Stochastic Environmental Research and Risk Assessment (2023) 37:2473–2493
https://doi.org/10.1007/s00477-023-02401-8

ORIGINAL PAPER

An automatic quality evaluation procedure for third-party daily rainfall observations and its application over Australia

Ming Li[1] · Quanxi Shao[1] · Joel Janek Dabrowski[2] · Ashfaqur Rahman[3] · Andrea Powell[1] · Brent Henderson[4] · Zachary Hussain[5] · Peter Steinle[5]

Accepted: 7 February 2023 / Published online: 24 February 2023

Abstract

Third-party rainfall observations could provide an improvement of the current official observation network for rainfall monitoring. Although third-party weather stations can provide large quantities of near-real-time rainfall observations at fine temporal and spatial resolutions, the quality of these data is susceptible due to variations in quality control applied and there is a need to provide greater confidence in them. In this study, we develop an automatic quality evaluation procedure for daily rainfall observations collected from third-party stations in near real time. Australian Gridded Climate Data (AGCD) and radar Rainfields data have been identified as two reliable data sources that can be used for assessing third-party observations in Australia. To achieve better model interpretability and scalability, these reference data sources are used to provide separate tests rather than a complex single test on a third-party data point. Based on the assumption that the error of a data source follows a Gaussian distribution after a log-sinh transformation, each test issues a *p*-value-based confidence score as a measure of quality and the confidence of the third-party data observation. The maximum of confidence scores from individual tests is used to merge these tests into a single result which provides overall assessment. We validate our method with synthetic datasets based on high-quality rainfall observations from 100 Bureau of Meteorology (BoM) of Australia stations across Australia and apply it to evaluate real third-party rainfall observations owned by the Department of Primary Industries and regional development (DPIRD) of Western Australia. Our method works well with the synthetic datasets and can detect 76.7% erroneous data while keeping the false alarm rate as low as 1.7%. We also discuss the possibility of using other reference datasets, such as numerical weather prediction data and satellite rainfall data.

Keywords Rainfall observations · Data quality · Third-party stations · Radar rainfields · Statistical method

1 Introduction

Rainfall is a primary component in the water cycle and it is arguably the most important climate variable that directly affects human society. Accurate and reliable rainfall observations are vital in many applications, such as water resources planning and management (Herman et al. 2020; Neupane and Guo 2019), hydrological forecasting (Li et al. 2013, 2016) and the study of climate trends and variability (Marengo et al. 2018; Neupane and Guo 2019). Furthermore, many weather-sensitive industries (such as mining, energy and agriculture) rely on accurate rainfall data for effective operation. For example, high-quality rainfall observations are required for index-based or parametric insurance policies in developing economies to reduce

✉ Ming Li
Ming.Li@data61.csiro.au

[1] CSIRO Data61, PO Box 1130, Bentley, WA 6102, Australia

[2] CSIRO Data61, GPO Box 2583, Brisbane, QLD 4001, Australia

[3] CSIRO Data61, Private Bag 12, Hobart, TAS 7001, Australia

[4] CSIRO Data61, GPO Box 1700, Canberra, ACT 2601, Australia

[5] Bureau of Meteorology, GPO Box 1289, Melbourne, VIC 3001, Australia

farmers' risk and increase average incomes (Clement et al. 2018; Greatrex et al. 2015).

The most accurate rainfall observations are obtained from official automatic weather stations (AWS) managed by government agencies. The installation and maintenance of these official AWS have followed the international standard set by, for example, the World Meteorological Organization (World Meteorological Organization, 2018). A quality control procedure is regularly applied to official AWS to identify and correct suspect rainfall data from human errors or equipment faults. However, official AWS are typically restricted to limited spatial coverage because the station locations are chosen based on a range of requirements and constraints. In Australia, the Bureau of Meteorology (BoM) managed $\sim$ 730 official AWS equipped with Tipping Bucket Rain Gauges (TBRG), though the number of official AWS differs from year to year and it can be a different number of observations available within a different year. Rain in Australia is well known to be highly variable (Murphy and Timbal 2008), and the current coverage of official AWS is inadequate to represent large natural rainfall variability in space.

Third-party rainfall stations could provide a potential solution to improve the current observation network for rainfall monitoring (Assumpcao et al. 2018; Buytaert et al. 2014; Muller et al. 2015; Zheng et al. 2018). The development of inexpensive sensors and communication technology has made AWS more affordable for the public and enabled connections with sensors in remote locations. Therefore, many third parties have installed their own weather stations with off-the-shelf equipment and obtained local observations with the advent of wireless technologies. The existing official rainfall observation network (which are usually limited in coverage) has been rapidly supplemented with third-party rainfall stations as well as remote sensing rainfall estimates from satellite and radar (which require validation from ground observations) (Bardossy et al. 2021). To the best of our knowledge, $\sim$ 8000 third-party AWS collect real-time rainfall observations in Australia. Some third-party AWS are owned by the general public in local communities and shared by online services such as Netatmo and Weather Underground, which enable near-real-time collection, integration and visualization of rainfall data. These third-party AWS are also known as private weather stations (Chen et al. 2021a, b) or citizen weather stations (Napoly et al. 2018) in the literature. Other third-party AWS are installed and serviced by local government agencies or organisations, which are referred to as non-private third-party weather stations. For example, the Department of Primary Industries and regional development (DPIRD) of Western Australia (WA) owns a network of $\sim$ 200 non-private third-party AWS mainly located in southwest WA to offer real-time local weather data for regional communities.

Although third-party AWS can provide enormous quantities of near-real-time data at specific locations, the quality of these data are more susceptible than official weather data because of (a) incorrect installations, (b) sensor failure or malfunctions, and (c) inadequate servicing leading to poor-quality data (Bell et al. 2015; Campbell et al. 2013; Chen et al. 2018). For example, "backyard" weather stations are often found to be installed too close to building walls or trees that can partially intercept rainfall and cause lower readings, and electronic sensors of AWS may malfunction or fail completely both by environmental phenomena (such as flooding, fire, lightning strikes and animal activity) and by malicious human activity (such as theft, vandalism and tampering). It is often not known if routine services and scheduled maintenance are performed to minimise external influences (such as a rain gauge being blocked by debris) on reading accuracy. Streaming weather data are often posted online with limited or no quality control and delivered in a raw form without any checks or evaluations. Therefore, end users have limited confidence to incorporate raw third-party rainfall observations in their decision-making.

To make the best use of the growing volume of third-party rainfall observations, there is an increasing need for automated and algorithm-based quality control (QC) methods to evaluate the data quality (Campbell et al. 2013; Muller et al. 2015; Zheng et al. 2018). Automated QC methods can improve the confidence of end users and enable prompt decision-making. Unlike quality assurance (QA), which is a proactive process to collect metadata and perform data maintenance, QC aims to improve the quality of data and provide confidence in the data. Traditionally, manual or semi-automated QC methods have been developed by meteorologists to detect errors in the observations for official weather stations. Traditional QC methods can be labour intensive, such as manually checking consistency with nearby sites, which may be challenging for a large volume of data in near real-time. If traditional QC methods are semi-automated, resource constraints limit the ability to apply them to third-party stations. For example, performing an intercomparison of redundant measurements taken at the same site is a common practice for some key official stations to maintain data quality but is not applicable for most third-party stations.

More recently, several research studies have developed automated QC methods applicable for rainfall observations collected from third-party AWS. de Vos et al. (2019) proposed a real-time applicable QC method for third-party rainfall measurements. Their method required no auxiliary data and consisted of four modules that identify and filter typical errors by checking spatial consistency. Chen et al.

(2021a, b) adopted the use of reputation systems to assign trust scores to crowdsourced third-party stations. Bardossy et al. (2021) proposed a two-fold approach to filter out suspicious rainfall measurements from third-party stations by checking whether they appear consistent with the spatial pattern of official weather stations. Several automated QC methods have been developed for third-party observations of other climate variables, such as air temperature (Beele et al. 2022; Chakraborty et al. 2020; Fenner et al. 2021; Meier et al. 2017; Napoly et al. 2018) and wind (Chen et al. 2021a, b; Droste et al. 2020) Almost all existing methods in the literature only considered the spatial consistency between third-party stations and primary gauges (e.g. de Vos et al. (2019) and Chen et al. (2021a, b)) or within third-party stations in a dense network (e.g. Bardossy et al. (2021). This limitation restricts the use of such methods in urban areas rather than regional areas.

Data quality is not an absolute concept and different applications may have different requirements for data quality. We cannot assume that observations, good enough for one application, are good in general. Given that the underlying true value is never known, any quality control can suffer from false positive error (i.e., good observations labelled as bad) and/or false negative error (i.e., bad observations labelled as good). End users may be willing to tolerate more on one type of error than the other. For example, this research is primarily motivated by facilitating parameter insurance with assessing third-party weather observations collected by farmers. In this application, parametric insurance companies typically consider minimal false positives as their first priority to avoid unnecessary lawsuits against them from farmers. As another example, climate scientists would prefer to include the most accurate observations in their research by controlling false negatives at first. Traditional quality control procedures often label each test observation with a fixed quality flag (e.g., wrong, ok or suspect) and consequently lead to fixed false positive and false negative rates, regardless of applications and user requirements. To overcome the limitation of traditional quality control, we report data quality by a continuous confidence score instead of a quality flag and allow users to decide whether to exclude or include in their specific application based on their risk preference. As this research is primarily motivated by facilitating parameter insurance with assessing third-party weather observations collected by farmers, our case studies will be evaluated in the context of this application. To satisfy the need for evaluating rainfall observations from all third-party stations on a large country scale with a mixture of urban and regional areas (such as Australia), we propose an automated data-driven quality evaluation procedure that can be implemented in an operational environment. The proposed procedure aims to (a) make the use of the best possible reference data, (b) assign a continuous confidence score to each test observation instead of a categorical quality flag, and (c) provide a parsimonious and interpretable model structure that can be easily extended with new data sources. We assume that the majority of third-party rainfall observations are of good quality and erroneous observations are only present in the form of outliers. Each type of reference data forms a prediction of daily rainfall at a specific location and is compared with the third-party observation at this location. The associated prediction uncertainty is estimated statistically based on distributional assumptions. If the difference between a third-party observation and the corresponding prediction from reference data is greater than the estimated prediction uncertainty, this third-party observation is likely to be wrong. To further confirm the quality of this third-party observation, we would like to compare it against additional sources of reference datasets, which are supposed to be mutually independent with the first reference data. If the differences between the observation and additional reference datasets are also sufficiently large, we have more confidence that the observation is of poor quality. Specifically, our method has two stages: (1) separate statistical tests to quantify the agreement between test observations and each type of reference data, including gridded rainfall analysis data and radar rainfall data (if available), and (2) an overall assessment by combining the results from individual tests and assigning a confidence score in a probabilistic framework.

We organize this manuscript as follows. Section 2 introduces the reference data used in the test methods. Section 3 provides the details of our test method. Section 4 validates the test method based on a synthetic data example and Sect. 5 provides a case study based on real third-party rainfall observations from the DPIRD network. We discuss other possible reference data in Sect. 6. Conclusions and discussion are made in Sect. 7.

2 Data

We have used the following two reference data sources to evaluate the data quality of daily rainfall observations collected from third-party weather stations in Australia.

2.1 AGCD

Australian Gridded Climate Data (AGCD) is the BoM's official dataset for Australian gridded rainfall analysis at daily and monthly time scales (Jones et al. 2009). AGCD applies state-of-the-art statistical modelling to combine available rainfall data and provides an accurate estimate of rainfall conditions in wider areas than rain gauges that provide rainfall point measurements. AGCD has incorporated the BoM's latest quality control and quality assurance to ensure that quality flagged and subsequently removed data do not affect the resulting analysis. AGCD of daily rainfall at 5 × 5 km resolution is available for the period since 1900 and can be requested directly from the BoM or NCI (NCI 2022). The BoM also provides a ten-fold cross-validated analysis to determine the accuracy of AGCD daily rainfall (Evans et al. 2020). The cross-validated analysis errors, measured by root-mean-squared error (RMSE), are available from a BoM OPeNDAP server (Bureau of Meteorology 2022b).

2.2 Radar rainfields data

Rainfields is the BoM current system for quantitative radar rainfall estimation, which performs several basic quality control checks on radar data, converts radar reflectivity to rainfall depths, and produces real-time, spatially and temporally continuous rainfall data (Seed et al. 2007). Rainfields data have been increasingly used for hydrometeorological applications such as flooding forecasting (May et al. 2013). There are a total of 63 rain radars available as of 2022 in Australia (Bureau of Meteorology 2022c). Rainfields data are available within a radius up to 256 km from a radar and at 1 km resolution with 5-min updates. Rainfields data can be requested directly from the BoM Climate Data Online (Bureau of Meteorology 2022a).

3 Methods

Three tests are performed to check the quality of daily rainfall observations from third-party weather stations and to the results from each individual test is combined to provide an overall assessment. In this section, we provide the detail of each individual test and the method to combine them.

3.1 Domain test

A domain test is designed to filter out obviously erroneous observations by checking whether observations are within physical limits. The physical lower limit for rainfall is zero, and an upper limit of 2000 mm per day is set for this study based on historical records. Any rainfall reading out of the physical limits is immediately identified as erroneous and is assigned a confidence level of zero.

3.2 AGCD test

AGCD test checks the agreement between daily rainfall observations from a third-party station and the corresponding AGCD rainfall estimates at the nearest gridded point to the target station. Because the magnitude of the different between station observations and gridded data is typically greater for high rainfall than low rainfall, we apply a data transformation and consider the difference in the transformed space. For a given third-party station, we denote the true underlying rainfall, the rainfall observation from this station and the corresponding AGCD rainfall estimate at time t by $R(t)$, $R_o(t)$ and $R_s(t)$, respectively. AGCD test essentially drives the predictive distribution of $R(t)$ conditional on $R_s(t)$ and calculates the degree of confidence of $R_o(t)$ based on this distribution. The presence of zero rainfall makes the distribution of $R(t)$ a mixture of continuous and discrete distributions. To conveniently deal with zero rainfall, we establish two separate models to derive the predictive distribution, one for $R_s(t) \leq ts$ and the other for $R_s(t) > ts$, where ts is a threshold for zero to small rainfall. In this study, we choose $ts = 2$ mm.

For the sake of simplicity, we denote the probability of true rainfall at time t less than or equal to x (which is the element of the conditional predictive distribution of $R(t)$) conditional on the AGCD rainfall estimate being $R_s(t)$ by $P\{R(t) \leq x | R_s(t)\}$. When $R_s(t) \leq ts$, the predictive distribution of $R(t)$ conditional on $R_s(t)$ is estimated directly from the empirical distribution of $R(k)$ conditional on $R_s(k)$ in a training period $k = 1, \ldots, n$:

$$P\{R(t) \leq x | R_s(t)\} = \frac{\sum_{k=1}^{n} I(R_o(k) \leq x \, and \, R_s(k) \leq ts)}{\sum_{k=1}^{n} I(R_s(k) \leq ts)} \tag{1}$$

where $I(.)$ is the indicator function, which is equal to 1 if the condition is met or to 0 otherwise. In fact, the right-hand side of Eq. (1) estimates the distribution of $R_o(t)$ conditional on $R_s(t)$. Because $R(t)$ is not observable and third-party observations are assumed to be mostly accurate, we approximate the true rainfall $R(t)$ by the corresponding third-party observation $R_o(t)$ and therefore $P\{R(t) \leq x | R_s(t)\}$ by $P\{R_o(t) \leq x | R_s(t)\}$.

When $R_s(t) > ts$, the predictive distribution of $R(t)$ conditional on $R_s(t)$ is derived from an error model which represents the statistical relationship between R_o and R_s. As the uncertainty of AGCD rainfall estimates often increases with higher rainfall, we apply the log-sinh transformation (Wang et al. 2012) to normalise the data, stabilise the variance and establish an error model on the transformed space that is Gaussian. Firstly, we define the following notations at time t:

$f(R) = b^{-1} log\{sinh(a + bR)\}$: the log-sinh transformation,

$Z(t) = f\{R(t)\}$: the transformed true rainfall,

$Z_o(t) = f\{R_o(t)\}$: the transformed third-party rainfall observation,

$Z_s(t) = f\{R_s(t)\}$: the transformed AGCD rainfall estimate,

$f^{-1}(z) = b^{-1}\left\{asinh\left(e^{bz}\right) - a\right\}$: the inverse log-sinh transformation,

where a and b are two log-sinh transformation parameters.

We apply the error model proposed by Li et al. (2016) and assume that the difference between $Z(t)$ and $Z_s(t)$ (i.e., the error of AGCD rainfall estimates in the transformed space) follows a Gaussian distribution:

$$Z(t) = \mu + Z_s(t) + \in (t) \tag{2}$$

where $\in (t) \sim N(0, \sigma^2)$ and μ and σ are the error model parameters. We treat $\in (t)$ as a surrogate of all sources of error in gridded reference data. One typical source of error is representativeness error (Janjic et al. 2018), which is the mismatch between the spatial scales represented by the reference field (e.g., AGCD or the radar Rainfields) and point observations. Because gridded reference data are available on a given grid, they are in general smoother than the true field and principally do not provide information on the scale smaller than the grid point distance. Other sources of error include interpolation error for AGCD and the approximation error of the reflectivity-rainfall relationship for the radar Rainfields. We do not impose any assumption on the spatial pattern of ε(t) in this study because we only perform the proposed quality control procedure for each individual station without using information at other locations. In this case, model parameters are station dependent, and no parameter estimates are shared across different stations. More discussion on the assumption of the spatial pattern of ε(t) will be made on Sect. 7.

Four parameters a, b, μ and σ are assumed to be station dependent and must be estimated prior to using this error model to perform the AGCD test assessment. Because of the presence of zero value, R and R_o follow a mixture of continuous and discrete probability distributions. To deal with the zero-inflated problem (Li et al. 2016) conveniently, we treat R and R_o as left-censored at ts. Based on these assumptions, the maximum likelihood estimation is used to jointly estimate four model parameters by minimising the following objective function calculated in a training period $k = 1, \ldots, n$:

$$\begin{aligned} L(a, b, \mu,\ \sigma) = &-\sum_{k=1}^{n} log\left[\Phi\left\{\frac{Z0 - \mu - Z_s(k)}{\sigma}\right\}\right] \\ &I\{R_o(k) = 0,\ R_s(k) > ts\} - \sum_{k=1}^{n} log\left[\Phi\left\{\frac{Z_o(k) - \mu - Z_s(k)}{\sigma}\right\}\right] \\ &I\{R_o(k) > 0,\ R_s(k) > ts\} + \sum_{k=1}^{n} log[tanh\{a + bR_o(k)\}] \\ &I\{R_o(k) > 0,\ R_s(k)\ > ts\} \end{aligned} \tag{3}$$

where $Z0 = f(0)$, Φ and ϕ are the cumulative distribution function and the density function of a standard normal distribution. Because the true rainfall is unknown and we assume that the third-party rainfall observations are mostly accurate, we use Z_o (or R_o) in place of Z (or R) in the estimation. To simplify notations, we use the same notations for the parameters and their corresponding estimates.

In presence of low-quality R_o and/or R_s (i.e., erroneous third-party observations and/or inaccurate AGCD rainfall estimates), we exclude problematic data from the maximum likelihood estimation to improve the robustness of the estimation. In practice, the observations satisfying any condition below are excluded from the parameter estimation:

- $|R_o(t) - R_s(t)| >\ 5$ mm, when $R_s(t) <\ 10$ mm;
- $|R_o(t) - R_s(t)| > 50\% \max\{R_s(t), R_o(t)\}$, when $R_s \geq$ 10 mm.

The predictive distribution of $R(t)$ conditional on $R_s(t)$ for $R_s(t) > ts$ is given by

$$P\{R(t) \leq x | R_s(t)\} = \Phi\left[\frac{f(x) - \mu - f\{R_s(t)\}}{\sigma}\right]. \tag{4}$$

To evaluate the degree of confidence of third-party observations, we propose a confidence score based on the two-sided p-value from a statistical hypothesis test. Specifically, the following statistical hypothesis test is proposed to decide whether sufficient evidence supports that a third-party observation $R_o(t)$ is of good quality:

$$H_0 : E\{R(t)|R_s(t)\} = R_o(t),$$

$$H_1 : E\{R(t)|R_s(t)\} \neq R_o(t)$$

The two-sided p-value p_2 can be expressed as a function of $R_o(t)$:

$$\begin{aligned} p_2 &= 2\min[P\{R(t) \leq R_o(t)|R_s(t)\}, P\{R(t) > R_o(t)|R_s(t)\}] \\ &= 1 - 2|p_1 - 0.5| \end{aligned} \tag{5}$$

where $p_1 = P\{R(t) \leq R_o(t)|R_s(t)\}$ is the one-sided (left) p-value. The confidence score (CS) of AGCD test based on the two-sided p-value is defined as:

$$CS_{AGCD} = 1 - 2|P\{R(t) \leq R_o(t)|R_s(t)\} - 0.5| \tag{6}$$

where $P\{R(t) \leq R_o(t)|R_s(t)\}$ can be calculated by Eq. (1) or (4) based on $R_s(t) \leq ts$ or $R_s(t) > ts$. To reduce false alarm rates, we do not attempt to detect any erroneous data with less than $2mm$ errors and conveniently force $CS_{AGCD} = 100\%$ if $|R_o(t) - R_s(t)| \leq\ 2$ mm. This practical consideration can satisfy most end-user needs.

The following assumptions and minimum requirements are made to ensure that the AGCD test is optimal:

- The training period should have at least two years (preferably four years) of historical observations to estimate the test model parameters.
- The correlation between historical observations and AGCD rainfall estimates in the training period should be at least 60%.
- The corresponding cross-validated AGCD rainfall RMSE should be less than $50\% R_s$ for $R_s > 10$ mm. This assumption implies that the AGCD test is not applicable if AGCD estimates for medium and high rainfall are associated with large uncertainty.

In this study, we perform independent quality check at different stations and that is why we only check the agreement between the historical time series from a test station and reference data at the corresponding location (as described the second minimum requirements above) instead of the agreement between the observational field from third-party stations and the estimation field from reference data. In the future research, it would be great to evaluate the quality of multiple stations jointly and a sanity check is required to ensure the consistency of the spatial correlation from different data.

3.3 Rainfields test

The Rainfields test is designed to compare third-party rainfall observations with radar Rainfields estimates of daily rainfall accumulation. The algorithm of the Rainfields test is the same as that of the AGCD test except for replacing AGCD rainfall estimates with radar Rainfields estimates. The confidence score of the Rainfields test, denoted by $CS_{Rainfields}$, can be similarly derived from Eq. (6). The assumptions and minimum requirements of the Rainfields test include.

- The training period should have at least two years (preferably four years) of historical observations to estimate the test model parameters.
- The correlation between historical observations and Rainfields rainfall estimates in the training period should be at least 60%.
- The target third-party station should be within 128 km of a rainfall radar.

3.4 Merged test

Both AGCD and radar Rainfields rainfall estimates are inevitably subject to uncertainty, particularly for medium and high rainfall. Combining the results from the AGCD test and Rainfields test minimises possible false alarms from each individual test. We combine the AGCD test and the Rainfields test by using the maximum confidence scores from the AGCD test and the Rainfields test as a combined confidence score. Specifically, the confidence score of the merged test, denoted by CS_{merged}, can be calculated by

$$CS_{merged} = \max\left(CS_{AGCD}, CS_{Rainfields}\right).$$

The overall confidence score from the merged test is greater than or equal to the confidence score from any individual test. Figure 1 illustrates how to merge the AGCD and Rainfields test by a flow chart. As a result, the false alarm rate from the merged test is expected to be smaller than or equal to any individual test. This is the main reason that we define the overall confidence score by taking the maximization rather than average. An adverse effect of this merged test is that the hit rate, indicating the ability to detect genuine erroneous data, is also lower than any separate test.

4 Model validation with synthetic data

We validate the test method described in Sect. 3 on a synthetically generated dataset. This dataset is generated by inserting random errors into rainfall observations for the period between 2016 and 2019 from 100 high-quality BoM stations in Australia (see Appendix A). We thus demonstrate our test performance with synthetic data, where the "true" rainfall value is assumed to be known. The synthetic data represents artificial third-party observations, under the assumption that observations from a third-party station will be resemble those from a BoM station with added random error.

Two synthetic datasets are considered in this study, a zero-rainfall synthetic dataset with random perturbation only at zero rainfall observations and a non-zero-rainfall synthetic dataset with random perturbation only at non-zero rainfall observations. Each synthetic dataset contains about 1% "erroneous" observations, which are raw rainfall observations with inserted random errors. The structure of random perturbation in these two synthetic datasets is given as follows:

- Zero-rainfall synthetic dataset: inserting random errors of 3–5 mm (positive only) at zero rainfall
- Non-zero-rainfall synthetic data:
  - ≤ 10 mm: inserting random errors of 3–5 mm (positive or negative)
  - > 10 mm: inserting random errors of 30–50% of raw observations (positive or negative)

A high performing test method should be able to effectively detect these "erroneous observations" in the

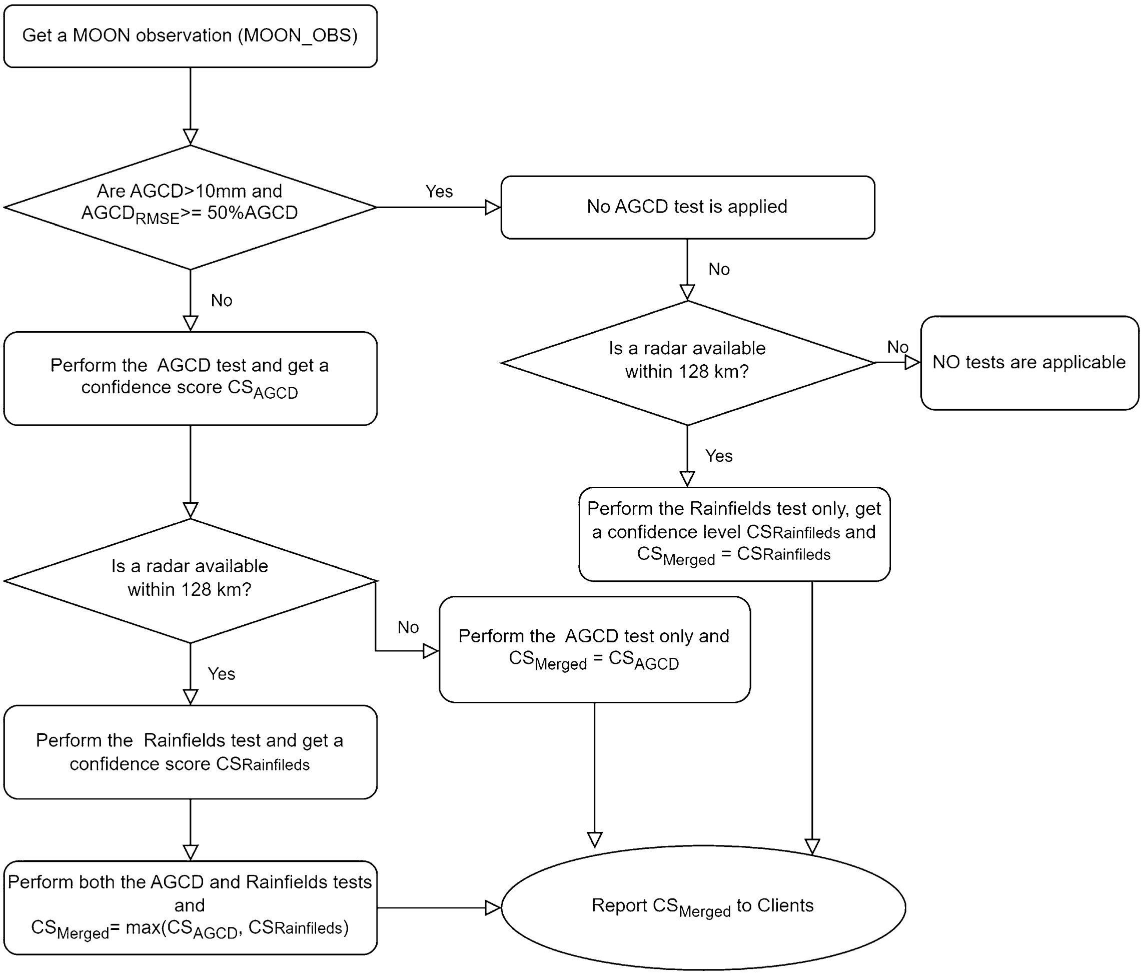

Get a MOON observation (MOON_OBS)
Are AGCD>10mm and AGCD$_{RMSE}$>= 50%AGCD
Yes
No AGCD test is applied
No
No
Perform the AGCD test and get a confidence score CS$_{AGCD}$
Is a radar available within 128 km?
No
NO tests are applicable
Yes
Perform the Rainfields test only, get a confidence level CSRainfileds and CS$_{Merged}$ = CSRainfileds
Is a radar available within 128 km?
No
Perform the AGCD test only and CS$_{Merged}$ = CS$_{AGCD}$
Yes
Perform the Rainfields test and get a confidence score CSRainfileds
Perform both the AGCD and Rainfields tests and CS$_{Merged}$= max(CS$_{AGCD}$, CSRainfileds)
Report CS$_{Merged}$ to Clients

Fig. 1 A flow chart to show how to combine the AGCD and Rainfields tests

synthetic data. Two evaluation metrics for binary classification, hit rates and false alarm rates, are used as summary statistics to evaluate test performance. In the context of this study, a hit rate (also known true positive rate) is defined by the number of erroneous observations with a confidence score less than a pre-defined threshold divided by the total number of erroneous observations in a synthetic dataset conditioning on a test is applicable. For a test satisfying all the minimum requirements, a false alarm rate (also known as false positive rate) is defined by the number of unmodified raw observations with confidence score less than a pre-defined threshold divided by the total number of unmodified raw observations. A good test method is indicated by a high hit rate but a low false alarm rate. There is usually a trade-off between a hit rate and a false alarm rate, such that a higher hit rate will mean a higher false alarm rate and vice versa. Both overall (i.e., pooling from all dates and stations) and station-wise (i.e., pooling from all dates for a particular station) hit rates and false alarm rates are considered in this study to demonstrate the overall performance and performance at each station. We further consider the percentage of stations with a hit rate of at least 80% and a false alarm rate of at most 10% as an indicator of the percentage of stations where a test method works well for the purpose of parametric insurance applications.

Table 1 provides the performance of the AGCD, Rainfields and merged tests for the zero-rainfall synthetic dataset. The hit rates from the AGCD and Rainfields tests are the same, but the false alarm rate from the Rainfields test is almost two times greater than that from the AGCD test. The merged test leads to a substantially lower hit rate than any of the AGCD and Rainfields tests but at the same time a slightly lower false alarm rate. Due to the minimum requirements for each test, the AGCD and Rainfields tests cannot be applied to all test observations. For example, the Rainfields test is only applicable for about one-third of observations. One of the advantages of the merged test is that it can be used in more situations than any individual test. With a confidence score threshold of 10%, the final assessment based on the merged test can detect nearly all erroneous observations (with about 99% hit rates) while keeping false alarm rates very low (at about 1%) in the

Table 1 Summary statistics for the test performance of the AGCD, Rainfields and merged tests applied to the zero-rainfall synthetic dataset when the confidence score threshold is chosen to be 10%

	AGCD	Rainfields	Merged
Hit rate (%)	99.2	99.2	98.9
False alarm rate (%)	2.4	4.7	1.8
The percentage of observations that a test can be applied to	98.6	34.9	99.1
The percentage of stations with > 80% hit rates and < 10% false alarm rates (%)	100	94.2	100

zero-rainfall synthetic dataset. This suggests that our method is fully capable of identifying erroneous data (with at least 3 mm errors) when no rainfall actually occurs. Such good performance to detect rainfall occurrence applies to every station considered in the dataset.

Table 2 summarises the performance of the AGCD, Rainfields and merged tests for the non-zero-rainfall synthetic dataset. The overall hit rates for the non-zero-rainfall synthetic dataset are significantly lower than those for the zero-rainfall dataset. It is consistent with our intuition that predicting rainfall amount is much more difficult than predicting rainfall occurrence. The Rainfields test achieves the highest hit rate, followed by the AGCD and merged tests. To understand test performance in more detail, we further consider five categories based on the true rainfall amount and calculate evaluation metrics for each category. We found that our test method works best for low rainfall (i.e., ≤ 10 mm), followed by high rainfall (i.e., > 30 mm), but relatively poorly for medium rainfall (i.e., 10–30 mm). The hit rate and false alarm rate of the merged test are satisfactory in the category of low rainfall. As the true rainfall amount increases, hit rates decrease and/or false alarm rates increase. Nevertheless, only less than 1% of true observations have more than 30 mm rainfall but more than 18% of erroneous data are associated with more than 30 mm rainfall. The synthetic data is designed to be challenging enough to understand the boundary of our test performance. Though it is rare to observe > 30 mm rainfall in Australia, a lack of skill for > 30 mm rainfall may be a major hurdle for the proposed method to real applications with interests on extreme rainfall. Apart from including more reliable reference data, the improvement of model performance at high rainfall could be obtained by checking the quality of sub-daily data in the future research. The overall false alarm rates for the non-zero-rainfall dataset are similar to those for the zero-rainfall dataset, because the two datasets are generated based on the same original dataset and are only different with only a small portion of modified data. The false alarm rates for > 10 mm rainfall are relatively high, suggesting the predictions from AGCD and Rainfields are subject to larger uncertainty.

We investigate the influence of the size of an inserted error on the overall hit rate based on the non-zero rainfall synthetic data. Because the error structure for true rainfall less than 10 mm different to the structure of rainfall greater

Table 2 Summary statistics for the test performance of the AGCD, Rainfields and merged tests applied to the non-zero-rainfall synthetic dataset when the confidence score threshold is chosen to be 10%

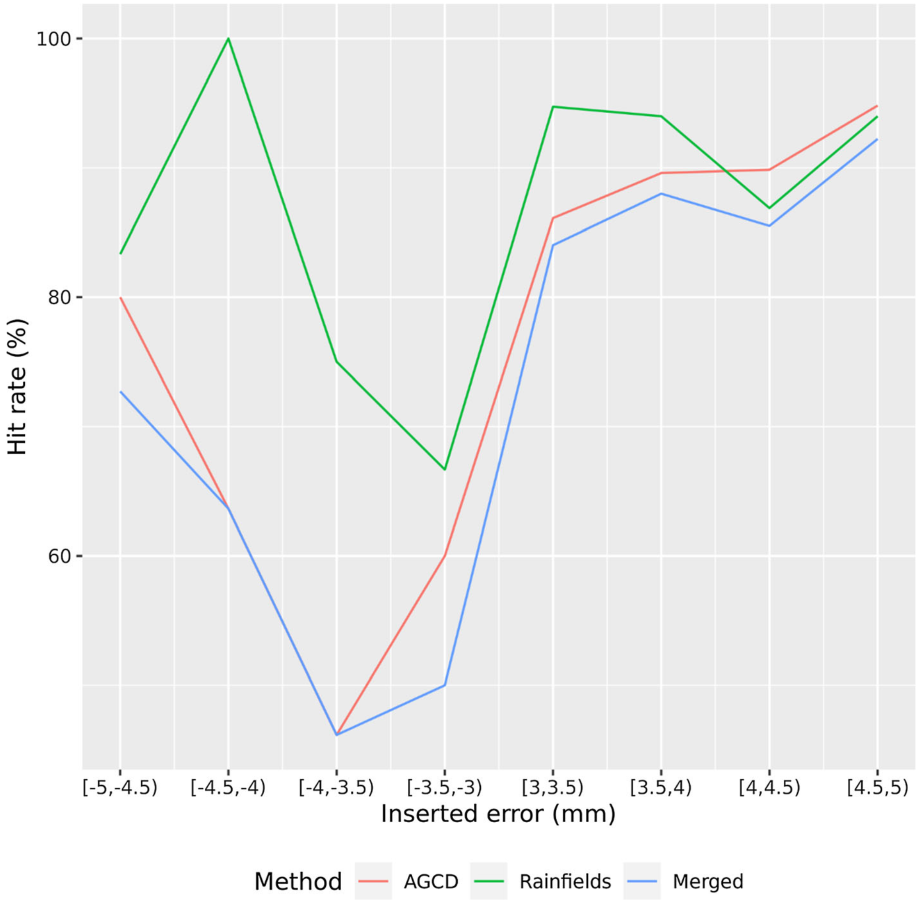
		AGCD	Rainfields	Merged
Hit rate (%)	All	79.6	87.6	76.7
	Non-zero rainfall	79.6	87.6	76.7
	zero rainfall	NA	NA	NA
	≤ 10 mm	87.8	91.2	85.1
	10–30 mm	67.3	83.2	64.7
	> 30 mm	80.7	87.6	78.1
False alarm rate (%)	All	2.2	4.5	1.7
	Non-zero rainfall	7.8	14.4	6.0
	zero rainfall	0.4	0.4	0.3
	≤ 10 mm	5.3	11.4	3.8
	10–30 mm	24.5	30.7	19.2
	> 30 mm	28.8	31.2	23.5
The percentage of observations that a test can be applied to		98.8	34.9	99.2
The percentage of stations with > 80% hit rates and < 10% false alarm rates (%)		56	71	49

than 10 mm, we show the relationship between hit rates and the size of an inserted error separately for two different rainfall ranges in Fig. 2 and 3. Given that the sign of inserted error is fixed as positive or negative, hit rates increase as the magnitude of an inserted error increases. The only exception is that the hit rate of the Rainfields test in Fig. 2 reaches the highest (i.e., 100%) with [− 4.5 mm, − 4 mm] inserted error. This is caused by the uncertainty from a small sample as the Rainfields test only is applicable for very few test observations. With the same magnitude of inserted errors, positive errors appear to more easily be detected than negative ones. The results of the synthetic error assessment are very susceptible to the magnitude of the inserted errors. The use of synthetic generators (Diez-Sierra et al. 2022) can help in this point in future works.

Fig. 4 shows how overall hit rates and false alarm rates relate to the confidence score threshold based on the non-zero rainfall synthetic data. Both hit rates and false alarm rates are increasing functions of the threshold. A method associated with a high hit rate also leads to a high false alarm rate. In this study, we choose the confidence score threshold to be 10% to keep a balance between false alarm rates and hit rates. To achieve a higher hit rate, the confidence score threshold may be set higher, such as 20%. This will result in a higher false alarm rate, especially in the range of high rainfall, and more effort is required to identify false alarms in further investigation.

Fig. 5 and Fig. 6 show the spatial distribution of station-wise hit rates and false alarm rates from the non-zero-rainfall dataset. The AGCD test can be applied to all test stations in the dataset and yields high hit rates and low false alarm rates on the western coast and central Australia. The Rainfields test only can be applied to a small portion of test stations, mainly on the coastal lines, and seems to work better than the AGCD test on the eastern coast. The merged test shares a similar spatial pattern of hit rates with the AGCD test, but improves the false alarm rates slightly, for example on the eastern coast. Darwin airport (station number: 14015) is one of the most challenging stations in the non-zero synthetic data with a hit rate of 44% and a false alarm rate of 5.8%. Fig. 7 shows the relationship between rainfall observations and the corresponding rainfall estimates from AGCD and Rainfields at the Darwin airport station. It is obvious that the distribution of rainfall observations at this station is extremely long tailed and both AGCD and Rainfields provide reasonably accurate rainfall estimates, even for an extreme rainfall event with

Fig. 2 Hit rate as a function of the inserted error for true rainfall less than 10 mm based on the non-zero rainfall synthetic data



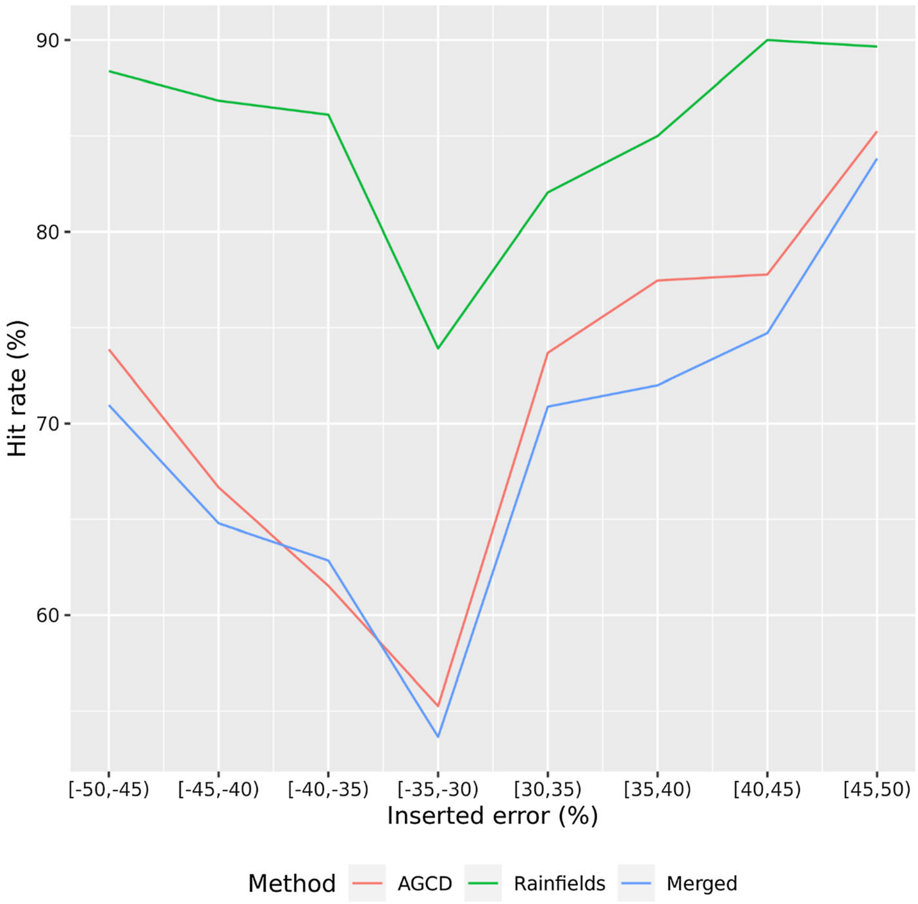


Fig. 3 Hit rate as a function of the inserted error (as a percentage of the true rainfall value) for true rainfall greater than 10 mm based on the non-zero rainfall synthetic data

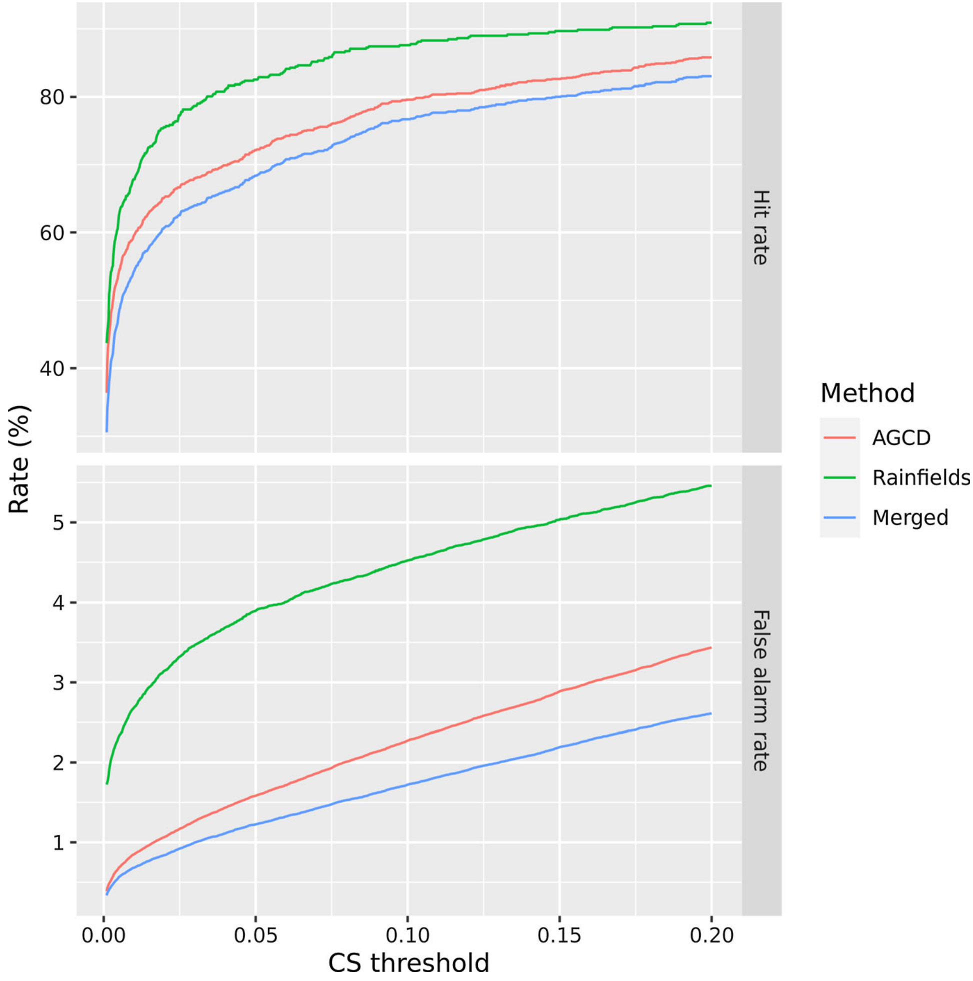
around 200 mm/day in the study period. The Darwin airport station and its surrounding stations are located in the tropical region and have highly variable seasonal and annual rainfall. In some cases, the inserted errors exceed the range of estimate errors, creating unrealistic synthetic data for at this station and in turn causing a low hit rate. The low hit rate is also caused by some modified observations (such as a modified observation of ~ 150 mm) that are very close to rainfall estimates from AGCD and/or Rainfields. In these situations, positive (or negative) inserted errors are applied when rainfall estimates are in fact over-estimated (or under-estimated) true rainfall.

We presume that some of these 100 stations are used in AGCD and in theory the synthetic data are not 100% independent from AGCD. Note that the list of the weather stations used for AGCD is not available for the general public and may vary from different days for many operational reasons. Nevertheless, the validation with synthetic data is, to our best knowledge, one of the most effective tools to understand the performance of the proposed procedure for three reasons: (1) we can assume that the true value and the data quality of each test observation are known and calculate performance indicators (such as hit rates and false alarm rates) conveniently; (2) AGCD does not honour the observations that are used for AGCD and the difference between AGCD and observations as a result of observation representativeness error and algorithm error is in general not zero; (3) the merged test is considered for a final assessment and AGCD test does not always dominate the merge test.

5 An application to real third-party rainfall data

We carry out a case study to evaluate the data quality of daily rainfall observations collected from third-party DPIRD weather stations. All DPIRD weather stations are located in WA and most of them are distributed in the Wheatbelt region and Southwest region. Because the DPIRD stations are operated or cooperated by a state agency, we believe that most the weather observations from DPIRD stations are of good quality and only a small portion of data is problematic. We evaluate one year of daily observations from 01/Jan/2019 to 31/Dec/2019 and estimate model parameters from three years of data in a separate period between 01/Jan/2016 and 31/Dec/2018. Based on data availability in evaluation and estimation

Fig. 4 The overall hit rate and false alarm rate as a function of the threshold of confidence score (CS threshold) based on the non-zero rainfall synthetic data



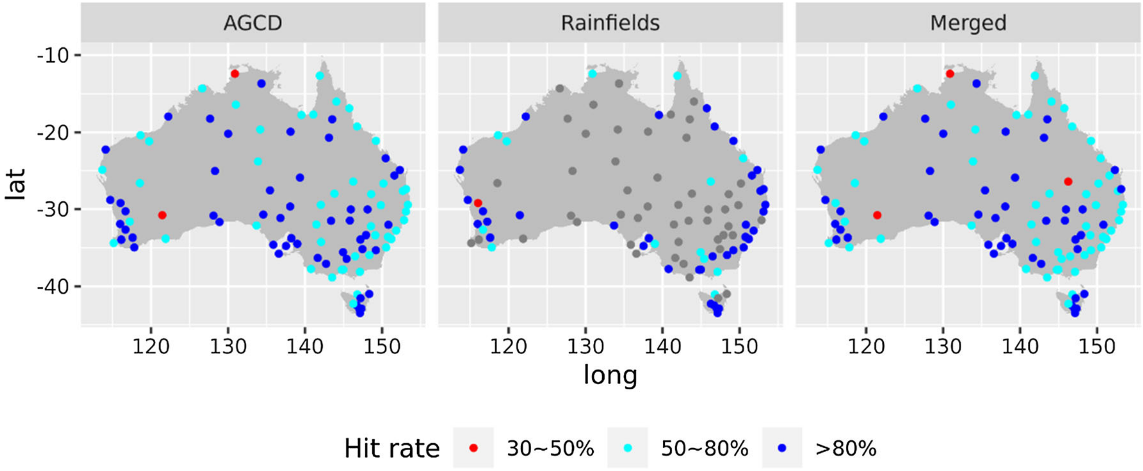


Fig. 5 Station-wise hit rates for different test methods

periods, we select and evaluate a total of 186 DPIRD stations with 68,777 daily rainfall observations (excluding 208 missing observations). DPIRD weather data are publicly available and can be accessed via an Application Programming Interface (API) from DPIRD (2022).

All of the DPIRD observations pass the domain test and we only perform AGCD and Rainfields tests if applicable. Table 3 presents the percentage of erroneous data flagged by two individual tests and the merged test. Due to the restriction of model assumptions and requirements, the

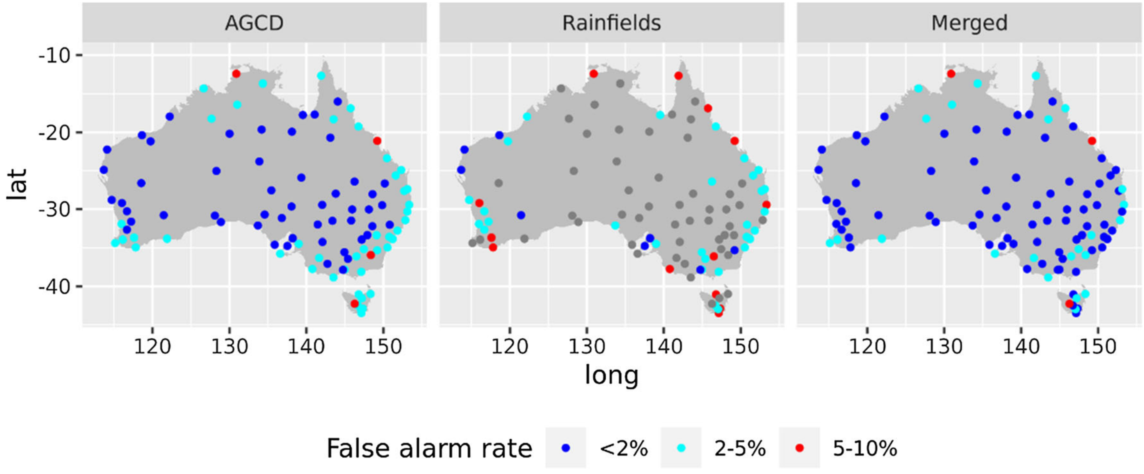
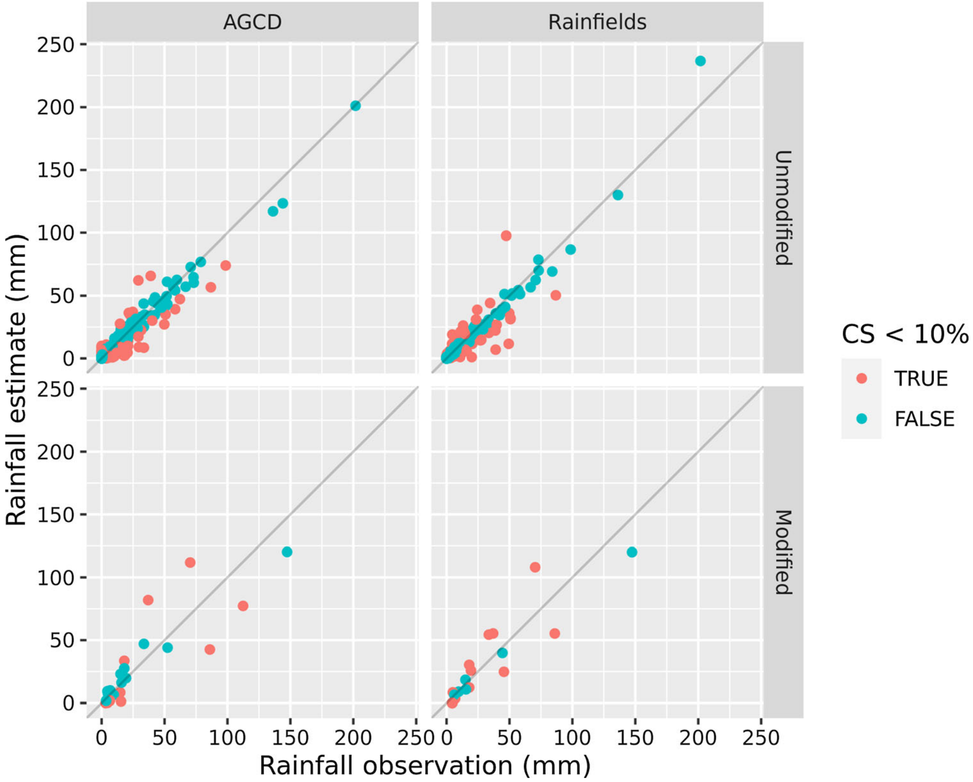


Fig. 6 Station-wise false alarm rates for different test methods

Fig. 7 A comparison of rainfall observations and estimates from AGCD and Rainfields. Modified (or unmodified) observations represent those observations with (or without) inserting random errors in the synthetic dataset. An observation associated with a confidence score (CL) less than (or not less than) 10% is labelled in red (or green)



AGCD, Rainfields and the merged tests can be applied to 149, 34 and 149 (out of 186) DPIRD stations respectively. The merged test flags 1054 DPIRD rainfall observations as suspect, which represents 2% of test data that can be evaluated. The rainfall observations with greater value tend to be more likely to be flagged. Because we have found that more false alarms occur with high rainfall from the experiment based on synthetic data, further investigation including a manual check is recommended to confirm whether those flagged high rainfall observations are indeed false alarms.

Fig. 8 shows the spatial distribution of the percentage of possible erroneous data at each individual DPIRD station. In general, more possible erroneous data are flagged at coastal stations than inland stations. None of our proposed tests can be applied to two clusters of stations, which are not covered by any rain radar and cannot be predicted well by AGCD rainfall estimates. Brunswick Junction (BJ) station is the DPIRD station with the highest percentage of flagged data (7.1%) in this study. This station departs from the nearest rain radar (Serpentine) by 88 km and there are three nearby primary rainfall stations operated by the BoM (including station 9982, station 9965 and station 109,507) within 20 km (Fig. 9). .

Table 4 provides detailed information on the flagged observations at the BJ station together with rainfall observations from the three closest primary stations. The Rainfields test is not applicable for this station because the correlation between radar Rainfields estimates and rainfall observations at the BJ station is just 20.6% In contrast, AGCD provides good rainfall estimates with a correlation of 94% and as a result the AGCD test is the only test that can be used for this station. Though true rainfall at this station is unknown, we attempt to further investigate the data quality from the supplementary information provided by nearby primary stations. We found that 13 observations are flagged from a 21-day period between 14/08/2019 to 03/09/2019 and most of them are significantly different from AGCD estimates and the nearby primary station observations. During this period, the BJ station reports non-zero rainfall on six days but AGCD and the nearby stations report zero (or nearly zero) rainfall. As the method is more capable of detecting rainfall occurrence than rainfall amount, these six non-zero rainfall observations are highly likely to be wrong. We also suspect that equipment failure may have happened during this period at the BJ station.

6 Other possible reference data

In this study, we only use two reference data (i.e., AGCD and Rainfields data) to test against third-party rainfall observations. In fact, three additional candidate reference data, including primary station observations, numerical weather prediction data and satellite rainfall estimates, have been considered in the method development. We decide to exclude these three possible reference data after initial investigation and exploratory analysis.

Table 3 The percentages of suspect observations for AGCD, Rainfields and merged tests in different categories

	AGCD	Rainfields	Merged
All	2.0% (53,778)	8.8% (10,900)	1.9% (53,809)
Non-zero rainfall	7.9% (12,504)	35.9% (2217)	7.7% (12,535)
zero rainfall	0.26% (41,274)	1.8% (8683)	0.2% (41,274)
$\leq$ 10 mm	5.1% (11,067)	29.6% (1989)	4.8% (11,072)
10–30 mm	28.5% (1283)	90.2% (216)	29.2% (1307)
$>$ 30 mm	35.7% (154)	100% (12)	36.5% (156)

A suspect observation is flagged by a confidence score being less than 10%. A number in parentheses denotes the total number of observations that can be assessed by a test

Primary station observations can provide the highest quality of rain estimates at point locations and have been used to form a spatial test in some operational QC applications. As we aim to design a fully automated quality control system to evaluate the data quality in near real time, only automatic weather stations (AWS) can be considered. However, we found that BoM operates a small number of AWS that can deliver accurate and near-real-time rainfall observations in Australia. For example, there are $\sim$ 730 tipping bucket rain gauge (TBRG) stations (170 of them also have a manual gauge which supersedes the TBRG as the primary record for the database) in Australia as of 2019. In addition to a lack of AWS, the quality of rainfall observations from AWS in Australia may have issues because BoM completes the full quality control for these observations some weeks after the end of the most recent month. The AGCD dataset is a better alternative for the purpose of this study. AGCD combines available rainfall station observations collected through electronic communication channels from all possible BoM stations (including TBRG stations, manual stations, hydrologic reference stations and those stations that are not publicly available) and performs initial quality control to screen for errors. Nevertheless, primary station observations can be applied to other regions or in the future when more AWS with prompt quality control become available.

Numerical weather prediction data and satellite rainfall estimates are other two interesting gridded datasets that may be served as possible reference data for rainfall. Numerical weather prediction provides weather forecasts on current weather conditions and satellite rainfall estimates measure precipitation from space from a constellation of research and operational satellites. Both datasets provide good spatial and temporal coverage, but their accuracy may be an issue. We have tested the accuracy of two candidate reference data in comparison with AGCD and Rainfields data based on the high-quality rainfall data used to generate the synthetic datasets in Sect. 4. The first one is the Australian Community Climate and Earth-System Simulator (ACCESS) NWP forecasts (Puri et al. 2013). The ACCESS NWP systems are based on the Unified Model/Variational Assimilation (UM/VAR) system developed by the United Kingdom Met Office (UKMO). This study uses the ACCESS system APS2 (Australian Parallel Suite version 2; see Bureau of Meteorology (2017) for reference) because APS2 data are available from 07/06/2016 to 25/09/2020 and cover the study period in the synthetic data example and the case study based on the DPIRD network. Though the latest version of ACCESS NWP (i.e., APS3) assimilates more observational inputs and has a higher resolution, we do not consider this dataset (only available from 23/07/2020 onwards) as it is unavailable for our study period. The second one is Global

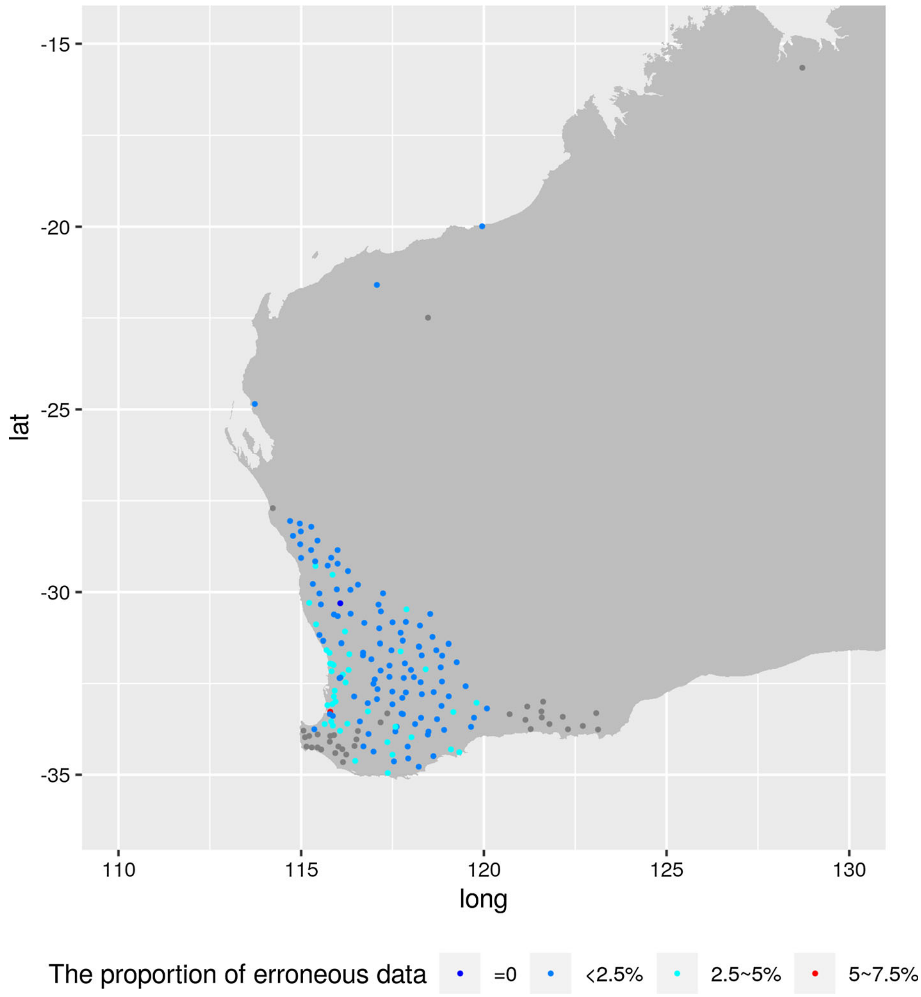


Fig. 8 The proportion of possible erroneous data at each DPIRD station. Gray dots denote the stations that the test method cannot be applied to

Precipitation Measurement Mission (GPM) (Hou et al. 2014) satellite rainfall estimates. In succeeding the Tropical Rainfall Measuring Mission (TRMM), GPM provides a new generation of precipitation measurements from space with an improved accuracy and higher space–time resolutions. The following four evaluation metrics are used to indicate the accuracy of rainfall estimates:

- Incorrect rainfall occurrence $= n^{-1}\sum_{i=1}^{n} I\{(O_i > ts, E_i > ts)OR(O_i \leq ts, E_i \leq ts)\} \times 100\%$,
- Relative bias $= \frac{\sum_{i=1}^{n}(E_i - O_i)}{\sum_{i=1}^{n} O_i} \times 100\%$,
- Relative RMSE $= \frac{\sqrt{n^{-1}\sum_{i=1}^{n}(E_i - O_i)^2}}{n^{-1}\sum_{i=1}^{n} O_i} \times 100\% = \frac{n^{\frac{1}{2}}\sqrt{\sum_{i=1}^{n}(E_i - O_i)^2}}{\sum_{i=1}^{n} O_i} \times 100\%$,
- Correlation $= \frac{\sum_{i=1}^{n}\left(E_i - \overline{E}\right)\left(O_i - \overline{O}\right)}{\sqrt{\sum_{i=1}^{n}\left(E_i - \overline{E}\right)^2 \sum_{i=1}^{n}\left(O_i - \overline{O}\right)^2}} \times 100\%$,

where O_i and E_i denote the observed and estimated rainfall, $\overline{O}$ and $\overline{E}$ denote the average of all observed and estimated rainfall, n is the sample size, I is the indicator function and ts is the threshold for rainfall occurrence. We set $ts = 2mm$, which is consistent with our choice in Sect. 3. Fig. 10 presents a comparison of four datasets (i.e., AGCD, Rainfields, ACCESS and GPM) against four evaluation metrics. It is evident that ACCESS and GPM are not as accurate as AGCD or Rainfields from all four metrics. Using ACCESS and GPM leads to an 8% incorrect rainfall occurrence and substantially large estimate errors suggested by large (negative) relative bias, large relative RMSE and low correlation. Because the relative RMSE of ACCESS and GPM are on average close to 100%, we doubt that these two datasets can be used to detect 30–50% inserted error in the synthetic dataset. Numerical weather prediction and satellite rainfall estimates may be valuable to evaluate third-party rainfall observation collected from other regions where gridded rainfall analysis data or rain radar data are not readily available.

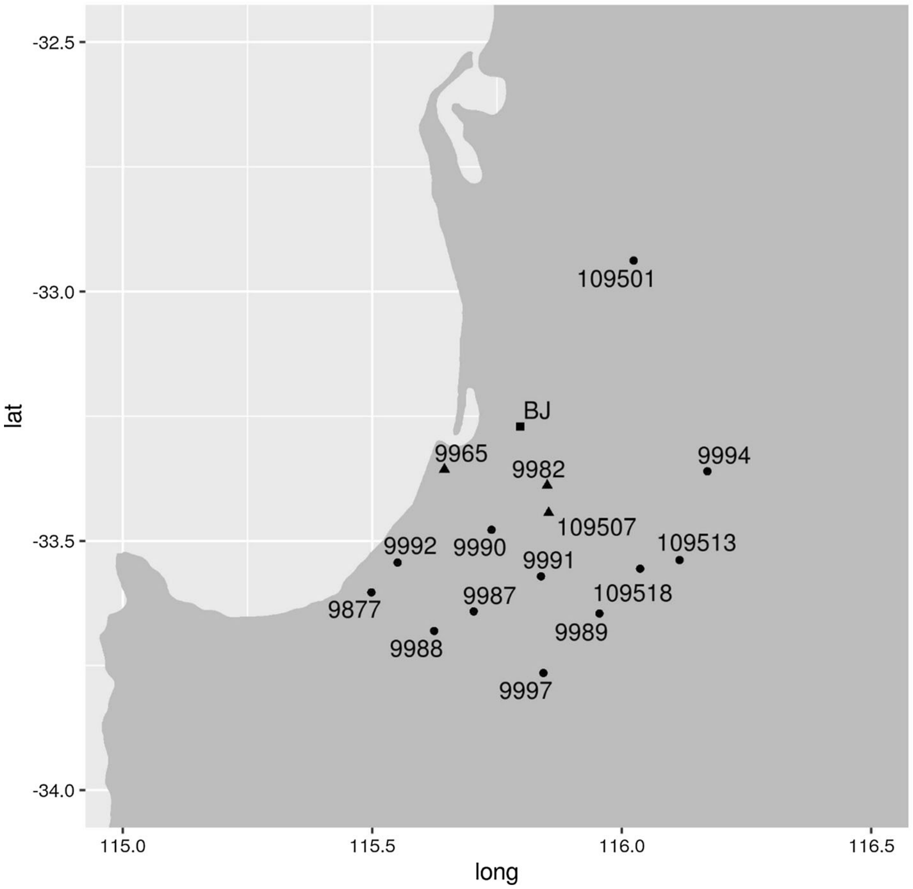


Fig. 9 The location of an example DPIRD station (BJ) (labelled with square) and the nearby primary rainfall stations operated by the BoM within 50 km. The closest three primary stations are labelled with triangles and the rest ones are labelled with circles

7 Conclusion and outlook

Third-party rainfall observations provide an improvement of the current observation network for rainfall monitoring in terms of spatio-temporal coverage. Although third-party weather stations can provide enormous quantities of near-real-time rainfall observations at specific locations, the quality of these data is susceptible due to a reduced focus on quality control. In this study, we develop a statistical method for an automated quality control system to evaluate daily rainfall observations collected from third-party stations in near real time. Two reliable reference datasets, including a rainfall analysis dataset (AGCD) and a radar rainfall dataset (Rainfields), have been identified to check against third-party observations in Australia. We treat rainfall observations as censored data to conveniently deal with a mixture of discrete and continuous distributions (as a result of many zero-rainfall readings) and apply a data transformation to model heterogeneous predictive errors from reference. We make the following conclusions based on these case studies:

(1) The merged test is extremely effective in detecting erroneous data with an incorrect rainfall occurrence.
(2) The merged test is less sensitive to evaluate the actual rainfall amount on a rainy day and can detect 76.7% of erroneous data with incorrect rainfall amount in the synthetic dataset.
(3) The merged test leads to a noticeably better (smaller) false alarm rate while the hit rate relative to the best individual test is only slight reduced.
(4) The performance of our proposed test can vary significantly across different stations.

Table 4 The detailed information on the flagged observations at the BJ station together with station observations from the three closest primary stations

Date	observation	estimate	CS$_{\mathrm{Merged}}$	AGCD	radar	CS$_{\mathrm{AGCD}}$	CS$_{\mathrm{Rainfields}}$	Stn9982	Stn9965	Stn109507
19/04/2019	11.2	15.4	8.2E-02	15.8	0	8.2E-02	NA	22	10.8	16
6/05/2019	5.8	10.3	2.8E-02	10.6	0	2.8E-02	NA	10.4	9.8	14.4
12/06/2019	5.2	2.3	9.0E-02	2.5	NA	9.0E-02	NA	1.2	0.6	0.6
22/06/2019	13.4	8.9	4.9E-02	9.3	NA	4.9E-02	NA	9.4	15.8	9
23/06/2019	52.4	35.0	1.5E-04	35.6	NA	1.5E-04	NA	36.2	44.8	43.6
27/06/2019	4.2	8.1	4.1E-02	8.4	NA	4.1E-02	NA	9.4	6	6
28/06/2019	13.6	6.6	1.4E-03	6.9	NA	1.4E-03	NA	5	5.8	8
29/06/2019	0.2	3.6	3.6E-02	3.8	NA	3.6E-02	NA	0	0	0
1/07/2019	0.8	3.5	9.8E-02	3.8	NA	9.8E-02	NA	9.2	1.2	4.4
4/07/2019	0	4.5	5.3E-03	4.8	NA	5.3E-03	NA	0	6	0.2
21/07/2019	9.4	4.0	5.5E-03	4.3	NA	5.5E-03	NA	5	4.2	1.6
14/08/2019	0.8	4.3	3.6E-02	4.5	NA	3.6E-02	NA	10	12.6	0.8
16/08/2019	3.4	0.0	1.8E-03	0.5	NA	1.8E-03	NA	0	0	0
17/08/2019	2.8	23.7	0.0E + 00	24.2	NA	0.0E + 00	NA	20.6	18.8	18.4
18/08/2019	2.8	0.0	1.1E-02	0.1	NA	1.1E-02	NA	0.2	0	0.4
19/08/2019	4.4	0.0	0.0E + 00	0.0	NA	0.0E + 00	NA	0	0	0
20/08/2019	7	0.0	0.0E + 00	0.0	NA	0.0E + 00	NA	0	0	0
23/08/2019	2.4	12.2	6.6E-07	12.6	NA	6.6E-07	NA	14.2	8.6	6.4
24/08/2019	3.8	0.0	1.8E-03	0.0	NA	1.8E-03	NA	0	0.2	0.2
25/08/2019	3.4	0.0	1.8E-03	0.0	NA	1.8E-03	NA	0.2	0	0
30/08/2019	3.8	34.7	0.0E + 00	35.3	NA	0.0E + 00	NA	34	17.4	27
31/08/2019	4.2	15.6	1.2E-07	16.0	NA	1.2E-07	NA	18.4	13	21.4
2/09/2019	2.8	12.5	1.2E-06	12.9	NA	1.2E-06	NA	14.2	18.6	18.4
3/09/2019	2.4	6.5	2.0E-02	6.8	NA	2.0E-02	NA	8.6	3.8	3.4
19/09/2019	0.4	6.3	5.4E-04	6.6	NA	5.4E-04	NA	9.2	5.2	6.2
1/11/2019	17	9.7	2.6E-03	10.0	NA	2.6E-03	NA	12.2	8.2	9.6

(5) The *p*-value based confidence score is a good measure of the quality of rainfall observations.

(6) The proposed test method is generic and can easily incorporate additional data sources (such as numerical weather prediction data and satellite rainfall estimates).

Quality evaluation and quality control are two related but different concepts. Quality control involves testing weather observations and determining if they are within the specifications for a given application. Quality evaluation also includes testing components but does not involve any decision-making procedure. In general, the outputs from quality evaluation can be served for the purpose of quality control together with benchmarks for data quality. Our proposed procedure is evaluated for the application to parametric insurance. In this context, the benchmarks for hit rates and false alarm rates are set as 80% and 10% respectively based on the application requirements from insurance companies. We would like to emphasise again that data quality that is good enough for one application may not in general good enough for other applications. End users are required to test our proposed procedure to understand model performance for their specific applications.

The error of reference datasets (e.g., AGCD rainfall estimates) can be correlated in space. For example, due to the presence of representativeness error, multiple third-

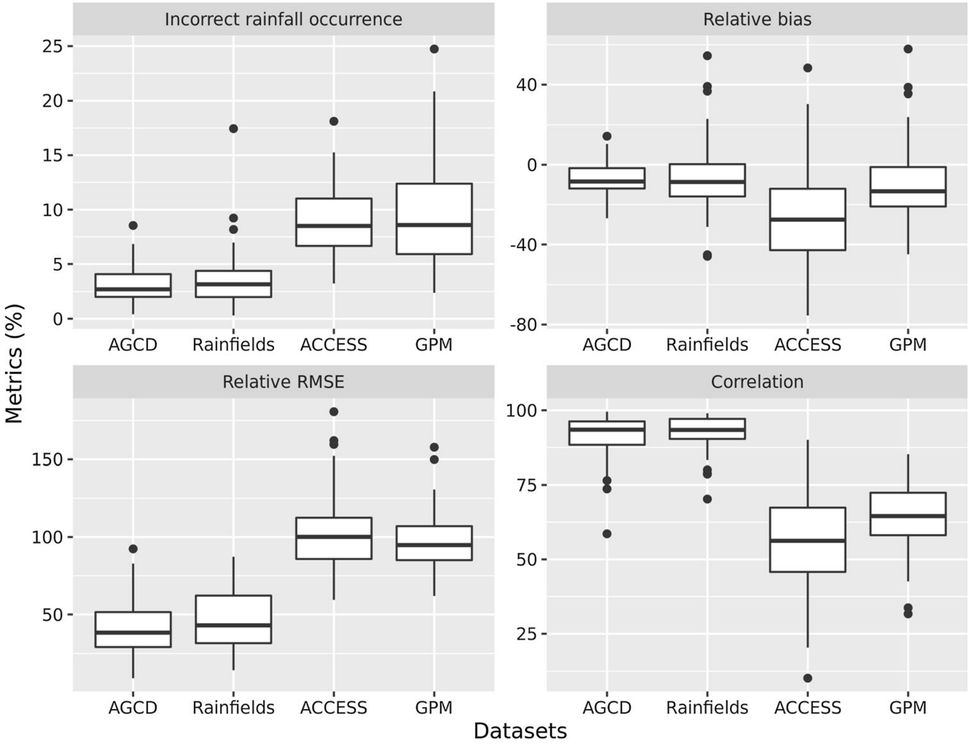


Fig. 10 A comparison of four datasets as rainfall estimates, including AGCD, Rainfields, ACCESS-R and GPM, based on four evaluation metrics

party stations in a very small region (such as the same city neighbourhood) correspond to the same AGCD grid point and always have the same AGCD rainfall estimate. As a result, the error term $\in(t)$ in Eq. (2) can be correlated across different stations. Because the proposed quality evaluation procedure is carried out at each station, the assumption on error dependence structure is not necessary. Though the presence of spatial pattern in $\in(t)$ does not reduce the effectiveness of the proposed method at each station, it may cause a spatial pattern in the model performance at different stations (e.g., some spatial pattern observed at the station-wise hit rates and false alarm rates in Figs. 5 and 6). In the early development, we assumed spatially correlated errors and trialled a spatial–temporal test (Shao et al. 2022). Because the spatial–temporal test is not easy for operational implementation, we decide to leave it for the future development.

We are working on extending the proposed QC procedure with the inclusion of third-party rainfall observations as reference data, which are particularly valuable for the locations where primary observations (used for AGCD) and radar information are insufficient. It still remains a challenging task because third-party observations are not 100% trusted and the observational errors in reference data have to be addressed carefully. Apart from this, further research work could consider (a) testing more complex algorithms to merge the assessment results from individual test, (b) selecting appropriate reference data, such as WFDE5 (Cucchi et al. 2020) and ERA5-Land (Hersbach et al. 2020), and extending the application of the proposed method to other countries in the world, (c) investigating the factors that may contribute to poor test performance, and (d) improving gridded rainfall estimates with the inclusion of quality-controlled third-party rainfall observations.

Appendix A: A full list of the weather stations used in the synthetic data examples

Number	Name	State	Latitude	Longitude	Height
1019	KALUMBURU	WA	− 14.296	126.645	23
2079	HALLS CREEK AIRPORT	WA	− 18.234	127.667	409.4
3003	BROOME AIRPORT	WA	− 17.948	122.235	7.42
4032	PORT HEDLAND AIRPORT	WA	− 20.373	118.632	6.4
4106	MARBLE BAR	WA	− 21.176	119.75	182.3
5007	LEARMONTH AIRPORT	WA	− 22.241	114.097	5
6011	CARNARVON AIRPORT	WA	− 24.888	113.67	4
7045	MEEKATHARRA AIRPORT	WA	− 26.614	118.537	517
8296	MORAWA AIRPORT	WA	− 29.204	116.025	271.4
8297	DALWALLINU	WA	− 30.276	116.671	324.5
8315	GERALDTON AIRPORT	WA	− 28.805	114.699	29.7
9021	PERTH AIRPORT	WA	− 31.928	115.976	15.4
9518	CAPE LEEUWIN	WA	− 34.373	115.136	13
9617	BRIDGETOWN	WA	− 33.949	116.131	178.66
9789	ESPERANCE	WA	− 33.83	121.893	25
9999	ALBANY AIRPORT	WA	− 34.941	117.816	68.4
10286	CUNDERDIN AIRFIELD	WA	− 31.622	117.222	216.7
10916	KATANNING	WA	− 33.686	117.606	320
10917	WANDERING	WA	− 32.672	116.671	275
11003	EUCLA	WA	− 31.68	128.896	93.1
11052	FORREST	WA	− 30.845	128.109	159
12038	KALGOORLIE-BOULDER AIRPORT	WA	− 30.785	121.453	365.3
13017	GILES METEOROLOGICAL OFFICE	SA	− 25.034	128.301	598
14015	DARWIN AIRPORT	NT	− 12.424	130.893	30.4
14627	BULMAN	NT	− 13.672	134.342	103.4
14825	VICTORIA RIVER DOWNS	NT	− 16.403	131.015	88.5
15135	TENNANT CREEK AIRPORT	NT	− 19.642	134.183	375.7
15590	ALICE SPRINGS AIRPORT	NT	− 23.795	133.889	546
15666	RABBIT FLAT	NT	− 20.182	130.015	340
16001	WOOMERA AERODROME	SA	− 31.156	136.805	166.6
16098	TARCOOLA AERO	SA	− 30.705	134.579	123
17043	OODNADATTA AIRPORT	SA	− 27.555	135.446	116.5
17126	MARREE AERO	SA	− 29.659	138.068	50
18012	CEDUNA AMO	SA	− 32.13	133.698	15.3
18192	PORT LINCOLN AWS	SA	− 34.599	135.878	8.5
21133	RAYVILLE PARK	SA	− 33.768	138.218	109.1
22031	MINLATON AERO	SA	− 34.748	137.528	32
22823	CAPE BORDA	SA	− 35.755	136.596	158
23373	NURIOOTPA PIRSA	SA	− 34.476	139.006	275
26021	MOUNT GAMBIER AERO	SA	− 37.747	140.774	63
27045	WEIPA AERO	QLD	− 12.678	141.921	17.96
28004	PALMERVILLE	QLD	− 16	144.075	203.8
29063	NORMANTON AIRPORT	QLD	− 17.687	141.073	18.418
29077	BURKETOWN AIRPORT	QLD	− 17.748	139.536	5.699
30124	GEORGETOWN AIRPORT	QLD	− 18.304	143.531	301.77
30161	RICHMOND AIRPORT	QLD	− 20.7	143.114	206.3
31011	CAIRNS AERO	QLD	− 16.874	145.746	2.22

Number	Name	State	Latitude	Longitude	Height
32040	TOWNSVILLE AERO	QLD	− 19.248	146.766	4.34
33119	MACKAY M.O	QLD	− 21.117	149.217	30.264
37010	CAMOOWEAL TOWNSHIP	QLD	− 19.923	138.121	231.2
38026	BIRDSVILLE AIRPORT	QLD	− 25.898	139.347	46.6
39066	GAYNDAH AIRPORT	QLD	− 25.617	151.616	110.9
39083	ROCKHAMPTON AERO	QLD	− 23.375	150.478	10.4
39128	BUNDABERG AERO	QLD	− 24.907	152.323	30.82
40004	AMBERLEY AMO	QLD	− 27.63	152.711	24.2
40842	BRISBANE AERO	QLD	− 27.392	153.129	4.51
42112	MILES CONSTANCE STREET	QLD	− 26.657	150.182	304.8
43109	ST GEORGE AIRPORT	QLD	− 28.048	148.596	198.5
44021	CHARLEVILLE AERO	QLD	− 26.414	146.256	301.6
45025	THARGOMINDAH AIRPORT	QLD	− 27.987	143.815	130.886
46012	WILCANNIA AERODROME AWS	NSW	− 31.519	143.385	94.3
46126	TIBOOBURRA AIRPORT	NSW	− 29.445	142.057	176.4
47048	BROKEN HILL AIRPORT AWS	NSW	− 32.001	141.469	281.3
48027	COBAR MO	NSW	− 31.484	145.829	260
48245	BOURKE AIRPORT AWS	NSW	− 30.036	145.952	107.3
50017	WEST WYALONG AIRPORT AWS	NSW	− 33.938	147.196	257
52088	WALGETT AIRPORT AWS	NSW	− 30.037	148.122	133
53115	MOREE AERO	NSW	− 29.49	149.847	213
58012	YAMBA PILOT STATION	NSW	− 29.433	153.363	27.4
59151	COFFS HARBOUR AIRPORT	NSW	− 30.319	153.116	3.5
60139	PORT MACQUARIE AIRPORT AWS (COMPARISON)	NSW	− 31.434	152.866	4.2
61078	WILLIAMTOWN RAAF	NSW	− 32.794	151.836	7.5
61363	SCONE AIRPORT AWS	NSW	− 32.034	150.826	221.4
63303	ORANGE AIRPORT AWS	NSW	− 33.377	149.126	944.65
65070	DUBBO AIRPORT AWS	NSW	− 32.221	148.575	284
65103	FORBES AIRPORT AWS	NSW	− 33.363	147.921	230.4
66214	SYDNEY (OBSERVATORY HILL)	NSW	− 33.859	151.205	43.37
67105	RICHMOND RAAF	NSW	− 33.6	150.776	19
68072	NOWRA RAN AIR STATION AWS	NSW	− 34.947	150.535	109
70351	CANBERRA AIRPORT	NSW	− 35.309	149.2	577.05
72150	WAGGA WAGGA AMO	NSW	− 35.158	147.458	212
72161	CABRAMURRA SMHEA AWS	NSW	− 35.937	148.378	1482.4
74258	DENILIQUIN AIRPORT AWS	NSW	− 35.558	144.946	94
76031	MILDURA AIRPORT	VIC	− 34.236	142.087	50
78015	NHILL AERODROME	VIC	− 36.309	141.649	138.9
79105	STAWELL AERODROME	VIC	− 37.072	142.74	235.364
81125	SHEPPARTON AIRPORT	VIC	− 36.429	145.395	113.9
82039	RUTHERGLEN RESEARCH	VIC	− 36.105	146.509	175
85072	EAST SALE	VIC	− 38.116	147.132	4.6
86338	MELBOURNE (OLYMPIC PARK)	VIC	− 37.826	144.982	7.53
87031	LAVERTON RAAF	VIC	− 37.857	144.757	20.1
90015	CAPE OTWAY LIGHTHOUSE	VIC	− 38.856	143.513	82
91293	LOW HEAD	TAS	− 41.055	146.787	3
91311	LAUNCESTON AIRPORT	TAS	− 41.548	147.216	166.9
92045	LARAPUNA (EDDYSTONE POINT)	TAS	− 40.993	148.347	19.7
94029	HOBART (ELLERSLIE ROAD)	TAS	− 42.89	147.328	50.5

Number	Name	State	Latitude	Longitude	Height
94198	CAPE BRUNY (CAPE BRUNY)	TAS	− 43.489	147.144	59.7
94220	GROVE (RESEARCH STATION)	TAS	− 42.984	147.076	65
95048	OUSE FIRE STATION	TAS	− 42.484	146.711	90
96003	BUTLERS GORGE	TAS	− 42.275	146.276	667

Acknowledgements We acknowledges the consistent support from the Australian Bureau of Meteorology. We thank Doerte Jakob and Liam Carroll from the BoM for the technical discussions and insight. We also thank Alister Hawksford, Boris Alvarenga, Paul Martin and Sunil Kokare from the BoM for managing this project and ensuring that we received necessary information on techniques, data, and client requirements. We also are grateful to two reviewers and the editor Salvatore Grimaldi for providing constructive comments and suggestions.

Author contributions ML Conceptualization, Methodology, Software, Validation, Formal analysis, Writing—Original Draft, Writing—Review & Editing, Visualization; QS Project administration, Investigation, Supervision, Funding acquisition, Contribution of knowledge, Writing—Review & Editing; JJD Investigation, Contribution of knowledge, Writing—Review & Editing; AR Investigation, Contribution of knowledge, Writing—Review & Editing; AP Investigation, Contribution of knowledge, Writing—Review & Editing; BH Project administration, Investigation, Writing—Review & Editing; ZH Resources, Data Curation, Investigation, Contribution of knowledge, Writing—Review & Editing; PS Resources, Data Curation, Investigation, Contribution of knowledge, Writing—Review & Editing.

Funding Open access funding provided by CSIRO Library Services.

Code and data availability It would be great to share all data and code with the public following the FAIR principles (Iturbide et al. 2022). However, due to commercial intellectual property protection, we are not able to release all code and data to the public yet. Please contact the authors for more detail.

Declarations

Conflict of interest The authors have not disclosed any competing interests.

References

Assumpcao TH, Popescu I, Jonoski A, Solomatine DP (2018) Citizen observations contributing to flood modelling: opportunities and challenges. Hydrol Earth Syst Sci 22(2):1473–1489. https://doi.org/10.5194/hess-22-1473-2018

Bardossy A, Seidel J, El Hachem A (2021) The use of personal weather station observations to improve precipitation estimation and interpolation. Hydrol Earth Syst Sci 25(2):583–601. https://doi.org/10.5194/hess-25-583-2021

Beele E, Reyniers M, Aerts R, Somers B (2022) Quality control and correction method for air temperature data from a citizen science weather station network in Leuven. Belgium Earth Syst Sci Data Discuss 2022:1–43. https://doi.org/10.5194/essd-2022-113

Bell S, Cornford D, Bastin L (2015) How good are citizen weather stations? Addressing Biased Opin Weather 70(3):75–84. https://doi.org/10.1002/wea.2316

Bureau of meteorology (2017) APS2 upgrade to the ACCESS-TC numerical weather prediction system. *BNOC* operational bulletin no 105. Retrieved from http://www.bom.gov.au/australia/charts/bulletins/APOB105.pdf. Acceesed: 10 May 2022

Bureau of meteorology (2022a) Climate data online. retrieved from http://www.bom.gov.au/climate/data/. Acceesed: 10 May 2022a

Bureau of meteorology (2022b) The cross validated error grids of AGCD rainfall. Retrieved from http://opendap.bom.gov.au:8080/thredds/catalog/agcd/precip/rmse/r005/01day/catalog.html. Acceesed: 10 May 2022b

Bureau of meteorology (2022c) Radar images. Retrieved from http://www.bom.gov.au/australia/radar/. Acceesed: 10 May 2022c

Buytaert W, Zulkafli Z, Grainger S, Acosta L, Alemie TC, Bastiaensen J et al (2014) Citizen science in hydrology and water resources: opportunities for knowledge generation, ecosystem service management, and sustainable development. Front Earth Sci 2:1–21. https://doi.org/10.3389/feart.2014.00026

Campbell JL, Rustad LE, Porter JH, Taylor JR, Dereszynski EW, Shanley JB et al (2013) Quantity is nothing without quality: automated QA/QC for streaming environmental sensor data. Bioscience 63(7):574–585. https://doi.org/10.1525/bio.2013.63.7.10

Chakraborty A, Lahiri SN, Wilson A (2020) A statistical analysis of noisy crowdsourced weather data. Ann Appl Stat 14(1):116–142. https://doi.org/10.1214/19-AOAS1290

Chen JY, Saunders K, Whan K (2021b) Quality control and bias adjustment of crowdsourced wind speed observations. Q J R Meteorol Soc 147(740):3647–3664. https://doi.org/10.1002/qj.4146

Chen AB, Behl M, Goodall JL (2018) Trust me, my neighbors say it's raining outside: ensuring data trustworthiness for crowdsourced weather stations. Paper Presented at the Proceedings of the 5th Conference on Systems for Built Environments https://doi.org/10.1145/3276774.3276792

Chen AB, Behl M, Goodall JL (2021a) Assessing the trustworthiness of crowdsourced rainfall networks: a reputation system approach. Water Resour Res. https://doi.org/10.1029/2021WR029721

Clement KY, Botzen WJW, Brouwer R, Aerts JCJH (2018) A global review of the impact of basis risk on the functioning of and demand for index insurance. Int J Disaster Risk Reduct 28:845–853. https://doi.org/10.1016/j.ijdrr.2018.01.001

Cucchi M, Weedon GP, Amici A, Bellouin N, Lange S, Schmied HM et al (2020) WFDE5: bias-adjusted ERA5 reanalysis data for impact studies. Earth Syst Sci Data 12(3):2097–2120. https://doi.org/10.5194/essd-12-2097-2020

de Vos L, Leijnse H, Overeem A, Uijlenhoet R (2019) Quality control for crowdsourced personal weather stations to enable operational rainfall monitoring. Geophys Res Lett 46(15):8820–8829. https://doi.org/10.1029/2019GL083731

Diez-Sierra J, Navas S, Jesus MD (2022) Neoprene: an open-source python library for spatial rainfall generation based on the neyman-scott process. SSRN

DPIRD (2022) The DPIRD weather v2 API. Retrieved from https://weather.agric.wa.gov.au/developer-api. Accessed 10 May 2022

Droste AM, Heusinkveld BG, Fenner D, Steeneveld GJ (2020) Assessing the potential and application of crowdsourced urban wind data. Q J R Meteorol Soc 146(731):2671–2688

Evans A, Jones D, Smalley R, Lellyett S (2020) An enhanced gridded rainfall analysis scheme for Australia (1925738124). Retrieved from http://www.bom.gov.au/research/publications/researchreports/BRR-041.pdf

Fenner D, Bechtel B, Demuzere M, Kittner J, Meier F (2021) CrowdQC+-A Quality-control for crowdsourced air-temperature observations enabling world-wide urban climate applications. Front Environ Sci. https://doi.org/10.3389/fenvs.2021.720747

Greatrex H, Hansen J, Garvin S, Diro R, Blakeley S, Le Guen M, et al (2015) Scaling up index insurance for smallholder farmers: Recent evidence and insights. (CCAFS report no. 14). Retrieved from www.ccafs.cgiar.org

Herman JD, Quinn JD, Steinschneider S, Giuliani M, Fletcher S (2020) Climate adaptation as a control problem: review and perspectives on dynamic water resources planning under uncertainty. Water Resour Res. https://doi.org/10.1029/2019WR025502

Hersbach H, Bell B, Berrisford P, Hirahara S, Horanyi A, Munoz-Sabater J et al (2020) The ERA5 global reanalysis. Q J R Meteorol Soc 146(730):1999–2049. https://doi.org/10.1002/qj.3803

Hou AY, Kakar RK, Neeck S, Azarbarzin AA, Kummerow CD, Kojima M et al (2014) The global precipitation measurement mission. Bull Am Meteor Soc 95(5):701–722. https://doi.org/10.1175/BAMS-D-13-00164.1

Iturbide M, Fernandez J, Gutierrez JM, Pirani A, Huard D, Al Khourdajie A et al (2022) Implementation of FAIR principles in the IPCC: the WGI AR6 Atlas repository. Sci Data. https://doi.org/10.1038/s41597-022-01739-y

Janjic T, Bormann N, Bocquet M, Carton JA, Cohn SE, Dance SL et al (2018) On the representation error in data assimilation. Q J R Meteorol Soc 144(713):1257–1278. https://doi.org/10.1002/qj.3130

Jones DA, Wang W, Fawcett R (2009) High-quality spatial climate data-sets for Australia. Aust Meteorol Oceanogr J 58:233–248. https://doi.org/10.22499/2.5804.003

Li M, Wang QJ, Bennett JC (2013) Accounting for seasonal dependence in hydrological model errors and prediction uncertainty. Water Resour Res 49(9):5913–5929. https://doi.org/10.1002/wrcr.20445

Li M, Wang QJ, Bennett JC, Robertson DE (2016) Error reduction and representation in stages (ERRIS) in hydrological modelling for ensemble streamflow forecasting. Hydrol Earth Syst Sci 20(9):3561–3579. https://doi.org/10.5194/hess-20-3561-2016

Marengo JA, Souza CA, Thonicke K, Burton C, Halladay K, Betts RA et al (2018) Changes in climate and land use over the amazon region: current and future variability and trends. Front Earth Sci. https://doi.org/10.3389/feart.2018.00228

May P, Protat A, Seed A, Rennie S, Wang X, Cass C, Murphy A (2013) The use of advanced radar in the Bureau of meteorology. Paper Presented at the 2013 International Conference on Radar https://doi.org/10.1109/RADAR.2013.6651952

Meier F, Fenner D, Grassmann T, Otto M, Scherer D (2017) Crowdsourcing air temperature from citizen weather stations for urban climate research. Urban Clim 19:170–191. https://doi.org/10.1016/j.uclim.2017.01.006

Muller CL, Chapman L, Johnston S, Kidd C, Illingworth S, Foody G et al (2015) Crowdsourcing for climate and atmospheric sciences: current status and future potential. Int J Climatol 35(11):3185–3203. https://doi.org/10.1002/joc.4210

Murphy BF, Timbal B (2008) A review of recent climate variability and climate change in southeastern Australia. Int J Climatol 28(7):859–879. https://doi.org/10.1002/joc.1627

Napoly A, Grassmann T, Meier F, Fenner D (2018) Development and application of a statistically-based quality control for crowdsourced air temperature data. Front Earth Sci. https://doi.org/10.3389/feart.2018.00118

NCI (2022) Australian Gridded Climate Data (AGCD). Retrieved from https://doi.org/10.25914/6009600304b02. Acceesed: 10 May 2022

Neupane J, Guo WX (2019) Agronomic basis and strategies for precision water management: a review. Agron-Basel. https://doi.org/10.3390/agronomy9020087

Puri K, Dietachmayer G, Steinle P, Dix M, Rikus L, Logan L et al (2013) Implementation of the initial ACCESS numerical weather prediction system. Aust Meteorol Oceanogr J 63:265–284. https://doi.org/10.22499/2.6302.001

Seed A, Duthie E, Chumchean S (2007) Rainfields: the Australian Bureau of meteorology system for quantitative precipitation estimation. Paper presented at the Proc of the 33rd Conf on Radar Meteorology, Cairns, Australia

Shao Q, Li M, Dabrowski J, Bakar S, Rahman A, Powell A, Henderson B (2022) An operational framework to automatically evaluate the quality of weather observations from third-party stations. Paper presented at the AI4Environment: First Australasian Symposium on Artificial Intelligence for the Environment, Perth, Australia, December 5–9 https://doi.org/10.48550/arXiv.2212.01998. https://arxiv.org/abs/2212.01998.

Wang QJ, Shrestha DL, Robertson DE, Pokhrel P (2012) A log-sinh transformation for data normalization and variance stabilization. Water Resour Res 48(5)

Zheng FF, Tao RL, Maier HR, See L, Savic D, Zhang TQ et al (2018) Crowdsourcing methods for data collection in geophysics: state of the art, issues, and future directions. Rev Geophys 56(4):698–740. https://doi.org/10.1029/2018RG000616