Dear Roberto Greco, dear anonymous referees,

We would like to thank you again for your very detailed comments, questions and suggestions. Below, we provide our response as direct answers to each comment and point out the respective changes to the manuscript. Please be aware, that the line numbers and chapters mentioned in the "Changes made" sections refer to the latest version of the manuscript.

We hope that our changes will be to your satisfaction.

Best, Lena Katharina Schmidt on behalf of all authors

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

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.
Changes made: *We have restructured and re-written the discussion section to better refer to the larger scope, as well as the conclusions.*

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.*
Changes made: *We have harmonized the use of tenses throughout the manuscript.*

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").*

Changes made: *We have improved the description (L 235 et seqq.).*

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.
Changes made: *We have improved the description of the aim of the study (L. 88 et seqq.).*

- 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.*
Changes made: *We have improved the description in the manuscript (L. 260 et seqq.).*

- 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.*
Changes made: *We have provided more details on QRF, by adding a description and explanation of the used predictors and improving the description of the optimization and the ancillary predictors (i.e. antecedent conditions) (section 3.2.). We also added a reference to the github repository, where the model version used in former studies can be found (section 3.2, L 237).*

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.*
Changes made: *We defined the term "mcp" in line 386 et seqq.*

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.*
Changes made: *We have improved the description in lines 380 et seqq.*

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.*
Changes made: *We have added this to the description in L 256 et seqq.*

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.*
Changes made: *We have emphasized the possibility of using the number of exceedances to evaluate the representativity of the training data set (conclusions, discussions (l. 629 et seqq.), and mentioned it in the abstract) and encouraged future studies to do so, especially with respect to future estimates.*

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 results 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.*

Changes made: *We have made clear that we refer to annual SSY and added a justification of the 250 realizations in lines 266 et seqq.*

- 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 <u>past</u> 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 $QRF_{2000/01}$) and validate it against 2019/20, we find that $QRF_{2000/01}$ performance is similar to $QRF_{2019/20}$ with respect to SSC and not as good as $QRF_{2019/20}$ with respect to SSY (see figure 1 below). $QRF_{2000/01}$ performance with respect to SSC is clearly better than SRC, performance.

Changes made: *We added the NSE and BE values for a model trained on the 2000 and 2001 data and validated against 2019 and 2020 in line 432 et seqq. and added the respective data points to figure 4 a).*

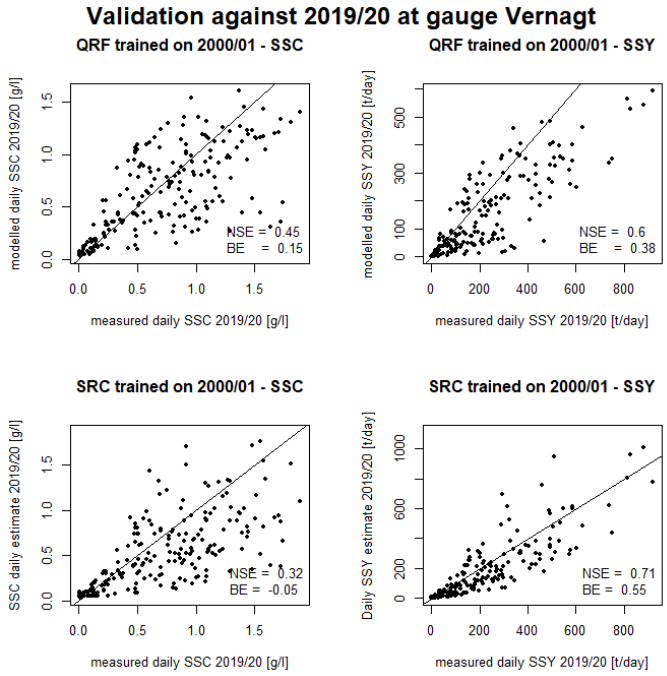

Figure 1 Validation of QRF models and sediment rating curves trained on 2000 and 2001 data against 2019/20 data. Top: QRF; Bottom: Sediment rating curve; Left: SSC estimates; right: SSY estimates.

- 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.*

Changes made: *We added a explanation in line 438 et seqq.*

- 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.*

Changes made: *We have moved the first paragraph to the methods (line 206 et seqq.).*

- L575: "independently": 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.
Changes made: *We improved the description in the methods, so that it becomes clear that only precipitation data at gauge Vent were corrected using data from gauge Vernagt and replaced the word "independently" by a more adequate description in lines 728 et seqq.*

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.*
Changes made: *We have clarified as suggested (L50).*

- 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).*
Changes made: *We have left the e.g., as suggested.*

- L78: long enough data -> long term data
Answer: *Thank you, we will change this.*
Changes made: *We have adjusted this, as suggested (L72).*

- L96: machine-learning -> machine learning; this term is never defined which would be beneficial for reader unfamiliar with it
Answer: *Thank you, we will add a definition.*
Changes made: *We have added a short definition and a reference for further reading (L89).*

- 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.*
Changes made: *We have adjusted this, as suggested (L91 et seqq.).*

- L102: data situations -> data availability
Answer: *Thank you, we will change this.*
- L103: bear -> 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.*
Changes made: *We have adjusted these issues, as suggested (L106 et seqq.), although some have become obsolete because we rewrote the sentence.*

- L106: with respect to trends, which -> for trends, some of which
Answer: *Thank you, we will change this.*
Changes made: *We have adjusted this, as suggested (L109).*
- 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.*
Changes made: *We have left this, as suggested (L148).*

- L173: in daily resolution -> at a daily resolution
Answer: *Thank you, we will change this.*
Changes made: *We have adjusted this, as suggested (L287).*

- L190-191: I would move "since 2006" after "turbidity has been measured"
Answer: *Thank you, we will change this.*
Changes made: *We have adjusted this, as suggested (L305).*

- 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.*
Changes made: *We have adjusted this, as suggested (L176).*

- L260: remove "truly"
Answer: *Thank you, we will do that.*
Changes made: *We have removed this, as suggested (L181).*

- L267: extraordinary -> rare
Answer: *Thank you, we will change this.*
Changes made: *We have adjusted this, as suggested (L188).*

- L269: benefit of the opportunities -> benefit from these opportunities
Answer: *Thank you, we will change this.*
Changes made: *We have adjusted this, as suggested (L190).*

- 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.*
Changes made: *We have harmonized this throughout the manuscript.*

- L279: repaired -> corrected; to match the language used in Fig. 2
Answer: *Thank you, we will adjust this.*
Changes made: *We have adjusted this, as suggested (L201).*

- 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.*
Changes made: *We have adjusted this throughout the manuscript.*

- 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.*
Changes made: *We have corrected 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*.
Changes made: *We have adjusted this, as suggested (L337).*

- 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.*
Changes made: *We have adjusted this, as suggested (L354 et seqq.).*

- L349: remove "As described earlier"
Answer: *Thank you, we will remove this.*
Changes made: *We have removed this, as suggested.*

- L350: in daily resolution -> at that resolution
Answer: *Thank you, we will change this.*
Changes made: *We have adjusted this, as suggested (L405).*

- 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.*
Changes made: *We have adjusted this, as suggested (L405 et seqq.).*

- L386: please add a reference to this statement since "it is known"
Answer: *Thank you, we will add a reference.*
Changes made: *We have added a reference (L208).*

- L418: A square exponent is missing in the units of the specific suspended sediment yield
Answer: *Thank you, we will correct this.*
Changes made: *We have corrected this, as suggested (L516 et seqq.).*

- L425-429: Should this two-sentence paragraph be merged with the previous paragraph?
Answer: *Thank you, we combine this paragraph with the following paragraph..*
Changes made: *We have combined the paragraphs, as suggested.*

- L468: where -> for which, remove "which was"
Answer: *Thank you, we will change this.*
Changes made: *This became obsolete, because we rewrote the paragraph to describe the newly added figure in the Appendix.*

- L472: remove "in the time"; not significant -> no significant
Answer: *Thank you, we will change this.*
Changes made: *We have rewritten this sentence (L531 et seqq.).*

- L506: before we discuss -> then we discuss
Answer: *Thank you, we will change this.*
Changes made: *We have decided to erase this paragraph in the course of restructuring the discussion.*

- 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.*
Changes made: *This sentence became obsolete during the rewriting and restructuring of the discussion, but we do not use the term "critical point" in the manuscript.*

- L518: extraordinary -> rare
Answer: *Thank you, we will change this.*
Changes made: *We have adjusted this, as suggested.*

- 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.*
Changes made: *We have adjusted this, as suggested in these two comments (L612 et seqq.).*

- L550: please add a reference to this statement since "it is known"
Answer: *Thank you, we will add a reference.*
Changes made: *We have added a reference (L627).*

- L641: gap of knowledge -> knowledge gap
Answer: *Thank you, we will change this.*
Changes made: *This sentence became obsolete during the rewriting and restructuring of the conclusions.*

RC2: 'Comment on egusphere-2022-616', Anonymous Referee #2, 18 Nov 2022

I appreciate the opportunity to review the manuscript, entitled 'Reconstructing five decades of sediment export from two glaciated high-alpine catchments in Tyrol, Austria, using nonparametric regression'. The topic is study is of great importance to not only the earth and environmental science community but also the policymakers and practitioners such as hydropower companies and water resource managers. This study presents an attempt to reconstruct the long-term suspended sediment export in alpine glacierized basins based on the available shorter records and machine learning. Despite some limitations, the proposed method is capable of reconstructing the sediment yield over the past decades with satisfactory performance.

Major comment 1: Based on modelling scheme in Figure 2, the model validation should target SSC, which is very reasonable and necessary. While, in the results section, the authors

only validate the performance of sediment discharge and sediment yield, which are the product of discharge and SSC. In your model (Quantile Regression Forest), discharge is also one of the model input variables and important predictors. The high validation coefficients (NSE and BE) could be only part of the story and maybe just because discharge appears in both input and output variables. Thus, I would kindly suggest the authors try to re-validate the model performance using SSC and replace both Qsed and sSSY in Figure 3-5 with SSC as shown in figure 2 if possible.

Answer: *Thank you for this comment. Indeed, we need to state more clearly, that e.g. the tuning of the models is performed on daily/hourly SSC (not daily Qsed). However, the quantity that we are ultimately interested in is (annual) sediment yield, as we want to understand whether the amount of sediment transported from the catchments changed over time. Adding to this, we find that yields are a more meaningful way to aggregate to annual resolution than mean annual SSC, because of the skewed nature of the concentration distribution. In mean annual SSC, low concentrations on days at the beginning and end of the season are given the same weight as high concentrations during the glacier melt season when discharge is also high – so actually, most of the sediment export happens during the glacier melt period. We believe that this can be captured better using sediment discharge and annual yields.*

Thus, we suggest to add NSE and BE calculated on SSC to the text. As you can see below, the values do not change substantially, if we use SSC instead of Qsed in validation A (hourly vs. daily model resolution at gauge Vernagt, figure 3a):

Hourly model:	*NSE(Qsed) = 0.98,*	*NSE(SSC) = 0.97*
	BE(Qsed) = 0.97,	*BE(SSC) = 0.95*
Daily model:	*NSE(Qsed) = 0.89,*	*NSE(SSC) = 0.82*
	BE(Qsed) = 0.84,	*BE(SSC) = 0.73*

In validation B (model trained on 2019/20 and validated against 2000/01 at gauge Vernagt), the NSE = 0.51 and BE = 0.33 still represent a satisfactory model performance (Moriasi et al., 2007; Pilz et al., 2019)*, as does model performance at gauge Vent (comparing SSC from turbidity to out-of-bag model estimates) with NSE = 0.6 and BE = 0.43. For mean annual SSC at gauge Vent, the NSE is even as high as for annual yields (NSE(SSC) = 0.825 vs. NSE(SSY) = 0.832).*

Changes made: *We have stated more clearly, why we focus on SSY instead of SSC (L162) and added the NSE and BE based on SSC to the text (L413 et seqq. and L431 et seqq.).*

In the introduction, the authors say that "Quantile regression forests (QRF) (Meinshausen, 2006) are a multivariate non-parametric regression technique based on random forests, that have performed favorably to sediment rating curves" (paragraph 95). Although it is proven in other publications, I think this statement still needs to be tested and evaluated in this study. If possible, I would suggest the authors compare the SSC simulations by QRF model and SSC simulations by sediment rating curves and explicitly demonstrate how much improvement can be done by the QRF model than sediment rating curves.

Answer: *Thank you for this valuable comment. When comparing daily SSC estimates using sediment rating curves (SRC) to QRF at gauge Vernagt (VF), we find that SRC estimates are in fact slightly better in validation B, i.e. when we train both QRF and SRC solely on SSC from 2019/20 at gauge Vernagt and compare modelled to measured SSC values in 2000/01 (see figure 2 below). However, when using the full dataset, SRC performance is worse than QRF performance, even though QRF performance considers out-of-bag estimates only. Thus,*

SRC performance gets worse with a larger training dataset, which already demonstrates that SRC cannot describe the variability in SSC as well as QRF.

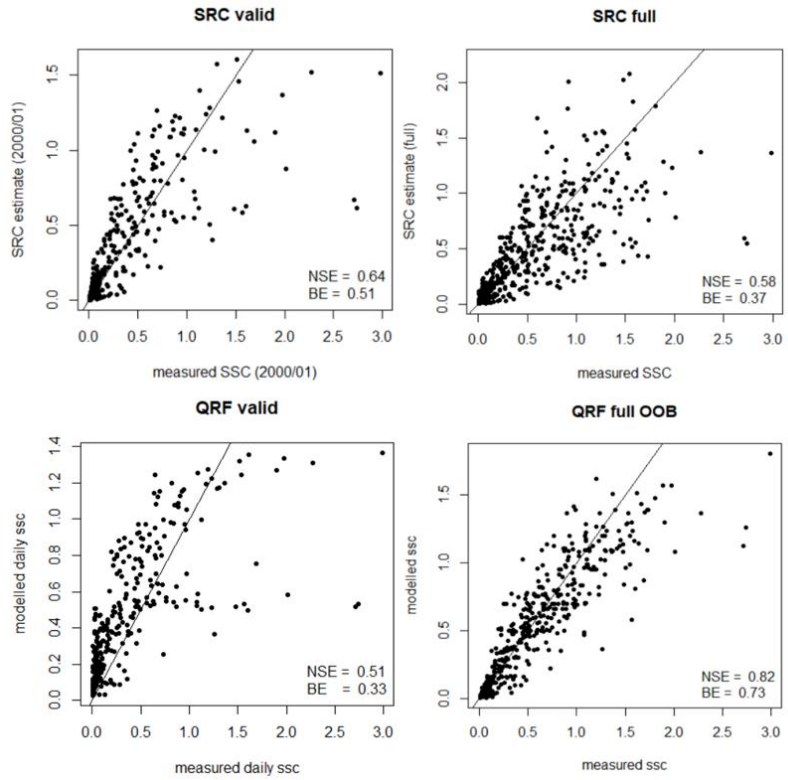

Figure 2 Comparison of sediment rating curve to QRF performance when both are trained solely on daily data from 2000/01 (left) and on all available training data (2000, 2001, 2019, 2020; right).

Likewise, mean daily SSC at gauge Vent is represented better by out-of-bag QRF estimates than by SRC (see figure 3 below). Adding to this, compared to gauge VF more years with turbidity measurements are available, so that performance with respect to annual yields can be evaluated (figure 3 c). Here, mean annual SSC estimated through SRC yields a negative NSE, indicating that the mean observed value would be a better predictor (Moriasi et al., 2007). *In contrast, annual values based on QRF show very good performance.*

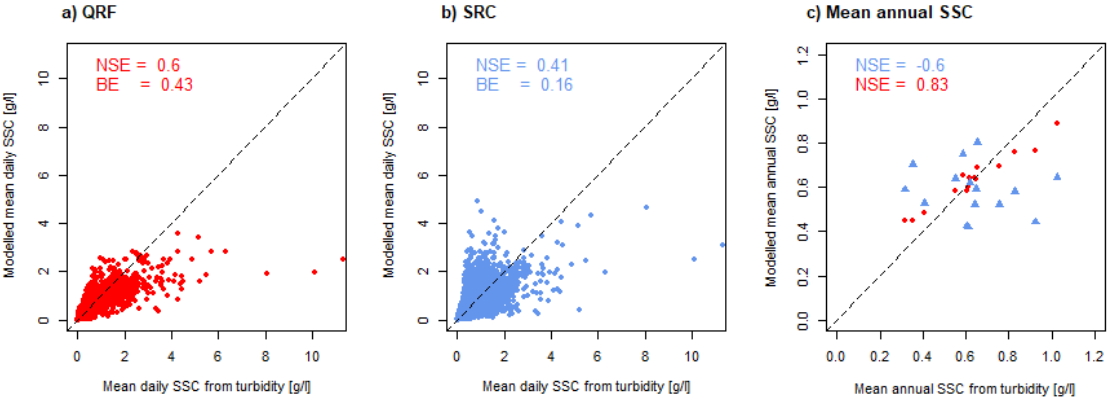

Figure 3 Performance of QRF (a) and sediment rating curves (b) compared to mean daily SSC derived from turbidity measurements at gauge Vent. Panel c) shows mean annual SSC estimates based on QRF (red circles) and SRC (blue triangles).

Changes made: *We have added a comparison to sediment rating curves (see table 1 and surrounding text) and extended figure 5 accordingly (L 510). This is of course explained in the methods (L211 et seqq.), and revisited in the discussion and the conclusions.*

Major comment 2: Usually, most of the annual sediment load is contributed by several extreme sediment events and they could cause severe socio-ecological-economic impacts. However, for the daily-scale model, such episodic high Qsed events are always underestimated, especially for the smaller nested basin Vent. Apart from the insufficient observations as training data as the authors discussed already, can this be also given rise to the different erosion and sediment transport processes during the episodic high-flow events and the threshold effect in sediment transport (see ref below)? If so, is that possible to re-fine such underestimation and consider the different transport mechanisms in Quantile Regression Forest Model? Zhang, T., Li, D., East, A.E. *et al.* Warming-driven erosion and sediment transport in cold regions. *Nat Rev Earth Environ* (2022). https://doi.org/10.1038/s43017-022-00362-0

Answer: *Thank you for this interesting question. Firstly, it is important to note that (unlike in many other fluvial systems), the majority of the annual sediment load in the Ötztal is not transported by several extreme events: on average, only about 21 % of the annual yield is transported by events associated with precipitation (Schmidt et al., 2022). The most extreme event captured in the measurements (i.e. from 2006 to 2020) was in August 2014, where 26 % of the annual yield were transported in 25 h. We assume that this event was associated with mass movements, unfortunately though there are no field data available from this instance. In August 2020, we observed a mass wasting event in the Vent catchment that lead to 13 % of the annual yield being transported at gauge Vent within 30 h. However, these events constitute exceptions.*

Secondly, since QRF is a statistical model, it is not possible to consider different transport mechanisms as such. However, the way the (ancillary) predictors were configured, is assuming that they can be proxies for certain processes; e.g. temperature as a proxy for melting processes or precipitation in time slices before the day to be modelled as a proxy for antecedent moisture conditions (see also (Francke et al., 2008)*). Unfortunately, to our awareness there are no other data available to re-fine the model further (such as thaw depths in permafrost etc., which could potentially describe these processes even better).*

Thus, we do already have some (presumed and observed) mass wasting events within the time series. This provides the opportunity for the model to learn that sediment yields are especially high under certain conditions (e.g. intense precipitation and high temperatures and/or high antecedent moisture conditions) and that precipitation (which translates to other transport mechanisms) might become a more important predictor at these times.

This represents an advantage compared to e.g. sediment rating curves, where such threshold effects cannot be described.

Adding to this, it is important to understand that figure 3 a) and 5 a), which we assume you are referring to, show out-of-bag data, i.e. the model prediction for such an extreme event, if this particular event is not part of the training data. So, underestimation is less severe in the full model. We will express this more clearly.

Changes made: *We have stated more clearly, that QRF can model threshold effects in principle, which is an advantage compared to sediment rating curves (L 28, L 677 et seqq., L 771); that the majority of annual yields are not transported by events (L655 et seqq.); that we refer to out-of-bag-data (L409 and L499); and added a detailed description and rationale of the predictors (L236 et seqq.).*

Major comment 3: As the authors introduced in Methods, Quantile Regression Forest Model is driven by discharge, temperature, and precipitation, and only a few years' sediment

observations are used for training the model. The reconstructed long-term sediment yield series is highly dependent on the input hydroclimatic predictors. Thus, I guess it's not surprising that the abrupt change in sediment yield coincides with the hydroclimatic abrupt change. Is that possible for the authors to collect any other relevant erosion, sedimentation, or landscape change data to independently prove the abrupt change in sediment transport in this region?

Answer: *Thank you for this question. However, unlike in sediment rating curves, it is not necessarily the case that we would observe an abrupt change in modelled sediment concentrations if there is one in the predictors, because with QRF there is not necessarily a linear or monotonous relationship between input and output. Adding to this, we will state more clearly, that the glacier mass balances were not part of the model predictors, so these already are relevant data that independently show an abrupt change, as you are referring to. We suggest to state this more clearly in the results / figure 7. Beyond that, to our knowledge there are no other long-term data from our catchment that could be used as continuous model drivers in daily resolution.*

Changes made: *We have reorganized figure 7, with the three primary predictors (Q, P, T) in the top and mass balances in the bottom to visualize more clearly that mass balances are not part of the predictors. We changed the order of the text in section 5.2 accordingly. We have also tried to make this more clear in the text (e.g. lines 546, 581). Adding to this, we have extended the description of the chosen predictors for the QRF model in section 3.2. as mentioned above.*

Specific comments:

1. The abstract can be substantially shortened with at most two paragraphs.
   Answer: *Thank you for this suggestion. We will streamline some parts of the abstract, but suggest to keep the indicated level of detail to provide a meaningful summary of the manuscript.*

   Changes made: *We have fundamentally rewritten and shortened the abstract.*

2. Introduction: there is a lack of acknowledging the existing literature on multi-decadal sediment observations in other high mountain areas and cold regions such as in the Tibetan Plateau, Andes, and the Arctic.
   Answer: *Thank you, we will integrate this.*

   Changes made: *We have included this (L75 et seqq.).*

3. Line 35: Considering the distinct underestimation of high sediment yield events. I would suggest the authors to be careful about the statement and clarify the possible insufficiency: "Our findings demonstrate that QRF performs well in reconstructing past daily sediment export".
   Answer: *Thank you for this suggestion. We will clarify this.*

   Changes made: *We have clarified this throughout the manuscript).*

4. Line 50: Impacts of sediment transport on hydropower production and reservoir sedimentation are also systematically elaborated in ref below: Li, D., Lu, X., Walling, D.E. *et al.* High Mountain Asia hydropower systems threatened by climate-driven landscape instability. *Nat. Geosci.* **15**, 520–530 (2022).

https://doi.org/10.1038/s41561-022-00953-y
Answer: *Thank you, we will include this.*

Changes made: *We have included this (L44).*

5. Line 60: The recent review systematically elaborates on the sediment dynamics and hydrogeomorphic processes in cold regions and discusses their complexity: Zhang, T., Li, D., East, A.E. *et al.* Warming-driven erosion and sediment transport in cold regions. *Nat Rev Earth Environ* (2022). https://doi.org/10.1038/s43017-022-00362-0
Answer: *Thank you, we will include this.*

Changes made: *We have included this (e.g. L54).*

6. For introduction and discussion: some of the other quantitative evaluations of the climate change impacts on sediment transport in high-mountain rivers based on decadal observations are listed below for further reading.
Zhang, T., Li, D., Kettner, A. J., Zhou, Y., & Lu, X. (2021). Constraining dynamic sediment-discharge relationships in cold environments: The sediment-availability-transport (SAT) model. Water Resources Research, 57, e2021WR030690. https://doi.org/10.1029/2021WR030690
Li, D., Lu, X., Overeem, I., Walling, D. E., Syvitski, J., Kettner, A. J., ... & Zhang, T. (2021). Exceptional increases in fluvial sediment fluxes in a warmer and wetter High Mountain Asia. Science, 374(6567), 599-603.
Answer: *Thank you, we will include them.*

Changes made: *We have included these references.*

7. Line 175: “see map” is unclear. do you mean "Fig. 1" or the other map?
Answer: Thank you, w*e will adjust this reference, it refers to fig.1.*

Changes made: *We have adjusted this as suggested.*

8. Line 165: the section numbering is quite confusing here. Please check this issue throughout the paper.
Answer: *Thank you, we will check that throughout the manuscript.*

Changes made: *We have corrected and checked this issue throughout the manuscript.*

9. Figure 3: the meaning of the black dash line should be explained in the caption. Besides, the actual sSSY values for the four observed years should be highlighted in Figure 3b, for evaluating the model performance.
Answer: *Thank you, we will add that (it is the 1:1 line) and highlight the points.*

Changes made: *We have explained the 1:1 line, but realized that highlighting the points is potentially misleading in this context, since the figure does not show modelled vs. measured values, but modelled values from the daily vs. modelled values from the hourly model.*

10. Line 240: the 5-fold cross-validation results are shown in any figures or tables or appendix. I would suggest the authors add at least one display item to show this result.

7. Figure 2: Why there is no validation for Vent station? It seems that the extrapolation ability at this station can be tested by the cross-validation.
Answer (to 10 and 11): *Thank you for these comments. We suggest to add a table to show cross-validation results at gauge Vent (see below). For gauge VF, we suggest to extend the evaluation of validation B (training on 2019/20 and validation on 2000/01) to training on 2000/01 and validation on 2019/20, which is more descriptive and reasonable given the limited temporal extent of the data (see also answers to reviewer 1).*

Table 1 Results of cross validation at gauge Vent with respect to mean daily SSC given as nash-sutcliffe efficiency (NSE) of out-of-bag predictions of the full model (Nash_OOB_full), for each 1/5 of the time series in the cross validation ("nash1" to "nash5") and the mean NSE of the 5 cross validation periods.

nash_OOB_full	nash1	nash2	nash3	nash4	nash5	meanNS
0.61	0.48	0.55	0.21	0.69	0.39	0.46

Changes made: *(With respect to the two comments above) We have added a table of the cross-validation results at gauge Vent (table 1, in section 4.2) and added a comparison with a cross-validation using sediment rating curves and the same periods.*

8. Figure 7c-d: the summer discharge trends are not shown, please add the summer discharge results and be consistent with the main text.
Answer: *We will add July discharge to figures 7c-d, consistent with temperature, and the main text.*

Changes made: *We realized that figure 7 would be too cluttered and unclear if we add the discharge data of several months. Instead, we added two figures showing trends and change points (if significant) for each month for the two gauges, to the Appendix, as this level of detail goes beyond the constraints of the results section.*

7. line 510: "satisfactory results" usually refer to the estimations with no significant overestimations and underestimations. Here, for accuracy, the authors should clarify that satisfactory results are found in annual sSSY estimations and there are underestimations for high Qsed events at the daily scale.
Answer: *Thank you, we will clarify that.*

Changes made: *We have clarified this throughout the manuscript (see also specific comment 3).*

9. Lines 580: an in-depth comparison with the world's cold regions would greatly enhance the discussion. For the sudden, tipping-point-like shifts of sediment transport in response to climatic changes have also been observed in the headwater of the Yangtze River on the Tibetan Plateau. The relative contributions of different factors can be also disentangled. Li, D., Li, Z., Zhou, Y., & Lu, X. (2020). Substantial increases in the water and sediment fluxes in the headwater region of the Tibetan Plateau in response to global warming. Geophysical Research Letters, 47, e2020GL087745. https://doi.org/10.1029/2020GL087745
Answer: *Thank you, we will add that to the discussion.*

Changes made: *We have added this in the restructuring and rewriting of the discussion (L725 et seqq.).*

RC3: 'Comment on egusphere-2022-616', Anonymous Referee #3, 21 Nov 2022

Review of the manuscript (egusphere-2022-616): Reconstructing five decades of sediment export from two glaciated high-alpine catchments in Tyrol, Austria, using nonparametric regression by Lena Katharina Schmidt, Till Francke, Peter Martin Grosse, Christoph Mayer, Axel Bronstert

Summary: In this manuscript, the authors apply quantile regression forest (QRF) to simulate suspended sediment concentration (SSC) at the outlet of two nested glacierized catchments in Upper Ötztal in the Tyrolean Alps, in Austria. As predictors, they use discharge, precipitation and temperature. The QRF model(s) are used to generate long-term (1967-2020, 1974-2020) time series of mean daily SSC and specific annual suspended sediment yield (sSSY), which are later analyzed for trend analysis and point change detection. To identify causality for such trends and abrupt changes, the authors apply the same statistical analysis to the observations of precipitation, temperature, discharge, and mass balance of the two largest glaciers within the study area.

General comments: I think that the aim of the manuscript of understanding the potential of machine learning techniques to model SSC in alpine catchments, including climate variables as predictors, is valuable. However, in my opinion several aspects require substantial revision.

Major revisions: The methodology is not sufficiently explained. The authors should clarify better mainly: (1) the quantile regression forests approach and the selection of the antecedent conditions for the predictors, (2) the procedure followed to fill-in the missing data (how did you compute the correction factors?) as well as (3) to disaggregate the data. Likewise, the availability of data and their resolution is quite confusing and requires clarification.

The authors frame some parts of the manuscript in a way that is conceptually questionable and potentially misleading. First, when the authors talk about 'reconstruction of sSSY', they should clarify well that the analysis of sSSY is based solely on simulations of SSC derived with a QRF model. Not only the QRF model cannot reproduce values outside of the range of values of the training dataset, but also the processes of sediment production and transport might have changed over time. Second, given the nature of the model, it is expected that trends and changes in the predictors lead to trends and changes in SSC. Therefore, I suggest that the authors discuss the trend analysis of the predictors before or together with the trend analysis of sSSY .

Answer: *Thank you for this valuable comment. To avoid being misread, we suggest to replace the term "reconstructing" with "estimating". Of course, the processes of sediment production and transport might have changed over time. That is why we designed the predictors in a way that they can be seen as proxies for these processes. For example, sediment production in these areas will be a function of temperature (glacier melt and movement, sub- and proglacial sediment transport, potential permafrost thaw), as well as potentially precipitation and antecedent moisture conditions (hillslope erosion, slope destabilization) and sediment transfer is tightly linked to discharge. As mentioned in the answers to reviewer 2, it is not necessarily the case that changes in the predictors lead to trends and changes in SSC, as QRF is not a linear model. Since we analyze trends and change points in the predictors but also in the glacier mass balances – which are independent data and not part of the predictors – we suggest to leave the order as it is, since it would otherwise be necessary to open a new chapter for the glacier mass balances.*

Changes made: *We have changed the terminology to "estimated" instead of "reconstructed" where possible. However, we also kept the term "reconstructed" in some places, because we*

find that it mode adequately describes the temporal connection, i.e. estimating past SSC / sSSY. To avoid misunderstandings, we have defined this at the beginning of the methods (L166). We have added a detailed explanation of the employed predictors (section 3.2) and changed the order in figure 7 and the according text in the results. Additionally, we have combined the discussion of trends and change points in the results and predictors as you suggested.

I think that it would be interesting to quantify the trends and shifts in SSC, to analyse how much the change in sSSY is related to a change in discharge, in SSC or in both. This would allow understanding if the increase in sediment load is due to an increase in transport capacity, in sediment supply or a combination of the two.

Answer: *Thank you for this comment. We agree that it is an interesting question. To put it briefly, we find roughly the same change points and trends in mean annual SSC as in sSSY (see figure 4 below): mcp yields change point around 1980/1981 for both locations, while the Pettitt test disagrees in Vernagt (likely due to its limitation with respect to the beginning and end of time series). However, mean annual SSC considers low concentrations in spring (when discharge and flux are also low) with the same importance as days in August (with higher SSC but also much higher discharge and higher fluxes). That is why we focus on annual yields (and changes therein), because we find them to be more adequate and meaningful to aggregate to annual resolution.*
Changes made: *We have explicitly stated and explained why we focus on SSY at the beginning of the methods section (L 162 et seqq.)*

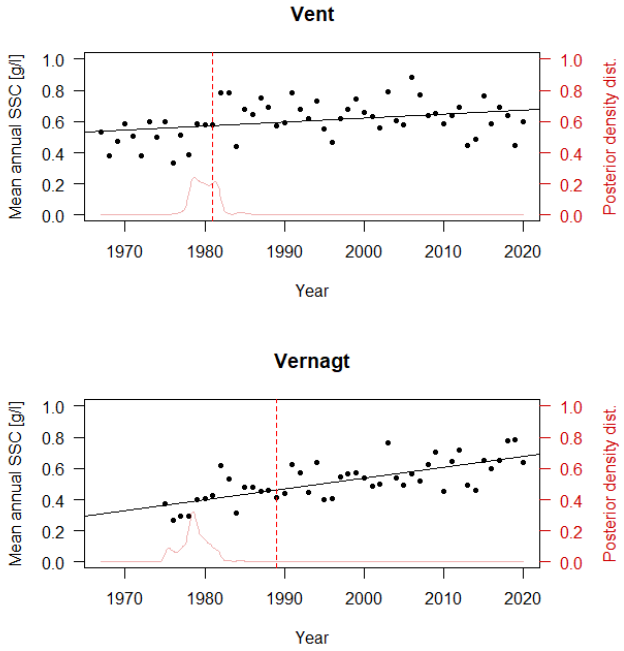

Figure 4 Mean annual SSC estimates over time with change points determined by Pettitt test (dashed) and with the mcp package (solid line).

In both Validation A and B, models fail to capture the largest SSC values. As discussed by the authors, this is likely related to the inherent limitation of QRF in extrapolating beyond the range of values of the training data. It would be important to quantify the impacts of this limitation on the total suspended sediment yield. I suggest that authors compute the fraction of total suspended sediment yield transported during these 'extreme' days.

Answer: *Thank you for this important point. It makes us aware, that we need to improve our explanations and point out more clearly, that both figures you are referring to show out-of-bag estimates, i.e. the model estimate for a day with very high SSC is based only on those trees, that do not "know" this day. Thus, it can be seen as a quite rigourous validation, and means that the performance of the full model for these days (or days with similar conditions) will be better.*

We calculated the difference between daily Q_{sed} based on turbidity and daily Q_{sed} from the out-of-bag model estimates of the full models, to assess the underestimation in annual SSY for the 10 days with the highest Q_{sed} in the turbidity time series. The underestimation on these 10 days represent 0.6 to 2.8 % of the annual SSY at gauge Vernagt and 1.7 to 19.1 % of annual SSY at gauge Vent. However, the 19 % underestimation stem from the most extreme event in the time series in August 2014, where 26 % of the annual SSY was exported within 25 h – likely associated with mass movements (Schmidt et al., 2022)*. The full (i.e. non-OOB) model estimate for this day only shows an underestimation of 6 %. We will add that to the discussion.*

Changes made: *We have added the analysis described in our answer above to L500 et seqq and refer to it in the discussion (L654).*

It is not clear to me, which is the added value of using P and T? This could be quantified by running the QRF models excluding either precipitation or temperature and evaluating their performances. Likewise, I think that it would be interesting to run the QRF model without discharge. This would contribute to understand the relevance of the predictors and to estimate the potential of using such models in ungauged catchments.

Answer: *Thank you for this interesting question. Variable importance can be analyzed for QRF models, by interpreting the so-called "variable importance" for the related RF-models, e.g. by quantifying the decrease in model performance, if a predictor is permuted (see Fig. 4 below). At both gauges, discharge is the most important predictor, but at gauge Vent, temperature and the derived predictors and the day-of-year are also above 10 % IncMSE (average increase in squared residuals if the variable is permuted). Precipitation and the derived predictors (such as precipitation of the antecedent 24 and 48 h) are less important. At gauge Vernagt, short-term precipitation is also less important, but long-term antecedent precipitation (up to 53 days) is the second most important predictor.*

However, the interpretation of these analyses is not straightforward because the predictors are partially correlated (as can easily be imagined with temperature and discharge, as glaciers melt) and thereby "share" some importance. That implies that if we perturb one predictor, some of the information would still be present in the correlated predictor. Secondly, predictor importance is also likely to vary thoughout the season, calling for a more elaborate analysis. Thus we suggest not to include this in the manuscript.

We would not recommend applying QRF in ungauged catchments, firstly, because discharge is a very important predictor, and secondly, because the model needs to be trained for each site, and needs training data, also and especially of suspended sediment concentrations.

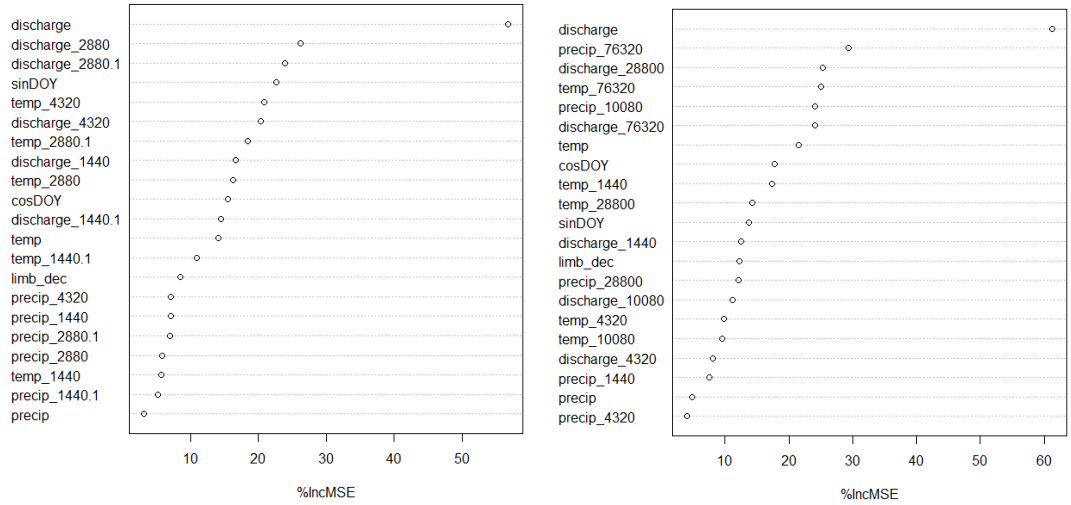

Figure 5

Left: Variable importance at gauge Vent. Right: Variable importance at gauge Vernagt

Changes made: *We added a detailed rationale of the employed predictors in section 3.2.*

Specific comments:

Ln. 166-171: Which is the resolution of the discharge data? Please, specify.
Answer: *The answer is a bit complex, there are different periods of times for the gauges where different temporal resolutions are available. That is why we did not state it here but prepared the table in the Appendix. We will add a reference to it here.*
Changes made: *We have added a reference to the table (L274 et seqq.).*

Ln. 174-176, Ln. 181-183: How did you compute the 'conversion factors'? Over which time period?
Answer: *We derived linear relationships between e.g. the available Temperature at gauge Vent and Vernagt for all dates when data were available at both measurement stations and used this linear model (as stated in the brackets in the text) for conversion. We will clarify this in the revised manuscript.*
Changes made: *We have added a description (L288 et seqq.).*

Ln. 179: which resolution?
Answer: *60 min resolution since 1974, 10 min in 2000 and 2001 and 5 minutes ever since. We stated this in the table in the Appendix and will add a reference to it here.*
Changes made: *We have added a short explanation and a reference to the table in the Appendix (L284 et seqq.).*

Ln. 234-244: Please, in addition to the reference to Zimmermann et al., 2012 provide clarification for the antecedent predictors.
Answer: *Thank you, we will make this more clear.*
Changes made: *We have added a more detailed explanation of the antecedent predictors (L237 et seqq.).*

Ln. 208-2010: How did you disaggregate the data? How did you use the 10-min data? In the gap-filling part?
Answer: *Thank you, we will provide details on how we disaggregated the data. The disaggregation only refers to the gap-filling model at gauge Vernagt, where precipitation and temperature data from 2000 and 2001 had to be disaggregated from 60 to 10 min. resolution.*

Discharge and temperature, that were given as hourly means before, were adopted as is for the 10 min timesteps and precipitation sums were divided by 6. Exactly, the 10-min data were used in the gap-filling model. We will clarify this.
Changes made: *We have clarified this (L325 et seqq.).*

Ln. 211-214: I find this paragraph confusing. Please, clarify.
Answer: *Thank you, we will clarify this.*
Changes made: *We have rewritten this paragraph (L 329 et seqq.).*

Ln. 259-265: Please, move this chapter to chapter 3.1
Answer: *You probably refer to the chapter up until line 285 and including figure 2 and are suggesting to first describe the general approach, then the data and then the model in detail? Thank you, we will do that.*
Changes made: *We have moved this part to section 3.1, in the beginning of the methods.*

Ln. 272-274: Does it make sense to first use a model to estimate the SSC data, and later use the modelled SSC to estimate a model? I think that it would be more correct to exclude from the QRF the time steps in which SSC is not available.
Answer: *Thank you for this question. For our purpose, these steps were indispensable to supply the model with the full range of values, which were originally lost in the limited turbidimeter data. Moreover, the modelled SSC from the gap-filling model are only a small part of the training data of the validation and reconstruction models, and we train the gap-filling model on different data, i.e. suspended sediment concentration samples instead of turbidity. These samples include times when the turbidity probe failed but also when it reached saturation. Keeping in mind the range-sensitivity of QRF, we argue that it is important to add these data.*
Changes made: *No changes.*

Ln. 277-278: Did you train the QRF models on all available data? Please, clarify.
Answer: *Thank you, we assume you are referring to the models in Validation A? Yes we did and we will clarify.*
Changes made: *We have clarified as suggested (L198).*

Ln. 281-282: Is this model different from the daily model of Validation A?
Answer: *Yes, it is different (see above), we will make this more clear.*
Changes made: *Sorry, we were mistaken in our previous answer, since we now realized you were referring to the final reconstruction model instead of the validation model in validation B. To clarify: the daily model in Validation A is the same as the final reconstruction model.*

Ln. 291: Please, clarify q-weighted.
Ln. 300: Please, clarify equation 2.
Answer: *Both comments refer to the Q-weighted SSC. We will clarify.*
Changes made: *we have clarified this (L336 et seqq.).*

Ln. 349-350: I understood that at gauge Vernagt predictors were available at hourly resolution (see ln. 259-260). Do you mean that the predictant, SSC, is daily? Please, clarify better in chapter 3.1. Data availability and resolution is very confusing.
Answer: *Thank you, we will clarify this. It is correct that predictors at gauge Vernagt are available in hourly resolution, but at gauge Vent, long-term data are only available in daily resolution. Since we wanted to ensure comparability between the gauges – and Validation A*

showed that the loss of model skill is small, we focused on daily resolution, which also helped keeping computational times reasonable.
Changes made: *We have clarified that we refer to gauge Vent here (L405), added references to the overview table in the Appendix (L276, L285, L295) and have explained that we use the daily resolution models at both gauges in L201 and L417.*

Ln. 369-370: I agree that NSE and BE are quite good, in the context of suspended sediment transport. However, I wonder how much the largest values, which are substantially underestimated by the model, contribute to the total suspended sediment yield. Quantifying this would help assessing the model performance.
Answer: *Thank you; we answered this question above (lines 730 et seqq. of this document).*
Changes made: *See above.*

Ln. 385-401: Please, move this chapter to the chapter about data (3.1).
Answer: *Thank you, this has also been suggested by reviewer 1. We will do that.*
Changes made: *We have moved this to chapter 3.1 as suggested.*

Ln. 471: CP is not defined previously.
Answer: *Thank you, we will define CP here.*
Changes made: *We have defined CP in the methods (L380) and results (L554).*

Ln. 476-477: please, describe more in details the mass balance record.
Answer: *Thank you, we will do that.*
Changes made: *We have added that these are glaciological mass balances and added a reference for further reading, which is openly accessible (L582 et seqq.).*