We want to thank the reviewer for the positive assessment 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.

Comment R2.1: "Why does constraining the model with snow-cover information not translate into improved streamflow simulations in snow-influenced alpine catchments?"

This is an excellent question and is now addressed in the expanded Discussion (see also Response to R1, Comment 6.2 above). The short answer is that in alpine catchments with sparse precipitation station networks, the model's forcing inputs are uncertain. Matching the observed MODIS snow cover under uncertain precipitation forcing trades off against matching the observed streamflow timing and magnitude. The satellite-only calibration error in alpine catchments is largely a forcing error, not a parameter error, and adding a snow cover constraint cannot correct a forcing error. A streamflow-only calibration compensates for the forcing error by adjusting runoff generation parameters, at the cost of unrealistic snow dynamics. The +sat variants find a compromise that partially corrects both but excels at neither.

Comment R2.2: "Is the deterioration in alpine runoff performance related to the degree of snow influence rather than simply the Alpine/lowland classification?"

This is an insightful suggestion. We have added an exploratory analysis relating the difference in runoff efficiency between Vmo and Vmo+sat to the fraction of annual precipitation falling as snow (estimated from the SPARTACUS dataset for each catchment). The results confirm that the performance degradation when adding satellite data is largest in catchments with the highest snow fraction, supporting the process-based interpretation. We report this in Section 5.1 (after the paragraph ending: "...delivers substantially better snow and soil moisture simulation in the large majority."): "**To investigate whether the degradation in runoff performance when satellite data are added is related to the degree of snow influence of individual catchments, we computed the Spearman rank correlation between the precipitation-weighted snow fraction (see Table 1; median 0.24 for alpine and 0.12 for lowland catchments) and the change in validation runoff efficiency ΔOQ = OQ(Vmo) − OQ(Vmo+sat) for each catchment. A positive ΔOQ means that adding satellite data degrades runoff performance. In alpine catchments, the correlation is strongly positive and highly significant (Spearman $\rho = 0.60$, $p < 0.001$, $n = 109$), indicating that the performance degradation when satellite data are added is substantially larger in more snow-dominated catchments. In lowland catchments, where snow fractions are uniformly low, the correlation is weak and not significant ($\rho = 0.14$, $p = 0.192$, $n = 87$). This result supports the process-based interpretation that the multi-objective trade-off between satellite snow constraints and runoff simulation is driven by the importance of snowmelt in the catchment water balance.**"

Comment R2.3:"The robustness of the internal consistency improvement should be discussed more carefully, as adding constraints to the calibration will inherently improve OSC and OSM."

The reviewer is correct that improving OSC and OSM by including them in the objective function is, to some extent, tautological. We have added the following explanation to Section 5.1:

"We note that the improvement in OSC and OSM for the +sat variants compared to streamflow-only variants is partly a direct consequence of including these metrics in the calibration objective. The more informative comparison is between the +sat variants and Vsat (which also includes OSC and OSM): the +sat variants add streamflow information on top of the satellite constraints, while Vsat uses satellite data only. The fact that all +sat variants achieve OSC and OSM values similar to Vsat confirms that the satellite constraints are maintained effectively even when a streamflow objective is added, without decreasing internal consistency. The validation period snow and soil moisture performance (Tables 4 and 5) further confirms that these improvements transfer to the out-of-sample period."

Comment R2.4: "L61/L83: Is high-flow focus also justified in a drought context?"

In response to this comment, we have added a sentence to the Introduction: **"While high-flow events are most informative for calibrating runoff-generation parameters (Pool et al., 2017), the combined NSE + NSElog objective function (Eq. 2) used here gives explicit weight to low and medium flows through the log-transformed term, which mitigates — though does not eliminate — the bias toward high-flow calibration in drought and low-flow periods."**

Comment R2.5: "Table 1: Fraction of annual precipitation falling as snow would be informative."

We have added a precipitation-weighted snow fraction to Table 1, computed directly from the SPARTACUS daily temperature and precipitation data used as model input. For each catchment, we calculated the fraction of total precipitation falling on days with area-weighted mean catchment temperature below 1°C (the lower bound of the TUWmodel rain/snow threshold parameter Tr, range 1–3°C). The median precipitation-weighted snow fraction is 0.24 for alpine catchments and 0.12 for lowland catchments, confirming the substantially greater importance of snow processes in the alpine group. We have updated the table footnote accordingly: "**Snow fraction is estimated as the fraction of total precipitation falling on days with area-weighted mean catchment air temperature below 1°C, computed from the SPARTACUS daily gridded dataset (Hiebl and Frei, 2016, 2018) for the period 2000–2014.**"

Comment R2.6: "L115: Investigate whether catchments with poorer performance are characterized by a sparser station network."

In response to this comment, we have revised the sentence "Interestingly, the best performance in cluster 1 is associated …" with "**The number of in-catchment precipitation stations (from the SPARTACUS network) differs across clusters. While all clusters have a median of 1 in-catchment station, the fraction of catchments with zero in-catchment stations increases systematically with cluster number: 24% in Cluster 1 (30 of 125 catchments), 33% in Cluster 2 (13 of 40), 35% in Cluster 3 (9 of 26), and 100% in Cluster 4 (5 of 5). All five catchments in Cluster 4 have no in-catchment precipitation stations, and their model forcing is derived entirely by extrapolating from stations located outside the catchment boundary. A Kruskal-Wallis test confirms that the distribution of station counts differs significantly across clusters (H = 12.67, df = 3, p = 0.005). Pairwise Wilcoxon tests show that Cluster 1 has significantly more in-catchment stations than Cluster 2 (W = 2935, p =**

0.045) and Cluster 4 (W = 550, p = 0.002). The comparison with Cluster 3 is not significant (W = 1946, p = 0.052), likely due to the small sample size of Cluster 3 (n = 26, all alpine). This pattern strongly suggests that not only parameter uncertainty but also uncertainty in orographic precipitation inputs contribute to the poor performance of the satellite-only model in Clusters 3 and 4. Adding even sparse streamflow observations can correct for forcing-induced biases in runoff, which partly explains why limited streamflow data provides larger improvements in these high-alpine clusters than in Cluster 1."

In response to this comment, we have also added the number of catchments in each cluster (after Figure 3 caption, a new sentence is added as follows): **"The alpine/lowland composition of each cluster is: Cluster 1 contains 54 alpine and 71 lowland catchments; Cluster 2 contains 24 alpine and 16 lowland catchments; Cluster 3 contains 26 alpine and 0 lowland catchments; and Cluster 4 contains 5 alpine and 0 lowland catchments (all alpine)."**

Comment R2.7: "L121: Why not use the newer MODIS version 6.1?"

The snow cover dataset used in this study was processed for version 6 products in Tong et al. (2020, 2021) and covers the full study period 2000–2014. Reprocessing to version 6.1 would require re-deriving all snow cover time series and recalibrating all 213 catchments across all eleven variants which is a substantial computational effort that is outside the scope of this revision. Version 6.1 introduced mainly minor improvements to cloud masking compared with version 6 and the differences for Austrian catchments are very small.

Comment R2.8: "Please provide uncertainty of the ASCAT-DIREX product."

Sleziak et al. (2024) evaluated the accuracy of ASCAT-DIREX for a small alpine catchment in Slovakia and found Pearson correlations with in-situ volumetric soil moisture of 0.55–0.72, depending on land cover and depth. We cite these figures in Section 2.3.2 and note that uncertainty in the satellite soil moisture product will propagate into OSM and may reduce the calibration benefit in areas with high retrieval uncertainty (dense forest, complex terrain).

Comment R2.9: "Section 3.1: Is clustering just a trick to aggregate and visualize results?"

In part, yes as we now explicitly state: the clustering is primarily diagnostic, providing a spatially coherent summary of where and why satellite-only calibration fails. It is not intended as a predictive tool for ungauged catchments, and we have clarified this.

Comment R2.10: "Table 2: How are parameter ranges chosen? Are calibrated parameters hitting the bounds?"

In response to this comment, we have analyzed the frequency of cases when parameters of different variants are close to the boundary of ranges. We have added this information into Section 3.2 (after the sentence "The model is calibrated from September 2000 to August 2010 ..": "**The parameter ranges follow those used in Tong et al. (2021) for the same dataset, ensuring direct comparability**

of results. An analysis of calibrated parameter distributions across all eleven variants is provided in Section 5.1."

And added the following evaluation into Section 5.1 (after the paragraph ending "...the benefit of satellite data for internal model consistency across both regions**."):**

"An inspection of the calibrated distributions of all 15 TUWmodel parameters across all eleven variants shows that the most noticeable pattern involves the degree-day melt factor DDF (par2; range 0–5 mm °C^{-1} d^{-1}). For the satellite-only variant (Vsat), DDF is strongly concentrated toward the upper boundary (median = 4.34 mm °C^{-1} d^{-1}; 46.9% of catchments near the upper bound), whereas for the streamflow-only variants DDF is substantially lower and centrally distributed (medians 1.8–2.2 mm °C^{-1} d^{-1}; fewer than 7% near the upper bound). The +sat variants are intermediate (medians 3.2–3.4, with 23–30% near the upper bound). An opposite but complementary pattern is found for the snow correction factor SCF (par1; range 0.9–1.5) where the satellite- containing variants push SCF toward the lower boundary (medians 0.91–0.93; 60–76% near the lower bound), while streamflow-only variants show more central values. Together, the simultaneous increase in DDF and decrease in SCF under satellite calibration reflects a compensatory mechanism: when precipitation inputs are underestimated in alpine catchments, the model can only match the observed MODIS snow cover extent by melting snow faster (high DDF) while reducing the snowfall correction (low SCF). The nonlinear runoff generation coefficient Beta (par8; range 0–20) shows equally strong polarisation. The satellite-containing variants yield very low Beta values (medians < 1.1; 66–89% near the lower bound), favouring rapid saturation-excess runoff, while streamflow-only variants show higher and more dispersed values (medians 3.1–5.3). The routing parameter Croute (par15; range 0–50 day^2 mm^{-1}) follows the same pattern, with satellite calibration producing substantially lower values. These systematic parameter differences confirm that the reduction in validation runoff efficiency when satellite data are added in alpine catchments is not a statistical artefact but reflects a change in the model's internal state, where the snow and soil moisture constraints impact the parameter set toward a fast-melt and rapid- saturation regime."

Comment R2.11:"L167: How is a snow pixel identified? MOD10A1 is a fractional snow product."

We have clarified in Section 2.3.1: **"MODIS NDSI values are converted to binary snow/no-snow using seasonally varying thresholds that differ by altitude and land cover class (Tong et al., 2020). The mapping approach, which accounts for the fractional nature of the NDSI product and reduces misclassification in forested areas, is described in detail in Tong et al. (2020)."**

Comment R2.12: "L171: Pearson correlation as calibration objective could lead to results biased in terms of Soil Moisture Content."

Correct, please see Response to R1, Comment 4.8. The Pearson correlation is sensitive to getting the temporal pattern right but is insensitive to mean bias in soil moisture levels. This is a known limitation of this metric and here we used for the consistency with Tong et al. (2021).

Comment R2.13: "Table 3: Are all available satellite observations used, or just one per month/year?"

All available satellite observations are used in the satellite objectives (OSC and OSM), subject to the cloud cover filter for snow (days with cloud cover < 60%) and data availability for ASCAT (from January 2007). The streamflow observations are the ones subsampled according to each variant. We have clarified this in a note to Table 3: "**Note: the streamflow observations are sampled as described. All cloud-free MODIS snow cover observations and all available ASCAT SWI observations are used in OSC and OSM regardless of variant.**"

Comment R2.14: "Fig. 2: Could the bad performance of Cluster 4 be due to its small size?"

The five catchments in Cluster 4 represent an extreme case, as they have the highest elevations, zero in-catchment precipitation stations, and the largest snowmelt-driven summer flows in the dataset. Their clustering together reflects real physical similarity. We have nonetheless added a note about the sample size: **"Cluster 4 should be interpreted with caution given the small sample (n=5); the patterns are consistent across all five catchments but cannot be generalized statistically."**

Comment R2.15: "Figure 3: Forest coverage, soil type, groundwater recharge, hydrological regime?"

The six physiographic attributes shown in Fig. 3 (area, elevation, precipitation, aridity, arable land, precipitation stations) were selected as those most relevant to the clustering drivers and most commonly reported in Austrian hydrology studies. Forest cover and soil type are correlated with elevation and arable land fraction and would not add substantially new information. Groundwater recharge is not a standard output of the TUWmodel. We have added this information into Section 4.1 (after the discussion of Figure 3):

“**The six physiographic attributes shown in Fig. 3 were selected as those most directly relevant to the clustering results and most commonly used in Austrian hydrology studies (Viglione et al., 2013; Tong et al., 2021). Forest cover and soil type are strongly correlated with mean catchment elevation and arable land fraction in Austria and the dominant land cover gradient runs from agricultural lowlands to forested mid-elevations to bare alpine terrain. Groundwater recharge and baseflow generation are implicitly represented through the model's runoff generation parameters (K1, K2, Cperc), but are not directly measured for these catchments and are not available as physiographic predictors in the dataset. The hydrological regime (snow-dominated versus rain-dominated) is captured by the combination of mean catchment elevation, mean annual precipitation, and aridity index already shown in Fig. 3.”**

Comment R2.16: "Fig. 10: Are Clusters 3 and 4 dominated by snowmelt? How does the hydrograph look?"

Yes, Clusters 3 and 4 are predominantly alpine catchments where runoff peaks from May to July, driven by snowmelt. The satellite-only model underestimates this peak because it fails to generate sufficient melt flux, which is consistent with the model receiving insufficient precipitation input in these catchments. We have added a sentence to Section 4.3: "**Clusters 3 and 4 are almost entirely composed of alpine catchments with strong snowmelt-driven summer peaks (June–July), and the negative monthly biases for Vsat during this period reflect a systematic underestimation of snowmelt-driven runoff, likely exacerbated by uncertain orographic precipitation forcing.**"

Comment R2.17: "Section 5.4 needs to be expanded."

Section 5.4 has been substantially expanded. For more details, please see the full response to reviewer#1 provided in the section Limitations.