Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A stochastic compartment model for the number of grid cells under heavy rain, with regime switching and Dirichlet-multinomial co-movement

Tirupathi Rao Padi
Pondicherry University
Muhammed Rais P K
pkmuhammedrais55@gmail.com

Pondicherry University

Research Article

Keywords:

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11608812/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A stochastic compartment model for the number of grid cells under heavy rain, with regime switching and Dirichlet-multinomial co-movement

Tirupathi Rao Padi[1] and Muhammed Rais P K[1*]

[1]Department of Statistics, Pondicherry University, Puducherry, 605014, India.

*Corresponding author(s). E-mail(s): pkmuhammedrais55@gmail.com;
Contributing authors: drtrpadi@gmail.com;

Abstract

On a day of heavy rain a regional relief plan needs the number of places affected, not the state of one station. We model the daily rainfall state of the 57 land cells of the India Meteorological Department (IMD) **0.25°** grid around Puducherry, on the Coromandel coast of India, by IMD categories Below Normal ($<$ **2.5** mm), Normal and Above Normal Rain ($\geq$ **64.5** mm), as a closed three-compartment stochastic system. With linear transfer rates the system is solved exactly: difference–differential equations with boundary cases, the generating-function equation and its product-of-multinomials solution, transition functions, the stationary multinomial law, moments, cumulants and a discrete-time twin. The record of 1991–2020 rejects the independence of cells on which this rests: in the North-East monsoon the variance of the daily number of Above Normal Rain cells is 18.6 times the independent-cell value. We prove that heterogeneity between independent cells can only lower this variance, and relax independence in two ways that keep the law exact: a hidden regime that switches the day's transition matrix, and a Dirichlet-multinomial emission with one concentration parameter per regime. Fitted freely, the emission misstates the mean level, so we derive the exact mean recursion and refit under a mean constraint tested by parametric bootstrap. In Monte Carlo checks the constrained models reproduce the mean and variance of the daily count in every season, and in the North-East monsoon also the probability that five or more cells receive Above Normal Rain (observed 0.056, model 0.061, independent cells 0.004).

Keywords: Probability generating function, Overdispersion, Hidden Markov model, Constrained maximum likelihood, Northeast monsoon, Coromandel coast

MSC Classification: 60J27 , 60J10 , 62M05 , 62P12

1 Introduction

A station record says how often heavy rain falls at one place. A district officer planning relief asks a different question first: on a day of heavy rain, how much of the region is under it? A cyclonic disturbance crossing the coast wets many places at once, and the burden on a relief plan is set by the number of places affected, not by the state of one gauge. On the Coromandel coast of south-east India this question is pressing. The North-East monsoon (October–December) brings most of the annual rain of the coast, its heavy rainfall events are tied to easterly waves and depressions from the Bay of Bengal (Sanap et al. 2019), the coast is among the most cyclone-exposed in India (Saravanan et al. 2023), and widespread extreme rainfall over south peninsular India has become more frequent in recent decades (Shahi and Rai 2023). This paper changes the unit of account accordingly. The population is not the days of a record but the grid cells of a region, and the quantity modelled is the number of cells in each rainfall state on a given day.

A natural tool for a population of units moving among states is the stochastic compartment model. For a closed linear system, Matis and Hartley (1971) showed that the contents of the compartments have a law that is a convolution of multinomial terms, one for each compartment of origin; the same product form appears for networks of infinite-server queues (Massey and Whitt 1993), and compartment models with births and interaction have been developed for biological populations (Becker 1970; Matis et al. 1995; Matis and Kiffe 2000). In rainfall modelling, by contrast, multi-site models describe the state of each site. Zucchini and Guttorp (1991) and Hughes et al.

(1999) let a hidden weather state govern rainfall occurrence at a network of stations, and the latter's non-homogeneous form has been used for daily rainfall in north-east Brazil (Robertson et al. 2004) and for Indian monsoon rainfall (Greene et al. 2011; Pal et al. 2015); Thompson et al. (2007) fitted a multi-site model with dry, light-rain and heavy-rain states to New Zealand data, and Heaps et al. (2015) combined hidden weather states with latent Gaussian variables. Such models must be simulated to say anything about the *number* of wet sites. A further fact is the overdispersion of rainfall counts: persistence and weather regimes inflate the variance of a count above the value for independent trials (Katz and Parlange 1998). Over space the effect is larger still, because one weather system covers many cells.

Three gaps follow. First, the exact solution of a closed compartment system is known, but we found no application of it to the number of places in a rainfall category, and so no exact distribution for that count. Second, the multi-site rainfall models we found do not give the law of the number of wet or heavy-rain sites in closed form, with its moments, cumulants and tail probabilities. Third, a compartment model of identical cells assumes that, given the rates, the cells move independently; the gridded record around Puducherry rejects this by a wide margin, and we found no compartment model of a weather population in which dependence between members is admitted and the law is kept exact.

The contributions are as follows. Section 2 describes the grid and the data. Section 3 formulates the daily rainfall state of the grid cells as a closed three-compartment system under Poisson postulates and derives its complete law: the difference–differential equations with their boundary cases, the generating-function equation and its solution, the transition functions, the mass function, the moments, cumulants and shape measures, the characteristics of a cell and a discrete-time twin on the daily scale. Section 4 relaxes independence in two ways that keep the law exact, a hidden regime that switches the transition matrix of the day and a Dirichlet-multinomial emission;

proves that heterogeneity between cells cannot explain overdispersion; and derives the exact mean of the counts under the most general model, with the reason why an overdispersed maximum-likelihood fit can misstate it. Section 5 gives estimation, with standard errors that do not assume independent cells, a likelihood fit under a constraint on the mean level and a Monte Carlo validation. Section 6 applies the models to the four seasons of 1991–2020 with hold-out years 2021–2023, Section 7 reports how the formulas were verified, and Section 8 concludes. Each result is labelled as exact (a closed form, checked symbolically and in exact arithmetic), asymptotic, or simulation-based.

2 Study area and data

The data are the IMD $0.25^\circ \times 0.25^\circ$ gridded daily rainfall (Pai et al. 2014) over the block 11.0–13.0°N, 78.5–80.5°E, which covers the Puducherry region of the Union Territory of Puducherry and the adjoining districts of Tamil Nadu (Fig. 1). The block has 81 grid points; the 24 that fall in the Bay of Bengal carry no value. The remaining 57 land cells are the population of this paper, $N = 57$, and they are observed on every one of the 19,358 days from 1 January 1971 to 31 December 2023: there is no missing cell-day. Each cell-day is classified by the 24-hour rainfall categories of the IMD (Mohapatra et al. 2023): with y the rainfall of the cell on the day, its state is

$$Z = \begin{cases} S_1, \text{ Below Normal Rain (B)}, & y < 2.5 \text{ mm}, \\ S_2, \text{ Normal Rain (N)}, & 2.5 \le y < 64.5 \text{ mm}, \\ S_3, \text{ Above Normal Rain (A)}, & y \ge 64.5 \text{ mm}, \end{cases} \tag{1}$$

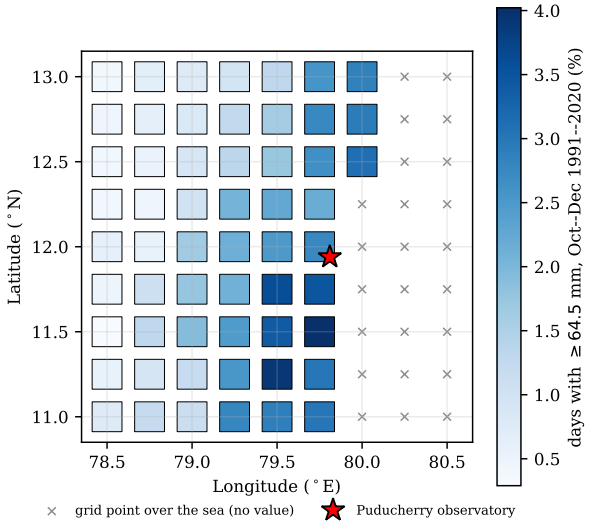

Fig. 1 The block of the IMD 0.25° grid used in this paper. Squares: the 57 land cells, shaded by the share of NE monsoon days of 1991–2020 on which the cell is in state A ($\geq$ 64.5 mm). Crosses: grid points over the Bay of Bengal, which carry no value. Star: the IMD observatory at Puducherry

State A is rainfall at or above the IMD heavy-rain threshold. The year is divided into four seasons, Winter (January–February), Pre-monsoon (March–May), South-West (SW) monsoon (June–September) and North-East (NE) monsoon (October–December), and every model is fitted separately for each season. Models are fitted on 1991–2020 and validated on the hold-out years 2021–2023; the grid ends in 2023. A pair of consecutive days is used only when both days lie in the same season of the same year.

3 The closed compartment model

3.1 Notation and postulates

Write $n_j(t)$ for the number of cells in S_j at time t and $\mathbf{n}(t) = (n_1(t), n_2(t), n_3(t))$ for the contents of the system. The population is closed, so

$$n_1(t) + n_2(t) + n_3(t) = N \qquad \text{for every } t, \tag{2}$$

and the state space is the lattice simplex $\mathcal{N} = \{(n_1, n_2, n_3) \in \mathbb{Z}^3_{\geq 0} : n_1 + n_2 + n_3 = N\}$, which has $\binom{N+2}{2}$ points. The system obeys four postulates.

1. *Transfer.* In $(t, t + \Delta t)$ one cell moves from S_i to S_j, $j \neq i$, with probability

$$n_i \lambda_{ij}\, \Delta t + o(\Delta t), \qquad \lambda_{ij} > 0, \tag{3}$$

   where λ_{ij} is the transfer rate per cell from S_i to S_j.
2. *Simplicity.* The probability of two or more transfers in $(t, t + \Delta t)$ is $o(\Delta t)$.
3. *Independence of intervals.* What happens in disjoint intervals is independent.
4. *Closure.* There is neither immigration into nor absorption out of the system, so (2) holds.

Because (3) is linear in the contents, the cells do not interact: each is a three-state Markov chain in continuous time with generator

$$\mathbf{Q} = \begin{pmatrix} -(\lambda_{12} + \lambda_{13}) & \lambda_{12} & \lambda_{13} \\ \lambda_{21} & -(\lambda_{21} + \lambda_{23}) & \lambda_{23} \\ \lambda_{31} & \lambda_{32} & -(\lambda_{31} + \lambda_{32}) \end{pmatrix}, \tag{4}$$

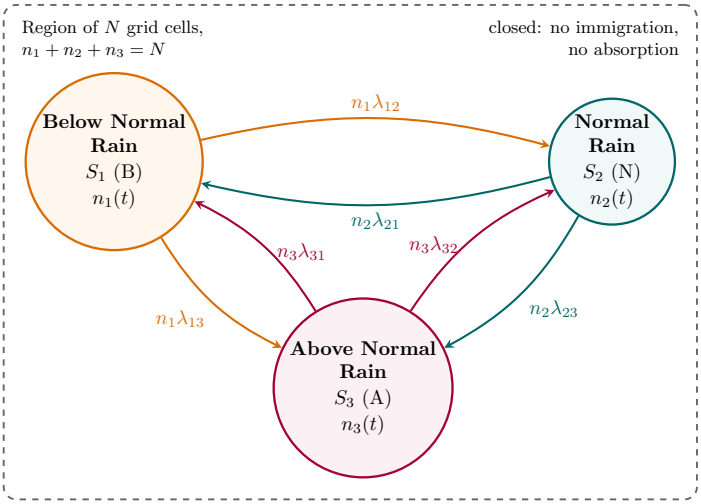

Fig. 2 Schematic diagram of the closed three-compartment model. The circles are the compartments B (S_1), N (S_2) and A (S_3), holding $n_1(t)$, $n_2(t)$ and $n_3(t)$ of the N cells of the region. An arrow from S_i to S_j carries the transfer rate $n_i\lambda_{ij}$ of (3). The dashed box is the closed region: the three contents always add to N, and no arrow enters or leaves it

whose rows sum to zero. Write $p_{ij}(t)$ for the probability that a cell is in S_j at t given that it was in S_i at 0, so that $\mathbf{p}(t) = (p_{ij}(t)) = e^{\mathbf{Q}t}$, and let $\mathbf{n}(0)$ be the known initial contents. The schematic diagram of the system is Fig. 2: every arrow stays inside the region, so the system has no source and no sink.

3.2 Difference–differential equations

Let $P(n_1, n_2, n_3; t) = P(\mathbf{n}(t) = (n_1, n_2, n_3))$. Collecting the ways the system can arrive in (n_1, n_2, n_3) during $(t, t + \Delta t)$, dividing by Δt and letting $\Delta t \to 0$ gives, at every interior point of $\mathcal{N}$,

$$\frac{\partial}{\partial t}P(n_1, n_2, n_3; t) = -\Big[\sum_{i=1}^{3} n_i \sum_{j\neq i} \lambda_{ij}\Big] P(n_1, n_2, n_3; t) + \sum_{i=1}^{3}\sum_{j\neq i} (n_i + 1)\lambda_{ij}\, P(\ldots, n_i + 1, \ldots, n_j - 1, \ldots; t), \quad (5)$$

where a term of the second sum is present only when the state it refers to lies in $\mathcal{N}$, that is when $n_j \geq 1$. At the three vertices of the simplex only two entry terms survive:

$$\begin{aligned}
\tfrac{\partial}{\partial t}P(N,0,0;t) &= -N(\lambda_{12}+\lambda_{13})P(N,0,0;t) \\
&\quad + \lambda_{21}P(N-1,1,0;t) + \lambda_{31}P(N-1,0,1;t), \qquad (6)\\
\tfrac{\partial}{\partial t}P(0,N,0;t) &= -N(\lambda_{21}+\lambda_{23})P(0,N,0;t) \\
&\quad + \lambda_{12}P(1,N-1,0;t) + \lambda_{32}P(0,N-1,1;t), \qquad (7)\\
\tfrac{\partial}{\partial t}P(0,0,N;t) &= -N(\lambda_{31}+\lambda_{32})P(0,0,N;t) \\
&\quad + \lambda_{13}P(1,0,N-1;t) + \lambda_{23}P(0,1,N-1;t). \qquad (8)
\end{aligned}$$

On an edge of the simplex, where exactly one n_j is zero, four entry terms survive; for example, with $N = 3$,

$$\begin{aligned}
\tfrac{\partial}{\partial t}P(2,1,0;t) &= -\big[2(\lambda_{12}+\lambda_{13}) + (\lambda_{21}+\lambda_{23})\big]P(2,1,0;t) \\
&\quad + 3\lambda_{12}P(3,0,0;t) + 2\lambda_{21}P(1,2,0;t) + \lambda_{31}P(1,1,1;t) + \lambda_{32}P(2,0,1;t). \qquad (9)
\end{aligned}$$

Summing the right-hand sides of (5)–(9) over $\mathcal{N}$ gives zero, so total probability is conserved.

3.3 The generating-function equation and its solution

Let

$$G(x_1,x_2,x_3;t) = \sum_{(n_1,n_2,n_3)\in\mathcal{N}} P(n_1,n_2,n_3;t)\, x_1^{n_1}x_2^{n_2}x_3^{n_3} \qquad (10)$$

be the probability generating function (PGF) of the contents. Multiplying (5) by $x_1^{n_1}x_2^{n_2}x_3^{n_3}$ and summing over $\mathcal{N}$ turns the whole system, boundary cases included,

into one first-order partial differential equation,

$$\frac{\partial G}{\partial t} = \sum_{i=1}^{3} \Big[\sum_{j \neq i} \lambda_{ij} \, (x_j - x_i) \Big] \frac{\partial G}{\partial x_i}, \qquad G(x_1, x_2, x_3; 0) = \prod_{i=1}^{3} x_i^{\, n_i(0)}. \tag{11}$$

The bracket is the drift of x_i: a cell leaving S_i for S_j exchanges one power of x_i for one power of x_j.

Proposition 1 (Matis and Hartley 1971) *The solution of* (11) *is*

$$G(x_1, x_2, x_3; t) = \prod_{i=1}^{3} \Big(\sum_{j=1}^{3} p_{ij}(t) \, x_j \Big)^{n_i(0)}, \qquad \mathbf{p}(t) = e^{\mathbf{Q}t}. \tag{12}$$

Proof The auxiliary equations of Lagrange's method are $\mathrm{d}x_i/\mathrm{d}t = -\sum_{j \neq i} \lambda_{ij}(x_j - x_i)$, that is $\mathrm{d}\mathbf{x}/\mathrm{d}t = -\mathbf{Q}\mathbf{x}$ for the column $\mathbf{x} = (x_1, x_2, x_3)^{\top}$, whose first integrals are the components of $e^{\mathbf{Q}t}\mathbf{x}$. Hence G is a function of $e^{\mathbf{Q}t}\mathbf{x}$, and the initial condition fixes it as $\prod_i \big((e^{\mathbf{Q}t}\mathbf{x})_i\big)^{n_i(0)}$, which is (12). Directly: the $n_i(0)$ cells that start in S_i are independent and each lands in S_j with probability $p_{ij}(t)$, and the PGF of a multinomial vector with $n_i(0)$ trials is the $n_i(0)$th power of $\sum_j p_{ij}(t) x_j$. □

The same product form appears for networks of infinite-server queues (Massey and Whitt 1993).

3.4 Transition functions, stationary law and relaxation time

Because $\mathbf{Q}$ has rank two, one eigenvalue is zero and the other two are the roots of

$$\rho^2 + s_1 \rho + s_2 = 0, \qquad s_1 = -\operatorname{tr} \mathbf{Q}, \qquad s_2 = \sum_{i<j} \det \mathbf{Q}\big[\{i,j\},\{i,j\}\big], \tag{13}$$

since $\det(\mathbf{Q} - \rho \mathbf{I}) = -\rho(\rho^2 + s_1 \rho + s_2)$; here $\mathbf{Q}[\{i,j\},\{i,j\}]$ is the principal 2×2 submatrix. Call the roots ρ_1, ρ_2; both have negative real part, so the system relaxes.

For distinct eigenvalues Sylvester's formula gives

$$\mathbf{p}(t) = e^{\mathbf{Q}t} = \mathbf{1}\boldsymbol{\pi}^* + \sum_{k=1}^{2} e^{\rho_k t} \frac{\prod_{m\neq k,\, m\in\{0,1,2\}} (\mathbf{Q} - \rho_m \mathbf{I})}{\prod_{m\neq k,\, m\in\{0,1,2\}} (\rho_k - \rho_m)}, \qquad \rho_0 = 0, \tag{14}$$

where $\boldsymbol{\pi}^*$ is the stationary law of one cell, the probability vector with $\boldsymbol{\pi}^*\mathbf{Q} = \mathbf{0}$, and $\mathbf{1}$ is the column of ones. As $t \to \infty$ every row of $\mathbf{p}(t)$ tends to $\boldsymbol{\pi}^*$ and (12) becomes

$$G(x_1, x_2, x_3; \infty) = \Big(\sum_{j=1}^{3} \pi_j^* x_j \Big)^N, \tag{15}$$

the PGF of $\mathrm{Mult}(N, \boldsymbol{\pi}^*)$: in the long run the cells are distributed over the three states as N independent draws from $\boldsymbol{\pi}^*$, whatever the start. The slowest term of (14) decays like $e^{-|\mathrm{Re}\,\rho_2| t}$, with ρ_2 the eigenvalue nearer zero, so the relaxation time

$$t_{\mathrm{relax}} = 1/|\mathrm{Re}\,\rho_2| \tag{16}$$

measures how long the system remembers its start.

3.5 Mass function, moments and shape

Corollary 2 *The contents are the sum of three independent multinomial vectors:*

$$P(n_1, n_2, n_3; t) = \sum_{\substack{\mathbf{m}_1+\mathbf{m}_2+\mathbf{m}_3=(n_1,n_2,n_3) \\ \mathbf{m}_i\cdot\mathbf{1}=n_i(0)}} \prod_{i=1}^{3} \binom{n_i(0)}{m_{i1}, m_{i2}, m_{i3}} \prod_{j=1}^{3} p_{ij}(t)^{m_{ij}}, \tag{17}$$

where m_{ij} is the number of cells that start in S_i and are in S_j at time t.

Proof Expand each factor of (12) by the multinomial theorem and collect the coefficient of $x_1^{n_1} x_2^{n_2} x_3^{n_3}$. □

Three special cases are useful. If all cells start in S_i, (17) is the single multinomial $\mathrm{Mult}(N, (p_{i1}(t), p_{i2}(t), p_{i3}(t)))$. If $\mathbf{n}(0) \sim \mathrm{Mult}(N, \boldsymbol{\pi}^*)$, the contents are $\mathrm{Mult}(N, \boldsymbol{\pi}^*)$

for every t. And one coordinate alone is a sum of independent binomials, $n_j(t) \sim \sum_{i=1}^{3} \mathrm{Bin}(n_i(0), p_{ij}(t))$. Differentiating (12) at $x_1 = x_2 = x_3 = 1$ gives

$$E[n_j(t)] = \sum_{i=1}^{3} n_i(0)\, p_{ij}(t), \tag{18}$$

$$\mathrm{Var}[n_j(t)] = \sum_{i=1}^{3} n_i(0)\, p_{ij}(t)\big(1 - p_{ij}(t)\big), \tag{19}$$

$$\mathrm{Cov}[n_j(t), n_k(t)] = -\sum_{i=1}^{3} n_i(0)\, p_{ij}(t)\, p_{ik}(t), \qquad j \neq k. \tag{20}$$

The means add to N and the covariance matrix is singular, as (2) requires; the covariances are negative, because within one day a cell counted in A is a cell not counted in B, and that is the only dependence the model admits. Putting $x_j = e^{\theta_j}$ in (12) gives the moment generating function (MGF) and the cumulant generating function (CGF),

$$M(\boldsymbol{\theta}; t) = \prod_{i=1}^{3} \Big(\sum_{j=1}^{3} p_{ij}(t) e^{\theta_j} \Big)^{n_i(0)}, \qquad K(\boldsymbol{\theta}; t) = \sum_{i=1}^{3} n_i(0) \log \sum_{j=1}^{3} p_{ij}(t) e^{\theta_j}, \tag{21}$$

and $x_j = e^{\mathrm{i}\theta_j}$ the characteristic function $\phi(\boldsymbol{\theta}; t) = M(\mathrm{i}\boldsymbol{\theta}; t)$. For one count, with $p_i = p_{ij}(t)$, the MGF is $M_{n_j}(u; t) = \prod_i (1 - p_i + p_i e^u)^{n_i(0)}$ and the first four cumulants are

$$\kappa_1 = \sum_i n_i(0) p_i, \qquad \kappa_2 = \sum_i n_i(0) p_i (1 - p_i), \tag{22}$$

$$\kappa_3 = \sum_i n_i(0) p_i (1 - p_i)(1 - 2p_i), \quad \kappa_4 = \sum_i n_i(0) p_i (1 - p_i)\big(1 - 6p_i + 6p_i^2\big), \tag{23}$$

with coefficient of variation, skewness and excess kurtosis

$$\mathrm{CV} = \sqrt{\kappa_2}/\kappa_1, \qquad \gamma_1 = \kappa_3/\kappa_2^{3/2}, \qquad \gamma_2 = \kappa_4/\kappa_2^2. \tag{24}$$

Table 1 Characteristics of the closed three-compartment system, $\mathbf{Q} = (q_{ij})$. All results exact

Characteristic	Expression	Reading
Mean sojourn in S_i	$1/(-q_{ii}) = 1/\sum_{j\neq i}\lambda_{ij}$	how long a cell stays in a state
Next state from S_i	$q_{ij}/(-q_{ii})$, $j \neq i$	where it goes when it leaves
Stationary law of a cell	$\boldsymbol{\pi}^*\mathbf{Q} = \mathbf{0}$, $\boldsymbol{\pi}^*\mathbf{1} = 1$	long-run share of time in each state
Stationary law of the population	$\mathrm{Mult}(N, \boldsymbol{\pi}^*)$	long-run law of the three counts
Stationary mean and variance of n_j	$N\pi_j^*$, $N\pi_j^*(1-\pi_j^*)$	the benchmark the data must beat
Relaxation time	$1/\lvert\mathrm{Re}\,\rho_2\rvert$	how long the start is remembered

For the count of cells in A, where every p_i is small, $\kappa_3 > 0$: the count is right-skewed, as a count of a rare event should be.

3.6 Characteristics of a cell and the discrete-time twin

Because the cells are independent, one cell describes the mechanism and the population multiplies it. Table 1 lists the characteristics; each follows from (4).

The record is daily. The discrete-time twin replaces (3) by a one-day stochastic matrix $\mathbf{P} = (P_{ij})$, with P_{ij} the probability that a cell in S_i today is in S_j tomorrow, so that

$$G(x_1, x_2, x_3; t) = \prod_{i=1}^{3} \Big(\sum_{j=1}^{3} \big(\mathbf{P}^t\big)_{ij} x_j \Big)^{n_i(0)}, \qquad t = 0, 1, 2, \ldots \tag{25}$$

Everything above carries over with $p_{ij}(t)$ replaced by $(\mathbf{P}^t)_{ij}$; the mean sojourn in S_i becomes $1/(1-P_{ii})$ days and the relaxation time $-1/\log|\mu_2|$, with μ_2 the second-largest eigenvalue of $\mathbf{P}$ in modulus. The two versions are linked by $\mathbf{P} = e^{\mathbf{Q}}$, so a fitted $\mathbf{P}$ gives continuous rates by $\mathbf{Q} = \log \mathbf{P}$ whenever that logarithm is a proper generator. The twin is the version fitted to the data.

4 Relaxing the independence of cells

Given the rates, the model of Section 3 makes the cells independent, and then (19) fixes the variance of a count once its mean is fixed. Section 6 shows that the record refuses this. This section gives two relaxations that keep the law exact (Fig. 3), proves

that heterogeneity between cells is not an alternative, and derives the mean level under the most general model.

4.1 A hidden regime that switches the matrix

Let $R_t \in \{1, \dots, M\}$ be a Markov chain with transition matrix $\mathbf{A} = (a_{rr'})$ and initial law $\boldsymbol{\psi}$, and let the one-day matrix of day t be $\mathbf{P}_{R_t} = (p_{ij}^{(R_t)})$. Given the regime path the cells are independent, so every result of Section 3 holds conditionally:

$$G(x_1, x_2, x_3; t \mid R_1, \dots, R_t) = \prod_{i=1}^{3} \Big(\sum_{j=1}^{3} \big(\mathbf{P}_{R_1} \mathbf{P}_{R_2} \cdots \mathbf{P}_{R_t} \big)_{ij} x_j \Big)^{n_i(0)}, \tag{26}$$

and the unconditional PGF is the average of (26) over regime paths, computed by a forward recursion.

Proposition 3 *Let p_R be the one-day probability that a cell is in A under regime R, and $\bar{p} = E[p_R]$. Then*

$$\operatorname{Var}[n_3] = E\big[\operatorname{Var}(n_3 \mid R)\big] + \operatorname{Var}\big[E(n_3 \mid R)\big] = N\bar{p}(1-\bar{p}) + N(N-1)\operatorname{Var}[p_R], \tag{27}$$

the indicators of two distinct cells $a \neq b$ being in A are positively correlated, with

$$\operatorname{corr}\big(\mathbb{1}\{a \in A\}, \mathbb{1}\{b \in A\}\big) = \frac{\operatorname{Var}[p_R]}{\bar{p}(1-\bar{p})}, \tag{28}$$

and the dispersion index is

$$D = \frac{\operatorname{Var}[n_3]}{E[n_3]} = (1-\bar{p}) + (N-1)\frac{\operatorname{Var}[p_R]}{\bar{p}}. \tag{29}$$

Proof Given R, $n_3 \sim \operatorname{Bin}(N, p_R)$, so $E[\operatorname{Var}(n_3 \mid R)] = NE[p_R(1-p_R)] = N\bar{p}(1-\bar{p}) - N\operatorname{Var}[p_R]$ and $\operatorname{Var}[E(n_3 \mid R)] = N^2 \operatorname{Var}[p_R]$; adding gives (27). For $a \neq b$ the two indicators are conditionally independent, so their covariance is $\operatorname{Var}[p_R]$, which gives (28); (29) is (27) divided by $N\bar{p}$. □

The regime adds a term of order N^2 to a term of order N, so with 57 cells a small spread of the regime-specific probabilities produces a large excess. For two regimes with probabilities ψ and $1-\psi$ and rates p_1, p_2, $\mathrm{Var}[p_R] = \psi(1-\psi)(p_1-p_2)^2$, and (27) collapses to the single-matrix variance when $p_1 = p_2$.

4.2 A Dirichlet-multinomial emission

A finite regime takes only M values, so (27) can reach a large variance only by adding regimes. The second relaxation lets the destination shares of a day be random. Given regime r and the n_i cells in S_i at the end of the day before, let

$$\mathbf{q}_i \sim \mathrm{Dirichlet}\big(\alpha_r\, p_{i1}^{(r)},\ \alpha_r\, p_{i2}^{(r)},\ \alpha_r\, p_{i3}^{(r)}\big), \qquad \alpha_r > 0, \tag{30}$$

and let the n_i cells move independently with the drawn shares $\mathbf{q}_i$. The row of transition counts $\mathbf{N}_i = (N_{i1}, N_{i2}, N_{i3})$ is then Dirichlet-multinomial (DM) (Mosimann 1962):

$$P(\mathbf{N}_i = \mathbf{m} \mid n_i, r) = \binom{n_i}{m_1, m_2, m_3} \frac{\Gamma(\alpha_r)}{\Gamma(n_i + \alpha_r)} \prod_{j=1}^{3} \frac{\Gamma\big(m_j + \alpha_r p_{ij}^{(r)}\big)}{\Gamma\big(\alpha_r p_{ij}^{(r)}\big)}. \tag{31}$$

Proposition 4 *Under (31), $E[N_{ij}] = n_i p_{ij}^{(r)}$ and*

$$\mathrm{Var}[N_{ij}] = n_i\, p_{ij}^{(r)}\big(1 - p_{ij}^{(r)}\big)\, \frac{n_i + \alpha_r}{1 + \alpha_r}. \tag{32}$$

The inflation factor tends to 1 as $\alpha_r \to \infty$, where the multinomial emission is recovered, and to n_i as $\alpha_r \to 0$, where all the cells of a row move together. Given the shares, the cells are independent, and for one count the mixture of the conditional product form is again a closed form:

$$E\big[(1 - q + qs)^n\big] = \sum_{k=0}^{n} \binom{n}{k} \frac{B\big(k + \alpha p,\ n - k + \alpha(1-p)\big)}{B\big(\alpha p,\ \alpha(1-p)\big)}\, s^k, \tag{33}$$

the PGF of the beta-binomial law, where $q \sim \mathrm{Beta}(\alpha p, \alpha(1-p))$ and $B(\cdot,\cdot)$ is the beta function.

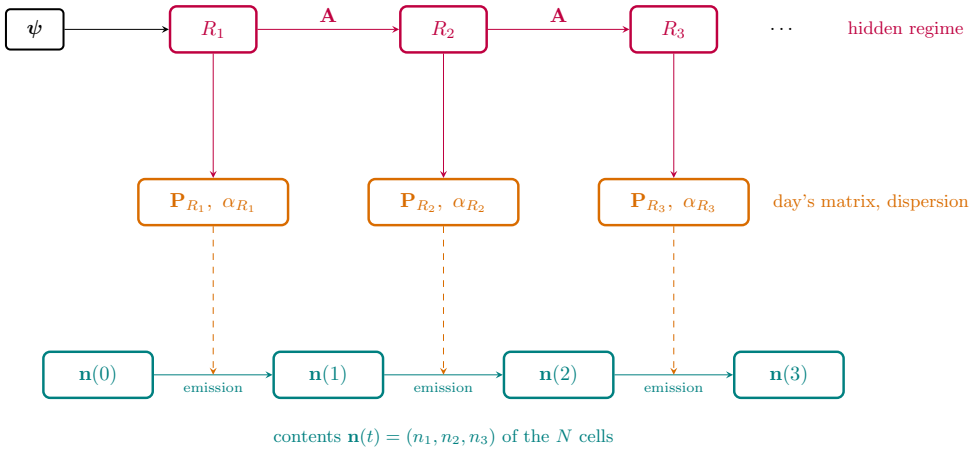

Fig. 3 The regime model and the Dirichlet-multinomial emission on the daily scale. The initial law ψ and the matrix $\mathbf{A}$ drive the hidden regime chain $R_1, R_2, \ldots$ (purple). The regime of day t selects that day's one-day matrix $\mathbf{P}_{R_t}$ and concentration α_{R_t} (orange), which act on the contents of the day before (teal) through the emission: independent cells, so that each row of transition counts is multinomial, or a random share of destinations (30), so that each row is Dirichlet-multinomial (31). With one regime and the multinomial emission the picture is the discrete-time twin (25)

Proof Conditioning on $\mathbf{q}_i$, $E[N_{ij}] = n_i E[q_{ij}] = n_i p_{ij}^{(r)}$ and $\mathrm{Var}[N_{ij}] = n_i E[q_{ij}(1-q_{ij})] + n_i^2 \mathrm{Var}[q_{ij}]$; with $\mathrm{Var}[q_{ij}] = p_{ij}^{(r)}(1-p_{ij}^{(r)})/(1+\alpha_r)$ this is (32). The marginal of one coordinate of a Dirichlet vector is beta, and integrating $\binom{n}{k} q^k (1-q)^{n-k}$ against the beta density gives the coefficients of (33). □

One parameter per regime therefore spans the whole range from independence to complete co-movement, and the multinomial model is the limiting case, so the two emissions are nested. Fig. 3 shows how the two relaxations sit together on the daily scale.

4.3 Why heterogeneous cells are not the explanation

The first explanation a reader reaches for is that the cells are not alike: some are coastal and some inland. The objection is right about the cells and wrong about the consequence.

Proposition 5 *Let the cells be independent, cell a being in A with probability p_a, and let $\bar{p} = N^{-1}\sum_a p_a$. Then*

$$N\bar{p}(1-\bar{p}) - \sum_{a=1}^{N} p_a(1-p_a) = \sum_{a=1}^{N} p_a^2 - \frac{1}{N}\Big(\sum_{a=1}^{N} p_a\Big)^2 \;\geq\; 0, \tag{34}$$

with equality only when all p_a are equal.

Proof Expand both sides; the inequality is the Cauchy–Schwarz inequality $(\sum_a p_a)^2 \leq N\sum_a p_a^2$. □

At a fixed mean, heterogeneity between independent cells *lowers* the variance of the count below the identical-cell value. It can never produce overdispersion, and adding it to the twin would widen the gap with the data. Overdispersion in a count of cells can come only from dependence between cells, which is why both relaxations keep the cells exchangeable and relax independence instead.

4.4 The mean level under both relaxations

Let $\mathbf{u}_t(r) = E\big[\mathbf{n}(t)\mathbb{1}\{R_t = r\}\big]$, a row vector over the three states.

Proposition 6 *Under the regime model with either emission,*

$$\mathbf{u}_1(r) = \psi_r\,\mathbf{n}(0)\,\mathbf{P}_r, \qquad \mathbf{u}_t(r) = \sum_{r'=1}^{M} a_{r'r}\,\mathbf{u}_{t-1}(r')\,\mathbf{P}_r \quad (t \geq 2), \qquad E[\mathbf{n}(t)] = \sum_{r=1}^{M} \mathbf{u}_t(r), \tag{35}$$

and the concentrations α_r do not enter.

Proof The regime chain does not depend on the counts, and given $R_t = r$ and $\mathbf{n}(t-1)$ the mean of the day's counts is $\mathbf{n}(t-1)\mathbf{P}_r$ under both emissions, by Proposition 4. Conditioning on R_{t-1} gives (35). □

Equivalently, the state of one cell and the regime together form a Markov chain on $3M$ points with transition probabilities $a_{r'r}\,p_{ij}^{(r)}$ from (i, r') to (j, r), and $E[\mathbf{n}(t)]$ is N

times the state margin of that chain started from the observed first day; with $M = 1$, (35) reduces to $E[\mathbf{n}(t)] = \mathbf{n}(0)\mathbf{P}^t$, which is (18) for the twin. The model therefore never changes the mean level through α_r. A *fit* can, for the following reason. Let $N_{ij}(t)$ be the number of cells in S_i on day $t-1$ and in S_j on day t. The counting estimator

$$\hat{P}_{ij} = \frac{\sum_t N_{ij}(t)}{\sum_t \sum_k N_{ik}(t)} \tag{36}$$

reproduces the observed totals identically,

$$\Big(\sum_t \mathbf{n}(t-1)\Big)\hat{\mathbf{P}} = \sum_t \mathbf{n}(t), \tag{37}$$

because the ith row total of the transition counts is $\sum_t n_i(t-1)$ and the jth column total is $\sum_t n_j(t)$. The maximum-likelihood estimator under (31) solves a different equation. For one row, with $\mathbf{p}$ taken as free coordinates and ψ_0 the digamma function, the score is

$$\begin{aligned}
\frac{\partial \log f}{\partial p_j} &= \alpha\big[\psi_0(m_j + \alpha p_j) - \psi_0(\alpha p_j)\big], \\
\frac{\partial \log f}{\partial \log \alpha} &= \alpha\Big[\psi_0(\alpha) - \psi_0(n+\alpha) + \sum_j p_j\big(\psi_0(m_j + \alpha p_j) - \psi_0(\alpha p_j)\big)\Big],
\end{aligned} \tag{38}$$

where f is (31). When n_i is large compared with α_r, a row carries little more information than one draw of the shares, so the estimator weights days rather than cell-days, whereas (36) weights each day by its number of cells. When the dependence has exactly the Dirichlet form the two estimators converge to the same matrix; when it does not, they converge to different limits. Even for one row of ten cells, all going to one state on the first day and splitting five and five on the second, the two estimating equations have different roots: the counting estimate is 0.75 and the Dirichlet-multinomial estimate at $\alpha = 1$ is 0.766. A gap between the level implied by a fitted DM model and the

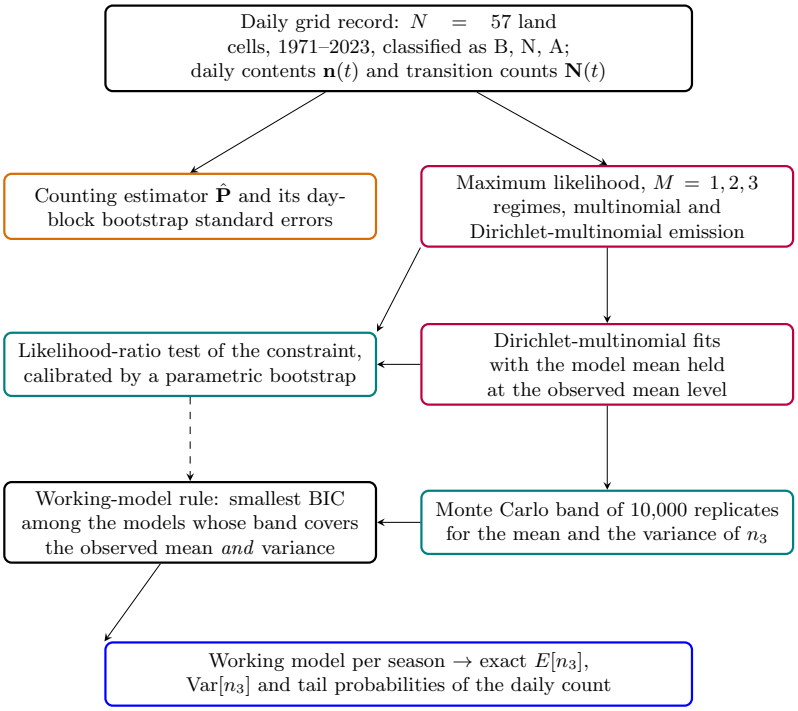

Fig. 4 Route of estimation and validation. Orange: the counting estimator of the single matrix and its day-block standard errors (Section 5.1). Purple: the likelihood models (Section 5.2), unconstrained and with the mean level held (Section 5.4). Teal: the test of the constraint and the Monte Carlo band (Sections 5.3 and 5.4). Solid arrows are computations; the dashed arrow marks a result that is reported but is not an input of the working-model rule

observed level is therefore a symptom that the *form* of the dependence is wrong, not evidence about the level. Section 5.4 restores the level by a constraint and measures what that costs.

5 Estimation and validation

Fig. 4 is the route of this section, from the daily record to the working model of each season.

5.1 Counting estimator and day-block standard errors

Let $\mathbf{N}(t) = (N_{ij}(t))$ be the 3×3 matrix of transition counts of day t, with $\sum_j N_{ij}(t) = n_i(t-1)$. Pooling over the cells and the season-years of the fitting window gives (36), which is the maximum-likelihood estimator when the cells are independent given the matrix and remains consistent for P_{ij} under the dependence of Section 4, because that dependence leaves the one-day mean unchanged. A standard error computed as if the cell-days were independent, $\sqrt{\hat{P}_{ij}(1-\hat{P}_{ij})/\sum_t n_i(t-1)}$, is far too small when the cells of a day move together. Standard errors are therefore obtained by a bootstrap (Efron 1979) that resamples *whole days* with replacement within each season-year, which keeps the dependence among the cells of a day intact; 2,000 resamples are used [simulation-based].

5.2 Likelihood of the regime and DM models

Given the day's regime, the matrix of transition counts has a known law $f_r(\mathbf{N}(t))$: the product of multinomials

$$f_r(\mathbf{N}) = \prod_{i=1}^{3} \binom{n_i}{N_{i1}, N_{i2}, N_{i3}} \prod_{j=1}^{3} \left(p_{ij}^{(r)}\right)^{N_{ij}} \tag{39}$$

for the multinomial (MN) emission, and the product of the DM laws (31) for the DM emission. The likelihood of a season-year $\mathbf{N}(1), \ldots, \mathbf{N}(T)$ is the forward recursion (Rabiner 1989)

$$L = \sum_{r_1,\ldots,r_T} \psi_{r_1} f_{r_1}(\mathbf{N}(1)) \prod_{t=2}^{T} a_{r_{t-1} r_t} f_{r_t}(\mathbf{N}(t)) = \boldsymbol{\psi}\, \mathbf{F}(1)\, \mathbf{AF}(2) \cdots \mathbf{AF}(T)\, \mathbf{1}, \tag{40}$$

with $\mathbf{F}(t) = \operatorname{diag}\big(f_1(\mathbf{N}(t)), \ldots, f_M(\mathbf{N}(t))\big)$; season-years are separate sequences and the recursion is scaled at every step. With $M = 1$ the recursion is a product and (36)

is the exact maximum for the MN emission. The MN models are fitted by the Baum–Welch algorithm, whose maximisation step is a weighted version of (36), and the DM models by direct numerical maximisation (L-BFGS-B on the log-odds of the matrices and $\log \alpha_r$), from several random starts; the DM maximisation was then repeated with the exact gradient (43) below, from the best of these and from six further random starts, which raised the maximum in all eight fits with two or three regimes by 0.01 to 8.3 in log-likelihood and left the single-matrix fits unchanged. The number of free parameters is $k = M(M-1)+(M-1)+6M$ for the MN emission and M more for the DM. Log-likelihoods include the multinomial coefficients, so that the two emissions are on one scale, and the gain from α_r is tested by the likelihood ratio (LR); the null value $\alpha_r = \infty$ lies on the boundary, so the asymptotic chi-square is conservative [asymptotic].

5.3 Model comparison and Monte Carlo validation

Models are compared by the Bayesian information criterion $\mathrm{BIC} = -2\ell + k \log n$ (Schwarz 1978) on the fitting days, with ℓ the maximised log-likelihood and n the number of days, and by the mean logarithmic score of the one-day-ahead prediction on the hold-out days,

$$S = \frac{1}{n_{\mathrm{hold}}} \sum_{t} \log f\big(\mathbf{N}(t) \mid \mathbf{N}(1), \ldots, \mathbf{N}(t-1)\big), \tag{41}$$

larger being better, with the predictive law taken from the filtered regime distribution of the day before. The test that matters here, however, is the *spread* of the daily counts. Each fitted model is used to simulate 10,000 replicates of the fitting window: every season-year starts from its observed first day, the regime path is drawn from $\mathbf{A}$, the shares from (30) where the model has them, and the cells then move independently. Four statistics are computed on every replicate: the mean and the variance of the

daily count n_3, the probability that $n_3 \geq 5$ on a day, and the lag-one autocorrelation of the daily count. A model passes on a statistic if the observed value lies inside the central 95% of the replicates [simulation-based]. The same simulation checks the code against the algebra: the simulated mean of $n_3(t)$ is compared with (35), and for the single-matrix model the simulated variance with (19).

5.4 Maximum likelihood under a constraint on the mean level

Let $\bar{\mathbf{m}}(\boldsymbol{\theta})$ be the mean count of each state over the fitting days implied by $\boldsymbol{\theta} = (\boldsymbol{\psi}, \mathbf{A}, \mathbf{P}_r, \alpha_r)$, computed exactly by (35) from the observed first day of every season-year, and $\bar{\mathbf{n}}$ the observed mean over the same days. The constrained estimator is

$$\tilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ \ell(\boldsymbol{\theta}) \quad \text{subject to} \quad \bar{m}_2(\boldsymbol{\theta}) = \bar{n}_2, \ \ \bar{m}_3(\boldsymbol{\theta}) = \bar{n}_3, \tag{42}$$

with ℓ the DM log-likelihood (40); the level of B follows from (2). The constraint fixes the level and leaves α_r free. The gradient of ℓ is exact, by the forward–backward recursion,

$$\frac{\partial \ell}{\partial \theta} = \sum_{t} \sum_{r=1}^{M} \gamma_t(r) \, \frac{\partial \log f_r(\mathbf{N}(t))}{\partial \theta} + \sum_{t \geq 2} \sum_{r', r} \xi_t(r', r) \, \frac{\partial \log a_{r'r}}{\partial \theta} + \sum_{r} \gamma_1(r) \, \frac{\partial \log \psi_r}{\partial \theta}, \tag{43}$$

where $\gamma_t(r)$ and $\xi_t(r', r)$ are the posterior probabilities of the regime on day t and of the pair of regimes on days $t-1, t$, and $\partial \log f_r / \partial \theta$ is (38) carried through the soft-max coding of the matrices; it agreed with central differences to a relative error below 10^{-5} at every fit. Problem (42) is solved by sequential least-squares quadratic programming (SLSQP), the gradient of the constraint being taken by central differences of (35), from the unconstrained optimum and six random starts; the constraint is met to 10^{-11} or better at every reported solution. Two numerical safeguards are needed: the ratios of gamma functions in (31) are computed as log rising factorials, $\log \Gamma(x+n) - \log \Gamma(x) =$

$\sum_{k<n} \log(x+k)$, by a series in $1/x$ when $x > 10^5$; and $\log \alpha_r$ is kept in $[-7, 25]$, the upper end being numerically indistinguishable from the MN limit. Both were checked against 50-digit arithmetic.

The cost of the constraint is $\mathrm{LR} = 2\{\ell(\hat{\boldsymbol{\theta}}) - \ell(\tilde{\boldsymbol{\theta}})\}$. The chi-square law with two degrees of freedom does not apply as it stands, because the constraint is set at the *sample* mean, so the null law is obtained by a parametric bootstrap from $\tilde{\boldsymbol{\theta}}$: the thirty season-years are simulated from their observed first days, both estimators are recomputed on each replicate with the constraint set at the replicate's own mean, and $p = (1 + \#\{\mathrm{LR}^* \geq \mathrm{LR}\})/(B+1)$ with $B = 99$ [simulation-based]. A small LR says that the Dirichlet form of the dependence is compatible with the observed level.

Working-model rule. The rule was fixed before any constrained fit was run. The candidates are the constrained DM models with $M = 1, 2, 3$. A candidate passes if the observed mean *and* the observed variance of the daily count n_3 both lie inside the central 95% of its 10,000 replicates; among passing candidates the one with the smallest BIC is the working model of the season. The lag-one autocorrelation is reported but not used.

6 Results

6.1 The cells move together

Table 2 gives the mean and variance of the three daily counts over 1991–2020, with the variance $N\bar{p}(1-\bar{p})$ that independent cells would give at the same mean. In the NE monsoon the number of cells in A on a day has mean 0.995 and variance 18.17, against 0.978 for independent cells: the observed variance is 18.6 times the value that the twin allows once its mean is fitted. The ratio is 8.9 in Winter, 8.1 in the Pre-monsoon and 4.8 in the SW monsoon, and it lies between 19.0 and 28.6 for the counts of B and N. Fig. 5 shows what this means for the law of the count: independent cells make days with more than a handful of cells in A practically impossible, while the

Table 2 Mean and variance of the daily number of cells in each state, 1991–2020, $N = 57$. "Indep." is $N\bar{p}(1-\bar{p})$, the variance of independent cells at the same mean; "Ratio" is the observed variance divided by it

Season	State	Days	Mean	Variance	Indep.	Ratio
Winter	B, n_1	1778	55.614	27.817	1.352	20.6
	N, n_2		1.360	26.172	1.328	19.7
	A, n_3		0.026	0.229	0.026	8.9
Pre-monsoon	B, n_1	2760	52.945	83.033	3.766	22.0
	N, n_2		3.998	78.860	3.717	21.2
	A, n_3		0.057	0.458	0.057	8.1
SW monsoon	B, n_1	3660	45.152	184.580	9.385	19.7
	N, n_2		11.698	176.785	9.297	19.0
	A, n_3		0.150	0.724	0.150	4.8
NE monsoon	B, n_1	2760	40.946	330.282	11.532	28.6
	N, n_2		15.059	278.825	11.080	25.2
	A, n_3		0.995	18.168	0.978	18.6

record has them in every season. The reading is physical: heavy rain over this region arrives with a disturbance that covers many cells at once. By Proposition 5 the gap cannot be closed by letting the cells differ; it requires dependence.

6.2 Transition matrices

Table 3 gives the counted cell-day transitions and the one-day matrix $\hat{\mathbf{P}}$ of the twin, with day-block bootstrap standard errors. In the NE monsoon a cell in A is followed by another day in A with probability 0.241 and by a day in N with probability 0.578, so heavy rain over a cell is persistent and decays through N; in the SW monsoon the same entry is 0.030 and in Winter 0.043. B is the most persistent state everywhere, from 0.842 in the NE monsoon to 0.984 in Winter, so a spell in B over a cell lasts $1/(1-0.842) = 6.3$ days in the NE monsoon and 63.3 days in Winter, and a spell in A lasts 1.3 days in the NE monsoon (Table 1). The day-block standard errors are 1.5 to 4.1 times the naive values that treat cell-days as independent, the largest ratios being for the moves out of B, where most cells are; the co-movement of the cells makes the effective sample many times smaller than the number of cell-days. In the NE monsoon, for example, $\hat{P}_{13} = 0.0025$ with a day-block standard error of 0.0003 and a central

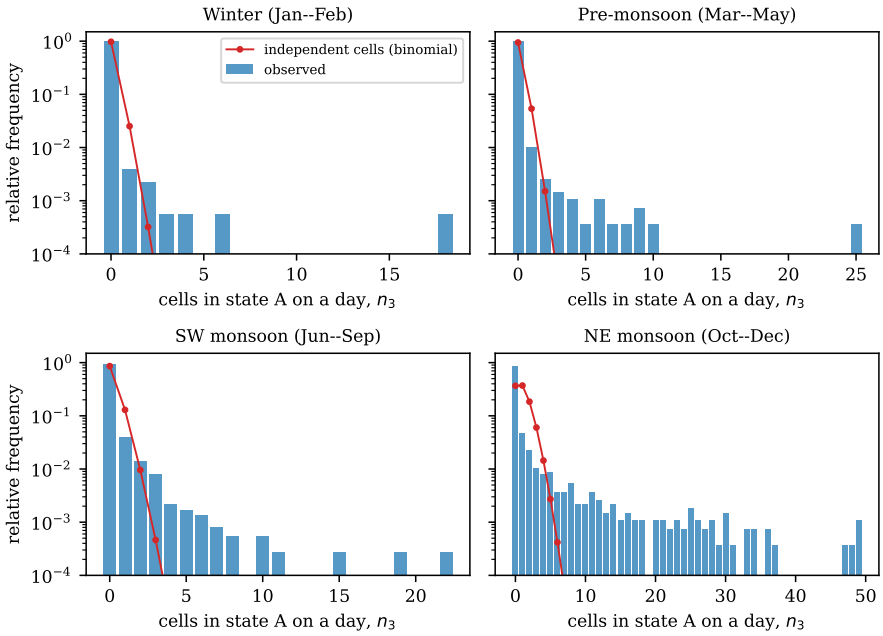

Fig. 5 Relative frequency of the daily number n_3 of cells in A ($\geq$ 64.5 mm), 1991–2020, by season (bars, logarithmic scale), against the binomial law of 57 independent cells with the same mean (line). Frequencies below 10^{-4} are not drawn

95% bootstrap interval of 0.0020 to 0.0031. Out of A in the drier seasons few cell-days are available: in Winter $\hat{P}_{33} = 0.043$ has a standard error of 0.057, so that entry is barely estimated [simulation-based].

6.3 The ladder of spatial models

Table 4 places the models on one scale. Adding regimes to the single-matrix MN model helps a great deal: in the NE monsoon the BIC falls from 89,467 at one regime to 48,353 at two and 38,380 at three, and the hold-out score rises from −18.88 to −7.68. But *one* parameter of overdispersion does more than two extra regimes, and does it in every season. The single-matrix DM model, with seven free parameters, beats the three-regime MN model, with twenty-six, on BIC and on the hold-out score in all four seasons: 30,808 against 38,380 in the NE monsoon, 37,706 against 42,615 in the SW monsoon, 13,927 against 15,289 in the Pre-monsoon and 3,669 against 3,902 in

Table 3 Counted cell-day transitions and the one-day matrix $\hat{\mathbf{P}}$ of the twin (25), by season, pooled over the 57 cells and the season-years of 1991–2020 by (36). In brackets, the standard error from 2,000 day-block bootstrap resamples, which keep the cells of a day together

		Counts to			$\hat{P}_{ij}$ (day-block SE) to		
Season	From	B	N	A	B	N	A
Winter	B	95,639	1,524	12	0.9842 (0.0015)	0.0157 (0.0015)	0.0001 (0.0001)
	N	1,637	748	30	0.6778 (0.0326)	0.3097 (0.0307)	0.0124 (0.0073)
	A	23	21	2	0.5000 (0.2234)	0.4565 (0.2139)	0.0435 (0.0565)
Pre-monsoon	B	137,140	7,433	47	0.9483 (0.0022)	0.0514 (0.0022)	0.0003 (0.0001)
	N	7,237	3,500	96	0.6681 (0.0168)	0.3231 (0.0162)	0.0089 (0.0026)
	A	67	76	14	0.4268 (0.0984)	0.4841 (0.0788)	0.0892 (0.0377)
SW monsoon	B	138,515	25,332	281	0.8439 (0.0034)	0.1543 (0.0033)	0.0017 (0.0002)
	N	25,064	16,929	247	0.5934 (0.0078)	0.4008 (0.0076)	0.0058 (0.0007)
	A	241	285	16	0.4446 (0.0398)	0.5258 (0.0371)	0.0295 (0.0116)
NE monsoon	B	93,873	17,343	279	0.8419 (0.0045)	0.1555 (0.0044)	0.0025 (0.0003)
	N	17,449	22,134	1,792	0.4217 (0.0091)	0.5350 (0.0085)	0.0433 (0.0037)
	A	494	1,585	661	0.1803 (0.0211)	0.5785 (0.0256)	0.2412 (0.0264)

Winter. In the NE monsoon it has a higher log-likelihood than every MN model with up to five regimes and a lower BIC than every one with up to six; only the five- and six-regime MN models, with 54 and 71 parameters, score slightly higher on the hold-out days (-6.56 and -6.38 against -6.63). The estimated concentrations are small, from $\hat{\alpha} = 1.23$ in the NE monsoon to 1.85 in the SW monsoon, so the inflation factor of (32) is large: for a row of 41 cells in B it is $(41 + 1.23)/(1 + 1.23) \approx 19$ in the NE monsoon, the order of the excess in Table 2. The LR test of MN against DM is decisive at every M in every season: in the NE monsoon twice the difference in log-likelihood is 58,667.0 on one parameter at $M = 1$, 19,758.0 on two at $M = 2$ and 10,714.9 on three at $M = 3$, and the smallest value in any season is 619.2, in Winter at $M = 3$ [asymptotic; at statistics of this size the boundary does not matter].

The two devices are not substitutes: the three-regime DM model has the lowest BIC and the highest hold-out score of the six fits in every season, and its concentrations spread out as the regimes separate. In Winter $\hat{\alpha} = (0.15, 5.18, 21.51)$: one regime behaves almost as a switch that moves all the cells of a row together, another is close to

Table 4 The ladder of spatial models, by season. MN, multinomial and DM, Dirichlet-multinomial emission; M regimes; k free parameters; ℓ maximised log-likelihood including the multinomial coefficients of (39); BIC on the fitting days of 1991–2020; S the mean log score (41) on the hold-out days of 2021–2023 (larger is better). The MN model with $M = 1$ is the discrete-time twin. Best BIC and best S of each season in bold

Season	Emission	M	k	ℓ	BIC	Hold-out S
Winter	MN	1	6	−5,833.9	11,712.6	−6.704
	MN	2	15	−2,265.2	4,642.5	−2.727
	MN	3	26	−1,853.8	3,901.7	−2.363
	DM	1	7	−1,808.3	3,668.8	−2.328
	DM	2	17	−1,594.5	3,316.0	−2.037
	DM	3	29	−1,544.2	**3,305.0**	**−1.962**
Pre-monsoon	MN	1	6	−19,796.8	39,641.1	−9.054
	MN	2	15	−9,663.1	19,444.9	−4.276
	MN	3	26	−7,541.7	15,289.1	−3.410
	DM	1	7	−6,935.8	13,926.9	−3.153
	DM	2	17	−6,406.1	12,946.7	−2.940
	DM	3	29	−6,183.8	**12,597.0**	**−2.810**
SW monsoon	MN	1	6	−44,860.6	89,770.4	−13.857
	MN	2	15	−25,474.9	51,072.8	−7.428
	MN	3	26	−21,201.1	42,615.4	−6.292
	DM	1	7	−18,824.2	37,705.7	−5.767
	DM	2	17	−17,758.9	35,657.2	−5.424
	DM	3	29	−17,409.5	**35,056.7**	**−5.296**
NE monsoon	MN	1	6	−44,709.9	89,467.2	−18.875
	MN	2	15	−24,116.9	48,352.5	−10.142
	MN	3	26	−19,087.0	38,379.7	−7.679
	DM	1	7	−15,376.3	30,808.1	−6.627
	DM	2	17	−14,237.9	28,610.3	−6.143
	DM	3	29	−13,729.6	**27,688.6**	**−5.914**

independent cells. A day's weather over the region has both a *kind*, which the regime carries, and a *coherence*, which α carries.

6.4 Monte Carlo validation

Code against algebra. For every fitted model the simulated mean of $n_3(t)$ was compared with the exact mean (35) in each of the 30 season-years and on each day: over the 1,770 to 3,660 (season-year, day) cells of a season the mean squared standardised difference lay between 0.93 and 1.10 for the thirty-three models other than the DM models of the two dry seasons, against 1 if the only difference is Monte Carlo error. Under those DM models the mean of a day comes from rare days on which a whole row of cells turns to A at once, the standardised differences are far from normal, and the mean

squared difference rises to 1.84 and the largest single one to 14.0. The robust check is the grand mean over all days, which pools those rare events: there the simulated and exact means agree to within 2.3 standard errors for every one of the 45 models, three of the 45 lying beyond 2, as chance alone would give about two. For the single-matrix model the median relative difference between simulated and exact variance (19) was 1.15% in the NE monsoon, 4.79% in Winter, 2.96% in the Pre-monsoon and 1.92% in the SW monsoon, all consistent with 10,000 replicates [simulation-based].

Models against data. Table 5 does not agree with Table 4. In Winter no MN model up to four regimes comes near the observed variance of 0.23, and all three DM models pass, the bands being wide in this nearly dry season. In the Pre-monsoon no MN model passes, their variance reaching only 0.14 against 0.458, and only the three-regime DM model passes on both mean and variance. In the NE monsoon the single-matrix DM model passes on variance, 20.43 with band 15.72 to 25.74 against 18.17, but not on the mean, 1.41 with band 1.21 to 1.62 against 0.995; among the MN models the variance climbs with the number of regimes, 0.99, 2.61, 4.83, 10.80, 14.58, 14.49, and those with five and six regimes pass on both. In the SW monsoon nothing passes: every MN model undershoots the variance, 0.15 to 0.26 against 0.724, and every DM model overshoots it, 1.94 to 6.61, while also overstating the mean, 0.26 to 0.47 against 0.150. The probability that five or more cells are in A on a day, the quantity a relief plan needs, is badly understated by independent cells: in the NE monsoon the observed 0.0562 against 0.0036 for the single-matrix MN model, and in the Pre-monsoon the observed 0.0036 against at most 0.0014 for every MN model up to five regimes. The day-to-day persistence of the count behaves the other way: in the NE monsoon the observed lag-one autocorrelation is 0.333, the MN models with regimes give 0.48 and 0.44 at two and three regimes and 0.37, 0.34 and 0.32 at four to six, and the DM models give 0.14 with one matrix and 0.21 and 0.26 with regimes.

Table 5 Monte Carlo test of the fitted models, 10,000 replicates of the thirty season-years of 1991–2020, each started from its observed first day. For each model, the mean over replicates of the statistic, with the central 95% of the replicate values in brackets; the first row of each season is the observed value. A tick means the observed mean, or variance, of the daily count n_3 lies inside the band. Lag-1 ac: lag-one autocorrelation of the daily count. MN rows with $M \leq 3$ and DM rows: the fits of Table 4; MN rows with $M \geq 4$: Baum–Welch fits

Season	Model	M	Mean of n_3 [95% band]	Variance of n_3 [95% band]	$P(n_3 \geq 5)$	Lag-1 ac	Mean	Var.
Winter	*Observed*		0.026	0.230	0.0011	0.12		
	MN	1	0.025 [0.018, 0.033]	0.03 [0.02, 0.04]	0.0000	0.02	✓	–
	MN	2	0.023 [0.014, 0.033]	0.03 [0.02, 0.05]	0.0000	0.09	✓	–
	MN	3	0.024 [0.012, 0.038]	0.04 [0.02, 0.08]	0.0001	0.12	✓	–
	MN	4	0.024 [0.010, 0.044]	0.06 [0.02, 0.14]	0.0006	0.13	✓	–
	MN	5	0.027 [0.008, 0.055]	0.10 [0.01, 0.25]	0.0015	0.14	✓	✓
	DM	1	0.066 [0.025, 0.127]	1.10 [0.11, 3.25]	0.0035	0.02	✓	✓
	DM	2	0.039 [0.014, 0.075]	0.37 [0.04, 1.22]	0.0023	0.03	✓	✓
	DM	3	0.032 [0.011, 0.062]	0.26 [0.03, 0.94]	0.0019	0.04	✓	✓
Pre-monsoon	*Observed*		0.057	0.458	0.0036	0.09		
	MN	1	0.057 [0.047, 0.067]	0.06 [0.05, 0.07]	0.0000	0.06	✓	–
	MN	2	0.052 [0.040, 0.064]	0.07 [0.05, 0.09]	0.0000	0.12	✓	–
	MN	3	0.055 [0.039, 0.071]	0.10 [0.07, 0.15]	0.0003	0.15	✓	–
	MN	4	0.052 [0.037, 0.071]	0.12 [0.07, 0.19]	0.0009	0.13	✓	–
	MN	5	0.054 [0.036, 0.075]	0.14 [0.08, 0.22]	0.0014	0.14	✓	–
	DM	1	0.145 [0.089, 0.213]	2.47 [0.90, 4.62]	0.0082	0.04	–	–
	DM	2	0.103 [0.062, 0.150]	1.17 [0.43, 2.31]	0.0066	0.05	–	✓
	DM	3	0.095 [0.057, 0.141]	1.04 [0.37, 2.06]	0.0061	0.05	✓	✓
SW monsoon	*Observed*		0.150	0.724	0.0060	0.08		
	MN	1	0.150 [0.138, 0.163]	0.15 [0.14, 0.17]	0.0000	0.02	✓	–
	MN	2	0.147 [0.132, 0.163]	0.19 [0.16, 0.21]	0.0000	0.10	✓	–
	MN	3	0.149 [0.131, 0.168]	0.24 [0.20, 0.28]	0.0003	0.11	✓	–
	MN	4	0.149 [0.130, 0.169]	0.26 [0.21, 0.30]	0.0006	0.11	✓	–
	MN	5	0.149 [0.130, 0.169]	0.26 [0.22, 0.31]	0.0006	0.11	✓	–
	DM	1	0.470 [0.388, 0.558]	6.61 [4.59, 8.98]	0.0291	0.01	–	–
	DM	2	0.321 [0.264, 0.383]	3.03 [2.06, 4.18]	0.0218	0.03	–	–
	DM	3	0.259 [0.212, 0.311]	1.94 [1.33, 2.69]	0.0179	0.05	–	–
NE monsoon	*Observed*		0.995	18.168	0.0562	0.33		
	MN	1	0.983 [0.937, 1.030]	0.99 [0.92, 1.06]	0.0036	0.21	✓	–
	MN	2	0.880 [0.777, 0.989]	2.61 [2.27, 2.97]	0.0520	0.48	–	–
	MN	3	0.896 [0.757, 1.047]	4.83 [3.92, 5.82]	0.0819	0.44	✓	–
	MN	4	0.960 [0.757, 1.182]	10.80 [7.82, 14.06]	0.0596	0.37	✓	–
	MN	5	0.965 [0.741, 1.213]	14.58 [9.75, 20.00]	0.0430	0.34	✓	✓
	MN	6	0.961 [0.738, 1.207]	14.49 [9.82, 19.79]	0.0429	0.32	✓	✓
	DM	1	1.410 [1.213, 1.617]	20.43 [15.72, 25.74]	0.0864	0.14	–	✓
	DM	2	0.942 [0.780, 1.112]	10.52 [7.80, 13.53]	0.0652	0.21	✓	–
	DM	3	0.980 [0.788, 1.191]	12.13 [8.82, 15.90]	0.0692	0.26	✓	–

6.5 The mean-constrained fits and the working models

The unconstrained DM fits overstate the mean level of n_3 in every season, most with a single matrix: the implied mean over the fitting days is 1.42 against the observed 1.00 in the NE monsoon, 0.47 against 0.15 in the SW monsoon, 0.15 against 0.058 in the Pre-monsoon and 0.066 against 0.025 in Winter, and with three regimes the gap shrinks but does not close outside the NE monsoon (0.26 against 0.15 in the SW monsoon). Section 4.4 explained why. Table 6 gives the models refitted under (42).

Holding the level costs little where the model has regimes, except in the SW monsoon: the LR is 45.2, 4.0 and 0.02 for one, two and three regimes in the NE monsoon, 14.8, 2.0 and 0.5 in Winter, 55.9, 12.4 and 8.6 in the Pre-monsoon, and 269.0, 111.5 and 49.1 in the SW monsoon. The bootstrap p-values are 0.01, the smallest attainable with 99 replicates, for every single-matrix model and every SW monsoon model; 0.01, 0.03 and 0.04 for the Pre-monsoon models with two and three regimes and the NE monsoon model with two; and 0.22, 0.54 and 0.88 for the Winter models with two and three regimes and the NE monsoon model with three [simulation-based]. With three regimes in the NE monsoon the Dirichlet form of the dependence is therefore compatible with the observed level, while in the SW monsoon form and level conflict at every number of regimes. Because the constrained models have two parameters fewer, their BIC is the lower in seven of the twelve fits; the hold-out score falls in every fit, by between 0.0001 and 0.09. The concentrations hardly move (in the NE monsoon $\hat{\alpha}$ goes from 1.23 to 1.29 with one matrix), and the constraint does not bring each entry of the matrix back to the counting estimate: in the NE monsoon $\hat{P}_{13}$ of the single-matrix model falls from 0.0110 to 0.0077, still above the counting value 0.0025, the other entries adjusting so that the level of the count is right.

Every constrained model reproduces the observed mean, and holding the level also lowers the simulated variance, because the variance of an overdispersed count grows with its mean. In Winter and the Pre-monsoon all three constrained models pass on

both mean and variance. In the SW monsoon the simulated variance falls from 1.94–6.61 unconstrained to 0.97–1.94, and the three-regime model passes, at 0.97 with band 0.60 to 1.45 against 0.724: the gap left by the unconstrained models was a symptom of the misstated level. In the NE monsoon the single-matrix model passes, at 14.20 with band 10.27 to 18.61 against 18.17.

By the rule of Section 5.4, the working model is the constrained DM model with three regimes in Winter, the Pre-monsoon and the SW monsoon, and with a single matrix in the NE monsoon; every season has one. For three of them, all but Winter, the bootstrap rejects the constraint at the 5% level, which we read as evidence that the Dirichlet form of the dependence is a good but not an exact description of how the cells move together. The working models reproduce the probability that five or more cells are in A on a day in three seasons: in the NE monsoon 0.061 with band 0.051 to 0.071 against the observed 0.056, and likewise in Winter and the Pre-monsoon; in the SW monsoon they overstate it, 0.0097 with band 0.0066 to 0.0131 against 0.0060 [simulation-based]. They do not reproduce the persistence of the count in the two monsoons: the lag-one autocorrelation is 0.12 with band 0.07 to 0.18 against 0.33 in the NE monsoon, and 0.03 with band 0.00 to 0.07 against 0.08 in the SW monsoon. In the NE monsoon the constrained models with regimes come closer, 0.22 and 0.26, but fail on variance, while the MN models with regimes overshoot (0.44 to 0.48). The overdispersion carries the co-movement of cells within a day, the regimes carry the persistence between days, and no model of the ladder carries both in the measure the record shows.

7 Verification of the formulas

Every closed form of Sections 3 and 4 was checked in two independent systems, in exact or symbolic arithmetic. In Python (SymPy and exact rationals), four scripts check:

Table 6 The DM models refitted under the mean constraint (42). LR: twice the log-likelihood lost to the constraint; p: its parametric-bootstrap p-value from 99 replicates (smallest attainable 0.01); BIC and S as in Table 4, with two parameters fewer; $\hat{\alpha}_r$: the constrained concentrations. Mean and variance of n_3 over 10,000 replicates with the central 95% band; the first row of each season is the observed value. "Working" marks the model chosen by the rule of Section 5.4

Season	M	LR	p	BIC	Hold-out S	$\hat{\alpha}_r$	Mean of n_3 [95% band]	Variance of n_3 [95% band]	Mean	Var.	Working
Winter							*Observed* 0.026	*Observed* 0.230			
	1	14.83	0.01	3,668.7	−2.383	1.67	0.026 [0.006, 0.060]	0.35 [0.01, 1.54]	✓	✓	
	2	1.97	0.22	3,303.1	−2.051	(3.07, 17.04)	0.026 [0.007, 0.055]	0.24 [0.02, 0.94]	✓	✓	
	3	0.51	0.54	3,290.5	−1.967	(0.15, 5.30, 21.11)	0.026 [0.008, 0.053]	0.20 [0.02, 0.82]	✓	✓	✓
Pre-monsoon							*Observed* 0.057	*Observed* 0.458			
	1	55.95	0.01	13,967.0	−3.180	1.81	0.057 [0.026, 0.097]	0.85 [0.15, 2.14]	✓	✓	
	2	12.36	0.01	12,943.2	−2.946	(2.74, 9.37)	0.057 [0.029, 0.092]	0.64 [0.16, 1.51]	✓	✓	
	3	8.63	0.03	12,589.8	−2.813	(3.08, 9.24, 24.40)	0.057 [0.029, 0.091]	0.59 [0.15, 1.36]	✓	✓	✓
SW monsoon							*Observed* 0.150	*Observed* 0.724			
	1	269.00	0.01	37,958.3	−5.858	2.00	0.150 [0.107, 0.197]	1.94 [0.97, 3.20]	✓	–	
	2	111.53	0.01	35,752.3	−5.462	(3.69, 6.86)	0.150 [0.114, 0.190]	1.32 [0.75, 2.09]	✓	–	
	3	49.10	0.01	35,089.4	−5.313	(5.44, 6.77, 11.30)	0.150 [0.117, 0.186]	0.97 [0.60, 1.45]	✓	✓	✓
NE monsoon							*Observed* 0.995	*Observed* 18.168			
	1	45.24	0.01	30,837.5	−6.658	1.29	0.995 [0.836, 1.161]	14.20 [10.27, 18.61]	✓	✓	✓
	2	4.00	0.04	28,598.5	−6.147	(2.56, 3.45)	0.996 [0.828, 1.178]	11.32 [8.48, 14.54]	✓	–	
	3	0.02	0.88	27,672.8	−5.914	(3.02, 4.32, 6.63)	0.995 [0.801, 1.207]	12.36 [8.97, 16.11]	✓	–	

the difference–differential equations with their vertex and edge cases, the generating-function equation, its solution, the spectral form (14) against the matrix exponential, the mass function (17) against a direct solution of the equations, the moments, covariances and CGF, the twin against an enumeration of every cell path, and the regime variance decomposition (25 checks); the DM theory (40 checks); the distributional results including (22)–(24), (33) and (34), the last also by a sweep of 20,000 random probability vectors (55 checks); and the mean recursion (35), the identity (37), the score (38), the soft-max chain rule and the forward–backward score identity behind (43), against brute-force enumeration of every regime path and count path (19 checks). The same results were checked independently in Wolfram Language and run in Mathematica, where every check returns True (21, 63 and 22 checks). The numerics of the constrained fit were checked against 50-digit arithmetic (6 checks). Every number in Section 6 was re-read by script from the saved outputs of the fits. All checks pass.

8 Discussion and conclusions

The daily rainfall state of a region's grid cells, written as a closed three-compartment stochastic system, has a complete exact law: the product-of-multinomials solution of Matis and Hartley (1971) gives the mass function, moments, cumulants and stationary law of the number of cells in each IMD category. The data, however, reject the independence of cells on which it rests, and by a wide margin: in the NE monsoon the variance of the daily number of cells in A is 18.6 times the independent-cell value. Two consequences follow for anyone modelling a count of places. First, heterogeneity is not the explanation: by Proposition 5 unequal but independent cells would lower the variance further, so the gap is evidence of dependence. Second, dependence can be admitted without losing the exact theory. A hidden regime adds a variance term of order N^2 (Proposition 3), and a DM emission inflates the variance of each row by $(n_i + \alpha_r)/(1 + \alpha_r)$ (Proposition 4); in both the cells are independent given the regime

or the shares, so the closed forms hold conditionally and the unconditional law is a mixture.

The data favour the DM device strongly: by BIC and hold-out score one concentration parameter does more than two extra regimes in every season, and in the NE monsoon it gives a higher likelihood than four extra regimes. But a freely fitted DM model misstates the mean level of the count, because its estimating equation weights days rather than cell-days (Section 4.4). Held at the observed level by the constrained estimator, its simulated variance falls into the observed range, and a working model that reproduces the mean and variance of the daily count is found in every season. For the Puducherry region this answers a question a station cannot. In the NE monsoon about one of the 57 cells is in A on an average day, but the count is strongly right-skewed: days with five or more cells in A occur with probability about 0.056, roughly one day in eighteen of the season, and such days are what a relief plan has to size. Independent cells would put that probability near 0.004, an order of magnitude too low; the working model puts it at 0.061, with a band that contains the observed value. The relevance of such counts is underlined by Shahi and Rai (2023), who define a widespread extreme event by the number of grid cells above a threshold and report that such events over south peninsular India have become more frequent.

The limitations are these. The grid ends in 2023, so the hold-out window is three years. A grid value is an area average, so the counts describe the gridded product, not a network of gauges. Both relaxations keep the cells exchangeable, so the model says how many cells are in a state and not which. The matrices are constant within a season and depend on yesterday only, through the cell's own state and the day's regime. Standard errors are given for the counting estimates; for the DM parameters the spread is judged by the simulation bands. The constraint restores the level of the count, not every entry of the matrix, and the bootstrap rejects it for three of the four working models. Finally, the working models do not reproduce the lag-one autocorrelation of

the count in the two monsoons: statements about runs of region-wide heavy rain over several days are not supported by them, and in the NE monsoon only the MN models with five and six regimes reproduce persistence, mean and variance together. A model with regimes that carries both the persistence between days and the co-movement within a day, and a cell-specific version that says which cells are wet, are the natural next steps.

Statements and Declarations

- **ORCID iD** Muhammed Rais P K: https://orcid.org/0009-0008-6722-6641.

- **Funding** No funds, grants, or other support was received.

- **Competing interests** The authors have no relevant financial or non-financial interests to disclose.
- **Ethics approval and consent to participate** Not applicable.
- **Consent for publication** Not applicable.
- **Data availability** The IMD 0.25° gridded daily rainfall is distributed by the India Meteorological Department, Pune (https://imdpune.gov.in/cmpg/Griddata/Rainfall_25_NetCDF.html). The extracted record of the 57 cells and the derived daily states and counts are available from the corresponding author on reasonable request.
- **Materials availability** Not applicable.
- **Code availability** The code used for the computations and the checks is available from the corresponding author on reasonable request.
- **Author contributions** Tirupathi Rao Padi: conceptualization, methodology, supervision, writing (review and editing). Muhammed Rais P K: methodology, data curation, formal analysis, software, writing (original draft). Both authors read and approved the final manuscript.

References

Becker NG (1970) A stochastic model for two interacting populations. Journal of Applied Probability 7(3):544–564. https://doi.org/10.2307/3211937

Efron B (1979) Bootstrap methods: another look at the jackknife. The Annals of Statistics 7(1):1–26. https://doi.org/10.1214/aos/1176344552

Greene AM, Robertson AW, Smyth P, et al (2011) Downscaling projections of Indian monsoon rainfall using a non-homogeneous hidden Markov model. Quarterly Journal of the Royal Meteorological Society 137(655):347–359. https://doi.org/10.1002/qj.788

Heaps SE, Boys RJ, Farrow M (2015) Bayesian modelling of rainfall data by using non-homogeneous hidden Markov models and latent Gaussian variables. Journal of the Royal Statistical Society: Series C (Applied Statistics) 64(3):543–568. https://doi.org/10.1111/rssc.12094

Hughes JP, Guttorp P, Charles SP (1999) A non-homogeneous hidden Markov model for precipitation occurrence. Journal of the Royal Statistical Society: Series C (Applied Statistics) 48(1):15–30. https://doi.org/10.1111/1467-9876.00136

Katz RW, Parlange MB (1998) Overdispersion phenomenon in stochastic modeling of precipitation. Journal of Climate 11(4):591–601. https://doi.org/10.1175/1520-0442(1998)011⟨0591:OPISMO⟩2.0.CO;2

Massey WA, Whitt W (1993) Networks of infinite-server queues with nonstationary Poisson input. Queueing Systems 13(1–3):183–250. https://doi.org/10.1007/BF01158933

Matis JH, Hartley HO (1971) Stochastic compartmental analysis: model and least squares estimation from time series data. Biometrics 27(1):77–102. https://doi.org/10.2307/2528929

Matis JH, Kiffe TR (2000) Stochastic Population Models: A Compartmental Perspective, Lecture Notes in Statistics, vol 145. Springer, New York, https://doi.org/10.1007/978-1-4612-1244-7

Matis JH, Zheng Q, Kiffe TR (1995) Describing the spread of biological populations using stochastic compartmental models with births. Mathematical Biosciences 126(2):215–247. https://doi.org/10.1016/0025-5564(94)00038-2

Mohapatra M, Chauhan A, Varshney A, et al (2023) Short to medium range impact based forecasting of heavy rainfall in India. Mausam 74(2):311–344. https://doi.org/10.54302/mausam.v74i2.6180

Mosimann JE (1962) On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. Biometrika 49(1–2):65–82. https://doi.org/10.1093/biomet/49.1-2.65

Pai DS, Sridhar L, Rajeevan M, et al (2014) Development of a new high spatial resolution ($0.25^\circ \times 0.25^\circ$) long period (1901–2010) daily gridded rainfall data set over India and its comparison with existing data sets over the region. Mausam 65(1):1–18. https://doi.org/10.54302/mausam.v65i1.851

Pal I, Robertson AW, Lall U, et al (2015) Modeling winter rainfall in Northwest India using a hidden Markov model: understanding occurrence of different states and their dynamical connections. Climate Dynamics 44:1003–1015. https://doi.org/10.1007/s00382-014-2178-5

Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2):257–286. https://doi.org/10.1109/5.18626

Robertson AW, Kirshner S, Smyth P (2004) Downscaling of daily rainfall occurrence over Northeast Brazil using a hidden Markov model. Journal of Climate 17(22):4407–4424. https://doi.org/10.1175/JCLI-3216.1

Sanap SD, Priya P, Sawaisarje GK, et al (2019) Heavy rainfall events over southeast peninsular India during northeast monsoon: role of El Niño and easterly wave activity. International Journal of Climatology 39. https://doi.org/10.1002/joc.5926

Saravanan S, Abijith D, Kulithalai Shiyam Sundar P, et al (2023) Multi-criterion analysis of cyclone risk along the coast of Tamil Nadu, India – a geospatial approach. ISPRS International Journal of Geo-Information 12(8):341. https://doi.org/10.3390/ijgi12080341

Schwarz G (1978) Estimating the dimension of a model. The Annals of Statistics 6(2):461–464. https://doi.org/10.1214/aos/1176344136

Shahi NK, Rai S (2023) An increase in widespread extreme precipitation events during the northeast monsoon season over south peninsular India. Scientific Reports 13:22757. https://doi.org/10.1038/s41598-023-50324-9

Thompson CS, Thomson PJ, Zheng X (2007) Fitting a multisite daily rainfall model to New Zealand data. Journal of Hydrology 340(1–2):25–39. https://doi.org/10.1016/j.jhydrol.2007.03.020

Zucchini W, Guttorp P (1991) A hidden Markov model for space–time precipitation. Water Resources Research 27(8):1917–1923. https://doi.org/10.1029/91WR01403