Proc. IAHS, 370, 9–14, 2015
proc-iahs.net/370/9/2015/
doi:10.5194/piahs-370-9-2015

Opportunities for multivariate analysis of open spatial datasets to characterize urban flooding risks

S. Gaitan and J. A. E. ten Veldhuis

Department of Water Management, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Stevinweg 1 room 4.75, 2628CN, Delft, the Netherlands

Correspondence to: S. Gaitan (s.gaitan@tudelft.nl)

Received: 11 March 2015 – Accepted: 11 March 2015 – Published: 11 June 2015

Abstract. Cities worldwide are challenged by increasing urban flood risks. Precise and realistic measures are required to reduce flooding impacts. However, currently implemented sewer and topographic models do not provide realistic predictions of local flooding occurrence during heavy rain events. Assessing other factors such as spatially distributed rainfall, socioeconomic characteristics, and social sensing, may help to explain probability and impacts of urban flooding. Several spatial datasets have been recently made available in the Netherlands, including rainfall-related incident reports made by citizens, spatially distributed rain depths, semidistributed socioeconomic information, and buildings age. Inspecting the potential of this data to explain the occurrence of rainfall related incidents has not been done yet. Multivariate analysis tools for describing communities and environmental patterns have been previously developed and used in the field of study of ecology. The objective of this paper is to outline opportunities for these tools to explore urban flooding risks patterns in the mentioned datasets. To that end, a cluster analysis is performed. Results indicate that incidence of rainfall-related impacts is higher in areas characterized by older infrastructure and higher population density.

1 Introduction

Cities are vulnerable to rainfall flooding risks; rainfall can affect electrical installations, household contents, road traffic, private and public assets, and business activities (ten Veldhuis et al., 2011; Ashley et al., 2005). Therefore, adaptation measures to better cope with those risks are required. Smart instrumentation and drainage maintenance (e.g. Gaitan et al., 2014; ten Veldhuis and Clemens, 2011), and emergency protocols (e.g. Melo et al., 2015), as well as retrofitting of existing infrastructure and urban redevelopment (e.g. Jacobs, 2012), are examples of such measures. However, their effectiveness depends on the available knowledge of the mechanisms leading to damage after heavy rains, which are still not completely understood (Gaitan et al., 2015; Spekkers et al., 2013, 2014; ten Veldhuis et al., 2011).

Previous studies indicated that currently implemented drainage models do not provide realistic predictions of local flooding occurrence during heavy rain events (Fontanazza et al., 2011; Ochoa-Rodriguez et al., 2014; Gaitan et al., 2015). Studying the spatial distribution of possible explanatory variables may help to explain how a heavy rainfall event triggers occurrence of flooding impacts in a city. Recent works have analyzed whether occurrence of citizens' complaints and insurance claim reports on rainfall-related incidents could be explained by variability in urban topography (Gaitan et al., 2015) and rainfall intensity (Spekkers et al., 2013). Response and explaining variables in those studies were selected a-priori. Their results show that for the considered cases of delta city conditions, pluvial flooding impacts cannot be explained solely by rainfall intensity or urban topography. The flat geography of delta areas may be part of the reason of the low explaining power of meteorological and topographic variables; this emphasizes the importance of exploring additional factors to further model urban flooding risks.

Spekkers et al. (2014) used a decision tree analysis for exploring the extent in which different environmental and socio-economic characteristics explain variances in insurance claim data due to rainfall-related damage. Due mostly to privacy issues, accessing them required explicit agreement with government and private parties. The decision tree ex-

plained around 25 % of variance in claim occurrence, improving beyond the 10 to 20 % of explaining power obtained via multiple regression models. Remaining unexplained variance was suggested to be related to a coarse level of spatial aggregation during the analysis and to the possible lack of key explanatory variables, which were not included in the study.

Multiple information sources describing urban characteristics have recently become publicly available under an Open Data policy in the Netherlands (Dutch Ministry of Interior and Kingdom Relations, 2014). This offers the opportunity to explore the potential of such data to provide insights about pluvial flood risks. Due to privacy issues, socioeconomic data in these sources has been aggregated, but their open availability enables flexible use of combinations of datasets, models, and different spatial aggregations, without the need of special permits of private or public parties.

These multivariate spatial datasets can be used to investigate whether certain urban characteristics trigger reports on rainfall-related incidents, which serve as a proxy to measuring flood impacts.

Multivariate spatial data analysis is widely applied in the field of ecology. In the field of community and landscape ecology, for instance, the main matrix of analysis is composed by sampled objects in the rows, and frequency or abundance of different species, or measurement of environmental variables, in the columns. Clustering, ordinations, and multiple regressions are part of the techniques used to analyze such matrices. Cluster analysis is used to classify associations between sampling sites, or between the composition of species inhabiting such sites. Ordination techniques, such as principal coordinates analysis (PCoA), are used to interpret community changes along environmental gradients in complex multivariate datasets. Regression analysis on singles species is also used to quantify possible relationships between species and environmental variables (Jongman et al., 1995).

While those methods are used by ecologists to characterize and model the environmental conditions in which species flourish, these techniques can be applied to open spatial data to explain occurrence of urban flooding impacts. A set of those methods is presented in this paper and their potential for flood risk analysis is discussed. Methods include spatial indexing, classification dendrograms, ordination via multidimensional scaling, and fitting of multiple regressions on the complaints.

This article is organized as follows: Sect. 2 explains the goals and procedures of the exploratory methods, results of the application of one of those methods are discussed in Sect. 3, and Sect. 4 draws conclusions and provides an outlook for the use of the described methods.

2 Information sources and multivariate analysis of spatial datasets

In order to explain how the methods presented in this paper can operate in big sets of open spatial data for the purpose of flood risk analysis, a number of publicly available datasets are considered. Datasets are collected for the case of a heavy rain event that hit Amsterdam on 28 July 2014, causing considerable flood damage. The data available includes socioeconomic, cadastral, and meteorological data. Reports about rainfall related incidents are also available, and can be used as indicators of urban flooding incidence. Data-sources and methods are described in the following subsections. An overview of data characteristics is shown in Table 1.

2.1 Maximum rainfall intensity data

Rainfall intensity measurements are based on a system of two C-band Doppler weather radars operated by the Royal Netherlands Meteorological Institute (KNMI, 2013). Rainfall depths, observed at 1.5 km above the ground, are provided with a temporal resolution of 5 min, on a grid of 1 km^2 spatial resolution with a custom geographic projection (Overeem et al., 2009). Information is available through a FTP server. Rainfall intensity was calculated at 15 min timestep. The highest intensity per radar cell during the rainfall event was used in Sect. 2.5.

2.2 Socioeconomic and cadastral data

Available socio-economic statistics include information about population density, age, and income. These datasets are updated every year. They also include housing characteristics and market-prices for the period 2012–2013. Statistics are provided on a geospatial vector data format: they are aggregated into hectare and 0.25 km^2 grids to avoid privacy issues (Centraal Bureau voor de Statistiek, 2013). Available cadastral information describes the floor area, geographic location, address and postal code, purpose, and age of construction of buildings in The Netherlands. Each building is identified by a unique number, and its building boundaries are provided on a geospatial vector data format. This cadastral database is updated monthly (Kadaster Nederland, 2013). Some areas in these datasets are masked so that general or no information is available, due to security or privacy reasons. Both datasets have been made available as part of the open data policy mentioned in Sect. 1.

2.3 Reports about rainfall-related incidents

The municipality of Amsterdam provides platforms for registering citizen complaints about incidents in the city, including those related to urban flooding impacts. Reports are recorded by personnel from the municipality, and include a unique identifier, a timestamp, address of reported incidents,

Table 1. Data sources used in this study. Total number of data points and mean refer only to case study area. Points with secret or not available data have been excluded from this table. Only buildings in use have been considered.

Variable	Spatial and temporal resolution	Data points	Metric and unit	Mean
Max. rainfall intensity	1 km^2, every 5 min	292	$mm\,h^{-1} \times km^{-2}$	40.1 ± 20.5
Inhabitants	1 Ha, static – 2013	6127	Individuals/Ha	131.2 ± 82.7
House market prices	0.25 km^2, static – 2012	6127	Average price (thousands of EUR) 0.25 km^{-2}	275.9 ± 155.9
Proportion of high income	0.25 km^2, static – 2012	453	% of homeowners earning at least EUR 46 950 yr^{-1}	45.2 ± 14.8
Proportion of low income	0.25 km^2, static – 2012	453	% of homeowners earning less than EUR 25 070 yr^{-1}	20.6 ± 12.9
Age of construction	Single building polygon, static – 2012	234 736	Years since built	105.6 ± 217.6
Rainfall-related incidents	Single address points, time-stamped	340	Register of a telephonic call including address	NA

an initial classification into general types of rainfall-related incidents (e.g. blocked gully, flooded basement, overloaded sewer) and a short description of the reported incident. Municipalities can be requested to provide these reports for scientific research.

2.4 Defining sample size and indexing spatial units

Data pre-processing starts with the definition of the spatial units of analysis or "sample sites". The selection of the shape and extent of each site was done according to three purposes: having a comparable area in every unit, facilitating geometric calculations and data re-arrangement, and keeping a balance between fine spatial resolution and availability of data. Using the grid of the radar data satisfies the stated purposes. Grid cells are simple geometries that do not demand too complex computations during spatial queries. As the Amsterdam area is around 200 km^2, the cell size in this grid (1 km^2) is big enough for ensuring that the number of resulting sampling sites is sufficient for performing statistical analysis. Higher resolution grids, on the other hand, result in finer units that can suffer from data scarcity: areas with masked socioeconomic and cadastral data can cover entire cells if they are too small. Also, if sampling cells are too small, the chances of having cells with none or very low occurrence of rainfall-related incidents become higher. The convenience of setting the 1 km^2-cell grid to sample the area of study can be further evaluated with a cluster analysis: if the aggregation is excessive, differences between samples become blurred, resulting in groupings with extremely high measures of similarity. Cluster analysis is described in the following section.

A second step in the pre-processing is to perform spatial queries and averages in each of the variables listed in Table 1 to determine their mean status in each sampling unit. The queries are used to intersect realizations of the variables with their underlying cells. The average value of the intersected realizations is assigned to each cell. Given the size of handled data, the implementation of a spatial index is required to reduce computational demands. As it greatly reduces the number of intersections that need to be made during the sampling of variables realizations at different sites, R-Tree indexing (Guttman, 1984) was applied to the sample sites in this study. Finally, the resulting realization averages per cell are stored in a matrix in which rows represent sampling sites and columns each of the variables.

Before describing the rest of the techniques presented by this paper, it is worthy to make some considerations about the usage of the variables described in Sects. 2.1 to 2.3. Reports of rainfall-related incidents are taken as a proxy to urban flooding impacts. For this reason the occurrence of those reports are considered the response variable in this study. The rest of the variables are assessed as explanatory variables.

2.5 Cluster analysis: grouping sites according to available data

Sampled sites can be grouped in terms of their similarities using a cluster analysis. Groups produced by this type of analysis are characterized by the similarity of values of the different variables. The goal of this classification is to discretize areas that share urban conditions, and to explore whether some of the obtained groups are more prone to rainfall-related impacts. The grouping is done in the matrix obtained according to Sect. 2.4, excluding the column of reports from the matrix, and considering only the sites with observed reports of incidents.

Cluster analysis can be done via a hierarchical grouping that maximizes similarities within groups, and differences between groups. The result of a hierarchical grouping applied to the sites is a tree-like structure, or dendrogram, displaying how close sites are to each other in terms of the values of the multiple variables under analysis (Jongman et al., 1995; Legendre and Legendre, 2012).

Clustering requires the computation of a square similarity matrix indicating how close are sites to each other in terms of all the considered variables. Selecting a distance metric

depends on the nature of data. For instance, non-parametric Spearman's r can be used in ranked variables with monotonic relationships, Jaccard or Sorensen for presence-absence data, Hamming distance for categorical data coded as integers, Euclidean or Bray-Curtis for measured or counted data, and Gower's coefficient for mixed data types (Legendre and Legendre, 2012; Gower, 1971). As the assessed variables in this study (see Table 1) have been measured, Spearman's r can be used to assess a distance based on simple correlations.

The clustering process can be made by setting weights on the objects being classified, and by establishing a priori similarity thresholds to define groups. Without previous knowledge about differences on the importance of sampling units or variables, the unweighted pair group method with arithmetic average (UPGMA) is preferred over the weighted pair group (WPGMA) or K-means methods. In UPGMA, the distance between groups is the average of distances between subgroups. The UPGMA clustering is built bottom-up, from the tips. The first group includes the two objects with the smallest distance. After this, the average distance from the two grouped objects to all the remaining ones is calculated. Then, a new group is made including the two objects or groups having the smallest updated distance. These tasks are iterated until all objects are grouped. Branching in the resulting dendrogram indicates the distance shared by pairs of objects or groups (Legendre and Legendre, 2012).

For the case of flood risk analysis, a likely outcome of a cluster analysis applied to different urban conditions would group similar sites into branches sharing urban characteristics such as population density, income, or infrastructure age. If such grouping also happens to differentiate groups of sites with remarkably different report incidence, which is not included in the clustering computation, then the environmental configurations with higher urban flooding risks can be discriminated. An expected result is that groups characterized by sites with higher urban densities, and thus bigger number of inhabitants, display higher incidence of reports (Gaitan et al., 2015). Even though the focus of this paper is to present an overview of methods to explore multivariate spatial datasets, results of the clustering test are discussed in Sect. 3.

2.6 Principal coordinate analysis: inspecting dimensionality reductions

Apart from the clustering of sample sites, patterns in the distribution of contextual variables can be projected on a Cartesian space of reduced dimensionality using Principal Coordinates Analysis (PCoA). This allows us to observe which variables explain most of the data variation and which ones are redundant. The representation produced by a PCoA is similar to the more well-known Principal Components Analysis (PCA).

The two methods differ in the way in which synthetic axes are calculated. In PCA, the computation of eigenvectors (i.e. the Components) is done assuming that Euclidean distances can be calculated directly from input data, and that analyzed variables correlate linearly. In a set of data having types different than quantitative (e.g., ranked or frequency data), PCA cannot be directly applied. PCoA, on the other hand, can be computed using different distance indexes (see indices described in previous section). Those distances are projected in Cartesian planes, and the resulting distance between all pairs of projected points is used to compute eigenvalues and eigenvectors (Legendre and Legendre, 2012; Jongman et al., 1995).

PCoA can visualize whether the incidence of rainfall-related impacts follows one of the theoretical environmental gradients proposed by the Coordinates. If so, the next step is to look if any of the analyzed variables is mostly aligned along that gradient. For example, one could expect that rainfall intensity aligns with one of the Principal Coordinates. One might also expect that points of average population density and building market prices gradients are displayed close to each other, suggesting some collinearity and redundancy. The distance index used to calculate the PCoA must always be considered when interpreting the resulting eigenvectors: the relationship between two variables of mixed data type might imply that the relations are not as simple as a correlation between two quantitative variables.

2.7 Multiple regression: identifying explaining factors

Multiple regression explores relations between a response variable and several explanatory variables. Among the methods presented in this paper, it is probably the most widely known. Multiple regression aims at describing possible links between explanatory and response variables, with a certain margin of error: the method can be used for estimating parameters ranks in which a response variable is more likely to occur. It also indicates the contribution of different explanatory variables to the variance of the response variable.

Special attention must be given to the type of data being analyzed as the method varies depending on it. Quantitative data can be modeled with straight lines, parabolas and Gaussian curves, using least-squares regressions and analysis of variance (ANOVA). Regressions on presence-absence data are modelled with logit models, using chi-tests, maximum-likelihood principle, and Gaussian logit (Jongman et al., 1995).

Results of multiple regression analysis should present some agreement with ordination analysis: the variables explaining most of the response variability in multiple regression (e.g. rainfall intensity) should be the ones most closely aligned to the main eigenvectors. Disagreements can arise, though, depending on the distance indexes used in the PCoA. The total variance explained by the significant parameters modeled in the regression determines how much variability it is explaining, indicating whether additional parameters need to be explored. In other words, multiple regression checks

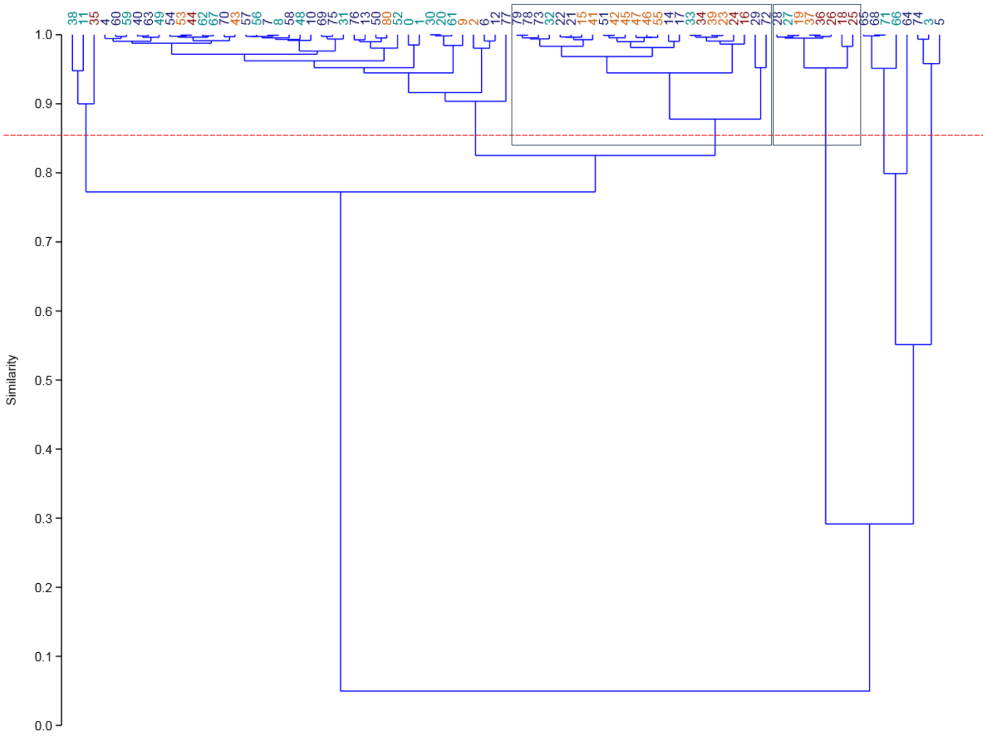

Figure 1. Cluster analysis of sampled sites. Groping is made via UPGMA.

for useful variables, and prompts for further exploration of variables if not enough response variance is explained.

3 Results and discussion of cluster analysis

After performing the aggregations described in Sect. 2.4, rainfall-related incidents were observed in a total of 81 sampling units. The cluster analysis was applied to these 81 sites. A first view of the resulting dendrogram (see Fig. 1) reveals the structure of similarities between analyzed sites. Being above 95 % of similarity, groupings at the leaves indicate that values of variables do not change drastically from one km^2 sampling unit to another at that branching level. However, branching at the root starts at only 5 % similarity. Two small groups, made of 8 sites each, were classified with a similarity of just 30 %. One of those groups is made of sites located in the Amsterdam city center. The remaining group is only branched at 70 % similarity, which means that the 65 sites in this group are closely related to each other. This suggests that the selected level of spatial aggregation is able to describe main differences between sites. To better differentiate groups made above 85 % similarity, either higher granularity on the spatial aggregation or additional variables are required though.

If the dendrogram is cut at 85 % of similarities (see red, dashed horizontal line in Fig. 1), seven groups are differentiable. Viewing them from left to right, third and fourth group (enclosed in black boxes) present a particularly high incidence of rainfall-related reports. The group in the right box is the one including eight sites from the city center. On the other hand, the group in the left box includes 24 sites from diverse city locations. The city center group is highly differentiated from the rest, and accounts for 80 incident reports: with 10 % of the sites it accounts for 24 % of the rainfall-related incidents. The left group box includes 127 reports which equal 37 % of incidents. The incidence of rainfall-related impacts is aprox. 5 km^{-2} for this group, and 10 km^{-2} for the city center group. Apart from these two groups the biggest remaining one includes 38 sites and 96 reported incidents; incidence of rainfall-related impacts is thus aprox. 3 km^{-2}. These results indicate that the configuration of environmental conditions in the city center group is accompanied by a relatively high incidence of incident reports.

When the average conditions in the three mentioned groups are compared, it is clear that the building age in the city center group ($\overline{x}$: 682.7 years, s: 174.4 years) is much higher than in the other two mentioned groups ($\overline{x}$: 62.9 and 52.0 years, s: 27.6 and 24.7 years, for the group in the left box and the mentioned biggest group respectively). The latter groups are mostly differentiated by their population density ($\overline{x}$: 154.5 and 78.7 Ha^{-1}, s: 47.5, 33.3 Ha^{-1}). Other variables do not differ greatly. These results indicate that urban environments characterized by older infrastructure and higher population density are more prone to rainfall-related incidents.

4 Conclusion

This paper presented open spatial data sources and multivariate exploratory methods that are available for research on urban flooding risks. These methods consisted of spatial indexing, clusters analysis, PCoA, and multiple regression. Spatial aggregation and indexing of data allow us to compile a matrix of sites and variables, in which the heterogeneous original information, which has references about geographic locations, can be rearranged for analysis. Cluster analysis using UPGMA provides a characterization of urban areas under study. PCoA reduces the dimensionality of multiple, heterogeneous datasets, suggesting the presence of environmental gradients. It also indicates which of the variables from the available explanatory variables, explains most of data variability. Multiple regression fits models to the explanatory variables to explain response variables, providing measures of the amount of explained power. The application of a cluster analysis to the available data indicated that the incidence of rainfall-related incidents is higher in areas characterized by older infrastructure and higher population density. Such information is useful for designing and implementing proper adaptation measures against urban flooding. Results from cluster analysis can be complemented by applying a PCoA and a multivariate analysis to the data, which will be the matter of future research.

Acknowledgements. We kindly thank the city of Amsterdam, for providing their complaint register for the purpose of this study, and Climate-KIC, for providing financial support for this research.

References

Ashley, R., Balmforth, D., Saul, A., and Blanskby, J.: Flooding in the future predicting climate change, risks and responses in urban areas, Water Sci. Technol., 52, 265–273, 2005.

Centraal Bureau voor de Statistiek: Statistische gegevens per vierkant – Statistischegegevenspervierkantupdatejuli2013.pdf, Kaart met statistieken per vierkant van 100 bij 100 meter, available at: http://www.cbs.nl/NR/rdonlyres/661D884F-CF5B-4192-8138-EA959D540EFE/0/Statistischegegevenspervierkantupdatejuli2013.pdf (last access: 26 March 2014), 2013.

Dutch Ministry of Interior and Kingdom Relations: Open Data NEXT in English – Data.overheid.nl: het opendataportaal van de Nederlandse overheid, Data.overheid.nl: het opendataportaal van de Nederlandse overheid [online] available at: https://data.overheid.nl/english (last access: 1 December 2014), 2014.

Fontanazza, C. M., Freni, G., La Loggia, G., and Notaro, V.: Uncertainty evaluation of design rainfall for urban flood risk analysis, Water Sci. Technol., 63, 2641–2650, 2011.

Gaitan, S., Calderoni, L., Palmieri, P., Ten Veldhuis, M.-C., Maio, D., and van Riemsdijk, M. B.: From Sensing to Action: Quick and Reliable Access to Information in Cities Vulnerable to Heavy Rain, IEEE Sensors J., 14, 4175–4184, doi:10.1109/JSEN.2014.2354980, 2014.

Gaitan, S., ten Veldhuis, J. A. E., and van de Giesen, N. C.: Spatial distribution of rainfall-related complaints along urban overland flow-paths in review, Water Resour. Manage., 2015.

Gower, J. C.: A general coefficient of similarity and some of its properties, Biometrics, 27, 857–871, doi:10.2307/2528823, 1971.

Guttman, A.: R-trees: a dynamic index structure for spatial searching, in: Proceedings of the 1984 ACM SIGMOD international conference on Management of data, New York, USA, June 1984, 14, 47–57, 1984.

Jacobs, J. C. J.: The Rotterdam approach: connecting water with opportunities, in Water Sensitive Cities, edited by: Howe, C. and Mitchell, C., IWA Publishing, 2012.

Jongman, R. H. G., Braak, C. J. F. T., and van Tongeren, O. F. R.: Data Analysis in Community and Landscape Ecology, Cambridge University Press, 1995.

Kadaster Nederland: BAG, Basisregistraties Adressen en Gebouwen (BAG), available at: https://www.kadaster.nl/bag (last access: 26 March 2014), 2013.

KNMI: KNMI radar gegevens, available at: http://www.knmi.nl/datacentrum/catalogus/catalogus/catalogus-gegevens-overzicht.html (last access: 12 February 2014), 2013.

Legendre, P. and Legendre, L. (Eds.): Chapter 8 – Cluster analysis, in: Numerical Ecology, vol. 24, 337–424, Elsevier, available at: http://www.sciencedirect.com/science/bookseries/01678892/24 (last access: 14 January 2015), 2012.

Melo, N., Santos, B. F., and Leandro, J.: A prototype tool for dynamic pluvial-flood emergency planning, Urban Water J., 12, 79–88, doi:10.1080/1573062X.2014.975725, 2015.

Ochoa-Rodriguez, S., Wang, L.-P., Gires, A., Pina, R. D., Rondinel, R. R., Bruni, G., Ichiba, A., Gaitan, S., Cristiano, E., van Assel, J., Kroll, S., Murla-Tuyls, D., Schertzer, D., Tchiguirinskaia, I., Onof, C., Willems, P., and ten Veldhuis, J. A. E. M.-C.: Impact of spatial and temporal resolution of rainfall inputs on urban hydrodynamic modelling output: A multi-catchment investigation, in review, J. Hydrol., 2014.

Overeem, A., Holleman, I., and Buishand, A.: Derivation of a 10-year radar-based climatology of rainfall, J. Appl. Meteorol. Clim., 48, 1448–1463, doi:10.1175/2009JAMC1954.1, 2009.

Spekkers, M. H., Kok, M., Clemens, F. H. L. R., and ten Veldhuis, J. A. E.: A statistical analysis of insurance damage claims related to rainfall extremes, Hydrol. Earth Syst. Sci., 17, 913–922, doi:10.5194/hess-17-913-2013, 2013.

Spekkers, M. H., Kok, M., Clemens, F. H. L. R., and ten Veldhuis, J. A. E.: Decision-tree analysis of factors influencing rainfall-related building structure and content damage, Nat. Hazards Earth Syst. Sci., 14, 2531–2547, doi:10.5194/nhess-14-2531-2014, 2014.

ten Veldhuis, J. A. E. and Clemens, F. H. L. R.: The efficiency of asset management strategies to reduce urban flood risk, Water Sci. Technol., 64, 1317, doi:10.2166/wst.2011.715, 2011.

ten Veldhuis, J. A. E., Clemens, F. H. L. R., and van Gelder, P. H. A. J. M.: Quantitative fault tree analysis for urban water infrastructure flooding, Struct. Infrastruct. E., 7, 809–821, doi:10.1080/15732470902985876, 2011.