Earth Syst. Sci. Data Discuss., referee comment RC2
https://doi.org/10.5194/essd-2021-8-RC2, 2021

Comment on essd-2021-8

Anonymous Referee #2

Referee comment on "The Bellinge data set: open data and models for community-wide urban drainage systems research" by Agnethe Nedergaard Pedersen et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2021-8-RC2, 2021

Summary

The submission is a timely contribution in the field of urban drainage, making use of the emerging possibility to publish articles on original research data. Presented data (and models) are expected to enable researchers to re-evaluate difficult-to-obtain data in context of urban drainage modelling, to study the influence of different sources of precipitation on hydraulic simulations, and to apply different data analysis techniques, e.g. to detect anomalous sensor recordings.

The manuscript is well structured and well written. A short review explains the underlying motivation for sharing data and models; reference is given to previously published data publications. The main part describes diverse data sets and corresponding hydraulic models related to a small Danish combined sewer network. The study area and its urban drainage system are very well described. The authors provide supplementary information on sewer infrastructure and hydraulics (models; photo documentation; very illustrative drawings on structures; geodata) allowing for an adequate interpretation of the complex drainage situation. Individual data sets are explained in the text and in a separate document accompanying the data package.

While it is clearly acknowledged that collecting and compiling this data set has been a great effort, and a publication of topic and data is principally recommended, I see the following key points that need to be addressed in a revised version:

- Data ownership: I am wondering if data, that is publicly available anyway (meteo data, topographic data) should be - at least - specifically labelled in order to allow a differentiation from data collected on purpose, such as in-sewer observations. It should be discussed (also in the community) how this should be handled, i.e. it needs clear

statements to clarify data ownership of original data.

- Hydraulic models: The authors attribute large parts of the manuscript (7 of 24 pages) to the comparison of two hydraulic model implementations that describe the same case study system (one being a modified export of the original). While comparing the effect of different conceptual approaches for surface runoff models may generally be an interesting aspect, the key focus (which is the data set) is - in my opinion - unnecessarily diluted by elaborating upon structural model uncertainty. I suggest streamlining the study here. This could be accomplished either by focusing on one model implementation only, by outsourcing the model comparison, and/or by discussing the models usefulness, e.g. to check the plausibility of observations.
- Only vague information is given on what to do with the presented data and models. The very last paragraph provides a glimpse and mention the "great potential in using data to a much greater extent than previously. Provide more concrete examples, i.e. ideas how to utilise the data. This should support/illustrate the value, uniqueness and usefulness of this research data publication.

Please find more elaborated comments, split in major and minor aspects, below.

Major points:

Handling the data and meta-data:

Data is provided in nine (9) individual packages (ZIPs) through a university hosted research data repository. Downloading, sorting, and renaming inconsistently named data packages takes a while. This should be organised in a more stringent manner, the file naming should revised and occasional redundancies be eliminated.

Ownership of the data

The data itself is actually a compilation of various data sets, of which some are acquired in own or contracted field measurement campaigns (water level, flow sensors, i.e. sensor data - #2; CCTV data - #5), some data stem from publicly available sources (radar data: VCS Denmark; Orthophoto, Digital Terrain Model: SDFE), or from sources where data typically need to be purchased (rain gauge data - Danish Meteorological Institute). One question I would like to put up for discussion here: is it scientifically innovative to publish compilations of different data sets that are, on the one hand, available (anyway) and, on the other hand, selectively undercut with own or specifically contracted field measurements?

Usability of data and models

There is a very short and unspecific section on what to do with this data in the beginning (line 81 ff.) and a more concrete paragraph in the conclusion (493 ff). The latter section would deserve a more in-depth elaboration in a previous chapter. More concrete examples should be provided to illustrate the value, uniqueness and usefulness of this data publication. For instance, what is the added value of providing CCTV inspection data?

Missing meta-data on sensor readings

In line 250 ff it is stated that "Exact documentation of sensor maintenance has not been a high priority over all the years, and it is therefore presently not possible to give an overview of when and where sensors have been repaired, replaced or received some sort of maintenance." That is, meta-data or log files are not provided. Comments such as, "The 0-point may have changed during the years, and there is no log-file with changes in SCADA settings in System2000…." (line below the Fig. 1 caption in Sensordata.pdf) are honest, but not very helpful. This is a drawback, which clearly limits the possibilities to interpret in-sewer sensor data. In a didactical example, the authors indeed provide two cases, which illustrate how this can effect sensor data interpretation. But, how can data from other sensors be interpret if I do not know zero-point has changed?

SWMM and MU model

A large part of the manuscript (7 of 22 pages) actually focuses on model-related issues, i.e. it discusses effects of conceptual differences in two urban drainage models. Since most data users would use the SWMM model for simulations, my comments mainly relate to the SWMM model implementation. While my general comment on the model comparison in the summary section remains the major point of critique, the following model-related aspects appear odd and need some clarification:

- It is not clear why two different versions of the MU model are provided. If the "Mike Urban model of the system anno 2020" represents the system "as it looked medio 2020, but it is a good representation of the system from 2010 and onwards.", and no significant land use changes were observed/assumed (cf. line 73 f.) it is not clear what the user should do with the old SWMM model implementation. In order to avoid ambiguities I suggest excluding irrelevant data and model files, such as the old model version. If still relevant, please explain why.
- In the model description document (pdf) it is mentioned that "the [SWMM model] parameters for the infiltration is currently set extremely high so that infiltration from green areas will not appear.". This is most likely a typo, since the sentence does not make sense in this context. Please correct the typo. NB. Generally it can be stated that tweaking the parameters of the SWMM implementation in such a way that runoff-efficient areas are reduced to only impervious areas can be critical when having an

average degree of imperviousness of 35 % (as it is the case here). It could further be discussed how this effects simulation results.

- In terms of plausibility of the SWMM model performance: a moderate rain event of about 13 mm h-1 leads to flooding of several nodes in the network (event early of 29-Jun-2012). Either the system is poorly designed, or the model is hydraulically incorrect. The potential overestimation of the flooding activity may also be discussed in the context of other peculiarities identified.
- Sewer infiltration is completely neglected, despite the fact an internal report says that it makes up 30 % of the hydraulic loading at the catchment outlet. This is a significant share, which IMO cannot be neglected when considering the model performance (dry and wet weather). Can you provide more information from the internal report on how the 30 % infiltration is quantified? Could you describe unsuccessful attempts to implement infiltration in the models? This could be useful for data users when trying to find alternative solutions.
- SWMM Infiltration parameters are tweaked to the extreme to match observations. At line 330 ff, the authors howver state "VCS has a philosophy of transparency in models, where understanding the system behaviour is more important than ensuring a perfect calibration with non-transparent parameter sets, meaning that VCS does not want to tune conceptual parameters to unrealistic values in order to fit models to observations.". While this is sustainable opinion, it is somewhat contradictive to the parameter tweaking. Please clarify!

Minor points:

- Data validation - Chapter 3.4: the section on data cleaning could be more elaborative; reference should be given to existing works, e.g. (Leigh et al. 2019). Five different methods are explained for data validation (cleaning), whereas two of them are subjective (“manual remove”; outlier detection for interim Danova sensor). Furthermore, it remains unclear why only gaps shorter than 5 minutes have been interpolated, why not up to 10, 15 minutes? Why would it be necessary to interpolate them at all?
- Meteorological variables from DMI should be referred to in Ch. 3 since these can also be considered as “observations”. Showing the 10 year time series in Fig. 2 illustrates the availability but has no added value.
- Figure 7: it is not clear what type of sensors the GF73F010 and GF72F040 are. Please specify in the Y-axis or caption.
- Figure 8: inconsistent caption formatting.
- Sentences like the one in line 330 ff. are rather opinions than solid research results. Please consider rewriting these sentences (without changing the valuable meaning) and provide references, if possible.
- Descriptions for some data sets are very sparse, for others they are sufficiently comprehensive. Some of the accompanying documents are somewhat sloppily prepared and need revision (e.g. Models.pdf).

References:

Leigh, C., Alsibai, O., Hyndman, R.J., Kandanaarachchi, S., King, O.C., McGree, J.M., Neelamraju, C., Strauss, J., Talagala, P.D., Turner, R.D.R., Mengersen, K. and Peterson, E.E. (2019) A framework for automated anomaly detection in high frequency water-quality data from in situ sensors. Science Of The Total Environment 664, 885-898.