EGUsphere, referee comment RC1
https://doi.org/10.5194/egusphere-2022-616-RC1, 2022

Comment on egusphere-2022-616

Anonymous Referee #1

Referee comment on "Reconstructing five decades of sediment export from two glacierized high-alpine catchments in Tyrol, Austria, using nonparametric regression" by Lena Katharina Schmidt et al., EGUsphere, https://doi.org/10.5194/egusphere-2022-616-RC1, 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. 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.

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.

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.

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.

- 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.

- 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. 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. 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). 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.

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.

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.

- 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)?

- L373: Why these percentiles were chosen?

- 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.

- 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.

Technical corrections

- L57: Please clarify for who the timescales are relevant; relevant for management?

- L75: remove e.g.

- L78: long enough data -> long term data

- L96: machine-learning -> machine learning; this term is never defined which would be beneficial for reader unfamiliar with it

- 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

- L102: data situations -> data availability

- L103: bear -> leads to

- L103: and taken together [...] -> so that, taken together, they give [...]

- L104: location -> catchment

- L106: with respect to trends, which -> for trends, some of which

- 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

- L173: in daily resolution -> at a daily resolution

- L190-191: I would move "since 2006" after "turbidity has been measured"

- L255: "developments": I am unsure what the authors mean here by developments: is it related to methods or evolution?

- L260: remove "truly"

- L267: extraordinary -> rare

- L269: benefit of the opportunities -> benefit from these opportunities

- L272: "fig. 2": the way figure are referenced is inconsistent: it is sometimes "fig", "Fig", or "figure". Please harmonize.

- L279: repaired -> corrected; to match the language used in Fig. 2

- L280: 2000/01 -> 2000-2001; and everywhere else where the authors use this notation instead of the full years separated by an hyphen

- L288: 3.2 Analysis of results: this section number is wrong as the previous section was already 3.3

- L291: [t/time]: use either dimension [mass/time] or units [t/day] not both; also consider replacing t by Mg

- 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.

- L349: remove "As described earlier"

- L350: in daily resolution -> at that resolution

- 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

- L386: please add a reference to this statement since "it is known"

- L418: A square exponent is missing in the units of the specific suspended sediment yield

- L425-429: Should this two-sentence paragraph be merged with the previous paragraph?

- L468: where -> for which, remove "which was"

- L472: remove "in the time"; not significant -> no significant

- L506: before we discuss -> then we discuss

- 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".

- L518: extraordinary -> rare

- L540: several reasons -> three reasons

- L541: Firstly -> First

- L542: Secondly -> Second

- L544: And thirdly -> Third

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

- L641: gap of knowledge -> knowledge gap