to 30% at 900 m. ~~At higher altitudes~~ Until 3300 m, we observe a decline in the percentage of massifs with a significant decreasing trend~~, which is non null up to~~ . Above 3300 m~~. However, for~~ , we do not find any significant decreasing trend. For both the relative decrease and the percentage of massifs with a decreasing trend, ~~even if~~ we notice a similar declining pattern~~, we also detect a slight growth above~~ . We also notice more decline between 3300 m and 3900 m than at 3000 m~~. This~~ , which echoes results from Lüthi et al. (2019), ~~which~~ who found that, in the Alps above 3000 m, the relative decrease for projected winter-mean of fresh ~~snow water equivalent~~ SWE is more marked than at 3000 m (see their Figure 8). We emphasize, however, that most meteorological observations used as input to the SAFRAN-Crocus reanalysis are situated below 2000 m. Therefore, trends beyond 2000 m altitude should be considered with great caution.

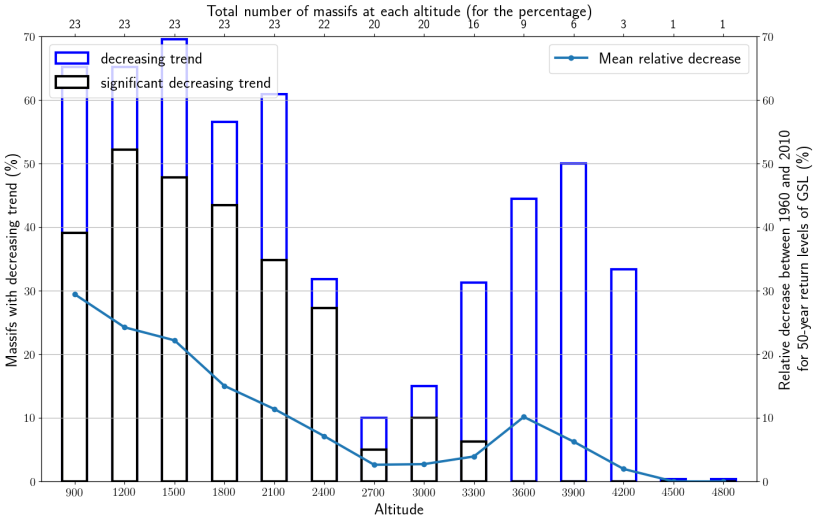

Figure 6. Temporal decreasing trend of 50-year return levels of ground snow load (GSL) between 900 m and 4800 m of altitude.

~~Left: Time series of annual maxima for ground snow load (GSL) for 3 massifs either at 300 m or 600 m of altitude. Right: Trend for annual maxima of ground snow load (GSL) at 300 m and 600 m of altitude. Markers show selected model $\mathcal{M}_N$.~~

Figure 7 illustrates that, for altitudes 300 m and 600 m, in general no ~~trend~~ trends are found except few decreasing trends at 600 m~~and 3~~ , and 2 time series (1 at 300 m, ~~2~~ 1 at 600 m) with ~~sometimes~~ important increasing trends (+100% for ~~the Parpaillon~~ one massif at 600 m). Despite this important increase in relative change, ~~recorded~~ annual maxima of snow load remain small (< 1 kN m^{-2}). Indeed, we found that these annual maxima correspond to snow load accumulated in few days, and thus are mainly driven by heavy ~~precipitations~~ precipitation rather than a ~~full season~~ seasonal snowpack accumulation. In particular, we hypothesize that the ~~2 massifs with important increasing trends (red color~~ important increasing trend observed in the South at 600 m (color red) might be ~~caused by a local phenomenoncalled "East return" which is a low pressure system coming from the~~

~~Mediterranean sea~~cause by a regional phenomenon, resulting from Mediterranean humid air masses flowing northward into the North of Italy and then westward to the eastern part of the French Alps, that might be intensifying with global warming ~~(Garavaglia et al., 2010; Faranda, 2019)~~(Garavaglia et al., 2010; Gottardi et al., 2012; Faranda, 2019).

To sum up trends in return levels of ground snow load: from 900 m to 4800 m, either no ~~trend or decreasing trend~~ trends or decreasing trends of 50-year return levels of GSL are found (Fig. 5, Fig. 6~~and~~ , Fig. A1)~~.~~ , while at 300 m and 600 m, no clear trends are found (Fig. 7).

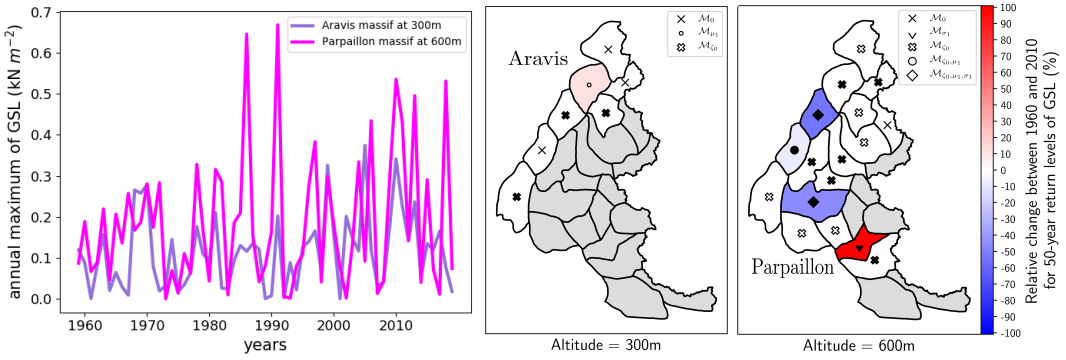

Figure 7. Left: Time series of annual maxima for ground snow load (GSL) for 2 massifs (Aravis and Parpaillon) either at 300 m or 600 m of altitude. Right: Trend for annual maxima of ground snow load (GSL) at 300 m and 600 m of altitude. Markers show selected model $\mathcal{M}_N$ while filled markers symbolize models that are significantly better than the Gumbel model $\mathcal{M}_0$ (Sect. 4.2). Grey areas denote either time series that were excluded (Sect. 4.4) or missing data, e.g. when the altitude considered is above the top altitude of the massif.

5.3 Comparison of return levels of ground snow load with French standards

~~Every 300 m of altitude, from 300 m to 1800 m, we compute~~ We compare 50-year return levels of GSL and their uncertainty ~~from GSL data~~ (Sect. 4.3) ~~, and compare them~~ to French standards first for 2 massifs (Fig. 8) then globally (Fig. 9). We consider GSL data from 300 m to 1800 m because standards are defined from 200 m to 2000 m (Sect. 3)~~for the 23 French Alps massifs~~.

Figure 8 illustrates these levels and their uncertainty for two massifs (Vercors and Beaufortain) associated ~~to~~ with different French standards regions. Standards are often exceeded at higher altitudes (e.g. at 1800 m). Also, Figure 8 exemplifies the impact of accounting for decreasing trends in return levels. Indeed, we observe that return levels from the stationary Gumbel model $\mathcal{M}_0$ (~~Left~~left) are often ~~above~~ larger than effective return levels in 2019 (last year of data) from the selected model $\mathcal{M}_N$ (~~Right~~right).

Figure 9 sums up the comparison between French standards and ~~our results~~50-year return levels for all 23 massifs. We display (i) the percentage of massifs whose return level ~~estimated from data exceeds return level from standards. The number of massifs with available data is equal to 11 at 300 m, 18 at 600 m, and 23 at 900 m and above~~exceeds standards, and (ii) the mean relative difference between return levels and standards. Limits of the confidence intervals ~~(black bars) are computed~~ are approximated as the percentage of exceedances (resp. mean relative difference) for the limits of return levels' 95% confidence

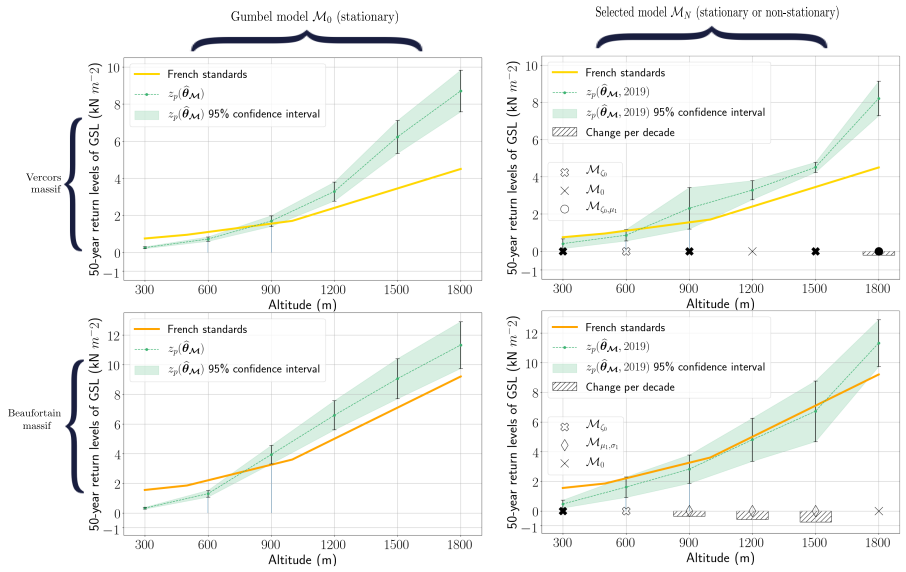

Figure 8. 50-year return levels of ground snow load (GSL) from altitude 300 m to 1800 m for Vercors (top) and Beaufortain (bottom) massifs. Return levels (green line) and their uncertainty (shaded green and black bars) are estimated from the data either with the stationary Gumbel model $\mathcal{M}_0$ (left) or with the selected model $\mathcal{M}_N$ (right). If $\mathcal{M}_N$ is a non-stationary model, the return level is the effective return level in 2019, and we display the change of return levels per decade (striped histogram), i.e. 10 times the time derivative of return level (Sect. 4.3).

interval (black bars ~~on~~ in Fig. 8) ~~..~~ and displayed with black bars (resp. shaded blue). The number of massifs considered is equal to 7 at 300 m, 17 at 600 m, and 23 at 900 m and above.

First, if we estimate return levels from data with the French standards method (Fig. 9 ~~Left~~left), i.e. with a stationary Gumbel model $\mathcal{M}_0$, and GSL data approximated with snow depth obtained from reanalysis and $\rho_{\text{SNOW}} = 150$ kg m^{-3}, then we observe few exceedances (~~always~~ less than 10%) and that on average return levels remains below standards, as the mean relative difference remains below zero. Thus, in this setting, estimations from our reanalysis are consistent with French standards.

However, if we consider the actual GSL, i.e. computed with ~~the snow water equivalent from~~ SWE from the reanalysis, then French standards drastically underestimate return levels. Indeed, with a stationary Gumbel model $\mathcal{M}_0$, then for altitudes above or equal to 900 m, French standards are exceeded for a majority of massifs (Fig. 9 ~~Center~~center). But, if we consider the selected model $\mathcal{M}_N$, i.e. if we account for the decreasing trend in 50-year return levels, we have less exceedances at all altitudes (Fig. 9 ~~Right~~right). In the latter case, at worst, i.e. at 1800 m, return levels exceed standards by 15% on average, and half of the massifs ~~still exceed French standardsreturn levels.~~

~~50-year return levels of ground snow load (GSL) from altitude 300 m to 1800 m for Vercors (Top) and Beaufortain (Bottom) massifs. Return levels (green line) and their uncertainty (shaded green and black bars) are estimated from the data either with the stationary Gumbel model $\mathcal{M}_0$ (Left) or with the selected model $\mathcal{M}_N$ (Right). If $\mathcal{M}_N$ is a non-stationary model, return~~

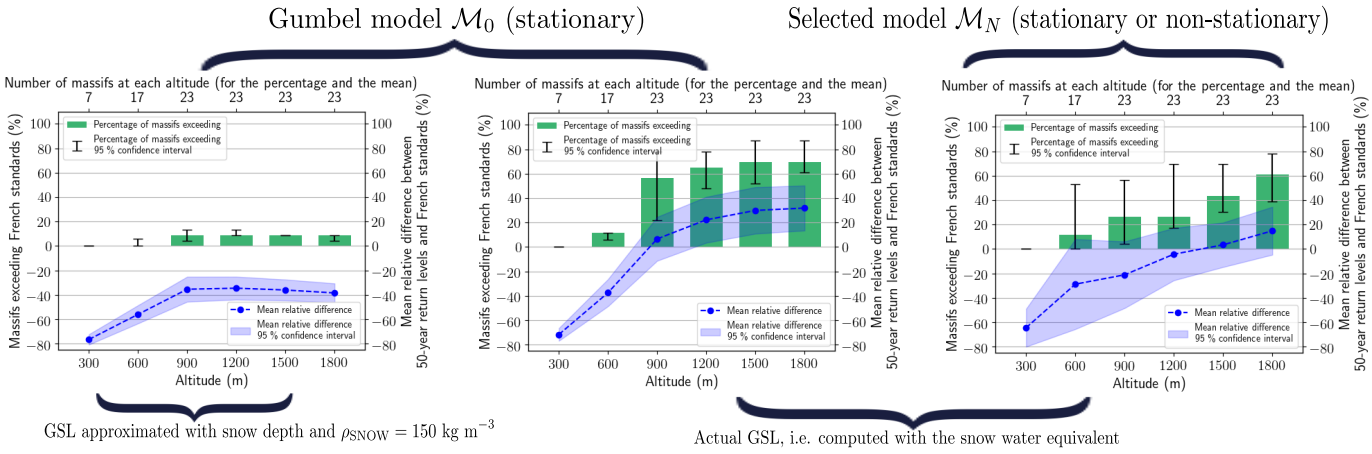

Figure 9. Comparison of 50-year return levels of ground snow load (GSL) with French standards between 300 m and 1800 m. We show the percentage of massifs (green histogram) whose return levels exceed French standards, and the mean relative difference (blue line) between return levels and standards. Left: similar to French standard estimation (stationary Gumbel $\mathcal{M}_0$, and GSL approximated with snow depth obtained from the reanalysis and $\rho_{\mathrm{SNOW}} = 150$ kg m^{-3}). Center: stationary Gumbel $\mathcal{M}_0$, and actual GSL, i.e. computed with SWE from the reanalysis. Right: selected model $\mathcal{M}_N$ (if $\mathcal{M}_N$ is non-stationary, the return level is the effective return level in 2019) and actual GSL.

~~level is the effective return level in 2019, and we display the change in return level per decade (striped histogram), i.e. 10 times the time derivative of return level (Sect. 4.3).~~

~~Percentage of massifs (green histogram) whose 50-year return levels of ground snow load (GSL) is above French standards between 300 m and 1800 m. Uncertainty (black bars) is approximated as the percentage of exceedances for return levels' 95% confidence interval limits. Left: similar to French standard estimation (stationary Gumbel $\mathcal{M}_0$, and GSL approximated with snow depth obtained from reanalysis and $\rho_{\mathrm{SNOW}} = 150$ kg m^{-3}). Center: stationary Gumbel $\mathcal{M}_0$, and actual GSL, i.e. computed with the snow water equivalent from reanalysis. Right: selected model $\mathcal{M}_N$ (if $\mathcal{M}_N$ is non-stationary, return level is the effective return level in 2019) and actual GSL.~~ (60%) exceeds standards.

Furthermore, despite that uncertainty intervals (black bars) can be large, it does not impact the main conclusions of this article. Indeed, in Figure 9 right at 1800 m, we still have between 40% and 80% of massifs exceeding French standards.

6 Discussion

6.1 Methodological considerations

We discuss in depth the statistical models chosen for this study. It is well-known that an annual maximum based approach can be wasteful in terms of data (Coles, 2001). However, since our objective is to estimate 50-year return levels and since we have 60 years of data, we still decide to rely on the annual maximum based approach (with the GEV distribution) rather than on the concurrent approach based on threshold exceedances (with the Generalized Pareto distribution). Also, with the

GEV distribution~~(and its particular case: the Gumbel distribution)~~ , our methodology is a direct extension of French building standards (Sect. 3).

For the non-stationary models, we focus on simple deterministic ~~function~~ functions of time ($\mu(t), \sigma(t)$, $\zeta(t)$) due to the limited length of time series. A linear non-stationarity seems preferable to a non-stationarity based on the Heaviside step function due to the continuous nature of climate change. We start the linear non-stationary at the initial year, i.e. 1959. ~~We tried to start the non-stationarity only after a most likely year (Blanchet et al., 2016) but our results lacked coherence both between close massifs (no clear spatial pattern for the most likely years) and w.r.t. the altitude (no clear trend in most likely years as altitude rises).~~

We decided to consider non-stationarity only for the location and scale parameter. Indeed, in the literature, a linear non-stationarity is considered sometimes only for the location parameter (Fowler et al., 2010; Tramblay and Somot, 2018) but more often both for the location and the scale (or log-transformed scale for numerical reasons) parameters (Katz et al., 2002; Kharin and Zwiers, 2004; Marty and Blanchet, 2012; Wilcox et al., 2018). ~~Another reason for which we considered~~ Also, we consider a non-stationarity for both parameters ~~is that we did not find~~ because the scale parameters ~~to be~~ were not proportional to the location parameters, which could have otherwise simplified our parametrization. ~~Also~~Finally, the shape parameter is typically considered constant ~~, with few exceptions~~ in the literature, and we follow this approach.

For time series containing zeros, French standards rely on a mixed discrete-continuous distribution. ~~We did not rely on this choice because for time series containing zeros considered in this study, i.e. with less than 10%~~ They fit both a Gumbel distribution on non-zero annual maxima and the probability of having a non-zero annual maxima. However, with our reanalysis data, this approach sometimes leads to fitting non-stationary extreme value models with less than 40 non-zero annual maxima. Therefore, we rather decide to exclude any time series with more than 10% of zeros (Sect. ~~5), we almost obtain the same~~ 4.4), to ensure that we fit models with more than 55 non-zero annual maxima. In practice, our approach gives 50-year ~~return levels with this distribution than with our approach~~ return levels close to the approach from French standards (absolute difference remains lower than 0.1 kN m^{-2}).

6.2 On the limitation to approximate annual maxima of ground snow load with annual maxima of snow depth

~~Snow water equivalent (SWE)~~ SWE times the gravitational constant equals GSL. However, most countries do not measure SWE but only have access to snow depth (HS) (Haberkorn et al., 2019). In that case, snow density is required to obtain SWE (and subsequently GSL) from HS (Sect. 1). In particular, French standards approximate annual maxima of GSL with annual maxima of HS and by assuming a constant snow density, equal to $\rho_{\text{SNOW}} = 150$ kg m^{-3}. ~~We~~ In Figure 10, we highlight limitations of such approaches with our reanalysis ~~(Sect. 2)~~ that provides, for the whole snowpack, daily values of SWE, HS, and thus of snow density.

We find that annual maxima of GSL ~~is~~ are always underestimated by French standards' approximation (Fig. 10 ~~Left). Main left).~~ The main reason is that, when annual maxima of GSL ~~is~~ are reached, snow density is ~~in~~ on average largely superior to $\rho_{\text{SNOW}} = 150$ kg m^{-3} (Fig. 10 ~~Center~~center). Indeed, we observe that at the time of the annual maxima of GSL the snow density is around ≈ 350 kg m^{-3} ~~in~~ on average at 2700 m, and close to ≈ 250 kg m^{-3} ~~in~~ on average at 900 m. Finally, despite

high variations along the years, we also notice that, when annual maxima of GSL ~~is~~ are reached, snow depth can be much lower than the annual maxima of snow depth (Fig. 10 ~~Right~~right), which is another argument against the use of snow depth maxima as a proxy for GSL maxima.

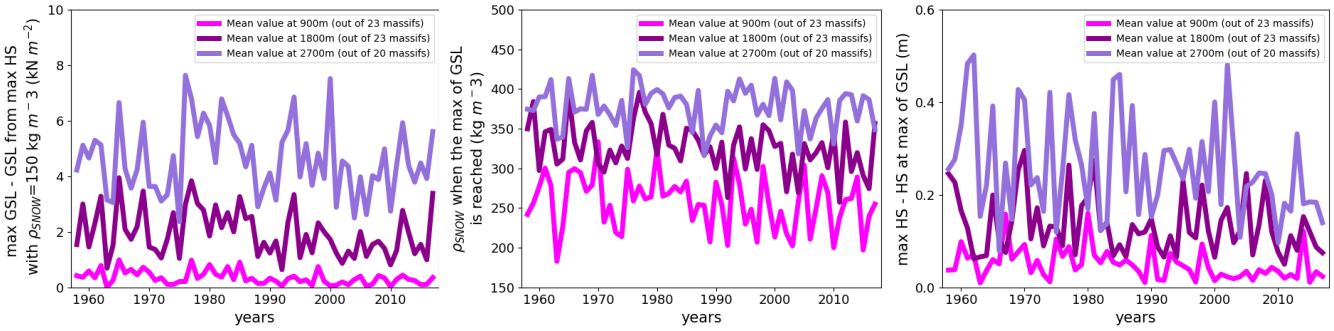

Figure 10. Limitation of approximating annual maxima of ground snow load (GSL) from annual maxima of snow depth (HS). Left: Difference between annual maxima of GSL and GSL computed from annual maxima of HS and $\rho_{\mathrm{SNOW}} = 150$ kg m^{-3}. Center: Snow density when annual maxima of GSL ~~is~~ are reached. Right: Difference between annual maxima of HS and HS when annual maxima of GSL ~~is~~ are reached.

7 Conclusions

Based on both a reanalysis and a snowpack model, we detect an overall temporal decreasing trend ~~w.r.t to~~ of 50-year return levels of ground snow load (GSL) between 900 m and 4200 m, which is significant until 2100 m in the Northwest of the French Alps. This confirms other studies in the Western Alps which also found overall decreasing trends in linked snowpack variables: ~~snow water equivalent~~ SWE and snow depth. ~~Despite decreasing return levels , in half of the massifs the 50-year return level in~~ The largest decrease is found at 900 m with -30% for return levels between 1960 and 2010. Despite these decreases, in 2019 ~~at 1800 m exceeds the stationary return level~~ return levels still exceed return levels designed for French building standards under a stationary assumption. At worst, i.e. at 1800 m, return levels exceed standards by 15% on average, and half of the massifs exceeds standards.

We hypothesize that this amount of exceedances might be due to an underestimation of GSL by French standards. Indeed, these standards were devised with GSL estimated from snow depth maxima and constant snow density equal to 150 kg m^{-3}, which underestimate typical GSL values for the ~~full~~ snowpack. Another reason for these exceedances might be ill-designed relationships between altitude and snow load. As shown ~~on~~ in Fig. 2, French standards return levels ~~augment linearly by parts w.r.t the altitude~~ increase linearly with altitude in three steps. Indeed, French standards (Biétry, 2005) follow previous national standards (AFNOR, 1996) that advised for a linear relationship between altitude and snow load instead of relying on European standards' results that showed a quadratic relationship for the Alpine Region (Sanpaolesi et al., 1998). Thus, at higher altitudes, French standards ~~might~~ underestimate actual return levels which might explain the augmenting percentage of exceedance ~~that we observe w.r.t~~ observed with the altitude (Fig. 9 ~~Right~~right).

Many potential extensions of this work could be considered. First, our methodology could be extended with more advanced definitions of non-stationary return levels (Rootzén and Katz, 2013; Serinaldi, 2015). Also, instead of considering time series of annual maxima as spatially independent, we believe that our analysis may benefit from an explicit modelling of the spatial dependence between extremes. Then, reanalyses are increasingly available at the European scale (e.g. ~~Soci et al. (2016)~~ Soci et al. 2016), which could be used for extending this work ~~at~~ to a wider geographical scale. This requires, however, to remain cognizant of the limitations of such reanalyses, in particular (i) the temporal heterogeneity of the meteorological data input to these reanalyses (Vidal et al., 2010), (ii) the lack of observations at high ~~elevations~~altitudes, requiring caution in analyzing trends for high ~~elevation~~ altitude locations and (iii) model errors (e.g. ~~,~~ snowpack model errors) which need to be taken into account when analyzing the results.

Finally, even if, according to our analysis, GSL exceeds French standards return levels in the French Alps, (Fig. 9 ~~Right~~right), few destructions related to snow loads actually occurred. Several reasons might explain that. First, French standards consider a coefficient that maps GSL return levels to roof snow load return level, i.e. multiplication by a coefficient that summarizes several roof features: shape, exposure and thermal transmission (Sanpaolesi et al., 1998)~~)~~. This coefficient might be overprotective. Also, following European standards, roof designers must add safety coefficients to ensure roofs~~reliability, designers~~ ' reliability. Indeed, they actually build roofs that withstand ~~return levels of~~ the sum of (i) the characteristic value of permanent action, i.e. self-weight, multiplied by a safety coefficient equal to 1.35 and (ii) the characteristic value of variable action, i.e. roof snow load return level, multiplied by a safety coefficient ~~(~~equal to 1.5 (Sanpaolesi et al. 1998 Eq. 8). Above all, French standards do not take into account that, after intense days of snowfall, the snow accumulated on the roof either slides off or is removed. In that case, the main risk lies in extreme snow events that might accumulate in few days enough snow to exceed French standards. Undeniably, most known snow load destructions resulted from such intense snow events, sometimes combined with liquid precipitation that often heavily increase snow load. The response of these short but extreme and complex snow events to climate change might be an interesting topic for future research.

Author contributions. E. L. R. performed the analysis and drafted the first version of the manuscript. All authors discussed the results and edited the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Data availability. The dataset can be download from AERIS website https://www.aeris-data.fr/catalogue/ (type "S2M" in the search bar).

Acknowledgements. ~~ELR~~ E. L. R. holds a PhD grant from INRAE. The S2M data are provided by Météo-France - CNRS, CNRM Centre d'Etudes de la Neige, through AERIS. We are grateful to Eric Gilleland for ~~its~~ his "extRemes" R package. Finally, we are indebted to Jacques Biétry for providing us the report on French standards w.r.t. ground snow load and for his explanations on their methodology.

Appendix A: Trends in return levels of ground snow load

In this section, we report, for every 300 m of altitude from 900 m to 4200 m, the map of the relative change ~~between 1960 and 2010 for the~~ of 50-year return levels of GSL between 1960 and 2010 (Fig. A1). Trends at 4500 m and 4800 m are not reported, since they only concern the Mont Blanc massif, where no significant trend is inferred at these altitudes.

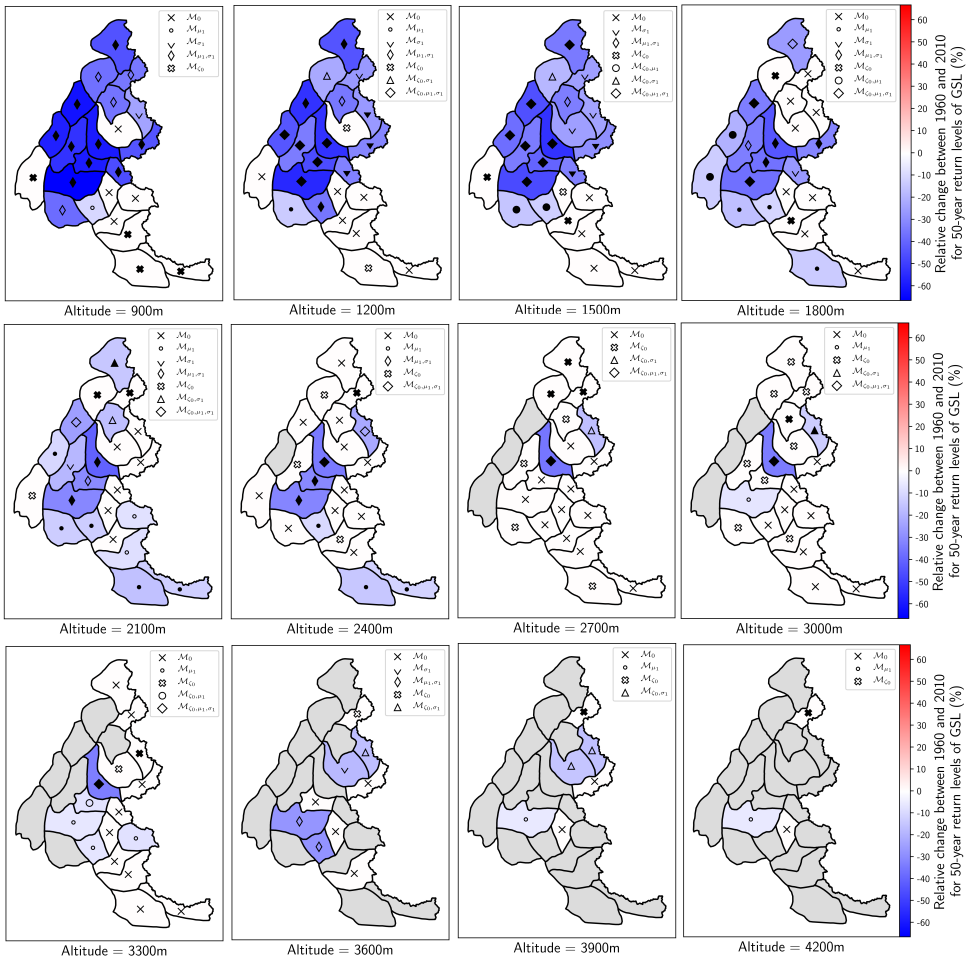

Figure A1. Trends in return levels of ground snow load (GSL) between 900 m and 4200 m of altitude. Markers show selected model $\mathcal{M}_N$ while filled markers symbolize models that are significantly better than the Gumbel model $\mathcal{M}_0$ (Sect. 4.2). Grey areas denote either time series that were excluded (Sect. 4.4) or missing data, e.g. when the altitude considered is above the top altitude of the massif.

Appendix B: Detailed methodology for the model validation

Quantile-quantile plot. Standard diagnosis tools for both stationary and non-stationary extreme value models (Coles, 2001; Richard W. Ka
rely on a probability integral transformation f to the standard Gumbel distribution, i.e. Gumbel$(0,1)$. Indeed, if $Z(t) \sim \text{GEV}(\mu(t),\sigma(t),\zeta(t)$
then $f(Z(t)) = \frac{1}{\zeta(t)}\log(1+\zeta(t)\frac{Z(t)-\mu(t)}{\sigma(t)}) \sim \text{Gumbel}(0,1)$. Thus, if $\boldsymbol{z} = (z_{1959},\ldots,z_{2019})$ represent a time series of annual maxima, then let $\tilde{z}_{1959} = f(z_{1959}),\ldots,\tilde{z}_{2019} = f(z_{2019})$.

Quantile-quantile (Q-Q) plot is a standard diagnosis based on the comparison of empirical quantiles (computed from the empirical distribution) and theoretical quantiles (computed from the expected distribution). On one hand, $\tilde{z}_{(1)},\ldots\tilde{z}_{(61)}$ are the empirical quantiles, which correspond to the ordered values of the $\tilde{z}_t$. On the other hand, $-\log(-\log(\frac{1}{62})),\ldots,-\log(-\log(\frac{61}{62}))$
correspond to the theoretical quantiles. Indeed, if $\tilde{Z} \sim \text{Gumbel}(0,1)$, then $P(\tilde{Z} \leq \tilde{z}) = \exp -e^{-\tilde{z}} = \frac{i}{62} \leftrightarrow \tilde{z} = -\log(-\log(\frac{i}{62}))$.
Thus, the Q-Q plot is comprised of the pairs $\{(-\log(-\log(\frac{i}{62})),\tilde{z}_{(i)}); i = 1,\ldots,61\}$.

In Figure B1, we display Q-Q plots for the three time series of annual maxima of GSL displayed in Fig. 1. We observe that the left and the right Q-Q plots show a good fit, as the points stay close to the line. However, for the center Q-Q plot, all points are close to the line, except the highest empirical quantile that is largely above the corresponding theoretical quantile. As a whole, when observing all Q-Q plots (not shown) most time series show a good fit, except few time series (less than 10) which have a pattern similar to the center Q-Q plot in Figure B1.

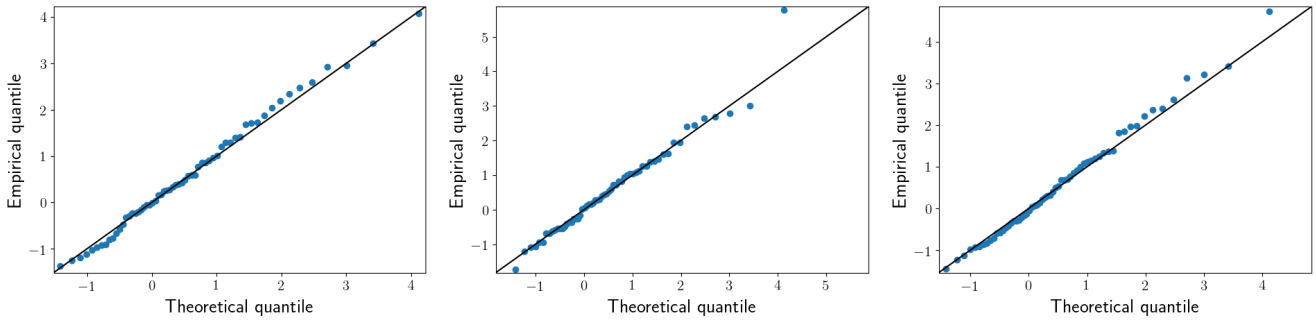

Figure B1. Q-Q plots of the selected models for the three time series displayed in Fig. 1. Left: Ubaye massif at 900m fitted with the model $\mathcal{M}_{\zeta_0}$. Center: Vercors massif at 1800m fitted with the model $\mathcal{M}_{\zeta_0,\mu_1}$. Right: Beaufortain massif at 2700m fitted with the model $\mathcal{M}_{\zeta_0}$.

Anderson-Darling test. Q-Q plot is a qualitative tool to validate the goodness-of-fit for probability models. For the quantitative validation of the goodness-of-fit of the selected models, we rely on the Anderson-Darling statistical test, which is the most powerful test for the Gumbel distribution according to the comparative study of Abidin et al. (2012).

In practice, with this test, we assess if the transformed annual maxima $\tilde{z}_{(1)}, ..., \tilde{z}_{(61)}$ are likely to be generated from a standard Gumbel distribution. Let $n = 61$ denotes the number of samples, and F_{emp} (resp. F_{gum}) denotes the cumulative distribution function of the empirical (resp. standard Gumbel) distribution. Then, Anderson-Darling test is based on the distance:

$$A^2 = n \int (F_{\text{emp}}(x) - F_{\text{gum}}(x))^2 w(x) dF_{\text{gum}}(x) \approx -\sum_{i=1}^{n} \frac{2i-1}{n} \{\log[F_{\text{gum}}(\tilde{z}_{(i)})] + \log[1 - F_{\text{gum}}(\tilde{z}_{(n+1-i)})]\} - n. \quad \text{(B1)}$$

where $w(x)$ places more weight on the tail of the standard Gumbel distribution. For details, we refer to Abidin et al. (2012).

We apply this test on the transformed data using Saeb (2018), and found that we cannot reject the null hypothesis (samples generated from the Gumbel model) at the 5% significance level for almost all our selected models (98%), justifying their good fit. As explained in Sect. 4.4, we exclude time series whose selected models do not pass this Anderson-Darling test.

References

Abidin, N. Z., Adam, M. B., and Midi, H.: The Goodness-of-fit Test for Gumbel Distribution: A Comparative Study, Matematika, 28, 35–48, http://www.matematika.utm.my/index.php/matematika/article/view/313, 2012.

AFNOR: DTU P06 006, Tech. rep., AFNOR, 1996.

Akaike, H.: A New Look at the Statistical Model Identification, IEEE transactions on automatic control, pp. 215–222, https://doi.org/10.1007/978-1-4612-1694-0_16, 1974.

BBC News: Arrests over Poland roof collapse, http://news.bbc.co.uk/2/hi/europe/5119424.stm, 2006.

Beniston, M., Farinotti, D., Stoffel, M., Andreassen, L. M., Coppola, E., Eckert, N., Fantini, A., Giacona, F., Hauck, C., Huss, M., Huwald, H., Lehning, M., López-Moreno, J.-I., Magnusson, J., Marty, C., Terzago, S., and Vincent, C.: The European mountain cryosphere: a review of its current state, trends, and future challenges, The Cryosphere, 12, 759–794, https://doi.org/10.5194/tc-12-759-2018, 2018.

Biétry, J.: Charges de neige au sol en France : proposition de carte révisée, Tech. Rep. 0, groupe de travail « Neige » de la commission de normalisation BNTEC P06A, 2005.

Blanchet, J. and Lehning, M.: Mapping snow depth return levels: Smooth spatial modeling versus station interpolation, Hydrology and Earth System Sciences, 14, 2527–2544, https://doi.org/10.5194/hess-14-2527-2010, http://www.hydrol-earth-syst-sci.net/14/2527/2010/, 2010.

Blanchet, J., Marty, C., and Lehning, M.: Extreme value statistics of snowfall in the Swiss Alpine region, Water Resources Research, 45, https://doi.org/10.1029/2009WR007916, http://doi.wiley.com/10.1029/2009WR007916, 2009.

Blanchet, J., Molinié, G., and Touati, J.: Spatial analysis of trend in extreme daily rainfall in southern France, Climate Dynamics, 51, 799–812, https://doi.org/10.1007/s00382-016-3122-7, 2016.

Bocchiola, D. and Diolaiuti, G.: Evidence of climate change within the Adamello Glacier of Italy, Theoretical and Applied Climatology, 100, 351–369, https://doi.org/10.1007/s00704-009-0186-x, 2010.

Cheng, L., AghaKouchak, A., Gilleland, E., and Katz, R. W.: Non-stationary extreme value analysis in a changing climate, Climatic Change, 127, 353–369, https://doi.org/10.1007/s10584-014-1254-5, 2014.

Coles, S. G.: An introduction to Statistical Modeling of Extreme Values, Springer Series in Statistics, p. 221, https://doi.org/10.1007/978-1-4471-3675-0, 2001.

Cooley, D.: Return Periods and Return Levels Under Climate Change, in: Extremes in a Changing Climate - Detection, Analysis & Uncertainty, pp. 97–114, Springer Science & Business Media, https://doi.org/10.1007/978-94-007-4479-0, 2012.

Croce, P., Formichi, P., Landi, F., Mercogliano, P., Bucchignani, E., Dosio, A., and Dimova, S.: The snow load in Europe and the climate change, Climate Risk Management, 20, 138–154, https://doi.org/10.1016/j.crm.2018.03.001, 2018.

Durand, Y., Laternser, M., Giraud, G., Etchevers, P., Lesaffre, B., and Mérindol, L.: Reanalysis of 44 yr of climate in the French Alps (1958-2002): Methodology, model validation, climatology, and trends for air temperature and precipitation, Journal of Applied Meteorology and Climatology, 48, 429–449, https://doi.org/10.1175/2008JAMC1808.1, 2009a.

Durand, Y., Rald Giraud, G., Laternser, M., Etchevers, P., Mé Rindol, L., and Lesaffre, B.: Reanalysis of 47 Years of Climate in the French Alps (1958–2005): Climatology and Trends for Snow Cover, Journal of applied meteorology and climatology, https://doi.org/10.1175/2009JAMC1810.1, 2009b.

Faranda, D.: An attempt to explain recent trends in European snowfall extremes, Weather Clim. Dynam. Discuss, https://wcd.copernicus.org/preprints/wcd-2019-15/, 2019.

Fisher, R. A. and Tippett, L. H. C.: Limiting forms of the frequency distribution of the largest or smallest member of a sample, Mathematical Proceedings of the Cambridge Philosophical Society, 24, 180, https://doi.org/10.1017/S0305004100015681, 1928.

Fowler, H. J., Cooley, D., Sain, S. R., and Thurston, M.: Detecting change in UK extreme precipitation using results from the climateprediction.net BBC climate change experiment, Extremes, 13, 241–267, https://doi.org/10.1007/s10687-010-0101-y, 2010.

Garavaglia, F., Gailhard, J., Paquet, E., Lang, M., Garaçon, R., and Bernardara, P.: Introducing a rainfall compound distribution model based on weather patterns sub-sampling, Hydrology and Earth System Sciences, 14, 951–964, https://doi.org/10.5194/hess-14-951-2010, 2010.

Gaume, J., Chambon, G., Eckert, N., and Naaim, M.: Relative influence of mechanical and meteorological factors on avalanche release depth distributions: An application to French Alps, Geophysical Research Letters, 39, 1–5, https://doi.org/10.1029/2012GL051917, 2012.

Gaume, J., Eckert, N., Chambon, G., Naaim, M., and Bel, L.: Mapping extreme snowfalls in the French Alps using max-stable processes, Water Resources Research, 49, 1079–1098, http://doi.wiley.com/10.1002/wrcr.20083, 2013.

Gilleland, E. and Katz, R. W.: extRemes 2.0: An Extreme Value Analysis Package in R, Journal of Statistical Software, 72, 1–39, https://doi.org/10.18637/jss.v072.i08, 2016.

Gnedenko, B.: Sur la distribution limite du terme maximum d'une série aléatoire, The Annals of Mathematics, 44, 423, https://doi.org/10.2307/1968974, 1943.

Gottardi, F., Obled, C., Gailhard, J., and Paquet, E.: Statistical reanalysis of precipitation fields based on ground network data and weather patterns: Application over French mountains, https://doi.org/10.1016/j.jhydrol.2012.02.014, 2012.

Haberkorn, A., Helmert, J., Leppänen, L., López-Moreno, J. I., and Pirazzini, R.: European Snow Booklet, https://www.dora.lib4ri.ch/wsl/islandora/object/wsl:20198, 2019.

Il Jeong, D. and Sushama, L.: Projected changes to extreme wind and snow environmental loads for buildings and infrastructure across Canada, Sustainable Cities and Society, 36, 225–236, https://doi.org/10.1016/J.SCS.2017.10.004, 2018.

IPCC: Managing the risks of extreme events and disasters to advance climate change adaptation, Cambridge University Press, https://doi.org/10.1596/978-0-8213-8845-7, 2012.

IPCC: Special Report: The Ocean and Cryosphere in a Changing Climate, In press, https://doi.org/https://www.ipcc.ch/report/srocc/, 2019.

Katz, R. W., Parlange, M. B., and Naveau, P.: Statistics of extremes in hydrology, Advances in Water Resources, 25, 1287–1304, https://doi.org/10.1016/S0309-1708(02)00056-8, 2002.

Kharin, V. V. and Zwiers, F. W.: Estimating extremes in transient climate change simulations, Journal of Climate, 18, 1156–1173, https://doi.org/10.1175/JCLI3320.1, 2004.

Kim, H., Kim, S., Shin, H., and Heo, J. H.: Appropriate model selection methods for nonstationary generalized extreme value models, Journal of Hydrology, 547, 557–574, http://dx.doi.org/10.1016/j.jhydrol.2017.02.005, 2017.

Klein Tank, A. M. G. and Können, G. P.: Trends in Indices of Daily Temperature and Precipitation Extremes in Europe, 1946–99, Journal of Climate, https://doi.org/10.1175/1520-0442(2003)016%3C3665:TIIODT%3E2.0.CO;2, 2003.

Krinner, G., Derksen, C., Essery, R., Flanner, M., Hagemann, S., Clark, M., Hall, A., Rott, H., Brutel-Vuilmet, C., Kim, H., Ménard, C. B., Mudryk, L., Thackeray, C., Wang, L., Arduini, G., Balsamo, G., Bartlett, P., Boike, J., Boone, A., Chéruy, F., Colin, J., Cuntz, M., Dai, Y., Decharme, B., Derry, J., Ducharne, A., Dutra, E., Fang, X., Fierz, C., Ghattas, J., Gusev, Y., Haverd, V., Kontu, A., Lafaysse, M., Law, R., Lawrence, D., Li, W., Marke, T., Marks, D., Ménégoz, M., Nasonova, O., Nitta, T., Niwano, M., Pomeroy, J., Raleigh, M. S., Schaedler, G., Semenov, V., Smirnova, T. G., Stacke, T., Strasser, U., Svenson, S., Turkov, D., Wang, T., Wever, N., Yuan, H., Zhou, W., and Zhu, D.: ESM-SnowMIP: Assessing snow models and quantifying snow-related climate feedbacks, Geoscientific Model Development, 11, 5027–5049, https://doi.org/10.5194/gmd-11-5027-2018, 2018a.

Krinner, G., Derksen, C., Essery, R., Flanner, M., Hagemann, S., Clark, M., Hall, A., Rott, H., Brutel-Vuilmet, C., Kim, H., Ménard, C. B., Mudryk, L., Thackeray, C., Wang, L., Arduini, G., Balsamo, G., Bartlett, P., Boike, J., Boone, A., Chéruy, F., Colin, J., Cuntz, M., Dai, Y., Decharme, B., Derry, J., Ducharne, A., Dutra, E., Fang, X., Fierz, C., Ghattas, J., Gusev, Y., Haverd, V., Kontu, A., Lafaysse, M., Law, R., Lawrence, D., Li, W., Marke, T., Marks, D., Ménégoz, M., Nasonova, O., Nitta, T., Niwano, M., Pomeroy, J., Raleigh, M. S., Schaedler, G., Semenov, V., Smirnova, T. G., Stacke, T., Strasser, U., Svenson, S., Turkov, D., Wang, T., Wever, N., Yuan, H., Zhou, W., and Zhu, D.: ESM-SnowMIP: Assessing snow models and quantifying snow-related climate feedbacks, Geoscientific Model Development, 11, 5027–5049, https://doi.org/10.5194/gmd-11-5027-2018, 2018b.

Lafaysse, M., Morin, S., Coléou, C., Vernay, M., Serça, D., Besson, F., Willemet, J.-M., Giraud, G., and Durand, Y.: Toward a new chain of models for avalanche hazard forecasting in French mountain ranges, including low altitude mountains, International Snow Science Workshop, pp. 162–166, 2013.

Laternser, M. and Schneebeli, M.: Long-term snow climate trends of the Swiss Alps (1931-99), International Journal of Climatology, 23, 733–750, https://doi.org/10.1002/joc.912, 2003.

Lüthi, S., Ban, N., Kotlarski, S., Steger, C. R., Jonas, T., and Schär, C.: Projections of Alpine snow-cover in a high-resolution climate simulation, Atmosphere, 10, 1–18, https://doi.org/10.3390/atmos10080463, 2019.

Martins, E. S. and Stedinger, J. R.: Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data, Water Resources Research, 36, 737–744, https://doi.org/10.1029/1999WR900330, 2000.

Marty, C. and Blanchet, J.: Long-term changes in annual maximum snow depth and snowfall in Switzerland based on extreme value statistics, Climatic Change, 111, 705–721, https://doi.org/10.1007/s10584-011-0159-9, 2012.

Marty, C., Tilg, A.-M., Jonas, T., Marty, C., Tilg, A.-M., and Jonas, T.: Recent Evidence of Large-Scale Receding Snow Water Equivalents in the European Alps, Journal of Hydrometeorology, 18, 1021–1031, https://doi.org/10.1175/JHM-D-16-0188.1, 2017.

Milly, P. C. D., Bentacourt, J., Falkenmark, M., Robert, M., Hirsch, R. M., Kundzewicz, Z. W., Letternmaier, D. P., and Stouffer, R. J.: Stationarity is dead: Whither water management?, Science, 319, 573–574, https://doi.org/10.1126/science.1151915, 2008.

Montanari, A. and Koutsoyiannis, D.: Modeling and mitigating natural hazards: Stationarity is immortal!, Water Resources Research, 50, 9748–9756, https://doi.org/10.1002/2014WR016092, 2014.

Naaim-Bouvet, F., Prat, M., Jacob, J., Calgaro, J. A., and & Raoul, J.: La neige : recherche et réglementation, Presses de l'École nationale des ponts et chaussées, 2000.

Nicolet, G., Eckert, N., Morin, S., and Blanchet, J.: Inferring Spatio-temporal Patterns in Extreme Snowfall in the French Alps Using Max-stable Processes, Procedia Environmental Sciences, 27, 75–82, https://doi.org/10.1016/j.proenv.2015.07.102, 2015.

Nicolet, G., Eckert, N., Morin, S., and Blanchet, J.: Decreasing spatial dependence in extreme snowfall in the French Alps since 1958 under climate change, Journal of Geophysical Research, 121, 8297–8310, https://doi.org/10.1002/2016JD025427, 2016.

Nicolet, G., Eckert, N., Morin, S., and Blanchet, J.: A multi-criteria leave-two-out cross-validation procedure for max-stable process selection, Spatial Statistics, 22, 107–128, https://doi.org/10.1016/j.spasta.2017.09.004, https://doi.org/10.1016/j.spasta.2017.09.004, 2017.

Nicolet, G., Eckert, N., Morin, S., and Blanchet, J.: Assessing Climate Change Impact on the Spatial Dependence of Extreme Snow Depth Maxima in the French Alps, Water Resources Research, 54, 7820–7840, https://doi.org/10.1029/2018WR022763, https://onlinelibrary.wiley.com/doi/abs/10.1029/2018WR022763, 2018.

O'Rourke, M. and Auren, M.: Snow loads on gable roofs, Journal of Structural Engineering, 123, 1645–1651, https://doi.org/10.1061/(ASCE)0733-9445(1997)123:12(1645), 1997.

Quéno, L., Vionnet, V., Dombrowski-Etchevers, I., Lafaysse, M., Dumont, M., and Karbou, F.: Snowpack modelling in the Pyrenees driven by kilometric-resolution meteorological forecasts, Cryosphere, 10, 1571–1589, https://doi.org/10.5194/tc-10-1571-2016, 2016.

Rajczak, J. and Schär, C.: Projections of Future Precipitation Extremes Over Europe: A Multimodel Assessment of Climate Simulations, Journal of Geophysical Research: Atmospheres, 122, 773–10, https://doi.org/10.1002/2017JD027176, 2017.

Revuelto, J., Lecourt, G., Lafaysse, M., Zin, I., Charrois, L., Vionnet, V., Dumont, M., Rabatel, A., Six, D., Condom, T., Morin, S., Viani, A., and Sirguey, P.: Multi-criteria evaluation of snowpack simulations in complex alpine terrain using satellite and in situ observations, Remote Sensing, 10, 1–32, https://doi.org/10.3390/rs10081171, 2018.

Richard W. Katz: Statistical methods for nonstationary extremes, in: Extremes in a Changing Climate - Detection, Analysis & Uncertainty, pp. 15–38, Springer Science & Business Media, https://doi.org/10.1007/978-94-007-4479-0, 2012.

Rootzén, H. and Katz, R. W.: Design Life Level: Quantifying risk in a changing climate, Water Resources Research, 49, 5964–5972, https://doi.org/10.1002/wrcr.20425, 2013.

Rózsás, A., Sýkora, M., and Vigh, L. G.: Long-Term Trends in Annual Ground Snow Maxima for the Carpathian Region, Applied Mechanics and Materials, 821, 753–760, https://doi.org/10.4028/www.scientific.net/amm.821.753, 2016.

Saeb, A.: gnFit R package, https://www.rdocumentation.org/packages/gnFit, 2018.

Sanpaolesi, L., Currie, D., Sims, P., Sacre, C., Stiefel, U., Lozza, S., Eiselt, B., Peckham, R., Solomos, G., Holand, I., et al.: Scientific support activity in the field of structural stability of civil engineering works: snow loads, Final Report Phase I. Brussels: Commission of the European Communities. DGIII-D3, 1998.

Schellander, H. and Hell, T.: Modeling snow depth extremes in Austria, Natural Hazards, 20, https://doi.org/10.1007/s11069-018-3481-y, https://doi.org/10.1007/s11069-018-3481-y, 2018.

Schöner, W., Koch, R., Matulla, C., Marty, C., and Tilg, A. M.: Spatiotemporal patterns of snow depth within the Swiss-Austrian Alps for the past half century (1961 to 2012) and linkages to climate change, International Journal of Climatology, 39, 1589–1603, https://doi.org/10.1002/joc.5902, 2019.

Serinaldi, F.: Dismissing return periods!, Stochastic Environmental Research and Risk Assessment, 29, 1179–1189, https://doi.org/10.1007/s00477-014-0916-1, 2015.

Serinaldi, F. and Kilsby, C. G.: Stationarity is undead: Uncertainty dominates the distribution of extremes, Advances in Water Resources, 77, 17–36, https://doi.org/10.1016/J.ADVWATRES.2014.12.013, 2015.

Soci, C., Bazile, E., Besson, F. O., and Landelius, T.: High-resolution precipitation re-analysis system for climatological purposes, Tellus, Series A: Dynamic Meteorology and Oceanography, 68, https://doi.org/10.3402/tellusa.v68.29879, 2016.

Strasser, U.: Snow loads in a changing climate: New risks?, Natural Hazards and Earth System Science, 8, 1–8, https://doi.org/10.5194/nhess-8-1-2008, 2008.

Terzago, S., Fratianni, S., and Cremonini, R.: Winter precipitation in Western Italian Alps (1926–2010), Meteorology and Atmospheric Physics, 119, 125–136, https://doi.org/10.1007/s00703-012-0231-7, 2013.

Tramblay, Y. and Somot, S.: Future evolution of extreme precipitation in the Mediterranean, Climatic Change, https://doi.org/10.1007/s10584-018-2300-5, 2018.

Vernay, M., Lafaysse, M., Hagenmuller, P., Nheili, R., Verfaillie, D., and Morin, S.: The S2M meteorological and snow cover reanalysis in the French mountainous areas (1958 - present), [Data set]. AERIS, https://doi.org/https://doi.org/10.25326/37, 2019.

Vidal, J. P., Martin, E., Franchistéguy, L., Baillon, M., and Soubeyroux, J. M.: A 50-year high-resolution atmospheric reanalysis over France with the Safran system, International Journal of Climatology, 30, 1627–1644, https://doi.org/10.1002/joc.2003, 2010.

Vigneau, J.-P.: 1986 dans les Pyrénées orientales : deux perturbations méditerranéennes aux effets remarquables, Revue géographique des Pyrénées et du Sud-Ouest, 58, 23–54, https://doi.org/10.3406/rgpso.1987.4969, 1987.

Vionnet, V., Brun, E., Morin, S., Boone, A., Faroux, S., Le Moigne, P., Martin, E., and Willemet, J.-M.: Geoscientific Model Development The detailed snowpack scheme Crocus and its implementation in SURFEX v7.2, Geosci. Model Dev, 5, 773–791, https://doi.org/10.5194/gmd-5-773-2012, 2012.

Vionnet, V., Dombrowski-Etchevers, I., Lafaysse, M., Quéno, L., Seity, Y., and Bazile, E.: Numerical weather forecasts at kilometer scale in the French Alps: Evaluation and application for snowpack modeling, Journal of Hydrometeorology, 17, 2591–2614, https://doi.org/10.1175/JHM-D-15-0241.1, 2016.

Vionnet, V., Six, D., Auger, L., Dumont, M., Lafaysse, M., Quéno, L., Réveillet, M., Dombrowski-Etchevers, I., Thibert, E., and Vincent, C.: Sub-kilometer Precipitation Datasets for Snowpack and Glacier Modeling in Alpine Terrain, Frontiers in Earth Science, 7, 1–21, https://doi.org/10.3389/feart.2019.00182, 2019.

Wilcox, C., Vischel, T., Panthou, G., Bodian, A., Blanchet, J., Descroix, L., Quantin, G., Cassé, C., Tanimoun, B., and Kone, S.: Trends in hydrological extremes in the Senegal and Niger Rivers, Journal of Hydrology, 566, 531–545, https://doi.org/10.1016/J.JHYDROL.2018.07.063, 2018.

Zhang, X., Zwiers, F. W., Li, G., Zhang, X., Zwiers, F. W., and Li, G.: Monte Carlo Experiments on the Detection of Trends in Extreme Values, Journal of Climate, 17, 1945–1952, https://doi.org/10.1175/1520-0442(2004)017<1945:MCEOTD>2.0.CO;2, 2004.