Restless climate: Lessons using data skills

Royal Geographical Society with IBG

Advancing geography and geographical learning

Lesson 2: Checking the quality of meteorological data

Lesson objectives

- To check the quality of daily maximum air temperature data that have been *automatically* collected at sites in the River Dove catchment, UK
- To investigate the occurrence of observer bias in the *manual* collection of daily precipitation data at a weather station in Dushanbe, Tajikistan

Setting the scene

Many things can (and do) go wrong between the measurement and use of meteorological data (Figure 1) – even for routinely collected variables such as air temperature and rainfall. For temperature, instruments may not be properly sited, shielded from direct sunlight, protected from artificial heat sources, or checked at the same time(s) each day. Instrumentation can change in time too. There is also the possibility of changes in the vicinity of the weather station (such as urban development or encroachment of vegetation), the site might be relocated, or temporarily closed. Even after weather data have been transmitted to the archive, changes in units of measurement, inconsistent use of missing data codes and rounding errors can accrue.

These are important considerations for students to learn about in relation to the use of quantitative data as a secondary source and also if they are collecting – and checking – first hand data that they have collected themselves. The latter might be particularly relevant for students who are undertaking an individual investigation during their A Level studies. See www.rgs.org/nea for further details.

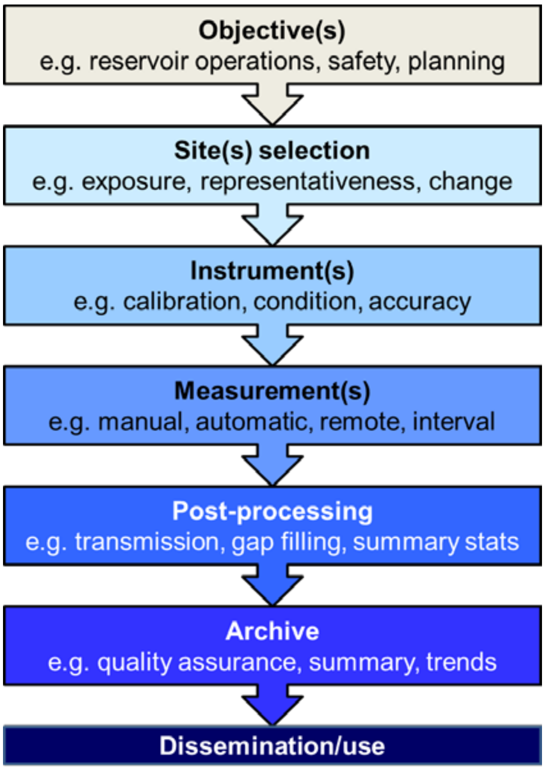


Figure 1. An information-flow that begins by setting project objectives and ends with meteorological data archiving, dissemination and use. Data biases and errors can enter the information-flow at any point in between. Source: Wilby et al. (2017)

For precipitation, the rain gauge may be poorly sited (either over- or under-exposed to wind), measured amounts may be depleted by evaporation losses or under-catch of snowfall, the observer may make errors reading the graduated collector or when transcribing values, observations may be taken at different times, missed or, even worse, fabricated.

Some errors can be 'trapped' by maintaining good meta-data and instrument logs. These document major changes at the site, or in equipment or measurement protocols. Other biases are harder to discern without careful inspection of the data. For example, rainfall observers may favour 'round' numbers ending in 5 or 10. Such anomalies may seem trivial, but they can affect design standards for extreme events (such as very heavy rainfall amounts) and hence the safety of dams, flood defences and drainage systems. This lesson demonstrates how a few simple statistical tests and comparisons can expose suspect values in meteorological data sets.

The data

Meteorological data are collected for various reasons. High quality weather information underpins national forecasting services provided to the public, aviation, shipping, agriculture and many other sectors. Data may also be gathered over decades to monitor long-term environmental change or over just a few seasons to years as part of intensive research programmes. The World Meteorological Organisation (WMO) provides detailed guidelines on best practices, beginning with how to choose a site for a meteorological station, as well as protocols for site maintenance and instrument use. Even so, standards vary enormously, and there is always a need for a 'critical eye' when working with meteorological records – even those that you may have collected yourself.

This lesson uses publicly available daily meteorological data from two sources:

(a) Daily maximum air temperature data for selected sites in the Loughborough University TEmperature Network (LUTEN). These data were collected as part of a research project looking at the long-term relationship between air and water temperature in the River Dove catchment. The project is helping to develop guidance on bankside vegetation management to protect freshwaters from climate change.: http://www.luten.org.uk/home

(b) Daily precipitation data for a site in Dushanbe, Tajikistan provided by the US National Climate Data Center (NCDC) Global Summary of the Day (GSOD): https://www7.ncdc.noaa.gov/CDO/cdoselect.cmd?datasetabbv=GSOD&resolution=40

Tasks

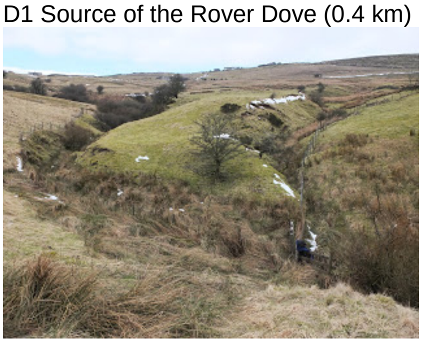
1. **What is the evidence of suspect air temperature measurements in the Dove?**
Download the Microsoft Excel file 'L2_Data_ LUTEN_Tmax.xlsx', the accompanying 'Datasheet 2a'. These daily maximum air temperatures (Tmax) were AUTOMATICALLY collected using thermistors (temperature probes) mounted on the north-facing side of trees at fixed locations along the River Dove (Figure 2).

Look at the air temperature summary statistics in Table 1. How are air temperatures expected to change with altitude? Which site(s) do not fit the expected pattern?

Table 1. Altitude (m) and air temperature for six sites in the River Dove.

Temperature statistics*	**Site code (see Figure 2)**					
* based on days with data at all sites	**D1** **394 m**	**D6** **348 m**	**D10** **244 m**	**D16** **214 m**	**D20** **180 m**	**D22** **163 m**
Mean (°C)	16.4	13.4	14.2	16.3	14.6	14.6
Maximum (°C)	33.3	24.8	26.5	29.8	25.7	26.0
Minimum (°C)	3.0	2.0	2.1	2.6	2.8	2.7
Standard deviation (°C)	6.2	4.5	4.7	5.8	4.5	4.5

Figure 2. Selected temperature monitoring sites in the River Dover, UK.
Values in brackets are the distance of the site from source of the river. Photos: Matt Johnson

D1 Source of the Rover Dove (0.4 km)


D6 Upstream of Stannery (5.0 km)


D10 Crowdecote (9.6 km)


D16 Hartington (19.0 km)


D20 Fishpond plantation (25.8 km)


D22 Milldale (27.8 km)


Click on 'Raw data' to view the daily records. Spend a few minutes scanning the data. Note any suspect values, recording in what way(s) these values are doubtful and how they might have arisen. Blank cells denote missing data (due to thermistor loss of malfunction).

Tip: Nearest neighbour comparison is routinely used to cross-check data collected at different locations as well as to infill missing values. Look for unexpectedly large differences between values on the same day at neighbouring sites. Begin by concentrating on data in the period May to July. Excel scatterplots for pairs of sites can also reveal large outliers.

2. **What site(s) should be prioritised for improved instrument siting/shielding?**
   Which of the six LUTEN sites selected (Figure 3) have the most suspect values? What checks and steps could be taken by the field technicians to improve the quality of future air temperature measurements along the banks of the River Dove?

   Take it further: Write down some 'rules' for automatically checking air temperature data collected at multiple sites. What should be done with suspect values *when* they are found?

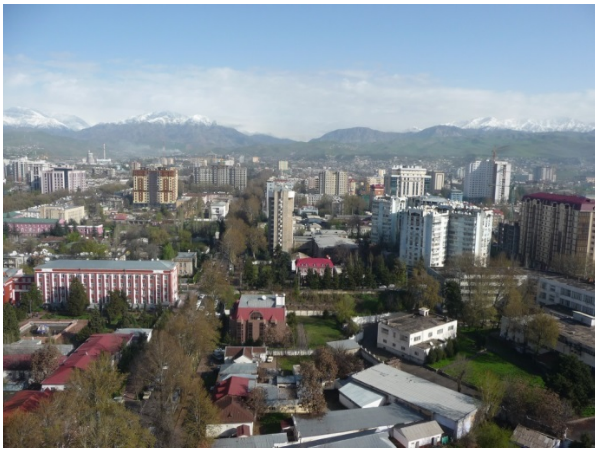


Figure 3. Dushanbe the capital city of Tajikistan. Photo: Rob Wilby

3. **What are the data quality concerns for the precipitation record in Dushanbe?**
   Download the Microsoft Excel file 'L2_Data_GSOD_Dushanbe.xlsx', the accompanying 'Datasheet 2b'. These data were MANUALLY collected. After reviewing the READ ME content, click on the 'Raw data' tab. There are 23 columns but daily precipitation amounts (Column T) and quality flags (Column U) are highlighted in yellow. Spend a few minutes browsing the record dates (Column C) and precipitation data. List any quality concerns.

4. **What is the evidence of observer measurement bias in the precipitation data?**
   Click on the 'Measurement bias' tab then compare the observed and expected frequencies for the different daily precipitation totals. The first 20 rows of data are shown in Table 2. What is the pattern in the data? How plausible are the very largest values in the file?

Table 2. Frequencies of observed and expected wet-day rainfall amounts in Dushanbe. Expected frequencies were estimated using a gamma distribution for daily precipitation.

Precipitation (inches)	Observed frequency	Expected frequency
0.01	187	319
0.02	234	193
0.03	43	144
0.04	351	116
0.05	24	98
0.06	16	86
0.07	0	76
0.08	315	69
0.09	17	63
0.1	10	58
0.11	3	54
0.12	198	50
0.13	8	47
0.14	7	44
0.15	3	42
0.16	178	40
0.17	12	38
0.18	10	36
0.19	1	34
0.2	134	33

Tip: Use the data in Table 2 to plot a bar chart to compare observed and expected frequencies in millimetre categories. [Note that GSOD precipitation amounts are stored as inches whereas the original data were collected in tenths of a millimetre – so apply a factor of 25 to the data in Table 2 to convert them back into millimetres].

Take it further: What statistical test could be used to evaluate the significance of the differences between frequencies of observed and expected wet-day rainfall amounts?

5. **What is the evidence of record errors amongst the precipitation values?**
Inspect the summary statistics for mean wet-day amount, dry-day frequency, mean temperature and visibility shown in Table 3. Use these data to plot bar charts of the values by day of week. What unexpected patterns occur?

Table 3. Mean daily precipitation, dry-days, temperature and visibility in Dushanbe.

Day	Mean precipitation (inches)	Dry days (number in whole period)	Mean temperature (Fahrenheit)	Mean visibility (miles)
Sunday	0.063	2167	59.9	1.3
Monday	0.074	2117	59.5	1.2
Tuesday	0.075	2129	59.8	1.2
Wednesday	0.065	2092	59.7	1.2
Thursday	0.062	2157	59.8	1.3
Friday	0.070	2107	59.9	1.2
Saturday	0.066	2113	59.7	1.3

Tip: More dry days and lighter rainfall on Sundays has also been documented for meteorological stations in Australia, the UK and U.S. Why is this?

Plenary

Return to the main lesson question: what are the methods by which instrument and observer errors can be detected in meteorological records?

Ask the students to list different methods of data checking and to rate each in terms of ease and effectiveness. Return to the key stages in an information flow (Figure 1). Overall, what are the most important sources of meteorological error to avoid?

Discuss the wider ramifications of errors in weather data for engineering design and public safety.

Further reading

For an expose of the widespread observer biases in the US Cooperative Observer Program weather station network, see: Daly C, Gibson WP, Taylor GH, Doggett MK, Smith JI. 2007. Observer bias in daily precipitation measurements at United States Cooperative Network Stations. *Bulletin of the American Meteorological Society*, **88**:899-912. http://journals.ametsoc.org/doi/abs/10.1175/BAMS-88-6-899

For a critique of a much wider suite of biases and errors in hydrological data, see: Wilby RL, Clifford NJ, De Luca P, Harrigan SO, Hillier JK, Hodgkins R, Johnson MF, Matthews TKR, Murphy C, Noone SJ, Parry S, Prudhomme C, Rice SP, Slater LJ, Smith KA, Wood PJ. 2017. The “dirty dozen” of freshwater science: Detecting then reconciling hydrological data biases and errors. *WIREs Water*, in press.