Reviewer 2:

Review for manuscript entitled "Seasonal variation in landcover estimates reveals sensitivities and opportunities for environmental models"

In this manuscript, the authors investigate the difference of classifying a land-use land cover (LULC) dataset during the growing vs. non-growing season. They use the Dynamic world dataset for this purpose and find differences in classification of 5 to 10% area in built and tree classes. In other words, there is 5 to 10% more(less) area classified as built(trees) in the non-growing season compared to the growing season. Subsequently, they investigate the effect of this difference on water quantity and quality modelling. For water quality modelling, the authors use both a statistical approach and the hydrologic model SWAT. The latter is also used for water quantity modelling. The selected 37 basins are located in the Eastern US and SWAT is set up for two small scale basins. Overall, the authors find small differences when using the classification during the growing season and non-growing season. The largest difference is found when SWAT parameters calibrated on streamflow and nitrogen yield data using the growing season data are transferred to using the non-growing season data. However, a calibration of SWAT in another catchment shows that similar performance can be achieved using the growing season and non-growing season data. This highlights that model parameters do compensate for differences in LULC maps in SWAT, an aspect that is not discussed in the manuscript. While the topic of LULC classification is very relevant because land cover information is a fundamental requirement for environmental models, the conclusions of this study are very limited and not of general nature because of two reasons. First, the limited spatial coverage only focusing on small basins in the Eastern US. Second, and more importantly, the choice of methods, i.e., using the SWAT model that only accounts for the dominant land cover class (l. 79) within a computational unit. Accounting for dominant land cover is not state-of-the-art. For example, state-of-the-art LSMs like the Community land model v5 use a mosaic approach to represent different land cover types and plant functional types within a grid cell (see also comment on CLM below). These models are also able to have transient land cover, which would account for the differences between growing and non-growing season. SWAT also does not account for relationships between model parameters and land cover. Within hydrologic modelling, regionalization methods have been developed that relate model parameters to land cover, which reduces the reduction in performance when transferring parameters from one land cover map to another (Samaniego et al. 2010). For these reasons, I do not find the results of this study are substantial enough to be interesting to a wider readership and recommend rejection.

The comments below might help the authors to further improve their manuscript in the continuation of their work.

> Response: We thank the reviewer for the thorough evaluation and the constructive comments for which we are using to improve the manuscript. The feedback has certainly helped us clarify the context of our study, and we are grateful for the insights about how to make the work more of interest to wider readership. However, we respectfully disagree

with the reviewer about their main conclusions for recommending rejection, each of which we address further below.

Point #1: The reviewer stated above “This highlights that model parameters do compensate for differences in LULC maps in SWAT, an aspect that is not discussed in the manuscript.” In our third case, we found that calibrating parameters for each separate LULC instance could account for differences in model performance due to different LULC data (e.g., 12% more forest cover in growing season), but would lead to different optimized parameter values. We discuss that different parameterizations could cause the model to respond differently when a transient LULC or climate change is being simulated.

“*Within reason, using separate calibration for each season may compensate for these inconsistencies, but lead to different model parameter optimizations*.” (page 1, lines 22-23 of preprint)

“*It is critical to consider that the differences in parameter values create the potential for the models to respond differently to future changes in LULC or climate change due to variations in unmeasured water balance outputs (Myers et al., 2021a).*” (page 13, lines 273-275 of preprint)

Point #2: The reviewer stated above “First, the limited spatial coverage only focusing on small basins in the Eastern US.” Although it is ideal to run water quality and quantity models across numerous watersheds across the globe, most studies performed modeling in one or multiple watersheds given limited time and resources. One example is the Samaniego et al. (2010) reference the reviewer mentioned above, which used one model on one watershed to introduce important concepts to the modeling community and has widely informed further studies. Frankly speaking, use of 37 watersheds is rare in watershed studies. But we agree with the reviewer that our study is not free of limitations, and discuss this limitation openly in our conclusion.

“*With a limited geographic scope (e.g., temperate watersheds) and small sample of models, our work does not intend to show definitively when, where, or in what model configurations these sensitivities would occur, but that they are a possibility that modelers should be aware of.*” (page 14, lines 288-291 of preprint)

Our work informs the modeling community that we should consider potential sensitivities using the high spatiotemporal resolution LULC data that can have illogical seasonal classification inconsistencies. For instance, recent work used one instance of Dynamic World land cover data to bias correct a global hydrologic model to be used for forecasting hydrologic extremes and potential emergencies in underdeveloped countries

(Hales et al., 2023). If that correction were used with a different instance of LULC data from a different season, our findings show that there could potentially be illogical LULC discrepancies for the temperate eastern United States watersheds. This is particularly important when models are being used for such meaningful societal purposes.

Point #3: The reviewer stated "Second, and more importantly, the choice of methods, i.e., using the SWAT model that only accounts for the dominant land cover class (l. 79) within a computational unit." We explored two set-ups of the SWAT model subbasin aggregation in this study: the dominant land cover aggregate in Case #2, and the full Hydrologic Response Unit detail in Case #3. Both are approaches used in watershed modeling, the first for computational efficiency and simplicity, and the second for high attention to detail and higher complexity. Either way, the illogical land cover changes due to seasonal classification inconsistencies in mixed temperate urban-forested watersheds are still a meaningful consideration. Incorporating potentially illogical land cover changes into modeling inputs could have implications to HESS readers, particularly when it is not clear which seasonal changes are real and should have logical effects on model outputs, or which are merely illogical errors that should be corrected before influencing model outputs. We discuss this with our second and third cases:

"*The differences observed between models using Dynamic World LULC were due to the 9% increase in built areas in non-growing season Dynamic World 2016 data, which have more impervious surfaces, a higher runoff curve number, and generate proportionally more water and nutrient runoff than the forested areas which were classified during the growing season. This could be particularly problematic when using computationally more efficient SWAT models that assign subbasin conditions based on the dominant HRU, as a change in dominant LULC type in a watershed could result in different subbasin conditions in the model greater than the proportional change in LULC.*" (page 11, lines 235-240 in preprint)

"*The difference in forests of 12% of watershed area between growing and non-growing season Dynamic World 2016 data for Difficult Run is as large a difference as real changes in forests that have been found to cause these sensitivities in model parameters (Li et al., 2019), but was likely caused by classification variation rather than an actual cycle from trees to built area and back (Hermosilla et al., 2018).*" (page 13, lines 270-273 of preprint)

Point #4: Referring to CLM5, the reviewer stated above "These models are also able to have transient land cover, which would account for the differences between growing and non-growing season." SWAT also simulates seasonal differences in land characteristics such as real crop and vegetation cycles. However, it is implausible that a large portion of the watersheds we studied had forest lands developed into built area for the non-growing season, then have the built area removed and reforested for the growing season. Considering the impacts of illogical classification inconsistencies in land cover inputs over seasonal time scales makes our findings relevant to wider readership. We explain this in our abstract and discussion:

"*Non-growing season LULC had more built area and less tree cover than growing season data due to seasonal impacts on classifications rather than actual LULC changes (e.g., quick construction or succession). In mixed-LULC watersheds, seasonal LULC classification inconsistencies could lead to differences in model outputs depending on the LULC season used, such as an increase in watershed nitrogen yields simulated by the Soil and Water Assessment Tool.*" (page 1, lines 18-22 of preprint)

"*Actual on-the-ground changes from built LULC to other types, or from other LULC types to trees (e.g., succession), are not likely to be occurring within the short (seasonal) time interval between our LULC composites (Cai et al., 2014).*" (page 9, lines 188-190 of preprint)

Point #5: The reviewer stated above "SWAT also does not account for relationships between model parameters and land cover." To the contrary, a large number of SWAT parameters are adjusted to account for different LULC at the subbasin and HRU scales. For example, the runoff curve number parameter (CN2) that was very sensitive in our study is calculated individually based on land cover inputs for each hydrologic response unit, and optimized using a relative (i.e., %) adjustment. We thank the reviewer for identifying this need for clarity, and updated our Table 1 footnote to make this clearer to readers.

"*† A 'v' indicates that the original parameter from QSWAT was replaced by the calibrated value globally, in the same unit. An 'r' indicates that the original parameter was modified relatively, multiplying it regionally by 1 + the calibrated value (e.g. a value of -0.2 reduces the original parameter by 20%).*" (to be updated at page 7, lines 155-157 of preprint)

Point #6: The reviewer stated above "For these reasons, I do not find the results of this study are substantial enough to be interesting to a wider readership and recommend rejection." In addition to the responses above, there are further reasons that we do believe that our results are substantial and interesting to a wider readership. First, evaluation of LULC products at high spatiotemporal resolution is an important research need with vast societal implications (Radeloff et al., 2024). Our study is one step toward achieving this goal, providing guidance to the modeling community for using high spatiotemporal resolution data such as Dynamic World. Second, we acknowledge that there are many hydrologic models being used by HESS readers and that we used SWAT in our experiments, but SWAT is a very common hydrologic model and is pertinent to the modeling community. We updated our text to make these points clearer to readers.

"*Evaluation of LULC products at high spatiotemporal resolution is an important research need with vast societal implications (Radeloff et al., 2024).*" (to be added at page 2, line 48 of preprint)

"*SWAT is the most common water quality model globally (Fu et al., 2019) and has been used in over 6,000 peer-reviewed studies (https://www.card.iastate.edu/swat_articles/, accessed 7 January, 2024).*" (to be added at page 5, line 110 of preprint)

Major comments

Overall, the results and discussion section is very short, only 120 lines! The authors should discuss how the differences in LULC classification that they find here would impact simulations using other methods than the ones used here (i.e., SWAT). To my understanding, the reduction in performance presented in this manuscript using SWAT provide an upper bound and different methods (i.e., CLM5 and methods outlined in Samaniego et al. 2010) would actually be less sensitive to the difference in LULC classification.

Response: We thank the reviewer for this feedback to improve the discussion of our results in relation with other modeling approaches. We have now expanded on our results and discussion section to include more details about illogical land cover classifications that could affect other models (3.1) and future directions (3.5). Specific text for those sections is included in our response to Reviewer 1, aside from the paragraph below. Concluding that our results present an upper bound of different modeling approaches would be outside the scope of our experiment, so we suggest a future model intercomparison study. However, we expect that our findings would not be an upper bound because we used seasonal LULC composites, while models using individual instances of 5-day temporal resolution LULC would likely encounter more extreme variation in LULC estimates.

"*Illogical LULC changes between data from different seasons could be pertinent to models beyond our cases of regressions and SWAT in the eastern United States, such as models for which accurate parameterization of LULC processes is essential for simulating the impacts of climate change (Glotfelty et al., 2021). For instance, potential seasonal variation in LULC estimates should be a consideration were an updated LULC layer to be used for modeling approaches such as Hales et al. (2023), which bias corrected a global hydrologic model GEOGloWS for extreme event forecasting in underdeveloped regions using a single instance of Dynamic World data. Our findings show that there is the potential for discrepancies at least for temperate watersheds in the eastern United States if the season of LULC update were not accounted for. These* *illogical LULC changes could also be pertinent for models that can use a mosaic approach to represent spatial variability of LULC within coarser grid cells (e.g., CLM5; Lawrence et al., 2019). The mosaic approach assumes that land surface properties (e.g., water fluxes) are homogeneously related to the LULC type (Li et al., 2013; Qin et al., 2023), in which case an illogical conversion of 12% area from forest to other types (our Case #3 example) could carry forward into the models, and potentially impact water and energy flux estimates or parameterizations similar to an actual LULC change. For instance, deforestation has previously been shown to alter heat and carbon fluxes and ecosystem productivity in CLM5 (Marufah et al., 2021; Luo et al., 2023). Variability within input data sub-grids has also been shown to influence model parameter optimization and performance simulating hydrology, making it an important aspect to account for (Samaniego et al., 2010). As models advance into higher spatiotemporal resolution following increasing computational resources and data availability (e.g., Hales et al., 2023), we encourage the modeling community to be cognizant of the potential impacts of illogical seasonal LULC change, such as we identified for mixed LULC areas of the eastern United States. The strength of the effect of the illogical seasonal LULC change on the model outputs and optimized parameters would depend on many factors including model processes and spatiotemporal extent. A model intercomparison study in this regard would likely be a meaningful contribution to the advancement of the field into higher spatiotemporal capabilities.”* (to be added after page 13, line 275 of preprint)

Case \#2: Parameters transferred from growing to non-growing season lead to substantial drop in model performance, but how do parameters calibrated to non-growing season perform when using the growing season data? The authors should include this case in their analysis.

Response: Our primary goal for the second case was to use a traditional calibration approach to growing season data and evaluate what would happen when non-growing season data is used. Although performing the reverse would be an interesting question, we have not done this analysis because we are not aware of any modeling approach designed for calibration to non-growing season data only. However, analysis-ready datasets such as Dynamic World can lead to non-growing season LULC data being built into modeling approaches (e.g., Hales et al., 2023), so this should absolutely be an avenue for future work building on our study.

L. 242.: The reference to Clark et al. (2015) is misleading. State-of-the-art land-surface models like CLMv5 are able to account for transient land use and land cover change (see chapter 27 in CLM5.0 technical note, downloaded from https://www2.cesm.ucar.edu/models/cesm2/land/CLM50_Tech_Note.pdf). The authors need to discuss how their findings here are relevant for land-surface models that are employing a mosaic approach to represent different land and plant functional types within a grid cell.

> Response: We thank the reviewer for this valuable feedback and improvement of the conceptualization of our study. We removed the mentioned sentence and combined discussion of the mosaic approach with our discussion of other modeling approaches, the text of which is provided above (the response to the first of the reviewer's major comments, two comments up). To summarize, we expect that these findings are relevant because they are likely to be based on illogical classification inconsistencies rather than real differences in land cover within a grid cell. Thus, they are pertinent for having accurate and consistent land cover representation within the grid cells over time. Although models may incorporate transient land cover change to best represent vegetation and crop characteristics, they are nonetheless vulnerable to errors caused by illogical land cover change estimates, which our study found could comprise a large proportion of watershed area in the eastern United States using high spatiotemporal resolution Dynamic World data from different seasons.

Table 4: The performance metrics are often higher in the validation than in the calibration period. Normally, it is expected that there is a substantial drop in model performance between calibration and validation period. The authors should explain this.

> Response: We respectfully disagree with the statement "Normally, it is expected that there is a substantial drop in model performance between calibration and validation period." To our interpretations, a well-performing model should be able to simulate the validation period consistent with the calibration period. A drop in performance during the validation period could indicate that a model is being overfit to the calibration period. The consistent performance of our models during the validation period could imply that our parameterization approach for this case, calibrating only the most sensitive parameters and using multiple optimization algorithms (Vrugt and Robinson, 2007), helped reduce model overfitting to the calibration period.

Minor comments:
Section 2.5: Not all readers will be familiar with SWAT parameters listed in Table 1. It would be helpful to add a short description in the Appendix.

> Response: We made the improvement as suggested, adding a table of additional descriptions for SWAT parameters as Table S4 (see below).

Table 3: The caption should mention that all results presented here are based on parameters calibrated with the growing season LULC (see l. 213f), even the results shown for the calibration period of the non-growing season. This could be misunderstood otherwise.

Response: We made the improvement as suggested and thank the reviewer for the helpful suggestion.

"***Table 3:*** *Model performance metrics for the calibrated Rock Creek hydrologic model (Case #2) for streamflow and nitrogen yield, based on Nash Sutcliffe Efficiency (NSE), mean absolute error (MAE), and percent bias (PBIAS, where <0 implies overestimation bias), at the monthly time step. In this case, model parameters were all calibrated to growing season Dynamic World 2016 data to investigate the impacts of simulating an LULC change using non-growing season data."* (to be updated at page 11, lines 226-228 of preprint)

Fig. 6: The visual quality of this Figure is low. In Figure 6a to 6c, the markers in the Scatter plots are overlapping and the underlying relationship between the simulated and observed discharge cannot be seen. Also the dpi resolution of the hydrographs is low leading to pixelization that does not allow to appreciate differences between different lines. I also suggest to use a black line instead of a yellow line and make the line widths slightly thicker.

Response: We thank the reviewer for these comments to improve our figure. We have modified color palettes and point transparencies to make underlying relationships in 6a-6c more visible. We also provide a higher (600) dpi resolution and modified line widths and types so that the observed and modeled discharge time series are more distinguishable. We also enlarged the figure.

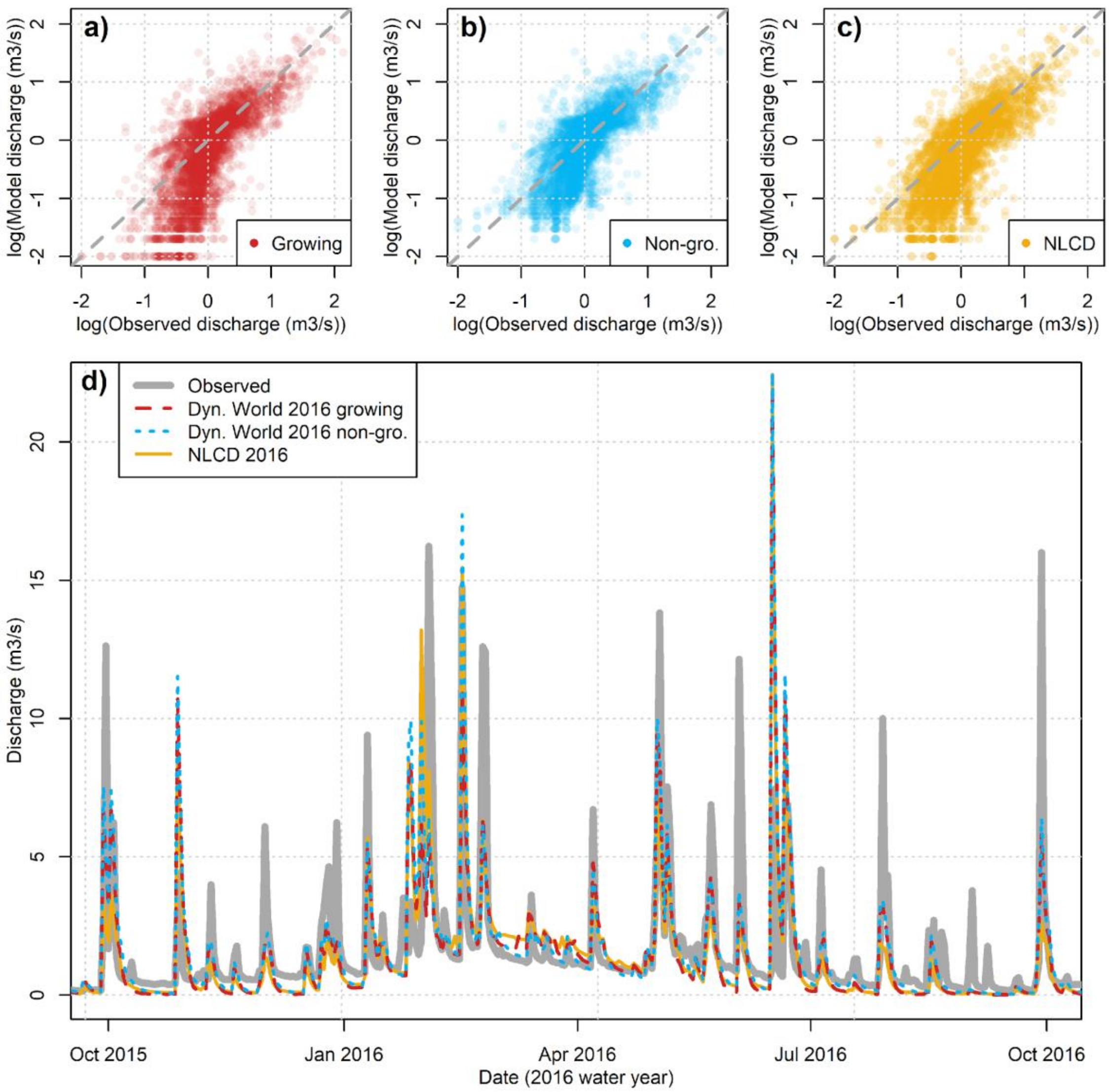

Figure 6: Daily discharge models for the Difficult Run Watershed displaying base-10 log, so that daily baseflows and low flows are visible, comparing independently calibrated models with a) Dynamic World 2016 growing season LULC, b) Dynamic World 2016 non-growing season LULC, and c) NLCD 2016. Also d) Time series of Difficult Run modeled discharge.

References:

Arnold, J. G., Kiniry, J. R., Srinivasan, R., Williams, J. R., Haney, E. B., and Neitsch, S. L.: Soil & Water Assessment Tool: Input/output documentation. version 2012, Texas Water Resources Institute, TR-439, 2013.

Cai, S., Liu, D., Sulla-Menashe, D., and Friedl, M. A.: Enhancing MODIS land cover product with a spatial-temporal modeling algorithm, Remote Sens Environ, 147, https://doi.org/10.1016/j.rse.2014.03.012, 2014.

Fu, B., Merritt, W. S., Croke, B. F. W., Weber, T. R., and Jakeman, A. J.: A review of catchment-scale water quality and erosion models and a synthesis of future prospects, Environmental Modelling & Software, 114, 75–97, https://doi.org/10.1016/J.ENVSOFT.2018.12.008, 2019.

Glotfelty, T., Ramírez-Mejía, D., Bowden, J., Ghilardi, A., and West, J. J.: Limitations of WRF land surface models for simulating land use and land cover change in Sub-Saharan Africa and development of an improved model (CLM-AF v. 1.0), Geosci Model Dev, 14, https://doi.org/10.5194/gmd-14-3215-2021, 2021.

Hales, R. C., Williams, G. P., Nelson, E. J., Sowby, R. B., Ames, D. P., and Lozano, J. L. S.: Bias Correcting Discharge Simulations from the GEOGloWS Global Hydrologic Model, J Hydrol (Amst), 130279, https://doi.org/10.1016/j.jhydrol.2023.130279, 2023.

Hermosilla, T., Wulder, M. A., White, J. C., Coops, N. C., and Hobart, G. W.: Disturbance-Informed Annual Land Cover Classification Maps of Canada's Forested Ecosystems for a 29-Year Landsat Time Series, Canadian Journal of Remote Sensing, 44, https://doi.org/10.1080/07038992.2018.1437719, 2018.

Lawrence, D. M., Fisher, R. A., Koven, C. D., et al.: The Community Land Model Version 5: Description of New Features, Benchmarking, and Impact of Forcing Uncertainty, J Adv Model Earth Syst, 11, https://doi.org/10.1029/2018MS001583, 2019.

Li, D., Bou-Zeid, E., Barlage, M., Chen, F., and Smith, J. A.: Development and evaluation of a mosaic approach in the WRF-Noah framework, Journal of Geophysical Research Atmospheres, 118, https://doi.org/10.1002/2013JD020657, 2013.

Li, Y., Chang, J., Luo, L., Wang, Y., Guo, A., Ma, F., and Fan, J.: Spatiotemporal impacts of land use land cover changes on hydrology from the mechanism perspective using SWAT model with time-varying parameters, Hydrology Research, 50, 244–261, https://doi.org/10.2166/NH.2018.006, 2019.

Luo, M., Li, F., Hao, D., Zhu, Q., Dashti, H., and Chen, M.: Uncertain spatial pattern of future land use and land cover change and its impacts on terrestrial carbon cycle over the Arctic–Boreal region of North America, Earths Future, 11, e2023EF003648, 2023.

Marufah, U., June, T., Faqih, A., Ali, A. A., Stiegler, C., and Knohl, A.: Implication of land use change to biogeophysical and biogeochemical processes in Jambi, Indonesia: Analysed using

CLM5, Terrestrial, Atmospheric and Oceanic Sciences, 32, https://doi.org/10.3319/TAO.2020.12.17.01, 2021.

Myers, D. T., Ficklin, D. L., Robeson, S. M., Neupane, R. P., Botero-Acosta, A., and Avellaneda, P. M.: Choosing an arbitrary calibration period for hydrologic models: How much does it influence water balance simulations?, Hydrol Process, 35, e14045, https://doi.org/10.1002/hyp.14045, 2021.

Qin, Y., Wang, D., Cao, Y., Cai, X., Liang, S., Beck, H. E., and Zeng, Z.: Sub-Grid Representation of Vegetation Cover in Land Surface Schemes Improves the Modeling of How Climate Responds to Deforestation, Geophys Res Lett, 50, https://doi.org/10.1029/2023GL104164, 2023.

Radeloff, V. C., Roy, D. P., Wulder, M. A., Anderson, M., Cook, B., Crawford, C. J., Friedl, M., Gao, F., Gorelick, N., and Hansen, M.: Need and vision for global medium-resolution Landsat and Sentinel-2 data products, Remote Sens Environ, 300, 113918, https://doi.org/10.1016/j.rse.2023.113918, 2024.

Samaniego, L., R. Kumar, and S. Attinger (2010), Multiscale parameter regionalization of a grid-based hydrologicmodel at the mesoscale,Water Resour. Res.,46, W05523, doi:10.1029/2008WR007327

Vrugt, J. A. and Robinson, B. A.: Improved evolutionary optimization from genetically adaptive multimethod search, Proceedings of the National Academy of Sciences, https://doi.org/10.1073/pnas.0610471104, 2007.

Table S4. Further explanations of SWAT model parameters mentioned in Table 1 of the text for the Difficult Run Watershed. Readers are directed to Arnold et al. (2013) for additional documentation.

Symbol	Definition	Description
CH_KII.rte	Channel hydraulic conductivity (mm/h) (v)	Effective hydraulic conductivity for alluvium in the main channel of the reach, describing relationships with groundwater.
ALPHA_BNK.rte	Bank flow recession constant (v)	Regulates bank flow to the reach using a recession constant.
CN_F.mgt	Runoff curve number (r)	Representation of soil permeability, landscape

		characteristics, and antecedent moisture conditions.
SNO50COV.bsn	Fraction of SNOCOVMX for 50% cover (v)	Fraction of complete snow cover which represents 50% snow cover in areal depletion curve.
ESCO.hru	Soil evaporation compensation coef. (v)	Modification to soil evaporative demand at different depths.
CH_NII.rte	Manning's n value for main channel (v)	Roughness coefficient for the main channel of the reach.
SOL_BD.sol	Soil moist bulk density (r)	Ratio of oven dry soil mass to total volume near field capacity.
SNOCOVMX.bsn	Snow depth above which is 100% cover (mm) (v)	Amount of snowpack needed for complete areal coverage.
SFTMP.bsn	Snowfall temperature threshold (°C) (v)	Temperature threshold that distinguishes snowfall from rainfall during a precipitation event.
SOL_AWC.sol	Available Water Capacity (r)	Plant available water capacity of the soil.