Integrated data management to prevent data loss and raise data quality

M. Pichler[1*], D. Camhy[1], A. König[1], D. Muschalla[1]

[1] *Institute of Urban Water Management and Landscape Water Engineering, Graz University of Technology, Austria*

**Corresponding author email: markus.pichler@tugraz.at*

Highlights

- An open-source and modular framework for processing and managing continuous measurement and simulation data is presented.
- The application of an adaptable automatic data validation approach for continuous measurement data can be used for a wide variety of use cases.
- The application for a precipitation gauge network including a weekly automated report is described.

Introduction

As in many scientific disciplines, data of many different types play a central role in urban drainage. Two of them are measurement data which describe the physical state over time and metadata of used sensors describing the attributes of the sensor and other environmental characteristics of the measurement location. Manually managing this data can be overwhelming and often leads to data loss or misinterpretation when original data set, metainformation, and primary source information are no longer available (Sonnenberg et al., 2013).
Measurement data are often used to calibrate and validate models or to monitor a specific system behavior in order to make conclusions, predictions, or decisions. The prerequisite for the most accurate application is high data quality. To achieve this, structured and, if possible, automated data management with error alerting is required to reduce gaps and speed up data validation (Branisavljevic et al., 2010).

Methodology

Data management

The data management presented here is based on the enhanced OpenSDM approach (Open Sensor Data Management, Camhy et al., 2014) which uses existing, sustainable, preferably universal, and open-source technologies. The core components are the time series database (influxDB, 2013), metadata database (FROST-server, 2016), data visualization and alerting tool (Grafana, 2013) and job management and automation tool (Jenkins, 2010). Each tool is open-source software and is already well tested and widely used especially in computer science.
For each measurement project, the project itself, the measurement location and the sensors used are stored in the metadata database as individual objects with unique identifiers, additionally the relationship between these objects are defined (see Figure 1). This metadata includes, for example, measurement range and uncertainties of the sensor, calibration parameters, contact info of the operator, connectivity to the sensor, location, maintenance, or known error period ranges. Metadata is organized using the SensorThings API (Liang et al., 2016), to enable easy data exchange and support long-term use.

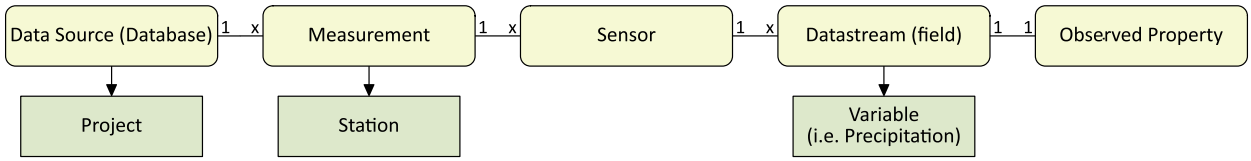


Figure 1. Data structure of the OpenSDM. The annotation of the connector describes the cardinality (i.e., one measurement, which is equal to a station, can have multiple sensors, but one sensor can only be assigned to one measurement at a time)

The raw sensor data is automatically transferred every six hours via a python script using a transfer protocol such as FTP and stored in raw format on a data server. The files are sorted by project and station (i.e., measurement site). Additionally, the data is differentiated between automatically and manually imported as well as measured and simulated data. Another script automatically detects newly added data sets and starts importing the new data into the time series database. During this process, the variables are renamed based on standardized conventions defined in the metadata. The ETL (extract, transfer, load) tasks are scheduled and executed with the automation tool Jenkins. All data in the time series can be visualized web-based. A fast aggregation algorithm facilitates viewing several decades of high-resolution measurements within seconds. It is possible to zoom into periods of interest or set predefined aggregations. An alert function can be used to send e-mail notifications of errors due to sensor failures, connectivity problems, or power failure.

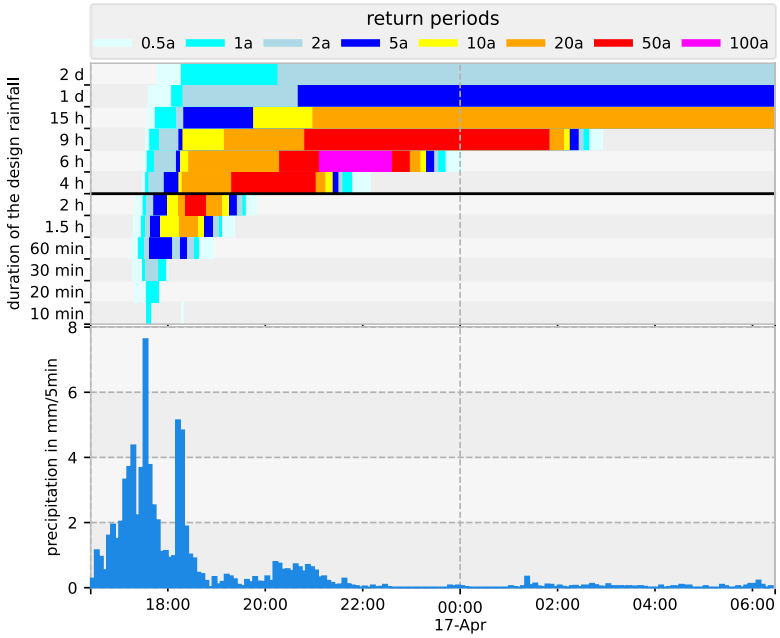
The data gets automatically validated after the import. The validation process includes for example identification of defects, instrument specific and climatological limits, temporal variability check, outlier detection and cross-correlation checks between different stations based on Maier et al. (2020). The validation results are saved as tags in the timeseries database.

Data reporting

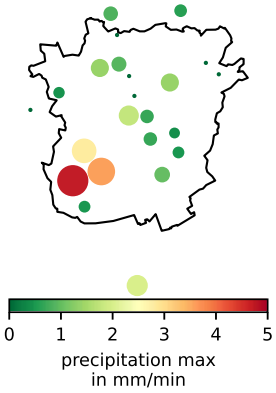
The data reporting is used to communicate with stakeholders and project partners and as preparation for manual validation. It consists of raw data summaries, data validation, and data analysis. Each reporting step is modular to address different requirements for different measurement applications. The incorporation of raw data and metadata allows for additional manual validation. The report is created on a regular basis to keep any data gaps short and to ensure that subjective observations are not too far in the past for additional manual validation.

Results and discussion

An example of the presented automated workflow is the precipitation report for the Graz city measurement network (GCMN—Maier et. al, 2020). The rain gauges of the currently 24 measuring stations of a total of five partners, which are combined in the GCMN, transmit their data to a central server. These data are further transmitted to all project partners in almost real-time and processed within the system presented here. A data report is generated weekly as well as event basis, which is sent as a static pdf via email and can be accessed online via web app. The report for each station includes basic metadata such as database identifier, location, owner, and operator, as well as a listing of recorded values for anomaly detection, observed recording frequency, dry period between rain events, duplicate values, data gaps, validation summary, overall data availability, aggregations, and event analysis (see Figure 2a). In addition, all stations are compared spatially to identify any correlations or anomalies (see Figure 2b).



(a) temporal distribution of a rainfall event and the return periods in relation to the duration



(b) spatial distribution of the precipitation event maximum at each of the 24 measurement locations

Figure 2. Example plots of the weekly precipitation report of the measurement network in Graz, Austria of an event from 16th April 2018 with a return period of greater than 100 years for a duration of six hours.

Conclusions and future work

The presented approach allows to generate data sets in high quality and with as little data losses as possible. Involving project partners, decision-makers, and supporting organizations in data validation and evaluation increases the confidence in the data. In addition, an exchange of experience is triggered which, among other things, standardizes the maintenance of the measuring equipment and thus contributes to a further increase in the quality of the data. Currently, this method is used for the described precipitation report and alerting system and an automated evaluation of a fully equipped combined sewer overflow. For a further project, collection of data for LIDs (Low-impact development) is prepared, where the calculation of water balances will be used to validate the measurements and to analyze the local urban water cycle.

References

Branisavljevic, N., Prodanovic, D., & Pavlovic, D. (2010). Automatic, semi-automatic and manual validation of urban drainage data. Water Science and Technology, 62(5), 1013–1021. https://doi.org/10.2166/Wst.2010.350

Camhy, D., Gruber, G., Steffelbauer, D. B., Hofer, T., & Muschalla, D. (2014). OpenSDM - An Open Sensor Data Management Tool. 11th International Conference on Hydroinformatics. HIC, New York City, USA.

FROST-Server (2016). Fraunhofer open-source SensorThings-server. Fraunhofer IOSB. https://github.com/FraunhoferIOSB/FROST-Server (accessed 04 August 2021)

Grafana (2013). Grafana Labs. https://grafana.com/oss/grafana/ (accessed 04 August 2021)

InfluxDB (2013). InfluxData. https://www.influxdata.com/time-series-platform/influxdb/ (accessed 04 August 2021)

Jenkins (2010). Jenkins. https://jenkins.io (accessed 04 August 2021)

Liang, S., Huang, C.-Y., & Khalafbeigi, T. (2016). OGC SensorThings API Part 1: Sensing. Open Geospatial Consortium. http://docs.opengeospatial.org/is/15-078r6/15-078r6.html

Maier, R., Krebs, G., Pichler, M., Muschalla, D., & Gruber, G. (2020). Spatial Rainfall Variability in Urban Environments—High-Density Precipitation Measurements on a City-Scale. Water, 12(4), 1157. https://doi.org/10.3390/w12041157

Sonnenberg, H., Rustler, M., Riechel, M., Caradot, N., Rouault, P., & Matzinger, A. (2013). Best data handling practices in water-related research. Water Practice and Technology, 8(3–4), 390–398. https://doi.org/10.2166/wpt.2013.039