

Dear Referee,

Thank you for your careful and constructive assessment of our manuscript entitled “A reference-basin based framework for identifying and transferring permafrost hydrological coupling regimes across China”. We have revised the manuscript thoroughly in response to your comments.

We address each comment point by point below. The referee’s comments are labeled as RC, and our responses are labeled as AR. Our responses are shown in **blue font**, and the corresponding changes in the revised manuscript are indicated with line numbers where appropriate.

Sincerely,

Lei Song, on behalf of all coauthors

College of Hydrology and Water Resources, Hohai University

Email: songlei530@126.com

Responses to the referee's major Comments

1. **RC1: Spatial interpretation of correlations with basin outlet runoff.** The spatial meaning of the DCD and *C* fields within the reference basin is not sufficiently clear. Soil ice is evaluated for individual spatial units at 5 km resolution, but the hydrological response variable is the same runoff record from the Tangnaihahi basin outlet. Consequently, the analysis does not directly measure a local runoff response at each grid cell. A high soil ice and runoff correlation at a particular location may indicate that soil ice variability there covaries with the integrated basin response, but it does not by itself demonstrate that this location or soil depth exerts dominant hydrological control on outlet runoff. This distinction is important because the manuscript subsequently interprets the spatial patterns in terms of shallow and deep hydrological control. The authors should state explicitly what can and cannot be inferred when spatially distributed soil ice series are related to a single basin integrated discharge series. The terminology in the Results and Discussion should also distinguish statistical association with basin runoff from local hydrological control.

AR1: We thank the reviewer for pointing out that the spatial interpretation of DCD and *C* required further clarification. In the revised manuscript, we now explicitly state that, for each 5 km spatial unit within the reference basin, the soil ice loss series was correlated with the basin-outlet runoff series observed at Tangnaihahi Station. DCD represents the weighted characteristic depth of the positive correlation contributions across the soil profile, whereas *C* represents the aggregated magnitude of these positive contributions across the soil layers and lag range considered.

We also reviewed and revised the relevant statements in the Results and Discussion. Throughout the manuscript, DCD and *C* are now consistently described as characterizing the statistical association structure between soil ice changes at different spatial locations and runoff at the basin outlet. Interpretations involving shallow and deep layer hydrological processes have accordingly been made more

cautious. In the Discussion, these processes are presented as possible physical explanations that are consistent with previous studies of permafrost hydrology.

2. RC2: Validation of the national transfer model. The reported overall accuracy of 58.15% and Kappa coefficient of 0.41 indicate moderate agreement, but the procedure used to obtain these statistics is not described clearly enough to assess the performance of the national transfer model. It is not clear whether these values are calculated from the observations used to fit the GAM, from a withheld subset, from cross validation, or from a spatially independent validation. This distinction is particularly important for gridded environmental data because neighboring cells are spatially correlated, and model performance estimate from randomly divided spatial samples may differ substantially from performance in geographically independent regions. The manuscript should state exactly how the predictions used to calculate accuracy and Kappa were generated. Class specific performance or a confusion matrix should also be reported because overall accuracy alone cannot show whether some archetypes are predicted substantially more reliably than others. Since the GAM produces class probabilities, information on the confidence of individual assignments would also help distinguish well supported regime cores from uncertain transition areas. With the reported performance, the national classification should be interpreted primarily at the regional scale rather than as a precise grid cell representation of subsurface hydrological state.

AR2: We thank the reviewer for requesting clarification of how the reported overall accuracy and Kappa coefficient were obtained. In the revised manuscript, we now specify the reference classes, the archetype assignments used for evaluation, and the spatial cross validation procedure. We also report the confusion matrix, performance for each class, and classification probability diagnostics.

The reference classes were the four trajectory archetype labels assigned to spatial units in the SRYR from their complete eight window DCD-C trajectories. Classification performance was evaluated using five-fold spatial cross validation. The 607 spatial units in the reference basin were divided into five folds based on

0.5° spatial groups. In each fold, the model was fitted using the other four folds and then used to generate class probabilities for the withheld spatial units. The archetype with the highest probability was taken as the assignment for each withheld unit. After all five folds were completed, the resulting out of fold archetype assignments for all spatial units were compared with their corresponding reference archetype labels. The pooled results were then used to calculate the overall accuracy and Cohen's kappa. This procedure yielded an overall accuracy of 58.15% and a Kappa coefficient of 0.41.

Following the reviewer's suggestion, we further added the complete confusion matrix together with precision, recall, and F1 score for each archetype. For Types 1-4, precision was 0.469, 0.553, 0.658, and 0.611, respectively; recall was 0.328, 0.699, 0.727, and 0.400; and F1 score was 0.386, 0.617, 0.691, and 0.484. The corresponding confusion patterns and performance for each class are summarized in the new Table S4.

We also added the maximum class probability, p_{max} , and the probability margin, Δp , defined as the difference between the largest and second largest class probabilities, to describe the within model decisiveness of individual archetype assignments. Across the national mapping domain, the median p_{max} was 0.461 and the median Δp was 0.136. Their distributions among archetypes and between units within and outside predictor support are shown in Fig. S7 and Table S3(b).

The revised manuscript therefore interprets the national classification primarily at the regional scale and uses more cautious language for assignments at individual grid cells and for fine spatial archetype boundaries.

Table S3. Predictor-space OOD threshold sensitivity and classification probability diagnostics.

(a) Sensitivity of predictor-space OOD coverage to the reference-support threshold

Threshold basis	Threshold	National OOD (%)	Area-weighted OOD (%)	QTP OOD (%)	NE OOD (%)	Agreement	Jaccard
99th percentile	4.122	66.2	65.0	57.3	100.0	0.708	0.558
99.5th percentile	4.755	59.1	57.7	48.4	100.0	0.778	0.625
Reference maximum	7.667	36.9	34.6	20.4	100.0	1.000	1.000

(b) Classification probability diagnostics by archetype and predictor support status.

Group	n	pmax median [IQR]	Probability margin median [IQR]
Type 1	4791	0.444 [0.379, 0.492]	0.075 [0.038, 0.126]
Type 2	7424	0.449 [0.381, 0.500]	0.116 [0.049, 0.211]
Type 3	2609	0.381 [0.346, 0.440]	0.126 [0.068, 0.201]
Type 4	9459	0.496 [0.436, 0.515]	0.186 [0.111, 0.215]
Type 4 within predictor support	1966	0.444 [0.385, 0.514]	0.138 [0.060, 0.254]
Type 4 OOD	7493	0.500 [0.459, 0.515]	0.188 [0.135, 0.210]

Note: The reference maximum (7.667) is used as the main OOD threshold. Agreement and Jaccard are calculated relative to this main mask. The probability margin equals the largest minus the second-largest class probability. Type 4 contains 1966 cells within predictor support, equivalent to 12.8% of all supported cells and 20.8% of all Type 4 cells.

Table S4. Confusion matrix and class-specific performance of the reference-basin classification evaluation.

(a) Confusion matrix. Rows indicate the reference trajectory archetypes, and columns indicate the out of fold archetype assignments.

Reference archetype	Classified as Type 1	Classified as Type 2	Classified as Type 3	Classified as Type 4	Support
Type 1	38	54	19	5	116
Type 2	31	146	20	12	209
Type 3	9	27	125	11	172
Type 4	3	37	26	44	110
Classified total	81	264	190	72	607

(b) Class-specific precision, recall, and F1 score.

Archetype	Precision	Recall	F1 score	Support
Type 1	0.469	0.328	0.386	116
Type 2	0.553	0.699	0.617	209
Type 3	0.658	0.727	0.691	172
Type 4	0.611	0.400	0.484	110

Note: The evaluation used out of fold archetype assignments obtained from five fold spatial cross validation for 607 spatial units in the reference basin, with folds constructed from 0.5° spatial groups. Each unit was classified only when its spatial group was withheld. Overall accuracy was 58.15%, and Cohen’s kappa was 0.41. Precision is the fraction of assignments to an archetype that were correct; recall is the fraction of reference units of that archetype that were correctly assigned.

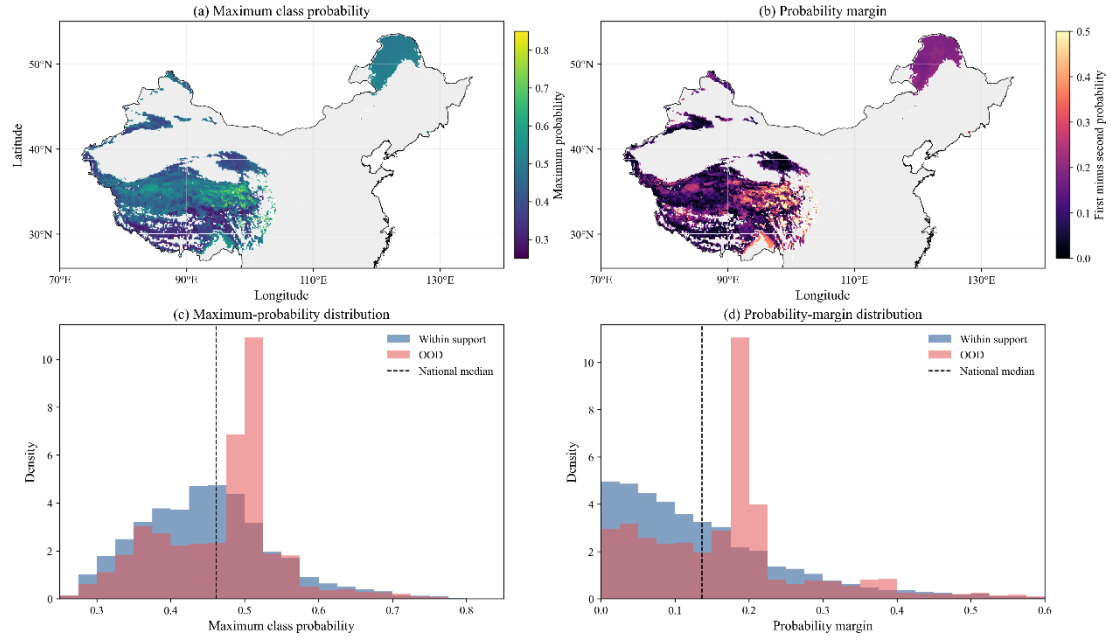


Figure S7. Spatial distribution and frequency distributions of classification probability diagnostics for the national archetype transfer. (a) Maximum class probability. (b) Probability margin between the largest and second-largest class probabilities. (c) Distribution of maximum class probability within and outside predictor support. (d) Distribution of probability margin within and outside predictor support. Larger values indicate a clearer model preference for the selected archetype and greater separation from the closest alternative, respectively.

3. RC3: OOD screening and the applicability boundary. The OOD analysis is central to the proposed applicability control, but the mathematical procedure used to define OOD conditions is not provided. The manuscript states that cells outside the environmental support represented by the reference basin are identified as OOD, but it does not specify how predictor space is standardized, what distance or range criterion defines environmental support, whether the predictors are considered independently or jointly, what numerical threshold is used, or how that threshold is selected. These details are necessary for reproducibility because the extent of the OOD domain depends directly on the definition of environmental support. I therefore recommend that the authors provide the complete OOD criterion and show how the mapped applicability boundary responds to reasonable changes in the selected threshold. The fraction of each regional domain classified as OOD would also be useful. The interpretation of predictions within OOD regions should remain clearly separated from predictions within the environmental support of the reference basin, since a numerical class can still be produced by the GAM even when the target conditions fall outside the predictor space represented by the training data.

AR3: We thank the reviewer for pointing out that the original manuscript did not provide sufficient detail on the mathematical definition, threshold selection, and sensitivity of the OOD screening. In the revised manuscript, we added the complete calculation procedure for OOD in predictor space, together with a threshold sensitivity analysis and regional OOD fractions.

For the environmental predictors used in the regional classification, each variable was standardized relative to its corresponding distribution in the reference basin. For target grid cell i , the OOD score was defined as

$$S_i = \max_{\omega, j} |z_{i\omega j}|$$

where ω denotes the eight time windows, j denotes the environmental predictors, and $z_{i\omega j}$ is the standardized value for the corresponding predictor and time window. This score represents the maximum standardized deviation of a target unit from the predictor space represented by the reference basin.

The primary OOD threshold was set to 7.667, the empirical maximum of S_i across the reference basin. We also evaluated the 99th and 99.5th percentiles of the reference S_i distribution as alternative thresholds, corresponding to 4.122 and 4.755, respectively.

Under the 99th percentile, 99.5th percentile, and reference maximum thresholds, the national OOD fractions were 66.2%, 59.1%, and 36.9%, respectively. The corresponding fractions for the Qinghai Tibetan Plateau were 57.3%, 48.4%, and 20.4%. The northeastern permafrost region was classified as 100% OOD under all three thresholds. The complete threshold sensitivity results are provided in Table S3(a) in AR2.

The revised Figure 3 shows the trajectory archetype classification together with the OOD environmental support overlay. The Results and Discussion also distinguish the distributions and interpretation of archetype assignments within predictor support and in OOD regions.

4. **RC4: Temporal transition analysis in Figure 4.** The procedure used to construct the early and late period classifications in Section 3.3 needs to be explained explicitly in the Methods. The four archetypes are initially defined by clustering complete DCD and C trajectories constructed from eight successive time windows. Figure 4 then assigns regime types separately to the first three and last three windows and interprets differences between these classifications as transitions such as Type 2 to Type 4 and Type 4 to Type 3. It is not clear how an archetype defined from a complete eight window trajectory is assigned to a shorter early or late subset. It is also unclear whether the period specific classifications are obtained directly from DCD and C , from the GAM using period specific environmental predictors, or through another procedure. These alternatives have different meanings. In particular, a change in statistically predicted class resulting from changes in environmental predictors is not equivalent to directly observing a transition between independently identified DCD and C trajectories. The exact procedure used to generate the early and late classifications and the transition percentages in Figure 4

should therefore be stated clearly, and the interpretation of these changes should match the actual classification procedure.

AR4: We thank the reviewer for requesting clarification of how the early and late stage assignments in Figure 4 were obtained. In the revised manuscript, we revised and clarified the stage diagnostic procedure used for Figure 4 and now provide its complete calculation steps.

The four archetypes were defined from the complete DCD – C trajectories in the reference basin. The stage diagnostic retained the Type 1 – 4 label system used in the full period national classification shown in Figure 3. For the Figure 4 analysis, we retained the time varying environmental predictors used in the national classification framework, namely TPD, TPI, VD, and precipitation, for each of the 8 time windows.

For each grid cell, the four predictors across eight windows formed a 32 dimensional environmental trajectory. Each predictor window dimension was standardized by subtracting its median and dividing by 1.4826 times its unscaled median absolute deviation, calculated across all valid mapped permafrost grid cells. The same dimension specific parameters were used for the full trajectories and both stage comparisons.

The standardized trajectories were then grouped according to the full period Type 1-4 assignments in Figure 3. For each archetype, the median value in each trajectory dimension was calculated to form a fixed full period environmental trajectory representative. The same four fixed representatives were used for both stages.

The early stage used W0-W2, whereas the late stage used W5-W7. Each stage therefore contained 12 dimensions, comprising four predictors over three windows. For each grid cell, the corresponding stage specific environmental trajectory was compared with the matching dimensions of the four fixed representatives using Manhattan distance,

$$d_{ij}^{(s)} = \sum_{k \in K_s} |z_{ik} - T_{jk}|$$

where s denotes the early or late stage, K_s is the set of environmental trajectory dimensions used for that stage, z_{ik} is the standardized value for grid cell i , and T_{jk} is the corresponding value of the fixed representative for archetype j . The archetype with the smallest distance was assigned to the grid cell for that stage.

Early to late changes were determined by comparing the stage assignments of the same grid cell. Transition percentages were calculated from cell counts using all mapped permafrost grid cells with valid assignments in both stages as the denominator. OOD status was overlaid only to indicate environmental support.

Responses to the referee's Minor Comments

5. RC5: The manuscript should provide the exact soil layer depths used to calculate C and DCD. Since C is calculated by summing contributions over the soil profile, the vertical discretization is relevant to the numerical value of the metric.

AR5: We thank the reviewer for pointing out that the vertical discretization used to calculate DCD and C should be stated more clearly. In the revised Methods, we now provide the exact soil depths used in the calculations and clarify how these discrete layers enter the definitions of the two metrics.

Both DCD and C were calculated over the 0-2 m soil profile simulated by WEB-DHM-pf. The discrete soil depths were 0, 20, 40, 60, 80, 100, 120, 140, 160, 180, and 200 cm, giving 11 depth levels at 20 cm intervals. C was defined as the discrete sum of positive correlations across all soil depths and the considered lag range:

$$C = \sum_z \sum_{\tau=0}^{\tau_{max}} r^+(z, \tau)$$

where

$$r^+(z, \tau) = \max[r(z, \tau), 0]$$

z denotes the discrete soil depth and τ denotes the lag in months. DCD was calculated as the depth weighted centroid of the same positive correlation

structure:

$$DCD = \frac{\sum_z z \sum_{\tau=0}^{\tau_{max}} r^+(z, \tau)}{\sum_z \sum_{\tau=0}^{\tau_{max}} r^+(z, \tau)}$$

Thus, the contributions to C and the depth weighting used for DCD were both based on the same set of discrete depths over the 0-2 m profile.

6. RC6: Figure 2 labels the horizontal axis as Time Window (0 to 7). Please provide the corresponding calendar periods in the caption or axis labels.

AR6: We thank the reviewer for the suggestion. We have added the corresponding calendar periods for Windows 0-7 to the caption of Figure 2:

W0: 1964–1983;

W1: 1969–1988;

W2: 1974–1993;

W3: 1979–1998;

W4: 1984–2003;

W5: 1989–2008;

W6: 1994–2013;

W7: 1999–2018.

All windows span 20 years, with a 5 year step between successive windows.

7. RC7: Please state the exact calendar periods included in the early and late periods shown in Figure4.

AR7: We thank the reviewer for the suggestion. We have clarified the time windows and corresponding calendar periods used for the early and late stages in the Methods and the caption of Figure 4.

The early period includes W0-W2, corresponding to 1964-1983, 1969-1988, and 1974-1993. The late period includes W5-W7, corresponding to 1989-2008, 1994-2013, and 1999-2018.

8. RC8: The terms “type,” “regime,” “coupling regime,” and “trajectory archetype” are used interchangeably. I suggest selecting one term and using it consistently.

AR8: We thank the reviewer for pointing out the inconsistent terminology. We have

reviewed and standardized the relevant terms throughout the manuscript.

In the revised manuscript, “trajectory archetype” is used primarily to refer to the four statistical archetypes identified from the complete DCD-C trajectories, and “archetype” is used as a shortened form where the context is clear. “Permafrost hydrological coupling regime” is retained only for the broader conceptual description of the study framework.

For Figure 4, the early and late results are described as “stage specific archetype assignments” to reflect their assignment relative to fixed full period trajectory representatives.

9. RC9: A compact table reporting the characteristic DCD, C, temporal behavior, and spatial proportion of the four reference basin archetypes would make the classifications easier to compare quantitatively.

AR9: We thank the reviewer for the suggestion. We have added Table 1 to provide a compact quantitative summary of the four reference basin trajectory archetypes, including their area weighted proportions, characteristic DCD and C values, ranges across the 8 time windows, and first to last changes.

Type 2 had the largest area weighted proportion at 34.3%, followed by Type 3 at 28.5%, Type 1 at 19.1%, and Type 4 at 18.1%. The mean DCD values for Types 1-4 were 1.057, 1.162, 1.276, and 1.205 m, respectively, while the corresponding mean C values were 1.261, 1.695, 1.740, and 1.357. Table 1 also reports the DCD and C ranges across the eight windows and their first to last changes, allowing the four archetypes to be compared directly in terms of spatial proportion, characteristic depth, correlation magnitude, and temporal behavior.

Table 1. Quantitative characteristics of the four DCD–C trajectory archetypes in the SRYR.

Type	Area-weighted proportion (%)	Mean DCD (m)	DCD window range (m)	Mean C	C window range	First-to-last change in DCD	First-to-last change in C
Type 1	19.1	1.057	0.924–1.177	1.261	0.758–1.666	−0.036	−0.380
Type 2	34.3	1.162	0.946–1.321	1.695	1.037–2.369	−0.005	−0.694
Type 3	28.5	1.276	1.174–1.379	1.740	1.014–2.480	−0.065	−1.096

Type	Area-weighted proportion (%)	Mean DCD (m)	DCD window range (m)	Mean C	C window range	First-to-last change in DCD	First-to-last change in C
Type 4	18.1	1.205	0.979–1.339	1.357	0.891–1.874	−0.092	−0.922