Brief communication: SWM: Stochastic Weather Model for precipitation-related hazard assessments using ERA5-land data

Melody G. Whitehead[1], Mark S. Bebbington[1]

[1]Volcanic Risk Solutions, School of Agriculture and Environment, Massey University, Palmerston North, 4472, New Zealand

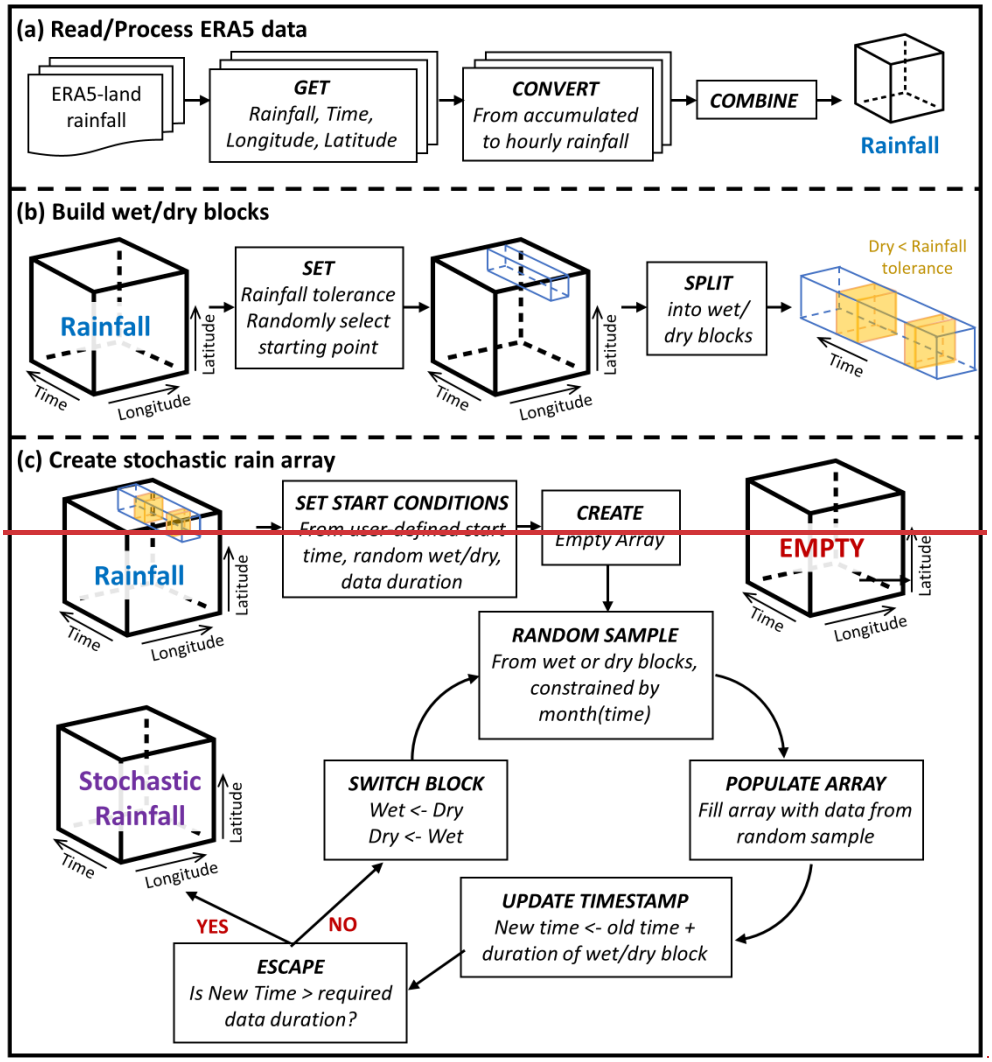
Correspondence to: Melody G. Whitehead (m.whitehead@massey.ac.nz)

Abstract

Long-term multi-hazard and risk assessments are produced by combining many hazard-model simulations, each using slightly different set of inputs to cover the uncertainty space. While most input parameters for these models are relatively well-constrained, atmospheric parameters remain problematic unless working on very short-time scales (hours to days). Precipitation is a key trigger for many natural hazards including floods, landslides, and lahars. This work presents a stochastic catchment-scale weather model that takes openly available ERA5-land data, and produces long-term, spatially varying precipitation data that mimics the statistical dimensions of real-data. This allows precipitation to be robustly included in hazard-model simulations. A working example is provided using 1981-2020 ERA5-land data for the Rangitāiki-Tarawera catchment, Bay of Plenty, New Zealand.

1 Introduction

Natural hazard and risk assessments are probabilistic by necessity. They must incorporate the intrinsic variability of natural systems, and the large number of unknown (but often data constrained) input parameters. To produce such assessments, many model simulations are run by sampling from a distribution for these parameters. The outputs are then combined (often overlain in a spatial context), to calculate hazard likelihoods across an area, and/or to produce risk maps, key for communicating hazards (Thompson et al., 2015; Hyman et al. 2019). The spatial extent of such hazards is key to such assessments, as is a robust approach to simulation design. Precipitation is causally linked to many natural hazards including floods, landslides, and lahars (Gill and Malamud, 2016). While several stochastic weather models exist in published literature, they either require detailed local rainfall information – which is rare over long timescales (Zhao et al. 2019; Muñoz-Sabater et al., 2021), or they are run for a single spatial reference point – which is insufficient for many hazard models (e.g., Floods: Arnaud et al., 2002; Landslides: Gao et al., 2017). The model provided here uses openly available ERA5-land data (Muñoz-Sabater, 2019) and produces realistic (statistically similar) precipitation patterns to improve the sampling strategy of atmospheric properties and support robust hazard assessments. This brief correspondence first presents algorithm construction, and then an example application using the Rangitāiki-Tarawera catchment, Bay of Plenty, New Zealand. All code is in R (R Core Team, 2021), and freely available.


(a) Read/Process ERA5 data
ERA5-land rainfall
GET
Rainfall, Time, Longitude, Latitude
CONVERT
From accumulated to hourly rainfall
COMBINE
Rainfall
(b) Build wet/dry blocks
Rainfall
Latitude
Time
Longitude
SET
Rainfall tolerance
Randomly select starting point
SPLIT
into wet/ dry blocks
Dry < Rainfall tolerance
(c) Create stochastic rain array
Rainfall
SET START CONDITIONS
From user-defined start time, random wet/dry, data duration
CREATE
Empty Array
EMPTY
RANDOM SAMPLE
From wet or dry blocks, constrained by month(time)
Stochastic Rainfall
SWITCH BLOCK
Wet <- Dry
Dry <- Wet
POPULATE ARRAY
Fill array with data from random sample
UPDATE TIMESTAMP
New time <- old time + duration of wet/dry block
YES
NO
ESCAPE
Is New Time > required data duration?

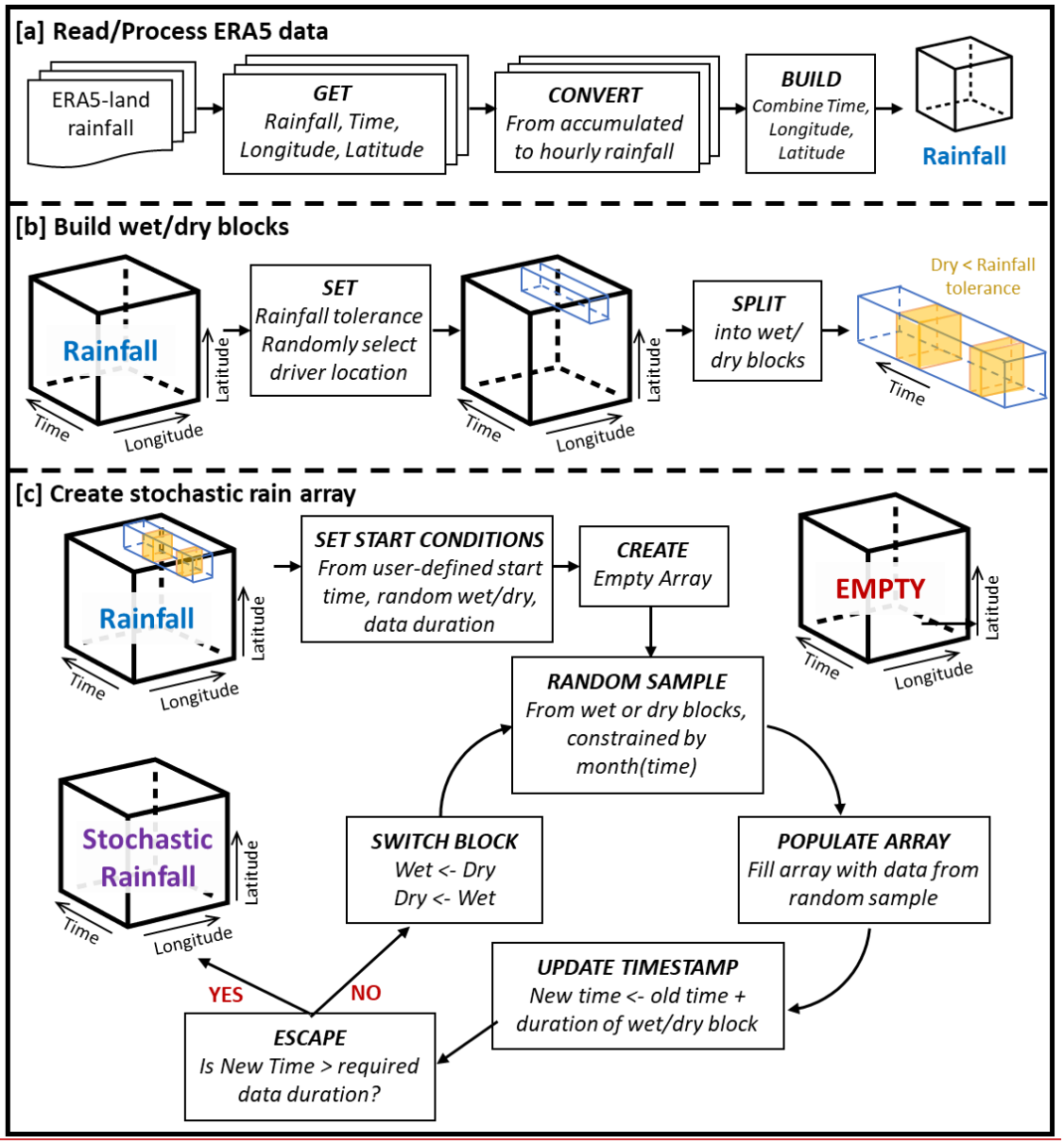


Figure 1: SWM algorithm flow diagram. (a) Read/Process ERA5 data, (b) Build wet/dry blocks, (c) Create stochastic array.

2 Algorithm construction

The stochastic weather model (SWM) comprises three steps: data conversion, block construction, and stochastic weather generation. Due to the relative simplicity of the model, and exploiting some coding efficiencies in the R package *dplyr* (Wickham et al., 2023), 10 years of hourly data can be generated at points on a 10 by 10 grid on a standard desktop computer in under five seconds. Before running SWM, data must be downloaded from ERA5-land data in netCDF format. In ERA5-

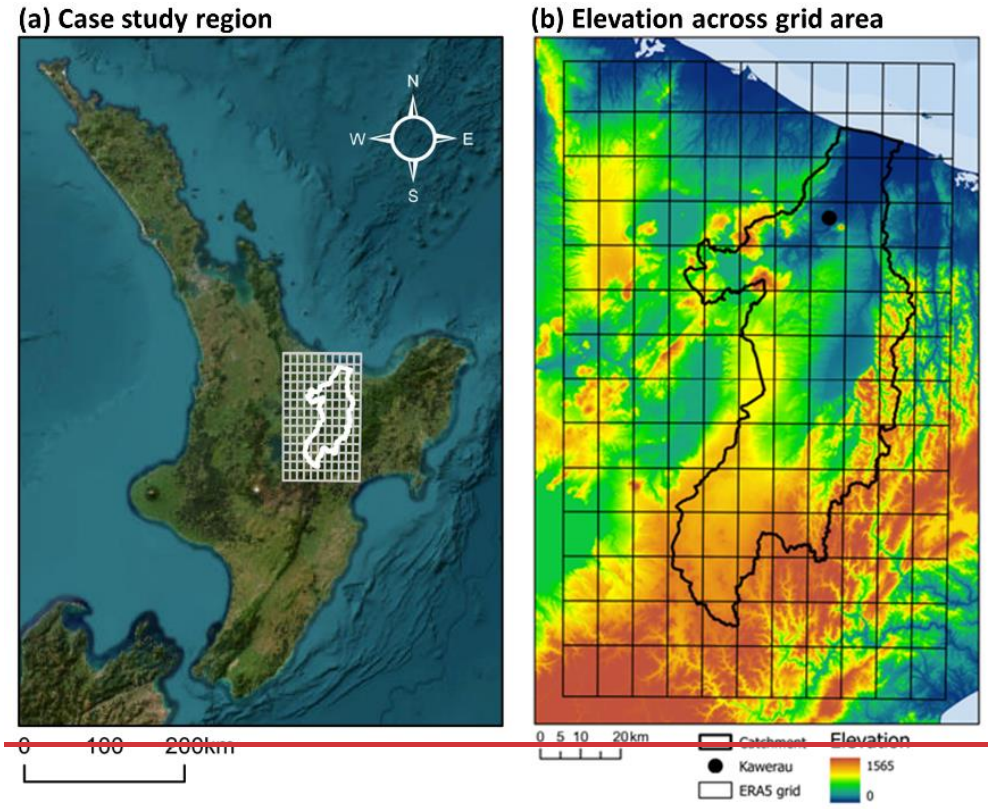
land, the variable is total precipitation *tp* in metres, and is the total amount of water accumulated over a particular time period, resetting every 24 hours (Muñoz Sabater, 2019). ~~SWM will run on as little as one day's worth of rainfall data however, the outputs will then only mimic the statistical properties of this day's weather. The recommendation is therefore to download at least the last 20 years of data (ERA5 data are available from 1960–current).~~

SWM first pulls time- and location-stamped precipitation data from the ERA5-land data and converts values from accumulated to hourly rainfall before combining all data into a single 3-D spatio-temporal array for analysis (Fig. 1a). A single point is selected at random from the locations used in the download of the original ERA5-land data (Fig. 1b). The precipitation data at this point is used to split the single array into periods of precipitation (*wet*), and no precipitation (*dry*) based on a user-defined rainfall tolerance (below which an hour is considered *dry*). The user must then define three start conditions: (1) the month and day from which stochastic data are to begin (noting that weather data are seasonal), (2) the length of time data are required for, and (3) how many sets of data are required. For example, 20 datasets for 10 years of data starting from 30th April. SWM builds an empty array, the spatial extent of which is based on the ERA5-land netCDFs, and the user-defined temporal extent. A starting block of wet or dry is randomly selected (constrained by the starting month) and inserted into the empty array, the timestamp is updated (starting time + block length), a check is made to see whether this timestamp exceeds the required data size, if yes, the algorithm stops (stochastic data already generated), if not, the type of block is switched (wet to dry, or dry to wet), and the loop begins again (Fig. 1c). The final output from SWM is a set of netCDFs identical in form to those of the ERA5-land data except that precipitation is hourly, rather than cumulative.

3 Example application: Rangitāiki-Tarawera catchment, Bay of Plenty, New Zealand

The Rangitāiki-Tarawera catchment is an area susceptible to many natural hazards including volcanic eruptions, flooding, and extreme weather events (ex-tropical cyclones). Hourly rainfall data across a 11x14 grid (Fig. 2) of longitude: {176° E, 176.1° E, …, 177.0° E} and latitude: {37.8° S, 37.9° S, …, 39.1° S}, for 40 years (1981 – 2020) were downloaded from ERA5-land. The ERA5-land data is then prepared for the Rangitāiki-Tarawera catchment through SWM by converting these individual netCDFs into a 11 (longitude steps of 0.1° E) by 14 (latitude steps of 0.1° S) by ~350,400 (time steps in hours) array of hourly data (24 hrs x 365 days x 40 yrs). The Rangitāiki-Tarawera array is then split into wet/dry time-stamped blocks. ERA5-land precipitation data at this catchment were commonly of very small (~10^{-18} m) but non-zero values (common for ERA4-land data, Muñoz-Sabater et al., 2021). The New Zealand climate report for the region (Chapell, 2013) provides average monthly rain and wet days at Kawerau (a town central to the catchment) of 112 wet days per year (~31 %), where wet days are defined as more than one mm per day. A rainfall tolerance level of zero resulted in ~54 % of data classified as *wet*, using the climate report criteria (less than one mm rainfall in 24 hrs = 4.12 x 10^{-5} m hr^{-1}) resulted in ~29 % of data defined as *wet*. For this exemplar, the latter was used. Stochastic precipitation data for the Rangitāiki-Tarawera catchment was then built using SWM to obtain 9~~5~~99 sets of 40 years' worth of hourly stochastic rainfall data across the region as 9~~5~~99 netCDFs. The number of

runs (9995) was chosen here to provide ninety-fifth percentile bounds for the statistical analyses (Sect. 4), but in practice can be set to any value, e.g., to match the number of downstream hazard simulations planned. This approach is common practice to assess statistical significance with non-parametric bootstrap methods (DiCiccio & Efron, 1996; Ramachandran & Tsokos, 2021).


(a) Case study region
N
W
E
S
0 100 200km
(b) Elevation across grid area
0 5 10 20km
Catchment
Kawerau
ERA5 grid
Elevation
1565
0
(c) January
N
W
E
S
Catchment
Kawerau
ERA5 grid
Monthly Rainfall
No Data
< 90 mm
90 - 100 mm
100 - 110 mm
110 - 120 mm
120 - 130 mm
130 - 150 mm
0 5 10 20km
(d) July
0 5 10 20km

75

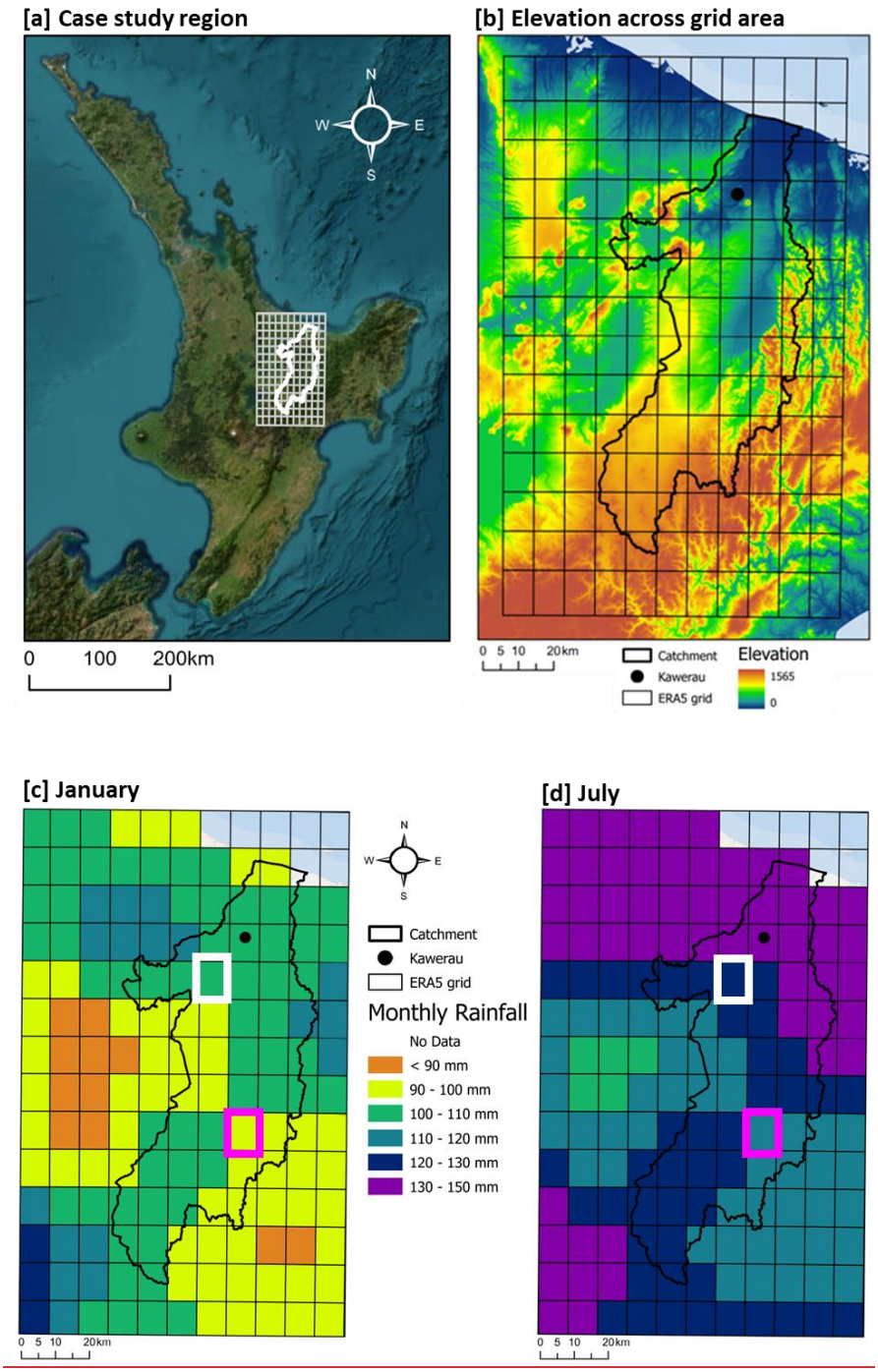


Figure 2: ERA5-land data across the case-study area. (a) Case study area, North Island, Aotearoa, basemap from Earthstar Geographics (https://www.terracolor.net/), (b) Elevation of case study area with ERA5-land grid, catchment of interest, and Kawerau township, (c) Mean monthly total rainfall for January (ERA5-land data: 1981 to 2020), (d) Mean monthly total rainfall for July (ERA5-land data: 1981 to 2020). Comparison locations in Fig. 3 are shown as white (Realisation 1) and pink (Realisation 2) boxes in (c) and (d).

4 ~~Statistical analyses~~Evaluation Method

Four sets of statistical analyses were undertaken to ensure that SWM simulated data are stochastically similar to ERA5-land data. For this, the 9~~5~~99 sets of 40 years of simulated data across the 11x14 grid are used, code for these analyses being in the GitHub repository, and all outputs for the exemplar in the supplementary material. The location at which tests are performed were selected randomly, and the whole process was run through twice to check ~~sample~~realisation sensitivity. The latter exercise was to check whether any patterns observed in the first 9~~5~~99 sets of simulated data were repeated in the second set. These are referred to as *Realisation*~~*Sample*~~ *1* and ~~*Sample*~~*Realisation* *2* in the results, with each ~~sample~~ realisation containing ~~95~~ 999 sets of simulated data. The analyses were:

(1) Monthly means and variance

Students t-test was used to compare the monthly mean rainfall between real and simulated data, the Shapiro-Wilks test is used to test normality to determine the appropriate equal variance test: Bartlett if data are normally distributed, or Levene if not (Fox and Weisburg, 2019).

(2) Significance of month and source for rainfall prediction

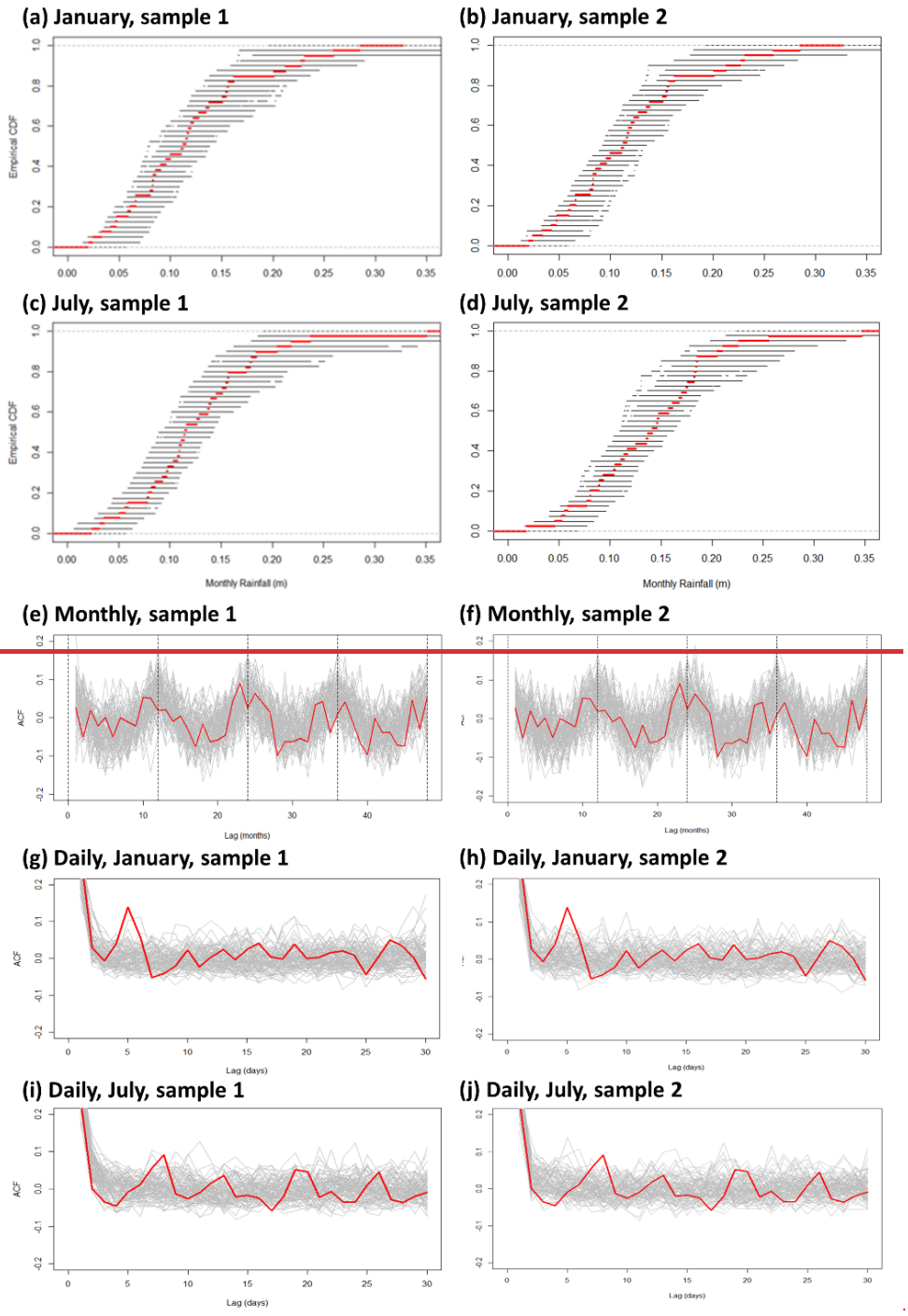
Tukey's Honest Significant Difference (HSD) (e.g., Miller, 1981) is used to determine whether source (real or simulated) is a significant factor in the prediction of total monthly rainfall.

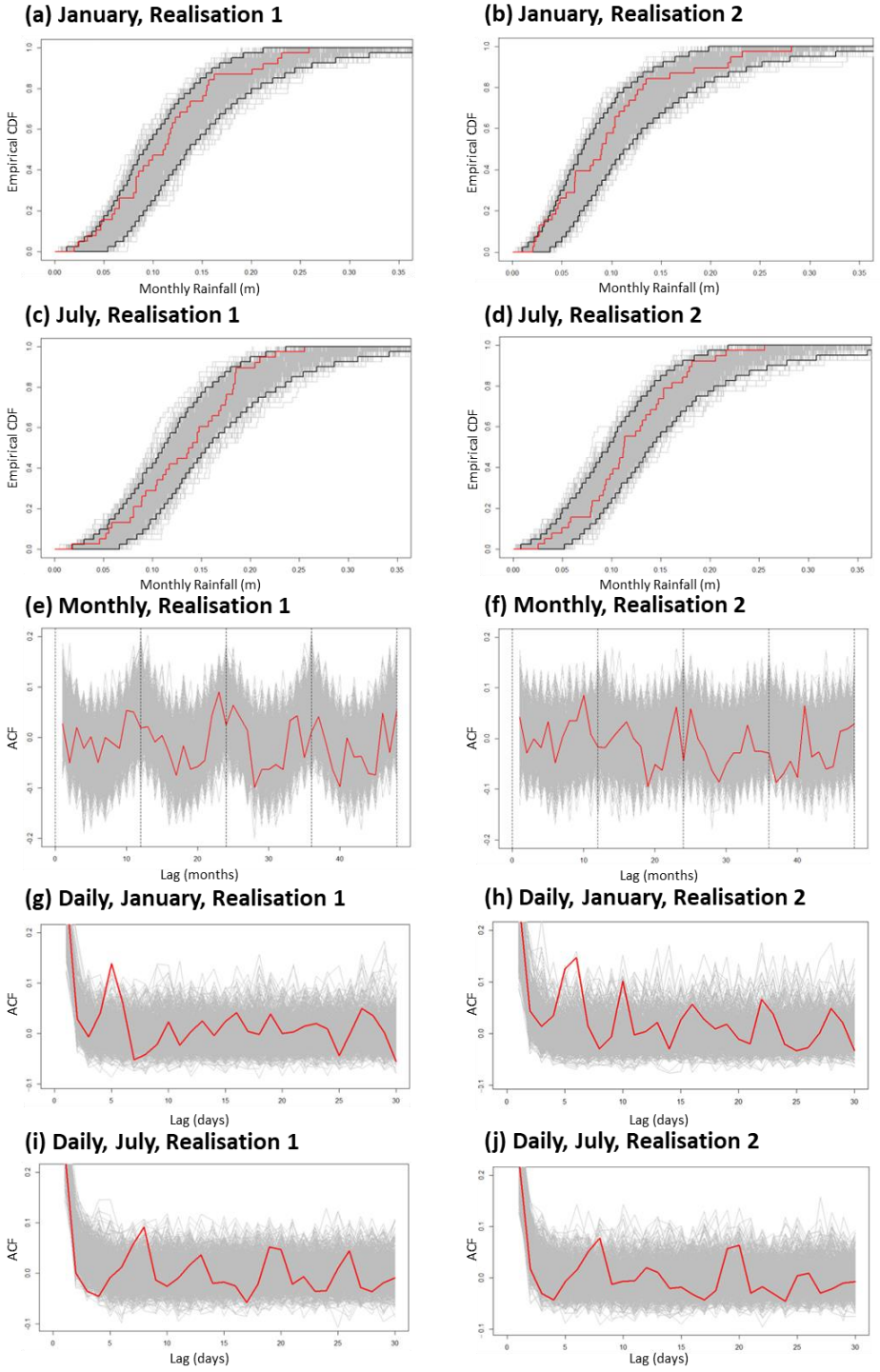
(3) Distributions of monthly rainfall totals

A non-parametric bootstrap method is applied whereby empirical cumulative distribution functions (eCDFs) for each simulated dataset representing the 95th percentile envelope (from 9~~5~~99 runs) is overlain by the real dataset.

(4) Temporal trends on daily and monthly timescales

Autocorrelations are calculated for each of the real and simulated data to compare any significant lags (Venables and Ripley, 2002). Tests require a continuous variable in discrete time so are run on cumulative daily and cumulative monthly rainfall.


(a) January, sample 1
(b) January, sample 2
(c) July, sample 1
(d) July, sample 2
(e) Monthly, sample 1
(f) Monthly, sample 2
(g) Daily, January, sample 1
(h) Daily, January, sample 2
(i) Daily, July, sample 1
(j) Daily, July, sample 2
Empirical CDF
Monthly Rainfall (m)
ACF
Lag (months)
Lag (days)



105 **Figure 3: Statistical analyses results, simulated data are shown as grey lines, real data are shown as red lines. (a) ~~Sample~~ Realisation 1: January empirical Cumulative Distribution Functions of rainfall totals (m), (b) Realisation~~Sample~~ 2: January empirical Cumulative Distribution Functions of rainfall totals (m), (c) ~~Sample~~Realisation 1: July empirical Cumulative Distribution Functions**

of rainfall totals (m), (d) RealisationSample 2: July empirical Cumulative Distribution Functions of rainfall totals (m), (e) SampleRealisation 1: Monthly cumulative rainfall autocorrelations, (f) SampleRealisation 2: Monthly cumulative rainfall autocorrelations, (g) SampleRealisation 1: January daily cumulative rainfall autocorrelations, (h) RealisationSample 2: January daily cumulative rainfall autocorrelations, (i) SampleRealisation 1: July daily cumulative rainfall autocorrelations, (j) SampleRealisation 2: July daily cumulative rainfall autocorrelations. 95th percentile bounds for (a) to (d) are represented by the black lines and built from the 999 runs of simulated data (grey lines). Results are shown for a single randomly selected grid point, here results are shown at locations identified in Fig. 2.

5 Results

Overall, while SWM passed all statistical tests for both samplesrealisations, some departure was noted in several combinations, the specifics of which are detailed below (see Supplementary Material for complete set of results):

(1) Monthly means and variance

Real data and simulated data failed the Shapiro-Wilks test of normality (p = 0.001), and realisations failed (p < 0.05) at a 41 % (Realisation 1) and 52 % (Realisation 2) of the 999 simulations respectively.similar rate (~52–58 %) for both samples, tThus, the Levene test for equal variance was always used.

SampleRealisation 1: Only eight 105 of 1140 11988 (12 x 40) tests showed a statistically significant difference (p < 0.05) in monthly means to the real data(p < 0.05); expected number under the null hypothesis of no difference being 57599 (11988 x 0.05). Fifty one679 pairs of real/simulated months (5.7 %) failed the test of equal variance (p < 0.05), of which 26 410 (more than half) were in JuneMay.

SampleRealisation 2: Only three 76 of 1140 11988 (12 x 40) tests showed a statistically significant difference in monthly means (p < 0.05). Forty three445 pairs (3.84 %) failed the test of equal variance (p < 0.05), of which 12 189 were in June, and 1263 in February.

(2) Significance of month and source for rainfall prediction

Two linear models were built with rainfall as the response variable, and both month and source (real or simulated) as predictor variables. One was built with an interaction term (m1), one without (m2). Both models for both samples realisations passed Tukey's HSD test, rejecting source as a statistically significant predictor (*SampleRealisation 1:* p = 0.8434.987; *SampleRealisation 2:* p = 0.9786989).

(3) Distribution of monthly rainfall totals

Samples Realisations 1 and 2: Envelopes were built for each sample realisation and for each month from eCDFs of the simulated data, these were then overlain by the real data, which consistently fell within this ninety-fifth percentile envelope (examples in Fig. 3a-db).

(4) Temporal trends on daily and monthly timescales

SamplesRealisations 1 and 2: Autocorrelations for cumulative daily rainfall for both real and simulated data over longer time lags (> 30 days) showed minimal trends through time, thus, although we cannot reject a conclusion of similar trends between the two, results remain inconclusive. Looking within monthly data only and over shorter time lags (< 30 days), there are potential departures of the real data more positively correlated than the simulated data for January at a lag of 5 days (Fig. 3g-

h), and for July at a lag of 8 days (Fig. 3i-j). Autocorrelations at the monthly scale show seasonal trends in both real and simulated data (Fig. 3e-fd) with strong positive correlations at the yearly level (every 12 months), and negative correlations at the six-monthly level (e.g., rainfall in January is negatively correlated with rainfall in July). A potential horizontal offset is noted between real and simulated data. The differences between real and simulated data here are attributed to multi-day rain or dry events. For example, in the simulated data, a large *wet* block representing a tropical cyclone in the original data occurring entirely in December, can feasibly start at 31 December in the simulated data and thus will run over into January. January and July results are provided here as they represent the most extreme weather months for the region (in terms of both drought and heavy rainfall) but also the most extreme departures from the simulated data. The remainder of the months do not show any notable differences in lags between simulated and real data for either ~~sample~~ realisation (see supplementary material for complete result set).

6 Conclusions

The method and code provided through this brief communication can be used to rapidly generate multiple sets of realistic, long-term, hourly precipitation data over a spatial region. While the outputs may not have the nuances that come with more complex models (e.g., Burton et al., 2008; Papalexiou, 2022), the efficient open-source code, written in an open-source language, and based on open source data facilitate~~This code provides~~ an easy to plug-in input for hazard simulations to support long-term, time and spatially varying, probabilistic risk assessments, uncertainty quantification, and multi-hazard models.

Data availability

All data were obtained from the Copernicus Climate Change Service (CS3) Climate data Store (CDS) and is published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

Code availability

Code is written in R (open-source software) and is freely available at https://github.com/MelWhitehead/SWM.

Supplement

The supplement related to this article is available online at: << link for nhess supplement to go here >>

Author contributions

Both authors conceptualised the model, MW built the model, statistical tests were guided by MB and coded by MW.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

This research was supported by Resilience to Nature's Challenges Multi-hazard Risk model Program, New Zealand (contract GNS-RNC043).

References

Arnaud, P., Bouvier, C., Cisneros, L., and Dominguez, R.: Influence of rainfall spatial variability on flood prediction, J. Hydrol., 260, 216-230, https://doi.org/10.1016/S0022-1694(01)00611-4, 2002.

Burton, A., Kilsby, C., Fowler, H., Cowpertwait, P., and O'Connell, P.: RainSim: A spatial–temporal stochastic rainfall modelling system. Environmental Modelling & Software, 23(12), pp.1356-1369, https://doi.org/10.1016/j.envsoft.2008.04.003, 2008.

~~Muñoz-Sabater, J.: ERA5-Land hourly data from 1950 to present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS), https://doi.org/10.24381/cds.e2161bac, 2019, last access: 23 August 2023.~~

Chappell, P.R.: The climate and weather of Bay of Plenty, 3rd Edition, NIWA Science and Technology Series, Number 62, ISSN 1173-0382, https://niwa.co.nz/static/BOP%20ClimateWEB.pdf, 2013, last access: 23 August 2023.

DiCiccio, T. and Efron, B.: Bootstrap Confidence Intervals, Stat. Sci., 11:3, 189-212, https://www.jstor.org/stable/2246110, 1996.

Fox, J. and Weisberg, S.: An R Companion to Applied Regression, Third edition, Sage Publications, Thousand Oaks CA, USA, 576 pp., https://socialsciences.mcmaster.ca/jfox/Books/Companion/, 2019.

Gao, L., Zhang, L., and Lu, M.: Characterizing the spatial variations and correlations of large rainstorms for landslide study, Hydrol. Earth Syst. Sci., 21, 4573–4589, https://doi.org/10.5194/hess-21-4573-2017, 2017.

Gill, J. C. and Malamud, B. D.: Hazard interactions and interaction networks (cascades) within multi-hazard methodologies, Earth Syst. Dynam., 7, 659–679, https://doi.org/10.5194/esd-7-659-2016, 2016.

Hyman, D. M., Bevilacqua, A., and Bursik, M. I.: Statistical theory of probabilistic hazard maps: a probability distribution for the hazard boundary location, Nat. Hazards Earth Syst. Sci., 19, 1347–1363, https://doi.org/10.5194/nhess-19-1347-2019, 2019.

Miller, R. G.: Normal Univariate Techniques, In: Simultaneous Statistical Inference, Springer Series in Statistics, Springer, New York, NY, 37-108, https://doi.org/10.1007/978-1-4613-8122-8_2, 1981.

Muñoz Sabater, J.: ERA5-Land hourly data from 1950 to present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS), https://doi.org/10.24381/cds.e2161bac, 2019, last access: 23 August 2023.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J.-N.: ERA5-Land: a state-of-the-art global reanalysis dataset for land applications, Earth Syst. Sci. Data, 13, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021, 2021.

Papalexiou, S.: Rainfall generation revisited: Introducing CoSMoS-2s and advancing copula-based intermittent time series modeling. *Water Resources Research*, *58*(6), https://doi.org/10.1029/2021WR031641, 2022.

R Core Team: R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/, 2021.

Ramachandran, K., and Tsokos, C.: Mathematical Statistics with Applications in R, Third Edition, Academic Press, Elsevier, https://doi.org/10.1016/B978-0-12-817815-7.01001-0, 2021.

Thompson, M.A., Lindsay, J.M., and Gaillard J.C.: The influence of probabilistic volcanic hazard map properties on hazard communication, J. Appl. Volc., 4, 1-24, https://doi.org/10.1186/s13617-015-0023-0, 2015.

Venables, W.N. and Ripley, B. D.: Modern Applied Statistics with S. Fourth Edition, Springer Series in Statistics and Computing, Springer-Verlag, New York, NY, 498 pp., https://doi.org/10.1007/978-0-387-21706-2, 2002.

Wickham, H., François, R., Henry, L., Müller, K., and Vaughan, D.: dplyr: A Grammar of Data Manipulation. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr, 2023.

Zhao, Y., Nearing, M.A., and Guertin, D.P.: A daily spatially explicit stochastic rainfall generator for a semi-arid climate, J. Hydrol., 574, 181-192, https://doi.org/10.1016/j.jhydrol.2019.04.006, 2019.