A high-resolution inland surface water body dataset for the tundra and boreal forests of North America

Yijie Sui[1], Min Feng*[1,2,3] , Chunling Wang[1,3], Xin Li [1,2,3]

[1]National Tibetan Plateau Data Center, Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing 100101, China

[2]CAS Center for Excellence in Tibetan Plateau Earth Sciences, Chinese Academy of Sciences, Beijing 100101, China

[3]University of Chinese Academy Sciences, Beijing 100049, China

Correspondence to: Feng M. (mfeng@itpcas.ac.cn)

Abstract. Inland surface waters are abundant in the tundra and boreal forests ~~in~~ of North America, essential to environments and human societies but vulnerable to climate changes. These high-latitude water bodies differ greatly in their morphological and topological characteristics related to the formation, type, and vulnerability. In this paper, we present a~~n~~ ~~inland surface~~ water body ~~inventory (SWBI)~~ dataset for the ~~tundra and boreal forests of~~ North American high latitudes (WBD-NAHL). Nearly 6.~~7~~ 5 million water bodies were identified, with approximately 6 million (~90%) of them smaller than 0.1 km^2. The dataset provides ~~geometry~~ area~~coverage~~ and morphological attributes for every water body. During this study, we developed an automated approach for detecting surface water extent and identifying water bodies in the 10-m resolution Sentinel-2 multispectral satellite data to enhance the capability for delineating small water bodies and their morphological attributes. The approach was applied to the Sentinel-2 data acquired in 2019 to produce the water body dataset for the entire tundra and boreal forests in North America. The dataset~~,~~ provided~~ing~~ a more complete representation of the region than existing regional datasets ~~in~~for North America, e.g., Permafrost Region Pond and Lake (PeRL). The total accuracy of the detected water extent by the ~~SWBI~~WBD-NAHL dataset was 96.36% by comparing to interpreted data for locations randomly sampled across the region. Compared to the 30-m or coarser resolution water datasets, e.g., JRC GSW yearly water history, HydroLakes, and Global Lakes and Wetlands Database (GLWD), the ~~SWBI~~WBD-NAHL provided an improved ability on delineating water bodies, and reported higher accuracies in the size, number, and perimeter attributes of water body by comparing to PeRL and interpreted regional dataset. This dataset is available ~~on~~ from the National Tibetan Plateau/Third Pole Environment Data Center (TPDC, http://data.tpdc.ac.cn): DOI: 10.11888/Hydro.tpdc.271021 (Feng et al., 2020).

1 Introduction

Inland surface waters include various types of water bodies, including rivers and streams; large and small lakes; reservoirs; and ephemeral ponds. Inland surface water occupies only 2% of the global land surface (Pekel et al., 2016), but it plays a critical role in terrestrial ecosystems. Surface water distribution varies across the landscape. More than 55% of global surface waters are located in high latitudes in the Northern Hemisphere (> 44°N), and these northern high-latitude waters are generally small and densely clustered. The high latitudes have warmed faster than other regions, with annual surface temperatures increasing > 1.4° C over the past century (IPCC 2014). The temperature of the Arctic, in particular, has risen twice as fast as the average global temperature (Graversen et al., 2008; Johannessen et al., 2004; Pachauri and Reisinger, 2007; Serreze and Francis, 2006; Li et al., 2020). This change in climate is driving changes in terrestrial ecosystems in the Arctic as well. For example, increases in vegetation productivity have been observed across the northern high latitudes (Forkel et al., 2016).

Meanwhile, high-latitude water bodies have started changing since the early 1970s (Carroll et al., 2011; Carroll and Loboda, 2017; Cooley et al., 2019; Smith et al., 2005; Fayne et al., 2020; Nitze et al., 2020). Although some changes are seasonal, and therefore temporary, permanent changes have been reported, and small lakes in permafrost regions are found to be more vulnerable to permanent changes in water extent (Carroll and Loboda, 2017; Karlsson et al., 2014).

With observed rising temperatures ~~As rising temperatures have been reported in permafrost~~ (Biskaborn et al., 2019), ~~its~~ permafrost thawing poses a threat to the stability of inland surface waters~~, especially in the high latitudes~~, especially in arctic lowland surface areas, where most of the ~~lakes~~water bodies could be thermokarst lakes (Jones et al., 2011; Olefeldt et al., 2016)~~where half~~large amount ~~of the lakes are thermokarst lakes~~distributed and have strong interactions with permafrost in the regions. Thawing permafrost not only leads to the formation of lakes and ponds of various sizes, but also leads to the release of organic carbon in the form of carbon dioxide (CO_2) and methane (CH_4) (Serikova et al., 2019). Changes in ~~thermokarst~~lake formation may result in concomitant changes to the extent and connectivity of surface water bodies, which can greatly impact the sustainability of aquatic ecosystems.

The ~~shapes~~morphology of the water bodies could ~~correlate~~be shaped by ~~to~~ the surrounding environment~~.The shapes of the water bodies correlate to the regulation of surrounding environment~~ (Grosse et al., 2013; Laird et al., 2003; Schilder et al., 2013; Sharma et al., 2019; Carpenter, 1983; Higgins et al., 2021). ~~The shapes of the water bodies correlate to suitability of surrounding ecosystems(Grosse et al., 2013; Laird et al., 2003; Schilder et al., 2013; Sharma et al., 2019; Carpenter, 1983; Higgins et al., 2021).~~ Shoreline complexity affects lake ice formation (Sharma et al., 2019). Lake connectivity affects fish migration (Laske et al., 2019; McCullough et al., 2019), fish habitats, and aquatic assemblages (Napiórkowski et al., 2019; Jiang et al., 2021)~~; improves~~Lake connectivity impacts, water self-purification and accelerates water cycling (~~Gli~~Glińska ~~GliÅ„ska~~- Lewczuk, 2009; Vaideliene & Michailov, 2008; Xiong et al., 2017). ~~Water density~~The d~~D~~ensity of water bodies impacts fish density and biomass (Sandlund et al., 2016; van Zyll de Jong et al., 2017; King et al., 2021). The shape and distribution of water bodies reflect ~~the reasons~~what led to the water body ~~formed~~ formation (Laurence C. Smith et al., 2007~~; Grosse et al., 2013; Laird et al., 2003; Schilder et al., 2013; Sharma et al., 2019; Carpenter, 1983; Higgins et al., 2021).~~~~).~~ Furthermore, information about lake area extent can improve arctic land surface modeling (Langer et al., 2016; van Huissteden et al., 2011). For these reasons, it is critical to ~~discern~~ quantify ~~the~~ high~~-~~-latitude surface water extent, as well as characterize related morphological and topological features, including size and shape.

In the past, inland surface water was mapped at sub-hectare (i.e., 30 ~~-~~m) resolution using satellite data (Feng et al., 2015; Pekel et al., 2016; Pickens et al., 2020), and these data provided unprecedented information about the global extent of inland waters ~~in the global extent~~, including their spatial distribution and temporal changes ~~of inland waters~~. These datasets provide data that delineates the extent of large and moderate sizes of water bodies but underrepresent or fail to include the large number of small water bodies. Coarse-resolution datasets also lead to underrepresentation in delineating complex shorelines and the shapes of surface water bodies, making it difficult to derive their morphological and topological attributes. Existing datasets containing information that describe water body shapes, such as the Global Lakes and Wetlands Database (GLWD) (Lehner and Döll, 2004) and HydroLAKES (Messager et al., 2016) are limited to water bodies larger than 0.1 km^2. In spite of these limitations, these datasets provide valuable information for improving the precision of mapping inland waters. Detecting the extent of inland surface water at finer spatial scale boosts our ability ~~for~~ to ~~mapping~~ map ~~the~~ small water bodies~~s~~ and improves the precision o~~f~~n delineating the shorelines of water bodies. This analysis then allows us to derive an inventory dataset of water bodies along with their morphological and topological attributes. The information allows scientists to analyze a water body as an object instead of a cluster of pixels, advancing our analysis and understanding of the water bodies' size, shoreline complexity, ecological effects, hydrological function, and vulnerability to natural and anthropogenic changes.

In this paper ,we present a higher resolution inland surface water body inventory (SWBI)dataset for the tundra and boreal forests of North American; high latitudes (WBD-NAHL). It The dataset was derived from by identifying the extent of inland waters using 10-m resolution Sentinel-2 multispectral data. The dataset provides the spatial extent and morphological attributes for each identified water body. It is the first inland water inventory dataset derived at this landscape scale with the capability of delineating inland surface waters as small as 0.001 km^2.

2 Spatial extent

The SWBIWBD-NAHL dataset covers all tundra and boreal forest biomes in North America (Figure 1), with the exception of the Arctic Archipelago and Baffin Island due to their long time of snow or ice covering over water bodies. The topography of the tundra and boreal forest in North America is extremely diverse, varying from mountains and rolling hills to plateaus and flat coastal plains. The mountains of the North American Cordillera are covered by numerous mountain glaciers, where also and also distributed a large number of glacieral lakes. A large number of thermokarst lakes were found in lLowlands in tundra areas, e.g., the Yukon dDelta and the Alaska Nnorth Sslope (Olefeldt et al., 2016). The vast Canadian Shield also consistshas ofa high density of lakes.The western mountains of North American Cordillera are covered by numerous mountain glaciers, where large amount of glacier lakes distributed. Lowlands in tundra, including the Yukon delta, the Alaska north slope and so on are mainly distributed by thermokarst lakes (Olefeldt et al., 2016). The vast eastern plateaus belong to the stable Canadian Shield, where lakes dominate. The eastern mountains of the Canadian Codillera are covered by numerous mountain glaciers and divide the region into east coastal plains and west plateaus. The long and narrow eastern coastal plain of this cordillera located near the Pacific Ocean is dominated by thermokarst landform and glacier lakes. The vast western plateaus belong to the stable Canadian Shield and are the result of glacial erosion. The climate of this study region is characterized by long, cold winters and short, cool summers. The summer season typically lasts from June to September. The plants in the northern tundra include lichen, moss, grass, sedge, and shrub. The southern boreal forest is dominated by evergreen forests (Ritter, 2006). Lakes are widely distributed in the study region and approximately 36% of the land surface is covered by water Lakes widely distributed in the study region with approximately 36% of land surface is covered with lakesLakes and ponds dominatewidely distributed in the landscapewhole study region and approximately 36% of land surface is covered with lakes. There are about 50% of the lakes and 30% of lakes by area in the total region (Messager et al., 2016counted by HydroLAKES). The distribution of lakes in this region is largely controlled by the presence of permafrost as well as glacial history (Mostakhov, 1973; Smith et al., 2007). The number of lakes in this region accounts for 50% of the global lakes and ponds , and the area of lakes accounts for 30% of the global lakes in the whole region, indicating the region to be one of the richest areas of surface water bodies (Messager et al., 2016counted by HydroLAKES). Various types of lakes, including organic lakes, fluvial lakes, meteorite lakes, volcanogenic lakes, and anthropogenic lakes, are distribute in the study region and featured with very different sizes and shapes. Among them, tThe water bodies formed by glacial erosion are abundant in the westerneastern wide flat Canadian Shield., where the shapes of water bodies usually are thin and complex. TheThe coastal lowlands are mainly consist of water bodies with circular shapes which were likely formed by the freezing and thawing effectThe coastal lowlands mainly are distributed by the nearly circular water bodies distributed on the east and north coast of which are mainly formed by freezing and thawing (Dranga et al., 2017).

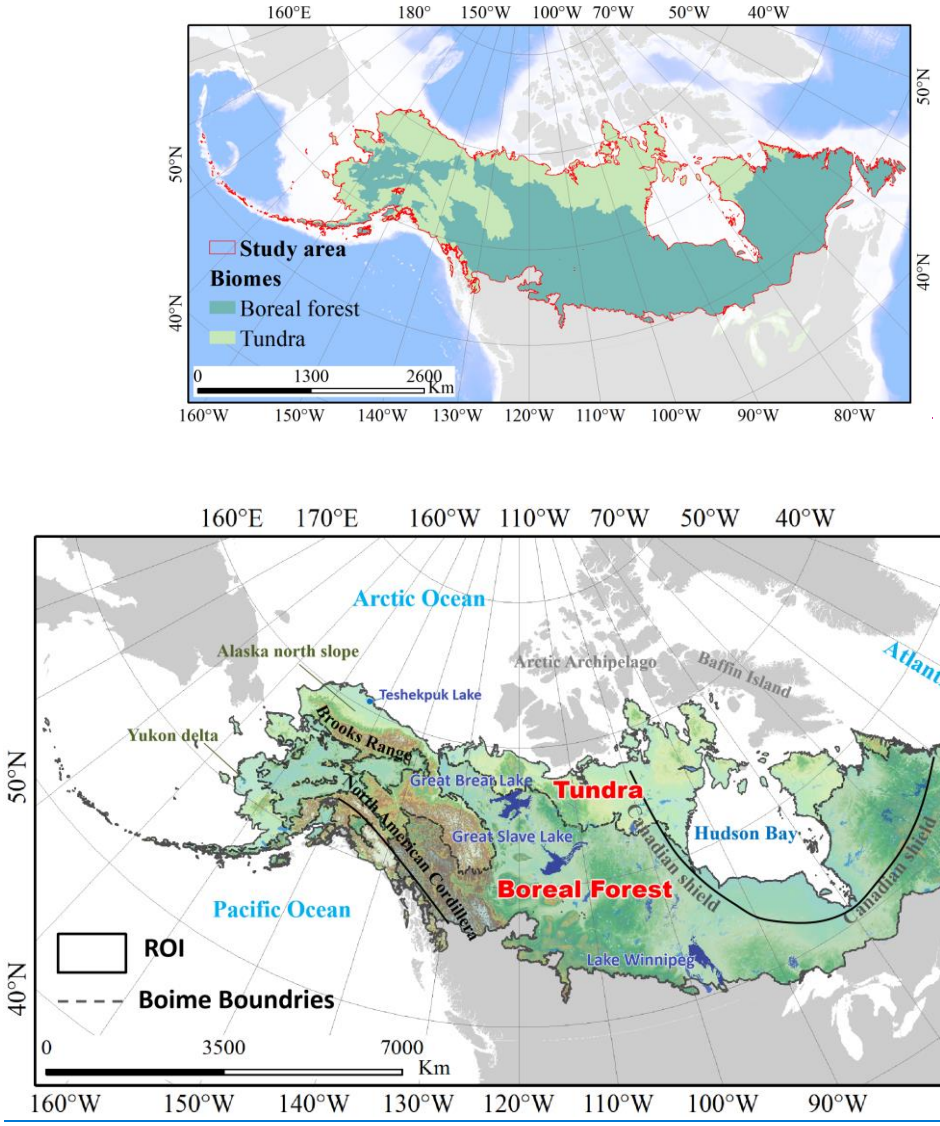


Figure 1: The extent of the study area, including the tundra and boreal biomes, in the North Americas continent, excluding the Arctic Archipelago and Baffin Island.

3 Data

3.1 Sentinel-2 A/B multi-spectral images

Sentinel-2 multi-spectral images were used to delineate surface water bodies in this study. ~~The~~ Sentinel-2 A/B provides a short revisit cycle (2-3 days) in the high latitudes, which is critical for detecting surface water during the short, snow-free season in the region. Sentinel-2 images were obtained using the United States Geological Survey (USGS) EarthExplorer client/server interface (https://earthexplorer.usgs.gov/, last access: 7 April 2021).

Each Sentinel-2 image consists of ~~12~~ 13 multispectral bands, including four bands at 10-m resolution, six bands at 20-m resolution, and ~~eight~~ three others at ~~20~~60-m resolution. Sentinel-2 data are distributed as collections representing different processing levels. We selected the Sentinel-2 Collection 2 data, which provides spectral bands of surface reflectance after atmospheric corrections. The 10-m Sentinel-2 bands were used for water detection to maximize spatial precision for delineating small water bodies. The 20-m Sentinel-2 bands were resampled to 10-m resolution to match the higher resolution bands~~capture the spectral properties of water bodies as much as possible~~. The "s2cloudless" (https://github.com/sentinel-hub/sentinel2-cloud-detector, last access: 7 April 2021) was applied to identify cloud-contaminated pixels, generating a probability of cloud and cirrus detection. This module includes a model generated by a Convolutional Neural Networks (CNN) trained with 6.4

million manually labeled samples. This model was validated to have 99% accuracy for identifying clouds and 84% accuracy for identifying cirrus in Sentinel-2 images (Zupanc, 2020).

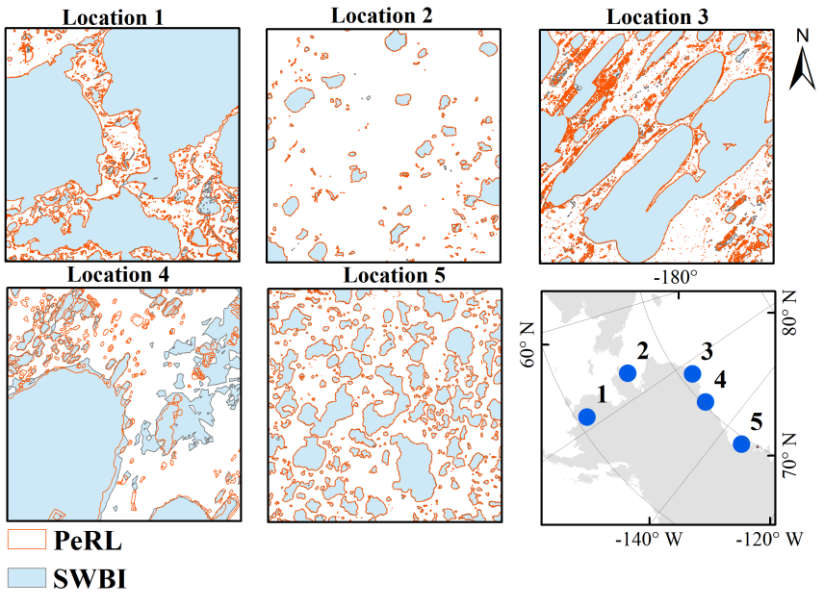
3.2 Joint Research Centre (JRC) ~~JRC (Joint Research Centre)~~ yearly water dataset

The JRC yearly water dataset (JRC GSW Yearly Water Classification History, v1.2, https://global-surface-water.appspot.com/) (Pekel et al., 2016) provides a delineation of permanent water, non-water, and seasonal water for global inland surface waters. The dataset was produced using long-term Landsat images, including Landsat TM, ETM+, and OLI images acquired from 1984 to 2019. Permanent water in the dataset was identified as water cover throughout the entire year, and seasonal water is identified based on occurrence during a single year.

The JRC yearly water dataset provides a reasonably accurate delineation of water distribution for the period 1984-2019, but its precision is limited by the 30-m spatial resolution of Landsat data. The dataset's accuracy at high latitudes is affected by the relatively poor return cycle of Landsat (16 days), cloudiness, and long periods of snow and ice in the region each year. The JRC dataset was used as a reference to overcome these limitations and improve our ability to identify and monitor inland surface water bodies, particularly small water bodies. The permanent water class in the JRC dataset was used in this analysis, while the seasonal water was excluded due to its reportedly low accuracy (Meyer et al., 2020). The maximum extent of permanent water bodies for the time period 1984-2019 were processed to fill gaps in individual years, which were then used as the reference in this study.

3.3 Permafrost Region Pond and Lake (PeRL)

The Permafrost Region Pond and Lake (PeRL) dataset was produced through a circum-Arctic effort to map ponds and lakes from modern (2002–2013) high-resolution aerial and satellite imagery with a resolution of 5 ~~-~~m or finer, including imagery from GeoEye, QuickBird, WorldView-1/2, the KOMPSAT-2, and TerraSAR-X. The PeRL dataset includes 69 small maps representing a wide range of environmental conditions in tundra and boreal biomes~~.~~ (Muster et al., 2017). There are 14 maps mainly distributed in five regions of North America~~.~~ (Figure 2). Because of the high-resolution data, the PeRL dataset is able to delineate water bodies as small as 10^{-7} km^2, which is valuable for validating satellite-derived water datasets for regions dominated by small water bodies.



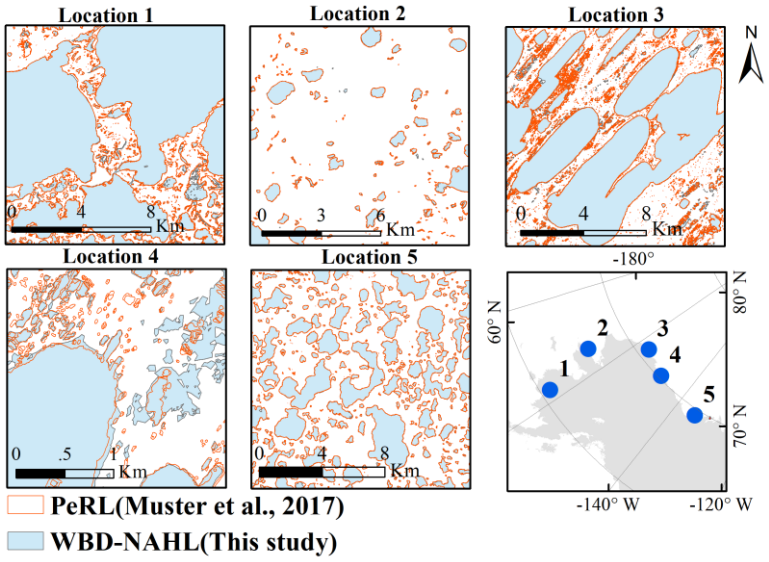


Figure 2: Water bodies identified in the ~~SWBI~~WBD-NAHL (~~T~~this study) and PeRL datasets~~,~~ (Muster et al., 2017), and the locations (blue dots) of the PeRL maps for the study region.

4 Methods

The 10-m resolution Sentinel-2 A/B multispectral data are the primary source used to identify small water bodies. An approach was developed to produce a~~n~~ water probability layer for 2019 by combining the water--sensitive indexes derived from the Sentinel-2 bands and the 30-m resolution JRC water dataset (section 4.1).~~,~~ ~~and a~~A~~n~~ machine learning model was trained to retrieve water extent from the Sentinel-2 images ~~in~~from possible water extent restricted by the water probability layer~~Machine learning models were built to detect surface water pixels in each Sentinel-2 image. The results were combined to produce a final 10-m resolution dataset of water extent for 2019~~ (~~see~~ section 4.2~~1~~) (Figure 3). Water bodies were finally identified from the ~~detected~~ water extent using an object-based algorithm to produce the final water body inventory (~~see~~ section 4.3~~2~~).

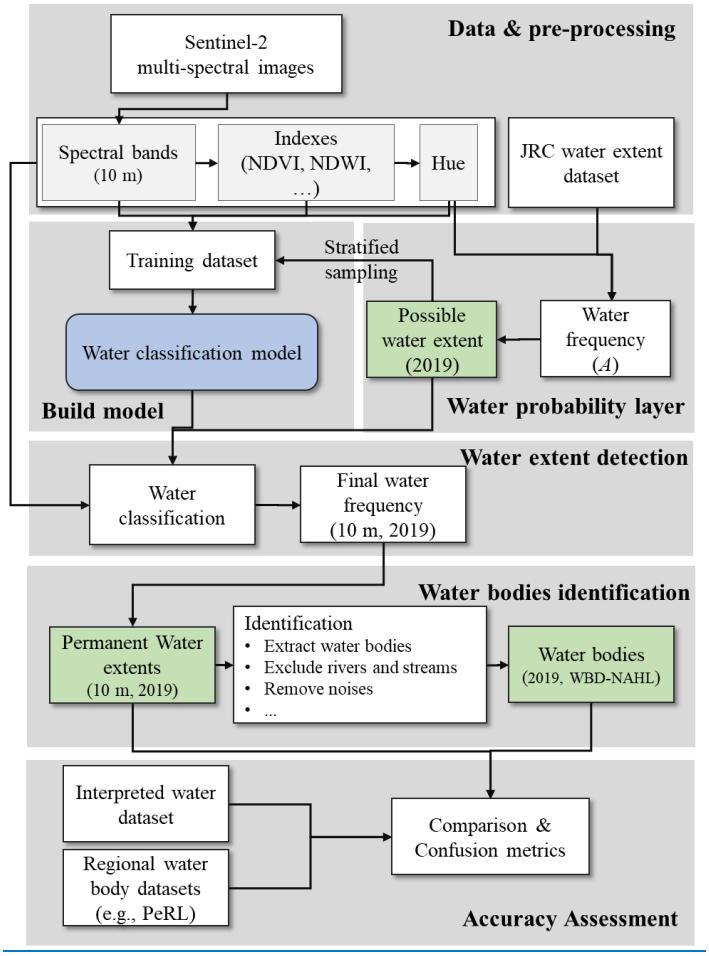


Figure 3: ~~The f~~Flowchart ~~of~~**or** processing water extent and identifying water bodies.

4.1 ~~Detect~~ W~~w~~ater probability layer~~extent~~

A water probability layer was derived to represent the likelihood ~~to~~of a pixel to correspond to ~~be~~ permanent water ~~in~~ during the summer of 2019. The 10-m resolution water-sensitive indexes calculated from the Sentinel-2 multispectral bands were used as the main input.~~, and~~ The other reference water dataset (e.g., the JRC water dataset) was adopted as a supplemental input and fused with the main input to produce the water probability estimate at each 10-m resolution pixel.

~~To enhance the information of water, a two-step water detection method was applied for the generation of water extent map (Figure 3). In the first step, a fusion of Sentinel-2 and JRC product was used to generate a possible water extent map. In the second step, the machine learning models were built to refine the possible water extent map.~~

To reduce effects ~~from~~ of snow cover, Sentinel-2 A/B images acquired between June and September 2019 were selected to represent the relatively snow-free season in North American tundra and boreal biomes. The pixels in each Sentinel-2 image with an estimated cloud probability higher than 65% were excluded to avoid the effects of cloud contamination.

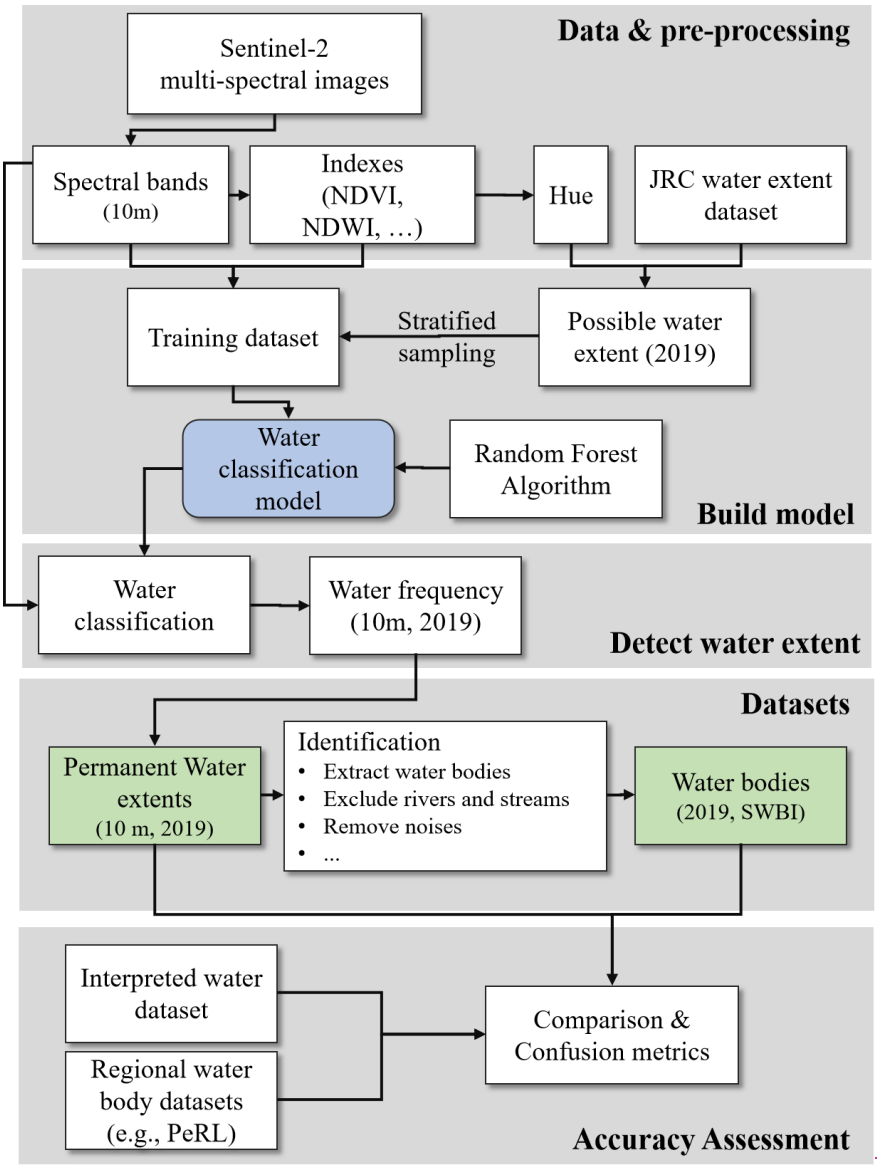
During pre-processing, multiple water-sensitive indexes were derived from each Sentinel-2 image to enhance the ability ~~on~~ to detect~~ing~~ water (Figure 3). To maximize the ability to separate water from non-water, especially vegetated land, three indexes were calculated to represent water and vegetation in each image: Normalized-Difference Water Index (NDWI) (~~Xu, 2006~~McFeeters, 1996), Normalized Difference Vegetation Index (NDVI) (Carlson and Ripley, 1997), and Modified Normalized-Difference Water Index (MNDWI) (~~McFeeters, 1996~~Xu, 2006). The three indexes were calculated as follows.

$$NDWI = (-(B_{green} - B_{nir})-)/(B_{green} + B_{nir}), \quad (1)$$

$$NDVI = (B_{nir} - B_{red})/(B_{nir} + B_{red}), \quad (2)$$

$$MNDWI = (B_{greenvir} - B_{swir})/(B_{greenvir} + B_{swir}), \quad (3)$$

Where B_{green}, B_{red}, B_{nir}, and B_{swir} are ~~the band of~~ green (band #3), red (band #4), near-infrared (band #8), and short-ware infrared (band #11), respectively. These bands ~~are~~have 10-m resolution except B_{swir}, which ~~is~~has 20-m resolution and was pan-sharpened using the À Trous Wavelet Transform (ATWT) algorithm as recommended by ~~(~~Du et al., (2016)~~xxxx.~~ An HSV color space conversion was used to combine the three indexes and produce a final index for identifying water. The HSV (hue-saturation-value) color space conversion is a non-trigonometric pair of transformations from a linear red-green-blue (RGB) color space to a perceived color space (Danielson and Gesch, 2011). This method converts the three input bands into hue (color), saturation, and value components. The three indexes (NDWI, MNDWI, and NDVI) were scaled by 255, converted to a byte value type, combined ~~into~~ into the RGB color space, and then converted to the HSV color space to derive a comprehensive index for identifying water.



Once the hue has been identified, an experimental threshold of < 0.45 was applied to ~~separate~~ identify the water pixels ~~from others~~. The same procedure was applied to derive temporal water extents from all selected Sentinel-2 images ~~to derive temporal water extents.~~, ~~which were~~All the water extents were then combined to calculate the water frequency (A_s) for the year. Potential water extent was then derived from the calculated water frequency data. The existing JRC water dataset provided complementary information for estimating possible water extent. The JRC permanent water records were resampled to 10-m resolution using the nearest neighbor ~~, and~~algorithm and combined with the Sentinel-2-derived water frequency dataset using a weighted linear combination~~:. A higher weight was assigned to the JRC in high elevations to reduce the effect of snow and ice on the Sentinel-2 derived hue over mountains.~~

$$A = W_s \cdot A_s + (1 - W_s) \cdot A_j, \quad (4)$$

where, A is the updated water frequency, W_s is the weight for the Sentinel-2-derived water frequency (A_s) and ~~was~~ set to 0.85~~, a higher weight~~ to ensure that thee 10-m measurements ~~to be~~were the main input for the final water probability estimate. However, W_s was ~~slightly~~ decreased to 0.65 in high elevations pixels (elevation > 1 km) to reduce the effect of snow and ice on the Sentinel-2-derived hue over mountains. ~~for locations with elevation < 1 km and 0.65 for higher elevations.~~ A_j is the JRC permanent water record, which was set to 1.0 for permanent water and to 0.0 for others. The final, combined ~~potential water extent~~ ~~(called~~ possible water extent ~~in the following)~~ was identified when $A > 0.5$.

4.2 Water extent detection

Although the possible water extent estimated the ~~likeliness~~likelihood of a pixel to correspond to ~~be~~ water, confusion with ~~it could still confuse with others (e.g.,~~ shadow, ice, ~~and~~or cloud contamination~~)~~ in area with complex environments is still possible due to the limitations of ~~the~~ water indexes ~~on~~with ~~synonyms~~similar ~~spectrum~~spectra ~~(~~(Isikdogan et al., 2017)~~).~~ A ~~machine learning~~random forest model was ~~built~~trained ~~from~~with points collected through visual interpretations to further detect water within the areas indicated as possible water.

~~Points were randomly selected across the possible water extent and then visually interpreted to~~ To provideduce ~~the training data for building a water body identification detection machine learning model, individual points were collected from the identified possible water extents.~~ ~~T~~To ~~enhance~~ensure the representation of water and other ~~the model's ability to separate water from oth~~land covers that~~– are~~can easily be confused ~~by~~as water ~~bodies have high chance er land cover types in the region,~~– ~~the potential water extent was divided into~~ five strata were introduced, i.e., ~~representing~~ water, glacier, mountain, vegetation, and cloud. ~~At this time~~Then, 250 points were randomly selected in each stratum, ~~and~~ for a total of 1,250 points ~~were collected.~~ ~~as the secondary source of training data besides the reference collected from the possible water extent~~ (Figure 4a~~)~~). ~~To enhance the model's ability to separate water from other land cover types in the region, the potential water extent was divided into five strata representing water, glacier, mountain, vegetation, and cloud.~~

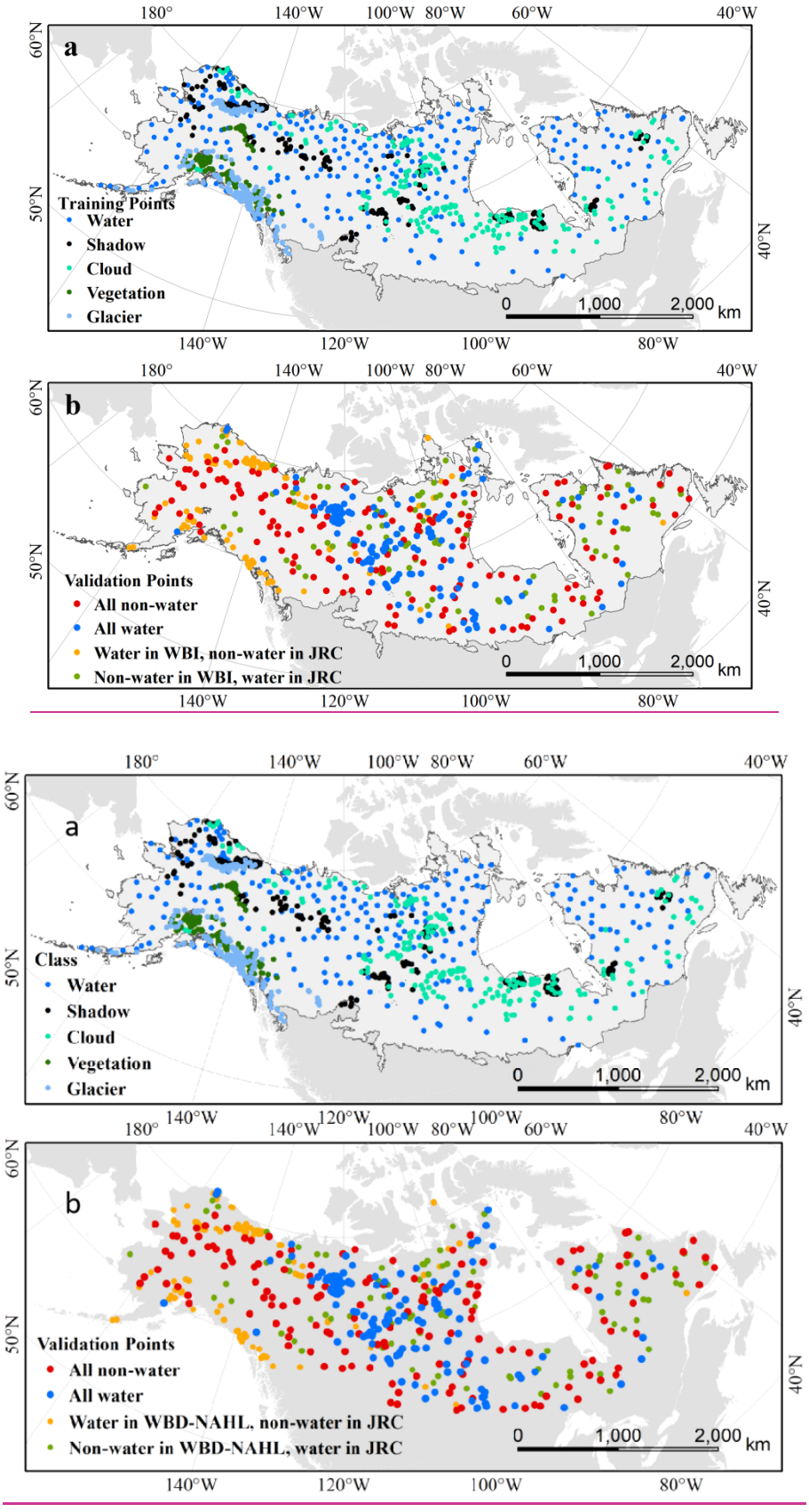


Figure 4: ~~The t~~Training samples for random forest model building (a) and points identified for validating the accuracy of the detected water extent (b).

The five strata were established using reference datasets or customized rules. ~~The water stratum was identified as possible water extent.~~ The glacier stratum was identified using the Global Land Ice Measurements from Space (GLIMS) dataset of 2017 (http://www.glims.org/, last access: 7 April 2021), which was a dataset of global glacier outlines including glacier area,

geometry, surface velocity, and snow line elevation and was produced from the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) and the Landsat Enhanced Thematic Mapper Plus (ETM+), as well as historical information derived from maps and aerial photographs. Vegetation was identified as areas with a positive mean NDVI value calculated from the June-September Sentinel-2 images. The cloud stratum was identified as having at least 20% of mean cloud probability calculated from the selected Sentinel-2 images. The mountain shadow stratum was identified as any elevation higher than 1 km and slope greater than or equal to 3°. The water stratum was identified as the remainingst area of possible water extent.

The selected points were interpreted by the team to provide training data. Although we only used Sentinel-2 images during from June to September 2019, points were matched with a randomly selected image at the location during the time period, providing representation for possible temporal variation. Each point was visually labeled by an interpreter after examining the image. Metrics for visible bands (red, green, and blue), NDWI, MDWI, NDVI, and hue were derived from each image to provide attributes for the point. These attributes were pooled to produce training data for building the machine learning model.

The scikit-learn Random Forest algorithm (Breiman, 2001) was adopted to build the model for surface water identificationdetection. This model was applied to the selected Sentinel-2 images to detect surface water pixels. The results were compiled temporally to produce a water frequency layer (*f*).

In this study, terrain shadows in the water frequency layer were removed with a terrain mask derived from the Global Multi-resolution Terrain Elevation Data (GMTED) (Danielson and Gesch, 2011). The mask was where the slope was greater than or equal to 7° and the elevation was over 1500 m1 km. The elevation threshold was used to minimize the impact of the slope threshold on rivers in lowlands. The method using slope to identify terrain shadows was verified to be more effective than using hill-shade (Carroll and Loboda, 2017).

4.3 Water bodies identification

Permanent water pixels were identified from the resulting water frequency layer (*f*) as being those pixels with at least 50% occurrence between June and September. The resulting water pixels were then converted to vector polygons using the "Raster to Polygon" tool in ESRI ArcMap 10.2. These water polygons provided the preliminary surface water body records.

An array of geometry metrics was calculated for each water body polygon using ArcMap in the Canada_Lambert_Conformal_Conic projection (datum D_North_American_1983 and Spheroid GRS80). These metrics include area, perimeter, and a shape index (*SI*), which estimates the complexity of a water body polygon. The *SI* was calculated as:

$$SI = P_{wateri}/P_{circlei}, \quad (5)$$

where P_{wateri} is the perimeter of the water body iI, $P_{circlei}$ is the perimeter of a circle that has the same area as water body i. *SI* equals 1 when a polygon is a perfect circle and greater than 1 when the polygon has a complex irregular shape.

At this point, tThe derived water body morphological metrics (i.e., the *SI* and area) and the HydroRIVERS were used to identify rivers and streams in the WBD-NAHL water bodies. Rivers and streams intend to have long, narrow, and linear shapes., wWe initially applied area thresholds of area > 5 km^2 and *SI* > 10 in combination with visual examination to exclude large rivers and streams in the WBD-NAHL. Considering the extreme difficulties onin distinguishing small rivers and streams, a more aggressive method was applied to further identify water bodies that could possibly be rivers and streams were further identified by selecting Then, long and linear (*SI* > 3) water bodies (*SI* > 3) located closed to (< 100 m) to tthe rivers and streams (< 100 m), as indicated by HydroRIVERS were identified as possible to be rivers and streams. Those water bodies (168,983) were marked in the attribute table (field "river")At this point, HydroRIVERS data and metrics (i.e., the *SI* and area) were used to

~~distinguish rivers and streams from lakes and ponds. Rivers and streams have long and linear feature, we initially applied thresholds of area > 5 km2 and SI > 10 to preliminarily separate them from lake and ponds. Then, these long and linear water bodies closed to the river lines of HydroRIVERS (within the 1km buffer of river lines) were detected as rivers and streams (marked as 1 in the field "river"). At last, labeled polygons were visually checked to confirm and correct misclassified water bodie~~At this point, ~~HydroRIVERS data and metrics (i.e., the SI and area metrics)~~ were used to distinguish rivers and streams ~~from lakes and ponds. Rivers and streams have long and linear feature, and we initially applied thresholds of area > 5 km^2 and SI > 10 to preliminarily separate them from lake and ponds. Then~~Then, when these long and linear water bodies closed to the ~~river lines of HydroRIVERS (within the 1km buffer of river lines), we detected them as rivers and streams (marked as 1 in the field "river"). At last, labeled polygons were visually checked to confirm and correct misclassified water bodies.~~

4.~~2~~ 4 Quality assessment

The accuracy and uncertainty of ~~the SWBI~~WBD-NAHL were assessed at two levels, i.e., pixel water extent and derived water bodies, to provide a comprehensive evaluation of the dataset. We randomly selected eight square blocks with a size of 10 km by 10 km in the North American tundra and boreal region (Figure 5). The selected blocks were visually interpreted by the team to identify all the water bodies within each using a high-resolution Google Earth image as reference for interpretation. Water bodies records from the PeRL were compared to the ~~SWBI~~WBD-NAHL water bodies to assess the number of water bodies and spatial area of each. The interpreted dataset was also compared to the JRC-derived water body records for 2019 to assess its accuracy in terms of representing water bodies. The JRC dataset provides water/nonwater~~y situation~~ map ~~for~~ at the 30-m resolution pixels, representing the distribution of water extent, but no information ~~of~~ in the spatial relationship between pixels and water bodies were provided, and we derived water bodies records from the JRC dataset using the same algorithm described in section 4.1.

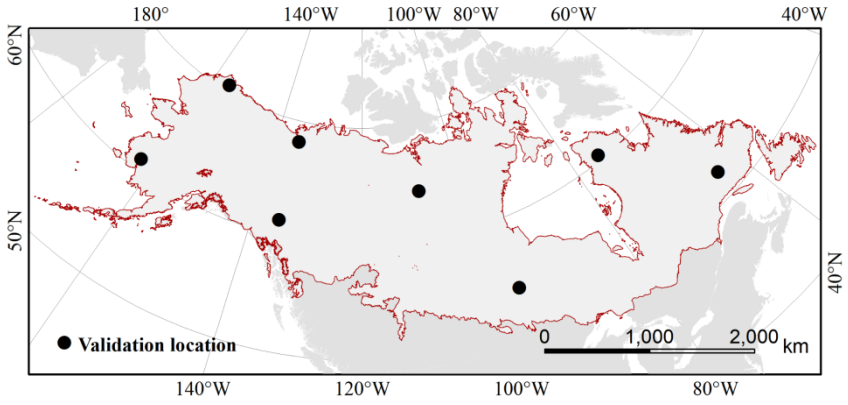


Figure 5: Locations of the ~~five~~5 regions selected and interpreted for assessing the accuracy of the indicators of water bodies.

The 14 regional PeRL maps were compared to the ~~SWBI~~WBD-NAHL water bodies. Although the PeRL maps were produced from high-resolution images acquired in 2002-2013, the maps show little temporal changes when comparing to the ~~SWBI~~WBD-NAHL dataset in the extents of the maps (Figure 2), and these maps were adopted as references for evaluating the ~~SWBI~~WBD-NAHL water bodies. The PeRL maps were produced from images with 5 m resolution or finer, we excluded all water bodies in PeRL smaller than 0.0003 km^2 to ensure comparability to the scale of the ~~SWBI~~WBD-NAHL dataset.

The water extent derived from the Sentinel-2 images were assessed by manually comparing specific points between the ~~SWBI~~WBD-NAHL dataset and the JRC surface water dataset. The points were collected using a stratified random sampling across the entire study region. To achieve higher sampling performance, the outcomes were divided into four strata that represent pixels that were agreed as water, disagreed as water, agreed as non-water, and disagreed as non-water. In each of the strata, 400 points were randomly selected from the dataset and manually assessed by examining the same point in the latest

Google Earth image. (Figure 4b) The results from the 1600 points were compared to the derived water extent. The confusion matrix was calculated from the results.

The sampling weights were included in the calculation of the metrics as following:

$$W_s = A_s / A_{all}, \quad (6)$$

where A_s is the area of stratum s, and A_{all} is the total area of the region.

Equations of the confusion metrics with weights:

$$OA = \sum_s^4 W_s * OA_s, \quad (7)$$

$$UA = \sum_s^4 W_s * UA_s, \quad (8)$$

$$PA = \sum_s^4 W_s * PA_s, \quad (9)$$

where OA, UA, and PA are the overall accuracy, user's accuracy and producer's accuracy of the entire dataset, OA_s, UA_s and PA_s are the concomitant accuracies in stratum s, and W_s is the sampling weight of ~~stratums~~strata.

5 Results

5.1 Water bodies in tundra and boreal forests of North America

More than 6.47 million (6,474,051)~~6.65 million (6,652,015)~~ surface water bodies were identified in the tundra and boreal forests of North America, while 90.4~~4~~3% of these water bodies~~–~~ (5,844,921)~~(6,015,484)~~ were smaller than 0.1 km^2. Those water bodies covered more than 0.8 million km^2, ~10.3% of the study area (Figure 6). The average size and perimeter of the identified water bodies were 0.12~~12~~ km^2 and ~~1.01~~1.01 km, respectively, and their average *SI* was 1.41~~2~~.

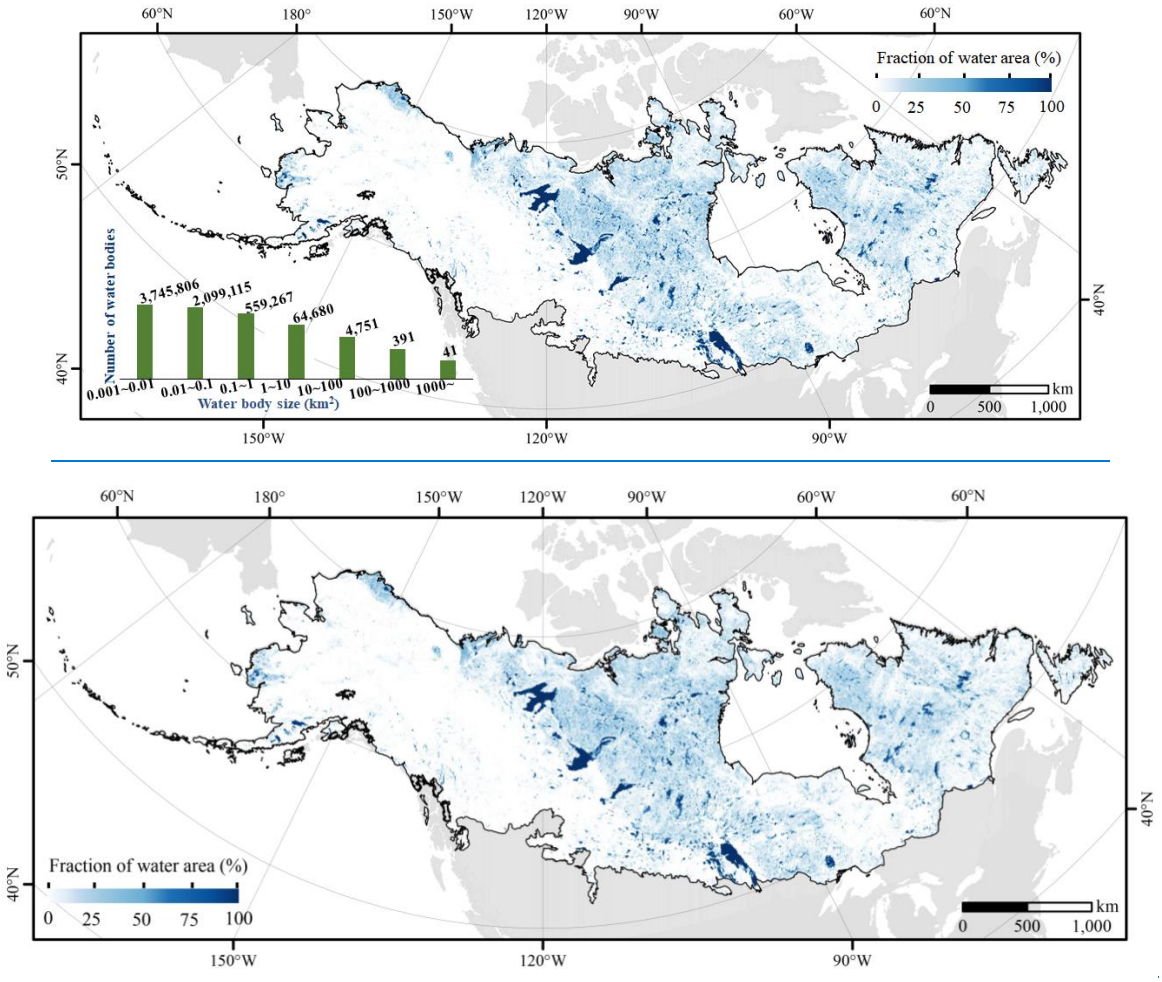
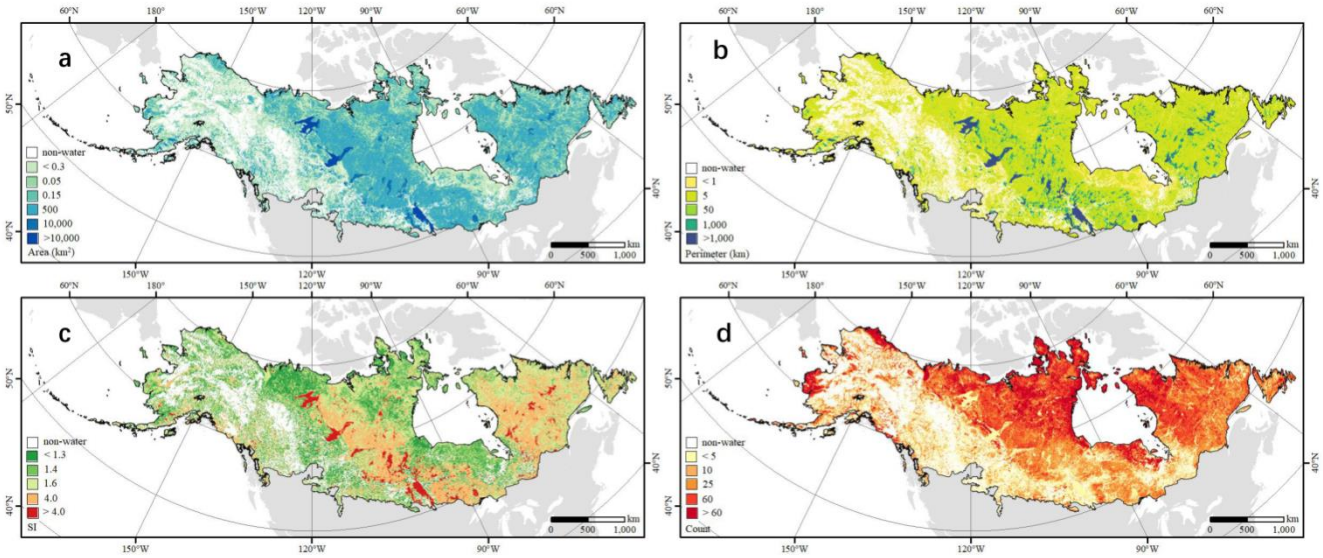


Figure 6: Percent of surface water (5 km × 5km grid) produced by aggregating the water extent for the tundra and boreal forests of North America as calculated using the ~~SWBI~~WBD-NAHL dataset.

All of the morphological indicators, including area, perimeter, and *SI*, of the identified water bodies showed great heterogeneity across the region (Figure 7). In general, the tundra biome consist~~sed~~ of ~~was dominated by~~a large number of densely packed small water bodies with regular shapes ~~formed by melting frozen ground (Grosse et al., 2013)~~. In contrast, the boreal forest biome consist~~eds~~ of a large number of ~~was dominated by~~ large water bodies with complex shapes ~~formed by glaciation (Smith et al., 2007)~~. The number of identified water bodies in the tundra (3.~~32~~ 24 million) and boreal forests (3.2~~3~~3 million) were nearly identical. However, the water extent in the boreal forest (0.57 million km^2; ~~70~~71% of total water area) is more than twice that found in the tundra (0.23 million km^2; ~~30~~29% of the total water area), indicating ~~suggesting again~~ that~~–~~ the average size of water bodies in the boreal~~tundra~~ are~~a~~ are ~~smaller~~larger than ~~than those~~ those in the tundra~~boreal~~. This finding was confirmed by reviewing the water body perimeters for the two biomes. ~~–~~The average perimeter of water bodies in boreal forests was 1.2 km, compared to a much smaller 0.8 km average perimeter for water bodies in the tundra. The average *SI* for water bodies in the boreal was 1.~~46~~45, longer than the 1.37 average *SI* for the tundra water bodies, suggesting that the boreal water bodies generally have much more complex shorelines, while the tundra water bodies are more circular.



341

342 **Figure 7:** The aggregated distribution of area (a), perimeter (b), and *SI* (c), and the number (d) of the identified water bodies in the
343 study area. The values at each 5 km × 5 km pixel in the grid ~~was~~were calculated by selecting the intersecting water bodies ~~interest~~
344 ~~with the extent of the pixel~~ and then either counting~~ed~~ or calculating~~ed~~ the mean of the targeted parameter (e.g., area, *SI*, and
345 perimeter) ~~from~~of these selected water bodies~~The average area, (a), perimeter, (b), *SI*, (c), and number (d) of identified water bodies~~
346 ~~in the study area aggregated to 5 km × 5 km grids for visualization~~.

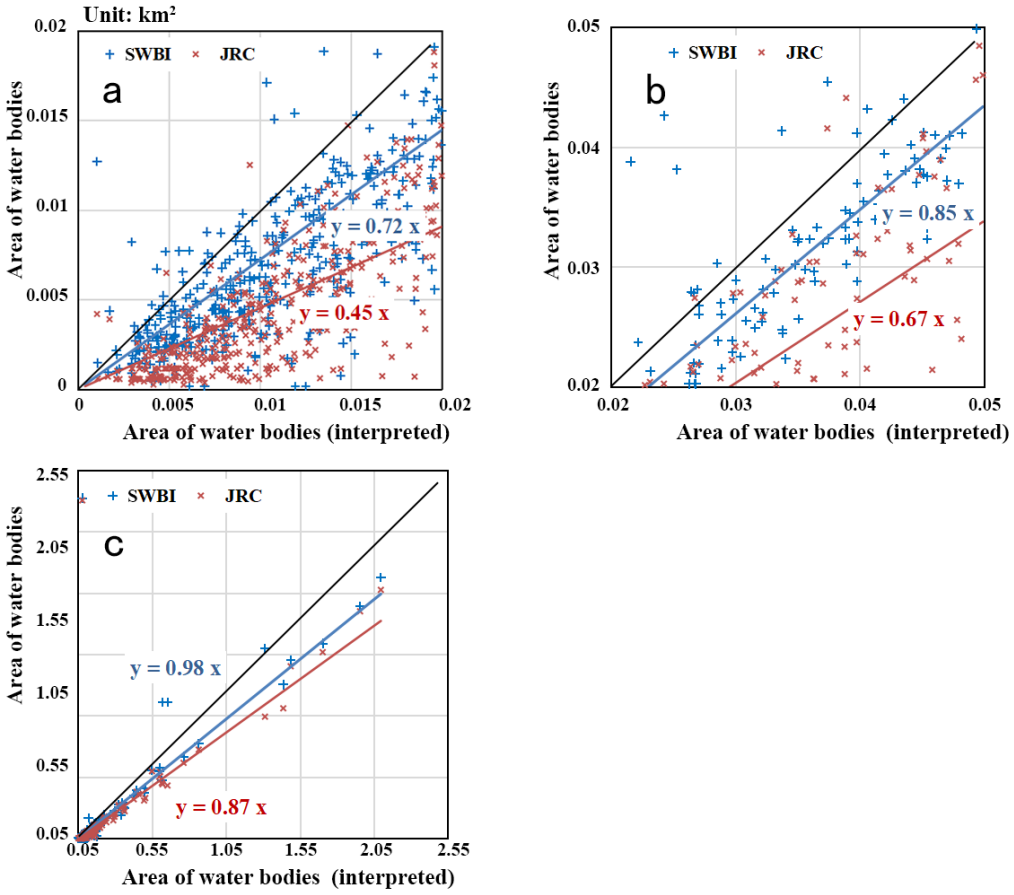
347

348 Inland water in the region is mainly concentrated in the Canadian Shield, i.e., about 0.~~79~~73 million km^2 of water (~~98~~92% of
349 water extent in the study region). In addition, most large water bodies were located in the Canadian Shield, including ~~75~~90%
350 of the identified large water bodies (sizes ~~≤~~≥ 1km^2). The shorelines of the water bodies in the Canadian Shield were also
351 more complex than those in other areas, especially south of the Laurentian Plateau near the Great Lakes.

352 **5.2 Accuracy assessment**

353 The overall accuracy of the ~~SWBI's~~WBD-NAHL's water extent was 96.36%, while the producer's accuracy was 99.9%, and
354 the user's accuracy was 96.36%. Misclassifications were primarily found in shadows of the Mackenzie Mountains, where the
355 east-west high-elevation mountain range cast constant shadows on the northern slopes.

356 Both the JRC and ~~SWBI~~WBD-NAHL datasets accurately identified the size of larger water bodies. For mixed water pixels,
357 the area estimates of ~~the two~~both datasets were~~are~~ more conservative than the reference data. However, the ~~SWBI~~WBD-NAHL
358 dataset performed better than the JRC.~~,~~ ~~and t~~The advantage of the ~~SWBI~~WBD-NAHL was demonstrated for smaller water
359 bodies (Figure 8). For small water bodies (size ≤ 0.02 km^2), the average area of the ~~SWBI~~WBD-NAHL water bodies was
360 72% of those manually digitized over high-resolution Google Earth images, compared to only 45% with the water area detected
361 by the JRC (Figure 8a). For medium water bodies (between 0.02 km^2 and 0.05 km^2), the average area of ~~SWBI~~WBD-NAHL
362 water bodies was about 85% times that of manually digitized water bodies, compared to 67% with the water area detected by
363 the JRC (Figure 8b). For water bodies larger than 0.05 km^2, the water areas of ~~SWBI~~WBD-NAHL were highly consistent
364 (98%) with that of manually digitized.~~,~~ w~~W~~hile the water area of JRC was slightly lower (about 87%) for water bodies in the
365 category (Figure 8c).



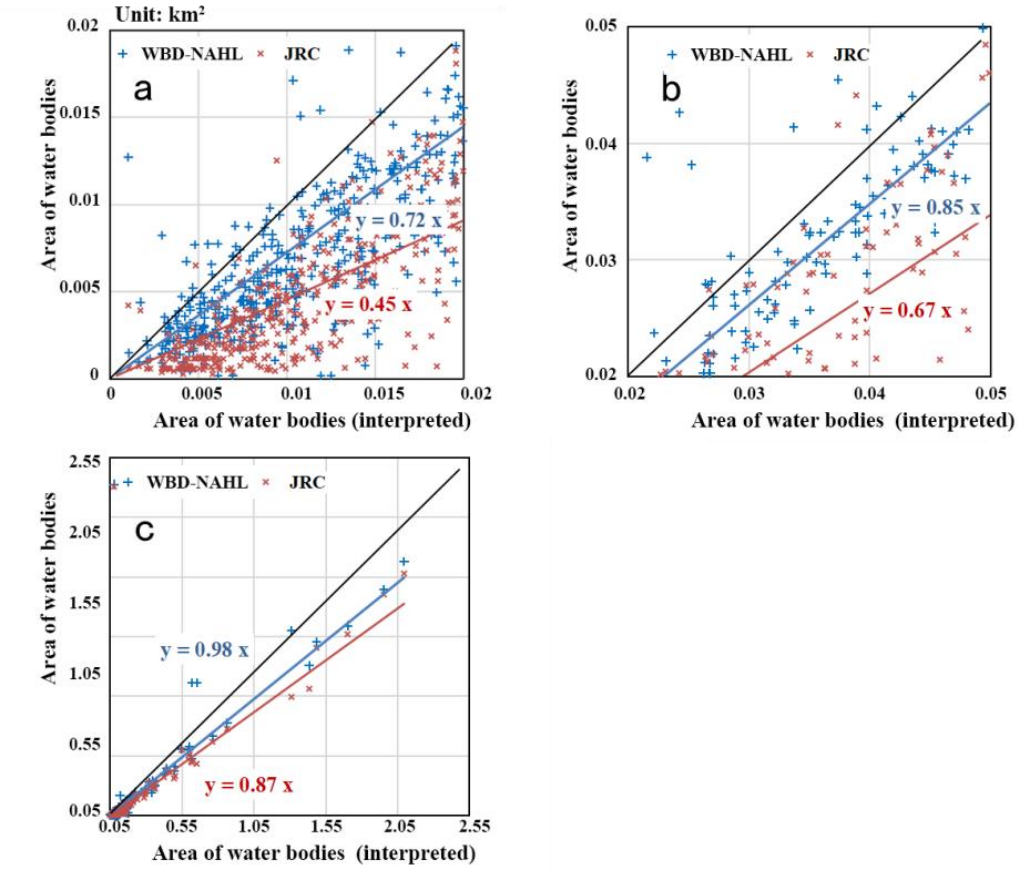


Figure 8: Comparisons of the water body area identified by the JRC, ~~SWBI~~WBD-NAHL, and interpreted water maps. The 1:1 lines are in black. The red crosses represent the JRC water bodies, and the blue pluses represent the ~~SWBI~~WBD-NAHL water bodies~~,~~ in comparison with the manually interpreted water bodies. The water bodies are compared in groups of sizes, i.e., (a) small water bodies with sizes < 0.02 km²; (b) medium water bodies with sizes between 0.02 km² and 0.05 km²; (c) large water bodies with sizes >

0.05 km². The R^2 for the ~~SWBI~~WBD-NAHL and JRC identified water bodies were similar, i.e., 0.6 for small water bodies, 0.5 for medium water bodies, and 0.9 for large water bodies.

The comparison between the water bodies identified by ~~SWBI~~WBD-NAHL and ~~the~~ PeRL were largely consistent for the derived indicators of water area, perimeter, and number (Figure 9). Linear correlations between the water bodies identified by ~~SWBI~~WBD-NAHL and ~~the~~ PeRL ~~water bodies reported~~had R^2 higher than 0.99 for all ~~the~~ three indicators. The slopes of the linear regressions ~~reported~~ indicated that the water area showed the least bias when compared to ~~the~~ PeRL (slope=0.98), followed by the number of water bodies (slope=0.78), and finally the perimeter of the water bodies (slope=0.62).

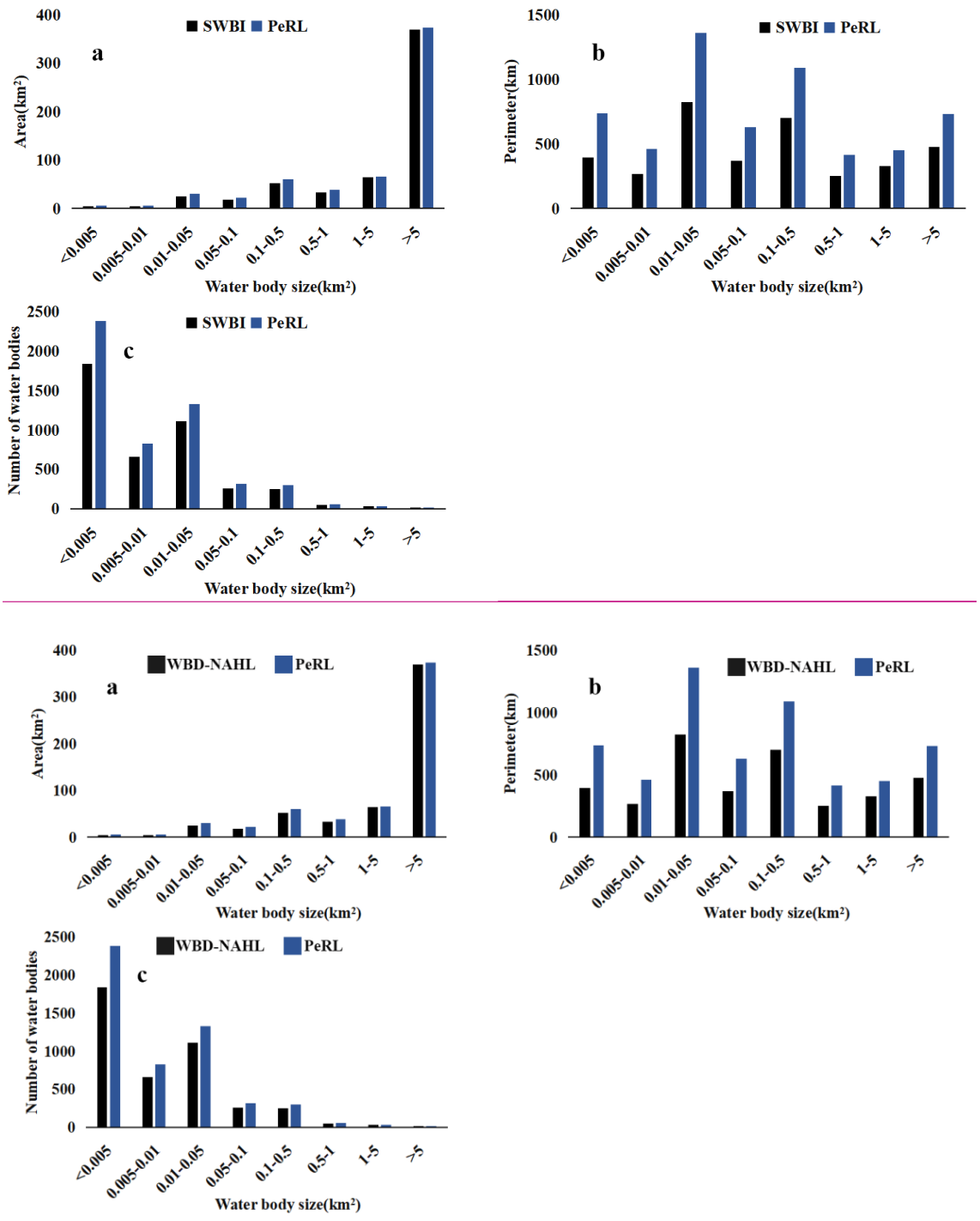


Figure 9: ~~The a~~Area, perimeter, and number of ~~the~~ water bodies identified by the PeRL and ~~SWBI~~WBD-NAHL dataset~~s~~.

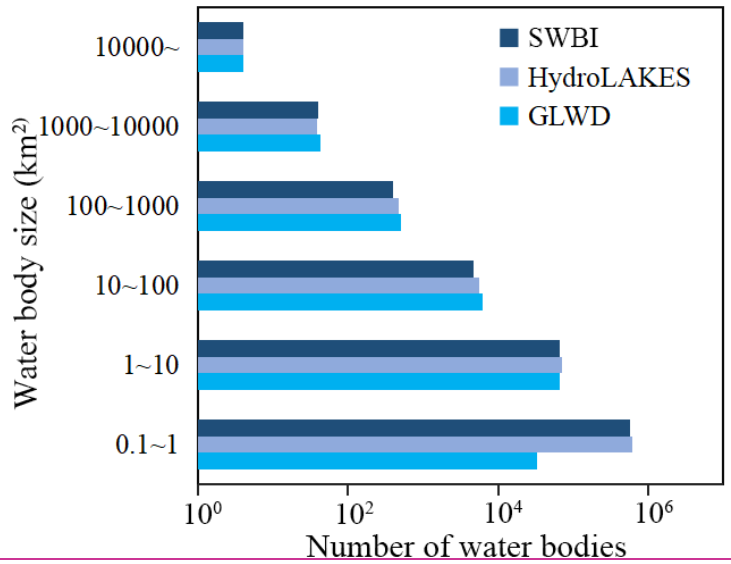
6 Discussion

6.1 A high-resolution water body dataset for the continental tundra and boreal

The ~~SWBI~~WBD-NAHL dataset provides the first known delineation of water bodies at 10-m resolution for the continental tundra and boreal forest of North America, which is one of the highest concentrations of the global inland water, especially the small sized water bodies. The dataset not only maps the extent of inland water during 2019 but also identifies the water bodies and their morphological metrics, which are critical for understanding and modeling freshwater lentic ecosystems (Downing, 2009; Heathcote et al., 2015; Kuhn and Butman, 2021; MacIntyre et al., 2009; Muster et al., 2013). The ~~SWBI~~WBD-NAHL was produced using Sentinel-2 satellite data to take advantage of the high resolution and 2-3-day revisit time of Sentinel-2 satellites. Sentinel-2's revisit time allows the ~~SWBI~~WBD-NAHL to have sufficient observations during the snow-free season, which is critical for mapping inland surface water in this high latitude region with long periods of snow coverage.

The ~~SWBI's~~WBD-NAHL's 10-m resolution ~~provided the capability for~~ enabled detecting water bodies as small as 0.001 km^2. The validation showed that the WBD-NAHL~~WBI~~ dataset had ~~a~~ high overall accuracy and significantly improved upon the ability of the existing global JRC water maps for detecting small water (e.g., smaller than 0.006 km^2) than the existing global JRC water maps. These small water bodies consist of nearly half the total water bodies in the tundra and boreal forest regions of North America, and generally experience faster cycling of water, material, and energy than larger water bodies (Winslow et al., 2014; Carroll et al., 2011; Messager et al., 2016). The improved ~~SWBI~~WBD-NAHL dataset may provide more accurate inputs for hydrological estimates, which are vital components for understanding and modeling the pan-Arctic hydrological, biochemical, and energy cycling.

The higher resolution of ~~SWBI~~WBD-NAHL also provides the ability to delineate the number, area, and shoreline complexity of water bodies. Our comparison confirmed that ~~SWBI~~WBD-NAHL-derived water areas and shorelines were similar to those from the regional 5-m or finer resolution PeRL dataset. Meanwhile, the number of water bodies identified in ~~the SWBI~~WBD-NAHL was consistent with those of other datasets, including HydroLAKES and GLWD (Figure 10). The number~~s~~ of water bodies larger than 1 km^2 ~~were~~ was roughly identical for ~~the SWBI~~WBD-NAHL, HydroLAKES, and GLWD ~~for water bodies larger than 1 km^2~~. For ~~the~~ water bodies between 0.1 and 1 km^2, ~~the SWBI~~WBD-NAHL and HydroLAKES reported similar numbers (Figure 10), but the number reported by GLWD was considerably lower, suggesting that the omission error of GLWD was higher for water bodies smaller than 1 km^2, as noted by Lehner and Döll (2004). Unfortunately, both the HydroLAKES and GLWD datasets only provide records ~~of~~ for water bodies larger than 0.1 km^2 (Messager et al., 2016; Lehner and Döll, 2004), and are thus missing records for what we estimate to be 90% of the total number of water bodies in the region. ~~The SWBI~~ The WBD-NAHL is able to extend these indicators to much smaller water bodies than HydroLAKES and GLWD, providing a much more complete record of water bodies in the region. This estimate of the number and extent ~~for~~ of small water bodies can improve our understanding of continental freshwater sources, stressing the importance of small water bodies in continental biochemical and energy cycling, potentially correcting a misconception that large lakes are most important (Downing, 2010).



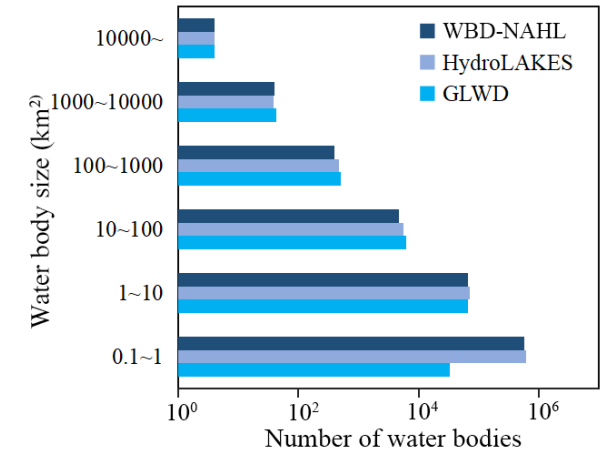


Figure 10: Comparing the number of water bodies identified by ~~the SWBI~~WBD-NAHL and ~~by~~ other datasets based on size ~~class~~.

6.2 Distribution of the water bodies

An empirical power-law distribution was found between lake areas and lake numbers (Messager et al., 2016; Downing et al., 2006), and the distribution ~~has been~~was applied to estimate the number of small lakes, which were used for estimating greenhouse gas emissions (Holgersson et al., 2016). According to the power-law distribution and HydroLAKES, the number of water bodies larger than 0.1 km^2 was estimated to be about 798,895, which was close to the 6~~36~~29,~~338~~130 water bodies reported by WBD-NAHL (Figure 11). However, the number of water bodies sized between 0.1 and 0.01 km^2 was estimated to be about 10.2 million, 4.8 times higher than ~~the~~estimated by WBD-NAHL. Furthermore, the water bodies sized between 0.01-0.001 km^2 ~~was~~were estimated to be about 126.1 million, 3~~5~~3.~~9~~6 times higher than ~~the~~what was estimated by WBD-NAHL, suggesting that the power-law distribution significantly overestimat~~ing~~es the number of small lakes.~~, and a~~A similar finding was reported by ~~which was also confirmed by~~ Seekell et al. (2016). ~~Considering the importance of the number of small water bodies to greenhouse gas emissions,~~ Estimating the number small water bodies using a ~~the~~ power-law distribution could ~~cause~~introduce considerable uncertainties in the estimation of the contribution~~y for~~ of small water bodies to greenhouse gas emissions.~~, and a~~Accurately identifying small water bodies could ~~contribute to~~ correct~~ing~~ thi~~s~~e overestimation and improv~~ing~~es ~~the~~ greenhouse gas emission estimates (Holgersson et al., 2016).~~According to this study, the water bodies 0.1-0.01km^2 is about 2.13 million, 4.8 times less than expected (extrapolated by HydroLAKES). The water bodies 0.01-0.001km^2 is about 3.52 million, 35.9 times less than expected (extrapolated by HydroLAKES). The large water bodies > 0.1 km^2 has a power law~~

~~distribution. The small water bodies < 0.1 km^2 deviate from the power-law distribution. The small water bodies will be overestimated with assumption that small water bodies have the same power-law distribution as large water bodies. This finding has been confirmed by the study of the global water body area distribution (Seekell et al., 2016). The number of small water bodies is important for estimates of Greenhouse Gas Emissions. The previous assessment on Greenhouse Gas Emission is based on databases that only included larger water bodies, along with assumptions of the number and area of smaller lakes (Holgersson et al., 2016). This study offered a observed value for small water bodies. This will contribute to estimates of greenhouse gas emissions in the tundra and boreal forests of North America.~~

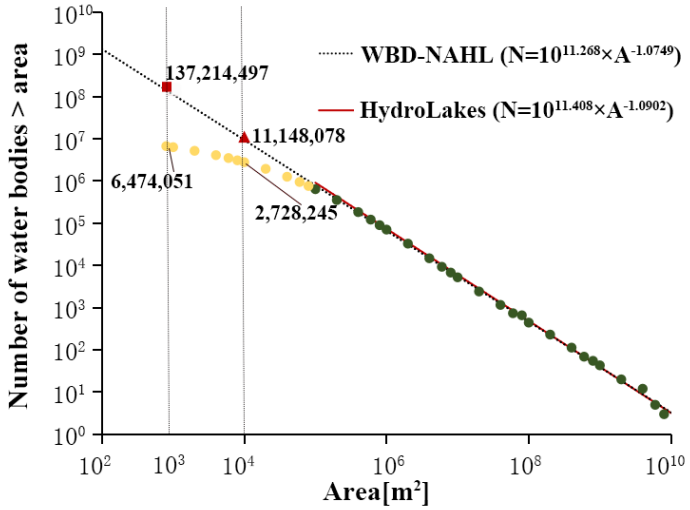


Figure 11: ~~The d~~Distribution of ~~the~~ total numbers of water bodies ~~correspondingin~~ relation to ~~the~~ areas ~~of water bodies in the~~ North American tundra and boreal forests water bodies~~of North America~~. The ~~round dots~~circles ~~presented~~represent the number~~s~~ of water bodies provided by ~~the~~ WBD-NAHL. The black line ~~was~~is the power-law distribution modeled using ~~the~~ water bodies > 0.1 km^2 from ~~the~~ WBD-NAHL. The red line ~~was~~is the power-law distribution modeled using ~~the~~ HydroLAKES in the study region. The red triangle and square ~~respectively~~ represent, respectively,~~ed~~ the extrapolated number~~s~~ of water bodies > 0.01 km^2 and > 0.001 km^2 based on the power-law distribution modeled from ~~the~~ HydroLAKES.

The largest and most complex water bodies are distributed primarily in the Canadian Shield. These lakes in the Canadian Shield formed through processes such as erosion and glaciation (Smith et al., 2007). Erosion and glaciation formed water bodies with complex shapes, which may contribute to the higher *SI* (1.48) reported by the ~~SWBI~~WBD-NAHL for the region. During the most recent Wisconsin glaciation, the Canadian Shield was covered by the Laurentide Ice Sheet, a giant, 3-km thick expanse of ice. When the ice sheet retreated north, it carved out the five Great Lakes as well as thousands of small lakes throughout the Canadian Shield (Dyke and Prest, 1987). Currently, ~~98~~92% of the water extent in the tundra and boreal forests are distributed in this particular region. For example, the largest lake in the region - Great Bear Lake - has a surface area of 30,227 km^2 with a long, complex shoreline (the perimeter is 5,705 km and the *SI* of the lake is 9.3). It was formed by ice erosion during the Pleistocene (Johnson, 1975).

The tundra, on the other hand, ~~distributed~~has a large number of ~~is dominated by~~ small, regularly--shaped water bodies, which could be related to the thick ~~overburden (e.g., the~~ peatland and~~, the~~ thermokarst landscape~~)~~~~The tundra, on the other hand, is dominated by small, regular shaped water bodies, which is related to the thawing and freezing of permafrost (Grosse et al., 2013). During the winter, water in the soil can freeze into ice. The freezing soil becomes puffy, forming a hilly structure. In the summer, this hilly structure melts and settles, forming a thermokarst lake. This hilly structure is small and regular, resulting in small, circular thermokarst lakes (Grosse et al., 2013). Numerous~~dominant thick ~~overburden which is either a result from being unglaciated (including aeolian deposits), or from being former seabottom that has been rising through isostatic rebound,~~

or by being located in regions with thick moraines or widespread peatlands.surficial geology whereThe thick overburden offers the developmental environment for round water bodies (e.g. thermokarst lake) can easily develop. DuringOver the past few decades, numerous thermokarst lakes are have been experiencing dramatic changes, which is are considered as an indicator for of permafrost degradation (Smith et al., 2005; Karlsson et al., 2012, 2014). The small thermokarst lakes were also found to experienceing stronger changes comparing than to the larger lakes (Karlsson et al., 2014; Carroll and Loboda, 2017). Monitoring water extent without discriminating by lake sizes could does not precisely accurately reflect those these strong changes in the small lakes due to the area dominance of large lakes. Additionally, the small thermokarst lakes are the primary source of permafrost carbon emissions (Kuhn et al., 2018; Walter Anthony et al., 2016; Yvon-Durocher et al., 2017), and the small water bodies were found to be a major source of uncertainty in estimating greenhouse gas emission estimates (Holgerson and Raymond, 2016). The SWBIWBD-NAHL dataset could provide critical information for investigating thermokarst lakes, especially the small thermokarst lakes and ponds, and estimating their effects on carbon emission and permafrost sustainability in the tundra and boreal forests in of North America. As reported by the analysis of the SWBIWBD-NAHL, 3.32 24 million small water bodies were found in the tundra in 2019, with an average size of 0.07 km^2 and average *SI* of 1.37, much smaller than the *SI* of the boreal lakes in the boreal. Teshekpuk Lake is the largest thermokarst lake in the world with and a relatively smooth shoreline (*SI* = 5.4), considerably smaller than the *SI* of the Great Bear Lake in the boreal region (Markon and Derksen, 1994).

The biome-based analysis provided an shallow insights into the distribution of the water body shapes across the study area; however, more complex relationships can be found between the shapes and the surficialsurface geology of the water bodies. For example, more circular-shaped lakes can be found in regions with thick overburden – eitherpossibly as a a result from beingremaining unglaciated, (includingfrom aeolian deposits), or from beingrising from the former sea bottom that has been rising through isostatic rebound; moreover, tThese circular-shaped lakes can be found located in regions with thick moraines or widespread peatlands in the boreal Hudson Bay lowlands and the Mackenzie River bBasin. The high-resolution WBD-NAHL could providehelp a key dataset for further exploreing the distribution of not only the water bodies but also these with specificby sizes or and shapes.

6.3 Limitations

The data and methods used to derive the 10-m resolution SWBIWBD-NAHL dataset are able to detect water bodies smaller than the 30-m or coarser--resolution satellite--derived datasets, but have difficulty identifying water bodies smaller than 0.001 km^2,, and theThis limitation capability can be further improved by incorporating higher resolution satellite data, such as from Planet, WorldView, QuickBird, and Gaofen (Veremeeva and Günther, 2017; Sun et al., 2020; Watson et al., 2016; Andresen and Lougheed, 2015). ErrorsDue to the lLimit eErrors in the satellite data provide substantial sources of uncertainty, including an inability to separate rivers and streams because the resolution is too coarse, bias in estimates of water extent resulting from temporal gaps in the data, and misclassifications resulting from spectral resolution. The misclassifications impacted by terrain (e.g., mountain shadows) still exist even though they have been substantially reduced during data processing. Further processing may be possible to further reduce these errors.

This dataset was produced using satellite data acquired in 2019, and it does not reflect changes of the water bodies in the region. The WBD-NAHL dataset was produced based on Sentinel-2 data acquired in the summer of 2019, and the result represents the distribution of surface water in the corresponding year. The mean of total cumulative precipitation in 2019 in the region was 438.5 mm, which was close to the historical average from 2010 to 2019 (mean: 435.9 mm, standard deviation: 11.5 mm) (HuffmanJin et al. 20219)(GPM IMERG). Although 2019 can be considered as a normal year of the past decade in terms of precipitation (Jin et al. 2021), the spatial extent of high-latitude water bodies, especially smaller water bodies, can still vary significantly both inter- and intra-annually in subregionslocally, and the number and distribution of water bodies could be

significantly different from 2019. The mean of total cumulative precipitation in 2019 in the region was 438.5 mm, which was close to the historical average from 2010 to 2019 (mean: 435.9 mm, standard deviation: 11.5 mm) (GPM IMERG). Nevertheless, it would be interesting to explore thewater bodies' changes using observations from multiple years in the future.The dataset was produced based on Sentinel-2 data acquired in the summer of 2019, and the result represent the distribution of surface water in the corresponding year. Although 2019 can be considered as a normal year of the past decade in terms of precipitation (Jin et al., 2021), the spatial extent of high-latitude water bodies, especially smaller water bodies, can still vary significantly both inter- and intra-annually, and the number and distribution of water bodies could be significantly different from 2019. Nevertheless, it would be interesting to explore the changes using observations from multiple years.To avoid the large deviation of water body area and number, we analyzed the annual cumulative precipitation from 2010 to 2019, the 2019 was a normal year.(The total cumulative precipitation in 2019 is 30,869 mm/hr. The historical average is 30,937 mm/hr. The mean square deviation is 1,701 mm/hr.) Further efforts can be carried out to produce an inland water dataset for multiple time periods using these methods to capture the seasonal and multi-year dynamics of inland water in the region. ThisThe WBD-NAHL dataset focused on the tundra and boreal forest regions inof North America. With the application of tThe methodology in North America, it would be interestingcan be to extended to Eurasia to provide a complete representation of the biomes in the future.

The biome-based analysis provided an insight into the distribution of the water body shapes across the study area; however, more complex relationship can be found between the shapes and the surficial geology of the water bodies. For example, more circular-shaped lakes were found in regions with thick overburden – either a result from being unglaciated (including aeolian deposits), or from being former sea bottom that has been rising through isostatic rebound, or by being located in regions with thick moraines or widespread peatlands. The distribution of these water bodies may not be limited to a specific biome, for example, circular-shaped lakes could be found in the extensive peatland regions of the Hudson Bay lowlands and the Mackenzie River basin in the boreal regions.This study analyzed the shape and size of water bodies based on ecoregions. This roughly explained the reason of formation by analyzing the dominant surficial geology. In the future, more detailed research combined with geology was needed to further reveal the formation reason of water body size and shape. This dataset focused on the tundra and boreal in North America. With the application of the methodology in North America, it would be interesting to extend to Eurasia to provide a complete representation of the biomes in the future.This dataset focused on the tundra and boreal in North America. Following the methodology, it would be interesting to cover the tundra and boreal in Eurasia to provide a complete representation of the biomes.

7 Data availability

This WBD-NAHL dataset can be accessed via the website of the National Tibetan Plateau/Third Pole Environment Data Center (TPDC, http://data.tpdc.ac.cn): DOI: 10.11888/Hydro.tpdc.271021 (Feng et al., 2020). The dataset is provided in ESRI Geodatabase format. The volume of this dataset is about 1.5 GB.

8 Conclusions

This study presents an inland surface water body dataset of tundra and boreal forest biomes offor the northern latitudes of North American. high latitudes. The SWBIWBD-NAHL dataset was generated using Sentinel-2 data with machine learning methods and an object-based algorithm. Three morphological metrics (area, perimeter, and *SI*) were calculated for each water body. The aAccuracy of the dataset was carefully assessed with respect to detecting inland surface water extent (or pixel level) and identifying water bodies. The dataset's overall accuracy for water extent reached 96.36%. In addition, the WBD-NAHLWBI showed a high consistency with high--resolution images in terms of water area, perimeter, and quantity.

To our knowledge, the ~~SWBI~~WBD-NAHL dataset provided the most complete inventory of inland surface water bodies for the tundra and boreal forest ~~forest~~ regions of North America. Overall, 6.~~65~~ 47 million water bodies were identified, covering 10.3% of the region. Small water bodies ~~were dominance~~dominate ~~in~~ the region, as ~~with~~ ~90.~~43~~% ~~were~~ have an area smaller than 0.1 km^2. ~~Results from an analysis of t~~The ~~SWBI~~WBD-NAHL indicates that the tundra biome is dominated by densely distributed small water bodies with regular shapes (the average *SI* was 1.37), while the boreal forest biome is dominated by large water bodies with complex shapes (the average *SI* was 1.~~46~~45). The WBD-NAHL~~WBI~~ is expected to be able to provide supporting data for modeling hydrologic, biochemical, and energy cycling in these areas.

Acknowledgements

This work was supported by Basic Science Center for Tibetan Plateau Earth System (BSCTPES, NSFC project No. 41988101)

Reference

Andresen, C. G. and Lougheed, V. L.: Disappearing Arctic tundra ponds: Fine-scale analysis of surface hydrology in drained thaw lake basins over a 65 year period (1948–2013), J. Geophys. Res. Biogeosciences, 120, 466–479, 2015.

Biskaborn, B. K., Smith, S. L., Noetzli, J., Matthes, H., Vieira, G., Streletskiy, D. A., Schoeneich, P., Romanovsky, V. E., Lewkowicz, A. G., and Abramov, A.: Permafrost is warming at a global scale, Nat. Commun., 10, 1–11, 2019.

Breiman, L.: Random forests, Mach. Learn., 45, 5–32, 2001.

Carlson, T. N. and Ripley, D. A.: On the relation between NDVI, fractional vegetation cover, and leaf area index, Remote Sens. Environ., 62, 241–252, 1997.

Carpenter, S. R.: Lake geometry: implications for production and sediment accretion rates, J. Theor. Biol., 105, 273–286, 1983.

Carroll, M. and Loboda, T.: Multi-Decadal Surface Water Dynamics in North American Tundra, Remote Sens., 9, https://doi.org/10.3390/rs9050497, 2017.

Carroll, M. L., Townshend, J. R. G., DiMiceli, C. M., Loboda, T., and Sohlberg, R. A.: Shrinking lakes of the Arctic: Spatial relationships and trajectory of change, Geophys. Res. Lett., 38, 2011.

Cooley, S. W., Smith, L. C., Ryan, J. C., Pitcher, L. H., and Pavelsky, T. M.: Arctic-Boreal Lake Dynamics Revealed Using CubeSat Imagery, Geophys. Res. Lett., 46, 2111–2120, https://doi.org/10.1029/2018gl081584, 2019.

Danielson, J. J. and Gesch, D. B.: Global multi-resolution terrain elevation data 2010 (GMTED2010), 2011.

Downing, J. A.: Global limnology: Up-scaling aquatic services and processes to planet Earth, Int. Ver. Für Theor. Angew. Limnol. Verhandlungen, 30, 1149–1166, 2009.

Downing, J. A.: Emerging global role of small lakes and ponds: little things mean a lot, Limnetica, 29, 0009–0024, 2010.

Dranga, S. A., Hayles, S., and Gajewski, K.: Synthesis of limnological data from lakes and ponds across Arctic and Boreal Canada, Arct. Sci., 4, 167–185, https://doi.org/10.1139/as-2017-0039, 2017.

Du, J., Kimball, J. S., Jones, L. A., and Watts, J. D.: Implementation of satellite based fractional water cover indices in the pan-Arctic region using AMSR-E and MODIS, Remote Sens. Environ., 184, 469–481, https://doi.org/10.1016/j.rse.2016.07.029, 2016.

Dyke, A. and Prest, V.: Late Wisconsinan and Holocene history of the Laurentide ice sheet, Géographie Phys. Quat., 41, 237–263, 1987.

Fayne, J. V., Smith, L. C., Pitcher, L. H., Kyzivat, E. D., Cooley, S. W., Cooper, M. G., Denbina, M. W., Chen, A. C., Chen, C. W., and Pavelsky, T. M.: Airborne observations of arctic-boreal water surface elevations from AirSWOT Ka-Band InSAR and LVIS LiDAR, Environ. Res. Lett., 15, 105005, 2020.

Feng, M., Sexton, J. O., Channan, S., and Townshend, J. R.: A global, high-resolution (30-m) inland water body dataset for 2000: first results of a topographic–spectral classification algorithm, Int. J. Digit. Earth, 9, 113–133, https://doi.org/10.1080/17538947.2015.1026420, 2015.

Feng, M., Sui, Y.: High resolution inland surface water dataset for the tundra and boreal in North America, https://doi.org/10.11888/Hydro.tpdc.271021, 24 October 2020.

Forkel, M., Carvalhais, N., Rödenbeck, C., Keeling, R., Heimann, M., Thonicke, K., Zaehle, S., and Reichstein, M.: Enhanced seasonal CO2 exchange caused by amplified plant productivity in northern ecosystems, Science, 351, 696–699, 2016.

Glińska-Lewczuk, K.: Water quality dynamics of oxbow lakes in young glacial landscape of NE Poland in relation to their hydrological connectivity, Ecol. Eng., 35, 25–37, https://doi.org/10.1016/j.ecoleng.2008.08.012, 2009.

Graversen, R. G., Mauritsen, T., Tjernström, M., Källén, E., and Svensson, G.: Vertical structure of recent Arctic warming, Nature, 451, 53–56, 2008.

Grosse, G., Jones, B., and Arp, C.: 8.21 Thermokarst Lakes, Drainage, and Drained Basins, in: Treatise on Geomorphology, edited by: Shroder, J. F., Academic Press, San Diego, 325–353, https://doi.org/10.1016/B978-0-12-374739-6.00216-5, 2013.

Han-Qiu, X.: A study on information extraction of water body with the modified normalized difference water index (MNDWI), J. Remote Sens., 5, 589–595, 2005.
Heathcote, A. J., del Giorgio, P. A., and Prairie, Y. T.: Predicting bathymetric features of lakes from the topography of their surrounding landscape, Can. J. Fish. Aquat. Sci., 72, 643–650, 2015.

Higgins, S., Desjardins, C., Drouin, H., Hrenchuk, L., and Van der Sanden, J.: The role of climate and lake size in regulating the ice phenology of boreal lakes, J. Geophys. Res. Biogeosciences, 126, e2020JG005898, 2021.
Holgerson, M. A. and Raymond, P. A.: Large contribution to inland water CO2 and CH4 emissions from very small ponds, Nat. Geosci., 9, 222–226, https://doi.org/10.1038/ngeo2654, 2016.
Huffman, G.J., E.F. Stocker, D.T. Bolvin, E.J. Nelkin, Jackson Tan (2019), GPM IMERG Final Precipitation L3 1 month 0.1 degree x 0.1 degree V06, Greenbelt, MD, Goddard Earth Sciences Data and Information Services Center (GES DISC), Accessed: 2022.5.24, 10.5067/GPM/IMERG/3B-MONTH/06
van Huissteden, J., Berrittella, C., Parmentier, F. J. W., Mi, Y., Maximov, T. C., and Dolman, A. J.: Methane emissions from permafrost thaw lakes limited by lake drainage, Nat. Clim. Change, 1, 119–123, https://doi.org/10.1038/nclimate1101, 2011.
Isikdogan, F., Bovik, A. C., and Passalacqua, P.: Surface Water Mapping by Deep Learning, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 10, 4909–4918, https://doi.org/10.1109/JSTARS.2017.2735443, 2017.
Jiang, X., Zheng, P., Cao, L., and Pan, B.: Effects of long-term floodplain disconnection on multiple facets of lake fish biodiversity: Decline of alpha diversity leads to a regional differentiation through time, Sci. Total Environ., 763, 144177, 2021.
Johannessen, O. M., Bengtsson, L., Miles, M. W., Kuzmina, S. I., Semenov, V. A., Alekseev, G. V., Nagurnyi, A. P., Zakharov, V. F., Bobylev, L. P., and Pettersson, L. H.: Arctic climate change: observed and modelled temperature and sea-ice variability, Tellus Dyn. Meteorol. Oceanogr., 56, 328–341, 2004.
Johnson, L.: The Great Bear Lake: its place in history, Arctic, 28, 231–244, 1975.
Karlsson, J., Lyon, S., and Destouni, G.: Temporal Behavior of Lake Size-Distribution in a Thawing Permafrost Landscape in Northwestern Siberia, Remote Sens., 6, 621–636, https://doi.org/10.3390/rs6010621, 2014.
Karlsson, J. M., Lyon, S. W., and Destouni, G.: Thermokarst lake, hydrological flow and water balance indicators of permafrost change in Western Siberia, J. Hydrol., 464–465, 459–466, https://doi.org/10.1016/j.jhydrol.2012.07.037, 2012.
King, K. B., Bremigan, M. T., Infante, D., and Cheruvelil, K. S.: Surface water connectivity affects lake and stream fish species richness and composition, Can. J. Fish. Aquat. Sci., 78, 433–443, 2021.
Kuhn, C. and Butman, D.: Declining greenness in Arctic-boreal lakes, Proc. Natl. Acad. Sci., 118, 2021.
Kuhn, M., Lundin, E. J., Giesler, R., Johansson, M., and Karlsson, J.: Emissions from thaw ponds largely offset the carbon sink of northern permafrost wetlands, Sci. Rep., 8, 9535, https://doi.org/10.1038/s41598-018-27770-x, 2018.
Laird, N. F., Walsh, J. E., and Kristovich, D. A.: Model simulations examining the relationship of lake-effect morphology to lake shape, wind direction, and wind speed, Mon. Weather Rev., 131, 2102–2111, 2003.
Langer, M., Westermann, S., Boike, J., Kirillin, G., Grosse, G., Peng, S., and Krinner, G.: Rapid degradation of permafrost underneath waterbodies in tundra landscapes—Toward a representation of thermokarst in land surface models, J. Geophys. Res. Earth Surf., 121, 2446–2470, https://doi.org/10.1002/2016jf003956, 2016.
Laske, S. M., Rosenberger, A. E., Wipfli, M. S., and Zimmerman, C. E.: Surface water connectivity controls fish food web structure and complexity across local- and meta-food webs in Arctic Coastal Plain lakes, Food Webs, 21, https://doi.org/10.1016/j.fooweb.2019.e00123, 2019.
Lehner, B. and Döll, P.: Development and validation of a global database of lakes, reservoirs and wetlands, J. Hydrol., 296, 1–22, https://doi.org/10.1016/j.jhydrol.2004.03.028, 2004.
Li, X., Che, T., Li, X., Wang, L., Duan, A., Shangguan, D., Pan, X., Fang, M., and Bao, Q.: CASEarth poles: big data for the three poles, Bull. Am. Meteorol. Soc., 101, E1475–E1491, 2020.
Lindgren, P. R., Farquharson, L. M., Romanovsky, V. E., and Grosse, G.: Landsat-based lake distribution and changes in western Alaska permafrost regions between the 1970s and 2010s, Environ. Res. Lett., 16, 025006, https://doi.org/10.1088/1748-9326/abd270, 2021.
MacIntyre, S., Fram, J. P., Kushner, P. J., Bettez, N. D., O'brien, W., Hobbie, J., and Kling, G. W.: Climate-related variations in mixing dynamics in an Alaskan arctic lake, Limnol. Oceanogr., 54, 2401–2417, 2009.
Markon, C. J. and Derksen, D. V.: Identification of tundra land cover near Teshekpuk Lake, Alaska using SPOT satellite data, Arctic, 222–231, 1994.
McCullough, I. M., King, K. B. S., Stachelek, J., Diaz, J., Soranno, P. A., and Cheruvelil, K. S.: Applying the patch-matrix model to lakes: a connectivity-based conservation framework, Landsc. Ecol., 34, 2703–2718, https://doi.org/10.1007/s10980-019-00915-7, 2019.
McFeeters, S. K.: The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features, Int. J. Remote Sens., 17, 1425–1432, 1996.
Messager, M. L., Lehner, B., Grill, G., Nedeva, I., and Schmitt, O.: Estimating the volume and age of water stored in global lakes using a geo-statistical approach, Nat. Commun., 7, 1–11, 2016.
Meyer, M. F., Labou, S. G., Cramer, A. N., Brousil, M. R., and Luff, B. T.: The global lake area, climate, and population dataset, Sci Data, 7, 174, https://doi.org/10.1038/s41597-020-0517-4, 2020.
Muster, S., Heim, B., Abnizova, A., and Boike, J.: Water body distributions across scales: A remote sensing based comparison of three arctic tundra wetlands, Remote Sens., 5, 1498–1523, 2013.

Napiórkowski, Bąkowska, Mrozińska, Szymańska, Kolarova, and Obolewski: The Effect of Hydrological Connectivity on the Zooplankton Structure in Floodplain Lakes of a Regulated Large River (the Lower Vistula, Poland), Water, 11, https://doi.org/10.3390/w11091924, 2019.
Nitze, I., Cooley, S. W., Duguay, C. R., Jones, B. M., and Grosse, G.: The catastrophic thermokarst lake drainage events of 2018 in northwestern Alaska: Fast-forward into the future, The Cryosphere, 14, 4279–4297, 2020.
Olefeldt, D., Goswami, S., Grosse, G., Hayes, D., Hugelius, G., Kuhry, P., McGuire, A. D., Romanovsky, V. E., Sannel, A. B. K., Schuur, E. a. G., and Turetsky, M. R.: Circumpolar distribution and carbon storage of thermokarst landscapes, Nat. Commun., 7, 13043, https://doi.org/10.1038/ncomms13043, 2016
Pachauri, R. K. and Reisinger, A.: IPCC fourth assessment report, IPCC Geneva, 2007, 2007.
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.
Pickens, A. H., Hansen, M. C., Hancher, M., Stehman, S. V., Tyukavina, A., Potapov, P., Marroquin, B., and Sherani, Z.: Mapping and sampling to characterize global inland water dynamics from 1999 to 2018 with full Landsat time-series, Remote Sens. Environ., 243, https://doi.org/10.1016/j.rse.2020.111792, 2020.
Ritter, M. E.: The physical environment: An introduction to physical geography, Date Visit. July, 25, 2008, 2006.
Sandlund, O. T., Eloranta, A. P., Borgstrøm, R., Hesthagen, T., Johnsen, S. I., Museth, J., and Rognerud, S.: The trophic niche of Arctic charr in large southern Scandinavian lakes is determined by fish community and lake morphometry, Hydrobiologia, 783, 117–130, https://doi.org/10.1007/s10750-016-2646-5, 2016.
Schilder, J., Bastviken, D., van Hardenbroek, M., Kankaala, P., Rinta, P., Stötter, T., and Heiri, O.: Spatial heterogeneity and lake morphology affect diffusive greenhouse gas emission estimates of lakes, Geophys. Res. Lett., 40, 5752–5756, 2013.
Serikova, S., Pokrovsky, O. S., Laudon, H., Krickov, I., Lim, A. G., Manasypov, R. M., and Karlsson, J.: High carbon emissions from thermokarst lakes of Western Siberia, Nat. Commun., 10, 1–7, 2019.
Serreze, M. C. and Francis, J. A.: The Arctic amplification debate, Clim. Change, 76, 241–264, 2006.
Sharma, S., Blagrave, K., Magnuson, J. J., O'Reilly, C. M., Oliver, S., Batt, R. D., Magee, M. R., Straile, D., Weyhenmeyer, G. A., Winslow, L., and Woolway, R. I.: Widespread loss of lake ice around the Northern Hemisphere in a warming world, Nat. Clim. Change, 9, 227–231, https://doi.org/10.1038/s41558-018-0393-5, 2019.
Smith, L. C., Sheng, Y., MacDonald, G. M., and Hinzman, L. D.: Disappearing arctic lakes, Science, 308, 1429–1429, 2005.
Smith, L. C., Sheng, Y., and MacDonald, G. M.: A first pan-Arctic assessment of the influence of glaciation, permafrost, topography and peatlands on northern hemisphere lake distribution, Permafr. Periglac. Process., 18, 201–208, https://doi.org/10.1002/ppp.581, 2007.
Sun, J., Wang, G., He, G., Pu, D., Jiang, W., Li, T., and Niu, X.: Study on the Water Body Extraction Using GF-1 Data Based on Adaboost Integrated Learning Algorithm, Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci., 42, 641–648, 2020.
Van Gerven, M. and Bohte, S.: Artificial neural networks as models of neural information processing, Front. Comput. Neurosci., 11, 114, 2017.
Veremeeva, A. A. and Günther, F.: Thermokarst lake and baydzherakh area changes on Yedoma uplands, Yakutian coastal lowlands: repeat inventory using high resolution imagery, 2017.
Vaideliene, A., & Michailov, N. (2008). Dam influence on the river self-purification. 748–757.
Walter Anthony, K., Daanen, R., Anthony, P., Schneider von Deimling, T., Ping, C.-L., Chanton, J. P., and Grosse, G.: Methane emissions proportional to permafrost carbon thawed in Arctic lakes since the 1950s, Nat. Geosci., 9, 679–682, https://doi.org/10.1038/ngeo2795, 2016.
Watson, C. S., Quincey, D. J., Carrivick, J. L., and Smith, M. W.: The dynamics of supraglacial ponds in the Everest region, central Himalaya, Glob. Planet. Change, 142, 14–27, 2016.
Winslow, L. A., Read, J. S., Hanson, P. C., and Stanley, E. H.: Lake shoreline in the contiguous United States: quantity, distribution and sensitivity to observation resolution, Freshw. Biol., 59, 213–223, 2014.
Xiong, G., Wang, G., Wang, D., Yang, W., Chen, Y., & Chen, Z. (2017). Spatio-temporal distribution of total nitrogen and phosphorus in Dianshan lake, China: The external loading and self-purification capability. Sustainability, 9(4), 500.
Yvon-Durocher, G., Hulatt, C. J., Woodward, G., and Trimmer, M.: Long-term warming amplifies shifts in the carbon cycle of experimental ponds, Nat. Clim. Change, 7, 209–213, https://doi.org/10.1038/nclimate3229, 2017.