Building an Open, Multi-Sensor, Dataset of Water Pollution of Ganga Basin and Application to Assess Impact of Large Religious Gatherings

Biplav Srivastava[1], Sandeep Singh Sandha[2], Vaskar Raychoudhury[4], Sukanya Randhawa[3], Viral Kapoor[5] and Anmol Agrawal[5]

biplavs@us.ibm.com, sandha@cs.ucla.edu, raychov@miamioh.edu, Sukanya_R@in.ibm.com, viral.kapoor21@gmail.com, anmol.agrawal17@gmail.com

[1]IBM, New York, USA; [2]University of California, Los Angeles, USA; [3]IBM Research - India
[4]Miami University, Ohio, USA; [5]Indian Institute of Technology, Roorkee, India

Abstract— **Water is a crucial pre-requisite for all human activities. Due to growing demand from population and shrinking supply of potable water, there is an urgent need to use computational methods to manage available water intelligently, and especially in developing countries like India where even basic data to track water availability or physical infrastructure to process water are inadequate. In this context, we present a dataset of water pollution containing quantitative and qualitative data from a combination for modalities - real-time sensors, lab results, and estimates from people using mobile apps. The data on our API-accessible cloud platform covers more than 60 locations and consists of both what we have ourselves collected from multiple location following a novel process, and from others (lab-results) which were open but hither-to difficult to access. Further, we discuss an application of released data to understand spatio-temporal pollution impact of a large event with hundreds of millions of people converging on a river during a religious gathering (Ardh Khumbh 2016) spread over months. Such unprecedented details can help authorities manage an ongoing event or plan for future ones. The dataset has been available on *BlueWater* platform since 2016 and anyone can use the data for their application and also contribute new data.**

Keywords—water quality monitoring, smart system, sensors, data analysis

I. Introduction

Water management deals with twin issues of *monitoring and quality control of potable water sources* and *discharge of domestic and industrial effluents*. These two objectives become related when effluents are directly dumped in rivers/lakes and such polluted water becomes the water source for downstream population. With a growing human population and shrinking supply of potable water, water resources are under intense stress world-wide and especially in developing countries where even basic data to track water availability or physical infrastructure to process water are inadequate. The general public is often aware of the upcoming water crisis and increasingly has access to computing resources, via mobile phones. Consider that *Abhay* may want to take a bath in the river during a religious festival and would want to know which banks (religious sites, i.e., *ghats*) of the river are feasible to go without getting sick. *Bina* may want to tap ground water or fetch from river water for household activities. *Chetan* may want to use river water for irrigating his fields over ground water. *Divya* may wonder if fishing or vegetable growing is promising on the river catchment area to supplement her family's earnings. These and other users are routinely taking decisions which can be driven by water pollution data if it were available and AI-researchers could create decision-support apps.

The situation gets further aggravated when large-scale gatherings take place on the banks of rivers and people indiscriminately use water in activities bathing and religious rites where they may discharge external additives to water like coins, milk, flowers and organically ashes. Ironically, water data has been collected by numerous agencies around the world (usually government, academic and non-profits) and individuals for many years on different stretches of the river measuring different parameters at different times using different technologies. However, they as a whole are not only insufficient in terms of quality and quantity but also inaccessible to wider public because they were not annotated along common usage scenarios.

Water data is available online from a few regions of countries like Bath in UK [14], New South Wales in Australia [15] and Hudson Valley in USA [16]. In India, CPCB maintains a website with data from tens of locations since 2015. An early example of support tool for water is Water Advisor [18].

In this context, our contributions in paper are that we: *(a)* Release a dataset of water pollution from Ganga basin in India that contains both quantitative and qualitative estimates obtained from a combination for modalities - real-time sensors, lab results, and estimates from people using mobile apps. It is hosted on *BlueWater* platform [1] and has thousands of data-points along with permissible parameter limits and their applicability for different purposes which can be accessed digitally via APIs and mobile apps. *(b)* Demonstrate usage of released data in a case study where hundreds of millions of people assemble along a river for over 4 months for religious rituals and understand its environmental impact. *(c)* Discuss methodological and AI issues in using the released data for other use-cases.

In the rest of the paper, we discuss the dataset and how it was created, followed by demonstration of its usage in understanding impact of a large event and concluding with a discussion on issues in using the data for other use-cases.

II. Data Set

In this section, we describe the quantitative and qualitative datasets released publicly by us and the methodology followed. We start with a background of water pollutants and how they are measured. We then describe the released data and its characteristics.

A. Water Quality Measurement Parameters

When monitoring water, a number of parameters are of interest to different stakeholders depending on their purpose. Environment agencies like EPA in US and CPCB in India recommend tens of parameters - CPCB recommends more than 30 parameters [12][13] in India. We describe the most important here - *pH*, *electrical conductivity*, *dissolved oxygen*, and *turbidity*. Agencies also recommend standards for different usage like drinking, irrigation - CPCB prescribes parameters of interest and their ranges for 22 industries [13]. A major challenge is overlapping specifications of multiple agencies within a government and also at multiple levels (national, state and international) which can be in conflict. We reconcile ranges for 25 parameters in the data released using CPCB guidelines on purpose and pollutants and provide it in machine processable JSON format [19].

- *pH*: pH is a measure of how acidic/basic water is. The range goes from 0-14, with 7 being neutral. The pH scale is a logarithmic one; so, water with a pH of 5 is ten times more acidic than water having a pH of 6.
- *Electrical Conductivity:* Pure water is not a good conductor of electricity. Conductivity of water increases as the concentration of ions increases. While drinking water has a conductivity of 0.005 – 0.05 S/m, for sea-water it is 5 S/m. Conductivity of water increases with increase in temperature and vice versa.
- *Dissolved Oxygen (DO):* It refers to the level of free, non-compound oxygen present in water or other liquids. Adequate DO is required to maintain good water quality. DO may increase with rising water temperature or decrease by mixing untreated or partially treated sewage with water. As DO level in water drops below 5.0 mg/l, aquatic life is endangered.
- *Turbidity:* It measures relative clarity of a liquid by the amount of light scattered by impurities (like, clay, silt, finely divided inorganic and organic matter, algae, soluble colored organic compounds, and plankton) suspended in the water. Highly turbid drinking water increases the chance of gastro-intestinal diseases.

Table I Water standards for different Activities

Activity	pH	Conductivity (µS/cm)	DO (mg/L)	Turbidity (FNU)
Drinking	6.5-8.5	≤ 2250	≥ 6	1-5
Bathing	6.5-8.5	--	≥ 5	--
Irrigation	6.0-8.5	≤ 2250	≥ 6	1-5

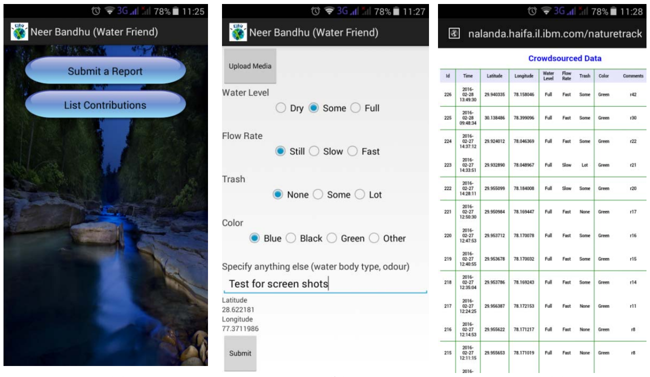
B. Methods of Collecting Water Pollution Data

- *Sample collection and lab tests for quantitative results:* The most accurate way of measuring water pollution is by collecting water samples from water bodies following standardized practices periodically, testing them for all parameters important for a purpose and publishing them. Such data may be available in printed form but not accessible digitally, is quite expensive to create due to cost of sample collection and lab testing, is time-consuming and quantitative. The scientific challenges are in testing the pollutants and collection of samples. We took published lab results from over 60 locations in India from different sources like universities and put them on our common platform. They covered 4-5 parameters at a time and the samples were collected at frequencies varying from once a month to yearly.
- *Real-time (RT) sensing:* We have used a state-of-art, portable, commercial multi-sensor kit from Hanna (HI9829[1]) which monitors up to 14 different water quality parameters including pH, ORP, conductivity, dissolved oxygen (DO), turbidity, ammonium, chloride, nitrate, and temperature. It is possible to log the data locally and to track location of measurement points using GPS (up to an accuracy of 10m). Following convention, the sensor kit is pre-calibrated to report the readings at 25C to remove the effect of temperature on the readings captured on different days and times. Note that RT sensors can only measure a subset of supported parameters (typically 4-7 based on probe deployed) at a time and the technology does not support measurement of all parameters an agency may recommend like Biological Oxidation Demand (BOD) and Faecal Count (FC). RT sensors, once procured, are inexpensive to operate and give results instantly. However, the technical challenge with them is determine a reliable process to measure under an environmental condition. We used our sensors on Ganga and its tributary, Yamuna, in India using a non-stationary platform to get higher spatial coverage, in contrast to the normal practice of attaching sensors to a fixed location. Further, we used a mobile app (explained below) to capture qualitative data in conjunction with the RT sensors at select landmarks to help with validation.
- *Crowd-sourced, qualitative, sensing using a mobile application*: We have released a mobile application (app) on Google play store called *Neer Bandhu* (NB), meaning 'Water Friend', which can be used by people to take the picture of a river area depicting water pollution (Fig. 1(a)) and to select a few options regarding the characteristic of water (Fig. 1(b)) which are specially derived from standards and common practices that exist in the water domain. The data that is being collected through this app is available within the app (Fig. 1(c)) and online. It has to be noted that the qualitative data is not precise as the quantitative data.

[1] [1] http://hannainst.com/

However, it offers evidence that can be used to interpret results from other modalities like RT sensors and also capture information that people care about but are hard to measure (e.g., water smell).

We used *Neer Bandhu* both independently as well as in conjunction with RT sensors.


78% 11:25
Neer Bandhu (Water Friend)
Submit a Report
List Contributions
78% 11:27
Neer Bandhu (Water Friend)
Upload Media
78% 11:28
nalanda.haifa.il.ibm.com/naturetrack
Submit

Water Level
○ Dry ● Some ○ Full

Flow Rate
● Still ○ Slow ○ Fast

Trash
● None ○ Some ○ Lot

Color
● Blue ○ Black ○ Green ○ Other

Specify anything else (water body type, odour)
Test for screen shots

Latitude
28.622181
Longitude
77.3711986

Crowdsourced Data

Id	Time	Latitude	Longitude	Water Level	Flow Rate	Trash	Color	Comments
226	2016-02-28 13:49:30	29.940335	78.158046	Full	Fast	Some	Green	r42
225	2016-02-28 09:48:34	30.138486	78.399096	Full	Fast	Some	Green	r30
224	2016-02-27 14:37:12	29.924012	78.046369	Full	Fast	Some	Green	r22
223	2016-02-27 14:33:51	29.932890	78.048967	Full	Slow	Lot	Green	r21
222	2016-02-27 14:28:11	29.955099	78.184008	Full	Slow	Some	Green	r20
221	2016-02-27 12:50:30	29.950984	78.169447	Full	Fast	None	Green	r17
220	2016-02-27 12:47:53	29.953712	78.170078	Full	Fast	Some	Green	r16
219	2016-02-27 12:40:55	29.953678	78.170032	Full	Fast	Some	Green	r15
218	2016-02-27 12:35:04	29.953786	78.169243	Full	Fast	Some	Green	r14
217	2016-02-27 12:24:25	29.956387	78.172153	Full	Fast	None	Green	r11
216	2016-02-27 12:14:53	29.955622	78.171217	Full	Fast	None	Green	r8
215	2016-02-27 12:11:15	29.955653	78.171019	Full	Fast	None	Green	r8
	2016-							

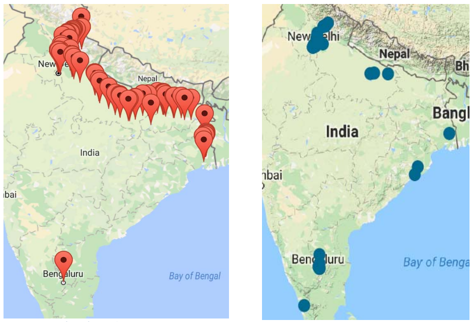
(a) (b) (c)

Fig. 1 Qualitative Data Collection using Neer Bandhu

C. Data Released and Exploration Tools

Figure 2 shows the location of quantitative and qualitative data on a map. We have 117 data points from lab tests, 10,506 from RT sensors and 78 points from NB. It covers the Ganga basin covering Ganga and its tributary, Yamuna. All data, except lab results, was collected by authors from Ganga (Feb-April 2016), Yamuna (Dec 2015) and Hindon, a tributary of Yamuna (Sep 2015).

We commissioned and release 5 of our lab results and the rest were contributed by other teams. The data is accessible via an Application Programmable Interface (API) to our cloud platform called *BlueWater* so that a consuming application can get all changes seamlessly and automatically. We have also released a mobile app, called *GangaWatch*, which uses the public APIs, to facilitate exploration of released data.



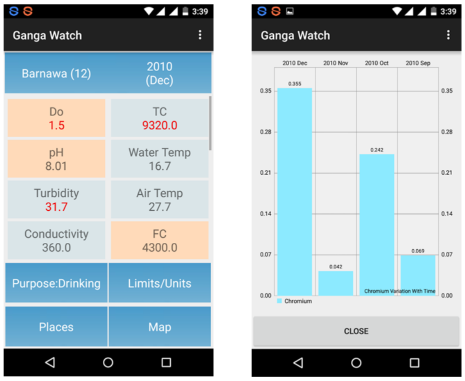
Quantitative (10,506 + 117 Points)

Qualitative (78 Points) via Neerbandhu

Fig. 2 Statistics and locations of pollution data

III. Application Case Study

In this section, we discuss our analysis of Ganga water quality during Ardh Kumbh 2016 held at Haridwar in the state of Uttarakhand. We also describe the relevant data collection process in more detail to help interpretation of the results.



Barnawa (12) 2010 (Dec)

Do 1.5	TC 9320.0
pH 8.01	Water Temp 16.7
Turbidity 31.7	Air Temp 27.7
Conductivity 360.0	FC 4300.0

Fig. 3 *GangaWatch* screenshots showing released data

Fig. 4 Religious Gathering during Ardh Kumbh 2016, Haridwar

A. Background of Khumbh – A Large-Scale Religious Event near a River Bank

In India, there are several river-centric large-scale religious gatherings of which the Kumbh festivals are most important. Here, millions of people gather at the nearest holy places and undergo ablution in the river on specific days and times with religious motifs. People indiscriminately use water for bathing and religious activities, such as, offering milk, ghee, flowers, coins, idols, ashes of departed ones, body hairs into water [4]. Depending on their activities, the river water gets affected impacting economic activities downstream.

Ardh Kumbh 2016 took place in Ganga river in Haridwar, Uttarakhand between Jan-April. An estimated 10 crore (100 Million) devotees took part overall with nine peak bathing days when 10s of millions devotees came[2].

B. Data Collection

The study involved collection of water quality parameters at a high resolution as detailed in Table II over 76 Km length along the Ganga river and its canal in Rishikesh, Haridwar and Roorkee regions of the Uttarakhand district in India. We selectively chose multiple locations and monitored them throughout the festival. The choice of location was made such that water quality *before*, *after* and *at* the bathing places could be studied. We collected data during various days in February to April 2016 (Table III).

In most of the places, we took readings by standing on the bank but in Rishikesh we used boats for data measurements in the river. We chose both peak and off-peak bathing days to see the variations caused by mass bathing. On peak bathing day(s),

[2] http://www.kumbhamela.net/kumbha-mela-haridwar.html

data was only collected at Haridwar since it is the center of almost all religious activities.

TABLE II DATA COLLETION LOCATION STATISTICS

Total no. of collection coordinates (repeating)	134 (only on the Ganga and its Canal)
Starting coordinate	(30.13848, 78.39982) – Near Shivpuri, Rishikesh
Last coordinate	(29.81985, 77.87248) – Asaf Nagar, Roorkee
Total Road distance	76.1 KMs

TABLE III DATA COLLECTION DATES

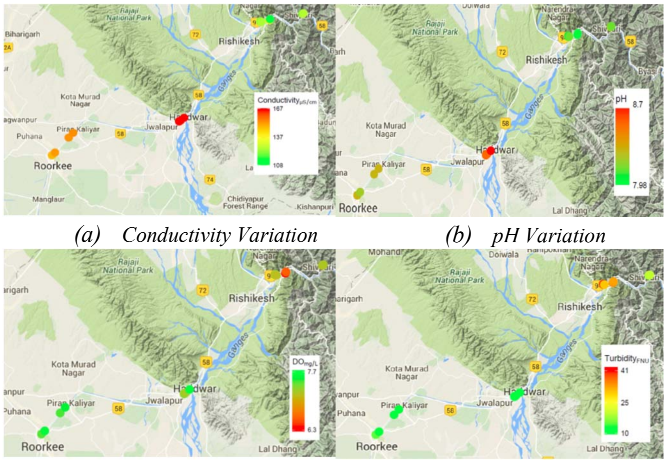
Date	Kumbh Bathing Day/ Other day	Place	Time
Feb. 27	Other day	Haridwar and Roorkee	3-7 PM
Feb. 28	Other day	Shivpuri, Rishikesh, Roorkee	11 AM - 6 PM
Mar. 07	Bathing Day	Haridwar	6-10 AM
Apr. 07	Bathing Day	Haridwar	1-4 PM
Apr. 15	Bathing Day	Haridwar	7-10 AM

IV. ANALYSIS AND INSIGHTS

In this section, first we show how the water quality parameters vary for the entire stretch from Rishikesh (Shivpuri) to Roorkee and then we shall show how the parameters vary during the bathing days at *Har-ki-Pauri* location, Haridwar, which is a prominent location midway.

A. Variations over the Entire Stretch

In order to understand the effect of bathing at Haridwar, we followed the river Ganga upstream and downstream. In Fig. 5, the mean of parameter values at different places is plotted. Haridwar stands out distinct from the upstream Rishikesh and downstream Roorkee. Ganga at Rishikesh (Shivpuri) is still at the Himalayas and has rocky bed and high velocity.



(a) Conductivity Variation *(b) pH Variation*

(c) DO Variation *(d) Turbidity Variation*

Fig. 5 Heat Maps of Parameter Variation (over entire stretch)

So, it carries large amount of silt, which affects its *turbidity* (Fig. 5(d)). At Haridwar, extensive use of water is affecting its *conductivity* and *pH*, as seen in the plots (Fig. 5(a) and (b)). It can be noted that, over the entire stretch the *DO* in Ganga was much within the safe limits, although it shows the variations across different *ghats* (religious site), places and banks (Fig. 5(c)).

B. Variation Only at the Bathing Ghats in Haridwar

The variations of different parameters are shown in Fig. 6 using heat maps. From top to bottom, each column has six heat maps showing day-wise variation of pH, conductivity, DO and turbidity, followed by Water Quality Index (WQI), and cluster of data points into good (Cluster 1) and poor (Cluster 2) quality ones. From the plots it can be seen that different bathing places at Haridwar are utilized to different extent. Minor variations along the stretch are clearly visible. As seen in the plots water quality on both banks of the rivers also shows significant variation. This variation is replicated on each bathing day, which clearly shows that one bank of the river is utilized more whereas other bank is utilized less. This is also expected because one bank of the river has *Har-ki-Pauri* (a place most sought after for religious baths) and other ghats, where there is less activity on the other bank of the river. Multiple readings (around 30) were collected at each place and mean of the values was used to create the plot. Different days also shows variations which indicates that each day variable amount of people attend the religious bathing. Our plots show changes which can be reasoned after understanding the effect of bathing as well as the river geological settings.

C. Water Quality Analysis during Ardh Kumbh 2016

Water Quality Index (WQI) defines the overall water quality calculated using several ways depicted in [17](Tyagi 2013). In order to compare the water quality at different places during the study on Ganges, we have calculated the **Weighted Arithmetic Water Quality Index** as shown in [17](Tyagi 2013). The calculation of WQI is influenced by the selection of limits of water quality parameters and their ideal values. The ideal value and limits of water quality parameters used in WQI calculation is as per Table I.

TABLE IV THE INTERPRETATION OF WQI (TYAGI 2013)

WQI Value	Rating of Water Quality	Grading
0-25	Excellent	A
26-50	good	B
51-75	poor	C
76-100	Very poor	D
Above 100	Unsuitable for Drinking	E

The WQI plots (Fig. 6(e)) depicts that relatively poor water quality is observed around *Har-ki-pauri* and rather cleaner water quality is available before and after *Har-ki-Pauri*. The WQI plots show that on different day(s), water quality varies at different Ghats. The variations are clearly highlighted by the heat map of the plots. These finer variations are captured by collecting water quality every second over these places then taking mean of the values.

In order to understand the variation of WQI at different places we used K-means clustering. So as to differentiate between regions having good water quality and poor water quality we computed two centers as shown in plots (Fig. 6 (f). The cluster plots show the places having similar WQI, in same color. The variations are as expected, with the sites which are used less frequently are clubbed together whereas the sites which are used more shows high WQI. The variations on different days are different. Also on each day, WQI varies between different sites as shown by the plots.

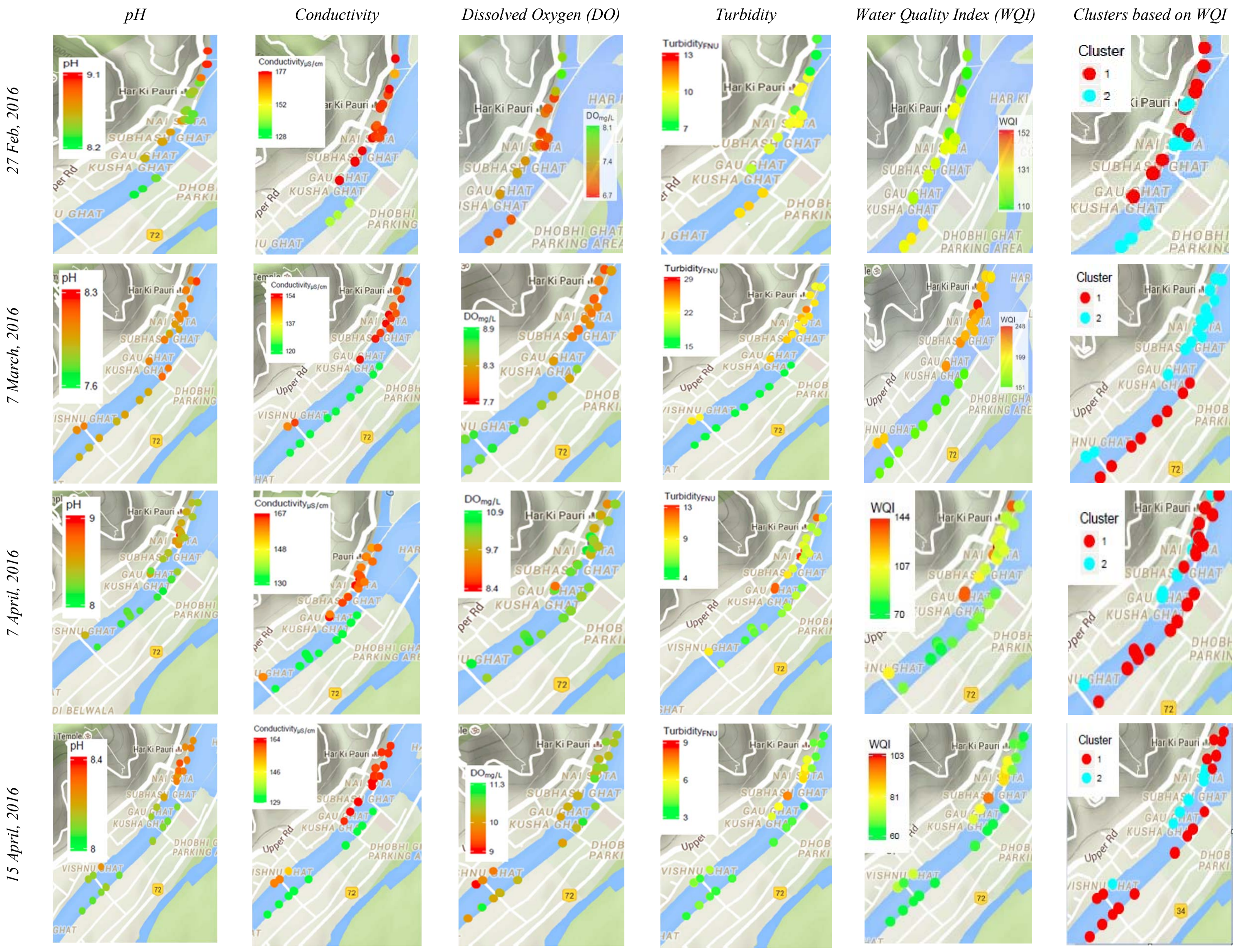


Fig. 6 Heat Maps to Show Day-wise variation of pH, Conductivity, DO, Turbidity, WQI, and Clusters during Ardh Kumbh 2016

V. Discussion and Conclusion

As one uses released data, one needs to overcome a few challenges:

(a) *Overlapping availability of usable pollution data by pollutants, data type and locations*. The data is large by relative volume but the sample size, frequency and measurement process vary over time and may even be under dispute when judged later as technology evolves. Many historical readings are once per month or quarter and the process used may vary. There is a wide diversity in the number of parameters recommended to be measured and the subset actually tested. Our stand is to consider all available data and correlate them (via statistics and machine learning) to establish data quality. Further, use of mobile phone cameras allow capture of qualitative perception of pollution from people which can help validate operation of sensors. This needs combination of techniques that can handle quantitative and qualitative data, respectively.

(b) *Ambiguity of purpose in using water pollution data*. The data publishers do not explicitly and consistently clarify what purpose, like drinking or distillery industries, their data is suitable to be used for. The consuming application should use all annotations where applicable, but should also consider raw sensor data when annotation is missing and label it.

(c) Ambiguity of permissible limits based on a purpose since there may be overlapping standards at local, national and international levels. We have resolved published data by CPCB standard but other standards may be relevant for some users.

(d) There was no easy access to pollution data via mobile apps and APIs, and *GangaWatch* and *BlueWater* fill the gap, respectively.

(e) Reconciling multiple parameters given sparsity of coverage to get aggregate view. In the Khumbh case, we could use WQI to get aggregate view since all parameters of interest were present.

In related work, human activities on rivers in general, and Kumbh festivals on Ganga in particular, have been regularly studied in the past ([2][5][6][7][8][9][10][11]). However, such studies were restricted to taking water samples at select locations and analyzing them in labs for local analysis. Our released data can be used for generating new insights like detailed spatio-temporal outlook of the river or projections at a place over time (as we demonstrated), validating data quality, manage an ongoing event or plan for future ones. A similar study has been carried out by our team [20] to spatially and temporally map river water quality based on in-situ data gathered (in some of India's major rivers) using mobile sensors.

More importantly, other researchers can use available data, both quantitative and qualitative, to create insights and decision aids in more use-cases. Further, they can themselves collect new data and contribute back via APIs to the platform so that it is available to the community.

References

[1] S.S Sandha, , S. Randhawa, and B. Srivastava, , *Blue Water: A Common Platform to Put Water Quality Data in India to Productive Use by Integrating Historical and Real-time Sensing Data*, IBM Research Report RI15002, 2015.

[2] A. Karar, *Impact of Pilgrim Tourism at Haridwar.* Anthropologist, 12(2): 99-105 (2010).

[3] S. Randhawa, et al. "*A multi-sensor process for in-situ monitoring of water pollution in rivers or lakes for high-resolution quantitative and qualitative water quality data*," In 14th IEEE/IFIP EUC, Aug. 24-26, 2016.

[4] H. Kulshrestha, and S. Sharma, *Impact of mass bathing during Ardhkumbh on water quality status of river Ganga*. Journal of Environmental Biology, 27(2): 437-440, 2006

[5] V. Sharma, et al. *Impact of Mass Bathing on water quality of Ganga River during Maha Kumbh-2010.* in Nature and Science of Sleep 10(6):1-5 · Jan. 2012

[6] D.R. Khanna, , et al. *Assessment of water quality of River Ganges during Kumbh mela 2010*. In Environmental Conservation 13(3):165-169, Jan. 2012.

[7] N.K. Arora, , et al. *Analysis of water quality parameters of River Ganga during Maha Kumbha Haridwar, 2010*. In Journal of Environmental Biology 34(4):799-803 · July 2013.

[8] S.K. Singh and A. Bisht, *Environmental Management in Mass Gatherings: A Case Study of Maha Kumbh Mela 2013 at Prayag India*. in–International Journal for Innovative Research in Science & Technology (IJIRST)| 1(7), December 2014.

[9] G. Matta, *A study on physico-chemical Characteristics to assess the pollution status of river Ganga in Uttarakhand*. in Journal of Chemical and Pharmaceutical Sciences (JCPS), Volume 7 Issue 3, July-September 2014.

[10] S. Panwar, et al. *Water Quality Assessment of River Ganga using Remote Sensing and GIS Technique*. in International Journal of Advanced Remote Sensing and GIS, North America, 4, Sep. 2015.

[11] R. Bhutiani, , et al. *DO-BOD modeling of River Ganga from Devprayag to Roorkee, India using BMKB model*. in Pollution, 2(1):25-34, 2016.

[12] CPCB, *Guide manual: water and wastewater analysis*. At: http://cpcb.nic.in/upload/NewItems/NewItem_17_guidemanualw&wwanalysis.pdf , 2014.

[13] CPCB, *Guidelines for online continuous monitoring system for effuents*. In Available at: http://www.inpaper.com/FinalGuidelinse.pdf , 2014.

[14] UK, UK's Bathing Data explorer, http://environment.data.gov.uk/bwq/explorer/index.html, accessed 14 Sep 2016.

[15] Australia, Australia's water quality in New South Wales (NSW), water quality data from continuous monitoring, for one river. http://waterinfo.nsw.gov.au/wq/darling.shtml, accessed 14 Sep 2016.

[16] USA, New York and Hudson's water quality http://www.riverkeeper.org/water-quality/hudson-river/, accessed 14 Sep 2016.

[17] S. Tyagi, et al. *Water Quality Assessment in Terms of Water Quality Index*, In American Journal of Water Resources, 1(3): 34-38, 2013.

[18] J. Ellis, B. Srivastava, R. Bellamy and A. Aaron, *Water Advisor - A Data-Driven, Multi-Modal, Contextual Assistant to Help with Water Usage Decisions*, AAAI 2018.

[19] B. Srivastava, Permissible limits for water usage, https://github.com/biplav-s/water-info, 2018.

[20] P. Hirani et al., *Using Cyber Physical Systems to Map Water Quality Over Large Water Bodies*, 2018 IEEE SENSORS, New Delhi, 2018, pp. 1-4.