Coarse-to-fine waterlogging probability assessment based on remote sensing image and social media data

Lei Xu\textsuperscript{a,b} and Ailong Ma\textsuperscript{a,b}

\textsuperscript{a}State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China; \textsuperscript{b}Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan, China

**ABSTRACT**

Urban waterlogging probability assessment is critical to emergency response and policymaking. Remote Sensing (RS) is a rich and reliable data source for waterlogging monitoring and evaluation through water body extraction derived from the pre- and post-disaster RS images. However, RS images are usually limited to the revisit cycle and cloud cover. To solve this issue, social media data have been considered as another data source which are immune to the weather such as clouds and can reflect the real-time public response for disaster, which leads itself a compensation for RS images. In this paper, we propose a coarse-to-fine waterlogging probability assessment framework based on multisource data including real-time social media data, near-real-time RS image and historical geographic information, in which a coarse waterlogging probability map is refined by using the real-time information extracted from social media data to acquire a more accurate waterlogging probability. Firstly, to generate a coarse waterlogging probability map, the historical inundated areas are derived from Digital Elevation Model (DEM) and historical waterlogging points, then the geographic features are extracted from DEM and RS image, which will be input to a Random Forest (RF) classifier to estimate the likelihood of hazards. Secondly, the real-time waterlogging-related information is extracted from social media data, where the Convolutional Neural Network (CNN) model is applied to exploit the semantic information of sentences by capturing the local and position-invariant features using convolution kernel. Finally, fine waterlogging probability map can be generated based on morphological method, in which real-time waterlogging-related social media data are taken as isolated highlight point and used to refine the coarse waterlogging probability map by a gray dilation pattern considering the distance-decay effect. The 2016 Wuhan waterlogging and 2018 Chengdu waterlogging are taken as case studies to demonstrate the effectiveness of the proposed framework. It can be concluded from the results that by integrating RS image and social media data, more accurate waterlogging probability maps can be generated, which can be further applied for inundated areas identification and disaster monitoring.

**ARTICLE HISTORY**

Received 12 May 2020
Accepted 16 August 2020

**KEYWORDS**

Remote sensing; social media; urban waterlogging; data fusion

1. Introduction

Since climate and environment change, flood disaster is becoming more frequent and irregular. As one of ten natural disasters, flood losses account for 34% of global disaster losses, which are much higher than that of other disasters such as droughts and fires. Meanwhile, with the progress of urbanization, the proportion of impervious surfaces in cities has been increased over the years (Weng and Lu 2008), which leads that urban drainage capacity is vulnerable to heavy or continuous precipitation. Therefore, urban waterlogging has the characteristics of suddenness, concentration, and destructiveness, causing injuries and property loss. How to evaluate urban waterlogging risk in a short time has become a widespread concern, which is critical to provide information for emergency response, thus minimize its negative effects and render valuable suggestions for policymakers.

As a long-range, large-scale, non-contact observation method, Remote Sensing (RS) can be applied to the monitoring of water bodies, forests, cities, and agriculture through the interpretation and classification of ground reflectance spectra. Intensive studies on waterlogging risk assessment using RS images have been conducted. Generally, flooded areas can be identified based on the difference of water body extraction results derived from the pre- and post-disaster RS images. The methods of water body identification from optical RS images can be divided into the single-band threshold, multiband spectral relationship, index calculation and classification method (Yang et al. 2011). A threshold was used to extract water body from RS images because of a strong absorption to water in the near-infrared and mid-infrared bands (Zhang et al. 2011). Since water reflectance in the green and red bands is higher than that of the near-infrared band and short infrared wave bands, a multiband spectral relationship algorithm was designed to identify water
body from Landsat TM/ETM+ (Yang et al. 2011). Meanwhile, the Normalized Difference Water Index (NDWI) was intensively used for remote sensing of vegetation liquid water from space (Gao 1996), and a series of modified water indexes were proposed based on NDWI (Xie et al. 2016; Xu 2006; Feyisa et al. 2014). Moreover, taking water as a category, the classification methods such as decision tree, Random Forest (RF) are also effective tools for water body detection (Mosavi, Ozturk, and Chau 2018; Jony et al. 2018; Feng, Liu, and Gong 2015; Sun, Yu, and Goldberg 2011). Although RS images can provide large-scale surface information and great progress has been made in flood monitoring, it is still limited by the quality of images and revisit cycle of satellites. To solve the above limitations, more waterlogging probability methods were put forward by considering geographic data from the ground as a compensatory, including Digital Elevation Model (DEM), streaming gauge recordings, precipitation and so on, which focus on the elevation in three-dimensional space more than the two-dimensional surface. DEM was proved to be a useful data source to identify the flood that occurred under the forest canopies, which cannot be extracted from optical RS images (Wang, Colby, and Mulcahy 2002), and it was also integrated with SAR images to generate the flooded areas at the peak time (Brivio et al. 2002). Meanwhile, to overcome the limitation of cloud cover, an algorithm integrating the images from Advanced Very High-Resolution Radiometer (AVHRR) and DEM was proposed for classifying the cloud covered pixels (Islam and Sado 2000). Furthermore, data including hydrology, water dynamics, meteorology, and urban sewerage systems could be taken into consideration for a more accurate flooded area mapping besides RS images. However, more complex hydrodynamic models need to be used involving the geographic data from the ground. In general, RS images may not meet the need of time resolution and quality in case of urban waterlogging probability estimation, which leads that the utilization rate of images is generally low. By adding the geographic data such as DEM and precipitation, the limitations mentioned above can be addressed partly, otherwise, accompanied by the unavailability of data source and the complex of methods.

Data from multiple social networks such as Sina Weibo and Twitter, which can track public focus and response, have been proved to be an emerging and valuable datasource for emergency response such as flood mapping (Ilieva and McPhearson 2018). To exploit the disaster-related social media data, the Convolutional Neural Networks (CNN) method was proved to be efficient in the heavy rainfall monitoring (Li et al. 2017), in the meantime, a multi-modal deep learning approach was proposed to extract related information not only from the text but also from photos (Lopez-Fuentes et al. 2017). Next, the usability of social media data for flood risk assessment was examined, meanwhile, the spatial relationship between social media data and catchments was analyzed, too (Smith et al. 2017; De Albuquerque et al. 2015). As for the utilization of social media data for flood mapping, the public response to flood disaster was studied and the active users were identified applying social network analysis (Xu 2015; Tyshchuk et al. 2012). Furthermore, intensive research has proved that the vital information from the ground is helpful to improve the situation awareness for flood disaster management (Herfort et al. 2014; Foresti, Farinosi, and Vernier 2015; Yin et al. 2012), thus the potential events could be detected using methods such as clustering, probabilistic topic extraction, and so on (Xu, Sugumaran, and Zhang 2015; Chae et al. 2012). Moreover, based on the theoretical research above, in recent years, a series of disaster visualization and events detection systems for Internet users have been developed, including “Did You Feel It?” (Atkinson and Wald 2007; Wald et al. 2012), “Toretter” (Sakai, Toriumi, and Matsuo 2011), “Twitcident” (Abel et al. 2012), “TweetAct” (Chowdhury et al. 2013), “CrisisTracker” (Rogstadius et al. 2013), “Ushahidi platform” (Okolloh 2009), “EARS” (Robinson, Power, and Cameron 2013a), “Twitter Earthquake Detector” (Robinson, Power, and Cameron 2013b), “Emergency Situation Awareness” (Power et al. 2014), and so on, which provide stable and reliable data streams from public response for disaster management. In a word, intensive research has proved that massive disaster-related information is involved in social media data which are real-time and easily obtained. Otherwise, social media data are usually presented as points and the geographical range cannot cover the whole study area, leading themselves inappropriate to assess the global flooded probability.

Social media data which regard citizens as local sensors, not only are immune to external factors such as clouds and fogs but also have the characteristics of real-time and fine-grained spatial distribution, which leads themselves compensation for RS images in space and time resolution. It is proved that the limitations of satellites can be overcome partially with the crowdsourced data by fusing these two data sources for the flooded areas mapping on a daily basis (Panteras and Cervone 2018). A data-driven approach was proposed for flood mapping, in which the Bayesian statistical model was used to quantify the contribution of each data source, including social media data, RS images and geographic data (Rosser, Leibovici, and Jackson 2017). Meanwhile, a flood inundation reconstruction model was proposed by fusing post-disaster RS imagery, real-time water gages data, and social media data which were manually verified to be disaster-related (Huang, Wang, and Li 2018a). To estimate the damage
to transportation infrastructure, a kernel-based method was used to generate the flood risk mapping by fusing the RS images and multiple sources of contributed data (Cervone et al. 2016). Moreover, to analyze the waterlogging probability between cities, a multi-view discriminant transfer learning method was proposed by fusing social media data, precipitation, road network, and DEM (Zhang et al. 2016). In the above research, usually, water body was detected from the post-disaster imagery and then fused with real-time contributed data, ignoring the historical geographic data which providelop knowledge about inundated-prone areas. On the other hand, it is crucial to determine the weight of each data source, which can be adjusted using a kernel interpolation; however, this method is sensitive to the bandwidth parameter, which causes the result less robust.

To solve the above problems, this paper proposes a coarse-to-fine waterlogging probability assessment framework based on multisource data. The main contributions of this paper are listed as follows:

1. A coarse-to-fine waterlogging probability assessment framework based on multisource data. In this framework, the multi-temporal data can be broken into historical, near real-time and real-time data. A Coarse Waterlogging Probability (CWP) map can be generated through historical geographic data and near real-time images. Historical inundated areas are generated from historical waterlogging points and DEM, which can be regarded as prior knowledge for detecting inundated-prone areas, and will be input to the next probability prediction model. Since the characteristics of large-scale, near real-time RS images are still useful information. ACWP map can be generated based on the above two data sources using RF in which the waterlogging probability of each pixel is calculated thus CWP is stored in raster format. Fine Waterlogging Probability (FWP) maps can be generated from the CWP and social media data. The multitemporal social media data (daily, hourly even minutely) can be compensation for RS images, which provides finer emergency information. In our proposed framework, real-time waterlogging-related social media data are taken as isolated highlight point and used to refine the CWP map by a gray dilation pattern considering the distance-decay effect, which eventually generates the FWP maps.

2. Information extraction from multisource data for waterlogging probability assessment. Disaster-related information or feature extraction can be divided into two parts, which are actually parallel in the proposed framework. In order to solve the lack of real-time information and the limitation of external factors of RS images, we exploit the geographic features on a large scale from RS image and DEM, meanwhile, the real-time emergency information from social media data, which can be concluded as following: (1) geographic feature extraction from multisource geographic data. To derive the CWP which can also be taken as an annual assessment for waterlogging probability, DEM, and historical waterlogging points are taken as sources of the prior information for inundated-prone areas, the related geographic features including slope, roughness of land surface and Fractional Vegetation Cover (FVC) are exploited from RS image and DEM considering water flow and accumulation as the direct indicators about waterlogging; (2) deep learning-based information extraction from real-time social media data. In general, the strategy for collecting social media data through Application Programming Interface (API) or crawler technology can be generated as “specific time + true location + right keywords.” However, keywords filter is a very rough method to extract the emergency information, ignoring the semantic information of social media text, which may lead that the false disaster-related information is involved. In order to solve this problem, in this paper, text classification method based on deep learning is used to further extract the emergency information. By encoding natural words as vectors, the local semantic meaning of a specific sentence matrix can be caught by a CNN-based network, which is helpful to identify the real-time waterlogging-related social media data with coordinates.

3. A morphology-based fusion method for point and raster data. As mentioned above, the CWP map is stored in raster and the real-time social media data point. To fuse the heterogeneous data, a morphology-based fusion method with Tobler’s First Law of Geography is used in this paper. Social media data generate the pixels need to be refined and the distance between social media data and pixel decides the refining degree, thus an improved gray dilation algorithm is applied based on the above theory.

The rest of the paper is organized as follows. The case and datasets we used are introduced in Section 2. Then we demonstrate the methodology of the proposed coarse-to-fine waterlogging probability assessment framework, in which three modules including preliminary
definitions and basic theory are presented in Section 3. We present the experiment design, results and analysis in Section 4 and conclude the paper in Section 5.

2. Case study and datasets

2.1. Case study 1: Wuhan

In order to verify the practicability of the coarse-to-fine waterlogging probability assessment framework, this paper presents a case of urban waterlogging of Wuhan in July 2016 based on the exploitation of RS image, DEM, and contributed data from Sina Weibo platform. Wuhan, the capital of Hubei province, is a city located in the central of China. As a city with 166 lakes, the water area of Wuhan is 2 117.6 km², accounting for 25.01% of Wuhan’s total area.

According to the data in 2017 from Wuhan Municipal Water Authority, the total precipitation of Wuhan was 1813.4 mm in 2016, which accounted for 46% more than the multi-year average value and took the second place since 1956. Especially in the summer of this year, the total precipitation reached 510.2 mm, which was much higher than that in other months. Figure 1 presents the trend of the precipitation in Wuhan city from 2009 to 2016 with an average value of 1240.6 mm from 1965 to 2010. According to the recordings of Wuhan National Basic Meteorological Station, the precipitation has accumulated to 580 mm from 20:00 on June 30th to 16:00 on July 6th, 2016, which accounted for 44% of the total of annual precipitation. During the disaster, at least 27 people lost their lives and the economic losses reached ¥5.7 billion (about 850 USD million). The red alerts for heavy rainfall were issued by authorities on July 2nd and July 6th, 2016, respectively.

Figure 2 presents the location and the river system of the study area in Wuhan. It is worth noting that our study area is the urban districts of Wuhan rather than the whole city. There are two main reasons why we choose the urban districts as the study area. One is that the urban districts are the hardest-hit areas, which will endanger more lives and cause higher economic losses. The other is that the number of contributed geotagged Weibo in urban districts is much larger than that in other districts. Using the real-time waterlogging-related Sina Weibo data as compensation for RS images and geographic data, our goal is to gain a more accurate waterlogging probability assessment and present the dynamic change of the spatial distribution of waterlogging probability during the emergency. Meanwhile, we will analyze the importance of each feature which contributes to urban waterlogging.

To illustrate the methodology and framework, multi-temporal sources of geographical data, RS images, and contributed data source were gained and used in this research. A summary of the metadata including collected

Figure 1. Trend of precipitation in Wuhan (2009–2016).

Figure 2. Study area in Wuhan.
time, sources and spatial resolution is available in Table 1. According to the collected time, the datasets we used during our research can be divided into four categories, which include real-time data, near real-time data, and historical data. Of course, accessary datasets including administrative division, river system are also collected.

**2.1.1. Real-time dataset**

The waterlogging disaster in Wuhan urban districts lasted more than a week and the red alerts were released by authorities on July 2nd and July 6th, so the geotagged Sina Weibo data were collected each day from June 30th to July 10th. Generally, the social media data should satisfy the time, location and keywords restraints. That is to say, the search tragedy for social media data is that the geotagged Weibo should be located inside the disaster-affected area, released during the emergency and at least one of the manually preset keywords should be included in the text. The real-time Weibo data used in this paper are provided by Jun Li’s research, which were collected by API provided by the Sina platform (Li et al. 2017). Sina platform provides access to geotagged Weibo centered on a given location within a specific radius. By calculating the coordinates, a total number of more than 8000 social media data located inside the urban districts could be downloaded. Figure 3(a) shows the geospatial distribution of social media data collected, which is centralized in the urban districts.

**2.1.2. Near real-time RS image**

The near real-time RS image was collected from Operational Land Imager (OLI) of Landsat 8, whose revisit cycle is 16 days. A series of RS images were checked and the multispectral image from OLI on July 23rd is the earliest cloud-free image provided by the United States Geological Survey (USGS). OLI has a 30-m spatial resolution in multispectral bands except that the resolution of band 8 is 15-m. Figure 3(b) presents the true color image of the urban districts by synthesizing band 2, 3, 4, which represent the blue-, green-, and red-band, respectively.

**2.1.3. Historical dataset**

Historical data can be broken into two parts, including historical waterlogging points and digital elevation model. In our research, historical waterlogging points are used to generate the inundated-prone areas, which can be regarded as prior information in waterlogging probability assessment. About 141 waterlogging points of 2013 were gathered from authority websites and official covers, and the spatial distribution is presented in

![Figure 3](image-url)
Figure 3(c), in which the points are mainly centralized in Wuchang, Jianghan, and Qiaokou districts. DEM is used to exploit terrain features, which affect the water flow and accumulation. The DEM at the same resolution as Landsat imagery was acquired from the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Digital Elevation Model (GDEM), whose geographic range can cover all land areas between 83°N and 83°S, reaching 99% of the Earth’s land surface. Figure 3(d) shows the DEM of the urban districts in Wuhan, in which the higher terrain is marked as red and the lower blue.

2.1.4. Accessory dataset
Other accessory data including city, country boundaries which were acquired from local authorities and the river system which was generated from RS image of Landsat 8 satellite at 30-m resolution.

2.2. Case study 2: Chengdu
In order to verify the effectiveness of social media data in our proposed framework, a case of Chengdu waterlogging in July 2018 was presented where each social media data were manually labeled and selected. Chengdu, the capital of Sichuan province, is a city located in the southwest of China. According to the Chengdu municipal water authority (2019), the average precipitation of Chengdu was 1434.2 mm in 2018, which accounted for 48.8% more than the multiyear average value and increased by 51.5% over the 2017’s. According to Chengdu National Basic Meteorological Station, the heavy rainfall attacked Sichuan province in July 2018 and the total precipitation was much higher than that in other months, which led to the occurrence of waterlogging in summer, causing huge economic losses. There were totally five waterlogging disasters of varying severity and the yellow alert was issued firstly on July 11th.

Similar to the case of the 2016 Wuhan waterlogging, our study area only covers the old urban districts of Chengdu rather than the whole city because of the higher risk which causes severe losses and the concentration of people who generate distributed Weibo data. Meanwhile, the dataset of 2018 Chengdu waterlogging used during our research can also be divided into four categories, including real-time data, near real-time data, historical data, and accessory data. Metadata such as collected time, sources, and spatial resolution are presented in Table 2. The two datasets used in our paper are from almost the same source, thus a briefer introduction will be given below.

2.2.1. Real-time dataset
The real-time social media data of Chengdu dataset were collected each day from July 1st to July 31st and the “specific time + true location + right keywords” restraints were applied, too. The function named “advancedsearch” of the Sinaplatform was used, in which the location was set as Chengdu and the keywords included “flood,” “rainstorm,” “rising water,” and so on to filter out irrelevant information. Social media data were manually checked and labeled in content and geographical location. Figure 4(a) shows the geospatial distribution of social media data collected, which is centralized in the urban districts.

2.2.2. Near real-time RS image
The near real-time RS image was collected from OLI of Landsat 8, too. After checking the images from Landsat 8, the multispectral image on June 15th is the most suitable cloud-free image with a 30-m resolution. Figure 4(b) presents the true color image of the urban area including Jingjiang, Qingyang, Jinniu, Wuhou, and Chenghua districts.

2.2.3. Historical dataset
Historical waterlogging points and digital elevation model are involved in this part. Figure 4(c) shows the spatial distribution of 101 waterlogging points of 2016 obtained from authority websites and official covers. The DEM at the same resolution as Landsat imagery was acquired from the Geospatial Data Clouds, which was processed based on the dataset of ASTER GDEM. Figure 4(d) shows the DEM data of the urban districts in Chengdu, in which the higher terrain is marked as red and the lower blue.

2.2.4. Accessory dataset
Other accessory data such as city, country boundaries were acquired from local authorities.

### Table 2. Dataset of 2018 Chengdu waterlogging.

| Data type            | Data                      | Collected time | Spatial resolution | Source                        |
|----------------------|---------------------------|----------------|--------------------|-------------------------------|
| Real-time data       | Sina Weibo                | 2018.7.1–2018.7.31 | –                 | Sina                          |
| Nearest Real-Time Data | Landsat 8                | 2018.6.15      | 30 m               | USGS                          |
| Historical data      | Waterlogging points       | 2016           | –                  | Official websites and covers  |
| Accessory data       | DEM                       | 2009           | 30 m               | Geospatial Data Clouds        |

3. Coarse-to-fine waterlogging probability assessment framework based on multisource data
To fuse multitemporal data introduced above, as Figure 5 shows, a coarse-to-fine waterlogging...
probability assessment framework based on multi-source data is proposed, in which three parts can be included: a CWP model based on RS and GIS data, emergency information extraction from social media data and a FWP model refined by social media data. Firstly, historical inundated areas which are generated from historical waterlogging points and DEM are used as prior information for inundated-prone areas. Then the geographical features extracted from DEM and RS image are input to a random forest classifier to estimate the waterlogging probability pixel by pixel, whose output is one of the mid-products, a CWP map. Meanwhile, to filter the noise information involved in social media data, a CNN model is used to extract the semantic information from social media text, whose output is another mid-product, multitemporal real-time and waterlogging-related social media data. At last, to refine the CWP with real-time waterlogging-related social media data, a morphology-based fusion method considering the first law of geography is applied to fuse the heterogeneous data of point and raster, whose final output will be the FWP maps chronologically, from which a daily dynamic change of waterlogging probability can be monitored and high-risk areas can be identified.

### 3.1. CWP model from multisource geographic data

CWP model is the first step of our main framework, in which the geographic information from the near real-time image, historical waterlogging points and DEM can be exploited fully and the result also could be regarded as an annual waterlogging probability map. The whole model could be broken into three parts including inundated areas generating, geographic feature extraction based on multisource data and waterlogging probability prediction based on geographic features.

#### 3.1.1. Inundated areas generating

Historical inundated areas can be taken as prior knowledge of inundated-prone areas, thus they will be helpful to choose training samples for the RF classifier, which will be used to predict the waterlogging probability for each pixel. Therefore in this paper we proposed an object-oriented geometric extraction method, in which the specific positions from historical waterlogging points and the elevation information from DEM are integrated to extract the historical inundated areas. Firstly, the rough DEM segmentation units were obtained easily through an object-oriented segmentation method. The second step is to judge whether each unit is inundated or not. A unit is judged to be inundated when at least one waterlogging point is inside a unit. The historical inundated areas can be identified quickly by repeating the above procedures.

#### 3.1.2. Geographic feature extraction based on multisource data

It is universally known that the elevation and material of land surface will decide whether the water will accumulate or not, thus a series of indexes are going to be exploited based on DEM and RS image, from which the parameters related to ground cover can be generated. In the following text, we will introduce the
method for topographical features extraction from DEM and FVC extraction from RS image.

Topographical features are features that can reflect the distribution of water accumulation and flow direction. The slope is the steepness of the surface, which is calculated by the ratio of the vertical height to the distance in the horizontal direction. The range of slope is 0 to infinity, in which the value of the flat surface is 0- and the 45-degree surface 100%, meanwhile, as the surface gets closer to vertical, the percentage becomes larger. The formula of the slope is presented as the following:

\[ \tan \alpha = \frac{h}{l} \times 100\% \quad (1) \]

where \( \tan \alpha \) represents the slope of a pixel and \( \alpha \) is the angle, \( h \) is the vertical height and \( l \) the distance in the horizontal direction. Roughness is an indicator for surface morphology which represents the ratio of the ground surface area to the projected area in a specific area, and the roughness is defined as follows:

\[ M = \frac{1}{\cos \alpha} \quad (2) \]

in which \( M = \frac{1}{\cos \alpha} \) represents the roughness for one pixel and the formula for \( \cos \alpha \) is defined as:

\[ \cos \alpha = \frac{S_{AC}}{S_{AB}} \quad (3) \]

where \( S_{AC} \) is the projected area of the pixel and \( S_{AB} \) the surface area. Considering the formula of the slope, the formula of roughness can be defined as the following:

\[ M = \frac{1}{\cos(\tan \alpha \times \pi/180)} \quad (4) \]

Figure 5. A coarse-to-fine waterlogging probability assessment framework based on multisource data.
The FVC is the proportion of the area covered by vegetation to that of the land surface, which can provide the quantitative information for the land permeability and the accumulation of water. The FVC is defined as the normalized form of the NDVI (Normalized Difference Vegetation Index), thus its range is 0–1. The formula of FVC is as following (Carlson and Ripley 1997):

\[ FVC = \frac{NDVI - NDVI_{\text{min}}}{NDVI_{\text{max}} - NDVI_{\text{min}}} \]  

(5)

where NDVI quantifies vegetation by measuring the difference between near-infrared which vegetation strongly reflects and red light which vegetation absorbs (Goward et al. 1991):

\[ NDVI = \frac{NIR - R}{NIR + R} \]  

(6)

in which NIR represents the reflection value of the near-infrared band and R that of the red band.

### 3.1.3. Waterlogging probability prediction based on geographic features

Based on the geographic features extracted from DEM and RS image, a prediction model can be established to estimate the waterlogging probability for each pixel, whose output is the CWP map. Interventional studies have testified that the ensemble classifier is an efficient prediction algorithm that performs better than a single classifier. The main focus of ensemble learning is to solve the inherent defects of a single classifier by aggregating multiple classifiers. The ensemble learning classifier builds a set of base classifiers from the training data and the final classification result can be generated by voting on the results of base classifiers. To achieve a more accurate classification, on the one hand, the base classifiers should be accurate enough, which is to guarantee the accuracy of the final results, on the other hand, there must be differences between the base classifiers, which means that the different features can be extracted and utilized from data sources.

According to the way in which the base classifiers are ensembled, the ensemble learning can be roughly divided into two categories, boosting, and bagging (Breiman 1996). “Boosting” is a serialization method that must be generated serially since there is a strong dependency between base classifiers. In contrast, providing the independency between base classifiers, “bagging” is a parallel optimization method that can be generated simultaneously, and RF is a typical method of bagging, which is intensively used in classification or regression. The RF classifier uses a series of decision trees as base classifiers, which are generated from a random dataset sampled independently from the whole input dataset, and the final class is defined to be the most popular result by voting among the results derived from each tree (Breiman 2001). There are mainly three different algorithms for generating decision trees including ID3, C4.5, and CART, which use information gain, information gain ratio and Gini Index as indicator for attribute selection, respectively. With Gini Index as the indicator for attribute selection in RF, for a given training set T, selecting one case (pixel) at random and the probability that it belongs to some class \( NDVI = \frac{NIR - R}{NIR + R} \), the Gini index can be written as (Pal 2005):

\[ \sum_{j \in T} \sum_{i \in T} \left( \frac{f(c, T)}{|T|} - \frac{f(c, T)}{|T|} \right)^2 \]  

(7)

where \( \sum_{j \in T} \sum_{i \in T} \left( \frac{f(c, T)}{|T|} - \frac{f(c, T)}{|T|} \right) \) is the probability that the selected case belongs to the class \( C_i \).

RF classifier can handle data of very high dimensions easily and avoid overfitting. Furthermore, an advantage of the tree-based integration algorithm is that the relative importance of the features used by the model can be output after the model training, which is convenient for us to evaluate the role of each variable in classification and understand which factor has a key influence on the result. In RF, simply, the closer a feature is to the average distance from the root among all trees, the more important this feature is in a specific classification or regression problem. Using the Gini coefficient to select the feature that splits down, the importance score of the feature \( C_i \) on a certain node \( m \) can be calculated as follows:

\[ VIM_{jm}^{(\text{Gini})} = CL_m - (CIL_m + CIR_m) \]  

(8)

where \( CL_m \) is the Gini coefficient of feature \( C_{IL} \), before splitting, \( GIL_m \) and \( GIL_m \) are the Gini coefficients of the left and right branches after splitting, respectively. Considering the number of nodes in a tree, the importance score of a feature in one certain tree can be represented as follows:

\[ VIM_{ij}^{(\text{Gini})} = \sum_{m \in M} VIM_{jm}^{(\text{Gini})} \]  

(9)

where \( M \) is the set that includes the nodes where the feature \( X_j \) was split in the decision tree \( X_i \), thus the importance score of a feature in random forest is defined as:

\[ VIM_{ij}^{(\text{Gini})} = \sum_{i=1}^{n} VIM_{ij}^{(\text{Gini})} \]  

(10)

where \( n \) denotes the number of trees in a random forest, to estimate the contribution of feature \( m \), the feature importance scores are normalized to obtain the importance score:

\[ VIM_j = \frac{VIM_j}{\sum_{i=1}^{n} VIM_i} \]  

(11)
3.2. Emergency information extraction from social media data

Massive disaster-related information is involved in social media data, thus intensive research has been conducted in which the disaster-related information was extracted from social media data using Natural Language Processing (NLP) method. Sentence classification is one of the important issues in the field of NLP. Generally, traditional sentence classifiers such as Support Vector Machine (SVM) and Naive Bayes use the “bag-of-words” model to represent the natural sentences, in which only the frequency of a specific word is considered rather than the sequence information. Recent studies have shown that neural network-based sentence classification algorithms can make full use of the contextual information and achieve better results in sentence classification (Nie et al. 2020). As one of the representative algorithms of deep learning, CNN is a type of feedforward neural networks and can capture local position-invariance features, which is suitable for social media text classification.

The most representative research about sentence classification with CNN is the Text-CNN model proposed by Yoon Kim of New York University, in which the multiple kernels of different sizes are applied to extract key information, therefore it can capture the local semantic meaning of sentences (Kim 2014). Furthermore, a sensitivity analysis of convolutional neural networks for sentence classification was presented, in which suggestions about tuning parameters were given (Zhang and Wallace 2015). Figure 6 shows the CNN model for sentence classification proposed by Yoon Kim. Firstly, the input layer is a matrix representing a sentence or document which is the concatenation of the word vectors, usually these word vectors can be encoded by not only the word embeddings derived from word2vec or GloVe model but also one-hot representation especially in the case of a small dataset. Assuming there are $n$ words in the sentence and the dimension of word vector is $k$, thus the sentence can be represented as a $n \times k$ matrix. In the meantime, the input has two channels including static and nonstatic methods. The static method indicates that the word vector does not change with the training process, in which a fixed-dimension encoding such as one-hot can be used. In contrast, the nonstatic method means the word vector is dynamic and can be taken as an optimizable parameter during the training process. Secondly, several feature maps with a column number of 1 can be obtained through convolution operation, assuming the size of the convolution window is $h \times k$, where $h$ is the number of words. In the pooling layer, max pooling is used to select the maximum value from

![Figure 6. CNN model for sentence classification.](image-url)
the feature map which is considered to be the most important feature after convolution. Finally, the output of the last step is concatenated together to form a single feature vector then the softmax layer is generated through the softmax function regularization, which can be set as the probability distribution on the final category. Actually, tricks such as dropout and L2 regularization are used on the fully connected layer to avoid overfitting.

The classification performance can be evaluated with precision, recall, and F-values. Precision is denoted as the ratio of true positives to the sum of true positives and false positives. Recall means the fraction of true positives in the sum of true positives and false negatives. F-value is a comprehensive indicator which is defined as the weighted harmonic average of precision and recall. The definitions of Precision, Recall, and F-value are given by the following formula:

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

\[
F-value = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

where TP, FP, and FN denote the number of true positives, false positives, and false negatives, respectively.

3.3. FWP model refined by social media data

For fusing the CWP map and real-time disaster-related social media data, a morphology-based fusion method for point and raster data is proposed. Social media data can be taken as compensatory information for the CWP layer, meanwhile, the method should also satisfy the universal spatial distribution of geographical entities. The real-time waterlogging-related social media data are taken as isolated highlight points which will refine the CWP map, thus pixels surrounding the social media data could be modified for the purpose of a FWP. According to the existing research, it is reasonable to assume that the fusion method follows a morphological dilation pattern and the dilation layer \(d(x, y)\) can be defined as following (Jackway and Deriche 1996):

\[
d(x, y) = (CWP + e)(x, y) = \max \{\text{CWP}(x - a, y - b) | (x - a, y - b) \in D_e; (a, b) \in De\}
\]

where \((x, y)\) denotes the location of a specific pixel, \(e\) denotes the structuring element, and \(D_e\) means a domain with a search radius centered at a geotagged Weibo.

Secondly, according to Tobler’s First Law of Geography (Miller 2004), pixels which are closer to a real-time waterlogging-related social media data are more likely to be inundated, thus a gray dilation pattern considering the distance-decay effect is used in our paper. At a pixel \((x, y)\), the FWP\((x, y)\) is calculated by introducing a distance-related coefficient \(c\) to the dilation layer and it can be defined as following (Huang, Wang, and Li 2018b):

\[
FWP(x, y) = \frac{[d(x, y) - \text{CWP}(x, y)]}{c} + \text{CWP}(x, y)
\]

(16)

where \(c\) is defined as following:

\[
c = \frac{r}{r - d}
\]

(17)

in which \(r\) denotes the radius of the search area and \(d\) denotes the distance from \((x, y)\) to the geotagged Weibo, and \(d\) will be assigned with zero when social media data are located in pixels, ignoring the distance-decay effect.

4. Experiments and discussion

4.1. Experiment design

To demonstrate the effectiveness of the proposed framework, the 2016 Wuhan waterlogging was taken as a case study, in which three parts will be illustrated in the following. Firstly, a CWP map is generated using RF classifier, meanwhile, the relationship between the spatial distribution of related features and inundated probability is analyzed. Secondly, the semantic information about emergency response is extracted from social media data with the Text-CNN model, from which the real-time waterlogging-related social media data are selected and labeled as “pos.” Last but not least, the FWP can be generated using a morphology-based data fusion algorithm considering the distance-decay effect. Furthermore, to verify the effect of social media data and the generalization of our framework, the experiment was performed on the 2018 Chengdu waterlogging dataset where the real-time and waterlogging-related social media data were checked manually, thus the second part of Text-CNN model will be ignored.

In this part, we will verify the effectiveness and robustness of the proposed framework comparing with a universal method for data fusion. The baseline is the kernel-based model using the RS images and social media data, in which the inundated areas are exploited through the difference of pre- and post-disaster RS images, then a kernel density interpolation is applied based on the two data sources. The kernel density function can be taken as the extension of the
histogram, in which the data range is divided with an equal interval and the probability is calculated as the ratio of the number of data in each group to the total number. The kernel estimator can be defined as the following form:

\[ f(x) = \frac{1}{n h} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \]  

where \( n \) is the number of neighborhood pixels within a specific distance from the target pixel \( x \), \( h \) is the smoothing parameter bandwidth, which controls the weight given to the pixels \( X_i \) surrounding pixel \( x \) based on their proximity. \( K \) means the kernel function and it must satisfy the following condition:

\[ \int_{-\infty}^{\infty} K(x)dx = 1 \]  

In general, there are two main parameters to be adjusted, one of which is the kernel function which reflects the spatial distribution pattern of the neighborhood pixels around the target pixel \( x \) and affects the shape rather than the degree of smoothing, the other one is the bandwidth \( h \) which mainly decide the final result, thus the kernel-based assessment model is most sensitive to the bandwidth. The waterlogging probability can be calculated by performing a kernel density interpolation on the two-dimensional data and then normalizing it pixel by pixel. For fusing different data sources, different bandwidth values were given, respectively. Narrower bandwidth means the larger weight for the data source, thus a narrower width will be given to authoritative data sources such as RS images, Unmanned Aerial Vehicles (UAV) and DEM since the credibility and accuracy.

All the experiments were implemented in Python and a uniform coordinate system was set for dataset to demonstrate in ArcGIS. In the process of generating a CWP map, a 3 \( \times \) 3 square window was used to extract the geographical features taking eight neighborhoods into consideration, the RF classifier was set with 10 estimators and the minimum number of samples required to split an internal node was 2. As mentioned before, a CNN model could identify the real-time and waterlogging-related text information from crawled social media data and it is essential to choose appropriate hyperparameters. The dimension of the sentence matrix was defined as 140 since at most 140 words could be published once on the SinaWeibo platform. We set the comma-separated filter sizes as 3, 4, 5, the dimensionality of character embedding as 128, the number of filters per filter size as 128, the dropout keep probability as 0.5, the batch size as 64 and the number of training epoch as 200. For fusing the above two data sources, we assumed the search radius of social media data as 1000m for Wuhan and 2000 m for Chengdu datasets in this paper and a square window with a size of 3 \( \times \) 3 was used to execute the morphology-based fusion method. Meanwhile, the search radius in the kernel-based model was also set as 1000 m, and the output resolution kept the same as the input data.

4.2. Experiment on Wuhan dataset

4.2.1. CWP model from multisource geographic data

A CWP map was generated in this part using the geographic features extracted from DEM and RS image as input to random forest classifier, which also could be taken as an annual waterlogging probability map. All the geospatial data used in this part were projected to a universal coordinate system which was WGS 1984 UTM Zone 49 N, at the same time, they had the same resolution at 30-m, thus there were totally 1.13 million grids in the study area. The spatial distribution of elevation, slope, roughness of DEM and FVC is presented in Figure 7.

The accuracy of the RF algorithm for predicting the waterlogging probability reached 98.1%, in which the historical inundated areas derived from historical waterlogging points and DEM were used as the training area. Figure 8 shows the CWP map of the study area with an interval of 0.2, meanwhile, the tree-based integration algorithm can generate the relative importance as Table 3 shows, thus it is convenient for us to discuss and analyze the contribution of each feature to the waterlogging probability prediction.

The average of waterlogging probability is 0.36, with a standard deviation of 0.19. It can be observed from Table 3 that the FVC contributes a lot to waterlogging probability and DEM takes the second place. The very low waterlogging probability areas are mainly located in the southeast of the study area, which are covered by forest and have a high elevation. Furthermore, more detailed figures are presented in Figure 8. Taking the “area a” as an example, the high waterlogging probabilities are mainly located in areas which are located inside the main ring roads and have low vegetation cover. This suggests that the high-risk areas are mainly concentrated on the residential and commercial districts, which are mainly covered by impervious layer. Similarly, the low-risk areas such as the “area b” are mainly mountainous, which have a high vegetation cover and high elevation. Therefore, FVC that decides whether the water will accumulate can be a critical indicator for assessing waterlogging probability.

4.2.2. Emergency information extraction from social media data

To exploit the emergency information involved in social media data, a CNN-based sentence classification method was used in our paper. When collecting the social media data, the collecting strategy whenever using API or crawler technology could be concluded as “specific time +
true location + right keywords,” in which the “specific time” means the published time of Weibo data should be during the disaster, the “true location” means its positioning should be located inside our study area, and the meaning of “right keywords” is that at least one preset keyword must be included in the text of social media data. However, keywords filter is a very rough method for the text of social media data, ignoring the semantic
information of the text, which may lead that the false disaster-related information is involved. For example, “Finally the rain stopped” or “it does not rain anymore” is not related to the rainfall although the keyword “rain” is involved. Another example of “it is pouring last night” in which the raining day is described can be given, although this sentence is related to the disaster, it still can not reflect the real-time information. Based on the examples above, the social media data must be real-time and disaster-related, which will be labeled as “pos.”

In the Wuhan dataset, a total number of 6945 social media data were collected in the urban districts of Wuhan from June 29th to July 10th, the number of samples used for training was 1461, in which there were 697 positive labels and 764 negative labels. After a series of processing like removing non-Chinese characters (html, json, number, emoji, URL and so on), Chinese encoding, word cut, removing stop-words and so on, the final training accuracy of Text-CNN for our dataset is 79.5%. Meanwhile, we find that there is a situation that will influence the model accuracy, where the tag rather than the following content is related to waterlogging. A comparative experiment was set to evaluate the negative influence of tags, and the accuracy of Text-CNN model for the same dataset is 75.7% when keeping tags, which is about 4% lower than that when removing tags.

Figure 9 demonstrates the number of classification result of each day during the disaster. Totally, the trend of occurrence of “pos” social media data is basically similar to that of “neg” social media data. It is notable that there are two peaks on July 2nd and July 6th, which corresponds to the actual case. Actually, the rainfall started on June 30th and became heavy in the next weeks. Wuhan authorities issued the red alerts on July 2nd and 6th, respectively. Especially on July 6th, the number of waterlogging-related points is much bigger than that of other days, which means the heavier rainfall leads to more serious public concerns, thus more useful information can be collected through the social media platform.

To demonstrate the trend of waterlogging-related social media data, Figure 10 shows the heat maps of heavy rainfall events that occurred in the urban districts of Wuhan form June 30th to July 10th, with an interval of two days. A uniform classification system rather than the Jenks natural breaks classification method is used since the convenience for observing and comparing the trend of waterlogging probability assessment. There is an increase of waterlogging probability from June 30th to July 2nd, and then a decrease on July 4th. However, a huge increase of waterlogging probability shows on July 6th since the heavy rainfall led the extensive inundation in the urban areas, which once caused traffic paralysis. Luckily most of the impounded water was evacuated in two days, and there is a distinguished decrease on July 8th and almost no inundation area till July 10th.

### 4.2.3. FWP model refined by social media data

FWP maps can be generated from the CWP map refined by social media data using a morphology-based fusion method with Tobler’s First Law of Geography. Social media data are taken as compensatory data to solve the limitations of RS images in quality. Therefore a series of FWP maps can be generated on a daily basis, in which the distance between the pixel and its nearest social media data is calculated. The pixels where the social media data are located have the distance variable assigned with zero meters.
Figure 10. Heat maps of heavy rainfall events occurred in the urban districts of Wuhan from June 30th to July 10th.

Figure 11. FWP maps in the urban districts of Wuhan from June 30th to July 10th.
Figure 11 shows the FWP maps in the urban districts of Wuhan from June 30th to July 10th, with an interval of two days. Similarly, a uniform classification system where the range is equally divided into five categories was used for the convenience of observing and comparing the trend of waterlogging probability assessment. It can be observed that an increase of waterlogging probability occurs on July 2nd, and then a decrease on July 4th. Next, there is a huge increase on July 6th when the heavy rainfall led the extensive inundation in the urban areas. Then there is a distinguished decrease on July 8th and almost no inundation area till July 10th. The day of July 6th can be taken as a representative of this disaster since its seriousness and destructiveness. By refining the CWP map with social media data on a daily basis, the waterlogging probability can be monitored chronologically, which just illustrates the processing of waterlogging probability assessment from coarse to fine in time scale.

To verify the correctness of our proposed framework, Figure 12 can be used for visual verification which presents the overlay of FWP map and true waterlogging points on July 6th, 2016 in the urban districts of Wuhan. Moreover, the affected roads are also labeled in the red line. It can be seen that the spatial distribution of true waterlogging points is consistent with that of the FWP map, which is in accord with the expectation and practical situation. Furthermore, the detailed diagram in Figure 12 shows the area located in Hongshan district, where the spatial distribution of not only the waterlogging points but also the affected roads are in accordance with the FWP map.

To verify the effectiveness of the proposed framework, in the meantime, a kernel-based model was tested as a baseline using the same dataset, in which different bandwidths were manually set for different data sources and a normalized kernel estimation was applied to fuse the multiple data sources. Social media data which are represented as points can be used for kernel interpolation directly, however, the inundated areas should be exploited from RS images. Based on the analysis above and the actual situation in Wuhan, the day of July 6th can be seen as a representative of the disaster, so the following experiments were performed only using the social media data on July 6th.

The near real-time RS image on July 23rd and the real-time social media data on July 6th were used to test the efficiency and robustness of the kernel-based model. The water body was extracted from RS image using the NDWI index, then the waterlogging probability could be generated by performing a normalization for kernel density interpolation result on the two-dimensional data. Figure 13 shows the kernel density interpolation results generated from only the water body derived from Landsat imagery and social media data, respectively, which are in totally different spatial distribution patterns.

According to the advice from previous research (Cervone et al. 2016), when choosing the appropriate parameters, the authoritative data source should be assigned with a narrower bandwidth which means the use of data from the initial stage of the disaster.
Figure 13. Kernel interpolation results.

Figure 14. Kernel-based fusion results of different bandwidths.

the larger weight and otherwise a wider one. Thus a series of bandwidths were set to evaluate the practicality of the kernel-based fusion method. As Figure 14 shows, in our experiment, the band widths of water body derived from RS image and social media data for kernel interpolation were set to 1:1, 1:1.5, 1:2, and 1:2.5, respectively. The waterlogging probability of urban districts in Wuhan city changes a lot as the ratio of bandwidth decreases. Thus it is hard to control the parameters setting and evaluate its practicality, which leads less robust for the model in case of urban waterlogging probability assessment.

There are two main reasons to explain why the kernel-based model lack practicality in the case of urban waterlogging probability assessment. On one hand, it can be observed from the CWP map that the waterlogging risk is independent of the distance from the river. Specifically, the occurrence of waterlogging in the urban area was not related to proximity to the nearest water channels and density of the hydrographic network per unit area. On the other hand, even if the catchment areas of the urban city should be derived from the pre – and after – RS images before inputting to the kernel-based model, the truth is the accumulated water was drained when the last cloud-free RS image can be obtained since the revisit cycle of satellites. Compared to the kernel-based fusion results of different bandwidths represented in Figure 14, obviously, the FWP map demonstrates a finer result for the whole study area. Furthermore, the river system in FWP map is labeled as very low-risk area, however, that in kernel-based fusion results is labeled as very high-risk area, which means the water body derived from RS image is a kind of distractive information in case of urban waterlogging probability assessment, thus the practicality of the coarse-to-fine waterlogging probability assessment framework is verified once again.

4.2.4. Discussion

(1) Performance analysis on topology type: a case of road

Road is one of the urban infrastructures easily affected by waterlogging disaster and directly related to traffic flow, thus it can be taken as a typical topology type in case of waterlogging risk assessment. By overlaying the road networks with FWP layer, a risk level of the roads located within the pixels can be acquired. Figure 15 shows the main roads which are predicted to
be affected by the waterlogging, in which the road networks in study area are marked with gray, the potential affected roads are marked with yellow, and the red points represent the true waterlogging points. Generally, the spatial distribution of potential affected roads is coupled with that of the true waterlogging points, proving the road assessment can be effectively performed using the FWP layer.

(2) Effectiveness of social media data

To prove the effectiveness of social media data quantitatively, a set of ablation experiments was performed. Figure 16 presents the number of pixels in the coarse and FWP map with an interval of 0.2, which are distributed in different patterns. Meanwhile, Table 4 shows the number of pixels and true waterlogging points with an interval of 0.2 in CWP and FWP. We can learn from the comparison between two tables that 0.6 can be taken as a threshold. The high-risk (>0.6) area of FWP can cover 25.5% of true waterlogging points more than that of CWP, especially in the range from 0.6 to 0.8 where the proportion is 24.9%. This indicates that CWP can be refined by fusing social media data.

4.3. Experiment on Chengdu dataset

The case of 2016 Wuhan waterlogging above has illustrated the effectiveness and robustness of our proposed framework for urban waterlogging probability assessment and monitoring, and the aim of this part is only to further verify the effectiveness of social media data and the generalization of the framework. Another case of 2018 Chengdu waterlogging includes the same kinds of datasets except that the real-time and waterlogging-related social media data were checked and labeled manually, thus semantic information extraction using Text-CNN model will be ignored.

Similarly, the features reflecting the water flow and accumulation were extracted from RS image and DEM, which would be input to random forest

---

**Table 4. Number of pixels and true waterlogging points in CWP and FWP of Wuhan.**

| Range  | No. of pixels | No. of true waterlogging points | No. of pixels | No. of true waterlogging points |
|--------|---------------|---------------------------------|---------------|---------------------------------|
| 0–0.2  | 113,532       | 6,337 (3.7%)                    | 85,465        | 2,123 (2.5%)                    |
| 0.2–0.4| 280,093       | 9,182 (3.2%)                    | 212,231       | 4,214 (2.0%)                    |
| 0.4–0.6| 290,912       | 10,069 (3.4%)                   | 321,363       | 6,093 (1.9%)                    |
| 0.6–0.8| 107,317       | 10,093 (3.7%)                   | 176,856       | 9,453 (5.3%)                    |
| 0.8–1.0| 19,209        | 4,214 (2.4%)                    | 15,148        | 5,148 (3.5%)                    |
| Total  | 811,063       | 165,176 (20.4%)                 | 811,063       | 165,176 (20.4%)                 |

**Figure 16. Number of pixels in waterlogging probability map of Wuhan.**
classifier. All geographic data at the same resolution of 30-m were projected to a universal coordinate system, which was WGS 1984 UTM Zone 48 N. Figure 17 shows the spatial distribution of elevation, slope, roughness of DEM and FVC in the Chengdu urban districts.

Figure 17. Spatial distribution of four features in the study area of Chengdu.

Figure 18. CWP map of Chengdu.
The accuracy of RF algorithm is 98.0% for estimating the waterlogging probability of Chengdu. The average and standard deviation of waterlogging probability in Chengdu are 0.27 and 0.25, respectively. Figure 18 presents the CWP map of Chengdu classified into five categories with a interval of 0.2, in which parts of the true waterlogging points are labeled as red and the 3rd ring road is marked as black. To analyze the contribution of each feature to the waterlogging probability prediction, Table 5 shows the relative importance of each feature. FVC is still the most important feature and DEM takes second place for urban waterlogging probability assessment. Generally, the waterlogging probability in the south of Chengdu is higher than that of the north, which is the opposite of the spatial distribution of DEM. However, the waterlogging probability within the central districts circled by the 3rd ring is low, which is clearly contrary to the practical situation.

The day of July 11th can be taken as the representative day of 2018 Chengdu waterlogging since its destructiveness and the issue of the first yellow alert, so Figure 19 shows the FWP map on July 11th generated by fusing CWP and social media data. The same classification as CWP is applied to FWP map, meanwhile, parts of the true waterlogging points are labeled as red and the third ring road is labeled as black, too. It can be observed from the CWP and FWP map that within the central districts circled by third ring road the waterlogging probability becomes high apparently, which is in accordance with the realism, and the true waterlogging points can be detected effectively, further verifying the generalization of our proposed framework.

To demonstrate the effectiveness of social media data, a quantitative analysis of the number of pixels and true waterlogging points in CWP and FWP can be presented in Table 6. Same as Wuhan city, 0.6 can be also taken as a threshold. It can be observed that the non-high-risk area of FWP covers 32.3% of true waterlogging points less than that of CWP, especially in the range from 0 to 0.2 where the proportion is 26.5%. This indicates that the addition of social media data can find the undetected waterlogging points, proving the contribution of social media data for refining the

| Feature         | Importance |
|-----------------|------------|
| FVC             | 0.38       |
| DEM             | 0.26       |
| Slope of DEM    | 0.18       |
| Roughness of DEM| 0.18       |

Figure 19. FWP map of Chengdu on July 11th, 2018.
5. Conclusions

In this paper, a coarse-to-fine waterlogging probability assessment framework is proposed, in which the multitemporal data such as historical, near real-time and real-time data are generated together to assess the waterlogging probability in the case of urban waterlogging. Meanwhile, a series of information extraction methods from multisource data especially social media data is introduced and performed to analyze their contributions. Furthermore, a morphology-based fusion method considering Tobler’s First Law of Geography is used in this paper for fusing the heterogeneous data (point and raster). Taking the 2016 Wuhan waterlogging and 2018 Chengdu waterlogging as case studies, the coarse-to-fine waterlogging probability assessment framework shows its accuracy and generalization by a series of statistics about true waterlogging points and assessing the performance on other topology type.

To be specific, the historical inundated areas can be regarded as prior knowledge of the inundated-prone areas, which are derived from historical waterlogging points and DEM using an object-oriented segment method. Since the characteristic of large-scale, geographic data such as RS images and DEM can provide useful information for a CWP map with RF classifier. Social media data, which are extremely timely and easily obtained, can improve the waterlogging probability mapping accuracy based on remote sensing image, finally generating the FWP maps for disaster monitoring.

In general, the proposed framework provides new thinking for urban waterlogging probability assessment using multisource data especially when the difference between the pre- and post-disaster RS images cannot be exploited. And the future work will be conducted as the following aspects: (1) since the primitive Text-CNN is proved to be effective, a more advanced text classification algorithm can be applied for selecting the real-time flood-related social media data, which will generate a more accurate FWP map logically; (2) since the unavailability of historical in undated areas derived from historical waterlogging points and DEM, a transfer learning method could be considered when the framework is applied to other urban cities.

Acknowledgments

The authors would like to thank the National Aeronautics and Space Administration (NASA) and the US Geological Survey (USGS) for acquiring and providing the Landsat images used in this study. The authors would also like to thank Jun LI (Sun Yat-sen University, Guangzhou, China) and Zhenyu LI(Shandong University of Science and Technology, Qingdao, China) for sharing social media data-sets in Wuhan and Chengdu, respectively.

Funding

This project was supported by the China Postdoctoral Science Foundation [grant number 2017M622522].

Notes on contributors

Lei Xu is a master student in State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University. Her research interests include the fusion of RS image and social media data and its application.

Ailong Ma is now working as Research Associate in Wuhan University. His major research interests are remote sensing-image processing, evolutionary computing, and deep learning.

References

Abel, F., C. Hauff, G.-J. Houben, R. Stronkman, and K. Tao. 2012. “Twitcident: Fighting Fire with Information from Social Web Streams.” Proceedings of the 21st International Conference on World Wide Web. Lyon, France, 305–308.

Atkinson, G. M., and D. J. Wald. 2007. “Did You Feel It?” Intensity Data: A Surprisingly Good Measure of Earthquake Ground Motion.” Seismological Research Letters 78 (3): 362–368. doi:10.1785/gssrl.78.3.362.

Breiman, L. 1996. “Bagging Predictors.” Machine Learning 24 (2): 123–140. doi:10.1007/BF00058655.

Breiman, L. 2001. “Random Forests.” Machine Learning 45 (1): 5–32. doi:10.1023/A:1010933404324.

Brivio, P. A., R. Colombo, M. Maggi, and R. Tomasoni. 2002. “Integration of Remote Sensing Data and GIS for Accurate Mapping of Flooded Areas.” International Journal of Remote Sensing 23 (3): 429–441. doi:10.1080/01431160010014729.

Carlson, T. N., and D. A. Ripley. 1997. “On the Relation between NDVI, Fractional Vegetation Cover, and Leaf Area Index.” Remote Sensing of Environment 62 (3): 241–252. doi:10.1016/S0034-4257(97)00104-1.

Cervone, G., E. Sava, Q. Huang, E. Schnebele, J. Harrison, and N. Waters. 2016. “Using Twitter for Tasking Remote-Sensing Data Collection and Damage Assessment: 2013 Boulder Flood Case Study.” International Journal of Remote Sensing 37 (1): 100–124. doi:10.1080/01431161.2015.1117684.
Chae, J., D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. S. Ebert, and T. Ertl. 2012. “Spatiotemporal Social Media Analytics for Abnormal Event Detection and Examination Using Seasonal-Trend Decomposition.” Proceedings of 2012 IEEE Conference on Visual Analytics Science and Technology (VAST), Seattle, 143–152.

Chowdhury, S., R. M. Imran, M. R. Asghar, S. Amer-Yahia, and C. Castillo. 2013. “Tweet4act: Using Incident-Specific Profiles for Classifying Crisis-Related Messages.” Proceedings of the International Conference on Information Systems for Crisis Response and Management, Baden-Baden.

De Albuquerque, J. P., B. Herfort, A. Brenning, and A. Zipf. 2015. “A Geographic Approach for Combining Social Media and Authoritative Data Towards Identifying Useful Information for Disaster Management.” International Journal of Geographical Information Science 29 (4): 667–689. doi:10.1080/13658816.2014.996567.

Feng, Q., J. Liu, and J. Gong. 2015. “Urban Flood Mapping Based on Unmanned Aerial Vehicle Remote Sensing and Random Forest Classifier—a Case of Yuyao, China.” Water 7 (12): 1437–1455. doi:10.3390/w7121437.

Feyisa, G. L., H. Melibly, R. Fensholt, and S. R. Proud. 2014. “Automated Water Extraction Index: A New Technique for Surface Water Mapping Using Landsat Imagery.” Remote Sensing of Environment 140: 23–35. doi:10.1016/j.rse.2013.08.029.

Foresti, G. L., M. Farinosi, and M. Verrnier. 2015. “Situational Awareness in Smart Environments: Socio-Mobile and Sensor Data Fusion for Emergency Response to Disasters.” Journal of Ambient Intelligence and Humanized Computing 6 (2): 239–257. doi:10.1007/s12652-014-0227-x.

Gao, B. 1996. "NDWI—a Normalized Difference Water Index for Remote Sensing of Vegetation Liquid Water from Space." Remote Sensing of Environment 58 (3): 257–266. doi:10.1016/S0034-4257(96)00067-3.

Goward, S. N., B. Markham, D. G. Dye, W. Dulaney, and J. Yang. 1991. “Normalized Difference Vegetation Index Measurements from the Advanced Very High Resolution Radiometer.” Remote Sensing of Environment 35 (2–3): 257–277. doi:10.1016/0034-4257(91)90017-Z.

Herfort, B., J. P. de Albuquerque, S.-J. Schelhorn, and A. Zipf. 2014. “Exploring the Geographical Relations between Social Media and Flood Phenomena to Improve Situational Awareness.” In Huerta J., Schade S., Granell C. (eds), Connecting a Digital Europe through Location and Place, 55–71. Berlin, Heidelberg: Springer.

Huang, X., C. Wang, and Z. Li. 2018a. “A near Real-Time Flood-Mapping Approach by Integrating Social Media and Post-Event Satellite Imagery.” Annals of GIS 24 (2): 113–123. doi:10.1080/19475683.2018.1450787.

Huang, X., C. Wang, and Z. Li. 2018b. “Reconstructing Flood Inundation Probability by Enhancing near Real-Time Imagery with Real-Time Gauges and Tweets.” IEEE Transactions on Geoscience and Remote Sensing 56 (8): 4691–4701. doi:10.1109/TGRS.2018.2835306.

Ilieva, R. T., and T. McPherson. 2018. "Social-Media Data for Urban Sustainability." Nature Sustainability 1 (10): 553–565. doi:10.1038/s41893-018-0153-6.

Islam, M. M., and K. Sado. 2000. “Flood Hazard Assessment in Bangladesh Using Noaa Avhrr Data with Geographical Information System.” Hydrological Processes 14 (3): 605–620. doi:10.1002/(SICI)1099-1085(20000228)14:3<605::AID-HYP957>3.0.CO;2-L.

Jackway, P. T., and M. Deriche. 1996. “Scale-Space Properties of the Multiscale Morphological Dilatation-Erosion.” IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (1): 38–51. doi:10.1109/34.476009.

Jony, R. L., A. Woodley, A. Raj, and D. Perrin. 2018. “Ensemble Classification Technique for Water Detection in Satellite Images.” Proceedings of Digital Image Computing Techniques and Applications, Canberra, 1–8.

Kim, Y. 2014. “Convolutional Neural Networks for Sentence Classification.” arXiv Preprint arXiv:1408.5882.

Li, J., Z. He, J. Plaza, S. Li, J. Chen, H. Wu, Y. Wang, and Y. Liu. 2017. “Social Media: New Perspectives to Improve Remote Sensing for Emergency Response.” Proceedings of the IEEE 105 (10): 1990–1912. doi:10.1109/JPROC.2017.2684460.

Lopez-Fuentes, L., J. van de Weijer, M. Bolanos, and H. Skinnemoen. 2017. “Multi-Modal Deep Learning Approach for Flood Detection.” MediaEval 17: 13–15.

Miller, H. J. 2004. “Tobler’s First Law and Spatial Analysis.” Annals of the Association of American Geographers 94 (2): 284–289. doi:10.1111/j.1467-8306.2004.00420005.x.

Mosavi, A., P. Ozturk, and K. Chau. 2018. “Flood Prediction Using Machine Learning Models: Literature Review.” Water 10 (11): 1536. doi:10.3390/w10111536.

Nie, L., Y. Li, F. Feng, X. Song, M. Wang, and Y. Wang. 2020. “Large-Scale Question Tagging Via Joint Question-Topic Embedding Learning.” ACM Transactions on Information Systems 38 (2): 1–23. doi:10.1145/3380954.

Okoloh, O. 2009. “Usahidi, or ‘Testimony’: Web 2.0 Tools for Crowdsourcing Crisis Information.” Participatory Learning and Action 59 (1): 65–70.

Pal, M. 2005. “Random Forest Classifier for Remote Sensing Classification.” International Journal of Remote Sensing 26 (1): 217–222. doi:10.1080/01431160412331269698.

Panteras, G., and G. Cervone. 2018. “Enhancing the Temporal Resolution of Satellite-Based Flood Extent Generation Using Crowdsourced Data for Disaster Monitoring.” International Journal of Remote Sensing 39 (5): 1459–1474. doi:10.1080/01431161.2017.1400193.

Power, R., B. Robinson, J. Colton, and M. Cameron. 2014. “Emergency Situation Awareness: Twitter Case Studies.” International Conference on Information Systems for Crisis Response and Management in Mediterranean Countries, 218–231, Toulouse, France.

Robinson, B., R. Power, and M. Cameron. 2013a. “An Evidence Based Earthquake Detector Using Twitter.” Proceedings of the Workshop on Language Processing and Crisis Information, Nagoya, 1–9.

Robinson, B., R. Power, and M. Cameron. 2013b. “A Sensitive Twitter Earthquake Detector.” Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, 999–1002.

Rogstadius, J., M. Vukovic, C. A. Teixeira, V. Kostakos, E. Karapanos, and J. A. Laredo. 2013. “Crisistracker: Crowdsourced Social Media Curation for Disaster Awareness.” IBM Journal of Research and Development. 57(5): 1–13. doi:10.1147/JRD.2013.2260692.

Rosser, J. F., D. G. Leibovici, and M. J. Jackson. 2017. “Rapid Flood Inundation Mapping Using Social Media, Remote Sensing and Topographic Data.” Natural Hazards 87 (1): 103–120. doi:10.1007/s11069-017-2755-0.

Sakaki, T., F. Toriumi, and Y. Matsuo. 2011. “Tweet Trend Analysis in an Emergency Situation.” Proceedings of the Workshop on Internet and Disasters, Tokyo, 1–8.
Smith, L., Q. Liang, P. James, and W. Lin. 2017. “Assessing the Utility of Social Media as a Data Source for Flood Risk Management Using a Real-Time Modelling Framework.” *Journal of Flood Risk Management* 10 (3): 370–380. doi:10.1111/jfr3.12154.

Sun, D., Y. Yu, and M. D. Goldberg. 2011. “Deriving Water Fraction and Flood Maps from Modis Images Using a Decision Tree Approach.” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 4 (4): 814–825. doi:10.1109/JSTARS.2011.2125778.

Tyszchuk, Y., C. Hui, M. Grabowski, and W. A. Wallace. 2012. “Social Media and Warning Response Impacts in Extreme Events: Results from a Naturally Occurring Experiment.” 2012 45th Hawaii International Conference on System Sciences,Washington, DC, 818–827.

Wald, D. J., V. Quitoriano, C. B. Worden, M. Hopper, and J. W. Dewey. 2012. “Usgs “Did You Feel It?” Internet-Based Macroseismic Intensity Maps.” *Annals of Geophysics* 54: 6.

Wang, Y., J. D. Colby, and K. A. Mulchahy. 2002. “An Efficient Method for Mapping Flood Extent in a Coastal Floodplain Using Landsat Tm and Dem Data.” *International Journal of Remote Sensing* 23 (18): 3681–3696. doi:10.1080/0143116011014484.

Weng, Q., and D. Lu. 2008. “A Sub-Pixel Analysis of Urbanization Effect on Land Surface Temperature and Its Interplay with Impervious Surface and Vegetation Coverage in Indianapolis, United States.” *International Journal of Applied Earth Observation and Geoinformation* 10 (1): 68–83. doi:10.1016/j.jag.2007.05.002.

Xie, C., X. Huang, W. Zeng, and X. Fang. 2016. “A Novel Water Index for Urban High-Resolution Eight-Band Worldview-2 Imagery.” *International Journal of Digital Earth* 9 (10): 925–941. doi:10.1080/17538947.2016.1170215.

Xu, F. 2015. “The Role of Social Media in Measuring Human Response to Urban Flash Flooding.” PhD thesis, Massachusetts Institute of Technology, +City, +Country.

Xu, H. 2006. “Modification of Normalised Difference Water Index (Ndwi) to Enhance Open Water Features in Remotely Sensed Imagery.” *International Journal of Remote Sensing* 27 (14): 3025–3033. doi:10.1080/01431160600589179.

Xu, Z., V. Sugumaran, and H. Zhang. 2015. “Crowdsourcing Based Spatial Mining of Urban Emergency Events Using Social Media.” Proceedings of the 1st ACM SIGSPATIAL International Workshop on the Use of GIS in Emergency Management, New York, 1–4.

Yang, H., Z. Wang, H. Zhao, and Y. Guo. 2011. “Water Body Extraction Methods Study Based on Rs and GIS.” *Procedia Environmental Sciences* 10: 2619–2624. doi:10.1016/j.proenv.2011.09.007.

Yin, J., A. Lampert, M. Cameron, B. Robinson, and R. Power. 2012. “Using Social Media to Enhance Emergency Situation Awareness.” *IEEE Intelligent Systems* 27 (6): 52–59. doi:10.1109/MIS.2012.6.

Zhang, F., B. Zhang, J. Li, Q. Shen, Y. Wu, and Y. Song. 2011. “Comparative Analysis of Automatic Water Identification Method Based on Multispectral Remote Sensing.” *Procedia Environmental Sciences* 11: 1482–1487. doi:10.1016/j.proenv.2011.12.223.

Zhang, N., H. Chen, J. Chen, and X. Chen. 2016. “Social Media Meets Big Urban Data: A Case Study of Urban Waterlogging Analysis.” *Computational Intelligence and Neuroscience* 2016: 3264587.

Zhang, Y., and B. Wallace. 2015. “A Sensitivity Analysis of (And Practitioners’ Guide To) Convolutional Neural Networks for Sentence Classification.” *arXiv Preprint arXiv:1510.03820*. 

