Major comments:

1. **The authors should provide additional details on the Landsat imagery used (line 96), including spatial/temporal resolution and how these images were selected.**

Response: We agree with the referee that providing these details improves the clarity of our methodology. In the revised manuscript, we will expand Section 2.2.1 to include specific details regarding the Landsat data.

Specifically, we used Landsat 5/7/8/9 surface reflectance imagery, which features a spatial resolution of 30 meters and a 16-day revisit cycle. Regarding the image selection process, we applied two strict criteria: first, to avoid the influence of ice and snow cover, we constrained our image selection strictly to the ice-free period (April to October) for each year; second, to ensure high-quality observations, we applied a strict cloud cover threshold (< 10%) to filter out heavily contaminated images.

2. **At line 107, the authors state that SR, SSR, and snowmelt were aggregated annually. How the remaining CEF variables were treated? Did they preserve their original temporal resolution?**

Response: The referee correctly points out the need for clarity regarding the temporal treatment of the driving factors. The original high temporal resolutions of CEF variables were not preserved, as our target variable (surface water area) was evaluated on an interannual (yearly) basis. The CEF variables were categorized and processed into three distinct temporal scales:

Cumulative/Flux variables (SR, SSR, Snowmelt, and SWA): These were aggregated as annual totals. In the high-latitude SHRZ, winter precipitation is stored as snowpack and melts in spring. This winter accumulation directly contributes to the surface water extent observed during our April – October window. Using annual totals ensures the complete hydrological input for the year is captured.

Surface environmental state variables (FVC, LST, and SSM): These were calculated as mean values strictly during the ice-free period (April – October). In high-latitude frozen ground regions, the surface system enters a state of physical and ecological dormancy during winter.

Subsurface storage variable (GWSA): GWSA was calculated as the annual mean anomaly. Even during the winter when the surface is ice-sealed, the groundwater system remains active and maintains continuous interactions with surface water. Therefore, using the annual mean is methodologically sound; it filters out strong seasonal noise and provides a reliable representation of the overall groundwater baseline for each year, which is necessary for evaluating the long-term trade-offs between surface water and groundwater.

We will detail these distinct temporal aggregation methods and their ecohydrological

justifications in Section 2.2.2 of the revised manuscript.

3. **All driver datasets begin after 2000, while Landsat images span 1988-2024. The temporal mismatch between trend estimation for surface water (1988–2024) and for drivers (2000–2020) should be addressed. The authors should either justify this inconsistency or harmonize the time periods. In addition, datasets in Table 1 have heterogeneous spatial resolutions. How were they spatially harmonized and resampled? Please, provide this information in the Materials and Method section.**

Response: We thank the referee for raising this important point regarding our temporal and spatial data framework.

Regarding the temporal mismatch: We will explicitly justify this dual-timeframe strategy in the revised methodology, as the two periods serve two distinct but complementary scientific objectives within our study: (1) Reconstructing the Long-Term Trajectory (1988-2024): Using the longest available observation record from Landsat ensures the robustness and reliability of our surface water area trend and breakpoint analysis. This extended timeframe is necessary to capture the full, long-term trajectory of the hydrological reversal from shrinkage to expansion around 2012. (2) Attribution Modeling (2000–2020): While we can reconstruct surface water area back to 1988, our quantitative attribution modeling (G-XGBoost) requires a comprehensive suite of multi-source spatial drivers, such as groundwater storage anomalies (GWSA), surface soil moisture (SSM), human footprint (HF), and irrigated cropland area (ICA). Because these key spatial datasets and satellite observations are restricted to the post-2000 era, the causal attribution analysis must be objectively constrained to the 2000–2020 period to ensure data consistency and model reliability. This approach of using a longer target record alongside shorter driver datasets follows previous remote sensing hydrological studies (We will add these references to the revised version).

Regarding spatial harmonization: To resolve the heterogeneous spatial resolutions (ranging from 30 m to 0.1°), all datasets were harmonized to a unified 1 km grid prior to trend calculation and modeling. Continuous variables (e.g., LST, SSM, and topography) were resampled using bilinear interpolation, while categorical variables (e.g., Soil Texture Type) were resampled using the nearest-neighbor method.

We will add a dedicated paragraph detailing these temporal justifications and spatial preprocessing steps to Section 2.3 in the revised manuscript.

4. **Line 124 introduces the WI+VI paradigm, which needs a brief conceptual explanation. What does WI+VI stand for, what is the extraction index, and why is this method chosen?**

Response: We agree that the "WI+VI" paradigm requires a clearer conceptual introduction. In

the revised manuscript, we will expand Section 2.3.1 to explicitly address these three points.

First, "WI+VI" stands for the combination of Water Indices (WI, e.g., MNDWI, NDWI) and Vegetation Indices (VI, e.g., NDVI, EVI).

Second, regarding why this method was chosen: our study area (the SHRZ) contains extensive croplands that share high spectral similarity with natural water bodies during certain phenological stages. We chose this paradigm because it relies on the physical principle that true open water typically exhibits higher WI values than VI values, whereas vegetation or mixed pixels (such as flooded croplands) exhibit the opposite. This makes it highly effective at suppressing agricultural noise.

Third, based on this paradigm, our specific "extraction index" is a composite logical rule. A pixel is extracted as water only when it satisfies strict lower bounds alongside the WI+VI condition: MNDWI > 0.1, NDWI > -0.1, EVI < 0.1, and (MNDWI > NDVI or MNDWI > EVI).

5. Several classification thresholds, such as the 10% cloud coverage or the 75% water-frequency threshold used to differentiate seasonal and permanent water, require justification. Which criteria did the authors use to determine these choices?

Response: We appreciate the referee highlighting the need for explicit justifications regarding these empirical thresholds. In the revised manuscript, we will incorporate robust literature support and the underlying rationale for these choices.

First, the 10% cloud coverage threshold is a widely adopted standard in optical remote sensing time-series analysis. We selected this specific threshold to strike an optimal balance: it rigorously minimizes cloud and cloud-shadow contamination while retaining a sufficient density of valid observations required to continuously monitor surface water dynamics during the ice-free period.

Second, the 75% water-frequency threshold used to differentiate permanent from seasonal surface water is also a widely adopted standard in long-term surface water mapping. In high-latitude frozen ground regions, employing a strict 100% frequency threshold to define permanent water is often impractical due to occasional valid observation gaps and short-term fluctuations at the ice-free margins. A widely accepted frequency of $\geqslant$75% effectively isolates water bodies that maintain water for the vast majority of the observable months, thereby reliably distinguishing stable structural water (permanent) from ephemeral or precipitation-driven inundation (seasonal).

We will explicitly state these rationales and add the corresponding citations in Sections 2.2.1 and 2.3.3 of the revised text to ensure full methodological transparency.

6. **Lines 159–160 state that “a significant excursion from zero” indicates a shift in hyrological regime. The authors should quantify what constitutes a “significant excursion”.**

Response: We thank the referee for pointing out this methodological ambiguity. We agree that the phrase "a significant excursion from zero" was overly descriptive and lacked explicit statistical criteria in the text.

To clarify, in our original analysis, the breakpoint was not determined arbitrarily; it was mathematically located using a binary segmentation algorithm to find the point of maximum cumulative deviation. However, we acknowledge that this mathematical inflection point lacked an explicit p-value to test its statistical significance.

To strictly quantify what constitutes a "significant excursion" and resolve this gap, we have now upgraded our methodology by implementing a Statistical CUSUM Test coupled with bootstrap permutation. The statistical significance of the maximum cumulative deviation is now mathematically quantified by randomly shuffling the time-series data by 1,000 bootstrap iterations to generate a null distribution. An excursion is now strictly defined as a "significant regime shift" only when the calculated permutation *p*-value is less than 0.05.

In the revised manuscript, we will remove the vague phrasing and update Section 2.3.4 to detail this rigorous quantitative standard.

7. **The authors should report the numerical slope ranges used to classify grid-cell trends as significant or insignificant increase/decrease, or as stable. It should also be clarified whether “insignificant” refers to statistical insignificance or to a change of very small magnitude.**

Response: We appreciate the referee pointing out this ambiguity. The term "insignificant" in our study strictly refers to statistical non-significance based on the Mann-Kendall test, rather than a change of small magnitude. To prevent any future misinterpretation, we will systematically replace the ambiguous term "insignificant" with "statistically non-significant" throughout the revised manuscript.

Furthermore, we will explicitly report the exact mathematical criteria used for the grid-cell trend classification in Section 2.3.4. For the Mann-Kendall test, the absolute value of the standardized test statistic ($|Z|>1.96$) corresponds to the significance level of $p<0.05$ (two-tailed test). Accordingly, our five spatial trend categories are rigorously defined by combining the Theil-Sen slope and the Mann-Kendall Z statistic as follows:

Significant increase: Slope > 0 and $|Z|>1.96$

Statistically non-significant increase: Slope > 0 and $|Z|\leqslant 1.96$

Stable: Slope = 0

Statistically non-significant decrease: Slope < 0 and |Z|⩽1.96

Significant decrease: Slope < 0 and |Z|>1.96

These explicit classification rules will be thoroughly integrated into the methodology section to ensure full transparency and reproducibility.

8. **The identification of the breakpoint around 2012 should be supported with statistical evidence, such as confidence intervals, or comparisons with alternative breakpoint detection approaches that justify selecting 2012 as a meaningful regime shift.**

Response: The referee makes an excellent point. Relying solely on the mathematical inflection point of the original CUSUM algorithm without an explicit p-value was indeed insufficient to robustly confirm 2012 as a regime shift. Taking the referee's valuable advice, we have conducted a rigorous dual-verification approach to provide solid statistical evidence.

Given that hydrological regime shifts in cold regions are often cumulative, long-term structural transitions, we retained the Statistical CUSUM Test (detailed in our response to Comment 6) as our primary analytical tool. Our newly computed permutation-based CUSUM test mathematically confirms that the maximum cumulative deviation for the total surface water area (TSW) occurring in 2012 is indeed statistically significant ($p = 0.0010$).

Furthermore, to address the referee's suggestion of using alternative approaches, we utilized the non-parametric Pettitt's test as an auxiliary verification method. The Pettitt test independently identified highly significant breakpoints for the permanent water components (TSW, PSW, S-PSW) tightly clustering between 2010 and 2013 ($p < 0.01$), which effectively cross-verifies the 2012 breakpoint derived from the CUSUM method.

Interestingly, our dual-verification testing also revealed a profound ecohydrological divergence: while permanent surface water (PSW) underwent a highly significant regime shift, the overall seasonal surface water (SSW) did not yield significant breakpoints in either test ($p > 0.05$). This aligns perfectly with the stochastic nature of seasonal water.

In the revised manuscript, we will update the Results section to include the exact p-values from the Statistical CUSUM test to justify the 2012 baseline. Additionally, the detailed results and comparative figures from the auxiliary Pettitt test will be provided in the Supplementary Information.

9. **The authors should expand their explanation of the G-XGBoost framework, including parameter choices, training-validation strategy, and how spatial heterogeneity is incorporated.**

Response: We appreciate the referee's constructive feedback. We agree that the original description of the G-XGBoost framework was too concise and lacked the necessary technical depth for full reproducibility. To address this, we will expand Section 2.3.5 in the revised manuscript to detail the algorithmic structure of G-XGBoost, specifically focusing on the three aspects requested:

(1) Incorporation of Spatial Heterogeneity:

Unlike the global XGBoost model that assumes spatial stationarity, G-XGBoost constructs a distinct local model for each spatial unit. Spatial heterogeneity is explicitly incorporated by using a bi-square spatial kernel to calculate spatial weights for neighboring observations. Rather than merely using these weights as input features, G-XGBoost integrates them directly into the algorithm's objective function. By multiplying the spatial weights with the first- and second-order gradients (Jacobian and Hessian matrices) during the node-splitting process, the model is mathematically compelled to focus on local data structures, thereby effectively capturing spatial non-stationarity.

(2) Parameter Choices:

The most critical spatial parameter in G-XGBoost is the bandwidth (i.e., the extent of the neighborhood), which defines the scale of local analysis. To ensure optimal performance, we independently calibrated the G-XGBoost models for both the permafrost region (P-PSW) and the seasonal frozen ground region (S-PSW). The optimal bandwidths were determined dynamically by minimizing the cross-validation (CV) criterion, and tree-specific hyperparameters were optimized via a grid search approach.

The resulting optimal hyperparameter configurations for both regions are summarized below:

Hyperparameter	Permafrost Region (P-PSW)	Seasonal Frozen Ground Region (S-PSW)
Spatial Units (N)	2,222	5798
Optimal Bandwidth (b)	77 (Adaptive, nearest neighbors)	77 (Adaptive, nearest neighbors)
n_estimators	500	200
learning_rate	0.01	0.01
max_depth	6	5
subsample	0.8	0.8
colsample_bytree	0.8	0.8
min_child_weight	1	1
gamma	0	0
reg_alpha (L1)	0	0
reg_lambda (L2)	1	1

(3) Training-Validation Strategy:

To ensure robust model evaluation and prevent overfitting, we adopted a nested cross-validation strategy. More importantly, to avoid data leakage and overestimated accuracies in local modeling, G-XGBoost excludes the central target point during the local training phase. The trained local model is then used to predict the value of this omitted central point, yielding an unbiased Out-Of-Bag (OOB) error estimation.

These technical details, along with the corresponding mathematical formulations of the spatially weighted objective function, will be integrated into the revised methodology section to ensure complete transparency.

10. **Lines 194-196 describe improvements in User's Accuracy (UA) and Producer's Accuracy (PA), but the underlying values are not reported. The authors should present the actual UA and PA metrics, not only the percentage changes.**

Response: We sincerely thank the referee for pointing out this omission. We agree that presenting the absolute accuracy metrics is necessary for a transparent evaluation.

As detailed in the full confusion matrices provided in our Supplementary Information (Tables S3 and S4), the User's Accuracy (UA) for surface water extraction using our IOWDM-ENC method increased from 84.73% (JRC-GSW) to 95.30% (an increase of 10.57%). Concurrently, the Producer's Accuracy (PA) increased from 54.29% (JRC-GSW) to 66.89% (an increase of 12.60%).

In the revised manuscript, we will update the text in Section 3.1 to explicitly report these actual UA and PA metrics alongside the percentage changes, and we will direct readers to the corresponding supplementary tables for the underlying data.

11. **Lines 198-199 state that the proposed method has a "distinct advantage in temporal continuity" relative to JRC-GSW. This appears overstated, as the current JRC-GSW coverage (ending in 2021) reflects its latest release rather than an inherent limitation, and the product is periodically updated. The authors should clarify whether the advantage pertains to near-real-time operational updating of IOWDM-ENC. It would also be valuable to compare detection accuracy across different water-body types against the JRC product.**

Response: We completely agree with the referee's assessment. The phrase "distinct advantage in temporal continuity" was poorly chosen and inadvertently overstated. We fully recognize that the JRC-GSW coverage ending in 2021 reflects its periodic release cycle rather than an inherent limitation. Following the referee's suggestion, we will revise the text in the manuscript to clarify that the primary advantage of the IOWDM-ENC framework lies in its capability for near-real-time operational updating. This enables us to perform on-demand processing up to the present year without waiting for the next periodic release of global datasets.

Regarding the comparison of detection accuracy across different water-body types, we agree this is a valuable addition. To address this, we are currently overlaying our 3,000 water validation points with our water-type classification map to conduct a stratified assessment. We will calculate the specific User's and Producer's accuracies for rivers, lakes, and reservoirs against the JRC-GSW product. These type-specific metrics will be incorporated into the Supplementary Information of the revised manuscript to provide a more comprehensive performance evaluation.

12. **The interpretation of Fig. 3 as visual confirmation of improved noise suppression should be revised. The comparison appears to rely on single-date imagery (Sentinel-2 and Jilin-1), which reflect specific acquisition times and may not capture seasonal or ephemeral surface water dynamics. As such, discrepancies between the imagery and the mapped products may reflect temporal sampling differences rather than classification performance. A more robust visual validation would require temporally consistent reference data (e.g., multi-temporal composites or independent water occurrence products) rather than single-scene imagery.**

Response: We appreciate this rigorous perspective. We completely agree that comparing dynamic water maps against single-date imagery can conflate temporal sampling differences with true classification performance.

We apologize for the lack of clarity in our original manuscript regarding the image sources. The reference images presented in Figure 3 are actually not single-date images, but rather annual cloud-free multi-temporal composites, which aligns with the referee's recommendation for robust visual validation.

Specifically, the Sentinel-2 imagery (Fig. 3a) is derived from the "Sentinel-2 cloudless" annual mosaic provided by EOX (https://s2maps.eu/), and the Jilin-1 imagery (Fig. 3b) is derived from the "Jilin-1 Global Map 2021" annual composite (https://www.jl1mall.com/rskit/).

Because these are annual composites, they successfully capture the stable underlying surface features for the respective years, eliminating the bias of single-date seasonal fluctuations. Therefore, the visual comparison in Figure 3 genuinely reflects our algorithm's effectiveness in suppressing static agricultural noise throughout the year, rather than temporal discrepancies.

To eliminate this ambiguity, we will explicitly detail these data sources and emphasize their multi-temporal composite nature in Section 3.1 and the caption of Figure 3 in the revised manuscript.

13. In Fig. 4, the authors should define CV and explain how it quantifies data fluctuation. Also, what is the cause of the peaks in surface water in 1998?

Response: We thank the referee for pointing out the need for further clarification on these details.

First, CV stands for the Coefficient of Variation, calculated as the ratio of the standard deviation to the mean. It is a standardized, dimensionless measure of dispersion. In our study, CV quantifies the degree of interannual fluctuation in surface water area relative to its average size. We used CV to fairly compare the relative volatility between seasonal surface water (SSW) and permanent surface water (PSW), as their absolute areas differ substantially. In the revised manuscript, we will add this full definition to the caption of Figure 4 to ensure it is self-explanatory.

Second, the peak in surface water area observed in 1998 was caused by the 1998 Songhua and Nen River flood, a well-documented extreme hydrological event in Northeast China. While we discussed the regional hydrological response to this flood later in the Discussion (Section 4.1), we agree that the cause of this peak should be contextualized when the time series is first introduced. Therefore, we will add a brief sentence in the Results section (Section 3.2) of the revised manuscript to link the 1998 peak to this flood event, providing immediate clarity for the readers.

14. The manuscript refers to groundwater depletion in both the abstract and results, but it is unclear how this inference is derived from the presented analysis. How is groundwater depletion inferred from the analysis? Were climate vs anthropogenic contributions separated?

Response: We thank the referee for this question, which highlights the need to clarify our analytical boundaries and refine our scientific terminology.

(1) Regarding terminology and how the trend is observed:

First, we agree that "groundwater depletion" is an overly strong term that may imply complete exhaustion of the aquifer. To be scientifically precise and avoid overstating the condition of the regional water resources, we will systematically replace "groundwater depletion" with "groundwater storage decline" (or "groundwater storage loss") throughout the revised manuscript, as the regional aquifers are experiencing a steady decline rather than absolute exhaustion.

This observed downward trend in the Groundwater Storage Anomaly (GWSA) dataset, which is derived directly from GRACE satellite gravity observations (Zhang et al., 2024, listed in Table 1) and presented in Figure 5c, serves as the direct observational evidence of this groundwater storage loss, rather than a secondary inference derived from our surface water

analysis.

(2) Regarding the separation of climate vs. anthropogenic contributions:

Our Geographical-XGBoost attribution model was specifically designed to separate climate and anthropogenic contributions for surface water dynamics, which is the primary objective of this study. We did not run a separate quantitative attribution model to partition the drivers of groundwater depletion itself, as that falls beyond the scope of this paper.

To resolve any potential confusion and ensure scientific precision, we will clarify in Section 3.3 of the forthcoming revised manuscript that the reported groundwater storage decline is a direct observation from the GRACE-derived GWSA dataset, and we will update the terminology accordingly.

15. **The manuscript occasionally interprets non-significant trends as meaningful increases or decreases. While the direction of estimated slopes can be reported, non-significant results ($p \geq 0.05$) should not be described as evidence of change. For example, in Fig. 5 panels a and b, all linear regressions are non-significant ($p \geq 0.05$), indicating that no statistically supported trend is detected. The authors should therefore revise the wording throughout the manuscript to clearly differentiate between statistically significant trends and non-significant directional tendencies, and avoid drawing interpretive conclusions from the latter.**

Response: We fully accept this criticism and apologize for the statistical ambiguity in our original phrasing. We agree that a p-value$\geqslant$0.05 indicates a lack of statistical evidence for a sustained trend, and such results should not be interpreted as meaningful directional changes..

In the revised manuscript, we will review the entire text to correct this terminology. Using Figures 5a and 5b as examples, we will remove descriptions such as "upward trend" or "decline" for any metrics where $p\geqslant$0.05. Instead, we will use accurate phrasing, such as "fluctuated without a statistically significant trend" or "showed no significant directional tendency."

Furthermore, we will revise our Discussion section to ensure that interpretive conclusions and mechanistic deductions are drawn exclusively from statistically significant results (p<0.05). We believe these revisions will enhance the statistical rigor of our manuscript.

Minor comments:

1. **In Fig. 1, the black boundaries delineating the permafrost and seasonal frozen ground regions appear open. The authors should clarify whether this is intentional or correct the boundaries so they are fully closed.**

   Response: We thank the referee for their keen observation. The boundaries for the permafrost and seasonal frozen ground regions are indeed fully closed in our spatial dataset. The apparent "open" gaps in Figure 1 are an unintentional visual artifact caused by the layer drawing order in our GIS software. Specifically, the blue river vector layer was placed above the frozen ground boundary layer, which obscured the black boundary lines where the rivers and boundaries intersect or overlap. In the revised manuscript, we will correct the map by adjusting the layer hierarchy (bringing the black frozen ground boundaries to the front) to ensure they are fully visible, continuous, and clearly closed.

2. **Lines 108–109 ("quantifying the depth… grid box") are unclear. The authors should revise the sentence to improve clarity.**

   Response: We agree with the referee that the original phrasing was unclear. To improve clarity, we have simplified and refined the sentence. In the revised manuscript, this sentence will read: "Specifically, SR, SSR, and snowmelt were derived as annual cumulative values from the ERA5-Land Daily dataset, representing the equivalent water depth over each grid cell."

3. **Acronyms such as MNDWI, NDWI, EVI should be defined upon their first appearance in the manuscript.**

   Response: We agree and will correct this omission. We will provide the full definitions for the modified normalized difference water index (MNDWI), normalized difference water index (NDWI), and enhanced vegetation index (EVI) upon their first appearance in Section 2.3.1.

4. **Line 127 ends with "and" which suggests that the sentence may be incomplete. The authors should revise or complete the sentence.**

   Response: We apologize for the confusion caused by the layout and line break in the original manuscript. The sentence is actually complete; the word "and" at the end of line 127 is immediately followed by the remaining condition (MNDWI > NDVI or MNDWI > EVI) on the next line. However, we agree that this awkward line break disrupts the reading flow and creates a visual disconnect. To improve readability, we will adjust the formatting and slightly refine the punctuation in the revised manuscript. The sentence will be reformatted to keep the logical conditions cohesively grouped,

reading as follows: "Candidate water pixels are identified only when they satisfy the newly imposed constraints: MNDWI > 0.1, NDWI > -0.1, EVI < 0.1, and (MNDWI > NDVI or MNDWI > EVI)."

5. **The caption for Fig. 2 could be more descriptive. It currently reiterates only the panel titles already shown within the figure. The authors should expand the caption to better explain the figure content.**

   Response: We agree. We will expand the caption for Figure 2 in the revised manuscript to provide a comprehensive summary of the methodological workflow, detailing the transition from data preprocessing and water detection to the G-XGBoost-SHAP attribution modeling.

6. **The caption for Fig. 5 should explain the meaning of the single and double stars used in the cross-correlation shown in panel d. It should also clarify the labels along the x-axis (e.g., what 1a, 2a, 3a represent).**

   Response: We will update the caption of Figure 5 accordingly. We will clarify that the single and double stars denote statistical significance levels (* $p < 0.05$, ** $p < 0.01$), and that the x-axis labels (1a, 2a, 3a) represent 1-year, 2-year, and 3-year time lags, respectively.

7. **Lines 260–269 refer to Figs. 9, 10, and 11, although based on their order in the text they should correspond to Figs. 6, 7, and 8. The authors should revise the figure numbering to follow the figures' order of appearance in the text.**

   Response: We thank the referee for their careful review of our manuscript. We have double-checked the figure numbering throughout the text and would like to respectfully clarify that the current numbering strictly follows the chronological order of their first appearance. Specifically, Figures 6, 7, and 8 are first introduced sequentially at the very beginning of Section 3.4 (Lines 248-250 in the original manuscript: "...exhibit spatially heterogeneity (Fig. 6-7). To ensure model robustness, variables with high multicollinearity were excluded prior to analysis (Fig. 8)."). Following these initial citations, Figures 9, 10, and 11 are subsequently introduced in the following paragraphs (Lines 260-269) to explain the local attribution effects. We apologize if the dense succession of these initial citations caused any visual interruption.

8. **The manuscript makes extensive use of acronyms, which become difficult to track. The authors should restate variable definitions when they first appear in the Results section and in figure captions to help reader comprehension.**

   Response: This is a very helpful suggestion for improving readability. In the revised manuscript, we will restate the full names of key variables (e.g., permanent surface

water [PSW], seasonal surface water [SSW], groundwater storage anomaly [GWSA]) at the beginning of the Results section, and we will ensure these acronyms are clearly defined in all relevant figure captions.

9. **More detailed captions should be provided for all figures to help readers interpret the visual information.**

   Response: We accept this advice. We will conduct a comprehensive review of all figure captions in the revised manuscript, expanding them to ensure that every figure is fully self-explanatory.

10. **The authors should explain the meaning of the single and double stars used in Fig. 8.**

    Response: Similar to Figure 5, we will add an explanatory note to the caption of Figure 8, specifying that the stars indicate statistical significance (* $p < 0.05$, ** $p < 0.01$).

11. **In line 306, the phrase "around 2009" is repeated twice; the authors should revise for conciseness.**

    Response: We apologize for this redundancy. We will delete the repeated phrase to improve conciseness in the revised text.