Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Evaluating stochastic rainfall models for hydrological modelling

Thien Huy Truong Nguyen [*], Bree Bennett, Michael Leonard

School of Architecture and Civil Engineering, University of Adelaide, North Terrace Campus, SA 5005, Australia

ARTICLE INFO

This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Simon Michael Papalexiou, Associate Editor.

Keywords:
Stochastic rainfall
Rainfall-runoff
Continuous simulation
Rainfall generation

ABSTRACT

Stochastic rainfall models are important tools for evaluating hydrological risks such as flooding and drought because of their ability to randomly generate alternative plausible climatic timeseries. The stochastic generation of climatic timeseries is not an end in itself, since they are typically applied to a catchment to determine the performance of water-related infrastructure systems, such as reservoirs or flood-control measures. This methodology typically involves a train of models to determine the end-of-system impact, yet the evaluation of stochastically generated rainfall timeseries is usually a stand-alone procedure focused on metrics directly related to the stochastic generator. This paper demonstrates discrepancies in this approach by evaluating two, daily-timestep, stochastic rainfall models in terms of rainfall metrics and their subsequently generated flow metrics after rainfall-runoff transformation. The two models are a Markov-based model and a latent-variable model, where each model is calibrated and evaluated showing 'overall good' performance. Stochastically generated timeseries, alongside observed rainfall timeseries are inputted to a calibrated catchment model (GR4J) to derive daily flow timeseries. Whereas the rainfall metrics typically showed 'good' performance, streamflow-based metrics are not necessarily 'good'. The procedure is repeated for 277 stations from Australia and 106 stations from the United States of America. Depending on the strictness of the flow-based comparison and region analysed, using the Markov-based model 12–26% of sites were classified as 'poor' performing, and 1%-9% of sites were classified as 'poor' using the latent-variable model. The results demonstrate that catchment-based performance of flow metrics is more holistic since it magnifies features of the rainfall not otherwise visible to rainfall-based evaluation.

1. Introduction

Floods and droughts are infrequent events, yet they have significant impact in terms of their economic disruption, damage to infrastructure, social upheaval, loss of life and environmental degradation (Leonard et al., 2014). The ability to determine the risk of extreme hydrological events is crucial for engineering design, disaster response, mitigation strategies, early warning systems, and long-term planning (Linsley and Crawford, 1974, Boughton and Hill, 1997, Blazkova and Beven, 2002, Lamb, 2005, Viviroli et al., 2009). Hydrological risks are notoriously difficult to estimate from streamflow records due to factors such as catchment change over time, and the limited availability of streamflow records along with the limited length of those records (Do et al., 2017). As with other hazards, a train of models (e.g. the climate-rainfall-runoff model-train) is typically needed to evaluate risks, wherein the significant challenge rests with establishing confidence in the end-of-system metric or variable of interest to decision making. The contribution of this paper is to emphasize end-of-system evaluation (i.e. hydrological evaluation) and the limitation of single component-wise evaluation, with the example of rainfall and streamflow models that are used for flood and drought risk estimation.

A common approach to hydrological risk evaluation is to simulate streamflow from observed rainfall using a hydrological model that has been established as an effective representation of catchment dynamics (Kuczera et al., 2006, Thyer et al., 2009, McInerney et al., 2018). This approach is appealing because rainfall records are typically longer, more widely available, and more homogenous than streamflow records. However, even with long observation records, there can be significant uncertainty in risk estimates given the focus on estimating low-frequency events (i.e. floods and droughts). Therefore, to further augment flood-risk or drought-risk assessment, the rainfall input may itself be derived from a stochastic rainfall model (SRM) (Baxevani and Lennartsson, 2015, Bennett et al., 2018, Evin et al., 2018, Grimaldi et al., 2022). SRMs are designed to mimic the features of rainfall records from

[*] Corresponding author.
E-mail address: truonghuythien.nguyen@adelaide.edu.au (T.H.T. Nguyen).

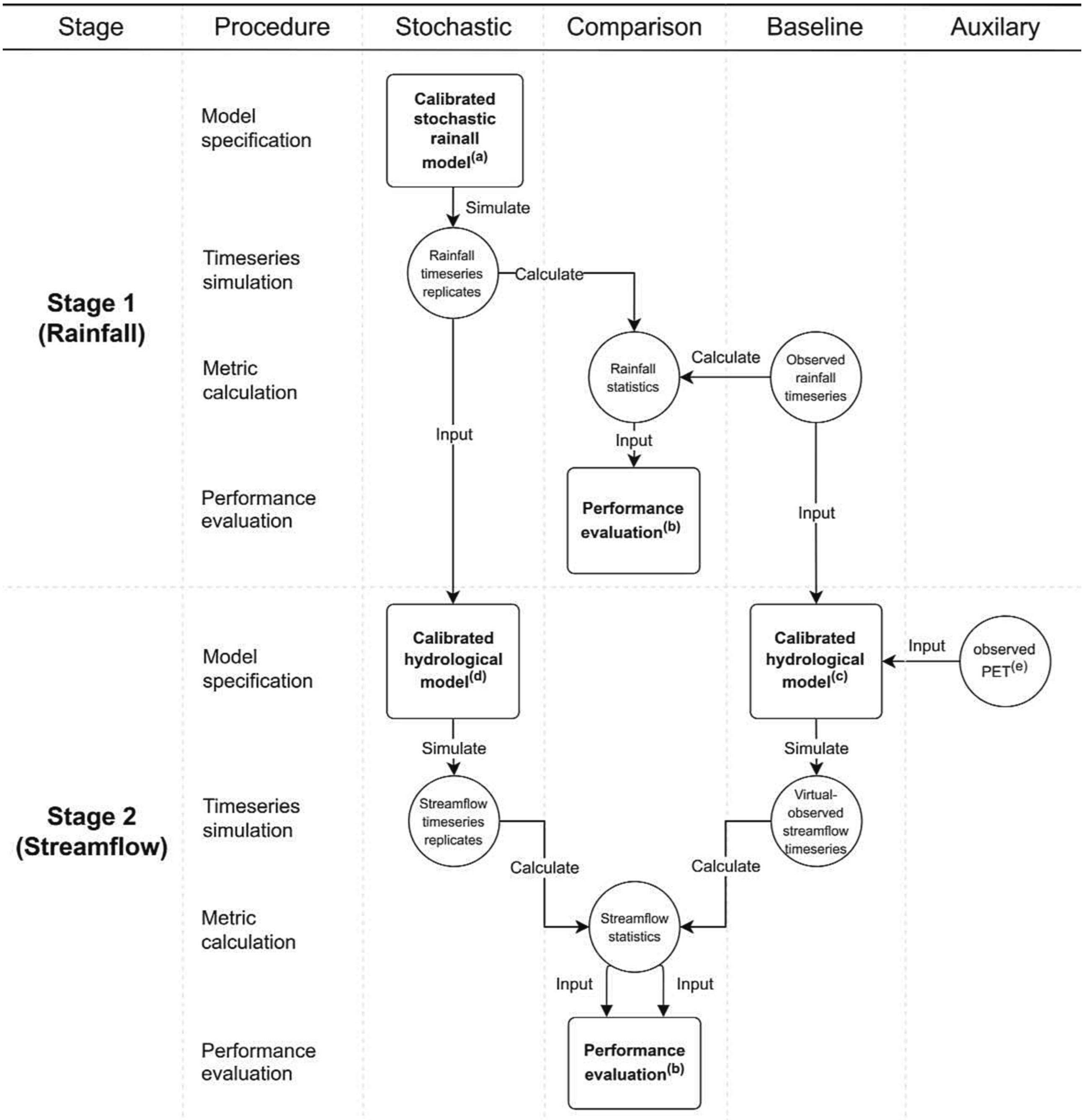
https://doi.org/10.1016/j.jhydrol.2023.130381
Received 23 April 2023; Received in revised form 1 August 2023; Accepted 14 September 2023
Available online 29 October 2023

Table 1
Typical rainfall statistics for stochastic rainfall model evaluation, where standard deviation is denoted as ’std. dev.’

Scale	Statistic	Richardson (1981)	Wilks (1998)	Rasmussen (2013)	Baxevani & Lennartsson (2015)	Evin et al. (2018)	Bennett et al. (2018)	Gao et al. (2020)	Papalexiou (2022)
Daily	Daily mean				✓	✓			
	Daily std. dev.					✓			
	Distribution of wet day amounts					✓			✓
	Autocorrelation of wet day amounts								✓
	Multi-day aggregations								✓
Monthly	Mean wet day amounts		✓	✓			✓	✓	
	Std. dev. wet day amounts		✓	✓			✓	✓	
	Skew wet day amounts			✓			✓		✓
	Mean number of wet days	✓					✓		✓
	Std. dev. number of wet days						✓		
	Wet and dry spells length			✓	✓	✓	✓		✓
	Wet and dry spells distribution			✓	✓	✓	✓		✓
	Mean total rainfall	✓		✓			✓	✓	
	Std. dev. total rainfall						✓		
	Total rainfall lower tail						✓		
	Total rainfall upper tail				✓	✓	✓		
Annual	Mean total rainfall						✓		
	Std. dev. wet day amounts						✓	✓	
	Total rainfall lower tail						✓	✓	
	Total rainfall upper tail						✓	✓	
	Mean wet day amounts						✓		
	Std. dev. wet day amounts						✓	✓	
	Mean number of wet days					✓	✓		
	Std. dev. number of wet days					✓	✓		
	Mean maximum consecutive dry days							✓	
	Mean maximum consecutive wet days							✓	
	Daily annual maxima					✓	✓	✓	✓
Spatial correlation	Joint probability of wet and dry events		✓	✓			✓		✓
	Cross-correlation occurrence-amount		✓	✓		✓			✓
	Continuity ratio		✓						

inter-annual timescales down to daily or sub-daily timescales, and in so doing generate plausible hypothetical alternative continuous sequences of rainfall variability. SRMs have been introduced over many decades (Richardson and Wright, 1984, Srikanthan and McMahon, 2001, Mehrotra, 2005) in terms of their first-order priority to directly reproduce rainfall features of interest, but seldom to evaluate their performance with reference to derived streamflow. This paper is therefore concerned with the fidelity of SRMs to produce reliable estimates of streamflow for hydrological risk assessment and asks, for a wide range of catchments and specified performance criteria, whether apparently ‘good’ modelled rainfall leads to ‘good’ modelled streamflow.

SRMs synthetically generate rainfall at a specified scale of interest (e.g. sub-daily, daily, monthly, annual, multi-annual) to have statistically similar properties to observed rainfall measured from rain gauges. There are a wide variety of SRMs and correspondingly a wide variety of rainfall-based evaluation metrics. The specific features of a SRM vary significantly and depend on the scale of interest and data sources (Srikanthan and McMahon, 2001). Models have been developed across a range of timescales including interannual (Thyer and Kuzera, 1999, Srikanthan and Pegram, 2009), monthly (Thompson, 1984), daily (Richardson and Wright, 1984, Sharma and Lall, 1999), and sub-daily (Gupta and Waymire, 1993, Cowpertwait, 2006, Papalexiou, 2022). Models can also be at a single site (Chowdhury et al., 2017, Gao et al., 2020), multiple sites (Evin et al., 2018, Wilks, 1998) or continuous in space (Leonard, 2010, Baxevani and Lennartsson, 2015, Bennett et al., 2018). The ambition of SRMs is that they reproduce key metrics across all relevant timescales (such as from sub-daily to inter-annual) and all elements of the distribution (lower/upper tails, mean, variability, wet-dry patterns, etc.). Table 1 provides a comparison of rainfall-based evaluations from a diverse sample of studies including single site, multi-site, daily and sub-daily models, with different underpinning simulation schemes. It shows that there is potentially a wide pool of statistics to consider in determining whether the SRM is performing adequately (wet-dry patterns, correlations, seasonal patterns, moments of the distribution, extremes, etc.), but also that there is a strong degree of variability in evaluation between studies. Even with an exhaustive set of evaluation metrics, the wide variety of performance across these metrics makes it difficult to establish the relative importance of any



[a] rainfall calibration processes not shown - i.e. parameters, observed data, optimisation method, objective function, loop.
[b] performance evaluation has aggregation parameters not shown in the flow chart (see Section 2.1).
[c] streamflow calibration processes not shown - i.e. parameters, observed data, optimisation method, objective function, loop.
[d] identical model specification and calibrated parameters, including same input PET timeseries (i.e. only rainfall input differs).
[e] bootstrap required to remove daily rainfall-PET correlation (see Section 2.5).

Fig. 1. Conceptual representation of evaluation procedure for stochastic rainfall model in terms of two stages: Stage 1, rainfall evaluation; and Stage 2, hydrological evaluation. The squares show three key models of interest: (i) the stochastic rainfall model ultimately being evaluated, (ii) a calibrated hydrological model necessary for streamflow generation and (iii) a specified model for performance evaluation. Circles show data (i.e. timeseries or metrics). Arrows show processes.

discrepancies or biases in the generated rainfall (Bennett et al., 2018). Even though SRMs can be evaluated against a variety of statistics (as in Table 1), there can remain elusive features of the rainfall that are not readily evaluated but may be hydrologically significant (e.g. the rainfall antecedent to an extreme event).

SRMs are typically used to perform a continuous simulation that generates streamflow, and thus they should be ultimately assessed in terms of resulting streamflow performance. Continuous simulation requires a rainfall-runoff model that receives input timeseries of rainfall and potential evapotranspiration or temperature (Beven, 2012), updates catchment infiltration and groundwater-fluxes, and together with the rainfall, determines streamflow. There are many different types of models whether physically based (Abbott et al., 1986, Liu et al., 2008), conceptual (Boughton, 2004, Croke et al., 2006, Perrin et al., 2003a) or statistical (Kingston et al., 2005, Adnan et al., 2019). Regardless of model type, the key observation is that the catchment properties together with the state of wetness in the catchment through time can either operate to dampen or amplify the transformation of rainfall into streamflow. Therefore, whenever a discrepancy exists in simulated rainfall, there is the potential for a hydrological model to amplify this discrepancy and cause the resulting streamflow to be statistically dissimilar to streamflow derived from observed rainfall (Bennett et al., 2019). Even with evaluation against a comprehensive set of rainfall metrics (Bennett et al., 2019), hydrological evaluation provides an additional and potentially greater assessment because it integrates the rainfall into a variable with closer connection to the end-of-system impacts and related decisions.

SRMs should be able to reproduce streamflow characteristics for practical hydrological application as well as preserving rainfall attributes. However, because catchments integrate rainfall over a region and over time, the causes of deficiencies in simulated streamflow are not simple to identify. For example, poorly simulated streamflow within a given month could be the result of rainfall deficiencies in a preceding month, and poorly simulated rainfall need not necessarily lead to poor streamflow (Bennett et al., 2019). The inconsistency in the quality of simulated rainfall and simulated streamflow has been reported in multiple studies (Bennett et al., 2019, Gao et al., 2020), indicating that the issue is not specific to a single type of SRM or an isolated catchment. Importantly, these papers demonstrate that the identification of 'poor' streamflow is not due to genuine lack of evaluation or calibration effort. For example, the study by Gao et al. (2020) showed a relatively large underestimation in the high streamflow range despite evaluating the rainfall model against numerous rainfall statistics including wet/dry spell distributions and values in the lower/upper tails. Despite these examples and despite the ultimate use of SRMs for the evaluation of hydrological risks, studies that evaluate SRMs in the context of continuous streamflow simulation are limited. Hence the aim of this paper is to emphasize the importance of hydrological evaluation and to demonstrate the performance of stochastically generated rainfall in terms of derived streamflow metrics.

Specifically, this paper systematically evaluates rainfall model performance in simulating rainfall and streamflow using an accessible method for hydrological evaluation (Section 2) that simplifies comparison across multiple sites, metrics, and models. While there are a few examples in the literature of hydrological evaluation on individual catchments, this paper advocates for hydrological evaluation as a standard evaluation practice when calibrating SRMs. To this end, the analysis has been designed using a flexible and broad framework (Section 2.1), and a large number of sites (277 sites from Australia and 106 sites from the United States of America – Section 2.3). The results are presented for selected rainfall and streamflow metrics (Section 2.5) to demonstrate the completeness of the analysis. Given the broad nature of the analysis, detailed diagnostic evaluation, or remedy of identified deficiencies in the rainfall timeseries is beyond the scope of this study. The paper is therefore confined to emphasizing the potential deficiency of SRMs for generating streamflow despite 'best-practice' methods of calibration and evaluation (Section 3), indicating the magnitude of the challenge along with possible pathways to address this problem (Section 4).

2. Methodology

To identify instances where potentially 'good' modelled rainfall degrades into inferior modelled streamflow, performance criteria for SRMs need to be established for both rainfall and streamflow statistics. The virtual hydrological evaluation framework proposed in Bennett et al. (2019) is used as the foundation of the evaluation framework of this paper (Section 2.1) because it introduces the concepts of comprehensive evaluation (via many statistics), systematic evaluation (via specified performance criteria) and hydrological evaluation. This paper extends the framework in Bennett et al. (2019) by utilising quantitative aggregation across sites, months, metrics, and models for any aggregation of interest, allowing an examination and comparison of 'overall performance'.

2.1. Evaluation framework

Fig. 1 illustrates the evaluation procedure for an individual catchment, following Bennett et al. (2019). The evaluation has two main stages: (Stage 1) rainfall-based evaluation and (Stage 2) hydrological evaluation. The outcome of each stage is an evaluation that is able to indicate the quality of the SRM's performance, whether in terms of rainfall metrics or streamflow metrics. The primary interest of the overall framework is the role of three different processes, indicated by black squares in the flow chart in Fig. 1, where circles show contributing data timeseries and metrics, and arrows show the process flow of data. The explanation of concepts and procedures in this paper is centred on: (i) the calibrated stochastic rainfall model ultimately being evaluated, (ii) the calibrated hydrological model necessary for streamflow generation, and (iii) a framework for performance evaluation. Note that the process representation of model calibration in Fig. 1 has been omitted, not because of unimportance, but to better highlight the procedure and objective of model evaluation without the additional clutter of this requisite step. The simplification is pragmatic given the extensive literature on calibration procedures for SRMs and hydrological models.

Traditional evaluation of SRMs has typically relied on non-systematic evaluation whereby the modeller decides that the model is 'reasonable' (Leonard et al., 2008), 'realistic' (Baxevani and Lennartsson, 2015), or 'fit-for-purpose' (Sadeghfam et al., 2021) according to their expert judgement based on an empirical review of statistics and visual inspection of summary graphs. The comprehensive and systematic framework (CASE) of Bennett et al. (2018) formalises the SRM evaluation by making explicit the performance evaluation model. The evaluation model explicitly specifies *a priori* rules by which the evaluation proceeds, the set of metrics over which the evaluation is based and the parameters (e.g. thresholds) that govern the comparison of simulated-to-baseline metrics (i.e. Fig. 1 – observed rainfall or virtual-observed streamflow). The benefit of the framework is that the evaluation is systematic and consistent and that it takes into account the variability of metrics via confidence intervals of the stochastic replicates (Bennett et al., 2018). The rainfall evaluation (Fig. 1 – Stage 1) and the streamflow evaluation (Fig. 1 – Stage 2) utilise the same evaluation model, only with difference in the metrics inputted for evaluation (though it is conceivable to specify different rules and parameters for the different stages if warranted). Following Bennett et al. (2018), the evaluation model categorises the SRM performance as either 'good', 'fair' and 'poor':

- 'Good' performance means that the baseline observed/virtual-observed metric lies inside the 90% probability limits of the simulated metric.
- 'Fair' performance means that the observed/virtual-observed metric lies outside the 90% probability limits of the simulated metric but within the 99.7% probability limits.
- 'Poor' performance is reserved for all other instances.

The framework introduced by Bennett et al. (2019) advises that the evaluation should be comprehensive due to the many degrees of freedom in simulating rainfall. Therefore, an evaluation should utilise many statistics, but avoid obvious redundancy. Table 1 outlines typical metrics for Stage 1 evaluation (the mean and standard deviation of rainfall totals at varying scales, properties of the wet-dry process such as wet-spell and dry-spell durations, correlations in time and space, and properties of extremes). While the suite of evaluation metrics can vary

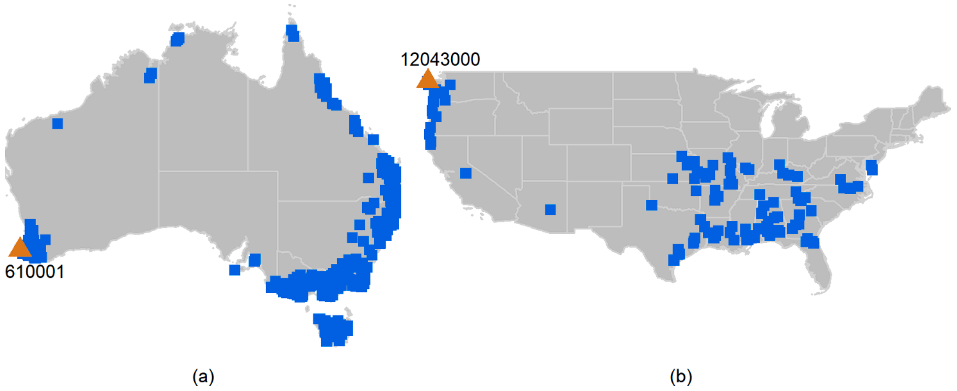


Fig. 2. Filtered catchment dataset and catchment locations in (a) Australia and (b) the US where NSE between observed and simulated streamflow is greater than 0.6 and the snow fraction is below 0.1. Orange triangles show the representative catchments discussed in detail in Section 3, other catchment aggregate evaluations shown in Supplementary Material B. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

significantly across studies, there are typically hundreds of metrics to compare given that performance is usually assessed at multiple scales (e.g. daily aggregates, monthly/annual variation, multiple sites). For Stage 2, Bennett et al. (2019) used the mean and standard deviation of streamflow totals at monthly and annual scales.

In this study the three performance categories are numerically indexed (hereafter referred to as the CASE index) for aggregation purposes in which ‘good’, ‘fair’, ‘poor’ are indexed as 1, 0.5 and 0, respectively. While it is conceivable to side-step the use of categories and rate performance on a continuous scale, this approach avoids significant complication of formally specifying test statistics according to the uncertainty of each metric. Although the specific values of the index weights are subjective (as is the specification of other aspects of the evaluation model, such as the selection of metrics, having three categories, and 90% limit thresholds), the framework forces a quantitative encoding of the ‘good/fair/poor’ trade-off that typically occurs heuristically in the evaluation of SRMs. Having specified numerical weights on the performance, the aggregate performance for a set of metrics of interest can be derived as a simple average with a resulting scale on the interval 0 to 1. It is assumed here that all specified metrics are weighted equally in the averaging process, although it is conceivable to have different weights to specify the relative importance to the end-user for different statistics. Having specified the numerical weights, a simple categorisation scheme for aggregate (overall) performance is applied to summarize:

- CASE index within 0 and 0.5 indicates ‘overall poor’ performance.
- CASE index within 0.5 and 0.75 indicates ‘overall fair’ performance.
- CASE index larger than 0.75 indicates ‘overall good’ performance.

In this study, the aggregation of performance typically occurs over 12 months for monthly statistics, over the number of temporal scales for aggregate statistics and over the number of years for annual statistics (Section 3.1). Overall performance for an individual catchment is the average performance of all statistics (Section 3.2).

2.2. Case study locations and performance metrics

This study considers a range of catchments in different hydro-climatic regions. A total of 1079 catchments are considered with 467 catchments located in the Australian continent and 671 catchments located in the contiguous United States (US). For Australian catchments, the daily streamflow, and meteorological data (daily rainfall and potential evapotranspiration) are sourced from the hydrologic reference database (Turner et al., 2012) and the SILO database (Jeffrey et al., 2001) with the daily rainfall and potential evapotranspiration sourced from the station nearest to the catchment outlet. For US catchments, the daily streamflow, and areal averaged meteorological data (daily rainfall, shortwave downward radiation, maximum and minimum temperature, humidity, and day length) are sourced from the widely used CAMELS dataset (Addor et al., 2017) and following Newman et al., (2015), the daily timeseries of potential evapotranspiration for US catchments are computed using the Priestly-Taylor method (Priestley and Taylor, 1972). The catchments cover a range of climates including tropical, subtropical, temperate, semi-arid and arid climates.

The set of catchments was filtered according to two conditions. The first condition was that the snow fraction was less than 0.1 (solely needed for US catchments), which would indicate that the rainfall-runoff relationship represents the majority of streamflow and that the relationship established by the hydrological model is suitable. The second condition was that the hydrological model had a relatively good fit between simulated and observed streamflow (comparable to instances of models used in practice), such that the Nash-Sutcliffe efficiency statistic used for calibration was 0.6 or greater (see Section 2.4). In other words, the set of catchments used for comparison was significantly reduced to mitigate the possibility of spurious and unrealistic rainfall-runoff relationships arising from poor calibration (while also noting that as the hydrological evaluation method is not an absolute comparison to observed streamflow this concern is moderated). Appling these criteria, the sample of catchments used to assess SRM performance reduces to 383 catchments, 277 catchments in Australia and 106 catchments in the US, as shown in Fig. 2. For these evaluated catchments average streamflow record lengths are 50 and 34 years for Australian and US catchments, respectively. The majority of catchments in both regions have a temperate climate. Details of each catchment are tabulated in Supplementary Material A.

Two catchments are selected to demonstrate the rainfall and streamflow evaluation of the two SRMs (Markov-based and latent-variable) (Fig. 2). The representative locations are chosen to provide greater detail in the manuscript, while summaries of overall performance for Australia and the US for each metric across all sites can be found in Supplementary Material B. Catchment 610001 is located in Western Australia with an area of 684 km^2 and has a temperate climate with winter-dominated rainfall and a dry, warm summer (annual average rainfall 1067 mm, seasonal temperatures from 13 °C to 30 °C). Catchment 12043000, is located in Washington with an area of 337 km^2, having a temperate climate with a warm summer (annual average rainfall 3096 mm, temperatures from −12 °C to 25 °C).

The performance of the two SRMs is evaluated on a range of statistics including rainfall and streamflow statistics at daily, monthly, annual scale with multiple thresholds and multi-day aggregations as well as selected streamflow event statistics. To avoid a high level of redundancy and potentially skewed summary, a correlation analysis of the candidate statistics was conducted and the resulting set of metrics for an individual catchment comprise 16 rainfall and 13 streamflow metrics as follows.

- Rainfall statistics at the daily scale: Evaluation of the mean, standard deviation of wet day amounts and number of wet days and the skewness of the wet day amounts to determine if the marginal distribution of the daily rainfall amounts and rainfall occurrence are being preserved. The wet and dry spell distributions ranging from 1-day to 10-day spells are also evaluated to ensure the models capture the intermittency of rainfall.
- Rainfall statistics at the monthly scale: Evaluation of the mean, median and standard deviation of the monthly total rainfall to ensure that the models preserve the seasonal characteristics of the rainfall. Rainfall extremes including the minimum and the 5th percentile (lower tail indicators), as well as the 90th percentile (upper tail indicators) of the monthly total rainfall are also evaluated as they are important features for flood and drought assessment.
- Temporal aggregation rainfall statistics: Evaluation of 1-day, 7-day annual maxima, proportion of dry days from 1 to 10-day aggregation, distribution of rainfall amount for 1,3,7-day aggregations as rainfall is highly structured across many scales.

The streamflow evaluation focuses on the following streamflow statistics:

- The mean and standard deviation of the monthly total streamflow to ensure that the model can represent the seasonal streamflow characteristics.
- A range of daily streamflow percentiles: the 5th, 50th, 70th and 90th, and the maximum of daily streamflow to ensure the model preserves the distribution of streamflow are evaluated on a month-wise basis. Due to the positive skewness of the streamflow distribution, flow percentiles from the 5th to the 50th are considered indicators of low flow while flow percentiles that are greater than the 90th percentiles are considered indicators of high flow.
- The mean and standard deviation of streamflow event metrics including peak, volume, and discharge. These event statistics are computed using the Lyne and Hollick filtering approach to baseflow separation (Lyne and Hollick, 1979). The procedure and parameters of the Lyne and Hollick baseflow separation follows the standard approach for daily streamflow suggested in Ladson et al. (2013).

For each catchment 100 replicates of rainfall and 100 replicates of streamflow are simulated of equal length to the observed input timeseries. The listed statistics are calculated for each replicate of the simulated rainfall and streamflow to compare with the observed rainfall and virtual-observed streamflow statistics.

2.3. Stochastic rainfall models

The SRM is calibrated to the catchment's observed rainfall timeseries and then used to generate a set of simulated rainfall replicates. The framework is flexible and can accommodate any type of SRM, with examples including Markov models (Katz, 1977, Wilks, 1998), autoregressive models (Rasmussen, 2013, Bennett et al., 2018) and non-parametric models (Mehrotra, 2005). Notably each SRM will have its own procedure for calibration, which is typically found alongside the introduction of the model in literature. A range of calibration methods are employed such as the method of moments (Rasmussen, 2013, Bennett et al., 2018) or likelihood techniques (Evin et al., 2018, Baxevani and Lennartsson, 2015), and a variety of parameterisation schemes may be adopted including varying parameters according to harmonics or monthly blocks. The subsequent evaluation framework assumes that the SRM has been genuinely calibrated using best-practice techniques to avoid spurious analyses.

To provide a comparison of model performance, this study utilises two different types of daily SRMs: a WGEN-type (WGEN) model (Richardson and Wright, 1984) and a latent-variable autoregressive (LV) model (Bennett et al., 2018).

The WGEN model is a daily two-part stochastic rainfall model where a Markov-chain is used to model the rainfall occurrence according to a number of wet-dry states (denoted by the order of the model) while the rainfall amount is sampled from a statistical distribution such as exponential (Woolhiser and Roldán, 1982), gamma (Richardson and Wright, 1984), mixed gamma (Yoo et al., 2005), generalised gamma (Papalexiou and Koutsoyiannis, 2012) and Weibull distribution (Wilks, 1989). This model is widely used and has shown a strong ability to reproduce wet-dry patterns along with good reproduction of totals and extremes (Soltani et al., 2000, Richardson and Wright, 1984). Variants of the Markov chain (first-order (2-state), second-order (4-state) and third-order (8-state)) were calibrated at each site and compared using the Bayesian Information Criterion (BIC) (Katz, 1981) to determine the best-performing variant at each site. The BIC is also used for the selection of rainfall amount distribution among 5 statistical distributions (exponential, gamma, mixed gamma, generalised gamma, and Weibull). This approach was adopted to mitigate potential concerns that the rainfall model being evaluated was conveniently 'simple' by default rather than optimally selected. In terms of the occurrence model, out of the 383 sites, 53 were selected as first-order models, 327 were second-order and 3 were third-order. In terms of the amount model, out of 383 sites, 26 were selected as exponential, 58 were gamma, 166 were mixed gamma, 3 were generalised gamma and 130 were Weibull distribution. Therefore, per month, WGEN has between 2 and 8 parameters for the occurrence process according to the Markov chain model order and an additional 1 to 5 parameters for the amount generation according to the various statistical distributions.

The LV model simulates daily rainfall by sampling from a latent Gaussian distribution. Values below zero of the latent-variable are truncated, while variables above zero are power-transformed to resemble the observed rainfall distribution (Rasmussen, 2013). A benefit of this model is that the wet-dry process and the amounts process come from the same distribution. The model has shown good performance in several case studies (Rasmussen, 2013, Bennett et al., 2018), but also some weakness in wet-dry patterns of multi-day totals (Bennett et al., 2018). The LV model is parsimonious and has 4 parameters per month (latent mean, latent standard deviation, power transform, temporal correlation) that are fitted to observed statistics (wet-day mean, wet-day standard deviation, proportion wet days and wet-day autocorrelation).

2.4. Hydrological model

A conceptual hydrological model is calibrated to the catchment's observed streamflow timeseries with ancillary observed meteorological timeseries as inputs (potential evapotranspiration derived from pan evaporation, temperature, solar radiation and/or other observations). Given that SRMs have many replicates to be evaluated, the hydrological framework is best suited to the use of conceptual hydrological models given their simulation speed (i.e. compared to physical models) and accessibility (compared to machine learning specifications). There is a wide range of conceptual hydrological models that vary in the specification of internal states and parameters (Abbott et al., 1986, Boughton, 2004, Liu et al., 2008, Perrin et al., 2003a, Moore, 2007). A variety of methods for calibrating hydrological models is available depending on the specification of the objective function (Thyer et al., 2009) and on the treatment of errors such as observed rainfall input errors, observed streamflow errors and structural model errors (Kuczera et al., 2006).

From Fig. 1, the hydrological evaluation adopted from Bennett et al. (2019) involves 'virtual-observed streamflow' rather than 'observed streamflow', where the term 'virtual' indicates streamflow derived from observed rainfall using a hydrological model. Using 'virtual-observed streamflow' enables attribution of discrepancies to the inputted stochastic rainfall since the hydrological model is common to the derived streamflow (Fig. 1). The framework does not obviate the need for

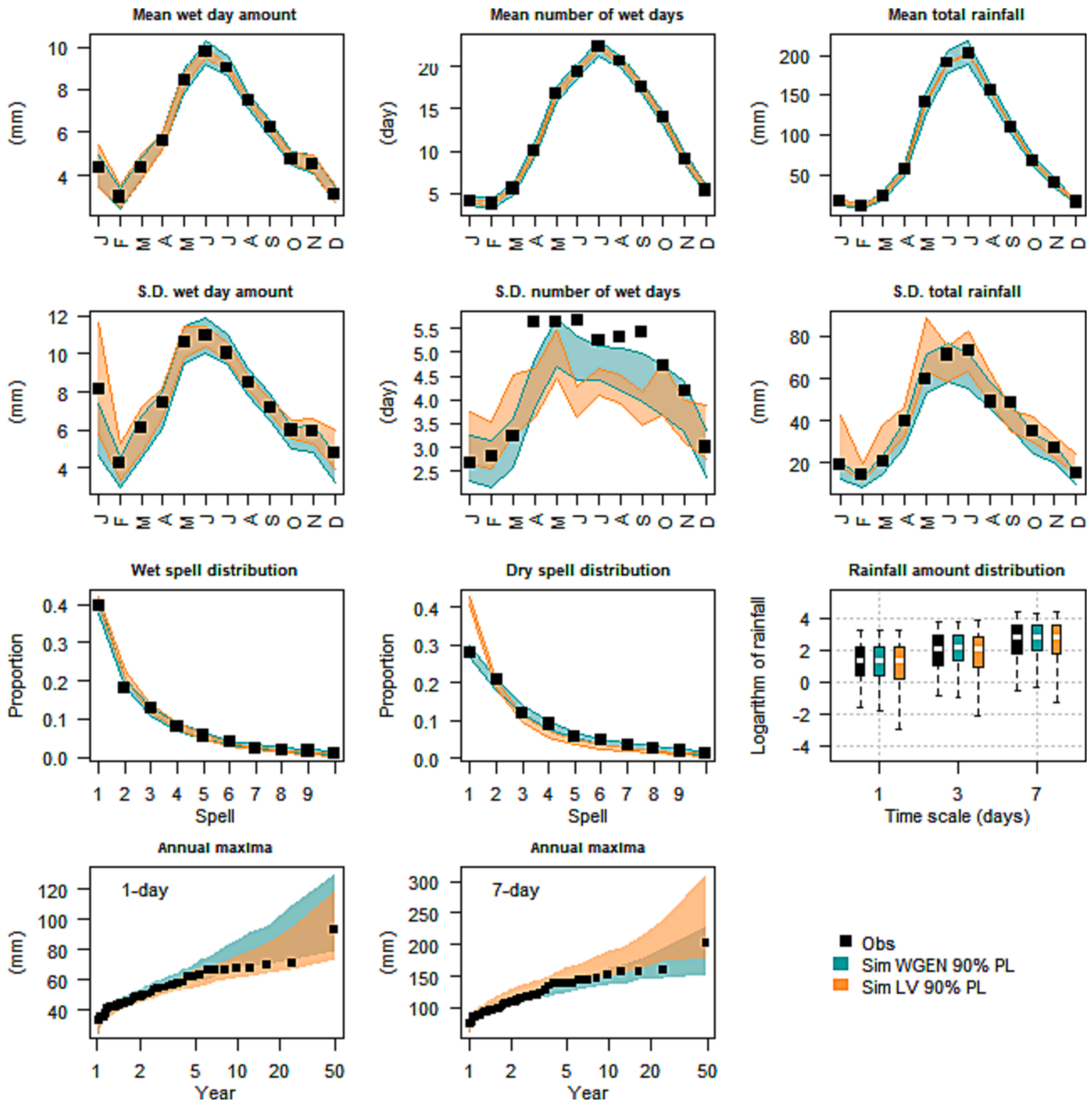


Fig. 3. WGEN and LV model performance in simulating monthly rainfall amounts, number of wet days, total rainfall, annual maxima and rainfall amount distribution at different temporal aggregation at catchment 610001, where standard deviations are denoted as 'S.D.' and percentile as 'perc.'. Coloured polygons and boxplot whiskers indicate the 90% probability limits of the simulated rainfall attributes from WGEN and LV model.

consideration of observed streamflow since it is required for calibration and evaluation of the hydrological model itself (Thyer et al., 2009). The hydrological model should be representative of the catchment of interest and utilise best-practice methods of calibration to ensure meaningful outputs from the ultimate application of the stochastic rainfall for decision making (Bennett et al., 2019). Although a well-calibrated hydrological model is a necessary condition for the hydrological evaluation the workflow for this task is not emphasized in Fig. 1 because it is outside the paper's scope, which is to evaluate the performance of the SRM.

The GR4J hydrological model (Perrin et al., 2003b) is used to generate streamflow from the respective stochastic rainfall and observed rainfall inputs. In this paper only one hydrological model is used for analyses to restrict the number of comparisons. GR4J is calibrated to the observed streamflow at each catchment using a two-year warmup period following the procedure proposed in Michel (1991) and considering the Nash-Sutcliffe Efficiency (NSE) (Nash and Sutcliffe, 1970).

2.5. Evapotranspiration

An issue with the framework, not previously raised, is that the streamflow model has two inputs: rainfall and evapotranspiration (Fig. 1). This paper has focussed on SRMs rather than the broader category of weather generators for sake of simplicity, since a model that stochastically generates both the rainfall and evapotranspiration involves more components and requires greater depth of analysis to isolate discrepancies. An issue arises when selecting evapotranspiration to be inputted alongside the stochastic rainfall: the virtual-observed streamflow will potentially have a daily cross-correlation between observed rainfall and observed evapotranspiration (typically negative – that is,

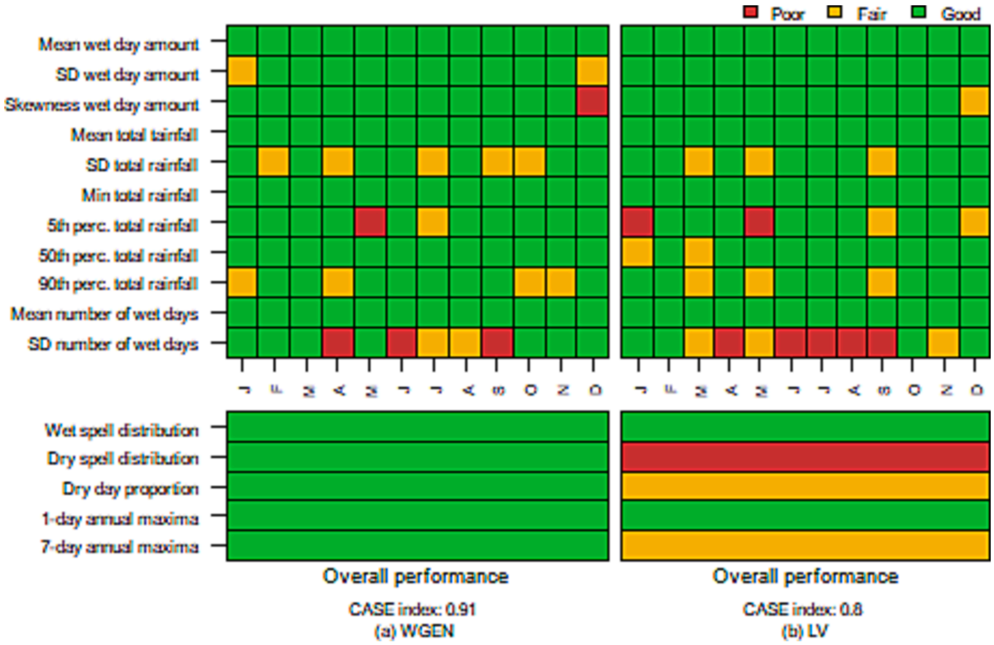


Fig. 4. Evaluation model output of rainfall performance for (a) WGEN model and (b) LV model at catchment 610001, where standard deviations are denoted as ‘S.D.’ and percentile as ‘perc.’.

higher rainfall days are associated with cooler conditions and less evapotranspiration), whereas the stochastic rainfall will not have daily cross-correlation with observed evapotranspiration.

To test whether daily-correlated evapotranspiration is a significant influence on streamflow generation, a bootstrap study was conducted by splitting and swapping the first and second half of the daily evapotranspiration records. This approach preserves the evapotranspiration monthly totals, annual totals, seasonal pattern, and daily autocorrelation (excepting the one breakpoint) while breaking the daily cross-correlation with rainfall. The flow-duration curve of the bootstrap replicate was compared to the original instance that preserves the rainfall-evapotranspiration cross-correlation. In the majority of

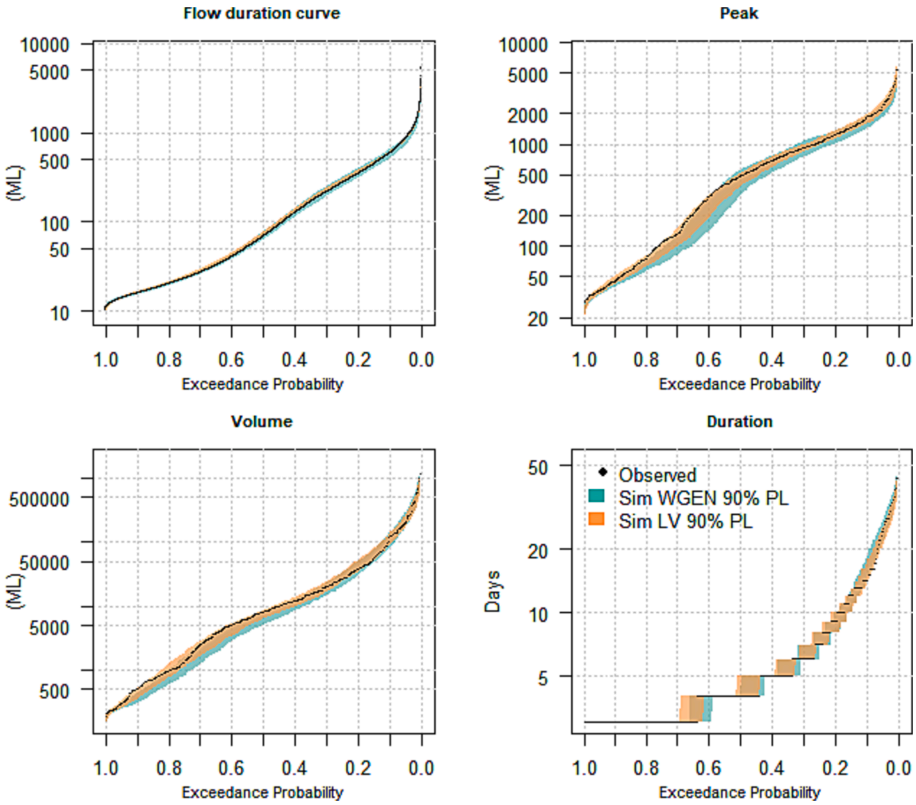


Fig. 5. WGEN and LV model performance in simulating the flow duration curve, flow event peaks, volume, and duration at catchment 610001. Coloured polygons indicate the 90% probability limits of the simulated runoff attributes from WGEN and LV model.

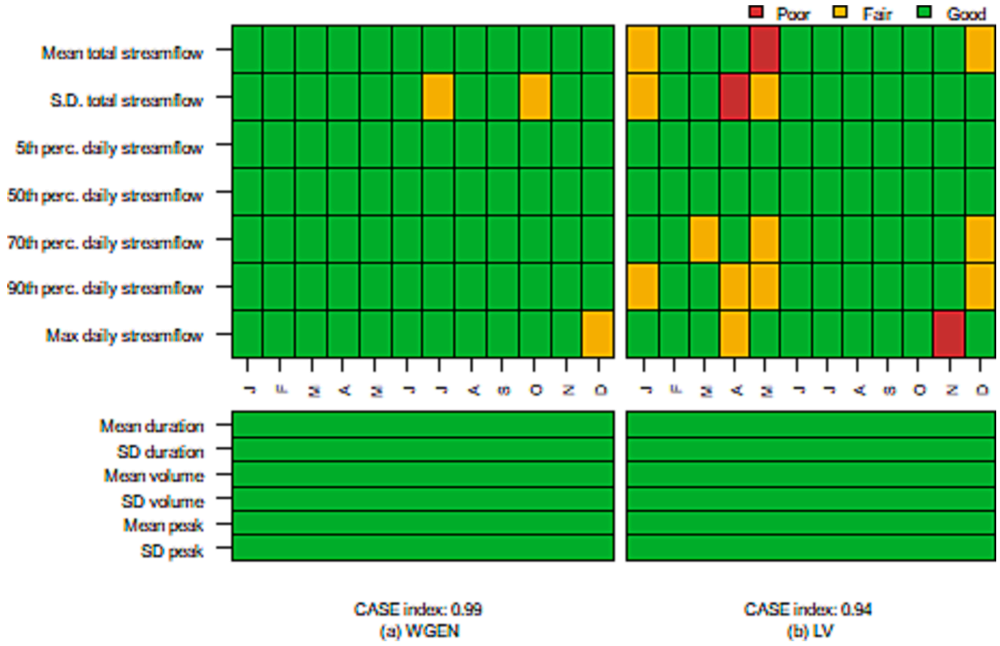


Fig. 6. Evaluation model output of streamflow performance for (a) WGEN model and (b) LV model at catchment 610001, where standard deviations are denoted as 'S.D.' and percentile as 'perc.'.

instances this procedure had minimal impact on the high flows (indicated by a low mean absolute relative error), but at numerous sites in Australia there was a difference to the lower flows (see Supplementary Material C). The impact of these differences depends on whether the evaluation is absolute or relative, given the large range of flows. Due to potential interest in the lower tail of the flow duration curve and to enable commensurate comparison at all sites, the outcome of this preliminary investigation is that both the stochastically simulated and the virtual-observed streamflow should also use the split evapotranspiration timeseries (i.e. without daily cross-correlation to the observed rainfall) to ensure consistency of comparison with the stochastic timeseries.

3. Results

The evaluation framework (Section 2.1) is employed to systematically assess the performance of the WGEN and LV model in simulating rainfall and streamflow. Three use cases of the framework in assessing SRM performance are established including individual catchment evaluation (Section 3.1), evaluation at a group of catchments level (Section 3.2) and identifying the relationship between rainfall and streamflow performance of SRMs (Section 3.3). While the quality of performance from rainfall evaluation to streamflow is maintained at some catchments, it deteriorates at others. The calibrated WGEN model and LV models are shown to have a majority of 'overall good' performance in reproducing target rainfall statistics for all evaluated catchments, yet this does not guarantee similar performance in terms of streamflow.

3.1. SRM evaluation for individual catchments

In this section, catchment 610001 is selected as a representative example that demonstrates consistent high-quality performance for both rainfall and streamflow metrics. Catchment 12043000 is selected to demonstrate anomalous streamflow performance despite 'overall good' rainfall evaluation.

3.1.1. Instance of 'good' streamflow

Catchment 610001 in Western Australia provides a representative example for the desirable case of stochastic rainfall that is 'overall good' translating to generated streamflow that is also 'overall good'. Fig. 3 shows the performance in simulating rainfall for both the WGEN and LV model. Both models have 'good' performance in preserving the monthly wet-day amount, number of wet days and total rainfall. However, neither model is perfect showing some 'poor' performance in the standard deviation of the number of wet days (April – September).

The rainfall attributes illustrated in Fig. 3 are summarised in the form of a heatmap (Fig. 4) for a wide set of 16 evaluated attributes. From this comparison the WGEN model outperforms the LV model in preserving the dry spell distribution, the proportion of dry days at different temporal scales, and the 7-day annual maxima.

Fig. 5 shows WGEN and LV model performance in simulating the entire flow duration curve, as well as the distribution of flow event metrics (i.e. peak, volume, and duration). As evidenced by the flow duration curve, the two models reproduce the streamflow at all quantiles with relatively small variation.

Fig. 6 summarises the performance of the WGEN and LV models in simulating 13 streamflow attributes including the mean, standard deviation of the monthly total streamflow and daily streamflow percentiles including the 5th, 50th, 70th, 90th percentile, the maximum of daily rainfall and the mean and the standard deviation of the flow event peak, volume, and duration. It is evident that both models show 'overall good' performance in simulating streamflow in catchment 610001.

3.1.2. Instances of 'poor' streamflow

Catchment 12043000 in Washington provides a representative example for the case of stochastic rainfall that is evaluated as 'overall good' yet translates to 'overall poor' streamflow. Fig. 7 shows the rainfall performance of the WGEN and LV models. Both models have typically 'good' performance in preserving the monthly wet day amount, number of wet days and total rainfall as well as the wet-day distribution at 1-, 3-, and 7-day aggregations. The performance in simulating the standard deviation of both the monthly number of wet days and monthly totals is mixed in both models. The WGEN model is shown to capture the 1-day annual maxima well. However, it fails to reproduce the 7-day annual maxima. Whereas the LV model exhibits the opposite performance in terms of reproducing the 1-day and 7-day annual maxima of rainfall. Fig. 8 summarises the performance across all metrics and shows that although both have 'overall good' performance (CASE indices are 0.75 and 0.79 for WGEN and LV models, respectively), the WGEN model

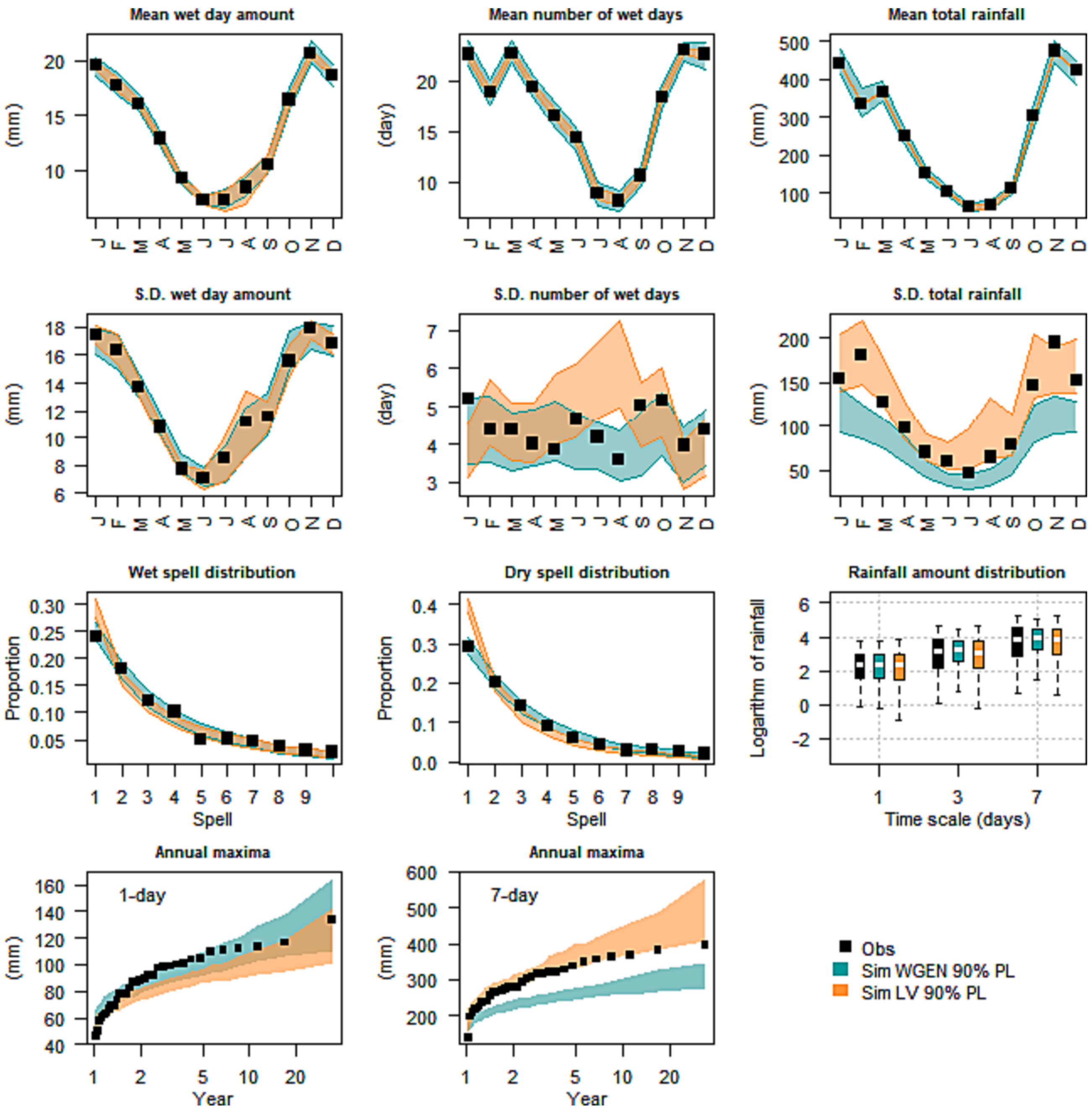


Fig. 7. WGEN and LV model performance in simulating monthly rainfall amounts, number of wet days, total rainfall, annual maxima and rainfall amount distribution at different temporal aggregation at catchment 12043000, where standard deviations are denoted as ‘S.D.’ and percentile as ‘perc.’. Coloured polygons and boxplot whiskers indicate the 90% probability limits of the simulated rainfall attributes from WGEN and LV model.

has difficulty in reproducing the standard deviation of the monthly total rainfall and the 7-day annual maxima, whereas the LV model has poorer performance in reproducing the wet and dry spell distributions and 1-day annual maxima.

Fig. 9 shows the WGEN and LV model performance in simulating the entire flow duration curve as well as flow event metrics (peak, volume, and duration). From the flow duration curve a bias in the upper tail and lower tail is noticeable for the WGEN model. Consequently, the WGEN model fails to preserve all 3 flow event statistics. Fig. 10 provides a heatmap summary of the performance of the two models in terms of streamflow which shows that the WGEN model has ‘overall poor’ streamflow performance while LV model has ‘overall good’ streamflow performance.

Reliably diagnosing the source of discrepancies in streamflow is a point for further discussion (Section 3.3.2). In this instance for the WGEN model, ‘poor’ performance in the standard deviation of monthly total rainfall and the 7-day annual maxima seems to be associated with ‘overall poor’ performance in streamflow. While in the instance for the LV model, having poor performance on the wet and dry distributions does not lead to ‘overall poor’ streamflow performance.

3.2. *Contrasting performance between the two SRMs*

A benefit of the evaluation method is it can be aggregated in numerous ways (e.g. across sites, statistics, or models) to identify common features of performance. In Fig. 11 and Fig. 12, the evaluation method is aggregated over all statistics at a site to summarise the performance of the SRMs in simulating rainfall and streamflow by a single

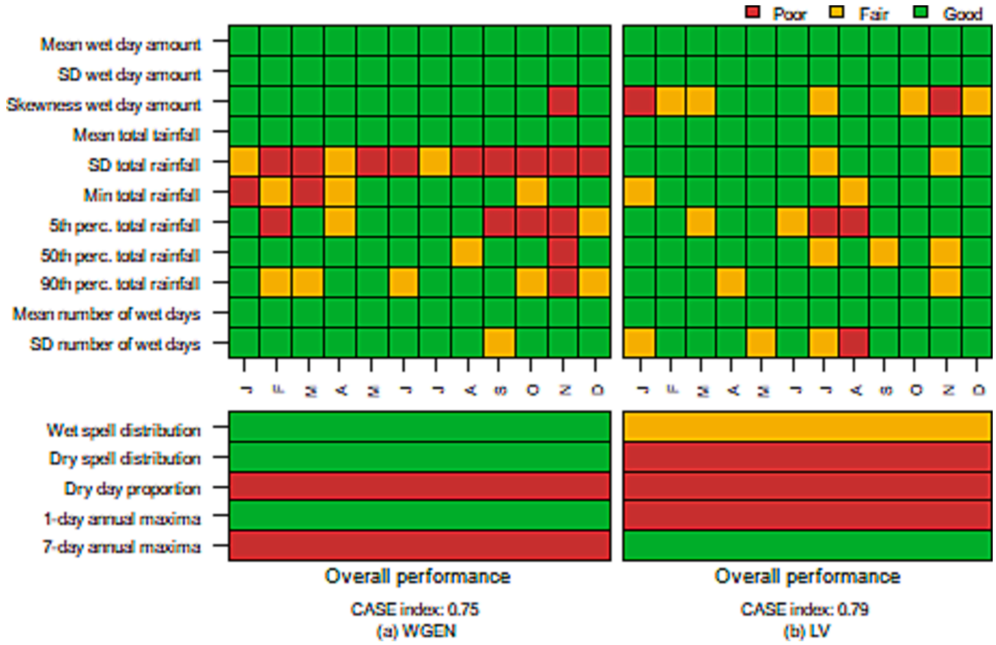


Fig. 8. Evaluation model output of rainfall performance for (a) WGEN model and (b) LV model at catchment 12043000, where standard deviations are denoted as ‘SD’ and percentile as ‘perc.’.

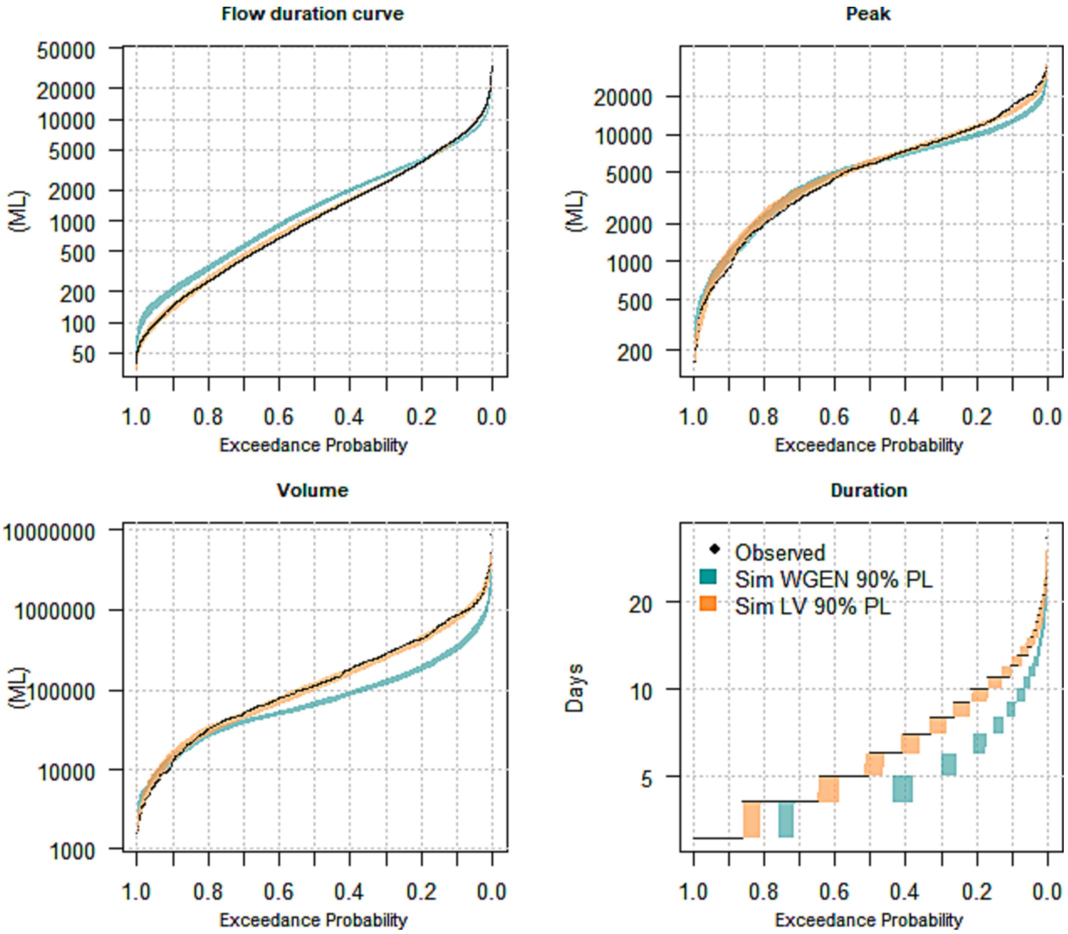


Fig. 9. WGEN and LV model performance in simulating the flow duration curve, exceedance probability of flow event peak, volume, and duration at catchment 12043000. Coloured polygons indicate the 90% probability limits of the simulated streamflow attributes from WGEN and LV model.

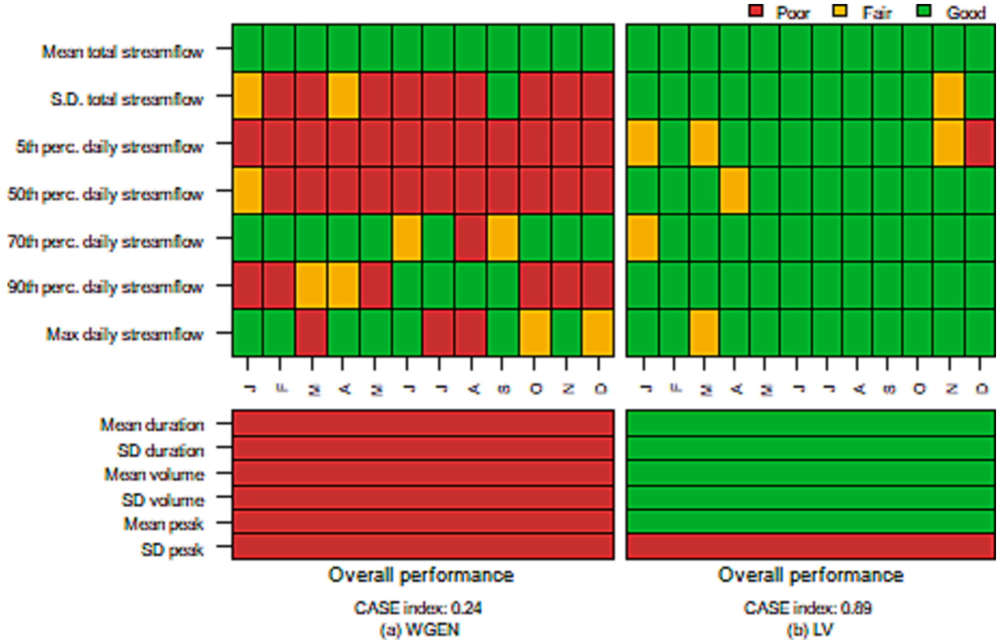
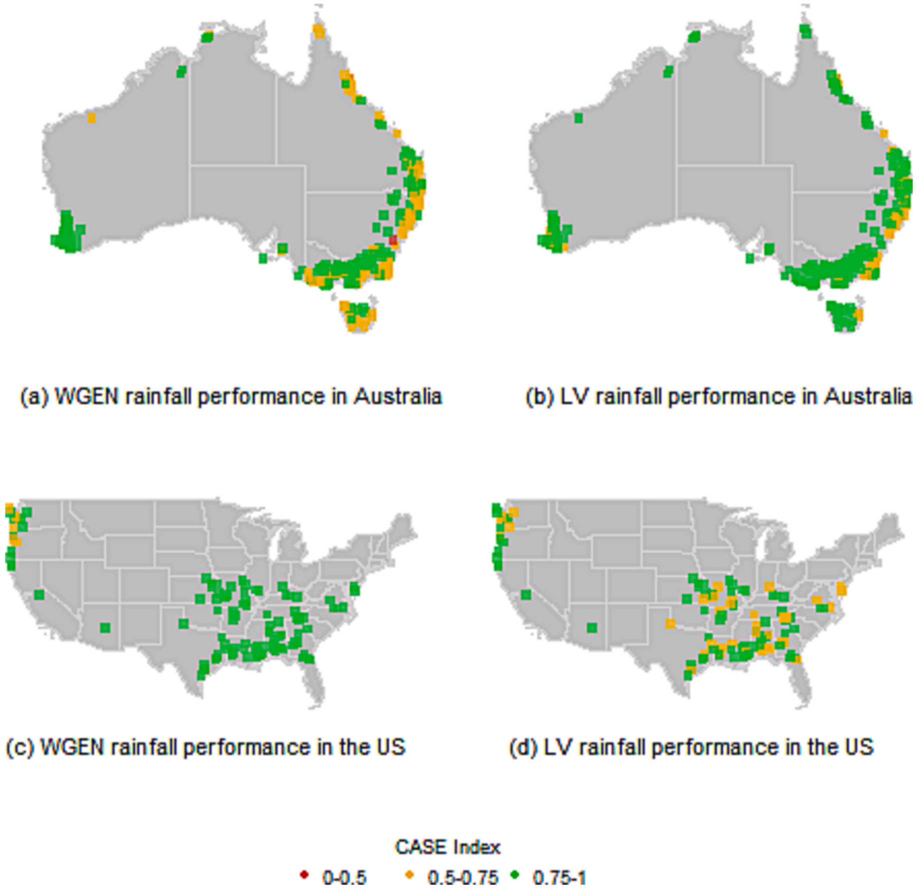


Fig. 10. Evaluation model output of streamflow performance for (a) WGEN model and (b) LV model at catchment 12043000, where standard deviations are denoted as 'SD' and percentile as 'perc.'.


(a) WGEN rainfall performance in Australia

(b) LV rainfall performance in Australia

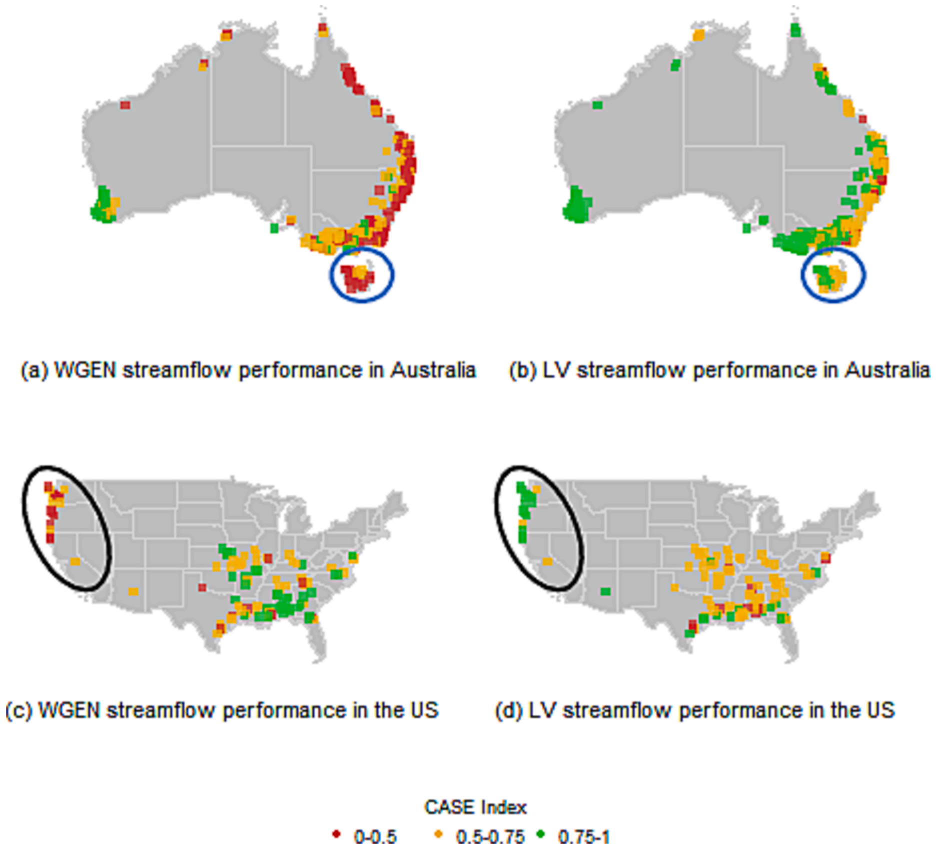
(c) WGEN rainfall performance in the US

(d) LV rainfall performance in the US

Fig. 11. Rainfall performance of the WGEN and LV models in Australian and US catchments.

value for each catchment. Fig. 11 shows the respective rainfall performance of the WGEN and LV models in Australia and the US. For Australia catchments, the WGEN model shows poorer performance when compared to the LV model. Whereas for the US catchments, the LV model shows poorer performance than the WGEN model.

Fig. 12 shows the respective streamflow performance of the WGEN and LV models in Australia and the US. It can be observed that comparatively poorer performing catchments are distributed across



(a) WGEN streamflow performance in Australia

(b) LV streamflow performance in Australia

(c) WGEN streamflow performance in the US

(d) LV streamflow performance in the US

Fig. 12. Streamflow performance of the WGEN and LV models in Australian and US catchments. The blue circles indicate the Tasmanian (Australia) catchment group while black circles indicate the West Coast (US) catchment group. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

different regions, but overall, that the LV model outperforms the WGEN model in simulating streamflow. For the highlighted regions in Fig. 12 (Tasmania in Australia, Upper West Coast in the US) 'overall poor' performance is determined for WGEN model (CASE index: 0.42 for Tasmania and 0.46 for Upper West Coast), and 'overall good' performance is seen for the LV model (CASE index: 0.76 for Tasmania and 0.84 for Upper West Coast).

It is possible to aggregate the evaluation over any grouping of sites to identify any common patterns of performance. Fig. 13 shows the rainfall and streamflow performance of the WGEN and LV models for the upper western US coast. Both models have 'overall good' performance in simulating rainfall despite some statistics not performing well (e.g. standard deviation of total rainfall and 7-day annual maxima for WGEN and wet day amount skewness and 1-day annual maxima for the LV model). Whereas WGEN has 'overall poor' performance in simulating multiple streamflow statistics, the LV model has 'overall good' performance.

3.3. Relationship between rainfall and streamflow performance

The deteriorated performance of streamflow relative to rainfall is prevalent in the WGEN model but is also noticeable at some sites for the LV model. This section further explores the relationship between rainfall and streamflow performance by analysing the sensitivity of evaluation thresholds and by analysing selected statistics.

3.3.1. Inconsistency between rainfall and streamflow performance

The CASE index is the numerical result for a single catchment by aggregating over all statistics (see Section 2.1) and enables broad comparison of streamflow-to-rainfall performance across models and regions. Comparing the CASE index for rainfall to the CASE index for streamflow (Fig. 14) shows that both models have degraded performance for streamflow (with more sites below the 1:1 line) and a stronger deterioration in performance for the WGEN model (denoted by the blue circles) in both Australian and US contexts.

It is instructive to test the sensitivity of the parameters used in the evaluation model to identify whether the apparent 'poor' performance is an artefact of the threshold selection or is indeed consistent regardless of their value. The results from Fig. 14 have assigned the values 1, 0.5, 0 to the respective classifications of 'good', 'fair', 'poor' performance. Table 2 shows the outcome when using alternative weights that represent increasingly stricter categorisation, summarised by the number (and percentage) of sites determined as exhibiting 'overall poor' streamflow performance despite having 'overall good' rainfall performance. While the stricter categorisations shown in Table 2 results in more sites showing 'overall poor' rainfall and streamflow performance, the pattern is consistent, that the WGEN model exhibits worse performance than the LV model and that Australian sites perform worse than US sites in terms of the WGEN model, while the opposite is true for the LV model. For the example of Australian catchments, when 'fair' is set as 0.25, 24 out of 135 (18%) of catchments show the WGEN model has 'overall poor' streamflow performance and when 'fair' is set as 0 this increases to 26% of the catchments. The LV model's performance is demonstrated to be far less sensitive to the classification with 1% of sites in Australia and no sites in the US showing poor performance. The analysis demonstrates that whether 'overall good' modelled rainfall translates to 'overall poor' modelled streamflow depends strongly on the specific region and model and less on the thresholds used for evaluation.

3.3.2. Indication of 'overall poor' streamflow performance

The relationship between the specific metrics of rainfall performance can be analysed to determine if there is a strong association with poor

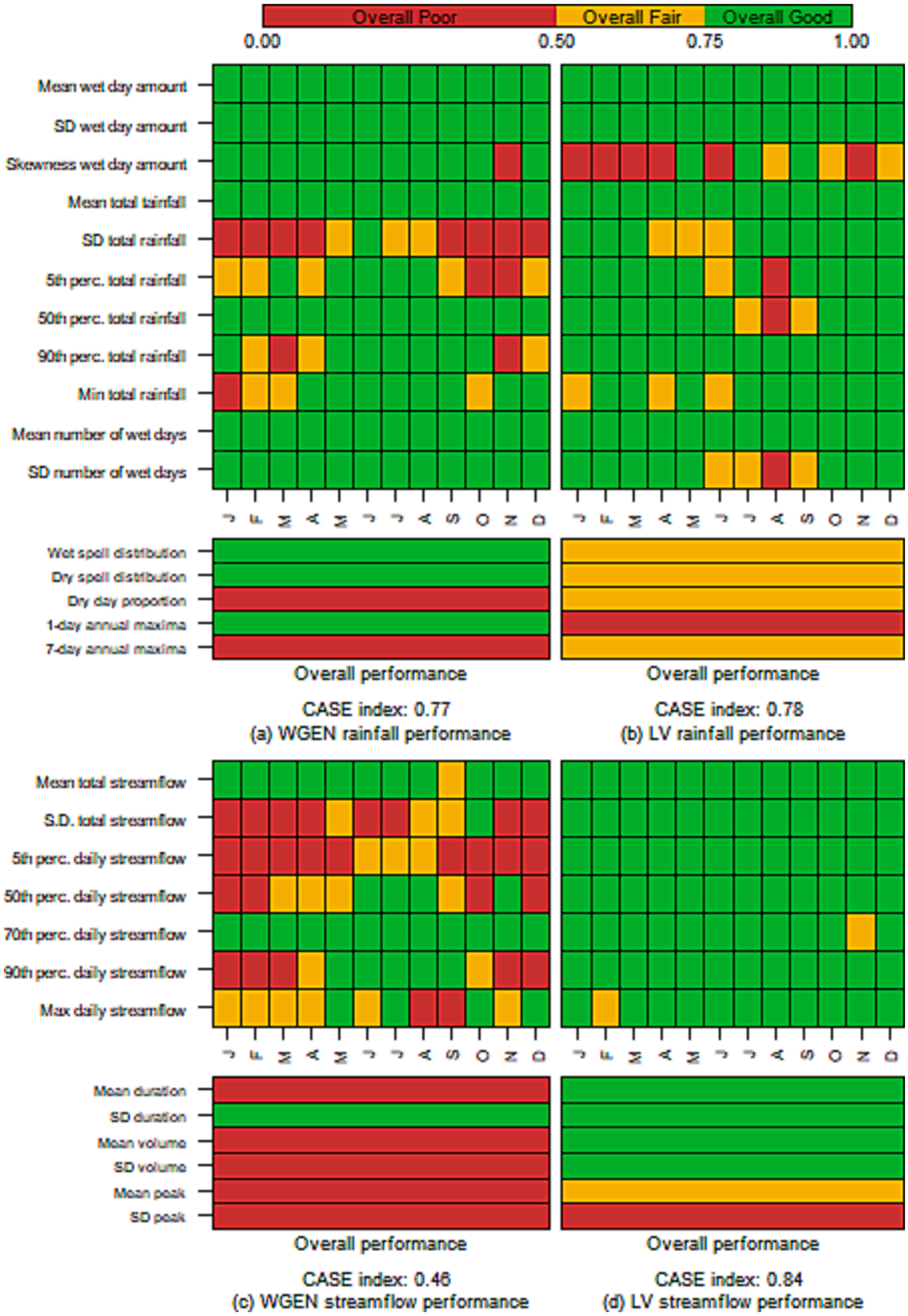


Fig. 13. Model performance summary for the West Coast catchment group for (a) WGEN model rainfall, (b) LV model rainfall, (c) WGEN model streamflow, and (d) LV model streamflow.

streamflow performance. Preliminary analyses (Section 3.1.2) identified that the standard deviation of monthly total rainfall and the 7-day annual maxima as candidate metrics for further analysis. Sub-setting the aggregation to those two candidate metrics derives a different value of the CASE index targeting rainfall variability at the monthly timescale and aggregation of extreme rainfall at the annual timescale. Fig. 15 compares the newly defined rainfall performance to the streamflow and shows evidence of a relationship between the standard deviation of monthly total rainfall, the 7-day annual maxima and streamflow (indicated by the regression line). Specifically, at sites showing 'poor' representation of monthly rainfall variability and 7-day annual maxima, there is higher likelihood they will have 'poor' streamflow performance. While both models show an association, the relationship between 'poor' standard deviation of monthly total rainfall and 'poor' streamflow performance is more consistent than that of the 7-day annual maxima. This is a useful outcome to help a stochastic rainfall modeller target improvements that are not otherwise obvious from a rainfall-only analysis and are not otherwise simple to isolate among the myriad statistics.

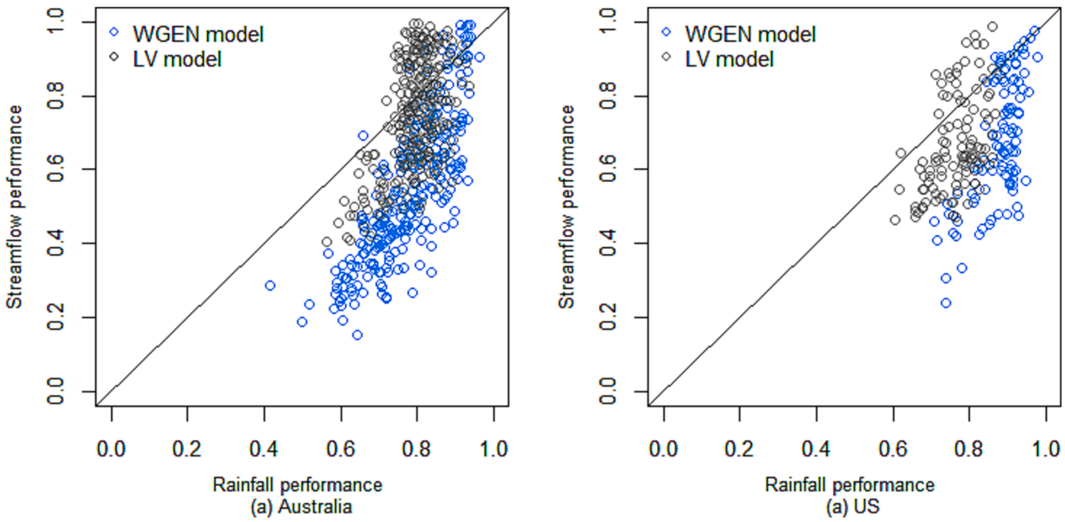


Fig. 14. WGEN model and LV model performance CASE index in simulating rainfall and streamflow in (a) Australia and (b) US when 'good', 'fair', 'poor' are defined as 1, 0.5, and 0.

Table 2
Number of catchments where the WGEN and LV models show 'overall good' rainfall performance (CASE index >0.75) translates to 'overall poor' streamflow performance (CASE index <0.5) over the number of catchments where WGEN and LV show 'overall good' rainfall performance.

CASE index setting	AU		US	
(Good – Fair – Poor)	WGEN	LV	WGEN	LV
1 – 0.50 – 0	32/173 (18%)	2/233 (1%)	12/100 (12%)	2/65 (3%)
1 – 0.25 – 0	24/135 (18%)	3/190 (2%)	12/94 (13%)	2/46 (4%)
1 – 0 – 0 (i.e., only good, or poor)	29/111 (26%)	4/138 (3%)	14/90 (16%)	2/23 (9%)

4. Discussion

This paper emphasized the importance of hydrological evaluation using a systematic comparison of stochastically generated rainfall in terms of derived streamflow metrics. The contribution and limitations of the generated results to the research aim are discussed in the following sections.

4.1. A systematic comparison of hydrological evaluation with rainfall-only evaluation

This paper has implemented a comparison of two SRMs in terms of their rainfall and streamflow performances spanning many locations (383) and climates. At a non-trivial portion of the catchments, the streamflow performance was significantly worse than the rainfall performance. The implication is that rainfall-only model evaluation can misrepresent the ultimate performance of a model, and in this case, overstate the performance. As a representative example, Fig. 8 showed the evaluation of the WGEN model was 'good' for the majority of rainfall metric evaluations, yet from Fig. 10 the resulting streamflow metrics were typically 'poor'. While there were indeed some rainfall metrics that were 'poor' (Fig. 8, SD total rainfall) it is not immediately obvious that these metrics should warrant special attention, and as the counter example, the LV model had some rainfall metrics that were 'poor' (Fig. 8, dry spell, dry proportion, 1-day annual maxima), yet the resulting streamflow was typically 'good'. This indicates that there are features of the stochastic rainfall that evade the (otherwise extensive) rainfall-based evaluation yet led to tangible discrepancies in the runoff. This observation is however dependent on the region and model, making it harder to *a priori* determine instances requiring additional care during calibration (Fig. 12).

An incidental outcome of the results is that the LV model performed better than the WGEN model (Fig. 12), showing the potential of the framework for consistent analysis that could be applied across a wider range of models. An intriguing question arises as to why the LV model outperformed the WGEN model. Further investigation identified that the WGEN model has a slightly stronger correlation between 'overall poor' performance in the standard deviation of monthly total rainfall to 'overall poor' streamflow performance compared to the LV model (Fig. 15). However, establishing a causal relationship between one rainfall attribute and overall streamflow performance is impractical because even with a plethora of metrics, there may always remain elusive features of rainfall not being scrutinized – as an esoteric example, the correlation of the rainfall antecedent to extreme events. As a more concrete example, it is not possible to determine from a rainfall-only evaluation whether it is important to seek improvement in the performance of the skewness of wet day amounts (see Fig. 13, the LV model has this deficiency, yet it did not affect streamflow), and as a result many false diagnoses for 'improved' calibration could be pursued. By analysing a wide range of sites, this paper has demonstrated the importance of streamflow-based evaluation for stochastic rainfall models due to its ability to magnify impactful features of the rainfall not otherwise visible to rainfall-based evaluation, thereby providing more a holistic evaluation.

4.2. Limitations of the evaluation framework

A potential limitation of the framework is that there are many assumptions underlying the evaluation, for example, the set of metrics used, or weighting applied to individual metrics. However, with traditional evaluation these assumptions are also present and are poorly articulated, which makes them less amenable to testing. The evaluation framework here has utilised a sensitivity analysis to identify consistency in its conclusion, specifically, the weight assigned to the 'fair' performance category (from 0.5 to 0.25, to 0) and the resulting analysis (Table 2) demonstrated the LV model performed consistently better than the WGEN model. Similar sensitivity analyses could be conducted on the mix of other statistics or weighting assigned to metrics used in the averaging process (e.g. to emphasize extremal behaviour more strongly over mean behaviour). While there will always be a level of subjectivity in the construction of an evaluation model and metrics, the framework's

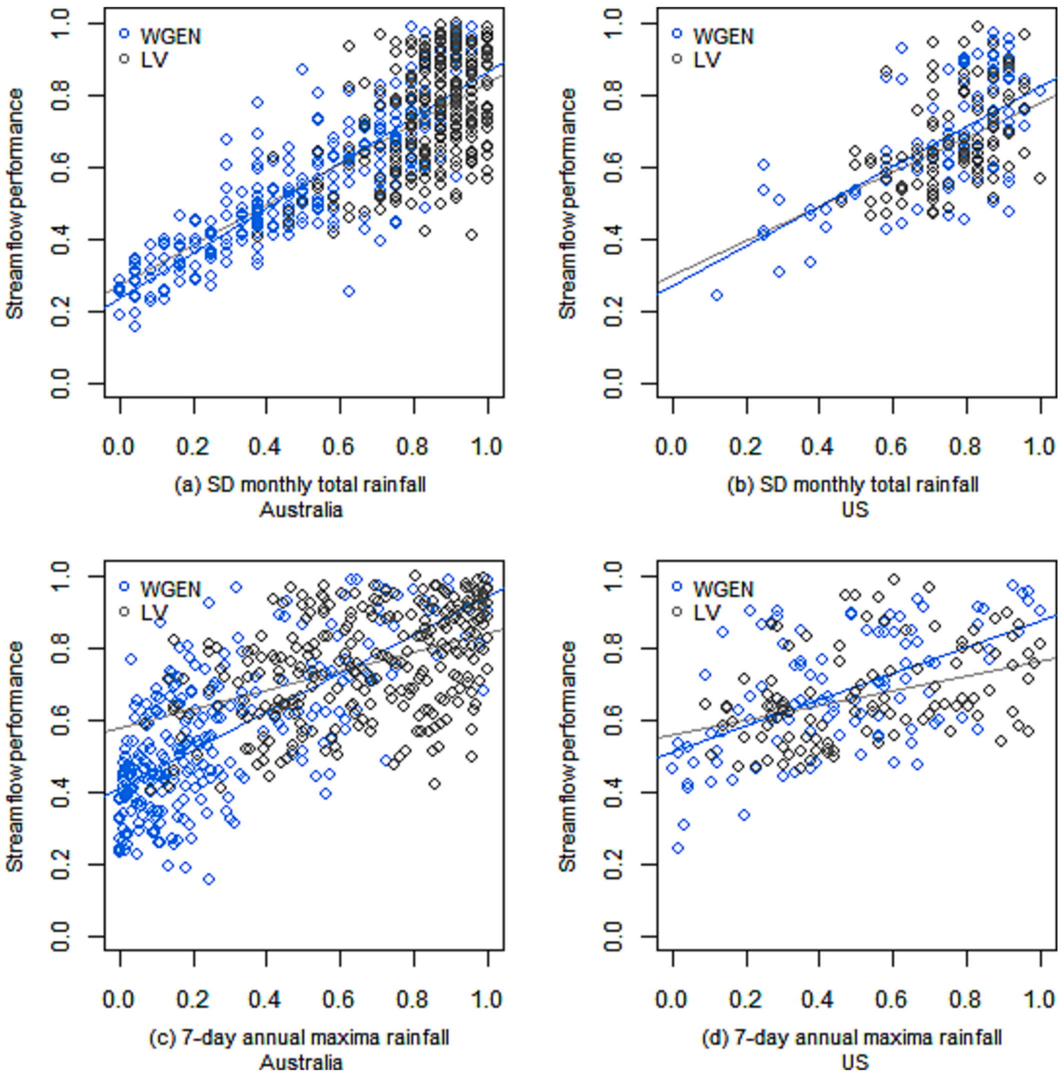


Fig. 15. WGEN model and LV model performance CASE index in simulating the standard deviation of monthly total rainfall (a and b), 7-day annual maxima rainfall (c and d) against aggregate streamflow performance in Australia and US when 'good', 'fair', 'poor' are defined as 1, 0.5, and 0 with corresponding regression line.

transparency is beneficial.

Another limitation of the framework rests with the requirement for suitable calibration of the SRM and hydrological models. For the naïve case of an inadequately calibrated SRM with consistently 'poor' rainfall, there is little point proceeding with detailed evaluation. The streamflow evaluation of SRMs is more meaningful for cases where despite 'best' efforts of calibration, some aspects of a model perform poorly relative to others (e.g. in certain months) and their relative impact on streamflow is unclear. For the case of a 'poorly' calibrated hydrological model or on the impact of different hydrological models on the evaluation, it is an important question, albeit beyond the scope of this paper. As the framework is entirely relative, it is conceivable to use any given set of hydrological model parameters for identifying discrepancies (even from an uncalibrated model or by applying model parameters from another location). For example, it might be of interest to specify parameters with different degrees of storage, yielding either fast or slow hydrological response, that influence the degree of damping and relative importance of extremes in the rainfall model. However, depending on the choices made for the hydrological model, they have the potential to either amplify or dampen discrepancies in generated stochastic timeseries and it is unclear whether using (multiple) alternatives for the hydrological model aids or distracts the evaluation of the stochastic model. For the interpretation to be meaningful at a location of interest, it is best to have representative parameters of the hydrological model.

5. Conclusions

This paper has demonstrated the need for hydrological evaluation through a systematic comparison of the performance of SRMs in terms of rainfall metrics and their subsequently generated streamflow metrics after rainfall-runoff transformation. The structured evaluation of the performance of two stochastic rainfall models (a two-state Markov model – WGEN and a Gaussian latent variable autoregressive model – LV) in 383 catchments across both Australia and the US provides a strong evidence base that potentially 'good' modelled rainfall does not necessarily translate to 'good' modelled streamflow. An extensive set of rainfall and streamflow metrics was utilised, relating to many aspects of the corresponding distributions across many aggregate scales: moments, extremes, wet-dry patterns, event-based metrics (peaks, volume, duration) and distribution quantiles. Compared to the performance of the stochastic rainfall models in terms of rainfall-only metrics, the hydrological evaluation indicates that both models show a decrease in streamflow performance. The framework also shows that the LV model outperforms the WGEN model in simulating streamflow consistently in both countries and was able to identify rainfall metrics associated with streamflow discrepancies, including the variability of monthly rainfall amounts and 7-day rainfall extremes. Rigorous evaluation of stochastic rainfall models requires that both rainfall and streamflow metrics are analysed for a comprehensive understanding of model performance.

Author contributions

THTN: Conceptualization, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, **BB and ML:** Conceptualization, Methodology, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

Thien Huy Truong Nguyen was supported by an Australian Government Research Training Program Scholarship.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhydrol.2023.130381.

References

Abbott, M.B., Bathurst, J.C., Cunge, J.A., O'connell, P.E., Rasmussen, J., 1986. An introduction to the European Hydrological System — Systeme Hydrologique Europeen, "SHE", 1: History and philosophy of a physically-based, distributed modelling system. J. Hydrol. 87, 45–59.

Addor, N., Newman, A.J., Mizukami, N., Clark, M.P., 2017. The CAMELS data set: catchment attributes and meteorology for large-sample studies. Hydrol. Earth Syst. Sci. 21, 5293–5313.

Adnan, R.M., Liang, Z., Trajkovic, S., Zounemat-Kermani, M., Li, B., Kisi, O., 2019. Daily streamflow prediction using optimally pruned extreme learning machine. J. Hydrol. 577.

Baxevani, A., Lennartsson, J., 2015. A spatiotemporal precipitation generator based on a censored latent Gaussian field: Spatiotemporal Stochastic Generator. Water Resour. Res. 51, 4338–4358.

Bennett, B., Thyer, M., Leonard, M., Lambert, M., Bates, B., 2018. A comprehensive and systematic evaluation framework for a parsimonious daily rainfall field model. J. Hydrol. 556, 1123–1138.

Bennett, B., Thyer, M., Leonard, M., Lambert, M., Bates, B., 2019. A virtual hydrological framework for evaluation of stochastic rainfall models. Hydrol. Earth Syst. Sci. 23, 4783–4801.

Beven, K.J., 2012. Rainfall-Runoff Modelling: The Primer. Wiley-Blackwell, Chichester, West Sussex.

Blazkova, S., Beven, K. 2002. Flood frequency estimation by continuous simulation for a catchment treated as ungauged (with uncertainty). *Water Resour. Res.,* 38, 14-1-14-14.

Boughton, W., 2004. The Australian water balance model. Environ. Model. Softw. 19, 943–956.

Boughton, W., Hill, P. 1997. A Design Flood Estimation Procedure Using Data Generation and a Daily Water Balance Model. Cooperative Research Centre for Catchment Hydrology.

Chowdhury, A.F.M.K., Lockart, N., Willgoose, G., Kuczera, G., Kiem, A.S., Manage, N.P., 2017. Development and evaluation of a stochastic daily rainfall model with long-term variability. Hydrol. Earth Syst. Sci. 21, 6541–6558.

Cowpertwait, P.S.P., 2006. A spatial–temporal point process model of rainfall for the Thames catchment, UK. J. Hydrol. (Amsterdam) 330, 586–595.

Croke, B.F.W., Andrews, F., Jakeman, A.J., 2006. IHACRES Classic Plus: A redesign of the IHACRES rainfall-runoff model. Environ. Model. Softw. 21, 426–427.

Do, H.X., Westra, S., Leonard, M., 2017. A global-scale investigation of trends in annual maximum streamflow. J. Hydrol. 552, 28–43.

Evin, G., Favre, A.-C., Hingray, B., 2018. Stochastic generation of multi-site daily precipitation focusing on extreme events. Hydrol. Earth Syst. Sci. 22, 655–672.

Gao, C., Booij, M.J., Xu, Y.-P., 2020. Development and hydrometeorological evaluation of a new stochastic daily rainfall model: Coupling Markov chain with rainfall event model. J. Hydrol. (Amsterdam) 589, 125337.

Grimaldi, S., Volpi, E., Langousis, A., Michael Papalexiou, S., Luciano De Luca, D., Piscopia, R., Nerantzaki, S.D., Papacharalampous, G., Petroselli, A., 2022. Continuous hydrologic modelling for small and ungauged basins: A comparison of eight rainfall models for sub-daily runoff simulations. J. Hydrol. 610.

Gupta, V.K., Waymire, E.C., 1993. A statistical analysis of mesoscale rainfall as a random cascade. J. Appl. Meteorol. 32, 251–267.

Jeffrey, S.J., Carter, J.O., Moodie, K.B., Beswick, A.R., 2001. Using spatial interpolation to construct a comprehensive archive of Australian climate data. Environ. Model. Softw. 16, 309–330.

Katz, R.W., 1977. Precipitation as a chain-dependent process. J. Appl. Meteorol. 16, 671–676.

Katz, R.W., 1981. On some criteria for estimating the order of a Markov chain. Technometrics 23.

Kingston, G.B., Maier, H.R., Lambert, M.F., 2005. Calibration and validation of neural networks to ensure physically plausible hydrological modeling. J. Hydrol. 314, 158–176.

Kuczera, G., Kavetski, D., Franks, S., Thyer, M., 2006. Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. J. Hydrol. 331, 161–177.

Ladson, A.R., Brown, R., Neal, B., Nathan, R., 2013. A standard approach to baseflow separation using the Lyne and Hollick filter. Aust. J. Water Resour. 17.

Lamb, R.O.B., 2005. Rainfall-Runoff Modeling for Flood Frequency Estimation. John Wiley & Sons Ltd, Chichester, UK.

Leonard, M., Lambert, M.F., Metcalfe, A.V., Cowpertwait, P.S.P., 2008. A space-time Neyman-Scott rainfall model with defined storm extent. Water Resour. Res. 44. W09402-n/a.

Leonard, M., Westra, S., Phatak, A., Lambert, M., Van Den Hurk, B., Mcinnes, K., Risbey, J., Schuster, S., Jakob, D., Stafford-Smith, M., 2014. A compound event framework for understanding extreme impacts. Wiley Interdiscip. Rev. Clim. Chang. 5, 113–128.

Leonard, M. 2010. *A Stochastic Space-Time Rainfall Model for Engineering Risk Assessment.*.

Linsley, R., Crawford, N., 1974. Continuous simulation models in urban hydrology. Geophys. Res. Lett. 1, 59–62.

Liu, Z.-Y., Tan, B.-Q., Tao, X., Xie, Z.-H., 2008. Application of a distributed hydrologic model to flood forecasting in catchments of different conditions. J. Hydrol. Eng. 13, 378–384.

Lyne, V. & Hollick, M. 1979. Stochastic time-variable rainfall-runoff modelling. In: Proceedings of the Hydrology and Water Resources Symposium. *Institution of Engineers National Conference Publication*, 89-92.

Mcinerney, D., Thyer, M., Kavetski, D., Bennett, B., Lerat, J., Gibbs, M., Kuczera, G., 2018. A simplified approach to produce probabilistic hydrological model predictions. Environ. Model. Softw. 109, 306–314.

Mehrotra, R., 2005. A nonparametric nonhomogeneous hidden Markov model for downscaling of multisite daily rainfall occurrences. J. Geophys. Res. 110.

Michel, C., 1991. Hydrologie appliquée aux petits bassins ruraux. Cemagref, Antony, France.

Moore, R.J., 2007. The PDM rainfall-runoff model. Hydrol. Earth Syst. Sci. 11, 483–499.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — A discussion of principles. J. Hydrol. 10, 282–290.

Newman, A.J., Clark, M.P., Sampson, K., Wood, A., Hay, L.E., Bock, A., Viger, R.J., Blodgett, D., Brekke, L., Arnold, J.R., Hopson, T., Duan, Q., 2015. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. Hydrol. Earth Syst. Sci. 19, 209–223.

Papalexiou, S.M., 2022. Rainfall generation revisited: introducing CoSMoS-2s and advancing copula-based intermittent time series modeling. Water Resour. Res. 58.

Papalexiou, S.M., Koutsoyiannis, D., 2012. Entropy based derivation of probability distributions: A case study to daily rainfall. Adv. Water Resour. 45, 51–57.

Perrin, C., Michel, C., Andréassian, V., 2003. Improvement of a parsimonious model for streamflow simulation. J. Hydrol. (Amsterdam) 279, 275–289.

Priestley, C.H.B., Taylor, R.J., 1972. On the assessment of surface heat flux and evaporation using large-scale parameters. Mon. Weather Rev. 100, 81–92.

Rasmussen, P.F., 2013. Multisite precipitation generation using a latent autoregressive model. Water Resour. Res. 49, 1845–1857.

Richardson, C.W., 1981. Stochastic simulation of daily precipitation, temperature, and solar radiation. Water Resources Research. Available at: https://doi.org/10.1029/WR017i001p00182.

Richardson, C.W., Wright, D.A., 1984. WGEN: A Model for Generating Daily Weather Variables. U.S Department of Agriculture. Agriculture Research Service ARS-8, p. 83.

Sadeghfam, S., Khatibi, R., Moradian, T., Daneshfaraz, R., 2021. Statistical downscaling of precipitation using inclusive multiple modelling (IMM) at two levels. J. Water Clim. Change 12, 3373–3387.

Sharma, A., Lall, U., 1999. A nonparametric approach for daily rainfall simulation. Math. Comput. Simul 48, 361–371.

Soltani, A., Latifi, N., Nasiri, M., 2000. Evaluation of WGEN for generating long term weather data for crop simulations. Agric. For. Meteorol. 102, 1–12.

Srikanthan, R., Mcmahon, T.A., 2001. Stochastic generation of annual, monthly and daily climate data: A review. Hydrol. Earth Syst. Sci. 5, 653–670.

Srikanthan, R., Pegram, G.G.S., 2009. A nested multisite daily rainfall stochastic generation model. Journal of hydrology (Amsterdam) 371, 142–153.

Thompson, C.S., 1984. Homogeneity analysis of rainfall series: An application of the use of A realistic rainfall model. J. Climatol. 4, 609–619.

Thyer, M., Kuzera, G. 1999. Modelling long-term persistence in rainfall time series: Sydney rainfall case study. In: *International Conference on Water Resources & Environment Research.* Brisbane, Qld.: Institution of Engineers, Australia.

Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S.W., Srikanthan, S., 2009. Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis. Water Resour. Res. 45.

Turner, M., Bari, M., Amirthanathan, G., Ahmad, Z., 2012. Australian network of hydrologic reference stations – Advances in design, development and implementation. 34th Hydrology and Water Resources Symposium. Engineers Australia, Sydney, Australia.

Viviroli, D., Mittelbach, H., Gurtz, J., Weingartner, R., 2009. Continuous simulation for flood estimation in ungauged mesoscale catchments of Switzerland – Part II: Parameter regionalisation and flood estimation results. J. Hydrol. 377, 208–225.

Wilks, D.S., 1989. Rainfall intensity, the Weibull distribution, and estimation of daily surface runoff. J. Appl. Meteorol. 28, 52–58.

Wilks, D.S., 1998. Multisite generalization of a daily stochastic precipitation generation model. J. Hydrol. (Amsterdam) 210, 178–191.

Woolhiser, D.A., Roldán, J., 1982. Stochastic daily precipitation models: 2. A comparison of distributions of amounts. Water Resour. Res. 18, 1461–1468.

Yoo, C., Jung, K.-S., Kim, T.-W., 2005. Rainfall frequency analysis using a mixed Gamma distribution: evaluation of the global warming effect on daily rainfall. Hydrol. Process. 19, 3851–3861.