**Response to The Comments from Reviewer #1**

The manuscript introduces a multi-temporal China annual river extraction framework, which includes a multi-data source-based water extraction module and an object-based hierarchical decision tree river extraction algorithm, and produces annual China river extent maps (CRED) from 2016 to 2023. However, the paper needs further improvement in terms of its structure and readiness for publication. The motivation and innovation of the research should be clarified.

**Response:**

Thank you for taking your precious time and making diligent efforts to review our manuscript. The valuable comments and constructive suggestions are definitely helpful, and we sincerely appreciate them for improving our paper. We have carefully studied the comments and revised the manuscript point-by-point. All modifications have been marked in the revised manuscript. One again, thank you for your valuable comments.

## Some major comments are as follows.

**1.** The motivation for using multisource datasets for water extraction should be better explained. The authors state that the choice of data sources (DW, EGLC, and Sentinel-2) is based on their availability, in that order. However, the river mapping results for China in 2016, primarily using Sentinel-2, show no significant differences compared to other years. Is the proposed method aimed at achieving higher extraction efficiency, or is it designed to enhance accuracy?

**Response 1**:

Thanks for your valuable comments. It is helpful to enhance our manuscript. We have made revisions and explanations for the motivation of using multisource datasets (DW, EGLC and Sentinel-2) in the manuscript.

areas had high mapping accuracy. Statistical analysis revealed that many regions in China had no or few good observations (Fig. S1). This occurred because the DW images were produced from Sentinel-2 images with less than 35% cloud coverage. Sentinel-2 images used in DW dataset may have could contaminations, which also resulted in missing observations. Thus, the complete coverage of the entire China to produce a complete China river map cannot be achieved by using DW dataset alone.

To facilitate image computation, we divided China into 52 tiles using a 5°×5° grid (Fig. S1). If a tile contains considerable pixels with fewer than 3 valid observations, it was defined as a data insufficient tile. For these tiles, we used the ESRI global land cover (EGLC) dataset to supplement the data (Karra et al., 2021). The EGLC dataset was produced on the basis of

The Dynamic World (DW) was a 10-m spatial resolution land use dataset from 2015 to the present, with a revision of 3-5 days. The high revisit frequency of DW allows it to effectively capture the seasonal variations of water bodies. Therefore, DW was chosen as the primary dataset for river classification. However, since the DW use only images with cloud coverage below 35% for classification, it exhibits significant data gaps in considerable regions of China. As shown in Figure T1, the value of each pixel represents the number of valid observations within a year, excluding bad observations affected by cloud contamination, cirrus, and cloud shadows. It is shown that there are significant data gaps in Southwest China in 2023, such as in tiles 14, 15 and 16. To address this issue, the EGLC was selected to substitute the DW dataset in these tiles with missing data, and were used to generate water maps from 2017 to 2023. However, the EGLC have no data in 2016.

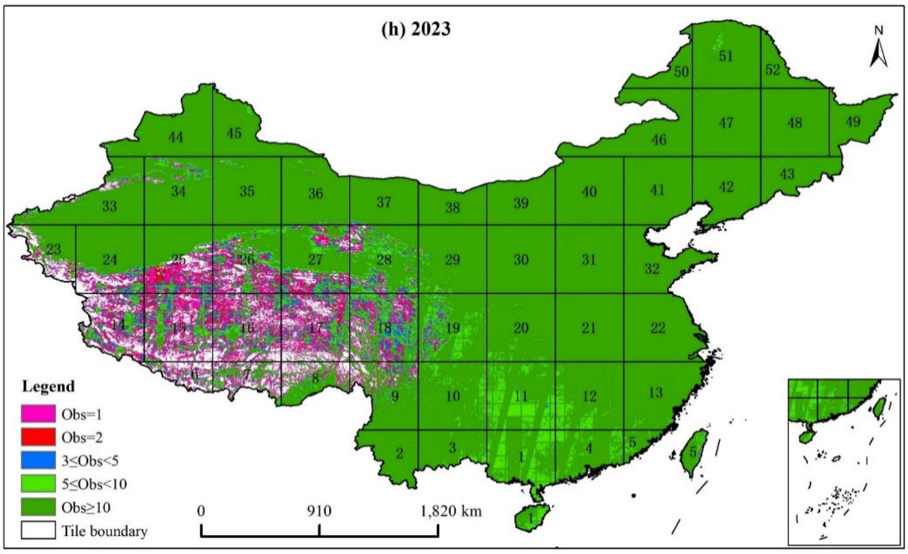


Figure T1. Valid observations for individual pixels in DW image of 2023. This figure was also displayed in Supplementary material

To make the best and maximize the use of Sentinel-2 imagery and extend the temporal span by producing an additional year of river map, we utilized the Sentinel-2 imagery to substitute the DW datasets with considerable invalid or insufficient observations from 2015-2016. We applied the multiple index water detection rule (MIWDR) to Sentinel-2 imagery to generate water time series and composited them into annual water map for 2016 using mode algorithm. In summary, the temporal span of three datasets is illustrated as Figure T2.

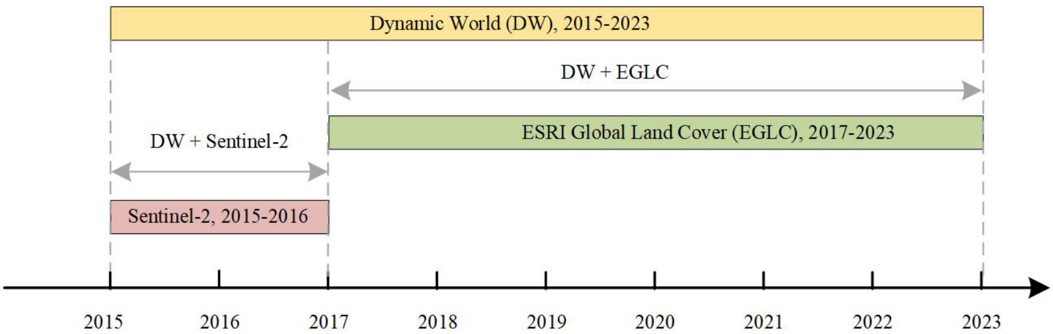


Figure T2. Time span of three datasets

**2.** In the proposed approach, the geometric rules for river extraction were based on the 2020 CNLUCC map. As shown in Fig. 8, the extraction results from CRED exhibit significantly higher spatial consistency with the CNLUCC map compared to the other two comparison datasets. Did the authors consider using different datasets during the geometric rule extraction or the result comparison process?

**Response 2:**

Thank you for your valuable comments. In our study, the CNLUCC were used to generate training samples. These samples were used to explore the geometric difference of water covers (e.g. river, lake and reservoir) and determine appropriate thresholds of each geometric features. Then, the rule set for river extraction was developed. The developed algorithm for river mapping is robust and effective. This algorithm is not supervised algorithm, and is independent of training samples. Its rules and thresholds were constant, did not change over time and regions.

The high consistency between our CRED and CNLUCC is mainly due to their high mapping accuracy of rivers. The CNLUCC was a 30-m dataset with detailed land use types, which was produced by human-computer interactive process. Its extensive manual interpretation and strict data production procedures ensure the high accuracy of the data. The CRED was produced using the accurate river mapping algorithm, and was further improved by post-processing operations. The CRED also achieved high accuracy of rivers. Thus, these two datasets had good consistency.

We did not use different datasets for the algorithm development. To further illustrate the accuracy, robustness and effectiveness of our algorithm, we implemented river mapping results using five data sources (i.e. DW, ELGC, Sentinel-2, ESRI, and JRC-GSW). The sensitive analysis was conducted in our manuscript. The detailed descriptions were added in sub-section 5.1.

**3.** In the statistical results for river areas from 2016 to 2023, the river area in 2016 was noticeably smaller than in other years. Was this phenomenon also observed in non-river water bodies? It would be helpful to include the accuracy of water extraction for each year.

**Response 3**:

Thanks for careful review and offering valuable comments. Indeed, the river area in 2016 was noticeable smaller than other years. This phenomenon was mainly attributed to two aspects. First, due to large gaps in DW datasets from 2015 to 2016, the Sentinel-2 image was used to extract waters using multiple index water detection rule (MIWDR) in 31 out of 52 tiles in China. However, due to its low observation frequency and the impact of cloud contaminations, the available Sentinel-2 images from 2015 to 2016 were relatively scare. This limitation may result in uncertainties for mapping river extents. Second, the MIWDR algorithm that applied to Sentinel-2 images exhibit different performance in term of water classification, compared with deep learning algorithm that adopted by DW and EGLC. Based on the sensitive analysis in sub-section 5.1, it was found that the MIWDR could well extract large and pure waters, while exhibited poor performance in seasonal waters or mixed pixels of waters. This characteristics of MIWDR may also lead to underestimations of river extents. We discussed this uncertainty in our Discussion sections.

annual water extents without seasonal information (Venter et al., 2022). The MIWDR algorithm misclassified some ice, snow and shadows as water, and exhibited poor performances for seasonal waters or mixed pixels of waters. Meanwhile, due to the low observation frequency and the impact of cloud contaminations, the available Sentinel-2 images from 2015 to 2016 were relatively scare, which may result in underestimations of waters. These issues lead to the river area in 2016 was noticeable smaller than that in other years. In addition, the extent and completeness of manual corrections varied across different years,

**More specific comments are as follows.**

**1.** Sensitivity analysis is required to validate the feasibility of the proposed method for extracting water body extents using different data sources across different tiles/periods.

**Response 1**:

Thanks for your valuable comments. As suggested by the comment, we have added sensitivity analysis in our manuscript.

To evaluate the feasibility of our river mapping algorithm, five data sources were collected to implement river mapping. The DW, EGLC and waters derived from Sentinel-2 images using MIWDR were selected. Additionally, the ESRI WorldCover (ESRI) and JRC Global Surface Water (JRC-GSW) were also chosen to apply our algorithm for river extraction. We implemented river classifications using these five datasets for tile 21 in 2021. To explicitly illustrate the sensitive analysis, a new section has been added in the manuscript.

### 5.1 Sensitivity analysis of our river extraction algorithms

To further evaluate the feasibility of our river mapping algorithm, we used different datasets to extract rivers for tile 21 in 2021. On the one hand, the three datasets used in our study, namely DW, EGLC and waters derived from Sentinel-2 images using MIWDR (MIWDR), were selected to extract rivers. On the other hand, ESRI land use dataset and JRC-GSW water dataset were additionally selected to apply our algorithm for river extraction. The spatial distribution of rivers in five datasets were shown in Figure 11. Results indicated that rivers in different water maps could be effectively extracted. Specifically, our algorithm could accurately extract the rivers shown in corresponding water maps.

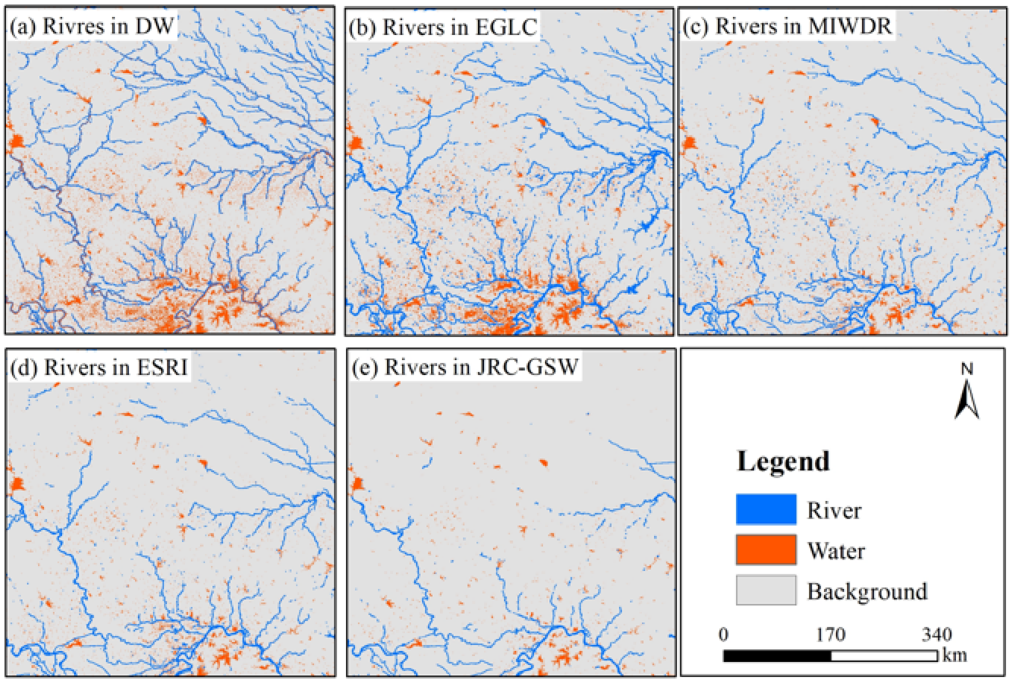


**Figure 11. Spatial distribution of rivers and waters in different data sources. The tile number was 21 and the year was 2021.**

However, due to differences of water extents in different datasets, the extracted river maps show significant variations. The DW and EGLC adequately mapped yearly water extents. Many complex rivers (e.g. narrow rivers) were well displayed in these datasets. Thus, river networks in DW and EGLC were dense and complete (Figure 11 (a) & (b)). In contrast, rivers in MIWDR and ESRI were relatively spare, due to considerable narrow waters were not identified (Figure 11 (c) & (d)). For the JRC-GSW, the large rivers were effectively identified, while narrow rivers were not detected ((Figure 11 (e)). This was primarily due to the limitation of its 30 m spatial resolution. The above regulations were also validated in two typical regions

(Figure 12). The DW and EGLC mapped more waters, and narrow rivers in these datasets were well extracted. For the MIWDR, ESRI and JRC-GSW, large rivers could be accurately extracted, while considerable narrow rivers were missed.

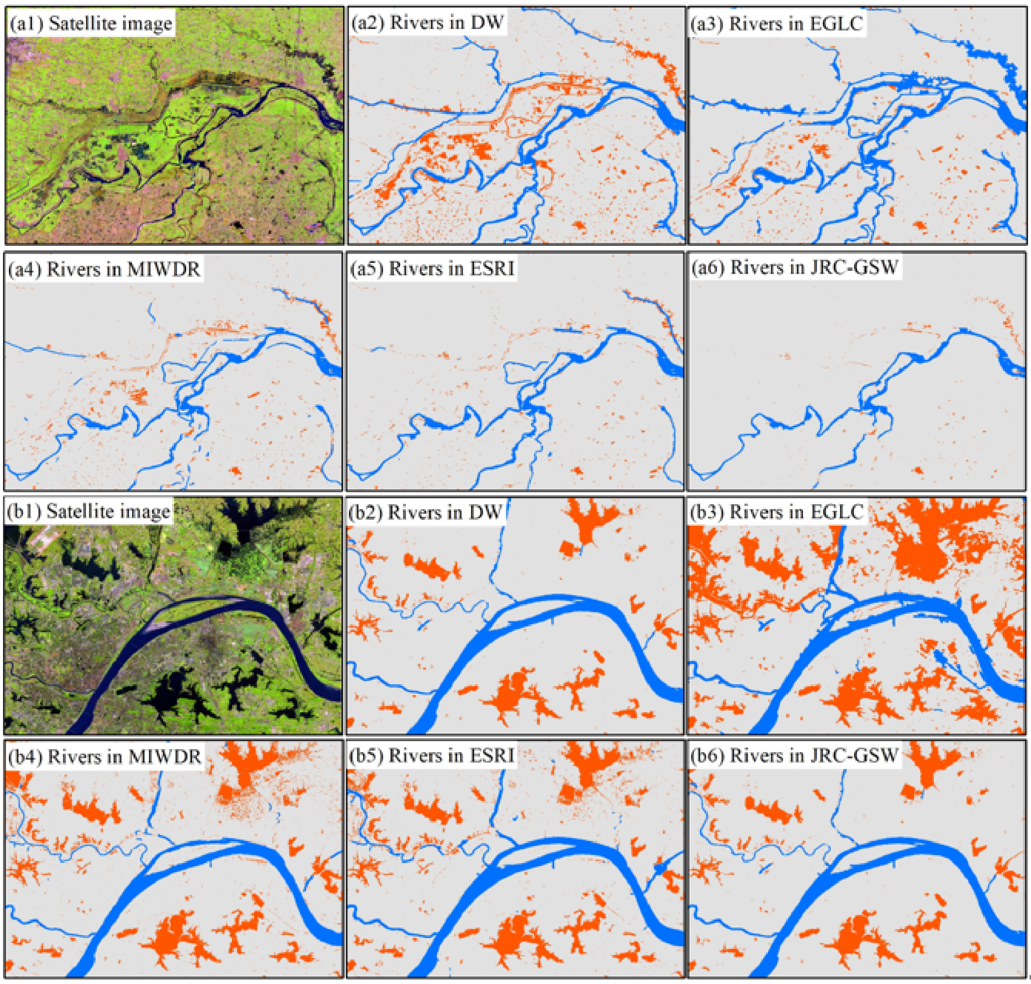


**Figure 12. River distributions of different data sources in two typical regions. The satellite images (a1) and (b1) were Sentinel-2 images that composited using median algorithm based on time series images within 2021.**

To quantitatively evaluate the accuracy of rivers from different datasets, we produced 100 test samples using random sampling method and visual interpretation. Accuracy validation indicated that rivers in DW, ELGC and ESRI achieved good accuracy, with UAs and PAs exceeding 87%. For the MIWDR and JRC-GSW, the UAs of river classification were low, which indicated many rivers, most of narrow rivers, were misclassified or omitted. In contrast, the high PAs of rivers in MIWDR and JRC-GSW shown that rivers displayed in corresponding water maps can be accurately mapped. The above conclusions were consistent with the spatial analysis of rivers in five datasets. It should be noted that the accuracy assessments in tile 21 were different with that of China's rivers. This mainly because the large rivers accounted for a significant proportion of rivers in China, and our algorithm was accurate and effective for rivers with large areas and spatial continuity.

**Table 3. Accuracy assessments of different datasets in 2021**

| | DW | ELGC | MIWDR | ESRI | JRC-GSW |
|---|---|---|---|---|---|
| PA | 83.83% | 90.32% | 67.74% | 80.35% | 48.21% |
| UA | 91.16% | 87.5% | 95.45% | 95.74% | 96.43% |
| F-score | 87.34% | 88.89% | 79.24% | 87.37% | 64.28% |

**2.** The water extraction section in Figure 2 could be clearer. Presenting data for all years together to generate the water time series may cause confusion and fails to adequately convey the meaning of "For areas where DW observations were missing" in line 92.

**Response 2:**

Thanks for your careful review and offer useful suggestions. We have corrected the Figure 2 in the manuscript. The modified Figure is also shown below.

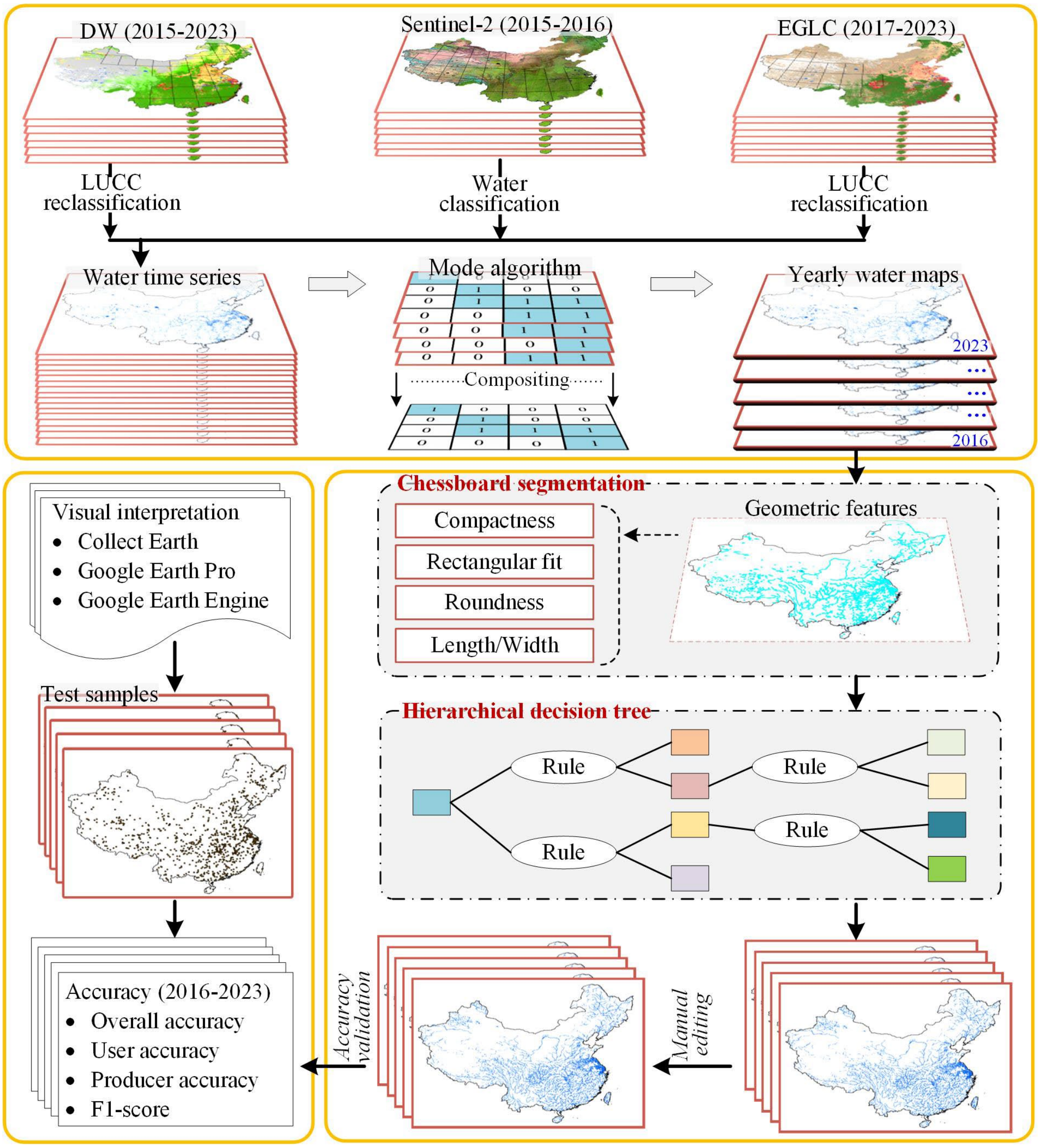


Figure T3. Workflow of annual river extraction. Three datasets marked by back boundaries was chose for river extraction. It should be noted that the boundaries of DW and EGLC varied across different years.

We have rewritten the sentence in line 92.

To facilitate image computation, we divided China into 52 tiles using a 5°×5° grid (Fig. S1). If a tile contains considerable pixels with fewer than 3 valid observations, it was defined as a data insufficient tile. For these tiles, we used the ESRI global land cover (EGLC) dataset to supplement the data (Karra et al., 2021). The EGLC dataset was produced on the basis of

**3.** In line 199, please clarify what "the rivers from 2020" refers to. If it refers to the extraction results from this study for 2020, please clarify the potential impact of generating validation samples based on extraction results on the randomness and representativeness of the samples.

**Resource 3:**

Thanks for your careful review and offer valuable comments. “the river from 2020” was the river map of CRED in 2020. We considered that CRED might omit some rivers, and generating random samples only within the CRED extent would make it difficult to evaluate these omission errors. Therefore, we spatially overlaid the 2020 CRED with the 2020 CNLUCC rivers. The union regions were used to create random river points. This procedure accounts for river identified in CNLUCC that are absent in CRED, allowing for more comprehensive assessment of omission errors.

For the generated random points, we conducted visual interpretation using imagery from different years between 2016 and 2023. These samples may have different attributes across different years, although their location remained unchanged. Thus, the generate samples were used not only to evaluate the accuracy of the 2020 CRED, but also for accuracy assessment of CRED in other years.

We have clarified the confused sentence, and provided more detailed description for generating validation samples.

Specifically, the rivers in the CNLUCC dataset for 2020 were extracted and then overlaid with the rivers from 2020 of CRED. In the union regions, river samples were created via random sampling. After that, non-river samples were produced within a 300 m outside buffer of the union regions. Second, all random samples were visually interpreted by combining Collect Earth (CE), Google Earth (GE) and the GEE platform (Peng et al., 2024b). The CE software enables user-friendly sample management, the GE provides high spatial resolution images, and the GEE offers median-composited Sentinel-2 images. We visually labelled these samples based on imagery from different year from 2016 to 2023. Correspondingly, their attributions may vary across different years, although their locations remain unchanged. Using these three platforms, river and non-river samples from 2016 to 2023 were produced.

**4.** The sample size for validation is unclear. Please provide details on the distribution and quantity of the validation samples in Section 4.1.

**Response 4:**

Thanks for your useful comments. This suggestion is definitely helpful to enhance our manuscript. We have revised this point in our manuscript. The sample size from 2016 to 2023 and their spatial distribution was displayed in supplementary materials.

from 2016 to 2023. We validated the CRED via test samples that were manually inspected via visual interpretation. The sample sizes from 2016 to 2023 were shown in Table S1, and their spatial distribution was shown in Fig. S2. The CRED achieved

Table S1. Sample size of river and non-river from 2016 to 2023 in China

| | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 | 2023 |
|---|---|---|---|---|---|---|---|---|
| River | 196 | 291 | 275 | 292 | 295 | 266 | 276 | 259 |
| Non-river | 652 | 597 | 585 | 572 | 542 | 586 | 567 | 596 |

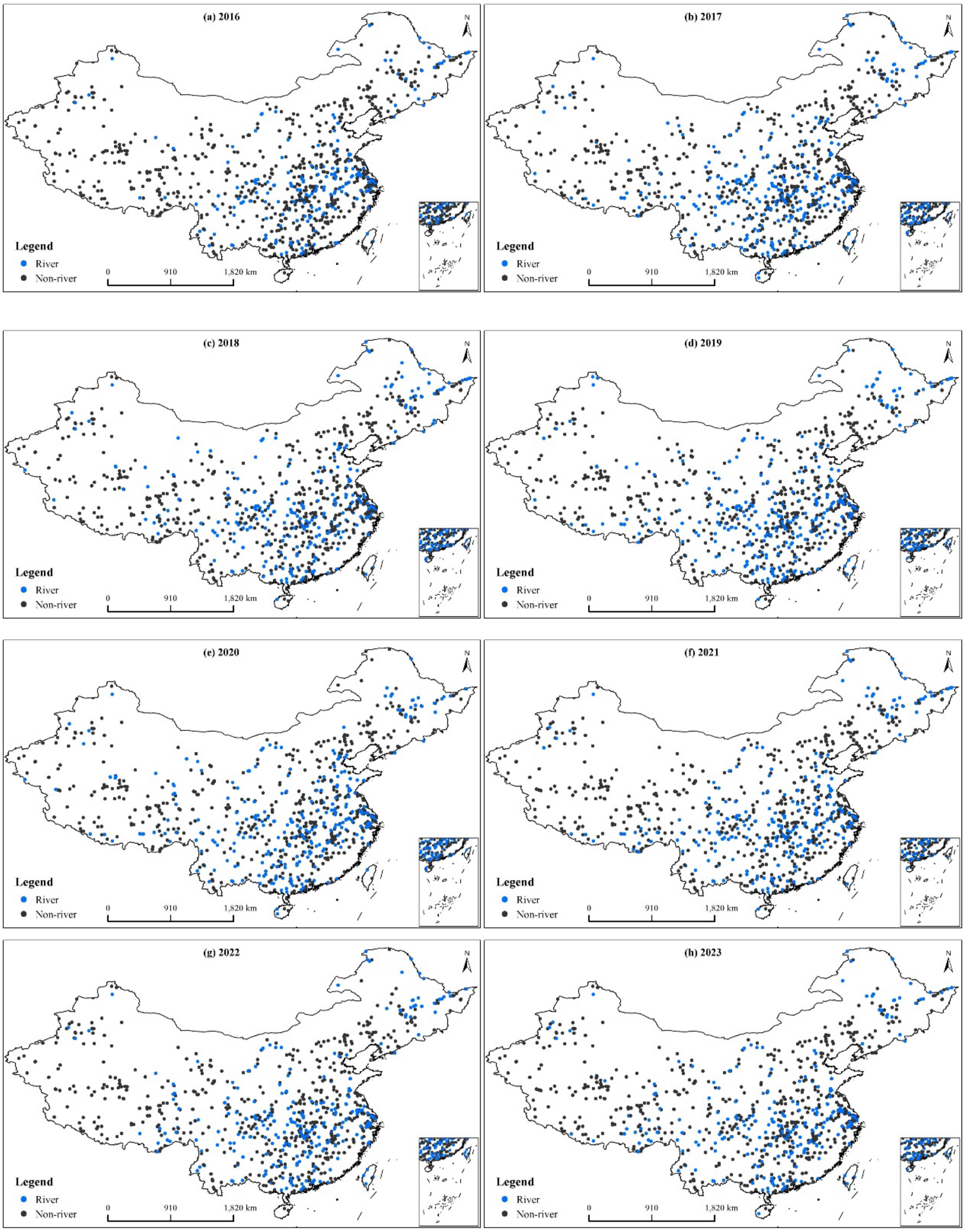

Fig. S2. Spatial distribution of river and non-river samples from 2016 to 2023

**5.** The resolutions of the three existing products used for comparing river extraction results are not exactly the same. Did the authors perform any resampling or other processing when comparing river areas to eliminate the area differences caused by resolution?

**Response 5:**

Thanks for your careful review and offer professional comments. We did not perform any resampling or other processing when make data inter-comparison. In our study, three datasets—CNLUCC, CWaC and EA_Wetlands—were used to make data inter-comparison. The spatial resolution of CWaC and EA_wetlands is same as our CRED, with a spatial resolution of 10 m.

The CNLUCC, 30-m resolution datasets, had a lower spatial resolution compared to our CRED. Indeed, due to this spatial limitation, some narrow rivers that can be identified in CRED are not detected in CNLUCC. However, given the relatively similar spatial resolutions and the high achieved through manual visual interpretation, we included the CNLUCC in our comparative analysis. This procedure may contain some uncertainties into the comparative analysis. We descripted this point in the Discussion section.

> as by using the maximum, minimum or medium water inundated areas, is a debated issue. For the data inter-comparison between CNLUCC and CRED, there may be some uncertainties due to differences of spatial resolution. For examples, some narrow streams are extracted in CRED but not in CNLUCC. Meanwhile, we did not use test samples to assess rivers' accuracy

**6.** In line 270, it is mentioned that the CRED dataset outperforms the existing most accurate products in extracting narrow rivers in mountainous areas. Did the authors consider providing a more precise definition of narrow rivers to highlight the advantages of this product?

**Response 6:**

Thanks for your careful review and offer specific suggestion. We added the definition of narrow rivers in the manuscript.

> their spatial disconnection. In our study, narrow rivers were defined as linear water bodies with a width greater than 10m. Generally, rivers with a width exceeding 30 m exhibited good spatial continuity in Sentinel-2 imagery, which could be automatically and accurately extracted by our algorithm. However, narrow rivers with width less than 30 m, due to spatial discontinuities, were challenging to be identified using geometric features. These narrow rivers were manually edited to improve our rivers' accuracy.

**7.** The area difference mentioned in line 273 between the river areas of CWaC and the CRED in 2020 is inconsistent with the visualization results in Fig. 8. Please provide more detailed comparisons of the water bodies extracted result.

**Response 7**:

Thanks for your careful review and offering valuable comment.

In Figure 8, the river extents shown by CWaC appears larger than that of our CRED, mainly due to the display of vector data under different scale. For the CWaC, there are large amount of fragmented river waters, most of which are smaller than 1 ha. When displayed at the national scale, these fragmented waters are stack together. Even with a very small line width set for these fragmented river waters, their display at national scale is still evident. To illustrate this phenomenon more clearly, we mapped rivers with areas smaller and larger than 1 ha separately. It was found that CWaC has large amount of fragmented river waters, and its spatially continuous rivers are fewer than CRED.

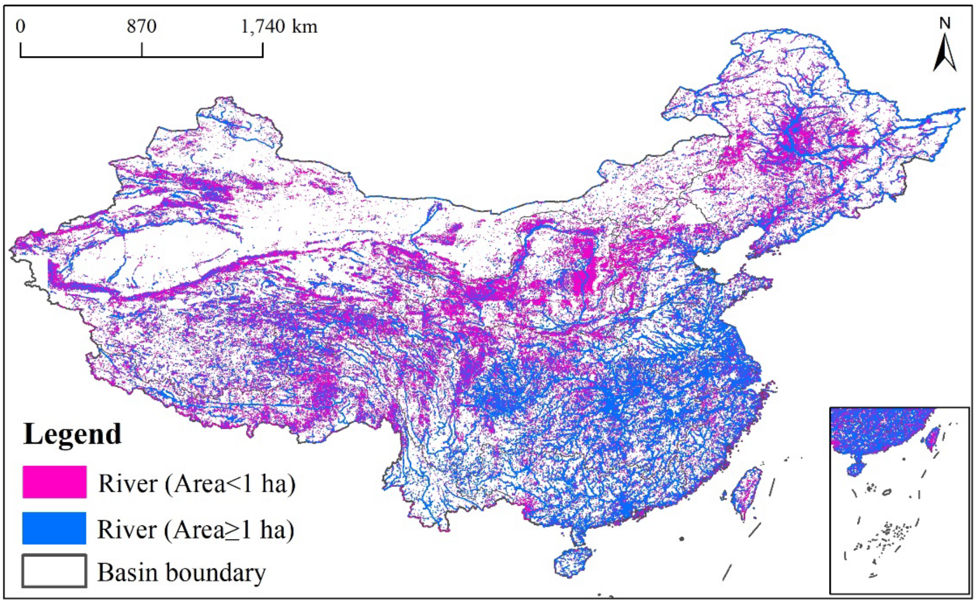


Figure T4. The spatial distribution of two group of rivers in CWaC (2020)

To better visualize these small rivers, we present the spatial comparison of CWaC and CRED at the region scale in typical areas. The results indicate that in large-scale maps, fragmented rivers are stacked together. In small-scale maps, these waters are shown as individual small-area patches. Due to the above phenomenon, the rivers in CWaC are visually larger than those in CRED, but the river area counted in CWaC is smaller than in CRED. We have added these two figures to the supplementary material and described this phenomenon in the revised manuscript.

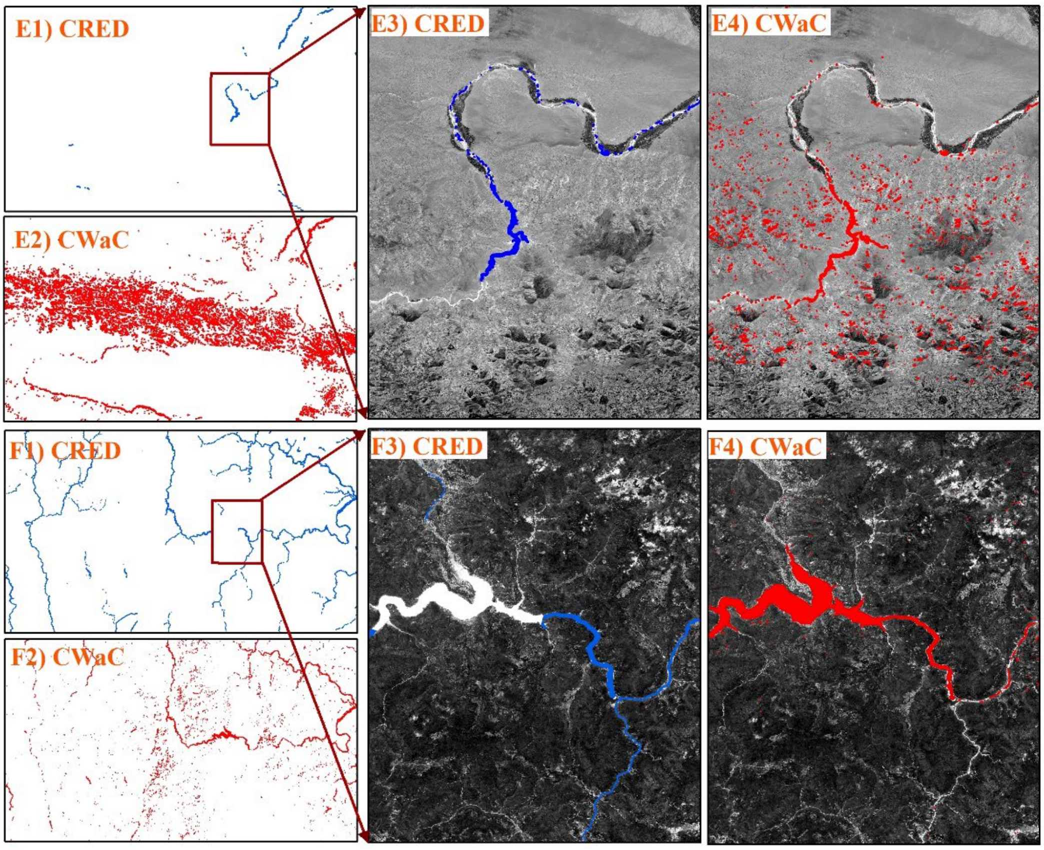


Figure T5. Spatial comparison of rivers between CRED and CEaC in 2020

For example, CWaC misclassified some reservoirs as rivers, whereas the CRED accurately excluded them (Figure S5 (F)). It should be noted that the rivers in CWaC appeared larger than these in CRED visually, mainly due to difference in display at different scale. To clearly illustrate this phenomenon, we mapped rivers of CWaC with areas less than and larger than 1 ha separately (Fig. S6).