

Spectral index-driven FCN model training for water extraction from multispectral imagery

Zhenshi Li, Xueliang Zhang^{*}, Pengfeng Xiao

Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, Key Laboratory for Land Satellite Remote Sensing Applications of Ministry of Natural Resources, School of Geography and Ocean Science, Nanjing University, Nanjing, Jiangsu 210023, China

ARTICLE INFO

Keywords:

Water delineation
Spectral index
Sentinel-2
Semantic labeling
Deep fully convolutional network

ABSTRACT

Terrestrial water is a fundamental component of the land surface. Accurate and robust water delineation of surface water is rather challenging due to the high intra-class variability and the spectral similarity with shadows and other dark surfaces. This study proposes a new method called the water index-driven deep fully convolutional neural network (WIDFCN) for high-accuracy water delineation with no need to collect samples manually. We formulate water delineation as a semantic labeling problem and solve it by training a deep fully convolutional network (FCN), whose capability of effectively extracting multilevel spatial and spectral features is exploited for discriminating water from complex surroundings. The main obstacle of using FCN, which requires a large volume of labeled training samples, is settled by utilizing the water recognition ability of the water spectral index (WI). Specifically, the training samples are automatically generated from WI by extracting a high-precision but incomplete water mask at first, which is then expanded to enhance the completeness. This strategy ensures the high quality of the automatically generated training samples and thus the water extraction performance of the trained FCN model. Twelve test sites from Sentinel-2 imagery with various water delineation challenges all over the world are used to assess the performance relative to that of supervised and unsupervised classification methods and water spectral index thresholding methods, in terms of Kappa coefficient, precision, recall, and F1-score. Overall, WIDFCN can achieve the highest precision and comparable recall, leading to the best water delineation accuracy with Kappa coefficient 0.9673 and F1-score 0.9696, as well as the lowest fluctuation in terms of various test sites. The results further demonstrate that WIDFCN can effectively deal with the scale and spectra variance of surface water, and has distinct robustness with respect to different kinds of shadows, including building, mountain, and cloud shadows. The findings in this study demonstrate a novel, robust, low-cost, and manual labor-free water delineation method that performs well in terms of both precision and completeness. Moreover, the core idea could provide a reference for extraction of geographic information by utilizing FCN models with no needs of manually labeling costs. Code and data will be available at <https://github.com/LZhenShi/WIDFCN>.

1. Introduction

Terrestrial water, as a main component of the land surface, plays an essential role for all creatures (Oki and Kanae, 2006). Delineating the spatial distribution of surface water is a significant and valuable task (Pekel et al., 2016) and can be utilized in many applications, such as water resource monitoring (Allen and Pavelsky, 2018; Chittoor Viswanathan et al., 2016; Güttler et al., 2013; Kiselev et al., 2015; Zou et al., 2017), natural hazard analysis (Huber et al., 2013; Yamazaki et al., 2014), and urban environment management (Chen et al., 2017; Marlow

et al., 2013). With the development of satellite technology, a variety of remote sensing data from spaceborne sensors provide opportunities for water mapping on a global scale at multiple spatial and temporal resolutions (Palmer et al., 2015).

Nevertheless, automating the delineation of water surfaces is by no means an easy task. On the one hand, the variability of water spectral signatures caused by different constituents, including phytoplankton, suspended sediments, and colored dissolved organic matter (CDOM), increases the difficulty of reliable water discrimination (Pekel et al., 2016, 2014). On the other hand, the commission errors from the

^{*} Corresponding author.

E-mail address: zxl@nju.edu.cn (X. Zhang).

<https://doi.org/10.1016/j.isprsjprs.2022.08.019>

Received 6 February 2022; Received in revised form 10 August 2022; Accepted 19 August 2022

0924-2716/© 2022 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.

similarity of water to shadows and low-albedo surfaces have all along been the main obstacles for water delineation (Feyisa et al., 2014; Wang et al., 2018; Yang et al., 2018).

Water spectral index (WI) combining with thresholding is a simple and useful water delineation method, which has been widely developed since the appearance of the normalized difference water index (NDWI) (McFeeters, 1996). With the simplicity of calculation, WI is nearly the most popular way for water extraction, which only needs a simple threshold to distinguish between water and non-water (Feyisa et al., 2014; Fisher et al., 2016). Xu (2006) proposed MNDWI, which replaces the near infrared (NIR) band in NDWI with the middle infrared (MIR) band, aiming at suppressing the noise from built-up areas. Subsequently, WIs with more bands are proposed, such as AWEI (Feyisa et al., 2014) and MBWI (Wang et al., 2018), to utilize more spectral features for greater robustness, which can discriminate water better from shadows. However, WI methods are always faced with difficulty in determining the optimal threshold. Yang et al. (2015) proposed to utilize fuzzy clustering method (FCM) for background homogenization to reduce threshold variability, yet the thresholding procedure is still required. Besides, the heterogeneity of water body and complexity of background make it difficult for applying WI to achieve high-performance water delineation. Consequently, the rule-based approach is proposed (Yang et al., 2020), in which expertise-driven prior knowledge of water and non-water regions is combined and multiple indexes are used together for higher accuracy. Nevertheless, rule-based methods also face the thresholding problem, which is always determined by human experience (Yang et al., 2020), and auxiliary data are also required (Pekel et al., 2016; Yamazaki et al., 2015; Yang et al., 2020).

Machine learning (ML) based supervised classification is also an effective approach for water delineation from remote sensing imagery (Lary et al., 2016). This kind of method, such as support vector machine (Nandi et al., 2017), random forest (Feng et al., 2015), multilayer perceptron (Mishra and Prasad, 2015), is independent of setting thresholds, and can use a variety of features, including WIs, for achieving higher water extraction accuracy (Jia et al., 2018). However, ML based methods rely on a large volume of training samples with plentiful diversity, whose collection is time-consuming, expensive, tedious, and even requires extensive field work (Bangira et al., 2019). Furthermore, it depends on the effectiveness of manually designed features, and is sometimes not robust enough to be universally applicable (Huang et al., 2018).

As a branch of ML techniques, deep learning (DL) has significantly forced progress in various scientific disciplines in recent years thanks to its strength to extract deep features and capture highly complex data distributions (Shen, 2018). It has also been widely employed in several hydrological applications (Sit et al., 2020) and has shown potential in water extraction (Bui et al., 2020; Hu et al., 2019; Wang et al., 2020). The fully convolutional network (FCN) is an end-to-end DL method for pixel-level image labeling (Long et al., 2015), which has dramatically prompted high accuracy geographical information extraction from remote sensing imagery (Chai et al., 2019; Diakogiannis et al., 2020; Kampffmeyer et al., 2016; Sun et al., 2018; Wurm et al., 2019). One of the obstacles of using FCNs is the requirement of a large number of pixel-level training samples, which requires a high level of labor and expertise (Li et al., 2021). The water masks (ground-truth) used for training FCN models for water delineation in existing studies are mostly from manual interpretation (Guo et al., 2020; Wang et al., 2020; Wang et al., 2020), which requires huge manual labeling costs, or referring to land cover classification datasets (Feng et al., 2019; Isikdogan et al., 2017), such as GLCF (Feng et al., 2016), whose resolution is insufficient for fine-grained training.

In this study, we propose water index-driven deep fully convolutional network (WIDFCN) for water extraction, which can combine the advantages of WI and DL, i.e. the identification ability of water bodies and feature extraction capability. We treat the water extraction task as a semantic labeling task (Hao et al., 2020), i.e., train an FCN for pixel-level

water delineation. Unlike previous methods that rely on visual interpretation or land cover classification datasets for ground-truth (Feng et al., 2019; Guo et al., 2020; Isikdogan et al., 2017; Wang et al., 2020; Wang et al., 2020), we generate water masks based on the water spectral index, which completely eliminates the costs of manual collection. To comprehensively examine the generality and robustness of WIDFCN, experiments were carried out on 12 representative sites with various water delineation challenges chosen around the world.

The contributions of this study are as follows: (1) We propose a new water delineation paradigm in which the WI is utilized to drive the training of FCN model. The proposed paradigm, called WIDFCN, can not only take advantage of the powerful DL, but also avoid the requirements of manually labeling training samples; (2) We propose a scheme for solving the inevitable inaccuracy and incompleteness of the training samples obtained from WI, composed of a mask generation method that ensures precision first and then expands for completeness, as well as utilizing median frequency balancing loss for class balancing; (3) We evaluate the accuracy and generality of WIDFCN in 12 test sites from different areas in the world, and demonstrate that WIDFCN can achieve high-performance water delineation, which can deal with water heterogeneity and shadow confusion; (4) We analyze the reason why the spectral index is effective for the training of FCN model for water delineation. We expect that the core idea of WIDFCN can provide a reference for information extraction for other surface features driven by spectral index.

2. Study areas and dataset

2.1. Satellite dataset

The Sentinel-2 mission is composed of two polar-orbiting satellites, Sentinel-2A and Sentinel-2B. Both satellites carry a multispectral instrument (MSI) that provides global coverage of the Earth's land surface with a high revisit frequency from 2 to 5 days (Drusch et al., 2012). The Sentinel-2 imagery consists of 13 spectral bands spanning from the visible (VIS) and NIR to the shortwave infrared (SWIR) bands at different spatial resolutions ranging from 10 m to 60 m. We choose Sentinel-2 imagery in this study because the high revisit frequency is of great importance to land cover dynamic mapping and monitoring, and its free use allows large-scale applications around the world. However, it is worth noting that the proposed WIDFCN is a general method and can be applied for other satellite datasets.

We use Sentinel-2B Level-2A imagery downloaded directly from the Sentinels Scientific Data Hub (<https://scihub.copernicus.eu/dhus/#/home>). The Level-2A dataset is the orthoimage bottom-of-atmosphere (BOA) corrected reflectance product generated by applying an atmospheric correction to top-of-atmosphere (TOA) Level-1C orthoimage products. Only 6 bands extracted from the original 13 bands are used in this study, including VIS, NIR, and SWIR bands. The two SWIR bands in 20 m were resampled to 10 m processed by the official software Sentinel Application Platform (SNAP) to maintain a unique spatial resolution, namely, 10 m.

2.2. Training dataset

Considering that the training dataset demands to cover as many challenging situations as possible to train a robust model, the middle and lower reaches of the Yangtze River are chosen to generate training samples. The water surfaces at this site have abundant hydrologic features and complex surroundings, which are suitable for generating rich training samples. Originating from the Qinghai-Tibet Plateau, the Yangtze River flows across 11 provinces of China from west to the east and finally into the East China Sea. Sixteen sites distributed in the middle and lower reaches of the Yangtze River are utilized, characterized by different surface water types, scales, landscapes, and topographies. The distribution of the 16 sites is illustrated in the extent of

training set in Fig. 1.

All 16 Sentinel-2 scenes are partitioned into a set of small non-overlapping images with a size of 256×256 pixels. In this way, 27,870 image patches are produced. These image patches combined with their masks, whose generation will be shown in Section 3.2, form

the training data of the water delineation model.

2.3. Test dataset

The test dataset consists of 12 sites with various challenges to

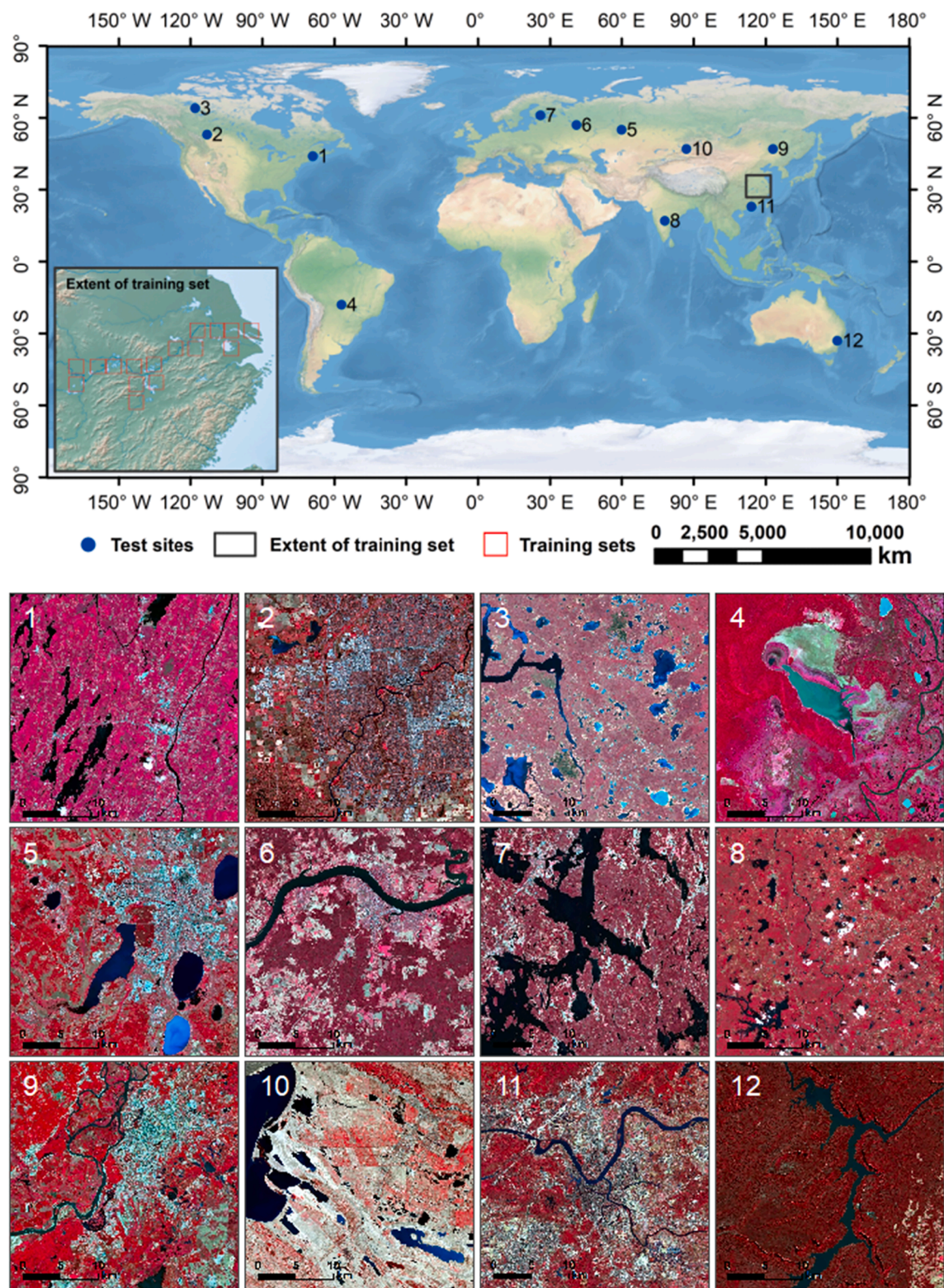


Fig. 1. Distribution of the training sites and test sites, in which the extent of the training set is presented, together with the Sentinel-2 images of the test set in false color composite of near infrared, red, and green bands. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

accurate water delineation (Table 1). These sites spread all over the world, with heterogeneous water conditions, multiple landscapes, and various water types, including linear or irregular rivers, different-sized lakes, ponds, reservoirs, etc. The dataset includes (1) Kennebec River, (2) North Saskatchewan River, (3) Great Bear Lake, (4) Paraguay River, (5) Smolino Lake, (6) Volga River, (7) Pajanne Lake, (8) Manjira River, (9) Nen River, (10) Ulungur Lake, (11) Pearl River, and (12) Burrigorang Lake. The Kennebec River located in America is a 270-km-long river, which collects water from hundreds of tributary streams and rivers. The north Saskatchewan River is a glacier-fed river characterized by a cool, humid continental climate, passing through Edmonton, Canada's fifth largest city. The Great Bear Lake is the largest lake in Canada, where northern continental climate is presented, surrounded by many lake clusters. The Paraguay River basin includes a sandy wetland, Pantanal, which is the world's largest wetland. Flooded vegetation and organic matter tend to be carried into the river. The Smolino Lake is a light-salted lake adjoining the Chelyabinsk City, with a humid continental climate. The Volga River is the longest river in Europe, with the city of Kineshma on the shore of it. The Pajanne Lake is the deepest and second largest lake in Finland, containing sandy eskers, isolated lagoons and small islets. The Manjira River is a major river in India, in which reservoirs are built on its tributaries. The Nen River locates in Northeast China and flows through the city of Qiqihar, in which building and building shadows are widely distributed. The Ulungur Lake is one of China's ten largest freshwater lakes, mainly formed and maintained by the seasonal replenishment of water from the melting of Altai mountain glaciers and snow. The Pearl River is the third-longest river in China, and the Huizhou City crossed by which provides numerous building shadows. The Burrigorang Lake is a major reservoir for Sydney, which fills the Burrigorang Valley.

All the images in the test dataset are clipped into 3000×3000 pixels (30×30 km) to achieve higher water percentages and lower visual interpretation costs (Jia et al., 2018). Two kinds of reference data are used to evaluate the water delineation accuracy. One is labeling all the pixels in the 3000×3000 maps, i.e. labeling the reference maps, which is the common-used way for evaluating the accuracies of semantic segmentation (Chai et al., 2019; Yue et al., 2019) and water delineation (Jia et al., 2018; Yang et al., 2015). The other is labeling the random reference samples selected by the probability sampling design (Stehman and Czaplewski, 1998), used for assessing the seasonal accuracies. Specifically, a geographic stratification is used to partition the map into 3×3 km cells to ensure the samples spatially well distributed (Stehman and Foody, 2019), which is a major criterion when choosing a sampling design (Stehman, 2009), followed by an extraction of 50 points within each cell in a simple random scheme. In total, 5000 points are selected in each map. Both reference data are digitized visually on screen by

comparing with the high-spatial resolution images in Google Earth.

3. Methodology

3.1. DeeplabV3-Plus as water delineation model

Semantic labeling (also called semantic segmentation), an end-to-end method to assign a category label to each pixel in an image, has been greatly promoted since the emergence of DL technology in the computer vision field (Hao et al., 2020). FCN (Long et al., 2015) is the pioneering research, and a sequence of improved methods have been developed, such as U-Net (Ronneberger et al., 2015), PSPNet (Zhao et al., 2017), and DeepLab (Chen et al., 2014). In this study, we adopt DeepLabV3-Plus (Chen et al., 2018b), a state-of-the-art method, as our baseline water delineation model to discover water features. DeepLabV3-Plus includes an atrous spatial pyramid pooling (ASPP) module (Chen et al., 2018a) to exploit the multiscale context information. Furthermore, DeepLabV3-Plus applies the encoder-decoder architecture (Long et al., 2015; Noh et al., 2015; Ronneberger et al., 2015), namely, to use DeepLabV3 (Chen et al., 2017) as the encoder module that gradually reduces the feature map resolution and captures higher semantic information, and add a simple decoder module that gradually recovers the spatial information to obtain fine-grained segmentation. We find this model effective for water extraction when facing various irregular and multiscale water bodies. It should be noted that the semantic labeling model can be any off-the-shelf algorithm. Water delineation with higher quality can be obtained with a more effective algorithm, which is not the focus of this study.

3.2. Spectral index-driven water mask generation

For the fully supervised semantic labeling task, ground-truths are pixel-level annotations, possessing the corresponding semantic class for every pixel in the image. In this section, we show how to generate appropriate water masks from WI, which serves as the ground-truth for training the water delineation model. The generation process is divided into two steps. First, a dual-threshold (dual-thre) is applied to the WI result to generate the base mask with highly precise water and non-water labels (Section 3.2.1). In the base mask, there are some ignored pixels without labels, whose spectral features could not be utilized for training FCN model. Therefore, the base mask is then iteratively expanded into the ex-mask, in which the ignored pixels can be gradually assigned the label of water, leading to a more complete water mask finally (Section 3.2.2). For the choice of WI, NDWI is adopted in this study. The overall pipeline of generating the water mask is demonstrated in Fig. 2.

Table 1

Detailed description of the 12 test sites, where SN refers to the serial number of the test sites.

SN	Country/continent	Name	Water type	Major challenge	Tile	Date
1	America/ North America	Kennebec River	Lake, pond, shallow river	Cloud shadow, phytoplankton	19TDK	05/24/2021
2	Canada/ North America	North Saskatchewan River	Lake, river, pond	Built-up, building shadow	11UQV	10/08/2021
3	Canada/ North America	Great Bear Lake	Lake cluster, water stream	Bare land	11WMM	08/01/2021
4	Brazil/ South America	Paraguay River	River, lake, pond	Phytoplankton, mud	21KVV	05/01/2021
5	Russia/Europe	Smolino Lake	Lake, water stream	Built-up, building shadow	41ULB	08/22/2021
6	Russia/Europe	Volga River	river	Built-up	37VFD	05/11/2021
7	Finland/Europe	Pajanne Lake	Lake cluster	Sandy beaches, built-up	35VMJ	05/12/2021
8	India/Africa	Manjira River	River, pond, reservoir	Cloud, cloud shadow, dark Surface	43QHV	10/25/2021
9	China/Asia	Nen River	River, pond	Paddy field, built-up, building shadow, mud	51TWN	07/20/2020
10	China/Asia	Ulungur Lake	Lake, pond	Built-up, dark surface	45TWN	06/12/2020
11	China/Asia	Pearl River	River, lake, reservoir	Mountain shadow, building shadow, built-up	49QHF	01/29/2021
12	Australia/Oceania	Burrigorang Lake	Lake, reservoir	Mountain shadow, built-up	56HKH	08/17/2021

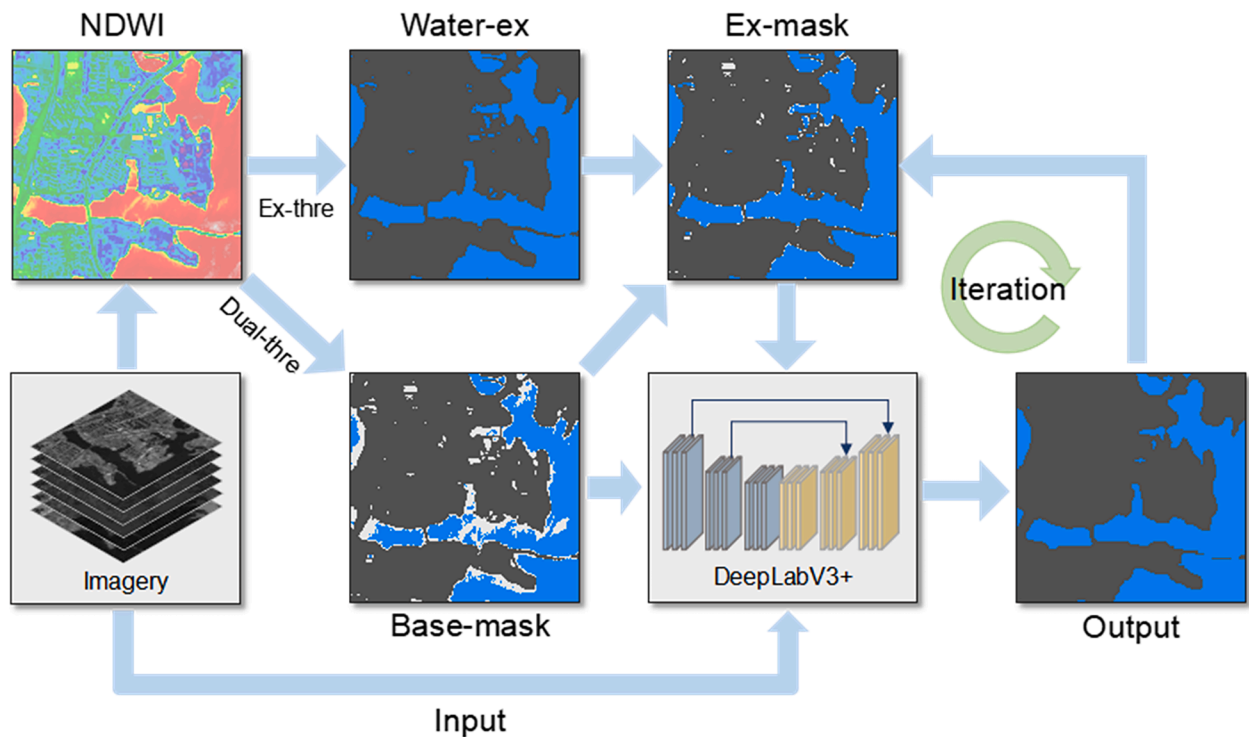


Fig. 2. The pipeline of WIDFCN. The gray, white, and blue in the base mask and ex-mask indicate non-water, ignored, and water, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.2.1. Base mask from dual-threshold

WI can generate different qualities of water masks depending on the adopted thresholds. In particular, when setting a higher threshold, the water region in the mask can be more precise, yet less adequate. The optimal threshold can lead to the best trade-off between precision and completeness, but it is hard to determine because it fluctuates dramatically when using different WIs or on different landscapes (Feyisa et al., 2014; Yang et al., 2015). In addition, a rigid one-cut threshold forces us to make a decision between higher precision yet lower completeness and higher completeness yet more noisy. Despite the noise tolerance ability of the DL semantic labeling model, the quality of the water mask can directly determine the model training (Li et al., 2021).

Instead of setting a hard single threshold as in previous works (Chen et al., 2020; Yang et al., 2015), we propose utilizing a dual-threshold method to generate base water masks (base mask). The dual-threshold is composed of a high threshold (h_t) and a low threshold (l_t), which satisfy $-1 < l_t < h_t < 1$. In addition, the base mask is generated according to Eq. (1). The objective of the dual-threshold is to prevent the effects from ambiguous pixels that are hard to recognize by WI, always being shadows or water with distinct features. To ensure the high precision of water and non-water in the base mask, we expect to set two safe thresholds in the dual-threshold, so the two values are much easier to determine compared to find the best trade-off. It is assumed that the more accurate the information in the base mask, the less noise will be introduced during the following expanding iteration, and the better the final mask can be obtained. Empirically, we found that setting h_t and l_t as 0.3 and 0, respectively, can be very qualified for Sentinel-2 data. When training the semantic labeling model, the ignored pixels are excluded for calculating losses to pretend negative effects from noise.

$$\begin{cases} \text{water,} & WI > h_t \\ \text{ignord pixel,} & l_t \leq WI \leq h_t \\ \text{non - water,} & WI < l_t \end{cases} \quad (1)$$

3.2.2. Iteratively expanding mask using ex-threshold

The base mask is initially used for training the delineation model, which can output water extraction with not only high precision but also superior recall than the base mask itself thanks to the self-learning power of the FCN model. However, there are still some ignored pixels without any labels in the base mask whose spectral features cannot be utilized for training, as shown in Fig. 2. To further improve the completeness of the water mask, the base mask is iteratively expanded in terms of the ignored pixels, whose procedure is shown in Fig. 2.

Specifically, we set up a threshold to WI for expansion, called ex-threshold (ex-thre), and the water bodies from applying ex-thre to WI, called water-ex, are utilized to update the base mask by combining with the water extraction from the model (trained with the base mask) output. For clear illustration, the detailed expansion process of an iteration is shown in Fig. 3. In particular, the water regions from the model output are taken as seeds. For each connected water domain in the water-ex, if there exists any pixel judged as water in the seeds, the whole connected water domain (expressed as ex-region in Fig. 3) is taken as water and will be used to update the corresponding position in the base mask. The update is only for the ignored pixels, and the ignored pixels can only be updated into water labels without consideration of non-water labels, which means that the non-water labels are determined from the original base mask with the purpose of maximizing the absence of noise. The updated base mask, called ex-mask, is used for training the delineation model successively, and the update continues iteratively. In this study, the ex-thre is set as 0.1, and the update operation performs every-five epochs when training the delineation model. The cross entropy loss is adopted for the iterative expansion.

The update procedure works due to two reasons. One is that the delineation model trained by the base mask (and by the updated mask) can generate very accurate but incomplete results. The other is that the water-ex possesses many complete water regions but also includes much noise. By complementing the highly precise seeds in the model output and the fairly complete water regions in the water-ex, the ex-mask can be improved continuously in iterations, and the ignored pixels will decrease gradually and finally with both high accuracy and

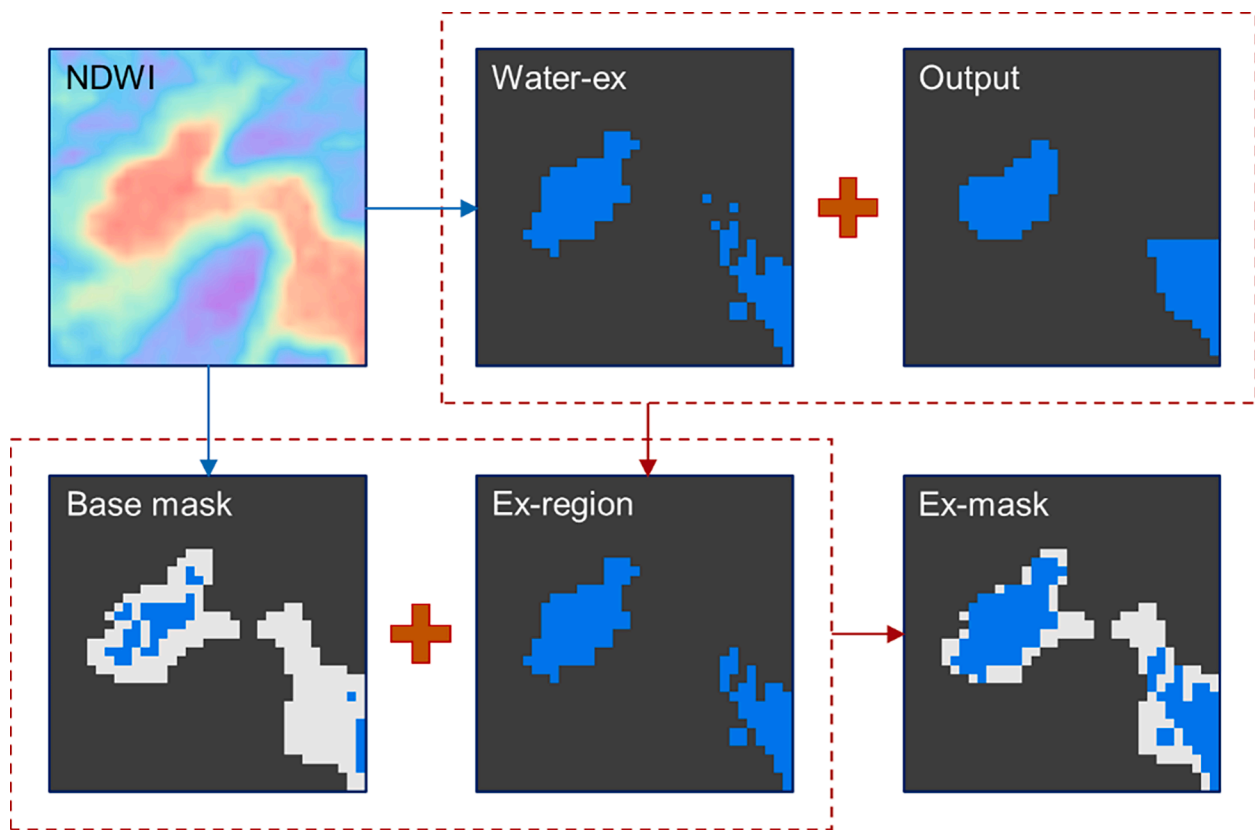


Fig. 3. Detailed illustration of a water mask expansion iteration, where the ‘base mask’ refers to the result of applying dual-threshold on the NDWI map, the ‘water-ex’ refers to the result of applying a single ex-threshold on the NDWI map, and the ‘output’ refers to the output of the trained water delineation model. The gray, white, and blue indicate non-water, ignored, and water, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

completeness.

For each expanding operation, the water bodies in the ex-mask can be more complete. The mask with complete water bodies can prompt the semantic labeling model to learn more water features and output more water bodies, which will be more complete in the next iteration. The ratio of water and non-water pixel numbers is calculated for each ex-mask in the update process, and the iteration stops when the difference of adjacent ratios is lower than a predefined threshold or the iteration times attain a predefined number.

3.3. Final training using MFB loss

After the water mask expansion, the ex-mask in the final iteration is taken as the final mask for training the final semantic labeling model (also using DeepLabV3-Plus). At this time, the ex-mask possesses a high completeness in terms of water labels. However, for the training dataset, sited on the middle and lower reaches of the Yangtze River, the water and non-water regions are imbalanced. Actually, this unbalanced phenomenon always occurs for inland water bodies. Therefore, we adopt a class-balanced weight for water and non-water in the cross-entropy loss when training the model with the final ex-mask, which may significantly improve the water recognition capability of the semantic labeling model and reduce the omission error.

In particular, the median frequency balancing (MFB) weight (Eigen and Fergus, 2015) is used as the class-imbalance weight, as shown in Eq. (2), where c is the class of water or non-water, $freq(c)$ is the pixel number of class c divided by the total number of pixels in images of training dataset, and $median_{freq}$ is calculated as the average value of $freq(water)$ and $freq(non-water)$.

$$W_c = median_{freq} / freq(c) \quad (2)$$

It is noted that MFB loss is only adopted for the final training of the semantic labeling model and is not used for the mask expansion operation. The reason is that although the introduction of MFB loss can greatly reduce the omission error, it can also induce a few commission errors. Even a minority of noise will affect the expanding operation to a great extent, which will lead to a dramatically inaccurate ex-mask after several iterations.

Overall, the whole method is in a two-step form, where the water mask serving as ground-truth is produced as described in Section 3.2, and the final water delineation model is trained using the water mask, i. e. the ex-mask, with MFB loss.

4. Results

4.1. Accuracy assessment

4.1.1. Accuracy assessment methods

Four accuracy metrics are adopted to evaluate the water delineation performance on the test dataset, including the Kappa coefficient, precision, recall, and F1-score (F1). The Kappa coefficient is an unbiased coefficient toward water and non-water, which can reflect the comprehensive accuracy. Precision and recall indicate the degree of accuracy and completeness, respectively, and F1-score takes the above two measures into account for comprehensive assessment (Zhang et al., 2015).

WIDFCN is quantitatively and qualitatively compared with several commonly used water extraction methods, including WI thresholding, unsupervised classification methods and supervised classification methods. It is noted that the term of ‘WIDFCN’ in this Section indicates the WIDFCN based on NDWI by default (WIDFCN-NDWI), if there is no

additional annotation. The adopted WIs for comparison include normalized difference water index (NDWI), modified normalized difference water index (MNDWI), and automated water extraction index (AWEI) consisted of AWEI_{sh} and AWEI_{nh} according to the shadow condition, the equations of which are shown below.

$$NDWI = \frac{G - NIR}{G + NIR} \quad (3)$$

$$MNDWI = \frac{G - SWIR1}{G + SWIR1} \quad (4)$$

$$AWEI_{nh} = \frac{4 \times (G - SWIR1) - (0.25 \times NIR + 2.75 \times SWIR2)}{G + NIR + SWIR1 + SWIR2} \quad (5)$$

$$AWEI_{sh} = \frac{B + 2.5 \times G - 1.5 \times (NIR + SWIR1) - 0.25 \times SWIR2}{B + G + NIR + SWIR1 + SWIR2} \quad (6)$$

where B, G, NIR, SWIR1, and SWIR2 indicate the reflectance value of spectral bands of Sentinel-2: band 2, band 3, band 8, band 11, and band 12. Two unsupervised classification methods, including k-means and ISODATA, and two supervised classification methods, random forest (RF) and support vector machine (SVM) are used for comparison. Otsu's thresholding algorithm (Otsu, 1975) is adopted to automatically obtain the thresholds for WIs (WIs-Otsu) at each test site. For the two unsupervised classification methods, 10 clusters are set and then merged into two classes of water and non-water. The base masks are used as samples for training the two supervised classification methods, in which all the water pixels (positive samples) are used, while the non-water pixels are randomly sampled to the total number of water pixels, aiming at ensuring the class balance and reducing the computer memory

requirements. Besides, water delineation results from the two supervised methods trained with the ex-mask are also analyzed, which will be shown in Section 4.1.3. Detailed comparison results with alternative methods will be shown in Section 4.1.2. In addition, the automatically acquired threshold using Otsu's method has difficulty achieving the best results, especially for complicated occasions. Thus, the ideally optimal threshold from traversal search is analyzed in Section 4.1.3. The temporal stability of the proposed method is also important, which will be analyzed in Section 4.1.4.

4.1.2. Comparison with alternative methods

The accuracies of alternative approaches for water body extraction on the test dataset is depicted in Fig. 4. It should be noted that the reference data used here are the reference maps with label on each pixel, as illustrated in Section 2.3. The accuracies in terms of the random reference samples are also calculated and present basically consistent with that of the reference maps, which are not displayed due to space limitation. In terms of Kappa coefficient and F1-score, WIDFCN can achieve the highest values (with Kappa coefficient 0.9673 and F1-score 0.9696), which outperforms other methods by a large margin, reflecting the best overall performance. Besides, it can be seen that WIDFCN has the lowest fluctuation, showing its applicability for all kinds of situations. In general, the two supervised classification methods perform inferior to WIDFCN and superior to the unsupervised classification methods and WIs-Otsu. RF performs better than SVM, with Kappa coefficient of 0.9391 and F1-score of 0.9432, which are the second highest among these methods. K-means and ISODATA act similarly, with Kappa coefficients of 0.8850 and 0.9017, respectively, which show apparent disparities compared to WIDFCN. The Kappa coefficients of the four WI-Otsu methods present great differences, in which the two AWEIs perform

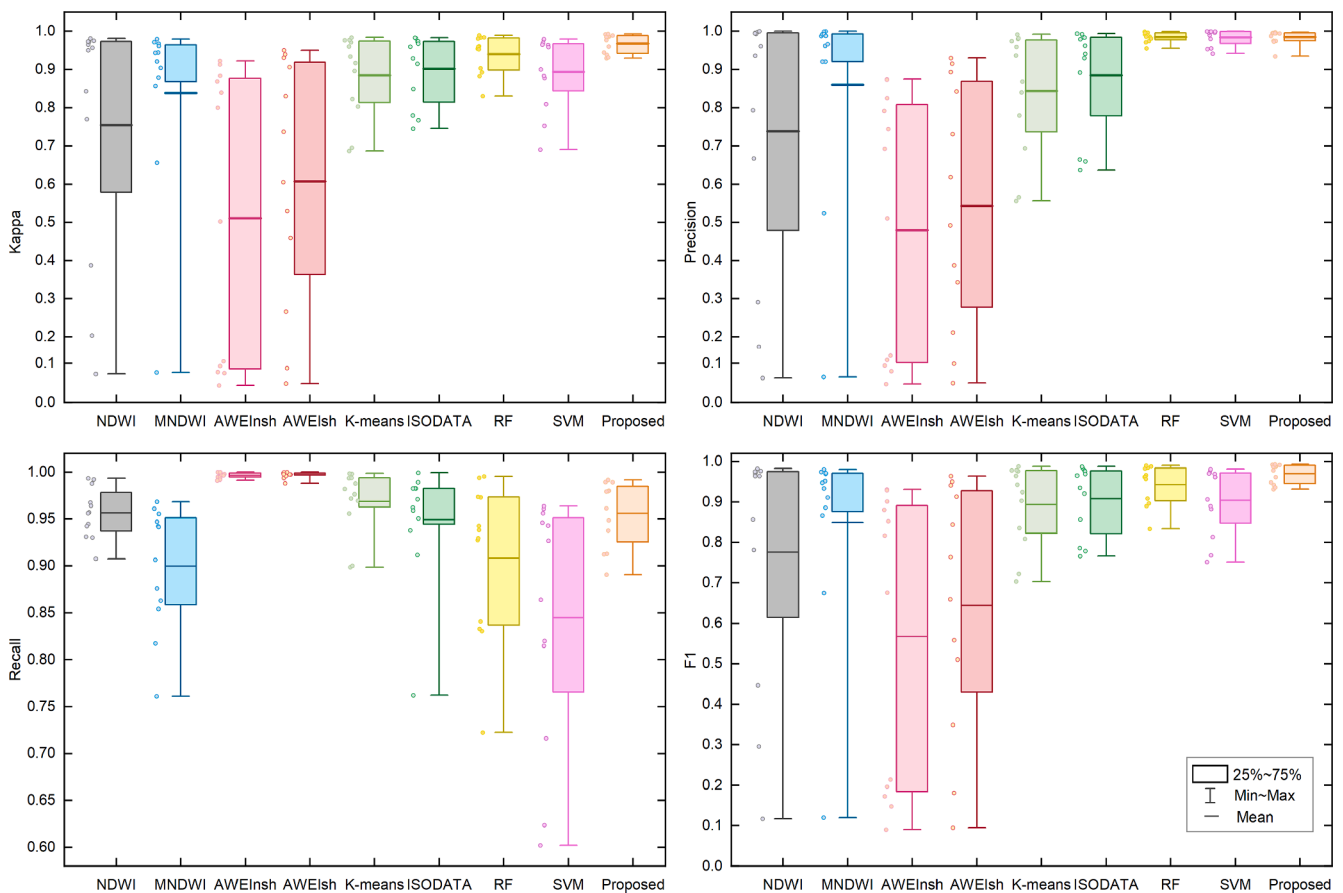


Fig. 4. Accuracies of WIDFCN compared with other methods, including four WIs thresholded using the Otsu's method, two unsupervised classification methods, and two supervised classification methods. The dots distributed on the left of the boxes represent accuracy values for each test site by the corresponding method.

the worst. This is mainly due to the improper thresholds from the Otsu's method, leading to the apparently low precision.

For precision of WIDFCN, the values of almost all the test sites are closely gathered, leading to a high mean precision (0.9840), as shown in Fig. 4. RF and SVM also produce high-precision results, because of the high-precision training samples. Interestingly, the precision distribution of the WIs-Otsu and the unsupervised classification methods in Fig. 4 shows a very similar trend with that of Kappa coefficient. This is because the recalls of these methods are relatively high, with the lowest mean value of 0.8995 achieved by MNDWI thresholding, and thus the precision has an important influence on the results. In terms of recall, the two supervised classification methods perform worst, which greatly affects their Kappa and F1. Although the two AWEI with Otsu's thresholding methods have the highest recall values, those results are hardly referential due to their low precision. The rest four methods (except for MNDWI thresholding) have similar recalls, in which WIDFCN has a mean value of 0.9561.

In summary, WIDFCN performs best on the test dataset according to the highest Kappa coefficient and F1-score, and is more stable than other methods with the least fluctuation. In addition, it can achieve apparently

high precision, although facing challenging test sites, as described in Table 1, and meanwhile with comparatively high recall. The supervised classification methods obtain high precision and low recall, while the unsupervised classification methods can achieve high recall and low precision. The WI-Otsu methods generate water delineations with severe noise because the Otsu's method can hardly assign appropriate thresholds for complicated sites.

4.1.3. Comparison with ideally optimal threshold for WIs

Though conveniently and commonly used (Chen et al., 2020; Cordeiro et al., 2021; Yang et al., 2018), the Otsu's method can hardly obtain the optimal threshold when facing complicated environments. Hence, we try to compute the optimal threshold for each test site by comparing the thresholding results with the reference map through traversal search, by which results with the highest Kappa coefficient can be generated. It should be noted that this optimal threshold is computed based on reference maps, so it is ideal and hardly practical for determining the threshold of WI because of the difficulty of predicting the optimal threshold in terms of accuracy and trade-off between precision and recall. However, it is still meaningful since it does reflect the ceiling

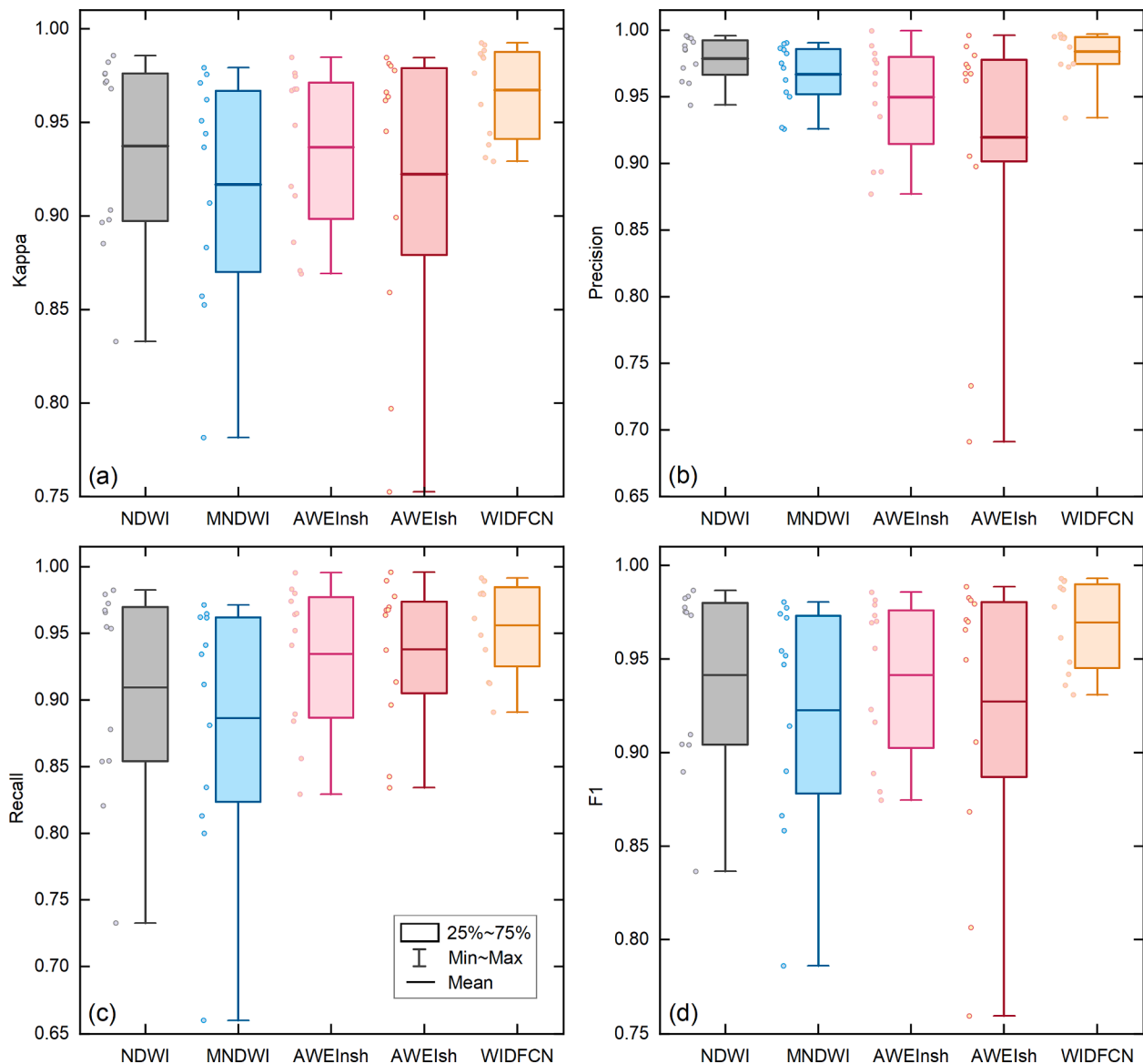


Fig. 5. Accuracies of WIDFCN compared with four WIs whose thresholds are calculated through traversal search for best Kappa coefficient. The dots distributed on the left of the boxes represent accuracies for each test site by the corresponding method.

of WI.

The accuracies of WIDFCN compared with that of four WI methods using optimal thresholds (WIs-Opti) is depicted in Fig. 5. Compared to thresholds from the Otsu's method in Section 4.1.2, WIs-Opti can achieve the trade-off between precision and recall, so their accuracy has astonishingly improved, which even surpasses those of the two unsupervised methods in Fig. 4. This phenomenon indicates the significance of determining a suitable threshold when using WI. In contrast, it does not need to consider the elaborate threshold setting in WIDFCN. For the training stage, the unique predetermined dual-threshold and ex-threshold are adopted for all scenes in the training set. Though simple and convenient when utilizing WI (NDWI in this study) in WIDFCN, the result from which is impressive and can achieve high precision and recall at the same time. As shown in Fig. 5, WIDFCN not only surpasses the WI-Opti methods by a large margin but also has higher values in terms of both precision and recall.

Furthermore, it can be seen in Fig. 5 that NDWI-Opti achieves the best accuracies among the four WI-Opti methods, reflecting that NDWI possesses the biggest application potential, which is the reason why using NDWI driven WIDFCN in this study.

4.1.4. Seasonal accuracy evaluation of WIDFCN

The test dataset covers different seasons as shown in Table 1, which

can preliminarily indicate the water delineation applicability of WIDFCN in different seasons. To further assess the temporal stability of WIDFCN, water delineation accuracies at the test site of Ulungur Lake in different months are calculated, in which the labeling method of random reference samples described in Section 2.3 is adopted for assessing accuracies. Due to the limitation of data acquisition and data quality, as well as the influence of snow cover, data from April to November are acquired to indicate the different seasonal conditions. The accuracies of WIDFCN in different months is shown in Fig. 6.

As shown in Fig. 6a, in April and May, the recalls of WIDFCN are relatively higher than the precisions, while it shows the opposite pattern from June to November. This can be attributed to the interference from many aquatic plants floating on the surface of water bodies. According to Fig. 6b–e, the water color changes in different months can be well responded by WIDFCN. The red areas in the water body may reflect the phytoplankton, as shown in Fig. 6d2 and e2. WIDFCN can deal well with the light red areas, with a high water delineation completeness. But in terms of the dark red regions in which too many phytoplankton exist, it is difficult to extract water completely. In general, WIDFCN shows its robustness to seasonal variation according to the stably high Kappa and F1, as shown in Fig. 6a, in which Kappa (F1) presents the variation range from 0.9832 (0.9859) to 0.9892 (0.9911).

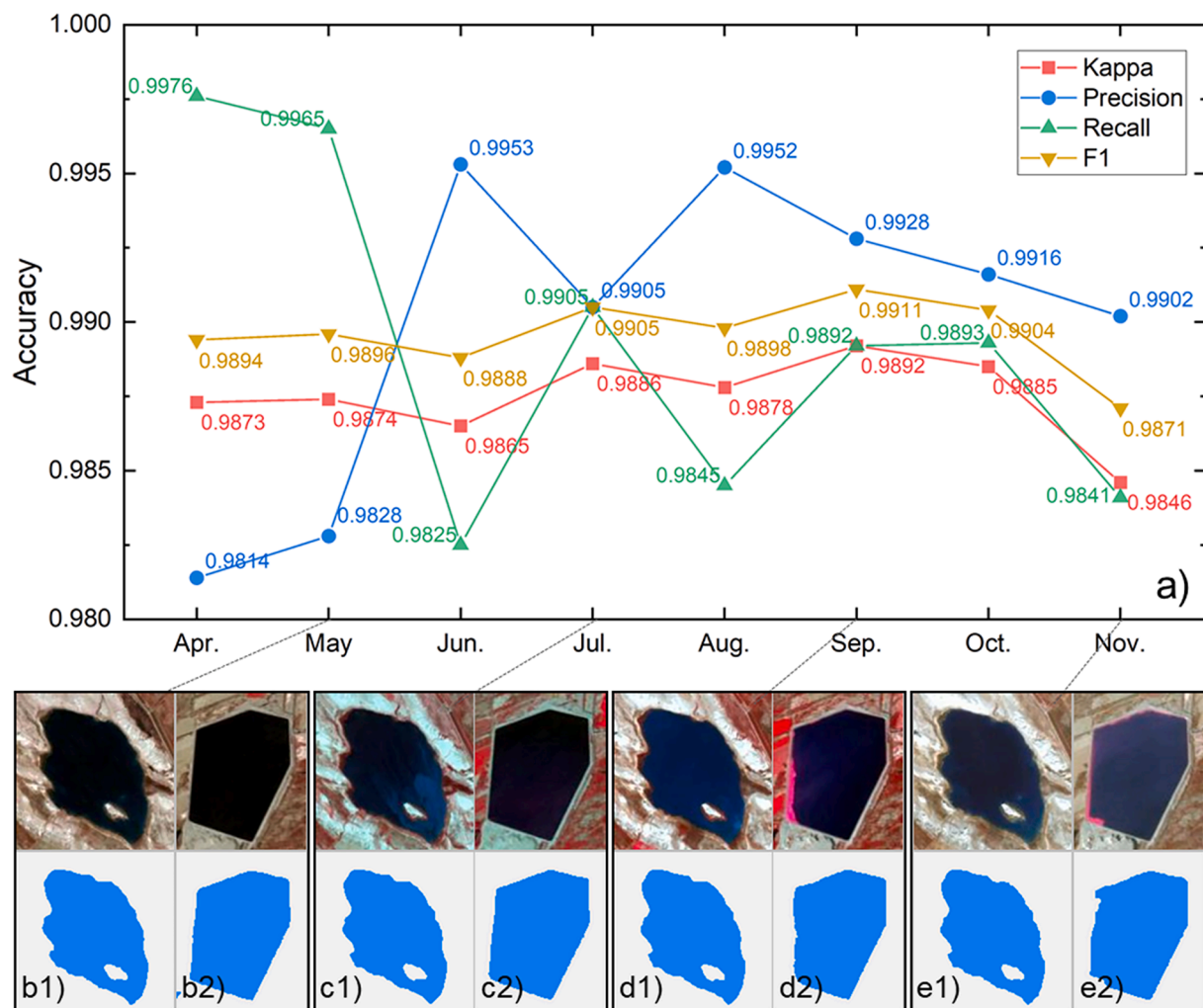


Fig. 6. Monthly accuracies and water delineation examples of WIDFCN. (a) represents accuracies of WIDFCN in different months. (b), (c), (d), and (e) represents the examples of water delineation in May, July, September, and November, respectively, in which the upper indicates Sentinel-2 images shown in false color composite of NIR, red, and green bands and the lower indicates the water delineation results by WIDFCN (blue for water and white for non-water). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.2. Performance of delineation on water with large variance

In this subsection, three test sites of Ulungur Lake, Paraguay River, and Great Bear Lake are selected for qualitatively and quantitatively demonstrating the performance of delineation on water with large variance, specifically, water with multiple scales and various spectra (indicating heterogeneity). For comparison, four methods are selected according to the analysis in Section 4.1, including ISODATA and RF, whose performance is better in the unsupervised and supervised classification methods, respectively, and MNDWI thresholding with Otsu's threshold (MNDWI-Otsu) and optimal threshold (MNDWI-Opti), in that MNDWI performs the best among the four WIs using Otsu's threshold.

The accuracies at the three test sites are shown in Table 2. The visual interpretation of the Ulungur Lake test site is shown in Fig. 7, and those of the Paraguay River and Great Bear Lake test site are presented in Fig. S1 and Fig. S2 in supplementary material. As shown in Table 2, all the five methods can obtain high-precision water delineations on the three test sites, so the accuracy depends more on the recall. For the Ulungur Lake test site, whose water bodies vary greatly in scale, WIDFCN can still generate good delineation, as shown in Fig. 7b. ISODATA also works well, but MNDWI and RF perform inferiorly in terms of recall.

The Paraguay River test site is composed of water bodies with large heterogeneity, presented as various colors, as shown in Fig. S1. Fig. S1b shows that the WIDFCN can cope well with the heterogeneity. ISODATA performs well in terms of recall, as shown in Table 2. MNDWI thresholding and RF still generate low recall, which is explicitly illustrated in Fig. S1f-h and Fig. S1l-n.

The Great Bear Lake test site is characterized by both spectral heterogeneity and scale variability, as shown in Fig. S2. For this region, ISODATA and RF perform inferior in terms of delineation completeness. MNDWI thresholding performs inferior in terms of precision due to the bright bare land.

Generally, WIDFCN is in a good position to extract multiscale water bodies and handle water heterogeneity, successively followed by ISODATA, RF and MNDWI thresholding. The high extraction completeness demonstrates its wide scope application capability for water delineation.

4.3. Resistance to different kinds of shadows

Commission error from the similarity between shadows and water is all along a great challenge for water mapping. In this section, three test sites of the Pearl River, Burrigorang Lake, and Manjira River with different kinds of shadows from buildings, mountains, and clouds are chosen to evaluate the shadow resistance of WIDFCN. The overall performance is shown in Table 3, with the water delineations visually shown in Fig. 8, Fig. S3 and Fig. S4.

Table 2

Accuracies of WIDFCN compared to ISODATA, RF, MNDWI-Otsu and MNDWI-Opti at three test sites of Ulungur Lake, Paraguay River, and Great Bear Lake.

Study area	Method	Kappa	Precision	Recall	F1
Ulungur Lake	WIDFCN	0.9844	0.9951	0.9792	0.9871
	MNDWI-Otsu	0.9207	0.9984	0.8760	0.9333
	MNDWI-Opti	0.9367	0.9853	0.9116	0.9470
	ISODATA	0.9726	0.9832	0.9713	0.9773
	RF	0.9541	0.9984	0.9274	0.9616
Paraguay River	WIDFCN	0.9763	0.9951	0.9613	0.9779
	MNDWI-Otsu	0.9434	0.9917	0.9064	0.9472
	MNDWI-Opti	0.9509	0.9751	0.9344	0.9543
	ISODATA	0.9735	0.9619	0.9892	0.9754
	RF	0.9586	0.9955	0.9295	0.9614
Great Bear Lake	WIDFCN	0.9854	0.9948	0.9802	0.9874
	MNDWI-Otsu	0.9440	0.9615	0.9423	0.9518
	MNDWI-Opti	0.9440	0.9632	0.9407	0.9518
	ISODATA	0.9592	0.9934	0.9378	0.9648
	RF	0.9604	0.9949	0.9383	0.9658

The accuracy gap between WIDFCN and other methods is enlarged to a great extent in Table 3 compared to the accuracies in Table 2, owing to the greater challenges from shadows. The performance of WIDFCN is still outstanding when facing this difficulty. For the Pearl River test site in which buildings and building shadows are complexly distributed, WIDFCN has good robustness to the noise of building shadows. ISODATA performs the worst because of the large commission error leading to low precision, which can be clearly illustrated in Fig. 8. For MNDWI, it seems that the delineation from MNDWI-Opti possesses a high precision according to Fig. 8m, in which the error detections (marked as red pixels) of MNDWI-Opti are fewer than WIDFCN. However, it should be noted that the missed detections (marked as yellow pixels) by MNDWI-Opti are also much more than WIDFCN. Combining the quantitative results in Table 3 in the manuscript, in which the recall of MNDWI-Opti (0.8103) is lower than that of WIDFCN (0.9377) in a large margin, it can be concluded that MNDWI is inferior than WIDFCN in terms of distinguishing shadows, because MNDWI-Opti has to set a high threshold and to sacrifice too much recall for high precision for the sake of the best Kappa. RF presents higher precision than WIDFCN, with the help of high-precision training samples, which instead leads to a low recall of 0.8329, as shown in Fig. 8n.

For the Burrigorang Lake test site in which the lake is surrounded by multiple mountain shadows, as shown in Fig. S3, WIDFCN can achieve a very high value in terms of not only precision but also recall. MNDWI performs well in precision but deficiently in recall, which is similar as the case of the Pearl River test site. ISODATA generates the lowest precision because of the commission errors on mountain shadows. RF can also achieve both high precision and recall, with a slightly lower recall than WIDFCN.

For the Manjira River test site, clouds and cloud shadows are the major noise, as presented in Fig. S4. MNDWI with the Otsu's threshold has poor performance, as shown in Table 3. It misclassifies not only cloud shadows but also clouds, because clouds also lead to a relatively high MNDWI value. With the optimal threshold, MNDWI can achieve higher precision (from 0.5243 to 0.9178) with the price of losing recall (from 0.9471 to 0.8062), which is similar to that of the Pearl River test site. For ISODATA, a large number of misclassifications of cloud shadows result in the low precision, as shown in Fig. S4e and k. Although RF can achieve high precision, its low delineation completeness greatly lowers the accuracy. WIDFCN can generate outstanding results with both high precision and recall, as shown in Fig. S4b, demonstrating its robustness toward clouds and cloud shadows.

In summary, WIDFCN has good resistance to shadows, assuring its precision of water delineation facing complicated shadow surroundings, successively followed by RF, MNDWI and ISODATA. RF can delineate water bodies with high precision due to the high-precision training samples, which also restricts its water delineation completeness. MNDWI can hardly achieve a trade-off between precision and completeness of water delineation when facing shadows or confused surfaces such as buildings and clouds, even with the ideally optimal threshold. Regarding ISODATA, although it can generate results with high recall, it has nearly no resistance to shadows of all kinds.

4.4. Thorough analysis of spectral index-driven learning

4.4.1. Effectiveness of spectral index-based training

As a WI-based method, it is necessary to know how NDWI influences model training and why that works. We compare the results from the semantic labeling models trained on water masks by applying different single NDWI thresholds with the results by directly applying the corresponding NDWI thresholds, which are shown in Table 4. It is noted that the threshold here is single, i.e., pixels with NDWI larger/lower than the threshold are taken as water/non-water. In addition, MFB loss is adopted when training to assure class balance. This comparison shows the accuracy differences caused by the FCN model, aiming to demonstrate the feasibility of this study, namely, the spectral index is able to drive

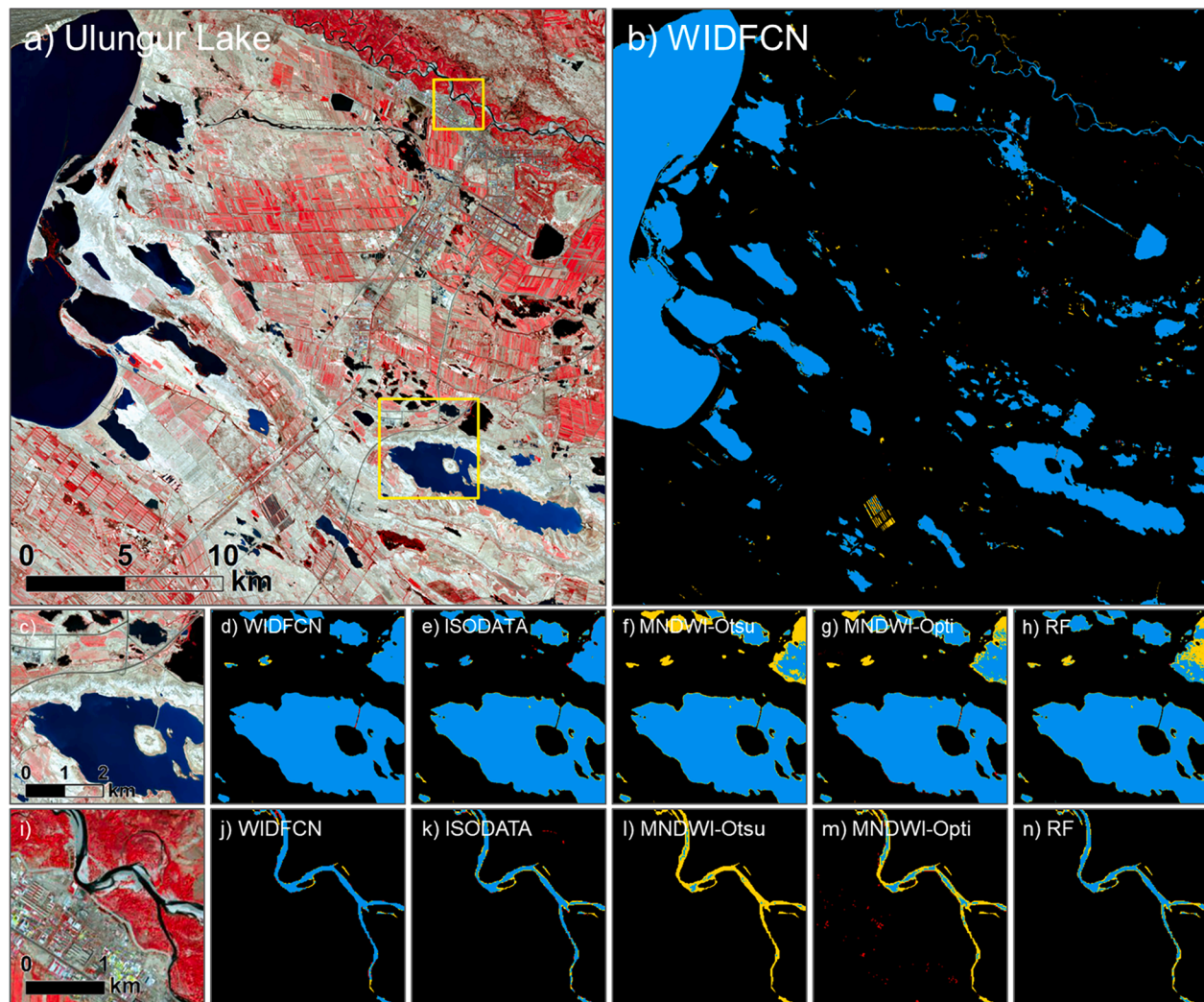


Fig. 7. Visual interpretation of WIDFCN compared with ISODATA, MNDWI-Otsu, MNDWI-Opti, and RF at the Ulungur Lake test site, whose water body is characterized by a multiscale distribution. The Sentinel-2 images (a, c, and i) are shown in false color composite of NIR, red, and green bands. The blue, yellow, and red in the water delineation results (b, d–h, and j–n) represent accurately detected water bodies, missed detections, and error detections, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3

Accuracies of WIDFCN compared to ISODATA, RF, MNDWI-Otsu and MNDWI-Opti at three test sites of the Pearl River, Burragorang Lake, and Manjira River.

Study area	Method	Kappa	Precision	Recall	F1
Pearl River	WIDFCN	0.9312	0.9341	0.9377	0.9359
	MNDWI-Otsu	0.8567	0.9204	0.8175	0.8659
	MNDWI-Opti	0.8572	0.9304	0.8103	0.8662
	ISODATA	0.7448	0.6372	0.9589	0.7657
	RF	0.8823	0.9550	0.8329	0.8898
Burragorang Lake	WIDFCN	0.9914	0.9947	0.9894	0.9920
	MNDWI-Otsu	0.9672	0.9991	0.9415	0.9694
	MNDWI-Opti	0.9757	0.9904	0.9647	0.9774
	ISODATA	0.9146	0.9296	0.9117	0.9206
	RF	0.9833	0.9961	0.9732	0.9845
Manjira River	WIDFCN	0.9596	0.9744	0.9486	0.9613
	MNDWI-Otsu	0.6561	0.5243	0.9471	0.6749
	MNDWI-Opti	0.8525	0.9178	0.8062	0.8583
	ISODATA	0.7671	0.6595	0.9507	0.7788
	RF	0.8923	0.9740	0.8304	0.8965

FCN training with no needs of manually labeling costs.

The spectral index-driven training is sensitive to labeling errors. As shown in Table 4, when setting the single threshold as -0.1 , the water extraction by NDWI possesses a high recall and a low precision. In this

case, the WIDFCN result is even worse than the NDWI result, because too many non-water samples are incorrectly viewed as water, and the errors lead to a poor delineation model.

However, the spectral index-driven training is able to improve the completeness by learning from relatively accurate labels. When NDWI thresholds increase from 0.0 to 0.3, the precision values of NDWI results are all relatively high, ranging from 0.9598 to 0.9983. Accordingly, the errors in the water masks are gradually reduced with the increasing thresholds, and the accuracies of WIDFCN trained by these masks are all better than those of their corresponding NDWI thresholding results, which shows the value of spectral index-driven training on improving the completeness. Additionally, as the precision of NDWI increases, the Kappa coefficients of NDWI-based training do not monotonically increase but increase first and then decrease, because the water masks are already very accurate but the amount of water sample continuously decreases when the threshold is fairly large, e.g., greater than 0.1. This indicates that to obtain an effective NDWI-based model, the precision and recall of training samples from NDWI should be trade-off.

4.4.2. Why using dual-threshold ?

The benefits of dual-threshold training compared to single-threshold training are illustrated in this subsection. Fig. 9 shows the accuracies of dual-threshold NDWI-based training (with MFB loss) compared with

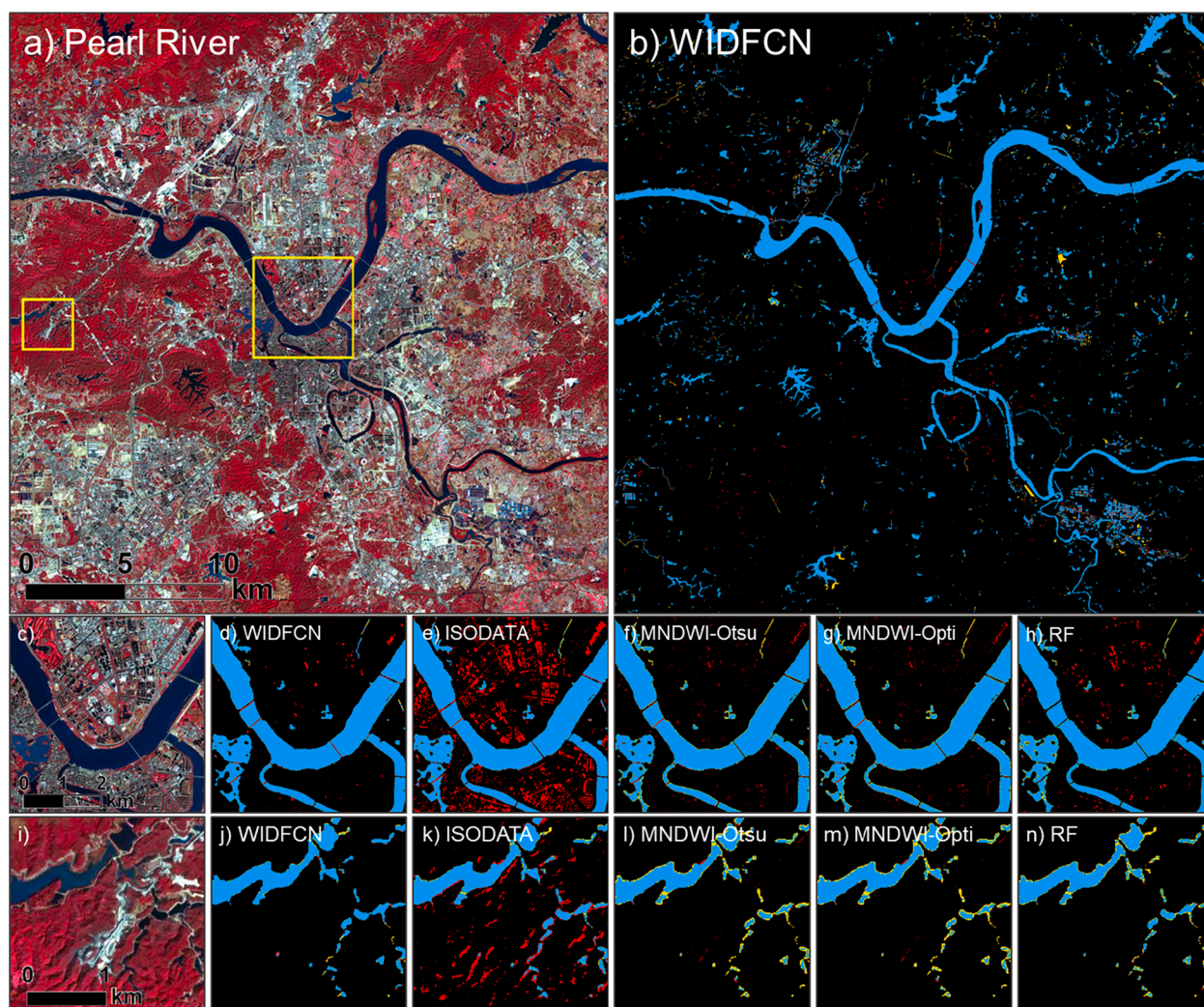


Fig. 8. Visual interpretation of WIDFCN compared with ISODATA, MNDWI-Otsu, MNDWI-Opti, and RF at the Pearl River test site, which possesses numerous buildings and building shadows. The Sentinel-2 images (a, c, and i) are shown in false color composite of NIR, red, and green bands. The blue, yellow, and red in the water delineation results (b, d–h, and j–n) represent accurately detected water bodies, missed detections, and error detections, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4

Accuracies of NDWI thresholding with different thresholds on the test sites compared with that of the WIDFCN results trained on the water masks generated by applying the corresponding NDWI thresholds on the training set.

Thre	NDWI thresholding				WIDFCN			
	Kappa	Precision	Recall	F1	Kappa	Precision	Recall	F1
−0.1	0.8713	0.8457	0.9392	0.8811	0.8584	0.8089	0.9740	0.8696
0.0	0.9247	0.9568	0.9077	0.9304	0.9426	0.9285	0.9683	0.9464
0.1	0.8920	0.9930	0.8354	0.9020	0.9622	0.9778	0.9530	0.9649
0.2	0.8068	0.9973	0.7339	0.8176	0.9529	0.9636	0.9496	0.9563
0.3	0.7003	0.9983	0.6108	0.7097	0.9381	0.9660	0.9218	0.9424

that of the corresponding single-threshold training. When using the water mask from the dual-threshold for training the model, the pixels whose NDWIs are between low threshold (lt) and high threshold (ht) will be seen as ignored pixels and do not contribute to loss computation. Three cases are shown here, in which ht is set as 0.1, 0.2 and 0.3. In these cases, lt is set as a unique value of 0, because that if lt is set as a high value, e.g. 0.1, there will be too many water pixels taken as background for the water mask, leading to water delineation with only low recall and thus undesirable accuracy. The single threshold training is identical to Section 4.4.1 by setting the thresholds from 0.0 to 0.3.

One can find in Fig. 9 that the dual-threshold training leads to a

better Kappa coefficient compared to the single threshold training with the same threshold. This is because labels are more accurate in the water mask from the dual-threshold. Ambiguous pixels are excluded for training models in case incorrect information is induced to the model. In contrast, the water mask from setting a hard single threshold may include noise for either water or non-water and disturb the model training. Generally, a lower single threshold tends to take more noise as water so that the precision is low yet the recall is high, while a higher single threshold will regard some water pixels as non-water so that the recall is low. However, the dual-threshold can avoid the above problems and lead to the better trained model.

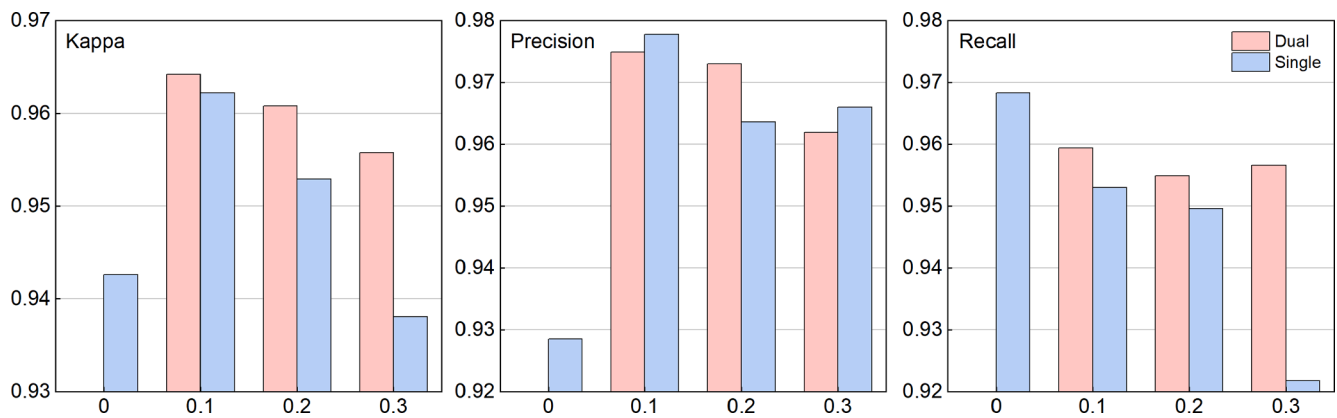


Fig. 9. Accuracies of training with water masks from dual-thresholds (dual) on NDWI compared to single thresholds (low and high). All It values are set as 0, and ht is set as 0.1, 0.2, and 0.3. The accuracies are calculated on the whole test dataset.

4.4.3. Influence of label expansion and class balancing

The dual-threshold is effective due to the less noise in the training samples, but the spectral features of the ignored pixels are not fully used so that they cannot make their contribution to the model training, which leads to water delineation with numerous omission errors. By filling up the ignored pixels and expanding water pixels in the water mask, as illustrated in Section 3.2.2, the water information will be more sufficient. For the base mask, ht is set as 0.3 to assure the precision of water mask, and It is set as 0 because we expect not to mistake water as non-water to assure a high precision of non-water samples.

As shown in the upper half part of Table 5, although the water delineation from WIDFCN trained directly using the base mask possesses a high precision, its delineation completeness is unsatisfactory (presented by the low recall). Compared with using the base mask for training the WIDFCN model, using the ex-mask can result in water delineation with a dramatic rise in recall and a slight increase in precision, which leads to a one percent higher Kappa coefficient from 0.9456 to 0.9577.

MFB loss considers the number of samples in each category and assigns higher weight to the category owing to the comparatively small sample volume, which plays a great role in such situations of extremely unbalanced samples for training FCN. Compared with using the base mask only for training WIDFCN, the import of MFB loss can alleviate the category imbalance, and the completeness can be greatly improved from 0.9227 to 0.9566, as shown in the upper half part of Table 5, leading to the rise of Kappa coefficient from 0.9456 to 0.9558. Although the ex-mask includes more water information than the base mask, it still faces the problem of class imbalance. Based on the ex-mask, the water delineation accuracy from WIDFCN can be further improved by considering class balancing, which leads to the best accuracy (Kappa coefficient 0.9673 and F1-score 0.9696).

Accuracies of SVM and RF trained by the base mask and the ex-mask are also calculated to further demonstrate the effectiveness of label expansion, as shown in the lower half of Table 5. When training the two

supervised classification models, all the water pixels (positive samples) in the base mask (or the ex-mask) are used, while the non-water pixels are randomly sampled to the total number of water pixels, aiming at ensuring the class balance and reducing the computer memory requirements. As shown in Table 5, compared with using the base mask, the ex-mask can improve the delineation accuracies for both SVM (Kappa 0.0233 higher) and RF (Kappa 0.0104 higher), demonstrating the effectiveness of label expansion. In general, SVM performs inferior than RF in terms of two masks, and both the two methods obtain lower accuracies than WIDFCN. This demonstrates the superiority of deep learning method for delineating water bodies compared to the classic supervised classification methods, even though this is not the focus of this study. Furthermore, both SVM and RF trained by the ex-mask can produce higher accuracies than other methods (as shown in Fig. 5), demonstrating the effectiveness of the ex-mask again.

5. Discussions

5.1. How does the spectral index drive DCNN training?

Spectral index is an approximate representation of the confidence of the ground feature except that it is noisy, such as shadows toward water. For WI, pixels with very high WI values are more likely to be pure water, and vice versa. Noise (e.g. shadows) comes mainly from the median range of spectral index values. Therefore, the base mask produced by the dual-threshold has relatively high precision, in which various pure water samples provide water features for learning and non-water samples provide discrimination ability.

Ambiguous ignored pixels can be roughly divided into three categories: noise with slightly high index values, water with moderate index values, and mixed pixels. To facilitate clarification, subsets of the final water mask from the training set and the corresponding water delineation of WIDFCN on the subsets are visually shown in Fig. 10. There are some shadows whose WI values are relatively low, regarded as negative samples, as shown in Fig. 10b1. In contrast, the number of shadows considered as positive samples is much lower. Therefore, the shadows in the ignored pixels are more likely to be judged as non-water by the trained model, as shown in Fig. 10c1, assuring the resistance of shadows for good precision.

Similarly, the water pixels in the ignored pixels have greater probabilities of being judged as water. Ignored water pixels (Fig. 10b2) possess water features that can be learned from the positive samples, while there are only a small quantity of pixels regarded as negative samples in the water mask. Hence, the ignored pixels can be accurately judged as water by WIDFCN, as shown in Fig. 10c2.

Table 5

Accuracies of water delineation from WIDFCN without (wo) or with (wi) MFB loss (the upper half) and that from SVM and RF (the lower half), trained by using the base mask (base) or ex-mask (ex).

Method/Mask	Kappa	Precision	Recall	F1
WIDFCN wo MFB/Base	0.9456	0.9788	0.9227	0.9494
WIDFCN wi MFB/Base	0.9558	0.9619	0.9566	0.9591
WIDFCN wo MFB/Ex	0.9577	0.9805	0.9425	0.9607
WIDFCN wi MFB/Ex	0.9673	0.9840	0.9561	0.9696
SVM/Base	0.8938	0.9835	0.8448	0.9036
RF/Base	0.9391	0.9846	0.9084	0.9432
SVM/Ex	0.9171	0.9676	0.8864	0.9236
RF/Ex	0.9495	0.9529	0.9540	0.9532

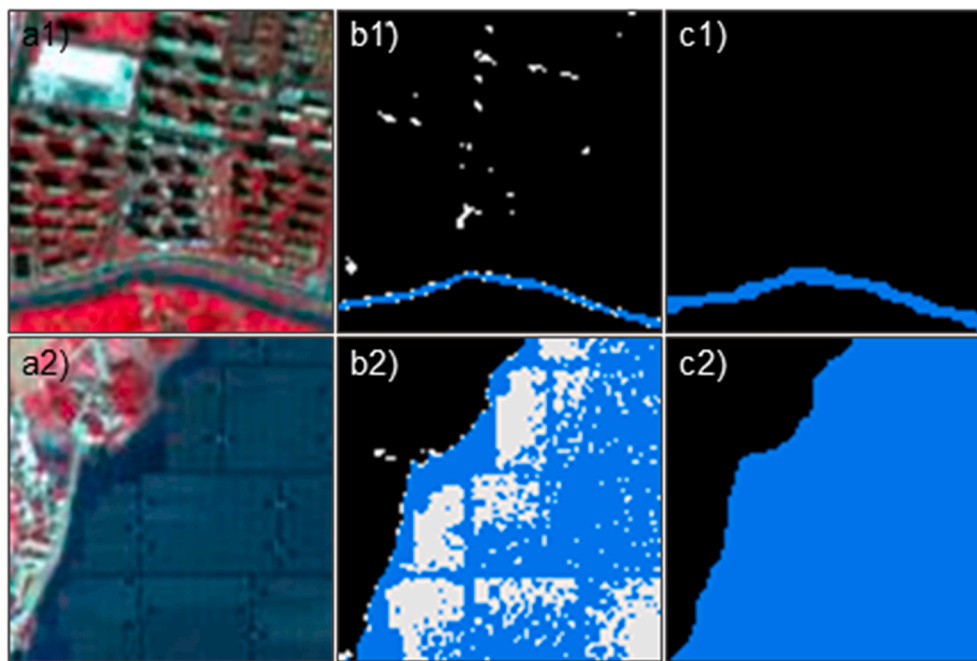


Fig. 10. Example of subsets on the training set. (a) Sentinel-2 images in false color composite of NIR, red, and green bands. (b) Final water masks of the subsets in which black, blue, and white represent non-water, water, and ignored, respectively. (c) Water delineations on the subsets using the trained model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.2. Learning process of the semantic labeling model

We visualize the outputs of the semantic labeling model during training to analyze the learning process of the model. Fig. 11 shows the visualization of the water channel of the softmax layer in different epochs, which represents the probability of belonging to water for each pixel. The images are from the test sites and with a size of 256×256 pixels.

As shown in Fig. 11b, at the beginning of training, the activation in the output heatmap is disorganized. After training for 0.01 epochs (more than 200 256×256 samples were used for training), the model has learned the basic ability to distinguish water bodies, as shown in Fig. 11c. However, the model right now is kind of ‘rough’, in that it can perceive not only rough outlines of water bodies but also noise of building, cloud, and mountain shadows. Further learning to epoch 0.1 makes the model able to perceive more accurate water bodies, which can

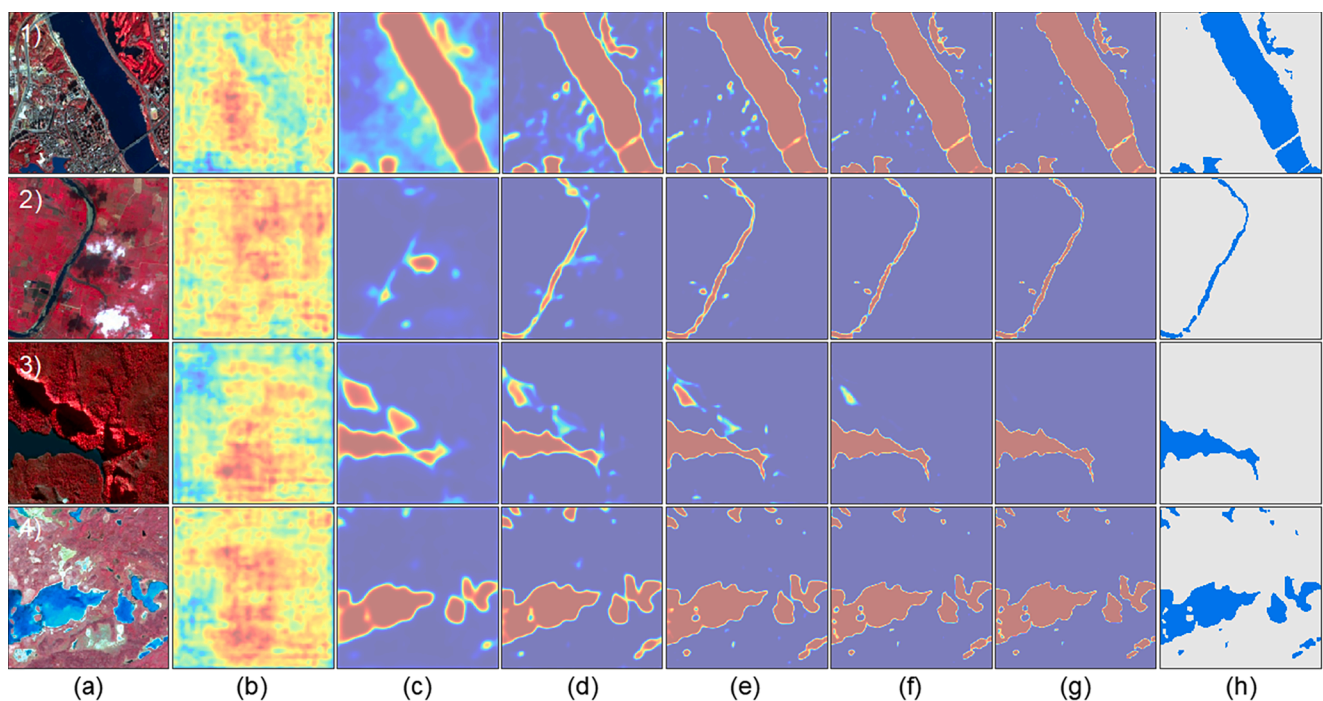


Fig. 11. Visualization of the softmax layer (water) on several test images (256×256 pixel) from different epochs in the training stage. (a) Sentinel-2 images in false color composite of NIR, red, and green bands. (b) Epoch 0. (c) Epoch 0.01. (d) Epoch 0.1. (e) Epoch 0.5. (f) Epoch 1. (g) Epoch 10. (h) Ground-truth. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

be clearly demonstrated in Fig. 11d2, and possess a preliminary ability to discriminate water from shadows, as shown in Fig. 11d1, d2, and d3. As the learning proceeds, the model can capture increasingly distinct water borders and be more discriminant to shadows, as shown in Fig. 11g.

5.3. Performance of MNDWI-driven WIDFCN

In this study, NDWI is utilized to drive the training of WIDFCN, the reason of which has been explicated in Section 4.1.3. However, it is still valuable to analyze the performance of WIDFCN based on different WIs, such as MNDWI, to understand the differences between WIDFCNs driven by different WIs. Hence, we conduct the MNDWI driven WIDFCN (WIDFCN-MNDWI) to compare with WIDFCN-NDWI in this subsection.

Accuracies of WIDFCN-MNDWI are shown in Table 6. In terms of the parameters, the lt, ht and ex-thre are set as 0, 0.5, and 0.2, respectively, which we find appropriate empirically. Compared to WIDFCN-NDWI, as shown in Table 5, the accuracies of WIDFCN-MNDWI are inferior in general, and both the precision and recall of WIDFCN-MNDWI are lower than those of WIDFCN-NDWI. To train a good FCN for water delineation, ideal background information extracted from WI should cover as much non-water information as possible and avoid water pixels at the same time, which is the main contribution to the shadow resistance ability of the trained model. However, the backgrounds extracted from MNDWI not only contain more water than that from NDWI, but also exclude less shadow information, which leads to a trained model with low precision. Besides, it is difficult for MNDWI to extract small water bodies with the ht, leading to a low delineation recall of the trained model as shown in Table 6.

To sum up, although WIDFCN-MNDWI performs better than other methods as shown in Fig. 4, its performance is inferior to WIDFCN-NDWI, which still needs further exploration.

5.4. Transferability and enlightenments

WIDFCN treats water delineation as a semantic labeling task, and the FCN model is trained by automatically collecting training samples without human labeling. The powerful capability of FCNs can be utilized, and the big-data-driven challenge is settled by the application of a water spectral index combined with appropriate training methods. However, this core idea is not necessarily limited to water extraction applications and has the potential to be applied to other geographical information. For example, other spectral indexes, such as vegetation, impervious area, and snow index, can potentially be used to drive the training of the FCN model.

Additionally, as a WI-based method, in addition to the advantage of no requirements of manually collecting training samples, it also avoids the subjectivity and uncertainty brought by setting thresholds. When applying this method for water extraction, the most convenient way is to directly train an FCN model with water masks generated by thresholding NDWI (or other WIs), as shown in Section 4.4.1, by which the trained model not only performs better than NDWI but also avoids the complexity of setting the optimal thresholds. Beyond that, the water extraction accuracy can be further improved by adopting several training techniques, such as setting a dual-threshold, using an ex-mask, and considering class balancing.

5.5. Limitations and future improvement

In this study, we proposed a novel solution for water delineation by utilizing the information provided by the spectral index for training a deep semantic labeling model, which avoids the manual labor of collecting training samples. However, it is difficult for FCN to deal with extremely fine-grade ground features due to its instinctive defect from which high-level features are extracted by down-sampling. Subtle water bodies could be lost during down-sampling and hardly recovered

Table 6

Accuracies of water delineation from WIDFCN-WNDWI trained by using the base mask and ex-mask with or without MFB loss.

Method	Kappa	Precision	Recall	F1
Base mask	0.9004	0.9567	0.8681	0.9072
Base mask + MFB	0.9219	0.9279	0.9305	0.9277
Ex-mask	0.9263	0.9485	0.9195	0.9315
Ex-mask + MFB	0.9403	0.9467	0.9444	0.9446

through up-sampling. The development of high-resolution FCN models could be useful for relieving this problem.

A second limitation is that WIDFCN can hardly extract water with plentiful sediment. For this case, the water feature may be greatly buried by the sediment, hindering the recognition of water by the model. To solve this problem, training samples for water with sediment should be purposively generated.

Additionally, it is hard to completely eliminate shadows. Though very resistant to shadows, there is still a little noise in the water delineation results. An appropriate noise-tolerant learning method may be a solution for this.

6. Conclusions

This study proposes a new water delineation method, named WIDFCN, which relies on no effort to manually collect samples. The problem is formulated as a semantic labeling task based on the training of an FCN, which allows multiscale spatial feature extraction and considers spectral features in multiple bands. As a powerful machine learning technology, FCN requires a large amount of training data and the corresponding ground-truths for a robust model, which requires high acquisition costs. We settle this problem by skillfully generating samples by utilizing the water index, e.g., NDWI in this study.

Twelve Sentinel-2 images with various challenges are chosen around the world to assess the performance of WIDFCN by comparison with several methods including WI thresholding, unsupervised and supervised classification methods. The results show that WIDFCN can achieve the best water delineation accuracy with Kappa coefficient 0.9673 and F1-score 0.9696 and the lowest fluctuation in terms of various test sites. This high-quality delineation accuracy comes from the excellent precision of 0.9840 and the comparative recall of 0.9561. Additionally, WIDFCN can deal with water with large variance and has resistance to different kinds of shadows, such as building, mountain, and cloud shadows.

In addition to the high precision and completeness performance, WIDFCN is convenient to employ because it avoids the subjectivity and complexity of determining thresholds when using water indexes for generating samples. This method also has the potential for high-accuracy large-scale water mapping due to the free costs of collecting abundant samples for training a robust model. This study demonstrates the feasibility of spectral index-driven DCNN model training for water information extraction. We expect the core idea of this study is not limited to the category of water and has the potential to be applied to information extraction for other ground surfaces.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 42071297, 41871235, 31860679), the Fundamental Research Funds for the Central Universities (Grant No. 020914380095),

and the High-level Innovation and Entrepreneurship Talents Introduction Program of Jiangsu Province of China.

Appendix A. Supplementary material

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.isprsjprs.2022.08.019>.

References

- Allen, G.H., Pavelsky, T.M., 2018. Global extent of rivers and streams. *Science* 361, 585–588.
- Bangira, T., Alfieri, S.M., Menenti, M., van Niekerk, A., 2019. Comparing Thresholding with Machine Learning Classifiers for Mapping Complex Water. *Remote Sens.* 11, 1351.
- Bui, Q.T., Nguyen, Q.-H., Nguyen, X.L., Pham, V.D., Nguyen, H.D., Pham, V.-M., 2020. Verification of novel integrations of swarm intelligence algorithms into deep learning neural network for flood susceptibility mapping. *J. Hydrol.* 581, 124379.
- Chai, D., Newsam, S., Zhang, H.K., Qiu, Y., Huang, J., 2019. Cloud and cloud shadow detection in Landsat imagery based on deep convolutional neural networks. *Remote Sens. Environ.* 225, 307–316.
- Chen, X., Chen, Y., Shimizu, T., Niu, J., Nakagami, K., Qian, X., Jia, B., Nakajima, J., Han, J., Li, J., 2017a. Water resources management in the urban agglomeration of the Lake Biwa region, Japan: An ecosystem services-based sustainability assessment. *Sci. Total Environ.* 586, 174–187.
- Chen, F., Chen, X., Van de Voorde, T., Roberts, D., Jiang, H., Xu, W., 2020. Open water detection in urban environments using high spatial resolution remote sensing imagery. *Remote Sens. Environ.* 242, 111706.
- Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2014. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *ArXiv14127062 Cs*.
- Chen, L.C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking Atrous Convolution for Semantic Image Segmentation.
- Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018a. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 834–848.
- Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018b. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), *Computer Vision – ECCV 2018*. Springer International Publishing, Cham, pp. 833–851.
- Chittoor Viswanathan, V., Jiang, Y., Berg, M., Hunkeler, D., Schirmer, M., 2016. An integrated spatial snap-shot monitoring method for identifying seasonal changes and spatial changes in surface water quality. *J. Hydrol.* 539, 567–576.
- Cordeiro, M.C.R., Martinez, J.M., Peña-Luque, S., 2021. Automatic water detection from multidimensional hierarchical clustering for Sentinel-2 images and a comparison with Level 2A processors. *Remote Sens. Environ.* 253, 112209.
- Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C., 2020. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm. Remote Sens.* 162, 94–114.
- Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., Bargellini, P., 2012. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sens. Environ.* 120, 25–36.
- Eigen, D., Fergus, R., 2015. Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-scale Convolutional Architecture. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 2650–2658.
- Feng, Q., Gong, J., Liu, J., Li, Y., 2015. Flood Mapping Based on Multiple Endmember Spectral Mixture Analysis and Random Forest Classifier—The Case of Yuyao. *China. Remote Sens.* 7, 12539–12562.
- Feng, M., Sexton, J.O., Channan, S., Townshend, J.R., 2016. 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.
- Feng, W., Sui, H., Huang, W., Xu, C., An, K., 2019. Water Body Extraction From Very High-Resolution Remote Sensing Imagery Using Deep U-Net and a Superpixel-Based Conditional Random Field Model. *IEEE Geosci. Remote Sens. Lett.* 16, 618–622.
- Feyisa, G.L., Meilby, H., Fensholt, R., Proud, S.R., 2014. Automated Water Extraction Index: A new technique for surface water mapping using Landsat imagery. *Remote Sens. Environ.* 140, 23–35.
- Fisher, A., Flood, N., Danaher, T., 2016. Comparing Landsat water index methods for automated water classification in eastern Australia. *Remote Sens. Environ.* 175, 167–182.
- Guo, H., He, G., Jiang, W., Yin, R., Yan, L., Leng, W., 2020. A Multi-Scale Water Extraction Convolutional Neural Network (MWEN) Method for GaoFen-1 Remote Sensing Images. *ISPRS Int. J. Geo-Inf.* 9, 189.
- Güttler, F.N., Niculescu, S., Gohin, F., 2013. Turbidity retrieval and monitoring of Danube Delta waters using multi-sensor optical remote sensing data: An integrated view from the delta plain lakes to the western–northwestern Black Sea coastal zone. *Remote Sens. Environ.* 132, 86–101.
- Hao, S., Zhou, Y., Guo, Y., 2020. A Brief Survey on Semantic Segmentation with Deep Learning. *Neurocomputing* 406, 302–321.
- Hu, R., Fang, F., Pain, C.C., Navon, I.M., 2019. Rapid spatio-temporal flood prediction and uncertainty quantification using a deep learning method. *J. Hydrol.* 575, 911–920.
- Huang, C., Chen, Y., Zhang, S., Wu, J., 2018. Detecting, Extracting, and Monitoring Surface Water From Space Using Optical Sensors: A Review. *Rev. Geophys.* 56, 333–360.
- Huber, C., Battiston, S., Yesou, H., Tinell, C., Laurens, A., Studer, M., 2013. Synergy of VHR pleiades data and SWIR spectral bands for flood detection and impact assessment in urban areas: Case of Krymsk, Russian Federation, in July 2012. In: 2013 IEEE International Geoscience and Remote Sensing Symposium - IGARSS, pp. 4538–4541.
- Isikdogan, F., Bovik, A.C., Passalacqua, P., 2017. Surface Water Mapping by Deep Learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 10, 4909–4918.
- Jia, K., Jiang, W., Li, J., Tang, Z., 2018. Spectral matching based on discrete particle swarm optimization: A new method for terrestrial water body extraction using multi-temporal Landsat 8 images. *Remote Sens. Environ.* 209, 1–18.
- Kampffmeyer, M., Salberg, A.-B., Jenssen, R., 2016. Semantic Segmentation of Small Objects and Modeling of Uncertainty in Urban Remote Sensing Images Using Deep Convolutional Neural Networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 680–688.
- Kiselev, V., Bulgarelli, B., Heege, T., 2015. Sensor independent adjacency correction algorithm for coastal and inland water systems. *Remote Sens. Environ.* 157, 85–95.
- Lary, D.J., Alavi, A.H., Gandomi, A.H., Walker, A.L., 2016. Machine learning in geosciences and remote sensing. *Geosci. Front.* 7, 3–10.
- Li, W., Hsu, C.-Y., Hu, M., 2021a. Tobler's First Law in GeoAI: A Spatially Explicit Deep Learning Model for Terrain Feature Detection under Weak Supervision. *Ann. Am. Assoc. Geogr.* 1–19.
- Li, Z., Zhang, X., Xiao, P., Zheng, Z., 2021b. On the Effectiveness of Weakly Supervised Semantic Segmentation for Building Extraction From High-Resolution Remote Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 14, 3266–3281.
- Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440.
- Marlow, D.R., Moglia, M., Cook, S., Beale, D.J., 2013. Towards sustainable urban water management: A critical reassessment. *Water Res. Urban Water Manage. Increase Sustainab. Cities* 47, 7150–7161.
- McFeeters, S.K., 1996. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *Int. J. Remote Sens.* 17, 1425–1432.
- Mishra, K., Prasad, P.R.C., 2015. Automatic Extraction of Water Bodies from Landsat Imagery Using Perceptron Model. *J. Comput. Environ. Sci.* 2015, 1–9.
- Nandi, I., Srivastava, P.K., Shah, K., 2017. Floodplain Mapping through Support Vector Machine and Optical/Infrared Images from Landsat 8 OLI/TIRS Sensors: Case Study from Varanasi. *Water Resour. Manag.* 31, 1157–1171.
- Noh, H., Hong, S., Han, B., 2015. Learning Deconvolution Network for Semantic Segmentation. In: Presented at the Proceedings of the IEEE International Conference on Computer Vision, pp. 1520–1528.
- Oki, T., Kanae, S., 2006. Global Hydrological Cycles and World Water Resources. *Science* 313 (5790), 1068–1072.
- Otsu, N., 1975. A Threshold Selection Method from Gray-Level Histograms. *Automatica* 11, 23–27.
- Palmer, S.C.J., Kutser, T., Hunter, P.D., 2015. Remote sensing of inland waters: Challenges, progress and future directions. *Remote Sens. Environ.* 157, 1–8.
- Pekel, J.F., Vancutsem, C., Bastin, L., Clerici, M., Vanbogaert, E., Bartholomé, E., Defourny, P., 2014. A near real-time water surface detection method based on HSV transformation of MODIS multi-spectral time series data. *Remote Sens. Environ.* 140, 704–716.
- Pekel, J.F., Cottam, A., Gorelick, N., Belward, A.S., 2016. High-resolution mapping of global surface water and its long-term changes. *Nature* 540, 418–422.
- Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241.
- Shen, C., 2018. A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists. *Water Resour. Res.* 54, 8558–8593.
- Sit, M., Demiray, B.Z., Xiang, Z., Ewing, G.J., Sermet, Y., Demir, I., 2020. A comprehensive review of deep learning applications in hydrology and water resources. *Water Sci. Technol.* 82, 2635–2670.
- Stehman, S.V., 2009. Sampling designs for accuracy assessment of land cover. *Int. J. Remote Sens.* 30, 5243–5272.
- Stehman, S.V., Czaplewski, R., 1998. Design and Analysis for Thematic Map Accuracy Assessment: Fundamental Principles. *Remote Sens. Environ.* 64, 331–344.
- Stehman, S.V., Foody, G.M., 2019. Key issues in rigorous accuracy assessment of land cover products. *Remote Sens. Environ.* 231, 111199.
- Sun, Y., Zhang, X., Xin, Q., Huang, J., 2018. Developing a multi-filter convolutional neural network for semantic segmentation using high-resolution aerial imagery and LiDAR data. *ISPRS J. Photogramm. Remote Sens.* 143, 3–14.
- Wang, Yi, Fang, Z., Hong, H., Peng, L., 2020a. Flood susceptibility mapping using convolutional neural network frameworks. *J. Hydrol.* 582, 124482.
- Wang, Y., Li, Z., Zeng, C., Xia, G.-S., Shen, H., 2020b. An Urban Water Extraction Method Combining Deep Learning and Google Earth Engine. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 13, 769–782.
- Wang, G., Wu, M., Wei, X., Song, H., 2020c. Water Identification from High-Resolution Remote Sensing Images Based on Multidimensional Densely Connected Convolutional Neural Networks. *Remote Sens.* 12, 795.
- Wang, X., Xie, S., Zhang, X., Chen, C., Guo, H., Du, J., Duan, Z., 2018. A robust Multi-Band Water Index (MBWI) for automated extraction of surface water from Landsat 8 OLI imagery. *Int. J. Appl. Earth Obs. Geoinformation* 68, 73–91.
- Wurm, M., Stark, T., Zhu, X.X., Weigand, M., Taubenböck, H., 2019. Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* 150, 59–69.

- Xu, H., 2006. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *Int. J. Remote Sens.* 27, 3025–3033.
- Yamazaki, D., Sato, T., Kanae, S., Hirabayashi, Y., Bates, P.D., 2014. Regional flood dynamics in a bifurcating mega delta simulated in a global river model: SIMULATION OF RIVER MEGA DELTA FLOWS. *Geophys. Res. Lett.* 41, 3127–3135.
- Yamazaki, D., Trigg, M.A., Ikeshima, D., 2015. Development of a global ~90m water body map using multi-temporal Landsat images. *Remote Sens. Environ.* 171, 337–351.
- Yang, Y., Liu, Y., Zhou, M., Zhang, S., Zhan, W., Sun, C., Duan, Y., 2015. Landsat 8 OLI image based terrestrial water extraction from heterogeneous backgrounds using a reflectance homogenization approach. *Remote Sens. Environ.* 171, 14–32.
- Yang, X., Qin, Q., Grussenmeyer, P., Koehl, M., 2018. Urban surface water body detection with suppressed built-up noise based on water indices from Sentinel-2 MSI imagery. *Remote Sens. Environ.* 219, 259–270.
- Yang, X., Qin, Q., Yésou, H., Ledauphin, T., Koehl, M., Grussenmeyer, P., Zhu, Z., 2020. Monthly estimation of the surface water extent in France at a 10-m resolution using Sentinel-2 data. *Remote Sens. Environ.* 244, 111803.
- Yue, K., Yang, L., Li, R., Hu, W., Zhang, F., Li, W., 2019. TreeUNet: Adaptive Tree convolutional neural networks for subdecimeter aerial image segmentation. *ISPRS J. Photogramm. Remote Sens.* 156, 1–13.
- Zhang, X., Feng, X., Xiao, P., He, G., Zhu, L., 2015. Segmentation quality evaluation using region-based precision and recall measures for remote sensing images. *ISPRS J. Photogramm. Remote Sens.* 102, 73–84.
- Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid Scene Parsing Network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6230–6239.
- Zou, Z., Dong, J., Menarguez, M.A., Xiao, X., Qin, Y., Doughty, R.B., Hooker, K.V., David Hambright, K., 2017. Continued decrease of open surface water body area in Oklahoma during 1984–2015. *Sci. Total Environ.* 595, 451–460.