Met Office

Quality-control of vehicle-based temperature observations and future recommendations

February 18, 2021

Weather science technical report 644

Zackary Bell[1], Sarah L. Dance[1,2], Joanne A. Waller[3], Katharine O'Boyle[4]

[1]Department of Meteorology, University of Reading, Reading, UK ;

[2]Department of Mathematics and Statistics, University of Reading, Reading, UK;

[3] Met Office@Reading, University of Reading, Reading, UK;

[4]Met Office, Exeter, UK

Met Office

Contents

Abstract

In numerical weather prediction, *datasets of opportunity* is a collective term used for meteorological observations obtained from unconventional data sources. This report presents the quality-control process designed for a vehicle-based dataset of opportunity containing 67959 observations obtained from a proof-of-concept trial run from 20th February until 30th April 2018 by the Met Office. In this trial, on-board diagnostic (OBD) dongles were used to transmit low precision dry bulb temperature measurements from a vehicle to the driver's phone which were subsequently uploaded to the Met Office cloud servers with time, location, and vehicle identification metadata using a Met Office phone app. The raw data from the trial was first filtered to remove observations with missing fields or invalid measurements and metadata. The resultant filtered dataset contained 32179 observations (47.4% of original dataset) which underwent further testing. The quality-control tests applied to the filtered dataset included a climatological range test, a stuck instrument test, and a GPS test that checks whether an observation location is physically consistent with the vehicle's previous location. A substantial number of observations were flagged by the GPS test due to the accuracy of smartphone GPS measurements, GPS location update app settings, and poor GPS signal, while the majority of observations passed the climatological range and stuck instrument test. The 19094 observations which passed all previous quality-control tests were put through a final sensor ventilation test to determine if the vehicle drove at a sufficient speed for the temperature sensor to be adequately ventilated. This test flagged 1669 observations with speeds below a predetermined sensor ventilation threshold. In total, the quality-controlled dataset consists of 17425 observations (25.6% of original dataset). The results of the quality-control process have shown that the observation location metadata can be inaccurate due to unsuitable app settings and poor GPS signal. Additionally, inadequate sensor ventilation can result in observations with a warm bias. Recommendations on future data collection include revising OBD dongle and app settings/features to correct observation GPS, methods to circumvent the need for vehicle identification in quality-control, and the use of higher precision instruments.

1 Introduction

The advancement of convection-allowing data assimilation requires a large number of observations of high spatio-temporal resolution relevant to the weather processes being modelled [Sun et al., 2014, Gustafsson et al., 2018, Dance et al., 2019]. Due to the enormous cost of installation, management, and maintenance, it may not be practical to extend traditional scientific observing networks. A potential alternative source of inexpensive, high-resolution meteorological observations to constrain convective-scale numerical weather prediction forecasts is from crowdsourced data which are currently receiving increased interest from the numerical weather prediction community (e.g. Nipen et al. [2019]).

In the context of numerical weather prediction, crowdsourced data collectively refers to reports and data generated by the public through use of privately owned equipment [Hintz et al., 2019a]. These observations will be inaccurate in comparison to those obtained from scientific observing networks but have the potential to far exceed the number of scientific observations currently produced. Application of

crowdsourced data is relatively new and studies into collection methods and observation characteristics are currently active areas of research (e.g., Bell et al. [2015], McNicholas and Mass [2018b], Hintz et al. [2019a], Nipen et al. [2019]).

Crowdsourced observations from citizen observing networks have been shown to successfully observe meteorological phenomena such as urban heat islands directly through digital and car thermometer observations [Knight et al., 2010] and amateur/citizen science weather stations observations [Steeneveld et al., 2011, Wolters and Brandsma, 2012, Chapman et al., 2017, Meier et al., 2017]. Additionally, urban heat islands have been shown to be observable through temperature measurements derived from smartphone battery temperatures [Overeem et al., 2013, Droste et al., 2017]. To ensure representative measurements from direct observations, several precautionary measures must be taken into consideration [Bell et al., 2015]. For instance, citizen weather stations need to be shielded from radiation and temperature sensors need to be located a sufficient distance from buildings and in a naturally ventilated area. Sensor ventilation is especially important to prevent large air temperature errors in circumstances of large radiative forcing (e.g. World Meteorological Organization [2008], Richardson et al. [1999], Nakamura and Mahrt [2005]). Though guidance on the proper use of meteorological instruments was provided in these studies, the overall credibility of the crowdsourced datasets must be evaluated through quality-control methods before the observed meteorological processes can be examined.

Quality-control is a vital process performed prior to data assimilation to reject observations that are likely to contain gross errors. The techniques which comprise the quality-control procedure usually include simple checks designed to test different aspects of the observed values [Zahumenskỳ, 2004, Fiebrich et al., 2010]. Quality-control of crowdsourced observations is a notably difficult task [Muller et al., 2015]. Non-traditional data sources may suffer from numerous issues that traditional scientific observations will not (e.g. calibration, user behaviour such as locations above/below ground, low precision data, sensor-specific measurement errors, etc). As an example, surface pressure observations from smartphones may not be the desired meteorological measurement due to user behaviour and inadequate location and elevation metadata. We note that while smartphone GPS accuracy may be degraded by urban environments, recent studies on horizontal position accuracy in urban areas have shown they can produce acceptable location-based metadata for crowdsourced observations (e.g. Merry and Bettinger [2019]). Due to the previously mentioned issues over half the smartphone observations examined in the work of Madaus and Mass [2017] and Hintz et al. [2019b] were removed through quality-control. However, smartphone observations have been shown to improve forecasts after bias correction and thorough quality-control had been applied to the data [McNicholas and Mass, 2018a,b, Hintz et al., 2019b].

Adverse weather is a leading contributor to severe congestion, large travel time delays and harmful incidents for surface transportation networks [Snelder and Calvert, 2016]. To combat this, vehicle-based meteorological observations have been utilised by the Clarus Initiative [Limber et al., 2010] and the Pikalert System [Boyce et al., 2017, Siems-Anderson et al., 2019] to provide improved road and

atmospheric hazard products to road maintenance operators and the travelling public. Examples of observations obtainable by vehicles include precipitation (e.g. Haberlandt and Sester [2010], Rabiei et al. [2013]), air quality (e.g. Devarakonda et al. [2013], Rada et al. [2016]), atmospheric pressure and temperature (e.g. Drobot et al. [2010], Chapman et al. [2010], Anderson et al. [2012]). Studies on the quality-control of vehicle-based temperature and pressure observations have been for idealised experiments where a set number of vehicles were driven along predetermined routes using fitted meteorological sensors with known error characteristics in order to generate weather statistics for road segments [Chapman et al., 2010, Drobot et al., 2010, Anderson et al., 2012]. This freedom over the experimental design allowed for Anderson et al. [2012] to use spatial consistency checks between neighbouring vehicles and nearby surface-stations as well as sensor range tests to check the meteorological instruments. Chapman et al. [2010] also used a nearby surface-station consistency check and a sensor range test but used a climatological range test instead of a spatial consistency test between neighbouring vehicles. In contrast, the small novel dataset examined in this report allowed for no such freedom in the experimental design and severely limits the quality-control tests applicable.

The objective of this report is to quality-control a vehicle-based observation dataset and to provide recommendations on future data collection methods. The structure of this report is as follows. In section 2 we provide an overview of the Met Office proof-of-concept vehicle-based observations trial. Preparation of the data obtained from the trial into a filtered dataset to be quality-controlled is discussed in section 3. The detailed description of the implementation and results of the quality-control tests is given in section 4. The results of the quality-control process highlight that the observation location metadata can be inaccurate due to poor GPS signal and app settings. Additionally, inadequate sensor ventilation can result in observations with a warm bias. Discussion of the quality-controlled dataset is given in section 5. A summary with recommendations is provided in section 6.

2 Met Office trial

The vehicle-based observations studied in this report are obtained from a trial by the Met Office from 20th February 2018 to 30th April 2018. In the trial, volunteer participants connected an on-board diagnostics (OBD) dongle to their vehicle engine management interface. The OBD dongle used in this trial is an inexpensive adapter which is inserted into the OBD port of a vehicle which transfers vehicle diagnostic data to a Bluetooth-connected computing device. Meteorological parameters and vehicle speed (given in km/h) selected on the app “Met Office OBD App 0909” are broadcast to the participant’s Android phone. This app decodes any data sent by the dongle and appends the corresponding date-time (given by the app as the date and 24 hour clock time) and GPS location metadata derived through the phone. In addition, a unique observation ID and sensor ID are appended to each observation. A sensor ID is used to determine if observations come from the same vehicle. However, the sensor ID is specific to the installed version of the app and so reinstalling the app used to record the observations

would result in a new identifier for the participant. The observations along with relevant metadata are then uploaded via 3G or 4G to the Met Office Weather Observations Website [Met Office, 2011]. The preset data collection frequency and GPS update period are set to 1 minute while the preset minimum distance for a GPS update is 500 metres. These settings can be changed by the participant through the app interface.

In total, 31 participants were successful in producing vehicle-based observations from journeys they undertook during the trial period. For clarity, we define a journey as a subset of data originating from the same vehicle over a fixed time interval. Throughout this trial, fewer observations were usually obtained during weekends than week-days. A "call for data" was made to obtain data for certain weather conditions on the 9th March (rainy), 22nd March (benign) and 16th April (sunny). In addition, a large number of observations were obtained on 27th March due to a few long journeys.

The observations of interest obtained through this trial include dry bulb temperature (°C), engine intake temperature (°C) and air pressure (hPa). Both temperature observations have low precision (1°C). The air pressure is precise to 10hPa and hence not useful in data assimilation. In this report we examine dry bulb temperature only. Engine intake temperature measures the dry bulb temperature in the vehicle engine which will not reflect the true atmospheric air temperature. However, a fault known to have occurred for some observations during this trial is for engine intake temperature to be recorded as dry bulb temperature.

3 The filtered dataset

In order to assess the quality of a dataset, each datum must contain information we are interested in examining. Any data which are obviously in error or do not have the relevant observation field we wish to examine will be discarded. The remaining data will be referred to as the *filtered dataset* and will undergo the quality-control process detailed in section 4.

3.1 Preparation of the filtered dataset

To obtain the filtered dataset we carry out some gross checks on the dry bulb temperature measurements and accompanying time and speed metadata of the complete dataset. In particular, we discard any datum which exhibit any of the following properties:

1. the vehicle speed is less than 0km/h,
2. there is an invalid date-time in the observation metadata,
3. there is no dry bulb temperature observation.

Data that do not exhibit any of these properties will form the filtered dataset.

In total, 35780 observations were discarded resulting in the filtered dataset containing 32197 observations. No observations were discarded due to an invalid date-time. Many of the discarded observations contain speeds with value -32768km/h which is the value used when the vehicle speed is unable to be recorded by the app. This is the minimum short signed integer for a 2-bit system. A single observation with speed 255km/h, over double the speed limit for dual carriageways and motorways in the UK [UK Government, 2015], is discarded as it is likely incorrect. There were 34681 discarded observations without a dry bulb temperature field.

The data that passed this data filtering test underwent further quality-control tests. We now examine the observations in the filtered dataset before we describe the implementation of the GPS and sensor ventilation tests detailed in sections 4.3 and 4.4 respectively.

3.2 Characteristics of the filtered dataset

To implement the quality-control process described in section 4 it is necessary to have knowledge of the characteristics of the filtered dataset. In this section we examine the dry bulb temperature observations (section 3.2.1), the temporal frequency of the observations (section 3.2.2) and the GPS measurements (section 3.2.3) of the filtered dataset.

3.2.1 Dry bulb temperature

The number of dry bulb observations for each month as well as their distribution is shown in figure 1. In total, 5684 observations were taken in February, 16211 in March, and 10284 in April. We note that, except the two outliers in April that had a dry bulb temperature of $-22°$C, there are no observations with dry bulb temperature less than $-8°$C. Any observations with a temperature greater than $20°$C occurred in April. Furthermore, the largest dry bulb temperature values (27-$34°$C) occurred for vehicle speeds less than 25km/h. As discussed in section 1, this is likely to be caused by inadequate sensor ventilation.

	February	March	April	All data
Number of observations	5684	16211	10282	32175
Mean	3.33°C	6.59°C	11.38°C	7.55°C
Standard deviation	3.99°C	3.48°C	4.34°C	4.84°C
Skew	−0.09	−0.44	0.67	0.27
Excess kurtosis	−0.74	0.27	1.13	1.14

Table 1: Summary of the descriptive statistics for the filtered dataset and split into each month. The two values of $-22°$C have been removed from the descriptive statistics calculations of “April” and “All data” as they are clearly outliers.

A summary of the descriptive statistics for the filtered dataset and each month is given in table 1. As expected, the distribution of dry bulb temperatures varies significantly with each month with mean temperatures increasing from February to April due to season. The distribution for February is the most

concentrated distribution (negative excess kurtosis) with slight asymmetry (negative skew). March is a less concentrated distribution than February with a longer left tail. The left tails for both February and March are likely to be caused by the Beast from the East (22/02/2018 to 05/03/2018 [Met Office, 2020]). Removing the extreme dry bulb temperatures from April reveals it is the least concentrated distribution with a longer right tail that is likely caused by inadequate sensor ventilation during more frequent sunny weather conditions and the April hot spell (18th-22nd April 2018 [Met Office, 2020]). Examining the variability of each month shows that April is the most variable month (highest standard deviation) and March the least variable. As the majority of the observations occurred in March and April, the characteristics of the combined temperature distribution share the most similarities with these months. Namely, the distribution is concentrated around similar temperatures to March and has a longer right tail like April. However, the combined temperature distribution has higher variability than each individual month.

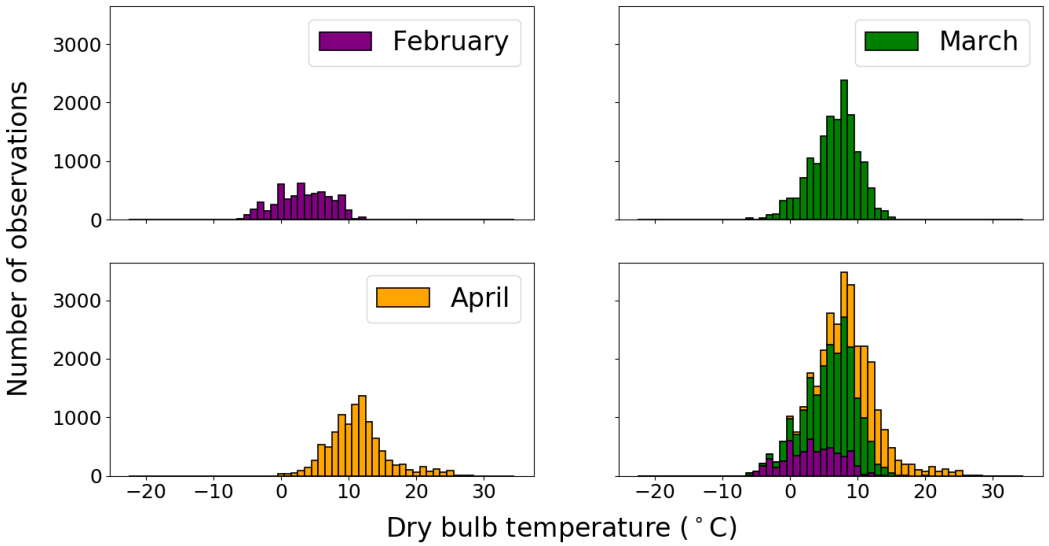

Figure 1: Distribution of dry bulb temperature observations for each month of the trial for the filtered dataset. The purple bar segments indicate the number of February 2018 observations, the green segments March 2018, and the orange segments April 2018. The combined distribution is a stacked histogram that shows the contribution from each month to the total number of observations for each dry bulb temperature.

3.2.2 Multiple reporting

The observations from this trial were designed to have a 1 minute temporal frequency. However, several observations are reported within a minute of the previous observation but retained a 1 minute temporal frequency with other similar observations. An example of this is shown in figure 2 which shows the dry bulb temperature observations for a journey segment from a single vehicle. The blue observations maintain a 1 minute temporal frequency with other blue observations while the red observations occur 7 seconds after a blue observation and maintain a 1 minute temporal frequency with other red

observations. An observation reported within a minute of the previous observation will be referred to as a *multiple reported observation* (MRO) and is suspected to be caused by phone or dongle hardware technical issues. We note that the GPS metadata may be inaccurate for MROs due to the short time between observations. This will be discussed further in section 4.3.

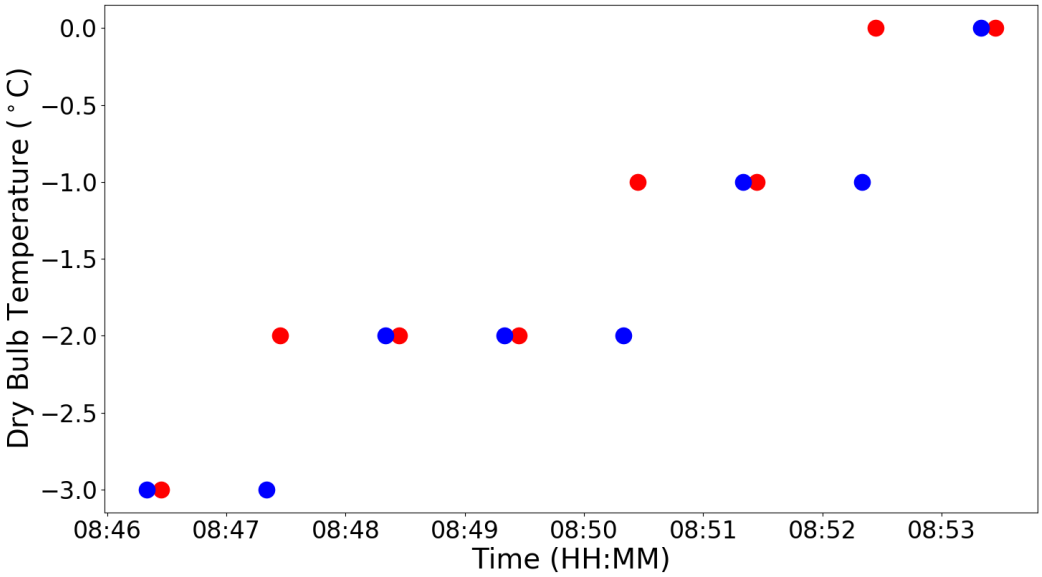

Figure 2: Time series segment of dry bulb temperature for a single vehicle taken on the 28th of February 2018. The blue observations retain a 1 minute temporal frequency with other blue observations. The red observations retain a 1 minute temporal frequency with other red observations but always occur 7 seconds after a blue observation.

Denoting the time-gap between consecutive observations from the same vehicle on a given day as Δt, a histogram of all Δt binned into one-minute intervals is shown in figure 3. Noting the log-scale, the vast majority of Δt have length 0-2 mins (i.e. the first two bins). The first bin contains all $\Delta t \in [0, 1)$ minutes which occurs 15364 times. This corresponds to the number of MROs in the filtered dataset. MROs form a large part of the filtered dataset and will need to be accounted for in the additional quality check tests we impose on this data. The second bin contains all $\Delta t \in [1, 2)$ minutes which occurs 15025 times with 14646 corresponding to the observations which maintain a 1 minute temporal frequency with the previous observation. Any $\Delta t \geq 60$ minutes are placed in the $60+$ minute bin. Any $\Delta t \in [2, 60)$ are caused by breaks in a vehicle journey, issues with the collection method such as loss of phone signal, or removal of data without the necessary information needed for quality-control.

3.2.3 GPS-lagged observations

The method of data collection in the Met Office trial used smartphones to obtain location-based metadata. A common occurrence in the filtered dataset is GPS location not updating due to poor GPS signal, the vehicle has not travelled far enough to trigger a GPS update or insufficient time between observa-

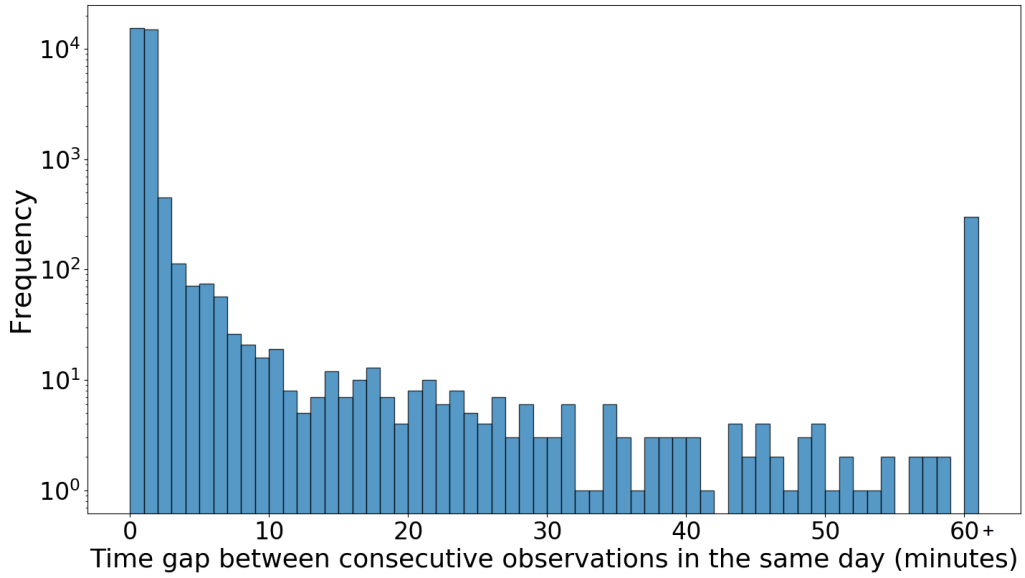

Figure 3: The distribution of time-gaps between consecutive observations from the same vehicle and day for the filtered data set binned into minute intervals. Note that a log-scale has been used for the time-gaps frequency. Any time-gaps greater than 60 mins are placed into the $60+$ minute interval.

tions. The default GPS update distance and period for the app are 500 metres and 60 seconds respectively. This results in some observations having identical GPS location to the previous observation taken by the same vehicle. These observations will be referred to as *GPS-lagged observations*.

Figure 4 shows data from the first half of a journey along the M5 motorway during the 25th of March 2018. We have plotted a time series of distance between consecutive observations calculated using the great circle distance,

$$d = 2r \sin^{-1}\left(\sqrt{\sin^2\left(\frac{\Phi_1 - \Phi_2}{2}\right) + \cos\left(\Phi_1\right)\cos\left(\Phi_2\right)\sin^2\left(\frac{\lambda_1 - \lambda_2}{2}\right)}\right), \tag{1}$$

where Φ_1 (λ_1), Φ_2 (λ_2), are the latitudes (longitudes) of the two locations and $r = 6371$km is the radius of the Earth. We also show the distance estimated using the time-gap between the observations and the speeds of the observations,

$$d^e_{max} = \mathsf{max}(v_1, v_2) \times \Delta t, \tag{2}$$

where v_1 and v_2 are the recorded speeds at the time of the two observations and Δt is the time-gap between them. We have used two estimates to provide reference for when a distance is realistic (i.e. $d \approx d^e_{max}$) or unrealistic (i.e. $d \approx 2d^e_{max}$). Instances of $d = 0$km indicate the observation at that time is a GPS-lagged observation. Almost all instances of $d = 0$km are immediately followed by $d^e_{max} < d \leq 2d^e_{max}$ which corresponds to the distance travelled between the observation before and after the GPS-lagged observation.

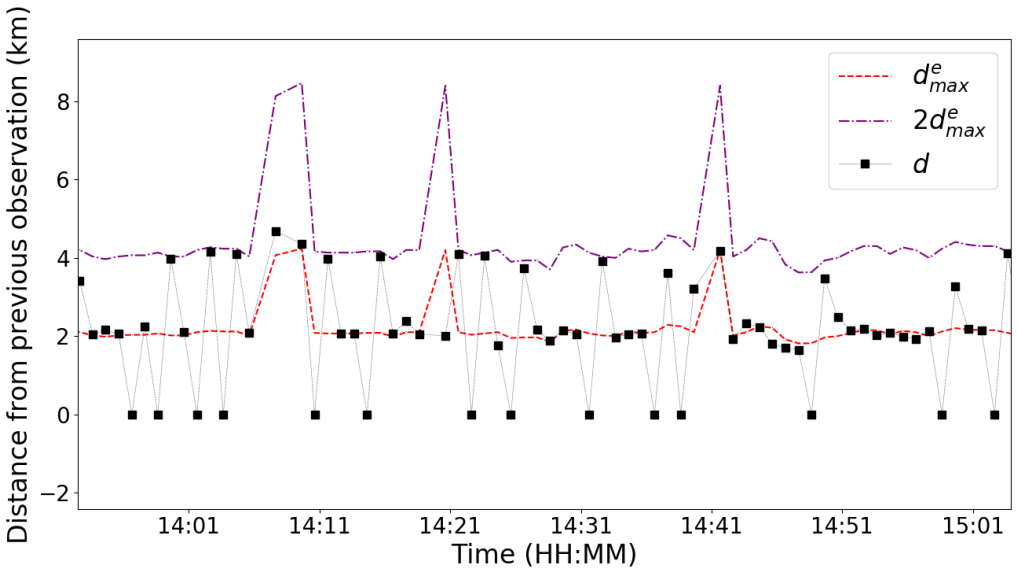

Figure 4: Time series of the distance between consecutive observations calculated using the great circle distance (black squares), denoted d, the distance estimated using the time-gap between the observations and the speeds of the observations, denoted d^e_{max} (red-dashed line) and $2d^e_{max}$ (purple dot-dashed line), for the first half of a journey along the M5 motorway during the 25th of March 2018. We have used two estimates to provide a reference for when a distance is realistic (i.e. close to the red line) or unrealistic (i.e. close to the purple line). Instances of $d = 0$km indicate the observation at that time is a GPS-lagged observation. Almost all instances of $d = 0$km are immediately followed by $d^e_{max} < d \leq 2d^e_{max}$ which corresponds to the distance travelled between the observation before and after the GPS-lagged observation.

4 Quality-control tests

In this section we describe the quality-control tests we use on the filtered dataset. A schematic showing the complete quality-control process applied to the trial data is shown in figure 5. The climatological range test (section 4.1), stuck instrument test (section 4.2) and GPS test (section 4.3) are applied in parallel to the filtered dataset obtained in section 3. Observations that have passed each quality-control test undergo a final sensor ventilation test (section 4.4). The final quality-controlled dataset will be comprised of observations passed by every quality-control test. Throughout this section we use the units that the data have been recorded in which are given in section 2.

4.1 Climatological range test (CRT)

The Climatological range test (CRT) identifies observations which fall outside of location-specific climatological ranges [Limber et al., 2010, Boyce et al., 2017].

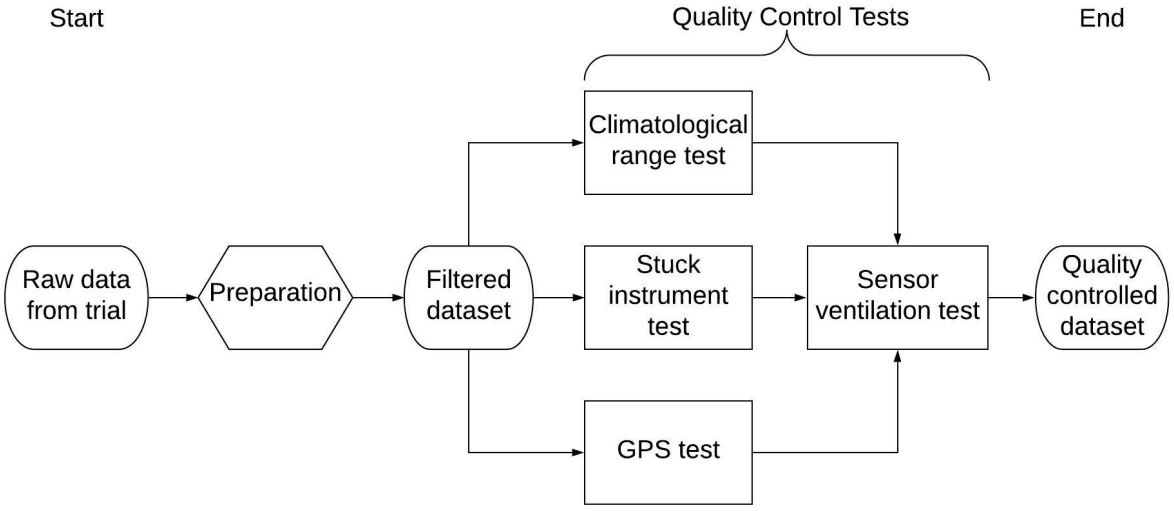

Figure 5: The complete quality-control process applied to the complete dataset obtained from the Met Office trial. The complete data is first prepared into the filtered dataset whereby all data without dry bulb temperature and necessary metadata are discarded. The climatological range test (section 4.1), stuck instrument test (section 4.2) and GPS test (section 4.3) are applied in parallel to the filtered dataset. Observations which have passed each quality-control test undergo a final sensor ventilation test (section 4.4). The final quality-controlled dataset consist of observations that pass the sensor ventilation test.

4.1.1 Test implementation

To implement this test, we use the Met Office integrated data archive system (MIDAS) daily temperature data [Met Office, 2006]. This dataset contains observations of the maximum and minimum temperatures over a specified time window (usually 12 to 24 hours) for various locations in the UK. MIDAS temperature data have a precision of $0.1°$C and an uncertainty of $0.2°$C.

To create our climatology dataset we use surface stations active during 2018. For each station, we obtain the minimum and maximum dry bulb temperature for February, March, and April using pre-2018 MIDAS data. We note that the climatology of each surface station will depend on when it was made operational and so the climatology length of each station will vary. Due to this, we remove the MIDAS stations with site IDs 62083 and 62119 from our April climatology dataset as they became active in 2016 and 2017 respectively. We also remove the MIDAS stations with site ID 6313 and 15365 from our April climatology dataset as they have implausibly small maximum temperatures despite becoming active in 1914 and 1988 respectively. While it is possible there exist other similarly problematic MIDAS stations, we have not found any further evidence to support removing any other stations from our climatology.

The CRT is performed by checking if the vehicle dry bulb temperature observations are within a predetermined tolerance of the minimum to maximum range for the nearest surface station. The nearest surface station is calculated through the use of the great circle distance (1). (The vehicle GPS measurement will be addressed in section 4.3).

For this test we use a 2°C tolerance to compensate for a number of factors. For example, dry bulb temperature would be expected to change with elevation in the surface layer [Stull, 1988]. Additionally, dry bulb temperature measurements taken on surfaces with higher albedo, such as grass, can produce noticeably different measurements from those taken on surfaces with lower albedo, such as asphalt (e.g. Huwald et al. [2009]). The purpose of the tolerance used in this test is to account for elevation and surface differences between a vehicle and its nearest MIDAS surface station. This tolerance will also partially compensate for extreme events that occurred during the trial such as the snow and low temperatures that occurred during late February and early March 2018 and a hot spell that occurred from 18th-22nd April 2018 [Met Office, 2020]. Operational MIDAS surface stations will likely be shielded from radiation by Stevenson screens but radiation errors may still occur for calm and/or sunny conditions due to poor ventilation [Harrison, 2015]. Vehicle dry bulb temperature sensors are not shielded from radiation and will be affected by re-radiated heat from road surfaces [Donegan, 2017].

The CRT applied to a single vehicle on the 25th of March 2018 is shown in figure 6. Observations taken during different segments of longer journeys are likely to be nearer to different MIDAS surface stations. This is seen by the jumps in the climatological maximum and minimum lines. Each dry bulb temperature observation plotted here would pass the CRT as each lies between the climatological minimum and maximum air temperature of the nearest MIDAS surface station.

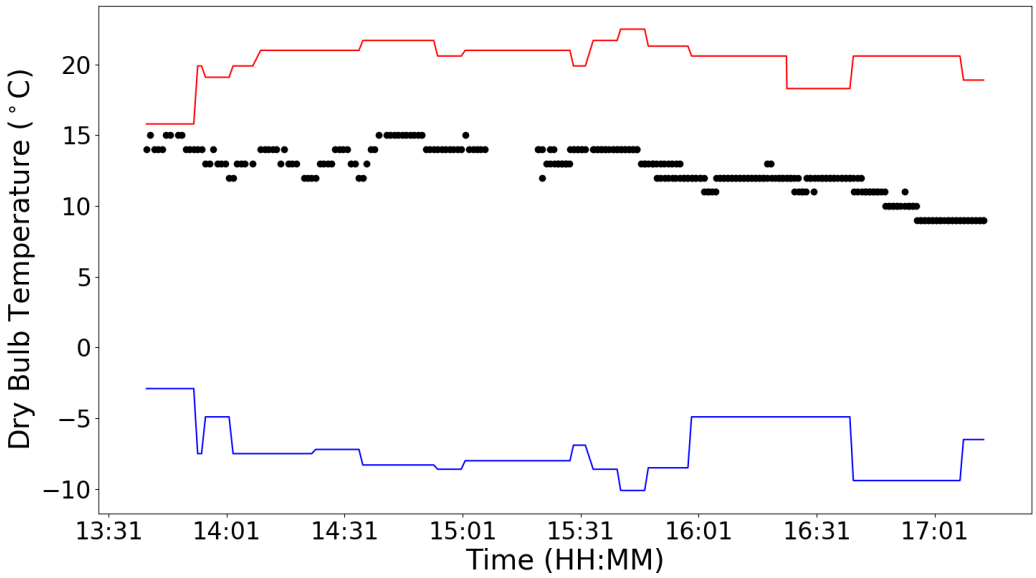

Figure 6: The dry bulb temperature observations (black) from one vehicle on the 25th of March 2018 and the climatological minimum (blue) and maximum (red) temperatures of the nearest MIDAS surface station. Any fluctuation in the minimum and maximum lines can be attributed to the tested observation being nearest to a surface station different from the previously tested observation. The two surface stations used to test the observations around 16.30 have the same minimum but different maxima. Each dry bulb temperature observation shown passes the CRT as each lies between the minimum and maximum temperature of the nearest MIDAS surface station.

4.1.2 Algorithm

The CRT algorithm is shown in algorithm 1. Here, $N_{filtered}$ is the total number of observations in the filtered dataset, y_i is the i-th observation in the filtered dataset with corresponding dry bulb temperature T_i, θ_{max} and θ_{min} are the maximum and minimum climatological air temperature observations of the nearest MIDAS surface station and tol is the tolerance.

Algorithm 1: Climatological range test pseudocode

for $i = 1$ ... $N_{filtered}$ **do**
  Determine corresponding MIDAS climatology to use (i.e. February, March or April)
  Find MIDAS station nearest to y_i to obtain θ_{max} and θ_{min}
  if $\theta_{min} - tol \leq T_i \leq \theta_{max} + tol$ **then**
    Pass y_i
  else
    Flag y_i
  end if
end for

4.1.3 CRT results

Using the settings described in section 4.1.1 we find that 32129 observations have been passed (over 99%) by the CRT and 50 observations have been flagged. Since few observations were flagged by this test we conclude that a 2°C tolerance is suitable for this dataset but larger tolerances may be more suitable for other vehicle-based observation datasets.

4.2 Stuck instrument test (SIT)

Persistence tests are a common quality-control test to determine whether an instrument is stuck and/or if the variability of the measurements of some meteorological field is physically realistic (e.g. Zahumenskỳ [2004], Drobot et al. [2011]). Standard persistence tests will be unsuitable for the filtered dataset due to a large number of short journeys undertaken by participants during the trial. Furthermore, insufficient knowledge on the variability of dry bulb temperature on the scales measured by vehicles, as well as whether this variability would be adequately reflected in low precision measurements, add to the infeasibility of persistence tests. We therefore implement a simplified form of a persistence test which only checks whether an instrument is stuck on some value. This test will be referred to as the stuck instrument test (SIT).

4.2.1 Test implementation

To implement the SIT for an observation y_{test} valid at time t_{test} we first create a sample of observations y_{sample} valid at times t_{sample} from the same vehicle such that

$$t_{test} - 15 \text{ mins} \leq t_{sample} \leq t_{test} + 15 \text{ mins}. \tag{3}$$

The accepted sample time window is chosen to be symmetric in t_{test} so that observations at the start and end of a journey can be tested. We note that a large time-window is chosen for our sample to compensate for the low precision of the dry bulb temperature observations. Provided at least one observation in the test sample has a different dry bulb temperature value to y_{test} then y_{test} is passed.

4.2.2 Algorithm

The SIT algorithm is shown in algorithm 2. Here, $N_{filtered}$ corresponds to the number of observations in the filtered dataset, y_{test} is the test observation taken at time t_{test} with dry bulb temperature T_{test} and sensor ID SID_{test} and y_i is the i-th observation taken at time t_i with dry bulb temperature T_i and sensor ID SID_i.

Algorithm 2: Stuck instrument test pseudocode

for i ... $N_{filtered}$ **do**
  Set $y_{test} = y_i$
  for i ... $N_{filtered}$ **do**
    if $y_{test} \neq y_i, t_i \in [t_{test} - 15 \text{ mins}, t_{test} - 15 \text{ mins}]$ and $SID_{test} = SID_i$ **then**
      Add T_i to sample S
    end if
  end for
  if the sample is empty **then**
    Cannot test y_{test}
  else if at least one dry bulb temperature doesn't equal T_{test} **then**
    Pass y_{test}
  else
    Flag y_{test}
  end if
end for

4.2.3 Results

There are 30124 observations passed by the SIT, 2008 observations flagged, and 47 observations which could not be tested as there were no other observations from the same vehicle in the sample window.

4.3 GPS test

Accurate vehicle navigation requires frequent location polling and smartphone GPS based positioning has been shown to be acceptable for vehicle tracking [Menard et al., 2011]. However, the accuracy of GPS measurements is known to be heavily affected by the smartphone and application used [Hess et al., 2012, Bauer, 2013]. The GPS test verifies whether the GPS metadata of individual observations provide physically plausible vehicle locations.

4.3.1 Test implementation

To quality check the GPS measurements associated with the vehicle data we examine the location of a test observation against the location of a reference observation. The reference observation will have been taken up to 30 minutes before the test observation and by the same vehicle. We note that the first observation taken by a vehicle and any observations with a time-gap larger than 30 minutes from the previous observation cannot be tested by this method as there will be no suitable reference observation to test against.

The GPS test first calculates the distance between the tested observation and reference observation, d_{test}, through the great circle distance (1). Next, using the metadata of the two observations, we calculate estimates of the minimum and maximum distances the vehicle could have travelled between the times the reference and test observation were taken. The maximum distance will be estimated by (2) where v_1 (v_2) is replaced by the speed of the test (reference) observation v_{test} (v_{ref}) and Δt is the time-gap between the test and reference observations. Similarly, the minimum distance will be estimated by

$$d^e_{min} = min\left(v_{test}, v_{ref}\right) \times \Delta t. \tag{4}$$

The estimates provided by (2) and (4) may not be reflective of the true distance traversed by the vehicle due to speed fluctuations and the route travelled between the observations. When estimating the maximum distance we must account for the vehicle having a larger speed than v_{test} and v_{ref} between the times the observations were taken. When estimating the minimum distance we must account for more factors than with the maximum distance. In addition to the vehicle having a lower speed than v_{test} and v_{ref} between the observation times, the journey may occur during heavy traffic congestion, the route travelled may not be a straight line, or the route traversed may be through a residential area. Additionally, MROs must be accounted for as it is possible their GPS will not have been updated because of the minimum time and distance update conditions on the app (see section 2).

The tested observation will pass the GPS test if $\Gamma_{min} \times d^e_{min} \leq d_{test} \leq \Gamma_{max} \times d^e_{max}$ where $\Gamma_{min} < 1$ and $\Gamma_{max} > 1$ are tolerance constants used to account for the uncertainty in d^e_{min} and d^e_{max} respectively. When an observation is passed by the GPS test, if $\Delta t \geq 1$ minute and $d_{test} > 0$km between this passed

observation and the reference observation it was tested against it becomes the reference observation for the next test observation. Otherwise, the reference observation is unchanged for the next test observation. This is to avoid GPS-lagged observations being reference observations as they are known to have inaccurate GPS (see section 3.2.3). If the next test observation is over 30 minutes from the reference observation, the test observation is unable to be tested and we set it to be the next reference observation.

For our implementation we set the tolerance constants as $\Gamma_{max} = 1.3$ and $\Gamma_{min} = 0.6$. To show the suitability of our choices for Γ_{max} and Γ_{min} in the GPS test we calculate the distances between observations and their respective reference observations denoted d_{test}. Figure 7 shows each d_{test} plotted against their corresponding d^e_{max} (black dots) and lines $d_{test} = d^e_{max}$ (red dashed line) and $d_{test} = 1.3 \times d^e_{max}$ (red solid line). The gradients of the two lines represent possible choices for Γ_{max}. By using $\Gamma_{max} = 1.3$, all points above the solid line will be flagged by the GPS test, since for these data, the vehicle appears to have travelled further than physically plausible since the previous reference observation. Points below the solid line are not flagged but must also pass a minimum distance test. Figure 8 shows each d_{test} plotted against their corresponding d^e_{min} (black dots) and lines $d_{test} = d^e_{min}$ (blue dashed line) and $d_{test} = 0.6 \times d^e_{min}$ (blue solid line). The gradients of the two lines represent possible choices for Γ_{min}. Similarly to figure 7, by using $\Gamma_{min} = 0.6$, most points below the solid line will be flagged by the GPS test, since for these data, the vehicle appears to have travelled less than physically plausible since the previous reference observation. Points above the solid line are not flagged but must also pass a maximum distance test. Additionally, we will set $\Gamma_{min} = 0$ when $v_{ref}, v_{test} < 25$km/h or $\Delta t < 1$ minute as we expect the test observation to be relatively near to the reference observation. (The specific choice of 25km/h is related to the sensor ventilation test which will discussed in section 4.4). Therefore, many observations beneath the solid line in figure 8 will not be flagged by the GPS test. We note that the horizontal threshold $d_{test} \approx 0.5$km in figures 7 and 8 is caused by the 500 metre default GPS update distance used by the app.

4.3.2 Algorithm

The GPS test algorithm is shown in algorithm 3. Here, N_{IDs} corresponds to the number of unique sensor IDs in the filtered dataset which determine if observations come from the same source (i.e. vehicle), N_{obs} is the number of observations for the i-th sensor ID and y_j is the j-th observation such

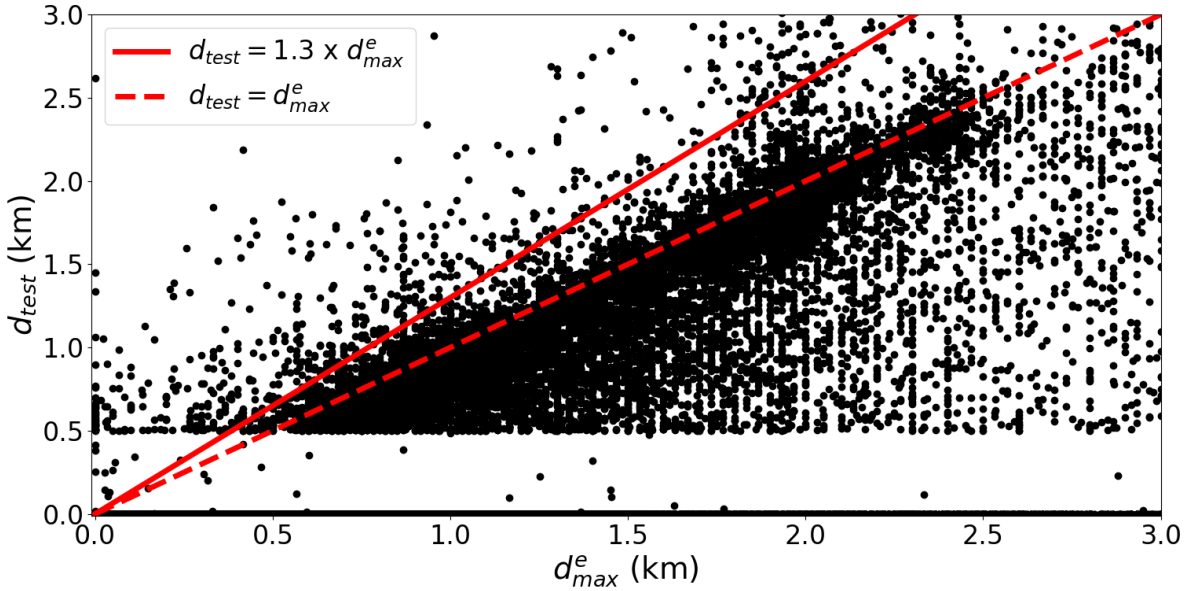

Figure 7: Scatter plot to show the distances between observations and their respective reference observation, denoted d_{test}, against their corresponding d^e_{max} (black dots) and the lines $d_{test} = d^e_{max}$ (red dashed line) and $d_{test} = 1.3 \times d^e_{max}$ (red solid line). The gradients of the two lines represent possible choices for Γ_{max}. We note that the threshold $d_{test} \approx 0.5$km is due to the 500 metre default GPS update distance of the app used in this trial. By using $\Gamma_{max} = 1.3$, all points above the solid line will be flagged by the GPS test, since for these data, the vehicle appears to have travelled further than physically plausible since the previous reference observation.

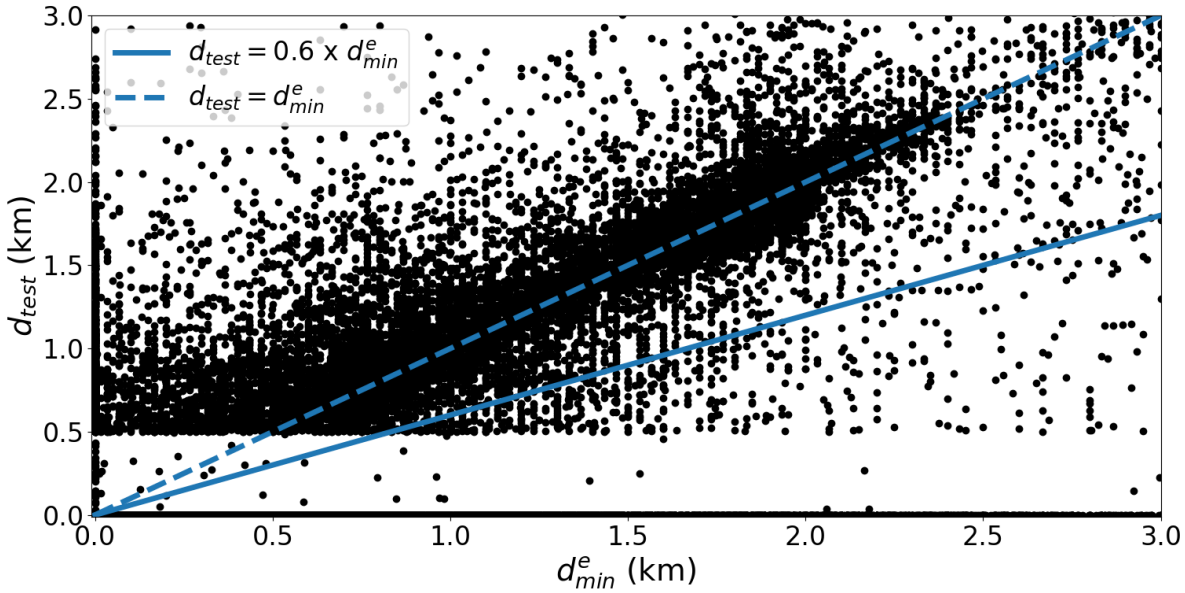

Figure 8: Scatter plot to show the distances between observations and their respective reference observation, denoted d_{test}, against their corresponding d^e_{min} (black dots) and the lines $d_{test} = d^e_{min}$ (blue dashed line) and $d_{test} = 0.6 \times d^e_{min}$ (blue solid line). The gradients of the two lines represent possible choices for Γ_{min}. We note that the threshold $d_{test} \approx 0.5$km is due to the 500 metre default GPS update distance of the app used in this trial. By using $\Gamma_{min} = 0.6$, all points below the solid line will be flagged by the GPS test, since for these data, the vehicle appears to have travelled less than physically plausible since the previous reference observation.

that y_j occurs before y_{j+1}.

Algorithm 3: GPS test pseudocode

for i ... N_{IDs} **do**
  Collect all observations y belonging to the i-th sensor ID into a test dataset
  Set $y_{ref} = y_1$
  for $j = 2$ ... N_{obs} **do**
    Calculate the time-gap Δt between y_{ref} and y_j
    if $\Delta t \geq 30$ mins **then**
      Set $y_{ref} = y_j$
    else
      if $\Delta t < 1$ minute or $v_{test}, v_{ref} < 25$km/h **then**
        $\Gamma_{min} = 0$
      else
        $\Gamma_{min} = 0.6$
      end if
      Calculate d^e_{min}, d^e_{max} and d_{test}
      if $\Gamma_{min} \times d^e_{min} \leq d_{test} \leq \Gamma_{max} \times d^e_{max}$ **then**
        Pass y_j
        if $\Delta t > 1$ minute and $d_{test} > 0$km **then**
          Set $y_{ref} = y_j$
        end if
      else
        Flag y_j
      end if
    end if
  end for
end for

4.3.3 Results

There are 20162 observations that pass the GPS test of which 9939 are MROs. There are 11181 flagged observations and 836 observations which could not be tested. The results of the GPS test are summarised in table 2. The majority of the observations that couldn't be tested were due to large time-gaps between the test and reference observations.

For the flagged dataset, 1331 (1429) observations were flagged for having $d_{test} > d^e_{max}$ ($d_{test} < d^e_{min}$). We note that it is possible many of these observations may have accurate location metadata but have inaccurate speed and time metadata resulting in the disagreement between d_{test} and d^e_{max} or d^e_{min}. The remaining 8421 flagged observations were GPS-lagged observations. The GPS-lagged observations

are primarily due to the 500 metre default GPS update distance of the app but also because of poor GPS signal resulting in more time being needed for a location update.

Determining the uncertainty of the GPS for this dataset would require examination of the GPS uncertainty caused by the smartphone app and each smartphone make used by the participants during the trial. Merry and Bettinger [2019] found an average horizontal position accuracy in urban areas of 7-13m for an iPhone 6. For convection-permitting NWP this would likely be an acceptable GPS uncertainty for vehicle-based observations.

A greater concern is the minimum GPS update distance used by the app (500m) which resulted in a substantial portion of the filtered dataset being flagged by the GPS test. For all crowdsourced observations accurate spatial location metadata are needed due to the scales of the atmospheric processes being observed. This is especially true for vehicles as their locations are non-stationary.

4.4 Sensor ventilation test (SVT)

The final quality-control test we apply in this report is the sensor ventilation test (SVT). As discussed in section 1, in order to produce realistic temperature measurements it is necessary for sensors to be adequately ventilated. For vehicle-based observations, sensor ventilation will be determined by the speed the vehicle is travelling at. From our examination of the dry bulb temperatures of the filtered dataset in section 3.2.1, we define the sensor ventilation threshold for the filtered dataset as $v_{sensor} = 25$km/h.

The SVT is used to check for adequate temperature sensor ventilation by removing any observations with speed less than the sensor ventilation threshold v_{sensor}. This test is implemented last as each observation can be tested individually and observations flagged by this test are still useful for the SIT or GPS test. We also note that $\Gamma_{min} = 0$ was used in the GPS test for observations with speed less than v_{sensor} as those observations will be flagged by the SVT regardless of the GPS test result.

	Number of tested observations	Number of passed observations	Number of flagged observations	Number of untested observations
Climatological range test	32179	32129	50	0
Stuck instrument test	32179	30124	2008	47
GPS test	32179	20162	11181	836
Sensor ventilation test	19094	17425	1669	0

Table 2: Summary table containing the results from all quality checking tests.

4.4.1 Algorithm

The algorithm for the SVT is shown in algorithm 4. Here, N is the number of observations that have passed all previous quality-control tests and y_i is the i-th observation with corresponding speed u_i.

Algorithm 4: Sensor ventilation test pseudocode

```
for i = 1 to N do
    if u_i ≥ 25 then
        Pass y_i
    else
        Flag y_i
    end if
end for
```

4.4.2 Results

The SVT is applied to the 19094 observations that have passed all other quality-control tests. In total, the SVT flags 1669 observations (8.7%). We note that a large number of observations relative to the number tested are flagged by this test due to the large number of observations that occurred when the vehicle was stationary (i.e. 0km/h speed).

5 The quality-controlled dataset

The dataset resulting from the quality-control process described in section 4 consists of 17425 observations (25.6% of the complete dataset). A summary of the quality-control test results is given in table 2. The number of observations that passed all tests, were flagged by at least one test, and could not be tested by all tests for each day is shown in figure 9.

A summary of the descriptive statistics for the quality-controlled dataset and each month is given in table 4. In total, 2833 observations were taken in February, 9012 in March, and 5580 in April. The ratio between the number of observations that occurred in each month in the quality-controlled dataset is relatively unchanged from that of the filtered dataset. The distributions of the dry bulb temperatures for each

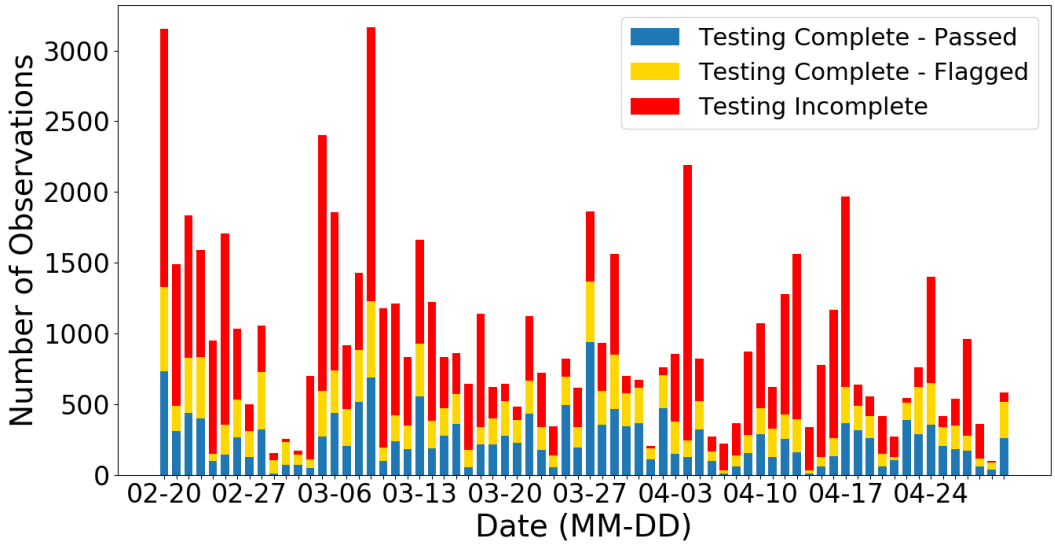

Figure 9: The observations that passed all tests (blue bar segments totalling 17425), were flagged by at least one test (yellow bar segments totalling 13897) and could not be tested by all tests (red bar segments totalling 36637). We note that the red bar segments include all observations removed in the initial filtering stage and those untested by any QC test.

month, shown in figure 10, retain the majority of the characteristics of their filtered dataset counterparts discussed in section 3.2.1. There has been a slight reduction in skew and kurtosis for February and April as higher temperatures have been removed by the SVT. Additionally, extreme temperature anomalies have been removed from April by the CRT.

	February	March	April	All data
Number of observations	2833	9012	5580	17425
Mean	3.52°C	6.67°C	11.29°C	7.63°C
Standard deviation	4.04°C	3.30°C	4.26°C	4.64°C
Skew	−0.22	−0.38	0.72	0.26
Excess kurtosis	−0.88	0.13	0.92	1.07

Table 3: Summary of the descriptive statistics for the quality-controlled dataset and split into each month.

	February	March	April	All data
Number of observations	2833	9012	5580	17425
Mean	3.52°C	6.67°C	11.29°C	7.63°C
Standard deviation	4.04°C	3.30°C	4.26°C	4.64°C
Skew	−0.22	−0.38	0.72	0.26
Excess kurtosis	−0.88	0.13	0.92	1.07

Table 4: Summary of the descriptive statistics for the quality-controlled dataset and split into each month.

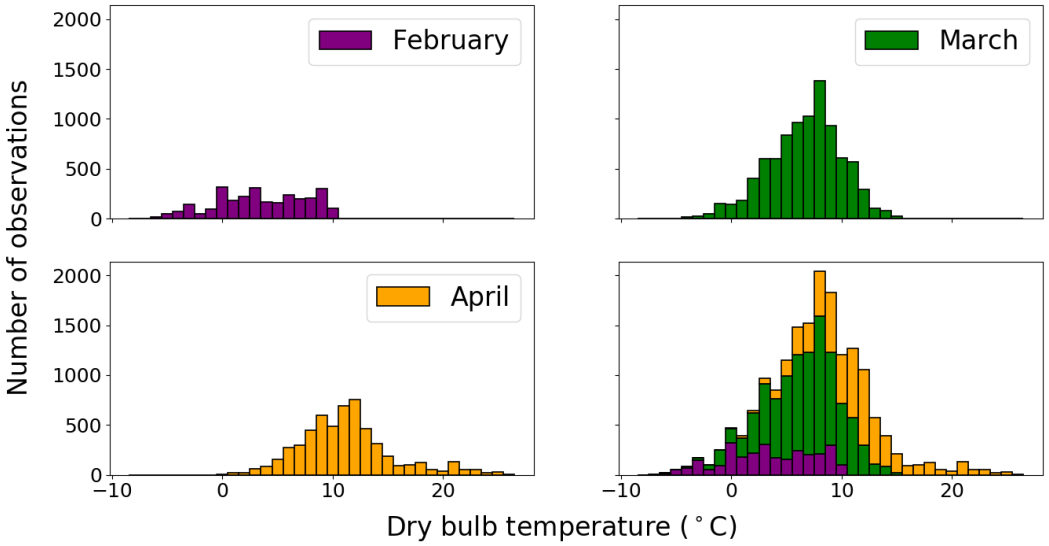

Figure 10: Distribution of dry bulb temperature observations for each month of the trial for the quality-controlled dataset. The purple bar segments indicate the number of February 2018 observations, the green segments March 2018, and the orange segments April 2018. The combined distribution is a stacked histogram that shows the contribution from each month to the total number of observations for each dry bulb temperature.

6 Summary and recommendations

The use of crowdsourced observations in numerical weather prediction is a new research area that is quickly receiving much attention from the meteorology community. Indeed, the high spatio-temporal resolution of the observations is particularly attractive for convection-permitting data assimilation where expansion and management of conventional scientific surface observing networks are too costly.

This report details the quality-control process applied to a novel low-precision vehicle-based observation dataset. In order to quality-control the raw dataset, we first needed to filter the dataset in order to determine which data had the necessary information. The climatological range test (CRT), stuck instrument test (SIT), and GPS test were applied in parallel to the filtered dataset. The sensor ventilation test (SVT) was applied to data that had passed the three previous tests. We note that as the CRT can be applied to each observation individually it may be more suitable to apply this test after the SIT or GPS test. The quality-control dataset consists of the 17425 observations (25.6% of original dataset) which have passed all quality-control tests. This is in stark contrast to the quality-control of scientific surface observing networks where approximately 10% of dry bulb temperature observations are flagged or discarded. Recommendations for future quality-control and data collection are given at the end of this section.

The CRT was used as a range validity test on the dry bulb temperatures. This type of quality-control test has been successfully used for smartphone observations [Hintz et al., 2019b] and vehicle-based observations [Chapman et al., 2010, Limber et al., 2010, Boyce et al., 2017]. To implement this test

we used monthly climatologies of the MIDAS surface stations active during the trial. Hence, this test is limited by the climatology of each MIDAS station. The CRT flagged the lowest number of observations out of any quality-control test and would be suitable for operational quality-control of vehicle-based observations.

The SIT was used as a simplified persistence test to determine if the sensor was stuck on some value. In order to implement this test, a vehicle identifier was required to determine if observations came from the same source. For operational persistence tests, it will be important to consider how observations taken from short finite journeys can be tested for persistence. We note that, despite the simplicity of this test, there is still a small amount of data in the filtered dataset unable to be tested. We also note that, because of the low precision of the data, it is also possible that valid observations have been flagged by this test. However, as the majority of observations able to be tested were passed, it is unlikely that our sample time-span was an issue for the implementation of this test on the filtered dataset.

The GPS test was performed to verify the plausibility of the GPS metadata. GPS accuracy has been an issue for crowdsourced smartphone observations where it is important to know the elevation of pressure observations [Madaus and Mass, 2017, Hintz et al., 2019b] but inaccuracies in the horizontal have not been a reported concern. Similarly to the SIT, vehicle identification is required to implement the GPS test. The GPS test involves comparing the distance between GPS coordinates of two observations with minimum and maximum estimates calculated through their metadata. As such, the results of this test are also dependent on the accuracy of the speed and time metadata of the observations. We also note that the starting observation of a journey or after a large time-gap was used as a reference observation in this test despite their GPS being untested and possibly inaccurate. The number of flagged observations is substantially greater than the number of observations flagged by the CRT, SIT and SVT combined. This is primarily due to the existence of GPS-lagged observations in the filtered dataset which are caused by poor GPS signal, insufficient distance travelled, or time between observations to trigger a GPS location update. As vehicles are able to traverse a greater distance in a short time-span than some current operational weather prediction model grid spacing it is important that lags in GPS be accounted for.

The SVT is the final quality-control test and is another filtering test where observations with speed less than $v_{sensor} = 25$km/h are flagged. A similar precaution was used by Knight et al. [2010] who found vehicle-based temperature observations to be reasonably accurate provided the vehicle was moving for a number of minutes prior to the observation time. Despite being applied to fewer observations than any other test, a large number of observations were flagged by the SVT. It is expected that a large amount of vehicle-based observations will be flagged due to traffic congestion or driving through residential areas.

Unfortunately, we were unable to perform a spatial consistency test on the filtered dataset due to a lack of observations occurring at similar locations and times. Additionally, due to the high spatial resolution of this dataset, we were unable to obtain an independent dataset which most observations from the filtered

dataset could be tested against. It is possible in the future that vehicle-based observations form vast observation networks that are densest in urban areas making spatial consistency tests more feasible. Additionally, the examination of any biases present in the observations was also infeasible due to a lack of data. Vehicle-based observations are likely to be biased due to the sensing instrument and heat from the car engine and road surface.

Recommendations for future data collection and quality-control of vehicle-based observations are as follows.

- A substantial amount of data was found without valid speed metadata. Correcting the OBD dongle or app settings/features causing this will result in fewer observations discarded prior to the quality-control process.
- Reduce the GPS update distance to reduce the number of GPS-lagged observations.
- The climatology datasets used in the CRT each contained around 400 MIDAS stations situated within the UK making this test a computationally expensive matching procedure. In operational settings, it may be more appropriate to use reduced datasets such as monthly regional climatologies. Additionally, the comparison of vehicle-based observations with WOW surface station data may provide another suitable quality-control test.
- The temperature observations obtained from the Met Office trial are all low-precision measurements. Because of this, a simplified persistence test was used as a quality-control test. If precision was increased to 0.1°C then more standard persistence tests can be used as a quality-control test. This will allow for testing of whether the sensor is stuck and whether the variability of the observed fields is physically plausible simultaneously.
- Both the SIT and GPS test could not have been implemented without a sensor ID. Due to data privacy, it may be unfeasible to have sensor IDs with potential vehicle-based observation sources. Using appropriate encryption techniques on vehicle sensor IDs may allow for the use of vehicle time series in quality-control (e.g. Verheul et al. [2019]). Alternatively, if a phone app is used in the data collection process it may be possible to quality-control locally on a smartphone before uploading to WOW servers. Such methods have been used for the quality-control and bias correction of smartphone observations [Hintz et al., 2019b].
- To check if a vehicle sensor is stuck without vehicle observation time series it may be suitable to record the amount of time since the vehicle observed a different value as metadata for that observation.
- To check GPS accuracy without vehicle location time series it would be useful to record GPS signal strength between a satellite and a phone, as well as how many satellites are involved in the GPS polling, as metadata for an observation. Having access to this information will also assist in determining the GPS measurement uncertainty. Additionally, provided enough observations are

available, spatial consistency tests may be capable of flagging some observations with incorrect GPS.

The Met Office proof-of-concept trial has shown it is possible to obtain vehicle-based observations from in-built vehicle sensors using smartphones and OBD dongles. However, the observations obtained through this trial leave much to be desired. While the quality-control procedure presented in this report may be a suitable reference point for other crowdsourced datasets, there is much improvement to be made in both data collection and quality-control for such observations to be utilisable. It is therefore necessary to conduct further trials possibly with alternative data collection methods which address the issues raised in this report.

Acknowledgements

Z. Bell was supported by a University of Reading PhD studentship. S. L. Dance and J. A. Waller were supported in part by UK EPSRC grant EP/P002331/1 (DARE). We would also like to thank Malcolm Kitchen and Ellie Creed for their assistance on this project.

References

A. R. Anderson, M. Chapman, S. D. Drobot, A. Tadesse, B. Lambi, G. Wiener, and P. Pisano. Quality of mobile air temperature and atmospheric pressure observations from the 2010 development test environment experiment. *Journal of Applied Meteorology and Climatology*, 51(4):691–701, 2012.

C. Bauer. On the (in-) accuracy of GPS measures of smartphones: a study of running tracking applications. In *Proceedings of International Conference on Advances in Mobile Computing & Multimedia*, page 335. ACM, 2013.

S. Bell, D. Cornford, and L. Bastin. How good are citizen weather stations? addressing a biased opinion. *Weather*, 70(3):75–84, 2015.

B. Boyce, G. Weiner, A. Anderson, and S. Linden. Pikalert® System Vehicle Data Translator (VDT) Utilizing Integrated Mobile Observations: Pikalert VDT Enhancements, Operations, & Maintenance. Technical Report No. FHWA-JPO-512, Booz Allen Hamilton, National Center for Atmospheric Research, 2017.

L. Chapman, C. Bell, and S. Bell. Can the crowdsourcing data paradigm take atmospheric science to a new level? a case study of the urban heat island of london quantified using netatmo weather stations. *International Journal of Climatology*, 37(9):3597–3605, 2017.

M. Chapman, S. Drobot, T. Jensen, C. Johansen, W. Mahoney III, P. Pisano, and B. McKeever. Diagnosing road weather conditions with vehicle probe data: Results from detroit intellidrive field study. *Transportation Research Record*, 2169(1):116–127, 2010.

S. L. Dance, S. P. Ballard, R. N. Bannister, P. Clark, H. L. Cloke, T. Darlington, D. L. Flack, S. L. Gray, L. Hawkness-Smith, N. Husnoo, et al. Improvements in forecasting intense rainfall: Results from the franc (forecasting rainfall exploiting new data assimilation techniques and novel observations of convection) project. *Atmosphere*, 10(3):125, 2019.

S. Devarakonda, P. Sevusu, H. Liu, R. Liu, L. Iftode, and B. Nath. Real-time air quality monitoring through mobile sensing in metropolitan areas. In *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*, pages 1–8, 2013.

B. Donegan. Why your car thermometer is wrong. https://weather.com/science/weather-explainers/news/car-thermometer-thermistor-temperature-wrong, 2017. Accessed: 27-01-2020.

S. Drobot, M. Chapman, E. Schuler, G. Wiener, I. William Mahoney, P. Pisano, and B. McKeever. Improving road weather hazard products with vehicle probe data: Vehicle data translator quality-checking procedures. *Transportation Research Record*, 2169(1):128–140, 2010. doi: 10.3141/2169-14. URL https://doi.org/10.3141/2169-14.

S. Drobot, M. Chapman, B. Lambi, G. Wiener, A. Anderson, et al. The vehicle data translator v3. 0 system description. Technical report, United States. Joint Program Office for Intelligent Transportation Systems, 2011.

A. Droste, J.-J. Pape, A. Overeem, H. Leijnse, G.-J. Steeneveld, A. Van Delden, and R. Uijlenhoet. Crowdsourcing urban air temperatures through smartphone battery temperatures in são paulo, brazil. *Journal of Atmospheric and Oceanic Technology*, 34(9):1853–1866, 2017.

C. A. Fiebrich, C. R. Morgan, A. G. McCombs, P. K. Hall Jr, and R. A. McPherson. Quality assurance procedures for mesoscale meteorological data. *Journal of Atmospheric and Oceanic Technology*, 27 (10):1565–1582, 2010.

N. Gustafsson, T. Janjić, C. Schraff, D. Leuenberger, M. Weissmann, H. Reich, P. Brousseau, T. Montmerle, E. Wattrelot, A. Bučánek, et al. Survey of data assimilation methods for convective-scale numerical weather prediction at operational centres. *Quarterly Journal of the Royal Meteorological Society*, 144(713):1218–1256, 2018.

U. Haberlandt and M. Sester. Areal rainfall estimation using moving cars as rain gauges-a modelling study. *Hydrology and Earth System Sciences 14 (2010), Nr. 7*, 14(7):1139–1151, 2010.

R. G. Harrison. *Meteorological measurements and instrumentation.* , chapter 5, pages 90–95. John Wiley & Sons, 2015.

B. Hess, A. Z. Farahani, F. Tschirschnitz, and F. von Reischach. Evaluation of fine-granular gps tracking on smartphones. In *Proceedings of the First ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems*, pages 33–40. ACM, 2012.

K. S. Hintz, K. O'Boyle, S. L. Dance, S. Al-Ali, I. Ansper, D. Blaauboer, M. Clark, A. Cress, M. Dahoui, R. Darcy, et al. Collecting and utilising crowdsourced data for numerical weather prediction: Propositions from the meeting held in copenhagen, 4–december 5, 2018. *Atmospheric Science Letters*, page e921, 2019a.

K. S. Hintz, H. Vedel, and E. Kaas. Collecting and processing of barometric data from smartphones for potential use in numerical weather prediction data assimilation. *Meteorological Applications*, 2019b.

H. Huwald, C. W. Higgins, M.-O. Boldi, E. Bou-Zeid, M. Lehning, and M. B. Parlange. Albedo effect on radiative errors in air temperature measurements. *Water Resources Research*, 45(8), 2009.

S. Knight, C. Smith, and M. Roberts. Mapping manchester's urban heat island. *Weather*, 65(7):188–193, 2010.

M. Limber, S. Drobot, T. Fowler, et al. Clarus quality checking algorithm documentation report. Technical Report No. FHWA-JPO-11-075, Research and Innovative Technology Administration, 2010.

L. E. Madaus and C. F. Mass. Evaluating smartphone pressure observations for mesoscale analyses and forecasts. *Weather and Forecasting*, 32(2):511–531, 2017.

C. McNicholas and C. F. Mass. Impacts of assimilating smartphone pressure observations on forecast skill during two case studies in the pacific northwest. *Weather and Forecasting*, 33(5):1375–1396, 2018a.

C. McNicholas and C. F. Mass. Smartphone pressure collection and bias correction using machine learning. *Journal of Atmospheric and Oceanic Technology*, 35(3):523–540, 2018b.

F. Meier, D. Fenner, T. Grassmann, M. Otto, and D. Scherer. Crowdsourcing air temperature from citizen weather stations for urban climate research. *Urban Climate*, 19:170–191, 2017.

T. Menard, J. Miller, M. Nowak, and D. Norris. Comparing the GPS capabilities of the Samsung Galaxy S, Motorola Droid X, and the Apple iPhone for vehicle tracking using FreeSim_Mobile. In *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 985–990. IEEE, 2011.

K. Merry and P. Bettinger. Smartphone GPS accuracy study in an urban environment. *Public Library of Science*, 14(7), 2019.

Met Office. MIDAS: UK daily temperature data. Data retieved from online CEDA archive https://catalogue.ceda.ac.uk/uuid/220a65615218d5c9cc9e4785a3234bd0, NCAS British Atmospheric Data Centre (Accessed on 16/06/2019), 2006.

Met Office. Weather observations website. https://wow.metoffice.gov.uk/, 2011.

Met Office. Past weather events. https://www.metoffice.gov.uk/weather/learn-about/past-uk-weather-events#y2018, 2020.

C. Muller, L. Chapman, S. Johnston, C. Kidd, S. Illingworth, G. Foody, A. Overeem, and R. Leigh. Crowdsourcing for climate and atmospheric sciences: current status and future potential. *International Journal of Climatology*, 35(11):3185–3203, 2015.

R. Nakamura and L. Mahrt. Air temperature measurement errors in naturally ventilated radiation shields. *Journal of Atmospheric and Oceanic Technology*, 22(7):1046–1058, 2005.

T. N. Nipen, I. A. Seierstad, C. Lussana, J. Kristiansen, and Ø. Hov. Adopting citizen observations in operational weather prediction. *Bulletin of the American Meteorological Society*, 2019.

A. Overeem, J. R. Robinson, H. Leijnse, G.-J. Steeneveld, B. P. Horn, and R. Uijlenhoet. Crowdsourcing urban air temperatures from smartphone battery temperatures. *Geophysical Research Letters*, 40 (15):4081–4085, 2013.

E. Rabiei, U. Haberlandt, M. Sester, and D. Fitzner. Rainfall estimation using moving cars as rain gauges-laboratory experiments. *Hydrology and Earth System Sciences 17 (2013), Nr. 11*, 17(11): 4701–4712, 2013.

E. C. Rada, M. Ragazzi, M. Brini, L. Marmo, P. Zambelli, M. Chelodi, and M. Ciolli. Perspectives of low-cost sensors adoption for air quality monitoring. In *Air Quality*, pages 29–40. Apple Academic Press, 2016.

S. J. Richardson, F. V. Brock, S. R. Semmer, and C. Jirak. Minimizing errors associated with multiplate radiation shields. *Journal of Atmospheric and Oceanic Technology*, 16(11):1862–1872, 1999.

A. R. Siems-Anderson, C. L. Walker, G. Wiener, W. P. Mahoney III, and S. E. Haupt. An adaptive big data weather system for surface transportation. *Transportation research interdisciplinary perspectives*, 3: 100071, 2019.

M. Snelder and S. Calvert. Quantifying the impact of adverse weather conditions on road network performance. *European Journal of Transport and Infrastructure Research*, 16(1), 2016.

G.-J. Steeneveld, S. Koopmans, B. Heusinkveld, L. Van Hove, and A. Holtslag. Quantifying urban heat island effects and human comfort for cities of variable size and urban morphology in the netherlands. *Journal of Geophysical Research: Atmospheres*, 116(D20), 2011.

R. B. Stull. *An introduction to boundary layer meteorology.* , chapter 1, pages 9–19. Springer Science & Business Media, 1988.

J. Sun, M. Xue, J. W. Wilson, I. Zawadzki, S. P. Ballard, J. Onvlee-Hooimeyer, P. Joe, D. M. Barker, P.-W. Li, B. Golding, et al. Use of NWP for nowcasting convective precipitation: Recent progress and challenges. *Bulletin of the American Meteorological Society*, 95(3):409–426, 2014.

UK Government. The highway code, road safety and vehicle rules. https://www.gov.uk/speed-limits, 2015.

E. Verheul, C. Hicks, and F. D. Garcia. Ifal: Issue first activate later certificates for v2x. In *2019 IEEE European Symposium on Security and Privacy (EuroS P)*, pages 279–293, 2019.

D. Wolters and T. Brandsma. Estimating the urban heat island in residential areas in the netherlands using observations by weather amateurs. *Journal of Applied Meteorology and Climatology*, 51(4): 711–721, 2012.

World Meteorological Organization. *Guide to meteorological instruments and methods of observation*. Secretariat of the World Meteorological Organization, Geneva, Switzerland, 7 edition, 2008.

I. Zahumenskỳ. Guidelines on quality control procedures for data from automatic weather stations. *World Meteorological Organization, Switzerland*, 2004.

Met Office
FitzRoy Road
Exeter
Devon
EX1 3PB
United Kingdom

Produced with the Met Office LaTeX template v3.0.0