Project

Delft Measures Rain

Analysing patterns in the rainfall that occurred in the city of Delft during the summer of 2020

By

Illias Timori

in partial fulfilment of the requirements for the degree of

Bachelor of Science

in Civil Engineering

at the Delft University of Technology

Supervisors:	Marie-Claire ten Veldhuis	TU Delft
	Martine Rutten TU Delft	TU Delft
	Sandra de Vries	TU Delft

Preface

This document is a thesis in the partial fulfilment for the title of Bachelor of sciences in Civil Engineering from the Technical university of Delft. It comprises my research performed at the faculty of Civil Engineering in collaboration with smartphones4water, which intends to integrate and apply the skills and knowledge gained during the Bachelor Civil Engineering at the Delft University of Technology. I would like to thank Marie-Claire ten Veldhuis, Sandra de Vries and Martine Rutten for their guidance, support and advises during the realization of this research and report.

Abstract

Delft Measures Rain is a collaboration project between Smartphones4water(S4W), Delft's technical university and the municipality of Delft. This research aims to elucidate patterns in the rainfall that occurred in the city of Delft during the summer of 2020.Ultimately, this project aims to have a better insight into rain patterns in the municipality of Delft and to communicate back to the citizen scientists what the results of their hard work are. This was achieved by analysing the results of the Citizen Science rain measurements provided by the citizens of Delft via the S4W ODK system, municipality of Delft rain data and the rain data of the KNMI. The measurements taken by the citizen scientists have been compared to the rain data of government institutions to have a better understanding of the differences and similarities between the three datasets collected by different organisations. The three datasets were outlined in an automated Python script. The results show higher spatial resolution that could not be detected by the government gauges. Furthermore, specific precipitation could be detected by the citizen science dataset. Even though there were many inconsistencies with data processing caused by the way in which the data was collected, this citizen science project could prove beneficial in the field of water management.

Table of contents

1. Introduction

1.1 Global rises in temperature

Climate change receives increasingly more attention world-wide as the temperatures are rising globally, which poses a wide scale of issues. One such problem is an increase in rainfall in certain parts of the world. Changes in the rainfall and other forms of precipitation have been directly linked to global climate change (O'Gorman, Precipitation Extremes Under Climate Change, 2015). Because atmospheric greenhouse gases increase due to industry and the production of livestock, the amount of downward infrared radiation has also increased. Downward infrared radiation is responsible for the heating of the earth's surface. Eventually, the increase in surface heating raises the global temperature, which culminates in higher evaporation rates and ultimately leads to an intensification of precipitation. This result of climate-change driven precipitation changes is more specific to countries located in the northern hemisphere (O'Gorman, 2012)While evaporation increases due to the higher surface temperatures, so does the atmosphere's water-holding capacity. Such increases in atmospheric water vapour holding capacity leads to a higher intensity, frequency and duration of extreme rainfall events (Allan & Soden, 2008)

Thus, if the atmosphere's water carrying capacity increases and there is increased evaporation, the actual atmospheric humidity is expected to increase, as has been observed in many places (Trenberth, Fasullo, & Smith, 2005). As a result of this, the precipitation will come in brief and severe forms (Trenberth, Fasullo, & Smith, 2005). This will cause a problem for populated cities where large quantities of water must be transported through rain pipes, drains and the sewage system in a short amount of time.

This is an ever-growing problem for cities worldwide, where the current urban drainage infrastructure does not meet the need for future rain intensity. The contemporary design of the drainage systems is based on a statistical analysis of past events. This does not take into account the rapid increase in the intensity and frequency of extreme rainfall events, which most probably results in more frequent floodings of cities (Notaro, Liuzzo, Freni, & La Loggia, 2015)

Therefore, the design criteria must be amended in a way that enables it to anticipate possible changes in the intensity of rainfall-induced climate change activities. If there is a better understanding of high precipitation intensity areas throughout the city during a rainfall event, it will make it easier to design the infrastructure of the future in a way that tackles the surcharge water events.

1.2 Delft Measures Rain project

Delft Measures Rain is an initiative aimed to localize and quantify rainfall through the city of Delft. It is a science project set up by TU-Delft Waterlab, based on the voluntary participation of the citizens of Delft. Held in the summer of 2020, the citizens of Delft measured and mapped out the city's daily rainfall using self-made rain gauges. Volunteer citizen scientists for the Delft measures Rain project were provided with a manual and a measurement kit from the department of the WaterLab at the TU-delft university. With these items each researcher was able to make their own rain gauge. The measurements were acquired using identical homemade rain gauges for all the citizen researchers to reduce any confounding factors based as a result of instrument-related measurement errors.

The self-made rain gauge was designed to be convenient and easy to make, requiring none other than a regular drinking bottle, cement and a ruler. Using these do-it-yourself gauges, the citizen scientists have gathered daily measurements of the precipitation at their residence. Their findings were recorded digitally using their smartphone, the access to which was also a requirement to take part in this experiment. The exact construction and use of these rain gauges will be further elaborated in chapter 2. In addition to all the materials, the researchers also received updates about the research via an online newsletter. Smartphones4water was responsible for collecting all the rain measurements from the citizen scientists. The data, which is the local daily precipitation expressed in mm, was recorded every morning around 08:00 in the morning with a smartphone and using pre-installed with the Open Data Kit (ODK) application. Additionally, it was required to fill out a delivered form with the personal details of the citizen scientist. When the citizens fill in the ODK form, a picture of the rain gauge with the GPS location and the time it was measured is uploaded so that a S4W scientist can validate the measurement's accuracy later. The post-validation of the data by S4W adds an extra layer of quality control to ensure that the provided data is error proof. The daily measurement data can be easily accessed through the servers of S4W on their website.

1.3 Background information Smartphones4water

SmartPhones4water (S4W) is an organisation that aims to strengthen our understanding of water management through the means of mobilising citizen researchers, using mobile technology and utilising citizen science (SmartPhones4Water, 2020). S4W provides simple instructions on making basic scientific measuring kits for ordinary citizens who would like to contribute to scientific research.

By utilising the advances made in mobile technology like the GPS and camera capabilities that nowadays everyone possesses, the data can be easily collected from each citizen. For the data collection, the Android application called Open Data Kit (ODK) is used to gather all the scientific data from multiple sources to a single location where it can go through an extra layer of quality control done by S4W (SmartPhones4Water, 2020).

This method aims to go beyond the tradition of gathering data through the means of using expensive installations at a few locations. However, it instead aims to use cheap scientific tools and volunteer citizen scientists to gather the same data on multiple locations to get greater coverage of the relevant area. The collected data of all the S4W projects in various countries are part of the public domain, thus open to being studied by everyone.

1.4 Relevance

The KNMI has an average of one rain gauge per 100 square kilometres throughout the Netherlands (KNMI, 2020). In the whole municipality of Delft, there is only one operating KNMI gauge (KNMI, 2020). This means that currently the average daily precipitation of the entire city is measured through this single rain gauge. The city of Delft is around 25km^2 and Delft municipality has five rain gauges (KNMI, 2020). The locations of the rain gauges are randomly placed over the city, making some rain gauges cover an area greater than 10km^2.

Each gauge must cover a vast area of space. There is much precipitation data lost due to the large distances between the rain gauges. Because of these broad unmeasured precipitation, data gaps left behind by the two primary forms of precipitation measurement methods, observing and understanding the raw data becomes a bit difficult in terms of spatial and temporal precipitation distribution (Kidd & et al, 2017).

The project of Delft Measures Rain aims to expand on the available data of daily precipitation through the means of citizen scientists by trying to provide a daily precipitation measurement of at least one square kilometre or less. This will, in turn provide a better picture by filling in the gaps that are left behind by the precipitation measurements done by the two government institutions.
Using citizen science, where citizens play an active role in the pursuit of the scientific process, we can expand our knowledge of the hydrological data for weather forecasting (Jollymore, Satterfield, & Haines, 2017). Exploring the spatial distribution and variation of precipitation over a period of time can contribute to ideas about water resources in the future (Yavuz & Erdogan, 2012). This in turn will give us a better scope of understanding on how to tackle future extreme short-period precipitation, help us in policymaking and make sure that our water transport infrastructure does not get overburdened.

1.5 Aims of the project

This project aims to elucidate rainfall patterns in the city of Delft from rainfall data acquired with the help of the citizen researchers, which will be compared to their rainfall measurements acquired by government institutions. We should ask ourselves what are the most useful kind of conclusions we can draw after analysing the provided rain data from the three different sources? The main goal of this thesis will be to investigate the reliability of the data acquired by citizen scientists by placing it against the data acquired by government institutions. While comparing these datasets any additional rain patterns that come forth from the higher spatial resolution that citizen science data provides will also be investigated.

To reach this goal, the project will aim at answering the following questions:

1) **What is the quality of the data collected from the citizen rain gauges?**
   This question will be answered by comparing the data (after a basic quality check by S4W) to data from two types of government owned rain gauges:
   - A totalling rain gauge operated by KNMI (daily observations)
   - Automatic rain gauges operated by City of Delft (sub-daily observations)

Sub questions:

- *What are the differences between citizen gauges and the KNMI gauge at daily/sub-daily scales?*
- *What are the differences between citizen gauges and the City of Delft gauges, at daily/sub-daily scale?*
- *How do differences between citizen gauges and KNMI / City of Delft gauges vary with distance from citizen gauge to the reference gauge (KNMI or City of Delft)?*

NB: if the answers to these questions give plausible values, this provides confidence in the citizen data quality. With the available datasets, it is not possible to do a full quality check.

2) **What is the spatial distribution of the rainfall during rainstorms across the city of Delft?**
   This question will be answered by comparing data for rainstorms that occurred during the measurement period (summer 2020) from citizen gauges at different locations in the city.

Sub questions:

- *What is the variability of measured rainfall across the city per storm?*
- *How does the measured rainfall change with the distance between rain gauges?*
- *Are there any hotspots in the city where rain gauges have recorded more rainfall than the average across all stations?*

2 Methodology and approach

This chapter examines the methods and approaches that were needed to answer the main and sub-questions. The research is conducted by first explaining how each of the three different methodologies of measuring precipitation work. In order to reach our goal of this project, more insight is needed into the measurement methodology of the citizen scientists, the KNMI and the municipality of Delft on the same subject of precipitation in Delft city. Furthermore, there will be an analysis on the quality of the data collected by S4W from the citizen scientists. This must determine how reliable the measurements provided by citizen scientists are, what they precisely depict, and what the individual performance per citizen is. Lastly, a quantitative research will be done using a Python script with similar metrics to compare the three datasets and answer the questions that were asked in the problem and goal section.

2.1 Measuring precipitation methods

2.1.1 The KNMI

Daily rainfall was measured at approximately 325 KNMI precipitation stations with a standard rain gauge (**Figure 1**). The measured amount of precipitation was daily communicated back by the volunteer observers, digitally or by an automatic invoice system, to a central location where all the data of the precipitations comes together (Neerslagmeting, 2020). One can easily access the data on daily precipitation that is measured at these rain gauges through the official website of the KNMI.

Figure 1: *KNMI rain gauge*

The standard gauges of the KNMI meet the international requirements set by the world meteorological organisation (WMO). A detailed methodology on setting up standard rain gauges following the WMO requirements and how to perform the measurements are described in the online library of the WMO, which can be found on their website (Measurement of precipitation, 2014).

The results of these measurements are gathered and saved in an open-access file containing the 24-hour manually rainfall estimation. This is the sum of measurements taken from 08:00 UTC on the previous day to 08:00 UTC the current day. In the government provided datafile, the data is set up as station number, the date, the amount of precipitation in 0,1mm and the code for the snow cover, which is irrelevant for this project's purposes. In this research, only the dates between 18-07-2020 and 10-09-2020 will be examined. Also, the measured precipitation will be divided by 10 to convert the units to millimetres.

2.1.2 The municipality of Delft

The municipality measures precipitation following the same principles as described in the WMO guidelines. Rain is collected using a tipping bucket rain gauge (TBRG) (**Figure 2**). There are a total of five TBRG sensors across the city of Delft, but only two of them will be examined in this research. The reason for including only two out of five sensors is because of practicality. The raw data within these two files was easier to work with compared to the other three, which contained missing datapoints, while the remaining data is more complicated. The two precipitation gauges that are included in this project can be identified with the following codes: S5241 and S5240.

Both these rain gauges use the Casella Cell system to measure precipitation. Within the Casella cell rainfall enters a funnel collector within the rain gauge and is then directed to the tipping bucket assembly. Two tipping bucket rainfall sensors are connected to a measuring system. One tipping bucket gives one impulse per 0.1 mm of precipitation, the other one impulse per 0.2 mm of precipitation.

Figure 2: *Delft municipality rain gauge*

The government provided dataset is set up using timestamps followed by the number of impulses, amount of precipitation in mm (0.1mm tipping bucket) and the amount of precipitation in mm (0.2mm tipping bucket). When an incremental amount of precipitation has been collected on the tipping point mechanism (0.2 mm or 0.1mm), the bucket assembly tips and activates registers the collected rain. The sample is discharged through the base of the gauge. The precipitation data is registered for every minute over a 24-hour period (Tipping Bucket Rain Gauge, 2020). For the purposes of this research, the data of the two files will be arranged in a way to make it similar to the provided KNMI values, namely: taking the measurements from 08:00 on the previous day to 08:00 the current day

2.1.3 Citizen scientist

Citizen gauges are constructed from clear see-through plastic bottles, concrete, a ruler, and glue. The online video tutorial on how to make a DIY rain gauge, which is used in this project, can be found on YouTube under the video name of "How to make your own rain gauge"(https://bit.ly/3nF7xV4).

The plastic bottle is cut in half on the top part, just where the bottle's diameter is about to decrease. From the top, concrete is poured in the lower parts of the see-through bottle up to the point where the five rounded bumps are full, and the uniform cross-section begins. This way, an obvious reference point is created within the bottle. The weight of the concrete at the bottom of the bottle also adds some stabilisation to the bottle. After the concrete has set, a ruler is glued vertically onto the outside of the bottle. It is important to glue the ruler in place in a manner that the zero mark is just precisely at the surface of the dried cement line inside the bottle. In order to minimise the variability of the measured errors, the concrete must be dried and set first before you glue the ruler. With a thick nail, a hole is made on the bottle's cap for the water's inflow. The inverted lid is placed upon the bottle in order to minimise the evaporation losses. The end product can be seen in **Figure 3**.

Figure 3: *Citizen scientist rain gauge*

Operating the rain gauge is relatively easy. The gauge must be put outside on a one-meter elevated platform, preferable in the middle of your yard, away from any high objects such as trees or walls and secured with some cord, so it does not fall with strong winds. Every day at 08:00 AM, measurements are done on how much precipitation has been collected since the previous day.

2.1.4 Data Collection S4W

Data is collected through an easily downloadable app which is available through the Google Play Store under the "Open Data Kit (ODK) Collect" provided by S4W. Once the app is installed, the citizen scientist can send their measurements digitally to the collection point of the S4W server. A one-time detailed form must be filled, where the participants must fill in personal details and answer questions relevant to the project and privacy concerns. This data is then used for matching the measurements, which are done daily, by using the Android smartphones of the respective citizen scientist.

After this, every time the citizen scientist submits the daily measured precipitation, a few basic questions have to be answered in the application regarding the weather. After answering these basic short questions, the measurements and an on-the-spot picture, which clearly depicts the amount of precipitation collected during the 24 hours, are sent as proof through the ODK application to the S4W aggregate server. Once the data arrives at the S4W servers, a quality check is done by looking at the gauge's picture while comparing it to the citizen scientist's precipitation. If measurements are done incorrectly, or if something is wrong with the gauge setup, an employee of S4W will notify the citizen through their Email.

Additional information regarding how the data is collected can be found on the S4W public Google Drive where all the steps are explained in detail, together with an explanation on how to send your daily measurements to the S4W servers (Data Collection Instructions, 2020). All the citizen data collected through the Android application, including their personal information, can be accessed and downloaded via the S4W server online, after requesting permission to access the data from S4W. For this research, all the measurements were downloaded in an Excel format to study and analyse the provided data.

2.2 Citizen science data analysis

The datasets for the KNMI and the Delft municipality are simple files with mostly two or three columns of information regarding the precipitation. It mainly consists of the time of the measurement and the precipitation in mm or 0.1mm. The citizen dataset on the other hand, has 188 columns, representing various gathered forms of information concerning the way the citizen scientists took the measurements. Many of the columns do not provide any added value to our research purposes, making them irrelevant. To be able to quantify the data, an exact arrangement of the data collection was needed for the input. The columns with useful information are the following:

- **AutoSiteID**: The rain gauges have a unique corresponding AutositeID. Using the AutoSiteID, the locations could be distinguished from each other. This column was used to determine at how many sites a measurement was collected on a specific date.
- **DeviceID**: Each citizen scientist received a unique traceable DeviceID referring to the device from which the pictures were taken. This column was used to trace each individual scientist to a distinctive location.
- **Msmt_datetime**: The date and time of every measurement taken.
  This column was used to compare the data on different days with each other.
- **Latitude & Longitude**: The latitude and longitude coordinates. This was used to determine the locations of each individual citizen scientist.
- **Precip_mm**: The rainfall in mm measured

This column was used to compare the measurements taken from each distinctive scientist.

- **QCFlag**: The extra layer of quality control of the measured data
  This column was used to quality check the data. Only the columns with CheckedGood and CheckedCorrected were considered since S4W approves those.

After going through the data collected from the citizen scientists, it was noticeable that the provided data was in good workable shape, even though it required some adjustments. A total of 1991 measurements were collected by the 95 citizen scientists from the period of 9th of July till 15th of September. Before the data could be used for analysis, a few adjustments had to be performed first. These adjustments varied from minor to severe. There were some faults in the way the data was collected in a portion of the measurements. Therefore, the first stage aimed to arrange the collected data in a way which made it easier to work with. This implies that certain manipulations of the current citizen dataset were needed for easier processing i.e., deleting faulty measurements and adding missing details to the incomplete columns. These faults were divided into two categories, namely: Solvable matters and unsolvable matters.

Solvable matters

In an effort to make the dataset more precise for analysis, the following issues had to be addressed first:

- **Missing AutoSiteID**: 655/1991 of the collected data did not have an AutoSiteID. Which made it harder to distinguish measurement locations from each other.
- **Missing MonID**: 556/1991 of the collected data did not have an MonID. Which made it harder to distinguish which citizen scientist has done the measurements.
- **Incorrect labelling**: Measurement locations close to each other were sometimes incorrectly assigned with the same AutoSiteID. That's because the ODK application mistook them as the same point based on their GPS location and close proximity to each other.
- **Geographical imprecision**: The accuracy of the longitude and the latitude varies with the measurement from the same locations.

To solve these issues, the data was cross-referenced with other available columns. For example, there were instances of measurements for which the AutositeID and MonID were missing, while the DeviceID's were similar. It is safe to assume that these measurements all belonged to the same person from the same location. The same principle of cross-referencing known values with the unknown values could be applied to fill-in and correct the other available data.

Unsolvable matters

These entail issues which were impossible to solve, because the measurements have been already done and the data is already collected. The information was still useable by:

- **Use of different forms of rain gauges** (**Figure 4**): While the vast majority of the citizen scientists have done their measurements with the reverted bottle cap on, while a minority has done measurements without the bottle cap. There were also some examples of citizen scientists who have used their own store-bought rain gauge.
- **Use of cumulative precipitation**: Some citizen scientists have done measurements regularly, but they have never emptied their bottles after each measurement. Making their precipitation measurements a cumulative value over the length of a period longer than 24-hour.

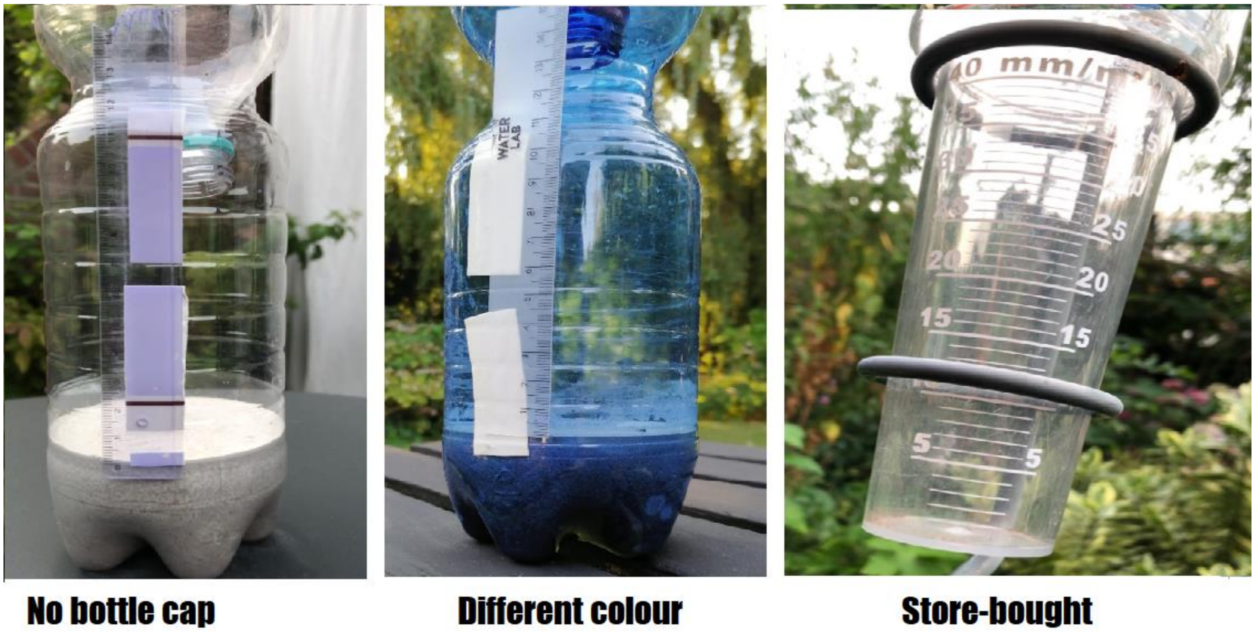

Figure 4: Variability in rain gauges

One of the first issues is that these differences in rain gauges could also lead to differences in the measured precipitation. To give an example, measuring without a reverted bottle cap could result in more loss of water due to evaporation. According to an experiment done with similar DIY rain gauges, the differences between bottles with an inverted cap and bottles without could be as high as $5mm^{-1}$ day depending on the temperature (Davids, 2019). The same thing could be said about the store-bought rain gauges. Parameters such the size, material and colour could have an influence on the measurements. Nonetheless, the use of different rain gauges and the measurements that go along with these gauges will not be treated any differently than the standard rain gauges.

The second issue is more complicated. The bottles are never emptied and every day the cumulative precipitation is communicated back to S4W, making the errors add up because of the cumulative nature of the measurements. These measurements were excluded from the data analysis.

2.3 Proposed analysis approach

2.3.1 Temporal approach

Once all the three forms of different datasets from the three distinctive sources were compiled, it was crucial to establish the rain measuring time intervals. The benefit of having rain data collected close to similar time periods is that your data can be easily compared to each other. For example, the KNMI precipitation measurement time periods are from 08.00 preceding day - 08.00 present-day measured in 0.1 mm. The precipitation measurements of the municipality of Delft are usually done every minute measured in 0.1 mm. The fact that it is measured every minute makes it possible to take the cumulative rain precipitation on a similar timeframe that coincides with the KNMI interval, namely from 08.00 UTC preceding day - 08.00 UTC present day.

The citizen science data on the other hand, was more inconsistent with the timeframe of the measurements. This makes comparing the three methods of rain measuring a bit more challenging. While most citizen scientist measurements were taken daily between the time periods of 07:00-09:00, with most of the data being close to the 08:00 time frame (**Figure 5**) a large portion of the measurements was taken during various other timeframes. To elaborate more on this problem, if the citizen scientists' measurements were done on various timeframes, it will be challenging to determine the mean rain precipitation for that given day. Suppose one citizen scientist gives us the precipitation measurements for a specific date for the time interval 08.00 preceding day – 08:00 present day, while the other citizen scientist provides the measurements for the time interval 08:00 preceding day – 12:00 present day. In that case, the precipitation results could be different if it rains between the later time frame.

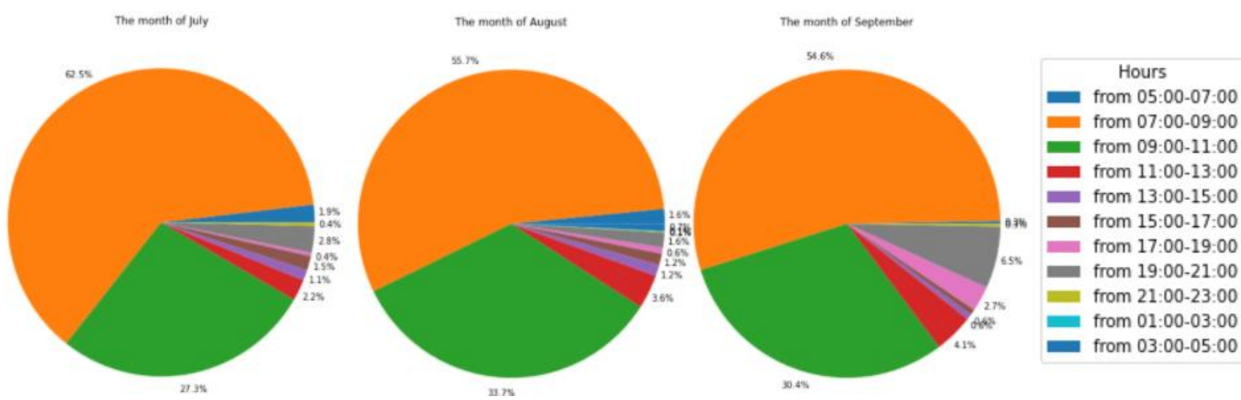

Figure 5*: Sampling timeframes Citizen Science data*

It is possible to work around these issues and decrease the margin of error in our comparison models if the measuring time intervals are increased. That is why the data will be analysed through two different methods:

- When the individual dates are compared between the datasets government datasets and citizen science data, only the citizen science precipitation data between 07:00-09:00 timeframes will be taken into our analysis.
- When a broader monthly comparison is made using the cumulative values between the government datasets and the citizen science data, all the precipitation measurements of the citizen science data on that given day will be taken into our analysis regardless of which timeframe.

The total time frame that these three datasets cover will be compared to each other spans from the 18th of July 2020 till the 10th of September 2020. The reason for this specific time frame is that we are limited by the usable citizen gauges measurements provided by citizen science data, even though we have measurements from the government institutions going back to the year 1961. The measurements from the citizen science dataset before the date 18th of July 2020 and after the date 10th of September 2020 will not be used for our analysis. Because the precipitation measurements on those dates in the citizen dataset are very irregular.

2.3.2 Spatial approach

The municipality of Delft and the KNMI rain gauges have their own distinctive GPS location, which is easily accessible since it is public information. There is also not much difficulty involved since there are just three GPS locations, one for the KNMI and two for Delft's municipality, used for the three rain gauges for this research.

Because only three GPS locations are involved see **Figure 6**. the spatial resolution of precipitation provided by these gauges is therefore also limited. Three sensors are not enough to cover the city of Delft for an adequate analysis. For these reasons, these three rain gauges alone will be insufficient to use data for characterising the spatial distribution of precipitation.

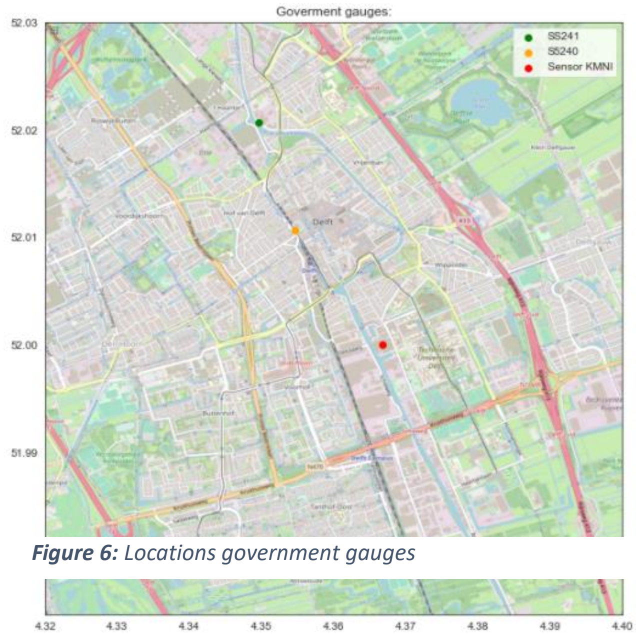

Figure 6: *Locations government gauges*

The citizen science dataset in contrast, provided us with 95 distinct rain gauge locations, with far better spatial coverage throughout the city of Delft. Each citizen scientist had his own gauge paired with his own unique GPS location. The vast distribution of the citizen gauges provided us better visualisation of the spatial coverage. This allowed us to do a short-range spatial variability analysis and a long-range spatial variability analysis. The Government gauges were used as a reference point while doing the spatial analyses. Given the time period mentioned above, three types of spatial analysis were performed given the three forms of datasets:

- A comparison of the citizen science dataset spatial precipitation distribution given the municipality of Delft rain gauges as reference.
- A comparison of the citizen science dataset spatial precipitation distribution given the KNMI rain gauge as reference.
- An analysis of the spatial precipitation distribution of citizen science dataset gauges during rainstorms.

2.4 Python script

The three precipitation datasets were analysed through a Python programming language. For the citizen science data, a workable framework was extracted from the excel file and used in the Python script. This was done by converting the raw Excel data into CSV. After errors were cleaned from the CSV file, several forms of plots and graphs were made. These include boxplots, scatter plots and maps of all the measurement points, with the aim of giving a better visualisation of the datasets and showing how large the outliers in the data are, as well as the variation of the data.

Once there was a clear understanding on the reliability of the citizen science data and what the variability between the rain gauges was, a comparison of the citizen science precipitation measurements was made with the KNMI/Delft municipality data, showing the daily and monthly differences.

3 Results

In this chapter the results of the research are presented. First, the findings of the citizen data are shared mainly focussing on the quality of the data. Second, citizen data is first compared with the KNMI precipitation and then with the of Delft visualising their similarities and differences. Lastly, the data of the three precipitation forms are compared municipality to each other focussing on the spatial distribution of the data.

3.1 Citizen science results

After processing the data and removing the faulty measurements, the data shows that there were a total of 95 citizen scientists who have participated in this project. Regretfully, two of the 95 citizen scientists had to be excluded from the analysis because they had given their cumulative values. There was not a clear follow up from S4W to warn these two participants to empty their bottles every 24-hour. They should have been alerted to adjust their measuring method. For the remaining citizens, the results can be seen in **Figure 7.** The vertical axis represents the number of measurements each citizen scientist has collected, which varies from 60 (highest) to 1 (lowest). The horizontal axis represents the corresponding AutoSiteID, the identificator for each participant. The bar graph is colour coded based on the time at which each measurement was performed, with blue being all the measurements done between the time-frame of 07:00-09:00 hour and orange being the measurements done on other hours of the day.

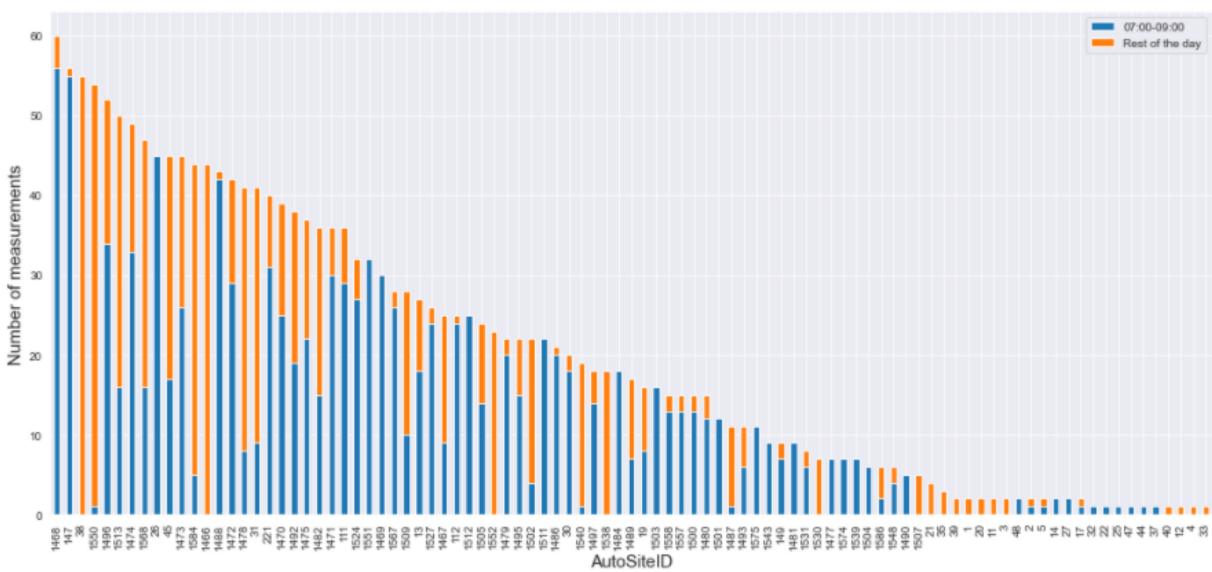

***Figure 7**: Number of records per autosite location (Blue=07:00-09:00 timeframe, Orange=Rest of the day)*

Figure 8 represents the daily mean measurement of all the recordings given the specific day in blue and the measurements done only between the 07:00-09:00 time period in orange. The difference between the two methods of sampling is visually noticeable. As stated in chapter 2.3.1, when we will be investigating the day to day comparisons between the citizen scientist provided data and the government institutions in the latter chapters, only the 07:00-09:00 will be used. In contrast, when monthly cumulative data of different datasets are compared, all of the provided citizen scientist measurements of the given day will be used. The reason for that is because error in sampling time of the citizen scientist data will be filtered out with this method.

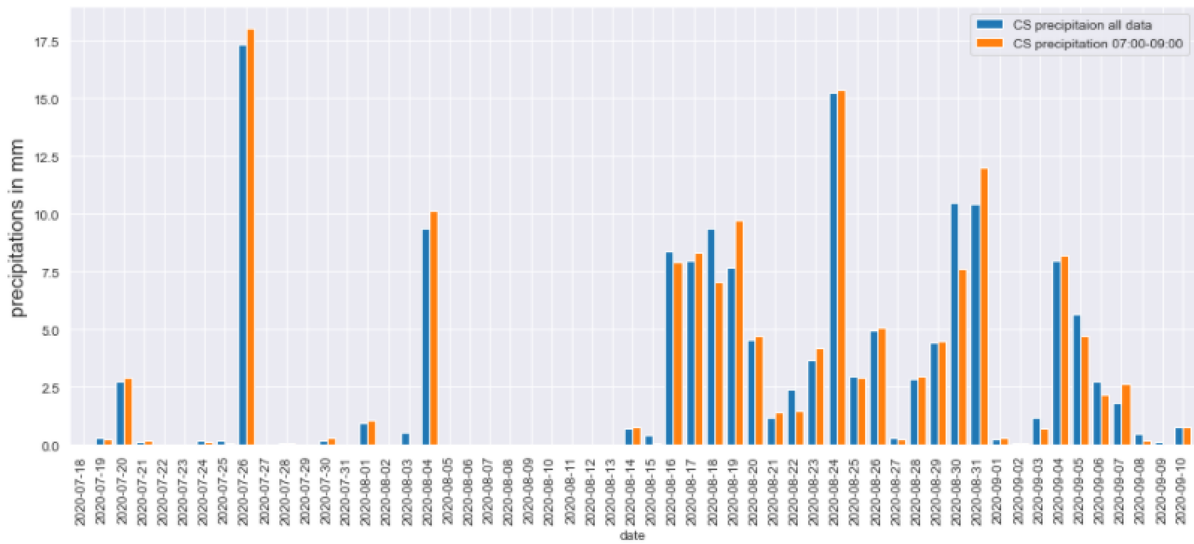

__Figure 8:__ Daily mean precipitation citizen science (Blue=all the data, Orange= 07:00-09:00 time frame) (precipitation in millimetre)

Figure 9 represents the boxplot values of the measurements. It is interesting to note here that the outliers within the boxplot graph doesn't necessarily mean that the outliers are wrong measurements. The outlier measurements could simply mean that in some parts of the city the cumulative 24-hour rainfall was more than the other parts, which we will examine further in chapter 3.3.

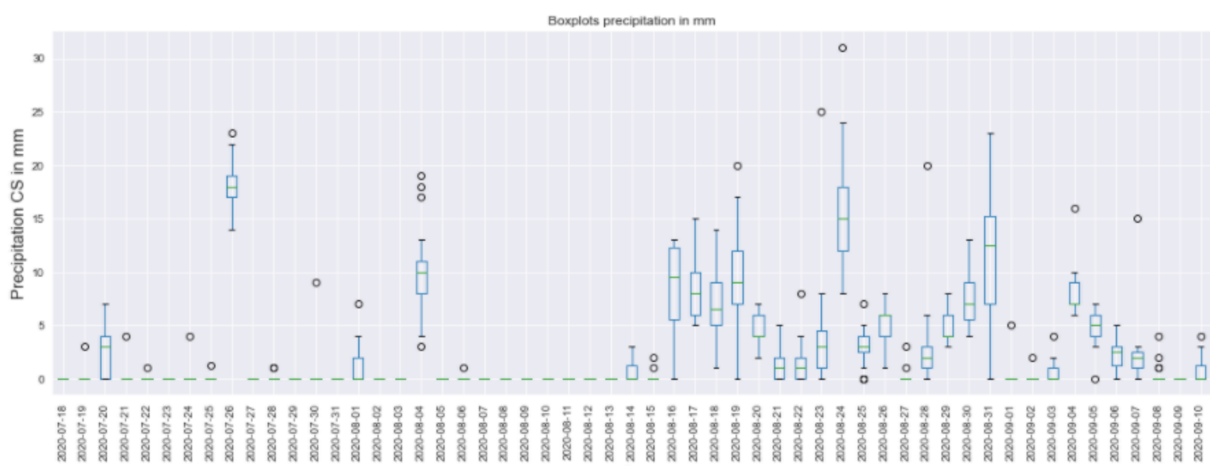

__Figure 9:__ Boxplots Citizen Science data (Timeframe 07:00-09:00) (precipitation in millimetre)

Table 1 gives a general view of all the different facets of the citizen data. The first column is the date where at least one of the 93 citizens has measured any rain. The second column "N" is counting how many measurements have been done on that specific day by various citizen scientists. The third column "Min" is the minimum value of precipitation in millimetre observed on the given day. The fourth column "Max" is the maximum value of precipitation in millimetre observed on the given day. The last column "Std" is the standard deviation, which measures the amount of variation of a set of values.

There was a sufficient amount of measurements on each given day over our 54 days long sampling time (table 1). The data is represented in detail in table 1, figure 8 and figure 9, providing a general understanding of the quality of the citizen scientist dataset. The values of the minimum, maximum and standard deviation of the mean are within the bounds of reason. No clear outliers can be detected visually from these datasets. After this first quality assessment of the citizen scientist data, it becomes clear that they can be used for further analysis.

Table 1: *Summary of the results from citizen science data (Timeframe 07:00-09:00) (N=sample size, Min=minimum measured value, Max=maximum measured value, Mean= the mean value, Std= standard deviation)*

Date	N	Min	Max	Mean	Std
2020-07-18	7	0.0	0.0	0.00	0.00
2020-07-19	10	0.0	0.0	0.00	0.00
2020-07-20	24	0.0	7.0	2.92	2.15
2020-07-21	21	0.0	0.0	0.00	0.00
2020-07-22	24	0.0	1.0	0.04	0.20
2020-07-23	23	0.0	0.0	0.00	0.00
2020-07-24	25	0.0	0.0	0.00	0.00
2020-07-25	15	0.0	0.0	0.00	0.00
2020-07-26	24	14.0	23.0	18.04	2.18
2020-07-27	29	0.0	0.0	0.00	0.00
2020-07-28	27	0.0	1.0	0.07	0.27
2020-07-29	30	0.0	0.0	0.00	0.00
2020-07-30	27	0.0	0.0	0.00	0.00
2020-07-31	27	0.0	0.0	0.00	0.00
2020-08-01	21	0.0	7.0	1.10	1.84
2020-08-02	13	0.0	0.0	0.00	0.00
2020-08-03	20	0.0	0.0	0.00	0.00
2020-08-04	28	3.0	19.0	10.14	3.64
2020-08-05	24	0.0	0.0	0.00	0.00
2020-08-06	20	0.0	0.0	0.00	0.00
2020-08-07	19	0.0	0.0	0.00	0.00
2020-08-08	15	0.0	0.0	0.00	0.00
2020-08-09	11	0.0	0.0	0.00	0.00
2020-08-10	18	0.0	0.0	0.00	0.00
2020-08-11	14	0.0	0.0	0.00	0.00
2020-08-12	21	0.0	0.0	0.00	0.00
2020-08-13	13	0.0	0.0	0.00	0.00
2020-08-14	22	0.0	3.0	0.76	1.04
2020-08-15	13	0.0	1.0	0.08	0.28
2020-08-16	17	0.0	13.0	7.95	4.87
2020-08-17	24	5.0	15.0	8.33	2.62
2020-08-18	22	1.0	14.0	7.05	2.98
2020-08-19	22	0.0	20.0	9.70	4.87
2020-08-20	19	2.0	7.0	4.74	1.63
2020-08-21	19	0.0	5.0	1.42	1.68
2020-08-22	15	0.0	8.0	1.50	2.16
2020-08-23	14	0.0	25.0	4.21	6.41
2020-08-24	25	8.0	31.0	15.36	5.63
2020-08-25	20	0.0	7.0	2.95	1.82
2020-08-26	19	1.0	8.0	5.05	1.96
2020-08-27	15	0.0	3.0	0.27	0.80
2020-08-28	19	0.0	20.0	3.00	4.33
2020-08-29	12	3.0	8.0	4.50	1.51
2020-08-30	11	4.0	13.0	7.64	2.66
2020-08-31	23	1.0	23.0	11.98	5.77
2020-09-01	15	0.0	0.0	0.00	0.00
2020-09-02	17	0.0	0.0	0.00	0.00
2020-09-03	17	0.0	4.0	0.71	1.05
2020-09-04	13	6.0	16.0	8.23	2.71
2020-09-05	11	0.0	7.0	4.73	2.00
2020-09-06	10	0.0	5.0	2.20	1.55
2020-09-07	14	0.0	15.0	2.64	3.67
2020-09-08	18	0.0	2.0	0.22	0.55
2020-09-09	11	0.0	0.0	0.00	0.00
2020-09-10	15	0.0	4.0	0.77	1.29

3.2 Quality comparison

3.2.1 Citizen scientists & KNMI

Figure 10 and **11** Represent the cumulative precipitation measurements of the KNMI and citizen scientists, but in two different ways. Figure 10 depicts the cumulative precipitation in millimetres of the KNMI dataset (red) and the citizen scientist dataset (blue) over time. Figure 11 on the other hand, is a version of a double mass graph in which the precipitation (mm) of the KNMI is used as a reference (red line) to which the corresponding precipitation values of the citizen scientists have been plotted (blue).

The measurements from the KNMI and the citizen scientists follow similar patterns of precipitation over time (figure 10) and double mass cumulative precipitation (figure 11).

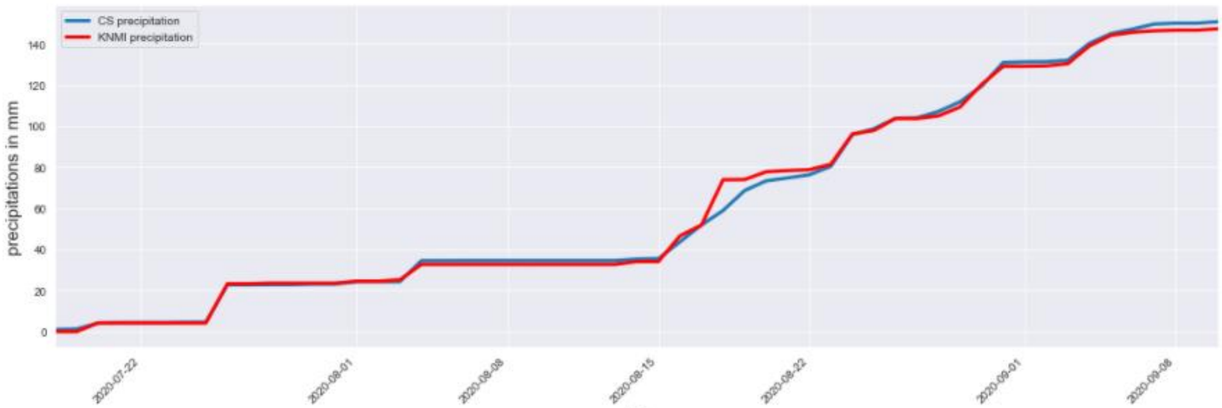

__Figure 10:__ Cumulative precipitation values Citizen Science (all values average) vs KNMI over time (precipitation in millimetre)

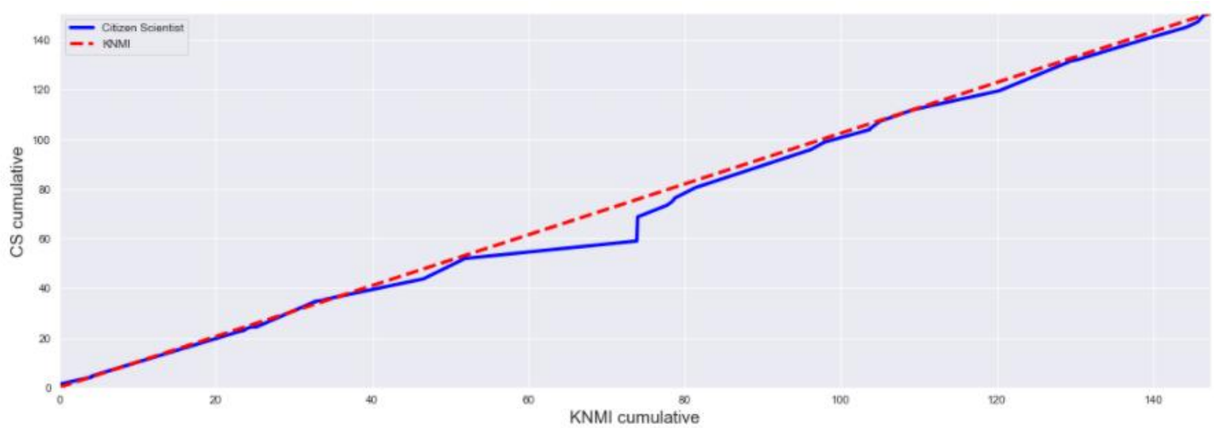

__Figure 11:__ Double mass precipitation Citizen Science (all values average) VS KNMI as reference (precipitation in millimetre)

Next, we visualized how the measurement accuracy changes in relation to the amount of precipitation on a given day. So, if on a given day the KNMI gauge measures a certain amount the graph displays what the other civilian gauges measure on that specific day. **Figure 12** represents a comparison of the measurements between the dataset of the citizen scientists and the KNMI. In this graph each dot represents a day for which we plotted the KNMI precipitation on the vertical axis and is also our reference value. On the vertical axis the corresponding citizen scientist measurements plotted in millimetres of precipitation.

Per KNMI reference measurements, the data was divided into three precipitation event ranges: 0–5 mm, 0-15mm and 0-35mm (highest KNMI value = 22mm). From these graphs we can observe that the correlation coefficient is low (*R*= 0.07, *N*=349) for the 0 to 4mm precipitation events. However, the correlation value increases in the higher 0-15mm and 0-35mm precipitation ranges (*R*= 0.62, *N*=615; *R*= 0.64, *N*=692 respectively) as measured by the KNMI gauges.

This could be due to the fact that the KNMI rain gauge also measures smaller events between 0mm to 1mm. Measuring such relatively small events could be difficult for the citizen scientist due to evaporation (between 0.1mm and 5mm per day) concrete soaking inside the do-it-yourself gauges (Davids & et al, 2019). Another possible explanation for this low correlation coefficient could be that some smaller rain events were not spatially uniformly distributed over the city. Allowing only a small subset of the gauges to pick up the precipitation, while the others did not.

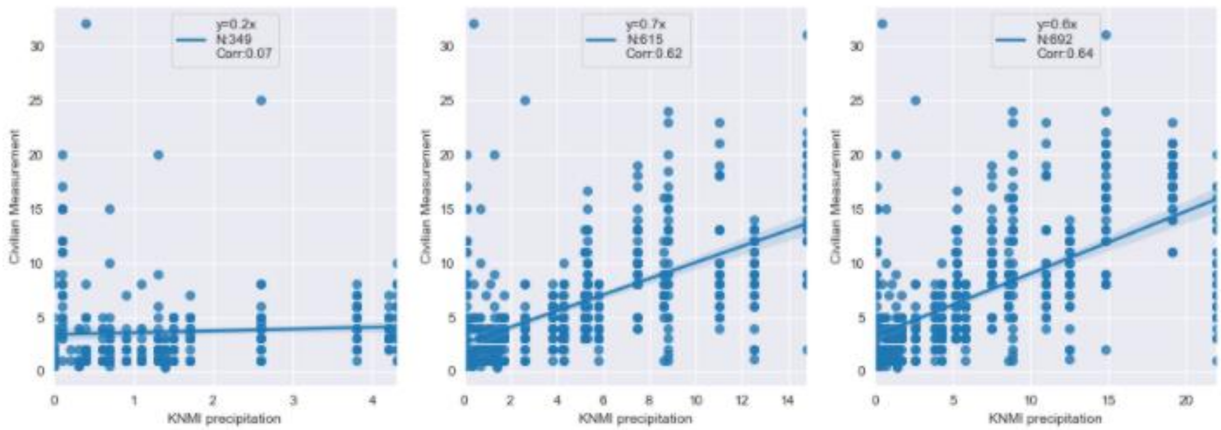

Figure 12*: Comparison of the precipitation data Citizen Science (Timeframe 07:00-09:00) vs KNMI (precipitation in millimetre)*

3.2.2 Citizen scientists & Delft City

A similar approach was used to compare the precipitation datasets obtained from the municipality of Delft to the citizen scientists (**Figure 13** and **14**). Figure 13 depicts the cumulative precipitation in millimetres of the Delft S5240 dataset (orange), the Delft S5241 dataset (green) and the citizen scientist dataset (blue) over time. A version of a double mass graph is shown in figure 14 in which the precipitation (mm) of the citizen scientists is used as a reference (blue dotted line) to which the corresponding precipitation values of the Delft S5240 (orange) and Delft S5241 (green) have been plotted.

Similar to the measurements from the KNMI (Figure 11), also the measurements from the Delft S5240 and Delft S5241 are very similar to the data of the citizen scientists, following similar patterns of precipitation over time (Figure 13) and double mass cumulative precipitation (Figure 14).

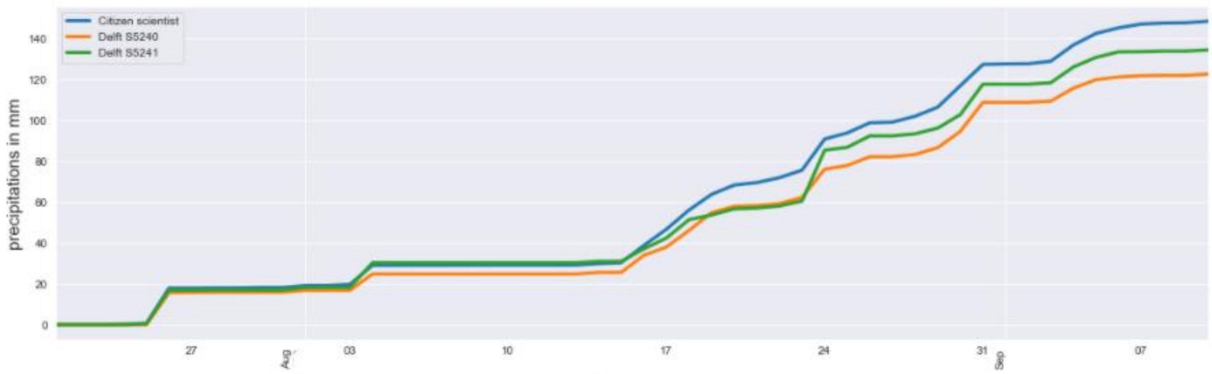

Figure 13:*Cumulative precipitation values Citizen Science (all values average) vs Delft gauges over time (precipitation in millimetre)*

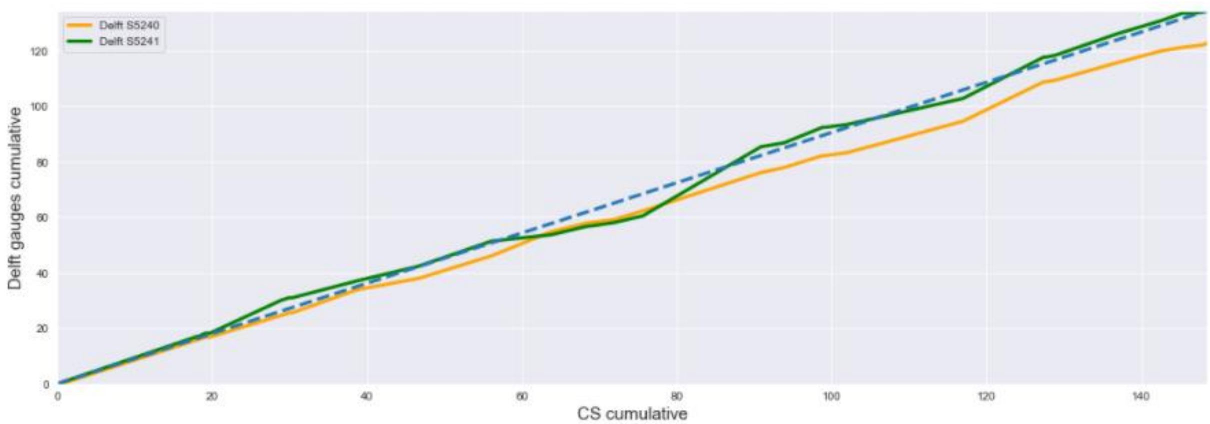

Figure 14: *Double mass precipitation Delft gauges vs Citizen Science (all values average) as reference (precipitation in millimetre)*

A correlation analysis between the precipitation that is measured by the Delft rain gauges and the citizen scientist data has been performed similar to the one shown in **Figure 15**. When investigating the Delft rain gauges, a similar pattern can be seen. Also, from these graphs we can observe that the correlation coefficient is low (R= 0.26, N=417) for the 0 to 4mm precipitation events. While, the correlation value increases in the higher 0-15mm and 0-35mm precipitation ranges (R= 0.70, N=664; R= 0.63, N=692 respectively) as measured by the S5240 gauge.

The S5241 gauge also exhibited stronger correlations in the higher precipitation events. While the correlation value was low for the 0 to 4mm precipitation event (R= 0.2, N=410), this was much higher for the middle 0-15mm and high 0-35mm ranges (R= 0.64, N=632; R= 0.64, N=692 respectively).

Most of the gauges showed an increase in precipitation, when including higher precipitation events. An exception of this was the S5240 rain gauge, which demonstrates a lower correlation in the high precipitation range of 0-35mm compared to the medium 0-35 precipitation range.

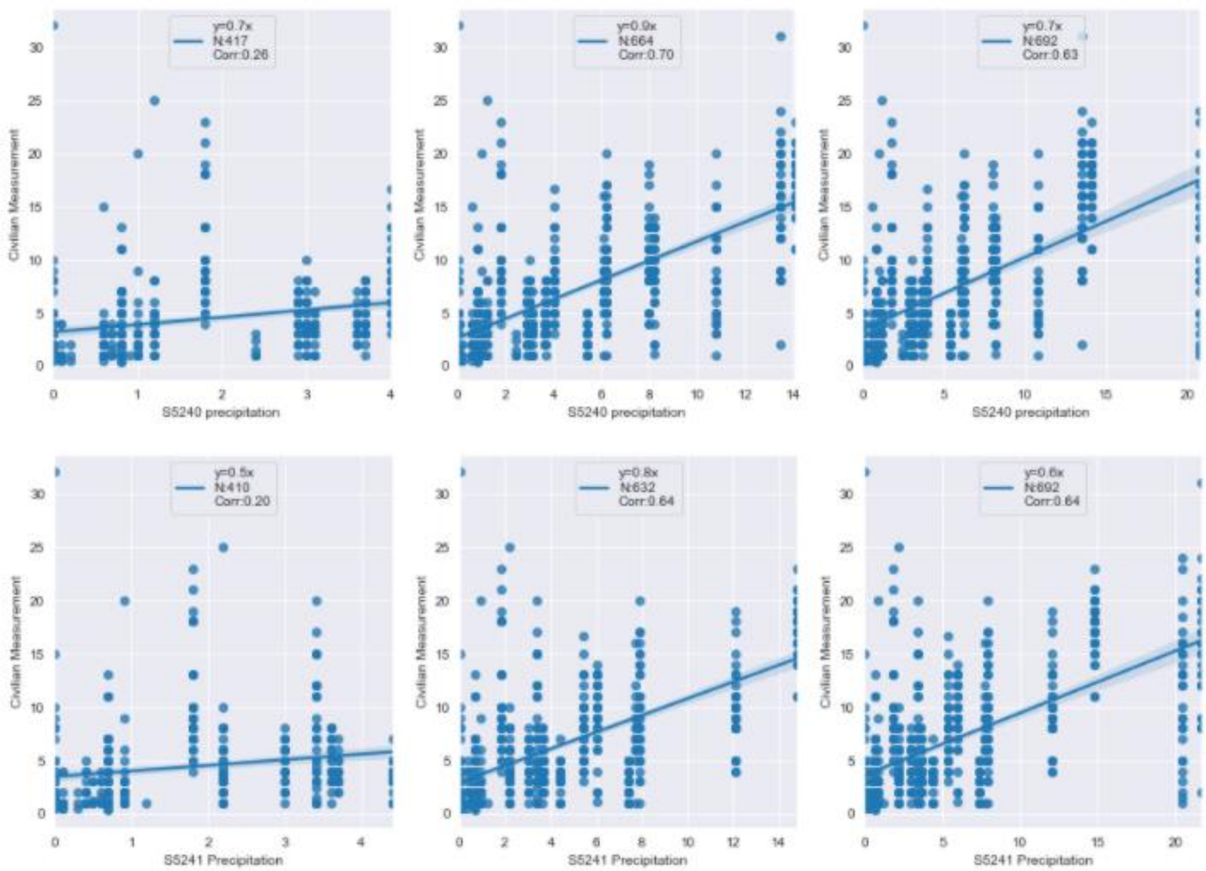

Figure 15*: Comparison of the precipitation data Citizen Science (Timeframe 07:00-09:00) vs Delft gauges (precipitation in millimetre)*

3.3 Spatial distribution comparison

3.3.1 Citizen scientists compared to KNMI

Figure 16 is the distance difference graphs of all the days when both the KNMI and citizen scientists (07:00-09:00) have measured rain. The days when citizen scientist gauges have measured no rain have been left out. A total of 30 out of 54 days where some level of precipitation has been measured by KNMI and one or more citizen gauges remained.

In these graphs, the difference between the measured precipitation of the KNMI and the measurement of the citizen scientists are plotted against the distance of the citizen scientist to the KNMI rain gauge. Visible on the vertical axis is the difference of rain in millimetres compared to the KNMI rain gauge and on the horizontal axis is the distance in kilometres compared to the KNMI gauge. Each plot shows the measures that have been collected at a certain day depicted in the title along with the N-value of the amount of citizen gauges with a positive precipitation value.

Although, there are some clear correlations between distance and difference in precipitation for example in the graph of 2020-08-04; where the differences tend to be higher when the distance increases. Other graphs can be observed where a linear correlation between distance and difference in precipitation measurements is absent. More complex patterns with single or multiple peaks are also present in the dataset. This is not entirely surprising given the fact that rainfall patterns are not spatially uniform. Within a single rain event the spatial rainfall distribution variability can be different depending on the place at the time of the measurement, effects of wind and direction of the precipitation which can all vary (Bacchi, Kottegodab, & Nick, 1995)

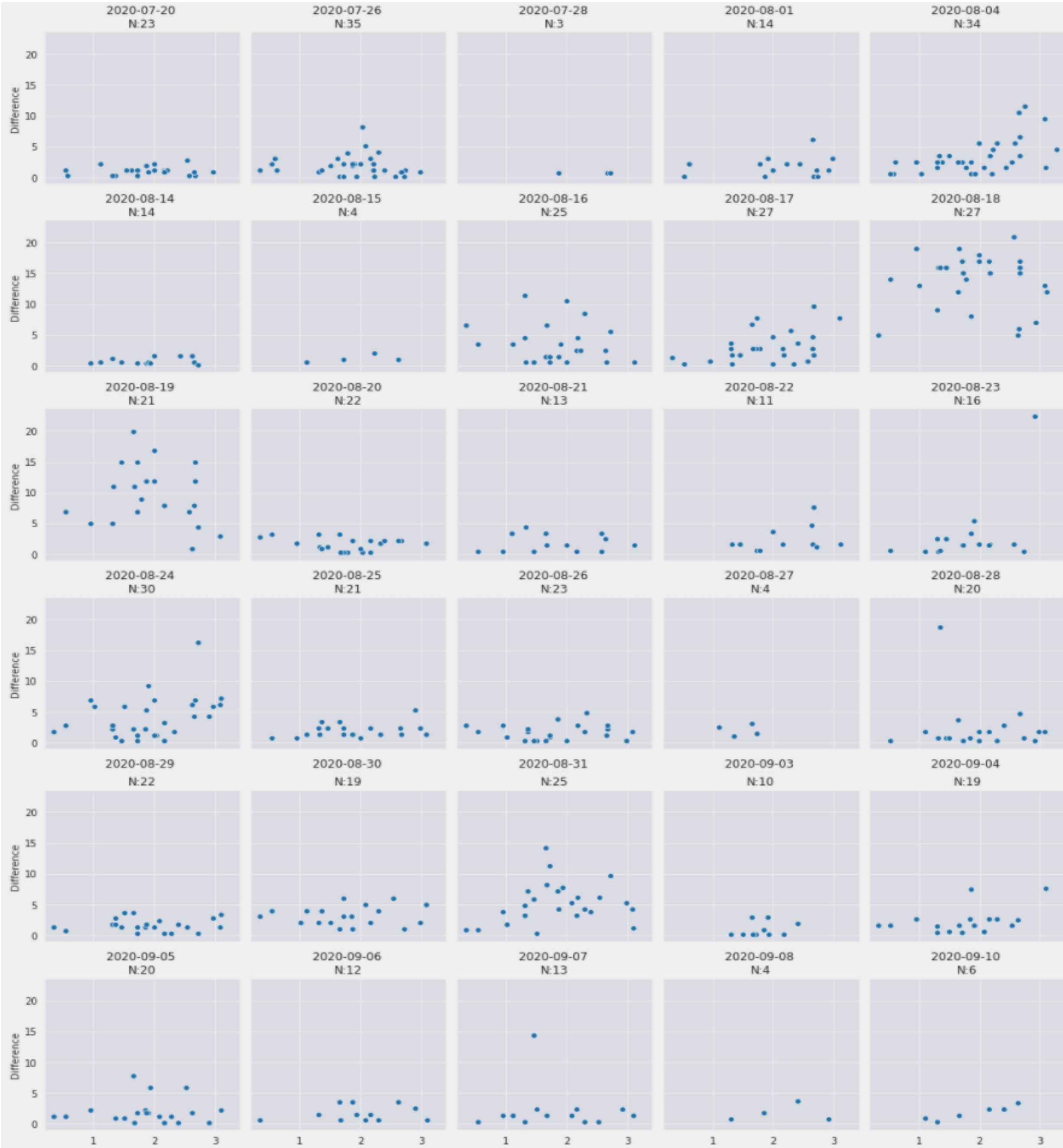

__Figure 16__: Distance difference comparison Citizen Science data (Timeframe 07:00-09:00) vs KNMI as reference (Distance in km, difference in mm)

Figure 17 represents the heatmap of the geographical location of the individual rain gauges after specific precipitation events. The date of when the precipitation measurements took place are used in the title. The vertical axis shows the latitude (y-axis) and the longitude (x-axis) measures of the location of each rain gauge. For this analysis, four consecutive rainy days were selected in the month of August where the KNMI precipitation was above 7mm per day.

The four heat maps show differences in precipitation of each citizen's rain gauge relative to the rain gauge of the KNMI. The relative differences are color-coded ranging from small differences in dark blue, to large differences in rain measurements in light-blue. These heatmaps show the more complex spatial-distribution over distance of the rainfall patterns. The dates 2020-08-04 and 2020-08-16, display the trend of increasing intensity in the north western regions of the map. The dates 2020-08-18 and 2020-08-30 on the other hand show that the gauges in the inner city of Delft have gathered less rain than the gauges at the outskirts of Delft.

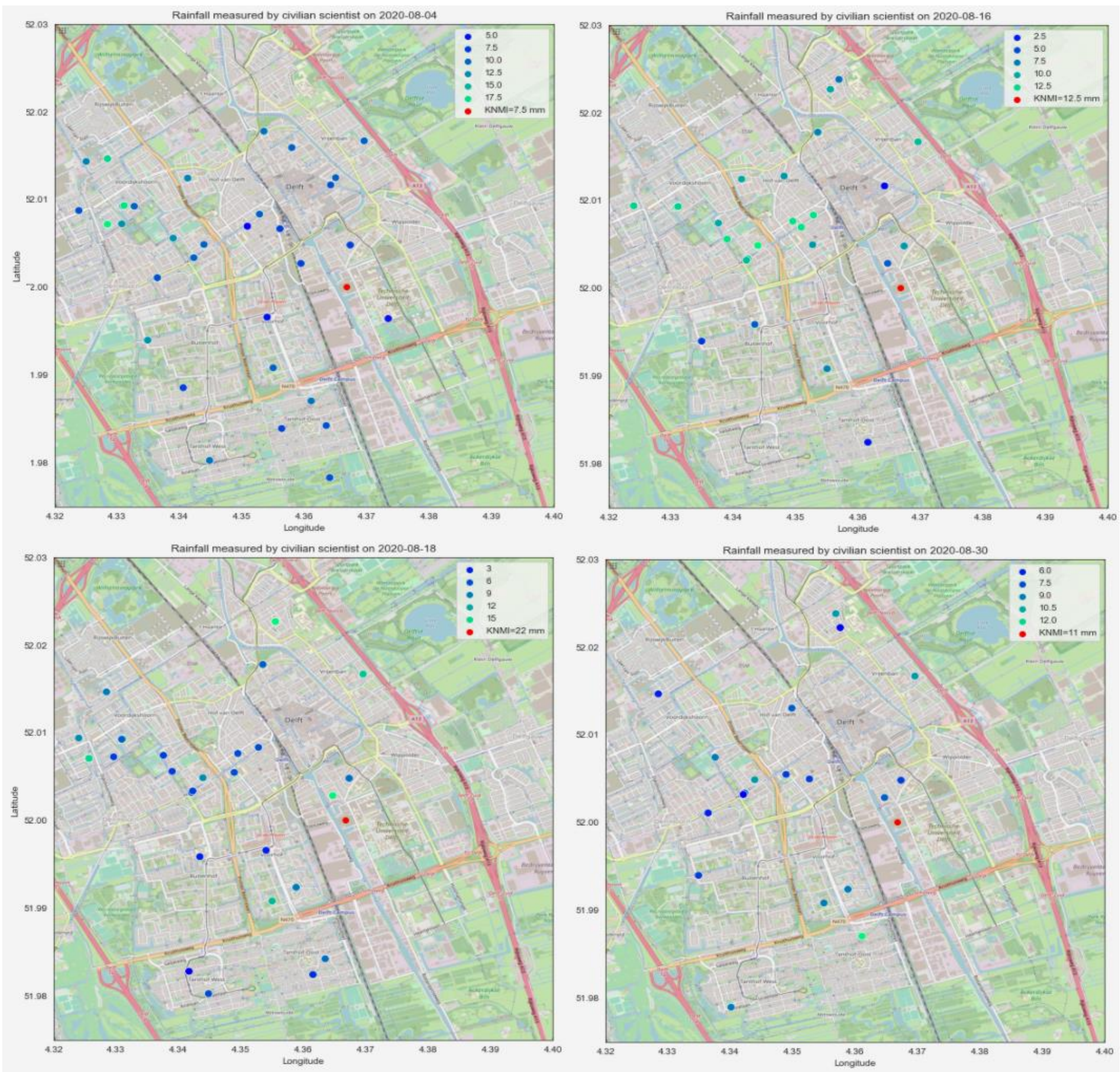

__Figure 17:__ Heatmap precipitation measured on a specific date Citizen Science data & KNMI (precipitation in millimetre)

Figure 18 represents the average true difference measured by the civilian scientists across the 30 days using the KNMI gauge as the reference. Using the true difference values, it is possible to observe which of the citizen gauges have collected the most rain and what the collective rainfall variability is in Delft city. With few exceptions, the rain gauges in the northern region, especially the north-western region of the city tend to collect more rain than the other gauges.

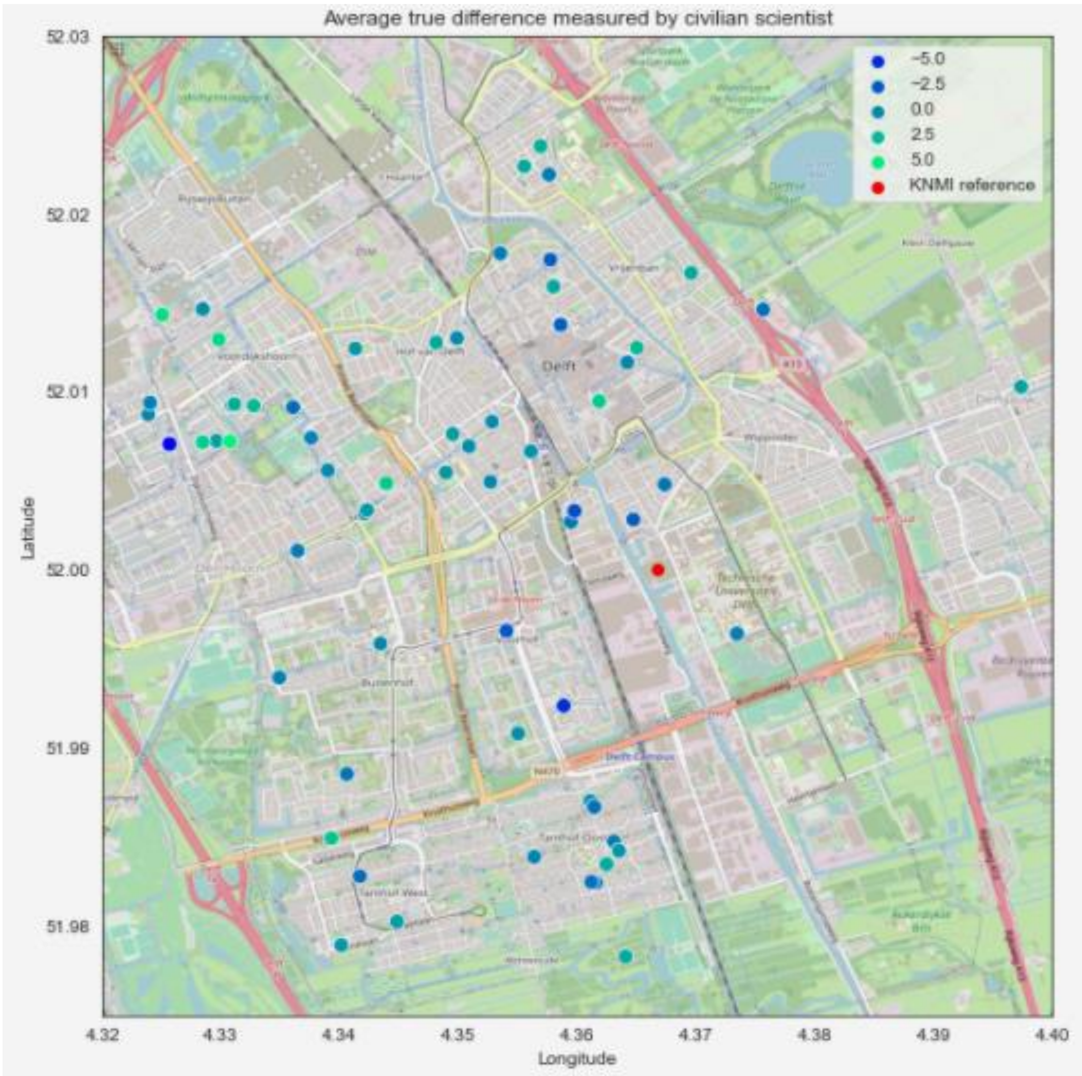

Figure 18: Heatmap average true difference measured by civilian scientists vs KNMI as reference (precipitation in millimetres)

3.3.2 Citizen scientists & Delft City

The same rain patterns mentioned in the paragraph 3.3.1 can similarly be observed for the rain gauges of Delft Municipality. An extra example of the date 2020-08-31 in **Figure 19** has been added to display the rain pattern which tends to be heavier with more precipitation in the northern parts of the city compared to the southern areas. The precipitation heatmap on a similar level provides us with insights into where the rain gauges collect more precipitation than the others. The distance difference graphs share a high resemblance with the KNMI measurements. The distance difference graphs of the two Delft gauges can be found in the appendix A. **Figure 20** is again the heatmap of average difference using Delft gauges as reference.

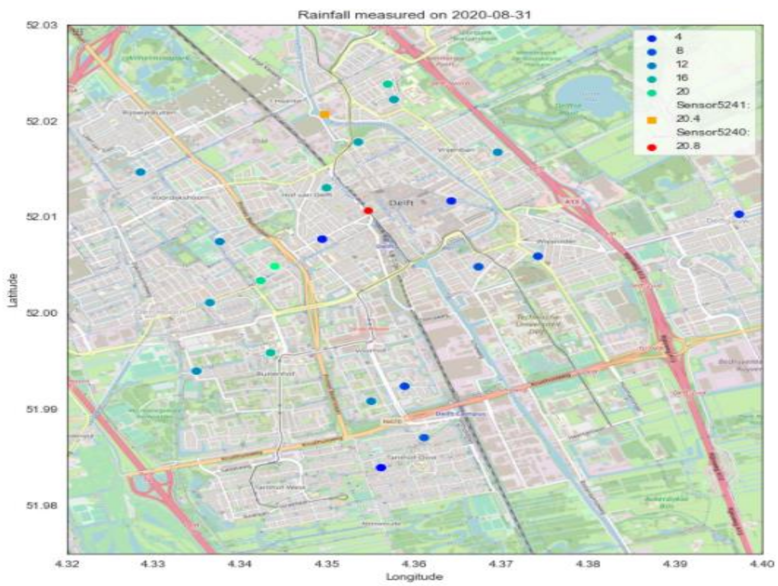

Figure 19: Heatmap precipitation measured on a specific date Citizen Science data & Delft Gauges (precipitation in millimetre)

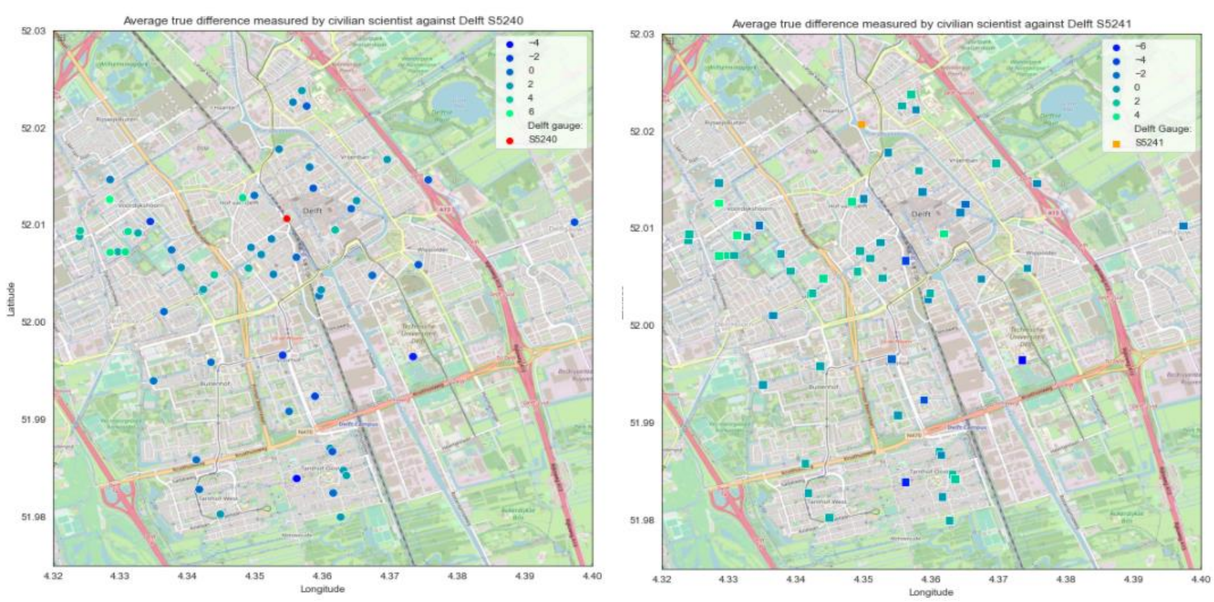

Figure 20: Heatmap average true difference measured by civilian scientists vs Delft gauges as reference (precipitation in millimetres)

4 Discussion

The goal of this thesis was to demonstrate the effectiveness of an efficient and low-cost system to map complex spatio-temporal patterns of rainfall. The results described in chapter 3 show that such a system adds value to the currently existing methodology of measuring precipitation. However, some issues must be addressed.

discuss result paragraph 1

During the data-cleaning and quality checking of the data, it became clear that not every citizen researcher complied with the time-window during which the measurements had to be acquired. Many datasets were unusable for day-to-day comparisons because of this. In future studies using a similar approach driven by volunteer scientists, compliance could be stimulated by sending reminders via texts, rewarding scientists after a streak of successful submissions.

Whereas the participation of the volunteers fluctuated over the period of this experiment, the data that was acquired by the scientist remained consistent. This demonstrates the effectiveness of the do-it-yourself rain gauge compared to the professional gauges that are employed by the KNMI and the municipality of Delft. On the other hand, there is a strong pattern in the correlation between the different open-access datasets and the citizen scientist measurement based on the amount of precipitation that has fallen. A possible explanation for these differences could be due to inter-subject variability in judging the amount of water in the gauge. Human error due to age can arise from deteriorated eye-sight or intrinsic differences in estimation capacity.

Another explanation could be that evaporation occurred before the measurement was acquired. The experiment and gauge were set up in a way to minimize this effect i.e., measurements were scheduled at a time when surface temperature is low while not interfering with the average human day-night cycle. The gauge itself was also designed to minimize the effect of evaporation which has been demonstrated in (Davids, 2019). A last explanation could be absorption by the concrete; however, this has also been investigated by Jeffrey C. Davids et al.

The outliers in the distance-difference graphs indicate that, even though there was a strict validation of the data submitted to S4W, some false measurements were approved. Again, human error could play a role in this. Therefore, a more automated approach using rulers with clear markings and software that detects the level of water could solve this issue by forming another layer validation.

The spatial distribution as demonstrated in chapter 3.3 shows the major benefit of the citizen science approach compared to governmental rain gauges. Patterns such as the ones demonstrated in this thesis cannot be shown with the limited spatial coverage of the governmental rain gauges.

Now that we have demonstrated the existence of localized “hotspots” of precipitation, these areas should be taken into account when reconstructing or building new sewers, drains and drainpipes. It is expected that these areas will suffer more from heavy rainfall. Climate theory also dictates that higher surface temperatures will lead to increased precipitation, making these hotspots even more vulnerable to heavy rain. These regions have been found in Delft due to experimental spatial restrictions, but throughout the Netherlands and globally any more of these hotspots could potentially be found. Programmes such as citizen science should therefore be stimulated and set up in different regions to validate the existence of patterns such as the one found in this thesis and to inform local authorities on the construction related to wastewater management.

Using citizen scientists for projects such as these could revolutionize our way of collecting data. Instead of using expensive machines we could use cheap materials, eager volunteer citizens who would love to do a scientific contribution and a good data collection methodology to achieve the same if not better results. Especially in the areas where we would like to know the spatial variability of patterns, citizen science would be a great addition.

There are a few areas where some things could be improved. There needs to be a more efficient system of letting citizen scientists know if their methodology is right or wrong with each measurement. The feedback and the follow-up to the citizens should be more efficient. Some citizens were not warned in time or not warned at all for using a different rain gauge or a wrong method of measuring. The amount of unusable data could have been decreased. The sampling time should be uniform or close to uniform. There needs to be a better method of registering the citizen data with each measurement. A lot of data manipulation had to take place before the citizen scientist dataset could be analysed.

5 Conclusion

The results in this thesis illustrates that the citizen data is not only just as reliable as the government institution gauges but it also provides a broader perspective of rain patterns that are overlooked if we just work with limited rain gauges. In addition, the variability of citizen scientists' data compared with the two government institutions, at all autosite locations, was slightly higher for smaller rainfall events. This could be due to short-range spatial variability of rainfall, evaporation, concrete soaking, the presence of other disturbing factors like bad placement, wind effects, human visual error could all affect the results. In contrast, for higher precipitation this was less the case.

Studying the spatial distribution variability of rain gained through citizen scientist data, will help us improve our city in a way to make it more efficient when building more target-based drainage infrastructure. It could fill in the data gaps that are left behind by orthodox scientific measuring tools. Government monetary funds from less rain impacted areas could be invested in places where the rainfall is relatively higher than normal.

Bibliography

Allan, R. P., & Soden, B. J. (2008). Atmospheric Warming and the Amplification of Precipitation Extremes. *Science*, Vol. 321, Issue 5895, pp. 1481-1484.

Data Collection Instructions. (2020). Retrieved from S4W Google Drive: https://docs.google.com/document/d/1u5BpbGPLps1gglgBUit6Kd05zbJnA21lVtF4v8VZxio/edit#

Bacchi, B., Kottegodab, N. T., & Nick. (1995). Identification and calibration of spatial correlation patterns of rainfall. *Journal of Hydrology, Issues 1–4*(Hydrology), Pages 311-348. Retrieved from https://www.sciencedirect.com/science/article/pii/0022169494025908

Davids, J. C., & et al. (2019). Soda Bottle Science—Citizen Science Monsoon Precipitation Monitoring in Nepal. *Front Earth Science*.

Jollymore, A., Satterfield, T., & Haines, M. J. (2017). Citizen science for water quality monitoring: Data implications of citizen perspectives. *Journal of Environmental Management*, 200:456-467.

Kidd, C., & et al. (2017). So, How Much of the Earth's Surface Is Covered by Rain Gauges? *American Meteorological Society*, 98 (1): 69–78.

KNMI. (2020). *Waarneemnetwerk KNMI*. Retrieved from Koninklijk Nederlands Meteorologisch Instituut: https://www.knmi.nl/kennis-en-datacentrum/uitleg/waarneemnetwerk-knmi

Measurement of precipitation. (2014). Retrieved from World Meteorological Organization: https://library.wmo.int/?lvl=notice_display&id=19661#.X4IEttAzaUk

Neerslagmeting. (2020). Retrieved from Koninklijk Nederlands Meteorologisch Instituut: https://www.knmi.nl/kennis-en-datacentrum/uitleg/vrijwillige-neerslagmeters

Notaro, V., Liuzzo, L., Freni, G., & La Loggia, G. (2015). Uncertainty Analysis in the Evaluation of Extreme Rainfall Trends and Its Implications on Urban Drainage System Design. *Water*, (7):6931-6945.

O'Gorman, P. A. (2012). Sensitivity of tropical precipitation extremes to climate change. *Nature Geoscience* , 5(10):697-700.

O'Gorman, P. A. (2015). *Precipitation Extremes Under Climate Change.* 1:49–59: Curr Clim Change Rep.

Trenberth, K. E., Fasullo, J., & Smith, L. (2005). Trends and variability in column-integrated atmospheric water vapor. *Climate Dynamics*, volume 24, pages741–758.

Yavuz, H., & Erdogan, S. (2012). Spatial Analysis of Monthly and Annual Precipitation Trends in Turkey. *Water Resources Management 26*, 609-621.

SmartPhones4Water. (2020). Retrieved from Make a Difference with Your Old Smartphone: https://www.smartphones4water.org/s4w/smartphones4water/

SmartPhones4Water. (2020). *Technology*. Retrieved from Water – Our Most Precious Resource: https://www.smartphones4water.org/s4w/technology/

Tipping Bucket Rain Gauge. (2020). Retrieved from Koenders Instruments: https://www.koenders-instruments.com/wp-content/uploads/2020/02/Tipping-Bucket-neerslagmeter-Casella.pdf

Appendix

Appendix A

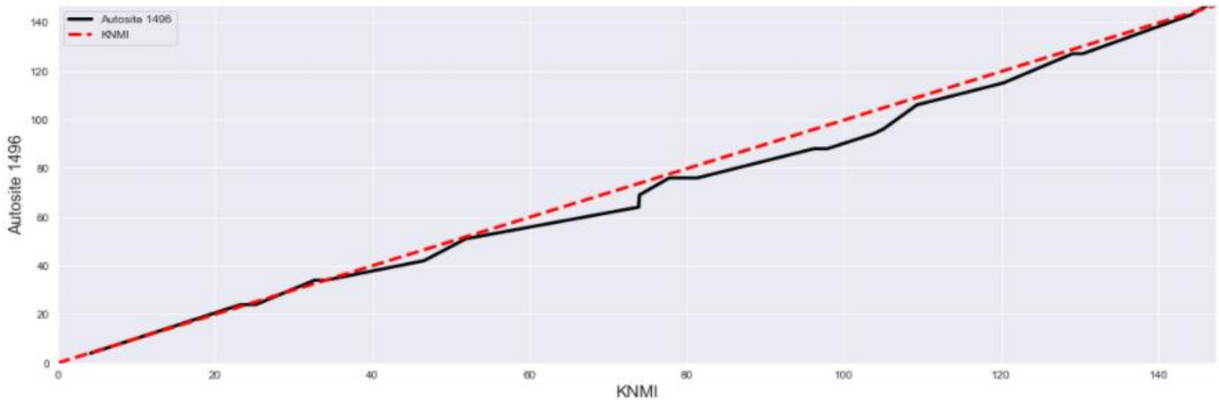

Figure 21: *Double mass precipitation Citizen Science autosite 1496 VS KNMI as reference (precipitation in millimetre)*

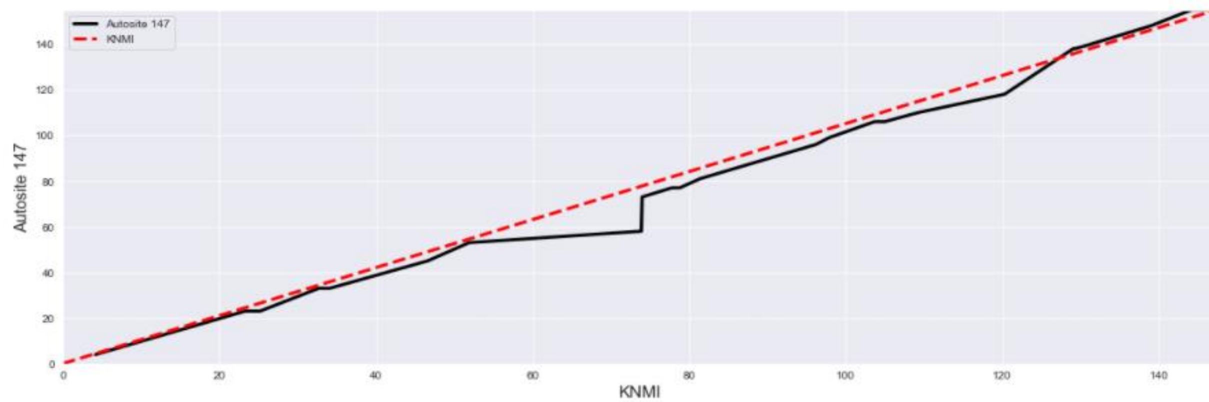

Figure 22: *Double mass precipitation Citizen Science autosite 147 VS KNMI as reference (precipitation in millimetre)*

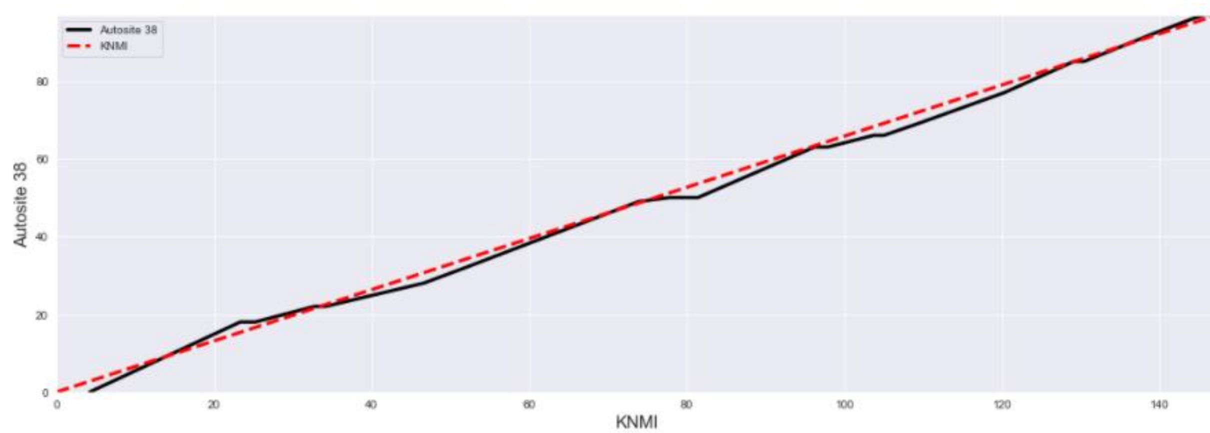

Figure 23: *Double mass precipitation Citizen Science autosite 38 VS KNMI as reference (precipitation in millimetre)*

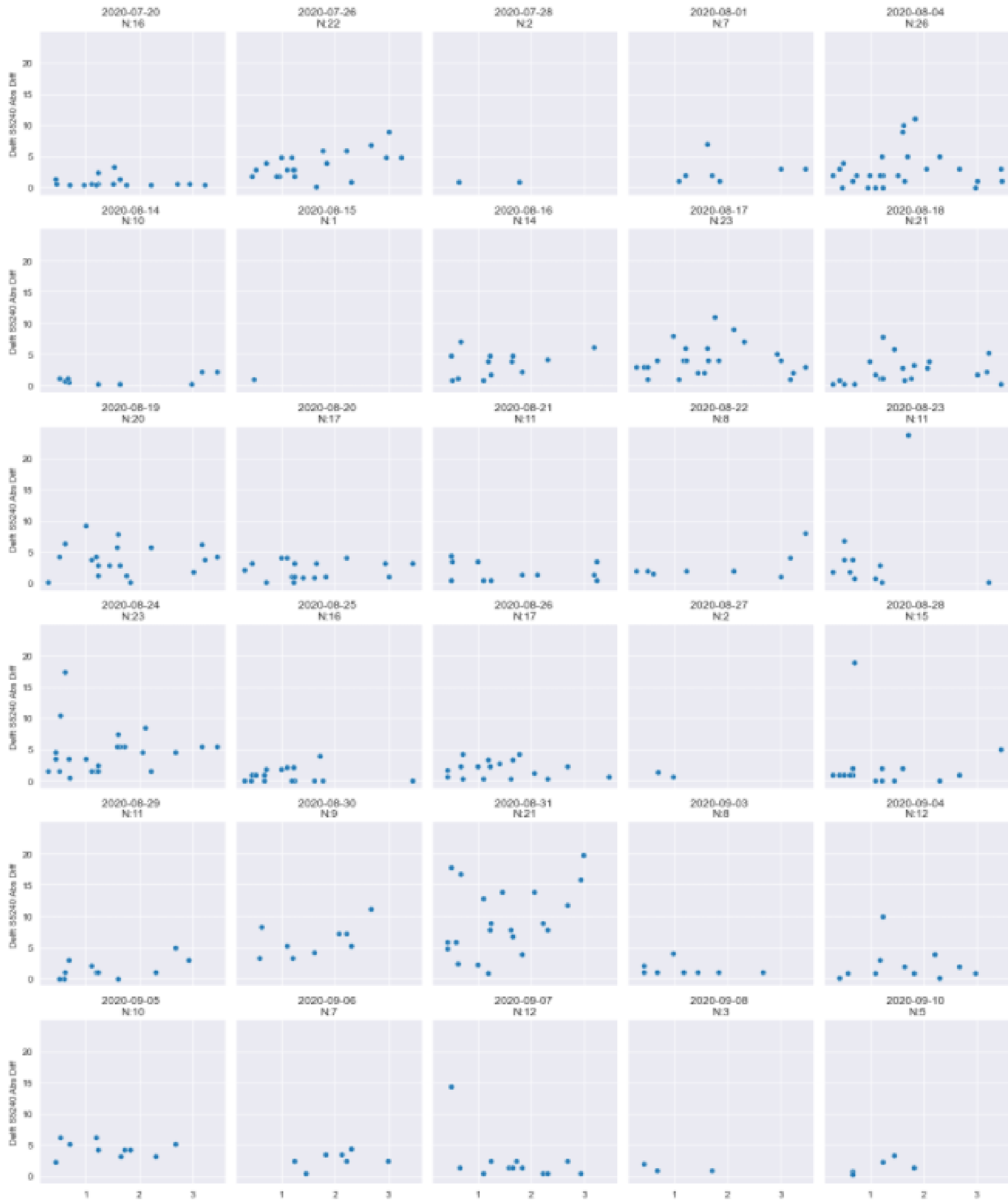

Figure 24 :*Distance difference comparison Citizen Science data (Timeframe 07:00-09:00) vs Delft gauge S5240 as reference (Distance in km, difference in mm)*

Figure 25: *Distance difference comparison Citizen Science data (Timeframe 07:00-09:00) vs Delft gauge S5241 as reference (Distance in km, difference in mm)*

Appendix B

Python code

```
]: import pandas as pd
   import matplotlib.pyplot as plt
   from matplotlib.dates import DateFormatter
   import datetime
   import seaborn as sns
   import matplotlib
   import matplotlib.pyplot as plt
   %matplotlib inline
   import copy
   import datetime
   import numpy as np
   from geopy.distance import distance
```

```
df = pd.read_excel("CitizenData.xlsx")
df["date"] = df["msmt_datetime"].str.slice(0, 10)
avg = df.groupby("date").mean()
sns.set_style('darkgrid')

fig, ax = plt.subplots(figsize=(15, 6))

avg.iloc[6:-6, :].plot(kind="bar", y="precip_mm", ax=ax)
ax.set_ylabel("precipitations in mm")
ax.legend(["CS precipitaion"])

avg.drop(columns=["AutoSiteID", "geopoint:Accuracy",
                  "geopoint:Latitude", "geopoint:Longitude"], inplace=True)
avg.reset_index(inplace=True)

plt.show()
```

```
df["hour"] = df["msmt_datetime"].str.slice(11, 13).astype(int)
sns.set_style('darkgrid')

dn = df[df["hour"].isin((7, 8))].copy()

avg_79 = dn.groupby("date").mean()
avg_79.drop(columns=["AutoSiteID", "geopoint:Accuracy",
                     "geopoint:Latitude", "geopoint:Longitude", "hour"], inplace=True)

avg_79.reset_index(inplace=True)

# merge
both = pd.merge(avg, avg_79, how="outer", on="date")
both.fillna(0, inplace=True)

fig, ax = plt.subplots(figsize=(15, 6))

both.iloc[6:-6, :].plot(kind="bar", x="date", ax=ax, width=0.8)
ax.set_ylabel("precipitations in mm", fontsize=15)
ax.legend(["CS precipitaion all data", "CS precipitation 07:00-09:00"])

plt.show()
```

Figure 26*: Python script used for figure 8*

```
df = pd.read_excel("CitizenData.xlsx")

n
df["date"] = pd.to_datetime(df["msmt_datetime"])
df["month"] = df["msmt_datetime"].str.slice(5, 7).astype(int)
```

```
months = sorted(df["month"].unique())

for month in months:

    # select the current month
    db = df[df["month"] == month]

    pie = {}
    N = len(db)

    # check all possible hours
    for h in range(1, 24, 2):
        start = f"{h:02}:00:00"
        end = f"{(h+2) % 24:02}:00:00"

        # filter by hours
        dc = db[db["date"].dt.strftime('%H:%M:%S').between(start, end)]

        # save the percentage
        ratio = len(dc) / N

        if ratio != 0:
            pie[f"from {start[:5]}-{end[:5]}"] = ratio


    for key, value in pie.items():
        if "from 05" not in key:
            del pie[key]
            pie[key] = value
        else:
            break


    fig, ax = plt.subplots(figsize=(18, 8))

    labels, x = list(pie.keys()), list(pie.values())
    ax.pie(x, labels=labels, labeldistance=None, autopct='%1.1f%%', pctdistance=1.1)

    # get month name
    name = datetime.datetime.strptime(str(month), "%m").strftime("%B")
    ax.set_title(f"The month of {name}")

    plt.legend(loc="center left",
               bbox_to_anchor=(1, 0, 0.5, 1),
               title="Hours",)
    plt.show()
```

Figure 27: *Python script used for figure 5*

```
df["day"] = df["msmt_datetime"].str.slice(0, 10)

fig, ax = plt.subplots(figsize=(18, 6))
df.iloc[7:-64, :].boxplot(by="day", column="precip_mm", ax=ax, rot=90)
```

```
dg = df[df["date"].dt.strftime('%H:%M:%S').between("07:00:00", "09:00:00")]
fig, ax = plt.subplots(figsize=(18, 6))
dg.iloc[1:-38, :].boxplot(by="day", column="precip_mm", ax=ax, rot=90)
plt.title('Boxplots precipitation in mm')
plt.ylabel('Precipitation in mm')
plt.show()
```

Figure 28: *Python script used for figure 9*

```
cit = pd.read_excel("CitizenData.xlsx")


cit["month"] = cit["msmt_datetime"].str.slice(5, 7).astype(int)
cit["day"] = cit["msmt_datetime"].str.slice(0, 10)

# filter the hours
hours = cit.copy()
hours["msmt_datetime"] = pd.to_datetime(citizens["msmt_datetime"])
seven_nine = hours[hours["msmt_datetime"].dt.strftime('%H:%M:%S').between("07:00:00", "09:00:00")]

by_day = seven_nine.groupby("day")
dc = pd.DataFrame(index=cit["day"].unique())
dc["mean"] = by_day["precip_mm"].mean()
dc["month"] = dc.index.str.slice(5, 7).astype(int)
dc.sort_index(inplace=True)
```

```
fig, ax = plt.subplots(figsize=(18, 6))

db = dc.copy()
db = db.reset_index()
db["index"] = pd.to_datetime(db["index"])

de = gov.copy()
db["Citizen site"] = db["mean"].cumsum()
de["Gov. site"] = de["gov_precip_mm"].cumsum()

db.plot(kind="line", x="index", y="Citizen site", ax=ax, linewidth=3)
de.plot(kind="line", x="date", y="Gov. site", ax=ax, color="red", linewidth=3)
ax.legend(["CS precipitation", "KNMI precipitation"])

ax.set_ylabel("precipitations in mm", fontsize=15)

ax.set_xlim(de.iloc[0, 0], de.iloc[-1, 0])

fig.autofmt_xdate(rotation=45)
plt.show()
```

Figure 29: Python script used for figure 10 & 13

```
fig, ax = plt.subplots(figsize=(18, 6))
db = dc.copy()
db = db.reset_index()
db["date"] = pd.to_datetime(db["index"])
de = gov.copy()

de["gov_cumsum"] = np.cumsum(de["gov_precip_mm"])
db["mean_cumsum"] = np.cumsum(db["mean"])
difference=(de["gov_cumsum"]-db["mean_cumsum"] )**2

# merge both sets
merged = pd.merge(db, de, on="date", how="inner")

ax.plot(merged["gov_cumsum"], merged["mean_cumsum"], color="blue",linewidth=3)
ax.plot(np.linspace(0, max(merged["gov_cumsum"]), 100),
        np.linspace(0, max(merged["mean_cumsum"]), 100), linestyle="--",color="red", linewidth=3)

ax.set_ylabel("CS cumulative", fontsize=15)
ax.set_xlabel("KNMI cumulative", fontsize=15)
ax.set_xlim(0, max(merged["gov_cumsum"]))
ax.set_ylim(0, max(merged["mean_cumsum"]))
ax.legend(["Citizen Scientist", "KNMI"])

plt.show()
print(np.sum(difference))
```

***Figure 30**: Python script used for figure 11 & 14*

```
hours = cit.copy()
hours["msmt_datetime"] = pd.to_datetime(citizens["msmt_datetime"])

seven_nine = hours[hours["msmt_datetime"].dt.strftime('%H:%M:%S').between("07:00:00", "09:00:00")]
rest_of_day = hours[~hours["msmt_datetime"].dt.strftime('%H:%M:%S').between("07:00:00", "09:00:00")]

counts_1 = seven_nine["AutoSiteID"].value_counts().reset_index()
counts_2 = rest_of_day["AutoSiteID"].value_counts().reset_index()
counts = pd.merge(counts_1, counts_2, on="index", how="outer")
counts.fillna(0, inplace=True)

# sorting
counts["total"] = counts["AutoSiteID_x"] + counts["AutoSiteID_y"]
counts.sort_values(by="total", ascending=False, inplace=True)
counts.drop(columns=["total"], inplace=True)

fig, ax = plt.subplots(figsize=(18, 8))
counts.plot(kind="bar", stacked=True, x="index", ax=ax)

ax.legend(["07:00-09:00", "Rest of the day"])
ax.set_xlabel("AutoSiteID", fontsize=15)
ax.set_ylabel("Number of measurements",fontsize=15)

plt.show()
```

***Figure 31**: Python script used for figure 7*

```
# adjust the x-axis
fig, ax = plt.subplots(figsize=(18, 6))

dn = df.copy()
dn["index"] = pd.to_datetime(dn["index"])

dn = dn.iloc[9:-6, :].copy()


dn["Citizen site"] = dn["citizens"].cumsum()
dn["s5240 site"] = dn["s5240"].cumsum()
dn["s5241 site"] = dn["s5241"].cumsum()

dn.plot(kind="line", x="index", y=["Citizen site", "s5240 site", "s5241 site"], ax=ax, linewidth=3)

ax.set_ylabel("precipitations in mm",fontsize=15)
ax.set_xlabel("The Date",fontsize=15)

ax.set_xlim(dn.iloc[0, 0], dn.iloc[-1, 0])
ax.legend(["Citizen scientist", "Delft S5240", 'Delft S5241'])


fig.autofmt_xdate(rotation=90)
plt.show()
```

***Figure 32**: Python script used for figure 13*

```
# adjust the x-axis


fig, ax = plt.subplots(figsize=(18, 6))

ax.plot(dn["Citizen site"], dn["s5240 site"], color="orange", label="Delft S5240",  linewidth=3)
ax.plot(dn["Citizen site"], dn["s5241 site"], color="green", label="Delft S5241", linewidth=3)

# dotted line
ax.plot(np.linspace(0, max(dn["Citizen site"]), 100),
        np.linspace(0, max(dn["s5241 site"]), 100), linestyle="--", linewidth=3)

ax.set_ylabel("Delft gauges cumulative",fontsize=15 )
ax.set_xlabel("CS cumulative", fontsize=15)
ax.set_xlim(0, max(dn["Citizen site"]))
ax.set_ylim(0, max(dn["s5241 site"]))

plt.legend()
plt.show()
```

***Figure 33**: Python script used for figure 14*

1

```python
cit_df = pd.read_csv('civ_data.csv')
cit_df = cit_df.astype({'precip_mm':float, 'AutoSiteID':str})

# seperate date and time for ease
cit_df['time'] = cit_df['DateAndTime'].str[11:]
cit_df['time'] = cit_df['time'].str[:2]
cit_df['time'] = cit_df['time'].astype('int32')
cit_df['DateAndTime'] = cit_df['DateAndTime'].str[:11]
cit_df['DateAndTime'] = cit_df['DateAndTime'].astype('datetime64[ns]')
cit_df['Date'] = cit_df['DateAndTime'].dt.date

# drop entries that are not between 7 and 9 am
Choose_date = '2020-08-01'
date = datetime.datetime.strptime(Choose_date,'%Y-%m-%d').date()
# cit_df = cit_df.drop(cit_df[(cit_df['time']< 7)|(cit_df['time']>8)].index)


Choose_date = '2020-07-17'
date1 = datetime.datetime.strptime(Choose_date,'%Y-%m-%d').date()
Choose_date = '2020-09-11'
date2 = datetime.datetime.strptime(Choose_date,'%Y-%m-%d').date()

SS5241_df = pd.read_csv('S5241.csv')
SS5241_df['DateAndTime'] = SS5241_df['DateAndTime'].astype('datetime64[ns]')
SS5241_df['precip'] = SS5241_df['precip']/10
SS5241_df = SS5241_df.groupby(pd.Grouper(key='DateAndTime',freq='24h',loffset='08h00min')).sum().reset_index()
SS5241_df = SS5241_df.set_index('DateAndTime')
SS5241_df['precip'] = SS5241_df['precip'].shift(1,freq='d')
SS5241_df = SS5241_df.reset_index()
SS5241_df['Date'] = SS5241_df['DateAndTime'].dt.date
SS5241_df = SS5241_df[(SS5241_df['Date'] > date1) & (SS5241_df['Date'] < date2)]
SS5241_df['cumsum'] = SS5241_df['precip'].cumsum()

SS5240_df = pd.read_csv('S5240.csv')
SS5240_df['DateAndTime'] = SS5240_df['DateAndTime'].astype('datetime64[ns]')
SS5240_df['precip'] = SS5240_df['precip']/10
SS5240_df = SS5240_df.groupby(pd.Grouper(key='DateAndTime',freq='24h',loffset='08h00min')).sum().reset_index()
SS5240_df = SS5240_df.set_index('DateAndTime')
SS5240_df['precip'] = SS5240_df['precip'].shift(1,freq='d')
SS5240_df = SS5240_df.reset_index()
SS5240_df['Date'] = SS5240_df['DateAndTime'].dt.date
SS5240_df = SS5240_df[(SS5240_df['Date'] > date1) & (SS5240_df['Date'] < date2)]
SS5240_df['cumsum'] = SS5240_df['precip'].cumsum()

# read in goverment measurements
sensorKNMI_df = pd.read_csv('sKNMI.csv')
sensorKNMI_df['DateAndTime'] = sensorKNMI_df['DateAndTime'].astype('datetime64[ns]')
sensorKNMI_df['Date'] = sensorKNMI_df['DateAndTime'].dt.date
sensorKNMI_df['precip'] = sensorKNMI_df['gov_precip']/10
sensorKNMI_df['cumsum'] = sensorKNMI_df['precip'].cumsum()

# Average geopoint per autosite
autosite_df = cit_df.groupby('AutoSiteID')['Latitude','Longitude'].mean()
#autosite_df.head()

# Isolate rainy days
rain_df = cit_df[(cit_df['precip_mm']>0)]
rain_df = rain_df[['AutoSiteID','Date','precip_mm']]
#if only one or two gauges differed that day consider it an outlier. These were manually removed form db

# join rainy days with average coordinate on autosite
joint_df = pd.merge(rain_df,autosite_df, on='AutoSiteID')
map_df = copy.deepcopy(joint_df)


closest_sensor = []
distances = []
true_diff = []
gov_r = []
measured_difference = []
sensor5240_coord = (52.010719,4.354635)
sensor5241_coord = (52.020734,4.349594)

for _, row in joint_df.iterrows():
    coord = (row['Latitude'], row['Longitude'])
    dist_sens5240 = distance(coord, sensor5240_coord).km
    dist_sens5241 = distance(coord, sensor5241_coord).km

    if dist_sens5240 > dist_sens5241:
        closest_sensor.append('Sensor5241')
        distances.append(dist_sens5241)
        sensor_row = SS5241_df[(SS5241_df['DateAndTime'] == str(row['Date']))]
        gov_r.append(float(sensor_row['precip']))
        measured_difference.append(np.abs(float(row['precip_mm'])- float(sensor_row['precip'])))
        true_diff.append(float(row['precip_mm'])- float(sensor_row['precip']))
    else:
        closest_sensor.append('Sensor5240')
        distances.append(dist_sens5240)
        sensor_row = SS5240_df[(SS5240_df['DateAndTime'] == str(row['Date']))]
        gov_r.append(float(sensor_row['precip']))
        measured_difference.append(np.abs(float(row['precip_mm']) - float(sensor_row['precip'])))
        true_diff.append(float(row['precip_mm'])- float(sensor_row['precip']))
```

2

```python
    if dist_sens5240 > dist_sens5241:
        closest_sensor.append('Sensor5241')
        distances.append(dist_sens5241)
        sensor_row = SS5241_df[(SS5241_df['DateAndTime'] == str(row['Date']))]
        gov_r.append(float(sensor_row['precip']))
        measured_difference.append(np.abs(float(row['precip_mm'])- float(sensor_row['precip'])))
        true_diff.append(float(row['precip_mm'])- float(sensor_row['precip']))
    else:
        closest_sensor.append('Sensor5240')
        distances.append(dist_sens5240)
        sensor_row = SS5240_df[(SS5240_df['DateAndTime'] == str(row['Date']))]
        gov_r.append(float(sensor_row['precip']))
        measured_difference.append(np.abs(float(row['precip_mm']) - float(sensor_row['precip'])))
        true_diff.append(float(row['precip_mm'])- float(sensor_row['precip']))

joint_df.insert(len(joint_df.columns),'Closest Sensor',closest_sensor)
joint_df.insert(len(joint_df.columns),'Closest Sensor Dist.',distances)
final_df_nonabs = copy.deepcopy(joint_df)
joint_df.insert(len(joint_df.columns),'Difference',measured_difference)
final_df_nonabs.insert(len(final_df_nonabs.columns),'Difference',true_diff)
final_df_nonabs = final_df_nonabs.sort_values('Date')
final_df = joint_df.sort_values('Date')

# drop measurements that are too far away
final_df = final_df.drop(final_df[final_df['Closest Sensor Dist.'] > 3.5].index)
final_df_nonabs = final_df_nonabs.drop(final_df_nonabs[final_df_nonabs['Closest Sensor Dist.'] > 3.5].index)


sens5240_dist = []
sens5241_dist = []
sensorKNMI_dist = []

sens5240_diff = []
sens5241_diff = []
sensorKNMI_diff = []

sens5240_precip = []
sens5241_precip = []
sensorKNMI_precip = []

sensor5240_coord = (52.010719,4.354635)
sensor5241_coord = (52.020734,4.349594)
sensorKNMI_coord = (52.000083, 4.366638)

for _, row in joint_df.iterrows():
    coord = (row['Latitude'], row['Longitude'])

    # for S5241
    dist_sens5241 = distance(coord, sensor5241_coord).km
    sens5241_dist.append(dist_sens5241)
    sensor_row = SS5241_df[(SS5241_df['DateAndTime'] == str(row['Date']))]
    sens5241_precip.append(float(sensor_row['precip']))
    sens5241_diff.append(float(row['precip_mm'])- float(sensor_row['precip']))

    # for sensor 5240
    dist_sens5240 = distance(coord, sensor5240_coord).km
    sens5240_dist.append(dist_sens5240)
    sensor_row = SS5240_df[(SS5240_df['DateAndTime'] == str(row['Date']))]
    sens5240_precip.append(float(sensor_row['precip']))
    sens5240_diff.append(float(row['precip_mm'])- float(sensor_row['precip']))

    # for sensorKNMI
    dist_sensKNMI = distance(coord, sensor5240_coord).km
    sensorKNMI_dist.append(dist_sensKNMI)
    sensor_row = sensorKNMI_df[(sensorKNMI_df['DateAndTime'] == str(row['Date']))]
    sensorKNMI_precip.append(float(sensor_row['precip']))
    sensorKNMI_diff.append(float(row['precip_mm'])- float(sensor_row['precip']))

map_df.insert(len(map_df.columns),'S5241 Dist.',distances)
map_df.insert(len(map_df.columns),'S5241 Precipitation',sens5241_precip)
map_df.insert(len(map_df.columns),'S5241 True Diff', sens5241_diff)
map_df.insert(len(map_df.columns),'S5241 Abs Diff', np.abs(sens5241_diff))

map_df.insert(len(map_df.columns),'S5240Dist.',distances)
map_df.insert(len(map_df.columns),'S5240 precipitation',sens5240_precip)
map_df.insert(len(map_df.columns),'S5240True Diff', sens5240_diff)
map_df.insert(len(map_df.columns),'S5240Abs Diff', np.abs(sens5240_diff))

map_df.insert(len(map_df.columns),'Sensor KNMI Dist.',distances)
map_df.insert(len(map_df.columns),'KNMI precipitation',sensorKNMI_precip)
map_df.insert(len(map_df.columns),'Sensor KNMI True Diff', sensorKNMI_diff)
map_df.insert(len(map_df.columns),'Sensor KNMI Abs Diff', np.abs(sensorKNMI_diff))

map_df = map_df.sort_values('Date')

# drop measurements that are too far away
map_df = map_df.drop(map_df[(map_df['S5240Dist.'] > 3.5) & (map_df['S5241 Dist.'] > 3.5 )].index)
```

```python
fig, axs = plt.subplots(1, 3,figsize=(15,5))
sns.set_style('darkgrid')

temp = multi_reg_df[multi_reg_df['KNMI precipitation']<=5]
N = temp.shape[0]
corr = np.corrcoef(temp['KNMI precipitation'],temp['Civilian Measurement'])[0,1]
slope, intercept, r_value, p_value, std_err = stats.linregress(temp['KNMI precipitation'],temp['Civilian Measurement'])
ax = sns.regplot(ax = axs[0], data = temp , x = 'KNMI precipitation', y = 'Civilian Measurement',line_kws={'label':"y={0:.1f}x\n
ax.legend(loc='upper center')

temp = multi_reg_df[multi_reg_df['KNMI precipitation']<=15]
N = temp.shape[0]
corr = np.corrcoef(temp['KNMI precipitation'],temp['Civilian Measurement'])[0,1]
slope, intercept, r_value, p_value, std_err = stats.linregress(temp['KNMI precipitation'],temp['Civilian Measurement'])
ax1 = sns.regplot(ax = axs[1], data = temp, x = 'KNMI precipitation', y = 'Civilian Measurement',line_kws={'label':"y={0:.1f}x\n
ax1.legend(loc='upper center')

temp = multi_reg_df[multi_reg_df['KNMI precipitation']<=35]
N = temp.shape[0]
corr = np.corrcoef(temp['KNMI precipitation'],temp['Civilian Measurement'])[0,1]
slope, intercept, r_value, p_value, std_err = stats.linregress(multi_reg_df['KNMI precipitation'],multi_reg_df['Civilian Measure
ax2 = sns.regplot(ax = axs[2], data = multi_reg_df, x = 'KNMI precipitation', y = 'Civilian Measurement',line_kws={'label':"y={0

ax2.legend(loc='upper center')

fig.suptitle('Sensor KNMI measured vs civilian measured precipitation')
#plt.tight_layout()
plt.show()
```

Figure 34*: Python script used for figure 12 & 15*

```python
sns.set_style('darkgrid')
# scatterplot per day: distance vs difference
g = sns.FacetGrid(final_df, col="Date", col_wrap=5, height=3)
g.map(sns.scatterplot, 'Distance', 'Difference')
axes = g.axes.flatten()

# count numbe of records for each date
N = final_df.groupby('Date').size()
i=0
for key,value in N.items():
    axes[i].set_title(str(key) + " \nN:" + str(value))
    i+=1

plt.tight_layout()
plt.show()
```

Figure 35*: Python script used for figure 16, 24 & 25*

```python
Choose_date = '2020-08-27'


date = datetime.datetime.strptime(Choose_date,'%Y-%m-%d').date()
for_day_df = final_df[(final_df['Date'] == date)]
for_day_df = for_day_df.reset_index()
gov_long = 4.366638
gov_lat  = 52.000083


plt.figure(figsize=(10,10))
sns.set_style("white")
img = plt.imread('map1.png')
BBox = [4.32,4.40,51.975, 52.03]
ax2 = sns.scatterplot(data = for_day_df, x='Longitude',y='Latitude', s=75, hue='precip_mm', alpha=1, palette='winter')
ax2.set_xlim(BBox[0],BBox[1])
ax2.set_ylim(BBox[2],BBox[3])
ax2.imshow(img, zorder=0, extent=BBox, aspect=1.5)
ax2.scatter(gov_long,gov_lat,color='red',label='KNMI')
ax2.set(title=('Rainfall measured by civilian scientist on ' + Choose_date))
ax2.legend()
plt.show()
```

Figure 36*: Python script used for figure 17 & 19*

```python
averages_df = final_df_nonabs.groupby(['AutoSiteID','Latitude','Longitude']).agg({'Difference':np.mean})
averages_df = averages_df.reset_index()
gov_long = 4.366638
gov_lat  = 52.000083


# plot averages differnce heatmap on street map
plt.figure(figsize=(10,10))
sns.set_style("white")
img = plt.imread('map1.png')
BBox = [4.32,4.40,51.975, 52.03]
ax2 = sns.scatterplot(data = averages_df, x='Longitude',y='Latitude', s=75, hue='Difference', alpha=1, palette='winter')
ax2.set_xlim(BBox[0],BBox[1])
ax2.set_ylim(BBox[2],BBox[3])
ax2.imshow(img, zorder=0, extent=BBox, aspect=1.5)
ax2.scatter(gov_long,gov_lat,color='red',label='KNMI reference')
ax2.set(title='Average true difference measured by civilian scientist')
ax2.legend()
plt.show()
```

Figure 37*: Python script used for figure 18 & 20*

```
summary_df = cit_df[['AutoSiteID','Date','precip_mm']]
summary_df = summary_df.groupby('Date').describe()
summary_df.columns = summary_df.columns.map('_'.join)
summary_df = summary_df[['precip_mm_count','precip_mm_min','precip_mm_max','precip_mm_mean','precip_mm_std']]
summary_df.columns = ['N', 'Min','Max','Mean','Std']
summary_df['N'] = summary_df['N'].astype(int)
summary_df.round(2)
```

Figure 38: *Python script used for table 1*