# Data Descriptor Template

**Scope Guidelines**

**Data Descriptors** submitted to *Scientific Data* should provide detailed descriptions of valuable research datasets, including the methods used to collect the data and technical analyses supporting the quality of the measurements. Data Descriptors focus on helping others reuse data, rather than testing hypotheses, or presenting new interpretations, methods or in-depth analyses. Relevant datasets must be deposited in an appropriate public repository prior to Data Descriptor submission, and their completeness will be considered during editorial evaluation and peer review. The data must be made publicly available without restriction in the event that the Data Descriptor is accepted for publication (excepting reasonable controls related to human privacy issues or public safety).

### Title

*Compilation of Meteorological Dataset for Suva, Fiji 2013-2019*

### Authors

Ravinesh Chand1, Ajal Kumar2, Dinesh Rao1 & Rohitash Chandra3

**Affiliations**

1. School of Computing, Information & Mathematical Sciences, The University of the South Pacific, Suva, Fiji.

2. School of Engineering & Physics, The University of the South Pacific, Suva, Fiji.

3. School of Mathematics and Statistics, The University of New South Wales, Sydney, Australia.

Corresponding author(s): Ravinesh Chand ([chand\_r@usp.ac.fj](mailto:chand_r@usp.ac.fj))

### Abstract

*170 words maximum*

The Abstract should succinctly describe the study, the assay(s) performed, the resulting data, and their reuse potential, but should not make any claims regarding new scientific findings. No references are allowed in this section.

To be completed

### Background & Summary

*700 words maximum*

The Background & Summary should provide an overview of the study design, the assay(s) performed, and the data generated, including any background information needed to put this study in the context of previous work and the literature, and should reference literature as needed. The section should also briefly outline the broader goals that motivated collection of the data, as well as their potential reuse value. We also encourage authors to include a figure that provides a schematic overview of the study and assay(s) design.

The need to collect meteorological data in real time that can be used in forecasting of climatological and hydrological processes is generally recognized as important for life and property safety. Nowadays, automated instruments are set up to collect meteorological data in real-time through weather stations across many countries around the world. Here we present real- time weather data obtained from School of Engineering & Physics (SEP) Automatic Weather Station located at The University of the South Pacific (USP), Laucala Campus, Suva, Fiji (Figure 1).

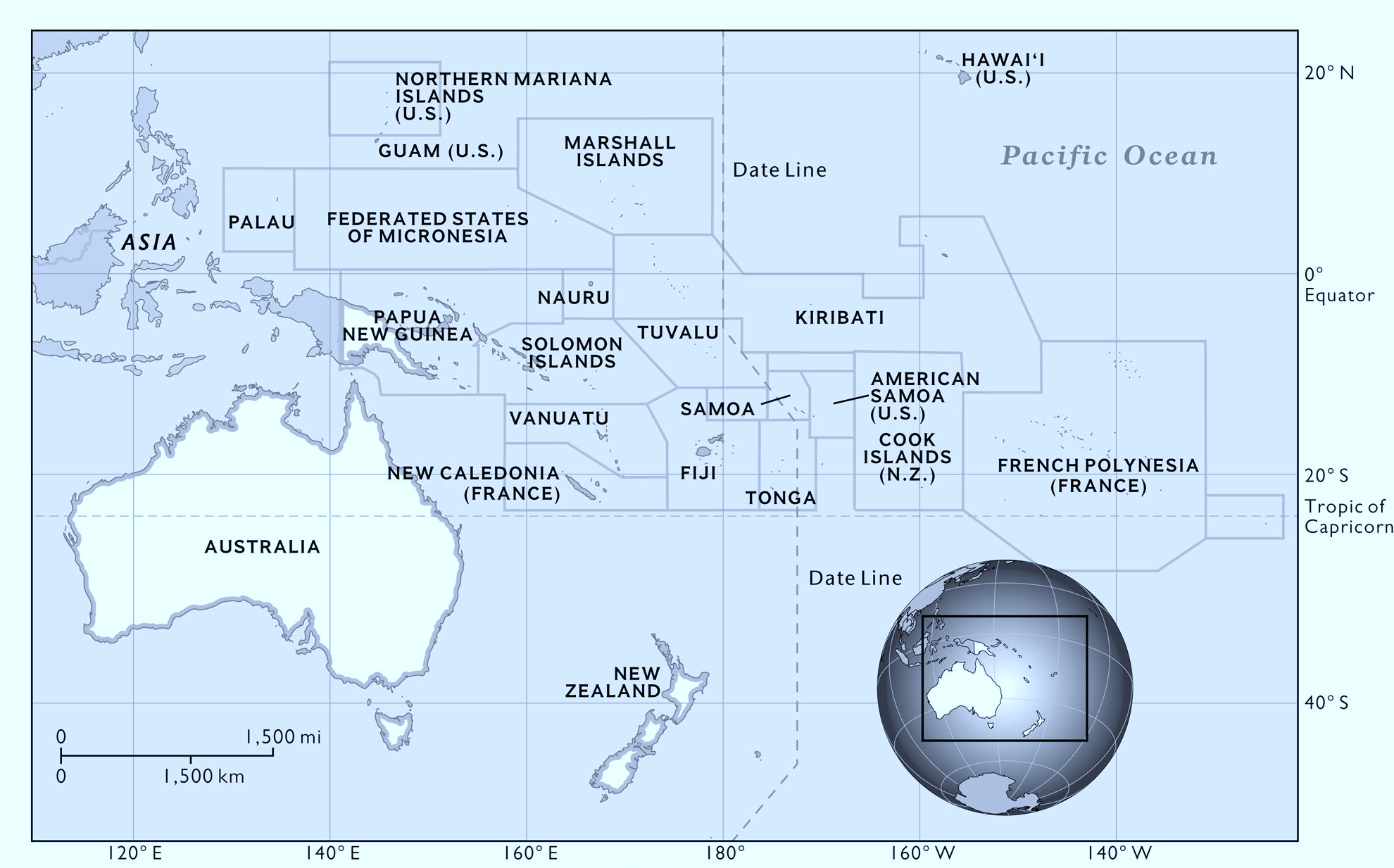
SEP's Automatic Weather Station was first established in 2000, replaced in 2006 and again replaced with the reliable, high- precision weather instruments in 2013. This weather station was set up mainly to collect real time data and be used exclusively for laboratory and/or project research by Physics Undergraduate students. It was also set up to collect long-term data for future research studies. Thus, the weather data presented in this paper was acquired at School of Engineering & Physics (SEP), USP, and includes daily average air temperature, relative humidity, vapor pressure, barometric pressure, solar radiation, horizontal wind speed, wind direction, soil moisture, soil temperature and total daily rainfall for the city of Suva (Lat.: 18.08°S, Long.: 178.45°E), Fiji. This dataset spans from 1 April 2013 to 31 December 2019.

Suva is situated on the southeastern part of the main island of Fiji, Viti Levu and the rainfall at Suva is heavily affected by the southeasterly trade winds by location  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
1. Thus, this southeastern part of Fiji's main island remains cool and rainy; while the western parts are mostly dry, rendering Suva one of the largest accumulations of rainfall from Fiji’s weather records  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
1. The general climate of the Fiji comprises of distinct rainy seasons (November – April) and dry seasons (May – October)  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
2, with cyclonic storms mostly occurring from November to January.

Fiji Meteorological Service (FMS) is an ISO accredited Aviation Meteorology & Climatology Services provider with headquarters located in Nadi, Fiji. FMS operates an observation network of 64 manned observation stations comprising 31 manual rainfall stations, 17 climate stations and 16 synoptic stations; and 29 automatic weather stations, 4 automatic weather observation stations designed for airports only (these are situated at Nausori, Nadi, Savusavu and Matei Airpots) and 8 TB3 Rainfall stations (records rainfall only). These stations are situated at different locations across the county (Figure 2). The weather data may also be acquired from FMS but cannot be released explicitly due to confidentiality agreement.

The dataset described in this paper can be used for many applications including modeling/forecasting of hydrological parameters such as rainfall  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
3–10 using various combinations of the other meteorological variables as an input to the model. Other climatological processes such as solar radiation  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
11 can also be modelled. The weather data can also be used for climatic monitoring processes, identification of extreme weather conditions and studying rainfall trend  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
12,13 and seasonality patterns for the city of Suva, Fiji. Data provided here might also be combined with various sources of other meteorological time series databases. A study by Onwubolu et al. (2007)  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
2 employed self- organizing data mining techniques in weather data forecasting including daily temperature, daily pressure and monthly rainfall using the weather data from SEP. The weather data acquired from SEP were validated by visiting FMS to ensure that SEP’s automatic weather station instruments were established in accordance with World Meteorological Organization standards. Below we describe the weather data, the missing data imputation methodology as well as with a cross- validation procedure.

**a**



**b**

****

Figure . A geographical map, showing (a) Fiji located in the South Pacific Region, (b) the measuring station, Suva, Fiji.

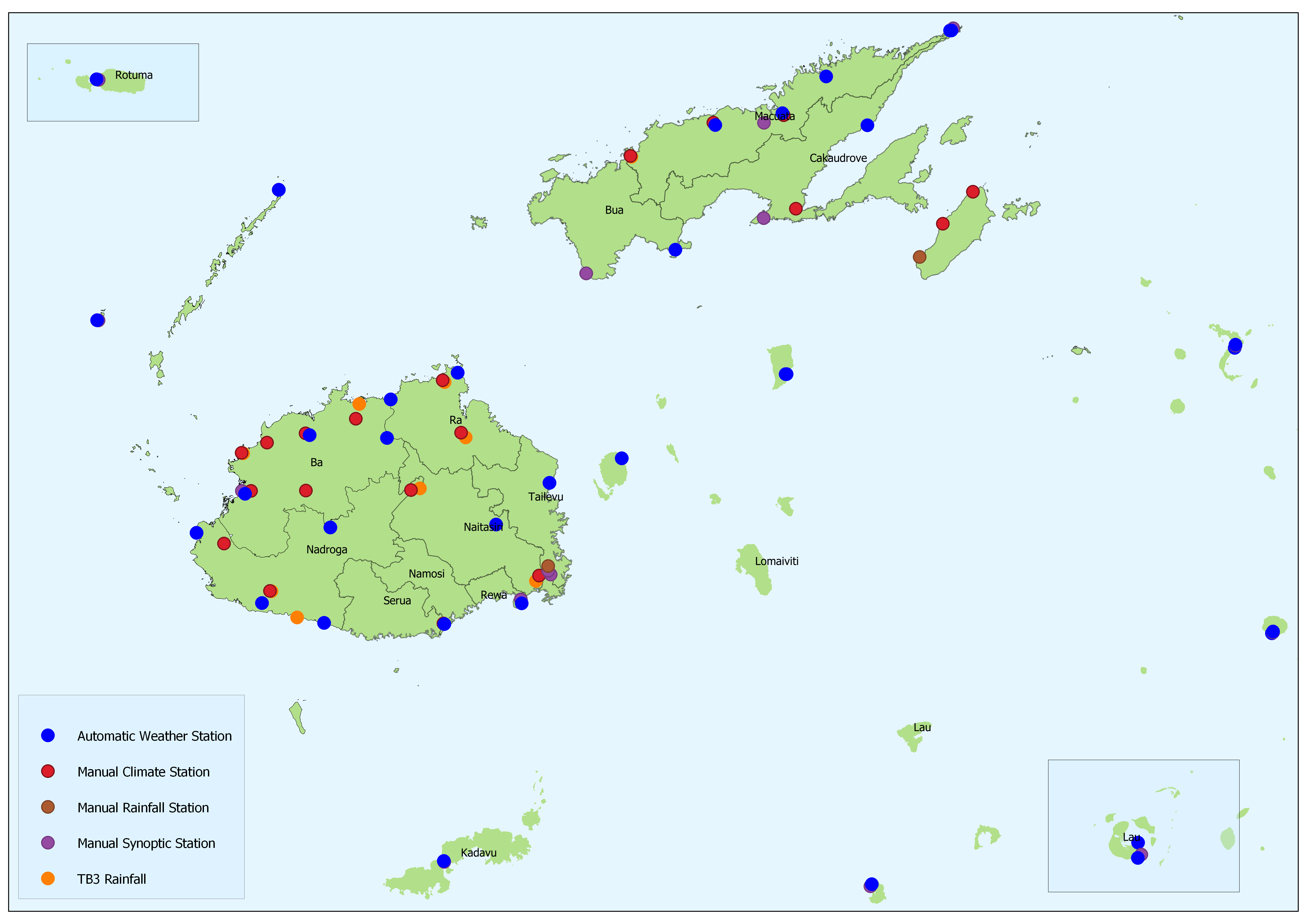


Figure . Fiji Meterological Service’s Observing Stations.

Source: <https://www.met.gov.fj/index.php?page=climateData>

### Methods

The Methods should include detailed text describing any steps or procedures used in producing the data, including full descriptions of the experimental design, data acquisition assays, and any computational processing (e.g. normalization, image feature extraction). See the [detailed section in our submission guidelines](https://www.nature.com/sdata/publish/submission-guidelines#sec-5) for advice on writing a transparent and reproducible methods section. Related methods should be grouped under corresponding subheadings where possible, and methods should be described in enough detail to allow other researchers to interpret and repeat, if required, the full study. Specific data outputs should be explicitly referenced via data citation (see Data Records and Citing Data, below).

Authors should cite previous descriptions of the methods under use, but ideally the method descriptions should be complete enough for others to understand and reproduce the methods and processing steps without referring to associated publications. There is no limit to the length of the Methods section.

Data collection & cleansing

School of Engineering & Physics automatic weather station is equipped with reliable, high-precision instruments and its setup is similar to FMS standards. The weather data described in this paper were compiled from 1 April 2013 to 31 December 2019. The data for January, February and March 2013 were unavailable as no data recording were done due to upgradation of the weather instruments. The weather instruments used for gathering weather data is shown in Table 1. All the variables were measured using automated instruments except vapor pressure which was computed from observed values of air temperature and relative humidity using the formula, as presented below

(Formula to be provided by Dr. Ajal)

Campbell Scientific® CR1000 data logger running on CRBasic® code currently being used as data acquisition system to capture weather data. At present, the logger stores data in the following specified output data tables; 10-minute, hourly and daily. These tables are populated for 19 days whereupon the oldest data is then overwritten by the latest. The Campbell Scientific® SC115 flash memory drive is used to retrieve data from the data logger and transferred to a designated computer in Physics laboratory fortnightly. The weather data from data logger is extracted in Data (.dat) files, it is then imported to ­Excel® and converted to spreadsheet (.xlsx) format, and renamed as appropriate. A senior Physics technician is in charge of the daily monitoring of the weather station.

All the data acquired from SEP were cross-checked for all those records that were missing for a time interval. Missing values are mainly due to hardware (weather instruments or data logger) fault, power supply issues, late data retrieval (where data table with old data values gets replaced after 19 days). It was noted that if an observed value is missing for a given time period, it was missing for all the measured variables (how to write this properly). All such records were identified and declared missing to ensure integrity of the data. The descriptive analysis of the weather data is shown in Table 2a. Each variable should consist of 2466 observations; however, only 93.1 percent of the data were available, with 169 (6.9 %) missing values for each variable. There were no missing values for all the variables from 1 August 2015 to 31 August 2017 while the highest percentage of missing data was recorded in 2013 with the greatest missing data period of 27 days in the same year for the month of May (Table 2b and Figure 3a). The time series plot of each variable over the record period is shown in Figure 3c.

|  |  |
| --- | --- |
| Variable | Instrument Used & Product Code |
| Air Temperature | *Vaisala*® HMP155A temp. & humidity probe  *Vaisala*® HMP155A temp. & humidity probe |
| Relative Humidity |
| Rainfall | *Hydrological Services*® TB3 Tipping Bucket Rain-gauge with 0.5 mm plastic bucket calibration |
| Solar Radiation | *LiCor*® LI200X pyranometer |
| Wind Speed | *Vector Instruments*® A101M pulse output anemometer |
| Wind Direction | *Vector Instruments*® W200PL potentiometer wind vane |
| Barometric Pressure | *Vaisala*® PTB110 barometric pressure transducer |
| Soil Moisture | *Aquaflex*® SI.99 soil moisture sensor  *Aquaflex*® SI.99 soil moisture sensor |
| Soil Temperature |

Table 1. Instrument used to collect the weather data with its product code.

**a**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | N | Missing | | Min. | Max. | Mean | Std.  Deviation |
| N | % |
| Air Temperature | 2297 | 169 | 6.9 | 19.46 | 30.11 | 25.31 | 1.68 |
| Relative Humidity | 2297 | 169 | 6.9 | 55.09 | 99.70 | 83.45 | 8.25 |
| Vapor Pressure | 2297 | 169 | 6.9 | 1.45 | 3.59 | 2.70 | 0.40 |
| Barometric Pressure | 2297 | 169 | 6.9 | 99.33 | 102.11 | 101.17 | 0.37 |
| Solar Radiation | 2297 | 169 | 6.9 | 2.31 | 285.20 | 129.44 | 63.54 |
| Horizontal Wind speed | 2297 | 169 | 6.9 | 0.05 | 4.42 | 1.38 | 0.69 |
| Wind Direction | 2297 | 169 | 6.9 | 31.94 | 231.41 | 144.74 | 35.81 |
| Soil Moisture | 2297 | 169 | 6.9 | 17.09 | 62.03 | 43.93 | 10.93 |
| Soil Temperature | 2297 | 169 | 6.9 | 18.37 | 27.94 | 23.45 | 1.67 |
| Total Rainfall | 2297 | 169 | 6.9 | 0 | 230 | 6.44 | 17.29 |

**b**

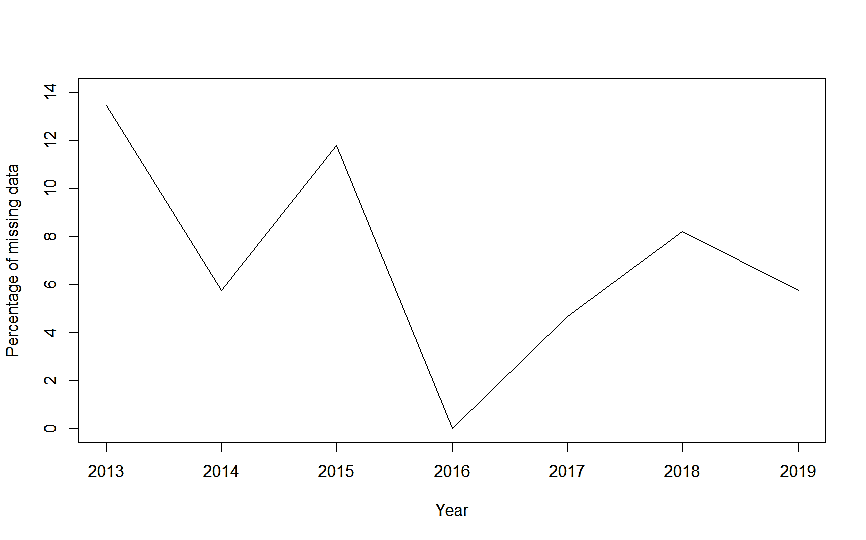
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | Missing values/month | | | | | | | | | | | | Missing values/yr  (%) |
| J | F | M | A | M | J | J | A | S | O | N | D |
| 2013 | NA | | | 3 | 27 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 37 (13.5%) |
| 2014 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 3 | 7 | 0 | 21 (5.8%) |
| 2015 | 0 | 0 | 1 | 21 | 0 | 13 | 8 | 0 | 0 | 0 | 0 | 0 | 43 (11.8%) |
| 2016 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 (0%) |
| 2017 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 15 | 17 (4.7%) |
| 2018 | 0 | 0 | 0 | 2 | 0 | 6 | 0 | 0 | 2 | 19 | 1 | 0 | 30 (8.2%) |
| 2019 | 0 | 0 | 3 | 0 | 0 | 1 | 0 | 2 | 15 | 0 | 0 | 0 | 21 (5.8%) |
|  | | | | | | | | | | | | | 169 (6.9%) |

**c (should this be added)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | T | Min. | Max. | Mean | Std.Deviation |
| Air Temperature | 2466 | 19.46 | 30.11 | 25.30 | 1.65 |
| Relative Humidity | 2466 | 55.09 | 99.70 | 83.50 | 8.13 |
| Vapor Pressure | 2466 | 1.45 | 3.59 | 2.70 | 0.39 |
| Barometric Pressure | 2466 | 99.33 | 102.11 | 101.17 | 0.37 |
| Solar Radiation | 2466 | 2.31 | 285.20 | 128.06 | 62.70 |
| Horizontal Wind speed | 2466 | 0.05 | 4.42 | 1.40 | 0.69 |
| Wind Direction | 2466 | 31.94 | 231.41 | 145.65 | 35.73 |
| Soil Moisture | 2466 | 17.09 | 62.03 | 44.01 | 10.88 |
| Soil Temperature | 2466 | 18.37 | 27.94 | 23.43 | 1.64 |
| Total Rainfall | 2466 | 0 | 230 | 6.42 | 16.89 |

Table . Missing data analysis. (a) shows descriptive analysis of the weather dataset. Where N is the total number of values available over the record period; n is the total number of missing values over the record period; min. and max. are the minimum and maximum values of the variable respectively. (b) shows the further analysis of the missing data by month and year through time. (c) shows the descriptive analysis of the completed weather dataset. Where T is the total number of values after imputation over the record period.

**a**



**b**

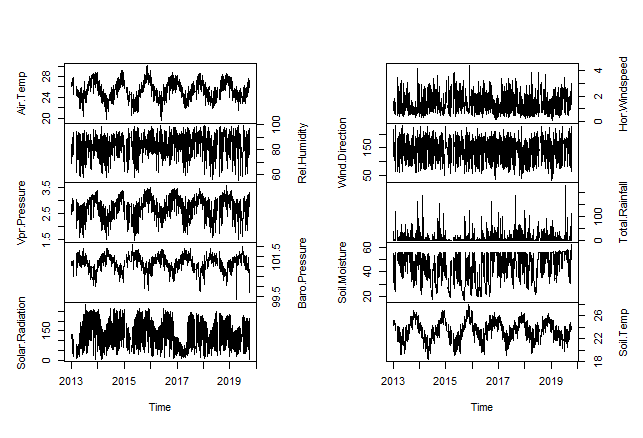
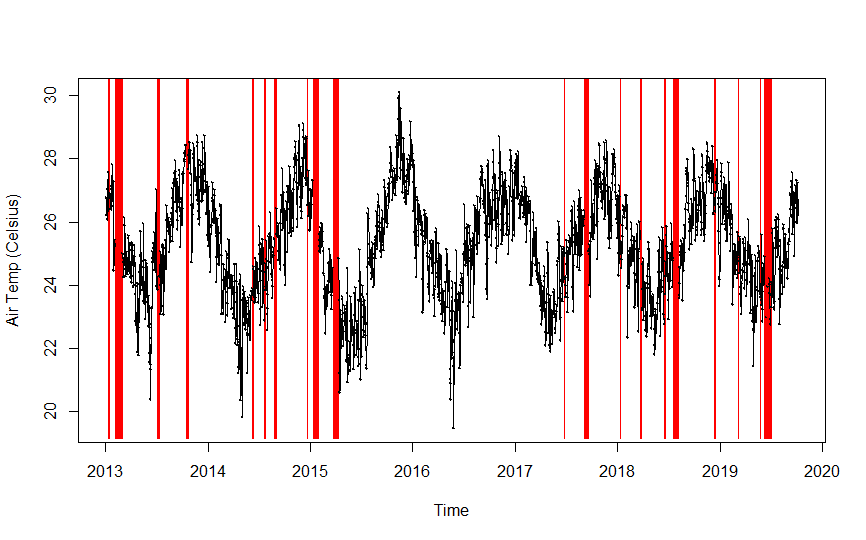


Figure . Missing data and time series plot. (a) shows the histogram of missing data for each variable over the entire record period, (b) shows the percentage of missing data through time, (c) shows the time series plot of the weather data with missing values.

Alternative Time series display (sequence of NA’s in Red)- which one is better?

****

Missing data imputation procedure

Missing data is a very common meteorological problem which affects the accuracy of the results that science studies can provide. Imputation was performed primarily to obtain a serially complete time series record for each variable. Typically, various methods are applied in time and space to interpolate and infill (or impute) weather data  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
14,15,  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
16. Meteorological variables such as rainfall and temperature often exhibit strong seasonal patterns. The forecast package in R software that has one advanced feature for time series with missing data imputation: na.interp() was used in this paper to impute missing data for each variable. na.interp function by default, “uses linear interpolation for non-seasonal series. For seasonal series, a robust STL decomposition is first computed. Then a linear interpolation is applied to the seasonally adjusted data, and the seasonal component is added back”  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
17. STL is an acronym for “Seasonal and Trend decomposition using Loess”, while Loess is a method for estimating nonlinear relationships. STL is a versatile and robust method for decomposing time series and was developed by Cleveland et al.  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
18. In particular, this function is supposed to be a good fit, where one expects a clear and strong seasonality in the time series data.

Extra information- should this be added?

STL has several advantages over the classical, SEATS and X11 decomposition methods: Unlike SEATS and X11, STL will handle any type of seasonality, not only monthly and quarterly data. The seasonal component is allowed to change over time, and the rate of change can be controlled by the user. The smoothness of the trend-cycle can also be controlled by the user. It can be robust to outliers (i.e., the user can specify a robust decomposition), so that occasional unusual observations will not affect the estimates of the trend-cycle and seasonal components. They will, however, affect the remainder component. On the other hand, STL has some disadvantages. In particular, it does not handle trading day or calendar variation automatically, and it only provides facilities for additive decompositions. It is possible to obtain a multiplicative decomposition by first taking logs of the data, then back-transforming the components. Decompositions between additive and multiplicative can be obtained using a Box-Cox transformation of the data with 0 < λ < 1. A λ = 1 value of corresponds to the multiplicative decomposition while λ = 0 is equivalent to an additive decomposition.

### Data Records

The Data Records section should be used to explain each data record associated with this work, including the repository where this information is stored, and to provide an overview of the data files and their formats. Each external data record should be cited as described below. A data citation should also be placed in the subsection of the Methods containing the data-collection or analytical procedure(s) used to derive the corresponding record.

Tables should be used to support the data records, and should clearly indicate the samples and subjects (study inputs), their provenance, and the experimental manipulations performed on each. They should also specify the data output resulting from each data-collection or analytical step, should these form part of the archived record.

The datasets released in this manuscript are included in Table 3. Individual datasets are available as CSV (Comma delimited) format (.csv) and can be downloaded from the online archive in figshare (Data Citation 1). These datasets include the original weather data (across the entire data, the missing values are identified with a blank) that was provided by the School of Engineering & Physics, USP, Laucala Campus, Suva, Fiji and the final completed time series weather data. The datasets include daily average air temperature, relative humidity, vapour pressure, barometric pressure, solar radiation, horizontal wind speed, wind direction, soil moisture, soil temperature and total daily rainfall for the city of Suva from 1 April 2013 to 31 December 2019.

|  |  |  |
| --- | --- | --- |
| Data Type | SI units | Data File |
| Air Temperature | Celsius | Weather\_Data\_Not\_Filled.csv  Final\_Weather\_Data\_Filled.csv |
| Relative Humidity | % RH |
| Vapor Pressure | kPa |
| Barometric Pressure | kPa |
| Solar Radiation | W m-2 |
| Horizontal Wind Speed | m/s |
| Wind Direction | Degrees |
| Soil Moisture | % |
| Soil Temperature | Celsius |
| Total Rainfall | mm |

Table . Datasets description accompanying this manuscript. Where: Data Type is the variable included in the dataset; SI Units is scientific units of the variable expressed in the dataset; Data File is the name of the dataset file.

### Technical Validation (Incomplete- Need assistance)

The Technical Validation section should present any experiments or analyses that are needed to support the technical quality of the dataset. This section may be supported by figures and tables, as needed. *This is a required section*; authors must provide information to justify the reliability of their data.

Possible content **may include:**

* experiments that support or validate the data-collection procedure(s) (e.g. negative controls, or an analysis of standards to confirm measurement linearity)
* statistical analyses of experimental error and variation
* phenotypic or genotypic assessments of biological samples (e.g. confirming disease status, cell line identity, or the success of perturbations)
* general discussions of any procedures used to ensure reliable and unbiased data production, such as blinding and randomization, sample tracking systems, etc.
* any other information needed for assessment of technical rigour by the referees

Generally, this **should not include:**

* follow-up experiments aimed at testing or supporting an interpretation of the data
* **statistical hypothesis testing** (e.g. tests of statistical significance, identifying differentially expressed genes, trend analysis, etc.)
* exploratory computational analyses like clustering and annotation enrichment (e.g. GO analysis).

**With reference to Scientific Data paper titled: Compilation of climate data from heterogeneous networks across the Hawaiian Islands (attached), for better validation, I need to select longest period of time series data with no missing values (in our case it is 1 August 2015 to 31 August 2017), and then I need to randomly create missing values and then apply the missing values technique and compare the imputed value with the original by calculating mean bias error, mean absolute error and root mean square error- comment on this?**

**Hypothesis test: Comparison between observed and imputed data for each variable- this may not be accepted**

Kolmogorov–Simonov (K–S) test results.

|  |  |  |
| --- | --- | --- |
| Variable | Observed Data vs. Completed Data | |
|  |  |
| Air Temperature | 0.0125 | 0.9922 |
| Relative Humidity | 0.0114 | 0.9980 |
| Vapor Pressure | 0.0105 | 0.9994 |
| Barometric Pressure | 0.0076 | 0.9999 |
| Solar Radiation | 0.0152 | 0.9464 |
| Horizontal Wind speed | 0.0133 | 0.9845 |
| Wind Direction | 0.0113 | 0.9981 |
| Total Rainfall | 0.0145 | 0.9632 |
| Soil Moisture | 0.0098 | 0.9999 |
| Soil Temperature | 0.0156 | 0.9354 |

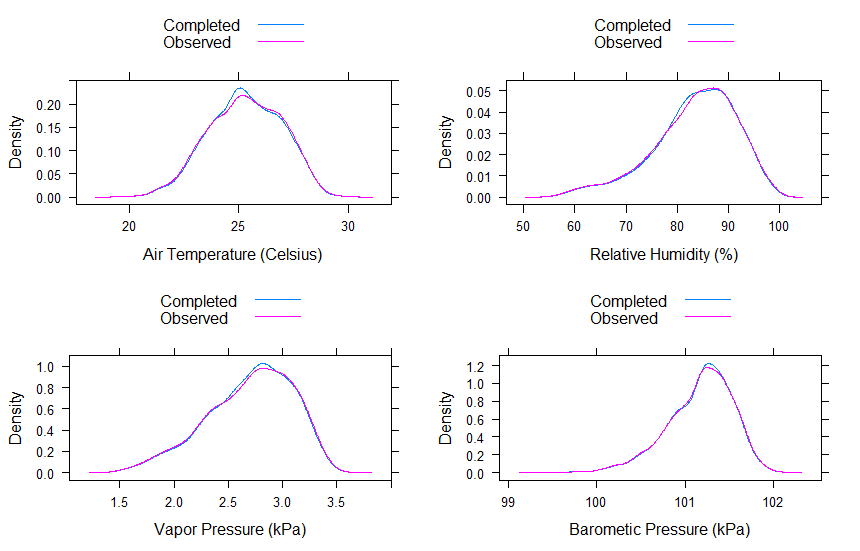
The empirical distributions of the observed and the imputed using the Kolmogorov-Smirnov test for each variable can be numerically compared, raising the flag when we find significant differences. The Kolmogorov–Simonov (K–S) test null hypothesis that suggests that the two samples were drawn from the same distribution and the alternative hypothesis dictates otherwise. If the p-value is greater than α= 0.05, the null hypothesis is confirmed; otherwise, the alternative hypothesis is accepted. The D statistic is the absolute maximum distance between the cumulative distribution functions of the two samples. The closer this number is to 0 the more likely it is that the two samples were drawn from the same distribution. Results from table reveals that the empirical distribution was not statistically different (p > 0.05), and hence the observed and imputed data do not differ.

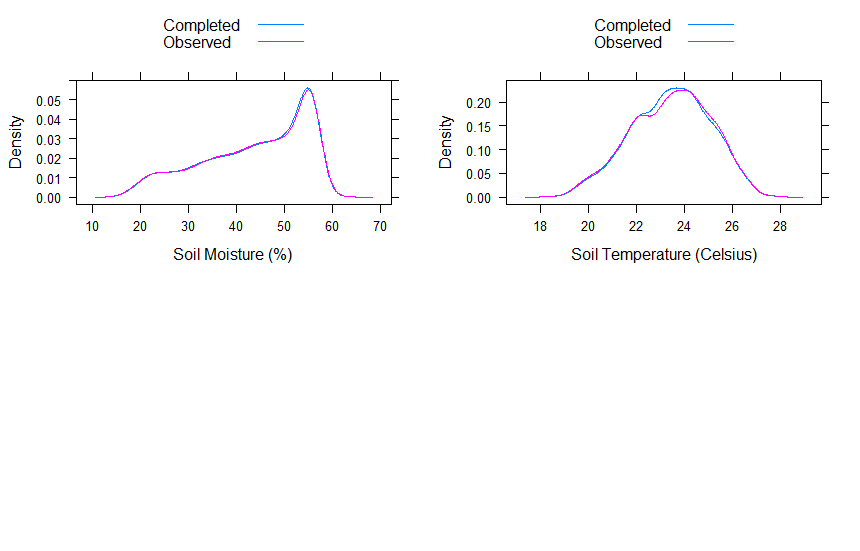
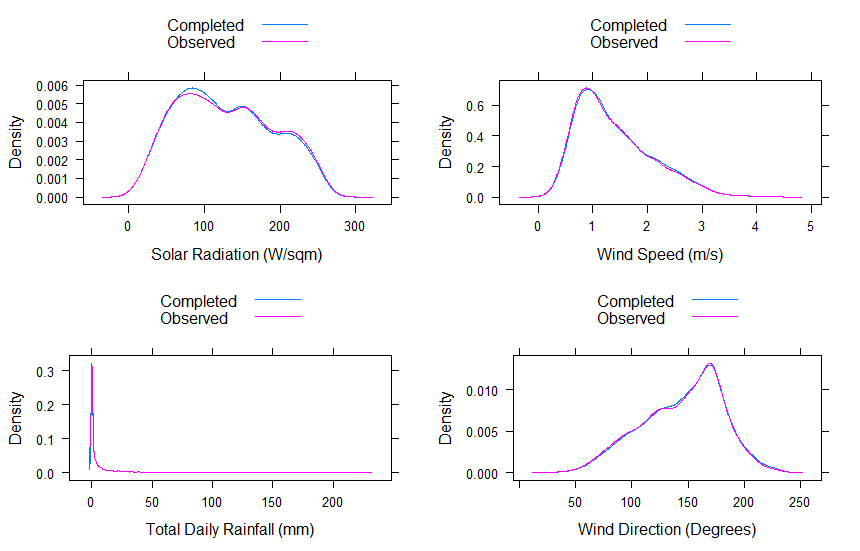
**Visual Inspection- this can be added**

**Paper reference:**19

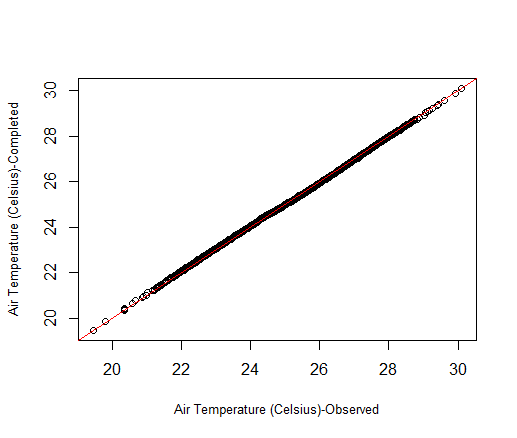
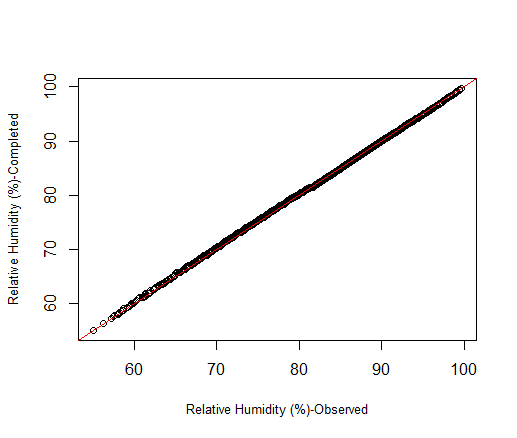
This can also be visually examined using empirical density plots. Figure below shows two plot types for comparing observed and imputed data for each variable. The kernel density and the quantile–quantile plots for each variable were examined to further support the K-S test above. The kernel density plot shows that there is no significant difference between the distributions of observed and imputed values for each variable. Similarly, for all the variables, the points remain close to the diagonal in the quantile-quantile plot and the hypothesis “the two samples come from the same distribution” is also confirmed.

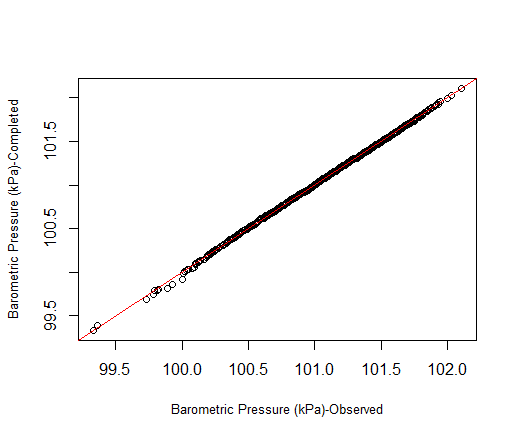
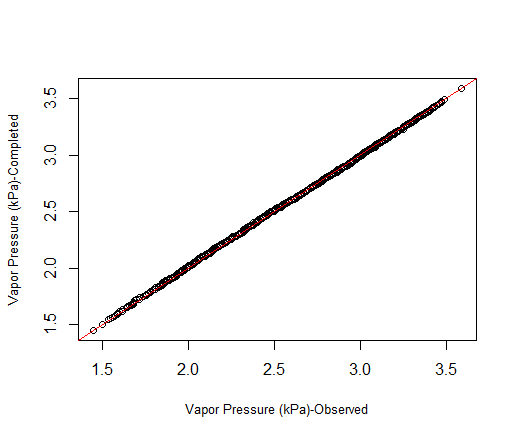
The kernel density plots

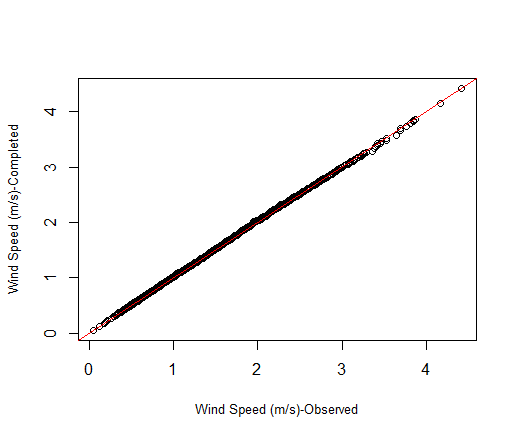
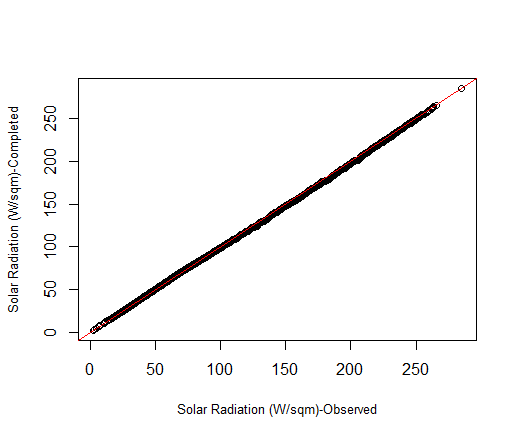


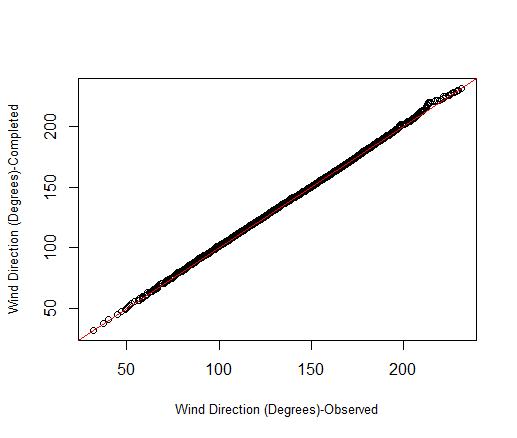
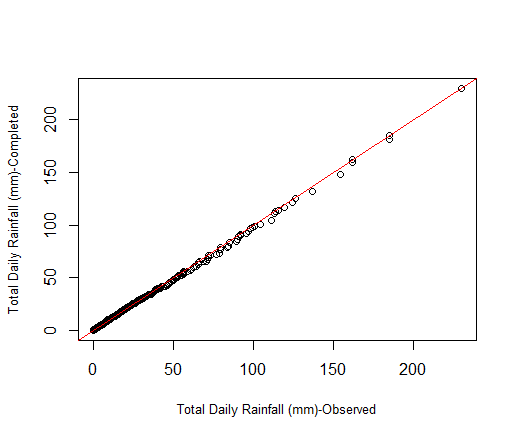


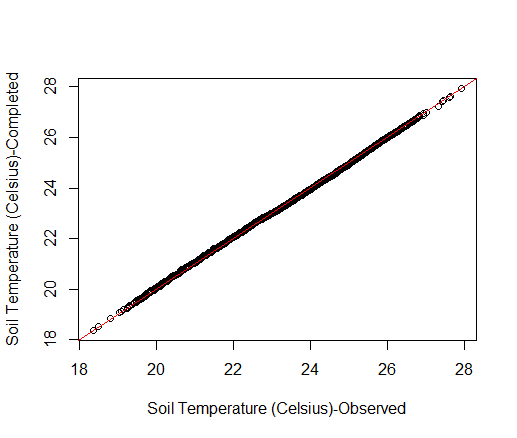
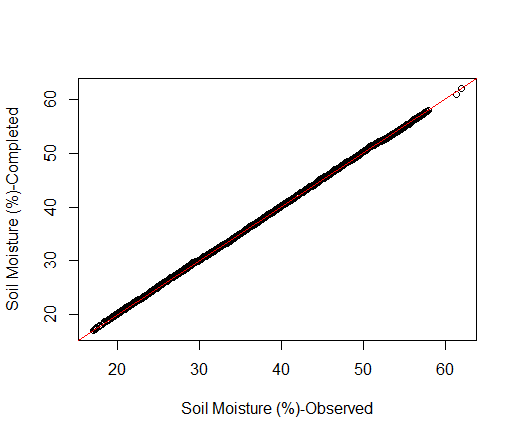
The quantile–quantile plots









### ~~Usage Notes~~

*~~This section is optional~~*

~~The Usage Notes should contain brief instructions to assist other researchers with reuse of the data. This may include discussion of software packages that are suitable for analysing the assay data files, suggested downstream processing steps (e.g. normalization, etc.), or tips for integrating or comparing the data records with other datasets. Authors are encouraged to provide code, programs or data-processing workflows if they may help others understand or use the data. Please see our~~ [~~code availability policy~~](http://www.nature.com/sdata/policies/editorial-and-publishing-policies#code-avail) ~~for advice on supplying custom code alongside Data Descriptor manuscripts.~~

~~For studies involving privacy or safety controls on public access to the data, this section should describe in detail these controls, including how authors can apply to access the data, what criteria will be used to determine who may access the data, and any limitations on data use.~~

### Code Availability

The missing data imputation was performed in a standard version of the R software, R.3.6.2, using the forecast package (this package is publicly available).

### Acknowledgements

The Acknowledgements should contain text acknowledging non-author contributors. Acknowledgements should be brief, and should not include thanks to anonymous referees and editors or effusive comments. Grant or contribution numbers may be acknowledged.

We thank Jasneel Chandra and Varanisese Vuniyayawa (Fiji Meterological Services). We also thank The University of the South Pacific for Staff Masters Sponsorship awarded to the first author as this manuscript arises from his master’s thesis (check?) Senior technician (check?)

### Author contributions

Each author’s contribution to the work should be described briefly, on a separate line, in the Author Contributions section.

All authors wrote the data descriptor; R. Chand did data cleaning, gap- infilling process and analysis of the data validation, and led the manuscript writing;

R. Chandra

A. K

D. R

(To be completed: Each author’s contribution to the work should be described briefly, on a separate line, in the Author Contributions section)

### Competing interests

A competing interests statement is required for all papers accepted by and published in *Scientific Data*. If there is no conflict of interest, a statement declaring this must still be included in the manuscript.

The authors declare no competing financial interests.

### References

1. Kumar, V. V., Deo, R. C. & Ramachandran, V. Total rain accumulation and rain-rate analysis for small tropical Pacific islands: a case study of Suva, Fiji. *Atmospheric Science Letters* **7**, 53–58 (2006).

2. Onwubolu, G. C. *et al.* Self-organizing data mining for weather forecasting. *IADIS European Conf. Data Mining* 81–88 (2007).

3. Ramirez, M. C. V., Campos Velho, H. F. de & Ferreira, N. J. Artificial neural network technique for rainfall forecasting applied to the Sao Paulo region. *Journal of hydrology* **301**, 146–162 (2005).

4. Hernández, E., Sanchez-Anguix, V., Julian, V., Palanca, J. & Duque, N. Rainfall prediction: A deep learning approach. *International Conference on Hybrid Artificial Intelligence Systems* 151–162 (2016).

5. Wu, C., Chau, K. & Fan, C. Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques. *Journal of Hydrology* **389**, 146–167 (2010).

6. Hong, W.-C. & Pai, P.-F. Potential assessment of the support vector regression technique in rainfall forecasting. *Water Resources Management* **21**, 495–513 (2007).

7. Hung, N. Q., Babel, M. S., Weesakul, S. & Tripathi, N. An artificial neural network model for rainfall forecasting in Bangkok, Thailand. *Hydrology & Earth System Sciences* **13**, (2009).

8. Júnior, J. M. P. & Barreto, G. A. Multistep-ahead prediction of rainfall precipitation using the NARX network. *ESTSP’08* 87 (2008).

9. Yen, M.-H., Liu, D.-W., Hsin, Y.-C., Lin, C.-E. & Chen, C.-C. Application of the deep learning for the prediction of rainfall in Southern Taiwan. *Scientific reports* **9**, 1–9 (2019).

10. Sumi, S. M., Zaman, M. F. & Hirose, H. A rainfall forecasting method using machine learning models and its application to the Fukuoka city case. *International Journal of Applied Mathematics and Computer Science* **22**, 841–854 (2012).

11. Remesan, R., Shamim, M. & Han, D. Model data selection using gamma test for daily solar radiation estimation. *Hydrological processes* **22**, 4301–4309 (2008).

12. Deo, R. C. On meteorological droughts in tropical Pacific Islands: time-series analysis of observed rainfall using Fiji as a case study. *Meteorological Applications* **18**, 171–180 (2011).

13. Bushra, P. *et al.* Analyzing trend and forecasting of rainfall changes in India using non-parametrical and machine learning approaches. *Scientific Reports (Nature Publisher Group)* **10**, (2020).

14. Gad, I., Manjunatha, B. R. & Doreswamy Performance evaluation of predictive models for missing data imputation in weather data. *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* 1327–1334 (2017).

15. Kashani, M. H. & Dinpashoh, Y. Evaluation of efficiency of different estimation methods for missing climatological data. *Stochastic Environmental Research and Risk Assessment* **26**, 59–71 (2012).

16. Plouffe, C. C., Robertson, C. & Chandrapala, L. Comparing interpolation techniques for monthly rainfall mapping using multiple evaluation criteria and auxiliary data sources: A case study of Sri Lanka. *Environmental Modelling & Software* **67**, 57–71 (2015).

17. Hyndman, R. J. *et al.* Package ’forecast. *Online] https://cran. r-project. org/web/packages/forecast/forecast. pdf* (2020).

18. Cleveland, R. B., Cleveland, W. S., McRae, J. E. & Terpenning, I. STL: A seasonal-trend decomposition. *Journal of official statistics* **6**, 3–73 (1990).

19. Moritz, S., Sardá, A., Bartz-Beielstein, T., Zaefferer, M. & Stork, J. Comparison of different methods for univariate time series imputation in R. *arXiv preprint arXiv:1510.03924* (2015).

### Citing Data

1. Chand, R. et al. Figshare (link)

note that I have already created an account with figshare and will be releasing

the data via that repository.

**Extra Information**

### Figures

Figure images should be provided as separate files and should be referred to using a consistent numbering scheme through the entire Data Descriptor. In most cases, a Data Descriptor should not contain more than three figures, but more may be allowed when needed. We discourage the inclusion of figures in the Supplementary Information – all key figures should be included here in the main Figure section.

For initial submissions, authors may choose to supply a single PDF with embedded figures.

Authors are encouraged to consider creating a figure that outlines the experimental workflow(s) used to generate and analyse the data output(s).

### Figure Legends

### Figure legends begin with a brief title sentence summarizing the purpose of the figure as a whole, and continue with a short description of what is shown in each panel and an explanation of any symbols used. Legends must total no more than 350 words, and may contain literature references. The first sentence of the legend will be used as the title for the figure. It should contain no references of any kind, including to specific figure panels, bibliographic citations or references to other figures or panels.

### Tables

Authors are encouraged to provide one or more tables that provide basic information on the main ‘inputs’ to the study (e.g. samples, participants, or information sources) and the main data outputs of the study; also see the additional information on providing metadata on page 6. Tables in the manuscript should generally not be used to present primary data (i.e. measurements). Tables containing primary data should be submitted to an appropriate data repository.

Authors may provide tables within the Word document or as separate files (tab-delimited text or Excel files). Legends, where needed, should be included in the Word document. Generally, a Data Descriptor should have fewer than ten tables, but more may be allowed when needed. Tables may be of any size, but only tables that fit onto a single printed page will be included in the PDF version of the article (up to a maximum of three).

Due to typesetting constraints, tables that do not fit onto a single A4 page cannot be included in the PDF version of the article and will be made available in the online version only. Any such tables must be labelled in the text as ‘Online-only’ tables and numbered separately from the main table list e.g. ‘Table 1, Table 2, Online-only Table 1’ etc.

### References

Bibliographic information for any works cited in the above sections, using the standard Nature referencing style.

In line with emerging [industry-wide standards for data citation](https://www.nature.com/articles/sdata2018259), references to all datasets described or used in the manuscript should be cited in the text with a superscript number and listed in the ‘References’ section in the same manner as a conventional literature reference. See ‘Citing Data’ below for further details.

Additional Formatting Information

**Referencing Figures, Tables, and other content**

**The Word document may reference Figures (e.g. Fig. 1), Tables (e.g. Table 1), online-only tables (e.g. Online-only Table 1) and Supplementary Information (e.g. Supplementary Table 1, or Supplementary File 2, etc.). When information from metadata documents must be referred to, it should also be included in the main manuscript as Tables, and formatted in a way that suits human readability. To refer to the ISA-Tab metadata records within the manuscript, use the phrase “see associated Metadata Record”.**

**Citation format**

All references should be numbered sequentially, first throughout the text, then in tables, followed by figures and, finally, boxes; that is, references that only appear in tables, figures or boxes should be last in the reference list. Only one publication is given for each number. Only papers that have been published or accepted by a named publication or recognized preprint server should be in the numbered list; preprints of accepted papers in the reference list should be submitted with the manuscript. Published conference abstracts, numbered patents, and archived code with an assigned DOI may be included in the reference list. Grant details and acknowledgments are not permitted as numbered references. Footnotes are not used.

Scientific Data uses standard Nature referencing style. All authors should be included in reference lists unless there are six or more, in which case only the first author should be given, followed by ‘et al.’. Authors should be listed last name first, followed by a comma and initials (followed by full stops, '.') of given names. Article titles should be in Roman text; only the first word of the title should have an initial capital and the title should be written exactly as it appears in the work cited, ending with a full stop. Book titles should be given in italics and all words in the title should have initial capitals. Journal names are italicized and abbreviated (with full stops) according to common usage. Volume numbers and the subsequent comma appear in bold. The full page range should be given where appropriate. See the examples below:

**Journal Article**:

1. Schott, D. H., Collins, R. N. & Bretscher, A. Secretory vesicle transport velocity in living cells depends on the myosin V lever arm length. *J. Cell Biol*. **156**, 35‐39 (2002).

**Book** ‐ Book titles should be given in italics and all words in the title should have initial capitals:

1. Hogan, B. *Manipulating The Mouse Embryo: A Laboratory Manual* 2nd edn (Cold Spring Harbor Laboratory Press, 1994)

**Publicly available preprint:**

1. Babichev, S. A., Ries, J. & Lvovsky, A. I. Quantum scissors: teleportation of single-mode optical states by means of nonlocal single photon. Preprint at http://arXiv.org/quant-ph/0208066 (2002).

**Code:**

1. Gallotti, R. & Barthélemy, M. Source code for: The multilayer temporal network of public transport in Great Britain. *Figshare* https://dx.doi.org/10.6084/m9.figshare.1249862.v1 (2014).

**Online material** ‐ Stable documents hosted on the web may be cited in the main reference list, using the format below. Websites or dynamic web resources should be cited by embedding the URL in the main article text:

1. Manaster, J. Sloth squeak. *Scientific American Blog Network* http://blogs.scientificamerican.com/psi-vid/2014/04/09/sloth-squeak (2014).

**Technical or government report:**

1. Akutsu, T. *Total Heart Replacement Device.* Report No. NIH-NHLI-69 2185-4 (National Institutes of Health, 1974).

## Citing Data

In line with emerging [industry-wide standards for data citation](https://www.nature.com/articles/sdata2018259), references to all datasets described or used in the manuscript should be cited in the text with a superscript number and listed in the ‘References’ section in the same manner as a conventional literature reference.

An author list (formatted as above) and title for the dataset should be included in the data citation, and should reflect the author(s) and dataset title recorded at the repository. If author or title is not recorded by the repository, these should not be included in the data citation. The name of the data-hosting repository, URL to the dataset and year the data were made available are required for all data citations. For DOI-based (e.g. figshare or Dryad) repositories the DOI URL should be used. For repositories using accessions (e.g. SRA or GEO) an [identifiers.org](https://identifiers.org/) URL should be used where available. For first submissions, authors may choose to include just the accession number. Scientific Data staff will provide further guidance after peer-review. Please refer to the following examples of data citation for guidance:

1. Zhang, Q-L., Chen, J-Y., Lin, L-B., Wang, F., Guo, J., Deng, X-Y. Characterization of ladybird Henosepilachna vigintioctopunctata transcriptomes across various life stages. figshare <https://doi.org/10.6084/m9.figshare.c.4064768.v3> (2018).
2. NCBI Sequence Read Archive <http://identifiers.org/ncbi/insdc.sra:SRP121625> (2017).
3. Barbosa, P., Usie, A. and Ramos, A. M. Quercus suber isolate HL8, whole genome shotgun sequencing project. GenBank<http://identifiers.org/ncbi/insdc:PKMF00000000> (2018).
4. DNA Data Bank of Japan <http://trace.ddbj.nig.ac.jp/DRASearch/submission?acc=DRA004814> (2016).

**Depositing your data to an appropriate repository**

Your *Scientific Data* manuscript will not be sent to review unless the dataset(s) described therein have been deposited in an appropriate public repository ([please see our list of recommended repositories](http://www.nature.com/sdata/policies/repositories)). Should a specific repository not be available for your field or data-type, or should the repository of your choice not permit confidential peer-review, you may upload your data to one of our [recommended generalist repositories](https://www.nature.com/sdata/policies/repositories#general). Integrated submission systems are available for both figshare and Dryad.