Dependence Geometry and Simulation Across Precipitation Fields

Tapiwa Edward Chida[1*], Nadarajah Ramesh[2] *and* Christian Onof [3]

[1,2]School of Computing and Mathematical Sciences, Queen Mary Building, University of Greenwich, Old Royal Naval College, Park Row, London, SE10 9LS, United Kingdom.

[3]Department of Civil and Environmental Engineering, Skempton Building, Imperial College London, Exhibition Road, London, SW7 2AZ, United Kingdom.

*Corresponding author(s). E-mail(s): T.E.Chida@greenwich.ac.uk ;
Contributing authors: n.i.ramesh@greenwich.ac.uk; c.onof@imperial.ac.uk;

Discussion and Further Diagnostics

Latent State Decoding

In order to analyse the state processes and gauge the models' performance against the empirical values we carry out a latent state decoding process. We borrow methods employed in speech processing from Zucchini and MacDonald (2009), starting with marginal posterior decoding. This is a method used to determine the most probable state for each day. The authors note the issue of numerical underflow so we calculated the forward and backward probabilities in the log space. The forward and backward recursions ($\alpha_t(j)$ and $\beta_t(j)$) are given by;

$$\log \alpha_t(k) = \log\big((\psi_t)_{kk}\big) + \log\left(\sum_{j=1}^{3} \alpha_{t-1}(j)\,\phi_{jk}\right) \qquad \log \beta_t(k) = \log\left(\sum_{j=1}^{3} \phi_{kj}\,(\psi_{t+1})_{jj}\,\beta_{t+1}(j)\right), \qquad k \in \{1,2,3\} \quad (1)$$

where ϕ and psi are as previously defined while $j, k \in \{1, 2, 3\}$, are hidden state indicies. Posteriors probabilities are then calculated as

$$\gamma_t(i) = \Pr\Big(Z_t = i \mid (y_1^{(1)}, \ldots, y_1^{(d)}), \ldots, (y_n^{(1)}, \ldots, y_n^{(d)})\Big) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{3} \alpha_t(j)\beta_t(j)} \quad (2)$$

On each day we choose the state with the optimal posterior mass, such that the decoded path at time n, the last day of the time series, is given by,

$$Z_{1:n}^{\text{post}} = \Big(\arg\max_i \gamma_1(i), \arg\max_i \gamma_2(i), \ldots, \arg\max_i \gamma_n(i)\Big). \quad (3)$$

For comparison, we also employ a Viterbi decoder, again borrowing from Zucchini and MacDonald (2009). Unlike the posterior counterpart, which focuses on the optimal daily posterior probability, the Viterbi algorithm is a global decoder which finds the most likely sequence of states across the whole time period. This is a dynamic maximisation problem for the probaility of various paths of length n in the log space,

$$\log P\Big(Z_{1:n}, (y_1^{(1)}, \ldots, y_1^{(d)}), \ldots, (y_n^{(1)}, \ldots, y_n^{(d)})\Big) = \log \pi_{z_1} + \log\big((\psi_1)_{z_1 z_1}\big) + \sum_{t=2}^{n} \Big[\log \phi_{z_{t-1} z_t} + \log\big((\psi_t)_{z_t z_t}\big)\Big] \quad (4)$$

where $Z_{1:n} = (z_1, \ldots, z_n)$ is a candidate hidden-state sequence, and $P\Big(z_{1:n}, (y_1^{(1)}, \ldots, y_1^{(d)}), \ldots, (y_n^{(1)}, \ldots, y_n^{(d)})\Big)$ is the joint probability of that state path with the observed data. The logarithm is taken so that products become sums, making the computation more numerically stable. The term π_{z_1} is the initial probability of starting in state z_1, taken

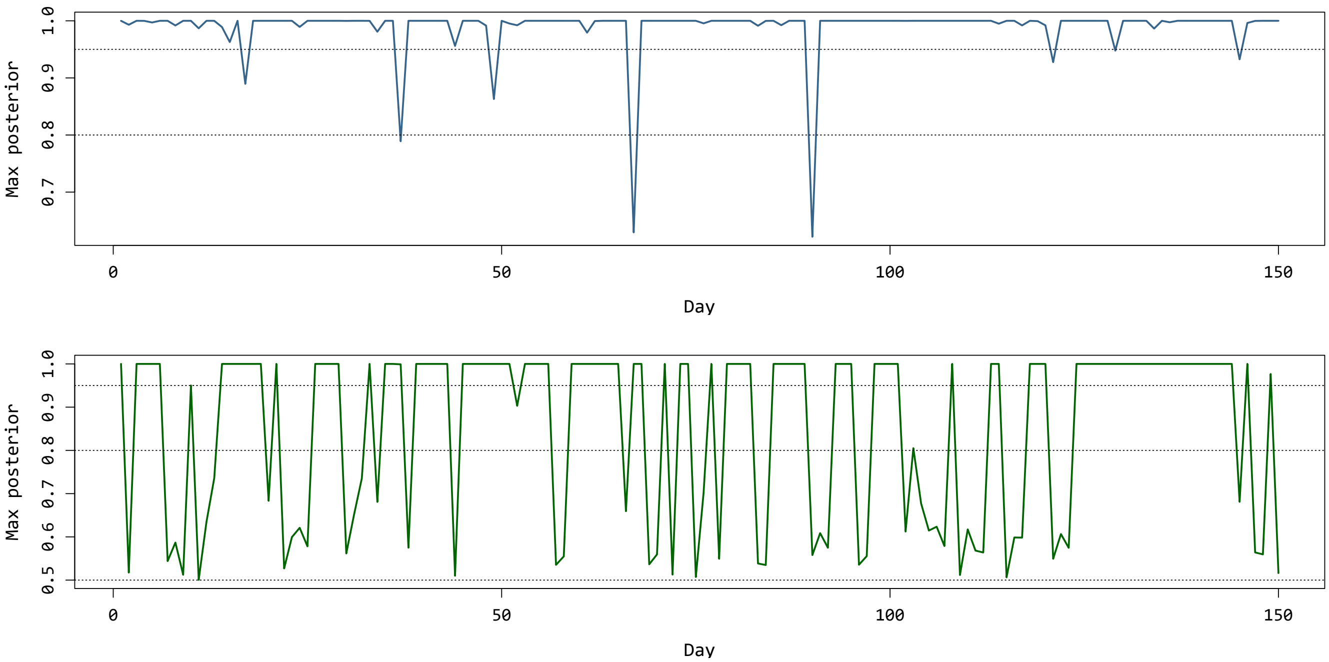

Fig. 1: Maximum posterior probabilities for the first 150 days using parameters from the EAC-HMM and CVC-HMM

from the stationary distribution $\pi = [\pi_1\ \pi_2\ \pi_3]$, $\phi_{z_{t-1}z_t}$ is the transition probability from state z_{t-1} to state z_t, and $(\psi_t)_{z_t z_t}$ is the diagonal entry of ψ_t corresponding to state z_t, representing the state-dependent contribution of the observation at time t. The probability of the best partial path ending in state k at time t is then

$$\delta_t(k) = \max_{z_1,\ldots,z_{t-1}} \log P\Big(Z_{1:t}, (y_1^{(1)}, \ldots, y_1^{(d)}), \ldots, (y_t^{(1)}, \ldots, y_t^{(d)})\Big)\,. \tag{5}$$

with the Viterbi recursion given by,

$$\delta_t(k) = \log\big((\psi_t)_{kk}\big) + \max_{j=1,\ldots,3}\left[\delta_{t-1}(j) + \log \phi_{jk}\right]. \tag{6}$$

The parameters from all fitted models were used to decode the state sequence, and the performance was tested against the state sequence generated from the simulation process. The average performance of each decoder for all models in the two domains are shown in Table 1. The posterior decoder comes out on top in most of the metrics.

Scale	Decoder	Accuracy	Prop. abs. error	Dwell abs. error	Recall_1	Recall_2	Recall_3
Local	Posterior	0.9234225	0.01417842	0.2447059	1.0000000	0.8068032	0.8508829
	Viterbi	0.9175265	0.02119745	0.4076231	1.0000000	0.7892090	0.8413933
Regional	Posterior	0.9542673	0.01588835	0.2273492	1.0000000	0.8965127	0.9080218
	Viterbi	0.9512077	0.01556629	0.2554676	1.0000000	0.8877081	0.9030308

Table 1: Decoding performance tested against state sequence of the simulations.

The proportional error or the mismatch in the number of switches compared to the reference path were lower. The Viterbi dwelled in the wrong states more often than required at both spatial scales. The posterior decoder correctly identified state two and three observations 80.68% and 85.09% of the time whilst for the viterbi, 78.92% and 84.14% of the time. Moving on with the superior decoder, Figure 1, we plot $\max_k \gamma_t(k)$, showing us how certain each decoder is based on the hidden Markov parameters from the two models. From the figure, the EAC-HMM state sequence is more defined. The states are clearly discerned from the data. The CVC-HMM fit points toward a suboptimal state separation. This can be attributed to the difference in the magnitudes of the parameter space; hence optimisation of the CVC-HMM involves more noise. Nested copulas, although not as flexible and highly restrictive, can be simplistic alternative. It is important to note that the plots to prove goodness-of-fit but rather how decisive the decode is under each model. Figure 2 then shows the empirical time series plot of the first 150 days at Michelstadt and the decoded hidden state dynamics from the CVC-HMM. Note that, since the CVC-HMM fit was on the regional scale, the spatial

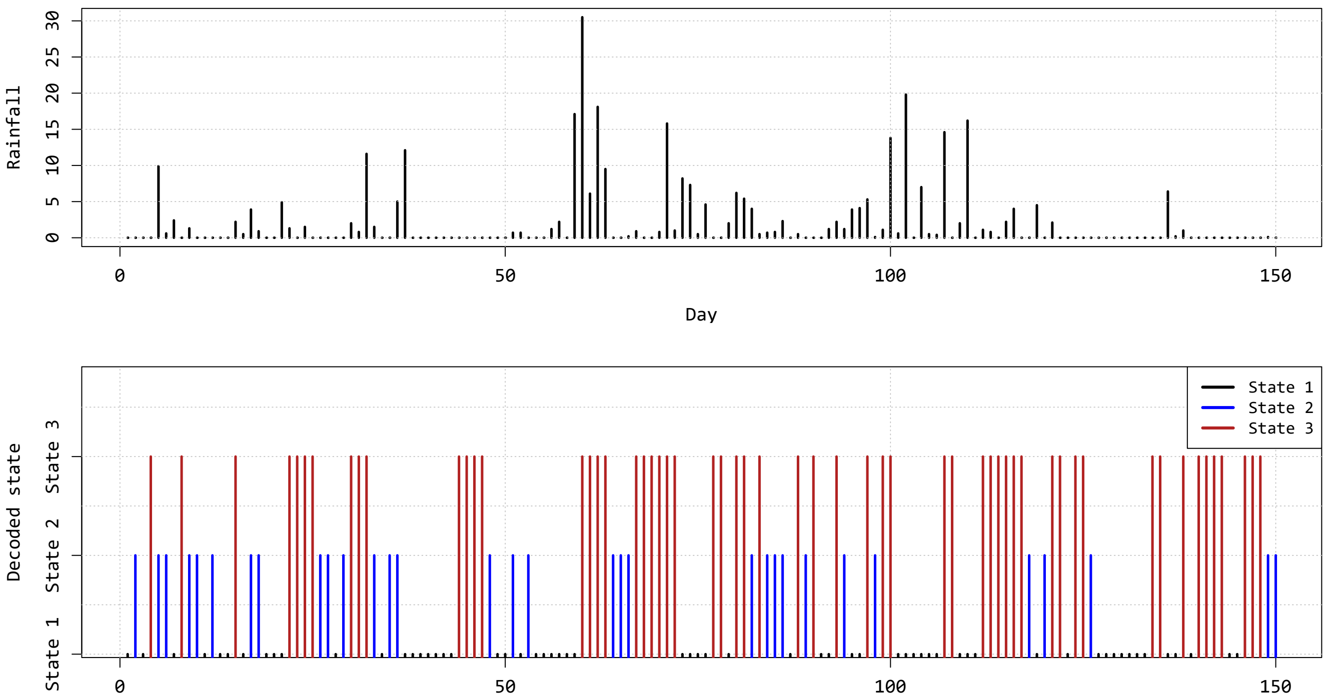

Fig. 2: Rollbach time series and its CVC-HMM posterior decoding for the first 150 days

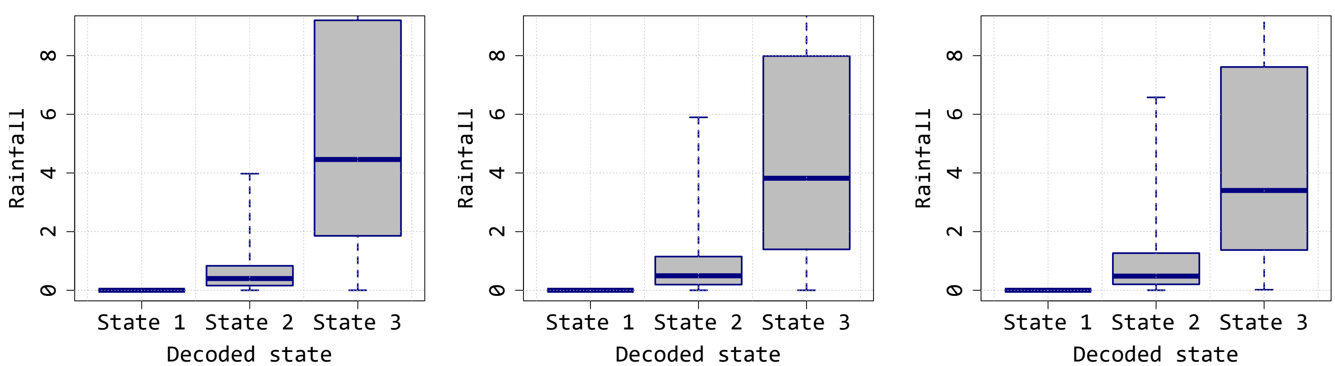

Fig. 3: EAC, PCC, CVC state composition from the decoding

heterogeneity leads to highly distinct rainfall readings. As a result, the decoded state shows the joint rainfall pattern rather than rainfall recordings at a one station. This is why on some days when there is very low rainfall the decoder is in state three, or high rainfall in state one and two; some of the times the decoder is just inaccurate. We get an approximation of the rainfall pattern in each state in Figure 3. The outline verifies the observation from the certainty plots (Figure 1), with the EAC-HMM states more defined. However, interquartile ranges for the states do not overlap across the other models, indicating that the state separations are interpretable.

Remarks

Numerical and Algorithmic Issues in Vine–Copula Hidden Markov Models

Fitting the vine copula HMMs brought about several numerical and computational challenges, especially for the covariate driven stochastic model. The first challenge is choosing the vine structure. A d-dimensional C-vine can have $(d-1)!$ possible root orderings, and the number of possible R-vine configurations grows even faster. In this study, a greedy algorithm based on the highest absolute Kendall's τ was used, which finds a locally optimal, but not globally optimal structure, introducing selection bias. Even with a fixed tree structure, the parameter space expands quickly and the log-likelihood surface becomes irregular as dependencies among pair copulas interact through the h-functions.

The Nelder-Mead optimiser regularly stopped in local maxima, especially for copulas components exhibiting weak tail dependence in the vine structure.

A second difficulty involves numerical instability in the distribution tails. For rotated Archimedean copulas, h-function values near $u, v = 0$ or 1 can overflow or underflow. This was mitigated using log densities though rounding errors still occasionally affected convergence. Employing automatic differentiation or higher-precision arithmetic could improve numerical stability in future applications.

A third challenge is weak parameter values, especially for conditional copulas in the vine structure, where dependence is small. This resulted in large uncertainty and occasional overfitting. Truncating the vine may help regularise estimation and reduce variance in such cases.

Finally, the computation time increases rapidly with sample size, as each likelihood evaluation involves many pair-copula calculations. This makes vine-copula HMMs computationally demanding, especially in high dimensions or long time series. Implementing more efficient, gradient-based optimisation algorithms could substantially reduce computation time and improve scalability. Overall, while these models effectively capture complex dependence structures, they remain limited by structural complexity, numerical instability and high computational cost.

References

Zucchini, W., & MacDonald, I.L. (2009). *Hidden markov models for time series: an introduction using r.* Chapman and Hall/CRC.

Funding

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

Contributions

T.E.Chida: Writing first draft, Formal analysis, Investigation, and Model fitting with guidance from N.Ramesh and C.Onof.
N.Ramesh: Conceptualisation, Methodology, Writing and editing manuscript.
C.Onof: Methodology, Writing and editing manuscript.
All authors read and approved the final manuscript.

Data Availability Statement

The daily precipitation datasets used in this study is available through the National Centre for Environmental Information. The web address is;

- https://www.ncdc.noaa.gov/cdo-web/datasets

The covariates data can be accessed from,

- https://cds.climate.copernicus.eu/

Competing interest

The authors have no relevant financial or non-financial interests to disclose.