Response: We thank both referees for their constructive feedback and the handling editor for their support. The comments will substantially improve the manuscript. Below we address each point in detail, indicating changes made or clarifications to be added in the revised version.

The manuscript introduces a novel, largely automated structure-from-motion (SfM) photogrammetric system for high-resolution monitoring of soil surface change over 3.5 years on agricultural hillslopes. Synchronized DSLR cameras (triggered by rainfall events and timers) capture daily imagery, which a custom Python workflow processes: it time-synchronizes photos, applies a convolutional neural network to detect ground control points under varying conditions, and runs Agisoft Metashape SfM to reconstruct daily 3D soil-surface point clouds. From these, daily digital surface models and change-of-surface (DoD) maps are derived at millimeter-scale resolution. The method is validated against terrestrial laser scanning (TLS) and UAV photogrammetry. The data from a freshly tilled loess field demonstrates detailed topographic changes following tillage and rainfall. Overall, this approach is innovative and promising for tracking erosion dynamics at high spatial and temporal resolution.

Major Comments

- The manuscript is very comprehensive, but some methodological sections (e.g. the multi-step GCP tracking and neural network architecture) are very detailed. To improve readability, consider moving deep technical specifics (such as the exact CNN architecture, clustering parameters, or algorithmic steps) to an appendix or supplementary material. This would shorten the main text while still making the detailed methods available for interested readers.

Response: We agree to shorten and streamline the methodological sections to improve readability. However, we prefer to keep essential technical aspects in the main text to ensure that our approach remains understandable and reproducible. Details will be made concise and thereby avoiding an appendix.

- The authors claim that "the resulting high-resolution datasets are valuable for analysing erosion dynamics and validating soil erosion models." However, the study is primarily a methodological demonstration rather than an explicit evaluation of model validation. The conclusions should be tempered to reflect this. If model validation is to be a major claim, the manuscript needs to either provide analyses that link the data directly to model performance or clearly state that testing such models is outside the current scope.

Response: Thank you for highlighting this potential misunderstanding. We adapted the abstract accordingly. The corresponding sentence has been changed: "The monitoring system and workflow are transferable, and the resulting high-resolution datasets are expected to be valuable for analyzing erosion dynamics and validating process-based soil erosion models." However, in the results and discussion we do not want to make further statements as we believe we are not promising anything that we cannot offer. Also, in the conclusion chapter, we solely state that future research should focus on the implementation of these new measurements into erosion modelling/validation.

- A critical limitation is that the system records all surface elevation changes, including non-erosional effects (e.g. soil compaction or tillage settling). The observed strong negative correlation between rainfall and elevation loss in the first week post-tillage indicates both erosion and compaction. However, without a method to distinguish these processes, the results cannot be interpreted purely as erosion. The authors note this challenge but provide limited quantitative separation techniques. The referenced approach by Epple et al. (2025) has not been applied at this temporal scale. The authors should discuss and, if possible, implement methods to discriminate erosional changes (e.g. mass removal) from other changes. For example, integration with sediment traps or other measurements (see below) could help verify when soil material has actually been removed by erosion versus simply redistributed or compacted.

This is indeed a valuable aspect, and one we would like to emphasize for future research. Unfortunately, our measurement method, which was tested with an overflow weir and a sampler at the slope foot, proved inefficient. Small events often led only to redistribution or sedimentation in front of the device, while large events overloaded it. In addition, fauna activity caused sampler failures during some events. Practical constraints also limited our set-up: the system had to be placed far enough from the lowest station to allow tractor access across the slope. This requirement resulted in slope-parallel tillage tracks at the shallow slope foot, which further promoted sedimentation in front of the device. As a result, our approach did not produce sufficiently reliable results, and we therefore excluded sediment yield measurements from the manuscript. Nevertheless, we agree that such systems can be advantageous and have added suggestions for potential measurement device designs in the discussion section, immediately following the paragraph on compaction and settlement. Future set-ups should consider integrating a sediment-collecting system, at least until settling rates can be reliably quantified using a data-driven approach (see Epple et al., 2025). However, these systems have their own limitations, such as providing only averaged changes across the slope. Moreover, the absence of sediment in a trap does not imply that measured height changes are purely non-erosional: SfM can detect small-scale erosion and accumulation, capturing localized yields that may not reach the slope-foot trap.

- The monitoring system achieved data on only 55–69% of survey days. Hardware issues (e.g. rigs knocked over by storms) and maintenance downtime caused significant gaps. This undermines the claim of “near-continuous” monitoring. The manuscript should acknowledge this limitation more explicitly. In particular, discuss how the 31–45% data loss affects the claimed continuity, and consider suggesting improvements for hardware robustness or system redundancy. The data gaps also impact the applicability for validating models like RUSLE that rely on annual or longer-term averages; this should be discussed. For example, if many events are missed, event-based model validation is compromised.

Response: In our setup, we indeed faced the challenge of larger data gaps. However, many of these gaps were the result of inexperience and are unlikely to occur to the same extent in the future. We anticipate that researchers reproducing our setup will benefit from our lessons learned and thus experience fewer interruptions. Data continuity was ensured by operating three rigs simultaneously, so that at least one usually recorded measurements, meaning that model validation should not be compromised. Furthermore, our approach is particularly aimed to be well suited for the validation of physically based models, as we measure event-based changes with high spatial resolution. Long-term average values are not the focus of our design for future model evaluation. Other, more substantial gaps occurred during winter, when no measurable changes were expected due to frost and snow cover. The hardware itself proved robust; most issues were related to power supply. We will elaborate on these aspects in the revised manuscript.

- The system's accuracy varies spatially, with errors growing toward the edges of the region of interest (5–15 mm). Such spatial bias can systematically affect erosion calculations, especially for analyses covering the whole area. The authors should discuss this limitation, for instance by quantifying how much area is affected by larger errors and whether analysis should be restricted to a smaller central zone for reliable measurements. Additionally, providing confidence intervals or error bars (rather than just mean errors) would clarify how spatial uncertainty influences change detection.

Response: Thank you for your comment. We would like to note that the suggested accuracy analysis for quantifying areas affected by larger errors cannot be conducted in such a systematic manner, as these errors depend on the specific setup conditions. Providing numerical thresholds could be misleading, as no fixed limits exist; instead, each site requires its own evaluation. For this reason, we consider the precision map values to provide a site- and event-specific level of detection. These already incorporate spatially resolved 95% confidence intervals, ensuring that only significant changes are considered. Consequently, we do not see additional benefit in also plotting standard deviations. We further emphasize that change detection should primarily focus on the central region, where accuracy is highest. Nevertheless, we deliberately present the less favorable regions to openly discuss the possible errors. We will refine the manuscript to make these points clearer.

- GCPs were noted to move due to animals, tillage, and storms. It is unclear how these movements propagate into final surface uncertainty. The manuscript should quantify or at least discuss the impact of GCP displacement on accuracy. For example, how often did GCPs shift beyond measurement error, and how does that translate into potential vertical error in the DSMs?

Response: If any GCPs were found to have moved, they were excluded from the bundle block adjustment and, where possible, re-measured using either a total station or UAV-based photogrammetry. As we had installed more GCPs than strictly necessary for georeferencing, the potential movement of some GCPs could be managed without impacting the results. Consequently, no issues arose regarding DSM height errors. Such movements occurred irregularly and were easily detected during the BBA, as the affected points showed error magnitudes significantly higher than those of stable points. The models listed in Table 1 as not being calculable due to GCP failures were therefore mostly the result of complete GCP losses, primarily after winter or following unfavorable tillage. Any remaining potential GCP movements were detected and corrected for. We will make this clearer in the revised manuscript.

- High-resolution temporal data are more suitable for some erosion models than others. The discussion should explicitly address which types of models can benefit. For instance, process-based models that use detailed spatial inputs (such as RillGrow or similar rill erosion models) can leverage sub-daily or event-scale data. In contrast, simple empirical models (e.g. RUSLE) operate on annual average erosion rates and cannot directly utilize such fine temporal detail. The authors should clarify that the proposed monitoring approach is especially useful for validating physically based, high-resolution models, and acknowledge the limited benefit for annual-scale models. Explicit examples (like RillGrow vs. RUSLE) would illustrate this point.

Response: We will explicitly state that our system is most suited for validating physically based, high-resolution models (e.g., LISEM, RillGrow), and less applicable to empirical annual-scale models such as RUSLE. This clarification will be added to both introduction and discussion.

- The innovative multi-camera synchronization needs clearer explanation. The temporal drift correction procedure, in particular, should be described in more detail to ensure reproducibility. It would also help to state whether camera internal calibration was performed only once before the campaign or periodically (e.g. after storms), and whether any re-calibration was necessary during the study. Details on camera maintenance and calibration will improve confidence in the results.

Response: As part of revising the method description for conciseness, we will also streamline the explanation of the synchronization. Camera calibration was performed once before the initial setup and again after reinstallation following the rig collapse. Further calibration was unnecessary, as only approximate values for the focal length and principal point were required; these parameters are re-estimated during each model calculation. Distortion parameters were estimated once and assumed to remain temporally stable.

- The use of a 3-sigma filter to remove outliers may be too aggressive for detecting subtle erosion features; the authors should justify this choice or consider less aggressive filtering. Additionally, the chosen 5 mm grid resolution for DSM interpolation should be justified: is this based on camera resolution, expected soil roughness scale, or processing considerations? Discussing why 5 mm is appropriate (and whether coarser or finer resolutions were tested) would strengthen the methods.

Response: We chose a strict filter because combining images that do not correspond correctly would result in erroneous 3D models. The 3-sigma threshold was selected as it is a common standard or rule of thumb in geodetic measurements. This threshold was particularly relevant for images captured during rainfall events, where offsets had a greater impact, while it was less critical for the daily datasets. Given that thousands of images were collected, the aggressive filtering did not lead to significant data loss. The 5 mm resolution was chosen based on the distance between the cameras and the area of interest. We aimed to achieve the highest possible resolution without introducing interpolation artifacts due to data gaps, which would have occurred at finer resolutions. Lower resolutions caused a loss of detail and were therefore not considered, as higher resolution was achievable without drawbacks. We will expand on this rationale in the revised manuscript.

- Accuracy is currently given in millimeters, which is precise but not easily interpretable in terms of soil loss or erosion impact. The authors should consider converting key error metrics into more meaningful units (e.g. tonnes per hectare per year) or comparing them to typical erosion rates. Furthermore, statistical details are needed: provide confidence intervals on accuracy estimates, and separate systematic versus random errors if possible. An analysis of how accuracy depends on surface roughness or vegetation cover would also be valuable (since rougher surfaces may degrade accuracy).

Response: Thank you for your suggestion. We prefer to report height changes in millimeters, as we are using SfM photogrammetry, an established method in erosion studies where height change is a valid and commonly reported metric. Converting height changes to sediment yield using bulk density can be misleading, as bulk density can vary spatially across the area of interest and may introduce a false sense of measurement certainty. Moreover, the process-based erosion models we aim to support typically output height changes in millimeters as well. Regarding accuracy, we already provide significant values through the M3C2-PM method, which incorporates confidence intervals. Similarly, our comparison of SfM data to TLS uses M3C2, accounting for the level of detection based on 95% confidence intervals.

Concerning roughness, we consider this beyond the scope of our current study, especially since defining roughness from 3D data remains an active research topic with open questions regarding scale and metrics. Regarding vegetation, we applied filtering to remove vegetation as much as possible and chose study areas with minimal vegetation to focus on surface changes of bare slopes. A study emphasizing vegetation cover, including larger extents as done in Onnen et al. (2020), would be an interesting topic for future research.

- Where possible, compare the SfM-based erosion estimates to traditional measurement methods (e.g. sediment traps, erosion pins) to demonstrate practical utility. This could also help in distinguishing erosional vs. non-erosional volume changes.

Response: Thank you for your comment. However, we do not have such observations available for comparison. Additionally, similar comparisons have been conducted in previous, smaller-scale studies (e.g., Eltner et al., 2017).

- The study area was a bare, conventionally tilled plot, which simplified reconstruction. Real agricultural fields often have growing vegetation, which can complicate SfM. The authors mention using a machine-learning method to filter vegetation in some cases. It would strengthen the paper to discuss how the system could be adapted for vegetated conditions (for example, using spectral filters, morphological editing, or annual timing). The transferability claims should be more specific: what additional challenges or modifications would other sites (with different climate, terrain, or vegetation) require?

Response: Thank you for the suggestion. We will add some more details regarding transferability considering vegetation affects to the revised manuscript.

- A valuable addition would be to suggest or demonstrate integration with other erosion measurements. For example, combining surface change data with sediment trap records or runoff measurements could help distinguish erosion from compaction. If changes in the soil surface do not correspond to transported sediment, one could infer non-erosional processes. The manuscript should at least discuss this possibility and how future work could integrate multiple datasets.

Response: Addressed in our response to comment 3.

Minor Comments

Line numbering should be continuous throughout for ease of review reference.

We follow the Copernicus template, which uses discontinuous numbering.

Some abbreviations are introduced without definition (e.g. RTC, IoT, LoD, M3C2). Define all acronyms at first use.

Response: All mentioned acronyms are already defined at first use; we will double-check.

The discussion of transferability would benefit from concrete guidance: for instance, recommended mounting improvements (e.g. sturdier rigs, solar power redundancy), or software alternatives (since Agisoft Metashape is proprietary, the authors might suggest open-source SfM tools for reproducibility).

Response: Thank you for the suggestions. We will refine the manuscript to provide more detailed guidance where appropriate. Regarding open-source software, currently, there is no solution that fully meets our requirements, which include ease of use, thorough code documentation, extensive error analysis options, and more. While MicMac is an option, it is not straightforward for users without photogrammetry expertise. Colmap could also be considered; however, it has limitations concerning georeferencing capabilities. Given that the scope of our study differs, we prefer not to discuss other software in detail here. Nonetheless, we acknowledge the importance of open-source solutions and plan to focus on this in future work.