**Reviewer 1**

**General**

This manuscript examines the use of streamflow and downscaled and bias-corrected satellite soil moisture data for calibrating a SWAT-VSA model in a small, saturation-excess dominated watershed. The topic is relevant and timely; however, some issues need to be addressed before the manuscript can be considered for publication.

RESPONSE: Thank you for your comments and for acknowledging our topic as relevant and timely. Below, we have provided our responses to your comments in blue text.

1. The Introduction does not sufficiently clarify the specific novelty of this study relative to existing literature. If the main contribution lies in calibrating hydrological models using downscaled and bias-corrected satellite soil moisture, the authors should more clearly review and position their work within existing studies on satellite soil moisture downscaling, bias correction, and reported performance improvements. Similarly, if the key contribution is the application of SWAT-VSA in saturation-excess dominated watersheds, the importance and added value of this choice should be more explicitly articulated. At present, the objectives and innovation of the study are not clearly distinguished from prior work.

RESPONSE: Thank you for this comment. We substantially revised the Introduction to explicitly articulate the novelty of this study relative to previous work. Specifically, we now clarify (i) how our approach differs from prior studies that used satellite soil moisture for calibration, (ii) why the small, saturation-excess watershed context is unique, and (iii) how the SWAT-VSA structure allows us to evaluate field-scale soil moisture performance, something not demonstrated in previous studies. The revised text also explicitly situates our downscaling and bias-correction approach within the current literature, highlighting the lack of studies using downscaled satellite soil moisture in watersheds < 100 km², and especially the absence of studies evaluating performance at field scale using TI-based HRU aggregation. We now clearly state the conceptual and methodological contributions and end the Introduction with explicit, concise research objectives:

"Although many studies have incorporated satellite soil moisture into hydrologic model calibration, most report trade-offs: improving soil-moisture fit often coincides with degraded streamflow skill under soil-moisture-only or multi-objective calibration (e.g., Rajib et al., 2016; López Ló pez et al., 2017; Dangol et al., 2023; Eini et al., 2023; Mei et al., 2023). These mixed outcomes—observed across different model structures and satellite products—underscore a persistent uncertainty about when and why satellite soil moisture provides useful calibration information. Moreover, the majority of applications are in large watersheds (typically > 700 km²), with comparatively few demonstrations in small basins (< 100 km²), where land–atmosphere coupling and source-area activation can differ substantially (Duethmann et al., 2022; Dangol et al., 2023).

"A second limitation of prior work is that satellite soil moisture, original or downscaled, does not consistently reproduce the sub-grid spatial variability governing runoff generation in saturation-excess landscapes. Recent evaluations show that downscaled products can capture regional temporal trends but still under-represent locally organized wetness patterns and event-scale peaks, particularly in topographically complex headwaters (e.g., Duethmann et al., 2022). This raises questions about their utility for constraining parameters in watersheds where topography and hydrologic connectivity dominate soil-moisture distribution and VSA activation (Easton et al., 2008; Lyon et al., 2004).

In this context, SWAT-VSA offers a targeted test of downscaled satellite soil moisture. By using a topographic-index (TI)–informed HRU definition, SWAT-VSA redistributes soil and hydrologic properties along terrain-derived similarity classes, improving the representation of saturation patterns and runoff source areas relative to standard HRU schemes (Easton et al., 2008; Fuka et al., 2016). This structure also enables field-scale soil-moisture evaluation through TI-weighted aggregation of in-situ measurements, an approach rarely attempted in previous satellite-SM studies. Further, we use a simple, independently determined bias-correction of the downscaled SM to enhance event responsiveness. while preserving longer-term structure, acknowledging the attenuation of peaks typical of pretrained downscaler data.

Objectives

“We evaluate whether a simple, empirical, and independently bias-corrected, downscaled satellite soil-moisture product can effectively calibrate a SWAT-VSA model under single-objective (soil moisture or streamflow) and multi-objective strategies in a small, saturation-excess watershed (~14.5 km²), where most prior work has not been conducted (Duethmann et al., 2022; Dangol et al., 2023).

We assess field-scale soil-moisture estimation by leveraging SWAT-VSA’s TI-based HRUs, enabling evaluation of model output against TI-weighted in-situ soil-moisture measurements collected within a 4.2-ha monitoring field; this explicitly tests whether gains in statistical fit translate into improved hydrologic realism at management scales (Easton et al., 2008; Fuka et al., 2016; Buell, 2022).

Together, these elements distinguish our work from earlier satellite-soil moisture calibration studies: we (i) operate in a small, VSA-dominated catchment, (ii) use TI-structured HRUs to bridge scales between downscaled soil moisture and field observations, and (iii) explicitly separate statistical improvement from process-level realism through independent, field-scale evaluation and water-balance diagnostics (Rajib et al., 2016; Duethmann et al., 2022; Mei et al., 2023).”

2. The study is conducted in a single small watershed (~14 km²) dominated by saturation-excess runoff and relies on the SWAT-VSA model. While this choice is physically appropriate for the study site, it substantially limits the generality of the conclusions. The manuscript should more explicitly discuss the limits of applicability of the proposed approach, and the conclusions **should be clearly framed as primarily applicable to small, saturation-excess runoff dominated catchments**.

RESPONSE: Thank you for noting the need to more clearly delimit the scope of our conclusions. We have added a Limitations section to explicitly emphasize that (i) the study was conducted in a single, small (<15 km²) watershed, (ii) the watershed is strongly dominated by saturation-excess runoff, and (iii) SWAT-VSA is specifically designed for such hydrologic settings. We now state clearly that the proposed downscaling and bias-correction approach, as well as the soil-moisture–based calibration findings, should be interpreted primarily within the context of small, humid, saturation-excess systems. We also highlight challenges that may arise in larger or more heterogeneous basins, including increased spatial variability in soils, land cover, and precipitation, and the potentially weaker topographic control on soil moisture patterns. We have added:

“4.5 Limitations

“This study was conducted in a single, small 14.5 km², saturation-excess-dominated watershed. While this setting aligns with the study objective of evaluating soil-moisture-driven calibration in terrain-controlled hydrologic systems, it limits the generalizability of the findings. In larger or more heterogeneous watersheds, greater spatial variability in precipitation, land use, soils, and hydrologic connectivity may influence both the behavior of the downscaled soil-moisture product and the performance of the calibration approaches. For example, the simple bias-correction method applied here relies on watershed-scale effective precipitation and may not capture spatially variable wetting responses in more complex basins.

The applicability of the SWAT-VSA model is inherently restricted to saturation-excess runoff environments, where shallow storage deficits and topographic convergence dominate runoff generation. In such settings, the topographic index (TI) provides a meaningful basis for distributing hydrologic response units and representing spatial soil-moisture patterns. Infiltration-excess systems, arid or semi-arid regions, or watersheds with weak terrain control may not benefit from the TI-based structure in the same way. Nonetheless, saturation-excess processes are widespread in humid and wet-temperate regions of the United States (Buchanan et al., 2018; Easton et al., 2009) and globally, suggesting that this approach remains relevant to a broad class of small headwater catchments.

The downscaled soil moisture product underestimates peak wetness during high-moisture periods. The bias-corrected product retains the large-scale temporal structure of the downscaling model while using a physically informed threshold-based adjustment to restore event responsiveness. Consequently, while the corrected data support improved calibration performance, they do not fully resolve the underlying spatial or physical limitations of the satellite product.

Overall, the approach, combining downscaled and bias-corrected satellite soil moisture with SWAT-VSA, shows promise for improving soil moisture and streamflow estimation in small, saturation-excess systems. However, careful consideration of watershed size, dominant runoff mechanism, topographic controls, and local soil-moisture variability is essential when applying this framework elsewhere. Future work should extend this analysis to multiple watersheds with contrasting geomorphology, climate, and runoff generation processes to assess the broader transferability of the method."

3. The study relies on a pretrained mlhrsm downscaling model and an empirical, precipitation-threshold-based bias correction scheme. However, the regional applicability of the pretrained downscaling model is not evaluated, and **no comparison is shown between original, downscaled, and bias-corrected soil moisture**. The authors should explicitly show these differences and discuss how the downscaling and bias correction steps influence the soil moisture signal used for calibration and, consequently, the model results.

RESPONSE: Thank you for highlighting the need for comparing the soil moisture products. We have added the following in the results and discussions sections:

"To evaluate how the bias-correction procedures modify the soil moisture signal used for calibration, we compared two products over the study watershed: (1) the 500-m downscaled soil moisture generated using the mlhrsm pretrained quantile random forest model, and (2) the bias-corrected downscaled product used in calibration.  The downscaled soil moisture underestimates peak wetness and the overall dynamic range relative to in-situ observations. Following bias correction, the soil moisture product exhibits (a) increased peak responses immediately following precipitation events, ~~(~~(b) improved temporal agreement with in-situ measurements (Figure X). The bias correction primarily affects wet-period responses and does not modify the lower envelope of the soil moisture signal. This confirms that the procedure enhances event responsiveness without artificially shifting long-term trends."

Correspondingly we add the following text in the discussion“The mlhrsm downscaling model relies on a pretrained quantile random forest trained at the continental scale using Sentinel-1 backscatter, Landsat optical/thermal products, DEM-derived terrain metrics, and soil texture and bulk density from POLARIS. Because the model is not locally retrained for this watershed, we evaluated its regional appropriateness by examining (1) consistency of predicted soil moisture ranges relative to known soil hydraulic properties, (2) temporal responsiveness to precipitation events, and (3) agreement in relative wetness ranking across terrain units.
The downscaled product reproduced regional-scale temporal variability and captured soil moisture increases following rainfall; however, it underestimated the magnitude of wetting events and the upper bounds of soil moisture relative to expected porosity values and in-situ observations.  The bias-correction procedure applied in this study enhanced event-scale responsiveness while preserving the overall temporal structure from the downscaling algorithm.”

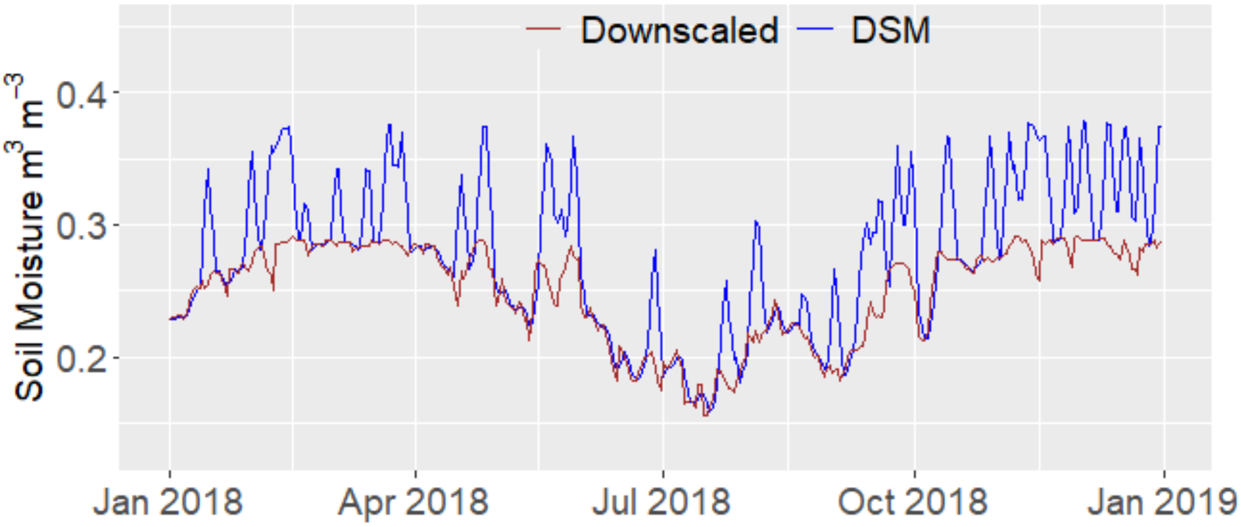

Figure X: Time series of downscaled (mlrhsm product) and downscaled-bias-corrected soil moisture (DSM) at the watershed scale. Only data from one year are shown for brevity and clarity.

4. A key concern relates to the representativeness of the soil moisture observations. The 25 soil moisture sensors are located within a 4.2 ha pasture, whereas the modeled watershed covers approximately 14 km². It is unclear whether such a limited area is sufficient for evaluating model performance at the watershed scale.

RESPONSE: Thank you for raising this important point regarding the representativeness of the field-scale soil-moisture observations. We have expanded the manuscript to clearly explain why the 4.2-ha monitoring field provides a hydrologically meaningful basis for evaluating model estimates in a 14.5-km² saturation-excess watershed. Specifically, we now emphasize that (i) soil moisture patterns in saturation-excess systems are governed primarily by topographic controls, as demonstrated in previous work (Easton et al., 2008; Lyon et al., 2004), (ii) SWAT-VSA uses topographic index (TI)–based HRUs, ensuring consistent representation of topographic wetness classes across the watershed and the monitored field, and (iii) all three TI classes present in the watershed are also represented within the field, allowing field-scale observations to capture the full range of hydrologic response types. We also clarify that model outputs were spatially extracted to the exact field boundary and aggregated using TI-weighted HRU area fractions, ensuring that evaluation is carried out at comparable functional units. Lastly, the small watershed size supports the assumption of relatively uniform precipitation and evapotranspiration inputs. These clarifications are now included in the manuscript:

“The soil-moisture monitoring plot covers 4.2 ha within the 14.5 km² watershed. Although this area is small relative to the watershed, it is hydrologically representative because the study watershed is saturation-excess dominated, with soil-moisture patterns strongly governed by topographic control rather than localized soil heterogeneity or microclimatic variation. Previous studies have shown that in such systems, soil moisture systematically varies along terrain-controlled gradients of wetness and hydrologic connectivity (Easton et al., 2008; Lyon et al., 2004). Consistent with this behavior, SWAT-VSA defines HRUs using topographic index (TI) classes, which capture relative wetness potential and runoff generation propensity.

In this study, the watershed was characterized by three TI classes, and all three classes are present within the monitored field, ensuring that the field observations span the full spectrum of hydrologic response types represented in the watershed. To ensure strict spatial comparability, modeled soil-moisture outputs were extracted specifically for the HRUs intersecting the field boundary and then aggregated using TI-class/HRU area-weighted averages, yielding a field-scale modeled soil-moisture estimate directly aligned with the sampling footprint.

Although land-use and soil variability contribute to watershed-scale moisture patterns, the small size of the watershed supports the assumption of near-uniform precipitation and evapotranspiration inputs, reducing the likelihood that climatic gradients confound field-to-watershed comparisons. Furthermore, previous work in the study area shows that spatial soil-moisture differences are dominated by topographic effects rather than local soil variation (Asfaw et al., 2025).”

5. An additional concern is that the evaluation of soil moisture performance throughout the manuscript is based primarily on soil moisture data that have undergone both downscaling and bias correction by the authors, without independent validation. While this preprocessing may be necessary, it raises the question of how the reliability of the calibrated model can be independently assessed.

RESPONSE: Thank you for this important comment. We agree that calibration against a processed soil-moisture product requires independent validation to ensure that model performance is not an artifact of the downscaling or bias-correction procedure. We now clarify in the manuscript that all model evaluation was carried out using independent observations that were not used in any stage of the downscaling or bias-correction process. Specifically: In-situ soil moisture measurements (0–120 mm, 25 points across a 4.2-ha field) were used solely for independent evaluation, not for calibration or for informing σ, k, or d in the bias-correction method. Daily streamflow observations were likewise completely independent of the satellite soil-moisture data and serve as an additional validation for the soil-moisture-calibrated model. The combination of these two independent datasets provides two orthogonal validation pathways, allowing us to determine whether statistical improvements during calibration reflect true hydrologic skill rather than agreement with the processed soil-moisture product.

These clarifications have been added to the manuscript:

“Although calibration uses the downscaled and bias-corrected satellite soil-moisture series, all model evaluation is performed using independent observations not involved in the preprocessing or calibration steps. In particular, in-situ soil moisture measurements collected at 25 locations within the 4.2-ha field (0–120 mm depth) were used exclusively for model evaluation, providing an independent benchmark for field-scale soil-moisture dynamics. These in-situ data were not used to compute the downscaling uncertainty (σ), the bias-correction factor (k), or the threshold precipitation depth (d), ensuring that no circularity is introduced into the calibration process.

In addition, independent streamflow observations were used to evaluate the hydrologic performance of the soil-moisture-calibrated models, providing a second, process-level assessment of model reliability that does not depend on the satellite product. Together, these two independent validation datasets allow us to determine whether improvements in soil-moisture calibration reflect genuine gains in hydrologic realism rather than improved agreement with a processed calibration target."

The manuscript would benefit from a **clearer explanation** of how the authors **ensure** that the improved **soil moisture performance reflects genuine model skill** rather than agreement with a processed target dataset. In particular, the role of independent in-situ soil moisture observations in validating the calibration results should be more explicitly discussed, and the potential circularity introduced by calibrating and evaluating against processed soil moisture data should be acknowledged.

RESPONSE: Although it is a common practice to use split sample calibration and evaluation in hydrology, i.e. a portion of a dataset is used for model calibration and the other portion for model evaluation. As rightly noted, since the soil moisture data is not a direct observation for the spatial scale in this study and that it had undergone several forms and rounds of processing, we explicitly used independently collected in-situ soil moisture data for model evaluation.

The primary model performance evaluation comes from the use of independent measurements, i.e. independent streamflow and in-situ soil moisture measurements. We  provide a clearer explanation in a revised Methods subsection ("Bias-Correction Rationale, Procedure, and Non-circularity") detailing exactly which data informed the correction, the physical basis for the precipitation threshold d why scaling by the downscaler-reported σ is appropriate, and an explicit non-circularity statement clarifying that in-situ measurements were not used to compute σ, k, or d or to tune the trigger:

"2.4.3 Bias-Correction Rationale and Procedure

We applied a parsimonious event-focused bias correction to the watershed-average downscaled soil moisture series to better represent short-term wetting after effective precipitation while preserving longer-term temporal structure. The correction is triggered only when daily effective precipitation exceeds a physically motivated threshold 5 mm, representing the depth required to fill the upper soil storage (from field capacity to near saturation) over the top 0–50 mm. When triggered, the downscaled mean is nudged by a multiple k of the model-reported uncertainty (the average σ, i.e., the standard deviation around the downscaled mean returned by the mlhrsm quantile model), thereby widening the upper tail without altering the dry-end envelope or mean seasonal cycle.

The d parameter approximates the effective rainfall needed to raise the top 0–50 mm from field capacity to near saturation, i.e., the short-term storage deficit that must be filled before the landscape expresses a saturation-excess response. This is consistent with the observed hydrology of TI-organized, saturation-excess systems, where event-scale runoff and rapid wetting require overcoming shallow storage first.

The mlhrsm downscaler provides an uncertainty spread around its mean estimate. Empirically, the uncorrected downscaled series under-represents peak wetness and event responsiveness. Scaling by σ (with a constant k) increases the dynamic range only when events are hydrologically plausible, thus correcting the documented attenuation of peaks without shifting dry conditions or the long-term mean.

In-situ soil moisture measurements were not used to compute σ, k or d nor to tune the trigger condition. In-situ data are used only (i) to independently evaluate model skill and (ii) qualitatively confirm that the corrected series exhibits a realistic peak range consistent with known soil hydraulic bounds. Parameter d is derived from soil hydraulic properties (porosity, field capacity) and target depth; σ is provided by the downscaling model; k is set a priori for parsimony and not fit to in-situ data."

We have also added a section on soil moisture layer compatibility among products to further clarify:

"To ensure a like-for-like comparison across data sources, we aligned depths for (i) the downscaled satellite soil moisture, (ii) SWAT-VSA soil-layer outputs, and (iii) in-situ observations (TDR). The SWAT soil profile was modified to output 0–50 mm and 50–120 mm layers, which we use to construct (a) a surface layer comparable to the satellite effective sensing depth and (b) a near-surface composite comparable to the field sampling depth. The downscaled product represents surface soil moisture (0-50mm). We therefore select the 0–50 mm model layer for calibration to the downscaled satellite series and reserve the 0–120 mm (composite from 0–50 and 50–120 mm) for evaluation against field measurements.

The downscaling (mlhrsm) preserves the depth of the surface-moisture product. Consequently, the downscaled time series is treated as an upper-layer signal; it is not intended to represent 0–120 mm conditions. The in-situ 0–120 mm measurements are used only for independent evaluation of field-scale soil moisture; the satellite-calibrated layer remains the 0–50 mm model output."

We intend to add a table like this to the Appendix.

| Source | Nominal / effective depth represented | How we aggregate to watershed/field | Model layer used for comparison | Role in this study |
|---|---|---|---|---|
| Downscaled satellite SM (mlhrsm from Enhanced SMAP) | Surface layer ( 0–50 mm ) | Average all 500 m pixels intersecting the watershed to a daily watershed mean | 0–50 mm SWAT layer | Calibration target in DSM and MO scenarios |
| SWAT-VSA model output | Discrete layers: 0–50 mm and 50–120 mm (plus standard deeper layers) | HRU outputs aggregated to watershed or field boundaries using TI-class/HRU area weights | 0–50 mm (for satellite comparison); 0–50 mm + 50–120 mm composite = 0–120 mm (for field comparison) | Calibration (0–50 mm) and Evaluation (0–120 mm) |
| In-situ TDR measurements (25 points) | 0–120 mm (field sampling depth) | TI-class-weighted average across the 4.2 ha pasture to obtain a field-scale mean | 0–120 mm model composite | Independent evaluation only |

Note: The 0–50 mm equivalent designation for the satellite product is an operational alignment for comparability; it does not imply that the satellite directly senses 50 mm uniformly. The purpose is to avoid mismatched depths during calibration and to keep field-scale evaluation at the measured 0–120 mm depth.

Finally, we have added more detail to the limitation section to address these points:

“While event responsiveness improves after correction, the method does not reconstruct sub-field spatial variability, nor does it alter longer-term biases inherited from the downscaler; thus, corrected series should be interpreted as temporally enhanced rather than fully bias-free.

6. In addition, the authors should clarify why downscaling to 500 m resolution is necessary if the soil moisture data are ultimately averaged to the watershed scale for calibration. It should be discussed whether this averaging undermines the advantages of using a spatially explicit, VSA-based distributed model.

RESPONSE: Thank you for raising this question. We have clarified in the manuscript why downscaling to 500 m remains valuable even though the soil-moisture series used for calibration is ultimately aggregated to the watershed scale. First, the native SMAP product has a 10-km footprint, close to an order of magnitude larger than our 14.5-km² watershed and therefore contains additional spatial variability from outside the target watershed. Without downscaling, the watershed-average series would inherit the spatial aggregation over the 10-km product and might not be representative of the specific conditions in the watershed.

Second, although the calibration uses the watershed-mean soil-moisture time series, SWAT-VSA uses TI-based HRUs internally to simulate spatially organized saturation patterns. The downscaled product does not need to match spatial patterns at 500 m resolution to be hydrologically useful; what matters is that the temporal signal driving calibration has realistic variability and amplitude. Finally, the 500 m resolution reflects the options available in the mlhrsm tool, which does not currently allow user-specified target resolutions.

These points are now clearly articulated in the revised manuscript:

“Although the soil-moisture data used for calibration were averaged to the watershed scale, downscaling the SMAP product to 500 m was still necessary because the native 10-km Enhanced SMAP footprint is far larger than the 14.5-km² watershed. The 500 m downscaled product contains daily time series constrained to the watershed boundary. These spatio-temporal improvements are essential for constraining soil-moisture-related model parameters at the watershed scale.

The choice of 500 m resolution was also constrained by the available pretrained models in the mlhrsm downscaling package, which does not currently provide alternative target resolutions. Importantly, spatial averaging of the downscaled soil-moisture fields does not undermine the value of using the SWAT-VSA model. SWAT-VSA relies on topographic-index-based HRUs to represent spatial patterns of saturation and runoff generation. While calibration uses a watershed-scale soil-moisture time series, the model internally maintains spatially explicit hydrologic processes using TI classes. HRU-weighted aggregation ensures that model-predicted soil moisture is scaled to the watershed or field boundary in a manner consistent with the model’s internal hydrologic structure.

Thus, the downscaled soil-moisture product enhances the temporal quality of the calibration target, while the SWAT-VSA HRU structure preserves the model’s spatial realism. The two components are complementary rather than redundant.”

7. The Discussion section could be streamlined by reducing repetitive literature comparisons and would benefit from a more critical assessment of which aspects of **soil moisture and streamflow dynamics remain poorly captured.** The authors should discuss whether these limitations reflect structural constraints of the SWAT model (e.g., simplified representation of unsaturated flow processes), and **whether the observed performance gains justify the additional complexity introduced by downscaling, bias correction, and multi-objective calibration**. A clearer distinction between statistical improvement and process-level improvement would strengthen the interpretation of the results.

RESPONSE: We have made an effort to incorporate suggested improvements to elucidate the performance gains in process-level improvements:

“Despite improvements in statistical performance, important aspects of the soil-moisture and streamflow dynamics remain imperfectly captured, and these patterns highlight structural limitations within SWAT-VSA as well as trade-offs among the different calibration strategies. Across all models, peak flows were underestimated, and the soil-moisture-calibrated (DSM) model showed a noticeably slower recession after storm events relative to the streamflow-calibrated (SF) and multi-objective-calibrated (MO) models. The MO and SF models captured the recession limb more realistically, whereas the DSM model tended to overestimate low-flow conditions, as illustrated by the flow-duration curve (Figure X), These patterns indicate that while soil-moisture-only calibration can improve the timing of near-surface wetting, it does not necessarily improve simulation of lateral redistribution or drainage processes that govern low-flow and recession behavior.

Water-balance diagnostics reinforce this distinction. The DSM model produced 13–16% lower ET and higher total streamflow than the SF and MO models, primarily because the DSM calibration converged on significantly lower AWC values. Lower AWC increases the proportion of precipitation available for quick runoff and lateral flow rather than storage and evapotranspiration. This reflects a compensation pattern rather than improved representation of the true hydrologic processes: the model reduces AWC to match the temporal variability of the satellite-informed soil-moisture target, but at the cost of producing unrealistic moisture storage and ET dynamics. By contrast, the SF and MO calibrations selected higher AWC and more moderate ESCO/EPCO values, resulting in ET fractions that are more consistent with regional expectations for humid, temperate watersheds (approximately 60% of annual water balance; Sanford et al., 2012).

These results underscore a key point: not all apparent improvements in statistical fit correspond to improved hydrological realism. The DSM calibration improves the variance structure and temporal alignment of surface soil moisture, a statistical improvement, but the associated ET underestimation, unrealistic dryness biases, and excessive partitioning into quickflow indicate that the internal process representation is less physically plausible. In contrast, the MO model balances soil-moisture fit with realistic streamflow recession behavior and produces water-balance components closer to independent regional estimates, suggesting that its improvements reflect both statistical and process-level gains.

Some performance limitations reflect structural constraints of the SWAT model itself. SWAT’s unsaturated zone representation is vertically lumped within layers and does not explicitly simulate lateral soil-moisture redistribution among HRUs; excess water leaves an HRU and is routed directly to stream reaches. This structural feature restricts the model’s ability to represent cross-slope moisture exchange, a key mechanism in saturation-excess systems, and likely contributes to the DSM model’s unrealistic ET–runoff partitioning. Similarly, SWAT’s simplified treatment of evaporation (ESCO, EPCO) limits the model’s ability to adjust vertical moisture redistribution without inducing compensatory parameter behavior. These constraints help explain why soil-moisture-only calibration does not automatically produce hydrologically consistent behavior, even when surface soil-moisture fit is improved.

Finally, the added complexity introduced by satellite-soil moisture downscaling, bias correction, and multi-objective calibration is justified only to the extent that it produces demonstrable process-level improvements rather than solely statistical gains. The present results suggest that downscaling + bias correction substantially improves the temporal quality of the soil-moisture calibration target, that DSM calibration alone yields statistical improvements but reduced physical realism, and that MO calibration provides the strongest balance between statistical performance and process fidelity.

Therefore, while the added complexity of downscaling and multi-objective calibration is warranted in this small, saturation-excess watershed, caution is needed in generalizing these findings to other settings or assuming that improved satellite-soil-moisture fit automatically implies improved hydrologic realism."

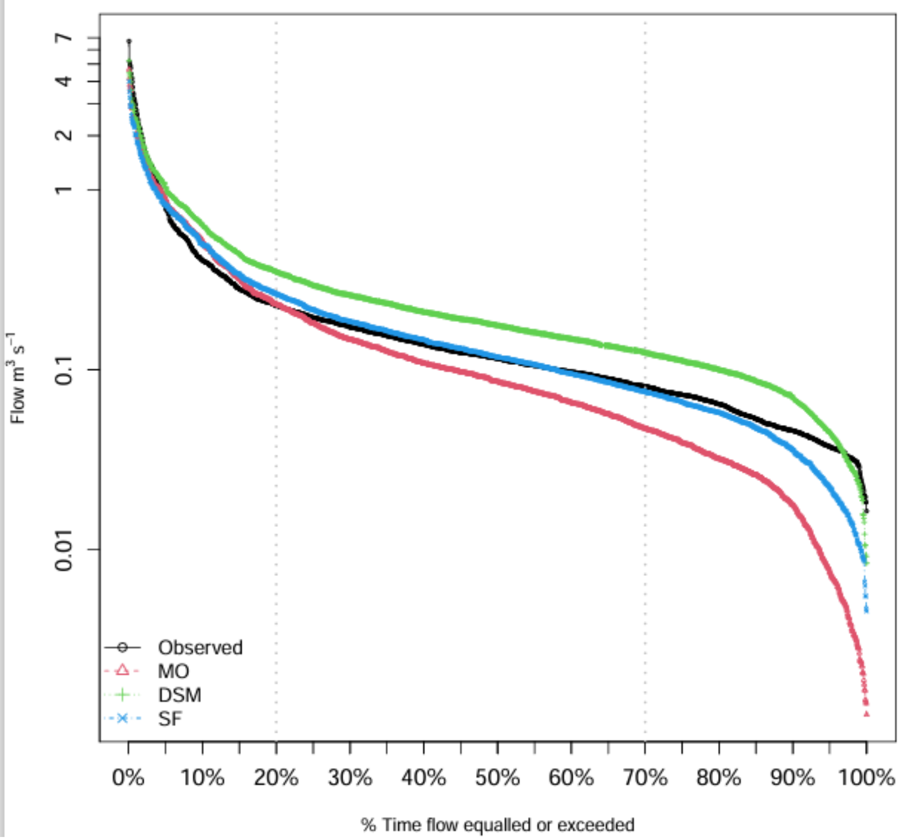

Figure X: Flow-duration curves comparing observed daily streamflow with simulations from the soil-moisture-only (DSM), streamflow-only (SF), and multi-objective (MO) calibration strategies. Differences among curves illustrate how calibration targets influence hydrologic realism: DSM captures surface wetting dynamics, but underestimates recession flows and high-flow peaks, whereas MO and SF better reproduce the observed distribution across both high-frequency (low flows) and low-frequency (peak flow) events.

Minor comments

1. Figures in the Appendix (e.g., Fig. A1 and Fig. A2) are essential for understanding the workflow and study area and should be moved to the main text.

RESPONSE: Figures have been moved.

2. The description in Section 2.2 and Fig. A2 may give the misleading impression that the study area is the full 57 km² Stroubles Creek watershed rather than the upper ~14.5 km² watershed.

RESPONSE: The description is updated as:

“Stroubles Creek is a 19.5 km long perennial stream in the New River Valley, VA, Fig. A2. In this study, the watershed outlet, 800 26’ 47’’ W and 370 12’ 10’’ N, is selected to define the study area. This outlet is about 1 km downstream of the flow monitoring station where the stream drains a 17.1 $km^2$ area.

The Virginia Tech StREAM Lab monitoring station, located at 80°26’ 42’’ W and 37°12’ 37’’ N, was the stream flow data source; at this monitoring station the creek drains the upper 14.5 $km^2$ watershed.”

3. The number of calibration iterations is reported in Section 2.5.4 (Parameter Uncertainty) but should be clearly stated in the Calibration Strategy section.

RESPONSE: We have edited as:

“The algorithm was implemented using the ‘DEoptim’ R package (Mullen et al., 2011). DEoptim was set up to evaluate 13 parameter vectors in each iteration for a total of 50 iterations per calibration. The number of iterations was decided sufficient after monitoring the successive diminishing returns in objective function minimization.”

4. The sensitivity analysis period (2015–2019) differs from the calibration and evaluation periods without sufficient justification. In addition, results in Fig. A3 are shown only for 2018, which requires explanation.

RESPONSE: We have restricted the results shown in Fig. A3 to 2018 for visual clarity. Similar results were obtained for the other simulated years, although the results were not shown for brevity.

The relative sensitivity values were calculated for model simulation from January 1, 2015, through December 31, 2019. This period adequately covers the range of streamflow variability and watershed responses, including during low-flow and high-flow periods (Figure 1), as well as the average soil moisture variability across the seasons.

5. Figure 1 and Figure 4 do not clearly convey differences among the three calibration strategies and could be improved for clarity.

RESPONSE: To improve clarity, the time span on the x-axis is now limited to one year. The full results are presented in the appendix for reference.

Citation: https://doi.org/10.5194/egusphere-2025-5813-RC1