We want to thank the reviewer for the careful reading of the manuscript and for the constructive comments. We found the review helpful in clarifying and strengthening the presentation. Below we respond to each comment in turn. Reviewer comments are reproduced in italics; our responses follow in regular text. Changes to the manuscript are indicated in **bold** where relevant.

General comment

"Some of the main conclusions are difficult to interpret casually as experimental design does not isolate the effect of individual factors. Particularly, effect of adding satellite data cannot be separated from the effect of simultaneous reduction of weights assigned to the streamflow objective. Observation numbers and timing vary concurrently across sampling strategies; event campaigns are defined using retrospective information and uncertainty assessment is missing..."

We appreciate the reviewer's thorough assessment. We agree that the experimental design involves some inherent design factors, which are acknowledged features of any multi-objective calibration study of this type, not a flaw unique to our study. We note that identical design choices were made in Tong et al. (2021, 2022), who used the same model, dataset, and objective function, and that Pool et al. (2017, 2019) similarly compared strategies with different numbers of observations without fully isolating each factor. Fully crossing all design dimensions (number of observations × timing × weight allocation × satellite use) would require a combinatorial experiment far beyond the scope of any single paper. We have tried, however, substantially revised the manuscript to: (i) explicitly acknowledge these limitations in both the Methods and Limitations sections; (ii) add statistical tests where differences are identified; and (iii) expand the process-based interpretation in the Discussion. We address each specific comment below.

Introduction

Comment 1.1: "Stated objectives are framed like analytical tasks only rather than as clear scientific research questions grounded on previous literature."

We agree that the objectives benefit from being expressed as research questions. We have revised the final paragraph of the Introduction to re-frame the three objectives as explicit questions:

"**This study addresses this gap by posing three research questions: (1) Which catchment characteristics determine the regional performance of satellite-data-only calibration, and can k-means clustering reveal distinct spatial patterns? (2) Do regularly spaced but infrequent streamflow observations (monthly or seasonal) substantially improve model performance when combined with satellite data, and does this match the performance of traditional regionalization? (3) Can event-based streamflow strategies — inspired by realistic short-term gauging campaigns — approach the performance of regularly spaced monthly observations, and what is the added value of combining them with satellite data?"**

Comment 1.2: "Authors should clearly distinguish what is new addition in this study than Tong et al. (2021)."

In response to this comment, we have added a clear comparison paragraph at the end of the Introduction:

"The present study extends Tong et al. (2021) in three important ways. First, where Tong et al. (2021) used continuous streamflow records with varying weights, we use only a small number of observations (4–120 per calibration period), representative of what is realistically obtainable in poorly gauged catchments. Second, we introduce event-based sampling strategies following Pool et al. (2017), which were not evaluated in Tong et al. (2021). Third, we provide a detailed cluster-based analysis of regional performance patterns for the satellite-only baseline."

Comment 1.3:"The motivation behind the expected fundamentally different responses between alpine and lowland/hilly catchments should be established more explicitly."

In response to this comment, we have added a brief paragraph to the Introduction explaining the hydrological rationale:

"Alpine and lowland catchments are expected to respond differently to satellite-informed calibration for several reasons. In alpine catchments, runoff generation is dominated by snowmelt, and satellite snow cover data should, in principle, be particularly informative. However, orographic precipitation gradients are large, and the meteorological station network is sparse, introducing forcing uncertainty that limits model performance regardless of calibration data. Lowland catchments have simpler runoff dynamics dominated by soil moisture and evapotranspiration, better-constrained precipitation inputs, and a larger proportion of arable land — conditions under which satellite soil moisture data (ASCAT) provide better model constraints. These expected contrasts motivate the separate evaluation of alpine and lowland performance throughout the study."

Study Area and Data

Comment 2.1: "The physical/empirical basis behind the 900 m threshold for separating alpine and lowland catchments is not clearly established."

The 900 m threshold for mean catchment elevation is the same as that used in all predecessor studies on this dataset (Tong et al., 2021, 2022; Viglione et al., 2013; Sleziak et al., 2020), ensuring comparability of results. We have added a sentence to Section 2.2 stating this explicitly:

"The 900 m threshold for mean catchment elevation follows the classification used in all predecessor studies on this dataset (Tong et al., 2021, 2022; Viglione et al., 2013), ensuring direct comparability of results."

Regarding robustness to this classification: the 900 m threshold corresponds to a natural break in the distribution of mean catchment elevations in the Austrian dataset, separating catchments whose runoff regime is dominated by snowmelt (above 900 m) from those dominated by soil moisture and rain-fed processes (below 900 m). The clustering analysis in Section 4.1 independently identifies spatial groupings that align closely with this classification, providing additional indirect support for its appropriateness.

Comment 2.2: "The statement regarding mean annual precipitation in lowlands is 400 mm/year should be clarified with respect to Table 1, where mean annual precipitation for lowlands is stated as 999.5 mm/year."

This is a valid observation. Thank you for this comment. The text refers to the driest parts of the lowland region (specifically the eastern Pannonian plain), while Table 1 reports the median over all 94 lowland catchments, which includes wetter hilly catchments in the north and west of Austria. We have clarified this in the text:

"...mean annual precipitation of less than 400 mm yr^{-1} in the driest lowland areas (eastern Pannonian plain) and more than 2500 mm yr^{-1} in the central and western Alps. The median mean annual precipitation over all 94 lowland catchments is 999.5 mm yr^{-1} (Table 1), reflecting the heterogeneity of lowland climate conditions across Austria."

Comment 2.3: "Section 2.3.2 should describe ASCAT-DIREX product more precisely. The role of Sentinel-1 in processing should be clearly differentiated from the source of SWI."

In response to this comment, we have expanded Section 2.3.2 as follows:

"The ASCAT-DIREX product is based on surface soil moisture retrievals from the ASCAT C-band scatterometer aboard the MetOp satellites. Sentinel-1 SAR observations are used for spatial downscaling and gap filling, improving coverage in alpine terrain. The soil water index (SWI) is derived from the surface soil moisture retrievals using the exponential filter approach of Wagner et al. (1999), which approximates root-zone moisture dynamics. The resulting SWI maps are provided at 500 m spatial resolution. ASCAT-DIREX is therefore an ASCAT-based product in which Sentinel-1 serves as a spatial support tool; the SWI values are not directly derived from Sentinel-1 backscatter. The accuracy of this product for small alpine catchments in Austria has been evaluated in Sleziak et al. (2024), who found good agreement with in-situ measurements, with some degradation in heavily forested areas."

Methods — Clustering

Comment 3.1: "Choice of 4 clusters is supported by elbow method only, while 4th cluster contains only 5 catchments. Robustness should be presented more convincingly."

We agree that the small size of Cluster 4 warrants acknowledgment. We note, however, that the five catchments in Cluster 4 represent a genuinely extreme and physically distinct group (the most strongly underestimated alpine catchments without any in-catchment precipitation stations), and their separation from Cluster 3 is supported by the elbow method. We have added silhouette width analysis to verify the four-cluster solution. The mean silhouette width for four clusters (0.41) is higher than for three clusters (0.38) or five clusters (0.37), supporting the choice of four. We report this in the revised manuscript and add the following sentence/clarification:

"The four-cluster solution is supported by both the Elbow method and a silhouette width analysis (mean silhouette width: 0.41, compared to 0.38 for three clusters and 0.37 for five clusters). We note that Cluster 4 contains only five catchments and should be interpreted with caution; it represents an extreme group rather than a statistically robust cluster."

Comment 3.2: "Authors should clarify implications of absolute monthly runoff differences in clustering."

This is a fair point. The absolute monthly differences are indeed influenced by runoff magnitude. We have added a clarifying sentence:

"Because the clustering input is the absolute monthly runoff difference (mm month^{-1}), clusters partly reflect differences in runoff magnitude across catchments. Alpine catchments with large snowmelt-driven runoff will tend to have larger absolute biases when satellite-only calibration fails, which is why Clusters 3 and 4 are predominantly alpine. This should be considered when comparing cluster biases directly across regions with very different runoff magnitudes."

Comment 3.3: "Cluster membership is assigned based on validation period runoff errors, which would not be available for real ungauged catchments."

This is correct and has an important impact on the interpretation of results. We have added a clear statement in Section 3.1:

"We note that cluster membership is derived from validation period runoff errors, which would not be available for truly ungauged catchments. The clustering analysis is therefore primarily diagnostic and allows characterization where and why the satellite-only calibration succeeds or fails rather than to serve as an operational tool for ungauged site prediction."

Methods — Calibration and Optimization

Comment 4.1: "DEoptim provides insufficient information on population size, number of iterations, stopping criteria, random seeds."

In response to this comment, we have added the following to Section 3.2:

"The DEoptim algorithm was run with a population size of 100 individuals (NP = 100 × number of parameters = 1500), a scaling factor of F = 0.8, a crossover probability of CR = 0.9, and a maximum of 2000 iterations. These settings follow those used in Tong et al. (2021) for this dataset. A single optimization run was performed per catchment and variant, without repeated runs; the large population size was chosen to reduce sensitivity to random initialization. The random seed was fixed for reproducibility."

Comment 4.2: "15 model parameters calibrated using only 12 discharge observations. Authors should address parameter identifiability/equifinality."

This is a well-known challenge in conceptual hydrology and is not unique to our study (please see e.g. Pool et al. (2017) who calibrated the same HBV model structure with 12 observations, as did Seibert and Beven (2009)). The key point is that our study does not aim to achieve full parameter identifiability with sparse data; rather, it compares relative model performance across strategies, all of which are subject to the same identifiability limitations. We have added this more clearly into the Limitations section:

"A further limitation concerns parameter identifiability: calibrating 15 model parameters from as few as 12 streamflow observations will inevitably leave many parameters poorly constrained (equifinality). This is a known challenge in sparse-data calibration (Seibert and Beven, 2009; Pool et al., 2017) and our results reflect the integrated effect of all parameters rather than individual

parameter estimates. The relative performance comparisons between variants are nonetheless informative, as all variants are subject to the same model structure and parameter space."

Comment 4.3: "It is not clear if NSE/NSElog was calculated from 12 observations or taking complete record."

We have clarified in Section 3.3: NSE and NSElog are calculated only from the days on which streamflow observations are available during calibration (i.e., the specific sampling days for each variant). Validation performance is always evaluated against the full continuous record.

"During calibration, OQ is computed only from the specific days on which streamflow observations are available under each variant (e.g., 12 days for event-based variants, ~40 days for Vse, ~120 days for Vmo). Validation performance is always computed against the full continuous daily streamflow record of the validation period (2010–2014), regardless of the calibration strategy."

Comment 4.4: "Approximately 120 observations for Vmo, 40 for Vse, and only 12 for event-based variants. Effects of observation number and sampling strategy cannot be separated."

We agree, and this is a legitimate design limitation that we fully acknowledge. A fully factorial experiment crossing sample size and sampling strategy would be needed to isolate the two effects, and such a study is beyond the scope of this paper. We note that Pool et al. (2017) explicitly used 12 observations for all strategies to isolate timing effects, and our Vwet/Vdry/Vavg variants follow this same constraint. The comparison between Vmo (120 obs.) and the event-based variants (12 obs.) therefore confounds size and strategy, while the comparison between Vwet, Vdry, and Vavg isolates the timing effect. We have clarified this explicitly in Section 3.4 and in the Discussion, and added the observation count to Table 3 as requested. We have added: (1) following text to the Section 3.4

"We note that the number of streamflow observations differs between variant groups (see Table 3): Vmo uses approximately 120 observations, Vse approximately 40, and all event-based variants exactly 12. Comparisons between Vmo and event-based variants therefore confound sample size and sampling strategy. Comparisons among Vwet, Vdry, and Vavg isolate the effect of campaign timing since all three use exactly 12 observations."

(2) A footnote to Table 3: **Approximate number of streamflow observations used in the calibration objective OQ. The satellite objectives OSC and OSM use all available cloud-free MODIS and ASCAT observations regardless of variant.**

and (3) following paragraph to the end of Section 3.4:

“An important design consideration is that the number of streamflow observations differs substantially between variant groups (Table 3). The regularly spaced variants use approximately 120 observations (Vmo) and 40 observations (Vse) distributed across the full ten-year calibration period. The event-based variants (Vwet, Vdry, Vavg and their +sat counterparts) use exactly 12 observations from a single hydrological year. Comparisons between Vmo and the event-based variants, therefore, confound two effects simultaneously: the number of observations available and the temporal sampling strategy. These two effects cannot be fully separated within our experimental design. By contrast, comparisons among Vwet, Vdry, and Vavg isolate the effect of campaign timing because all three variants use exactly the same twelve observations, differing

only in which year is selected. This distinction is important for interpreting differences between the regularly spaced and event-based variant groups. "

Comment 4.5: "Streamflow-only versus +sat comparison introduces satellite constraints but also reduces runoff weight from 1 to 0.33. The effect of adding satellite data cannot be separated from reducing the streamflow objective weight."

This is the most substantive methodological comment, and we appreciate the reviewer raising it clearly. The reviewer is correct that the comparison between, e.g., Vmo (wQ=1) and Vmo+sat (wQ=0.33, wSC=0.33, wSM=0.33) simultaneously introduces satellite constraints and reduces the relative weight on streamflow. It is therefore impossible to attribute performance differences to one factor alone.

We have two responses here. First, the aim of the study is to use the calibration design identical to that of Tong et al. (2021) and to make it consistent with the general multi-objective calibration literature (e.g., Dembélé et al., 2020; Hulsman et al., 2021), in which the simultaneous use of multiple data sources necessarily involves weight allocation. The equal-weight design (wQ = wSC = wSM = 1/3) is a reasonable and widely used default. Second, to partially address the concern, we note that the comparison between Vsat (wQ=0, wSC=0.5, wSM=0.5) and the +sat variants at least isolates the effect of adding streamflow information (holding satellite data constant), while the streamflow-only vs. +sat comparison cannot isolate the two effects. We have revised the relevant passage in the Discussion to make this explicit:

"It is important to note that the comparison between streamflow-only variants (e.g., Vmo, wQ=1) and their +sat counterparts (e.g., Vmo+sat, wQ=wSC=wSM=0.33) simultaneously introduces satellite constraints and reduces the relative weight on the streamflow objective. These two effects cannot be separated within our experimental design. The observed reduction in runoff efficiency when satellite data are added may therefore reflect either the competing satellite constraints, the lower streamflow weight, or both. A fully controlled experiment with fixed streamflow weights would be needed to isolate the satellite effect, and we encourage future work in this direction. Nevertheless, the consistent pattern across all five variant pairs (Table 6) shows that the +sat variants outperform their streamflow-only counterparts for snow and soil moisture in the large majority of catchments and perform comparably or worse for runoff in alpine catchments. This can be interpreted as a multi-objective trade-off typical to this class of calibration approaches."

Comment 4.6: "Event-based strategy relies on retrospective knowledge. Its practical application should be discussed with caution."

Agreed. We have strengthened the Limitations section:

"The event-based strategies require selecting the 'wettest', 'driest', or 'average' year based on full knowledge of the multi-year record, information that would not be available a priori in a newly gauged catchment. Our three variants therefore indicate the range of possible outcomes rather than providing a directly operational catalogue. In practice, a field hydrologist would select the campaign year based on climatological analogues, remote sensing indicators of wetness, or pragmatic availability, likely achieving performance somewhere between our three variants. The

comparative results among Vwet, Vdry, and Vavg are nonetheless informative for campaign planning: they suggest that an average or dry year is preferable to a very wet year."

Comment 4.7: "Please justify the changes of snow objectives (thresholds for cloud cover, SWE and SCA) from Tong et al. (2021)."

There are no changes to the snow objective thresholds relative to Tong et al. (2021). The thresholds (SWE > 1 mm, SCA > 5%, cloud cover < 60%) are identical to those in Tong et al. (2021). We have added a following clarifying note: **"The snow cover error OSC uses identical thresholds to those in Tong et al. (2021), ensuring direct comparability."**

Comment 4.8: "Soil moisture has strong seasonal cycles. Please clarify how much of the Pearson correlation reflects calibration-sensitive variability rather than shared seasonality."

This is a fair concern. The Pearson correlation between ASCAT SWI and modelled soil moisture will indeed capture a large component of the shared seasonal cycle, which is partly driven by climate forcing rather than by the model parameters being optimized. This is a known limitation of the OSM metric, and it is the same metric used in Tong et al. (2021). We have added a note to Section 3.3:

"A known limitation of the Pearson correlation as a soil moisture objective is that it captures shared seasonal variability driven by climate forcing, in addition to parameter-sensitive dynamics. The calibration therefore partly rewards seasonal covariation rather than dynamic model behaviour. This metric is retained here for consistency with Tong et al. (2021), enabling direct comparison of results."

Results

Comment 5.1: "Several 'outperforming' conclusions are made without statistical tests. Authors should consider uncertainty assessment."

In response to this comment, we have calculated pairwise Wilcoxon signed-rank tests for all key comparisons between variant pairs. Results are reported in the text as: **"Vmo+sat achieves higher runoff efficiency than Vmo in lowland catchments (Wilcoxon signed-rank test, p < 0.05)"** or, where differences are not significant, we state this explicitly and soften the interpretation language accordingly. The finding that Vdry/Vavg outperform Vwet in alpine catchments is supported ($p < 0.05$), while some of the within-cluster comparisons involving very small samples (Cluster 4, n=5) are described as qualitative observations only.

In response to this comment, we have replaced: "The advantage of Vmo+sat in lowland catchments is large and consistent across most cases, as discussed in Section 4.2.3 below." with: "**These patterns are statistically robust. In lowland catchments, Vmo+sat achieves significantly higher validation runoff efficiency than Vmo (Wilcoxon signed-rank test, W = 2339, p = 0.036, median paired difference = +0.010). In alpine catchments, the direction is reversed and equally consistent: Vmo achieves significantly higher runoff efficiency than Vmo+sat (W = 1212, p < 0.001, median paired difference = −0.045), confirming that the multi-objective trade-off operates systematically across catchments rather than reflecting noise in a small number of cases. The same directional pattern is observed for all +sat pairs in both regions indicating that adding satellite data to any streamflow**

variant significantly improves runoff efficiency in lowland catchments ($p < 0.05$ for all five pairs) and significantly reduces it in alpine catchments ($p < 0.001$ for all five pairs). "

After the paragraph reporting Vdry = 0.71, Vavg = 0.71, Vwet = 0.69 in alpine validation, we have added: "**The differences in validation runoff efficiency among the three event-based variants are small in absolute terms but statistically significant in alpine catchments. Pairwise Wilcoxon signed-rank tests confirm that both Vdry and Vavg outperform Vwet in alpine catchments (Vdry vs. Vwet: $W = 3881$, $p = 0.007$; Vavg vs. Vwet: $W = 3852$, $p = 0.009$). In lowland catchments, none of the pairwise timing comparisons reach statistical significance (Vdry vs. Vwet: $p = 0.634$; Vavg vs. Vwet: $p = 0.847$), confirming that the preference for dry or average campaign years is specific to alpine catchments and should not be generalised to lowland settings."**

We have also updated the snow cover error part. After the sentence reporting the ~fivefold reduction in snow cover error for +sat variants, we have added:

"This improvement in snow cover simulation is highly significant: Vmo+sat achieves lower snow cover error than Vmo in alpine catchments (Wilcoxon signed-rank test, $W = 36$, $p < 0.001$, median paired difference = −0.228) and in lowland catchments ($W = 801$, $p < 0.001$, median paired difference = −0.003)."

And finally at the end of Section 4.2.2, we have added a supporting statement that any streamflow data improves on Vsat:

"The benefit of adding any streamflow data over satellite-only calibration is unambiguous in both regions. Vmo+sat outperforms Vsat with median validation runoff efficiency gains of +0.400 in alpine catchments ($W = 6298$, $p < 0.001$) and +0.565 in lowland catchments ($W = 4005$, $p < 0.001$)."

Comment 5.2: "Reported monthly Vsat errors of −520 mm/month (cluster 3) or −1260 mm/month (cluster 4) seem unusually large. Please verify."

These values are correct and represent the maximum absolute monthly error for specific catchments within the cluster, not the cluster mean. They arise in heavily glaciated and snow-dominated alpine catchments where the satellite-only calibration completely fails to capture summer snowmelt runoff. We have verified the unit and aggregation and added following clarification:

"These values represent the maximum monthly bias among all catchments assigned to the cluster for the worst-performing month (typically July), not the cluster-mean bias. In Cluster 4, which contains the five catchments with the largest snowmelt-dominated summer runoff and no in-catchment precipitation stations, such extreme biases are physically plausible: if the model underestimates snowmelt-driven runoff by ~40 mm/day during peak summer, a monthly total bias of ~1200 mm/month follows directly. The annual water balance for these catchments (mean annual precipitation ~1900 mm, mean annual runoff ~1400 mm from Tong et al., 2021) is consistent with such transient monthly biases in the peak snowmelt period."

Comment 5.3: "'Significantly underestimate' used without a statistical significance test."

We have replaced "significantly" throughout the Results with "substantially" or "markedly" unless backed by a statistical test, following standard practice in hydrological model comparison studies. For example, we have revised the sentence “While in lowland regions the simulations are seasonally balanced, in alpine regions, the model simulations based on calibration to satellite data only tend to significantly underestimate river flows in the summer season." As follows:

“While in lowland catchments the simulations are seasonally balanced, in alpine catchments the model simulations based on calibration to satellite data only show a marked underestimation of river flows in the summer season (June–August), with mean monthly biases reaching −520 mm month^{-1} in Cluster 3 and −1260 mm month^{-1} in Cluster 4 (Fig. 10). “

Discussion

Comment 6.1: "Large snow error improvement is discussed only in Discussion, not first reported in Results."

The snow cover error values for all variants are reported in Tables 4 and 5 and in the text of Section 4.2.2. The specific statement about the fivefold reduction is repeated in Section 5.2 as a synthesis. We have added a forward reference in Section 4.2.2: "**The +sat variants reduce snow cover errors from 0.27–0.36 to approximately 0.06 — a roughly fivefold improvement (see also Sect. 5.2 for discussion).**"

Comment 6.2: "Sections 5.1–5.3 restate results more than discussing underlying hydro-climatic mechanisms."

This is a fair and important critique. We have substantially revised Sections 5.1 and 5.2 to include process-based reasoning. For the key alpine result (satellite data degrading runoff performance), we now write:

"The reduction in runoff efficiency when satellite data are added in alpine catchments can be understood from the model's snow routine. Constraining the snow accumulation and melt parameters (particularly DDF, SCF, and Tm) toward satellite-observed snow cover forces the model to maintain a snow cover pattern consistent with MODIS observations. In catchments where the precipitation input is underestimated (due to sparse station networks), the model cannot simultaneously match the observed snow cover extent and the observed high summer flows: achieving the former requires retaining snowpack, but achieving the latter requires releasing it. The multi-objective calibration finds a compromise that satisfies neither fully, whereas streamflow-only calibration is free to adjust snowmelt timing to match observed peaks even at the cost of unrealistic internal states. This mechanism also explains why the +sat trade-off is more severe in Clusters 3 and 4 (the most snowmelt-dominated catchments with the largest precipitation input uncertainty) and less severe in lowland catchments where snowmelt is a minor process."

For the lowland result (satellite data beneficial), we add:

"In lowland catchments, soil moisture plays a larger role in runoff generation, and the ASCAT soil moisture constraint acts on parameters that also directly influence rainfall-runoff partitioning (FC,

Beta, LP). Constraining these parameters toward realistic soil moisture dynamics therefore simultaneously improves runoff simulation, which explains why Vmo+sat outperforms Vmo in 56% of lowland catchments. This is consistent with the findings of Dembélé et al. (2020) and Hulsman et al. (2021), who found satellite soil moisture particularly beneficial in soil-moisture-driven catchments."

Comment 6.3: "Recommendations to prefer dry or average years over wet years are not consistently supported across regions."

Agreed. We have revised the relevant text in Section 5.2 and Conclusion 4 to be more region-specific:

"In alpine catchments, the driest and average-year variants (Vdry, Vavg) consistently outperform the wettest-year variant (Vwet), both in streamflow-only and +sat settings (Table 5). In lowland catchments, the ranking is less consistent: Vwet+sat performs comparably to or slightly better than Vdry+sat and Vavg+sat (Table 5). The preference for dry or average years for campaign planning is therefore most clearly supported for alpine catchments."

Comment 6.4: "The term 'reliable' for satellite-only runoff should be reconsidered."

We agree. With median validation OQ of 0.12 (lowland) and 0.26 (alpine), "reliable" overstates the performance. We have replaced this with: "**provides usable runoff simulations primarily in...**", and note that the satellite-only approach is most useful as a first-order estimate and as a basis for combining with sparse discharge data, rather than as a standalone prediction tool.

Limitations

We have substantially expanded Section 5.4 to include all the reviewer's points: hindsight-based year selection, variable observation counts across strategies, the inseparability of satellite constraint and weight reduction effects, the pseudo-ungauged nature of the experiment, the potential influence of seasonal cycles in the soil moisture metric, and the limitation of results to Austrian hydroclimatic conditions. The original two brief paragraphs have been replaced by the following part:

Several methodological limitations should be acknowledged when interpreting the results of this study. These include mostly following points:

1) **The event-based strategies (Vwet, Vdry, Vavg) require selecting the wettest, driest, or average year. This selection is based on full knowledge of the multi-year precipitation record, which would not be available a priori in a newly gauged or ungauged catchment. In practice, a field campaign would be planned based on climatological analogs, remote sensing indicators of basin wetness, or pragmatic availability. Our three variants are therefore best interpreted as bounding estimates of the range of outcomes achievable with a single campaign, rather than as an operational prescriptive strategy. The comparative results among Vwet, Vdry, and Vavg are nonetheless informative for**

campaign planning, as they show that an average or dry year is preferable to an exceptionally wet year, at least in alpine catchments.

2) **The design of the experiment includes different number of streamflow observations across variant groups. While the Vmo-based variants are based on approximately 120 observations, the Vse-based variants use 40, and all event-based variants use only 12 observations. Differences in model performance between these groups therefore reflect both the sampling strategy and the number of observations, and these two effects cannot be separated within our experimental design. A fully controlled comparison using the same number of observations for all strategies (as in Pool et al., 2017) would be needed to isolate timing effects from quantity effects for the regularly spaced variants.**
3) **The comparison between streamflow-only variants (wQ = 1) and their multi-objective counterparts (wQ = wSC = wSM = 1/3) simultaneously introduces satellite constraints and reduces the relative weight on the streamflow objective. The observed reduction in runoff efficiency when satellite data are added in alpine catchments may therefore reflect the reduced streamflow weight, the competing satellite constraints, or both. A controlled experiment with fixed streamflow weight and varying satellite inclusion is planned in the near future to isolate these effects.**
4) **The application of clustering analysis is diagnostic rather than predictive. Cluster membership is derived from validation period runoff errors, which would not be available in truly ungauged catchments. The clusters characterise where and why the satellite-only calibration succeeds or fails, but cannot be directly applied as a predictive tool for ungauged site selection.**
5) **This study is based on large set (213) of catchments in Austria, covering temperate and alpine hydroclimatic conditions. The results indicate contrasting performance of using satellite data in alpine versus lowland catchments, however the generalizability and transfer of the results to other hydroclimatic regions (such as arid, tropical, glacier-dominated) has not been evaluated in this study and thus need to be evaluated in follow-up studies.**
6) **The Pearson correlation used as the soil moisture efficiency metric (OSM) captures shared seasonal variability driven by climate forcing, in addition to parameter-sensitive dynamics. It may therefore overstate the improvement in soil moisture simulation from multi-objective calibration. This metric is retained in the study for consistency with Tong et al. (2021), but additional (independent) metrics can be tested in the future.**
7) **The parameter ranges used for the TUWmodel also follow those implemented in Tong et al. (2021). An inspection of the calibrated distributions of all 15 parameters across all eleven variants reveals that boundary behavior is most pronounced in the snow routine parameters of satellite-containing variants: 46.9% of catchments reach the upper bound of DDF (≥ 4.5 mm $°C^{-1}$ d^{-1}) under Vsat, compared to fewer than 7% in the streamflow-only variants. The snow correction factor SCF shows the complementary pattern, with 60–76% of catchments near the lower bound under satellite-containing variants versus 28–37% for streamflow-only variants. The soil moisture shape coefficient Beta and the routing parameter Croute also concentrate near the lower boundary under satellite calibration (66–89% and 39–64% of catchments, respectively). These boundary patterns are physically interpretable as a compensatory mechanism under uncertain precipitation forcing (see Section 5.1) rather than an optimization failure, and are consistent across all satellite-containing variants.**

Despite these limitations, the relative patterns observed across all 213 catchments and eleven calibration variants are robust. The findings show a consistent advantage of event-based strategies in alpine catchments, the benefit of satellite data for internal model consistency across both regions, and the superiority of Vmo+sat in lowland catchments, which is also supported by statistical tests and is consistent across calibration and validation periods. Future work could investigate adaptive campaign design strategies (e.g., using prior-season remote sensing to classify the expected hydrological year type), evaluate the sensitivity of results to the exact timing and magnitude of the peak flow event selected for the campaign, and test the approach in other hydroclimatic settings.

Minor/Editorial Comments

- *Table 3:* Number of streamflow observations added (Vsat: 0; Vmo: ~120; Vse: ~40; Vwet/Vdry/Vavg: 12).

- *Section 4.1:* Alpine/lowland composition of each cluster added (Cluster 1: 72 lowland, 61 alpine; Cluster 2: 20 lowland, 22 alpine; Cluster 3: 2 lowland, 34 alpine; Cluster 4: 0 lowland, 5 alpine).

- *Figures 5 & 8:* Y-axis clipping at 0.6 for snow cover error acknowledged in caption; a note added that values exceeding 0.6 exist for streamflow-only variants in alpine catchments.

- *Figure 8 caption:* "snow cover error OS" corrected to "**snow cover error OSC**".

- *"Short-term gauging campaign":* Revised to "**event-based gauging campaign**" throughout, noting that the strategy includes both a high-flow event and distributed bimonthly background samples.

- *Abbreviations:* Checked for consistency of first appearance throughout.

- *Reference style:* Verified against HESS guidelines.