Response to Reviewer

We sincerely thank the reviewer for the constructive comments and valuable suggestions on our manuscript. We have carefully revised the manuscript according to the reviewer's comments. In the revised version, major changes were made to improve methodological transparency, clarify the model training and validation strategy, moderate overinterpretation of machine-learning and SHAP-based results, and improve the presentation of figures and supplementary materials. Below, we provide a point-by-point response to each comment. The reviewer's comments are reproduced first, followed by our responses.

Reviewer's general comment:

This manuscript presents a machine learning-based framework to investigate the spatially heterogeneous drivers of the Budyko parameter across hydroclimatic transition zones in the Yellow River Basin, combining clustering techniques, ML models, and SHAP interpretation methods. The manuscript is generally well structured, the figures are of good quality, and the study addresses a relevant topic for the HESS community.

However, several aspects require substantial clarification and revision before the manuscript can be considered for publication:

1. The methodological description currently lacks sufficient detail to ensure full transparency and reproducibility, especially regarding the clustering and model validation strategies.
2. Several interpretations and conclusions appear overstated relative to what is actually supported by the analyses and figures. The robustness and generalizability of some statistical inferences should be discussed more carefully given the relatively limited number of catchments considered in the study. I therefore encourage the authors to moderate some claims and strengthen the methodology and result interpretation sections.

Additional detailed comments and annotations are provided directly in the annotated PDF attached to this review.

Response: We sincerely thank the reviewer for the constructive and encouraging assessment of our manuscript. We appreciate the positive comments on the overall structure, figure quality, and relevance of the topic to the HESS community. We also agree that several aspects of the original manuscript required further clarification and revision, particularly regarding methodological transparency, the clustering procedure, the model training and validation strategy, and the interpretation of machine-learning- and SHAP-based results.

In response to these comments, we have substantially revised and expanded the manuscript. Major revisions were made throughout the Methods, Results, Discussion, and Conclusions. Specifically, we expanded the Methods section to provide a more transparent and reproducible description of the datasets, preprocessing procedures, Budyko parameter estimation, PCA-based hierarchical clustering, zone-specific model training and validation, hyperparameter settings, and SHAP implementation. We also revised the methodological framework, figure captions, and result interpretation to make the workflow and the meaning of the model outputs clearer. In addition, we added supplementary materials, revised several figures, and provided further explanations to clarify the methodological basis, robustness, and interpretive limitations of the functional zoning, zone-specific modeling, and SHAP-based interpretation. We also revised the manuscript title from a driver-oriented framing to "Zone-dependent machine-learning interpretation of Budyko parameter variability in a hydroclimatic transition region" to better reflect the exploratory and model-inferred nature of the study.

Furthermore, we moderated several statements throughout the Abstract, Results, Discussion, and Conclusions to avoid overinterpreting model-inferred associations as direct physical causality. We also discussed more explicitly the limitations related to the relatively small number of catchments, potential predictor collinearity, and limited spatial generalizability and transferability. Detailed responses to the reviewer's annotated comments are provided below.

Comment 1：In the Abstract, the reviewer noted that the labels C1, C2, and C3 were not sufficiently clear when first used in the results sentence, and suggested reordering the sentence. The reviewer also suggested adding the unit for LAI and correcting the unit of GDP to be consistent with Table 1.

Response: We thank the reviewer for this helpful suggestion. We agree that although the original Abstract stated that three hydrological functional zones were identified, the labels C1, C2, and C3 were introduced too abruptly in the following results sentence. In the revised Abstract, we now explicitly state that the three hydrological functional zones are denoted as C1, C2, and C3 before presenting the zone-dependent results. We also reordered the results from

C1 to C3 to improve clarity. In addition, we added the unit for LAI as m²/m² and corrected the GDP density unit to USD/km², consistent with Table 1. To avoid overstatement, we also replaced "critical thresholds" with "approximate SHAP-inferred transition points."

Comment 2: The reviewer noted two extra closing parentheses in the Introduction and pointed out that references within the same parentheses should be ordered alphabetically by first author name.

Response: We thank the reviewer for carefully pointing out these formatting issues. We have removed the two extra closing parentheses in the Introduction. We also checked the in-text citations throughout the manuscript and reordered references within the same parentheses alphabetically by the first author's surname where applicable.

Comment 3: The reviewer noted that it was difficult to clearly associate the four integrated steps in Fig. 2 with the corresponding methodological sections, and suggested linking the key steps to their corresponding sections to improve readability and comprehension of the manuscript.

Response: We thank the reviewer for this helpful suggestion. We agree that the original description of the methodological framework did not clearly link the workflow in Fig. 2 with the corresponding methodological sections. In the revised manuscript, we rewrote Sect. 2.2 to explicitly describe the framework as four integrated steps and linked each step to its corresponding methodological section. Specifically, the first step, "Data and Preprocessing," now includes the estimation of the time-varying Budyko parameter ω in Sect. 2.3 and the construction of the 13 predictor variables in Sect. 2.4. The second step, "Objective Functional Zoning," corresponds to Sect. 2.5, where PCA and hierarchical clustering are used to identify three clusters that are subsequently interpreted as hydrological functional zones. The third step, "Zone-specific Modelling," corresponds to Sect. 2.6, where four machine learning algorithms are benchmarked within each identified functional zone. The fourth step, "SHAP-based Interpretation," corresponds to Sect. 2.7, where the selected zone-specific model is interpreted in terms of feature-level contributions, nonlinear response patterns, approximate SHAP-inferred transition points, grouped SHAP contributions, and directional alignment patterns among predictor groups. We also revised Fig. 2 and its caption accordingly so that the visual workflow, section structure, and methodological sequence are now clearly aligned.

Comment 4: The reviewer found Fig. 2 helpful but noted that some information and formatting in the figure were unclear. Specifically, several acronyms, such as POP and WC, were only defined later in the manuscript, and the four integrated steps were difficult to associate with the corresponding boxes in the figure. The reviewer also noted that "Data and Preprocessing" was positioned between two boxes, "Budyko Parameter (ω)" used a similar font style to the step heading, and "Interpretation & Mechanism" was not formatted consistently with the other integrated steps. The reviewer suggested revising Fig. 2 to improve readability and clarify the correspondence between the workflow steps and their analytical components.

Response: We thank the reviewer for this constructive suggestion. We agree that the original version of Fig. 2 did not present the four methodological steps and their corresponding analytical components clearly enough. In the revised manuscript, we redesigned Fig. 2 to make the workflow more explicit and visually consistent. The framework is now organized into four clearly separated steps: (1) Data and Preprocessing, (2) Objective Functional Zoning, (3) Zone-specific Modelling, and (4) SHAP-based Interpretation. We also adjusted the position and formatting of the key elements so that the Budyko parameter estimation, predictor construction, functional zoning, model benchmarking, and SHAP-based interpretation can be more easily associated with the corresponding steps. In addition, we revised the terminology in Fig. 2 to be consistent with the revised manuscript. For example, "Cluster-based Modeling" was changed to "Zone-specific Modelling," and "Interaction & Coupling" and "Synergy vs. Trade-off" were replaced by "Grouped SHAP Contributions" and "Directional Alignment," respectively. These changes avoid overinterpreting SHAP results as formal interaction effects and better reflect the revised interpretation framework. We also added the definitions of the main abbreviations in the figure caption, including POP, WC, LAI, ECV, and AWSC, so that the figure can be understood without waiting for later sections. The caption of Fig. 2 was also rewritten to explain the four steps and their links to Sections 2.3–2.7 more clearly.

Comment 5: The reviewer suggested that typical values of ω and guidance on their interpretation should be provided to facilitate the analysis and improve readability for the reader.

Response: We thank the reviewer for this helpful suggestion. We agree that a clearer interpretation of the Budyko parameter ω would improve the readability of the manuscript, especially before presenting its spatiotemporal variation and SHAP-based interpretation. In the revised manuscript, we added a schematic figure (renumbered as Fig. 3 in the revised manuscript) showing Fu-type Budyko curves under representative ω values to illustrate how changes in ω affect the evaporative ratio under the same climatic condition. We also expanded the explanation of ω in Sect. 2.3 by clarifying that it is a dimensionless shape parameter reflecting integrated catchment characteristics, and by explaining the hydrological meaning of higher and lower ω values in terms of evapotranspiration partitioning and runoff coefficient. In addition, we added the range of estimated ω values in this study and clarified that "low" and "high" ω are interpreted relative to the study-specific range rather than as universal physical thresholds.

Comment 6: The reviewer noted that the assumption of negligible terrestrial water storage change (ΔS) over multiyear periods should be better justified, preferably with supporting literature, to clarify why the long-term water balance can be approximated as E = P − R for the studied catchments.

Response: We thank the reviewer for this helpful suggestion. We agree that the assumption of negligible terrestrial water storage change should be more explicitly justified. In the revised manuscript, we first present the full catchment water balance equation, $P = E + R + \Delta S$, where ΔS denotes the change in terrestrial water storage over the accounting period. We then clarify that Budyko-type analyses are generally applied at long-term or multiyear catchment scales, where ΔS is commonly assumed to be negligible because positive and negative interannual storage anomalies may partly offset each other over sufficiently long periods. We also added supporting references for this assumption. Therefore, for the purpose of estimating the Budyko parameter ω, the water balance can be approximated as $E = P - R$.

Comment 7: The choice of an 11-year moving window appears somewhat arbitrary and should be justified. Have the authors evaluated the sensitivity of the results to the selected window length?

Response: We thank the reviewer for this important comment. We agree that the choice of the 11-year moving window should be better justified. In the revised manuscript, we expanded Sect. 2.3 to clarify that the 11-year moving window was selected with reference to previous Budyko-based studies in the Loess Plateau and Yellow River region, where the same 11-year window has been adopted to estimate time-varying Budyko parameters. These studies indicate that an approximately decadal window is effective for reducing the influence of interannual climate variability and terrestrial water storage change (ΔS), while still retaining the temporal evolution of catchment characteristics. In particular, previous research has shown that, for most catchments in this region, ignoring ΔS at approximately decadal scales has relatively small effects on the Budyko framework, and that the errors of Budyko-simulated streamflow are relatively low and tend to stabilize after about 10 years. Therefore, the 11-year moving window was considered a reasonable and literature-supported choice for balancing the reduction of short-term storage and climate anomalies with the preservation of long-term variations in the Budyko parameter ω. Although we did not conduct an exhaustive comparison of all possible window lengths, the revised manuscript now provides a clearer methodological justification for the selected window length and discusses its implication as a reasonable approximation rather than an arbitrary choice.

Comment 8: The reviewer noted that important information regarding the datasets used in the analysis was missing from the main Methods section. In particular, the reviewer asked the authors to clearly specify the data sources for variables such as P and T, whether they were derived from reanalysis or station-based products, how E_0 was calculated, and the spatial and temporal resolutions of the datasets. The reviewer also noted that although some information was partly provided in the Data Availability Statement, these details should be included in the main Data and Methods section.

Response: We thank the reviewer for this helpful suggestion. We agree that the data sources and preprocessing procedures of the predictor variables should be more clearly described in the main Methods section. In the revised manuscript, we substantially expanded Sect. 2.4 to clarify the sources, spatial and temporal resolutions, and preprocessing procedures of the climatic, anthropogenic, and landscape variables. Specifically, we clarified that the climatic variables were derived from the CN05.1 daily gridded meteorological dataset at a spatial resolution of 0.25° × 0.25°. Daily precipitation was summed to obtain annual P, while daily mean temperature was averaged to obtain annual T. We also clarified that potential evapotranspiration E_0 was calculated at the daily scale using the standard Penman–Monteith equation and then aggregated to annual totals. All annual climatic variables were spatially averaged within each catchment boundary. We also added detailed descriptions of the anthropogenic variables, including the GlobPOP population dataset, gridded GDP dataset, HSWUD water use dataset, and CLCD land-cover dataset, together with their normalization or aggregation procedures. In addition, we clarified the sources and processing of the landscape variables, including ASTER GDEM, the monthly LAI dataset, and the calculation of AWSC from plant-available water content and effective rooting depth. We further stated that the access links, temporal coverage, spatial resolution, and availability information for all datasets are provided in the Data availability section. These revisions improve the transparency and reproducibility of the variable construction and preprocessing procedures.

Comment 9: It would be good to include more details about the PCA and the clustering analysis in Appendix or SI. For example, given the relatively small sample size (12 catchments) compared to the number of explanatory variables (13), it would be useful to discuss the robustness of the identified clusters.

Response: We thank the reviewer for this helpful suggestion. We agree that the PCA and hierarchical clustering procedures should be described in more detail, especially considering the relatively small number of catchments compared with the number of predictor variables. In the revised manuscript, we expanded Sect. 2.5 to provide a clearer and more reproducible description of the functional zoning procedure. Specifically, we clarified that the zoning analysis was based on the long-term mean values of the 13 predictor variables across the 12 catchments, and that the Budyko parameter ω was not used in the zoning procedure. The predictor variables were first standardized using z-score standardization to ensure comparability among variables with different units and scales. PCA was then performed on the standardized 12 × 13 predictor matrix. In this study, the first two principal components, PC1 and PC2, were retained for clustering because they explained most of the total variance. Hierarchical clustering was subsequently conducted on the PC1 and PC2 scores using Euclidean distance and Ward's minimum variance method. We also clarified how the number of clusters was selected. The average silhouette coefficient was used to compare alternative cluster numbers from $k = 2$ to $k = 5$. In the revised Results section, we report that PC1 and PC2 explained 61.1% and 24.0% of the total variance, respectively, accounting for 85.10% in total. The average silhouette coefficient reached its highest value when $k = 3$ (0.676), supporting the three-cluster solution. Following the reviewer's suggestion, we added Supplementary Tables S1–S2 and Supplementary Fig. S1. Table S1 reports the PCA explained variance ratios, Table S2 reports the loading matrix of PC1 and PC2, and Fig. S1 presents the average silhouette coefficients for $k = 2$ to $k = 5$. These additions improve the transparency and reproducibility of the PCA and clustering analysis. Finally, we added a statement in the Discussion to acknowledge the limitation associated with the relatively small number of catchments. We clarified that the identified zones should be interpreted as an objective grouping within the selected 12 representative catchments rather than as a universal regional classification. We also noted that future studies involving more catchments or independent validation regions would be useful for further testing the robustness and transferability of the zoning scheme.

Comment 10: The reviewer suggested that more details on the selected model should be included in the Appendix or Supplementary Information to ensure transparency and reproducibility, such as the final hyperparameter values, the period used for training and validation, and the performance metrics. The reviewer also asked whether the models were trained independently for each catchment or simultaneously across all catchments. In addition, the reviewer noted that the modeling strategy is reminiscent of recent hydrological emulator frameworks, such as single-site emulators and large-sample emulators, and suggested discussing these approaches in the Introduction.

Response: We thank the reviewer for this important and constructive comment. We agree that the original description of the model training and validation strategy was not sufficiently explicit, especially regarding whether the models were trained at the individual-catchment level or across multiple catchments. In the revised manuscript, we clarified that the models were not trained independently for each individual catchment. Instead, we adopted a cluster-informed, zone-specific modeling strategy. Specifically, for each hydrological functional zone identified by PCA-based hierarchical clustering, samples from all catchments belonging to that zone were pooled to train one zone-specific model. Thus, three independent zone-specific models were developed for C1, C2, and C3, respectively. This strategy is different from both individual-catchment modeling and a fully pooled basin-wide model. It allows catchments within the same hydrological functional zone to share information, while avoiding the assumption that all 12 catchments follow a single homogeneous relationship. We also revised Sect. 2.6 to provide a clearer description of the training and validation procedure. Within each functional zone, 70% of the samples from each catchment were randomly selected for model training, while the remaining 30% were retained as an independent test set. Hyperparameters were optimized using 5-fold cross-validation on the training set. Model performance was evaluated using the coefficient of determination (R^2) and root mean square error (RMSE). In addition to evaluating the overall predictive performance on the aggregated test set of each functional zone, each trained zone-specific model was further applied to the test samples of each catchment within the corresponding zone to assess catchment-level robustness under intercatchment heterogeneity. The optimal model was selected based on both zone-level accuracy and catchment-level stability. To improve reproducibility, we added a supplementary table reporting the final selected hyperparameter values of the benchmarked machine-learning models. The revised manuscript now refers to this table in Sect. 2.6.

Following the reviewer's suggestion, we also added a brief discussion of recent hydrological emulator frameworks in the Introduction. We use these studies as background examples showing the growing role of machine learning in cross-catchment hydrological modeling and in addressing spatially heterogeneous hydrological behavior. However, we also clarified the distinction between these emulator-based calibration studies and the present work. Single-site and large-sample emulators are mainly designed to approximate process-model responses or performance metrics for parameter calibration and regionalization. In contrast, our study does not aim to emulate or calibrate a process-based hydrological model. Instead, machine learning is used here as a diagnostic tool to identify zone-dependent associations between climatic, anthropogenic, and landscape predictors and the Budyko parameter ω across hydrological functional zones.

Comment 11: The reviewer noted that implementation details of the SHAP analysis were missing and should be added to improve reproducibility and interpretation. Specifically, the reviewer asked which SHAP implementation was used, to which machine-learning model SHAP was applied, whether SHAP values were computed on training or validation data, and how SHAP values should be interpreted. The reviewer also suggested briefly discussing the potential influence of multicollinearity among hydroclimatic predictors on SHAP values.

Response: We thank the reviewer for this constructive suggestion. We agree that the original description of the SHAP analysis did not provide sufficient implementation details, which may have limited reproducibility and interpretability. In the revised Sect. 2.7, we clarified that SHAP analysis was applied to the zone-specific XGBoost models selected in Sect. 2.6. Because the final explanatory analysis was based on XGBoost, we used the TreeSHAP implementation, which is designed for tree-based ensemble models. We further clarified that SHAP values were computed separately for the independent test samples of each functional zone using the corresponding trained zone-specific XGBoost model, rather than using the training samples or a single pooled model. To improve interpretation, we expanded the explanation of the SHAP decomposition. We clarified that the baseline, or bias, term returned by TreeSHAP represents the expected model output before assigning contributions to individual predictors. For each sample, the sum of this baseline prediction and all predictor-specific SHAP values equals the final XGBoost prediction. We also added a short numerical example to explain the meaning of positive and negative SHAP values. For example, in a simplified case with only three predictors, if the baseline predicted ω is 2.50 and the SHAP values of precipitation, temperature, and LAI for a given sample are +0.15, −0.05, and +0.10, respectively, the final model prediction is calculated as $2.50 + 0.15 - 0.05 + 0.10 = 2.70$. In this example, precipitation and LAI make positive contributions to the model-predicted ω, whereas temperature makes a negative contribution. This example was added to emphasize that the final model-predicted ω is determined by the combined SHAP contributions of all predictors, rather than by any single predictor alone.

In addition, we clarified the group-level SHAP analysis used to summarize the contributions of climatic, anthropogenic, and landscape predictor groups. For each sample and each predictor group, the aggregated SHAP value was calculated by summing the SHAP values of all predictors assigned to that group. A positive aggregated SHAP value indicates a positive net group-level contribution to the model-predicted ω relative to the zone-specific baseline, whereas a negative aggregated SHAP value indicates a negative net contribution. We also clarified that positive and negative SHAP values within the same group may offset each other.

We further revised the description of the pairwise comparison among predictor groups. Instead of interpreting these results as directional synergy or trade-off effects, we now describe them as pairwise directional alignment patterns based on the signs of aggregated SHAP values. Positive directional consistency indicates that two predictor groups both have positive aggregated SHAP values, negative directional consistency indicates that both groups have negative aggregated SHAP values, and divergence indicates opposite signs. We explicitly stated that these categories describe directional consistency or opposition in aggregated SHAP contributions and should not be interpreted as formal SHAP interaction effects or direct physical interactions.

Finally, following the reviewer's suggestion, we added a cautionary statement on the potential influence of predictor correlations on SHAP interpretation. We now state that correlations among hydroclimatic, anthropogenic, and landscape predictors may influence SHAP values because the contributions of correlated variables can be partially shared or redistributed among related predictors. Therefore, SHAP values are interpreted as model-based associations and relative contributions within the fitted zone-specific XGBoost models, rather than as independent causal effects of individual predictors or predictor groups.

Comment 12: The reviewer asked whether it was expected that neighboring catchments within a similar climatic context, such as catchments g, h, and i, exhibit very different ω values as well as markedly different temporal variability.

Response: We thank the reviewer for this insightful comment. We agree that the contrasting ω

values and temporal variability among neighboring catchments, such as g, h, and i, deserve further explanation. In the revised Sect. 3.1, we added a discussion to clarify that such differences are not necessarily unexpected. Although these catchments are geographically close and share a broadly similar climatic context, the Budyko parameter ω is not a purely climatic indicator. Rather, it represents an integrated catchment water-partitioning parameter that may also reflect catchment-specific landscape and anthropogenic characteristics. We therefore clarified that differences in topography, vegetation conditions, soil and storage properties, land use, and human water regulation may lead to distinct evaporative partitioning behavior and temporal stability, even among neighboring catchments. The observed contrasts among catchments g, h, and i suggest that local landscape and anthropogenic characteristics may modulate the climatic influence on catchment water partitioning. This point also provides additional motivation for the subsequent PCA-based functional zoning and SHAP-based interpretation, which were designed to examine zone-dependent associations between climatic, anthropogenic, and landscape predictors and the Budyko parameter ω.

Comment 13: The reviewer noted that the violin plots in Fig. 3 may be misleading because they imply a smoothed distribution extending beyond the observed annual values. The reviewer therefore suggested removing the violin plots, as the boxplots and individual data points already provide sufficient information.

Response: We thank the reviewer for this helpful suggestion. We agree that the violin plots may visually imply a continuous smoothed distribution extending beyond the observed values, especially given the limited number of moving-window estimates for each catchment. This could potentially lead to a misleading impression of the range and density of the estimated ω values. Following the reviewer's suggestion, we revised Fig. 3 (renumbered as Fig. 4 in the revised manuscript) by removing the violin plots. The lower panel now retains only the boxplots and individual data points, which provide a more direct and transparent representation of the distribution of ω across the 12 catchments. We also revised the figure caption accordingly by removing the description of violin plots and clarifying that the bottom panel shows the statistical distribution using boxplots and individual moving-window estimates.

Comment 14: The reviewer noted that several variables, such as P, AI, LAI, and AWSC, appear to be strongly correlated, and suggested discussing the potential impact of collinearity on both the PCA interpretation and the clustering results. The reviewer also suggested highlighting that the identified clusters are broadly consistent with the climatic zones shown in Fig. 1.

Response: We thank the reviewer for this insightful comment. We agree that several variables, particularly P, AI, LAI, and AWSC, show similar loading directions in the PCA space, suggesting potential collinearity among hydroclimatic, vegetation, and soil-storage variables. In the revised Sect. 3.2, we clarified that these variables should not be interpreted as independent single-variable effects. Instead, they jointly contribute to an integrated hydroclimatic–vegetation–soil-storage gradient that helps characterize differences among catchments. We further clarified that PCA reduces redundancy among correlated predictors by transforming the standardized variables into orthogonal principal components before clustering. However, correlated variables may still strengthen the influence of the dominant shared hydroclimatic–landscape gradient on the clustering results. Therefore, the PCA loadings and the resulting clusters are interpreted as reflecting combined environmental gradients rather than independent effects of individual predictors or classifications determined by a single variable. Following the reviewer's suggestion, we also revised Sect. 3.2 to highlight the broad consistency between the identified functional zones and the climatic zones shown in Fig. 1. Specifically, C1 corresponded approximately to the transition zone, C2 to the semi-humid zone, and C3 to the semi-arid zone. We clarified that this agreement is meaningful because the climatic-zone labels were not used as categorical input in the PCA-based clustering; instead, the three zones emerged from the long-term hydroclimatic, anthropogenic, and landscape predictors. We further explained this correspondence using the zone characteristics shown in Fig. 4d (renumbered as Fig. 5d in the revised manuscript). C2, located mainly in the semi-humid zone, was characterized by higher precipitation, vegetation cover, AI, and AWSC, indicating more favorable water availability. In contrast, C3, corresponding broadly to the semi-arid zone, showed stronger water limitation, with lower precipitation, lower AI, lower AWSC, and sparse vegetation. C1, located mainly in the transition zone, represented an intermediate hydroclimatic condition but was further distinguished by complex topography and relatively low temperature. These patterns indicate that the identified functional zones were broadly aligned with the climatic background, while their distinct topographic and anthropogenic characteristics suggest that the zoning results were not determined by climate alone. This consistency strengthens the interpretation of the identified clusters as hydrological functional zones rather than purely statistical groupings.

Comment 15: The reviewer asked why most predictors in Fig. 4a contribute in roughly similar proportions, between approximately 6.5% and 9%.

Response: We thank the reviewer for this helpful comment. We agree that the similar contribution proportions shown in Fig. 4a (renumbered as Fig. 5a in the revised manuscript) require further clarification, and we have clarified this point in the revised Sect. 3.2. The contribution values in Fig. 4a (renumbered as Fig. 5a in the revised manuscript) represent the relative PCA contribution of each standardized predictor to the PC1–PC2 space, rather than the importance of each predictor in explaining or predicting the Budyko parameter ω. Since PCA was performed on the z-score standardized predictor matrix, all variables entered the analysis with comparable initial variance. Therefore, differences in units or numerical ranges among variables did not cause any single predictor to dominate the PCA results. The contribution values were derived from the loading strength of each variable in the PC1–PC2 space. The fact that most contribution values ranged between approximately 6.5% and 9% indicates that many predictors had comparable representation in the first two principal components, rather than one predictor being particularly dominant. This pattern is consistent with our interpretation of PC1 and PC2 as integrated catchment-characteristic gradients. In other words, the PCA captured combined environmental gradients formed by multiple hydroclimatic, anthropogenic, and landscape attributes, rather than gradients determined by a single predictor. We also clarified that similar PCA contribution values do not mean that the variables have identical physical roles, nor do they imply equal importance for predicting ω. Instead, they indicate comparable representation in the reduced PCA space, while the different loading directions in Fig. 4a (renumbered as Fig. 5a in the revised manuscript) still provide useful information on how different predictors contribute to the integrated environmental gradients used for functional zoning.

Comment 16: The reviewer noted that the original interpretation of Fig. 5 was somewhat overstated, particularly the use of subjective terms such as "exceptional robustness" and "high accuracy". The reviewer also asked for clarification of what distributions were represented by the violin plots and requested that the number of points in each violin plot be indicated.

Response: We thank the reviewer for these helpful comments. We agree that the original wording used to describe model performance was somewhat subjective, and that the violin plots in Fig. 5 (renumbered as Fig. 6 in the revised manuscript) did not clearly show the underlying catchment-level sample size. In the original figure, the violin plots represented the distributions of catchment-level test performance values for each model within each functional zone. Each point corresponded to one individual catchment, with 5 catchments in C1, 3 catchments in C2, and 4 catchments in C3. Given these relatively small sample sizes, we agree that violin plots may imply smoothed probability distributions that are not fully supported by the available catchment-level observations. To improve clarity, we redesigned Fig. 5 (renumbered as Fig. 6 in the revised manuscript) by replacing the violin plots with catchment-level paired difference plots relative to XGBoost. In the revised Fig. 5 (renumbered as Fig. 6 in the revised manuscript), each point represents one catchment-level paired comparison, and the number of catchments is explicitly indicated for each functional zone: C1 ($n = 5$), C2 ($n = 3$), and C3 ($n = 4$). The top row shows ΔR^2, calculated as the R^2 value of XGBoost minus the corresponding R^2 value of RF, ANN, or SVM for each catchment. The bottom row shows ΔRMSE, calculated as the RMSE value of RF, ANN, or SVM minus the corresponding RMSE value of XGBoost for each catchment. Therefore, positive values indicate better performance of XGBoost for both metrics. We also revised Sect. 3.3 and the caption of Fig. 5 (renumbered as Fig. 6 in the revised manuscript) to explain the paired comparison more clearly and to use more neutral wording. Specifically, we removed subjective expressions such as "exceptional robustness" and "high accuracy" and instead provided quantitative support. The revised text reports that 35 of the 36 ΔR^2 values and 31 of the 36 ΔRMSE values were positive, indicating that XGBoost generally achieved higher R^2 and lower RMSE than the other benchmarked models across most catchment-level comparisons. At the same time, we clarified that the paired differences were not uniformly positive for all catchment-model pairs, particularly for RMSE. Therefore, the relative advantage of XGBoost is interpreted as a general tendency across most comparisons rather than as a universal improvement in every single case. Overall, the revised Fig. 5 (renumbered as Fig. 6 in the revised manuscript) provides a more transparent catchment-level comparison and supports the selection of XGBoost for the subsequent SHAP-based interpretation.

Comment 17: The reviewer suggested reminding readers that the reference ("actual") ω values were numerically inverted by matching the observed runoff with the theoretical runoff derived from Fu's equation. The reviewer also asked us to clarify what each individual dot represents in Fig. 6, and whether the values correspond to catchment-specific moving-window samples or cluster-averaged values. In addition, the reviewer noted that the statement "XGBoost was selected as the optimal regressor" was somewhat redundant with the previous sentence.

Response: We thank the reviewer for these helpful comments. We agree that the original caption and text did not sufficiently clarify the meaning of the "actual" ω values and the individual points in Fig. 6 (renumbered as Fig. 7 in the revised manuscript). In the revised caption of Fig. 6 (renumbered as Fig. 7 in the revised manuscript), we clarified that the "Actual Values" on the x-axis refer to the reference ω values numerically inverted from Fu's equation by matching observed runoff with theoretical runoff for each 11-year moving-window period, rather than directly observed ω measurements. We also clarified that each point represents one catchment-specific 11-year moving-window sample in the independent test set, rather than a cluster-averaged or zone-averaged value. Therefore, the points shown in each panel are pooled independent test-set samples from all catchments within the corresponding hydrological functional zone. We also revised the related main-text interpretation to avoid redundancy and overstatement. Specifically, we removed the repeated statement that XGBoost was selected as the optimal regressor, because model selection had already been stated in the preceding sentence. We further replaced stronger expressions such as "confirm this choice", "tightly clustered", and "ensures a solid physical basis" with more neutral wording. The revised text now states that the validation scatter plots further evaluated the predictive performance of the selected zone-specific XGBoost models and that the results provide additional support for using XGBoost in the subsequent SHAP-based interpretation.

Comment 18: The reviewer noted that the SHAP interpretation may be sensitive to collinearity among predictors, and that it was unclear whether the SHAP analyses were derived from models trained separately for each functional zone, from a single global model, or from catchment-specific models. The reviewer also pointed out that the positive SHAP associations of GDP and impervious surfaces appear counterintuitive from a hydrological perspective, because increasing impervious surfaces and human development would generally enhance runoff generation and reduce infiltration. The reviewer therefore suggested clarifying whether these relationships reflect direct physical mechanisms, indirect correlations with other environmental variables, or potential artifacts related to the limited sample size and predictor collinearity.

Response: We thank the reviewer for these important and constructive comments. We agree that the original manuscript did not sufficiently clarify the source of the SHAP analyses or the interpretive limitations of SHAP under predictor dependence. We also agree that the positive SHAP associations of GDP and impervious surface proportion in C3 require a more cautious and detailed explanation. First, in the revised Methods section and Sect. 3.4, we clarified that the SHAP analysis was conducted separately for the XGBoost model trained in each hydrological functional zone. Therefore, the SHAP values for C1, C2, and C3 were derived from the corresponding zone-specific XGBoost models using independent test-set samples. They were not derived from a single global model pooling all catchments, nor from separate catchment-specific models. This clarification makes the SHAP analysis consistent with the zone-specific modeling strategy described in Sect. 2.6. Second, following the reviewer's suggestion, we added a cautionary statement on the potential influence of predictor collinearity. Hydroclimatic, anthropogenic, and landscape predictors are not fully independent in real catchments. For example, precipitation, vegetation conditions, soil water storage, land use, socioeconomic development, and water-management activities may covary across space. Under such predictor dependence, SHAP values may partly share or redistribute contributions among correlated predictors. Therefore, we now interpret SHAP results as conditional, model-based associations and relative contribution patterns within the fitted zone-specific XGBoost models, rather than as direct physical mechanisms or independent causal effects of individual predictors. Third, we revised the interpretation of the positive SHAP associations of GDP and impervious surface proportion in C3. We agree with the reviewer that, from a local or event-scale hydrological perspective, impervious surfaces would generally be expected to reduce infiltration and enhance quick runoff generation. Therefore, a positive SHAP association between impervious surface proportion and model-predicted ω should not be interpreted as evidence that impervious surfaces directly increase ω through local infiltration or runoff-generation processes. Instead, we interpret this result from a long-term catchment-scale Budyko perspective. The ω analyzed in this study is a catchment-scale Budyko parameter inferred from long-term water-balance partitioning, rather than an event-scale metric of infiltration or runoff response. In this context, GDP and impervious surface proportion are better understood as indicators of broader anthropogenic modification and catchment development status. Catchments with higher GDP or greater impervious surface proportion may also differ in infrastructure construction, land-use intensity, water-resource regulation, ecological restoration, soil-conservation investment, irrigation practices, and other human-modified land – water conditions. These co-occurring conditions may influence long-term evapotranspiration–runoff partitioning and may be captured by the model as positive SHAP associations with model-predicted ω.

We therefore revised Sect. 3.4 and further discussed this issue in Sect. 4.2 to avoid a direct mechanistic interpretation. The positive SHAP associations of GDP and impervious surface proportion in C3 are now described as catchment-scale, model-inferred associations that may partly reflect indirect relationships among co-varying anthropogenic, landscape, and water-management conditions. At the same time, we acknowledge that the limited number of catchments and potential predictor dependence introduce uncertainty into this interpretation. Thus, these results should not be taken as independent causal evidence that socioeconomic development or impervious surfaces directly increase ω, but rather as an indication that anthropogenic modification is strongly associated with model-predicted long-term water-partitioning behavior in the C3 catchments.

Comment 19: The reviewer noted that Fig. 7 may lead to misinterpretation because the scales differ between subplots. In particular, the reviewer suggested that the x-axis ranges and color-bar scales should be kept consistent across clusters to facilitate a fair visual comparison.

Response: We thank the reviewer for this helpful comment. We agree that differences in x-axis ranges and color-bar scales among the SHAP summary plots could lead to visual misinterpretation if readers directly compare the apparent SHAP-value spread or color intensity across functional zones. In the revised Fig. 7 (renumbered as Fig. 8 in the revised manuscript) and its caption, we clarified how the SHAP summary plots should be interpreted. The SHAP analyses were conducted separately for the zone-specific XGBoost models of C1, C2, and C3. Because the predictor ranges and SHAP-value ranges differ among the three functional zones, using separate x-axis ranges helps preserve within-zone visual resolution and allows the directional SHAP patterns of each zone-specific model to be displayed more clearly. At the same time, we agree that this design limits direct visual comparison of apparent SHAP-value magnitudes across zones. To avoid misinterpretation, we revised the Fig. 7 (renumbered as Fig. 8 in the revised manuscript) caption to explicitly state that the x-axis ranges are scaled separately for each functional zone to improve within-zone visual resolution. We also clarified that cross-zone comparisons should be based on the reported predictor rankings and numerical mean absolute SHAP values, rather than on the apparent visual spread of SHAP values in the subplots. We further clarified the interpretation of the color bars. The color scale in each SHAP summary plot represents the relative feature-value distribution within the corresponding functional zone, rather than a common absolute feature-value scale across all zones. Therefore, colors should be interpreted within each zone-specific panel and should not be used for direct cross-zone comparison of absolute predictor values. These revisions make the interpretation of Fig. 7 (renumbered as Fig. 8 in the revised manuscript) more cautious and reduce the risk of overinterpreting differences among zones based solely on subplot scales.

Comment 20: The reviewer noted that the threshold values inferred from the SHAP dependence plots may be uncertain because of the limited number of points, potential sampling uncertainty, and the influence of the fitted smoothing curves. The reviewer

pointed out that the interpretation appeared somewhat overconfident, especially for variables such as POP and Slope in C2, where very few points are located near SHAP = 0. The reviewer therefore suggested that the reported threshold values should be interpreted more cautiously as approximate transition ranges rather than robust physical thresholds.

Response: We thank the reviewer for this helpful and important comment. We agree that the threshold values in the original manuscript were presented with excessive confidence, especially given the limited number of samples near the SHAP = 0 line for some predictors. In the revised manuscript, we replaced terms such as "critical tipping points", "threshold effects", "physical thresholds", and "optimal windows" with the more cautious expression "approximate SHAP-inferred transition points". We also clarified that these transition points refer to the approximate locations where the fitted SHAP dependence curves cross the SHAP = 0 line, indicating a change in the direction of the model-inferred SHAP contribution relative to the zone-specific baseline prediction. We further revised the interpretation of Fig. 8 (renumbered as Fig. 9 in the revised manuscript) to emphasize that these transition points are descriptive reference values derived from the fitted SHAP dependence patterns, rather than robust physical thresholds or deterministic hydrological tipping points. Their exact locations may be affected by the limited number of samples, the distribution of data points around SHAP = 0, potential sampling uncertainty, and the smoothing method used to visualize the SHAP dependence curves. This uncertainty is particularly relevant for variables such as POP and Slope in C2, where few samples are located near the SHAP = 0 line and the exact zero-crossing position is therefore difficult to determine robustly. We also revised the text to avoid causal or deterministic wording when describing these transition points. Instead of stating that a predictor "promotes", "suppresses", or "exerts an effect on" ω, we now describe whether its SHAP contribution is positive or negative relative to the corresponding zone-specific baseline prediction. Thus, the reported values for T in C1, LAI in C2, and GDP in C3 are interpreted as approximate SHAP-inferred changes in contribution direction, not as universal hydrological thresholds.

Finally, we revised the caption of Fig. 8 (renumbered as Fig. 9 in the revised manuscript) to make this interpretation clearer. The caption now states that the fitted curves are used to visualize general SHAP dependence patterns and that the marked transition points represent approximate SHAP-inferred changes in contribution direction. We also clarified that these points should be interpreted cautiously because their precise locations may be influenced by sample distribution and smoothing uncertainty. These revisions make the interpretation of the SHAP dependence plots more cautious and better aligned with the exploratory, model-inferred nature of the SHAP analysis.

Comment 21: The reviewer noted that several SHAP-based results were followed by proposed physical interpretations, such as explanations related to soil saturation, agricultural drainage, urbanization, or the urban heat island effect. The reviewer pointed out that SHAP does not directly reveal physical mechanisms, but rather highlights model-learned correlations that may be compatible with such hypotheses. The reviewer also asked how these interpretations might compare with those derived from a process-based hydrological model.

Response: We thank the reviewer for this thoughtful and important comment. We agree that the original manuscript sometimes linked SHAP dependence patterns too directly to specific physical mechanisms. SHAP values explain how predictors contribute to the output of the fitted XGBoost model, but they do not by themselves establish hydrological causality or identify process-level mechanisms. In the revised Sect. 3.5, we therefore removed or substantially weakened mechanistic statements following the SHAP dependence plots, including explanations related to soil saturation, agricultural drainage, intensified urbanization, and the urban heat island effect. We also revised causal or deterministic expressions such as "possibly because of", "likely because of", "promotes", "suppresses", and "exerts an effect on" ω. The revised text now describes these results as model-inferred SHAP dependence patterns and approximate SHAP-inferred transition points, rather than as direct evidence of specific physical mechanisms.

We also appreciate the reviewer's point regarding comparison with process-based hydrological models. A process-based hydrological model would provide a different and more mechanistic type of interpretation because it can explicitly represent hydrological fluxes and state variables, such as infiltration, soil moisture, evapotranspiration, runoff generation, groundwater exchange, routing, and water-resource regulation, depending on the model structure. Such models are therefore better suited for testing whether a proposed mechanism can physically reproduce observed hydrological responses. The objective of the present study is different. We do not aim to simulate event-scale runoff generation or to resolve the internal hydrological processes of the catchments. Instead, our analysis focuses on the Budyko parameter ω, which is an integrated long-term parameter describing catchment-scale evapotranspiration–runoff partitioning. In this sense, ω is process-informed but not process-resolving: it reflects the combined outcome of climatic, landscape, and anthropogenic influences on long-term water partitioning, but it does not separate individual processes such as infiltration, drainage, groundwater flow, routing, or water regulation. Within this framework, the XGBoost–SHAP analysis is used as a diagnostic tool to identify nonlinear, zone-dependent associations between predictors and the model-predicted ω. The physical explanations suggested by the SHAP patterns should therefore be viewed as hypotheses that are compatible with the data-driven associations, rather than as mechanisms directly demonstrated by SHAP. We have clarified this distinction in the revised manuscript and noted that future work combining SHAP-based diagnostics with process-based hydrological modeling would be valuable for testing whether the inferred associations correspond to physically consistent mechanisms.

Comment 22: The reviewer provided several suggestions regarding Fig. 8. Specifically, the reviewer suggested adding cluster labels (C1, C2, and C3) above each column to improve readability, using consistent y-axis limits for SHAP values and consistent color-bar limits for ω values across subplots to facilitate fair visual comparison, and further discussing the different SHAP response patterns of AI between C1 and C2.

Response: We thank the reviewer for these helpful suggestions. We agree that the readability and interpretation of Fig. 8 (renumbered as Fig. 9 in the revised manuscript) should be improved. First, following the reviewer's suggestion, we revised Fig. 8 (renumbered as Fig. 9 in the revised manuscript) by adding the labels C1, C2, and C3 above the three columns. This makes the functional-zone structure of the figure clearer and helps readers identify which SHAP dependence plots correspond to each zone-specific XGBoost model. Regarding the y-axis limits and color-bar limits, we agree that consistent scales can facilitate direct visual comparison across panels. However, the main purpose of Fig. 8 (renumbered as Fig. 9 in the revised manuscript) is to show the within-zone SHAP dependence pattern of each top-ranked predictor, especially the shape of the fitted dependence curve and the approximate location where it crosses the SHAP = 0 line. Because the SHAP-value ranges differ substantially among predictors and functional zones, using a single common y-axis limit for all panels would compress several dependence curves and make their nonlinear patterns and approximate SHAP-inferred transition points more difficult to discern. We therefore retained panel-specific y-axis limits to preserve the readability of the within-panel SHAP dependence patterns.

Similarly, we retained zone-specific color-bar limits for ω because the distributions of ω differ among the three functional zones. Applying a single global color scale would reduce the contrast within some zones and make it more difficult to distinguish samples with relatively higher or lower ω values within each zone. We clarified in the revised caption that the point colors provide supplementary information on the corresponding ω level of each sample, reflecting the combined predictor conditions, rather than the isolated contribution of the predictor shown on the x-axis. To reduce possible misinterpretation, we revised the Fig. 8 (renumbered as Fig. 9 in the revised manuscript) caption to explicitly state that the y-axis limits and ω color-bar limits are scaled separately to improve within-panel readability. We also clarified that the panels should be interpreted primarily in terms of within-zone SHAP dependence patterns and approximate SHAP-inferred transition points, rather than through direct visual comparison of axis ranges or color intensity across zones.

We also revised Sect. 3.5 to further discuss the different AI dependence patterns between C1 and C2. In this study, AI is defined as P/E_0 and represents the balance between precipitation supply and atmospheric evaporative demand. In C1, which corresponds broadly to the transition zone, AI displayed a nonmonotonic hump-shaped SHAP dependence pattern, with positive SHAP contributions mainly occurring at moderate AI values and negative contributions occurring at both lower and higher AI values. In contrast, in C2, which is mainly located in the semi-humid zone, the fitted AI dependence curve shifted toward negative SHAP contributions at higher AI values, with an approximate transition point around 0.62. This contrast suggests that the model-inferred contribution of AI depends on the functional-zone context. In other words, the same water–energy balance indicator did not show a uniform SHAP dependence pattern across zones. We therefore revised the text to interpret the AI transition values in Fig. 8

(renumbered as Fig. 9 in the revised manuscript) as zone-specific, approximate SHAP-inferred transition points within the corresponding fitted models, rather than as generally transferable hydrological thresholds. These revisions improve the readability of Fig. 8 (renumbered as Fig. 9 in the revised manuscript) and provide a more cautious interpretation of the contrasting AI response patterns between C1 and C2.

Comment 23: This part remains somewhat unclear to me. In particular, it is difficult to interpret what a "positive" or "negative" contribution practically implies in terms of hydrological behavior or changes in ω. In addition, many of the reported percentages in Fig. 9a appear relatively close to 50%, suggesting that positive and negative contributions occur with nearly similar frequencies. Similarly, in Fig. 9b, many interaction percentages are close to 33%, which may indicate an almost equiprobable distribution between positive, negative, and divergent cases. It would therefore be useful to discuss whether these differences are actually meaningful, how robust these interaction patterns are, if that was expected, and what they imply.

Response: We thank the reviewer for this important and constructive comment. We agree that the original interpretation of Fig. 9 (renumbered as Fig. 10 in the revised manuscript) was not sufficiently clear, especially regarding the practical meaning of positive and negative group-level SHAP contributions and the interpretation of percentages that are close to 50% or one third. We also agree that terms such as "interaction", "synergy", and "trade-off" could lead to overinterpretation. We have therefore revised the SHAP methodology, Fig. 9 (renumbered as Fig. 10 in the revised manuscript) caption, and the corresponding results text to clarify the meaning and limitations of these patterns. First, we clarified what positive and negative contributions mean in Fig. 9a (renumbered as Fig. 10a in the revised manuscript). They refer to the sign of the aggregated SHAP value of a predictor group, rather than to a direct physical increase or decrease in the observed or inferred ω. For each independent test sample, SHAP values were summed within the climatic, anthropogenic, and landscape predictor groups. A positive aggregated SHAP value indicates that the corresponding predictor group made a positive net contribution to the model-predicted ω relative to the baseline prediction of the corresponding zone-specific XGBoost model, whereas a negative aggregated SHAP value indicates a negative net contribution. In practical terms, these signs indicate whether a predictor group was associated within the fitted model with a higher or lower predicted ω for a given sample. However, the final model-predicted ω is determined by the baseline prediction plus the combined SHAP contributions of all predictor groups, rather than by any single group alone. Second, we clarified that the percentages in Fig. 9a (renumbered as Fig. 10 in the revised manuscript) represent the frequency of positive or negative net aggregated SHAP contributions across the independent test samples, not contribution magnitudes or physical effect sizes. Therefore, percentages close to 50% should not be interpreted as strong directional dominance. For example, when the positive and negative proportions of a predictor group are nearly balanced, this indicates mixed group-level contribution directions and sample heterogeneity, rather than a stable positive or negative tendency. We revised the text to make this interpretation explicit. Third, we revised the interpretation of Fig. 9b (renumbered as Fig. 10b in the revised manuscript). The original wording may have implied formal interaction effects, but Fig. 9b (renumbered as Fig. 10b in the revised manuscript) is not a formal SHAP interaction analysis. It is a pairwise directional-alignment analysis based on the signs of aggregated SHAP values. Positive directional consistency means that two predictor groups both had positive aggregated SHAP values in the same sample; negative directional consistency means that both groups had negative aggregated SHAP values; and divergence means that the two groups had aggregated SHAP values with opposite signs. We revised the caption and text to clarify that these categories describe directional consistency or opposition in net aggregated SHAP contributions, rather than formal SHAP interaction effects or direct physical interactions among predictor groups.

Regarding the robustness and meaning of these patterns, we now interpret Fig. 9 (renumbered as Fig. 10 in the revised manuscript) as an exploratory summary of zone-dependent and sample-dependent group-level SHAP contribution directions. Differences among C1, C2, and C3 are meaningful insofar as they show that climatic, anthropogenic, and landscape predictor groups did not have identical positive–negative contribution frequencies or pairwise directional-alignment patterns across functional zones. However, proportions close to 50% in Fig. 9a (renumbered as Fig. 10a in the revised manuscript) or close to one third in Fig. 9b (renumbered as Fig. 10b in the revised manuscript) do not support strong or robust directional regimes. This outcome is not unexpected because the predictor groups are not independent, and each group contains multiple predictors whose SHAP values may have different or even opposing signs. Thus, near-balanced cases indicate mixed group-level contributions and sample heterogeneity, rather than deterministic hydrological interactions. Overall, we revised the description and interpretation of Fig. 9 (renumbered as Fig. 10 in the revised manuscript) to emphasize that these patterns should be understood as model-inferred, group-level associations with higher or lower predicted ω, rather than as direct causal mechanisms. The main implication is that the combined contribution directions of climatic, anthropogenic, and landscape predictors were functional-zone dependent, while some predictor groups or group pairs showed weak or balanced directional tendencies rather than robust dominance.

Comment 24: The reviewer noted that the interpretation of Figs. 5 and 10 was not fully clear. The reviewer pointed out that the two figures appear to rely on related catchment-level performance information, with Fig. 5 mainly stratifying the results by functional zone. The reviewer also noted that the low-performance cases or outliers still appear to be present in Fig. 5, although the zoning helps identify that they are primarily associated with specific functional zones. Therefore, the reviewer suggested that the figures should be interpreted as showing structured performance heterogeneity rather than a clear disappearance of "failure points" or a robust validation of the proposed framework.

Response: We thank the reviewer for this insightful comment. We agree that our original interpretation was overstated. The comparison between Figs. 5 and 10 should not be interpreted as evidence that low-performance cases or "failure points" completely disappeared after functional zoning, nor as a definitive validation of the proposed framework. We have therefore revised Sect. 4.1 to provide a more cautious and accurate interpretation. In the revised manuscript, we clarified the different roles of Figs. 5 (renumbered as Fig. 6 in the revised manuscript) and 10 (renumbered as Fig. 11 in the revised manuscript). Fig. 5 (renumbered as Fig. 6 in the revised manuscript) presents catchment-level paired performance differences among the benchmarked algorithms within the zone-specific modeling framework, whereas Fig. 10 (renumbered as Fig. 11 in the revised manuscript) summarizes the catchment-level performance variability of models trained on the pooled dataset of all 12 catchments. Thus, the two figures both involve catchment-level performance information, but they are used to diagnose different aspects of model behavior under different modeling settings.

We also revised the interpretation of Fig. 10 (renumbered as Fig. 11 in the revised manuscript). Rather than describing it as showing a decomposition of model performance or as revealing “failure points”, we now describe it as a pooled-model benchmark that illustrates the variability of model performance across individual catchments. Although the pooled models may achieve satisfactory aggregate performance, their performance is not uniform across catchments, and several lower-performance outliers indicate that aggregate accuracy can mask structured catchment-level heterogeneity. We further clarified the role of Fig. 5 (renumbered as Fig. 6 in the revised manuscript) and the corresponding catchment-level performance results. We agree that relatively lower-performance cases were still present under the functional-zone-based framework. However, the catchment-level performance values indicate that these lower-performance cases were generally less severe than the low-performance outliers observed under the pooled-model setting. Therefore, the main implication is not that functional zoning eliminated all low-performance cases, but that it helped organize performance heterogeneity according to hydrological functional zones and reduced the severity of lower-performance cases. Accordingly, we revised Sect. 4.1 to avoid overstrong expressions such as “critical failure”, “failure points”, “sharp contrast”, and “robustly validate”. The revised text now interprets Figs. 5 and 10 (renumbered as Figs.6 and 11 in the revised manuscript) as complementary diagnostics: the pooled-model benchmark highlights catchment-level performance heterogeneity, while the zone-specific modeling results indicate that such heterogeneity is partly structured by functional-zone differences and that the lower tail of model performance can be improved under the functional-zone-based framework. This interpretation better reflects the evidence provided by the figures without overstating the robustness or generality of the proposed framework.

Comment 25: The reviewer suggested presenting Fig. 10 as two panels, one for each performance metric, and adding reference lines to indicate the optimal values of the metrics, namely R^2 = 1 and RMSE = 0.

Response: We thank the reviewer for this helpful suggestion. We revised Fig. 10 (renumbered as Fig. 11 in the revised manuscript) accordingly to improve the clarity of the performance comparison. In the revised figure, the two performance metrics are now shown in separate panels. The left panel presents the distribution of catchment-level R^2 values, and the right panel presents the distribution of catchment-level RMSE values. This separation avoids placing two metrics with different meanings and optimal directions on the same axis and makes the interpretation of each metric clearer. Following the reviewer's suggestion, we also added dashed horizontal reference lines to indicate the optimal value of each metric, with $R^2 = 1$ in the R^2 panel and RMSE = 0 in the RMSE panel. These reference lines help readers more directly assess the relative distance between the model performance and the optimal value for each metric. In addition, we revised the caption of Fig. 10 (renumbered as Fig. 11 in the revised manuscript) to improve clarity. The revised caption now explains that the boxplots summarize catchment-level performance variability under the pooled-model setting. We also clarified that the black dots indicate catchment-level outliers with relatively low predictive performance, and revised wording such as "failed significantly" to avoid overinterpreting these outliers as categorical model failures.

We again thank the reviewer for the constructive and detailed comments. These comments helped us improve the methodological transparency, figure presentation, terminology, and interpretation of the machine-learning and SHAP-based results. We hope that the revised manuscript and the detailed responses above adequately address the reviewer's concerns.