Dataset Guide

Data Rationale

Schools Weather and Air Quality (SWAQ) is a citizen science project funded by the Department of Industry, Innovation and Science as part of its Inspiring Australia - Citizen Engagement Program. SWAQ is equipping public schools across Sydney, Australia with research-grade meteorology and air quality sensors, enabling students to collect and analyse research quality data through curriculum-aligned classroom activities. The network includes twelve automatic weather stations (Vaisala WXT536) and seven automatic air quality stations (Vaisala AQT420). SWAQ data provides urban canopy layer observations of the intra-urban heterogeneity and inter-parameter dependency of all major urban climate and air quality variables, valuable across diverse urban disciplines. Quality control procedures are designed to ensure observations of extreme episodes are not excluded. Beyond research purposes, SWAQ is a citizen-centered network, conceived to promote valuable STEM (science, technology, engineering, mathematics) skills among citizens and students

Geolocalization

SWAQ stations are located where there are gaps in existing government networks, and focus on Sydney's western suburbs, where the highest urbanization rate is taking place. The network stretches from -33.5995° to -34.0424° latitude and from 150.6918° to 151.2706° longitude. The average spacing is 10.2 km. Optimum site allocation was determined by undertaking a multi-criteria weighted overlay analysis to ensure data representativeness and quality. All SWAQ sensors are installed:

- in homogenous urban regions, without sections of anomalous variation in the regional urban makeup and aspect-ratio, and without large, concentrated heat/pollution sources or sinks;
- in areas falling into the WMO Class 4 with no electromagnetic sources;
- at a constant height of 2 - 3.5 m above ground level.

Station locations, codes, geographical coordinates, and corresponding monitoring stations are tabulated below.

Table 1.

Location	Code	Latitude	Longitude	Monitoring Stations
OEH Supersite Chullora	OEHS	-33.8915297	151.0460133	WXT536 + AQT420

UNSW Campus	UNSW	-33.916105	151.232912	WXT536 + AQT420
Brookvale Public School	BROO	-33.7610924	151.2706266	WXT536 + AQT420
Glenorie Public School	GLEN	-33.5994777	151.0069077	WXT536 + AQT420
Kurnell Public School	KURN	-34.0099519	151.2046102	WXT536 + AQT420
Leppington Public School	LEPP	-33.9593216	150.8106364	WXT536 + AQT420
Luddenham Public School	LUDD	-33.8814147	150.6918441	WXT536 + AQT420
Dulwich Hill Public School	DULW	-33.905453	151.1399031	WXT536
Kellyville Public School	KELL	-33.7109338	150.9578669	WXT536
Narellan Public School	NARE	-34.042402	150.734044	WXT536
Taren Point Public School	TARE	-34.0188092	151.1230662	WXT536
Newtown Public School	NEWT	-33.8998603	151.179179	WXT536

Time window

Observations and metadata are available from September 2019 for WXT536 + AQT420 stations and from October 2019 for WXT536 stations (refer to Table 1), thus encompassing the Black Summer bushfire and the COVID-19 lockdown period. Data is sampled at 20 minute intervals.

Measured Variables

Six meteorological parameters (dry-bulb temperature, relative humidity, barometric pressure, rain, wind speed, and wind direction) and six air pollutants (SO_2, NO_2, CO, O_3, $PM_{2.5}$, and PM_{10}) are recorded. Units, symbols and accuracies are included in the following table (n.s.=not specified).

Table 2.

Symbol	Parameter	Units	Accuracy
T	Air temperature (dry bulb)	°C	± 0.3
RH	Relative Humidity	%	±3 at 0-90 % ±5 at 90-100 %
p	Air Pressure (mean above sea level)	hPa	±0.5 at 0-30 °C

			±1 at -52-0, 30-60 °C
ws	Wind speed	m/s	±3 % at 10 m/s
wd	Wind direction	°	±3.0 at 10 m/s
Rain	Rainfall rate	mm/h	n.s.
SO_2	Sulphur dioxide (air concentration)	ppm	±0.05
CO	Carbon monoxide (air concentration)	ppm	±0.2
NO_2	Nitrogen dioxide (air concentration)	ppm	±0.025
O_3	Ozone (air concentration)	ppm	±0.06
PM_{10}	Mass fraction of particulate matter less than 10 µm in diameter	$\mu g/m^3$	n.s.
$PM_{2.5}$	Mass fraction of particulate matter less than 2.5 µm in diameter	$\mu g/m^3$	n.s.

Dataset Description and Quality Control

All data files are in a Comma Separated Value format (CSV) and contain measurements from all stations, concatenated horizontally. The columns' headers take the general form of "STAT_Variable", where "STAT" is the four-character station code in capital letters as displayed in Table 1 and "Variable" indicates the measured parameter (see "Symbol" in Table 2). An ISO-8601-compliant date-time index is provided in the first column, under the heading "Time".

Two versions of the same dataset are provided, both quality controlled.

The first file is called "YYYY-MM-DD_Raw", where YYYY-MM-DD represents the date of the last recorded value in year-month-day format. This dataset contains the raw data, followed by an equal number of columns, headed "STAT_Variable_Flags". These columns contain the semicolon-separated list of Quality Control Flags per each variable and station. The flag represents the result of quality control algorithms. The SWAQ dataset is checked against 6 different filters, summarized in Table 3.

Table 3.

1: Continuity Test
Definition: identification of temporal misalignments and gaps Parameters: all
2: Fixed Range Test
Definition: test data against physical and instrumental limits Parameters: all
3: Internal Consistency Test
Definition: Identify consecutive identical readings over 3 hours Parameters: ws, wd: (ws=0 → wd=0)
4: Persistence Test
Definition: Compare variables against each other based on known relationships

	Parameters: all but Rain
5: Dynamic Range Test	
	Definition: for each station and parameter, verify if the measured value is an outlier with respect to the monthly dataset of all stations that passed the previous steps, for any month having >90 % level of completeness (Outlier definition: <p25-1.5IQR or >p75+1.5IQR) Parameters: all but Rain, RH, wd
6: Step Test	
	Definition: for each station and parameter, verify if the step from previous measurement (absolute values) is an outlier with respect to the steps of all stations across the same month, for any month having >90 % level of completeness (Outlier definition: <p25-3IQR or >p75+3IQR) Parameters: all but Rain, wd

Two flagging systems are proposed (see Table 4). A Single Test Flag (STF) is first applied, following the sequence and the coding in Table 3. Then, a Combinational Flag (CF) is applied by combining different STFs. According to the combinatorial system, only data points that fail both the step and the dynamic range tests are flagged (and potentially removed), as these represent isolated perturbations and sensor spikes. The CF system is more conservative and is thus recommended for studies focused on weather and pollution extremes.

Table 4.

STF FlagCode	**Description**	**CF FlagCode**	**Description**
STF0	Good	CF0	Good
STF1.1	Suspect (dynamic range test)	CF1	Suspect (step test AND dynamic range test)
STF1.2	Suspect (step test)		
STF2.1	Erroneous (fixed range test)	CF2	Erroneous (fixed range test OR internal consistency test)
STF2.2	Erroneous (internal consistency test)		
STF3	Failure (persistence test)	CF3	Failure (persistence test)
STF4.1	Missing	CF4	Missing (missing or insufficient data)
STF4.2	Warning: insufficient data for dynamic range or step test		

Alongside, comes a second csv file called “YYYY-MM-DD_Cleaned”. This is a ready-to-use dataset, quality controlled as recommended by SWAQ’s technicians. The SWAQ QC procedure aims at trading off accuracy and data preservation. It is described in Table 5.

Table 5.

1: Replace all negative pollutant values with zero
2: Replace RH and wd values slightly crossing the physical boundaries with the boundaries themselves
3: Remove all data points failing the instrumental fixed range test
4: Remove all data flagged as CFx, with x>1

Metadata

Individual and detailed metadata files (pdf) are included for every site, showing the list of stations and components, the dominant land use and land cover, average height and type/materials of trees and buildings in a 500m-radius neighbourhood, notes on the presence of water/heat and pollutants sources/sinks, traffic intensity, changes in the local landscape, sky view factor. The percentage of built features, vegetated cover and water features is detailed within three different radii of the station (20 km, 500 m, and 50 m). Graphical representations include: satellite images, street-view maps, cardinal direction photographs, panoramic photos, and close-up photos of sensors, solar panels and connections. The files are named "Metadata_Code.pdf", where Code is the station name as in Table 1.