Stochastic Environmental Research and Risk Assessment
https://doi.org/10.1007/s00477-025-03037-6

ORIGINAL PAPER

Spatial and temporal dependence framework in multi-site precipitation modelling

Tapiwa Edward Chida[1] · **Nadarajah Ramesh**[1] · **Christian Onof**[2] · **Iris Yip**[1]

Accepted: 31 May 2025
© The Author(s) 2025

Abstract
Multi-site stochastic models consist of a rich class of models that can be utilised to analyse environmental data and provide a range of possible inputs to hydrological models to quantify uncertainty and assess risk in environmental systems. We develop a class of multi-site hidden Markov models that incorporate a copula to capture the characteristics of the daily precipitation process across a network of stations. The construction of the likelihood function of the proposed multi-site precipitation models is described. A copula with appropriate dependence structure is selected from the family of Archimedean copulas. The maximum likelihood method is used to estimate the parameters of the models. The proposed class of models is used to analyse twelve years of daily rainfall data from four weather stations in London, England. The copula-embedded multi-site models captured the properties of the daily rainfall well and reproduced the correlation structure of the daily precipitation better than the other hidden Markov models.

Keywords Copula-embedded hidden Markov model · Covariate · Dependence · Multivariate · Precipitation modelling

1 Introduction

Stochastic modelling of daily precipitations over a network of stations is a fruitful area of research as the multi-site models provide a basis for capturing the properties of rainfall across a region and to understand and manage water resources, environmental and ecological systems to which the rainfall is an input. The multi-site stochastic models can provide a number of possible multivariate time-series inputs to hydrological models, to quantify uncertainty and assess risk in environmental systems. Hidden Markov Models (HMMs) can be used to provide a modelling framework for this and many other environmental processes. Although their ability to accommodate sufficient temporal dependence over long durations is somewhat limited, they can be modified in several ways to provide a rich class of flexible models that are useful in many environmental applications. One example of that, amongst many others, is the class of additional dependence models described by Ramesh and Onof (2014) to model regional rainfall data. In this paper, we extend that idea to a multivariate set-up and develop a class of multi-site hidden Markov models that incorporate a copula to capture the characteristics of the precipitation process across a network of stations.

Monthly, seasonal, and inter-annual precipitation modelling and simulation are fairly widespread practices in precipitation dependent regions. The stochastic precipitation simulators are constructed to produce time series of synthetic data which replicate the statistical characteristics of the rainfall in the region. Mean, variance, autocorrelation, dry and wet spells, and extremal events are among some of the key statistical characteristics which reflect on the validity of the simulations and hence the model's correctness. Enhancement of the models to a multivariate framework

✉ Tapiwa Edward Chida
T.E.Chida@greenwich.ac.uk

Nadarajah Ramesh
n.i.ramesh@greenwich.ac.uk

Christian Onof
c.onof@imperial.ac.uk

Iris Yip
i.yip@greenwich.ac.uk

[1] School of Computing and Mathematical Sciences, Queen Mary Building, University of Greenwich, Park Row, London SE10 9LS, UK

[2] Department of Civil and Environmental Engineering, Skempton Building, Imperial College London, Exhibition Road, London SW7 2AZ, UK

Published online: 03 July 2025

adds multidimensional correlation to the properties of interest. This is particularly useful for hydrological simulations over large areas or very heterogeneous ones. This type of stochastic precipitation models are frequently employed in water allocation, risk planning, resource management, flood mitigation infrastructure design, etc. Additionally, the simulations are used in dynamical and statistical downscaling to evaluate the significance of climate change in economic domains (Lee 2018). Burton et al. (2010), Frost et al. (2011) and Charles et al. (1999) give comprehensive information on the applications.

Our area of interest is in multi-site stochastic precipitation models. In recent years, various stochastic models have been proposed in the literature, with many incorporating either an autoregressive component or a copula to address spatial dependencies. Some of the recent work on dependency-based models include; Lee (2018), Majumder et al. (2020), Härdle et al. (2015), Bárdossy and Pegram (2009). Ailliot et al. (2009) took a different route, modelling dependency with a censored power-transformed multivariate Gaussian distribution embedded within a Hidden Markov Model (HMM). Despite capturing the dependence structure better than a distance-based model, the model failed to reproduce the lag 1 autocorrelation of the rainfall volume time series. A distinct approach would be to incorporate dependence at a temporal level by allowing the distribution of variables to be dependent on the preceding values or climatological factors. Ramesh and Onof (2014) have constructed models based on this viewpoint, taking means of exponential distributions to depend on precipitation measures from the previous three days. Similar to Ailliot et al. (2009), the models underestimated the autocorrelation structure of the precipitation data, despite capturing other statistical characteristics, such as mean and variance. Another important type of multi-variate precipitation model uses Generalised Linear Models to represent the occurrence and depth of rainfall, as well as other hydrologically relevant variables at several sites (Chandler 2020). This approach is different in that it relies upon variables related to climate, geographical location, season, etc. to drive the generation of rainfall. The spatial dependence can be represented by modelling the distribution of the total number of sites with non-zero rainfall on any given day, or using the correlation structure of latent Gaussian processes (e.g. Ambrosino et al. (2014)). Such models are potentially very useful, but require careful handling in particular to avoid overfitting.

Copula-based multivariate distributions offer a great deal of flexibility, allowing us to model the marginal distribution functions separately from the correlation structure which is often of interest in multi-site precipitation modelling. The multi-site precipitation models allow the freedom to test marginal distributions with a different degree of heavy-tailedness and varying correlation structures. Among the multitude of copulas, the following have been employed in precipitation modelling; Gaussian copula by Majumder et al. (2020), Hierarchical Archimedean copula by Härdle et al. (2015), vine copulas by Gao et al. (2021), Gumbel copula by Chen et al. (2015).

We extend the copula-based multi-site partially hidden Markov model framework to better capture the statistical properties of rainfall series. To achieve this, we propose the use of an Archimedean copula, which is well-suited for modeling rainfall data collected from nearby stations. Building on the methodology proposed by Wilks (1998) and Wilks (2009) on the use of a double exponential distribution to capture within-state precipitation, a similar approach is taken. However, to alleviate the computational burden potentially associated with optimising the likelihood function over a high-dimensional parameter space, we shift to a more conventional one-parameter exponential margin.

The remainder of this paper is presented as follows: Sect. 2 presents the mathematical background of the generalised model for dimensions greater than 2 with a subsection on the dependence structure, and concluding with the inference. In Sect. 3, the methods are applied to daily rainfall volume time series from four weather stations in London, England.We discuss the results and key observations in Sect. 4, and conclude the study in Sect. 9.

2 A stochastic precipitation model

A Hidden Markov model has two main defining characteristics. The first characteristic is that it assumes the observations from each precipitation recording station at time t were generated by a state process hidden from the observer. The other characteristic is that the dependence of a future state, j_{t+1}, upon the previous states is entirely encapsulated in its dependence upon j_t: this is known as the Markov property. Suppose then we have d localised weather stations, $s = 1,2,...,d$, with daily rainfall recordings $y_t^{(s)} = y_t^{(1)}, y_t^{(2)}, \ldots, y_t^{(d)}$ respectively, on day t. The rainfall process is characterised by an irreducible three state Markov chain, $\{Z_t\}$ with, on day t; state 1 for no rainfall, state 2 for moderate rainfall and state 3 for heavy rainfall. The state classification on the wet days is determined by a random Markov process, so we can not identify the state on each rainy day. However, on dry days $y_t^{(s)}$ will equate to zero meaning state 1 is observed on the day t, hence the name *partially* hidden Markov model. We consider a multivariate distribution for the wet states, with a copula and one parameter marginal distributions, the exponential distribution $y_t^{(s)} \sim Exp(\lambda)$ that depends on the current state of the

underlying Markov chain. Given *m* hidden Markov states, a generalisation of the *d*-dimensional multivariate density function, for state *j = 1,...,m* at time *t*, is given as;

$$h_j(y_t^{(1)}, \ldots, y_t^{(d)}) = c_{\theta|(\mathbf{d})\mathbf{j}} \prod_{s=1}^{d} f_j(y_t^{(s)}) \tag{1}$$

where $c_{\theta|(\mathbf{d})}$ is a *d*-dimensional copula density with parameter θ, and $f_j(y_t^{(1)}), \ldots, f_j(y_t^{(d)})$ are the marginal distributions of rainfall volume from each station *s*.

2.1 The copula

In order to capture within-state spatial correlation, a copula is incorporated into the Markov process, using the transition probability matrix parameters to identify the hidden parameters of a multivariate distribution with a copula and Exponential margins. We explore the Archimedean family of copulas to identify a copula that effectively captures the uniform dependence and zero-inflated nature of precipitation volume data. Given the diverse range of Archimedean copulas, each capturing distinct types of dependence, a suitable choice should exhibit strong lower-tail dependence while still being capable of representing moderate upper-tail dependence. This aligns with the empirical characteristics of the most precipitation dataset, indicating the co-occurrence of dry days alongside moderate clustering of extreme heavy rainfall events. A *d*-dimensional Archimedean copula can be defined in terms of a single scalar generator function φ, given by,

$$\begin{aligned} C(u_1, \ldots, u_d; \theta) = & \varphi^{-1}\left(\varphi(u_1; \theta)\right. \\ & \left. + \varphi(u_2; \theta) + \cdots + \varphi(u_d; \theta)\right) \end{aligned} \tag{2}$$

where $\varphi(\cdot; \theta)$ is a continuous, strictly decreasing convex function satisfying $\varphi(1; \theta) = 0$ and φ^{-1} is the pseudo-inverse (Naifar 2011). The parameter θ determines the degree and direction of dependence captured by the copula.

The correlation coefficient provides a simplified way to characterise strength and direction of association of the variables of interest. A Kendall's τ of 0 describes independent variables, whilst on the extreme ends, Kendall's τ values of 1 and -1 represent absolute positive and negative dependence respectively. We can visualise the nature of some Archimedean copulas in Fig. 1: there are 1000 simulated samples from the Clayton, Gumbel-Hougaard and Frank copulas at Kendall's τ correlation coefficient values of 0.5 and 0.85. The characteristics of the copulas are more visible at higher correlation estimates: at low Kendall's τ values the effect of the copulas is reduced. In multi-site precipitation modelling the weather stations are most likely to be closer. This means that the precipitation volume measurements will be jointly low or high, driving us towards symmetric dependence. The Frank copula may be suitable for such observations because of its co-dependence structure, however, it does not capture tail dependence. In contrast, the Gumbel copula is better at modeling upper-tail dependence (strong correlation in extreme rainfall conditions), while the Clayton copula is more suited for lower tail dependence (strong correlation during low rainfall periods).

2.2 Temporal dependence model

In this section, we introduce a Hidden Markov model that incorporates covariate-driven dependencies at the observation level, utilising the moving average of daily temperature and pressure as key factors. This model assumes that the distribution of rainfall observations is influenced not only by the current state of the Markov chain but also by the smoothed covariates. We model rainfall amounts in each state using an exponential distribution, where the parameter λ depends on the moving averages of temperature and sea-level pressure.

The model is defined as follows, at station *s*, with $\lambda_j^{(s)}$ representing the parameter of an exponential distribution for the observation $y_t^{(s)}$ at time *t*, given that the Markov chain is in state *j*:

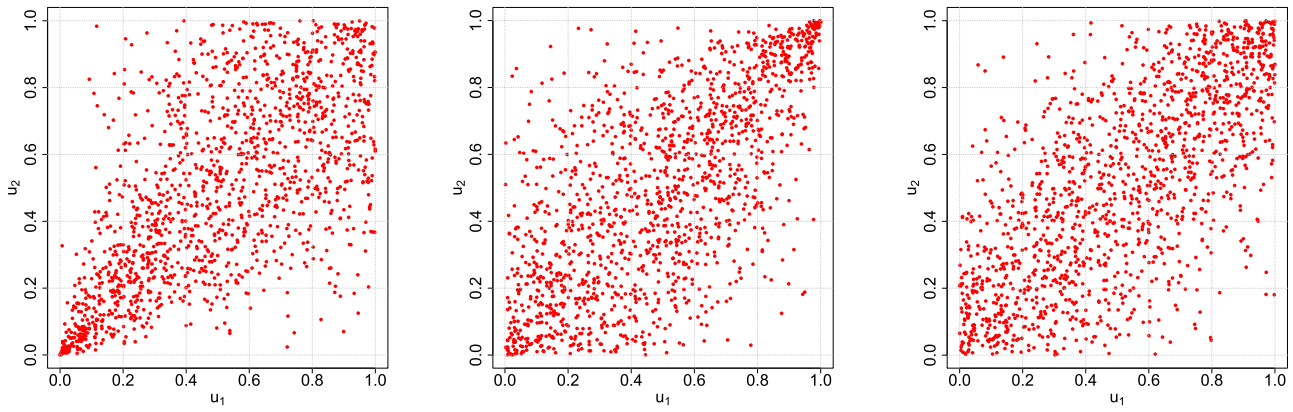
$$\lambda_j^{(s)} = exp\left(\beta_{0j}^{(s)} + \beta_1 \bar{T}_t + \beta_2 \bar{P}_t\right) \tag{3}$$

Here, $\bar{T}_t$ and $\bar{P}_t$ denote the moving averages of daily temperature and pressure, respectively, at time *t*. To integrate these covariate-driven dependencies into the basic Hidden Markov model framework without the copula component, we assume that the singular covariate effect is the same in any state *j* and at any station *s*. By incorporating the moving averages of temperature and pressure, we filter out the noise from random fluctuations of the covariates. This aids the model in providing a more refined and accurate representation of the rainfall process, capturing how these smoothed environmental factors influence the patterns. The optimal interval of the moving average dependends on the daily precipitation data, and a sensitivity analysis would help to make the optimal selection.

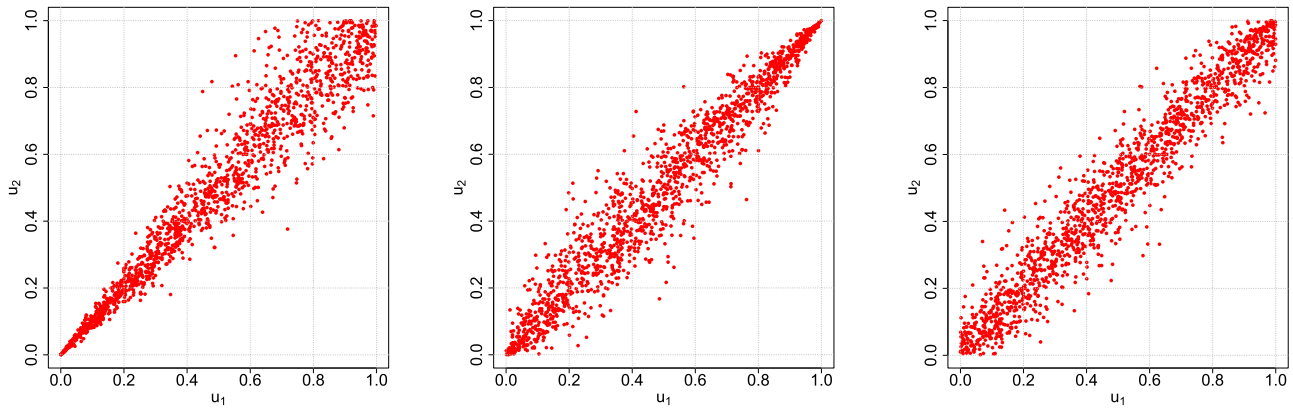
2.3 Maximum likelihood estimation

To formulate a likelihood function for the copula-embedded model, three matrices are brought forward: The equilibrium probabilities row matrix for the Markov process $\{Z_t\}$,

$$\pi = [\pi_1 \quad \ldots \quad \pi_m]$$



(a) Clayton, τ=0.5 (b) Gumbel-Hougaard, τ=0.5 (c) Frank, τ=0.5



(d) Clayton, τ=0.85 (e) Gumbel-Hougaard, τ=0.85 (f) Frank, τ=0.85

Fig. 1 Simulated bivariate dependence structures from Archimedean copulas at Kendall's $\tau = 0.5$(moderate association) and $\tau = 0.85$(strong association), showing the copula's capacity to model varying spatial correlation strengths

the m-state transition probability matrix,

$$\Phi = \begin{bmatrix} p_{11} & p_{12} & \dots & p_{1m} \\ p_{21} & p_{22} & \dots & p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \dots & p_{mm} \end{bmatrix}$$

and a diagonal matrix with the copula-embedded state distribution given in Eq. (1)

$$\Psi = \begin{cases} diag\,[1, \quad 0, \quad \dots, \quad 0] & y_t^{(s)} = 0 \\ diag\left[0, \quad h_2\left(y_t^{(1)}, \dots, y_t^{(d)}\right), \quad \dots, \quad h_m\left(y_t^{(1)}, \dots, y_t^{(d)}\right)\right] & y_t^{(s)} \neq 0 \end{cases}$$

where $h_j\left(y_t^{(1)}, \dots, y_t^{(d)}\right)$ is the multivariate density function for state j, where $j \in \{2, \dots, m\}$.

The likelihood function is then given as,

$$L\left(y_t^{(1)}, y_t^{(2)}, \dots, y_t^{(d)} | \Phi, \Psi\right) = \pi \left[\prod_{t=1}^{n} \Phi\Psi\right] 1_{m\times 1} \tag{4}$$

The matrices can be adapted for the covariate models described in Sect. 2.2 by replacing the multivariate density function, $h_j\left(y_t^{(1)}, \dots, y_t^{(d)}\right)$, with the product of corresponding Exponential functions,

$$\prod_{s=1}^{d} f_j(y_t^{(s)}) \tag{5}$$

Some technical issues that usually arise in the numerical optimisation of Eq. (4) are numerical underflow and overflow, and local minima. A scaling method proposed in Zucchini and MacDonald (2009) can be used to resolve the issue of underflow and overflow. There is no straightforward method we can use to determine whether the numerical optimisation

algorithm has reached a global minima. The strategy we use, proposed in Zucchini and MacDonald (2009), involves trial and error until we identify a range of starting values giving a similar minima.

3 Modelling and simulation

In this section, we fit the calibrated spatial dependence model, the survival Gumbel HMM, to daily precipitation data. Alongside this, we also fit a multivariate HMM assuming contemporaneous independence, and a temporal dependence model, the covariate HMM. The outcomes of these models are presented through comprehensive tables and detailed diagnostic plots, with each result followed by explanations to aid interpretation. Where appropriate, comparisons are made across the three models in order to highlight differences in goodness-of-fit and spatio-temporal dependence capture.

3.1 Geospatial scope: London Borough of Hillingdon

The proposed model is tested on daily precipitation recordings from the London Borough of Hillingdon in West London. The map of the region is provided in the Fig. 2, highlighting the weather stations of interest; located across Brent, Pinn, Yeading East and Yeading West. These sites have been selected to provide comprehensive coverage of microclimates within the Greater London area. Each station plays a critical role in capturing the unique weather patterns influenced by the broader atmospheric conditions. Although the stations are relatively close, with the largest distance approximately 8 km, we feel that a joint probability density approach is important for capturing spatial dependencies, particularly in the tails of the distribution. Analysis of the scatterplots in Fig. 3 shows that even at short distances, differences in tail behaviour can occur, such that one station might record a heavy rainfall event while a nearby station records moderate rainfall or a dry day. These discrepancies are likely due to variations in factors such as urbanisation, land cover, and microclimatic influences, which can lead to significant differences in extreme precipitation events in fine spatial ranges. A joint probability model maintains spatial associations, therefore, ensuring that rainfall patterns in the synthetic series accurately represent the co-occurrence of extremes at both the lower and upper tails. This helps in the reliable prediction of flood risk and other extreme events with adverse effects on infrastructure. Unlike univariate methods, our framework quantifies mutual information between stations across the region, capturing dependencies not only during intense rainfall episodes, but also during moderate rainfall events. This also addresses a challenge in Yang et al. (2005) of representing spatial dependence at small scales relative to weather systems.

3.2 Data

The data analysed are for a 12 year period, from January 2000 to December 2011. These are measurements of daily precipitation volume (millimetres) from the four weather stations located in the London Borough of Hillingdon; the summary statistics of the precipitation volume data are given in Table 1. While the mean daily rainfall volumes are relatively close (ranging from 1.83 to 1.87 mm), larger discrepancies emerge in the extremes. Yeading East records the highest maximum daily rainfall at 50.6 mm, notably higher than Brent's maximum of 43.6 mm, suggesting stronger localised events at certain sites. This variation is further reflected in the skewness and kurtosis values, with Yeading

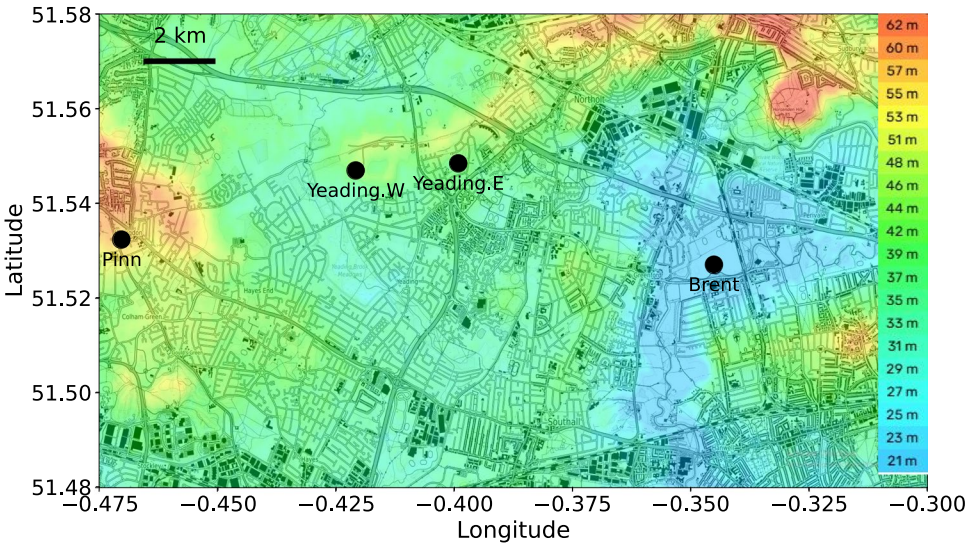


Fig. 2 Topographic map of the London Borough of Hillingdon showing station locations with rainfall data recorded between 2000 and 2011. The elevation scale (far right) indicates colour coded terrain heights, while the 2 km bar provides spatial reference

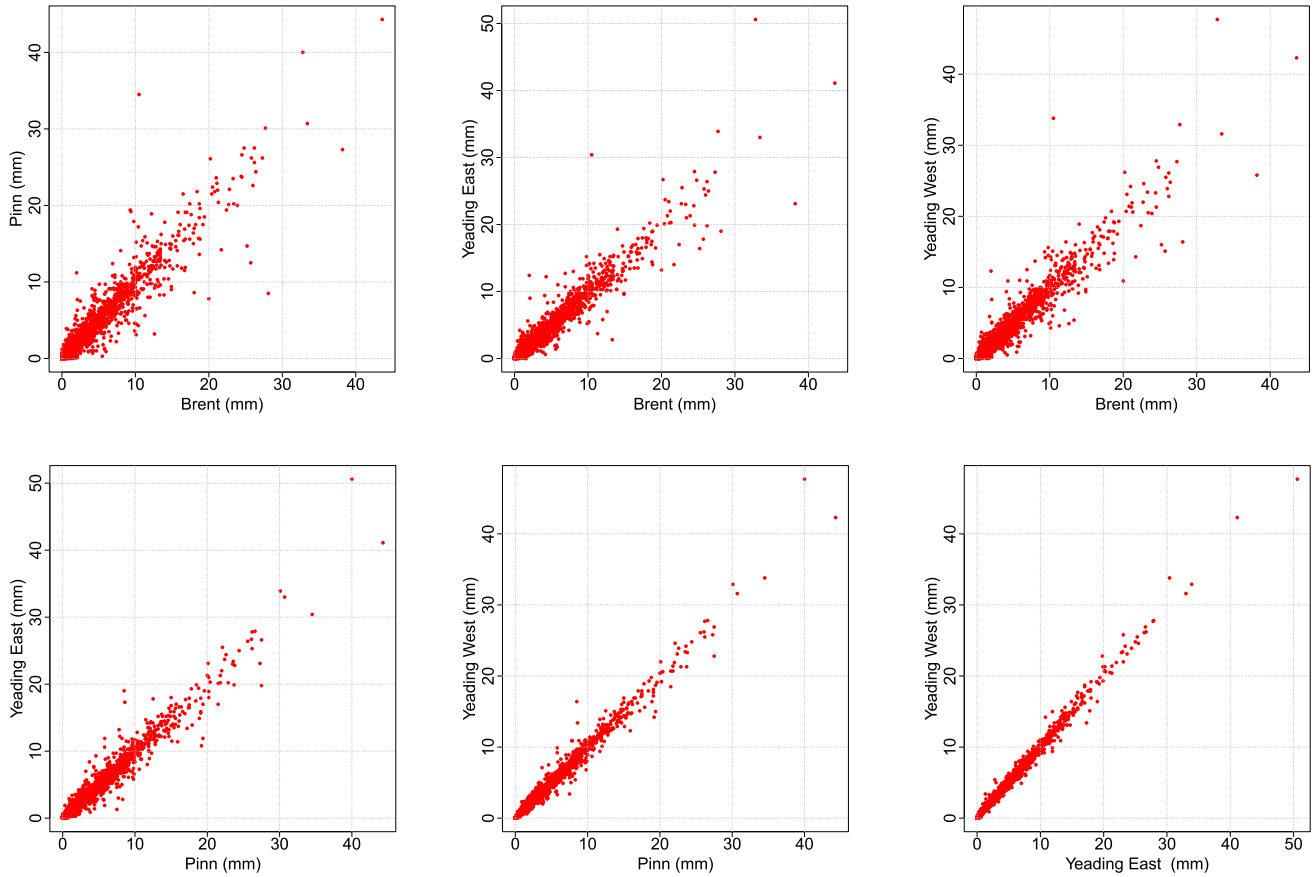


Fig. 3 Scatter plots of daily rainfall volumes for all six pairwise combinations of the four stations, illustrating spatial dependence across the region

Table 1 Descriptive statistics for rainfall volume data recorded at the four stations in the London Borough of Hillingdon, where mAOD stands for Meters Above Ordnance Datum

Variable	Mean	SD	p90 Alt (mAOD)	Latitude	Logitude	Max	Skewness	Kurtosis
Brent	1.868	3.8018	104.5000	51.527032	– 0.34509418	43.6000	3.4689	16.7450
Pinn	1.8540	3.7818	87.5000	51.532329	– 0.47033532	44.3000	3.5071	17.4638
Yeading East	1.8321	3.7668	65.1000	51.548464	– 0.39912319	50.6000	3.6483	20.2992
Yeading West	1.8462	3.7801	63.5000	51.546960	– 0.42080934	47.7000	3.5992	19.2682

East displaying the most pronounced skewness (3.65) and heaviest tail (kurtosis 20.30), indicating a greater propensity for extreme rainfall compared to the other stations. Differences in 90th percentile altitude also likely contribute to these patterns: Brent, the highest site at 104.5 mAOD, experiences somewhat lower extremes than Yeading West (63.5 mAOD) and Yeading East (65.1 mAOD), consistent with the general tendency for lower-lying areas to accumulate more intense rainfall events under convective conditions. These variations in both marginal and extreme behaviour across a relatively compact geographical area reinforce the necessity of joint rainfall modelling frameworks that can account for spatial dependence and heterogeneity in distributional properties, rather than treating stations as independent. As much of the rainfall in London is recorded during the winter season, we are also interested in modelling the daily precipitation during this period; a joint time series of the months November, December and January for the twelve years. London's weather is shaped by a combination of maritime influences and urban factors. Weather systems from the Atlantic Ocean often bring rain to the city, with frontal systems causing prolonged periods of steady rain, particularly during the autumn and winter months. Due to London's relatively low elevation and proximity to the Thames River, the city experiences slightly milder temperatures compared to surrounding areas. Coastal influences from the English Channel and the North Sea can also affect local weather patterns, though to a lesser extent than in more exposed coastal regions.

Temperature and atmospheric pressure play significant roles in London's weather variability. During autumn and winter, lower pressure systems are more common,

contributing to frequent rain and cooler temperatures, with average daily temperatures often dipping to single digits (Celsius). In contrast, spring and summer generally see higher pressure systems, leading to drier conditions, although London remains prone to convective showers and thunderstorms, particularly during warmer spells.

The urban heat island effect can increase temperatures, especially in summer, leading to warmer nights and occasionally influencing the intensity and distribution of rainfall. Seasonal variations are evident, with winter and autumn being wetter and cooler, while spring and summer are drier and warmer, although still prone to occasional rain.

3.3 Pre-analysis tests

Multisite daily precipitation dependence structure

Daily rainfall data are characteristically zero inflated, with the UK experiencing a pronounced rainy season from December to January and only intermittent moderate rainfall during other months. This results in an exponential-like precipitation process, with a modal peak at zero (representing dry days) that is consistently observed at localised stations. While most spatial dependence modelling methods effectively capture the associations in the lower tail, it is in the upper-tail where extreme events occur that significant variations emerge, particularly as the distance between stations increases. For stations in proximity, extreme events tend to occur concurrently, indicating that upper-tail dependence must be substantial. Scatter plots of daily rainfall volumes for all the six paired combinations of the four stations support this pattern, alongside strong lower-tail dependence as shown Fig. 3. However, selecting an appropriate copula to replicate this complex correlation behaviour presents a significant challenge. In this section, we elaborate the rationale behind why certain copulas may be unsuitable for modelling dependence in such rainfall series and discuss the criteria for choosing copulas that can more accurately capture strong lower tail dependence and the simultaneous occurrence of extreme rainfall events across spatially correlated sites.

A multivariate test of radial symmetry, based on Kojadinovic (2017), on the London precipitation data indicate that there is strong evidence of invariance under rotation around the origin. In the Archimedean family, only three copulas exhibit such a characteristic; the Frank copula, Ali-Mikhail-Haq copula and the independence copula. Only the Frank copula is a viable choice from these three, offering a balanced approach to dependence modelling, but without any tail dependence. In this setting, it is advantageous because it can capture the overall dependence structure without over-amplifying the effects of dry days introduced by the HMM, where a dry day is simultaneously assigned across all stations. This balanced dependence is particularly useful for moderate rainfall days, where the overall empirical correlation is high. However, its only shortcoming is hard to ignore especially in the upper-tail, where there is a strong under-representation of the probability of joint heavy rainfall events. Evidence from a multivariate test of extreme value dependence based on Kojadinovic (2017), where the null hypothesis of the empirical copula being an element of extreme value copulas, suggests that there is strong evidence of extreme tail dependence in London data. This eliminates a Clayton copula, although, nearly half of the dataset comprises of dry days. This copula type is problematic for heavy rainfall events, where the co-occurrence of extremes across the stations is of particular interest. Additionally, the zero inflation due to the HMM's treatment of dry days can exaggerate the lower-tail dependence, potentially leading to an over-reliance on a copula that fails to sufficiently represent the dynamics of the wet states.

The only upper-tail copulas within the Archimedean family are the Gumbel-Hougaard and Joe copulas but, none of these exhibit strong lower-tail dependence. In contrast, survival versions of these Archimedean copulas, can be adapted by rotating the original copula functions to capture dependence in the opposite tail. The survival Gumbel copula, derived from the standard Gumbel which has strong upper-tail dependence, retains some of that dependence structure, thereby modelling the co-occurrence of heavy rainfall events across stations. On the other hand, the survival Joe copula tends to show less upper-tail dependence due to its inherent dependence structure. This difference makes the survival Joe copula potentially more moderate and closer to a Clayton copula when describing extreme co-occurrences. We will not consider nested copulas such as the BB1 and BB7 because they are an extension of the pure Archimedean family, incorporating extreme value behaviour through additional parameters. Because they blend features of both Archimedean and extreme-value copulas, they cannot be calibrated in a straightforward way. We would then also have to consider extreme value copulas, some of which are a generalisation of the Gumbel copula, like the Tawn copula. Combining these generalisations of the Archimedean copulas with an HMM not only increases the computational cost, but also increases the risk of over-fitting, making the optimisation less stable, therefore affecting the accurate capture of other statistical properties (mean and variance). A one parameter copula in each wet state makes the model parsimonious, while keeping precipitation generation process robust.

We compared the Archimedean copulas using standard metrics such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for different pairwise combinations of the stations (Brent-Yeading West,

Table 2 AIC and BIC values for the fitted bivariate copula models across station pairs

Combination	Copula	AIC	BIC
Brent–Yeading West	Clayton	– 13212.37	– 13205.99
	Survival Gumbel	– 14439.82	– 14433.43
	Frank	– 12900.34	– 12893.96
	Survival Joe	– 13310.80	– 13304.41
Pinn–Yeading East	Clayton	– 15510.87	– 15504.49
	Survival Gumbel	– 16733.02	– 16726.63
	Frank	– 14754.91	– 14748.52
	Survival Joe	– 15589.43	– 15583.04
Yeading East–Yeading West	Clayton	– 19488.52	– 19482.13
	Survival Gumbel	– 20540.42	– 20534.03
	Frank	– 16893.86	– 16887.48
	Survival Joe	– 19618.58	– 19612.19

Table 3 Copula model performance and statistical comparisons based on likelihood-ratio methods

Copula	Log-likelihood	Parameter
Survival Gumbel	24535.05	11.5227
Clayton	23462.44	17.5654
Frank	22946.87	40.4348
Survival Joe	22566.94	16.9948
Vuong Test (Survival Gumbel vs Clayton)		
Test Statistic	3.0647	
p-value	0.0022	

Pinn-Yeading East, and Yeading East-Yeading West); the results are presented in Table 2. The survival Gumbel consistently achieved the lowest AIC and BIC values, indicating superior fit compared to the other copulas. The survival Gumbel copula had the lowest values of both the AIC and BIC for every other pairwise combination.

We further supported these results by performing a four-dimensional Vuong test, a likelihood-ratio-based method that evaluates which model is significantly closer to the empirical copula. The test results also pointed to the survival Gumbel as the best copula for this analysis, as shown in Table 3.

We move a step further by fitting the Archimedean copulas to all the pairwise combinations of the precipitations variables; assessing the performance against a nonparametric estimator of the copula, which according to Hofert et al. (2018) is given as;

$$\begin{aligned} C_N(u) &= \frac{1}{N}\sum_{i=1}^{N} 1(U_{i,N} \le u) \\ &= \frac{1}{N}\sum_{i=1}^{N}\prod_{s=1}^{d} 1(U_{i,s,N} \le u_s), u \in [0,1]^d \end{aligned} \tag{6}$$

where $U_{i,N} = (U_{i,1,N}, \ldots, U_{i,d,N})$, $i \in 1, \ldots, N$, are uniformly distributed and the inequalities $U_{i,N} \le u$, are component-wise. Equation (6) is also known as the empirical copula. The contour plots of the fitted survival Gumbel, Frank, survival Joe and Clayton copulas overlaid by those of the empirical copula for the Brent-Pinn combination is represented in Fig. 4. In the mid-range of the distribution (approximately $u = 0.6 - 0.7$), both the Clayton and survival Joe copulas trace the empirical contours quite well. Beyond this band, however, their fits weaken: neither copula adequately reproduces the strong upper-tail dependence inherent in the data for $u > 0.8$. The Frank copula performed well, however, it still falls short in the tails (e.g, $u = 0.9$). In contrast, the survival Gumbel copula consistently fits both the mid-range and the upper-tail, aligning with empirical contours for all probabilities above 0.6. The pronounced lower-tail empirical contours in the plots simply reflect the zero-inflated nature of daily precipitation; all the copulas, including the survival Gumbel, capture these extreme low-probability events satisfactorily. Overall, these findings confirm that the survival Gumbel copula strikes the best balance between capturing the upper and lower tail dependencies present in the precipitation data, making it the most suitable choice for analysing spatial dependence. It should be noted that contour plots for the copula fit on the winter season precipitation volume data yielded similar results.

Sensitivity analysis of moving average interval

We conducted a sensitivity analysis, testing the most optimal moving average period for this analysis. The averaged, state- and temporal-dependent, exponential parameter, given in Eq. (3), was calculated using 3-day, 5-day and 7-day moving averages. As shown in Table 4, the parameter estimates derived from the 3-day and 5-day moving averages are very similar, indicating that switching between these two would not result in significant differences in the model output. However, when using a 7-day moving average, the parameter values begin to increase, leading to a consequential decrease in the state means.

The overall means computed from the parameter estimates of each moving average period, as shown in Table 5, revealed that the 5-day moving average produces means that are closest to the empirical observations at each station.

These results suggest that the 5-day averaging window provides the best balance, it is long enough to smooth out short-term fluctuations while still capturing the temporal dynamics. Therefore, we have adopted the 5-day moving average for our subsequent analyses, confident that it gets parameter estimates consistent with the observed data.

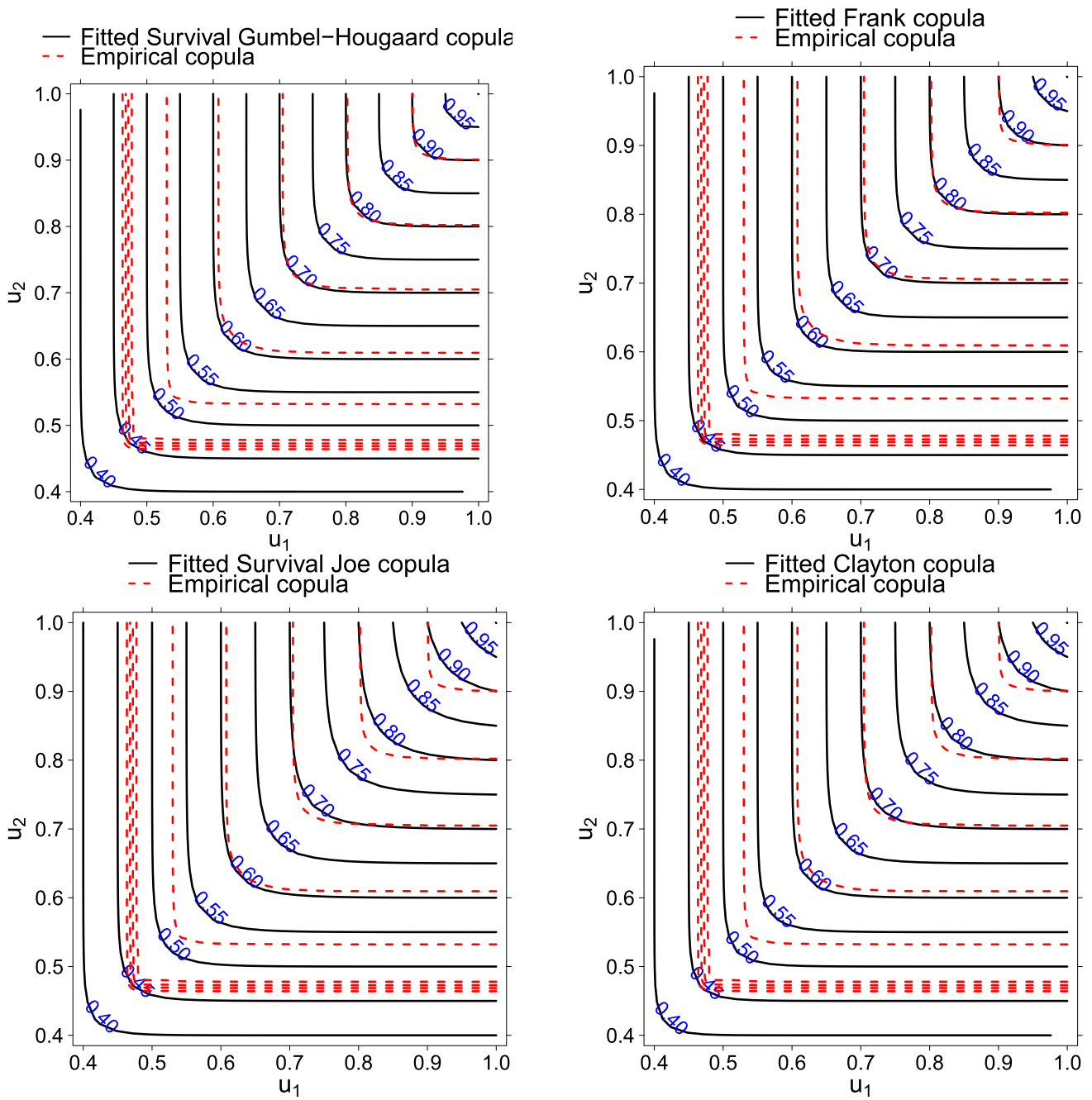


Fig. 4 Contour plots comparing theoretical Archimedean copula distributions (fitted) to the empirical copula for the Brent-Pinn precipitation series

Table 4 Estimated values of λ_2 and λ_3 over 3-day, 5-day, and 7-moving averages

	Brent		Pinn		Yeading East		Yeading West	
	λ_2	λ_3	λ_2	λ_3	λ_2	λ_3	λ_2	λ_3
3-day	0.3218	0.3388	0.3480	0.3452	0.1902	0.2155	0.1903	0.1910
5-day	0.3332	0.3261	0.3410	0.3314	0.1886	0.1888	0.1889	0.1888
7-day	0.3937	0.4064	0.4267	0.4140	0.2278	0.2315	0.2383	0.2400

Table 5 Empirical means and means derived from the moving average parameters at each stations

	Brent	Pinn	Yeading East	Yeading West
Empirical	1.8677	1.8541	1.8318	1.8462
3-day	1.8038	1.6272	1.7642	1.7633
5-day	1.8263	1.8559	1.8126	1.8280
7-day	1.4917	1.4613	1.4115	1.4168

3.4 Spatial and temporal multi-site precipitation models

3.4.1 Two-site models

In this section, we explore three hidden Markov models for precipitation modelling in a two-dimensional setting. The first model is a Bivariate Exponential Hidden Markov Model (B.HMM), with a relatively simple structure and assuming contemporaneous independence. The second model is a Bivariate Covariate Exponential Hidden Markov Model (BC.HMM), enhanced to capture temporal dependence by incorporating a 5-day simple moving average of the daily temperature and pressure as covariates, allowing the model to adapt to changes in the daily rainfall patterns effectively. The third model is a Bivariate survival Gumbel copula-embedded Hidden Markov Model (BSG.HMM) with Exponential margins, designed to capture state dependencies between the sites. By allowing the bivariate distribution of observations at time t to depend on the unobserved state of a Markov chain, the basic bivariate HMM can accommodate serial dependence (Zucchini and MacDonald 2009). In this regard, for our application, the inclusion of state dependent copula components to the bivariate distributions in the two wet states adds to the temporal aspect of the model. This approach leverages the copula-embedded hidden Markov model to capture complex spatial relationships between sites, while accounting for temporal dependence within sites as the underlying Markov chain shifts through different states in time. We anticipate that the enhancements made to each model will improve their ability to accurately capture the statistical characteristics of the data.

The general three-state transition probability matrix for the models is given as,

$$\Phi = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix}$$

where each row sums up to 1, and p_{jj} is obtained by subtracting the sum of the estimated parameters in the row from one. The diagonal matrices of the state distributions for the BC.HMM are given by,

$$\Psi_{BC.HMM} = \begin{cases} diag\,[1, \quad 0, \quad 0] & y_t^{(s)} = 0 \\ diag\left[0, \quad \prod_{s=1}^{d} f_2\left(y_t^{(s)} | \beta_{02}^{(s)}, \beta_1, \beta_2\right), \quad \prod_{s=1}^{d} f_3\left(y_t^{(s)} | \beta_{03}^{(s)}, \beta_1, \beta_2\right)\right] & y_t^{(s)} \neq 0 \end{cases}$$

where $\beta_{0j}^{(s)}$ are the state j intercept parameters at station s; β_1 and β_2 are the temperature and pressure covariate coefficient parameters respectively. The effect of the covariates is assumed to be identical across both states and at all stations s. This approach ensures consistency in how covariates influence the model across different states, allowing us to balance outcomes without expanding the parameter space.

Moving on to the BSG.HMM, the diagonal matrices of the state distributions are given by,

$$\Psi_{BSG.HMM} = \begin{cases} diag\,[1, \quad 0, \quad 0] & y_t^{(s)} = 0 \\ diag\left[0, \quad h_2\left(y_t^{(s)} | \lambda_2^{(1)}, \lambda_2^{(2)}, \theta_2\right), \quad h_3\left(y_t^{(s)} | \lambda_3^{(1)}, \lambda_3^{(2)}, \theta_3\right)\right] & y_t^{(s)} \neq 0 \end{cases}$$

where $\lambda_j^{(s)}$ are the Exponential margin parameters for station s in state j, and θ_j is the copula parameter in state $j \in \{2, 3\}$. By removing dependence parameters θ_2 and θ_3 in each state, we obtain the diagonal matrices for the B.HMM. Parameter estimates of all three models are given in Table 6.

Since the wet states from the model are latent it is not possible to know the empirical count of data in each state. Instead, we estimated the fitted stationary distribution from the model. For instance, the estimated stationary distribution ($[0.5822, 0.1708, 0.2470]$) from 4838 days of daily precipitation gives us roughly 826 observations used in state 2 and 1195 in state 3. While State 2 has fewer observations, the sample size remains sufficient for modelling the 4-dimensional Archimedean survival Gumbel copula due to the strong dependencies (Kendall's τ of 0.9111 and 0.8865 in State 2 and 3 respectively), and the copula goodness-of-fit, confirmed by rigorous testing in Sect. 3.3.

In copula modelling, ensuring stable parameter estimation typically requires a large sample size relative to the number of parameters. While specific recommendations vary, some studies suggest that a higher ratio of observations per parameter is beneficial, especially in high-dimensional settings (Oh and Patton 2017). The appropriate sample size also depends on the complexity of the copula model and the estimation method used, with strategies such as factor copulas and iterative estimation techniques helping to improve stability (Krupskii and Joe 2013). A 4-dimensional Archimedean copula (the survival Gumbel copula in our case) commonly requires one dependence parameter, making it more data-efficient compared to vine or Gaussian copulas.

We evaluated the performance of bivariate models, focusing on their ability to reproduce key statistical characteristics of the daily rainfall for the Brent-Pinn yearly and seasonal time series. The simulations from the fitted models demonstrated consistent reproduction of critical metrics, ensuring robustness across different temporal scales. The

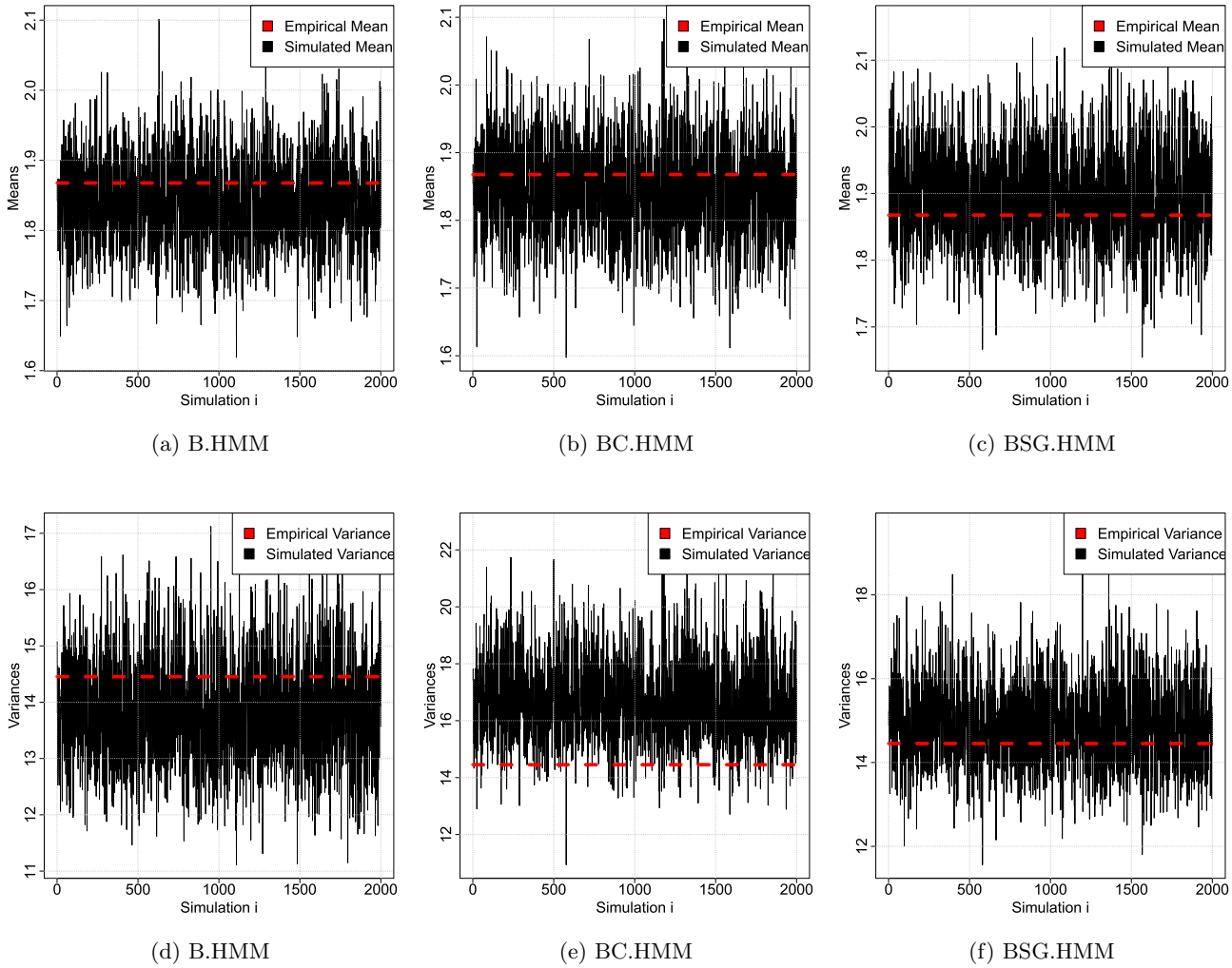
Table 6 Maximum likelihood estimates from bivariate fits on yearly rainfall

Model												
B.HMM	0.1796	0.1751	0.4278	0.2644	0.2779	0.2044	0.7598	0.1853	0.7703	0.1887	–	–
	p_{12}	p_{13}	p_{21}	p_{23}	p_{31}	p_{32}	$\beta_{02}^{(1)}$	$\beta_{03}^{(1)}$	$\beta_{02}^{(2)}$	$\beta_{03}^{(2)}$	β_1	β_2
BC.HMM	0.1484	0.1663	0.3699	0.2804	0.2849	0.1978	−0.1246	−1.5675	−0.1826	−1.5632	−0.0658	0.0702
	p_{12}	p_{13}	p_{21}	p_{23}	p_{31}	p_{32}	$\lambda_2^{(1)}$	$\lambda_3^{(1)}$	$\lambda_2^{(2)}$	$\lambda_3^{(2)}$	θ_2	θ_3
BSG.HMM	0.1073	0.1284	0.3803	0.2833	0.2925	0.2060	0.0.3220	0.1805	0.3245	0.1795	11.2612	8.8116

mean and variance statistics derived from 2000 simulations of each of the fitted models are illustrated in Fig. 5. Among the models tested, the B.HMM and BSG.HMM provided an equal performance in capturing the observed data's variability and distribution patterns. The B.HMM marginally underestimates the mean and variance; the opposite is true for the BSG.HMM. While the BC.HMM produced mean estimates that closely matched empirical observations, it exhibited significant overestimation of variance. This discrepancy arises from the BC.HMM's daily-varying exponential parameters, which enhance temporal dependence modelling at the cost of inflated variance.

Comparisons of the theoretical and empirical autocorrelation presented in Fig. 6 highlight that the statistic is captured within the upper and lower simulation bounds for the BC.HMM. The B.HMM cannot capture this well: models that assume contemporaneous independence often fail to capture autocorrelation. At lag three, an uncharacteristic increase in empirical autocorrelation compared to lag two suggests that this maybe due to the climate dynamics of the region. London's climate, characterised by the frequent passage of frontal systems where warm and cold air masses converge, can lead to periodic precipitation over several days. These fronts typically move slowly, with heavy rainfall occurring on the first day (lag zero), a reduction in rainfall on the third day (lag two) as the front weakens, followed by a resurgence of rainfall on the fourth day (lag three) as the front moves further across the region. This behaviour may explain the elevated autocorrelation at lag three. Autocorrelation estimates from the BSG.HMM offer a slight improvement over those from the B.HMM. This suggests that the influence of the copula component on temporal dynamics seems to diminish when modeling precipitation measurements for these two sites. Including additional sites, and thus employing a higher-dimensional state density function may help overcome this. The proficiency of a bivariate copula model in capturing the underlying spatial correlation structure is presented in Fig. 7. Lag zero estimates of the cross-correlation, generally underestimated by the BC.HMM (with the B.HMM providing a similar performance), are overestimated from 2000 multivariate simulations. The limitation of the copula component in accurately capturing the empirical dependence structure is evident in the lag one and lag three cross-correlations, where the model underestimates the empirical values. This, in part, is influenced by the discrepancies between the empirical copula and the survival Gumbel copula, as shown in Fig. 4.

Distributions for the wet and dry periods are derived from the observed and the simulated precipitation series. A dry spell in this context is a consecutive time period (days) in which the rainfall volume equates to zero, preceded and succeeded by a wet days: the opposite is true for a wet spell.



(a) B.HMM (b) BC.HMM (c) BSG.HMM

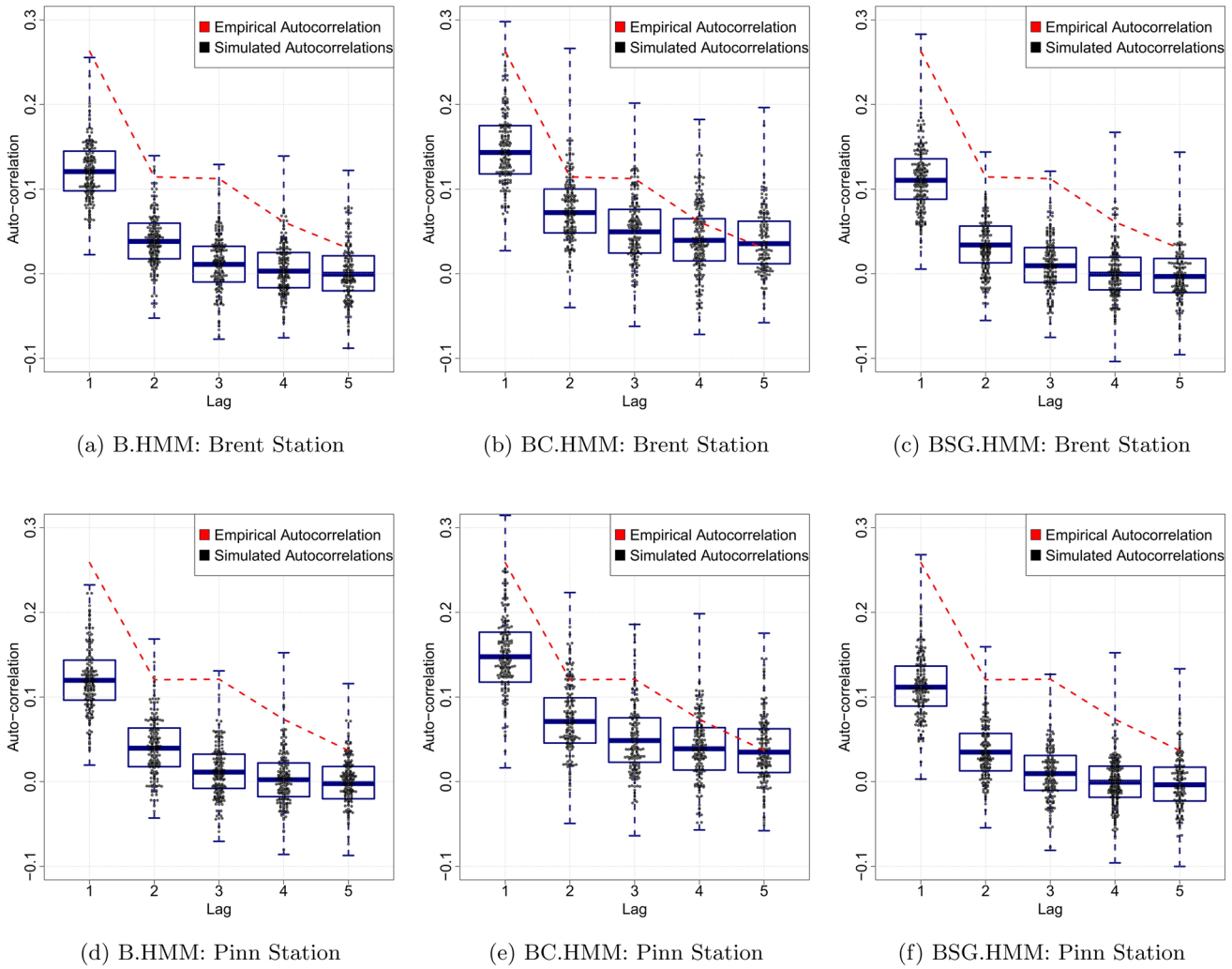
(d) B.HMM (e) BC.HMM (f) BSG.HMM

Fig. 5 Comparison of theoretical (black) and empirical (red) mean and variance estimates for the Brent station's yearly rainfall series

Figure 8 presents a density plot of dry spell durations for Brent and Pinn stations, overlaid with corresponding plots generated from model simulations. The performance of the models was impacted by the methodology used in calculating the log-likelihood. Specifically, when the daily two-site observation, represented as a 1-by-2 row vector Y for day t, contains a zero in any of its elements–indicating a dry day at one station while the other station records precipitation–the model classifies the entire day as dry in both stations. This treatment leads to discrepancies in the simulation of dry spells across the stations. Dry state stationary probabilities derived from the maximum likelihood estimates of the transition probability matrix in Table 6 are; 0.4901, 0.5041 and 0.5822 for the B.HMM, BC.HMM and BSG.HMM respectively. For further context, across the two stations, approximately 47% of the days during the 12-year period were classified as dry.

In Fig. 9, the fitted distributions for wet spell durations demonstrate an impressive performance from the B.HMM and the BC.HMM. Based on the observations from the dry spell distribution, one might anticipate a higher empirical density compared to the model-based densities. However, the Hidden Markov Model framework accounts for this discrepancy, resulting in a closer alignment between the empirical and theoretical wet spell distributions. The MSG. HMM underestimates one-day wet spells but overestimates those lasting five to nine days because its strong wet-state persistence delays returns to dryness. When it finally does switch, it most often reverts directly to the dry state, so isolated wet days become too rare, and medium-length wet runs are over-represented.

The log-survival plots presented in Fig. 10 demonstrate that each of the models successfully replicates the marginal distributions, closely matching the characteristics of the empirical survivalfunctions. Simulation bands in the plots correspond to the minimum and maximum values across simulations, capturing the full range of variability around the model estimates. The strong alignment suggests that



(a) B.HMM: Brent Station (b) BC.HMM: Brent Station (c) BSG.HMM: Brent Station

(d) B.HMM: Pinn Station (e) BC.HMM: Pinn Station (f) BSG.HMM: Pinn Station

Fig. 6 Box plots comparing theoretical autocorrelation (with jittered black dots) and empirical autocorrelation (red) for the seasonal rainfall series at Brent (top) and Pinn (bottom) stations

the models are effective in capturing the tail behaviour and distributional properties of the data. We visually assess the underlying Markov chain by means of a dynamic programming algorithm (Viterbi algorithm), which can determine the most likely sequence in the state space (Zucchini and MacDonald 2009). Figure 11 displays the Viterbi-decoded state sequences for the first three months of the 12-year period at Brent and Pinn stations, overlaid with daily rainfall volumes shown as vertical lines. While the true states are unobservable, the Viterbi paths shows an agreement with rainfall patterns: heavy rainfall events tend to coincide with state 3, moderate rainfall events with state 2 and dry spells correspond to the dry state. This suggests that the model captures key structures in the rainfall dynamics. This optimal prediction would also suggest that the Hidden Markov Model is capturing the temporal dynamics of the rainfall process effectively.

This visualisation is important because it offers a direct, day-by-day decoding of the latent states, allowing us to interpret how the model captures persistence and switches between wet and dry conditions. By overlaying the Viterbi path on the rainfall series, we can qualitatively validate the model's temporal dynamics without relying only on summary statistics. It should be noted that this exercise is not a comparison between the models but, a demonstration of how well the MSG.HMM decodes its own hidden state sequence. As such, it provides insight into the model's internal coherence rather than a performance evaluation.

The three proposed stochastic models demonstrated strong performance across all evaluated metrics. Each model, characterised by its unique structural features, successfully reproduced the theoretical elements most closely aligned with the empirical data for at least one of the metrics. When comparing model fit based on the AIC and BIC values (Table 7), the BSG.HMM emerged as the best fit for

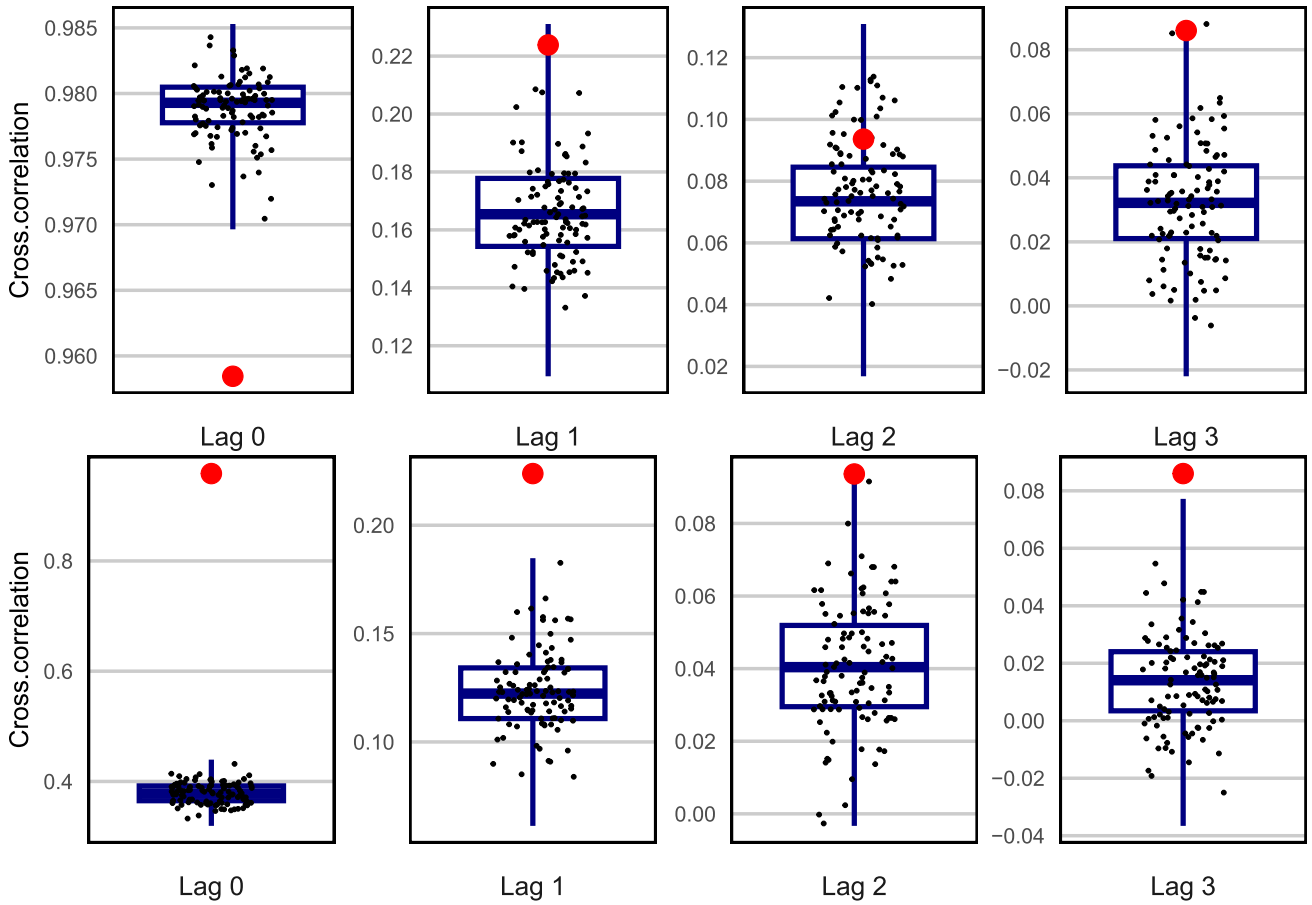
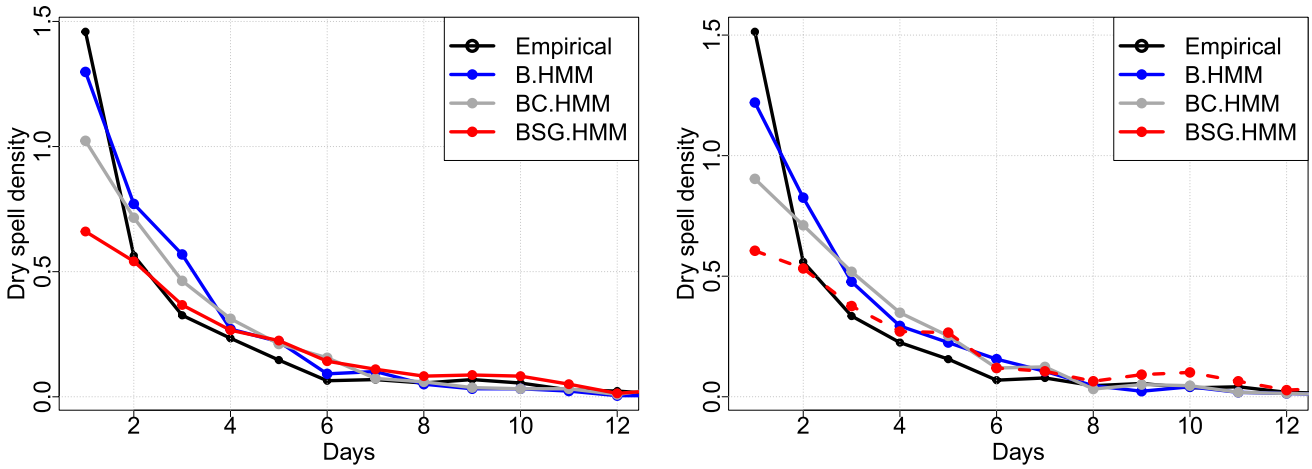


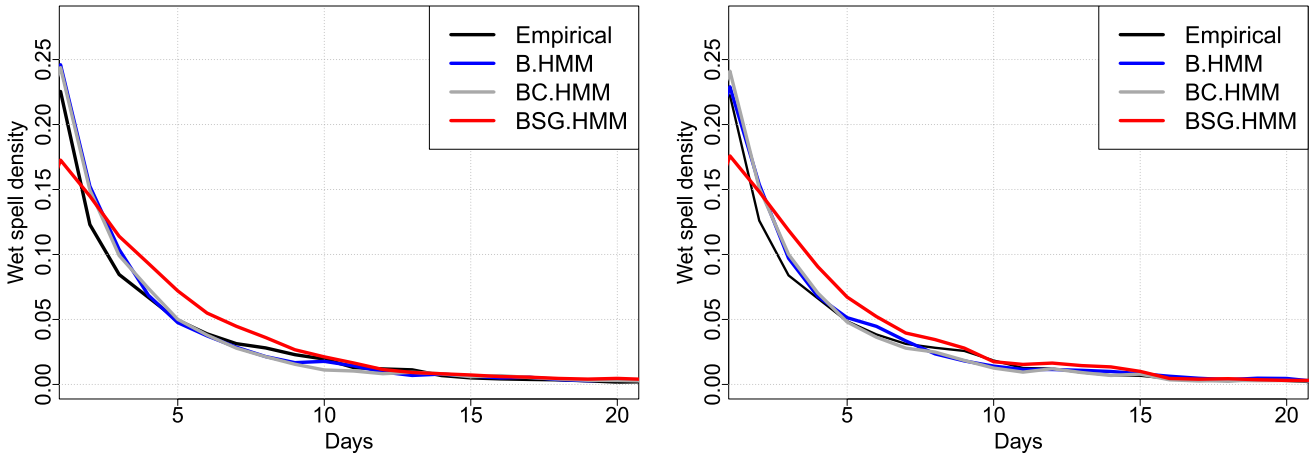
Fig. 7 Box plots comparing theoretical cross-correlation values (jittered black dots) from BSG.HMM (top) and BC.HMM (bottom) simulations with empirical cross-correlation estimates (red dots) for the Brent-Pinn station yearly rainfall series



(a) Brent (b) Pinn

Fig. 8 Density plots comparing empirical and model-predicted dry spell durations for Brent (left) and Pinn (right) stations

the yearly rainfall series. The superior performance of the BSG.HMM suggests that its ability to capture both spatial dependency using the copula framework, and temporal dependence, though to a lesser extend, through the Hidden Markov Chain allows for a more accurate representation of the rainfall process. This makes the BSG.HMM well-suited for applications where joint modelling of sites is important for understanding complex precipitation dynamics.



(a) Brent (b) Pinn

Fig. 9 Density plots comparing empirical and model-predicted wet spell durations for Brent (left) and Pinn (right) stations

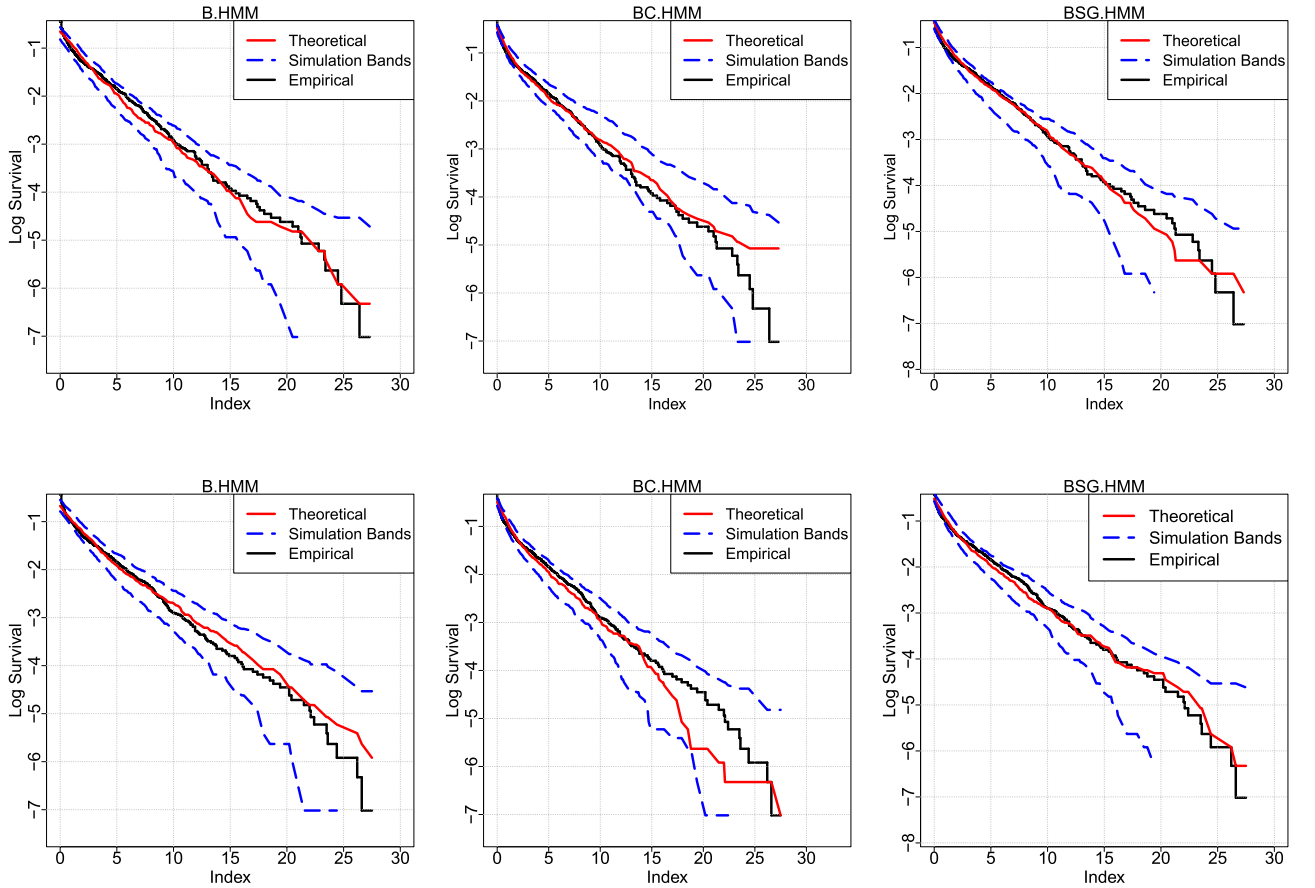


Fig. 10 Log-survival function plots comparing theoretical and empirical values for Seasonal precipitation series at Brent (top) and Pinn (bottom) stations

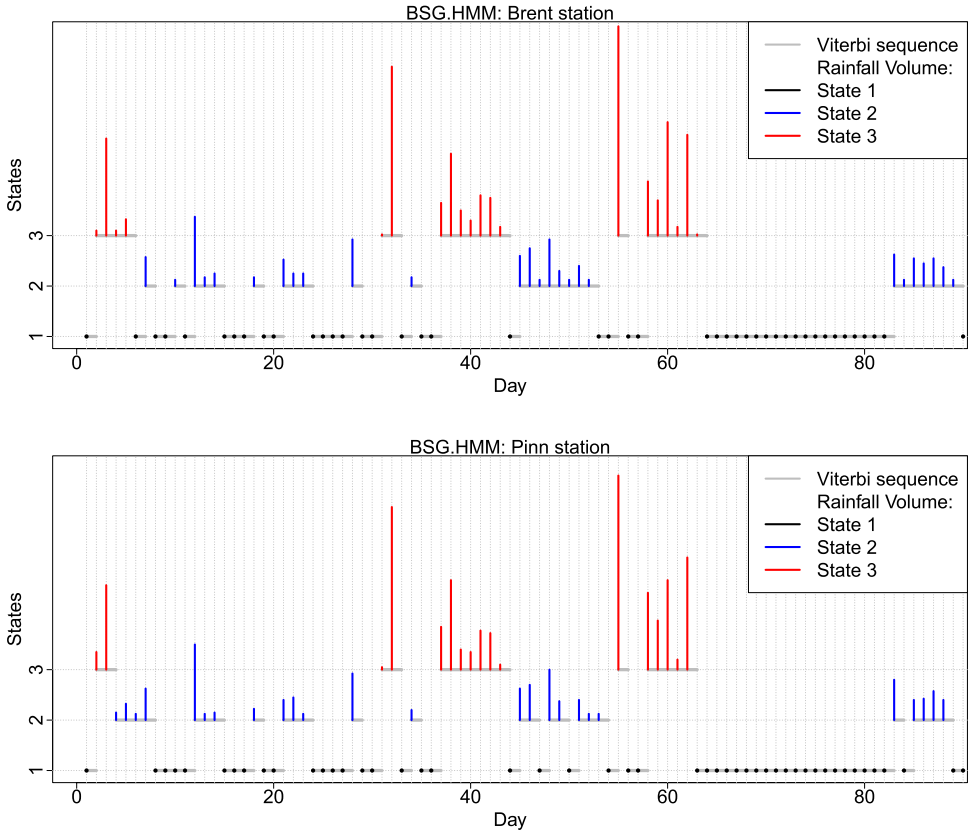


Fig. 11 Predicted Viterbi sequence derived from the BSG.HMM's estimated parameters and observed rainfall volume at Brent and Pinn station

Table 7 AIC and BIC values for the fitted bivariate models

Model	AIC	BIC
B.HMM	1677.861	1741.716
BC.HMM	1681.861	1758.487
BSG.HMM	1634.599	1711.225

3.4.2 Correlation structures

We now turn our focus to the correlation structures of the fitted models. Since the wet states are unobserved, it is not feasible to compare the empirical correlation with the theoretical correlation coefficients derived from the maximum likelihood estimates of the copula parameters. However, by leveraging the maximum likelihood estimates for the copula parameters within each hidden state, we can examine the dependence structure associated with different rainfall states. Figure 12 presents the copula density and corresponding contour plots for the bivariate models fitted to the Brent and Pinn station combination in state 2. As expected, the survival Gumbel copula, capturing strong dependence, shows a clustering of observations along the diagonal line, reflecting its characteristic asymmetrical dependence structure. This pattern indicates a high degree of dependence in the moderate rainfall state, where the copula model effectively captures the joint variability between the two stations. The bivariate density plots, also displayed in Fig. 12, exhibit similar behaviour, confirming the strong interdependence between the stations in this wet state.

In contrast, Fig. 13 depicts the density and contour plots for state 3, associated with heavier rainfall events. The survival Gumbel copula concentrates observations in the lower-left corner, indicating the heavy-rainfall correlation patterns. The strength of its upper-tail dependence is even more pronounced, with noticeably higher joint probabilities for co-occurring extreme events.. This suggests that the model effectively captures the increased synchronisation of heavy rainfall events across the two stations, a key feature in the modelling of extreme precipitation behaviour. These findings emphasise the capacity of copula-based models, particularly the survival Gumbel copula, to account for the complex dependence structures inherent in multivariate rainfall data across different states.

The bivariate case allows us to analyse an information based measure of multivariate association known as mutual information. Mutual information measures the average quantity of information communicated in one random variable about another, and can be derived from a copula-based joint density function (Zeng and Durrani 2011) as,

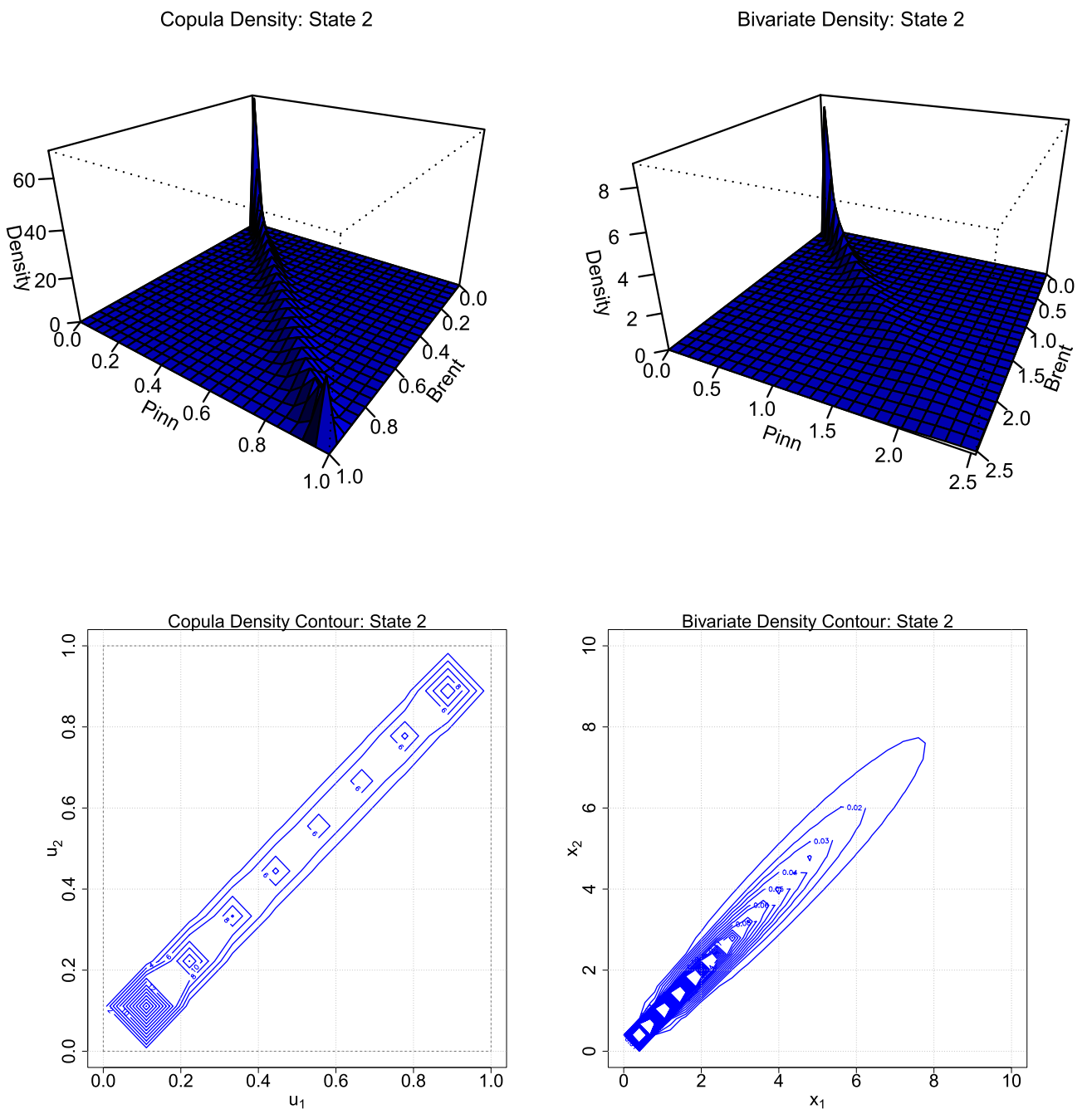


Fig. 12 Copula and bivariate density plots for State 2 (moderate rainfall) under the BSG.HMM

$$mi\left(y_t^{(1)}, y_t^{(2)}\right) = \iint\limits_{y_t^{(1)}, y_t^{(2)}} f\left(y_t^{(1)}, y_t^{(2)}\right) \ln\left[\frac{f\left(y_t^{(1)}, y_t^{(2)}\right)}{f\left(y_t^{(1)}\right) f\left(y_t^{(2)}\right)}\right] dy_t^{(2)} dy_t^{(1)} \tag{7}$$

A high estimate of mutual information suggests a high reduction in entropy. Indeed, it measures the difference between the entropy of one variable on its own and of that variable conditional upon the other (Latham and Roudi 2009). Therefore, by observing the precipitation process at one station, we reduce the uncertainty of the precipitation process at another station. According to Zeng and Durrani (2011), estimates of mutual information derived from copula-based multivariate distributions are reliable and are close to standard estimates from the observations' underlying distributions. The author also demonstrates the importance of choosing the best fitting copula in order to obtain precise estimates of the statistic. In Fig. 14, we examine

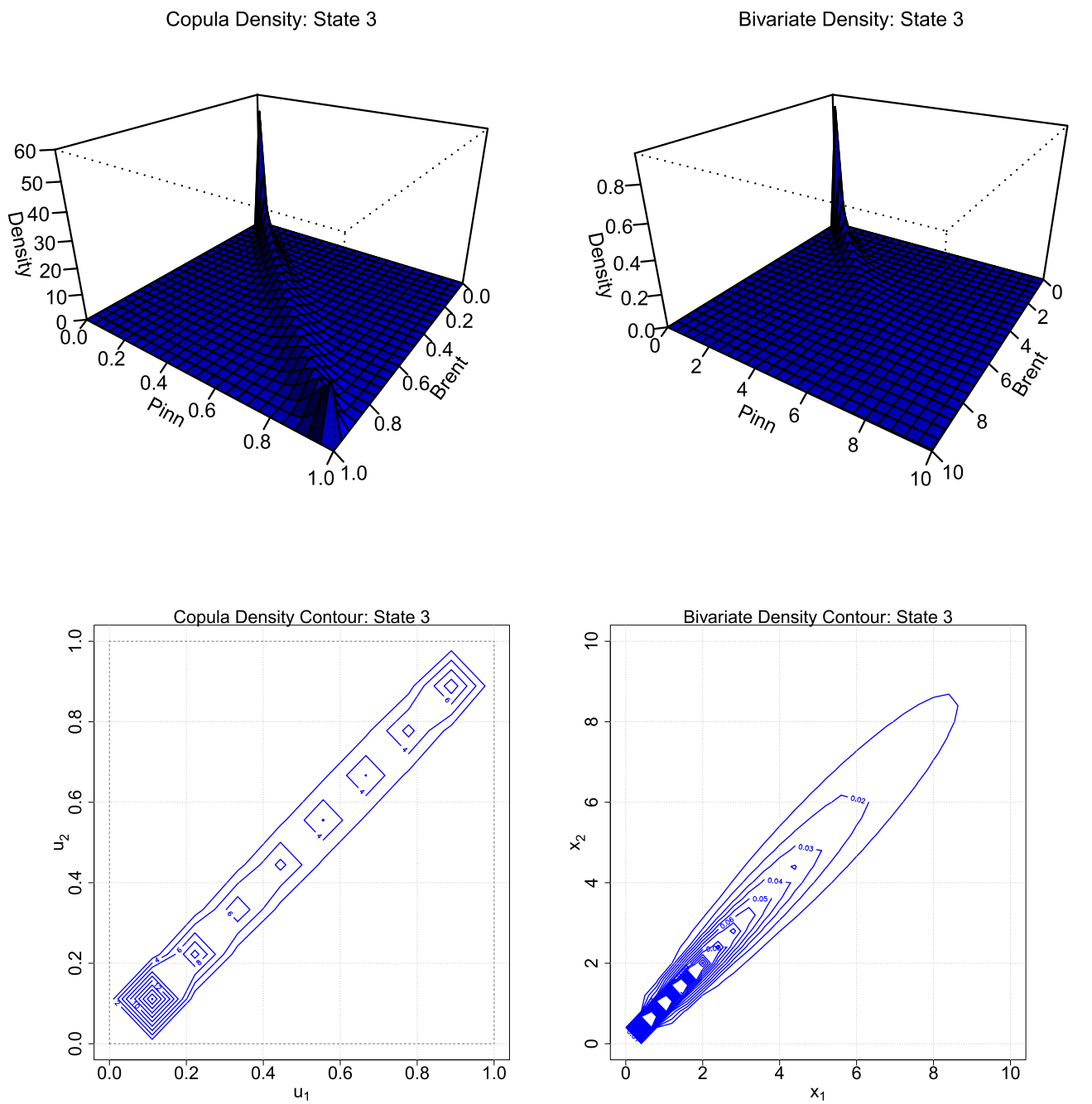


Fig. 13 Copula and bivariate density plots for State 3 (heavy rainfall) under the BSG.HMM

mutual information within wet states across varying levels of Kendall's τ. The maximum likelihood-estimated copula parameters range from 1 to 20, from which the correlation coefficients for the plot are derived. The mutual information in state 3 is higher than in state 2 for high correlation coefficients ($\tau > 0.45$), indicating that we are likely to gain more information about the rainfall process at another station when heavy rainfall occurs at the current station, compared to when moderate precipitation is observed. A high Kendall's tau estimate in both states, and consequently high mutual information, implies strong predictive performance of the precipitation series following the distribution. For $\tau < 45$, state 2 has marginally higher mutual information compared to state 3, signifying a poor predictive performance in either rainfall regimes.

3.4.3 Four-site models

Building on the bivariate models, we now extend the analysis to a multivariate framework, incorporating data from

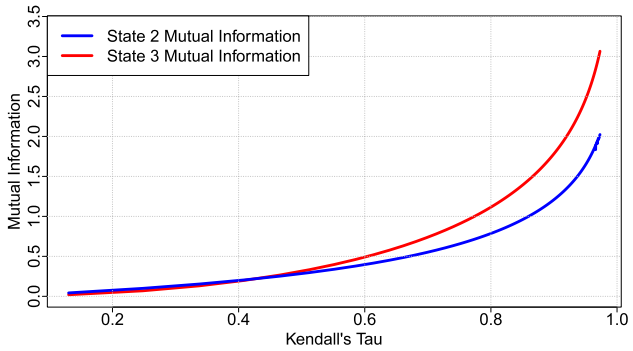


Fig. 14 Copula-derived mutual information between Brent and Pinn stations, quantifying the strength of dependence across all rainfall intensities

four different sites. The first model in this multivariate setting is the Multivariate Exponential Hidden Markov Model (M.HMM), which, like its bivariate counterpart, assumes contemporaneous independence between the sites but now operates across a higher dimensional space. The second model, the Multivariate Covariate Exponential Hidden Markov Model (MC.HMM), introduces temporal dependence by incorporating a 5-day moving average of daily temperature and pressure as covariates, allowing the model to dynamically adjust to shifts in rainfall patterns at all four locations. Lastly, the third model, the survival Gumbel Copula-embedded Multivariate Hidden Markov Model (MSG.HMM), utilises the copula framework to capture spatial and temporal dependence as described in Sect. 3.4.1. This model does not incorporate the covariates, as the combination of covariates and copula in the HMM increases the parameter space, making the optimisation unstable.

These multivariate adaptations are expected to offer a more nuanced representation of the statistical characteristics of the precipitation data, improving upon the bivariate models' performance by modelling both spatial and temporal interdependencies more comprehensively.

The generalised likelihood, in Eq. (6), for the models is adopted for the multivariate case, with corresponding modifications to the parameter sets. We introduce additional parameters for the marginal distributions to account for the new data dimensions. The modified diagonal matrices of the state distributions for the MC.HMM ($\Psi_{MC.HMM}$) and the MSG.HMM ($\Psi_{MSG.HMM}$) are defined as follows,

$$\Psi_{MC.HMM} = \begin{cases} diag\,[1, \quad 0, \quad 0] & y_t^{(s)} = 0 \\ diag\left[0, \quad \prod_{s=1}^{d} f_2\left(y_t^{(s)} | \beta_{02}^{(s)}, \beta_1, \beta_2\right), \quad \prod_{s=1}^{d} f_3\left(y_t^{(s)} | \beta_{03}^{(s)}, \beta_1, \beta_2\right)\right] & y_t^{(s)} \neq 0 \end{cases}$$

$$\Psi_{MSG.HMM} = \begin{cases} diag\,[1, \quad 0, \quad 0] & y_t^{(s)} = 0 \\ diag\left[0, \quad h_2\left(y_t^{(s)} | \lambda_2^{(1)}, \lambda_2^{(2)}, \lambda_2^{(3)}, \lambda_2^{(4)}, \theta_2\right), \quad h_3\left(y_t^{(s)} | \lambda_3^{(1)}, \lambda_3^{(2)}, \lambda_3^{(3)}, \lambda_3^{(4)}, \theta_3\right)\right] & y_t^{(s)} \neq 0 \end{cases}$$

where $\beta_{0j}^{(s)}$ are the state j intercept parameters at station s; β_1 and β_2 are the temperature and pressure covariate coefficient parameters respectively. To obtain the diagonal matrices of the Multivariate Hidden Markov Model (M.HMM), we simply remove the copula parameters θ_2 and θ_3 from $\Psi_{MSG.HMM}$.

These formulations extend the models' capacity to handle the multivariate data structure, incorporating both spatial and temporal dependencies across all four sites. Figure 15 presents a comparison between the empirical density plots for the Brent station and the theoretical densities generated from the M.HMM and the MSG.HMM. The figure illustrates the extent to which the models capture the observed data distribution, with the empirical densities overlaid on the theoretical curves. The models performed very well (including the MC.HMM), capturing the empirical density on both the tails and the body of the precipitation series. The models capture the upper tails of precipitation density plots in Fig. 15 by using state-switching dynamics and state-specific parameterisation. While exponential distributions are light-tailed, the HMM combines their outputs through transitions between states, therefore creating a mixture of exponentials with distinct rate parameters ($\lambda_2 > \lambda_3$). State 3 is governed by a smaller λ_3, and it dominates the upper tail

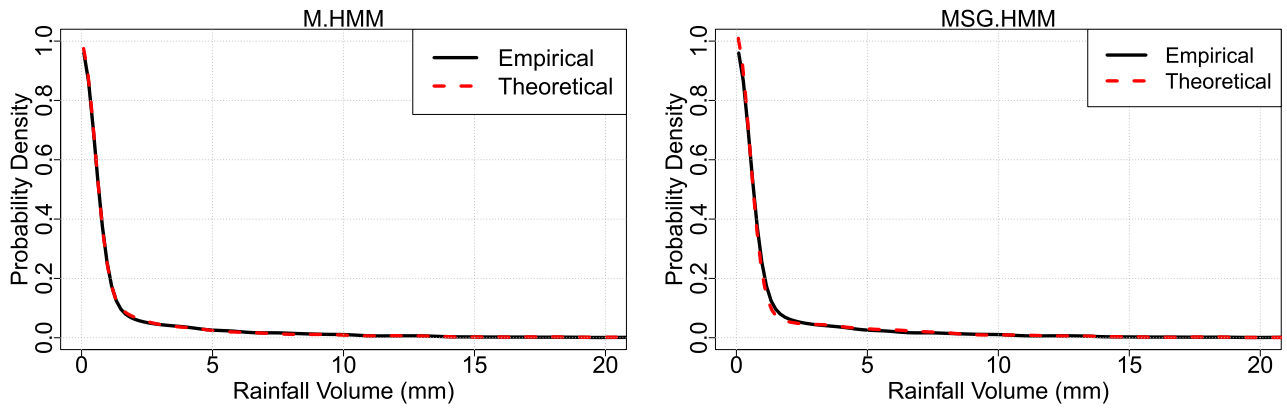


Fig. 15 Empirical and model-predicted density plots for daily rainfall at Brent station across the three models considered

Table 8 Statistical properties of Pinn station from the models' fit on yearly time series

	Mean	Variance	Autocorrelation lag				
			1	2	3	4	5
Empirical	1.8541	14.3022	0.2180	0.0980	0.0970	0.0530	0.0470
M.HMM	1.8010	13.7800	0.1373	0.0505	0.0186	0.0072	0.0021
MC.HMM	1.8430	13.9500	0.1733	0.0797	0.0386	0.0183	0.0087
MSG.HMM	1.8890	15.0500	0.1700	0.0764	0.0341	0.0155	0.0069

Table 9 Statistical properties of Pinn station from the models' fit on seasonal time series

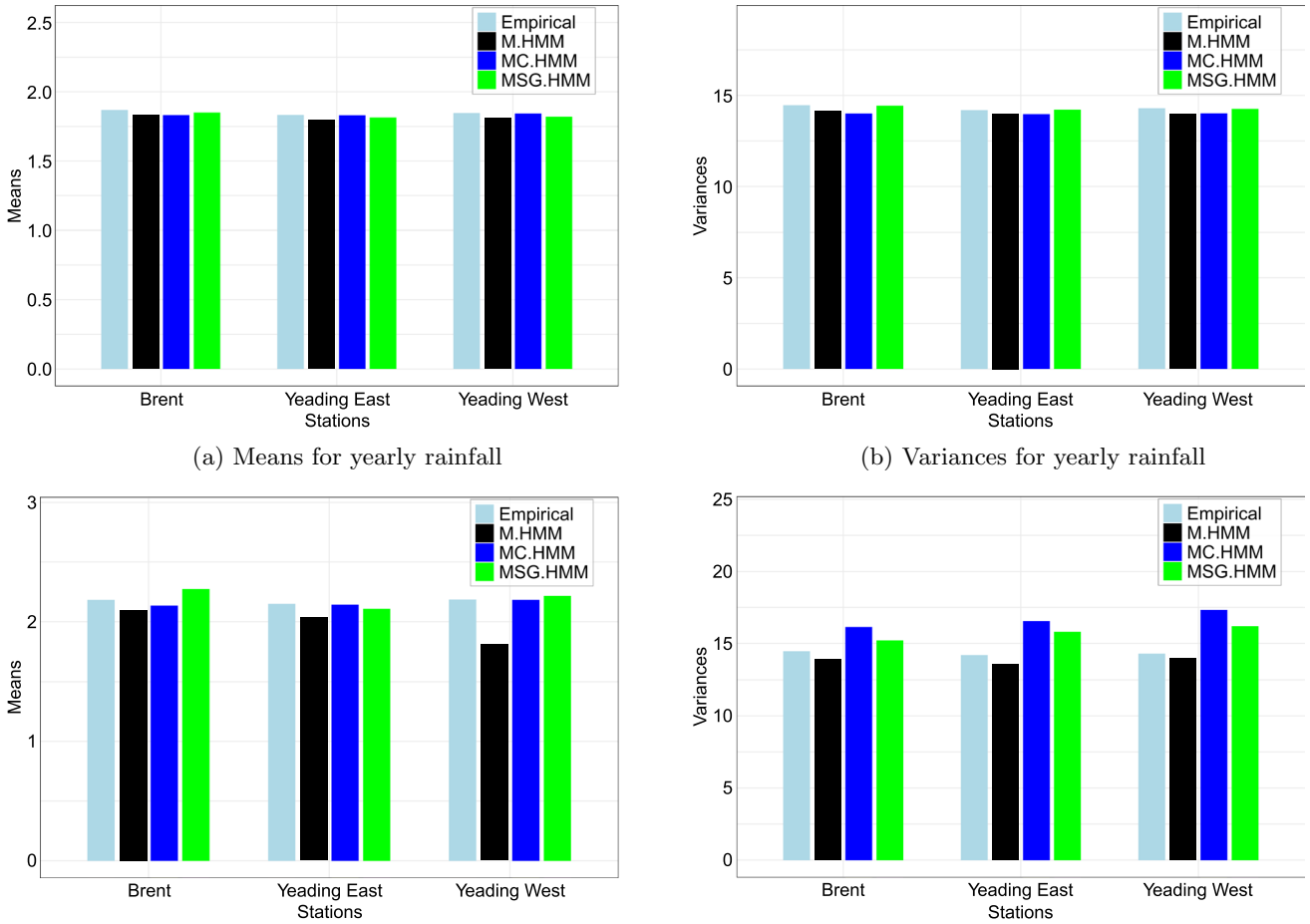
	Mean	Variance	Autocorrelation lag				
			1	2	3	4	5
Empirical	2.2205	16.7895	0.2590	0.1200	0.1210	0.0740	0.0370
M.HMM	2.1650	15.5000	0.07604	0.0172	0.0022	0.0011	– 0.0005
MC.HMM	2.1720	17.8000	0.1239	0.0484	0.0250	0.0170	0.0145
MSG.HMM	2.0499	14.0317	0.1275	0.0483	0.0189	0.0069	0.0024

behavior, as its slower exponential decay $(e^{-\lambda_3 y_t})$ allows prolonged periods of heavy rainfall. This mimics empirical heavier-tailed behavior, even if the tails are being captured by an exponential distribution. By isolating extreme rainfall into state 3 and regulating transitions, the HMM aligns simulated tail behaviour with the empirical tails, as shown in Fig. 15. Although not characteristically heavy-tailed, the model's state-specific flexibility and persistence enable it to capture tail behaviour by clustering high rainfall events into a state.

We simulated 2000 data sets, each equal in length to the daily precipitation records of the 12-year period. Summary statistics for each simulation were compiled, averaged, and presented in tabular and graphical form. A comparison of the empirical and theoretical statistical properties of Pinn station are provided in Table 8. The mean and variance were satisfactorily captured by most of the models; as expected, we noticed discrepancies when it came to autocorrelation. Despite the difference, the theoretical estimates are an improvement on the results, with the same precipitation data, based on the bivariate models.

For a precipitation simulation model, accurately capturing mean and variance statistics is fundamental before addressing any correlation structures. For instance, the Pinn mean estimate from the MSG.HMM in Table 9 significantly underestimates the true mean, with no notable improvement in other statistics. Such a model is unreliable when applied to a shorter seasonal series, as evidenced by the results in Table 9, which show even greater deviations from empirical values. In contrast, the M.HMM and MC.HMM exhibit much smaller deviations in mean and variance estimates from the seasonal empirical values, indicating their greater accuracy. Theoretical means and variances for other stations, alongside their empirical counterparts, are presented in Fig. 16. Accurately reproducing the mean and variance of the seasonal series posed a significant challenge for the quadrivariate models, and as before, this is likely a result of the volatility inherent in the much shorter seasonal time series. Based on a visual assessment, the MSG.HMM exhibited marginally better performance compared to the other models for the yearly series, while all the models performed poorly in modelling the seasonal series. This observation is further supported by the AIC and BIC values presented in Table 10. Both AIC and BIC values decrease as model complexity increases, both on yearly and seasonal timescales, indicating that more complex models, such as the MSG.HMM, provide a better fit to the respective series. This trend reinforces the conclusion that incorporating additional dependencies and structural enhancements improves model accuracy in capturing the underlying rainfall patterns.

Further analysis focuses on correlation measures of the simulations generated by each model. The rudimentary precipitation stochastic model assuming contemporaneous independence satisfactorily captures the mean and variance statistics, however, the model typically distorts critical information on the correlation framework; as evidenced by autocorrelation estimates from Tables 8 and 9. Ramesh and Onof (2014) gave proof of the positive influence of a dependence inclusive model in capturing autocorrelation. The authors' autoregressive-based model offered an improvement, though marginal, upon rudimentary models. The spatial and temporal dependence models we have constructed, with copula and covariate components chosen based on the nature of precipitation series, offer a significant improvement. Figure 17 shows the average autocorrelations from 2000 simulations of the four stations' yearly precipitation series; these are from an MSG.HMM fit. Among the 2000 simulations, 300 simulations were randomly sampled and their autocorrelation estimates are 'jittered' on the boxplots, showing the variability of estimates captured with the model. Plots of the empirical autocorrelation are superimposed on the boxplots for each station. Lag 3 autocorrelation proved a challenge, with none of the empirical estimates captured within the min-max simulation bands. We believe that this



(a) Means for yearly rainfall

(b) Variances for yearly rainfall

(c) Means for seasonal rainfall

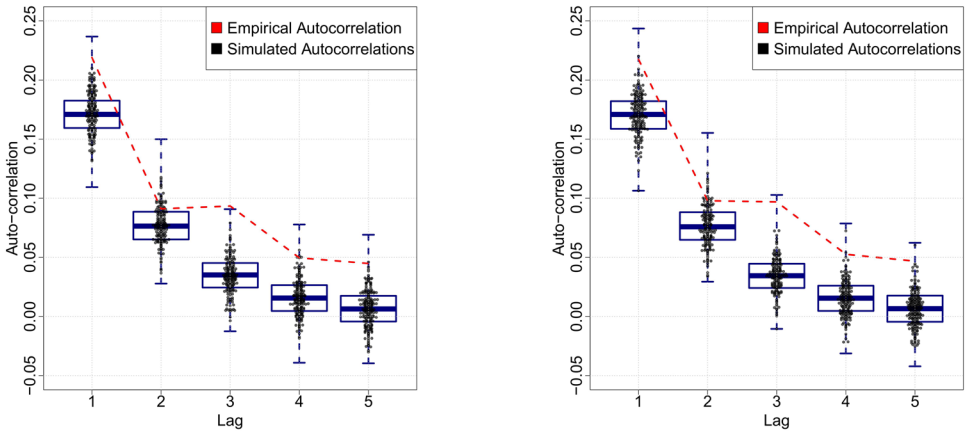
(d) Variances for seasonal rainfall

Fig. 16 Comparison of means and variances from the fitted models against observed values for yearly and seasonal precipitation series

Table 10 AIC and BIC values for the fitted quadrivariate models

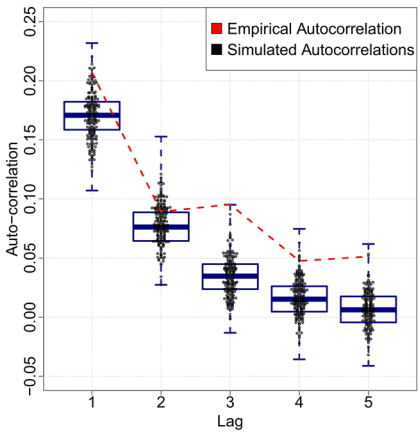
Model	Yearly Series		Seasonal Series	
	AIC	BIC	AIC	BIC
M.HMM	3573.9810	3663.3780	1087.1890	1157.4340
MC.HMM	3209.5690	3311.7360	953.0349	1033.3150
MSG.HMM	2242.482	2344.65	630.6725	630.6725

is attributed more to the data than the model itself, because a similar pattern was observed from the MC.HMM fit. All but the lag-2 series at Yeading East failed to reproduce the observed serial correlation in the inter-quartile range because the exchangeable Archimedean copula imposes a one-size-fits-all dependence pattern that erodes the heterogeneity driving temporal patterns. By forcing every station pair to share the same association, the model eliminates temporal memory effects that vary from site to site. In addition, because the copula was applied to pre-fitted marginal distributions, any small mis-specification in the marginal fits further erodes the ability of the joint model to recover the true lagged correlation. It should also be noted that despite the improvement on Ramesh and Onof (2014), the model-based simulations inflate the dry state stationary probability by 10%. Multivariate likelihood functions are more susceptible to the issue of zero-inflation, which can significantly affect model outcomes. This leads to an overestimation of dry days, as reflected in the parameter estimates for all models (Table 11). The models performed exceptionally well for the seasonal series, with the MC.HMM providing autocorrelation estimates closest to the empirical values (Fig. 18); even the lag 3 autocorrelation was within range of the min-max simulation bands. We can point to the increase in the number of wet days to justify this improvement, as the functionality of the correlation structure is optimised. However, it should be noted that further efforts to minimise the dry spell, for example by taking the monthly precipitation series, will increase the variability, thus making it difficult to capture statistical properties. The MSG.HMM failed to reproduce the lag 1 autocorrelation at all stations, a likely consequence of the increased variability, which in turn distorts the copula component's capacity to capture temporal dependence. Maximum likelihood estimates from the fit on seasonal precipitation data are given in Table 12.

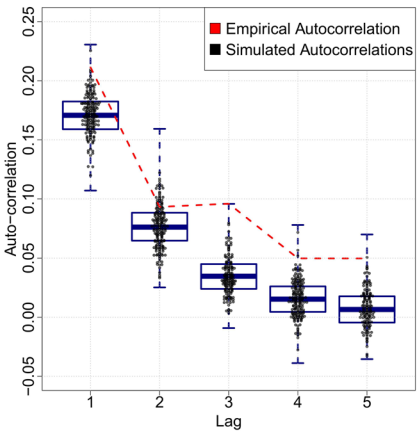


(a) Brent

(b) Pinn



(c) Yeading East



(d) Yeading West

Fig. 17 Box plots comparing theoretical autocorrelation (with jittered black dots) and empirical autocorrelation (red) for the yearly rainfall series at the four stations

Correlation coefficients derived from maximum likelihood estimates of the copulas (Tables 11 and 12) point towards an approximately equal spatial dependence in state 2 and state 3 of the underlying Markov chain for both yearly and seasonal rainfall series. Kendall's τ values derived from copula parameter estimates in state 2 and 3 are 0.9070 and 0.9073 respectively for the yearly series; 0.9106 and 0.9105 respectively for the seasonal series. This suggests that spatial dependence remains consistently strong across different rainfall intensities, likely due to the close proximity of the stations and shared weather systems. While heavy rainfall events are typically widespread, moderate rainfall volumes can still align at stations under similar atmospheric conditions. Minor variations, such as localised drizzle at one site while another experiences moderate rain, may occur but are not significant enough to reduce the overall dependence.

We also assessed the cross-correlation of different pairwise combinations of the four stations. Cross-correlations from each of the 2000 model-based simulations were compared against the empirical estimates. Figure 19 shows the Pinn-Yeading East and Pinn-Yeading West combinations for the MSG.HMM fit on the yearly series. Lag zero correlations estimates for the Pinn-Yeading West combination underestimated the empirical spatial dependence likely due the uniform state correlations imposed on the different pairwise combinations, a model feature that ignores pair-specific variations. Inspite of this, estimates aligned more closely with the empirical values than those produced by the alternative models, indicating a relatively better fit despite the constraint. Non-spatial models generally underestimate the lag zero cross-correlation, it is only when we move to lag one that they start to accurately reproduce the statistic. We revert to the likelihood estimation technique we mentioned before to justify the substantial contrast in the lag zero theoretical and empirical values. The model could be tuned so that if there is only one dry station among the four we could pass it on as a wet day, however, this would also have an adverse impact on the scaling of numerical underflow and overflow. The copula component of the MSG.HMM enhances its capability to capture spatial dependence as shown in Fig. 19. We move on to analyse other performance metrics in literature, comparing empirical values

Table 11 Maximum likelihood estimates from the fit on yearly rainfall

Model	p_{12}	p_{13}	p_{21}	p_{23}	p_{31}	p_{32}	$\lambda_2^{(1)}$	$\lambda_3^{(1)}$	$\lambda_2^{(2)}$	$\lambda_3^{(2)}$	$\lambda_2^{(3)}$	$\lambda_3^{(3)}$	$\lambda_2^{(4)}$	$\lambda_3^{(4)}$	–	–
M.HMM	0.15	0.17	0.40	0.27	0.28	0.20	0.71	0.18	0.73	0.19	0.77	0.19	0.74	0.19	-	-
	p_{12}	p_{13}	p_{21}	p_{23}	p_{31}	p_{32}	$\beta_{02}^{(1)}$	$\beta_{03}^{(1)}$	$\beta_{02}^{(2)}$	$\beta_{03}^{(2)}$	$\beta_{02}^{(3)}$	$\beta_{03}^{(3)}$	$\beta_{02}^{(4)}$	$\beta_{03}^{(4)}$	β_1	β_2
MC.HMM	0.11	0.12	0.37	0.28	0.29	0.19	−1.09	−1.66	−1.11	−1.66	−1.07	−1.66	−1.09	−1.66	−0.02	0.02
	p_{12}	p_{13}	p_{21}	p_{23}	p_{31}	p_{32}	$\lambda_2^{(1)}$	$\lambda_3^{(1)}$	$\lambda_2^{(2)}$	$\lambda_3^{(2)}$	$\lambda_2^{(3)}$	$\lambda_3^{(3)}$	$\lambda_2^{(4)}$	$\lambda_3^{(4)}$	θ_2	θ_3
MSG.HMM	0.10	0.12	0.38	0.28	0.29	0.20	0.29	0.19	0.33	0.18	0.34	0.19	0.33	0.19	10.75	10.79

with the outcomes from the fitted values. One such metric, is the dry and wet spell distribution for a station. In Fig. 20 we observe that the three models overestimate the short duration (1–5 days) wet spells for the seasonal series at Yeading West. The yearly wet spell distribution at Yeading West station is adequately reproduced by the three models during both short and long durations.

Despite the contrasting day one densities, discrepancies between the observed and fitted dry spell distributions are minimal as the days increase in Fig. 21. As mentioned before, this is a result of the methods adopted to estimate the likelihood value.

The log transformation of the survival function, commonly known as the log-survival function, for the observed and fitted series is presented in Fig. 22. Min–max simulation bands derived from 1000 simulations have also been included for a clear assessment. The three models perform exceptionally well, producing fitted values similar to the empirical log-survival function in the yearly series. The MC.HMM failed to capture the seasonal empirical log-survival, meaning it could not adequately reproduce the observed seasonal frequency and magnitude of extreme rainfall events. This may stem from the covariates lacking seasonal sensitivity and/or the exponential distribution being ill-suited for seasonal extremes.

Figure 23 presents the predicted state sequence of the Yeading West precipitation series and the scaled rainfall volume for the first three months of the series. The underlying hidden-state sequence is then decoded via the Viterbi algorithm, using the maximum-likelihood estimates of the transition probability matrix, the state-conditional distribution parameters, and the observed rainfall series. In practice, once the HMM parameters are estimated, Viterbi finds the single most likely path of wet- and dry-state assignments that could have generated the data. We then overlay the decoded states on the observed rainfall volumes, plotting daily rainfall as vertical bars color-coded by the Viterbi-inferred state. As Fig. 23 demonstrates, this approach captures the timing and intensity of precipitation events exceptionally well, with the state-decoded dry and wet periods aligning closely with the actual rainfall record.

4 Discussion

4.1 Evaluation of model performance in capturing rainfall characteristics

Our progression from the two-site to the four-site Hidden Markov frameworks highlights that realistic rainfall simulation requires a balance of marginal flexibility, temporal memory, and a strong spatial link. In the bivariate setting,

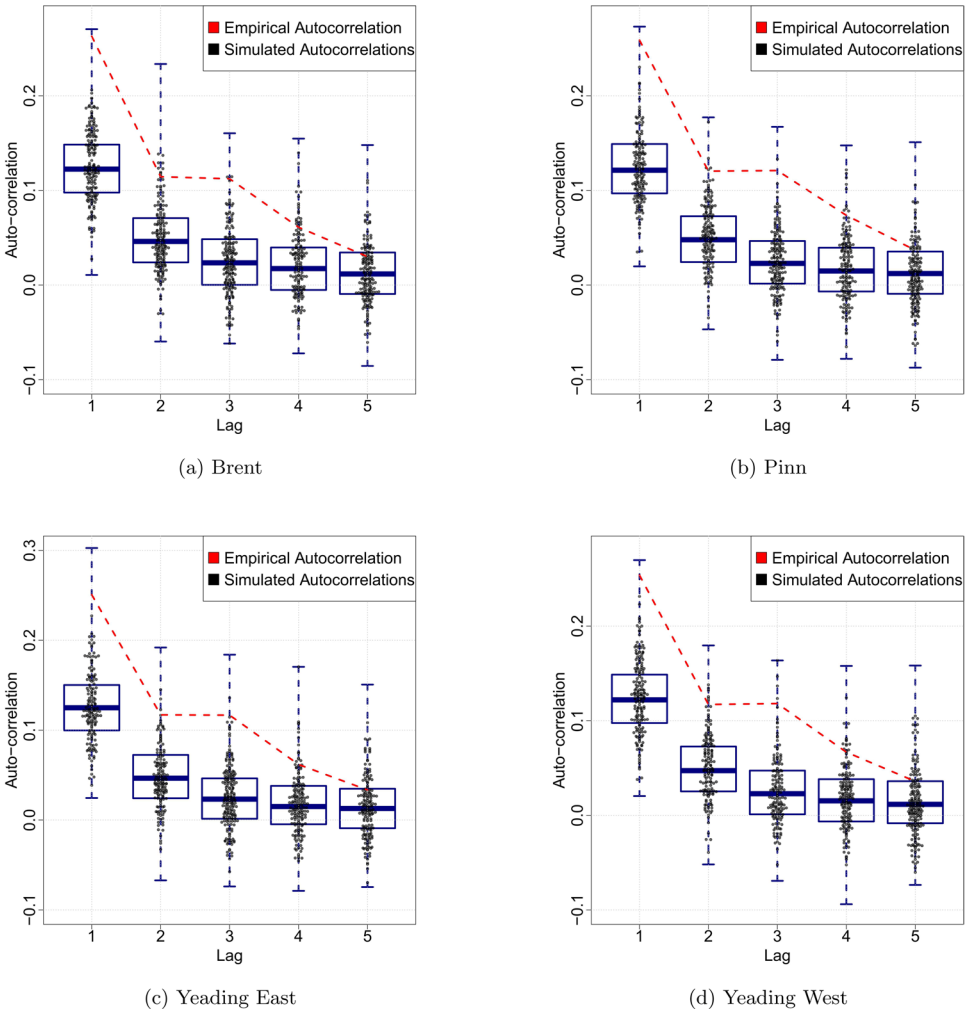


Fig. 18 Box plots comparing theoretical autocorrelation (with jittered black dots) and empirical autocorrelation (red) for the seasonal rainfall series at the four stations

the BSG.HMM, embedded with a survival Gumbel copula, excelled at reproducing co-occurrences of extreme rainfall at Brent and Pinn. This is crucial for flood risk assessment, since heavy-rain events tend to propagate regionally along synoptic fronts (Berry et al. 2011). The asymmetric upper-tail dependence that the survival Gumbel captures aligns with findings by Salvadori and De Michele (2004), reinforcing that symmetric copulas systematically understate joint extremes.

Scaling up the dimensionalty, the MSG.HMM brought those same tail-dependence advantages to a four-station network, confirming the findings from Banerjee et al. (2023) that intense storm systems imprint a coherent extreme-value signature across multiple sites. Although introducing covariates alongside the copula (in a single model) proved numerically challenging, the pure copula-HMM already improved joint tail fit over models assuming contemporaneous independence.

Turning to the temporal dynamics, a bivariate BC.HMM, with daily-varying exponential parameters, satisfactorily captured observed autocorrelation within min-max simulation bands. This suggests that, in moderate rainfall regimes, flexible marginals can sometimes mimic deeper state-persistence effects. Yet when we explicitly conditioned transitions on a 5-day moving average of temperature and pressure in the MC.HMM, the four-site autocorrelation at lags 1–3 aligned even more closely with empirical data. This underscores that temperature and pressure covariates increase the realism about frontal passages. In the MSG.HMM, the inability of all but the lag-2 series at Yeading East to capture serial correlation in the inter-quartile range can be traced to three issues: the exchangeable Archimedean copula's one-size-fits-all assumption of identical dependence across station pairs erases the moderate, site-specific memory effects driven by local micro-climates and catchment characteristics; its limited flexibility prevents it from strengthening site specific correlations. Any misfitting marginal distributions, where observations are most concentrated, further undermines the joint model's ability to recover true lagged structure. Only at Yeading East lag 2 do the local temporal dynamics happen to align with these

Table 12 Maximum likelihood estimates from the fits on seasonal rainfall

Model	p_{12}	p_{13}	p_{21}	p_{23}	p_{31}	p_{32}	$\lambda_2^{(1)}$	$\lambda_3^{(1)}$	$\lambda_2^{(2)}$	$\lambda_3^{(2)}$	$\lambda_2^{(3)}$	$\lambda_3^{(3)}$	$\lambda_2^{(4)}$	$\lambda_3^{(4)}$	–	–
M.HMM	0.25	0.27	0.42	0.27	0.32	0.17	0.72	0.20	0.76	0.19	0.85	0.20	0.87	0.19	-	-
	p_{12}	p_{13}	p_{21}	p_{23}	p_{31}	p_{32}	$\beta_{02}^{(1)}$	$\beta_{03}^{(1)}$	$\beta_{02}^{(2)}$	$\beta_{03}^{(2)}$	$\beta_{02}^{(3)}$	$\beta_{03}^{(3)}$	$\beta_{02}^{(4)}$	$\beta_{03}^{(4)}$	β_1	β_2
MC.HMM	0.18	0.18	0.40	0.28	0.32	0.19	−0.86	−1.62	−0.85	−1.65	−0.81	−1.64	−0.83	−1.65	−0.05	0.05
	p_{12}	p_{13}	p_{21}	p_{23}	p_{31}	p_{32}	$\lambda_2^{(1)}$	$\lambda_3^{(1)}$	$\lambda_2^{(2)}$	$\lambda_3^{(2)}$	$\lambda_2^{(3)}$	$\lambda_3^{(3)}$	$\lambda_2^{(4)}$	$\lambda_3^{(4)}$	θ_2	θ_3
MSG.HMM	0.14	0.19	0.29	0.32	0.26	0.20	0.57	0.18	0.59	0.20	0.66	0.19	0.63	0.18	11.18	11.17

restrictive assumptions, allowing that one series to retain its expected serial correlation.

The analysis of wet and dry spell length patterns highlights key differences between the models. The MSG.HMM reduced the usual problem in basic Markov chains where wet or dry periods are often too long. It gave a more realistic mix of short one-day rainfall events and longer wet spells at two stations. The M.HMM, which covers four sites, was accurate in matching the average rainfall and its variability, but it struggled to represent how wet and dry spells actually occurred. Adding temperature and pressure covariates to the MC.HMM slightly improved the timing of rainfall events. However, only the MSG.HMM, could accurately reproduce how wet and dry spells played out across the sites.

Spatial dependence proved to be quite uniform in the wet regimes: Kendall's τ floated around 0.91 in both moderate and heavy precipitation states for two stations, and the MSG.HMM's single-parameter copula delivered closer lag-0 cross-correlations than the independence and temporal dependence models in the quadrivariate case. In meteorological terms, this reflects the dominance of slow-moving frontal systems that synchronise rainfall throughout the Hillingdon region. Yet applying a uniform dependence structure across all pairwise combinations of the stations also masks stronger or weaker associations influenced by the local topography, highlighting the need for vine copulas or hierarchical constructions to capture these differences.

While the exponential distribution is inherently light-tailed, the HMM framework demonstrates its practicality in modelling rainfall extremes by leveraging state-switching dynamics. Isolating heavy rainfall into a state 3 with a smaller rate parameter makes the model mimic heavy-tailed behavior, therefore aligning the tail behaviour of simulated densities with empirical densities in the yearly series. However, the exponential distribution is limited in the seasonal context. This suggests that while state-specific exponential distributions can approximate tail behavior through state switching dynamics, they remain light-tailed and struggle with seasonal series, which has volatile extremes. For robust modeling of the heavy-tails, employing distributions like the Weibull, Gamma or Pareto in the wet states may be necessary, though the HMM's flexibility may be suited for yearly rainfall patterns.

We noticed that the autocorrelation showed an increase at lag 3, that is, when comparing rain today to rain three days ago. This pattern likely reflects a common weather cycle, where a strong rain front brings heavy rain, then there's a break or lull in rainfall, followed by more rain a few days later as the frontal system circles back. This 3-day cycle is typical in frontal systems that pass through a region. While the MC.HMM introduces covariate-driven transitions to account for atmospheric variability, it still failed to

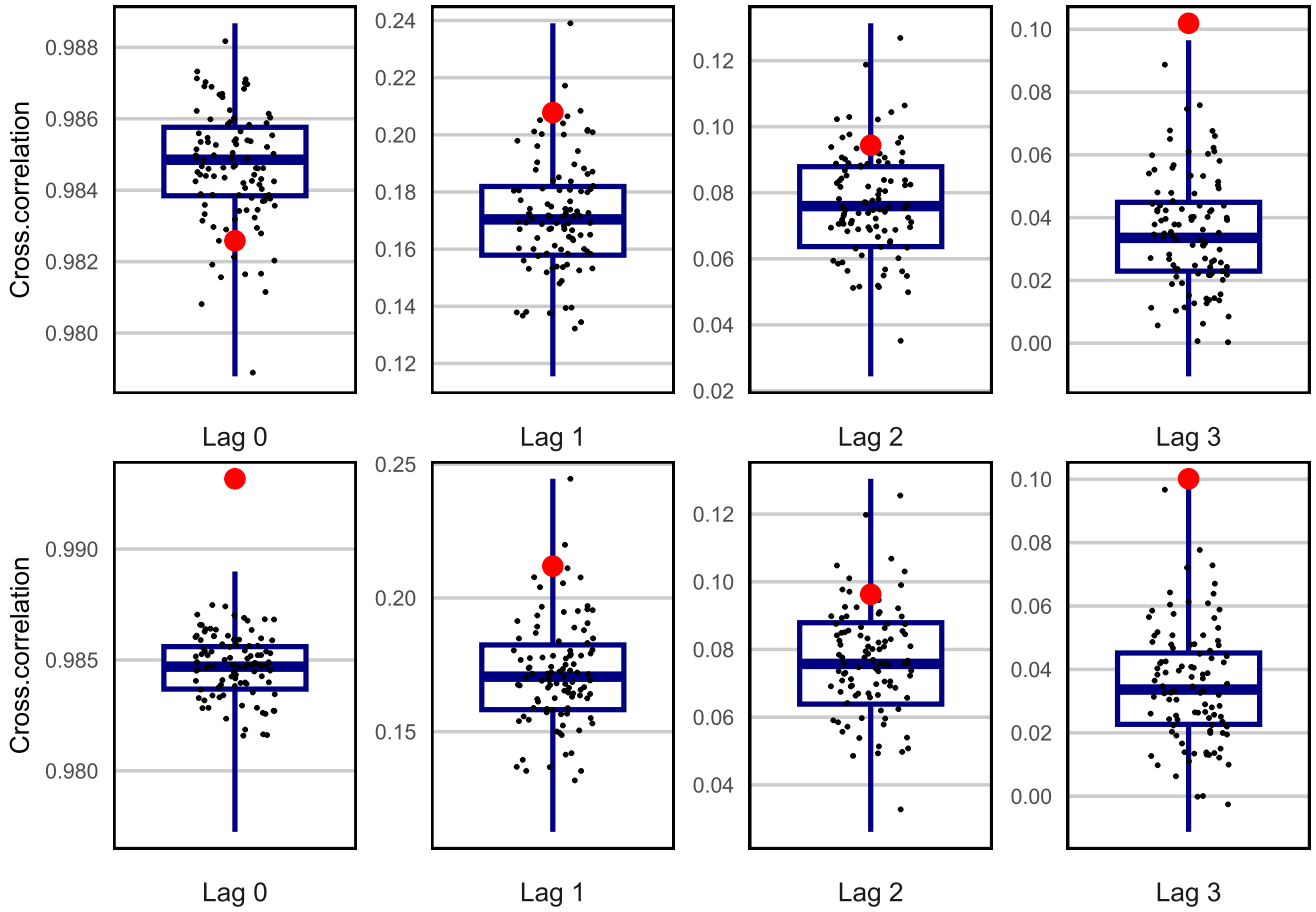
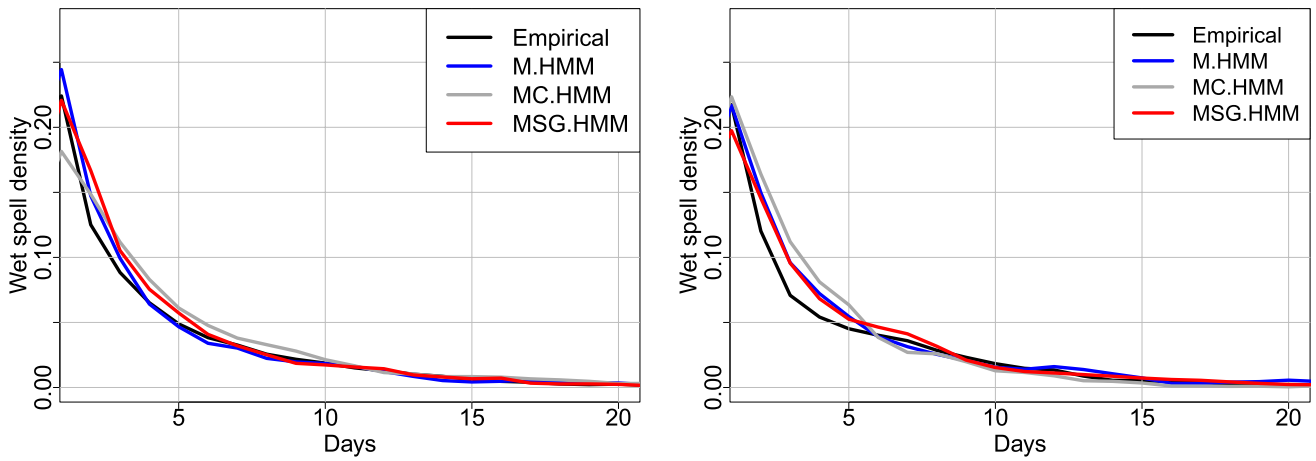


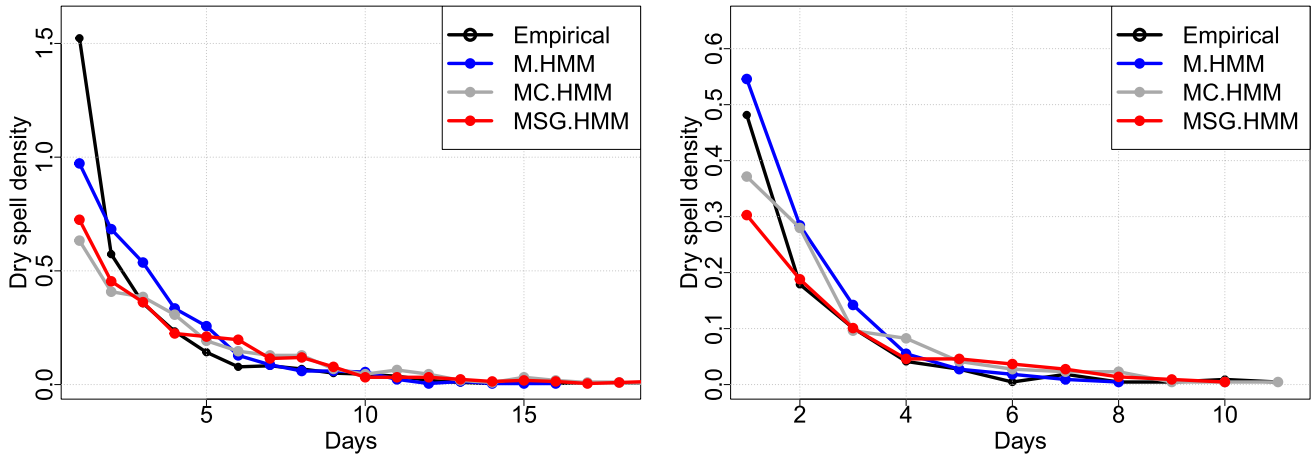
Fig. 19 Box plots comparing theoretical cross-correlation values (jittered black dots) derived from the MSG.HMM with empirical cross-correlation estimates (red dots) for the Brent-Pinn (top) and Brent-Yeading West (bottom) yearly rainfall series



(a) Yearly

(b) Seasonal

Fig. 20 Density plots comparing empirical and theoretical wet spell durations for Yearly (left) and Seasonal (right) precipitations series at Yeading West station



(a) Yearly (b) Seasonal

Fig. 21 Density plots comparing observed and fitted dry spell durations for Yearly (left) and Seasonal (right) precipitation series at Yeading West station

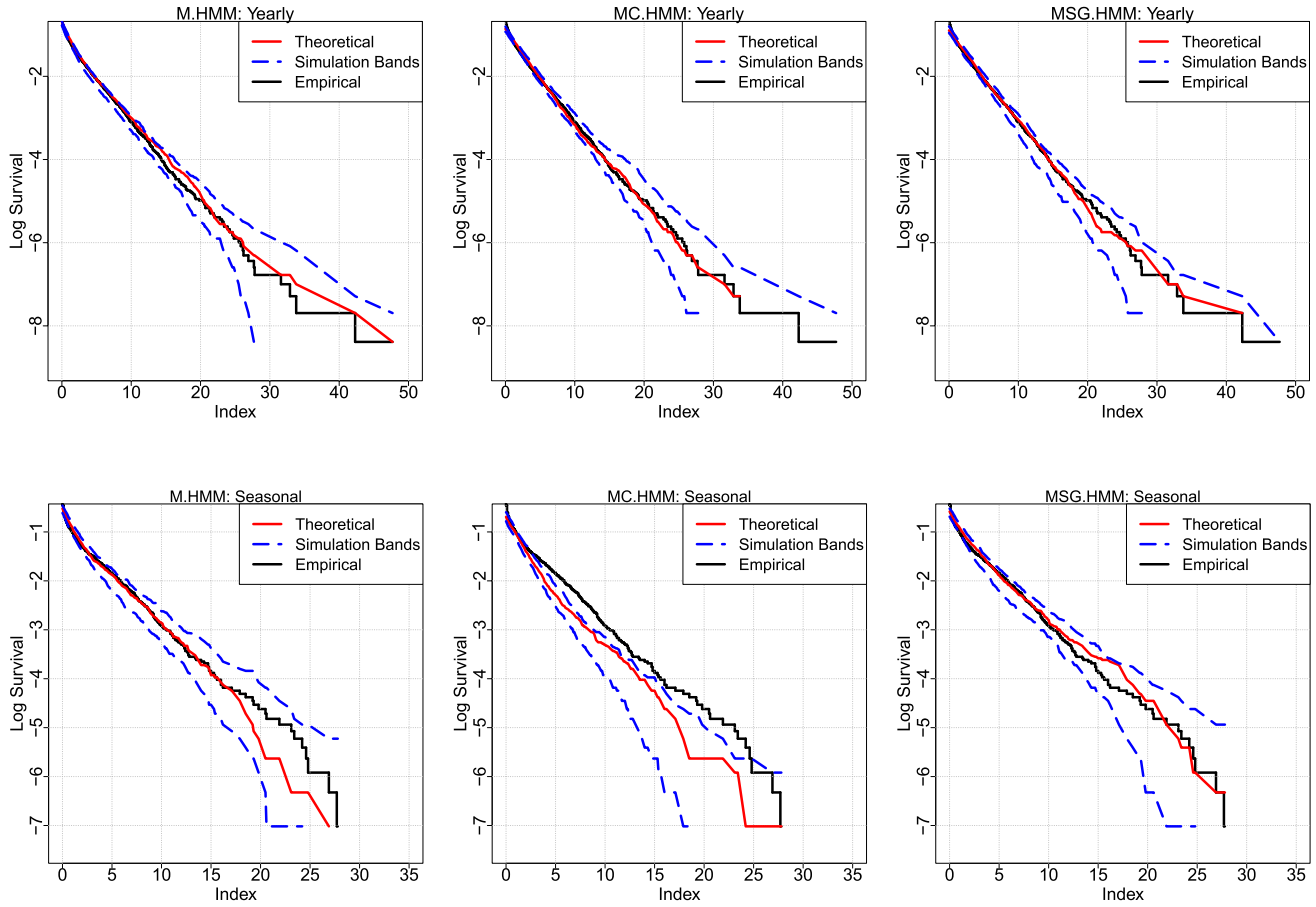


Fig. 22 Log-survival function plots comparing theoretical and empirical values for Yearly (top) and Seasonal (bottom) precipitation series at Yeading West station

reproduce the empirical bump at lag 3. This suggests that the chosen covariates, though useful for general modelling, may not capture temporal dynamics of precipitation. Such lag anomalies likely require specific and detailed weather indicators, like fine scale data or cloud cover, humidity, wind speed etc.

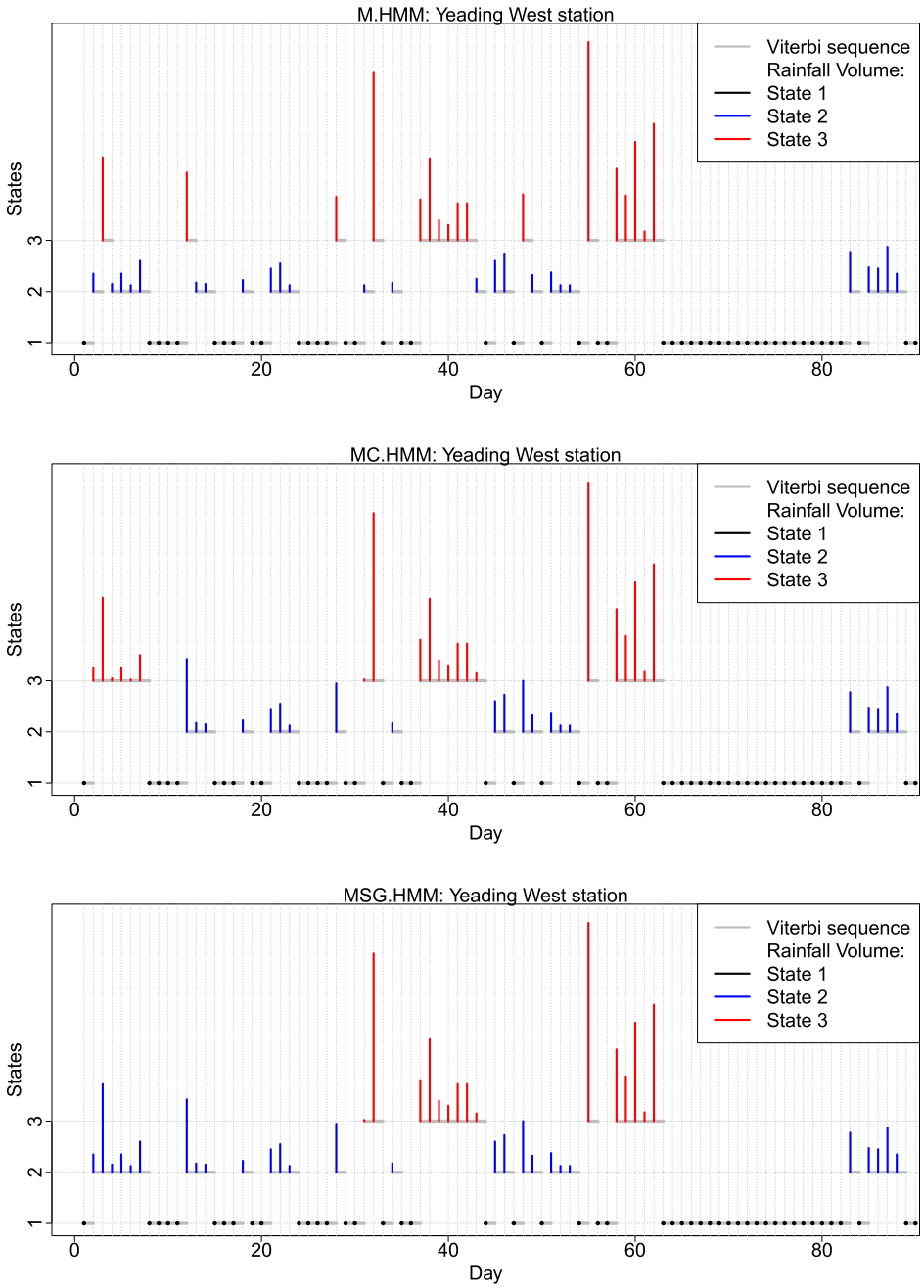


Fig. 23 Predicted Viterbi sequence and observed rainfall volumes at Yeading West station for the three models considered

4.2 Benchmarking the survival Gumbel-HMM against generalised linear models

The generalised linear model (GLM) framework, proposed by Yang et al. (2005), Asong et al. (2016), Chandler (2020), etc., is one of the popular methodologies for generating synthetic daily rainfall data. The approach is computationally efficient; however, its simplified structure introduces limitations in capturing the full complexity of rainfall dynamics. By decoupling the precipitation process into two components (the occurrence process where logistic regression governs rainfall occurrence and the intensity process where a gamma distribution models conditional rainfall amounts), the framework imposes an independence assumption between the probability of rainfall events and the associated magnitudes. This separation ignores common precipitation events, such as the transition from long dry spells to heavy rainfall events, particularly during autumn and winter in England. The copula-embedded HMM integrates the rainfall occurrence and intensity processes by employing latent

state-specific multivariate density functions, which capture multivariate dependencies within each hidden state.

GLMs assume that daily rainfall depends only on selected covariates, whose effect is linear, and ignores any unusual precipitation patterns, such as those caused by monsoons. The hidden states of the proposed model follow a Markov chain, which means that the state on day t depends on the state on day $t-1$. This creates temporal dependence through the non-linear state sequence, indirectly linking rainfall across days. By leveraging a copula, the model captures rainfall intensity within each state through a multivariate probability distribution that explicitly incorporates spatial dependencies across the region. In practice, we found this to significantly improve the realism of simulated rainfall, especially in preserving co-occurring rainfall intensities across multiple stations. Although the copula itself does not directly add to temporal dependence, its role in reducing variability means that the time series generated by the HMM better reflects the natural persistence and changes found in actual weather records. Because of these features, the copula-embedded HMM can reproduce clusters of wet days, such as monsoons-influenced storms.

Extensions of GLMs for capturing spatial dependence often involve transforming precipitation data to a Gaussian scale via the Anscombe transformation (Yang et al. 2005; Ambrosino et al. 2014), which forces the data into a symmetric, thin-tailed distribution. Gaussian random fields, however, lack tail dependence and therefore cannot capture the co-occurrence of extreme events across sites, an important aspect when modelling convective storms. Even advanced models like those proposed by Asong et al. (2016), which incorporate topographic attributes and seasonal basis functions to represent regional variations, still rely on a Gaussian assumption that smooths out the sharp transitions observed in actual rainfall patterns. A copula-embedded HMM decouples the marginal from the dependence: the HMM models temporal state transitions while the copula captures spatial dependence, including on the tails. Within the copula framework we can choose a copula whose tail activity match the empirical extreme co-occurrences. In our results, the bivariate and multivariate survival Gumbel HMM was able to replicate observed clustering of extremes, including simultaneous heavy rainfall events. These patterns were severely underestimated in the Gaussian-transformed GLM variants, underscoring the importance of tail modelling in extreme event simulation. The resulting synthetic time series from the survival Gumbel HMMs preserved both the joint spell structures, bringing simulations closer to the empirical multivariate behavior.

4.3 Framework adaptability in diverse climates and multidimesional settings

The London region exhibits a temperate maritime climate, characterized by mild winters, cool (but not hot) summers, and evenly distributed moderate rainfall. The model is suited for regions with similar climates or even those experiencing more wet days, such as tropical rainforest climates, because more data allows it to capture the nuances of spatial dependence for a synthetic rainfall generation process closely resembling the empirical activity. In contrast, applying the model to regions with less rainfall (e.g., Mediterranean, semi-arid, or desert climates) would require more data, and its efficiency might be decreased since precipitation co-occurrences are less common.

One of the main limitations of the model is the zero-inflation issue, arising when even one station records a dry day, causing the HMM to classify the entire region as dry for that day. This results in fewer wet days available for the copula to model spatial dependence, which in turn limits the copula's ability to accurately capture the complex daily precipitation patterns at high dimensions. When too many days are classified as dry, the model has insufficient data on wet days to reliably estimate the intricate dependencies between stations, especially in regions with low rainfall. In such regions, the frequent occurrence of dry days makes it challenging to simulate realistic spatial correlations, therefore reducing the model's effectiveness in capturing the overall precipitation patterns. Another limitation of the copula-embedded HMM is its inability to capture spatial intermittency, which refers to the irregular distribution of rainfall where nearby stations might exhibit contrasting states, some in wet conditions while others are dry at the same time.

As the dimensionality increases, or as we incorporate more sites, more data is required, and the parameter estimation process can become unstable. In higher dimensions, assuming the same dependence structure across widely separated stations does not reflect real-world variability. Therefore, it is advisable to shift to a vine copula structure for high-dimensional settings. Vine copulas allow for flexible, pair-copula constructions that can capture varying levels of dependence between stations at different distances. Additionally, a more robust parameter estimation procedure would be necessary to handle the increased complexity and ensure stable, reliable results.

4.4 Computational and numerical challenges of hybrid hidden Markov models

Parameter initialisation proved to be a computationally delicate task, especially in the covariate models. The introduction of dynamic weather indicators like temperature and

pressure expanded the parameter space, making it the modelling complex with instability. The increased dimensionality, along with parameters for transition probabilities and state-dependent distributions, significantly increased the risk of converging to local minima during optimisation. This complexity was one of the primary reasons why we did not pursue a hybrid model that combines copulas and covariates within the hidden Markov framework. While it is theoretically appealling, in potential to capturing both temporal and spatial dependence to a high degree, its implementation would require robust optimisation procedures, beyond the capabilities of the Nelder-Mead which we used in R.

The optimisation of the copula components produced additional challenges, especially when modelling tail dependence. In the tails of the distribution, numerical underflow and overflow become critical issues, especially during likelihood computation, where values can become extremely small or large. The survival Gumbel copula, for instance, while effective in capturing lower-tail dependence and the upper-tail dependence to a certain degree, is bounded in structure, which limits its flexibility. This complicates optimisation, as gradients become unreliable and the estimates' precision is reduced. To handle such issues, more advanced and robust numerical frameworks would be necessary.

5 Conclusion

In conclusion, our analysis has provided valuable insight into the modelling and simulation of precipitation data using spatial and temporal multivariate models. Through a comprehensive exploration of various statistical characteristics and modelling techniques, we have gained a deeper understanding of the complex dependence structures inherent in precipitation time series data. Our investigation began with a review of conventional practices in precipitation modelling, highlighting the importance of capturing key statistical characteristics such as mean, variance and autocorrelation. We then delved into the development of stochastic precipitation simulators, which are essential tools for replicating the statistical properties of rainfall in a given region. The copula model emerged as particularly promising frameworks for modelling multivariate dependencies, offering flexibility in capturing the correlation structure of precipitation data across multiple sites. Throughout our analysis, we examined the performance of different copula-based models, considering their ability to reproduce empirical features of precipitation data.

From the fitting of bivariate to quadrivariate models, we observed variations in correlation structure across different rainfall states, with the copula providing valuable insights into the varying levels of dependence. The copula-derived mutual information correlation measure offered a valuable approach for understanding the relationships within spatial precipitation processes. By assessing the dependencies between rainfall series from different stations, mutual information provides insight crucial for accurate rainfall predictions: The statistic enables a deeper understanding of the interaction and influence of rainfall occurrence and intensity.

In evaluating model performance, we considered metrics such as AIC and BIC values, as well as visual assessments of simulated and empirical distributions. While each model exhibited strengths and weaknesses, the copula-embedded model stood out as the best fit for yearly and seasonal rainfall series, in both the bivariate and quadrivariate cases, demonstrating superior performance in capturing key statistical properties. Overall, our findings emphasise the importance of incorporating copula-based approaches in precipitation modelling, particularly for capturing the complex dependence structures present in multivariate rainfall data. By leveraging copula constructs, researchers and practitioners can enhance their understanding and modelling of precipitation processes, ultimately contributing to improved water resource management, risk planning, and climate change assessment in precipitation-dependent regions.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00477-025-03037-6.

Author Contributions T.E.C: Writing first draft, Formal analysis, Investigation, and Model fitting with guidance from N.R and C.O. N.R: Conceptualization, Methodology, Writing and editing manuscript. C.O: Methodology, Writing and editing manuscript. I.Y: Methodology, Writing and editing manuscript. All authors reviewed the manuscript.

Funding This work is partly supported by a Vice-Chancellors scholarship from the University of Greenwich for T.E.Chida.

Data availability The daily precipitation dataset used in this study is available through the UK National River Flow Archive. The specific web addresses for each dataset are given below: 1. https://nrfa.ceh.ac.uk/data/station/meanflow/39131. 2. https://nrfa.ceh.ac.uk/data/station/meanflow/39137. 3. https://nrfa.ceh.ac.uk/data/station/meanflow/39145. 4. https://nrfa.ceh.ac.uk/data/station/meanflow/39098
The covariates data can be accessed from, 5. https://www.kaggle.com/datasets/emmanuelfwerr/london-weather-data

Declarations

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless

indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Ailliot P, Thompson C, Thomson P (2009) Space-time modelling of precipitation by using a hidden Markov model and censored gaussian distributions. J R Stat Soc Ser C Appl Stat 58(3):405–426

Ambrosino C, Chandler RE, Todd MC (2014) Rainfall-derived growing season characteristics for agricultural impact assessments in South Africa. Theor Appl Climatol 115:411–426

Asong ZE, Khaliq M, Wheater H (2016) Multisite multivariate modeling of daily precipitation and temperature in the Canadian prairie provinces using generalized linear models. Clim Dyn 47:2901–2921

Banerjee A, Kemter M, Goswami B, Merz B, Kurths J, Marwan N (2023) Spatial coherence patterns of extreme winter precipitation in the us. Theoret Appl Climatol 152(1):385–395

Bárdossy A, Pegram GGS (2009) Copula based multisite model for daily precipitation simulation. Hydrol Earth Syst Sci 13(12):2299–2314

Berry G, Reeder MJ, Jakob C (2011) A global climatology of atmospheric fronts. Geophys Res Lett. https://doi.org/10.1029/2010GL046451

Burton A, Fowler HJ, Kilsby CG, O'Connell PE (2010) A stochastic model for the spatial-temporal simulation of nonhomogeneous rainfall occurrence and amounts. Water Resour Res. https://doi.org/10.1029/2009WR008884

Chandler RE (2020) Multisite, multivariate weather generation based on generalised linear models. Environ Model Softw 134:104867

Charles SP, Bates BC, Hughes JP (1999) A spatiotemporal model for downscaling precipitation occurrence and amounts. J Geophys Res 104(D24):31657–31669

Chen L, Singh VP, Guo S, Zhou J, Zhang J (2015) Copula-based method for multisite monthly and daily streamflow simulation. J Hydrol 528(9):369–384

Frost AJ, Charles SP, Timbal B, Chiew FHS, Mehrotra R, Nguyen KC, Kent DM (2011) A comparison of multi-site daily rainfall downscaling techniques under Australian conditions. J Hydrol 408(1–2):1–18

Gao C, Guan X, Booij MJ, Meng Y, Xu Y-P (2021) A new framework for a multi-site stochastic daily rainfall model: coupling a univariate Markov chain model with a multi-site rainfall event model. J Hydrol 598(9):126478

Härdle WK, Okhrin O, Wang W (2015) Hidden Markov structures for dynamic Copulae. Econ Theory 31(5):981–1015

Hofert M, Kojadinovic I, Mächler M, Yan J (2018) Elements of copula modeling with r. Springer, Berlin

Kojadinovic I (2017) Some copula inference procedures adapted to the presence of ties. Comput Stat Data Anal 112:24–41

Krupskii P, Joe H (2013) Factor copula models for multivariate data. J Multivar Anal 120:85–101

Latham PE, Roudi Y (2009) Mutual information. Scholarpedia 4(1):1658. https://doi.org/10.4249/scholarpedia.1658 (revision #186917)

Lee T (2018) Multisite stochastic simulation of daily precipitation from copula modeling with a gamma marginal distribution. Theor Appl Climatol 132:1089–1098

Majumder R, Mehta A, Neerchal NK (2020) Copula-based correlation structure for multivariate emission distributions in hidden Markov models

Naifar N (2011) Modelling dependence structure with Archimedean copulas and applications to the iTraxx CDS index. J Comput Appl Math 235(8):2459–2466

Oh DH, Patton AJ (2017) Modeling dependence in high dimensions with factor copulas. J Bus Econ Stat 35(1):139–154

Ramesh NI, Onof C (2014) A class of hidden Markov models for regional average rainfall. Hydrol Sci J 59(9):1704–1717

Salvadori G, De Michele C (2004) Frequency analysis via copulas: theoretical aspects and applications to hydrological events. Water Resourc Res. https://doi.org/10.1029/2004WR003133

Wilks DS (1998) Multisite generalization of a daily stochastic precipitation generation model. J Hydrol 210(1–4):178–191

Wilks DS (2009) A gridded multisite weather generator and synchronization to observed weather data. Water Resour Res. https://doi.org/10.1029/2009WR007902

Yang C, Chandler R, Isham V, Wheater H (2005) Spatial-temporal rainfall simulation using generalized linear models. Water Resour Res. https://doi.org/10.1029/2004WR003739

Zeng X, Durrani T (2011) Estimation of mutual information using copula density function. Electron Lett 47(8):493–494

Zucchini W, MacDonald IL (2009) Hidden Markov models for time series: an introduction using r. Chapman and Hall/CRC, Boca Raton

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.