General comments from both referees

Below, we detail the specific changes made to the manuscript in response to the general comments provided by both referees. Please note the following colour coding: **[Orange]** indicates the referee's comment, **[Blue]** our response and action taken, **[Black]** the citations from the preprint, and **[Violet]** indicates the text as it now appears in the revised manuscript.

This section focuses exclusively on the resulting modifications to the text. Please see our responses regarding the general comments in our response letters from the open discussion phase (https://doi.org/10.5194/egusphere-2025-3391-AC1 and https://doi.org/10.5194/egusphere-2025-3391-AC2).

Anonymous; Referee #1, general comment 1:

In terms of the scientific advancement which are offered for soil erosion modelling, I find the conclusion to reiterate what is perhaps the most fundamental concept of the USLE - that temporal aggregation is required to generate acceptable predictions according to the central limit theorem - which isn't neccessarily surprising. Reiterating this point is nevertheless useful, however one could argue that "why" is more relevant than "if". While the study correctly demonstrates that the model fails at the annual timestep, it offers limited insight into the reasons for this failure. A key value of testing a model outside its intended use case is to diagnose its structural weaknesses; this potential is not fully realised here.

Thank you again for your comment. We agree that analysing the structural causes of model failure is crucial for scientific advancement. As we stated already in the previous response, we believe this was already discussed in our preprint.

In the revised manuscript, we have expanded the Discussion to identify structural, temporal and spatial limitations in more detail. Specifically, we made changes in Section 4.2, 4.5, Conclusion and added a new Section 4.4 to discuss the implications of spatial aggregation.

Preprint, L447-475: "The model's annual time step fails to capture these critical temporal coincidences, a limitation that becomes more pronounced when such events are infrequent."

Section 4.2; Revised manuscript; L512-513: "The model's annual time step fails to capture these critical temporal coincidences, a structural limitation that becomes more pronounced when such events are infrequent."

Section 4.2; Revised manuscript; L516-517: "This deficit could not be compensated using high-resolution input data (daily soil cover, high resolution and on-site rainfall measurements)."

New section 4.4; Revised manuscript; L578-589: "4.4 Spatial aggregation

For the structure-dominated watersheds, spatial aggregation was critical for overcoming watershed-specific model failures. While the model failed to produce any behavioural realisations for W06 (even temporally lumped), the spatially aggregated group achieved 1.35% behavioural realisations. This aligns with the scale-dependency concepts reviewed by De Vente and Poesen (2005), who note that process dominance shifts with spatial scale, often allowing models to perform adequately at larger scales even if they miss finer processes.

For field-dominated watersheds, the spatial aggregation acted primarily as an averaging of varying crop states, smoothing out the heterogeneity of cover conditions, and any given watershed peculiarities. In contrast, for structure-dominated watersheds, spatial aggregation facilitates the identification of behavioural parameter spaces by masking local structural inadequacies, such as the deposition dynamics in grassed waterways. Consequently, the differences between spatially aggregated field-dominated and structure-dominated watersheds can be attributed to their structural components."

Preprint, L447-475: "The current static connectivity and transport capacity parameters (p_{con}, b_{dep} and $k_{TC/G}$) cannot adequately capture these temporal variations and flux-dependent relationships, suggesting the need for a more dynamic parameterisation approach that accounts for both seasonal changes and influx response."

Section 4.5; Revised manuscript; L629-632: "The current static connectivity and transport capacity parameters (p_{con}, b_{dep} and $k_{TC/G}$) cannot adequately capture these temporal variations and flux-dependent relationships, suggesting the need for a more dynamic parameterisation approach that accounts for both seasonal changes and influx response if the model is applied to an annual time-step."

Preprint, L568-572: "The model was unable to produce behavioural realisations at annual timesteps based on our strict limits of acceptability criterion despite the small absolute prediction errors (eight-year MAE = 0.12 t ha^{-1} yr^{-1} for field dominated and eight-year MAE = 0.16 t ha^{-1} yr^{-1} for structure-dominated watersheds). For the field-dominated watersheds, the model particularly struggled with the simulation of annual sediment yields when individual extreme events dominated the annual sediment production."

Conclusion; Revised manuscript; L641-647: "The model was unable to produce behavioural realisations for watersheds optimised for soil conservation and sediment transport reduction at annual time steps based on our strict limits of acceptability criterion despite the small absolute prediction errors over all model realisations (eight-year MAE = 0.14 t ha^{-1} yr^{-1} for field-dominated and eight-year MAE = 0.29 t ha^{-1} yr^{-1} for structure-dominated watersheds). For the field-dominated watersheds, the model particularly struggled with simulating annual sediment yields. This is because soil conservation practices reduced the number of erosion events, yet some events (e.g., after potato harvest) retained a similar magnitude as in conventional cultivation."

Anonymous; Referee #1, general comment 2:

Linking to this broader point is the main methodological critique I have of the manuscript. Despite using an implementation design which is intended to understand model drawbacks and accept only a subset of simulations, the authors do not consider uncertainty on the individual USLE parameters but instead opt for the use of an error surface on the gross erosion predictions. Given that a subset of these parameters are propagated into the transport capacity (i.e. L, S, R and K), the setup potentially ignores important parameter interactions which may impact both the uncertainty quantification and the insights into the model's shortcomings. Considering the simplistic design of the model, parameter lumping seems avoidable.

You are correct that we opted to lump the uncertainty from the ABAG factors into a single error surface rather than sampling each factor in the Monte Carlo simulation. In our manuscript, we justified the use of the e_{sur} parameter in more detail to address the referees concern about parameter lumping.

Section 2.7; Revised manuscript; L322-327: “The decision to aggregate the uncertainty of the ABAG factors into a single error surface (esur), rather than sampling individual factors within the Monte Carlo simulation, stemmed from the nature of our input data. We assumed that parameterisation errors (apart from the P factor) were negligible due to the exceptionally high-quality monitoring data used as input for calculating the ABAG factors (section 2.4). However, as the ABAG is based on regressions that carry residual error, we pragmatically used an error surface to evaluate inherent model biases.”

Anonymous; Referee #1, general comment 3:

The choice of approach also induces a circularity into the argumentation of the study, where the lack of acceptable model realisations at the annual timestep is attributed to unconsidered uncertainty in the input factors (e.g. the (bio)physical impacts of conservation tillage on overland flow and erosion), which could have otherwise been considered in the modelling approach.

Thank you again for raising this point. As we noted during the open discussion phase, we respectfully disagree that our argumentation is circular. In our framework, the use of temporal aggregation functions somewhat similarly to a sensitivity analysis. Because temporally static USLE factors (K and LS factors) remain constant across all simulations, they cannot be driving the variability when results are aggregated into an eight-year average. This means that the model's variability is inherently linked to temporal dynamics (R and C factors), rather than unconsidered uncertainty in the input factors.

Because the model's performance is tied to these dynamic temporal factors, the core issue lies in how the model simulates specific events and their aggregation. We have refined our argumentation in Section 4.2. Rather than just attributing the lack of acceptable realisations to input uncertainty, we detailed that the annual timestep structurally fails to capture the extreme, event-driven nature of erosion. Aggregating over time smooths out these temporal coincidences, which explains the improved performance.

Preprint, L474-479: “The model's annual time step fails to capture these critical temporal coincidences, a limitation that becomes more pronounced when such events are infrequent. This temporal limitation aligns with findings by Risse et al. (1993), who demonstrated that USLE's model efficiency diminishes at the annual scale. When averaging over the eight-year study period, these extreme events are smoothed out, which explains the model's improved performance at longer timescales (Tab. 2). This observation supports the basic assertion that the USLE was designed to compute long-term soil losses (Wischmeier and Smith 1978).”

Section 4.2; Revised manuscript; L512-520: “The model's annual time step fails to capture these critical temporal coincidences, a structural limitation that becomes more pronounced when such events are infrequent. This temporal limitation aligns with findings by Risse et al. (1993), who demonstrated that USLE's model efficiency diminishes at the annual scale. When averaging over the eight-year study period, these extreme events are smoothed out, which explains the model's improved performance at longer timescales (Tab. 2). This deficit could not be compensated using high-resolution input data (daily soil cover, high resolution and on-site rainfall measurements). However, it is important to note that the episodic nature of erosion is always difficult to capture even with event-based models, and hence aggregation over time tends to improve any kind of erosion model. This is especially true if events are rare and the overall erosion values are small (Nearing, 1998).”

Joris Eekhout; Referee #2, general comment 1:

The objectives should be better defined. The first objective focusses on testing of the model's capabilities. This does not seem to be too ambitious. No matter what is the outcome, this objective will always be achieved. So please refine this objective to make it more ambitious. The second objective seems related to the GLUE approach, I suggest to explicitly include the GLUE approach in this objective. The last objective is similarly not too ambitious (either testing or analysing something will always be achieved).

We appreciate your suggestion to increase the ambition and clarity of our goals. We have rewritten the objectives in the Introduction to explicitly integrate the limits-of-acceptability approach and the GLUE framework.

Preprint, L98-103: "In this study we employ this limit-of-acceptability approach based on the GLUE framework, focusing on three main objectives: (i) testing WaTEM/SEDEM's capability to simulate sediment yields in micro-scale watersheds either characterised by in-field soil conservation or by in-field soil conservation plus linear landscape features designed to trap sediments, (ii) analysing the behaviour of model parameters that control erosion and sediment transport processes, and (iii) assessing the model's performance across different spatiotemporal resolutions through data aggregation."

Introduction; Revised manuscript; L93-105: "Here we employ a rejectionist limits-of-acceptability approach within the GLUE framework to test the widely used WaTEM/SEDEM model for representing soil erosion, transport, and deposition in soil conservation optimised agricultural watersheds, i.e. featuring a combination of in-field practices and sediment transport control structures. Specifically, we aimed to: (i) identify limits of acceptability of model error derived from measurement uncertainty in order to reject non-behavioural model realisations; (ii) develop a two-stage model conditioning process to test the fitness for purpose of WaTEM/SEDEM for representing the effects of both in-field conservation practices and sediment control structures on erosion, transport, and deposition; and (iii) test WaTEM/SEDEM under different levels of temporal (annual vs. eight-year means) and spatial aggregations (individual vs. grouped watersheds). We accomplish these objectives using a comprehensive, long-term monitoring dataset from Southern Germany, which provides high resolution model inputs (e.g. precipitation, crop-specific daily soil cover, etc.), as well as continuous surface runoff and sediment flux data for six micro-scale watersheds under optimised soil conservation and reduced sediment transport (Auerswald et al., 2001; Auerswald and Fiener, 2019; Fiener et al., 2019a)."

Joris Eekhout; Referee #2, general comment 2:

The concept of aggregating the data using different spatiotemporal resolutions has not been mentioned in the Introduction. I was expecting that this would go in some direction of using different spatial and temporal resolutions (different cell sizes and time steps, for instance). However, this is totally not the case. The authors instead use the long-term median model outcome, instead of the annual outcomes (the way USLE-type of models should actually be used). And the spatial aggregation is related to the two different conservation types considered. I'm not sure if this requires to be included in an objective.

We appreciate the referee's comment regarding the spatiotemporal analysis. We agree that clarifying the scope of aggregation is important. We believe that the concept of evaluating the model across different scales is now included in our revised third objective in the Introduction.

Introduction; Revised manuscript; L99-101: "(iii) test WaTEM/SEDEM under different levels of temporal (annual vs. eight-year means) and spatial aggregations (individual vs. grouped watersheds)."

We believe this sufficiently introduces the concept of aggregation in the Introduction, while the detailed description of the aggregation methodology is described in the Methods:

Section 2.8; Revised manuscript; L351-356: "Model outputs were evaluated at multiple spatiotemporal scales through sequential aggregation steps to analyse short-term dynamics against its intended long-term design: First, we temporally aggregated the sediment yields by calculating eight-year means for each individual watershed. Second, we spatially aggregated the simulated sediment yields by calculating their means for each watershed group (field- and structure-dominated), but keeping an annual resolution. Third, eight-year means for each watershed group were calculated (spatial and temporal aggregation)."

General comments from both referees:

Anonymous; Referee #1:

Secondly, the study implements a replication of the WS code but performs neither benchmarking against the standard model nor releases the source code openly. I am in favour of replications of models such as WS in popular programming languages such as Python (which although on average slower, permit easy data integration and parallelization as mentioned by the authors), but without showing at least a benchmarking use-case the current implementation lacks reproducibility and good modelling practice.

Joris Eekhout; Referee #2:

Lines 161-164: The authors included a code availability statement saying that the code is available on reasonable request. I highly suggest to make the code publicly available through an open-source repository such as GitHub or Zenodo.

We have addressed the concerns regarding code verification and availability. In Section 2.3, we added a statement confirming that the Python implementation was rigorously verified against the original Delphi codebase to ensure identical outputs. Regarding availability, we have provided the most recent WaTEM/ SEDEM code to the reviewers and published the R-code. Further, we updated the code availability statement accordingly.

Section 2.3; Revised manuscript; L171-173: "To ensure reproducibility and accuracy, we compared the Python implementation against the original Delphi codebase at each individual step of the translation process, verifying that it produced identical outputs for these test cases."

Code availability; Revised manuscript; L673-674: "The R code used to compute individual factors and statistics is available at https://zenodo.org/records/18714865. The Python WaTEM/SEDEM code is available upon reasonable request."

Data availability; Revised manuscript; L677-679: "The specific data used to compute individual factors and statistics can be found at https://zenodo.org/records/18714865."

Detailed comments of Anonymous; Referee #1:

We'd like to thank the referee for the very detailed and well-structured comments on our manuscript. We appreciated your insights, which helped us to improve the manuscript by incorporating them. Please note the following colour coding: **[Orange]** indicates the referee's comment, **[Blue]** our response and action taken, **[Black]** the citations from the preprint, and **[Violet]** indicates the text as it now appears in the revised manuscript.

No.	Referee comments	Authors responses
1	Introduction: No introduction on temporal resolution is given, and how it influences the model assumptions, constrains the model parameters, and influences equifinality. WaTEM/SEDEM simulates the central tendency not the temporal variability. Finer timescales can mean more variability through time compared to through space (i.e. in the long-term annual average), which has obvious implications for the (required) parameter sensitivity. So using it for a case in which it is tested on the annual dynamics of sediment load should be clarified, and also a mention of the assumed impact on model equifinality compared to the long-term simulation.	We agree with the referee's assessment that WaTEM/SEDEM is designed to simulate central tendencies rather than an annual sediment yield. However, explicitly testing the model against these annual dynamics was a deliberate choice to evaluate where and why the model fails when pushed beyond its intended long-term design. We believe that the intent to test the model outside its standard use case is sufficiently established in our third objective, which explicitly aims to "test WaTEM/SEDEM under different levels of temporal (annual vs. eight-year means) and spatial aggregations (individual vs. grouped watersheds)." (Introduction; Revised manuscript; L99-101). Rather than elaborating on these constraints in the Introduction, we address the referee's specific points regarding model assumptions, constraints, and equifinality in the following sections: **Temporal variability and central tendency:** **Section 4.2; Revised manuscript; L512-535:** "The model's annual time step fails to capture these critical temporal coincidences, a structural limitation that becomes more pronounced when such events are infrequent. This temporal limitation aligns with findings by Risse et al. (1993), who demonstrated that USLE's model efficiency diminishes at the annual scale. When averaging over the eight-year study period, these extreme events are smoothed out, which explains the model's improved performance at longer timescales (Tab. 2).

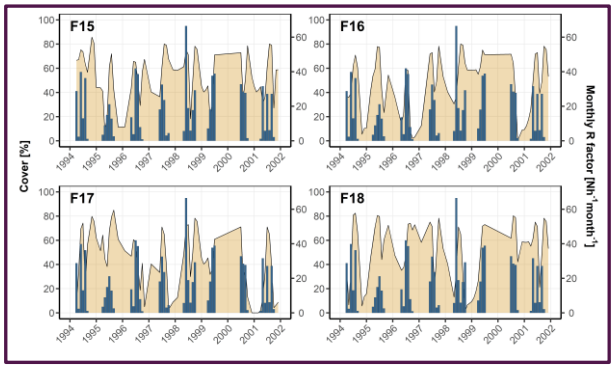
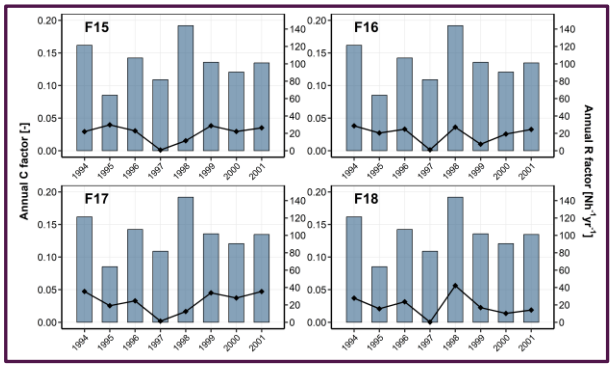
		Parameter constraints and equifinality: **Section 4.2; Revised manuscript; L523-525:** "For the temporally aggregated eight-year means, there was no single parameter set that produced behavioural model realisations across all field-dominated watersheds simultaneously when applying our limits of acceptability criterion. This indicates a limitation in parameter transferability within our study context." **Uncertainty regarding the magnitude of single events:** **Section 4.2; Revised manuscript; L530-535:** "At such fine scales, WaTEM/SEDEM may struggle to accurately represent the complex interactions between soil conservation practices and erosion processes. However, it is important to note that these difficulties may stem from the low-magnitude nature of the erosion events observed in our study. As demonstrated in paired-plot experiments, natural variability is higher for small events (Wendt et al., 1986). Therefore, areas with higher erosion rates typically yield data that is less noisy and inherently easier for models to reproduce (Nearing, 1998)." **Implication for the model use:** **Conclusion; Revised manuscript; L648-659:** "Aggregating model outputs in time and space worked best for field-dominated systems, which compensated for the underestimation of soil conservation in controlling soil erosion and the model's inability to capture extreme events within an annual time step. While WaTEM/SEDEM is generally better suited for long-term erosion modelling, our findings confirm that this is especially the case for watersheds with optimised soil conservation and reduced sediment transport." Further: **Conclusion; Revised manuscript; L668-670:** "For long-term, large-scale soil conservation planning in which the effects of single erosive events on individual fields are less relevant for representing the system behaviour, WaTEM/SEDEM seems to be fit for purpose within our testing conditions." We hope that this structure sufficiently addresses the referee's point.

2	L54-57: Can you add evidence regarding the most used model? I suggest adding a citation.	**Preprint, L54-57:** "The most widely used model for soil conservation planning is the Universal Soil Loss Equation (USLE) (Wischmeier and Smith, 1978) and its revisions and regional adaptations, like the revised USLE (RUSLE) (Renard, 1997) and the German ABAG (Allgemeine Bodenabtragsgleichung, German for Universal Soil Loss Equation; Din-Normenausschuss, 2022; Schwertmann et al., 1987)." Thank you for spotting the missing reference. We rewrote this section and added sources to our statement: **Introduction; Revised manuscript; L56-61:** "As indicated by erosion modelling reviews (Batista et al., 2019; Borrelli et al., 2021), the most widely used erosion model is still the empirical Universal Soil Loss Equation (USLE; Wischmeier and Smith, 1978) and its revisions and regional adaptations, such as the revised USLE (RUSLE; Renard, 1997) and the German ABAG (Allgemeine Bodenabtragsgleichung, German for Universal Soil Loss Equation; Din-Normenausschuss, 2022; Schwertmann et al., 1987)."
3	L66-68: I suggest being more specific and mentioning what conservation measures haven't been evaluated. It should be stated what differences there are compared to grass and non arable elements which are commonly represented in the model. Or is it that they are not typically evaluated with real data in studies?	**Preprint L66-68:** "However, to the best of our knowledge, the suitability of WaTEM/SEDEM for representing soil erosion, transport, and deposition processes within soil conservation settings combined with measures to reduce sediment connectivity, which can minimise sediment redistribution, has not been thoroughly tested." You are correct that the original phrasing was too general. We rewrote this section to be more specific and to clarify the differences between our approach and other approaches in testing conservation settings: **Introduction; Revised manuscript; L70-73:** "However, to the best of our knowledge, the suitability of WaTEM/SEDEM for representing soil erosion, transport, and deposition processes within soil conservation settings (e.g. no-till, cover crop, optimised crop rotations) combined with sediment transport control structures to reduce sediment connectivity (e.g. grassed waterways, retention ponds), has not been thoroughly tested against measured data."

4	L73-75: This lumps measurements at vastly different spatial scales (e.g. erosion plots to large watersheds) into the same context, despite having considerably different scale-related implications when considering sediment delivery.	**Preprint, L73-75:** "Such outlet-based measurements do not allow for testing a model's representation of internal erosion and deposition patterns, as they provide little information on the spatial distribution of sediment sources and sinks within the landscape." It is true that lumping erosion plots and large watersheds is an oversimplification, as the sediment delivery processes and their implications are highly scale dependent. Our intended point was not to equate these different scales, but to state that all single-point outlet measurements, by their nature, integrate the effects of both in-field soil conservation and sediment transport control structures into one lumped measurement. This makes it difficult to disentangle their individual contributions to (dis)connecting the sediment cascade, which is the core of the equifinality problem. We revised this paragraph to further clarify this point: **Introduction; Revised Manuscript; L74-81:** "One difficulty is that most plot-scale data do not account for landscape features, as they are typically not included in plots. Conversely, watershed outlet data may integrate the effects of both in-field soil conservation and sediment transport control structures into one lumped measurement, which makes it difficult to disentangle their individual contributions to (dis)connecting the sediment cascade. Long-term monitoring data from micro-scale watersheds (1-10 ha) offer the opportunity to evaluate in-field soil conservation practices separately from sediment transport control structures implemented at field to the landscape scale (Choudhury et al., 2022; Fiener and Auerswald, 2018). However, such datasets are rare (Fiener et al., 2019a), and erosion and sediment delivery models have hardly been tested under these conditions."

5	L69-77: This paragraph would benefit from a consideration of the practical considerations in the modelling process, since the implications depend on the objective of the modeller. Many modelling efforts seek acceptable sediment yield predictions, and use models with this predictive target but producing intermediate spatially distributed estimates. Others do require accurate spatial estimations with an acceptable level of uncertainty at their representative spatial and temporal scale.	**Preprint, L69-77:** "Testing the ability of spatially distributed erosion models to simulate the combined effects of in-field soil conservation and landscape features trapping sediments is inherently challenging. Observational data for model calibration and validation are typically restricted to measurements of sediment yields at the outlet of a system (Batista et al., 2019), which typically consist of small erosion plots, meso-scale watersheds, or large-scale catchments. Such outlet-based measurements do not allow for testing a model's representation of internal erosion and deposition patterns, as they provide little information on the spatial distribution of sediment sources and sinks within the landscape. This exacerbates the equifinality problem (Beven, 2006), and models may achieve accurate outputs while incorrectly representing the spatial patterns of erosion and deposition processes within watersheds." Thank you for your contribution to the discussion. While we acknowledge that some modelling efforts are conducted with a focus on outlet predictions, we argue that this approach is scientifically problematic. Our paper's central assumption is that a model achieving an "acceptable" outlet prediction for the wrong internal reasons (the equifinality problem) is not a reliable tool for any purpose. It provides a false sense of understanding and predictive power, which can lead to poor or even counter-productive management decisions, even for applications that seem "outlet-focused".
6	Methods: Currently I don't see a justification for the selection of the priors given the driving processes of erosion. Why are error distributions considered uniform for all parameter distributions at all considered time scales? Is it not the case that driving events may be driven by low probability rainfall events or high intensity bursts? I suggest including justifications which match the nature of the driving processes, particularly in the case of changing temporal scales. In such a well-measured watershed, is it not possible to constrain the uncertainty components? It is later discussed	As Beven (2009) recommended, we used uniform distributions in our parameter sampling (for the field-dominated watersheds) within the Monte Carlo framework because we have no evidence to justify other distributions. While the driving processes (e.g. rainfall) may not be uniform distributed, as you mentioned, the distribution of the sampled model's parameters (such as $k_{TC,g}$, e_{sur}, p_{con}) remains unknown. But we agree that generating synthetic data could yield further insights, we added a sentence stating this:

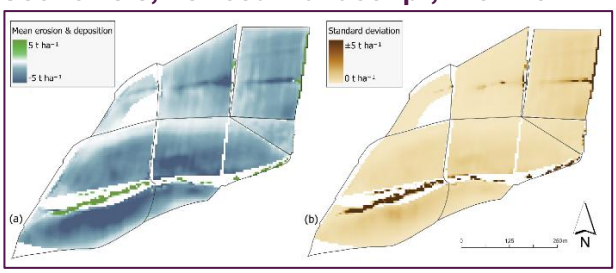
	that short windows of coincidence between bare soil and heavy rainfall can be critical, which would manifest as high uncertainty in the C-factor and R-factor. This is arguably the advantage of generating synthetic data.	**Section 4.2; Revised manuscript; L516-522:** "This deficit could not be compensated using high-resolution input data (daily soil cover, high resolution and on-site rainfall measurements). However, it is important to note that the episodic nature of erosion is always difficult to capture even with event-based models, and hence aggregation over time tends to improve any kind of erosion model. This is especially true if events are rare and the overall erosion values are small (Nearing, 1998). Future studies could employ synthetic data generation (Srikanthan and Mcmahon), allowing for an assessment of the model's sensitivity beyond our high-resolution input data."
7	As mentioned above, lumping everything into an error parameter on gross erosion poorly represents the individual contributions of sub-parameters, their interactions, and identifiability. At present, I miss a justification for this. What about the contribution (combinations of) sub-factors and their contribution to erosion and sediment transport realisations?	Please see our earlier response during the open discussion phase: Thank you for this critical methodological point. You are correct that we opted to lump the uncertainty from the ABAG factors into a single error surface rather than sampling each factor in the Monte Carlo simulation. This was a methodological choice driven by two primary considerations: (i) our study's specific objectives and data quality, (ii) computational costs: (i) Our primary objective focused on testing WaTEM/SEDEM's transport component within micro watersheds under soil conservation, and not the ABAG, which was developed for Southern Germany and where it has been extensively evaluated. Moreover, our study is based on an exceptionally high-quality monitoring dataset from the Scheyern experimental farm, which provides site-specific, highly detailed field measurements of rainfall, soil, and crop and management data that are used as input for calculating the ABAG factors (which were experimentally calibrated precisely for our study region - Schwertmann et al. (1987)). Hence, we assume that parameterisation errors are negligible, but as the ABAG is essentially built upon regressions that will always carry residual errors, we pragmatically used an error surface (esur) to account for this.

		(ii) Representing the uncertainty in each ABAG factor separately is not as straightforward as sampling different values in a Monte Carlo simulation. This is because these factors are not parameters being calibrated or conditioned against observations but are more like variables that are estimated from input data. Uncertainty estimation in this case requires complex approaches for representing uncertainty in the input data used for calculating input factors and propagating them through the model (as we have done in previous work - Batista et al. (2021); Batista et al. (2022)). We did not feel like this was justified because of the quality of our input data, as we explain in point (i) above.
8	Regarding the general model implementation, a German USLE formulation is used in place of the typical RUSLE formulation. Can the authors show time series data of the annual parameter inputs used? It would be helpful for the reader to know the distributions of the input values. I would also suggest a discussion on what impact this may have on the model, plus the consequences for comparing parameter values (e.g. ktc) with other studies given the parameter compensation effects.	Thank you for this suggestion. The distribution of the input data is already shown in a monthly timestep in the manuscript in Figure 2, which we refined to exclude doubled information on the y-axis (**L225, Revised manuscript**):  Additionally, we created an annual figure (Fig. 3) for the computed C factors for all fields and R factors for the entire timeframe (**L243-244, Revised manuscript**):  *Figure 1: Annual C factor (black dots and lines) and R factor (blue bars) for the individual fields F15 to F18 over eight years.*

		All other input data were either directly measured (K factor, DEM) or derived from input data (kTC map, L and S factors). Regarding the use of the German USLE (ABAG) instead of the standard RUSLE, we agree with the referee that the specific USLE formulation influences the magnitude of the calculated potential erosion. However, we argue that the ABAG is the most appropriate model for this study, as it is specifically adapted to Central European soils and climatic conditions. Furthermore, it represents the standard methodology used by regional governmental institutions (e.g., the Bavarian State Office for Agriculture / LfL). While we believe a detailed comparative discussion between ABAG and RUSLE is outside the scope of this study, we have clarified the reasoning for this choice and the specific differences in the **Methods section (Section 2.4)** to address the referee's concerns: **Section 2.4; Revised manuscript; L184-186:** "This approach is specifically adapted to the soils and climatic conditions of Central Europe and represents the standard methodology for the study region (Schwertmann et al., 1987; Din-Normenausschuss, 2022)." **Section 2.4; Revised manuscript; L193-195:** "Following the German adaptation of the USLE, rainfall events were considered erosive if they met at least one of two criteria: (i) total rainfall amount ≥ 10 mm (in contrast to the 12.7 mm threshold of the standard USLE, Wischmeier and Smith, 1978) or (ii) maximum 30-minute intensity ≥ 10 mm h-1."
9	What about stream initiation and transition to channelised flow? In WS, there are various ways to consider the stream channel initiation by digitizing channels or considering a flow accumulation threshold. I would recommend mentioning this.	Thank you for pointing towards to the models' internal features. The stream initiation by flow accumulation threshold is implemented in the translated Python version of the model. Despite that, we decided not to use this feature as we didn't expect a stream inside our micro watersheds. In the revised manuscript, we added a sentence mentioning the not used, but implemented stream initiation feature:

		Preprint, L164-165: "Although the Python implementation includes tillage erosion calculations, this component was not utilised in the present study." **Section 2.3; Revised manuscript; L173-175:** "Although the Python implementation includes tillage erosion and stream initiation calculations, these components were not utilised in the present study."
10	L197: The word seasonal is ambiguous in this case. I would also suggest using a mathematical formulation for the C-factor, showing how the SLR is generated and combined with rainfall erosivity and at what time scale. It's also of general interest to the reader to know what these SLR and C-factor values are for both arable and grassland, and how they change through time. The literature reference for the SLR formulation is also grey literature, so more details are justified.	**Preprint, L197-198:** "The annual crop factor (C factor) was calculated by combining seasonal rainfall erosivity with temporal changes in soil coverage (Schwertmann et al., 1987)." We agree that this section needs refinement in the revised manuscript, specifically by elaborating on the calculation of the C factor and by including the C factor equation. We completely revised the whole section. This is shown in the **detailed comment section of Joris Eekhout, Referee #2, no. 14.**
11	L297-304: How does the calculation of likelihoods vary between temporal aggregations? For individual years is this done by comparing the time series or individual simulations?	**Preprint, L297-304:** "For these behavioural simulations, we calculated likelihoods by rescaling the mean absolute error (*MAE*) (Brazier et al., 2000): $L_i = \frac{1}{MAE_i} / \sum \frac{1}{MAE_i}$, (6) *with:* $MAE_i = \lvert Sim_i - Obs_i \rvert$, (7) where L_i is the likelihood of one realisation i (dimensionless), MAE_i is the mean absolute error of realisation i (t ha^{-1} yr^{-1}), Sim_i is the simulated values for behavioural runs of realisation i (t ha^{-1} yr^{-1}), and Obs_i is the observed sediment value for realisation i (t ha^{-1} yr^{-1})." The likelihood calculation (Preprint, Eq. 6 and 7) was only applied to the model realisations that passed our limits of acceptability criterion based on the error margins of the measured data (Only the eight-year aggregated values). We clarify this further in the modified section:

		Section 2.7; Revised manuscript; L330-333: "Model realisations were classified as behavioural if the simulated sediment yield values fell within the established limits of acceptability (error margins) for the observed data. Likelihoods were calculated only for the spatiotemporal aggregated data for which behavioural model realisations were identified. For these behavioural simulations, we calculated likelihoods by rescaling the mean absolute error (MAE) (Brazier et al., 2000)"
12	L316-324: Can the authors justify the use of the median? This would assume an underestimation of the total sum due to the positively skewed nature of sediment yield, which would have obvious implications for watershed management. Typical applications of WaTEM/SEDEM are applied to the mean.	**Preprint, L316-321:** "Model outputs were analysed at multiple spatiotemporal scales through sequential aggregation steps: First, we calculated an eight-year median of the sediment yield for each individual watershed. Second, we spatially aggregated the watersheds based on their dominant erosion characteristics (field- and structure-dominated) while maintaining an annual resolution. Third, we aggregated the median values over the eight-year monitoring period for these spatially aggregated groups. To further analyse relative errors, the percent bias (PBIAS) was calculated by: $PBIAS = \left(\frac{Sim_i - Obs_i}{Obs_i}\right) * 100,$ (8) Where Sim_i is the simulated values for behavioural realisation i (t ha^{-1} yr^{-1}), and Obs_i is the observed sediment value for realisation i (t ha^{-1} yr^{-1})." Thank you for your comment regarding the use of the median instead the mean. We acknowledge that the mean is often used in USLE-based applications, particularly for assessing sediment deliveries and watershed management. The USLE was designed to estimate long-term average annual soil loss. Therefore, computing the mean aligns with the intrinsic logic of the model. Additionally, this allows for better comparability with other studies, where the mean is typically used for assessing sediment delivery and watershed management. We decided to change our statistics using the mean consistently, instead of the median. Please note that we changed the numbers in the manuscript, which is detailed in the **detailed comment section of Joris Eekhout, Referee #2, no. 21.**

13	Results: The results are in general concisely presented. However, the spatial analysis section is overly brief. Do the multiple model runs which were made to address equifinality not give significantly more spatial information in addition to the median? What is the spatial variability of the behavioural predictions? Only the median is currently given but arguably one of the advantages of multiple realisations in WS is that you can get some idea of the variability in the spatial patterns from acceptable simulations. This is also useful to know the added spatial information which can be achieved for land management.	We do agree that we should elaborate more on the spatial distribution of the model output. We added two maps showing the relative uncertainty and the 95% prediction width of the behavioural sediment yield: **Section 3.3; Revised manuscript; L454-457:**  "Figure 8: (a) The mean of simulated potential erosion and deposition of behavioural model realisations over the eight-year period. Negative values indicate erosion and positive values deposition. (b) The cell-wise standard deviation of behavioural model realisations over the eight-year period." Further, we added multiple extensions to the manuscript, describing and discussing the calculation and spatial distribution in the Methods, Results, Discussion and Conclusion section: **New Section 2.9; Revised manuscript; L361-364:** "2.9 Spatial analysis Spatial analysis was performed using R-Studio (R 4.4.2; R-Studio 2024.12.1 Build 563). To quantify the spatial distribution of sediment yield and the associated uncertainty, the cell-wise mean and standard deviation were calculated across all behavioural model realisations." **Section 3.3; Revised manuscript; L443-453:** "3.3 Spatial analysis In field-dominated watersheds, substantial deposition was primarily confined to retention ponds, while other areas outside arable lands showed minimal deposition, except for W03 (Fig. 8a). In W04, negligible to no deposition was observed. Conversely, structure-dominated watersheds exhibited considerably more intense erosion-deposition dynamics. The grassed waterway showed a clear deposition pattern, with W06 exhibiting the most pronounced deposition patterns leading toward the retention pond at the outlet.

		The map of the standard deviation (Fig. 8b) also displays a spatial shift in model uncertainty between the two watershed groups. In field-dominated watersheds, standard deviation exhibits concentrations along in-field flow pathways, the retention ponds and the small structures next to their outlets. In structure-dominated watersheds a substantial increase in standard deviation is visible along the flow pathways towards and along the grassed waterway." **New Section 4.3; Revised manuscript; L566-577:** "4.3 Spatial dynamics of erosion and deposition The main reason for the reduction of sediment yield at the outlet within structure-dominated watersheds is visible in the eight-year mean map of behavioural model realisations (Fig. 8a). Depositional patterns are prominent along the flux pathway of the grassed waterway, particularly in W06. Because WaTEM/SEDEM is sensitive to the parameter controlling deposition inside these soil conservation structures ($k_{TC/G}$), the standard deviation in these areas is exceptionally high (Fig. 8b). However, high uncertainty is not limited to the grassed waterways. It is also evident in depositional areas of field-dominated watersheds (most pronounced in W03). This suggests that deposition is generally prone to high uncertainties, regardless of the dominant structures. Furthermore, field-dominated watersheds exhibit high standard deviations within the fields themselves. This likely stems from parameter uncertainty regarding sediment supply, specifically the error surface parameter (e_{sur}, see Fig. 6b), which directly influences the sediment supply generated within the arable land."

		Conclusion; Revised manuscript; L653-657: "The GLUE framework revealed specific patterns in the sampled parameter space, particularly the compensation mechanism between $k_{TC/A}$ and e_{sur} values for field-dominated watersheds, and the narrow behavioural parameter range of $k_{TC/G}$ values (7.5-15 m) for structure-dominated watersheds. Spatially, this sensitivity is mirrored by high standard deviations concentrated along the grassed waterways. In contrast, the standard deviation within arable fields shows the influence of sediment supply parameterisation (e_{sur}) in field-dominated systems (Fig.8b)."
14	Discussion: A typical explanation for the lack of global ktc parameters is the existence of unconsidered processes. Is this the case, or is it more of an inadequacy to capture the system behaviour? Can the field data give insights on this?	Thank you for these interesting questions. Our study demonstrates that watersheds with nearly equal conditions regarding soils, management, and crops still required different behavioural parameter configurations. This highlights a challenge in parameter transferability and suggests that the model, in its current form, has an inadequacy to capture the large spatiotemporal variability in the behaviour of each individual micro-watershed. For the structure-dominated watersheds, the analysis showed a strong concentration of high-likelihood values in a very narrow and low range of $k_{TC/g}$. We interpret this as the $k_{TC/g}$ parameter compensating for re-infiltration, along with unidentified processes. This influence of the grassed waterway is supported by Fiener and Auerswald (2003), who reported that between 1994 and 2000, the system reduced runoff by 90% in the upper sub-watershed (W06) and 10% in the lower sub-watershed (W05), resulting in sediment trapping efficiencies of 97% and 77%, respectively. WaTEM/SEDEM's model structure does not account for the loss of runoff in the grassed waterway, which our field observations suggest is a critical mechanism for trapping sediment in Scheyern. Because the model cannot capture this mechanism, it inherently overestimates sediment transport, especially in W05 and W06. To compensate, the optimisation forces $k_{TC/G}$ to a low value, enhancing deposition to fit the observations.

15	L459-467: This is somewhat difficult to follow. Is including this uncertainty in the input parameters not the purpose of using GLUE?	**Preprint, L459-467:** "The model simulated the very low sediment yields resulting from well-established in-field soil conservation practices in field-dominated watersheds, comparable to the measured data. In general, observed sediment yields were overestimated, which can be attributed primarily to difficulties in accurately representing the specific C factors of this conservation system, particularly unique practices such as mustard sown onto autumn-built dams where potatoes were later directly planted (Fiener and Auerswald, 2003). Such unconventional approaches are not adequately captured in the SLR values for no-till systems as evaluated in the German adaptation of the USLE (ABAG; Schwertmann et al., 1987; DIN-Normenausschuss, 2022), even with the use of very low soil loss ratios in the parameterisation of the C factor, which represent the continuous soil cover through the crop rotation in the experimental farm (Fig. 2)." This part of the discussion is our interpretation of the GLUE results. We believe that the analysis was successful, as it demonstrated that high-likelihood behavioural runs necessitated a negative e_{sur} (preprint, Fig. 5b), indicating a systematic overestimation bias in the potential erosion input. The text explains the source of this bias. We argue it is not a simple parameter uncertainty in the C factor, more a fundamental problem of the ABAG as erosive events are driven by single events.

16	"Conservation landscapes" is combining multiple physical characteristics of the agricultural watersheds together, which have differing roles in soil erosion and sediment delivery through on-site and off-site effects. Is it due to grassed areas or conservation tillage? Indeed, grassed areas are commonly applied in the model through land use elements and grass buffer strips. So one could argue that they are indeed commonly applied in the model, but the impacts of conservation tillage on erosion and overland flow generation less so. It would help to separate conservation landscapes into their specific elements.	We added a sentence to differentiate between in-field soil conservation (encompassing tillage and cover) and structural conservation practises when discussing model limitations, rather than using the broad term "conservation landscapes". **Section 4.2; Revised manuscript; L525-532:** "While Van Rompaey et al. (2001) recognized technical limitations of WaTEM/ SEDEM in model transferability related to grid size and routing methods, our findings suggest additional challenges in accurately representing processes within micro-scale watersheds with specific in-field soil conservation practices (e.g. no-till farming). The need for watershed-specific calibration, even within relatively homogeneous landscapes with similar crop and soil properties, indicates that parameter calibration compensates for inherent model or data limitations. At such fine scales, WaTEM/SEDEM may struggle to accurately represent the complex interactions between soil conservation practices and erosion processes."
17	Can the authors elaborate on the effect of using the USLE formulation for Germany versus the typical RUSLE formulation used in WaTEM/SEDEM. Indeed I expect the model to be better calibrated for Germany agri-environmental conditions on which it was developed, however there are differences compared to the RUSLE formulation which are worthwhile to mention.	This is an important point which relates to your earlier **comment no. 8**. You are correct that using the German ABAG formulation instead of the internationally used RUSLE is a key methodological choice. As we explained in our previous response, we believe a detailed comparative discussion between ABAG and RUSLE is outside the scope of this study.
18	Conclusion: L584-585: I didn't see this point addressed in the manuscript. Is it not the case that the gross erosion estimates from the USLE factors overestimate the rates based on the most likely error surface values?	**Preprint, L584-585:** "Ultimately, our study demonstrates that WaTEM/SEDEM can simulate the very low sediment yields observed from soil conservation agricultural systems, provided that high spatiotemporal resolution input data and locally adapted USLE factors (e.g., the ABAG for Southern Germany) are available."

		You are correct, and this is a central finding of our uncertainty analysis. Our results for the e_{sur} parameter, which showed higher likelihoods near -0.5, demonstrate that the standard ABAG factors do overestimate erosion in case of optimised soil conservation at an annual time step. We included the e_{sur} parameter precisely to account for known biases in USLE-type models (Nearing, 1998; Risse et al., 1993; Kinnell, 2007). The key finding is that WaTEM/SEDEM can generally simulate the correct magnitude of the very low sediment yields at this specific setting as shown by the low MAE values for the entire set of model realisations (0.14 t ha^{-1} yr^{-1} for field-dominated and 0.29 t ha^{-1} yr^{-1} for structure-dominated watersheds).

Detailed comments of Joris Eekhout; Referee #2:

Thank you for the well-structured and helpful comments on our preprint. We appreciate your insights, which helped us to improve the manuscript by incorporating them. Below, we address each of your points and clarify how we intend to incorporate them into the manuscript. Please note the following colour coding: **[Orange]** indicates the referee's comment, **[Blue]** our response and action taken, **[Black]** the citations from the preprint, and **[Violet]** indicates the text as it now appears in the revised manuscript.

No.	Referee comments	Authors responses
1	(1) Lines 43-44: I was expecting the two strategies already in this sentence. I suggest to add after the colon "(i) in-field control measures and (ii) off-site sediment transport control structures". It would be even better to make a clearer distinction, such as on-site and off-site measures. (2) Lines 43-46: In-field measures also frequently have the aim to increase infiltration and reduce runoff generation or to increase surface roughness to reduce flow velocities. This definition of in-field measures can be made a bit broader. (3) Line 47: Replace the first "and" with a comma.	**Preprint, L43-47:** "Effective soil conservation relies on two complementary strategies: (i) In-field control measures that increase soil surface cover by vegetation and hence prevent soil detachment by raindrop impact and sheet flow. Such measures include optimised crop rotations, using cover crops, and soil residue management (Andersson and D'souza, 2014). (ii) Off-site sediment transport control structures along the runoff pathway that increase infiltration and foster sediment trapping and minimise sediment connectivity." We agree with you in every sense, as this will enhance the quality of the manuscript. In the revised manuscript, we have rewritten this paragraph to: 1. List the two strategies (on-site and off-site) at the beginning of the section for better clarity. 2. Broaden the definition of in-field measures including surface roughness and infiltration to reduce runoff generation. 3. Correct the punctuation by replacing the first "and" with a comma." **Introduction; Revised manuscript; L44-49:** "Overall, effective soil conservation relies on two complementary strategies: (i) in-field soil conservation and (ii) sediment transport control structures along the flow pathways. In-field practices focus on increasing soil surface cover by vegetation to prevent soil detachment by raindrop impact and overland flow. Such practices include optimised crop rotations, using cover crops, and soil residue management (Andersson and D'souza, 2014). Sediment transport control practices consist of structures installed along the runoff pathway to increase infiltration, sediment trapping, and hence minimise sediment connectivity."

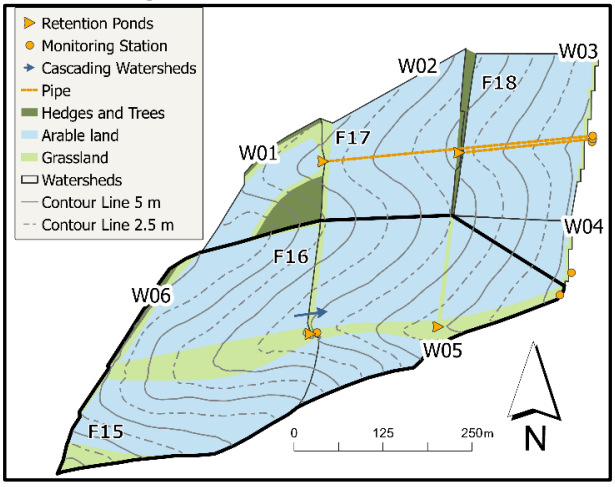
2	Lines 50-51: Field demonstration would be highly feasible at the scale the authors are working and likely more convincing for stakeholders. Models are indeed valuable tools, but likely more for scenario evaluation, for instance, for different configurations of on-site and off-site measures.	**Preprint, L50-51:** "Soil erosion models are potentially valuable tools for identifying high erosion risk areas and evaluating intervention needs, enabling stakeholders to effectively implement soil conservation strategies." We agree that at the micro-scale, field demonstrations are feasible and convincing. We acknowledge that an additional advantage of modelling at this scale lies in scenario evaluation, specifically the ability to test and compare different configurations of conservation measures before implementation. We have refined the sentence to emphasise this aspect, as suggested: **Introduction; Revised manuscript; L52-54:** "Soil erosion models are potentially valuable tools for identifying erosion-prone areas and developing what-if scenarios, allowing stakeholders to assess different configurations of on- and off-site soil conservation practices. This enables the identification of optimal intervention strategies before implementation."
3	Lines 72-73: What is the difference between meso-scale watersheds and large-scale catchments? Please clarify in the text.	**Preprint, L72-73:** "Observational data for model calibration and validation are typically restricted to measurements of sediment yields at the outlet of a system (Batista et al., 2019), which typically consist of small erosion plots, meso-scale watersheds, or large-scale catchments." Thank you for asking for this clarification. In revising this section (also in response to **detailed comment section Anonymous, Referee #1, no. 4**), we agreed that the term meso-scale introduced unnecessary ambiguity without adding value to the manuscript. We removed the term meso-scale and focused on the common limitation of outlet-based measurements across all scales.

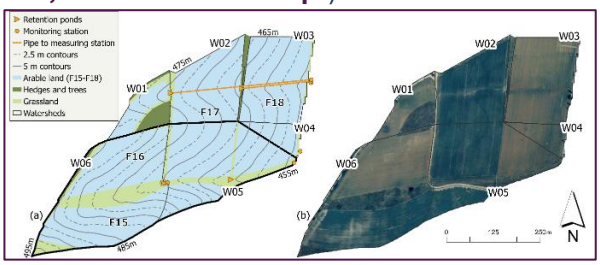
		Introduction; Revised Manuscript; L74-81: "One difficulty is that most plot-scale data do not account for landscape features, as they are typically not included in plots. Conversely, watershed outlet data may integrate the effects of both in-field soil conservation and sediment transport control structures into one lumped measurement, which makes it difficult to disentangle their individual contributions to (dis)connecting the sediment cascade. Long-term monitoring data from micro-scale watersheds (1-10 ha) offer the opportunity to evaluate in-field soil conservation practices separately from sediment transport control structures implemented at field to the landscape scale (Choudhury et al., 2022; Fiener and Auerswald, 2018). However, such datasets are rare (Fiener et al., 2019a), and erosion and sediment delivery models have hardly been tested under these conditions."
4	(1) Line 80: Replace "at the" with "in a". (2) Line 81: Replace "in large-catchment sediment yield observations" with "at larger scales".	**Preprint, L79-81:** "This is because soil erosion and sediment connectivity processes that are distinguishable at the micro-scale watershed are not represented in small plots or get diluted in large-catchment sediment yield observations. " Thank you for finding this typo and the suggestion. The cited lines are not in the revised manuscript anymore.
5	Line 87: Replace the first "and" with a comma.	**Preprint L86-87:** "Notwithstanding the spatial extent of (long-term) soil erosion monitoring, measurement uncertainties arise from instrumental precision and temporal instrument malfunctioning, data handling and processing. " Thank you for finding this typo. We revised this section in our manuscript: **Introduction; Revised manuscript; L82-84:** "Regardless of the spatial scale in which erosion is monitored, it is important to note that perfect observational data do not exist. All measurements include errors stemming from instrumental precision, temporary malfunctioning, and data handling and processing."

6	Lines 94-96: These two sentences are a bit difficult to understand. For instance, what do the authors mean with "These behavioural models", behavioural models were not mentioned in the previous sentence.	**Preprint, L94-96:** "Within the GLUE framework, limits of acceptability are defined to identify which model runs fall within the uncertainty bounds of the measurements (Beven and Lane, 2022). These behavioural models are retained, while non-behavioural models are rejected." Thank you for your comment. We have revised this section completely: **Introduction; Revised manuscript; L87-92:** "GLUE acknowledges that it is not possible to identify a single calibrated parameter set as "correct". Rather, all parameter combinations that produce results within given limits-of-acceptability cannot be rejected (Beven and Lane, 2022). Contrarily, if not a single model realisation encompasses the uncertainty bounds of the observational data, non-behavioural models or model structures can be rejected, which might lead to improvements in terms of understanding and modelling."
7	(1) Lines 101-102: The second objective refers to a sensitivity analysis? Or is this related to the application of the GLUE method? Please clarify in the text. (2) Lines 102-103: The third objective seems unrelated to the information provided in the Introduction or is this also related to the GLUE method (seems unlikely)? If not, please provide a short introduction on how differences in spatiotemporal resolutions impact soil erosion model outcomes.	**Preprint L101-102:** "(ii) analysing the behaviour of model parameters that control erosion and sediment transport processes, and (iii) assessing the model's performance across different spatiotemporal resolutions through data aggregation. " Thank you for bringing this up. These comments relate directly to the general comment part (**General comments section, Joris Eekhout; Referee #2; no. 1**). In that section, we have already changed the objectives in the manuscript.
8	Line 105: Replace "from" with "for".	**Preprint, L103-106:** "We accomplish these objectives using a comprehensive dataset from a long-term, farm-scale monitoring in Southern Germany, which provides continuous precipitation, surface runoff and sediment flux data from six micro-scale watersheds under optimised soil conservation (Auerswald et al., 2001; Auerswald and Fiener, 2019). " Thank you for finding this typo. We changed it in the manuscript accordingly:

		Introduction; Revised manuscript; L101-105: "We accomplish these objectives using a comprehensive, long-term monitoring dataset from Southern Germany, which provides high resolution model inputs (e.g. precipitation, crop-specific daily soil cover, etc.), as well as continuous surface runoff and sediment flux data for six micro-scale watersheds under optimised soil conservation and reduced sediment transport (Auerswald et al., 2001; Auerswald and Fiener, 2019; Fiener et al., 2019a)."
9	Lines 121-125: The main difference between the two different systems seems to be that W05 and W06 include grass strips, while the other study areas don't. The other study areas also include retention ponds, which I consider to be a structural conservation measure. Please clarify in the text.	**Preprint, L120-125:** "The six watersheds exhibit different landscape connectivity characteristics: W01-W04 (0.8 to 4.2 ha) are classified in this study as field-dominated systems due to their structure, with most of their area covered by agricultural fields and minimal landscape structures along sediment flux pathways. In contrast, W05 and W06 are classified as structure-dominated systems due to their configuration, featuring more complex landscapes. The Watershed W06 (5.7 ha) constitutes the upper part of the larger watershed W05 (7.8 ha) (Fiener et al., 2019b). " You are right that this section might be a bit misleading, as the group field-dominated also includes watersheds with retention ponds (W01 and W02). We rewrote this section to make the distinction clearer: **Section 2.1; Revised manuscript; L120-125:** "The six watersheds exhibit different landscape connectivity characteristics: W01–W04 (0.8 to 4.2 ha) are classified as field-dominated systems. Despite the presence of retention ponds at the outlets of W01 and W02, these watersheds have no internal linear landscape structures, resulting in sediment flux pathways governed primarily by topography and the management of the arable fields. In contrast, W05 and W06 are classified as structure-dominated systems due to the presence of a grassed waterway along the thalweg. Watershed W06 (5.7 ha) constitutes the upper part of the larger watershed W05 (7.8 ha) (Fiener et al., 2019a)."

10	Line 127: Looking at Figure 1, it does not seem that the fields are arranged parallel to the contour lines (assuming that the curved lines indicate the contour lines). Please clarify in the text.	**Preprint, L126-128:** "Three key conservation measures were implemented to minimise hydrological and sedimentological connectivity: (i) optimised field layout with fields arranged parallel to contour lines, (ii) retention ponds at field borders, and (iii) a grassed waterway along the main thalweg of W05 and W06." **Referred figure, L138:** Please see **next comment no. 11, Figure 1**. Thank you for pointing out this inconsistency. Indeed, looking closely at the field arrangement in Figure 1 not all fields were arranged parallel to contour lines. As Fiener et al., 2019b describes, the field layout was influenced by the watershed boundaries, which served as borders between the farming systems. The optimisation mentioned in the text referred to a reduced field size adapted to the steep slopes, rather than a strict contour-parallel arrangement. We apologise for this mistake. We changed that in the manuscript: **Section 2.1; Revised manuscript; L131-133:** "Three key conservation practices were implemented to minimise hydrological and sedimentological connectivity: (i) an optimised field layout with reduced field sizes adapted to the steep slopes, (ii) retention ponds at watershed outlets, and (iii) a grassed waterway along the main thalweg of W05 and W06 (Fiener et al, 2019a)."
11	Line 135: What is meant with F15-F18? These are different configurations of the crop rotation? Please clarify. Figure 1: I had to study the figure quite a bit to figure out which study area belonged to which field. I suggest to include another smaller panel where the fields are better indicated, with different colours, for instance. It would also be useful to get some more information about the contour lines, to give the reader an idea about the slopes in the study area. Moreover, the differences in crop	**Preprint, L134-135:** "All fields within the watersheds were managed using no-till practices with a crop rotation of winter wheat, maize, winter wheat, potatoes, whereas the rotation was shifted between the fields (F15-F18, Fig. 1). "

	rotation are not that clear from this figure. To which fields do the different F-codes belong?	**Referred figure, L138:**  Figure 1: Land use and topography of the experimental farm in Scheyern, Bavaria, with flow direction from west to east. Note: watershed W05 (thick line) includes the upslope watershed W06. We apologize for the confusion regarding the F-codes. F15–F18 refer to the four agricultural fields, not to specific crop rotation configurations. As described in the methods, all fields follow the same crop rotation, but the phases are staggered temporally so that different crops are present in the landscape in any single year. We rewrote this section to make this clearer: **Section 2.1; Revised manuscript; L139-142:** "All fields within the watersheds (F15 to F18 in Fig. 1) were managed using no-till practices with a crop rotation of winter wheat (*Triticum aestivum* L.), maize (*Zea mays* L.), winter wheat, and potatoes (*Solanum tuberosum* L.). This rotation was staggered across the fields, meaning that while the sequence was identical, the specific crop grown each year varied between fields." Regarding Figure 1, we agree that the fields and topography needed better definition. However, we refrained from adding distinct colours for each field, as we believe this would overcrowd the figure and obscure the watershed boundaries and land use details. Instead, to address your concerns and improve readability, we have:

		1. Repositioned the F-labels (F15-F18) to be more central and distinct. 2. Updated the legend to explicitly link arable land with F15-F18. 3. Added elevation labels to the contour lines to provide better information on the slopes. 4. Added the orthophoto taken in 2022 to clarify the land-use structures. This figure was added to the manuscript (**L140-144; Revised manuscript**):  Figure 1: (a) Schematic land use and topography of the experimental farm in Scheyern, Bavaria, with flow direction from west to east. W01-W06 are abbreviations for six watersheds; F15-F18 are the fields located in these watersheds. Note: Watershed W05 (thick line) includes the upslope watershed W06, as W06 is cascading into W05. (b) Aerial photograph of the study area shows the land use patterns and field boundaries on the Scheyern farm in 2002.
12	Line 151: With “aliquot” the authors mean “sample”?	**Preprint, L150-152:** “The device continuously diverted an aliquot of approximately 0.5 % from the total flow that left the watersheds through underground-tile outlets with a diameter of 15.6 cm and 29 cm (Fig. 1). “ Thank you for this comment. You are right, effectively it is a sample. However, we chose the term “aliquot” to emphasize that the device extracts a specific, representative fraction (approx. 0.5%) of the total flow volume, which is the mechanism of the Coshocton wheel. We believe that aliquot is the technically precise term here, but “sample” is also applicable. We would like to stick to aliquot, as this is the term used in the standard literature for Coshocton wheels.

13	Lines 148-156: How were runoff and sediment totals estimated using this system? Was this continuously monitored or estimated after each event? Were the sediment samples further analysed on grain size distribution? Please clarify in the text.	**Preprint, L148-156:** "For model testing, we used continuous sediment delivery data from the six micro-scale watersheds between 1994 and 2001. Runoff and suspended-sediment loads were monitored with a measuring system based on a Coshocton-type wheel sampler (precision ± 10%; Carter and Parsons, 1967; Fiener and Auerswald, 2003). The device continuously diverted an aliquot of approximately 0.5 % from the total flow that left the watersheds through underground-tile outlets with a diameter of 15.6 cm and 29 cm (Fig. 1). At lower rates (< 0.5 L s^{-1}) the system slightly over-estimated runoff, but these small events contributed negligibly to the cumulative water and sediment budgets. Under sampling during very high flows was avoided by (i) employing large wheels (Ø 61 cm) and (ii) the flow-dampening effect of the retention ponds situated immediately upstream of each outlet (Fiener and Auerswald, 2003). " Thank you for pointing towards the sampling strategy of the experimental farm in Scheyern. We do agree that more information should be included in the methodology. We have clarified the text regarding the runoff and sediment quantification. Runoff and sediment totals were determined based on the volume collected by the Coshocton-type wheel samplers, which continuously diverted an aliquot (sample) of the runoff into storage tanks. The volume in these tanks was measured after each event (or during very large events). Sediment totals were calculated from these runoff volumes and the sediment concentration. To determine the concentration, the tank content was mixed to homogenise the suspension before sampling, and samples were subsequently dried at 105°C. Grain size distributions of the suspended sediment samples were not analysed.

		Section 2.2; Revised manuscript; L155-162: "The device continuously diverted an aliquot of approximately 0.5 % from the total flow that left the watersheds through underground tile outlets with a diameter of 15.6 cm and 29 cm (Fig. 1) into storage tanks (1.0 – 3.5 m3). Runoff volumes were measured after each event. Sediment yield was calculated from runoff volumes and sediment concentrations derived from homogenised tank samples dried at 105°C. At lower rates (< 0.5 L s^{-1}) the system slightly over estimated runoff, but these small events contributed negligibly to the cumulative water and sediment budgets."
14	(1) 197-223: I suggest to restructure this subsection, especially the first paragraph introduces several concepts that are later on described in more detail. This might be confusing for many readers. Please add 1-2 sentences where is explained what will follow in this subsection. (2) Lines 197-202: This seems to be a bit confusing. What do the authors mean by "combining seasonal rainfall erosivity with temporal changes in soil cover"? The rainfall erosivity is used to calculate the crop factor? How is the SLR calculated and how is the SLR related to the crop factor? Please clarify in the text. (3) Lines 203-205: Change "bi-weekly measurements" to "bi-weekly crop and residue cover measurements" and remove the sentence in line 205. (4) Lines 205-207: How were the bi-weekly measurements translated to daily cover values? What is meant by standardised crop development? Please clarify.	**Preprint, L197-223:** "The annual crop factor (C factor) was calculated by combining seasonal rainfall erosivity with temporal changes in soil coverage (Schwertmann et al., 1987). The soil loss ratio (*SLR*) quantifies the protective effect of soil coverage by comparing potential soil loss under a given vegetation condition to that under standardised fallow conditions (Schwertmann et al., 1987; Wischmeier and Smith, 1978). While the *SLR* traditionally considers five crop growth stages, from bare soil (0% cover) to full canopy coverage (75-100% cover), we also considered crop residue cover. From 1994 to April 1997, direct bi-weekly measurements during growing seasons and monthly measurements during autumn and spring were conducted, with additional observations before and after soil management operations. These field measurements included both crop and residue cover. From these field measurements, standardised daily crop development and residue cover were established and used for the subsequent period from April 1997 onwards (Auerswald et al., 2019; Fiener et al., 2019b). The support practices factor (P factor) was not specifically parametrised for contour-seeding because of field heterogeneity, i.e. not all parts of a single field were contour-seeded, and/or the absence of specific P factor values for structures such as the potato dams. However, we accounted for the uncertainty stemming from this lack of parameter representation as part of the model conditioning process (see section 2.4 below). Total soil cover was calculated with residues protecting portions of the otherwise exposed soil according to:

		$Co_{tot} = Co_{crop} + (100 - Co_{crop}) * Co_{res}$, (2) With Co_{tot} is the total soil cover (%), Co_{crop} the cover of the growing crop on the respective field (%), and Co_{res} the measured soil cover of the residues (%). Figure 2 illustrates the total soil cover on the respective fields with monthly rainfall erosivity. Determining field specific *SLR* values involved categorising soil cover into the five growth stages and assigning corresponding *SLR* values. As no-till was applied at the research farm, lower *SLR* values were assigned than in conventional systems due to increased soil surface protection. These *SLR* values were obtained from Schwertmann et al. (1987) and adapted based our expert knowledge regarding the soil conservation practices in the Scheyern experimental farm (Fiener and Auerswald, 2007; Fiener et al., 2019b). The support practices factor (P factor) was not specifically parametrised for contour-seeding because of field heterogeneity, i.e. not all parts of a single field were contour-seeded, and/or the absence of specific P factor values for structures such as the potato dams. However, we accounted for the uncertainty stemming from this lack of parameter representation as part of the model conditioning process (see section 2.4 below)." We thank the reviewer for this constructive feedback regarding the structure and clarity of the C factor subsection. We agree that the original text introduced concepts too early, which created confusion. To address this, we have completely restructured the subsection to follow a more logical flow. We now start with the data acquisition (measurements), move to the data processing (deriving daily values), then describe the SLR assignment, and finally the C factor calculation. Please see the changes we made in the manuscript below: **Section 2.4; Revised manuscript; L209-244:** "To account for the temporal dynamics of soil protection, we calculated the C factor based on soil cover and rainfall erosivity. This calculation involved deriving continuous daily soil cover data from field measurements, determining the corresponding soil loss ratios (SLR), and weighting these by the seasonal rainfall erosivity.

From 1994 to April 1997, bi-weekly crop and residue cover measurements were conducted during growing seasons, with monthly measurements during autumn and spring and additional observations before and after soil management operations. To obtain continuous time series, the point data was linearly interpolated to generate daily cover values (Auerswald et al., 2019b). For the subsequent period (April 1997–2001), we applied standardised daily crop development and residue cover curves, which were derived from the daily crop and residue cover values observed during the detailed monitoring phase in combination with management information (mainly sowing and harvest dates) observed between 1997 and 2001 (Auerswald et al., 2019b; Fiener et al., 2019a). Total soil cover was calculated with residues protecting portions of the otherwise exposed soil according to:

$$Co_{tot} = Co_{crop} + \left(100 - Co_{crop}\right) * \frac{Co_{res}}{100}, \qquad (2)$$

Where Co_{tot} is the total soil cover (%), Co_{crop} the cover of the growing crop on the respective field (%), and Co_{res} the measured soil cover of the residues (%).

Figure 2 illustrates the total soil cover on the respective fields with monthly rainfall erosivity.

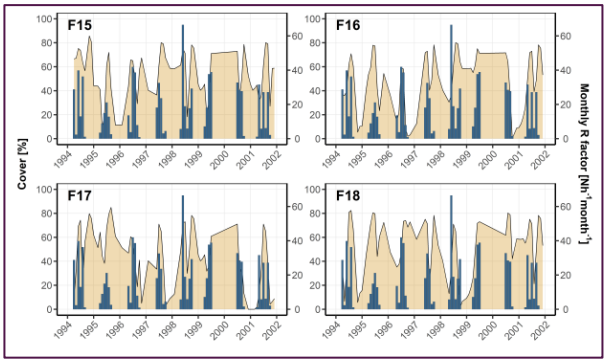


Figure 2: Each field's total soil cover (Residues and crops). Blue bar plots (monthly sum) show monthly R-factors

The soil loss ratio (SLR) quantifies the protective effect of soil cover by comparing potential soil loss under a given vegetation condition to that under standardised fallow conditions (Schwertmann et al., 1987; Wischmeier and Smith, 1978). While the SLR traditionally considers five crop growth stages, from bare soil (0% cover) to full canopy coverage (75-100% cover), we also considered crop residue cover. Determining field-specific SLR values involved

categorising soil cover into the five growth stages and assigning corresponding SLR values. As no-till was applied at the research farm, lower SLR values were assigned than in conventional systems due to increased soil surface protection. These SLR values were obtained from Schwertmann et al. (1987) and adapted based on our expert knowledge regarding the soil conservation practices in the Scheyern experimental farm (Fiener and Auerswald, 2007; Fiener et al., 2019a).

The annual C factor was calculated by multiplying the monthly proportion of the R factor with the average SLR value of the respective month:

$$C = \sum_{month=1}^{12} SLR_{month} * R_{prop,month} \quad (3)$$

Where C is the cover management factor (dimensionless), SLR_{month} the average soil loss ratio value of the respective month and $R_{prop,month}$ the proportion of the annual rainfall erosivity factor for the respective month (dimensionless).

Figure 3 shows the annual C and R factor throughout the entire study period.

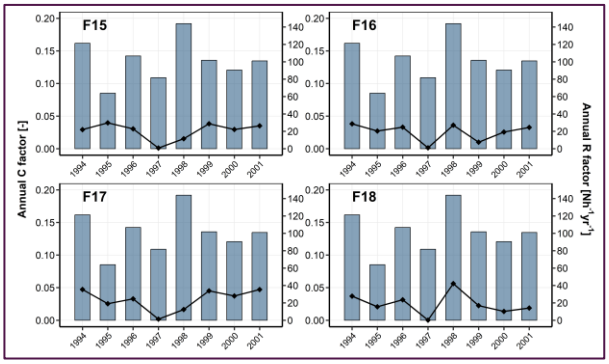


Figure 3: Annual C factor (black dots and lines) and R factor (blue bars) for the individual fields F15 to F18 over eight years."

Please note that additional changes have been made, as anonymous referee #1 also pointed towards this section (please see **detailed comments section of Anonymous, Referee #1, no. 8 and 10**).

15	Lines 238-240: These values were obtained from literature or included in a further analysis, e.g. GLUE. Please clarify.	**Preprint, L237-240:** "The transport capacity coefficient depends on surface roughness and therefore differs according to land use and management. In our model parameterisation, we distinguish between higher values for arable (kTC/A) land and lower values for grassland (kTC/G; along field borders and in grassed waterways). " We realised that introducing the k_{TC} parameters without specifying how they will be used in our methodology might create confusion. The k_{TC} parameters are model parameters sampled and conditioned within the GLUE framework (as detailed in Section 2.7 and Table 1), rather than fixed values taken from literature. In this paragraph, we added a clarification to avoid confusion: **Section 2.5; Revised manuscript; L266-268:** "In our model parameterisation, we distinguish between higher values for arable land ($k_{TC/A}$) and lower values for grassland ($k_{TC/G}$; along field borders and in grassed waterways), which were subjected to the GLUE-based analysis (see Section 2.7)."
16	Lines 260-261: How was this standard deviation applied? Please clarify.	**Preprint, 260-261:** "The standard deviation across all watersheds (± 13.7 %) was applied to account for measurement error in the trapping efficiency values. " We thank the referee spotting this inconsistency. The application of the standard deviation is described in the next section (Section 2.6). We clarified this in the manuscript: **Section 2.5; Revised manuscript; L288-289:** "The standard deviation across all watersheds (± 14%) was applied to account for measurement error in the trapping efficiency values within the GLUE framework (see Section 2.6)."
17	Line 270: So the runoff samples are collected in a barrel. This has not been described under Data.	**Preprint, L269-271:** "These included Coshocton wheel measurement errors (± 10%, Fiener and Auerswald, 2003), runoff collector barrel sampling errors (estimated ± 10 %), and retention pond uncertainties (± 14 %). "

		This comment refers to the referee's comment **no. 13**. To be consistent with the section above, we changed the word "barrel" to "storage tank" throughout the manuscript: **Section 2.6; Revised manuscript; L297-302:** "These included Coshocton wheel measurement errors (± 10%, Fiener and Auerswald, 2003), runoff collector storage tank sampling errors (estimated ± 10%), and retention pond uncertainties (± 14 %). For events with data collection issues (flagged in the data set), we assigned an additional ± 50 % error margin. However, for events flagged as "storage tank overflown", we introduced only an upper error boundary since the measurement taken from the storage tank represents a minimum possible sediment yield during a rainfall event."
18	Lines 317-319: Here the authors mean that the study areas were subdivided into field- and structure-dominated systems? But that was already defined much earlier in this section. Why is there a need to repeat that here? Please clarify.	**Preprint, L316-319:** "First, we calculated an eight-year median of the sediment yield for each individual watershed. Second, we spatially aggregated the watersheds based on their dominant erosion characteristics (field- and structure-dominated) while maintaining an annual resolution. " We agree that the watershed classification was already defined in Section 2.1. However, the intent of this sentence in Section 2.8 was not to re-define the groups, but to describe the specific data processing step where model outputs were statistically aggregated across these groups. To avoid the impression of repetition, we have refined the wording to emphasize that we are aggregating the results (sediment yields) rather than the watersheds themselves. **Section 2.8; Revised manuscript; L353-356:** "Second, we spatially aggregated the simulated sediment yields by calculating their means for each watershed group (field- and structure-dominated), but keeping an annual resolution."

<table>
<tr>
<td>19</td>
<td>Lines 319-320: How did the authors aggregate the median values between field- and structure-dominated systems, by taking the average? Please clarify.</td>
<td>Preprint, L319-320: “Third, we aggregated the median values over the eight-year monitoring period for these spatially aggregated groups. “

You are right that this sentence needs more clarification. Please note that we made statistical changes regarding the median and the mean in the manuscript (Discussed in the detailed comment section Anonymous, Referee #1, no. 12). We changed the sentence in the manuscript:

Section 2.8; Revised manuscript; L355-356: “Third, eight-year means for each watershed group were calculated (spatial and temporal aggregation).”</td>
</tr>
<tr>
<td>20</td>
<td>Lines 343-344: With “model performance” the authors mean “simulations”?</td>
<td>Preprint, L343-344: “When evaluated using eight-year median values, model performance showed better agreement with observations. “

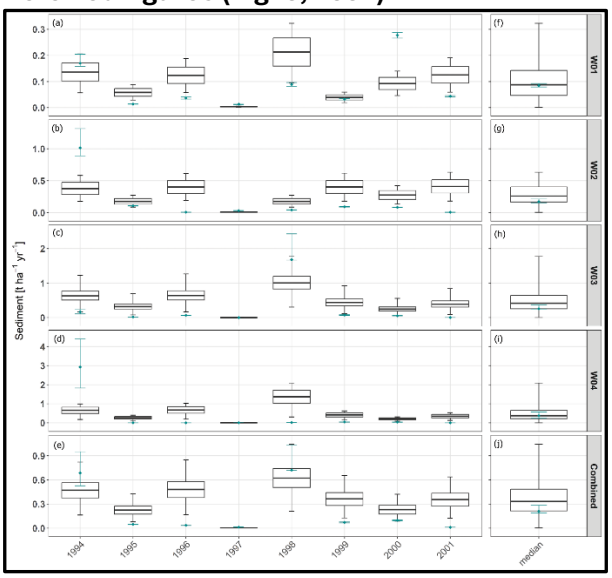
Thank you for finding this typo. We changed it in the manuscript:
Section 3.1; Revised manuscript; L386-387: “When evaluated using eight-year mean values per watershed, simulations showed better agreement with observations. “</td>
</tr>
<tr>
<td>21</td>
<td>Lines 344-345: It seems that the median of the median is higher than 0.3 t/ha/yr (based on the boxplot in panel j of Figure 3), but here the authors suggest 0.24 t/ha/yr. Please explain where this value is based on.

Lines 345-347: Similar to the previous comment. The simulated median of the median is higher than 0.5 t/ha/yr (panel d of Figure 4). Please explain where the 0.15 t/ha/yr is based on.

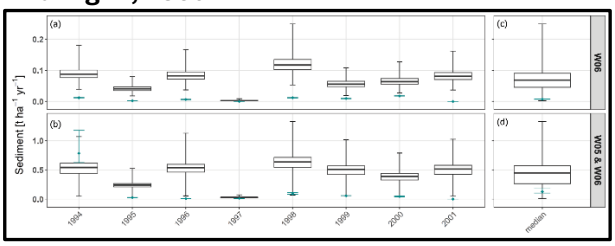
Table 2: If the Simulated SY is indeed the median of the median, then these values do not align with what is shown in Figures 3 and 4. See previous comments about this.</td>
<td>Preprint, L344-347: “The eight-year median modelled sediment yield across field-dominated watersheds (W01-W04) was 0.24 t ha^{-1} yr^{-1}, closely aligning with the measured eight-year median of 0.21 t ha^{-1} yr^{-1}. For structure dominated watersheds W05 and W06, we simulated an eight-year median of 0.15 t ha^{-1} yr^{-1} (Fig. 4c, d), against a measured median of 0.13 t ha^{-1} yr^{-1}. “

Referred table (Table 2, L376):
<table>
<tr><th>Unit of measure</th><th></th><th>Field-dominated</th><th>Structure-dominated</th></tr>
<tr><td>Behavioural realisations [%]</td><td></td><td>30.04</td><td>1.33</td></tr>
<tr><td>Measured SY [t ha^{-1} yr^{-1}]</td><td>Med.</td><td>0.21</td><td>0.13</td></tr>
<tr><td>Simulated SY [t ha^{-1} yr^{-1}]</td><td>Med.</td><td>0.24</td><td>0.15</td></tr>
<tr><td>MAE [t ha^{-1} yr^{-1}]</td><td>Min.
Med.
Max.</td><td>$4.21*10^{-6}$
0.03
0.07</td><td>$5.76*10^{-5}$
0.03
0.05</td></tr>
<tr><td>PBIAS [%]</td><td>Min.
Med.
Max.</td><td>-10.79
15.35
35.38</td><td>-17.70
20.15
42.64</td></tr>
</table>
</td>
</tr>
</table>

Referred figures (Fig. 3, L337):



And Fig. 4, L350:



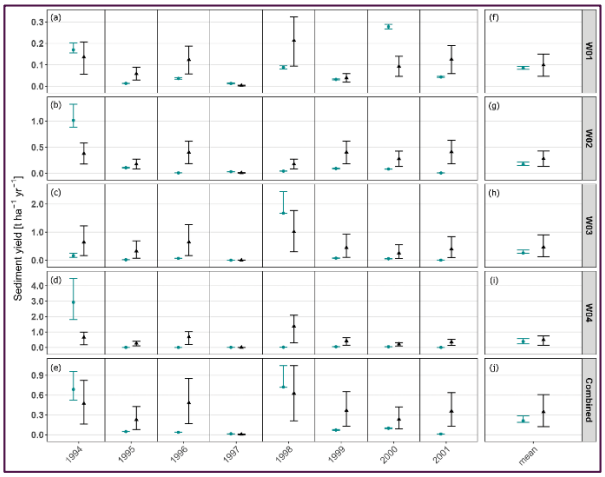
Thank you for highlighting this discrepancy. You are entirely correct that there was a misalignment between the values in the text and the figures. Initially, this stemmed from a confusion between the median of the behavioural model iterations versus the full set of 25,000 iterations.

Upon recalculating the medians to address your comment, we identified a coding error that had affected the calculation of all follow-up computations. We would like to thank the referee again, as we likely would not have found this issue without your review. We have re-run the analysis and updated the values. Please note, that we also changed to mean values as referee #1 addressed the discrepancy between mean and median in soil erosion modelling (see detailed comments section of the **Anonymous, Referee #1, no. 12**). We believe the new results still support our arguments well. Further, we restructured our results section to have a scale-depended line of argumentation:

Abstract; Revised manuscript; L24-27: "Model performance improved substantially when outputs were averaged over the eight-year monitoring period, with mean absolute errors of 0.14 t ha^{-1} yr^{-1} for field-dominated and 0.29 t ha^{-1} yr^{-1} for structure-dominated watersheds."

Section 3.1; Revised manuscript; L367-372:
"The modelled annual sediment yields for field-dominated watersheds (W01-W04) were within the same order of magnitude of the measurements. However, the model was not considered behavioural for predicting annual sediment yields according to our pre-established acceptability criterion. The simulated annual sediment yields were predominantly overestimated (22 out of 32 cases; Fig. 4a-d), occasionally underestimated (4 out of 32 cases; in the year 1997 and 2000 in W01; 1994 in W02; 1994 in W04; Fig. 4a, b and d), with only a small portion of simulations overlapping the observational data (6 out of 32 cases; Fig. 4a-d)."

Figure 4, L378-385; Revised manuscript:



"Figure 4: Annual and eight-year mean sediment yields in field-dominated watersheds. (a-d) Annual sediment yields: Black triangles indicate the mean and the full range from 25,000 model realisations (black whiskers), while cyan dots represent mean measured sediment yields with computed error ranges (cyan whiskers). (f-i) The watershed-specific eight-year mean measured sediment yields (cyan) and eight-year mean simulated yields (black). (e) Spatially combined annual watershed sediment yields. (j) Spatially aggregated eight-year mean yields. Note: In some years (e.g., 1998), cyan whiskers show larger uncertainties above the mean values; this

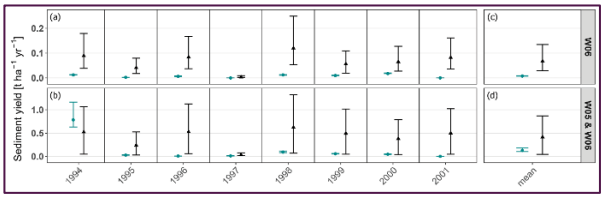
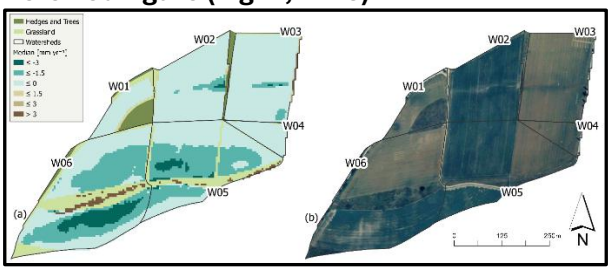
		is because storage tank overflow contributes only to higher uncertainties (see Section 2.6)." **Section 3.1; Revised manuscript; L387-391:** "The temporal aggregation revealed varying proportions of behavioural model realisations across individual watersheds. W04 had the highest amount with 69 % of all realisations, while other watersheds exhibited lower proportions (W01: 13 %, W02: 22 %, W03: 22 %). W05 exhibited minimal behavioural realisations of 1 %. In W06 (Fig. 5c), none of the actual model realisations matched the observational data including measurement errors." **Figure 5, L396-400; Revised manuscript:**  "Figure 5: Annual and eight-year mean sediment yields in structure-dominated watersheds. (a-b) Annual sediment yields: Black triangles indicate the mean and the full range from 25,000 model realisations (black whiskers), while cyan dots represent mean measured sediment yields with computed error ranges (cyan whiskers). (c-d) The watershed-specific eight-year mean measured sediment yields (cyan) and eight-year mean simulated yields (black)." **Section 3.1; Revised manuscript; L402-404:** "Across the entire set of 25,000 realisations, the mean MAE values were 0.14 t ha-1 yr-1 for field-dominated watersheds and 0.29 t ha-1 yr-1 for structure-dominated watersheds, with maximum MAE values of 0.40 t ha-1 yr-1 and 0.74 t ha-1 yr-1, respectively." **Section 3.1; Revised manuscript; L405-408:** "The eight-year mean modelled sediment yield across field-dominated watersheds (W01-W04) was 0.35 t ha^{-1} yr^{-1}, compared to the measured eight-year mean of 0.21 t ha^{-1} yr^{-1}. For structure-dominated watersheds W05 and W06, we simulated an eight-year mean of 0.41 t ha^{-1} yr^{-1} (Fig. 5c, d), against a measured mean of 0.13 t ha^{-1} yr^{-1}. "

		Table 2, L411-413; Revised manuscript: "Table 2: Comparison of model performance metrics between micro-scale watershed groups based on eight-year mean of behavioural realisations, including mean sediment yield (SY) as well as error statistics (MAE, PBIAS) with maximum (Max.), minimum (Min.) and mean values." **Conclusion; Revised manuscript; L641-644:** "The model was unable to produce behavioural realisations for watersheds optimised for soil conservation and sediment transport reduction at annual time steps based on our strict limits of acceptability criterion despite the small absolute prediction errors over all model realisations (eight-year MAE = 0.14 t ha^{-1} yr^{-1} for field-dominated and eight-year MAE = 0.29 t ha^{-1} yr^{-1} for structure-dominated watersheds). "
22	Lines 356-358: This means that the model is performing better in W04 and worse in W05?	**Preprint, L356-358:** "W04 had the highest amount with 57 % of all realisations, while other watersheds exhibited lower proportions (W01: 13 %, W02: 21 %, W03: 23 %). W05 exhibited minimal behavioural realisations of 1 %. " We thank the referee for this question. We would like to clarify that the percentage of behavioural realisations refers to the size of the valid parameter space, rather than the model's performance.

Unit of measure		Field-dominated	Structure-dominated
Behavioural realisations [%]		28.70	1.35
Measured SY [t ha^{-1} yr^{-1}]	Mean	0.21	0.13
Simulated SY [t ha^{-1} yr^{-1}]	Mean	0.24	0.15
MAE [t ha^{-1} yr^{-1}]	Min.	$4.21*10^{-6}$	$5.76*10^{-5}$
	Mean	0.04	0.03
	Max.	0.08	0.05
PBIAS [%]	Min.	-11.51	-17.70
	Mean	15.19	15.96
	Max.	36.69	42.29

		The high percentage of behavioural model realisations in W04 (69%) indicates that it is easier to find parameter sets that fit the observations, which could be seen as a higher degree of equifinality (many different parameter combinations yield similar results). In contrast, the low percentage in W05 (1%) indicates that the system is highly constrained, meaning only very specific parameter combinations can reproduce the complex processes observed there. While finding these parameters is more difficult, the resulting simulations are not worse than those in W04; they are simply more sensitive to parameter selection.
23	Figure 7: It seems that negative values are erosion and positive values deposition, please indicate this in the figure caption.	**Referred figure (Fig. 7, L413):**  "Figure 7: (a) The median of simulated potential erosion and deposition of behavioural model realisations over the eight year period. (b) An aerial photograph of the study area shows the land use patterns and field boundaries on the Scheyern experimental farm in 2002." We have updated the figure caption to explicitly state that negative values indicate erosion, and positive values indicate deposition. Please note that the figure has entirely been changed to address comments from Referee #1 (see **referee #1, comment no. 13**) regarding uncertainty visualization. **Section 3.3; Revised manuscript; L455-457:** "Figure 8: (a) The mean of simulated potential erosion and deposition of behavioural model realisations over the eight-year period. Negative values indicate erosion and positive values deposition. (b) The cell-wise standard deviation of behavioural model realisations over the eight-year period."

24	Lines 476-479: In that sense, it would be logical to also apply the model using long-term average rainfall erosivity values for the different study areas, instead of taking the median of the annual results.	**Preprint, L475-479:** "This temporal limitation aligns with findings by Risse et al. (1993), who demonstrated that USLE's model efficiency diminishes at the annual scale. When averaging over the eight-year study period, these extreme events are smoothed out, which explains the model's improved performance at longer timescales (Tab. 2). This observation supports the basic assertion that the USLE was designed to compute long-term soil losses (Wischmeier and Smith 1978). " We agree with the reviewer that for standard conservation planning, using long-term average rainfall erosivity is the intended and most logical approach for USLE-type models. However, the specific objective of this study was to utilise our exceptionally detailed dataset to test the model outside its actual purpose. We have addressed this already in response to the **detailed comment no. 1, Anonymous, Referee #1.**
25	Line 520-521: The TC is controlled by any value of kTC/A, not only high values. Please revise the sentence accordingly.	**Preprint, L520-521:** "The *TC* within agricultural fields is primarily controlled by a high transport coefficient *kTC/A* (Van Rompaey et al., 2001). " Thank you for finding this mistake. You are right. We changed the sentence accordingly: **Section 4.5; Revised manuscript; L591-592:** "The *TC* within agricultural fields is primarily controlled by the transport coefficient $k_{TC/A}$ (Van Rompaey et al., 2001). "

26	Lines 520-526: What is exactly the point the authors want to make here? That TC remains high enough to transport all sediment, without causing deposition. The question should be if this coincides with the observations or does the inclusion of retention ponds has a large influence on the modelled processes? (Ok, this is further explained in the subsequent paragraphs. I suggest to add 1-2 sentences explaining the likely reasons for this behaviour in the model, which you subsequently explain in the following paragraphs.)	**Preprint, L520-526:** "The *TC* within agricultural fields is primarily controlled by a high transport coefficient *kTC/A* (Van Rompaey et al., 2001). Lower *kTC/A* values reduce TC, promoting in-field deposition and consequently decreasing sediment yield at the watershed outlet. Our analysis revealed behavioural model realisations across the full *a priori* selected range of *kTC/A* values, with no clear pattern for field-dominated watersheds, demonstrating no sensitivity even at very low *kTC/A* values near 1 or very high *esur* values of 0.5 (Fig. 5a). This lack of sensitivity may be attributed to the implementation of retention ponds in W01 and W02 and by the very low simulated erosion values, as *TC* remained sufficiently high to transport the generally low sediment fluxes even with very low *kTC/A* values. " The point we intend to make is that due to the effective soil conservation measures, the sediment supply is so low that it rarely exceeds the transport capacity (*TC*), even when the transport coefficient is set to low values. Therefore, the model is insensitive to $k_{TC/A}$ because the system on the hillslopes effectively becomes supply-limited rather than transport-limited. Following the reviewer's advice, we have added a sentence to explicitly describe this process mechanism and the resulting dominance of retention ponds in controlling sediment yield in these specific watersheds. **Section 4.5; Revised manuscript; L595-598:** "This lack of sensitivity may be attributed to the retention ponds in W01 and W02 and the very low simulated erosion values. Since TC remained sufficiently high to transport the low sediment fluxes even with very low kTC/A values, sediment transport is thus supply-limited rather than constrained by transport capacity within the fields."

Further changes in the revised manuscript:

Preprint	Revised manuscript
L10: "essential"	**L10:** "important"
L13: "in"	**L13:** "for"
L14: "In this study"	**L14:** "Here"
L15-16: "[...] simulate sediment yields in six micro-scale watersheds ranging from 0.8 to 7.8 ha, monitored over eight years from 1994 to 2001."	**L15-16:** "[...] simulate sediment yields in six highly instrumented micro-scale watersheds ranging from 0.8 to 7.8 ha, monitored over eight years from 1994 to 2001, in Southern Germany"
L16: "comprised"	**L17:** "composed"
L19-20: "This setup enabled a separate analysis of model performance for both watershed groups."	**L20-21:** "Arable fields in both watershed groups were managed for soil conservation, including no-till and optimised crop rotations."
L22: "generally"	**Deleted**
L24-25: "However, the WaTEM/SEDEM´s performance improved substantially when model realisations were aggregated across the eight-year monitoring period and over the two watershed groups [...]"	**L24-25:** "Model performance improved substantially when outputs were averaged over the eight-year monitoring period [...]"
L27-30: "Our findings demonstrate that the model can represent the influence of soil conservation measures on reducing soil erosion and sediment delivery but performs better for long-term conservation planning at larger scales than for precise annual predictions in individual micro-scale watersheds with specific conservation practices."	**L27-31:** "Our findings demonstrate that WaTEM/SEDEM can represent the influence of soil conservation practices on reducing soil erosion and sediment yield in our study area. However, the model is fit for long-term conservation planning at larger spatial scales and not for precise annual predictions for individual micro-scale watersheds with specific conservation practices even if high-resolution, high-quality input data are available for parameterisation."

L38-42: This is particularly problematic in regions where the intensification of agriculture, exemplified by the historical increase in the size and weight of agricultural machinery that has led to increased soil compaction levels (Brus and Van Den Akker, 2018; Keller et al., 2019), and the increase in frequency and intensity of extreme precipitation events due to climate change (Auerswald and Fiener, 2024; Hosseinzadehtalaei et al., 2020; Myhre et al., 2019) is likely exacerbating the erosion risk.	**L39-43:** "This is particularly problematic in regions where agricultural intensification (e.g. field consolidation, soil compaction Brus and Van Den Akker, 2018; Keller et al., 2019; Foucher et al., 2014; Wang et al., 2022) and the increase in frequency and intensity of extreme precipitation events due to climate change (Auerswald and Fiener, 2024; Hosseinzadehtalaei et al., 2020; Myhre et al., 2019) are exacerbating the erosion hazards."
L49: "[…] or a generally optimised layout […]"	**L51:** "[…] or generally optimised layout […]"
L51-54: "Diverse models have been developed and applied for this purpose, ranging from empirical and conceptual to process-oriented model types (e.g. Eekhout et al., 2018; Smith et al., 2018; Nearing, 2013; Dymond et al., 2010; Hessel and Tenge, 2008)."	**L54-56:** "Diverse models have been developed and applied for this purpose, ranging from empirical or conceptual to process-oriented (e.g. Eekhout et al., 2018; Smith et al., 2018; Nearing, 2013; Dymond et al., 2010; Hessel and Tenge, 2008)."
L62: "RUSLE"	**L66:** "USLE-technology"
L64-66: "sediment delivery in small rivers of mesoscale catchments (e.g. Batista et al., 2022; Rehm and Fiener, 2024), or long-term erosion and deposition patterns derived from radionuclides (e.g. Van Oost et al., 2000; Wilken et al., 2020)."	**L68-70:** "sediment yield in small rivers of mesoscale catchments (e.g. Batista et al., 2022; Rehm and Fiener, 2024), and long-term erosion and deposition patterns derived from radionuclides (e.g. Van Oost et al., 2000; Wilken et al., 2020)."
L86-97: "Notwithstanding the spatial extent of (long-term) soil erosion monitoring, measurement uncertainties arise from instrumental precision and temporal instrument malfunctioning, data handling and processing. The uncertainties in observational data have important implications for erosion modelling, as models cannot be expected to be better than the observational data (Beven and Lane, 2022; Beven, 2019). The Generalized Likelihood Uncertainty Estimation (GLUE) framework (Beven and Binley, 1992) allows for testing environmental models while accounting for the uncertainty in both models and the observational data. In light of inherent measurement uncertainties, GLUE acknowledges that it is not possible to identify a single parameter set as "correct". Rather, all parameter combinations that	**L82-92: "**Regardless of the spatial scale in which erosion is monitored, it is important to note that perfect observational data do not exist. All measurements include errors stemming from instrumental precision, temporary malfunctioning, and data handling and processing. These uncertainties have important implications for evaluating erosion models, which cannot be expected to be better than the observational data used for model conditioning and testing (Beven and Lane, 2022; Beven, 2019). One approach for evaluating (uncertain) environmental models is the Generalized Likelihood Uncertainty Estimation (GLUE) framework (Beven and Binley, 1992). GLUE acknowledges that it is not possible to identify a single calibrated parameter set as "correct". Rather, all parameter combinations that produce results within given limits-of-acceptability cannot be

produce results within the observational uncertainty cannot be rejected. Within the GLUE framework, limits of acceptability are defined to identify which model runs fall within the uncertainty bounds of the measurements (Beven and Lane, 2022). These behavioural models are retained, while non-behavioural models are rejected. This limits-of-acceptability GLUE approach thus provides a systematic methodology to evaluate model performance with uncertain testing data."	rejected (Beven and Lane, 2022). Contrarily, if not a single model realisation encompasses the uncertainty bounds of the observational data, non-behavioural models or model structures can be rejected, which might lead to improvements in terms of understanding and modelling."
L109: "southern Germany"	**L109:** "Southern Germany"
L131: "sediment delivery"	**L132:** "sediment yield"
L137: "the down-frozen mustard"	**L138-139:** "the frost-killed mustard"
L141: "2.2 Data"	**L145:** "2.2 Erosion monitoring data"
L148-149: "For model testing, we used continuous sediment delivery data from the six micro-scale watersheds between 1994 and 2001."	**L152-153: "**For model testing, we used continuous sediment yield data from the six micro-scale watersheds (W01-W06) between 1994 and 2001 (Fig. 1)."
L158-161: "The WaTEM/ SEDEM version used in this study consists of two main components: (i) WaTEM, which implements a spatially distributed German adaption of the USLE , and (ii) SEDEM, which incorporates a transport capacity (TC) equation (Eq. 3) and a routing algorithm for sediment re-distribution based on a DEM (Verstraeten et al., 2002; Van Rompaey et al., 2001; Van Oost et al., 2000)."	**L165-168: "**The WaTEM/ SEDEM version used in this study consists of two main components: (i) WaTEM, which implements a spatially distributed German adaption of the USLE (Schwertmann et al., 1987; Din-Normenausschuss, 2022), and (ii) SEDEM, which incorporates a transport capacity (TC) equation (Eq. 4) and a routing algorithm for sediment re-distribution based on the DEM (Verstraeten et al., 2002; Van Rompaey et al., 2001; Van Oost et al., 2000)."
L168-169: "While these months contributed 10.7 % of the total measured sediment delivery (Fiener et al., 2019b), our analysis focused on the dominant water erosion period during heavy rainfall months."	**L178-179: "**While the colder months contributed 10.7% of the total measured sediment yield (Fiener et al., 2019b), our analysis focused on the dominant water erosion period during heavy rainfall months."
L171: "2.4 Potential Erosion"	**L181:** "2.4 Potential erosion"
L177: "rainfall erosivity factor in"	**L188: "**rainfall erosivity factor"

L220-223: "The support practices factor (P factor) was not specifically parametrised for contour-seeding because of field heterogeneity, i.e. not all parts of a single field were contour-seeded, and/or the absence of specific P factor values for structures such as the potato dams. However, we accounted for the uncertainty stemming from this lack of parameter representation as part of the model conditioning process (see section 2.4 below)."	**L245-251: "**The support practices factor (P factor) was not specifically parametrised for contour-seeding because of field heterogeneity, i.e. not all parts of a single field were contour-seeded, and/or the absence of specific P factor values for structures such as the potato dams. Furthermore, the field geometries often result in high L factors that often exceed the critical slope length limit for effective contouring defined in the German USLE (Schwertmann et al., 1987; Din-Normenausschuss, 2022). Hence, the effective P-factor converges towards 1.0. We accounted for the uncertainty stemming from this lack of parameter representation as part of the model conditioning process (see section 2.4 below)."
L224-226: "2.5 Sediment Transport and Deposition The Transport Capacity (TC) quantifies the maximum amount of sediment transported through a grid cell without deposition."	**L252-254: "**2.5 Sediment transport and deposition Transport capacity (TC) quantifies the maximum amount of sediment transported through a grid cell without deposition."
L229-231: Equation numbers 3 and 4. **L290:** Equation number 5. **L299-301:** Equation numbers 6 and 7. **L322:** Equation number 8.	**L257-259:** Equation numbers 4 and 5. **L318:** Equation number 6. **L334-336:** Equation numbers 7 and 8. **L358:** Equation number 9.
L246: "Parcel Connectivity (p_{con})"	**L274: "**parcel connectivity (p_{con})"
L249: "it affects the TC"	**L277: "**it affects TC"
L262: "2.6 Generalised Likelihood Uncertainty Estimation (GLUE) framework"	**L290: "**2.6 Generalised Likelihood Uncertainty Estimation (GLUE)"
L264: "sets"	**L292:** "spaces"
L277: "served as criterion for behavioural model realisations"	**L305-306:** "served as criterion for identifying behavioural model realisations"
L280: "2.7 Model evaluation"	**L308:** "2.7 Model conditioning and evaluation"
L311: "the"	**Deleted**
L314: "erosion control measures"	**L349:** "erosion control practices"
L326: "3.1 Model Performance Across Scales"	**L366:** "3.1 Model performance across scales"

L332-334: "The tendency to overestimate sediment yield is more pronounced in watersheds W05 and W06. Only in 1994 the model underestimated measured sediment yields in watershed W05 (Fig. 4b)."	**L372-374: "**The tendency to overestimate sediment yield is more pronounced in watersheds W05 and W06. Only in 1994 the model had the tendency to underestimate measured sediment yields in watershed W05 (Fig. 5b)."
From L384: All figure numbers one lower.	**From L370**: All figure numbers one higher.
L378: "watershed groups"	**L415:** "model outputs"
L386: "s_{cor}"	**L423:** "e_{sur}"
L392-393: "In structure-dominated watersheds, the analysis focused on parameters controlling sediment transportation and deposition in grassland ($k_{TC/G}$, b_{dep}, and p_{con})."	**L429-430:** "In structure-dominated watersheds, the analysis focused on parameters controlling sediment transport and deposition in grasslands and landscape structures ($k_{TC/G}$, b_{dep}, and p_{con})."
L395-396: "showed no sensitivity, displaying relatively"	**L433:** "displayed relatively"
L418: "4.1 GLUE Framework and Uncertainties"	**L459: "**4.1 GLUE framework and uncertainties"
L420: "approach"	**L461:** "procedure"
L440-449: "This is particularly relevant for our study area, where the median measured sediment yield of 0.16 t ha^{-1} yr^{-1} was substantially lower than erosion rates which can exceed 10 t ha^{-1} yr^{-1} in the Bavarian Tertiary hill region (Auerswald et al., 2009). To investigate the modelling under/over prediction issue, we used an error surface (esur) multiplied with the erosion calculated by the USLE (Eq. 5). The esur parameter served three purposes: (i) adjusting potential erosion to investigate the USLE's inherent biases, (ii) analysing the biases by looking at the behaviour of esur, and (iii) representing the uncertainty stemming from measurement errors for the USLE factors and the lack of parameterisation for the P factor. The analysis of behavioural model realisations revealed a concentration of likelihood values near small esur values, reducing sediment by up to 50 % (Fig. 5b). This indicates that in our study WaTEM/SEDEM overestimates soil erosion in landscapes with implemented conservation measures."	**L481-488: "**This is particularly relevant for our study area, where the implemented watershed-wide soil conservation and sediment trapping resulted in a measured mean sediment yield of only 0.16 t ha^{-1} yr^{-1}, which is substantially lower than erosion rates typically range between 3-10 t ha^{-1} yr^{-1} in the Bavarian Tertiary hill region (Auerswald et al., 2009). To investigate the modelling under/over prediction issue, we used an error surface (esur) multiplied with the erosion calculated by the USLE (Eq. 6). The analysis of behavioural model realisations revealed a concentration of likelihood values near small esur values, reducing sediment by up to 50 % (Fig. 6b). This indicates that in our study WaTEM/SEDEM overestimates soil erosion in landscapes with implemented conservation practices."

L451-457: "4.2 Model Performance and Limitations WaTEM/SEDEM correctly simulated the magnitude of the very low sediment yields in micro-scale watersheds under optimized soil conservation, with annual values closely aligning with measured data (Fig. 3a-d, 4a-b). Despite this achievement, the model did not consistently meet our strict limits of acceptability for annual realisations and therefore was rejected for making precise annual simulations. However, the model's performance improved notably when applied to longer-term medians and larger spatial units, where more behavioural model realisations were identified."	**L490-496:** "4.2 Model performance and limitations WaTEM/SEDEM correctly simulated the magnitude of the very low sediment yields in micro-scale watersheds optimised for soil conservation and reduced sediment transport, with annual values closely aligning with measured data (Fig. 4a-d, 5a-b). Nonetheless, the model did not consistently meet our limits of acceptability for annual realisations and therefore was rejected for making precise annual simulations. The model's performance improved notably when applied to longer-term means and larger spatial units, where more behavioural model realisations were identified."
L460-463: "In general, observed sediment yields were overestimated, which can be attributed primarily to difficulties in accurately representing the specific C factors of this conservation system, particularly unique practices such as mustard sown onto autumn-built dams where potatoes were later directly planted (Fiener and Auerswald, 2003)."	**L508-511: "**. In general, observed sediment yields were overestimated, which can be attributed primarily to difficulties in accurately representing this conservation system, particularly unique practices such as mustard sown onto autumn-built dams where potatoes were later directly planted (Fiener and Auerswald, 2003)."
L461-462: "the specific C factors"	**deleted**
L516-517: "long-term runoff and sediment yield reduction was about 90% and 97%"	**L563-564:** "long-term runoff and sediment yield reductions were respectively about 90% and 97%"
L519: "4.3 Distribution of behavioural model parameter values"	**L590: "**4.5 Distribution of behavioural model parameter values"
L551: "An additional limitation of the current parameterisation approach is its static nature."	**L623-624:** "An additional limitation of the current parameterisation approach, particularly for grassed waterways, is its static nature."

L583-588: “Ultimately, our study demonstrates that WaTEM/SEDEM can simulate the very low sediment yields observed from soil conservation agricultural systems, provided that high spatiotemporal resolution input data and locally adapted USLE factors (e.g., the ABAG for Southern Germany) are available. However, capturing the effects of linear landscape features like grassed waterways where concentrated runoff occurs remains challenging for WaTEM/SEDEM, primarily due to the model's inability to represent re-infiltrating processes that are critical for sediment trapping in such structures.”	**L660-665:** “Ultimately, our study demonstrates that WaTEM/SEDEM can simulate the magnitude of very low sediment yields observed from soil conservation agricultural systems, provided that high spatiotemporal resolution input data and locally adapted USLE factors (e.g., the ABAG) are available. However, capturing the combined effects of low in-field erosion and linear landscape features like grassed waterways, where concentrated runoff occurs, remains challenging for WaTEM/SEDEM, primarily due to the model's inability to represent re-infiltrating processes that are critical for sediment trapping in such structures.”

Sources:

Andersson, J. A., & D'Souza, S. (2014). From adoption claims to understanding farmers and contexts: A literature review of Conservation Agriculture (CA) adoption among smallholder farmers in southern Africa. AGR ECOSYST ENVIRON, 187, 116-132. https://doi.org/10.1016/j.agee.2013.08.008

Auerswald, K., Kainz, M., Scheinost, A., & Sinowski, W. (2001). The Scheyern Experimental Farm: Research methods, the farming system and definition of the framework of site properties and characteristics. In (pp. 183-194). https://doi.org/10.1007/978-3-662-04504-6_10

Auerswald, K., Fiener, P., & Dikau, R. (2009). Rates of sheet and rill erosion in Germany — A meta-analysis. Geomorphology, 111, 182-193. https://doi.org/10.1016/j.geomorph.2009.04.018

Auerswald, K., & Fiener, P. (2019). Soil organic carbon storage following conversion from cropland to grassland on sites differing in soil drainage and erosion history. SCI TOTAL ENVIRON, 661, 481-491. https://doi.org/10.1016/j.scitotenv.2019.01.200

Auerswald, K., Fiener, P., Gerl, G., & Wilken, F. (2019). Land use and land management data from the Scheyern experimental farm covering 14 small adjacent watersheds and their surroundings https://doi.org/10.13140/RG.2.2.26172.49285

Auerswald, K. & Fiener, P. (2024). Assessing the impact of climate change on soil erosion by water, in: Understanding and preventing soil erosion, Burleigh Dodds Science Publishing Limited, London, 51-76, https://doi.org/10.19103/AS.2023.0131.05

Batista, P. V. G., Davies, J., Silva, M. L. N., & Quinton, J. N. (2019). On the evaluation of soil erosion models: Are we doing enough? *EARTH-SCI REV*, 197, 102898. https://doi.org/10.1016/j.earscirev.2019.102898

Batista, P. V. G., Laceby, J. P., Davies, J., Carvalho, T. S., Tassinari, D., Silva, M. L. N., Curi, N., & Quinton, J. N. (2021). A framework for testing large-scale distributed soil erosion and sediment delivery models: Dealing with uncertainty in models and the observational data, Environmental Modelling & Software, 137, 104961, https://doi.org/10.1016/j.envsoft.2021.104961

Batista, P. V. G., Fiener, P., Scheper, S., & Alewell, C. (2022). A conceptual-model-based sediment connectivity assessment for patchy agricultural catchments, HYDROL EARTH SYST SC, 26, 3753-3770, https://doi.org/10.5194/hess-26-3753-2022

Beven, K. (2006). A manifesto for the equifinality thesis. *J Hydrol*, 320, 18-36. https://doi.org/10.1016/j.jhydrol.2005.07.007

Beven, K. (2006). Environmental Modelling: An Uncertain Future?. CRC Press, ISBN: 9780415457590

Beven, K. (2019). Towards a methodology for testing models as hypotheses in the inexact sciences, P ROY SOC A-MATH PHY, 475, 20180862, https://doi.org/10.1098/rspa.2018.0862

Beven, K., & Binley, A. (1992). The future of distributed models: Model calibration and uncertainty prediction. HYDROL PROCESS, 6(3), 279-298. https://doi.org/10.1002/hyp.3360060305

Beven, K., & Lane, S. (2022). On (in)validating environmental models. 1. Principles for formulating a Turing-like Test for determining when a model is fit-for purpose. HYDROL PROCESS, 36(10), e14704. https://doi.org/10.1002/hyp.14704

Borrelli, P., Alewell, C., Alvarez, P., Anache, J. A. A., Baartman, J., Ballabio, C., Bezak, N., Biddoccu, M., Cerdà, A., Chalise, D., Chen, S., Chen, W., De Girolamo, A. M., Gessesse, G. D., Deumlich, D., Diodato, N., Efthimiou, N., Erpul, G., Fiener, P., Freppaz, M., & Panagos, P. (2021). Soil erosion modelling: A global

review and statistical analysis, Science of The Total Environment, 780, https://doi.org/10.1016/j.scitotenv.2021.146494

Brazier, R. E., Beven, K. J., Freer, J., & Rowan, J. S. (2000). Equifinality and uncertainty in physically based soil erosion models: application of the GLUE methodology to WEPP-the Water Erosion Prediction Project- for sites in the UK and USA. *Earth Surf Processes*, 25, 825-845. https://doi.org/10.1002/1096-9837(200008)25:8<825::AID-ESP101>3.0.CO;2-3

Brus, D. J. & van den Akker, J. J. H. (2018). How serious a problem is subsoil compaction in the Netherlands? A survey based on probability sampling, SOIL, 4, 37–45, https://doi.org/10.5194/soil-4-37-2018

Choudhury, B. U., Nengzouzam, G., & Islam, A. (2022). Runoff and soil erosion in the integrated farming systems based on micro-watersheds under projected climate change scenarios and adaptation strategies in the eastern Himalayan mountain ecosystem (India). J ENVIRON MANAGE, 309, 114667. https://doi.org/10.1016/j.jenvman.2022.114667

Carter, C. E., & Parsons, D. A. (1967). Field tests on the coshocton-type wheel runoff sampler. T ASAE, 10(1), 133-135. https://doi.org/10.13031/2013.39613

de Vente, J., & Poesen, J. (2005). Predicting soil erosion and sediment yield at the basin scale: Scale issues and semi-quantitative models. Earth-Science Reviews, 71(1), 95-125. https://doi.org/10.1016/j.earscirev.2005.02.002

DIN-Normenausschuss, W. (2022). Soil quality Predicting soil erosion by water by means of ABAG. https://doi.org/10.31030/3365455

Dymond, J. R., Betts, H. D., & Schierlitz, C. S. (2010). An erosion model for evaluating regional land-use scenarios. Environmental Modelling & Software, 25, 289-298. https://doi.org/10.1016/j.envsoft.2009.09.011

Eekhout, J. P. C., Terink, W., & De Vente, J. (2018). Assessing the large-scale impacts of environmental change using a coupled hydrology and soil erosion model. Earth Surface Dynamics, 6, 687-703. https://doi.org/10.5194/esurf-6-687-2018

Fiener, P., & Auerswald, K. (2003). Effectiveness of grassed waterways in reducing runoff and sediment delivery from agricultural watersheds. *J Environ Qual*, 32, 927-936. https://doi.org/10.2134/jeq2003.9270

Fiener, P., & Auerswald, K. (2007). Rotation effects of potato, maize, and winter wheat on soil erosion by water. SOIL SCI SOC AM J, 71(6), 1919-1925. https://doi.org/10.2136/sssaj2006.0355

Fiener, P. & Auerswald, K. (2018). Grassed waterways. In: American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, 131-150. https://doi.org/10.2134/agronmonogr59.c7

Fiener, P., Wilken, F., & Auerswald, K. (2019b). Filling the gap between plot and landscape scale – eight years of soil erosion monitoring in 14 adjacent watersheds under soil conservation at Scheyern, Southern Germany. Advances in Geosciences, 48, 31-48. https://doi.org/10.5194/adgeo-48-31-2019

Foucher, A., Salvador-Blanes, S., Evrard, O., Simonneau, A., Chapron, E., Courp, T., Cerdan, O., Lefèvre, I., Adriaensen, H., Lecompte, F., & Desmet, M. (2014). Increase in soil erosion after agricultural intensification: Evidence from a lowland basin in France. Anthropocene, 7. https://doi.org/10.1016/j.ancene.2015.02.001

Hessel, R., & Tenge, A. (2008). A pragmatic approach to modelling soil and water conservation measures with a catchment scale erosion model. Catena, 74. https://doi.org/10.1016/j.catena.2008.03.018

Hosseinzadehtalaei, P., Tabari, H., & Willems, P. (2020). Climate change impact on short-duration extreme precipitation and intensity–duration–frequency curves over Europe. J HYDROL, 590, 125249. https://doi.org/10.1016/j.jhydrol.2020.125249

Keller, T., Sandin, M., Colombi, T., Horn, R., & Or, D. (2019). Historical increase in agricultural machinery weights enhanced soil stress levels and adversely affected soil functioning. Soil and Tillage Research, 194, 104293. https://doi.org/10.1016/j.still.2019.104293

Kinnell, P. I. A. (2007). Runoff dependent erosivity and slope length factors suitable for modelling annual erosion using the Universal Soil Loss Equation. *Hydrol Process*, 21, 2681-2689. https://doi.org/10.1002/hyp.6493

Myhre, G., Alterskjær, K., Stjern, C. W., Hodnebrog, Ø., Marelle, L., Samset, B. H., Sillmann, J., Schaller, N., Fischer, E., Schulz, M., & Stohl, A. (2019). Frequency of extreme precipitation increases extensively with event rareness under global warming. SCI REP-UK, 9, 16063. https://doi.org/10.1038/s41598-019-52277-4

Nearing, M. (1998). Why soil erosion models over-predict small soil losses and under-predict large soil losses. *Catena*, 32, 15-22. https://doi.org/10.1016/S0341-8162(97)00052-0

Nearing, M. (2013). Soil erosion and conservation. In: Environmental Modelling, edited by: Wainwright, J., and Mulligan, M., 365-378. https://doi.org/10.1002/9781118351475.ch22

Rehm, R., & Fiener, P. (2024). Model-based analysis of erosion-induced microplastic delivery from arable land to the stream network of a mesoscale catchment. SOIL, 10, 211-230. https://doi.org/10.5194/soil-10-211-2024

Renard, K. G. (1997). Predicting soil erosion by water: A guide to conservation planning with the Revised Universal Soil Loss Equation (RUSLE). Agricultural handbook, US Department of Agriculture, Agricultural Research Service.

Risse, M., Nearing, M. A., Laflen, J. M., & Nicks, A. D. (1993). Error assessment in the Universal Soil Loss Equation. *Soil Sci Soc Am J*, 57, 825-833. https://doi.org/10.2136/sssaj1993.03615995005700030032x

Schwertmann, U., Vogl, W., & Kainz, M. (1987). Bodenerosion durch Wasser: Vorhersage des Abtrags und Bewertung von Gegenmassnahmen. Stuttgart: Ulmer.

Smith, H. G., Peñuela, A., Sangster, H., Sellami, H., Boyle, J., Chiverrell, R., Schillereff, D., & Riley, M. (2018). Simulating a century of soil erosion for agricultural catchment management. Earth Surface Processes and Landforms, 43, 2089-2105. https://doi.org/10.1002/esp.4375

Srikanthan, R. & McMahon, T. A. (2001). Stochastic generation of annual, monthly and daily climate data: A review. Hydrology and Earth System Sciences, 5. https://doi.org/10.5194/hess-5-653-2001

Van Oost, K., Govers, G., & Desmet, P. (2000). Evaluating the effects of changes in landscape structure on soil erosion by water and tillage. LANDSCAPE ECOL, 15, 577-589. https://doi.org/10.1023/A:1008198215674

Van Rompaey, A. J. J., Verstraeten, G., Van Oost, K., Govers, G., & Poesen, J. (2001). Modelling mean annual sediment yield using a distributed approach. EARTH SURF PROCESSES, 26(11), 1221-1236. https://doi.org/10.1002/esp.275

Verstraeten, G., Van Oost, K., Van Rompaey, A., Poesen, J., & Govers, G. (2002). Evaluating an integrated approach to catchment management to reduce soil loss and sediment pollution through modelling. SOIL USE MANAGE, 18, 386-394. https://doi.org/10.1111/j.1475-2743.2002.tb00257.x

Wang, S., Szeles, B., Krammer, C., Schmaltz, E., Song, K., Li, Y., Zhang, Z., Blöschl, G., & Strauss, P. (2022). Agricultural intensification vs. climate change: what drives long-term changes in sediment load? Hydrology and Earth System Sciences, 26. https://doi.org/10.5194/hess-26-3021-2022

Wendt, R. C., Alberts, E. E., & Hjelmfelt, A. T. (1986). Variability of Runoff and Soil Loss from Fallow Experimental Plots. Soil Science Society of America Journal, 50. https://doi.org/10.2136/sssaj1986.03615995005000030035x

Wilken, F., Ketterer, M., Koszinski, S., Sommer, M., & Fiener, P. (2020). Understanding the role of water and tillage erosion from 239+240Pu tracer measurements using inverse modelling. SOIL, 6, 549-564. https://doi.org/10.5194/soil-6-549-2020

Wischmeier, W. H., & Smith, D. D. (1978). *Predicting rainfall erosion losses: A guide to conservation planning*. Agriculture Handbook, 537, Department of Agriculture, Science and Education Administration.