Item-by-item responses to all review comments

NOTE: To facilitate the evaluation of our responses, original comments are listed first in their originals (in black), followed by our itemized responses (in red). An annotated version of revised manuscripts is attached.

Referee's Comments

This manuscript investigates the relationship between vegetation restoration, with particular emphasis on forest stand structure, on groundwater storage changes in the Three-North Shelterbelt Forest Region of China. The topic is relevant and timely, as understanding the trade-offs between ecological restoration and groundwater sustainability is important for dryland management. The consideration of forest age and species composition represents an interesting contribution that extends previous studies based primarily on vegetation greenness or land-cover changes.

Overall, the study is potentially suitable for publication in HESS. However, the manuscript currently presents several methodological and presentation issues. Important assumptions are insufficiently justified, several conclusions appear stronger than the evidence presented, and many results are difficult to evaluate because they are either insufficiently explained or not adequately supported by figures or quantitative analyses. I therefore recommend major revisions before the manuscript can be considered for publication.

Response: First of all, we would like to thank you for your constructive comments to improve our manuscript. We really appreciate your precious time and great efforts on reviewing the manuscript. We have carefully addressed all the comments, which have greatly improved the quality of this manuscript. Please find below our detailed responses to the comments point by point.

Specific comments:

1. The validation strategy should be described more clearly. It is not entirely clear how the comparison between the GRACE-derived GWSA and groundwater observations was performed, particularly considering the scale mismatch between point observations and the 0.25° GRACE product. Since the validation represents an important component of the study, I also recommend moving at least part of the figures currently presented in the Supplementary Information (Figs. S2–S5) into the main manuscript.

   Response: Based on the comments, we have provided details on the comparison procedure between GRACE-derived GWSA and groundwater observations. To mitigate the influence of local noise, monitoring wells located within the same 0.25° grid cell were first averaged, yielding a subset of 21 stations for validation (Fig. S1). The observed GWSA for each grid was then derived by multiplying the groundwater table depth anomaly by the corresponding specific yield. To address the spatial scale mismatch between localized well observations and the GRACE-derived estimates, validation was conducted using the regional-averaged, normalized annual mean GWSA series across all 21 stations. Furthermore, we evaluated the consistency of the mean long-term trends between the 21 stations and GRACE-derived estimates. This approach focuses on evaluating interannual variability and long-term regional consistency rather than direct point-to-grid matching.
   Please see Lines 137–146: "*GRACE-derived GWSA was compared with the GWSA derived from monitoring data. To reduce local heterogeneity, monitoring wells located within the same 0.25° grid cell were averaged, resulting in* a subset of *21 stations from the Geological Environment Monitoring Institute of the China Geological Survey, spanning the period from 2005 to 2016 (Fig. S1). The observed GWSA can be calculated by multiplying groundwater table depth anomaly (GWTDA) by specific yield (Sy) (Bhanja et al., 2016; Gong et al., 2018), as shown in Eq. (2):*

   $$GWSA = -GWTDA \times Sy, \quad (2)$$

   where GWTDA represents the groundwater table depth anomaly, and Sy is the specific yield, with data sourced from Sutanudjaja et al. (2018). To account for the scale mismatch between observations and GRACE-derived GWSA, validation was performed using normalized annual mean series averaged across these 21 stations, alongside a comparison of their mean regional trends, focusing on the consistency of interannual variability and long-term trends."

   Moreover, we moved the figures that are critical for supporting the main conclusions (e.g., Fig. S3 and S5) from the Supplementary Materials to the main text and renamed them as Fig. 4 and Fig. 5. Please see the revised Section 2.3.2, Section 3.1, and Figs 4-5 in the manuscript.

2. Several methodological choices require further justification. For example, please explain the rationale for selecting precipitation, evapotranspiration and temperature simultaneously as climatic predictors, particularly given the close relationship between temperature and evapotranspiration. Likewise, Equation (3), which downscales provincial socioeconomic statistics to the grid scale, requires a much clearer explanation. In particular, the definition of the reference gridded dataset (Gridref) and the assumptions underlying this downscaling approach should be described in greater detail.

   Response: Based on the comments, we have added the physical rationale for selecting climatic predictors and provided a detailed clarification of Equation (3), including variable definitions and underlying spatial assumptions. Please see Lines 221–227 and Lines 202–211.

   Lines 221–227:"*Precipitation (PRE), temperature (TEM), and evapotranspiration (ET) were simultaneously included as climatic predictors. Precipitation constitutes the primary water input, ET represents the dominant water loss pathway, and TEM acts as a key energy driver that regulates both evaporative demand and cryospheric processes such as snowmelt and glacier melt (Adeyeri et al., 2024). The Normalized Difference Vegetation Index (NDVI) was incorporated to reflect vegetation regulation on water and energy exchanges, while irrigation (IRR), gross domestic product (GDP), coal consumption (COAL), and population (POP) were included to account for anthropogenic modifications to the water cycle.*"

   Lines 202–211:"*where $Grid_{human}$ represents the downscaled gridded human activity dataset (including irrigation, population, and GDP), $Grid_{ref}$ denotes the reference gridded dataset that provides high-resolution spatial distribution patterns for each variable (Zhu et al., 2023). $S_{province,ref}$ is the provincial total aggregated from the reference gridded dataset by summing all grid cells within each provincial boundary, while $S_{province}$ is the corresponding annual statistical value obtained from the China Statistical Yearbook.*

   This downscaling approach assumes that the spatial allocation pattern of human activities remains consistent between the reference year of the gridded datasets and the study period, and that the relative spatial distribution of population, irrigation, and economic intensity captured by $Grid_{ref}$ serves as a reliable proxy for distributing provincial totals. By normalizing Gridref with the ratio of statistical to reference totals, we preserve both the official provincial magnitudes and the fine-scale spatial heterogeneity."

3. Several conclusions presented throughout the Results are insufficiently supported

by figures or quantitative analyses. For example: (i) the locations of groundwater gain and loss regions are not clearly shown; (ii) greening hotspots are discussed without explaining how they were identified; (iii) trends for grassland expansion areas are mentioned but not illustrated; (iv) the methodology used to derive the trends shown in Fig. 8b is unclear, and their statistical significance is not reported. Please ensure that all major conclusions are directly supported by figures, quantitative results, or statistical analyses.

Response: Based on the comments, we have carefully revised the manuscript to ensure that the major conclusions are supported by corresponding figures, quantitative analyses, or statistical tests. Specifically:

(i) The description of groundwater gain and loss regions has been clarified in the caption of Fig. 3. Blue colors indicate increasing groundwater storage (gain), whereas yellow-to-red colors indicate decreasing groundwater storage (loss), with darker colors representing larger trend magnitudes.

(ii) The description of greening hotspots has been revised. Instead of referring to "greening hotspots", we now describe the spatial distribution of areas with increased NDVI from 2003 to 2023, which were mainly distributed in the Loess Plateau, Northeast China, and the central Tianshan Mountains. Please see Lines: 318–319 and Lines 435–436.

Lines 318–319: "*The areas showing increased NDVI from 2003 to 2023 were mainly distributed in the Loess Plateau, Northeast China, and the central Tianshan Mountains.*"

Lines 435–436: "*Fig 8a-c shows that agricultural irrigation and coal mining activities were mainly concentrated in the North China Plain and northwestern region.*"

(iii) The spatial distribution of grassland expansion areas has been added to the Supplementary Materials (Figure S4).

(iv) The statistical significance of the trends shown in Fig. 12b has been evaluated using the Mann–Kendall test. Significant trends are represented by filled symbols, whereas non-significant trends are represented by open symbols.

4. In several places, the Discussion appears to draw stronger conclusions than are directly supported by the analyses. In particular, statements implying that differences in groundwater decline are directly driven by specific forest types or tree species should be interpreted more cautiously, as the analyses presented are largely observational and do not fully separate the effects of forest structure from

other spatially varying factors (e.g., climate, topography, irrigation, or geology). I encourage the authors to better discuss the limitations of their approach and moderate some of the causal interpretations.

Response: Based on the comments, we have moderated the causal language throughout the Discussion. Please see Lines 425–426, 429–430, 434–435, 439–440, 442–443, 447–448 and 450–452.

Lines 425–426:

"*Our results revealed notable associations between forest structure and GWS changes in the TNSFR.*"

Lines 429–430:

"*One possible ecological explanation is that intraspecific competition is, on average, 4 to 5 times stronger than competition between species (Adler et al., 2018).*"

Lines 434–435:

"*Beyond vegetation restoration, GWS decline in the TNSFR may also be associated with anthropogenic activities, particularly agricultural irrigation and coal mining.*"

Lines 439–440:

"*However, aquifer-system disruption related to mine dewatering and industrial water consumption has been linked to substantial GWS losses (Xie et al., 2018)*"

Lines 442–443:

"*Despite these policy interventions, the legacy effects of decades of intensive mining and irrigation may have contributed to persistent aquifer-system stress.*"

Lines 447–448:

"*Our results suggest that species composition and stand age structure are closely associated with spatial variations in groundwater storage dynamics.*"

Lines 450–452:

"*This simplification prevents models from adequately representing the differences in water consumption among various forest types when simulating GWS changes, and may therefore underestimate groundwater depletion risks associated with afforestation.*"

5. The forest structure analysis relies on National Forest Inventory data collected

during 2014–2018, whereas groundwater storage trends are calculated for 2003–2023. Please discuss the implications of this temporal mismatch and justify the assumption that forest age, species composition, and stand structure during the inventory period are representative of the entire groundwater trend period.

Response: Based on the comments, we have added further discussion regarding the temporal mismatch between the forest inventory period and the groundwater trend period. This mismatch may introduce uncertainties because forest stands could have experienced growth, forest management, or disturbances during the groundwater analysis period. However, this potential impact on our conclusions is effectively mitigated for three major reasons:

According to the Technical Regulations for National Forest Continuous Inventory, key species in northern China are classified using broad age intervals (e.g., spruce plantations are grouped in 20-year intervals). Because this 20-year age interval directly aligns with our 2003–2023 study period, using single-period inventory data has little to no effect on the age-class categorization of spruce forests.

For poplar forests, over 80% of stands fall within the 20–40 year age range (classified mainly as mature or over-mature forests). Although poplar uses a 10-year age interval, our core objective is to infer the relationship between GWSA trends and the spatial heterogeneity of species and age groups, rather than relying on precise annual age sequences.

The 9th NFI provides standardized, ground-based observational covariates. Replacing these field measurements with annually continuous, satellite-derived age products would introduce substantial secondary uncertainties from remote-sensing inputs, disturbance detection algorithms, and age-estimation biases (Besnard et al., 2025).

Please see the revised text in Section 4.

Lines 488–501: "*This temporal mismatch may introduce uncertainties, as forest stands could have experienced growth, forest management, or disturbances during the groundwater analysis period. However, the influence of this limitation is likely limited. The use of single-period rather than multi-period forest inventory data may affect the accuracy of stand age determination, this limitation can be effectively mitigated given the characteristics of the dominant tree species in the TNSFR. The predominant species are spruce and poplar. According to the Technical Regulations for National Forest Continuous Inventory, spruce plantations in northern China are grouped in 20-year age intervals. Since this aligns with the 2003–2023 study period, using single-period inventory data has little effect on the age classification of spruce forests. For poplar forests, over 80% of stands fall within the 20–40 years age range, with the vast majority classified as mature or over-mature forests. Although the age group interval for poplar is 10 years, the core objective of this* *research is to infer the relationship between GWSA and the spatial heterogeneity of tree species and stand age, rather than relying on precise annual age sequences. The 9th NFI data provide standardized, ground-based observational covariates rather than experimental treatments (Musavi et al., 2017). Replacing them with annually resolved, model-derived forest-age products would not necessarily improve the analysis and could introduce additional uncertainties from remote-sensing inputs, disturbance detection, and age-estimation biases (Besnard et al., 2025).*"

6. The Discussion focuses primarily on the results related to forest structure and the SEM analysis, whereas other important findings (e.g. groundwater trends, climatic controls and validation) receive limited discussion. I encourage the authors to provide a more balanced discussion that places all major findings into the context of previous literature and discusses the limitations of the study.

   Response: Based on the comments, we have substantially expanded the Discussion section (Section 4) to ensure a more balanced and comprehensive reflection of all major findings. Specifically, we have added in-depth discussions addressing: (1) climatic controls and hydrological mechanisms; (2) comparisons with previous literature and validation evidence; (3) study limitations and future directions.

   Please see Lines 406–423 and 470–477.

   Lines 406–423: "*P-ET is jointly controlled by precipitation and evapotranspiration and is an important component of terrestrial water resource conditions (Padrǿn et al., 2020). Our results suggest different patterns of GWS decline between arid and semi-arid regions. In the arid region, both precipitation and evapotranspiration showed non-significant decreases, leaving P-ET largely stable. As most greening occurred over bare land and sparsely vegetated areas, vegetation may have relied mainly on soil water rather than groundwater, which could explain the relatively limited GWS decline and the absence of rapid aquifer depletion despite local irrigation (Ma et al., 2021; Wu et al., 2021). In contrast, greening in the semi-arid region was mainly associated with grassland restoration, forest recovery, and grassland-to-forest conversion, and covered a larger area. Although precipitation increased, much of this gain was likely consumed by significantly enhanced evapotranspiration, resulting in only minor changes in P-ET (An et al., 2025). Together with anthropogenic water withdrawals from intensive agriculture and mining, this pattern may be associated with the stronger GWS decline in the semi-arid region, where the rate of decline was approximately 1.86 times that in the arid region (Feng et al., 2022a).*

   However, ET and GWS products are subject to retrieval uncertainties. In addition, this study did not quantitatively separate the contributions of irrigation, mining, and natural vegetation water use, and thus only qualitatively identified *anthropogenic influences. Future work could integrate additional data to quantify the relative contributions of different drivers. Nevertheless, studies using Independent Component Analysis, grey slope relational analysis, and partial least squares regression have reported that vegetation restoration, agricultural irrigation, and coal mining can jointly exacerbate GWS decline in semi-arid regions such as the Loess Plateau, consistent with the stronger GWS decline observed in the semi-arid region in this study (Chen et al., 2020; Feng et al., 2022b; Han et al., 2020b).*"

Lines 470–477: "*To further evaluate the reliability of the derived GWSA trends, we compared them with direct and indirect validation evidence. The normalized annual GRACE-derived GWSA series was generally consistent with groundwater monitoring wells, indicating that the estimates captured the regional long-term groundwater signal despite the scale mismatch between point observations and GRACE grids. The evaluation of GLDAS-derived SMSA against ESA CCI, ERA5, and in situ soil moisture observations further supported the GWSA calculation. The GWSA trend in the TNSFR was* -5.47 *mm* yr^{-1}*, comparable to the GRACE-based estimate of* -5.03 *mm* yr^{-1} *reported for Xinjiang, which accounts for approximately 40 % of the TNSFR (Liang et al., 2025). A GWSA depletion rate of* -7.7 *mm* yr^{-1} *has also been reported for the Yellow River Basin, whose upper and middle reaches overlap with the TNSFR and include major semi-arid regions such as the Loess Plateau (Liu et al., 2026).*"

7. Several figures would benefit from improvements. Figure 1 could better illustrate the location of the study area (e.g. by slightly zooming out the overview map) and panels b–c could be enlarged. Figure 2 requires a much more informative caption explaining the different steps of the workflow. In addition, some figure references appear out of sequence (e.g. Fig. 9 before Fig. 6; lines 256–259), and certain figure panels (e.g. Fig. 6g) are difficult to interpret because the captions provide insufficient explanation.

   Response: Based on the comments, we have revised the figures and captions accordingly.

   Specifically, Figure 1 has been modified by slightly zooming out the overview map and enlarging panels (b–c) to improve the visualization. Please see Lines 79–82."

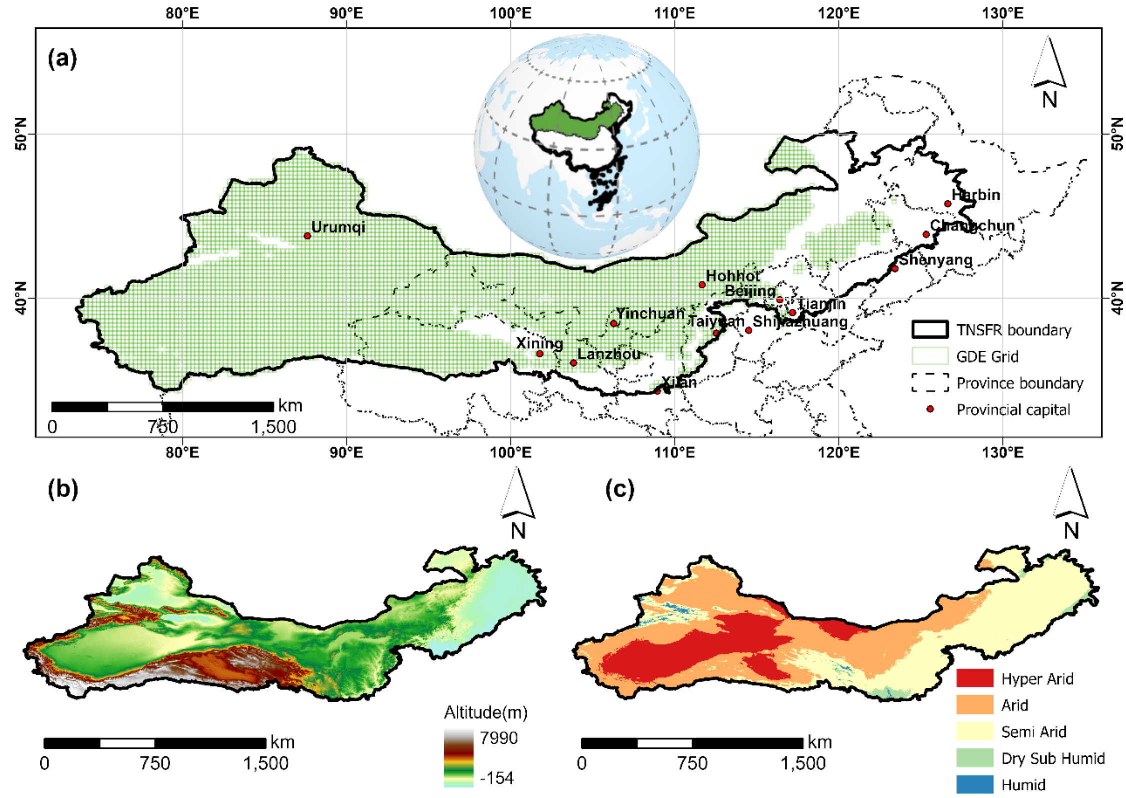

Figure 1 Overview of the Three-North Shelter Forest Region (TNSFR). (a) Geographical location of TNSFR. (b) Elevation map. (c) Climate zone map."

The caption of Figure 2 has been expanded to provide a more detailed description of the workflow steps. Please see Lines 130–134: " *Figure 2 Flowchart illustrating the analytical procedures carried out in this study. Multi-source datasets were collected and processed to estimate and validate GWSA using a GRACE/GLDAS residual water-balance approach. The derived GWSA was then used to assess groundwater trends, vegetation and land-cover changes, and forest stand-structure effects. Finally, structural equation modelling was applied to quantify the direct and indirect effects of climatic, vegetation, and anthropogenic factors on GWSA.*"

We have also checked and corrected the sequence of figure citations throughout the manuscript. Specifically, the previous Fig. 9 has been moved forward and renamed as Fig. 8, while the previous Fig. 6 has been placed later and renamed as Fig. 9.

In addition, the description of the corresponding panel (Fig. 9g) has been improved to enhance its interpretation. Please see Lines 327–329: "*(g) Land-cover transitions from 2003 to 2023. Each colored flow represents the transition from one land-cover class to another, with the color indicating the land-cover class. (i) Percentage of land use change. Relative contribution of the net change rate of each land use type to the total land use change rate from 2003 to 2023.*"

8. Lines 77–78: Please provide additional information on how the aridity classification shown in Fig. 1c was obtained. Although the manuscript refers to the Aridity Index (AI), it does not explain how the index was calculated (or whether it was taken directly from the referenced dataset) nor the threshold values used to define the five climate classes. Including this information in the Methods would improve the reproducibility and interpretation of the analysis.

   Response: Based on the comments, we have added a description to explicitly state the source of the Global Aridity Index (AI) dataset and the specific threshold values used to define the five climate zones shown in Fig. 1c.

   Please see Lines 75–77: "*For subsequent analysis of climatic impacts, the TNSFR was classified into five zones based on Global Aridity Index (AI) form Zomer et al. (2022): hyper-arid (AI<0.03), arid (0.03"≤" AI<0.2), semi-arid (0.2"≤" AI<0.5), dry sub-humid (0.5"≤" AI<0.65) and humid (AI"≥" 0.65) (Fig. 1c).*"

9. Line 104: Please specify the temporal resolution of the China Land Cover Dataset. Does the dataset provide annual land cover maps for the entire 2003–2023 period, or were only selected years available?

   Response: Based on the comments, we have updated the description to explicitly state that the China Land Cover Dataset (CLCD) provides continuous annual land cover maps for every year throughout the entire 2003–2023 period.

   Please see Lines 108–109: "*China Land Cover Dataset provides annual land cover data for China at a 30 m spatial resolution from 2003 to 2023 (Yang and Huang, 2021).*"

10. Line 210: The manuscript states that 77.06% of the study area experienced "significant negative trends" in GWS. Please clarify whether this percentage refers to statistically significant trends (e.g. $p < 0.05$) identified using the Mann–Kendall test. In addition, the reported values of −40 mm and −21.9 mm (line 212) are referred to as "rates" and should therefore be expressed with appropriate time units (e.g. mm yr^{-1} or mm $month^{-1}$). Please ensure that trend units are reported consistently throughout the manuscript.

    Response: We thank the reviewer for pointing out these important details regarding statistical significance and unit consistency.

    (1) Regarding statistical significance (77.06%): Yes, the 77.06% specifically refers to areas showing statistically significant negative GWS trends ($p < 0.05$) evaluated using the Mann－Kendall (MK) test. As described in Section 2.3.3 (Lines 142–150), the non-parametric MK test was consistently applied across the study region to determine trend significance.

Please see Lines 155–163: "*2.3.3 Trend analysis*

Long-term trends in GWS and related driving factors were identified using a combination of the Theil–Sen slope estimator (Sen, 1968) and the Mann–Kendall (MK) trend test (Hirsch et al., 1982; Kendall, 1948). These non-parametric methods are less sensitive to seasonal variability and outliers, and are therefore well suited for detecting long-term trends in hydroclimatic time series.

The Theil–Sen slope was employed to capture change trends. The Mann–Kendall test was applied to assess trend significance, with $p < 0.05$ *indicating a significant trend. Based on the sign of the GWS trend, the study area was further classified into groundwater gain and loss regions for subsequent analyses of climatic change, vegetation restoration, and anthropogenic influences on groundwater dynamics.*"

(2) Regarding trend units and consistency: The units of the reported trend rates (-40 and -21.9) have been corrected to mm yr^{-1}, please see Lines 246–247: "*The maximum annual decline rates of GWSA reached -40 mm* $year^{-1}$ *in the Tianshan Mountains and -21.9 mm* $year^{-1}$ *in North China, respectively.*"