https://doi.org/10.5194/essd-2026-396
Preprint. Discussion started: 7 September 2026

Open Access Earth System Science Data Discussions

A Monthly 30m Dataset of Lakes Larger than 1 km^2 on the Tibetan Plateau from 2001 to 2025

Aashish Tiwari [1, 2], Qiang Liu [1,2], Liqiao Liang [3]

[1.] State Key Laboratory of Wetland Conservation and Restoration, School of Environment, Beijing Normal university, Beijing 100875, China;

[2.] Key Laboratory for Water and Sediment Sciences, Ministry of Education, School of Environment, Beijing Normal University, Beijing 100875, China;

[3.] State Key Laboratory of Tibetan Plateau Earth System, Environment and Resources (TPESER), Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing 100101, China;

Correspondence: Qiang Liu (qiang.liu@bnu.edu.cn)

Abstract

Long-term monthly records of lake dynamics are critical for understanding hydrological and climatic change on the TP. However, existing products are constrained by factors such as cloud cover, snow and ice contamination, terrain shadow, observational gaps, and scale-related bias. Here we present a new monthly dataset of lakes larger than 1 km^2 across the TP for 2001-2025 at 30 m spatial resolution. The dataset was generated using TSRFD framework integrating MODIS, Landsat and terrain variables, with refinement steps to improve mapping accuracy and boundary consistency. Technical validation against three reference datasets showed strong spatial agreement (IoU = 0.80-0.92) and very high consistency in lake area estimates (R^2 > 0.99). Uncertainty analysis indicated that most lake-months were classified as having low to moderate uncertainty. From 2001 to 2025, lakes on the TP expanded in both number increasing by 20.6% and total area expanding by 22.9%, despite interannual variability. Small lakes contributed to the increase in lake number, while larger lakes were responsible for most to area expansion. Both metrics showed clear seasonal peaks in late summer. This dataset provides a valuable long-term resource for investigating lake dynamics, hydrological variability, and climate-lake interactions in high-altitude regions.

1. Introduction

Lakes on the Tibetan Plateau (TP) are critical components of the cryosphere and hydrosphere, and are widely regarded as sensitive indicators of climate change (Zhang & Duan, 2021; Leng et al., 2024). Consequently, monitoring lake dynamics is essential for assessing regional environmental

https://doi.org/10.5194/essd-2026-396
Preprint. Discussion started: 7 September 2026

and climatic changes. Such observations contribute to a better understanding of regional cryosphere variations and their implications for ecosystems, energy balance, hydrological processes, and human livelihoods (Zhang et al., 2021; Zhao et al., 2022).

Table 1: Limitations and drawbacks of existing monthly lake area mapping methods

References	**Lake sizes (km^2)**	**Methods**	**Limitation/Drawbacks in TP**
Pekel et al. (2016) (GSWD)	> 0.01	Rule-based expert-system classifier	Gap in lake observation
Zhao et al. (2022).	> 0.1	Reconstructed form GSWD	Better for larger lakes
Zhao et al. (2025).	> 10	ML-based water indices	Biased in lake area due to MODIS 500 m scale data
Wu et al. (2025)	> 1	NDWI	Gap before Sentinel-2 data
Li et al. (2025)	> 0.1	U-Net spatio-temporal fusion (GSWD + MODIS)	Inherit biases from GSWD and computational expensive and complex

Lake area mapping studies for the TP have predominantly focused on annual-scale inventories of large lakes (Zhang et al., 2019; Zhao et al., 2022; Zhou et al., 2024; Wang et al., 2024), while available monthly products are still hindered by substantial observational gaps arising from persistent cloud cover, snow and ice contamination, and terrain shadowing (Pekel et al., 2016; Wu et al., 2025). Moreover, estimates derived at the MODIS spatial resolution data are susceptible to scale-dependent bias (Zhao et al., 2025). Although Li et al. (2025) partially addressed these limitations by using a deep learning-based spatio-temporal fusion framework using U-Net, and produced a global monthly lake-surface-extent dataset, their approach remains computationally intensive and methodologically complex. More importantly, there is a regional gap persists for the TP, where a dedicated framework is required to better resolve the combined effects of cloud contamination, snow and ice interference, terrain shadowing, and the need for a continuous long-term monthly record. These challenges collectively imply that existing monthly datasets still suffer from significant uncertainties and methodological constraints, as summarized in Table 1.

https://doi.org/10.5194/essd-2026-396
Preprint. Discussion started: 7 September 2026

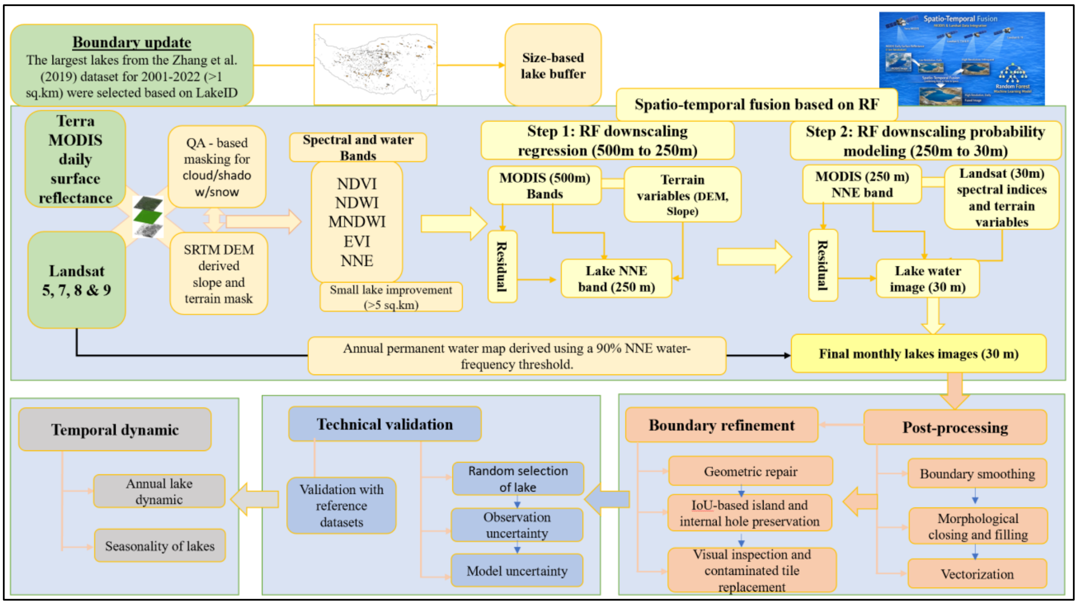

Figure 1: Overview of the workflow used to generate monthly 30 m lake maps over the Tibetan Plateau based on a two-step random forest downscaling framework integrating MODIS-Landsat data. Step 1: applies RF regression to downscale MODIS data from 500 m to a 250 m NNE water band, and Step 2: applies an RF probability model to further downscale the 250 m product to a 30 m lake water image. Both steps incorporate residual correction and are constrained by spectral indices and terrain variables. The framework further includes boundary updating, boundary refinement, post-processing, technical validation, and assessment of annual and seasonal lake dynamics.

In light of the limitations and drawbacks of existing monthly lake-area datasets, we employed a two-step Random Forest downscaling model with MODIS-Landsat integration (TSRFD), adapted from Wen et al. (2025). within contrast to more computationally intensive approaches with Li et al. (2025), this framework is relatively simple, efficient, and well suited for implementation in Google Earth Engine. The random forest (RF) approach has proven to be an effective method for lake-water classification and has shown strong performance in spatiotemporal fusion-based water mapping (Tesfaye & Breuer, 2025; Wen et al., 2025; Zhao et al., 2025). The TSRFD framework employs an NNE-based labeling strategy, in which water pixels are identified using the NDWI-NDVI/EVI criteria (NDWI > NDVI or NDWI > EVI) following the NDWI-NDVI-EVI algorithm of Li et al. (2023). Thus, the TSRFD-based approach helps fill the gap in long-term monthly lake dynamics by providing a more continuous and detailed record, thereby enabling deeper understanding of the underlying hydrological drivers and processes across the TP. This study

https://doi.org/10.5194/essd-2026-396
Preprint. Discussion started: 7 September 2026

presents a long-term monthly lake dynamics dataset for lakes larger than 1 km^2 on the TP from 2001 to 2025.

2. Study area and data

2.1 Study Area

The TP, located in the center of Asia, is the world's second-largest plateau, with a mean elevation of about 4,000 m above sea level. Most of its area lies above 2,500 m, covering a total area of 3,080,565 km^2 (Fig. 2). is the plateau is divided into 12 river basins, including 9 exorheic basins like Amu Darya, Brahmaputra, Ganges, Hexi, Indus, Mekong, Salween, Yangtze, and Yellow; and 3 endorheic basins as Tarim, Inner TP, and Qaidam (Wan et al., 2016; Zheng et al., 2025).

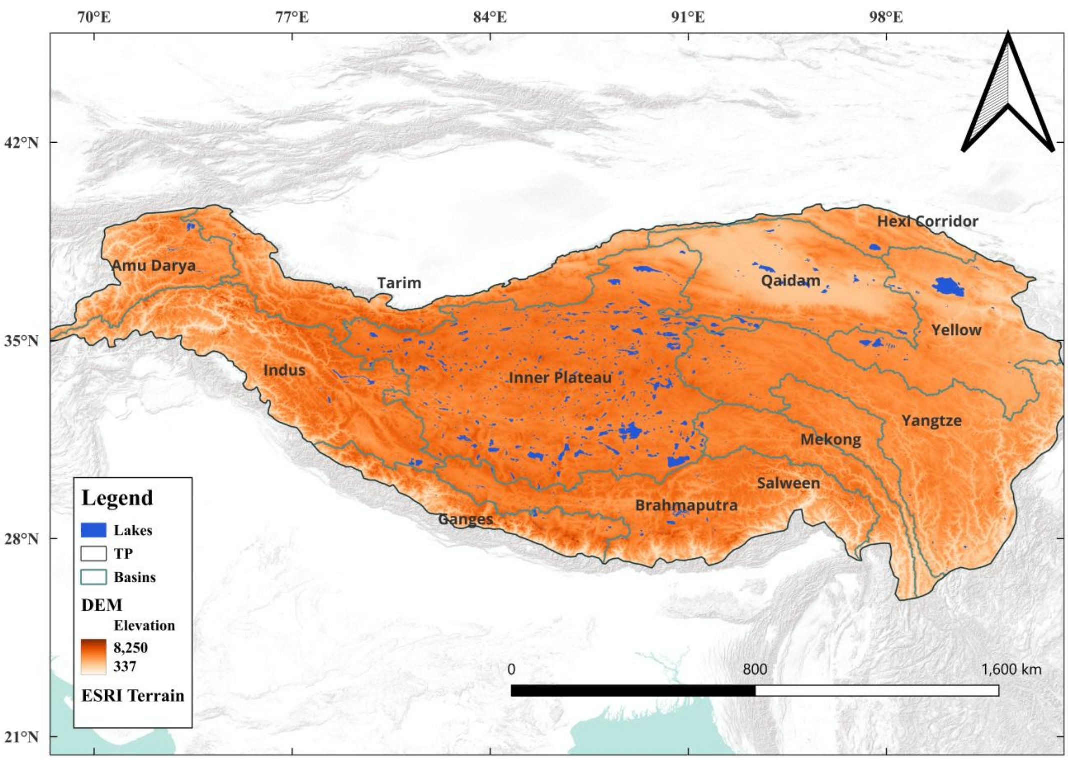

Figure 2: Spatial distribution of lakes across the Tibetan Plateau, with lake extent in blue, the plateau boundary and elevation in shades of orange, and the 12 major river basins in green.

2.2 Data sources

Table 2 summarizes the main satellite and elevation datasets used in this study. The analysis was primarily relied on multi-source remote sensing data, including MODIS daily surface reflectance

https://doi.org/10.5194/essd-2026-396
Preprint. Discussion started: 7 September 2026

at 500 m spatial resolution, Landsat 5, 7, 8, and 9 surface reflectance data at 30 m resolution, and a digital elevation model (DEM) obtained from the USGS (Table 2).

Table 2: Satellite, elevation and other data sources utilized in this study.

SN	Data	Source/references
1	MODIS daily reflectance (500m)	http://modis-sr.ltdri.org
2	Landsat 5, 7, 8 & 9 surface reflectance (30m)	https://www.usgs.gov/
3	TP DEM	https://www.usgs.gov/
4	River Basin of TP	https://data.tpdc.ac.cn/

3. Method

3.1 Monthly Lake Dataset Generation

The dataset was generated using a TSRFD that integrates MODIS temporal continuity with Landsat spatial detail for monthly 30 m TP lake mapping, following the workflow outlined below and illustrated in Figure 1.

3.1.1 Determining lake boundaries

The multi-year lake inventory of Zhang et al. (2019) as used to define the study lakes and quantify monthly lake area for lakes larger than 1 km^2. For each lake, the maximum recorded extent associated with a given LakeID was adopted as the reference boundary to establish the analysis domain. To account for differences in lake size and seasonal variability in shoreline position, a size-dependent buffering strategy was applied around each reference polygon. The workflow was implemented in a tile-based manner to enable efficient processing over the TP while minimizing edge effects during modelling and export.

3.1.2 Preprocessing

Daily 500 m MODIS surface reflectance and 30 m Landsat surface reflectance imagery from Landsat 5, 7, 8, and 9 were used to derive monthly lake boundaries. Both MODIS and Landsat data were preprocessed to remove clouds, cloud shadows, snow, and other low quality observations using QA-based masking procedures. Additionally, terrain variables derived from the SRTM DEM, such as elevation, slope, and hillshade-related screening, were incorporated to reduce topographic noise and minimize misclassification in mountainous terrain, following the general logic of Li et al. (2023). Separate terrain masks were used for model training and prediction,

https://doi.org/10.5194/essd-2026-396
Preprint. Discussion started: 7 September 2026

with a stricter mask for training to exclude unreliable samples and a comparatively looser mask for prediction to preserve shoreline information.

3.1.3 Indices and Water Band

For each month and region, MODIS observations were aggregated to generate monthly composite layers of NDVI, NDWI, MNDWI, EVI, and an NNE-based water indicator, and the resulting bands were renamed as mNDVI, mNDWI, mMNDWI, mEVI, and mWater. For Landsat, all valid cloud-masked observations from Landsat 5/7/8/9 within a given month were merged to generate monthly predictor layers. In this step, median composites were used for Landsat spectral indices, while additional water summary layers describing monthly water occurrence were also derived.

The NNE water indicator followed the methodology outlined by Li et al. (2023), where water pixels were identified using the condition (NDWI > NDVI) or (NDWI > EVI). This NNE-based rule was adopted because it has been shown to perform effectively for lake-water identification on the TP, outperforming alternative like MNE and IMNE (see Li et al., 2023 for details). Additionally, a 30 m resolution DEM and slope data were further added as auxiliary predictor variables for the fine resolution model. Unlike a simple temporal majority rule approach, monthly lake area in this workflow was estimated through a TSRFD procedure in which MODIS monthly water information and Landsat-derived predictors were jointly used to derive the final 30 m monthly water map.

3.1.4 Small lakes improvement

The delineation of small lakes was enhanced through a series of size aware adjustments tailored for lakes smaller than 5 km^2. This adjustment included the use of median-based Landsat composites, water-any labeling to preserve weak water signals, size aware core sampling strategies, lake focused training masks, lower probability thresholds for final classification. Additionally, coarse-fine scaling was excluded for small lakes, and stronger size dependent post-processing during boundary extraction. Together, these modifications reduced omission errors, improved the retention of fragmented or weak water features, and enhanced shoreline continuity for small lakes.

3.1.5 RF based two-step downscaling

TSRFD framework was used to delineate monthly lake boundaries by integrating MODIS and Landsat observations, following a method similar to that of Wen et al. (2023). The workflow consisted of three main steps.

https://doi.org/10.5194/essd-2026-396
Preprint. Discussion started: 7 September 2026

3.1.5.1 Step 1: Downscaling from 500m to 250m using RF regression with residual correction

In the first step, monthly MODIS water information at 500 m spatial resolution was downscaled to 250 m using an RF regression model. Predictor variables included MODIS-derived NDVI, NDWI, MNDWI, and EVI, along with terrain variables such as elevation and slope. The MODIS monthly water layer served as the target variable. To ensure consistency with the original coarse resolution observations, residuals between RF-predicted and observed MODIS water values were calculated, spatially interpolated, and then added back to the regression output, thereby refining the downscaled 250 m water layer.

3.1.5.2 Step 2: Downscaling from 250 m to 30 m using RF probability modelling with residual correction and conditional coarse-fine consistency

In the second step, the intermediate 250 m water layer was integrated with Landsat-derived spectral indices, terrain variables, and Landsat validity information to estimate water probability at 30 m resolution. A two-stage RF modelling strategy was employed, consisting of a probability-based classification model followed by a regression model to correct residual prediction errors. Training samples were derived from strict water and land labels identified from Landsat water occurrence, spectral constraints, and terrain screening, while separate masking strategies were used for training and prediction. For lakes larger than or equal to 5km^2, the fine resolution predictions were further adjusted to preserve consistency between aggregated 30 m water probability and the MODIS-derived 250 m water fraction. This coarse-fine consistency adjustment was not applied to lakes smaller than 5 km^{2}in order to avoid suppression of weak but meaningful small-water signals.

3.1.5.3 Step 3: Integration with yearly permanent water

The monthly RF-derived water maps were further stabilized using an annual permanent water mask derived from Landsat observations. Annual water frequency was calculated from all valid Landsat observations within each year, and pixels identified as water in at least 90% of valid observations were classified as permanent water. Zhao et al. (2022) and Zhou et al. (2022) demonstrated that, while truly permanent water should theoretically have a frequency of 100%, a lower threshold better account for missed detections caused by cloud, shadow, and image quality limitations; in their case, a 75% threshold was adopted for permanent water identification. In contrast, Li et al. (2023) defined permanent water more strictly as pixels with a 100% water

https://doi.org/10.5194/essd-2026-396
Preprint. Discussion started: 7 September 2026

frequency while 25-75% used to identify seasonal water. Considering these differences, a 90% threshold was adopted here as a balanced criterion to retain persistent water bodies while reducing omission errors. The permanent water pixels were then merged with the monthly RF results to improve temporal stability and reduce the omission of stable water bodies.

3.1.6 Lake boundary extraction and post-processing

Lake boundaries were extracted through a series of post-processing steps to obtain the final monthly lake area. First, to improve the geometric consistency of the predicted 30 m water mask, morphological smoothing was applied within buffered lake regions. A circular kernel with a radius of one pixel was used to perform dilation ($focal_{max}$) followed by erosion ($focal_{min}$), which helped reconnect fragmented water pixels and remove isolated noise. For lakes smaller than $5km^2$, a slightly stronger dilation with two iterations was applied to better preserve small and fragmented lake boundaries.

After smoothing, internal holes within the buffered lake regions were identified as non-water pixels fully surrounded by water. Connected components smaller than 100 pixels (approximately $0.09km^2$) were treated as artefacts and filled to produce a continuous lake water mask (waterFilled_30m), thereby reducing residual salt-and-pepper noise and improving the topological consistency of lake boundaries (Chi et al., 2017; Zhao et al., 2025). Finally, the hole-free raster water mask was converted to vector polygons at 30 m resolution, and only polygons classified as water were retained. A spatial join was then performed to assign each polygon to the corresponding LakeID from the reference lake dataset (Zhang et al., 2019), with polygons lacking a valid LakeID excluded from the final output.

3.1.7 Boundary refinement and geometric correction

The obtained monthly vector lake polygons were further refined through geometric repair and LakeID-based dissolving to close gaps within polygon features and correct topological anomalies. The repaired lake boundaries were then enhanced using IoU-based boundary control against the reference layer of Zhou et al. (2024), which helped identify and preserve internal holes and islands before export as a corrected global lake vector dataset. Additionally, the final dataset was manually inspected to ensure boundary reliability. Since the dataset was generated using a tile-based workflow, visually contaminated tiles were identified and replaced with data from the nearest error-free month.

https://doi.org/10.5194/essd-2026-396
Preprint. Discussion started: 7 September 2026

Earth System Science Data
Open Access Discussions

3.2 Technical validation

I. Uncertainty quantification

To evaluate the reliability of the monthly lake mapping results, a lake size-based uncertainty framework was developed for each lake month. The framework considers two complementary sources of uncertainty: observation support uncertainty, which reflects the strength of direct Landsat support for the mapped lake extent, and model confidence uncertainty, which reflects the reliability of the RF-assisted MODIS-Landsat classification workflow under data limited conditions. Since the product was derived from cloud masked optical imagery and a multi-step downscaling and classification procedure, uncertainty may arise from both incomplete valid observations and reduced model robustness under difficult mapping conditions. These two components were represented by five workflow-based variables: i) Scene Count, ii) Valid Pixel Fraction, iii) Mean RF Probability, iv) Fallback Dependence, and v) Permanent Water Merge Fraction. All variables were first normalized to a common scale ranging from 0 to 1, where 0 represents low uncertainty and 1 represents high uncertainty. They were then integrated into sub-indices and a final composite uncertainty index. For clearer interpretation, the continuous uncertainty index was subsequently classified into five ordinal categories: very low (<0.20), low (0.20 to <0.40), moderate (0.40 to <0.60), high (0.60 to <0.80), and very high (≥0.80). These categories provide a simple and intuitive summary of lake-level mapping reliability for each monthly result.

3.2.1.1 Random selection of lakes

Lakes from different parts of the TP were selected across all directions, stratified by size class. A total of 20 lakes were chosen from smaller size classes (1-5km^2), while 10 lakes were selected from each of the three larger classes (5-10 km^2, 10-50km^2, and >50 km^2). Monthly uncertainty was then estimated using the method described below for the years 2001, 2005, 2010, 2015, 2020, and 2025. The sample-based uncertainty assessment characterizes overall product reliability across lake sizes, seasons, and satellite data periods rather than uncertainty for every individual lake month polygon.

3.2.1.2 Observation support uncertainty

Observation support uncertainty was represented by Scene Count and Valid Pixel Fraction, which describe the temporal and spatial strength of direct Landsat support for each lake month. Scene

https://doi.org/10.5194/essd-2026-396
Preprint. Discussion started: 7 September 2026

Count was defined as the maximum number of valid Landsat observations within the buffered lake region of interest (ROI) during the target month after masking clouds, cloud shadows, and snow. Higher scene availability generally improves temporal support and reduces uncertainty in optical water mapping, whereas sparse clear sky observations increase sensitivity to data gaps and scene specific effects (Pekel et al., 2016; Mullen et al., 2021). Scene count uncertainty was defined as 1 when the number of available scenes was 1 or fewer, indicating high uncertainty, and as 0 when the number of available scenes was 6 or more, indicating low or negligible uncertainty (Equation 1). Valid Pixel Fraction described the spatial completeness of usable Landsat observations within the buffered ROI after removal of contaminated pixels (Equation 2), following the logic used in Landsat-based surface water studies (Yue et al., 2023; Hanly et al., 2025). Together, these two variables provide complementary measures of direct observational support. These variables can then be estimated using the following equations 1 to 3.

- Scene Count Uncertainty was defined as:

$$U_{scene} = \frac{1-(SceneCount-1)}{6-1} \qquad (if\ 1 < SceneCount < 6) \text{ ....eq. 1}$$

- Valid Pixel fraction was calculated as:

$$Valid\ Pixel\ Fraction = \frac{Number\ of\ valid\ Landsat\ pixels}{Total\ pixels\ within\ the\ lake\ buffer} \text{ ....eq. 2}$$

And its uncertainty term was defined as:

$$U_{valid} = 1 - Valid\ Pixel\ Fraction \text{ ....eq. 3}$$

3.2.1.3 Model confidence uncertainty

Model confidence uncertainty was represented by Mean RF Probability, Fallback Dependence, and Permanent Water Merge Fraction, which characterize the confidence of the RF-assisted classification process and the extent to which the final result depended on indirect or supplementary logic. Mean RF Probability was derived from the final RF water-probability surface before thresholding and used as a direct indicator of classification confidence (Equation 4), consistent with previous uncertainty studies based on class-probability outputs (Shadman Roodposhti et al., 2019). Fallback Dependence indicated whether simplified processing was required when valid Landsat observations were unavailable or RF training could not be completed, which is conceptually consistent with continuity preserving MODIS-Landsat integration

https://doi.org/10.5194/essd-2026-396
Preprint. Discussion started: 7 September 2026

approaches (Wu et al., 2023). Permanent Water Merge Fraction quantified the proportion of final monthly water extent contributed by the yearly persistent water layer beyond the RF-only result (Equation 5), reflecting reliance on annual occurrence information to improve completeness when monthly observations were insufficient (Pekel et al., 2016; Zhao & Gao, 2018). Then, these variables can be estimated using the following equations 4 and 5.

- Mean RF probability uncertainty was defined as:

$$U_{rf} = 1 - MeanRFProbabilty \text{ .... eq. 4}$$

- Fallback dependence was represented as a binary lake-month indicator: $U_{fallback} = 1$, if fallback was used; otherwise, 0
- Permanent-water merge fraction-based uncertainty was calculated as:

$$PermanentWaterMergeFraction\ (U_{merge}) = \frac{Final\ water\ area - RF\ only\ water\ area}{Final\ water\ area} \text{ .... eq. 5}$$

3.2.1.4 Final composite uncertainty index

The observation-support sub-index was estimated by assigning slightly greater importance to the valid-pixel fraction than to scene count, as it more directly reflects the amount of usable spatial evidence. The model confidence sub-index was then derived by giving the highest weight to mean RF probability, followed by fallback dependence and the permanent water merge fraction, because RF probability provides the strongest indication of model certainty. The overall uncertainty index was calculated by assigning slightly greater weight to observation support than to model confidence, reflecting the primary role of monthly optical observations while also accounting for the influence of RF classification and fallback logic. In its expanded form, the final composite uncertainty index integrates the contributions of scene count, valid pixel fraction, RF probability, fallback dependence, and permanent-water merging in a balanced framework (Equation 6), with all weights summing to 1.00 to ensure straightforward interpretation.

$$U_{final} = 0.22U_{scene} + 0.33U_{valid} + 0.225U_{rf} + 0.135U_{fallback} + 0.09U_{merge} \text{......................eq. 6}$$

https://doi.org/10.5194/essd-2026-396
Preprint. Discussion started: 7 September 2026

3.2.2 Validation with reference datasets

Accuracy was assessed using a one-to-one lake matching approach between mapped and reference polygons based on the Intersection over Union (IoU). For each mapped lake, the reference polygon with the highest IoU was identified, and only matches exceeding the selected IoU threshold (0.70) were retained. For the matched lake pairs, spatial agreement was further evaluated using IoU, Dice coefficient, area bias, and agreement in lake area estimates, including R^2, RMSE, and bias. Among the available datasets, we selected three reference datasets Zhang et al. (2019), Zhou et al. (2024), and Zhao et al. (2025) for validation and comparison.

4. Data set description

The dataset is organized into 25 annual folders, each containing 12 monthly shapefiles as “MonthlyDataset_larger1sqkm\Year\Year_Month.shp". The attribute table for each monthly file includes LakeID, LakeArea, and RiverBasin for every mapped lake.

5. Technical Validation

5.1 Uncertainty quantification

The Uncertainty (U_{final}) results from Equation 1 to 6 indicate that the monthly lake-mapping product is generally reliable, with most lake-months classified as low to moderate uncertainty. Uncertainty decreases systematically with increasing lake size, from the 1-5 km^2class to the >50 km^2 class, demonstrating that larger lakes are mapped more robustly than smaller lakes (Fig. 3). A pronounced seasonal cycle is also evident, with uncertainty highest in winter-early spring and lowest in late summer-autumn (Fig. 3a). Interannual patterns further reveal relatively higher uncertainty in the earlier years (2001 and 2005) and improved reliability in the later years (2015, 2020, and 2025) (Fig. 2b). These patterns characterize the sampled lakes and should not be interpreted as lake month specific uncertainty estimates for the complete dataset.

https://doi.org/10.5194/essd-2026-396
Preprint. Discussion started: 7 September 2026

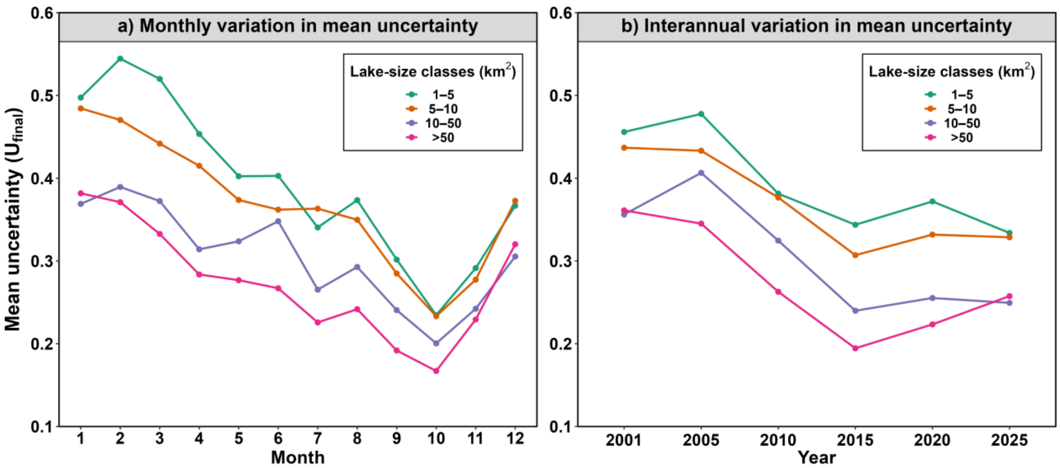

Figure 3: Temporal variation in dataset mean uncertainty (U_{final}) across the sampled lakes, showing (a) monthly variation by lake-size class and (b) variation among the six selected benchmark years. The results represent the stratified uncertainty sample rather than every lake month in the complete dataset.

5.2 Comparison with reference datasets

We evaluated the accuracy of our dataset against three reference datasets, and the results are presented in Figure 4. After filtering all reference datasets to retain lakes >1km^2, the Zhang et al. (2019) and Zhou et al. (2024). datasets were compared with the October observations from our dataset, whereas the comparison with Zhao et al. (2025) was limited to lakes >10 km^2. Overall, the validation indicates strong spatial agreement, with IoU values ranging from 0.803 to 0.923 (Fig. 4) and very high consistency in lake-area estimates ($R^2 > 0.99$). Among the three reference datasets, the lowest RMSE (4.32 km^2) was obtained for Zhang et al. (2019), indicating the closest agreement, whereas Zhao et al. (2025) showed the highest RMSE (24.97 km^2), indicating the largest discrepancy in lake area estimates. Comparisons with Zhang et al. (2019) and Zhou et al. (2024) showed particularly high agreement, with overall IoU values exceeding 0.923 and 0.909 (Fig. 4a1, b1) and mean area bias below 0.80 km^2 and 1.32 km^2 (Fig. 4a2, b2), respectively. By contrast, the comparison with Zhao et al. (2025) showed relatively lower agreement (IoU = 0.803; Fig. 4c1) and a substantially higher area bias (13.46 km^2; Fig. 4c2), suggesting differences in mapped lake extent that may reflect methodological or scale-related effects. Overall, these results demonstrate that the dataset is robust and reliable for monitoring monthly lake dynamics and other climate-hydrology research across the TP.

https://doi.org/10.5194/essd-2026-396
Preprint. Discussion started: 7 September 2026

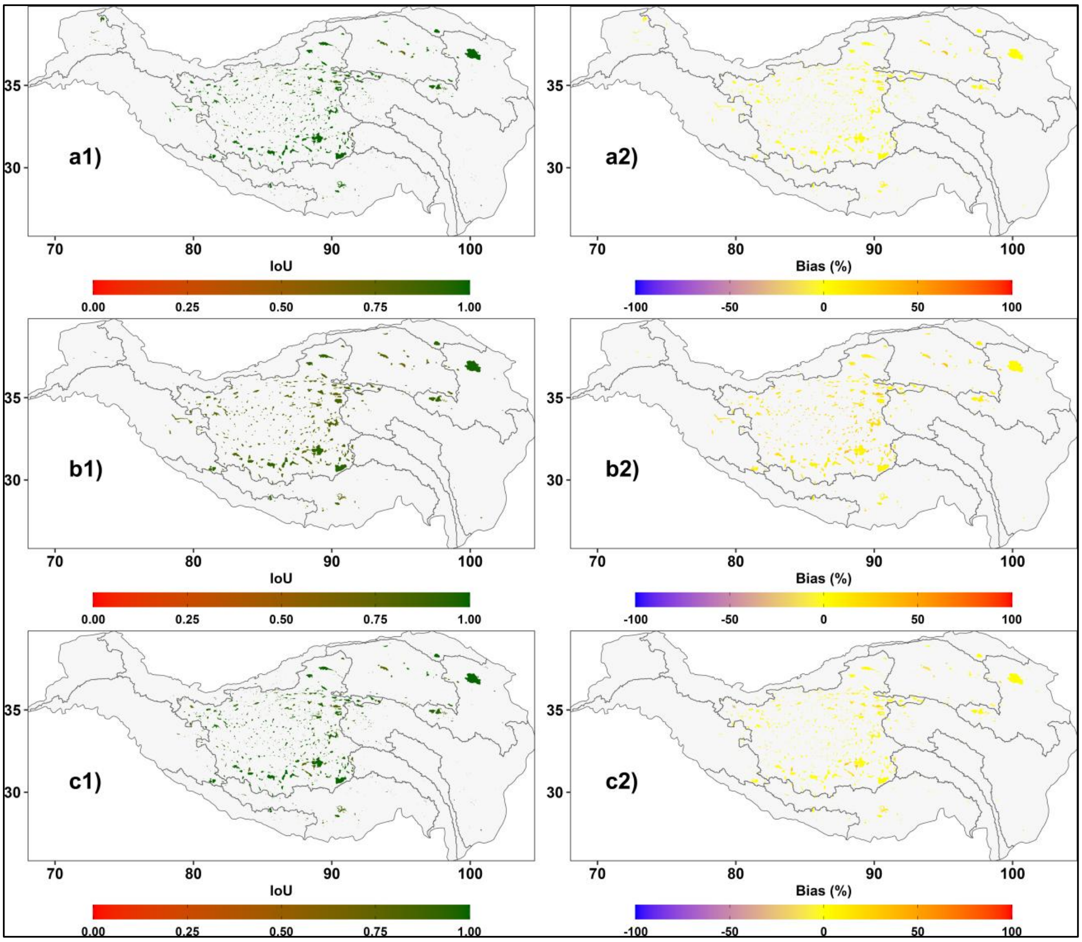

Figure 4: Spatial accuracy maps of the monthly lake dataset evaluated against Zhang et al. (2019), Zhao et al. (2025), and Zhou et al. (2024) across different river basins, where panels a1-a2, b1-b2, and c1-c2 show the comparison results for each reference dataset, with IoU in the left subpanels and bias (%) in the right subpanels.

6. Temporal Dynamic of Lakes

Figure 5 shows that lakes on the TP expanded in both number and area during 2001-2025, despite interannual variability (Fig. 5a). the mean lake number increased from 1,191 to 1,436 (20.6%), while the mean total lake area increased from 42,739.49 km^2 to 52,537.00 km^2 (22.9%). When grouped by size class, small lakes (1-10 km^2) dominated lake number, whereas very large lakes (≥200 km^2) accounted for the largest share of total lake area (Fig. 5b), indicating that the increase in lake number was mainly driven by small lakes, while the increase in total lake area was primarily associated with larger lakes. Annual change rates varied among years and size classes, but positive

https://doi.org/10.5194/essd-2026-396
Preprint. Discussion started: 7 September 2026

changes were generally stronger than negative ones, consistent with the long-term expansion trend (Fig. 5c). The strongest increases occurred around 2017, particularly for the 1-10 km^2 and 10-50 km^2 classes, whereas some classes showed short-term declines during 2020-2021. Seasonally, both lake number and lake area were lowest in February and increased through spring and summer, with lake area peaking in August and lake number in September, before declining toward winter (Fig. 5d).

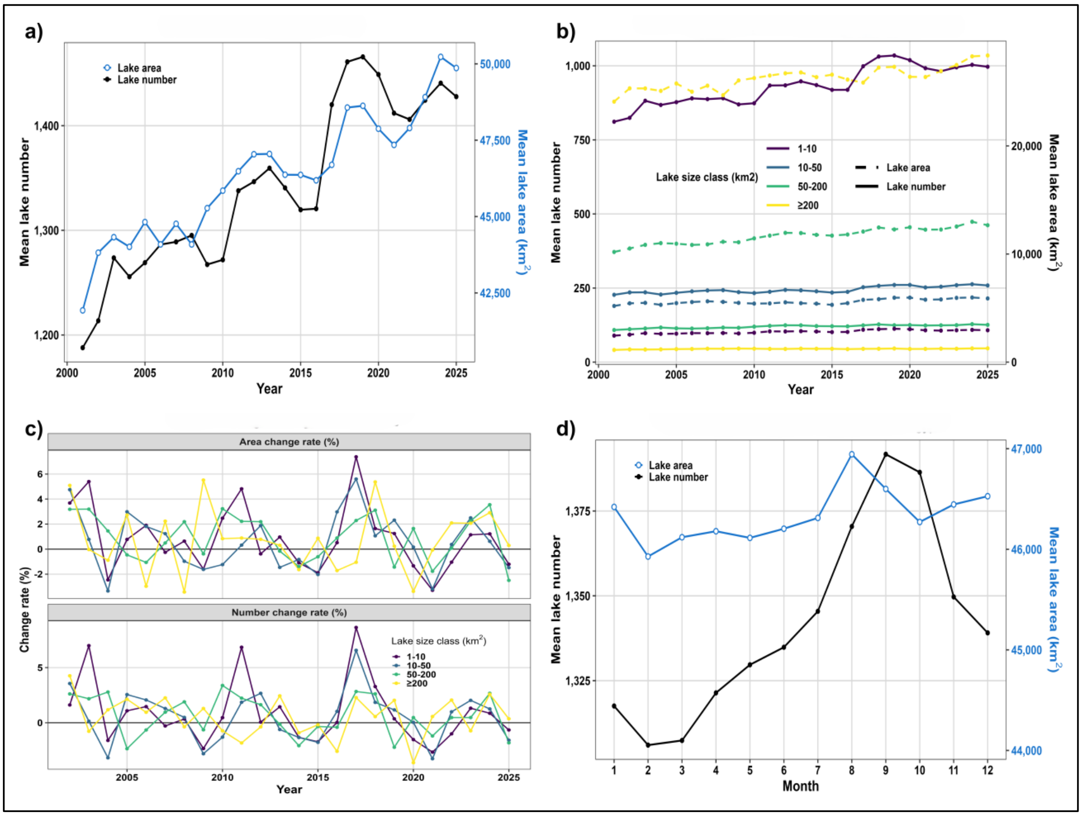

Figure 5: Four-panel combined figure showing lake dynamics on the Tibetan Plateau: (a) annual variation in total lake area and lake number; (b) annual lake area and lake number by lake size class; (c) annual change rates in lake area and lake number by lake-size class; and (d) seasonal variation in total lake area and lake number.

7. Data Availability

The generated monthly lake polygon dataset provides monthly lake polygon for lake larger than 1 km^2 across the TP from 2001 to 2025. The dataset covers TP lakes within 25°-41°N, 67°-105°E and is provide as monthly shapefile under the dataset name *MonthlyDataset_larger1sqkm*. The

https://doi.org/10.5194/essd-2026-396
Preprint. Discussion started: 7 September 2026

supporting files include *Comparison_ReferenceDataset.xlsx*, which reports validation results against three reference datasets, and *Uncertainty_result.xlsx*, which provides sample-based uncertainty results across lake size classes, spatial sectors, months, and benchmark years. The uncertainty results represent an overall diagnostic of dataset reliability rather than record specific uncertainty for every lake month. The dataset and its supporting file are freely accessible at **https://doi.org/10.5281/zenodo.22107953** **(**Tiwari et al., 2026) with entitled as "*A Monthly 30m Dataset of Lakes Larger than 1 km^2 on the Tibetan Plateau from 2001 to 2025*".

8. Conclusion

In this study, a two-step random forest downscaling framework (TSRFD) was developed to integrate the temporal continuity of MODIS with the spatial detail of Landsat for monthly 30 m mapping of TP lakes larger than 1 km^2. The uncertainty assessment indicated that the resulting monthly lake mapping product was generally reliable. Uncertainty decreased with increasing lake size, while higher uncertainty was observed during winter-early spring and lower uncertainty during late summer-autumn, mainly due to seasonal differences in observation availability and quality. Validation against Zhang et al. (2019) and Zhou et al. (2024) showed high spatial agreement, particularly in terms of overall IoU. In contrast, comparison with Zhao et al. (2025) showed relatively lower agreement and higher area bias, likely reflecting differences in mapping methods, spatial resolution, or scale-related effects. Overall, the proposed framework provides a useful basis for future spatiotemporal lake monitoring using emerging higher-resolution satellite observations, such as Sentinel imagery, and for advancing long-term lake mapping and modelling in high-altitude regions.

Lakes on the TP expanded in both number and area during 2001-2025, despite clear interannual variability. The strongest increases occurred around 2017, while short-term declines were observed during 2020-2021. Mean lake number increased from 1,191 to 1,436, representing a 20.6% increase, while mean total lake area increased from 42,739.49 km^2 to 52,537.00 km^2, representing a 22.9% increase. By size class, small lakes (1-10 km^2) dominated lake counts, whereas very large lakes (≥200 km^2) contributed the largest share of total lake area. Seasonally, both lake number and lake area were lowest in February and increased through spring and summer, with lake area peaking in August and lake number peaking in September before declining toward winter. This dataset can support hydrological assessment, climate-change attribution, lake-evolution analysis,

https://doi.org/10.5194/essd-2026-396
Preprint. Discussion started: 7 September 2026

and model evaluation, while also providing an important input for water-resource planning and regional environmental-change studies at finer temporal resolution.

Author contributions

Aashish T. conceived the study, designed the dataset framework, developed the lake mapping workflow, processed the satellite imagery, generated the monthly 30 m lake dataset, performed data validation and uncertainty assessment, prepared the figures and tables, and wrote the main manuscript text. Qiang L. supervised the study, contributed to the research design and methodological development, provided scientific guidance, and reviewed and edited the manuscript. Liqiao L. contributed to data interpretation, validation design, manuscript revision, and scientific discussion. All authors reviewed and approved the final manuscript.

Acknowledgement

This study was financially supported by the National Natural Science Foundation of China (No. 42271141, 42571159), and the Key Research and Development Program of the Tibet Autonomous Region (No. XZ202401ZY0050).

References

Chi, Y., Lai, R., Yan, Q., and Li, M.: Detection and extraction of mountain shadow information based on Landsat 8 OLI data, J. Mt. Sci., 35, 580–589, https://doi.org/10.16089/j.cnki.1008-2786.000256, 2017.

Hanly, P. J., Webster, K. E., and Soranno, P. A.: LAGOS-US LANDSAT: Remotely sensed water quality estimates for U.S. lakes over 4 ha from 1984 to 2020, Sci. Data, 12, 1315, https://doi.org/10.1038/s41597-025-05600-w, 2025.

Leng, X., Feng, X., Feng, Y., Sun, C., Liu, X., Zhang, Y., Zhou, C., Wang, Y., and Fu, B.: Imbalance in lake variability but not embodying driving factors on the Qinghai–Tibetan Plateau calls on heterogeneous lake management, J. Environ. Manage., 351, 119887, https://doi.org/10.1016/j.jenvman.2023.119887, 2024.

Li, C., Zhang, S., Zhang, D., and Zhou, G.: An intra-annual 30 m dataset of small lakes of the Qilian Mountains for the period 1987–2020, Sci. Data, 10, 365, https://doi.org/10.1038/s41597-023-02285-x, 2023.

Li, L., Long, D., Wang, Y., and Woolway, R. I.: Global dominance of seasonality in shaping lake-surface-extent dynamics, Nature, 642, 361–368, https://doi.org/10.1038/s41586-025-09046-3, 2025.

Li, X., Ling, F., Cai, X., Ge, Y., Li, X., Yin, Z., Shang, C., Jia, X., and Du, Y.: Mapping water bodies under cloud cover using remotely sensed optical images and a spatiotemporal dependence model, Int. J. Appl. Earth Obs. Geoinf., 103, 102470, https://doi.org/10.1016/j.jag.2021.102470, 2021.

Mullen, C., Penny, G., and Müller, M. F.: A simple cloud-filling approach for remote sensing water cover assessments, Hydrol. Earth Syst. Sci., 25, 2373–2386, https://doi.org/10.5194/hess-25-2373-2021, 2021.

Pekel, J.-F., Cottam, A., Gorelick, N., and Belward, A. S.: High-resolution mapping of global surface water and its long-term changes, Nature, 540, 418–422, https://doi.org/10.1038/nature20584, 2016.

Shadman Roodposhti, M., Aryal, J., Lucieer, A., and Bryan, B. A.: Uncertainty assessment of hyperspectral image classification: Deep learning vs. random forest, Entropy, 21, 78, https://doi.org/10.3390/e21010078, 2019.

Tesfaye, M. and Breuer, L.: Remote sensing with machine learning for multi-decadal surface water monitoring in Ethiopia, Sci. Rep., 15, 12444, https://doi.org/10.1038/s41598-025-96955-y, 2025.

Tiwari, A., Liu, Q., & Liang, L. (2026). A Monthly 30m Dataset of Lakes Larger than 1 km^2 on the Tibetan Plateau from 2001 to 2025 [Data set]. Zenodo.22107953

Wan, W., Long, D., Hong, Y., Ma, Y., Yuan, Y., Xiao, P., Duan, H., Han, Z., and Gu, X.: A lake data set for the Tibetan Plateau from the 1960s, 2005, and 2014, Sci. Data, 3, 160039, https://doi.org/10.1038/sdata.2016.39, 2016.

Wang, X., Jin, R., Wang, W., Xu, F., Geng, L., and Shao, D.: Interannual variation in lake areas over 50 km² on the Tibetan Plateau from 1986 to 2020 based on remote sensing big data, Int. J. Digit. Earth, 17, 2300308, https://doi.org/10.1080/17538947.2023.2300308, 2024.

Wen, J., He, Y., Yang, L., Wan, P., Gu, Z., and Wang, Y.: A two-step downscaling model for MODIS land surface temperature based on random forests, Atmosphere, 16, 424, https://doi.org/10.3390/atmos16040424, 2025.

Wu, G., Chen, C., Liu, Y., Fan, X., Niu, H., and Liu, Y.: Developing a high-resolution seamless surface water extent time-series over Lake Victoria by integrating MODIS and Landsat data, Remote Sens., 15, 3500, https://doi.org/10.3390/rs15143500, 2023.

Wu, J., Ke, C.-Q., Cai, Y., Wei, H.-Y., Tang, J.-P., Xiao, H.-P., and Liu, Z.: Machine learning revealed monthly change characteristics of lakes on the Tibet Plateau over the past two decades, J. Hydrol. Reg. Stud., 62, 102839, https://doi.org/10.1016/j.ejrh.2025.102839, 2025.

https://doi.org/10.5194/essd-2026-396
Preprint. Discussion started: 7 September 2026

Earth System Science Data
Open Access
Discussions

Yue, L., Li, B., Zhu, S., Yuan, Q., and Shen, H.: A fully automatic and high-accuracy surface water mapping framework on Google Earth Engine using Landsat time-series, Int. J. Digit. Earth, 16, 210–233, https://doi.org/10.1080/17538947.2023.2166606, 2023.

Zhang, G. and Duan, S.: Lakes as sentinels of climate change on the Tibetan Plateau, Geogr. Sustain., 2, 275–276, https://doi.org/10.1080/27669645.2021.2015870, 2021.

Zhang, G., Ran, Y., Wan, W., Luo, W., Chen, W., Xu, F., and Li, X.: 100 years of lake evolution over the Qinghai–Tibet Plateau, Earth Syst. Sci. Data, 13, 3951–3966, https://doi.org/10.5194/essd-13-3951-2021, 2021.

Zhang, G., Luo, W., Chen, W., and Zheng, G.: A robust but variable lake expansion on the Tibetan Plateau, Sci. Bull., 64, 1306–1309, https://doi.org/10.1016/j.scib.2019.07.018, 2019.

Zhang, G.: The lakes larger than 1 km² in Tibetan Plateau (v3.1) (1970s–2022), National Tibetan Plateau / Third Pole Environment Data Center [data set], https://doi.org/10.1016/j.scib.2019.07.018, 2019.

Zhao, G. and Gao, H.: Automatic correction of contaminated images for assessment of reservoir surface area dynamics, Geophys. Res. Lett., 45, 6092–6099, https://doi.org/10.1029/2018GL078343, 2018.

Zhao, G., Li, Y., Zhou, L., and Gao, H.: Evaporative water loss of 1.42 million global lakes, Nat. Commun., 13, 3686, https://doi.org/10.1038/s41467-022-31125-6, 2022.

Zhao, R., Fu, P., Zhou, Y., Xiao, X., Grebby, S., Zhang, G., and Dong, J.: Annual 30 m big lake maps of the Tibetan Plateau in 1991–2018, Sci. Data, 9, 164, https://doi.org/10.1038/s41597-022-01275-9, 2022.

Zhao, S., Zhao, X., Zhao, J., Zhang, X., Liu, X., and Yao, C.: TPLake-MED: A monthly extent dataset for lakes on the Tibetan Plateau, Earth Syst. Sci. Data Discuss. [preprint], https://doi.org/10.5194/essd-2025-649, in review, 2025.

Zheng, C., Jia, L., Hu, G., Menenti, M., and Timmermans, J.: How much water vapour does the Tibetan Plateau release into the atmosphere? Hydrol. Earth Syst. Sci., 29, 485–506, https://doi.org/10.5194/hess-29-485-2025, 2025.

Zhou, Y., Liu, B., Cui, Y., Wang, X., Cao, M., Zhang, S., Xiao, X., and Dong, J.: Annual improved maps to understand the complete evolution of 9 thousand lakes on the Tibetan Plateau in 1991–2023, ISPRS J. Photogramm. Remote Sens., 217, 134–148, https://doi.org/10.1016/j.isprsjprs.2024.08.012, 2024.