Response to Reviewer #2:

First, we would like to thank reviewer #2 for the feedback on our manuscript. Reviewer #2 provided general as well as specific comments which we will address below. While the original review comments are shown in italics, our responses are given in regular font.

***General Comments:***

*This well written paper describes a set of seven flume experiments in a sand box in order to mimic conditions and controls of fill-terrace formation. The main controls explored are changes in water Qw and sediment Qs discharge and changes in base level. The paper gives a nice and consistent description of current terrace formation theories, models and controls. It gives a clear description of the experiments and relates them in a transparent way to current model insights on fluvial dynamics. The derived conclusions are supported by the sand box experimental evidence but the translation to field evidence is not equally well considered and not always supported by evidence (there a quite some constraints related to the physical experiments). The main limitation of this investigation is that all results and relationships found are only valid for a flume sand box system which cannot be linearly scaled up to real world system without some critical considerations and reflections.*

*First of all is the sand box experiment dealing with a relatively short and steep fluvial system with Qw,in = Qw,out. The setup resembles, in a qualitative way, more an alluvial fan system than a large mature fluvial system that are usually studied in the cited terrace studies.*

First, we thank the reviewer for the positive feedback on our work.

We further acknowledge that upscaling is common problem when transferring experimental results to natural settings. Therefore, we will include a paragraph in which we discuss potential scaling problems and why the results from experimental work are still useful.

We agree that the channel system is relatively short and steep, but includes the fundamental feedbacks: water and sediment inputs, base level, and a channel that responds to these forcings. And we do believe that our setup differs from an alluvial fan setting because it has a narrow, defined outlet, which ensures that the river stays within a confined valley. Typical processes observed on alluvial fans during aggradation are gradual channel migration and avulsion (sudden changes in channel position), which results in an overall widening of the actively reworked alluvial fan area in downstream direction. This can be seen for example in experimental setups from Whipple et al. (1998), Kim et al. (2006a,b) and Martin et al. (2009). In our experimental setup, however, the confined outlet forces the river to stay ‘in place’ and limits avulsions. In addition, alluvial fans are often characterized by superelevation, i.e. elevation in the central part can be higher compared to the fan margins. As our main purpose is to study process-behavior of a system, we assume that the preservation of processes (e.g., dominance of lateral migration over avulsion) is more important than the absolute scaling of the slope. The slope of the river is a function of the sediment supply and water discharge. As such, we could have chosen a gentler slope of the river. The reason for the stepper references slopes was to produce pronounced differences in channel geometry within all the different settings.

Also, the results of the experiment are qualitatively similar to those of the numerical alluvial channel simulations of Wickert & Schildgen 2019.

*Secondly, is the ‘fluvial system’ studied a braided system only, while many studied and cited terrace systems are thought to be initiated when the fluvial system switched from a braided to (more) meandering state (and back).*

Within our review section (section 2), we indeed cite field studies of terraces that were formed within braided as well as within meandering channel systems. The purpose of this section is to give an overall overview of the different processes of terrace formation.

However, a large number of the studies cited refer to terraces that were formed in braided channels system only. These studies include for example: Scherler et al. (2015), Schildgen et al. (2016), Tofelde et al. (2017), Norton et al. (2015), Faulkner et al. (2016), Fuller et al. (1998), Malatesta et al. (2018), Malatesta and Avouac (2018), Bookhagen et al. (2016), McPhillips et al. (2014), Dey et al. (2016), Steffen et al. (2009), Steffen et al. (2010), Hu et al. (2017) and Litty et al. (2016).

Nevertheless, we agree that our experimental approach is restricted to terrace formation in braided systems only. We will clarify the text accordingly.

*Finally has the used methodology the issue of reproducibility. If we would repeat the same experiments in the same sand box would we get the same terraces (properties) and results? This is crucial to know because the laser scanning allows us to measure very small changes (with known uncertainties) but if there is significant other uncertainty ('noise') in the sand box data of a higher magnitude we might be over interpreting the data. As long as we do not know the 'noise' in the experiments we should be reluctant to draw too many conclusions from relative minor changes in elevation. I recommend to address these potential limitations in the discussion in a separate section.*

This is a good point and we agree that reproducibility is crucial. In the set of experiments contained within this manuscript we only repeated the control experiment (*Ctrl_1 and Ctrl_2*). The purpose of the control experiments was to investigate 'noise' within the system. We only interpret changes in morphology that are beyond the variability within the control experiments as externally driven adjustments.

Although we did not repeat the experiments that included external perturbations with exactly the same settings, we consider the last phase of the two experiments during which we performed two changes ($DQ_w\_IQ_w$ and $IQ_{s,in}\_DQ_{s,in}$) as repetition of the experiments with only one perturbation ($IQ_w$ and $DQ_{s,in}$), although with different absolute values of $Q_w$ and $Q_{s,in}$. The comparison of those experiments with each other reveals that the trajectories of channel evolution (longitudinal profiles, slope, width (Fig., 4 and 5)) is robust. In addition, the final $Q_s$ and $Q_w$ settings of the experiments with two changes ($DQ_w\_IQ_w$ and $IQ_{s,in}\_DQ_{s,in}$) were equal to the reference settings (*Ctrl_1*, *Ctrl_2* and 'spin-up' time setting of all experiments but *BLF*). When comparing the slope values to which all those sub-experiments evolve, the final slope values are very similar (around 0.07). Although not being exact repetitions of the same experiments, the evolution to the same equilibrium conditions is an indicator that the results are reproducible.

However, we agree that despite the apparent repeatability based on two different experiments, our number of repeat experiments is very limited. Moreover, whereas the behavior of channel morphology evolution is repeated, we lack knowledge on the reproducibility of terrace properties, especially the lag-times. We will expand the discussion to include these limits.

*Having raised these concerns I do believe the experiments generate an interesting set of criteria and hypotheses that could and should be more rigorously tested on real world systems and be evaluated in numerical models. I will certainly test some of the proposed relationships on existing terrace field evidence and with numerical modelling. I therefore recommend to publish this publication after revisions.*

Thank you.

***Specific comments:***

*The validity of the results and relationships observed are certainly more valid for fluvial fan type settings where also transport distances are relatively short and gradients are steep and we only observe braided behavior. In such real world systems we actually do observe differences in gradients between different fill type terraces. The large and longer fluvial systems are often characterized by almost parallel gradients of* *preserved terraces. Often terrace formation and preservation is linked to tributaries causing reach specific changes in Qs and Qw, something that has not been evaluated in the experiments.*

As already discussed above, we think that our setting is not entirely representative of an alluvial fan system due to the confined outlet, the absence of superelevation and the dominance of lateral migration over avulsion. In real world systems, terraces along the main stem can be parallel to the active channel (e.g., Hanson et al., 2006; Faulkner et al., 2016), but they can also vary in gradient (e.g., Tofelde et al., 2017; Poisson and Avouac, 2004; Baker and Gosse, 2009; Burgette et al., 2017). As terrace sequences along the main stem can be up to tens of kilometers in extent, changes in slope might not be so obvious locally and can only bet determined by detailed surface elevation surveys of those terraces.

We agree that many terraces are preserved at confluences of tributary channels and the main stem. Within this set of experiments, we only focus on terraces that form along the main stem to keep the setting as simple as possible and investigate the direct effects of changes in $Q_s$, $Q_w$ or base-level on changes in bed elevation and terrace formation. Adding a tributary channel adds another level of complexity due to possible internal feedback mechanism between the main stem and the tributary. We also have performed experiments in which we focus on the interaction of a tributary and the main stem. This work is currently in preparation. We think that including another set of experiments, with a detailed focus on tributary-main stem interactions, would overload this manuscript and also draw the focus in a different direction. However, for clarification, we will add an explaining sentence that this study only investigates terrace formation along the main stem.

*The link between landscape dynamics and Qs,in is another scaling challenge. Landscapes often display a delay between environmental changes and sediment flux responses. These response lags can be even an order magnitudes larger than the lag-times within the fluvial system itself. This is related to coupling and decoupling of hillslope dynamics to the fluvial system.*

We agree that changes in sediment supply from the hillslopes to the channels can lag behind any changes in environmental conditions that might cause an adjustment of the supply rate. In this study, we only investigate the response of the fluvial part to variations in input conditions, and we do not have the ability to address lag times between environmental forcing and hillslope responses. Following pioneers like Stanley Schumm and Philipp Allen, we consider a sedimentary source-to-sink system as systems that can be subdivided into three zones – the erosion zone, the transfer zone and the deposition zone. Each of those zones has its own responses and response timescales to external perturbations. We only investigate the transfer sub-system of a source-to-sink sediment transport system. The transfer sub-system connects the erosion zone (hillslopes) with the final deposition zone (e.g. a terrestrial or marine basin). As such, we only investigate response or lag-times of the transfer sub-system and do not investigate delays between sediment supply from hillslopes to river channels. Although we have stated this in the manuscript (p. 2 l. 2, p.2 l. 17-19, p. 6 l. 4), we will add additional clarification that our experimental set-up only allows us to investigate the response of the transfer sub-system to external perturbations.

*The autogenic dynamics analysis requires more thought. We can only discard them if they do not occur after longer repeated runs under 'stable' conditions. It seems there is more autogenic dynamics related in the transient response of channel width, an aspect in the model results that are not as detailed analyzed as the terrace profiles, surface slopes and signal propagation.*

We apologize, but this seems to be a misunderstanding. We do not discard autogenic terrace formation. On p. 11 l. 3-10 we state that we did not observe any autogenic terrace formation after the 'spin-up' time, but that the absence of such terraces does not mean that autogenic terraces do not exist. We also state that most mechanisms of autogenic terrace formation, which were summarized in section 2, could not be tested with our experimental setup. For clarification, we will adjust the section (p.11 l. 3-10) to clearly state that we do not discard autogenic terrace formation.

*I like the prediction that net deposition along the channel leads to the majority of the grains at the outlet being freshly delivered from hillslopes (assuming hillslope coupling). While during incision older material is reworked in the outlet material, potentially yielding older ages (with cosmogenics).*

Thank you.

*In terms of the boundary conditions of the physical experiments I have the following remarks/questions: How realistic is a constant Qs,in input? In reality sediments are released as sediment waves into the fluvial system.*

We agree that sediment supply from hillslopes to the channel (erosion zone to transfer zone) can be highly variable. The further downstream transport of the sediment in the river, however, is then limited by the availability of water. As alluvial rivers are limited by their transport capacity and not by the availability of sediment, we consider the $Q_{s,in}$ for a given channel reach within the transfer zone as less variable compared to sediment supply from hillslopes to the channel itself. For clarification, we will adjust the text such that our experiments only investigate the geomorphic response of the transfer zone of a source-to-sink system (see comment above). The constant water discharge prescribed in the experiments is also a difference to natural channels that are dominated by variable discharges. In a way, we are 'compressing time' and assume that the experiments integrate over a number of large floods in natural channels; therefore, the timescales cannot be scaled directly.

*How important are the initial conditions? (referring to initial channel and 'spin-up' phase).*

We assume that the initial conditions play a minor role as we only look at changes in the system once the system is close to equilibrium. If the initial conditions were different, we expect the time to reach steady conditions to be longer or shorter (depending on the initial conditions). The two experiments during which we performed two changes ($IQ_s\_DQ_s$ and $DQ_w\_IQ_w$), both result at the initial slope value after the conditions have been changed back to reference conditions (Fig. 5C, E). As such, we expect the initial conditions mainly affect the 'spin-up' time required to reach stable conditions.

*What is the effect of stopping the experiment for the laser scanning? Doesn't this 'disturb' the experiment? A comparison between two equal runs with and without stopping could answer this issue? If this has been investigated before, please cite the relevant literature on this.*

Unfortunately, it is not possible to scan the surface without stopping the experiment for two reasons: (1) The laser scanner is mounted directly above the setting and it scans the surface in five lines parallel to the flow direction. Those five lines largely overlap and are merged after finishing the scans. The scanning of all five lines requires about 5 min. A continuation of the experiment would alter the surface morphology during the scanning time, such that the overlapping parts could not be merged anymore. (2) The water supplied to the experiments is dyed blue (Fig. 2). The reason is to enable the automatic detection of wet and dry pixels from the overhead photos. For the automatic detection, significant color differences between the water and the surrounding sand is necessary. The laser scanner, however, cannot penetrate the dyed water. As such, the experiments have to be stopped to be able to scan the surface topography. For the two reasons listed above, a comparison as suggested by the reviewer is unfortunately not possible.

However, the experiments have also been stopped overnight. In those cases, a laser scan was performed after stopping the experiment in the evening and before starting it again in the morning. The DEM of difference (DOD) between those scans reveal no major changes in topography for example through drying of the surface and collapse of channel banks. Finally, the time to drain the system took only a few minutes, and therefore does not leave a lot of time for significant reworking of the surface. Our approach is common for these type of experiment, and so far, there is no indication in the literature that it causes significant problems.

*You give temporal lags in measured time. How would you scale this up to reality? (see fig 5)*

As already mentioned above, our experiments are simple in a way that sediment supply and water discharge are constant through time, such that we assume that the experiments integrate over a number of large floods in natural channels. This makes an absolute scaling of channel response time and lag-times between perturbation and terrace abandonment complicated. Rather, we see the advantage of our approach that we can observe the form of the response (e.g. decrease in slope follows and exponential pattern and not a linear one). As such, we can differentiate whether a terrace was abandoned instantly after the onset of perturbation or rather later during the transient channel response phase.

*A difference between the Qw and Qs experiments compared to the base level change scenarios is the there is far less accommodation space in the upper part for terrace preservation (a narrow steep incision) compared to the downstream section and its response to base level change. Shouldn't this not be included in the impact analysis of perturbations?*

This is an interesting point. We agree that if the channel widens downstream (as channels tend to do in real systems), there is indeed more "space" to accommodate terraces downstream than upstream. That might be reflected in the width of terraces formed upstream and downstream. Because of the fixed outlet, we have a limited capacity of the system to widen downstream, and therefore are not sure we can make a strong statement about downstream changes in accommodation space with our setup.

*I fully agree with the statement that simulating long-profile evolution requires an improved understanding of the transient response of channel width. I presume that the Wickert and Schildgen, 2018 relationship between S, Qs ,in and Qw are also only valid for braided sand box systems under transport limited conditions?*

Wickert and Schildgen (2019) derive a general set of equations for gravel-bed river long-profile evolution -- meaning that flows are bedload-dominated and lack bedforms. They also note that transient width response is a needed direction of future research, and limit their approach to the assumption that such channels will tend to have a near-equilibrium width (e.g., Parker, 1978), which is appropriate for gradual changes discharge or other drivers of width change. This equilibrium width is set such that the Shields stress against the bank is equal to the critical Shields stress for initiation of motion, which is also appropriate for experiments such as ours, in which the banks are not held together by cohesive forces. For further questions on this study, we refer the reviewer to the final (2019) published paper.

*This also implies uniform 'bedrock' lithology. In reality (all cited real world examples) tectonic stability doesn't exist, nor do uniform lithologies or transport limited conditions. I am not suggesting to exclude the comparison but be more sensitive of the differences.*

In our experimental setup, we only study alluvial rivers. Therefore, uniform bedrock lithologies are not of major concern compared to studies of bedrock channels, in which a lowering in slopes requires the erosion of bedrock, which indeed is influenced by lithology. In our case, the material that needs to be moved is sediment, and we consider its lithology of minor importance.

As already noted above, we make the assumption that the system is always in 'transport-limited' conditions. The same conditions characterize some of the cited real world examples. Several of the cited field studies refer to braided, alluvial rivers in mountain basins that are characterized by massive alluvial fills (Tofelde et al. (2017), Schildgen et al. (2016), Dey et al. (2016), Malatesta et al. (2018), Malatesta and Avouac (2018), Scherler et al. (2015), Burgette et al. (2017), Huntington (1907), Litty et al. (2016), Steffen et al. (2009, 2010)). In those settings, the amount of sediment that is transported out of the basin is restricted by the transport capacity of the river. As such, we consider those sites to be in transport-limited conditions. Concerning tectonic stability, we agree that it is unlikely to be maintained over very long time periods, but even over the millennial timescales that many alluvial features are formed, it is not uncommon to find areas where there is no substantial change in tectonic forcing.

*The view of terraces/floodplains as temporal storage space is a realistic one. The percentage of Qs,in is in temporary storage during experiment in total in time, in Fig 5 could be used to quantify this effect and the possible effect on cosmogenic age.*

Thanks for the suggestion. Indeed, as the absolute $Q_{s,in}$ values are known, we can quantify the percentage of sediment discharge ($Q_{s,out}$) that has been supplied from upstream and that has been remobilized from within the channel. We will include those quantitative information to Fig. 5 and adjust the text accordingly.

An absolute quantification of the effect on any cosmogenic nuclide concentration is not possible, as this would also depend on the age and initial concentration of the remobilized sediment, much of which, in these experiments, comes from the initial basin fill from before the sediment input started (as opposed to from alluvial fill formed during the experiment).