NOTE:

Reviewers' original comments are in Cambria type with yellow highlighting.

Text from the revised manuscript is in bold italic font.

Page/Line/Figure numbers given by reviewers were left unchanged in their text and refer to the originally submitted version of the manuscript

[Line XXX] indicates line number in the annotated manuscript

Reply to Reviewer #1

I. Summary of the most important scientific findings

(1) First, I would like to express that I very much like the subject of the paper and the research questions that have been investigated. The regionalization of hydrological model parameters based on catchments attributes has been a long history in the PUB initiative, and, over the last decade with the developments of deep learning tools gained some new momentum.

(2) This paper addresses this area of research, by using symbolic regression methods and combination with genetic programing, to not only map the (transfer) relationship of model parameters with physio-geographic properties of the catchments, but also make these relations interpretable in terms of concrete mathematical expressions that are derived.

(3) In this way, the paper provides a valuable approach towards the development of "explainable AI" in hydrology and eater management.

II. Novelty and contributions of this paper

(4) In detail, the paper investigates how the single parameter N of the Smooth-Minima-approach to perform base flow separation can be estimated and regionalized from tracer-based measurements and physio-geographical data for 855 catchments in the US.

(5) Optimal algebraic equations and parameters were estimated using a genetic programming approach and are compared to standard literature formulations demonstrating superior performance.

Thank you for this positive and encouraging summary of our work. We appreciate the reviewer's recognition of the relevance of the research question, the methodological contribution of symbolic regression, and the potential of the proposed approach for developing interpretable and explainable AI tools in hydrology.

III. Critical comments and recommendations to the editor

While I very much like the general scope of the paper as stated before, here are a number of critical questions, concerns or suggestions I would like to make and that are listed in the following:

(6) Why did you use an genetic programming approach? It is well known that this approach is limited to a few number of variables and levels of branches. You are generating a discrete optimization problem that is in my opinion very difficult to handle. In recent work by e.g. Feigl et al. (2020, cited by you) this is approached/solved by generation a grammar and "compressing" it with a VAE (and some constrains) into a latent space which is continuous and therefore usable for "gradient-based" search methods. Why didn't you use such an approach?

Thank you for this insightful and constructive comment. We used the relatively simple genetic programming approach instead of more complex grammar-based encoding

combined with variational autoencoders because the search space of our problem is restricted in both dimensionality and structural complexity. First, prior to symbolic regression, we applied a mutual information–based feature selection procedure to reduce the candidate predictors to nine relevant variables. Second, the set of mathematical operators was constrained to addition, subtraction, multiplication, division, power-law, and logarithm. Nested operators were not allowed except for addition and subtraction. The internal complexity of expressions within the power-law and logarithmic operators was limited to a maximum of 3 and the total expression complexity was capped at 20. These constraints substantially reduce the search space and mitigate the limitations of GP in highly complex symbolic structures. In addition, the GP-based symbolic regression is computationally manageable and have been successfully applied in other studies (Dorgo et al., 2021; Pumo and Noto, 2023; Razaq et al., 2016).

To assess whether the derived SR expressions depend on the choice of optimization methods, we implemented the reinforcement learning–guided search. The resulting expressions are: $A^{0.22} + 3.70$, $(A + 901)^{0.24} + {K_{sat}}^{0.28}$, $0.31 * (A + 549)^{0.31} + {K_{sat}}^{0.34} + 3.15 f_{SWE} + 1.84$, with R^2 values of 0.47, 0.52, and 0.55, respectively. These expressions exhibit high similarity and comparable predictive performance to those obtained using the GP-based approach. This consistency suggests that the identified SR expressions are robust and independent from the symbolic regression optimization methods.

(7) You used N=5 (FD) and N=1.6*A^0.2 (FPL) as benchmark models taken as typical parameterizations from the literature. I think that is ok to see the limitations and the lower end of performance. But why didn't you use that equation structure (a*A^b) and looked ho well it worked in comparison?

We appreciate this insightful comment. We tested a calibrated version of $N = a \cdot A^b$ as suggested. The calibrated power-law model achieves an R^2 of 0.42, lower than those of the SR-derived ones (F_1, F_2, and F_3 ranging from 0.48-0.55). This implies the limitation in the $N = a \cdot A^b$ function, highlighting the advantage of using SR framework for the prediction. We added the calibrated power-law results in Figure 4 and some descriptions in section 3.1, 4.2, and 4.3.

[Section 3.1, lines 168-171]

To further examine the explanatory capacity of the areal-based power-law relationship, we considered a calibrated form of $N = a \cdot A^b$ (denoted as F'_{PL}), where the coefficients a and b are estimated from the data by minimizing the squared error between the reference and predicted Ns.

[Section 4.2, lines 329-332]

The calibration with respect to SEC data improves the performance for F'_{PL} to median R^2 values of 0.43 and 0.42 for the training and testing sets, respectively, and R^2 of 0.42 for the combined testing set.

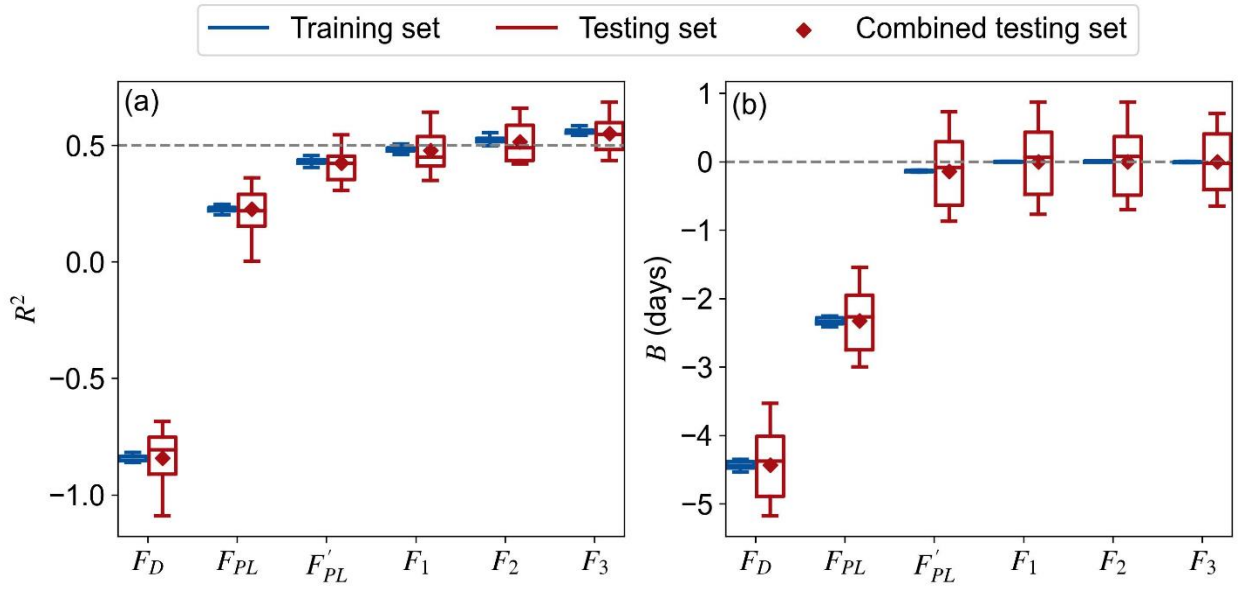


Figure 4. Performance of N predictions using the constant (F_D), power-law (F_{PL}), calibrated power-law (F'_{PL}), and the three SR formulas (F_1, F_2, and F_3) for the ten training and testing sets of the 10-fold cross-validation: coefficient of determination R^2 (a) and mean bias B (b).

[Section 4.3, lines 402-404]

The calibrated power-function (F'_{PL}) reveals improved performance to a similar level with the SR formulas except for the $A < 100$ km^2 bin, indicating the necessity regional calibration.

[Section 4.3, lines 412-416]

In the intermediate ranges (5–25 mm/hour), the two regionally fitted power-law formulas (F_1 and F'_{PL}) outperform the conventional power-law formula (F_{PL}). Although F_1 and F'_{PL} do not explicitly include K_{sat}, their performance remains comparable to F_2 and F_3, which incorporate K_{sat}. This indicates that regional fitting may compensate the effects of K_{sat} for small and medium values.

[Section 4.3, lines 423-425]

In contrast, for $f_{SWE} < 0.4$, F'_{PL}, F_1, and F_2 perform similarly to F_3, demonstrating that regionally calibrated coefficients can partially offset the lack of explicit consideration of snow-related variables.

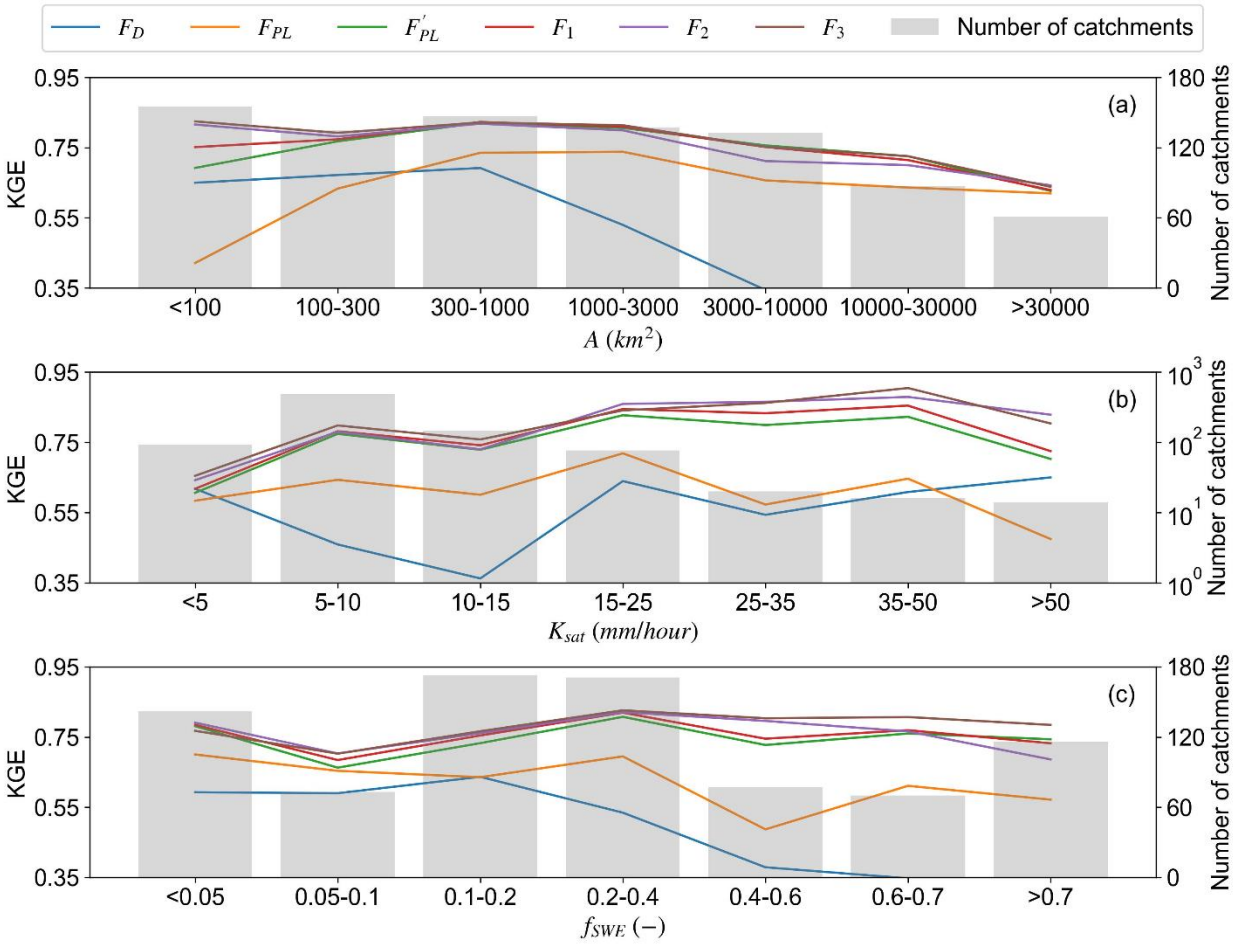


Figure 6. Performance of baseflow separation for different ranges of catchment area (a), catchment-averaged saturated hydraulic conductivity (b), and snow day fraction (c). The right y-axis for panel b is in logarithmic scale.

(8) For such benchmark experiments, we often use a ML approach (e.g. Xgboost) in order to analyze the "maximum" possible performance given the calibration data and catchment properties, to see how well the CF approach does in comparison.

Thank you for this valuable suggestion. We evaluated a random forest (RF) model, which achieved a relatively high predictive performance (R^2 = 0.80). This result is reported in the Discussion to provide a reference. However, since predictive optimization is not the primary focus, we did not further expand on additional ML models or extensive benchmarking experiments. The primary objective is to identify generalizable and physically interpretable relationships that can inform parameter regionalization and hydrological understanding. While RF achieves higher predictive skill, its tree-based ensemble structure does not provide an explicit functional form linking catchment attributes to N. In contrast, SR yields closed-form equations that enables derivative analysis. For instance, differentiating F_3 gives the marginal effects of each attribute on N: $\frac{\partial N}{\partial A} = a_3 d_3 \ (A + c_3)^{\ a_3 - 1}$, $\frac{\partial N}{\partial K_{\mathrm{sat}}} = e_3 \ {K_{\mathrm{sat}}}^{\ e_3 - 1}$, and $\frac{\partial N}{\partial f_{\mathrm{SWE}}} = f_3$. These derivatives can quantify

the contributions from the different geomorphological, edaphic, and climatic factors on baseflow separation parameter. This mathematical functions are particularly valuable for understanding the physical relationshp between catchment attributes and baseflow processes. Therefore, RF and SR should be viewed as complementary rather than competing approaches: RF provides a benchmark for achievable predictive performance, whereas SR offers structural transparency that facilitates theoretical interpretation and model integration.

We clarified this scope and the complementary roles of RF and SR in the introduction and discussion sections.

[Section 1, lines 107-111]

This study should not be viewed as an effort to assert a superior utility of SR over other machine learning models in the regionalization of baseflow parameters. Instead, the SR formulas serve as post-hoc interpretability tools to complement other black box models, enhancing the transparency of the underlying relationship between hydrological signatures and catchment attributes (Rudin, 2019).

[Section 5.1, lines 471-494]

In this study, we used SR to derive mathematical expressions for the predictions of N using 9 catchment attributes. Across ten cross-validation iterations, the identified expressions exhibited consistent structures, predictors, and nearly identical regression coefficients, indicating that SR can yield stable functional relationships between catchment attributes and N. Compared to the RF-based predictions reported by Lin et al. (2026), the SR-based approach showed lower predictive skill (R^2 = 0.54 vs. 0.80), reflecting the trade-off between predictive accuracy and interpretability. While RF achieves superior predictive performance, it functions as a 'black-box' ensemble, offering no explicit functional form to clarify whether environmental controls operate additively, multiplicatively, or through nonlinear transformations. In contrast, SR provides structural transparency by yielding a closed-form equation, facilitating direct analytical insights (Häfner et al., 2023; Karpatne et al., 2024). This explicit representation enables rigorous sensitivity analysis via differentiation; for instance, the marginal effects derived from equation F_3 quantify how geomorphic and climatic factors jointly govern N. By trading a degree of predictive skill for parsimony, SR transforms the problem from simple estimation into a hypothesis-generating exercise, providing compact transfer functions that are easily integrated into regionalization frameworks (Feigl et al., 2022; Samaniego et al., 2010). Therefore, RF and SR should be viewed as complementary rather than competing approaches: RF provides a benchmark for predictive performance, while SR offers structural transparency that facilitates theoretical interpretation and model integration.

(9) While Appendix A is given some text book information on the principle of streamflow separation, I would be much more interested in how the separation was performed and especially what kind of uncertainties are involved.

Thank you for this important comment. We added descriptions on the SMM procedures for baseflow separation in Section 3.1.

[Section 3.1, lines 157-162]

The SMM procedure involves partitioning daily streamflow into non-overlapping N-day intervals and identifying the minimum value within each segment. These minimum points (Q_1, Q_2, ..., Q_i, …) are

then screened using a filtering coefficient (M): a point is discarded if $M \cdot Q_i$ exceeds the value of either adjacent minimum. Finally, the baseflow series is constructed by linearly interpolating the remaining minima. The method involves two key parameters: the segment length parameter (N) and the filtering coefficient parameter (M).

Studies found that the block length parameter of SMM constitute the primary source of uncertainty (Lin et al., 2026; Stoelzle et al., 2020). That is why we proposed to regionalize the SEC-optimized N parameters using SR instead of adopting the default values. Using SEC-optimized parameter in baseflow separation ensures the different streamflow components are consistency with the tracer dynamics.

Some technical questions of concern are the following:

(10) The equations in Table 2 look to me strongly depend on what units you are choosing for Ks and A. Also, in F2 and F3 you add A and Ks which makes no sense in terms of units! What are the impacts of different choices on the overall estimation process. Is there guidance needed for any other user application?

Thank you for this insightful comment. We acknowledge that the additive structure $A + K_{sat}$ lacks hydrological justification and its appearance may be the coincidence of unit inconsistency of K_{sat}. In the previous version, we didn't take out the $\times 100$ scaling factor from K_{sat}, which may be the cause of this dimensionally inconsistency. To address this issue, we revised the SR training configurations in the following ways: A) the unit of saturated hydraulic conductivity (Ksat) was converted to mm/hour; B) multiplication, division, power-law, and logarithmic operators were not allowed to be nested within operators of the same type; C) the internal complexity of expressions inside power-law and logarithmic operators was restricted to a maximum value 3; and D) recursive formulations were not allowed. These new configurations result in more clear SR expressions: $N = A^{a_1} + b_1$, $N = d_2(A + c_2)^{a_2} + (K_{sat})^{e_2} + b_2$, and $N = d_3(A + c_3)^{a_3} + (K_{sat})^{e_3} + f_3 \cdot f_{SWE} + b_3$. Compared to the previous expressions, the new ones reveal no direct additions between variables with different units and every variable is in a single term with clear form.

We updated Table 2 and paragraphs in section 3.2 to reflect the changes.

[Section 3.2, Lines 186-188]

The nine representative catchment attributes in original units (Table 1) were used as predictors without normalization to maintain interpretability for the SR expressions.

[Section 3.2, Lines 191-206]

To control the search space and ensure physically interpretable expressions, several structural constraints were imposed in SR model training. Multiplication, division, power-law, and logarithmic operators were not allowed to be nested within operators of the same type. The internal complexity of expressions inside power-law and logarithmic operators was restricted to a maximum value of 3. The maximum allowable total complexity was set to 20. Expression complexity is defined as the sum of the complexity index assigned to each component in the equation. Take $N = 1.6 \times A^{0.2}$ as an example, if multiplication and power-law operators are

each assigned a complexity of 2 and constants and input variables are assigned a complexity of 1, the total complexity of the expression is calculated as $2 + 2 + 1 + 1 + 1 = 7$. In this study, all operators were assigned a uniform complexity index of 1 to avoid bias toward specific functional forms. Recursive formulations (i.e., expressions where the output variable appears as an input to itself) were not permitted to ensure model interpretability and avoid trivial or ill-posed solutions.

[Table 2]

Table 2. The calibrated F'_{PL} and the three stable forms of SR expressions across all ten iterations of the cross-validation. The numbers within the brackets of the first row are the complexity indices of the expressions.

Iteration	F'_{LR} (5)	F_1 (5)	F_2 (13)	F_3 (17)
1	$3.00 \times A^{0.15}$	$A^{0.23} + 3.66$	$0.40(A + 529)^{0.31} + {K_{sat}}^{0.30} + 2.39$	$0.27(A + 640)^{0.33} + {K_{sat}}^{0.35} + 3.71 f_{SWE} + 1.79$
2	$3.10 \times A^{0.15}$	$A^{0.23} + 3.76$	$0.62(A + 642)^{0.28} + {K_{sat}}^{0.32} + 1.37$	$0.40(A + 652)^{0.30} + {K_{sat}}^{0.36} + 3.38 f_{SWE} + 1.26$
3	$3.10 \times A^{0.15}$	$A^{0.23} + 3.74$	$0.47(A + 526)^{0.30} + {K_{sat}}^{0.28} + 2.19$	$0.32(A + 575)^{0.32} + {K_{sat}}^{0.33} + 3.55 f_{SWE} + 1.80$
4	$3.25 \times A^{0.14}$	$A^{0.23} + 3.86$	$0.39(A + 334)^{0.30} + {K_{sat}}^{0.31} + 2.78$	$0.26(A + 361)^{0.33} + {K_{sat}}^{0.35} + 3.50 f_{SWE} + 2.32$
5	$3.17 \times A^{0.15}$	$A^{0.23} + 3.79$	$0.39(A + 371)^{0.31} + {K_{sat}}^{0.31} + 2.64$	$0.35(A + 547)^{0.31} + {K_{sat}}^{0.35} + 3.60 f_{SWE} + 1.61$
6	$3.03 \times A^{0.15}$	$A^{0.23} + 3.76$	$0.56(A + 602)^{0.28} + {K_{sat}}^{0.29} + 1.77$	$0.41(A + 754)^{0.30} + {K_{sat}}^{0.34} + 3.63 f_{SWE} + 1.21$
7	$3.14 \times A^{0.15}$	$A^{0.23} + 3.83$	$0.44(A + 401)^{0.30} + {K_{sat}}^{0.29} + 2.58$	$0.30(A + 443)^{0.32} + {K_{sat}}^{0.33} + 3.28 f_{SWE} + 2.18$
8	$3.16 \times A^{0.15}$	$A^{0.23} + 3.85$	$0.44(A + 421)^{0.30} + {K_{sat}}^{0.31} + 2.49$	$0.31(A + 491)^{0.32} + {K_{sat}}^{0.35} + 3.54 f_{SWE} + 1.95$
9	$3.13 \times A^{0.15}$	$A^{0.23} + 3.81$	$0.46(A + 481)^{0.30} + {K_{sat}}^{0.28} + 2.42$	$0.30(A + 525)^{0.32} + {K_{sat}}^{0.33} + 3.38 f_{SWE} + 2.06$
10	$3.14 \times A^{0.15}$	$A^{0.23} + 3.79$	$0.43(A + 545)^{0.30} + {K_{sat}}^{0.31} + 2.29$	$0.29(A + 627)^{0.32} + {K_{sat}}^{0.35} + 3.57 f_{SWE} + 1.86$

(11) Looking at Fig. 2, I find some information on the derived equations, but I do not see what kind of function space is actually generated to be used in the genetic programming approach – does the GE approach in your case for recursion? Please clarify.

We thank the reviewer for pointing this out. We added the function space used in the symbolic regression framework in section 3.2. The candidate expressions were constructed from input variables (catchment attributes), free constants, and a predefined operator set (addition, subtraction, multiplication, division, power-law, and logarithm). Recursive formulations were not allowed. In addition, structural constraints were imposed to prevent excessive nesting of nonlinear operators and to maintain physical interpretability.

[Section 3.2, lines 188-206]

The function space is consisted of the catchment attributes, free constants, and a set of mathematical operators: addition (+), subtraction (−), multiplication (×), division (÷), power-law (power), and logarithm (log).

To control the search space and ensure physically interpretable expressions, several structural constraints were imposed in SR model training. Multiplication, division, power-law, and logarithmic operators were not allowed to be nested within operators of the same type. The internal complexity of expressions inside power-law and logarithmic operators was restricted to a maximum value of 3. The maximum allowable total complexity was set to 20. Expression complexity is defined as the sum of the complexity index assigned to each component in the equation. Take $N = 1.6 \times A^{0.2}$ ***as an example, if multiplication and power-law operators are each assigned a complexity of 2 and constants and input variables are assigned a complexity of 1, the total complexity of the expression is calculated as*** $2 + 2 + 1 + 1 + 1 = 7$***. In this study, all operators were assigned a uniform complexity index of 1 to avoid bias toward specific functional forms. Recursive formulations (i.e., expressions where the output variable appears as an input to itself) were not permitted to ensure model interpretability and avoid trivial or ill-posed solutions.***

Overall, I would recommend to the editor to accept the manuscript after some (major) revisions.

IV. Minor and specific comments

L95: Does the data set included nested catchments/information?

Thank you for the comment. Yes, the dataset may include some nested catchments. In this study, all catchments were treated as independent units, and no explicit consideration of upstream–downstream relationships was incorporated in the modeling, as the focus is on deriving generalizable relationships at the catchment scale.

L117: Units of the variables in the equation?

We thank the reviewer for pointing this out. The units for all variables in the equation were added.

[Section 3.1, lines 165-167]

In the literature, N is often default to 5 days or predicted using a power-law relationship with catchment area, namely $N = 1.6 \times A^{0.2}$, where A is the catchment area in km^2 and N is expressed in days.

L199: In Fig. 3 the individual (10) lines are hardly visible or distinguishable

Thank you for this comment. The individual lines correspond to the consistent SR formulas of the ten cross-validation folds. They appear closely clustered because the coefficients of the formulas are highly similar. This behavior reflects the stability of the SR-derived relationships rather than a visualization limitation. We added some descriptions to clarified this point.

[Section 4.1, lines 287-299]

Figure 3 further illustrates the behavior of these formulas. For F_1, the nearly identical exponents (~0.23) and intercepts (3.66–3.86 days) result in almost overlapping curves (Figure 3a), indicating a stable power-law relationship between N and A, with a diminishing rate of increase. For F_2, the similarly constrained exponents (0.28–0.32) and intercepts (1.37–2.78 days) produce tightly clustered response curves (Figure 3b-c), showing that both A and K_{sat} contribute positively to N. F_3 extends F_2 by introducing f_{SWE} as a linear term. The replicated formulas still exhibit closely grouped slopes (3.28–3.71 days) and intercepts (1.21–2.32 days), which explains the clustering of curves in Figure 3d–f. The marginal relationships of N with A and K_{sat} in F_3 remain consistent with those in F_2, whereas increasing f_{SWE} leads to an approximately linear increase in N, at a rate of about 0.3–0.4 days per 0.1 increment in snow fraction. Overall, SR identifies A as the most influential factor in predicting N, as evidenced by its presence in all SR-derived formulas. The narrower ranges of predicted Ns in Figure 3b and d also suggest that A exerts greater influence than K_{sat} and f_{SWE}.

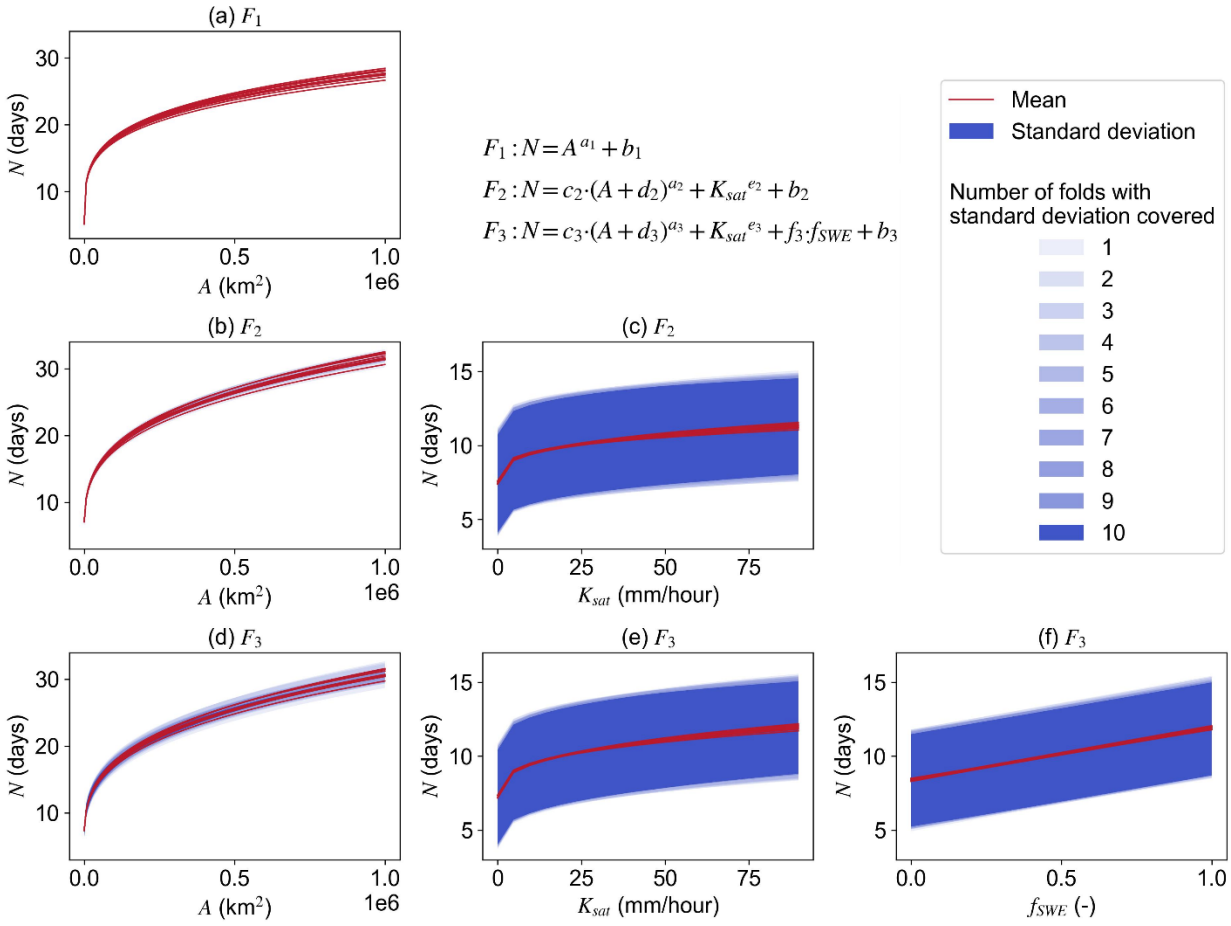


Figure 3. Marginal relationship of N on different predictors (A, K_{sat}, and f_{SWE}) that consist of the SR expressions (F_1, F_2, and F_3). Each line represents one of the ten instances of F_1, F_2, and F_3. Panels a, b, and d are for A; panels c and e are for K_{sat}; panel f is for f_{SWE}.

L268: I am very surprised by the relatively small spread of Ks values that is observed in the catchments?

Thank you for this comment. The K_{sat} values were derived by taking spatial average over the catchment, which smoothes the extremes, resulting in a narrow value range. In addition, these K_{sat} values (ranging from 0 to 80 mm/hour) are in consistent with the ranges reported by others (Choulga et al., 2024; Gupta et al., 2021).

L336-340: Can you provide some reference for the statements.

Thanks for this suggestion. We added two references to support the statements.

[Section 5.3, lines 540-544]

Moreover, the effects of K_{sat} and f_{SWE} are more evident when A is smaller than 100 km^2 (Figure 6a). As catchment area increases ($A > 100$ km^2), their influence is outweighed by A, highlighting the

dominant role of drainage area in shaping the streamflow response time (McGlynn et al., 2004; Sólyom and Tucker, 2007).

References

Choulga, M. et al., 2024. Technical note: Surface fields for global environmental modelling. Hydrol. Earth Syst. Sci., 28(13): 2991-3036. DOI:10.5194/hess-28-2991-2024

Dorgo, G., Kulcsar, T., Abonyi, J., 2021. Genetic programming-based symbolic regression for goal-oriented dimension reduction. Chemical Engineering Science, 244: 116769. DOI:https://doi.org/10.1016/j.ces.2021.116769

Feigl, M. et al., 2022. Automatic Regionalization of Model Parameters for Hydrological Models. Water Resources Research, 58(12): e2022WR031966. DOI:https://doi.org/10.1029/2022WR031966

Gupta, S., Lehmann, P., Bonetti, S., Papritz, A., Or, D., 2021. Global Prediction of Soil Saturated Hydraulic Conductivity Using Random Forest in a Covariate-Based GeoTransfer Function (CoGTF) Framework. Journal of Advances in Modeling Earth Systems, 13(4): e2020MS002242. DOI:https://doi.org/10.1029/2020MS002242

Häfner, D., Gemmrich, J., Jochum, M., 2023. Machine-guided discovery of a real-world rogue wave model. Proceedings of the National Academy of Sciences, 120(48): e2306275120. DOI:doi:10.1073/pnas.2306275120

Karpatne, A., Jia, X., Kumar, V., 2024. Knowledge-guided machine learning: Current trends and future prospects. arXiv preprint arXiv:2403.15989.

Lin, Y. et al., 2026. Regionalization of Optimal Baseflow Separation using Catchment-scale Characteristics. Water Resources Research, 62(3).

McGlynn, B.L., McDonnell, J.J., Seibert, J., Kendall, C., 2004. Scale effects on headwater catchment runoff timing, flow sources, and groundwater-streamflow relations. Water Resources Research, 40(7). DOI:https://doi.org/10.1029/2003WR002494

Pumo, D., Noto, L.V., 2023. Exploring the use of multi-gene genetic programming in regional models for the simulation of monthly river runoff series. Stochastic Environmental Research and Risk Assessment, 37(5): 1917-1941. DOI:10.1007/s00477-022-02373-1

Razaq, S.A. et al., 2016. Prediction of Flow Duration Curve in Ungauged Catchments Using Genetic Expression Programming. Procedia Engineering, 154: 1431-1438. DOI:https://doi.org/10.1016/j.proeng.2016.07.516

Rudin, C., 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nat Mach Intell, 1(5): 206-215. DOI:10.1038/s42256-019-0048-x

Samaniego, L., Kumar, R., Attinger, S., 2010. Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. Water Resources Research, 46(5). DOI:https://doi.org/10.1029/2008WR007327

Sólyom, P.B., Tucker, G.E., 2007. The importance of the catchment area–length relationship in governing non-steady state hydrology, optimal junction angles and drainage network pattern. Geomorphology, 88(1): 84-108. DOI:https://doi.org/10.1016/j.geomorph.2006.10.014

Stoelzle, M., Schuetz, T., Weiler, M., Stahl, K., Tallaksen, L.M., 2020. Beyond binary baseflow separation: a delayed-flow index for multiple streamflow contributions. Hydrology and Earth System Sciences, 24(2): 849-867. DOI:10.5194/hess-24-849-2020

Reply to Reviewer #2

The manuscript titled addresses the challenge of accurately estimating the segment length parameter (N) for the Smooth Minima Method (SMM) in baseflow separation. By employing Symbolic Regression (SR) on a dataset of 855 catchments across the Contiguous United States, the authors aim to overcome the limitations of fixed default values or simple power-law functions traditionally used in the literature. The study successfully identifies three interpretable mathematical expressions that link parameter N to catchment area, saturated hydraulic conductivity, and snow day fraction. The validation results demonstrate that the SR-derived formulas significantly improve predictive performance (improving R2 from 0.23 to 0.54) and achieve higher Kling-Gupta Efficiency in baseflow separation compared to existing methods. Additionally, the use of Specific Electrical Conductance (SEC) mass balance serves as a robust independent validation of the proposed regionalization approach. Overall, this work provides valuable insights into the physical controls of baseflow processes and demonstrates the utility of interpretable machine learning in hydrology.

However, I have several concerns regarding the physical consistency of the derived formulas, the justification for choosing SR over other methods, and the robustness of the performance evaluation thresholds. These issues, along with comments on terminology and methodology, are detailed below.

Thank you for this comprehensive summary and for the positive evaluation of our work. We appreciate the reviewer's recognition of the motivation of this study, the use of symbolic regression to derive interpretable mathematical relationships between the SMM parameter N and catchment characteristics, and the improvement in predictive performance achieved by the proposed approach. We also thank the reviewer for the constructive comments. These important points have been carefully considered, and detailed responses and corresponding revisions have been made in the manuscript, as described below.

1. The abstract states that "N increases exponentially with catchment area...". However, the derived formulas are power-law functions, not exponential functions.

Thank you for pointing this out. We corrected the statement in the abstract and throughout the manuscript.

[Abstract, lines 30-33]

The three expressions reveal that N increases following a power-law relationship with catchment area (A) and catchment-averaged soil saturated hydraulic conductivity (K_{sat}) with decreasing rates, while it increases linearly with snow day fraction (f_{SWE}).

[Section 4.1, lines 287-289]

For F_1, the nearly identical exponents (~0.23) and intercepts (3.66–3.86 days) result in almost overlapping curves (Figure 3a), indicating a stable power-law relationship between N and A, with a diminishing rate of increase.

2. The introduction of SR in the current manuscript is relatively brief (lines 61-71). While the authors highlight the interpretability of SR compared to Random Forest (RF), there is a lack of review regarding the application of SR in the broader field of hydrology. It would strengthen the rationale of the study if the authors could briefly mention successful applications of SR in other hydrological contexts (Chadalawada et al., 2020) before narrowing down to baseflow separation. This would demonstrate that SR is a proven tool in the domain.

We thank the reviewer for this constructive suggestion. We added some descriptions in the Introduction to address the successful applications of SR in hydrological model parameter regionalization.

[Section 1, lines 92-102]

In recent years, SR has been increasingly applied in hydrology to uncover governing relationships in complex environmental systems, owing to its ability to balance predictive performance and interpretability (Chadalawada et al., 2020; Sheta et al., 2023). One example is the use of SR to extract explicit functional relationships between catchment attributes and hydrological model parameters for ungagged catchments (Feigl et al., 2022; Li et al., 2024). Unlike a "black-box" model such as RF, SR derives explicit and concise equations that identify underlying data patterns, while mitigating overfitting through complexity control (Kronberger et al., 2022; Wilstrup and Kasak, 2021). This structural transparency enables direct interpretation of how catchment attributes govern baseflow parameter values, which in terms influence the partitioning of streamflow (Feigl et al., 2020; Klotz et al., 2017; Sheta et al., 2023).

3. The Introduction highlights the role of environmental tracer data in parameter optimization (Line 48) but lacks specificity. I suggest briefly listing common tracers, specifically including SEC, in this paragraph. Since the reference dataset used in this study is SEC-optimized (as detailed in Section 2), explicitly mentioning SEC early on would provide the reader with necessary context regarding the physical basis of the ground truth parameters.

Thank you for this helpful suggestion. We listed several commonly used tracers (e.g., specific electrical conductivity, turbidity, and stable isotopes) in the Introduction and explicitly mentioned SEC.

[Section 1, lines 52-54]

Commonly used tracers include specific electrical conductivity (SEC), turbidity, and stable isotopes, among which SEC is the most widely applied due to its routine availability in many monitoring programs (Mei et al., 2024).

4. The statement in Line 84 that optimal N shows “no clear spatial patterns” is ambiguous, particularly since the subsequent analysis links N to spatially structured attributes. I suggest clarifying this phrasing to avoid apparent contradiction.

Thank you for this comment. We refined our descriptions to indicate the multifactor controls on N.

[Section 2, lines 123-126]

The optimal N parameters for these catchments are depicted in Figure 1d with most values smaller than 17 days. Larger N values tend to occur in mountainous regions characterized by drier climates and greater snow persistence with implicit spatial patterns, which may be attributed to the multifactor controls of N (Lin et al. 2026).

5. Section 3.1 describes the parameters N and M well but briefly lacks a description of the algorithmic procedure. Adding one or two sentences explaining how the method iterates (e.g., dividing the hydrograph into blocks of N days, identifying minima, and comparing adjacent minima using M) would make the method section more self-contained.

Thank you for this helpful suggestion. We added the main steps of SMM as requested.

[Section 3.1, lines 157-161]

The SMM procedure involves partitioning daily streamflow into non-overlapping N-day intervals and identifying the minimum value within each segment. These minimum points (Q_1, Q_2, ..., Q_i, ...) are then screened using a filtering coefficient (M): a point is discarded if $M \cdot Q_i$ exceeds the value of either adjacent minimum. Finally, the baseflow series is constructed by linearly interpolating the remaining minima.

6. In Section 3.2, the manuscript does not specify whether the input variables (catchment characteristics) were normalized or standardized prior to SR training. Based on the discussion in Section 4.1 (where specific thresholds like A < 100 km2 are mentioned), it implies that raw data with physical units were used.

If raw data were used, the term \left(A+K_{sat}\right) in Formulas F_2 and F_3 is dimensionally inhomogeneous, adding Area (L^2) to Hydraulic Conductivity (L/T). This makes the formula valid only for the specific combination of units chosen in this study (km2 and mm/h) and undermines the claim of physical meaningfulness.

Even if the variables were normalized (rendering them dimensionless), the authors should interpret the physical meaning of the additive structure \left(A+K_{sat}\right). This structure implies a direct substitutability between catchment size and soil permeability in generating baseflow delay. What is the hydrological justification for this specific interaction form, rather than, for example, a multiplicative interaction \left(A+K_{sat}\right) which is more common in hydraulic process laws?

Thank you for this insightful comment. We acknowledge that the $A + K_{sat}$ structure lacks hydrological justification and its appearance may be the coincidence of unit inconsistency of K_{sat}. In the previous version, we didn't take out the $\times 100$ scaling factor from K_{sat}, which distorted its magnitude and contributed to the emergence of this dimensionally inconsistent form during the SR search.

To address this issue, we retrained the SR models using K_{sat} (mm/hr) in its actual magnitude. In addition, we imposed stricter structural constraints to guide the search toward physically interpretable forms. Specifically, multiplication, division, power-law, and logarithmic operators were not allowed to be nested within operators of the same type. The internal complexity of expressions within power-law and logarithmic operators was limited to a maximum value of 3, and the overall expression complexity was capped at 20. Recursive formulations were prohibited to avoid trivial or ill-posed solutions.

With these modifications, the updated SR expressions no longer contain direct additions between variables with different units. The revised formulas are: $N = A^{a_1} + b_1$, $N = d_2(A + c_2)^{a_2} + (K_{sat})^{e_2} + b_2$, and $N = d_3(A + c_3)^{a_3} + (K_{sat})^{e_3} + f_3 \cdot f_{SWE} + b_3$. In these expressions, A and K_{sat} contribute through separate terms, ensuring dimensional consistency and improving physical interpretability. From a hydrological perspective, catchment area reflects the characteristic spatial scale of storage and drainage organization, whereas K_{sat} represents soil permeability and subsurface flow transmissivity. Their additive contributions represent independent influences on the characteristic delay parameter *N*.

For the comment on if SR should be trained with normalized variables, we prefer to retained the units and the magnitudes to preserves their physical meanings for better interpretation of the resulting expressions.

[Section 3.2, lines 186-188]

The nine representative catchment attributes in original units (Table 1) were used as predictors without normalization to maintain interpretability for the SR expressions.

[Section 3.2, lines 191-206]

To control the search space and ensure physically interpretable expressions, several structural constraints were imposed in SR model training. Multiplication, division, power-law, and logarithmic operators were not allowed to be nested within operators of the same type. The internal complexity of expressions inside power-law and logarithmic operators was restricted to a maximum value of 3. The maximum allowable total complexity was set to 20. Expression complexity is defined as the sum of the complexity index assigned to each component in the equation. Take $N = 1.6 \times A^{0.2}$ as an example, if multiplication and power-law operators are each assigned a complexity of 2 and constants and input variables are assigned a complexity of 1, the total complexity of the expression is calculated as $2 + 2 + 1 + 1 + 1 = 7$. In this study, all operators were assigned a uniform complexity index of 1 to avoid bias toward specific functional forms. Recursive formulations (i.e., expressions where the output variable appears as an input to itself) were not permitted to ensure model interpretability and avoid trivial or ill-posed solutions.

[Table 2]

Table 2. The calibrated F'_{PL} and the three stable forms of SR expressions across all ten iterations of the cross-validation. The numbers within the brackets of the first row are the complexity indices of the expressions.

Iteration	F'_{LR} (5)	F_1 (5)	F_2 (13)	F_3 (17)
1	$3.00 \times A^{0.15}$	$A^{0.23} + 3.66$	$0.40(A + 529)^{0.31} + {K_{sat}}^{0.30} + 2.39$	$0.27(A + 640)^{0.33} + {K_{sat}}^{0.35} + 3.71 f_{SWE} + 1.79$
2	$3.10 \times A^{0.15}$	$A^{0.23} + 3.76$	$0.62(A + 642)^{0.28} + {K_{sat}}^{0.32} + 1.37$	$0.40(A + 652)^{0.30} + {K_{sat}}^{0.36} + 3.38 f_{SWE} + 1.26$
3	$3.10 \times A^{0.15}$	$A^{0.23} + 3.74$	$0.47(A + 526)^{0.30} + {K_{sat}}^{0.28} + 2.19$	$0.32(A + 575)^{0.32} + {K_{sat}}^{0.33} + 3.55 f_{SWE} + 1.80$
4	$3.25 \times A^{0.14}$	$A^{0.23} + 3.86$	$0.39(A + 334)^{0.30} + {K_{sat}}^{0.31} + 2.78$	$0.26(A + 361)^{0.33} + {K_{sat}}^{0.35} + 3.50 f_{SWE} + 2.32$
5	$3.17 \times A^{0.15}$	$A^{0.23} + 3.79$	$0.39(A + 371)^{0.31} + {K_{sat}}^{0.31} + 2.64$	$0.35(A + 547)^{0.31} + {K_{sat}}^{0.35} + 3.60 f_{SWE} + 1.61$
6	$3.03 \times A^{0.15}$	$A^{0.23} + 3.76$	$0.56(A + 602)^{0.28} + {K_{sat}}^{0.29} + 1.77$	$0.41(A + 754)^{0.30} + {K_{sat}}^{0.34} + 3.63 f_{SWE} + 1.21$
7	$3.14 \times A^{0.15}$	$A^{0.23} + 3.83$	$0.44(A + 401)^{0.30} + {K_{sat}}^{0.29} + 2.58$	$0.30(A + 443)^{0.32} + {K_{sat}}^{0.33} + 3.28 f_{SWE} + 2.18$
8	$3.16 \times A^{0.15}$	$A^{0.23} + 3.85$	$0.44(A + 421)^{0.30} + {K_{sat}}^{0.31} + 2.49$	$0.31(A + 491)^{0.32} + {K_{sat}}^{0.35} + 3.54 f_{SWE} + 1.95$
9	$3.13 \times A^{0.15}$	$A^{0.23} + 3.81$	$0.46(A + 481)^{0.30} + {K_{sat}}^{0.28} + 2.42$	$0.30(A + 525)^{0.32} + {K_{sat}}^{0.33} + 3.38 f_{SWE} + 2.06$
10	$3.14 \times A^{0.15}$	$A^{0.23} + 3.79$	$0.43(A + 545)^{0.30} + {K_{sat}}^{0.31} + 2.29$	$0.29(A + 627)^{0.32} + {K_{sat}}^{0.35} + 3.57 f_{SWE} + 1.86$

7. Section 3.2 outlines the SR setup but omits key hyperparameters required for reproducibility, such as population size, number of generations, and mutation/crossover rates used in the PySR configuration. I suggest including these details, perhaps in the Appendix.

We appreciate the reviewer's suggestion. We added detailed descriptions regarding the genetic programming hyperparameters used for modeling.

[Section 3.2, lines 188-212]

The function space is consisted of the catchment attributes, free constants, and a set of mathematical operators: addition (+), subtraction (−), multiplication (×), division (÷), power-law (power), and logarithm (log).

To control the search space and ensure physically interpretable expressions, several structural constraints were imposed in SR model training. Multiplication, division, power-law, and logarithmic operators were not allowed to be nested within operators of the same type. The internal complexity of expressions inside power-law and logarithmic operators was restricted to a maximum value of 3. The maximum allowable total complexity was set to 20. Expression complexity is defined as the sum of the complexity index assigned to each component in the equation. Take $N = 1.6 \times A^{0.2}$ ***as an example, if multiplication and power-law operators are each assigned a complexity of 2 and constants and input variables are assigned a complexity of 1, the total complexity of the expression is calculated as*** $2 + 2 + 1 + 1 + 1 = 7$***. In this study, all operators were assigned a uniform complexity index of 1 to avoid bias toward specific functional forms. Recursive formulations (i.e., expressions where the output variable appears as an input to itself) were not permitted to ensure model interpretability and avoid trivial or ill-posed solutions.***

The SR search process was configured with the following hyperparameters: a population size of 33, populations of 15, the crossover rate of 0.066, and evolved over 40 generations. The goodness of fit between the reference and the predicted Ns is evaluated using the mean squared error (MSE)

8. The legend in Figure 7b includes categories like F_1&F_2, implying a tie in performance. However, the manuscript lacks an explicit definition of the threshold used to categorize these ties (e.g.,\Delta R^2<0.02?). Crucially, the choice of this threshold is intrinsic to the discussion on the complexity-accuracy trade-off (Section 5.3). A strict threshold (e.g., \Delta R^2=0) favors complex models F3 even for negligible gains, while a practical threshold (e.g., \Delta R^2<0.01) might reveal that the simpler F1 is good enough for a much larger portion of the CONUS.Therefore, beyond simply clarifying the current threshold, I suggest the authors perform a brief sensitivity analysis on this threshold. For instance, how does the spatial pattern of the best formula change if the tolerance for a tie is set to 0.01 vs. 0.03? This would provide a more robust and transparent visualization of where the added complexity of F3 yields substantial benefits versus marginal statistical gains.

We thank the reviewer for this insightful comment. Instead of applying an arbitrary threshold to quantify performance differences, we applied the Diebold–Mariano test (Diebold and Mariano, 1995) to define ties. We ranked the R^2s of F_1, F_2, and F_3, and performed Diebold–Mariano tests for the pairwised R^2 differences at the 0.01 significance level. If the largest R^2 is significantly different than the second-largest R^2, the SR formula associated with the largest R^2 is the single best formula of the catchment. If the difference between the largest and the second-largest R^2s is insignificant but that between the second and the third ones is significant, the first two SR formulas are tied. Otherwise, the three SR are tied. We believe that this statistical approach offers a more objective quantification than using arbitary R^2 as thresholds.

The descriptions regarding DM and how we applied it to our task were added to Section 3.3 and Test S2. In addition, Figure 7 and the corresponding descriptions were updated accordingly.

[Section 3.3, lines 253-263]

To assess whether the differences in SR-based SEC prediction performance are statistical significant, we apply the Diebold–Mariano (DM) test (Diebold and Mariano, 1995). Details on procedure and test statistics are provided in Text S2. For each catchment, pairwise DM tests are performed among the three SR formulas to determine whether their SEC R^2 differences are statistical significance at the 0.01 significance level. If the largest R^2 is significantly different from the second-largest R^2, the SR formula associated with the largest R^2 is the single best formula of the catchment. If the difference between the largest and the second-largest R^2s is insignificant but that between the second and the third ones is significant, the first two SR formulas are tied. Otherwise, the three SR are tied. Based on these catchment-scale rankings, the best-performing formula for each region is determined as the one most frequently ranked as the best (including ties) among all catchments within the region.

[Supporting Information, Text S2]

To evaluate whether the predictive performance of different formulas is statistically distinguishable, we applied the Diebold–Mariano (DM) test (Diebold and Mariano, 1995). The DM test compares predictive accuracy based on the loss differential between two competing models. First, the loss differential at time t is defined as

$$d_t = L(e_{A,t}) - L(e_{B,t}) \ , \qquad (4)$$

where $e_{A,t}$ and $e_{B,t}$ denote the prediction errors of formulas A and B at time t, respectively, and $L(\cdot)$ is the loss function. In this study, the squared error was used to test for the significant differences between different R^2 values:

$$L(e_t) = e_t^2 \ . \qquad (5)$$

The DM test statistic is then computed as

$$DM = \frac{\bar{d}}{\sqrt{\widehat{Var}(\bar{d})}} \ , \qquad (6)$$

where $\bar{d}$ is the mean loss differential and $\widehat{Var}(\bar{d})$ is the long-run variance of $\bar{d}$. Under the null hypothesis of equal predictive accuracy, the DM statistic asymptotically follows a standard normal distribution. A two-sided test was used to determine whether the predictive performance of two SR formulas differed significantly at the 0.01 significance level.

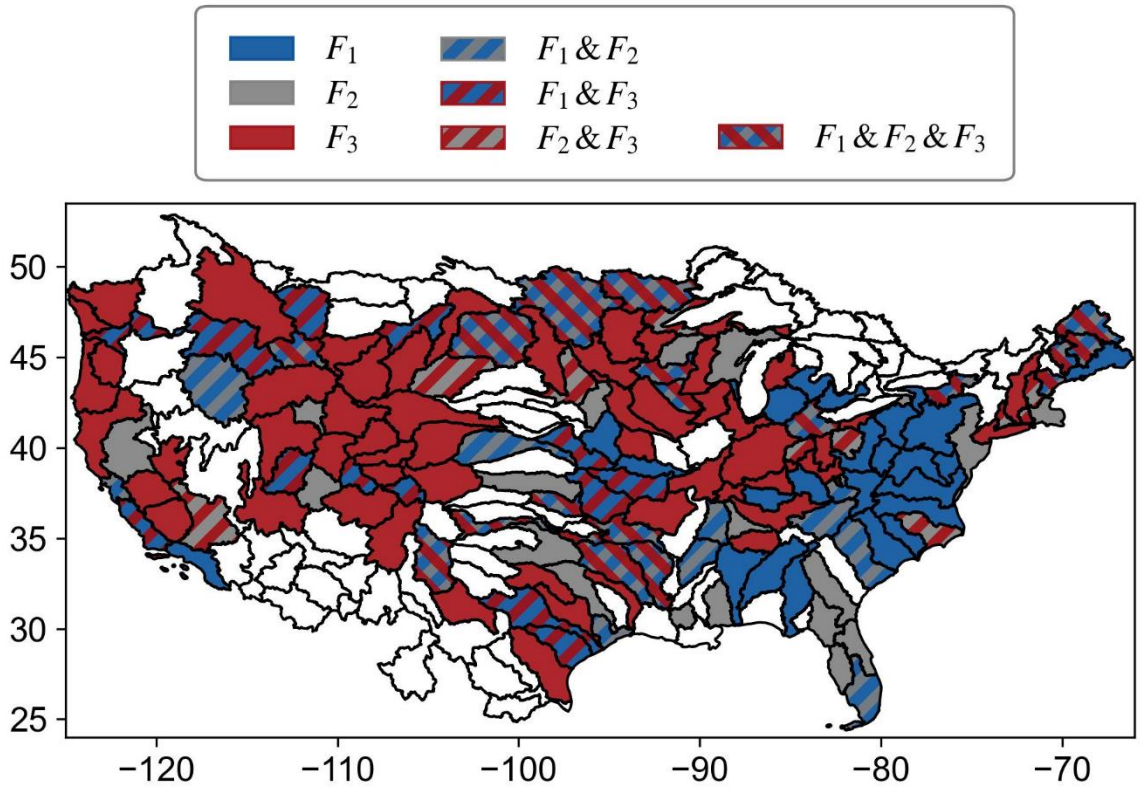


Figure 7. Spatial distribution of the best-performing formulas for SEC estimations across different HUC4 regions. At the HUC4 scale, the best-performing formula is the one(s) appears most frequently as the best among all catchments within the region.

9. In Section 5.1, the manuscript motivates the use of SR by contrasting it with the black box nature of RF. I suggest the authors provide a more precise definition of what constitutes interpretability in this context.

In hydrological literatures (Chen et al., 2022; Guillon et al., 2020), tree-based models are often considered interpretable via tools like SHAP values, which can quantify driver-response relationships much like the derivatives in SR. Given the significant drop in predictive skill (RF R2=0.80 vs. SR R2=0.54), it would be helpful to clarify why the explicit mathematical form of SR is preferred over the higher accuracy of RF. Is the interpretability here strictly defined as having a closed-form equation? A brief discussion about the definition of interpretability would provide a more balanced perspective.

References

Chadalawada, J., Herath, H.M.V.V., Babovic, V., 2020. Hydrologically Informed Machine Learning for Rainfall-Runoff Modeling: A Genetic Programming-Based Toolkit for

Automatic Model Induction. Water Resour. Res. 56, e2019WR026933. https://doi.org/10.1029/2019WR026933

Chen, Y., Li, D., Zhao, Q., Cai, X., 2022. Developing a generic data-driven reservoir operation model. Adv. Water Resour. 167, 104274. https://doi.org/10.1016/j.advwatres.2022.104274

Guillon, H., Byrne, C.F., Lane, B.A., Sandoval Solis, S., Pasternack, G.B., 2020. Machine Learning Predicts Reach-Scale Channel Types From Coarse-Scale Geospatial Data in a Large River Basin. Water Resour. Res. 56, e2019WR026691. https://doi.org/10.1029/2019WR026691

Thank you for this insightful comment. We explicitly defined interpretability as structural transparency, meaning that the relationship between predictors and predictands is expressed in an analytical equation. Although tree-based models can be interpreted using post-hoc tools such as SHAP values and partial dependence plots, they only provide approximate, model-dependent interpretations. Moreover, they typically describe average effects while overlooking higher-order interactions and may be sensitive to feature correlation, potentially leading to unstable or misleading explanations (Apley and Zhu, 2020; Sundararajan and Najmi, 2020). In contrast, the SR-derived equation provide explicit expression for N, making the model transparent and analyzable (Häfner et al., 2023; Karpatne et al., 2024). For example, differentiating F_3 gives the marginal effects of each attribute on N: $\frac{\partial N}{\partial A} = a_3 d_3 \, (A + c_3)^{\,a_3 - 1}$, $\frac{\partial N}{\partial K_{sat}} = e_3 \, {K_{sat}}^{\,e_3 - 1}$, and $\frac{\partial N}{\partial f_{SWE}} = f_3$. These derivatives make it possible to quantify parameter sensitivity directly and to clarify how geomorphic and climatic factors jointly influence baseflow separation. Overall, RF and SR should be viewed as complementary rather than competing approaches: RF provides a benchmark for achievable predictive performance, whereas SR offers structural transparency that facilitates theoretical interpretation and model integration.

We revised the introduction and discussion sections to clarify these points.

[Section 1, lines 78-86]

Despite higher prediction accuracy, the RF-based regionalization model does not provide an explicit analytical expression linking catchment attributes to parameter N (Rudin, 2019). Although tree-based models can be interpreted using post-hoc tools such as SHAP values and partial dependence plots, they provide only approximate, model-dependent interpretations rather than explicit functional relationships (Makke and Chawla, 2024; Rudin, 2019). Moreover, they typically describe average effects while overlooking higher-order interactions and may be sensitive to feature correlation, potentially leading to unstable or misleading explanations (Apley and Zhu, 2020; Sundararajan and Najmi, 2020).

[Section 1, lines 107-111]

This study should not be viewed as an effort to assert a superior utility of SR over other machine learning models in the regionalization of baseflow parameters. Instead, the SR formulas serve as post-hoc interpretability tools to complement other black box models, enhancing the transparency of the underlying relationship between hydrological signatures and catchment attributes (Rudin, 2019).

[Section 5.1, lines 471-494]

In this study, we used SR to derive mathematical expressions for the predictions of N using 9 catchment attributes. Across ten cross-validation iterations, the identified expressions exhibited consistent structures, predictors, and nearly identical regression coefficients, indicating that SR can yield stable functional relationships between catchment attributes and N. Compared to the RF-based predictions reported by Lin et al. (2026), the SR-based approach showed lower predictive skill (R^2 = 0.54 vs. 0.80), reflecting the trade-off between predictive accuracy and interpretability. While RF achieves superior predictive performance, it functions as a 'black-box' ensemble, offering no explicit functional form to clarify whether environmental controls operate additively, multiplicatively, or through nonlinear transformations. In contrast, SR provides structural transparency by yielding a closed-form equation, facilitating direct analytical insights (Häfner et al., 2023; Karpatne et al., 2024). This explicit representation enables rigorous sensitivity analysis via differentiation; for instance, the marginal effects derived from equation F_3 quantify how geomorphic and climatic factors jointly govern N. By trading a degree of predictive skill for parsimony, SR transforms the problem from simple estimation into a hypothesis-generating exercise, providing compact transfer functions that are easily integrated into regionalization frameworks (Feigl et al., 2022; Samaniego et al., 2010). Therefore, RF and SR should be viewed as complementary rather than competing approaches: RF provides a benchmark for predictive performance, while SR offers structural transparency that facilitates theoretical interpretation and model integration.

References

Apley, D.W., Zhu, J., 2020. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. Journal of the Royal Statistical Society Series B: Statistical Methodology, 82(4): 1059-1086. DOI:10.1111/rssb.12377

Chadalawada, J., Herath, H.M.V.V., Babovic, V., 2020. Hydrologically Informed Machine Learning for Rainfall-Runoff Modeling: A Genetic Programming-Based Toolkit for Automatic Model Induction. Water Resources Research, 56(4): e2019WR026933. DOI:https://doi.org/10.1029/2019WR026933

Diebold, F.X., Mariano, R.S., 1995. Comparing Predictive Accuracy. Journal of Business & Economic Statistics, 13(3): 253-263. DOI:10.1080/07350015.1995.10524599

Feigl, M., Herrnegger, M., Klotz, D., Schulz, K., 2020. Function Space Optimization: A Symbolic Regression Method for Estimating Parameter Transfer Functions for Hydrological Models. Water Resour Res, 56(10): e2020WR027385. DOI:10.1029/2020WR027385

Feigl, M. et al., 2022. Automatic Regionalization of Model Parameters for Hydrological Models. Water Resources Research, 58(12): e2022WR031966. DOI:https://doi.org/10.1029/2022WR031966

Häfner, D., Gemmrich, J., Jochum, M., 2023. Machine-guided discovery of a real-world rogue wave model. Proceedings of the National Academy of Sciences, 120(48): e2306275120. DOI:doi:10.1073/pnas.2306275120

Karpatne, A., Jia, X., Kumar, V., 2024. Knowledge-guided machine learning: Current trends and future prospects. arXiv preprint arXiv:2403.15989.

Klotz, D., Herrnegger, M., Schulz, K., 2017. Symbolic Regression for the Estimation of Transfer Functions of Hydrological Models. Water Resources Research, 53(11): 9402-9423. DOI:10.1002/2017wr021253

Kronberger, G., de Franca, F.O., Burlacu, B., Haider, C., Kommenda, M., 2022. Shape-Constrained Symbolic Regression—Improving Extrapolation with Prior Knowledge. Evolutionary Computation, 30(1): 75-98. DOI:10.1162/evco_a_00294

Li, Q. et al., 2024. Advancing symbolic regression for earth science with a focus on evapotranspiration modeling. npj Climate and Atmospheric Science, 7(1). DOI:10.1038/s41612-024-00861-5

Mei, Y. et al., 2024. Optimal Baseflow Separation Through Chemical Mass Balance: Comparing the Usages of Two Tracers, Two Concentration Estimation Methods, and Four Baseflow Filters. Water Resources Research, 60(7). DOI:10.1029/2023wr036386

Sheta, A., Abdel-raouf, A., Fraihat, K., Baareh, A.K., 2023. Evolutionary Design of a PSO-Tuned Multigene Symbolic Regression Genetic Programming Model for River Flow Forecasting. International Journal of Advanced Computer Science and Applications, 14: 2023. DOI:10.14569/IJACSA.2023.0140489

Sundararajan, M., Najmi, A., 2020. The Many Shapley Values for Model Explanation. In: Hal, D., III, Aarti, S. (Eds.), Proceedings of the 37th International Conference on Machine Learning. PMLR, Proceedings of Machine Learning Research, pp. 9269--9278.

Wilstrup, C., Kasak, J., 2021. Symbolic regression outperforms other models for small data sets. arXiv preprint arXiv:2103.15147.

Reply to Reviewer #3

This manuscript presents an innovative approach to the regionalization of the baseflow separation parameter N by bridging the gap between "black-box" machine learning and traditional physical hydrology. By utilizing Symbolic Regression (SR) across 855 catchments in the US, the authors successfully identified explicit and interpretable mathematical formulas that incorporate key catchment attributes such as drainage area, soil conductivity, and snow processes. The study is well motivated, and the use of SEC for validation adds significant scientific rigor. While the work is technically sound and highly relevant to the community, there are several conceptual and methodological points that require clarification and revision before publication.

Thank you for this positive and encouraging assessment of our work. We appreciate the reviewer's recognition of the novelty of using symbolic regression to bridge data-driven approaches and physical interpretability in hydrology, as well as the value of deriving explicit mathematical relationships between the baseflow separation parameter and key catchment attributes. We also thank the reviewer for the constructive comments. These points have been carefully considered, and the manuscript has been revised accordingly. Detailed responses to each comment are provided below.

1. A major concern is the mathematical structure of F2 and F3, where catchment area (A) is directly added to saturated hydraulic conductivity (Ksat). Please clarify the physical or mathematical justification for adding variables with different dimensions and how this affects the transferability of the formulas across different unit systems

Thank you for this important comment. In the previous SR model training, the $\times 100$ scaling factor in K_{sat} was preserved, causing this dimensional inconsistency. To address this issue, we retrained the SR models using the correct K_{sat} magnitude, together with some additional structural constraints to improve the physical interpretability of the resulting expressions. Specifically, multiplication, division, power-law, and logarithmic operators were not allowed to be nested within operators of the same type. The internal complexity of expressions within power-law and logarithmic operators was limited to a maximum value of 3, and the overall expression complexity was capped at 20. In addition, recursive formulations (i.e., expressions where the output variable appears as an input to itself) were not permitted to ensure model interpretability and avoid trivial or ill-posed solutions. The new SR expressions are consisted of single variable terms only, which are more physically interpretable and avoid dimensionally inconsistent combinations.

[Section 3.2, Lines 191-206]

To control the search space and ensure physically interpretable expressions, several structural constraints were imposed in SR model training. Multiplication, division, power-law, and logarithmic operators were not allowed to be nested within operators of the same type. The internal complexity of expressions inside power-law and logarithmic operators was restricted to a maximum value of 3. The maximum allowable total complexity was set to 20. Expression complexity is defined as the sum of the complexity index assigned to each component in the equation. Take $N = 1.6 \times A^{0.2}$ as an example, if multiplication and power-law operators are

each assigned a complexity of 2 and constants and input variables are assigned a complexity of 1, the total complexity of the expression is calculated as $2 + 2 + 1 + 1 + 1 = 7$. In this study, all operators were assigned a uniform complexity index of 1 to avoid bias toward specific functional forms. Recursive formulations (i.e., expressions where the output variable appears as an input to itself) were not permitted to ensure model interpretability and avoid trivial or ill-posed solutions.

746 [Table 2]

747 ***Table 2. The calibrated F'_{PL} and the three stable forms of SR expressions across all ten iterations of the cross-validation. The***
748 ***numbers within the brackets of the first row are the complexity indices of the expressions.***

Iteration	F'_{LR} (5)	F_1 (5)	F_2 (13)	F_3 (17)
1	$3.00 \times A^{0.15}$	$A^{0.23} + 3.66$	$0.40(A + 529)^{0.31} + {K_{sat}}^{0.30} + 2.39$	$0.27(A + 640)^{0.33} + {K_{sat}}^{0.35} + 3.71 f_{SWE} + 1.79$
2	$3.10 \times A^{0.15}$	$A^{0.23} + 3.76$	$0.62(A + 642)^{0.28} + {K_{sat}}^{0.32} + 1.37$	$0.40(A + 652)^{0.30} + {K_{sat}}^{0.36} + 3.38 f_{SWE} + 1.26$
3	$3.10 \times A^{0.15}$	$A^{0.23} + 3.74$	$0.47(A + 526)^{0.30} + {K_{sat}}^{0.28} + 2.19$	$0.32(A + 575)^{0.32} + {K_{sat}}^{0.33} + 3.55 f_{SWE} + 1.80$
4	$3.25 \times A^{0.14}$	$A^{0.23} + 3.86$	$0.39(A + 334)^{0.30} + {K_{sat}}^{0.31} + 2.78$	$0.26(A + 361)^{0.33} + {K_{sat}}^{0.35} + 3.50 f_{SWE} + 2.32$
5	$3.17 \times A^{0.15}$	$A^{0.23} + 3.79$	$0.39(A + 371)^{0.31} + {K_{sat}}^{0.31} + 2.64$	$0.35(A + 547)^{0.31} + {K_{sat}}^{0.35} + 3.60 f_{SWE} + 1.61$
6	$3.03 \times A^{0.15}$	$A^{0.23} + 3.76$	$0.56(A + 602)^{0.28} + {K_{sat}}^{0.29} + 1.77$	$0.41(A + 754)^{0.30} + {K_{sat}}^{0.34} + 3.63 f_{SWE} + 1.21$
7	$3.14 \times A^{0.15}$	$A^{0.23} + 3.83$	$0.44(A + 401)^{0.30} + {K_{sat}}^{0.29} + 2.58$	$0.30(A + 443)^{0.32} + {K_{sat}}^{0.33} + 3.28 f_{SWE} + 2.18$
8	$3.16 \times A^{0.15}$	$A^{0.23} + 3.85$	$0.44(A + 421)^{0.30} + {K_{sat}}^{0.31} + 2.49$	$0.31(A + 491)^{0.32} + {K_{sat}}^{0.35} + 3.54 f_{SWE} + 1.95$
9	$3.13 \times A^{0.15}$	$A^{0.23} + 3.81$	$0.46(A + 481)^{0.30} + {K_{sat}}^{0.28} + 2.42$	$0.30(A + 525)^{0.32} + {K_{sat}}^{0.33} + 3.38 f_{SWE} + 2.06$
10	$3.14 \times A^{0.15}$	$A^{0.23} + 3.79$	$0.43(A + 545)^{0.30} + {K_{sat}}^{0.31} + 2.29$	$0.29(A + 627)^{0.32} + {K_{sat}}^{0.35} + 3.57 f_{SWE} + 1.86$

749

2. Why was a standard Genetic Programming-based SR approach chosen over more recent continuous search methods or grammar-based optimization, such as those discussed by Feigl et al. (2020)?

Thank you for this insightful comment. In our study, the search space was deliberately restricted in both dimensionality and structural complexity to ensure tractability and robustness. Prior to symbolic regression, we applied a mutual information–based feature selection procedure, which reduced the candidate predictors to nine physically relevant variables. In addition, the set of mathematical operators was explicitly constrained to a small and interpretable subset, including addition, subtraction, multiplication, division, power-law, and logarithmic functions. To further control structural complexity, we imposed additional restrictions: a) multiplication, division, power-law, and logarithmic operators were not allowed to be nested within operators of the same type; b) the internal complexity of expressions within power-law and logarithmic operators was limited to a maximum of 3; and c) the total expression complexity was capped at 20. These constraints substantially reduce the search space and mitigate the known limitations of GP in handling highly complex symbolic structures. Under these controlled conditions, the advantages of more advanced approaches (e.g., their ability to efficiently explore high-dimensional and flexible grammars) are less critical. Instead, our objective is not to exhaustively search for highly complex symbolic forms, but to identify low-complexity, physically interpretable, and structurally stable relationships that are consistent across cross-validation runs. For this purpose, GP-based symbolic regression provides a good balance between search efficiency, interpretability, and implementation simplicity, and remains well aligned with our study goals. Similar GP-based SR approaches have also been successfully applied in other studies (Dorgo et al., 2021; Pumo and Noto, 2023; Razaq et al., 2016).

To further assess whether the derived relationships depend on the choice of optimization paradigm, we additionally implemented an alternative symbolic regression framework based on deep learning with reinforcement learning–guided search. The resulting expressions are: $A^{0.22} + 3.70$, $(A + 901)^{0.24} + {K_{sat}}^{0.28}$, $0.31 * (A + 549)^{0.31} + {K_{sat}}^{0.34} + 3.15 f_{SWE} + 1.84$, with corresponding R^2 values of 0.47, 0.52, and 0.55, respectively. These expressions exhibit highly similar functional forms and comparable predictive performance to those obtained using the GP-based approach. This consistency suggests that the identified relationships are robust across different symbolic regression optimization paradigms.

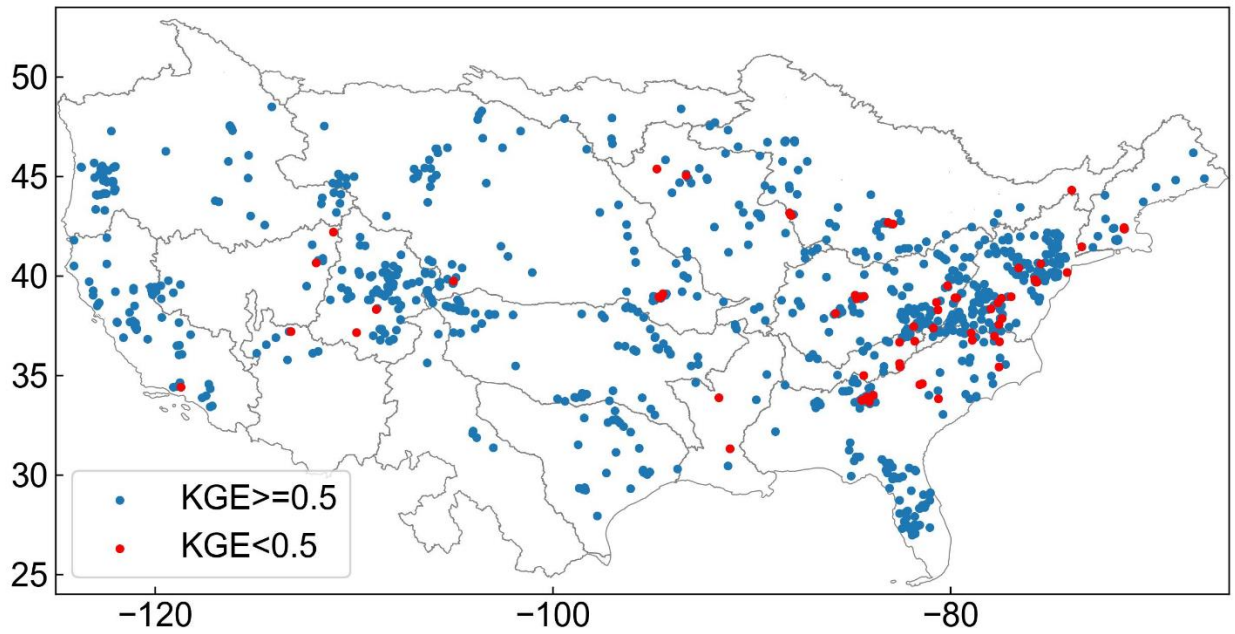
3. There is a mismatch in the units of Ksat between Table 1 (10-2 mm/h) and Figure 3 (cm/day). Please unify the units throughout the manuscript and figures for consistency.

Thank you for pointing this out. The units were unified to mm/hr throughout the manuscript.

4. Catchments with a KGE below 0.5 were excluded from the analysis. Please clarify if these excluded catchments belong to a specific hydrologic regime (e.g., arid or ephemeral

streams), and discuss the potential limitations this imposes on the global applicability of your formulas.

Thank you for this insightful comment. Mei et al. (2024) shows that the performance of baseflow separation had no explicit spatial pattern. Therefore, the excluded gages with KGE<0.5 are unlikely from specific hydrologic regimes. We also examined the spatial distribution of the excluded gages (63 in total) by plotting them with the others. The figure reveal no explicit spatial clusters, which stands as a proof of our claim.



Spatial distribution of the 855 selected catchments with Kling Gupta Efficiency between estimated and observed SEC below and above 0.5, respectively.

5. You report an R2 of 0.54 for SR versus 0.80 for RF. This is a substantial loss in predictive power. I request a more detailed discussion on whether the gain in interpretability justifies such a high "accuracy cost," particularly for practical water resource management applications.

Thank you for raising this important point. We clarify that the choice of SR is not intended to replace high-accuracy models such as RF, but to complement them by providing structural interpretability. While RF achieves higher accuracy, its ensemble structure does not provide an explicit functional form linking predictors to N, such as whether controls operate additively, multiplicatively, or through specific nonlinear transformations. In contrast, the principal advantage of SR lies in its ability to produce an explicit closed-form equation. This feature is particularly valuable in hydrology, where the goal is often not only to predict a quantity of interest, but also to understand how catchment characteristics jointly control that quantity. For instance, differentiating F_3 gives the marginal effects of each attribute on N: $\frac{\partial N}{\partial A} = a_3 d_3 \ (A + c_3)^{\ a_3 - 1}$, $\frac{\partial N}{\partial K_{\text{sat}}} = e_3 \ {K_{\text{sat}}}^{\ e_3 - 1}$, and $\frac{\partial N}{\partial f_{\text{SWE}}} = f_3$. These derivatives make it possible to quantify parameter sensitivity directly and to clarify how geomorphic and climatic factors jointly influence baseflow separation. For practical water

resource management, understanding systematic controls and mechanistic relationships can be as valuable as incremental improvements in predictive accuracy. An additional advantage of SR is parsimony. The final model is a compact equation rather than a large ensemble of trees, which makes it easier to communicate and potentially integrate into regionalization frameworks. Therefore, RF and SR should be viewed as complementary rather than competing approaches: RF provides a benchmark for achievable predictive performance, whereas SR offers structural transparency that facilitates theoretical interpretation and model integration.

We revised the introduction and discussion sections to clarify this point and to emphasize that RF and SR serve complementary roles.

[Section 1, lines 107-111]

This study should not be viewed as an effort to assert a superior utility of SR over other machine learning models in the regionalization of baseflow parameters. Instead, the SR formulas serve as post-hoc interpretability tools to complement other black box models, enhancing the transparency of the underlying relationship between hydrological signatures and catchment attributes (Rudin, 2019).

[Section 5.1, lines 471-494]

In this study, we used SR to derive mathematical expressions for the predictions of N using 9 catchment attributes. Across ten cross-validation iterations, the identified expressions exhibited consistent structures, predictors, and nearly identical regression coefficients, indicating that SR can yield stable functional relationships between catchment attributes and N. Compared to the RF-based predictions reported by Lin et al. (2026), the SR-based approach showed lower predictive skill (R^2 = 0.54 vs. 0.80), reflecting the trade-off between predictive accuracy and interpretability. While RF achieves superior predictive performance, it functions as a 'black-box' ensemble, offering no explicit functional form to clarify whether environmental controls operate additively, multiplicatively, or through nonlinear transformations. In contrast, SR provides structural transparency by yielding a closed-form equation, facilitating direct analytical insights (Häfner et al., 2023; Karpatne et al., 2024). This explicit representation enables rigorous sensitivity analysis via differentiation; for instance, the marginal effects derived from equation F_3 quantify how geomorphic and climatic factors jointly govern N. By trading a degree of predictive skill for parsimony, SR transforms the problem from simple estimation into a hypothesis-generating exercise, providing compact transfer functions that are easily integrated into regionalization frameworks (Feigl et al., 2022; Samaniego et al., 2010). Therefore, RF and SR should be viewed as complementary rather than competing approaches: RF provides a benchmark for predictive performance, while SR offers structural transparency that facilitates theoretical interpretation and model integration.

6. In Section 5.2, the formulas indicate that N (flow event duration) increases with Ksat. Physically, higher soil conductivity often implies faster drainage. While you attribute this to a shift toward slower subsurface paths, please provide more quantitative evidence or literature support to explain why this "slowing" effect dominates over the expected increase in drainage efficiency.

Thank you for this insightful comment. It is important to clarify that the parameter N represents a catchment-scale flow duration, reflecting the integrated hydrological response rather than local-scale drainage velocity. Its relationship with K_{sat} therefore emerges from

the combined effects of runoff generation, flow partitioning, and subsurface transport processes. In highly permeable catchments, enhanced infiltration reduces rapid overland flow generation and increases the contribution of subsurface flow pathways. This shift in runoff partitioning can modify the hydrograph shape and potentially extend the effective flow duration.

To further examine this mechanism, we analyzed the statistical relationships between Ksat, the recession coefficient, and N across all study catchments. The Spearman correlation between Ksat and the recession coefficient is positive ($r = 0.17$, $p < 0.01$), indicating that catchments with higher K_{sat} tend to exhibit slower recession behavior (i.e., longer recession times). In addition, the correlation between K_{sat} and N is $r = 0.24$ ($p < 0.01$), consistent with the positive dependency identified in the derived formula. These correlations support the interpretation that higher K_{sat} is associated with hydrological responses characterized by relatively longer effective flow duration. This interpretation is also consistent with previous studies showing that highly permeable geological settings can promote prolonged recession behavior (Carlier et al., 2018; Krakauer and Temimi, 2011; Tague and Grant, 2004). Therefore, while higher K_{sat} enhances local drainage efficiency at the soil profile scale, its integrated effect at the catchment scale appears to be mediated through runoff partitioning and increased subsurface contributions, which can outweigh the expected acceleration of depletion and result in longer effective flow durations.

7. Line 230, Performance is notably poor in HUC 12. Given that your input data includes reservoir storage, please clarify if the SMM method's fundamental assumption of "natural" baseflow is even applicable in these highly regulated and irrigated basins.

Thank you for this important comment. The relatively poor performance in HUC 12 may be related to its distinct hydrological characteristics. Many catchments in this region are relatively arid and exhibit weak storage effects, flashy hydrographs, and strong sensitivity to short-duration, high-intensity rainfall events (Feng et al., 2020; Kratzert et al., 2019). These fast hydrological responses may not be captured by SMM, which often reveals smooth baseflow dynamics (Stewart, 2015). Similar challenges in accurately predictions of baseflow and streamflow were also reported by other studies for the region (Feng et al., 2020; Xie et al., 2022). We added this discussion to Section 4.3.

[Section 4.3, lines 367-371]

In contrast, the lowest performance for all three formulas occurs in HUC 12, where median KGE values fall below 0.65 and more than 25% of catchments show KGE values below 0.5. This may be related to the relative arid climate and flashy hydrological response of HUC 12 (Feng et al., 2020; Kratzert et al., 2019), which is difficult for SMM to capture. Note that SMM is more skillful for smooth baseflow dynamics (Stewart, 2015).

Regarding the concern about human regulation, we examined the relationships between baseflow separation performance and two indicators of anthropogenic disturbance, namely reservoir storage and the areal fraction of irrigation, across the 855 study catchments. The correlation coefficients between R^2 of baseflow and the two attributes are both smaller than 0.05 ($p > 0.1$), indicating no statistical significance. This suggests that the baseflow

separation performance of SMM does not appear to related to the reservoir storage and the areal fraction of irrigation factors.

8. In Figure 3, the ten individual lines representing different cross-validation folds are extremely difficult to distinguish. Please improve the visualization to more clearly show the variance between folds.

Thank you for this comment. The individual lines represent the SR-derived formulas obtained from the ten cross-validation folds. They appear closely clustered because the corresponding coefficients are highly similar across folds. This behavior reflects the robustness and stability of the SR-derived relationships, rather than a limitation of the visualization. Therefore, we decide to retain the current presentation of this figure. Additionally, we have added clarifying descriptions in the revised manuscript to better explain this point.

[Section 4.1, lines 287-299]

Figure 3 further illustrates the behavior of these formulas. For F_1, the nearly identical exponents (~0.23) and intercepts (3.66–3.86 days) result in almost overlapping curves (Figure 3a), indicating a stable power-law relationship between N and A, with a diminishing rate of increase. For F_2, the similarly constrained exponents (0.28–0.32) and intercepts (1.37–2.78 days) produce tightly clustered response curves (Figure 3b-c), showing that both A and K_{sat} contribute positively to N. F_3 extends F_2 by introducing f_{SWE} as a linear term. The replicated formulas still exhibit closely grouped slopes (3.28–3.71 days) and intercepts (1.21–2.32 days), which explains the clustering of curves in Figure 3d–f. The marginal relationships of N with A and K_{sat} in F_3 remain consistent with those in F_2, whereas increasing f_{SWE} leads to an approximately linear increase in N, at a rate of about 0.3–0.4 days per 0.1 increment in snow fraction. Overall, SR identifies A as the most influential factor in predicting N, as evidenced by its presence in all SR-derived formulas. The narrower ranges of predicted Ns in Figure 3b and d also suggest that A exerts greater influence than K_{sat} and f_{SWE}.

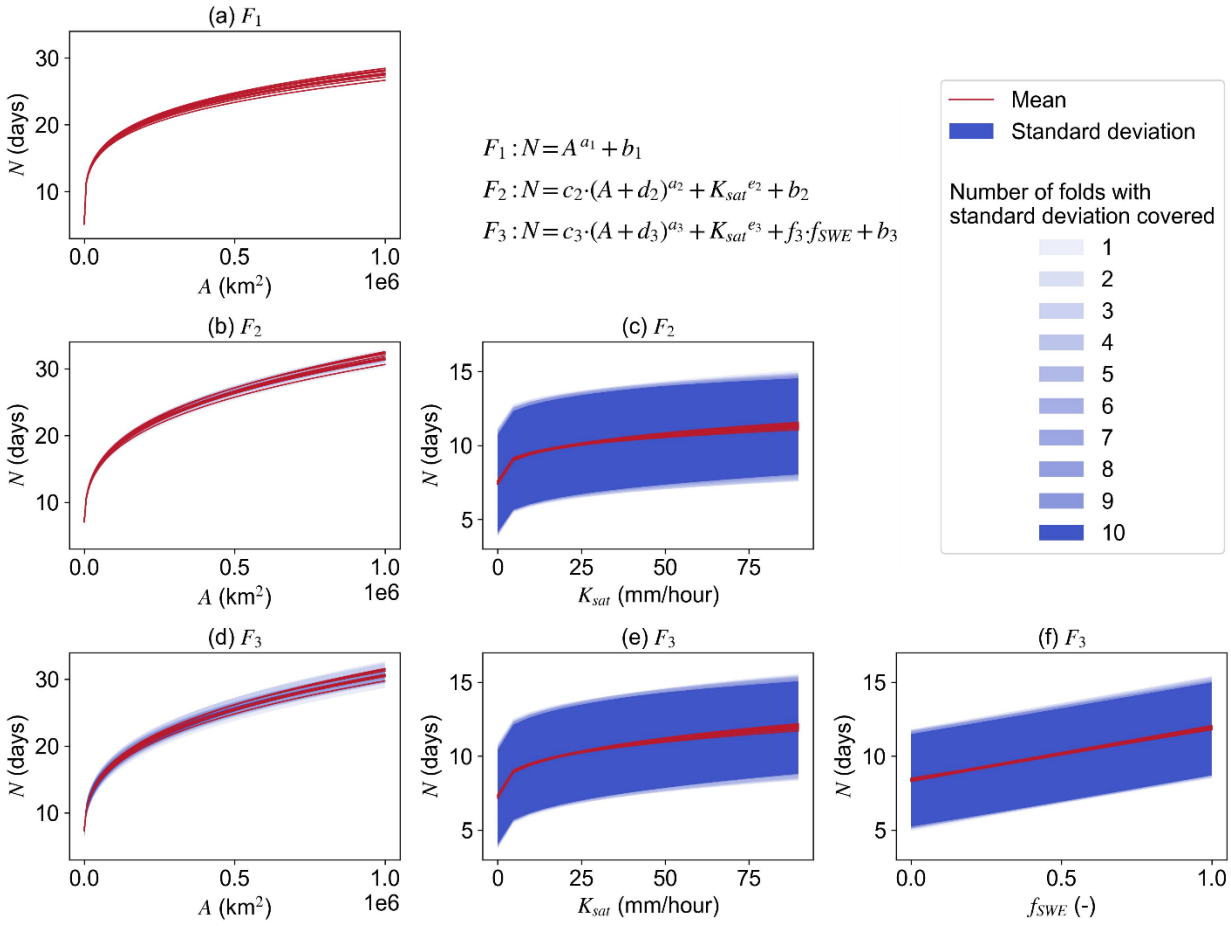


Figure 3. Marginal relationship of N on different predictors (A, K_{sat}, and f_{SWE}) that consist of the SR expressions (F_1, F_2, and F_3). Each line represents one of the ten instances of F_1, F_2, and F_3. Panels a, b, and d are for A; panels c and e are for K_{sat}; panel f is for f_{SWE}.

9. The use of multiple hatching patterns and overlapping colors in Figure 7b results in a cluttered visual presentation that is difficult for the reader to interpret.

Thank you for pointing this out. The combined color–pattern encoding was intentionally designed to preserve compositional information, allowing readers to immediately identify which formula(s) participate in the best-performing category, rather than only indicating a single dominant formula. We explored alternative visualization schemes, including a triangular color spectrum to represent combinations of formulas. However, these approaches reduced interpretability, as it became difficult to visually trace the contribution of individual formulas across regions and to distinguish shared dominance (i.e., ties) from single-formula dominance. To improve readability, we have refined the figure in the revised manuscript by adjusting hatching density and enlarging legend. In addition, a higher-resolution version will be provided upon publication to allow readers to better examine regional details.

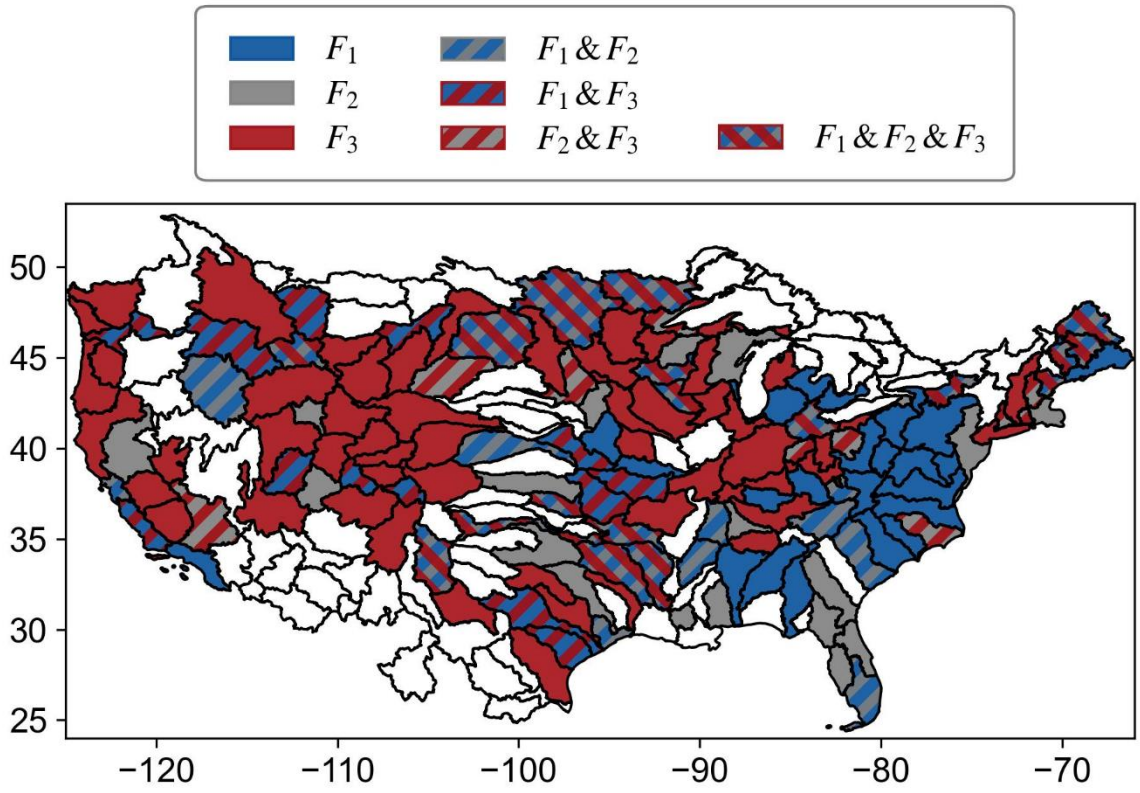


Figure 7. Spatial distribution of the best-performing formulas for SEC estimations across different HUC4 regions. At the HUC4 scale, the best-performing formula is the one(s) appears most frequently as the best among all catchments within the region.

References:

Carlier, C., Wirth, S.B., Cochand, F., Hunkeler, D., Brunner, P., 2018. Geology controls streamflow dynamics. Journal of Hydrology, 566: 756-769.

Dorgo, G., Kulcsar, T., Abonyi, J., 2021. Genetic programming-based symbolic regression for goal-oriented dimension reduction. Chemical Engineering Science, 244: 116769. DOI:https://doi.org/10.1016/j.ces.2021.116769

Feigl, M. et al., 2022. Automatic Regionalization of Model Parameters for Hydrological Models. Water Resources Research, 58(12): e2022WR031966. DOI:https://doi.org/10.1029/2022WR031966

Feng, D., Fang, K., Shen, C., 2020. Enhancing Streamflow Forecast and Extracting Insights Using Long‐Short Term Memory Networks With Data Integration at Continental Scales. Water Resources Research, 56(9): e2019WR026793. DOI:10.1029/2019wr026793

Häfner, D., Gemmrich, J., Jochum, M., 2023. Machine-guided discovery of a real-world rogue wave model. Proceedings of the National Academy of Sciences, 120(48): e2306275120. DOI:doi:10.1073/pnas.2306275120

Karpatne, A., Jia, X., Kumar, V., 2024. Knowledge-guided machine learning: Current trends and future prospects. arXiv preprint arXiv:2403.15989.

Krakauer, N.Y., Temimi, M., 2011. Stream recession curves and storage variability in small watersheds. Hydrology and Earth System Sciences, 15(7): 2377-2389.

Kratzert, F. et al., 2019. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. Hydrology and Earth System Sciences, 23(12): 5089-5110. DOI:10.5194/hess-23-5089-2019

Lin, Y. et al., 2026. Regionalization of Optimal Baseflow Separation using Catchment-scale Characteristics. Water Resources Research, 62(3).

Pumo, D., Noto, L.V., 2023. Exploring the use of multi-gene genetic programming in regional models for the simulation of monthly river runoff series. Stochastic Environmental Research and Risk Assessment, 37(5): 1917-1941. DOI:10.1007/s00477-022-02373-1

Razaq, S.A. et al., 2016. Prediction of Flow Duration Curve in Ungauged Catchments Using Genetic Expression Programming. Procedia Engineering, 154: 1431-1438. DOI:https://doi.org/10.1016/j.proeng.2016.07.516

Rudin, C., 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nat Mach Intell, 1(5): 206-215. DOI:10.1038/s42256-019-0048-x

Samaniego, L., Kumar, R., Attinger, S., 2010. Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. Water Resources Research, 46(5). DOI:https://doi.org/10.1029/2008WR007327

Stewart, M.K., 2015. Promising new baseflow separation and recession analysis methods applied to streamflow at Glendhu Catchment, New Zealand. Hydrol. Earth Syst. Sci., 19(6): 2587-2603. DOI:10.5194/hess-19-2587-2015

Tague, C.L., Grant, G.E., 2004. A geological framework for interpreting the low-flow regimes of Cascade streams, Willamette River Basin, Oregon. Water Resources Research, 40.

Xie, J. et al., 2022. Estimating Gridded Monthly Baseflow From 1981 to 2020 for the Contiguous US Using Long Short‐Term Memory (LSTM) Networks. Water Resources Research, 58(8). DOI:10.1029/2021wr031663