Comment on “Hydrometeorological Triggers of Periglacial Debris Flows in the Zermatt Valley (Switzerland) Since 1864” by Michelle Schneuwly-Bollschweiler and Markus Stoffel

M. Heiser1,2, M. Schlögl1,2, C. Scheidl1, and S. Fuchs1

1Institute of Mountain Risk Engineering, University of Natural Resources and Life Sciences, Vienna, Austria, 2Department for Earth Observation and Geoinformation, Zentralanstalt für Meteorologie und Geodynamik, Vienna, Austria

Abstract Generalizing properties of samples derived via data analysis to advance understanding of underlying processes constitutes the essence of statistical inference. This is especially true in geophysical research, where general findings often emerge from case studies. Consequently, statements about trends in case studies should naturally be grounded on a comprehensive and robust analysis of available data. Here, we focus on trends reported for a debris flow time series of a well-known case study in the Zermatt Valley, Switzerland (Schneuwly-Bollschweiler & Stoffel, 2012: https://doi.org/10.1029/2011JF002262). We elaborate that we were not able to confirm the findings of the original study with respect to changes in debris flow seasonality. On the contrary, we show that confusion between data analytic question types (exploratory vs. inferential) can lead to the confirmation of trends that do not necessarily hold beyond the data set, and we demonstrate how the reported effect could be an artifact of incomplete data. In doing so we employ a range of methods that can be used for adequately addressing the original question.

Plain Language Summary Mountain hazards such as debris flows repeatedly cause high loss in the European Alps. Time series of past (historical) events are often the most reliable information used for hazard mitigation. These time series, however, are subject to uncertainties in data collection and are likely to be incomplete. Taking a 150-year data set derived in the Zermatt region, Switzerland, we show the pitfalls of confusing data analytic question types (exploratory vs. inferential) and of analyzing trends in time series without taking the completeness of the time series into account. Failure to avoid these pitfalls can lead to potentially erroneous conclusions about the drivers of changes in debris flow occurrence, such as climate change, land-use change, or mitigation measures.

1. Comment: Fallacies in Debris Flow Trend Analysis

One of the most important and pervasive challenges in natural hazard research concerns the analysis of event time series (Fuchs et al., 2015; Pitt et al., 2020; Schlögl et al., 2021; Telesca, 2007). Particularly in mountain regions such as the European Alps, knowledge on the occurrence and thus on the frequency of rare but often high-magnitude events such as debris flows is crucial for tailored risk management (Jakob et al., 2020; Perov et al., 2017; Turkington et al., 2016). Information on past (historical) events is often the most reliable information to characterize frequency, magnitude, and seasonality of natural hazard processes, and it also provides indication on the performance of hazard mitigation in managed catchments. This is illustrated by Schlögl et al. (2021), who conducted trend analysis of torrential events from 1962 to 2017 in Austria and elucidated the effectiveness of existing protection measures but also a need for their maintenance. Similarly, Sebald et al. (2019) analyzed time series of both annual forest cover and observed torrential events for the last 30 years in Austria and showed that forest cover and patch density have a strong negative effect on the occurrence and frequency of torrential hazards. Moreover, analyzing time series of torrential flow events, compiled for instance from documentations, stratigraphic analysis, or analysis of geomorphological field evidence, is necessary to better understand characteristics of and changes in alpine headwater regimes. Known historical findings—which are mainly available in terms of case studies on a local (catchment) or regional (valley) scale—provide a major indication for drawing general conclusions about changes in torrential event frequencies, magnitude, and seasonality in alpine regions. Given the overall limited amount of studies, however, statements about trends in case studies should naturally be grounded on a comprehensive and robust analysis of available data.
Our comment refers, by way of example, to an analysis of debris flow activity published by Schneuwly-Bollschweiler and Stoffel (2012), which is based on a localized but long and continuous record. In their study, the authors presented an assessment using data from eight torrents in the Zermatt Valley (Valais, Switzerland) at elevations between 1900 m a.s.l and 4600 m a.s.l. The data set was based on dendrochronological reconstructions and daily records of meteorological and gauge measurements over a period of almost 150 years. With reference to the collected and reconstructed event data, the authors concluded that “the debris flow season at these high-altitude sites now [2007] is much longer (May to October) than it used to be in the late nineteenth century when activity was limited to June–September” (Schneuwly-Bollschweiler & Stoffel, 2012).

Based on extensive inferential data analysis, we were not able to reproduce the findings of Schneuwly-Bollschweiler and Stoffel (2012) with respect to changes in debris flow seasonality in the Zermatt Valley. By contrast, we show in accordance with Leek and Peng (2015) that the study confuses data analytic question types by mistaking exploratory data analysis for an inferential analysis. In addition, we demonstrate how the reported effect could be an artifact of incomplete data.

Conducting a completeness analysis in terms of reporting rate of the “Zermatt Valley” event data set therefore represented a first and essential step (Heiser et al., 2019). The completeness analysis comprised a series of statistical tests for homogeneity in the reporting rate, employing methods based on the cumulative number of events (using different loss functions) as well as change-point analysis (using parametric, nonparametric, and Bayesian variants). These methods differ in how they account for different data-generating processes.

While the change-point methods indicated that the time series could be regarded homogeneous for the entire period (beginning in the year 1864), the methods based on the cumulative number of events suggested that the time series was complete only from 1916 onwards (Figure 1). Both the entire (1864–2007) and the shortened time series (1916–2007) were examined for trends in the number of events per year and for trends in seasonality.

In a second step, we assessed the presence of trends in event frequency (Figure 1) following Schlögl et al. (2021), again based on a plurality of methods (Appendix A). While no trends in frequency could be detected in the full data set, the null hypothesis of no trend in the time series could only be rejected in the case of one test on the reduced time series (Appendix B).

In a final step we analyzed the existence of trends in seasonality, using quantile regression and the Theil-Sen slope, again following Schlögl et al. (2021). Additionally we investigated the stability of seasonality between 1864 and 2007 by analyzing the temporal concentration of debris flow events in the course of a year (cf. Appendix C). The temporal concentration, that is, the magnitude of seasonality, was expressed as the mean resultant length, which takes values between zero and one (Jammalamadaka & Sengupta, 2001). A value close to one indicates low spread (i.e., high seasonality), while a value close to zero indicates high spread (i.e., uniformity in time).

Since the event dates were partially uncertain, a Monte Carlo simulation was performed to account for event date uncertainty. For all uncertain time points, the day was perturbed randomly (uniformly distributed) in the range of

![Figure 1](image_url)

**Figure 1.** Panel A shows the number of debris flows per year for the whole event series. Panel B shows the number of events for the reduced event series, starting at $t_d = 1916$. The red dashed line is the local polynomial regression with 95% confidence interval (red shaded area) based on 1,000 bootstrap samples. The span was chosen for each series individually by means of generalized cross-validation. The gray shaded area corresponds to 95% of all smoothers estimated on 10,000 permuted samples. The dashed black lines are the 95% confidence interval of the mean number of events per year. Panel C shows the cumulative number of events per year. Also shown is the average yearly increase of $\beta = 2$ events per year for the reduced event series.
±7 and 14 days, respectively. The width of the interval depended on the quality of the observation as stated by the authors in the supplementary material. The regression models were estimated for the data generated in this way. The entire computation procedure was repeated 1,000 times.

For the entire event series, no trend was observed in the day of event occurrence (upper panel Figure 2). Both the quantile regression, yielding a slope of 0.05 (−0.19, 0.20), and the Theil-Sen slope with a value of 0.01 (−0.12, 0.04) included 0 as a possible slope value within the confidence interval. Furthermore, the null hypothesis of no trend in seasonality could not be rejected by any of the trend tests (Table B2).

Similarly, the reduced event series also did not provide indications of significant trends in seasonality. Both the quantile regression, yielding a slope of −2 (−7, 1), and the Theil-Sen slope with a value of −2 (−4, 0.8) included 0 as a possible slope value within the confidence interval. Given the fact that 0 just lies within the upper bound of the confidence interval, a possible shift of 1–2 days per 10 years cannot be ruled out, since such slight trends simply might not yet be detectable due to the lack of statistical power. However, such a slight negative trend would be low in magnitude in any case, compared to the apparent variability in the event series, which exhibits a standard deviation of 38 (33, 42) days (lower panel Figure 2). Furthermore, the null hypothesis of no trend in seasonality could not be rejected by any of the trend tests (Table B2).

The temporal concentration, which was high across the whole event series with values consistently above 0.75, showed a periodical behavior (see Figure 3). Compared to the concentration estimated from the base period suggested by Schneuwly-Bollschweiler and Stoffel (2012) (1864–1900), which exhibits a value of 0.92 (0.87, 0.97), the event series showed a decreasing concentration. However, if the base period was based on the first 30 years after completeness (1916–1945), no departure from the expected concentration of 0.85 (0.77, 0.92) was visible. Hence, if stricter completeness criteria were applied, the change in concentration, which would imply an extension of the debris flow season in summer, disappeared.

The analysis suggests that the debris-flow event time series, compiled by Schneuwly-Bollschweiler and Stoffel (2012), does not in fact constitute compelling evidence that would refute the hypotheses of (a) no change in the number of events per year, as well as (b) no change in seasonality in the period between 1864 and 2007. Solely
based on the available data set, it is reasonable to conclude that the number of events and seasonality of debris flow occurrence in small, steep, permafrost-influenced catchments at high-altitude sites in the Zermatt Valley did not change between 1864 and 2007.

Consequently, we argue that the data set does not hold enough empiric evidence to warrant the conclusion that “the debris flow season at these high-altitude sites now [2007] is much longer (May to October) than it used to be in the late nineteenth century when activity was limited to June–September” (Schneuwly-Bollschweiler & Stoffel, 2012). In the original study, this conclusion on seasonal shifts was based on an exploratory analysis (in essence a percent stacked barplot) across four equally long time segments. However, a formal statistical analysis that could have quantified whether the observed trend pattern is significantly different from patterns that are to be expected under the assumption of no trend was lacking. Therefore, the study confuses data analytic question types by mistaking exploratory data analysis (real question type) for an inferential analysis (perceived question type; Leek & Peng, 2015). While exploratory data analysis “seeks to make discoveries, but can rarely confirm those discoveries,” inferential data analysis “quantifies whether an observed pattern will likely hold beyond the data set in hand” (Leek & Peng, 2015).

From a classical inferential point of view, the lack of significant results in the trend tests on seasonality does challenge this conclusion of Schneuwly-Bollschweiler and Stoffel (2012), as does the variability in the concentration (cf. Figure 3) from a purely exploratory perspective.

Our comment illustrates the caveat of selection biases as well as the intricacies of statistical inference. Properties of the underlying data need to be thoroughly reviewed and analyses need to be conducted with care in order to avoid faulty conclusions. Consequently, careful consideration of the robustness of results and conclusions is of prime importance, especially in light of possible confirmation bias in the climate change discussion. We are convinced that local studies do provide important insights into the dynamics of debris flow frequency and seasonality due to climate change, yet we advocate that data analysis has to be performed with caution.

Appendix A: Trend Tests

A number of (nonparametric) trend tests were conducted for assessing the presence of trends in the data set.

A1. Mann-Kendall Trend Test

The Mann-Kendall trend test can be used to test for a monotonic trend in a time series \( z(t) \) based on the Kendall rank correlation \( r \) of \( z(t) \) and \( t \) (Kendall, 1975; Mann, 1945). Specifically, it can be used to test the null hypothesis that the data are i.i.d. (independent and identically distributed), that is, do not follow a monotonic trend.

The Mann-Kendall test statistic \( S \) is calculated according to

\[
S = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} \text{sgn} (x_j - x_k)
\]

(A1)
with sgn denoting the signum function

\[ \text{sgn}(x) = \begin{cases} 
-1 & \text{if } x < 0, \\
0 & \text{if } x = 0, \\
1 & \text{if } x > 0.
\end{cases} \quad (A2)\]

in case of positive S it can be assumed that observations obtained later in time tend to be larger than observations made earlier, and vice versa.

The statistic S is approximately normally distributed. Using continuity correction, the z score can be computed as:

\[ z = \frac{\text{sgn}(S) (|S| - 1)}{\sigma} \quad (A3) \]

the mean of S is \( \mu = 0 \), and the variance (including the correction term for ties) is

\[ \sigma^2 = \frac{1}{18} \left\{ n(n - 1)(2n + 5) - \sum_{j=1}^p t_j (t_j - 1)(2t_j + 5) \right\} \quad (A4) \]

where \( p \) denotes the number of the tied groups in the data set and \( t_j \) is the number of data points in the \( j \)th tied group.

The critical region can be obtained from the quantiles of the standard normal distribution. In the two-sided case, \( \pm z_{1-\alpha/2} \) is the critical region. \( \alpha \) denotes the type I error rate, that is, probability of falsely rejecting the null hypothesis.

### A2. Bartels Test

The Bartels test (Bartels, 1982) is a rank version of von Neumann’s ratio test for randomness. It tests the null hypothesis of randomness in a sequence of observations. The Bartels test uses the ranks \( r_1, \ldots, r_n \) of the observations \( X_1, \ldots, X_n \) and employs the rank version of von Neumann’s ratio (RVN) as a test statistic:

\[ \text{RVN} = \frac{\sum_{i=1}^n (r_i - r_{i+1})^2}{\sum_{i=1}^n (r_i - \bar{r})^2} \quad (A5) \]

the computation of critical values depends on the size of the observation series. For sample sizes in the range of \( 10 \leq n < 100 \) a beta approximation over the range \( 0 \leq x \leq 4 \) is used. The density function of the beta distribution is

\[ f(x; p, q) = \frac{(B(p, q))^{-1} x^{p-1}(4 - x)^{q-1}}{4^{p+q} p(q)} \quad (A6) \]

with parameters

\[ p = q = \frac{5n(n + 1)(n - 1)^2}{2(n - 2)(5n^2 - 2n - 9)} - \frac{1}{2} \quad (A7) \]

for sample sizes \( n \geq 100 \), a normal approximation with \( \mathcal{N}(2, 20/(5n + 7)) \) is used.

### A3. Wald-Wolfowitz Test

The Wald-Wolfowitz runs test (Wald & Wolfowitz, 1940) can be used to test the null hypothesis of randomness in a time series. It is based on transforming the series into a dichotomous vector, depending on whether each value is above or below a given threshold. Usually, the sample median is used as threshold, and values above it are coded as positive, while values below it are coded as negative. Observations equal to the median are dropped. A run is a subset of the transformed sequence, which is characterized by adjacent equal (i.e., positive or negative) elements.

The number of runs \( r \) in a series of \( n = n_+ + n_- \) elements has a conditional distribution given the observation of
n_+ positive values and n_- negative values, which is asymptotically normally distributed under the null hypothesis of randomness (cf. Central limit theorem).

The mean is estimated using the expected number of runs \( \tilde{r} \)

\[
E(r) = \tilde{r} = \frac{2n_+ n_-}{n_+ + n_-} + 1 \tag{A8}
\]

and the variance is estimated using the variance of the number of runs \( s_r^2 \)

\[
s_r^2 = \frac{2n_+ n_- (2n_+ n_- - n_+ - n_-)}{(n_+ + n_-)^2 (n_+ + n_- - 1)} = \frac{(\tilde{r} - 1)(\tilde{r} - 2)}{n - 1} \tag{A9}
\]

the test statistic is the standard score

\[
Z = \frac{r - \tilde{r}}{s_r} \tag{A10}
\]

with \( \pm z_{1-\alpha/2} \) as critical region in the two-sided case.

A4. Theil-Sen Slope

The Theil-Sen estimator is a method for robustly estimating the slope in simple linear regression. The original method uses the median of a set of slopes \( (y_j - y_i)/(x_j - x_i) \) for all pairs of \( n \) two-dimensional points \((x_i, y_i)\) to estimate the slope coefficient (Sen, 1968; Theil, 1950). In this comment we employ the repeated medians method of Siegel (1982), which is more robust and maintains a 50% breakdown point. This method is based on first determining the slopes of all lines between each observation \((x_i, y_i)\) and all other observations, and then computing the median slope \( m_i \) related to each of the \( n \) observations. The final slope estimator is eventually determined as the median of all these \( n \) medians \( m_i \).

Appendix B: Trend Test Results

Quantile regression and the Theil-Sen slope on both the entire (1864–2007) and the shortened time series (1916–2007) data set resulted in a slope of 0. Furthermore, the mean trend of the local polynomial regression is included in the range of an ensemble of 10,000 local regression smoothers, which were calculated on the permuted data set (Figure 1). According to the Bartels test, the Mann-Kendall trend test and the Wald-Wolfowitz test, the null hypothesis of no trend in the time series could only be rejected in the case of the Wald-Wolfowitz test on the reduced time series (Table B1).

| Period       | Observations | Bartels Statistic | Bartels p-value | Mann-Kendall Statistic | Mann-Kendall p-value | Wald-Wolfowitz Statistic | Wald-Wolfowitz p-value | Theil-Sen Slope (Percent per Year) |
|--------------|--------------|-------------------|-----------------|------------------------|----------------------|--------------------------|------------------------|----------------------------------|
| 1864–2007    | 118          | -1.65             | 0.10            | 0.09                   | 0.19                 | -1.85                    | 0.07                   | 0 (0.00, 0.01)                    |
| 1916–2007    | 86           | -1.77             | 0.08            | -0.05                  | 0.51                 | -4.34                    | 0.00                   | 0 (−0.02, 0.01)                   |
Table B2

| Period   | Observations | Bartels Statistic | Bartels p-value | Mann-Kendall Statistic | Mann-Kendall p-value | Wald-Wolfowitz Statistic | Wald-Wolfowitz p-value | Theil-Sen Slope (Percent per Year) |
|----------|--------------|------------------|-----------------|------------------------|----------------------|--------------------------|------------------------|-----------------------------------|
| 1864–2007 | 112          | −0.55            | 0.58            | −0.03                  | 0.60                 | −0.38                    | 0.70                   | 0.01 (−0.12, 0.04)                |
| 1916–2007 | 80           | −0.32            | 0.75            | −0.10                  | 0.18                 | −0.23                    | 0.82                   | −0.2 (−0.4, 0.08)                 |

Appendix C: Temporal Concentration

The temporal concentration in the course of the year, that is, the magnitude of seasonality, is given by the length of the vector \( \langle r \rangle \) (Equation C1), with \( \langle r \rangle \in [0, 1] \).

\[
\langle r \rangle = \sqrt{\langle x \rangle + \langle y \rangle}
\]  

(C1)

A value of \( \langle r \rangle \) close to unity indicates low spread that is, high seasonality, while a value close to zero indicates high spread, that is, uniformity in time. Note that a cyclic pattern also results in \( \langle r \rangle \) being near 0. \( \langle x \rangle \) and \( \langle y \rangle \) are defined as

\[
\langle x \rangle = n^{-1} \sum_{i=1}^{n} \cos(\theta_i)
\]  

(C2)

\[
\langle y \rangle = n^{-1} \sum_{i=1}^{n} \sin(\theta_i)
\]  

(C3)

where \( \langle x \rangle \) and \( \langle y \rangle \) are the coordinates of the mean occurrence date and \( \theta_i \) is the angle of the \( i \)th event according to Equation C4. The transformation, projecting the days of the years on the unit circle, is given by Equation C4.

\[
\theta_i = t_i^{(*)} \left( \frac{2\pi}{365} \right)
\]  

(C4)

where \( t_i^{(*)} \) is the Julian day at which event \( i \) happened.

Data Availability Statement

The data that support the findings of this study are openly available as supporting information (Table S1) in Schneuwly-Bollschweiler and Stoffel (2012). The code used for data preparation, processing, and statistical analysis is available on GitLab at https://gitlab.com/Rexthor/debris-flow-trends-zermatt.

References

Bartels, R. (1982). The rank version of von neumann’s ratio test for randomness. *Journal of the American Statistical Association*, 77(377), 40–46. https://doi.org/10.1080/01621459.1982.10477764

Fuchs, S., Keiler, M., & Zischg, A. (2015). A spatiotemporal multi-hazard exposure assessment based on property data. *Natural Hazards and Earth System Sciences*, 15(9), 2127–2142. https://doi.org/10.5194/nhess-15-2127-2015

Heiser, M., Hübl, J., & Scheidl, C. (2019). Completeness analyses of the Austrian torrential event catalog. *Landslides*, 16(11), 2115–2126. https://doi.org/10.1007/s10346-019-01218-3

Jakob, M., Mark, E., McDougall, S., Friele, P., Lau, C.-A., & Bale, S. (2020). Regional debris-flow and debris-flood frequency–magnitude relationships. *Earth Surface Processes and Landforms*, 45(12), 2954–2964. https://doi.org/10.1002/esp.4942

Jammalamadaka, S. R., & Sengupta, A. (2001). *Topics in circular statistics*. (No. v. 5). World Scientific.

Kendall, M. G. (1975). *Rank correlation methods*. (4th ed.). Charles Griffin.

Leek, J. T., & Peng, R. D. (2015). What is the question? *Science*, 347(6228), 1314–1315. https://doi.org/10.1126/science.aaa6146

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica*, 13(3), 245. https://doi.org/10.2307/1907187

Perov, V., Chernomoretov, S., Badarina, O., Savennyuk, E., & Leontyeva, T. (2017). Debris flow hazards for mountain regions of Russia: Regional features and key events. *Natural Hazards, 88*(S1), 199–235. https://doi.org/10.1007/s11069-017-2841-3

Pitt, D., Tröck, S., van den Honert, R., & Wong, W. W. (2020). Modeling risks from natural hazards with generalized additive models for location, scale and shape. *Journal of Environmental Management*, 275, 111075. https://doi.org/10.1016/j.jenvman.2020.111075

Acknowledgments

Open access funding provided by University of Natural Resources and Life Sciences Vienna (BOKU).
Schlögl, M., Fuchs, S., Scheidl, C., & Heiser, M. (2021). Trends in torrential flooding in the Austrian Alps: A combination of climate change, exposure dynamics, and mitigation measures. *Climate Risk Management*, 32, 100294. https://doi.org/10.1016/j.crm.2021.100294

Schneuwly-Bollschweiler, M., & Stoffel, M. (2012). Hydrometeorological triggers of periglacial debris flows in the Zermatt valley (Switzerland) since 1864. *Journal of Geophysical Research: Earth Surface*, 117(F2), 12. https://doi.org/10.1029/2011JF002262

Sebald, J., Senf, C., Heiser, M., Scheidl, C., Pflugmacher, D., & Seidl, R. (2019). The effects of forest cover and disturbance on torrential hazards: Large-scale evidence from the Eastern Alps. *Environmental Research Letters*, 14(11), 114032. https://doi.org/10.1088/1748-9326/ab4937

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall’s tau. *Journal of the American Statistical Association*, 63(324), 1379–1389. https://doi.org/10.1080/01621459.1968.10480934

Siegel, A. F. (1982). Robust regression using repeated medians. *Biometrika*, 69(1), 242–244. https://doi.org/10.1093/biomet/69.1.242

Telesca, L. (2007). Time-clustering of natural hazards. *Natural Hazards*, 40(3), 593–601. https://doi.org/10.1007/s11069-006-9023-z

Theil, H. (1950). A rank invariant method of linear and polynomial regression analysis, i, ii, iii. *Proceedings of the Koninklijke Nederlandse Akademie Wetenschappen, Series A Mathematical Sciences*, 53, 1397–1412.

Turkington, T., Remaire, A., Ettema, J., Hussin, H., & van Westen, C. (2016). Assessing debris flow activity in a changing climate. *Climatic Change*, 137(1–2), 293–305. https://doi.org/10.1007/s10584-016-1657-6

Wald, A., & Wolfowitz, J. (1940). On a test whether two samples are from the same population. *The Annals of Mathematical Statistics*, 11(2), 147–162. https://doi.org/10.1214/aoms/1177731909