Reply on RC1
Lena Katharina Schmidt et al.

Author comment on "Reconstructing five decades of sediment export from two glaciated high-alpine catchments in Tyrol, Austria, using nonparametric regression" by Lena Katharina Schmidt et al., EGUsphere, https://doi.org/10.5194/egusphere-2022-616-AC1, 2022

RC1: 'Comment on egusphere-2022-616', Anonymous Referee #1, 04 Oct 2022

Dear anonymous Referee #1,

We would like to thank you for the very thoughtful and detailed comments, questions and suggestions. Below, we provide our response as direct answers to each comment and hope that our suggestions will be to your satisfaction. We also provide a figure in the attached pdf for better understanding.

Best, Lena Katharina Schmidt on behalf of all authors

# General comments

In this manuscript, the authors applied machine learning to reconstruct sediment discharge records in two catchments in the Austrian Alps. After validating the reconstructed record, the authors identified trends and regime shifts with various change point detection methods. They identify the early 1980s as a turning point for the sediment dynamics and suggest links with temperature-driven glacier dynamics.

This is a valuable contribution showcasing the application of modern, data-driven methods to a field where they are yet to be routinely applied. However, beyond its technical value, the paper falls short from connecting its methods and results to the wider literature and addressing how such methods could be applied to other areas of study. For example, the discussion section would benefit from circling back to the larger scope and scientific questions mentioned in the introduction. Overall, the paper is well-structured easy to follow, but key information is missing from the Methods section for readers both familiar and unfamiliar with the techniques applied (see specific comments below).

# Specific comments
## Inconsistent verb tenses

In Methods and Results section, verb tenses switch between past and present. Some authors prefer to use present all along, while some prefer to use past to describe all past actions including methods and results. This is the authors’ choice, but it has to be consistent. For example, L152, the authors use “we train” to describe past training, then L157 the authors use “we applied” to describe past application. This is inconsistent and is found in a number of places.

**Answer:** Thank you. We will harmonize the use of tenses.

## Differences in precipitation gradients

The authors mentioned L126 that the precipitation gradient is 0.05 per 100. At L175, the correction factor between P(Vent) and P(VF) is P(Vent) = 1/1.3 * P(VF) = 0.769 * P(VF). Using the elevation from gauges at Vent (1891 m) and Vernagt (2635) leads to an elevation difference of 744 m. The correction factor calculated from the previously cited precipitation gradient is then 744 / 100 * 0.05 = 0.372 and equals roughly half of the reported value. I understand that the authors used the recorded data to derive their value, but I am curious for the large difference between the value reported and the one cited.

**Answer:** Thank you for this interesting question. Schöber et al. (2014) state 4 – 5 % per 100 m for the area, but that includes a neighbouring valley (around Obergurgl) as well. However, Vent receives considerably less precipitation than Obergurgl, due to its shielded location between the highest mountain in Tyrol (Wildspitze 3770m) and Ramolkogl (3550) and because it is located further away from the alpine ridge (luv/lee effects). This may be why the difference in measurement time series is larger than expected from the gradient.

## Any ensemble of models can assess model uncertainty

- L230-232: I disagree with that statement. The quantification of the uncertainties that the authors attribute to QRF is a result from ensembles of model with a random component. One could get a distribution of predicted values from an ensemble of neural networks with random initialization, or random partitions between training and testing. Ensemble of neural networks is not uncommon: in deep learning literature, results for new neural networks are often reported from a 10-fold cross-validation for which 10 models are trained, and, sometimes, the ensemble of these 10 models used for predictions. I would suggest the authors clarify the advantage of QRF if I misunderstood it, or be more nuanced in this statement and back it to QRF ensemble process rather than to QRF itself.

**Answer:** Thank you for this comment. It seems we have to be more clear about the QRF approach, which inherently includes ensemble processes (to produce a “forest” of regression trees). If we understand it correctly, this is not inherent to the other methods you mentioned. We suggest to improve the description in this segment and add “traditional” (i.e. “compared to traditional fuzzy logic or ANN”).

## Key information missing when describing QRF, too much information for change point detection

Key information is missing when describing QRF:

- L320: The authors mention here that the time series used as predictors show autocorrelation. Is there also some correlation between the time series? If so, this could be leveraged by methods like ARIMA or NARX to perform the predictions. In general, it is not best practice for machine learning approaches to only use one approach, and tree-
based approach are not often the go-to algorithm(s) to perform time series predictions. I recommend that the authors better justify their choice of using only one algorithm, and specifically QRF. This may be done summarizing the cited literature, but is at the moment insufficient by itself.

**Answer:** Thank you for these suggestions. It seems we have to express more clearly that the scope of the study was to test QRF specifically in the alpine catchments (as it had been applied to sediment dynamics successfully in the past) and interpret the results – rather than identifying the best possible method in a comparison. Although there might be other applicable methods, we find that QRF works sufficiently well with the presented data.

To our knowledge, there are no studies directly comparing QRF to other approaches for sediment concentration modelling – except the one we already mentioned: Compared to other methods, that are traditionally applied for suspended sediment concentration modelling, QRF performance was superior (Francke et al., 2008). As reviewer 2 suggested to compare QRF to sediment rating curves – a very simple and traditional approach for estimating sediment concentrations – we will add that to compare QRF with it. However, a study comparing random forest (which QRF is based on) to support-vector machines and artificial neural networks for suspended sediment concentration modelling (Al-Mukhtar, 2019) concluded that performance of random forest was superior. A study on the prediction of lake water levels (i.e. not with respect to sediment concentrations, but at least hydrological timeseries) came to the same conclusion (Li et al., 2016).

We suggest to improve the description of the aim of the study.

- L243: The authors mention here that they used a 5-fold cross validation. While cross-validation is often performed with 5 or 10 folds, it is also common practice to perform repeated cross-validation to have more robust statistics on model performance. It would be beneficial if the authors justified the number of folds (i.e. why 5 instead of 10), and the choice of not doing any repeats.

**Answer:** Thank you for this detailed question. We will point out more clearly that - unlike "usual" cross validations - we use temporally contiguous blocks of our data for the cross-validation, to avoid unrealistically good performance simply though autocorrelation. This would be an issue if we just allowed to pick individual days for the cross-validation. Thus, ours is a rather strict approach and repeats in the classical sense are not as easily possible.

Beyond that, the number of folds is indeed always arbitrary to some extent. We tried to find a compromise between too selective test data and too few training data. Choosing 5-fold cross validation as a compromise roughly corresponds to the number of complete seasons included in the shortest time series at VF.

- L325-339: The level of details provided here for change point detection departs from the level of details provided in the section detailing QRF. In particular, the QRF section does not mention any implementation details. I deem these details to be unessential. In particular, the names of the R packages are unnecessary.

**Answer:** We do not fully agree here, since the stating of the R packages, which in our view is common practice, promotes reproducibility and acknowledges the work of others. With the respect to the implementation details of QRF, we build upon other publications and
published the code alongside the manuscript, which we hope facilitates reproducibility.

Nonetheless, the term "mcp" is used throughout the paper but never defined; please provide a clear definition of it and use an uppercase acronym instead of the package name.

Answer: Thank you, we will do that.

Beyond the justification of using the Mann-Kendall tests, there is a lack of references justifying the use of these specific change point detection methods, and a reader with a different perspective may ask why the authors did not use another method (for example, the Fisher Information; https://doi.org/10.3390/w14162555 for a recent example in hydrologic sciences).

Answer: Thank you for this suggestion. Indeed, there are many available change point detection methods. We intended to apply an established, often-applied method (Pettitt, e.g. by Costa et al., 2018) and – in contrast to most studies, that only use one method - counter-balance its weaknesses (no uncertainty quantification, low detection probability if change point is located near the beginning or end of the time series) by using another approach with complementary advantages, i.e. mcp, which is being applied in an increasing number of studies and research fields (e.g. Veh et al., 2022; Yadav et al., 2021; Pilla and Williamson, 2022). We will improve the description to make this decision more easily understandable to the readers.

Furthermore, the choice of hyper-parameters for the QRF is crucially missing and should be reported. It seems that the authors have not performed any tuning of the hyper-parameters which should also be justified.

Answer: The two most important hyper-parameters are the number of trees in a “forest” and the number of selected predictors at each node (“mtry” parameter). The latter is optimized in the modelling process (and is hardly sensitive). A larger number of trees increases robustness (i.e. reduces the effect of the heuristic nature of QRF) – at the expense of computation time. We set the number of trees to 1000, which is twice the default value, to ensure robustness. We will add this to the description.

## Limits to applicability and links to introduction context and questions

L551-559: In this paragraph, the authors could start discussing implications of the applicability of their method. For example, how lucky were the authors in finding such limited out-of-domain observations during the period for which they wanted to apply their model? Was that expected? Is that expected in the future if extreme conditions are more likely (e.g. increased temperature, increased precipitation)? How does this impact the applicability of the same approach in other catchments, or over different timescales? In particular, could this be used at all for forecasting future evolution of sediment dynamics? All of these questions are interesting, and I suggest that the authors address at least a few of them to explain to the wider audience the limits of their approach. Specifically, this could be mentioned in the Outlook section 6.4 to circle back to the wider themes of the introduction.

Answer: Thank you for this interesting question. We do not think that the number of out-of-domain observations is a question of "luck". Naturally, for data-driven approaches, datasets must be “sufficiently large”- and the larger and more varied the training dataset,
the less likely occurrences of out-of-domain observations will be. Thus, this rather gives some indication on the representativity of the training data and therefore also the credibility and limits of the model results. However, we agree that we should emphasize the need to assess this for future studies on other catchments and/or future evolution.

## Minor specific comments

- L245: "250 Monte-Carlo realizations": at this point in the manuscript, it is unclear on which random variable the Monte-Carlo simulation is performed. It became clear to me at L340, but the authors should probably add some clarification before that point. The number of Monte-Carlo simulation should also be justified. Why 250 iterations were chosen? If the authors used a convergence criterion, it should be reported and justified.

Answer: We will improve the description in L245. Generally, a higher number of iterations will result in a more robust estimate of the mean annual suspended sediment yield. In practice however, this is one of the main points that will increase computation time. The chosen number of 250 iterations yields sufficiently good results. This can e.g. be seen in the confidence intervals of the mean estimates, that are ca ± 1.25 % of the mean.

- L280: Is there a reason for choosing the partition of the data between data from 2019-2020 for training and data from 2020-2021 for validation. Why not the other combination too (2020-2021 for training, 2019-2020 for validation)?

Answer: There seems to be a misunderstanding, it is not 2020/21 but 2000/01. Since we wanted to assess how well the model can reproduce past suspended sediment yields and dynamics, this seemed more relevant than using past data to reconstruct years that are more recent. Moreover, this choice results in a stricter evaluation, because there are less training data available from 2019/20 than from 2000/01. If we train (and tune) the QRF model based on the 2000/01 data (hereafter QRF\textsubscript{2000/01}) and validate it against 2019/20, we find that QRF\textsubscript{2000/01} performance is similar to QRF\textsubscript{2019/20} with respect to SSC and not as good as QRF\textsubscript{2019/20} with respect to SSY (see figure 1 in added pdf file). QRF\textsubscript{2000/01} performance with respect to SSC is clearly better than SRC performance.

- L373: Why these percentiles were chosen?

Answer: We chose these percentiles because they are more robust than the extremes (i.e. min and max), and because they cover 95 % of all estimates, which is common in our perception.

- L385-401: This 4.3 section seems like it should be mentioned in the Methods. I would suggest to place appropriate mentions of this in the Methods section, before such an important validation check on the methods is reported as a result.

Answer: We agree. We will move the first paragraph to the methods.
I question the independence that the authors refer to here. One catchment is nested within the other, and the data at one location was used to correct the data at the other location. This introduces some level of dependence between the two datasets thus they cannot be described as independent.

Answer: Thank you. What we tried to express here, is that we deem it unlikely that e.g. changes in measurements could have caused these shifts at both locations at the same time. The two gauges are nested, but the annual discharge at gauge Vernagt is only about 15 % of the annual discharge in Vent, so if the increase had only occurred at gauge Vernagt, it would not necessarily be visible at gauge Vent, much less to this extent. Also, we need to clarify that only precipitation data at gauge Vent were corrected using precipitation data from gauge Vernagt. Discharge data and temperature time series were measured and used completely independently.

We agree that “independently” is not be the right word here and will correct that, yet we do not think this changes out conclusions.

# Technical corrections

- L57: Please clarify for who the timescales are relevant; relevant for management?
  
  Answer: Thank you, we will clarify that we are referring to relevant timescales for investigating changes associated with anthropogenic climate change.

- L75: remove e.g.
  
  Answer: There are more factors and we only named the most relevant ones for our case, which is why the e.g. makes sense here. More information can then be found in the cited paper (Huss et al., 2017).

- L78: long enough data -&gt; long term data
  
  Answer: Thank you, we will change this.

- L96: machine-learning -&gt; machine learning; this term is never defined which would be beneficial for reader unfamiliar with it
  
  Answer: Thank you, we will add a definition.

- L97: In past studies: QRF has not only been used in geomorphology. I would suggest adding a qualifier here to narrow the scope of the sentence
  
  Answer: Thank you, we will do that.

- L102: data situations -&gt; data availability
  
  Answer: Thank you, we will change this.

- L103: bear -&gt; leads to
  
  Answer: Thank you, we will change this.
- L103: and taken together [...] -> so that, taken together, they give [...] 
Answer: Thank you, we will change this.

- L104: location -> catchment 
Answer: Thank you, we will change this.

- L106: with respect to trends, which -> for trends, some of which 
Answer: Thank you, we will change this.

- L145: The legend for Figure 1 refers to gauge then catchment for the two areas of interest; it would be clearer if only one type was mentioned 
Answer: We attempted to describe it in the hydrologically correct way, thus we suggest leaving it as it is.

- L173: in daily resolution -> at a daily resolution 
Answer: Thank you, we will change this.

- L190-191: I would move "since 2006" after "turbidity has been measured" 
Answer: Thank you, we will change this.

- L255: "developments": I am unsure what the authors mean here by developments: is it related to methods or evolution? 
Answer: We are referring to long-term changes in catchment dynamics. We will clarify this.

- L260: remove "truly" 
Answer: Thank you, we will do that.

- L267: extraordinary -> rare 
Answer: Thank you, we will change this.

- L269: benefit of the opportunities -> benefit from these opportunities 
Answer: Thank you, we will change this.

- L272: "fig. 2": the way figure are referenced is inconsistent: it is sometimes "fig", "Fig", or "figure". Please harmonize. 
Answer: Thank you, we will do that.

- L279: repaired -> corrected; to match the language used in Fig. 2 
Answer: Thank you, we will adjust this.

- L280: 2000/01 -> 2000-2001; and everywhere else where the authors use this notation instead of the full years separated by an hyphen
Answer: Thank you, we will change this.

- L288: 3.2 Analysis of results: this section number is wrong as the previous section was already 3.3
Answer: Thank you, we will correct this.

- L291: [t/time]: use either dimension [mass/time] or units [t/day] not both; also consider replacing t by Mg
Answer: Thank you, we will change this to mass/time.

- L302: When introducing the Nash-Sutcliffe efficiency, it would be beneficial if the authors provide its range and directionality so that readers unfamiliar can interpret the following figures more easily by knowing that a value of one relates to good performance
Answer: Thank you, we will add this.

- L349: remove "As described earlier"
Answer: Thank you, we will remove this.

- L350: in daily resolution -> at that resolution
Answer: Thank you, we will change this.

- L350-351: rewrite this sentence; right now it reads as if the loss is crucial whereas it is the information or the impact of its loss that is
Answer: Thank you, we will change this.

- L386: please add a reference to this statement since "it is known"
Answer: Thank you, we will add a reference.

- L418: A square exponent is missing in the units of the specific suspended sediment yield
Answer: Thank you, we will correct this.

- L425-429: Should this two-sentence paragraph be merged with the previous paragraph?
Answer: Thank you, we combine this paragraph with the following paragraph..

- L468: where -> for which, remove "which was"
Answer: Thank you, we will change this.

- L472: remove "in the time"; not significant -> no significant
Answer: Thank you, we will change this.

- L506: before we discuss -> then we discuss
Answer: Thank you, we will change this.

- L511: the term "critical point" has very precise meaning in the study of dynamical
system, I would advise using "significant change point" rather than "critical point".

Answer: Thank you, we will adjust this.

- L518: extraordinary -> rare

Answer: Thank you, we will change this.

- L540: several reasons -> three reasons

Answer: Thank you, we will change this.

- L541: Firstly -> First, - L542: Secondly -> Second, - L544: And thirdly -> Third

Answer: Thank you, we will change this.

- L550: please add a reference to this statement since "it is known"

Answer: Thank you, we will add a reference.

- L641: gap of knowledge -> knowledge gap

Answer: Thank you, we will change this.

References

Al-Mukhtar, M.: Random forest, support vector machine, and neural networks to modelling suspended sediment in Tigris River-Baghdad, Environ. Monit. Assess., 191, 673, https://doi.org/10.1007/s10661-019-7821-5, 2019.

Francke, T., López-Tarazón, J. A., and Schröder, B.: Estimation of suspended sediment concentration and yield using linear models, random forests and quantile regression forests, Hydrol. Process., 22, 4892–4904, https://doi.org/10.1002/hyp.7110, 2008.

Li, B., Yang, G., Wan, R., Dai, X., and Zhang, Y.: Comparison of random forests and other statistical methods for the prediction of lake water level: a case study of the Poyang Lake in China, Hydrol. Res., 47, 69–83, https://doi.org/10.2166/nh.2016.264, 2016.

Pilla, R. M. and Williamson, C. E.: Earlier ice breakup induces changepoint responses in duration and variability of spring mixing and summer stratification in dimictic lakes, Limnol. Oceanogr., 67, S173–S183, https://doi.org/10.1002/lno.11888, 2022.

Veh, G., Lützow, N., Kharlamova, V., Petrakov, D., Hugonnet, R., and Korup, O.: Trends, Breaks, and Biases in the Frequency of Reported Glacier Lake Outburst Floods, Earths Future, 10, e2021EF002426, https://doi.org/10.1029/2021EF002426, 2022.

Yadav, V., Ghosh, S., Mueller, K., Karion, A., Roest, G., Gourdji, S. M., Lopez-Coto, I., Gurney, K. R., Parazoo, N., Verhulst, K. R., Kim, J., Prinzivalle, S., Fain, C., Nehrkorn, T., Mountain, M., Keeling, R. F., Weiss, R. F., Duren, R., Miller, C. E., and Whetstone, J.: The Impact of COVID-19 on CO2 Emissions in the Los Angeles and Washington DC/Baltimore Metropolitan Areas, Geophys. Res. Lett., 48, e2021GL092744, https://doi.org/10.1029/2021GL092744, 2021.
Please also note the supplement to this comment:
https://egusphere.copernicus.org/preprints/2022/egusphere-2022-616/egusphere-2022-616-AC1-supplement.pdf