Article, Published Version

Yang, Pan; Niu, Geng; Du, Erhu; Zheng, Yi

Boosting urban hydrologic research with citizen collected data

Hydrolink

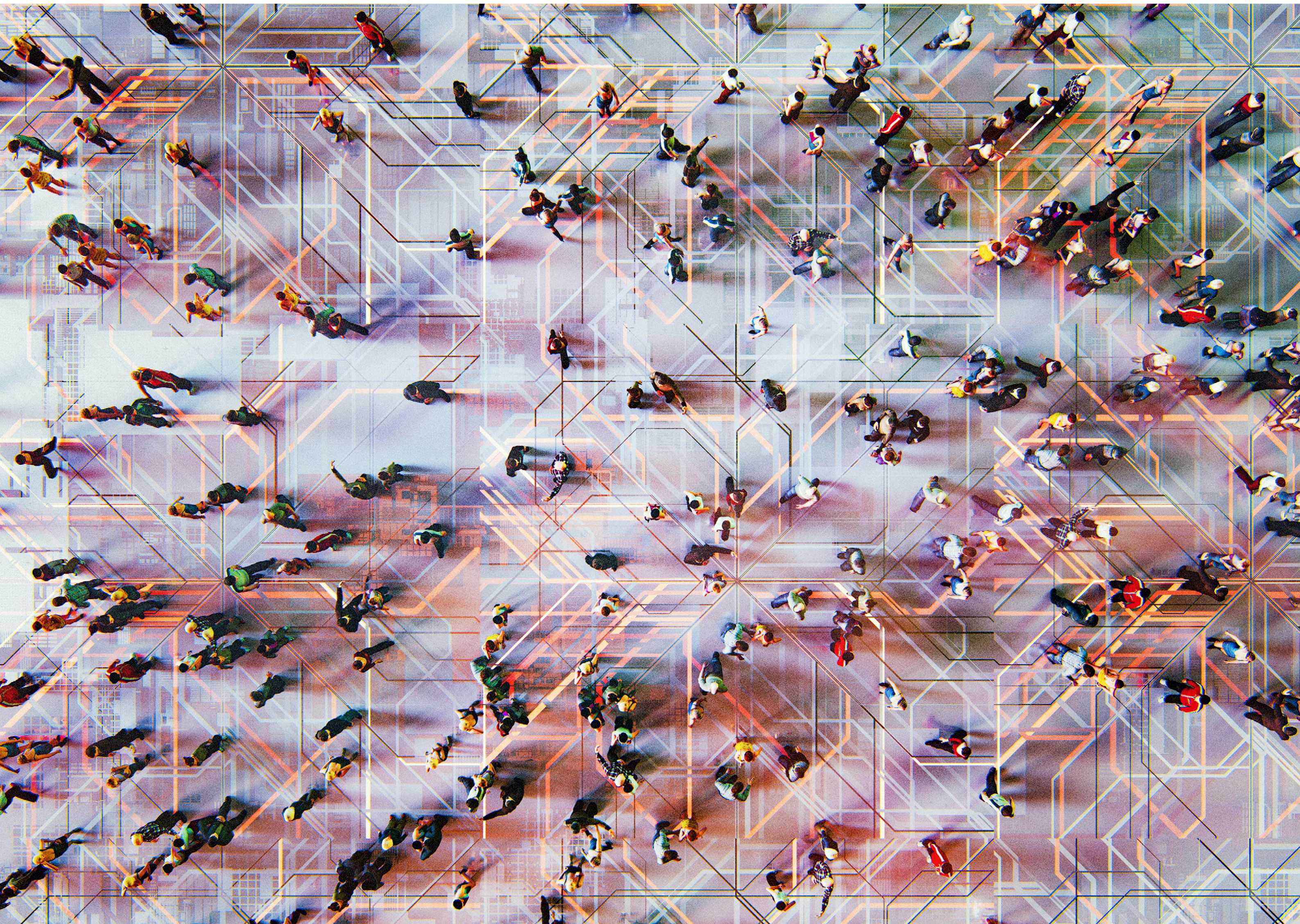
Boosting urban hydrologic research with citizen collected data

By Pan Yang, Geng Niu, Erhu Du and Yi Zheng

Citizen science, described as the method that involves citizens or communities in the scientific research process, has been used in a wide range of studies since the term was first introduced in the 1980s. In the era of big data, citizen science is increasingly attracting attention across a wide variety of scientific disciplines, as indicated by the appearance of vast reports, projects, and peer-reviewed publications. It significantly expands data collection means and scope across temporal and spatial scales, especially in urban hydrologic research.

With the dramatic popularity and implementation of social media apps and IoT (Internet of Things) in recent years, the crowdsourcing approach has been an emerging method to collect hydraulic data at a fraction of the cost of traditional ways. Through various applications in water-related issues, the crowdsourcing approach facilitated the observation of traditional methods that were difficult to quantify and thus improved the statistical power of datasets. It offers a new research paradigm and has been an important component of citizen science.

In urban hydrology research, more specifically, crowdsourcing has been seen as an innovative and promising method of environmental monitoring and data collection through the general where individual citizens or groups are encouraged to provide uploaded data through social media, mobile applications, or IoT platforms. While crowdsourced data is usually associated with a relatively high level of observation error, its tremendous volume of data could largely offset this disadvantage and provide useful information for hydrologic research. Crowdsourcing methods are thus considered to have the potential to provide high-density measurement data and augment related project scope across temporal and spatial scales. This is critical for studies such as flood management or urban stormwater modeling in which high spatial-temporal resolution monitoring data are required.



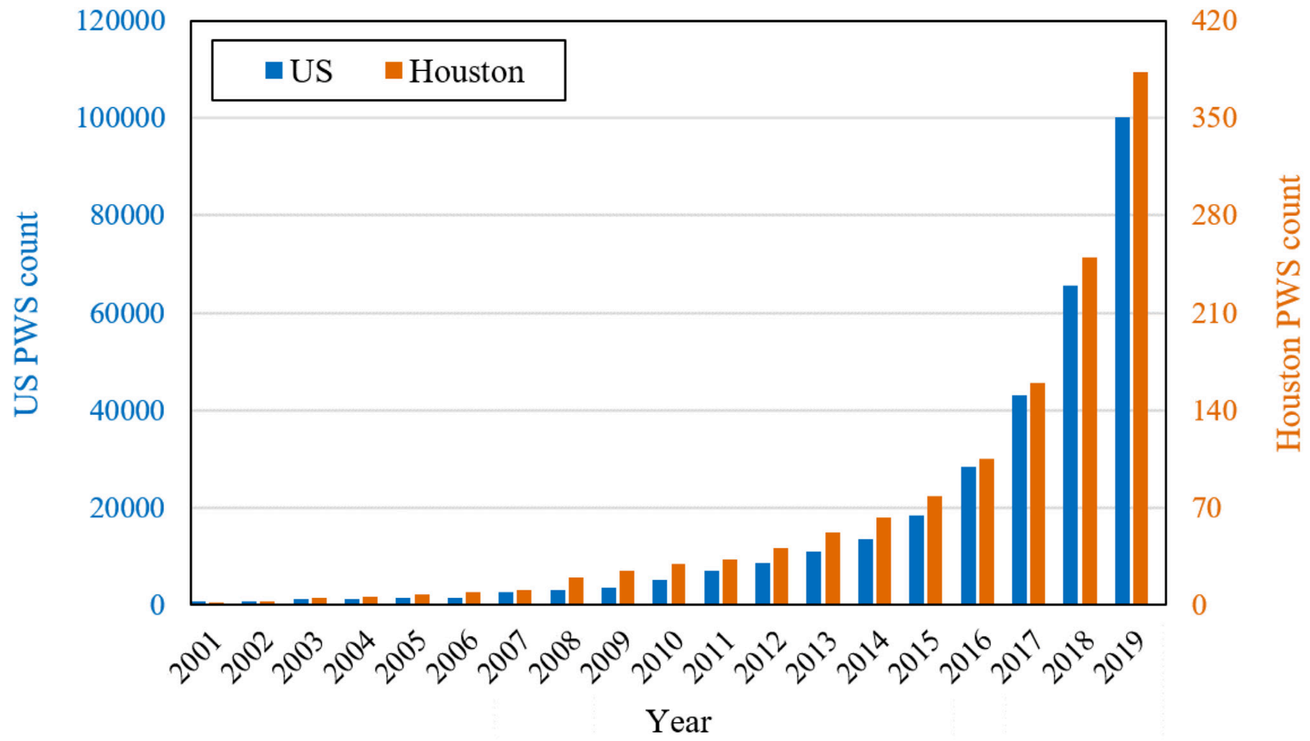


Figure 1 | In the past two decades, the number of Personal Weather Stations (PWSs) in the US and Houston has been growing exponentially. (Reproduced from Chen *et al.*, 2021)[1].

1 | Data acquisition methods

The contribution of crowdsourced data for urban hydrologic research is maximized when the data volume is large, which depends heavily on the choice of method for data acquisition. Ideally, the data acquisition method should have the potential to reach a large number of citizens or to be integrated with other existing devices (e.g., smartphones or surveillance cameras).

I Citizens

Collecting and reporting by individual citizen participation represent the most straightforward manner for sourcing data. Crowdsourced data collected via citizen participation usually is qualitative or quantitative regarding weather conditions, geospatial data, and stream stage via low-cost or homemade sensors. Take rainfall estimation as an example, during the past two decades, the number of personal weather stations (PWSs) in the US has increased exponentially from nearly 7,000 in 2010 to almost 100,000 in 2019 (**Figure 1**). Data collected from those PWSs have contributed to a much-improved estimate of rainfall in urban regions.

II Images/Videos

In recent years, image/video-based methods have been widely used in urban hydrology research. Based on report by IHS Markit[2], this planet is estimated to have over one billion surveillance cameras by 2021. With the aid of various data processing techniques, images captured by ordinary surveillance cameras or in-vehicle cameras have been undertaken to extract information about rainfall[3], flood inundation[4], and water pollution.

III Social Media

With the widespread use of various social media apps, it has been employed to assist urban hydrology research with the aid of data mining techniques. Crowdsourced data from social-media is usually attributed to unintentional type, as the information source is not shared for purposes of hydrology research. Examples include mining text information from social media, such as geo-tagged tweets about floods and water, to map flooding extents and estimate water levels.

IV IoT sensors

Crowdsourcing weather data can be collected from sensors integrated with portable devices, vehicles, mobile phones, smart home equipment, and telecommunication infrastructure. For instance, audio clips from built-in sensors of smartphones have been utilized to detect precipitation intensity. Moreover, wireless antenna signals could be used to monitor a variety of weather data, from fog to precipitation. Based on the principle that precipitation will attenuate the electromagnetic signal between antennas, the precipitation intensity is measured according to the relationship between electromagnetic signal attenuation and precipitation intensity.

Currently, there is not enough study showing the exact figures of the data volume contributed by each of the data acquisition methods. However, it is considered that passive crowdsourcing methods (where sensors can passively collect data without human interference) could potentially provide a much larger size of data than active crowdsourcing methods (where active data collection actions are needed to procure the data).

2 | Data processing

Despite their large data volume, crowdsourced data are usually heterogenous and unstructured in nature, and prone to a variety of noises (including observation and sampling error, incorrect data report, equipment failure, etc.). Adequate data processing is necessary to ensure the quality of crowdsourced data for urban hydrologic research.

Different data processing and cleaning approaches have been proposed. Some compared crowdsourced data with gold-standard data sets or by expert judgments; some assessed the trustworthiness of crowdsourced networks by reputation system; others identified noisy data from crowdsourced observations with preset rules[5]. More recently, a machine learning-based crowdsourced data quality control model was proposed to identify and filter noise from general crowdsourced rainfall observations automatically[6]. Since crowdsourced data is usually in large volume (e.g., obtained from PWSs) and discontinuous in both time and space, a machine learning-based processing method can automate the noise detection and removal process. As shown in Fig. 2, the machine learning based method can detect noisy data in spatially and temporally discrete crowdsourced observations coming from both fixed points (e.g., surveillance cameras) and moving sensors (e.g., moving cars/pedestrians), and can significantly reduce the overall rainfall estimate errors.

3 | Data utilization

Crowdsourcing can be used directly, or indirectly via urban stormwater models, to assist urban water management, flood evacuation, and other urban hydrologic management practices. Since performances of stormwater models rely heavily on the quality of input data (i.e., 'garbage in, garbage out'), a concern regarding their use of crowdsourced data is the relatively poor data quality. However, researchers have shown that low-quality crowdsourced data could contribute positively to these models, via specific mechanisms[7].

Specifically, the error propagation property of the stormwater model and the error structure of crowdsourced data ensures 'gold' (good stormwater modeling result) derived from 'garbage' (low-quality crowdsourced data). As shown in Fig. 3, because of the relatively low observation density, rain gauge estimated rainfall fields are usually associated with systematic under-estimation/overestimation of rainfall intensity at the storm center, where the hydrologic response is most sensitive to errors in rainfall data. In contrast, though the individual errors of crowdsourcing rainfall data are high, their high spatial density avoids systematic error at the storm center. As a result, stormwater flow simulated with erroneous crowdsourcing rainfall data can outperform that of more accurate traditional rain gauge data[7].

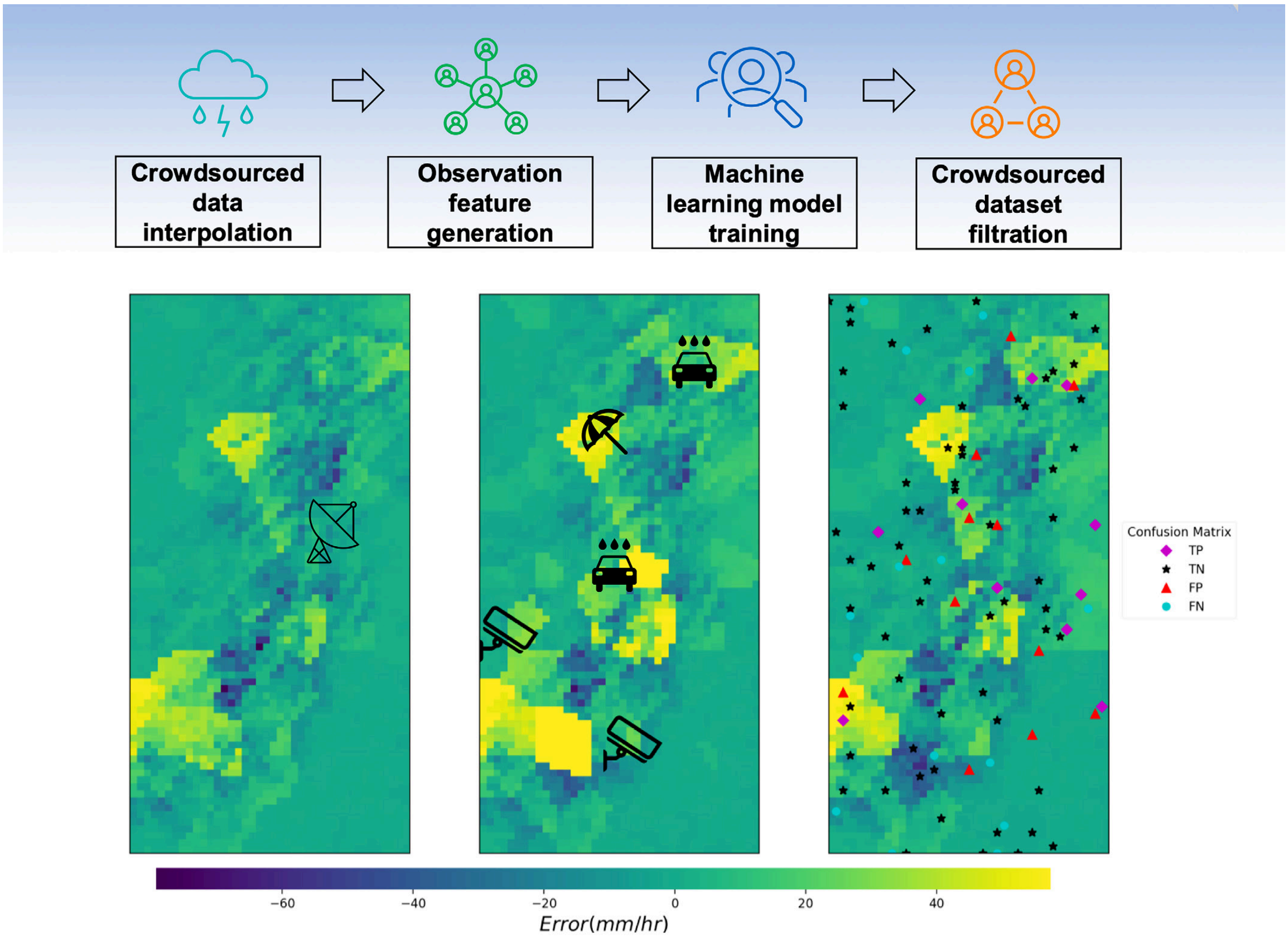


Figure 2 | Machine learning-based processing approach for automatic crowdsourced data quality control. (Modified from Niu *et al.*, 2021)[6].

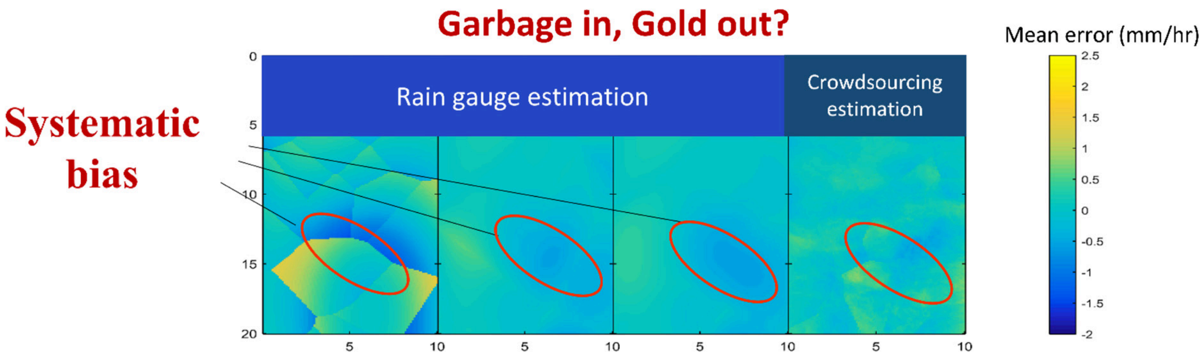


Figure 3 | Comparison of rainfall estimation error structure (derived from a synthetic but realistic rainfall simulation with known rainfall intensities) between rain gauge and crowdsourced data explains the superior stormwater modeling performance of the crowdsourced data. (Modified from Yang and Ng, 2017)[7].

The higher density of data and improved performance of high-resolution stormwater modeling enhance our ability to manage complex urban drainage systems, which are interrelated with many other urban infrastructures such as traffic networks, power grids, and water supply systems. A simple application crowdsourcing urban hydrological data is flash flood management, in which the real-time control of pumps and weirs can be improved with crowdsourced data. For the secondary damages caused by urban flash floods (e.g., traffic jams and power outrage), crowdsourced hydrological data can be used to develop early warning and management systems. An example is the planning of flood evacuation, which can be greatly improved with precise information on the location and timing of flood inundation provided by crowdsourced data. Such efforts can be integrated to assist the development of the smart city and digital twins of urban infrastructures, which aim to improve government service and citizen welfare with information and communication technologies.

4 | Citizen involvement

Crowdsourcing hydrological data can be broadly categorized into passive and active types, and active crowdsourcing is considered more challenging in terms of ensuring enough data volume and adequate data quality as continuous efforts from participants are needed. Based on past experiences, it is suggested that the simplicity of the application and immediate feedbacks are critical elements for a successful crowdsourcing project. The design of feedback must relate to the primary motivators for participating in a crowdsourcing project, which may include monetary reward, the expectation of reciprocal activities by others, the feeling of competition, and pure altruism. Passive crowdsourcing is considered less challenging in citizen involvement, but it may also bring inequality problems. As passive crowdsourcing projects usually require citizens to own specific sensors, e.g., personal weather stations, wealthy communities benefit more from such projects as they can afford more sensors to improve the monitoring quality. Researchers have now started combining efforts from both social science and hydrological science to improve participation in crowdsourcing projects. For example, Yang *et al.* (2019)[8] developed an agent-based model (ABM) which integrates individual participant behavior rules regarding crowdsourcing reward-action relationship with a rainfall field estimation model, and the integrated model was further applied to investigate the performances of several reward allocation strategies.

There can be ethical or legal barriers to the collection of crowdsourced data, especially regarding the data privacy issue. The use of personally owned sensors in crowdsourcing projects, especially smartphones, raises concerns regarding the collection and use of sensitive personal data. Currently, such an issue has not been widely discussed by the research community, which poses a challenge in the development of laws and regulations regarding the use and governance of citizen science data. Recently, the European Union (EU) developed a General Data Protection Regulation (GDPR) for citizen data privacy protection, which highlights and regulates the risks of accidental or unlawful destruction, loss, alteration, unauthorized disclosure of, or access to, personal data. To avoid ethical and legal issues regarding data privacy, anonymous task distribution, anonymous data reporting, privacy-aware data processing, as well as access control and audit of data utilization, can be used as suggested by relevant researchers.

Conclusions and outlook

Data acquisition remains challenging in urban hydrologic research. With the fast development and exponential adoption of low-cost sensors, citizen science, especially crowdsourcing, provides a promising direction to (at least partially) address the data challenge. Developing crowdsourcing in urban hydrologic research requires efforts from electronic engineering, hydrology, computer science, and social science. This article briefly introduces four aspects of crowdsourcing: data acquisition methods, data processing, data utilization, and citizen involvement. To move forward, there still exist many research opportunities in this relatively new field, e.g., the motivation of citizens, the integration with smart city development, and the development of standardized protocols. Among them, the integration of crowdsourcing hydrologic data with professionally collected data can be a promising direction. Combining both sources of data could overcome their disadvantages of low coverage (professional data) and low accuracy (crowdsourced data) (see an example in Yang and Ng, 2019)[8], and it provides an opportunity to re-design our urban hydrologic monitoring system with the aids of both professionals and amateurs. Another promising field is the development of digital twin modeling of urban watersheds, which can be substantially benefited from the vast volume of crowdsourcing hydrological data.

References

1 | Chen A B, Behl M, Goodall J L. Assessing the Trustworthiness of Crowdsourced Rainfall Networks: A Reputation System Approach. Water Resources Research, 2021, 57(12): e2021WR029721.

2 | Oliver Philippou (2019). Video Surveillance Installed Base Report - 2019. https://omdia.tech.informa.com/OM005353/Video-Surveillance-Installed-Base-Report---2019

3 | Jiang S, Babovic V, Zheng Y, *et al.* Advancing opportunistic sensing in hydrology: A novel approach to measuring rainfall with ordinary surveillance cameras. Water Resources Research, 2019, 55(4): 3004-3027.

4 | Wang R Q, Ding Y. Semi-supervised Identification and Mapping of Surface Water Extent using Street-level Monitoring Videos. arXiv preprint arXiv:2202.00096, 2022.

5 | De Vos, L. W., Leijnse, H., Overeem, A., & Uijlenhoet, R. (2019). Quality control for crowdsourced personal weather stations to enable operational rainfall monitoring. Geophysical Research Letters, 46(15), 8820 8829. https://doi.org/10.1029/2019GL083731

6 | Niu G, Yang P, Zheng Y, *et al.* Automatic quality control of crowdsourced rainfall data with multiple noises: A machine learning approach. Water Resources Research, 2021, 57(11): e2020WR029121.

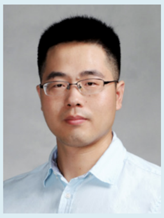
7 | Yang P, Ng T L. Gauging through the crowd: A crowd sourcing approach to urban rainfall measurement and storm water modelling implications. Water Resources Research, 2017, 53(11): 9462-9478.

8 | Yang P, Ng T L, Cai X. Reward based participant management for crowdsourcing rainfall monitoring: An agent based model simulation. Water Resources Research, 2019, 55(10): 8122-8141.

Pan Yang
Pan Yang received his Ph.D. at Hong Kong University of Science and Technology and now works as an associate professor at School of Ecology, Environment and Resources at Guangdong University of Technology. Dr Yang's research interests include crowdsourcing rainfall monitoring, water resources systems analysis, coupled nature-human systems, and environmental big data analytics. He has published over 20 peer-reviewed papers in top journals such as Review of Geophysics, Environmental Science & Technology, and Water Resources Research.

Geng Niu
Geng Niu received his Ph.D. degree in Environmental Science from Peking University, China. Now he works as a researcher at China Institute of Water Resources and Hydropower Research. His research interests include Eco-hydrological Modelling, coupled natural-human systems, and big data and machine learning in water sciences.

Yi Zheng
Dr Yi Zheng received his Ph.D. degree at University of California, Santa Barbara in 2007. His research interests include watershed system modeling, human-nature nexus, and environmental big data and AI. He is currently an associate editor of Water Resources Research and Journal of Hydrologic Engineering-ASCE. He has published over 130 peer-reviewed papers, many in top journals such a s Nature Food, Nature Geoscience, Nature Sustainability, Geophysical Research Letters, Environmental Science & Technology, and Water Resources Research.



Erhu Du
Erhu Du received his Ph.D. at University of Illinois Urbana-Champaign and now works as a professor at Yangtze Institute for Conservation and Development at Hohai University. Dr Du's research interests include water resources management, integrated surface water-groundwater simulation, urban flood management and environmental policy.