Challenges and benefits from crowdsourced atmospheric data for urban climate research using Berlin, Germany, as testbed

Fred Meier, Daniel Fenner, Tom Grassmann, Britta Jänicke, Marco Otto, Dieter Scherer

Chair of Climatology, Department of Ecology,
Technische Universität Berlin, Germany, fred.meier@tu-berlin.de

dated: 28 August 2015

1. Introduction

Provision of atmospheric data from observational networks at high spatial resolution and over long time periods remains a challenge in urban climate research. Classical observational networks are designed for detection of synoptic atmospheric conditions, and thus are rarely suitable for city-specific and intra-urban analysis. Therefore, using citizens as data provider offers huge potentials, especially in urban areas due to high population density.

The concept of citizen science is not new, especially in the field of ecology (Dickinson et al. 2012). This concept relies on active participation of citizens to contribute to research. A number of efforts have been made in recent years concerning atmospheric applications, e.g. mapping of atmospheric aerosols with smartphones (Snik et al., 2014) or involving citizens in observational networks such as “CoCoRaHS” (Community Collaborative Rain, Hail and Snow Network, http://www.cocorahs.org/) or the “Citizen Weather Observer Program” (http://wxqa.com). Another approach to acquire huge amounts of data is the concept of crowdsourcing, defined by Dickinson et al. (2012) as “…getting an undefined public to do work, usually directed by designated individuals or professionals…” For instance, Overeem et al. (2013) took battery-temperature measurements from smartphones to derive urban air temperatures by using data from the smartphone application ‘OpenSignal’ (opensignal.com), while Mass and Madaus (2014) exploited air-pressure measurements from another application called ‘pressureNET’ (pressurenet.cumulonimbus.ca) to simulate an active convection event in the United States.

The netatmo urban weather stations (www.netatmo.com) act as an intermediate between active citizen science and crowdsourcing of passively acquired data. The netatmo company develops and distributes weather stations around the world for interested citizens for monitoring the atmospheric conditions inside and outside their buildings. The netatmo weather station is cost-efficient, and Wi-Fi connection serves for data transfer, storage and visualisation via application software. These smart devices upload automatically their data to the netatmo server. They belong to the ‘Internet of things’, which plays an important role for recent innovations in data mining and crowdsourcing (Muller et al. 2015). While netatmo weather stations offer huge potentials due to dense spatial coverage in many urban areas, the question remains if and how crowdsourced data from this source could be suitable for urban climate research. What are the key challenges and benefits? The focus of this contribution is on crowdsourced air temperature (Ta) records.

2. Material and methods

2.1 Study site

Berlin (52° 31′ N, 13° 24′ O) is the largest city in Germany, with about 3.5 million inhabitants. We have chosen Berlin and its surroundings as test bed, because its urban climate is not directly influenced by mountains or seas, which could interfere with formation of urban heat islands (UHI). Furthermore, the urban climate observation network (UCON) of our department and the observational network of the German Weather Service can be used to evaluate crowdsourced atmospheric data. In particular, we use data from the UCON stations Dessauer Straße (DESS, LCZ 5: Open midrise) and Dahlemer Feld (DAHF, LCZ B: Scattered trees). These observational sites are described in detail in Fenner et al. (2014).

2.2 Acquisition of atmospheric data through crowdsourcing

We use the getpublicdata method, part of the PUBLIC API (application programming interface) developed by netatmo, for acquisition of public data provided by netatmo weather stations in a given area. Owners of netatmo stations can decide whether their data are public or not. The following data may be provided to the public: air temperature and humidity observations by netatmo outdoor modules, air pressure records by indoor modules, and precipitation records from rain gauges. Indoor devices also measure air temperature, humidity, and furthermore CO2 concentration and noise level, but these records are not available to the public. Metadata include timestamps of each record, as well as latitude, longitude and altitude of the station. Our crowdsourced raw data do have a temporal resolution of one hour. All records and metadata are stored in a MySQL database. For this study, we have analysed the crowdsourced data of a period of six months (Jan-Jun 2015).

2.3 Comparative measurements

Comparative measurements were conducted in the climate chamber of our department. We analysed the accuracy of eight netatmo outdoor modules. Fig. 1 (left) shows the differences between Ta as measured by netatmo sensors and a reference sensor (Vaisala HMP155) for pre-defined temperature levels between 0 and 30 degree Celsius. The results show that the netatmo sensors fulfil the specified accuracy of +/-0.3 K for the lower temperature range. There is a cold bias (0.5 - 1.0 K) for the higher temperature range.

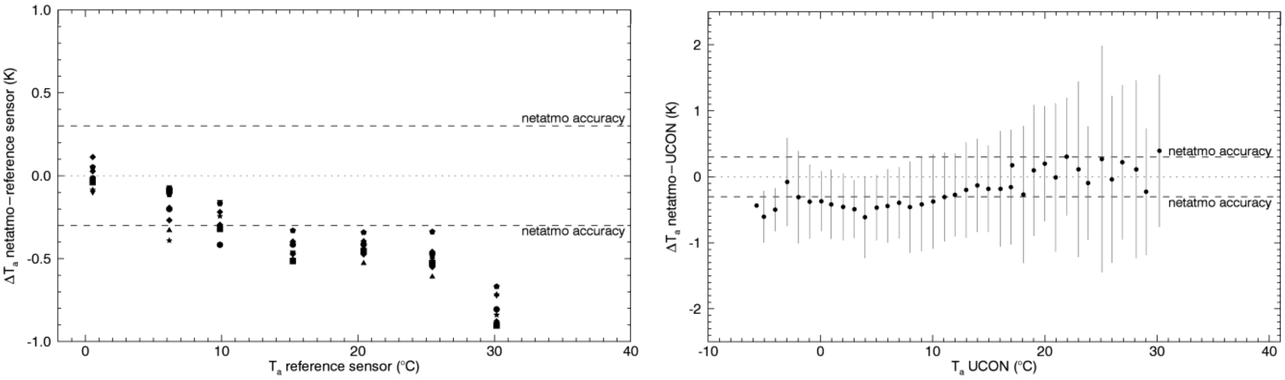

Fig. 1: Air temperature (Ta) differences between eight netatmo sensors (outdoor modules) and a reference sensor (Vaisala HMP155) for seven air temperature levels in a climate chamber (left). Ta differences between a netatmo sensor (outdoor module placed in a Stevenson screen) and a UCON sensor (Campbell CS215) observed at UCON site Rothenburgstraße in the garden of the Department of Ecology; values are binned to 1 K intervals; vertical lines represent standard deviations (right).

Offset and gain values from the climate chamber measurements were applied to Ta records obtained from our comparative measurements in the field over a period of six months. We obtain reasonable values when the netatmo sensor is situated in shadow like in the used Stevenson screen. However, even after calibration, we observe a slight cold bias for the lower temperature range. The reference sensor (Campbell CS215) belongs to UCON and is set up in a white radiation shield, which is actively ventilated during sunlit periods, otherwise naturally ventilated.

3. Results and discussion

3.1 Spatial and temporal data availability

Fig. 2 shows the spatio-temporal availability of crowdsourced raw data. We collected 4’076’734 records during the whole study period of six month. These air temperature observations belong to more than 1000 netatmo stations. From January to June the number of available stations increased from ca. 900 to 1100.

Approximately 75% of netatmo stations are located within the city of Berlin. The remaining stations are in the surrounding area. The stations are located almost exclusively in built-up areas (grey-shaded areas, CORINE Land Cover). There are no stations in urban parks, forests, grassland, or other non-built-up areas. Only occasionally, we found stations in small settlements that are not classified as built-up area in the CORINE data.

Stations with high data availability were installed by their owners before January 2015. The other stations were installed later, broken down, are not approved by their owners, or metadata have been changed by them. Once the user modified the metadata (latitude and longitude), we terminated this time series and create a new ID in our netatmo database for this device.

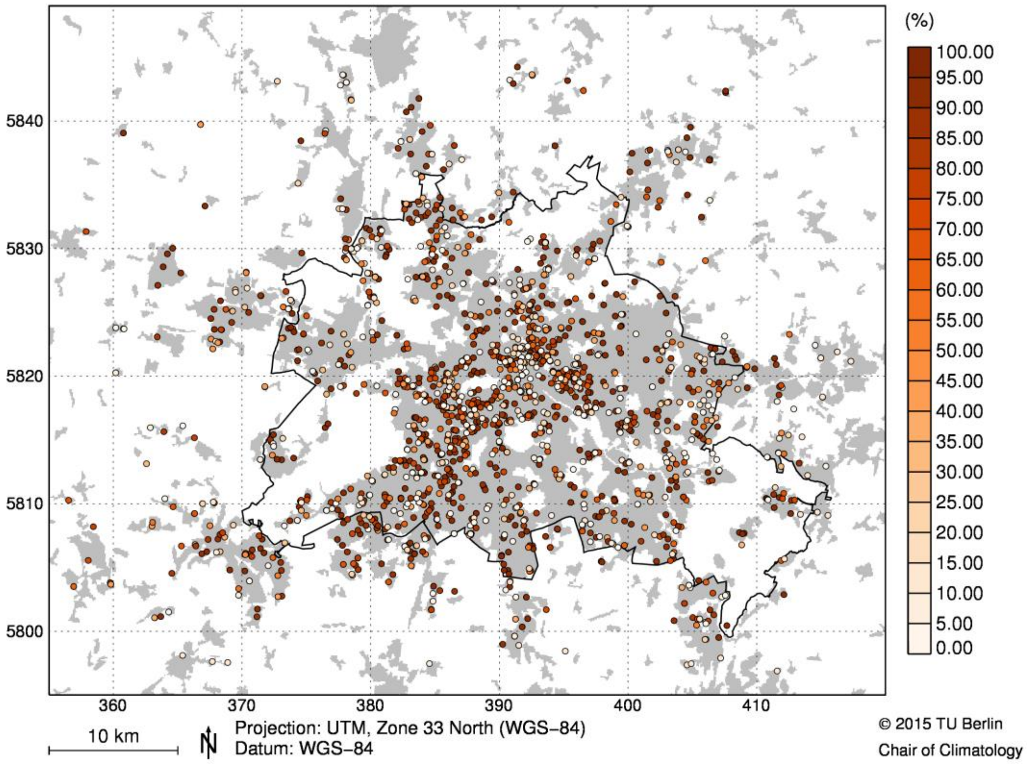

Fig. 2: Spatial and temporal data availability of air temperature records (raw data) from netatmo stations for the study period. Grey-shaded areas mark built-up areas in the CORINE land cover data set.

3.2 Challenges

Crowdsourced data can hardly be used directly to yield usable information about the state of the urban atmosphere. Assessment of data quality is a real challenge within the analysis of crowdsourced data. We classified data quality into five levels based on available station-specific metadata (latitude and longitude, timestamp), data availability and the atmospheric data themselves. Table 1 lists the defined netatmo data quality levels for air temperature records and gives a short description of the selection criteria for data filtering and possible error sources.

Table 1: Data quality level, filter criteria and potential error sources for netatmo air temperature records.

Quality Level	Data Filter Criteria for each station	Potential Error Sources
0	Crowdsourced raw data	Netatmo API and server limits
1	Invalid metadata (latitude, longitude, altitude, timestamp)	User-specific operating error
2	> 80 % hourly data per day	Intermittent failure of wireless network; loss of battery power at netatmo sites
3	> 80 % daily data per month	See above
4	Monthly average of daily minimum netatmo Ta within range of UCON thresholds (UCON DAHF = lower limit and UCON DESS = upper limit)	User-specific installation error (misuse), netatmo device has been set up indoor
5	Difference between daily maximum netatmo Ta and UCON DAHF < 3 Kelvin	Netatmo device has not been set up in shadow (no radiation shield)

During our work with crowdsourced netatmo data we identified different classes of error sources. At first, we had to deal with hard- and software limitations. In contrast to traditional observational networks, we have to manage a variable number of stations for which data are returned for each hourly request via the netatmo API. The reasons are the changing number of netatmo users, as well as query and caching limits of the netatmo server. The getpublicdata method provides near real-time data, i.e., data of previous periods are not supplied. Therefore, any failure of our netatmo MySQL database, our crowdsourcing server or the PHP-script fetching the data, results in missing values. The outdoor module wirelessly sends its measurements to the indoor module using a radio signal. Intermittent failure of this connection due to loss of battery power and obstacles in the transmission path are further reasons for missing values.

Further problems can arise through user-specific operating errors. There are no standard guidelines on how to use the devices, and therefore inappropriate installation of netatmo sensors can lead to problems. Data availability is slightly higher during day-time than during night-time. Probably, the Wi-Fi connection of some stations is switched off during night-time by the user. Furthermore, we detected invalid metadata values (latitude, longitude, altitude, timestamp) in raw data, which are not correctly set by the user. Fig. 3 shows the raw data of hourly air temperature records from January to June 2015. Several time series of Ta show no clear diurnal course and stay around 20 degree Celsius. This problem in the crowdsourced data set is related to a kind of misuse by netatmo users. Outdoor modules may measure indoor atmospheric conditions. We compared the monthly averages of daily minimum Ta with UCON values in order to filter out these stations (quality level 4).

The next class of error sources is related to user-specific and sensor-specific measurement errors due to insufficient radiation protection, as well as insufficient ventilation. An accuracy (+/- 0.3 K) of the air temperature sensor is provided by the manufacturer specifications, which could be proved by our climate chamber measurements. A problem is overestimation of air temperature during day-time (see Fig. 3, upper panel). Radiation errors during day may become very strong, since some of the netatmo stations measure not permanently in shadow, and thus are not installed correctly. We used daily maximum values of our UCON station DAHF in order to filter out these erroneous values (quality level 5). The lower panel of Fig. 3 shows hourly netatmo values at data quality level 5 in relation to Ta observed at two UCON sites. In May, our crowdsourcing server was down for several days, and therefore, there are no data available for this month with quality level 5. Linear correlation coefficients between UCON and netatmo time-series vary between 0.80 and 0.99 (histogram distribution's peak: 0.98) at quality level 5. For days without a server failure, the daily number of available netatmo station varies between 350 and 831 for this quality level.

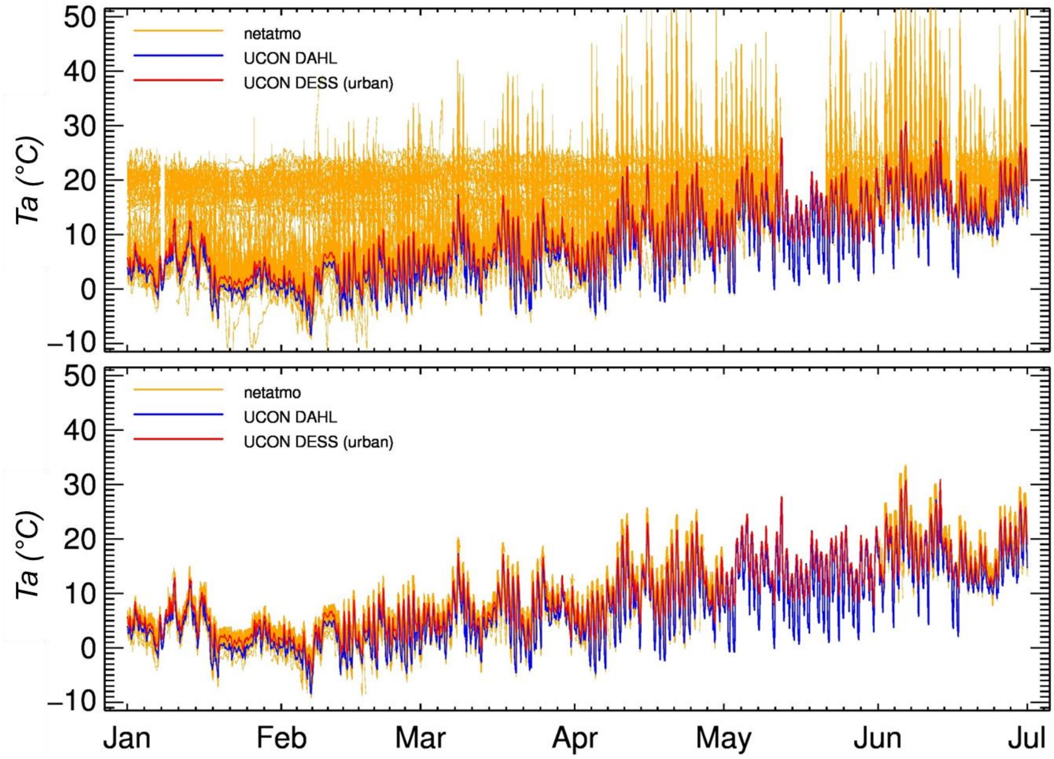

Figure 3: Hourly time-series of crowdsourced air temperature from netatmo stations in Berlin and surrounding areas (orange) and observed air temperature at UCON stations Dessauer Straße (DESS, red) and Dahlemer Feld (DAHF, blue). The upper panel shows hourly values of crowdsourced raw data (data quality level 0) and the lower panel shows hourly netatmo values (data quality level 5).

3.3 Benefits from crowdsourced atmospheric data

Sufficient spatial coverage of various urban morphologies is important for detailed investigation of urban heat island (UHI) effects. For each netatmo site within the city of Berlin, we calculated mean sky-view factor (SVF) for a radius of 250 m, including vegetation effects. We used the SOLWEIG model (Lindberg and Grimmond 2011) and the ‘‘Building and Vegetation Heights 2010’’ data set provided by Berlin Senate Department for Urban Development and Environment (2010) for SVF calculation. The trunk zone of the vegetation was set to a fraction of 0.25 of the vegetation height for the calculation. Furthermore, we calculated the complete aspect ratio for each netatmo site within the city. The results show that netatmo stations cover a wide range of urban morphologies in Berlin, even after data filtering (Fig. 4). For instance, at data quality level 4 the number of available stations within the city of Berlin amounts to 630 for July.

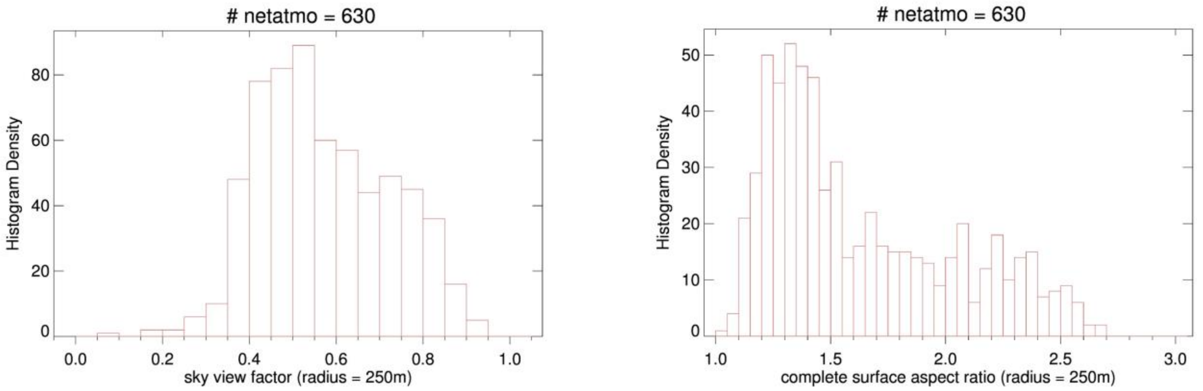

Figure 4: Histogram of the sky-view factor (left) and complete surface aspect ratio (right) for available netatmo stations within the city of Berlin (data quality level 4). The binsize of each histogram is 0.05.

The highest UHI intensities, defined as the difference between air temperatures observed at the urban UCON site DESS and the reference UCON site DAHF, were recorded during summer nights (Fenner et al. 2014). Therefore, we used netatmo Ta values of data quality level 4 for calculation of night-time Ta differences between netatmo sites and the reference UCON site DAHF. We did not use data quality level 5 for this analysis, because day-time overestimation of air temperature does not influence night-time air temperatures. Regression analysis revealed no relation between higher day-time values (due to radiation errors) and night-time values. The map displayed in Fig. 5 shows the spatial distribution of monthly maximum night-time UHI intensities during June for 794 netatmo weather stations. The netatmo UHI intensities correspond very well to the values derived from the UCON and DWD observations. The netatmo data show distinctive UHI patterns for Berlin. The inner city neighborhoods are clearly warmer than built-up areas in the outskirts during the night. A further advantage of the netatmo system is that all devices use the same type of air temperature sensor.

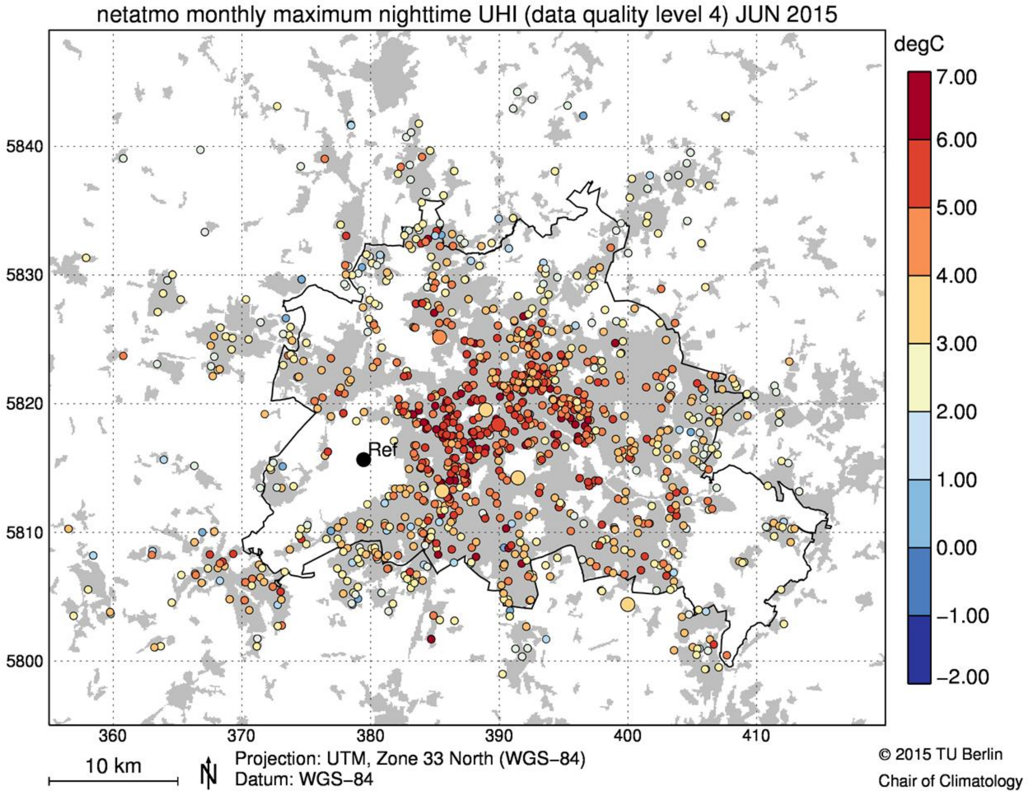

Figure 5: Spatial distribution of urban heat island (UHI) intensities i.e. monthly maximum night-time UHI intensities for 794 netatmo weather stations within the city of Berlin and surrounding area. Large circles represent UHI intensities observed at three DWD and three UCON sites. The large black circle represents the reference UCON site Dahlemer Feld, which is used for UHI calculation. The circles overlap partially because the density of netatmo stations is remarkable higher in the city centre. Grey-shaded areas mark built-up areas in the CORINE land cover data set.

4. Conclusions and outlook

Our study showed that crowdsourced atmospheric data can contribute to urban climate research. It is possible to explore the urban atmosphere with crowdsourced air temperatures from netatmo weather stations and available metadata (geographic and timestamp information). Our results show a distinctive UHI pattern in Berlin during the night. The spatial density of available stations in Berlin exceeds that of our classical monitoring networks by far, but observations at urban and rural sites of standardized, calibrated and quality-checked networks are essential in order to validate such crowdsourced data. How many netatmo stations there are in other cities and how are they distributed throughout the city?

Quality assurance is a huge challenge. In any case, we must take into account sensor-specific characteristics and user-specific peculiarities of the growing netatmo network. We showed that it is possible to filter out erroneous air temperature records. These filter techniques will be improved (for instance by the use of regression statistics) in order to be applicable to crowdsourced data from other cities. The map in Fig. 6 gives an insight into future work with netatmo data from Berlin and surrounding areas by showing the monthly sum of precipitation derived from crowdsourced records of 346 netatmo rain gauges. We will continue to explore the potential of crowdsourced atmospheric data for urban climate research in Berlin and other cities.

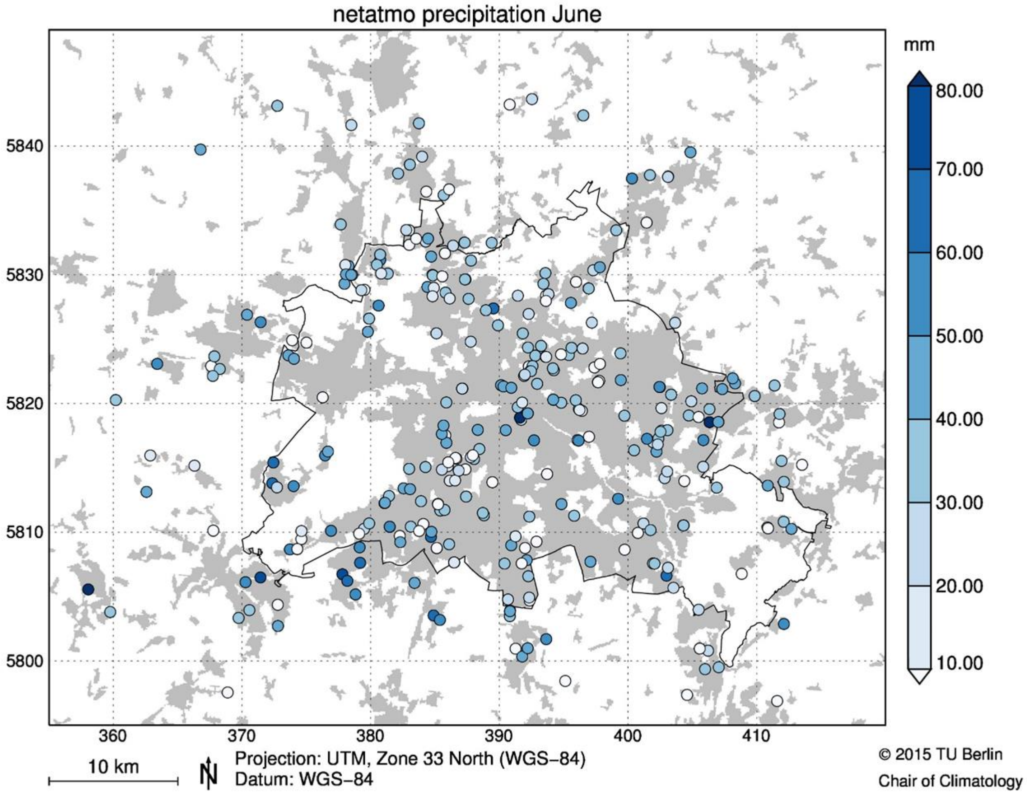

Figure 6: Spatial distribution of monthly sum of precipitation observed by 346 netatmo weather stations during June 2015 within the city of Berlin and surrounding area. The grey shaded area marks the built-up areas in the CORINE data set.

Acknowledgment

The authors wish to thank all owners of netatmo stations for providing their data to the public. The study is part of the Research Unit 1736 “Urban Climate and Heat Stress in mid-latitude cities in view of climate change (UCaHS)” (www.UCaHS.org) funded by the Deutsche Forschungsgemeinschaft (DFG) under the codes SCHE 750/8-1, SCHE 750/9-1.

References

Berlin Senate Department for Urban Development and the Environment, Urban and Environmental Information System, Environmental Atlas: Building and Vegetation Heights 2010

Dickinson J.L., Zuckerberg B., Bonter D.N., 2010: Citizen science as an ecological research tool: challenges and benefits. *Annu. Rev. Ecol. Evol. Syst.*, **41**, 149-72.

Fenner D., Meier F., Scherer D., Polze A., 2014: Spatial and temporal air temperature variability in Berlin, Germany, during the years 2001-2010. *Urban Climate*, **10**, 308-331.

Lindberg F., Grimmond C.S.B., 2011: The influence of vegetation and building morphology on shadow patterns and mean radiant temperatures in urban areas: model development and evaluation. *Theor. Appl. Climatol.*, **105**, 311-323.

Mass C.F., Madaus L.E., 2014: Surface Pressure Observations from Smartphones: A Potential Revolution for High-Resolution Weather Prediction? *Bull. Amer. Meteor. Soc.*, **95**, 1343-1349.

Muller C.L., Chapman L., Johnston S., Kidd C., Illingworth S., Foody G., Overeem A., Leigh R.R., 2015: Crowdsourcing for climate and atmospheric sciences: current status and future potential. International *Journal of Climatology*, doi:10.1002/joc.4210.

Overeem A., Robinson J.C.R., Leijnse H., Steeneveld G.J., Horn B.K.P., Uijlenhoet R., 2013: Crowdsourcing urban air temperatures from smartphone battery temperatures. *Geophys. Res. Lett.*, **40**, 4081–4085.

Snik F., Rietjens J.H.H., Apituley A., Volten H., Mijling B., Di Noia A., Heikamp S., Heinsbroek R.C., Hasekamp O.P., Smit J.M., Vonk J., Stam D.M., van Harten G., de Boer J., Keller C.U., 3187 iSPEX citizen scientists, 2014: Mapping atmospheric aerosols with a citizen science network of smartphone spectropolarimeters. *Geophys. Res. Lett.*, **41**, 1-8.