RESEARCH ARTICLE

Selecting Markov chain orders for generating daily precipitation series across different Köppen climate regimes

Sarah Wilson Kemsley[1] | Timothy J. Osborn[1] | Stephen R. Dorling[2] |
Craig Wallace[3] | Joanne Parker[4]

[1]Climatic Research Unit, School of Environmental Sciences, University of East Anglia, Norwich, UK

[2]School of Environmental Sciences, University of East Anglia, Norwich, UK

[3]EarthSystemData Ltd, The Enterprise Centre, Norwich Research Park, Norwich, UK

[4]Atkins Global, Western House (Block C), Peterborough Business Park, Peterborough, UK

Correspondence
Sarah Wilson Kemsley, Climatic Research Unit, School of Environmental Sciences, University of East Anglia, Norwich NR4 7TJ, UK.
Email: s.wilson-kemsley@uea.ac.uk

Funding information
Atkins Global; NERC, EnvEast Doctoral Training Partnership, Grant/Award Number: NE/S007334/1

Abstract

Markov chain models are a commonly used statistical technique to generate realistic sequences of precipitation, but the choice of model order can strongly affect their performance. Although it is widely accepted that a first-order Markov chain reproduces precipitation occurrence in temperate latitudes quite well, it is also well known that first-order models have several shortcomings. These include a limited memory of rare events and inaccurately reproducing the distribution of dry-spell lengths, and their performance outside of temperate regions is less well understood. We present, therefore, the first assessment of model-order optimization which is both global in extent and which uses four evaluation methods: the Bayesian information criterion (BIC) and each model-order's ability to reproduce wet- and dry-spell lengths, and the interannual variability of precipitation occurrence. As well as a global analysis, we also assessed Markov chain performance and model-order selection separately within five climate regimes based on the Köppen classification system: tropical, dry, temperate, continental and polar. These metrics were used to determine the best performing model-order to generate realistic time series of precipitation across the five different climate regimes. We find that the choice of model order is most sensitive to the performance metric and less dependent on the climate regime. Across all regimes, we show that a first-order model performs best when evaluated with BIC and for generating realistic wet-spell distributions across all climate regimes except tropical, where third order performs best. We also find that a third-order model reproduces observed dry-spell distributions the best and second order commonly reproduces the interannual variability of precipitation occurrence across all regimes except tropical, where third order once again performs best. Our findings highlight the benefits of a flexible and tailored approach to the choice of Markov chain order for constructing precipitation series.

KEYWORDS
global < 2.scale, Markov chain, precipitation, stochastic weather generator

1 | INTRODUCTION

Stochastic weather generators are a technique used to produce synthetic rainfall time series with high spatial and temporal resolutions. Computationally inexpensive tools, they can be used to produce long time series for use in hydrological and agricultural risk assessments when the record length or quality of representative observational data are inadequate. For example, measurement time series are often too short to robustly estimate the probability of extreme events, such as long wet- or dry-spells. Stochastic weather generators were initially developed to address these issues (Semenov et al., 1998), though they have since been applied to a broader range of problems (e.g., perturbing their parameters so they produce synthetic time series under future climates (Eames et al., 2011)) and to other climate variables.

Precipitation is one of the most important variables for assessing risks affecting crop growth and the hydrological cycle. Precipitation is also often considered as the primary variable when stochastically modelling other weather variables (e.g., solar radiation, maximum and minimum temperatures) which can be conditioned on the precipitation status (Richardson, 1981). It is important that impact and risk assessors have access to the most accurate high-resolution models (Dubrovský, 1997), as generated data is often used in place of insufficient observed records as an input to hydrological, ecological, and agronomic studies (Larsen and Pense, 1982). Therefore, it is particularly important to ensure accurate modelling of daily precipitation.

Markov chains are a commonly used stochastic approach to modelling daily precipitation. Simulated occurrence of rain is conditional on the previous day(s)' precipitation status. The order of a Markov chain model refers to the number of previous days considered, that is, a first-order model conditions precipitation status on the status of one previous day. Several of the widely used Markov-type weather generators (e.g., WGEN (Richardson and Wright, 1984), SIMMETEO (Soltani and Hoogenboom, 2003a), and AAFC WG (Qian et al., 2005)) use the same model order regardless of geographical location. Furthermore, many of these weather generators have been designed, implemented and tested for climates local to where they were produced, meaning they may not be optimal at generating realistic rainfall time series in different climates (Semenov et al., 1998). Despite the common use of first-order models in climate and hydrological impact studies, their limited memory of extremes has been criticized (Semenov and Barrow, 1997), and though it is accepted that different sites require different orders, first-orders remain prevalent (Lennartsson et al., 2008).

However, the stochastic dependence on previous days' precipitation is dependent on particular meteorological drivers (Chin, 1977), with first-order models commonly misrepresenting important meteorological properties (Ailliot et al., 2015).

Although Markov-chain models are the focus of this study, there are several other methods used to stochastically model precipitation occurrence, including Markov renewal processes (MRPs) (Foufoula-Georgiou, 1987) and series-based approaches (e.g., LARS-WG (Qian et al., 2005)). Both methods use observed wet or dry-spell lengths to generate synthetic precipitation time series differently to Markov-chain models. In MRP models, the probability of precipitation depends on the number of days since the last precipitation event, whereas series-based weather generators draw lengths of wet and dry spells from semi-empirical distributions (Semenov et al., 1998).

Akaike (Tong, 1975) and Bayesian information criteria (BIC; Schwarz, 1978) are commonly used to select model order, balancing the model fit with the number of parameters. There have been several local or regional studies which have assessed the best model order, including across the United States (Schoof and Pryor, 2008), Costa Rica (Harrison and Waylen, 2000), England (Gates and Tong, 1976), Sweden (Lennartsson et al., 2008), Canada, Israel, India and Nigeria (Jimoh and Webster, 1996) using information criteria and other methods, including spell length analysis (Figure 1). Although many of these studies reiterate that a first-order model is adequate, seasonal and spatial variations in model-order choice were identified in all of them. Further studies critique this first-order assumption (Gates and Tong, 1976) and the use of an information criterion as the sole method of model-order selection (Hosseini et al., 2011).

Further evidence suggests that using a first-order model, regardless of location, may not always reproduce observed weather accurately. Low-order models are known to underestimate interannual and inter-seasonal precipitation variances (suffering from overdispersion (Katz and Parlange, 1998)), despite often being chosen by information criteria (Harrison and Waylen, 2000). Higher order models may more accurately reproduce variances in weather, though the risk of overfitting becomes more prominent.

A common use of precipitation generators is to estimate the probabilities of extreme events, such as long wet or dry spells, due to their ability to simulate long time series (Liao et al., 2004). Although information criteria are the most common methods used to assess model-order performance, they do not test the ability to accurately reproduce the distribution of wet- and dry-spell lengths, despite their importance for many applications. It is known (Lennartsson et al., 2008) that

FIGURE 1 Global locations of previous Markov chain model-order assessments used to inform this study [Colour figure can be viewed at wileyonlinelibrary.com]
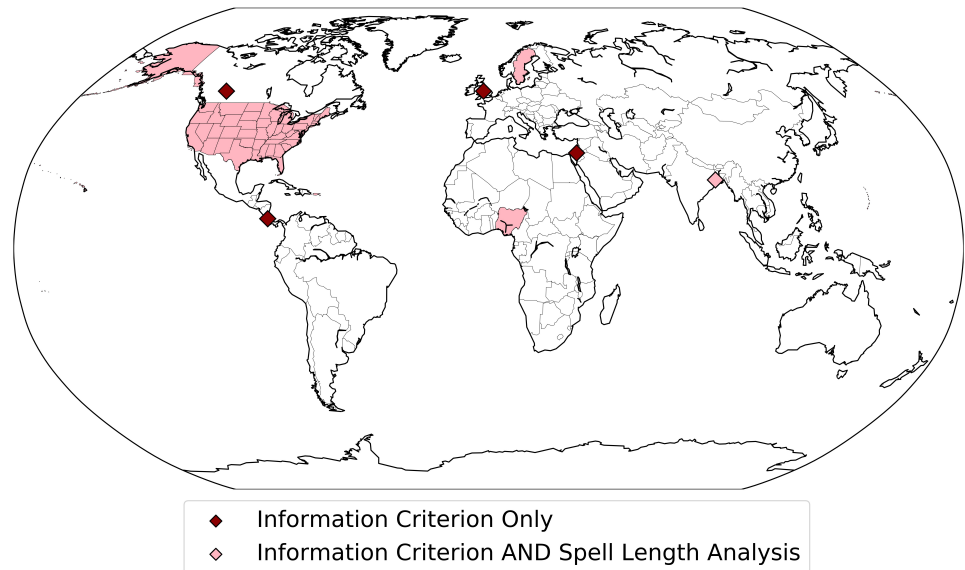
first-order models often do not reproduce dry spells accurately even though they are favoured by information criteria. This inability is an established issue with many Markov-type weather generators (Semenov and Barrow, 2002). Wet spells are important for hydrological modelling, impacting flood risk and soil erosion, while knowledge of dry-spell behaviour impacts agricultural and environmental planning, preparation for drought and irrigation infrastructure (Ochola and Kerkides, 2003). Therefore, a method to assess the ability of different model orders to reproduce observed wet- and dry-spell distributions is needed.

This study extends the existing knowledge base in several respects that are important for providing better guidance to users and developers of Markov chain weather generators. First, we undertake an assessment of preferential model order using *multiple* metrics of model performance. Alongside the BIC, we also quantify the ability of different model orders to reproduce observed distributions of wet- and dry-spell length, and the interannual variability (IAV) of precipitation occurrence. In each case, we consider the relative performance of models with orders 0, 1, 2 and 3. Higher orders are not considered in this study due to their increased risk of overfitting and greater computational demand, reducing usability. Second, we undertake a global assessment using more than 44,000 weather stations across most land areas. Third, using the global assessment, we consider whether the choice of model order is systematically dependent on the climate regime of each location.

Climate patterns can be classified into regimes according to several different characteristics, such as annual precipitation, interannual temperature and

precipitation variances, cloud cover and so forth (Belda *et al.*, 2014). Here, we use the Köppen climate classification which is one such classification in widespread use (Köppen, 1900), using monthly temperatures and precipitation to identify different climatic regimes across the world (Figure 2).

2 | DATA

We used daily precipitation data from the Global Historical Climatology Network Daily (GHCN Daily) to fit Markov chain properties and evaluate the models. GHCN-D quality control procedures flag potentially inaccurate or inconsistent records, estimated to affect approximately 0.3% of the data. Flagged data was removed, and only weather stations with at least 20 cumulative years of daily precipitation remaining (Soltani and Hoogenboom, 2003b) were used, resulting in a total of 44,071 stations. Historical records were capped at their most recent 30 years, to reduce any artefacts that may arise from using records of different lengths.

Each weather station was allocated a climate classification, determined by the centre of the Köppen grid cell which the longitude and latitude of the weather station was closest to. Locations are allocated to one of five overarching classes: tropical, dry, temperate, continental or polar. Within each class, regimes can be subcategorized further based on climatic behaviour. Our study applied each evaluation method to observed data in each overarching regime to determine the most appropriate model order for each climatic zone. Köppen classification data was taken from Chen and Chen (2013) on a $0.5° \times 0.5°$ grid.

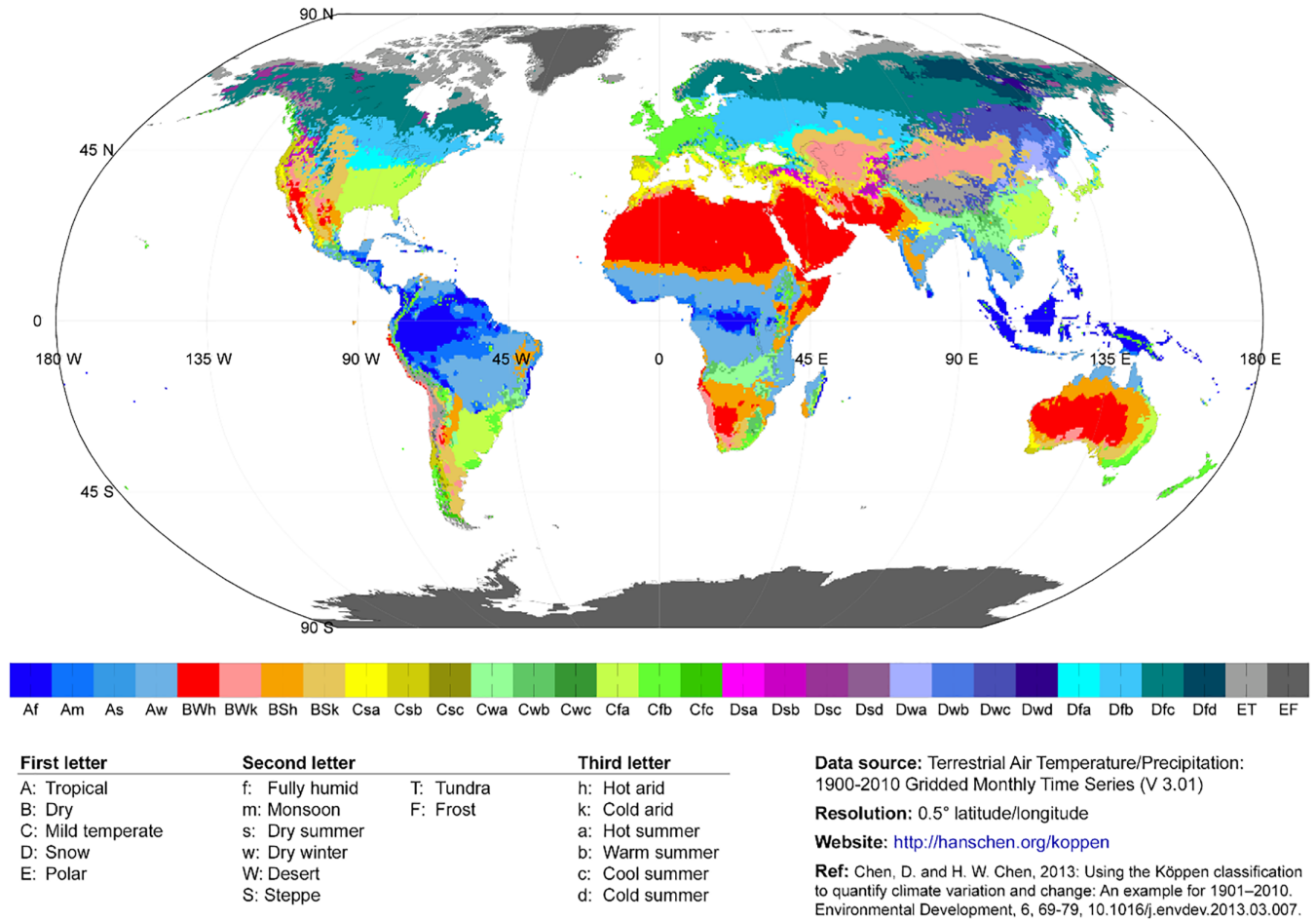## World map of Köppen climate classification for 1901–2010

First letter
A: Tropical
B: Dry
C: Mild temperate
D: Snow
E: Polar

Second letter
f: Fully humid
m: Monsoon
s: Dry summer
w: Dry winter
W: Desert
S: Steppe

T: Tundra
F: Frost

Third letter
h: Hot arid
k: Cold arid
a: Hot summer
b: Warm summer
c: Cool summer
d: Cold summer

Data source: Terrestrial Air Temperature/Precipitation: 1900-2010 Gridded Monthly Time Series (V 3.01)

Resolution: 0.5° latitude/longitude

Website: http://hanschen.org/koppen

Ref: Chen, D. and H. W. Chen, 2013: Using the Köppen classification to quantify climate variation and change: An example for 1901–2010. Environmental Development, 6, 69-79, 10.1016/j.envdev.2013.03.007.

FIGURE 2 The Köppen climate classification according to Chen and Chen (2013). There are some differences in terminology used for each class; in the current study, we denote class C as "temperate" and class D as "continental." Figure from Chen and Chen (2013) [Colour figure can be viewed at wileyonlinelibrary.com]
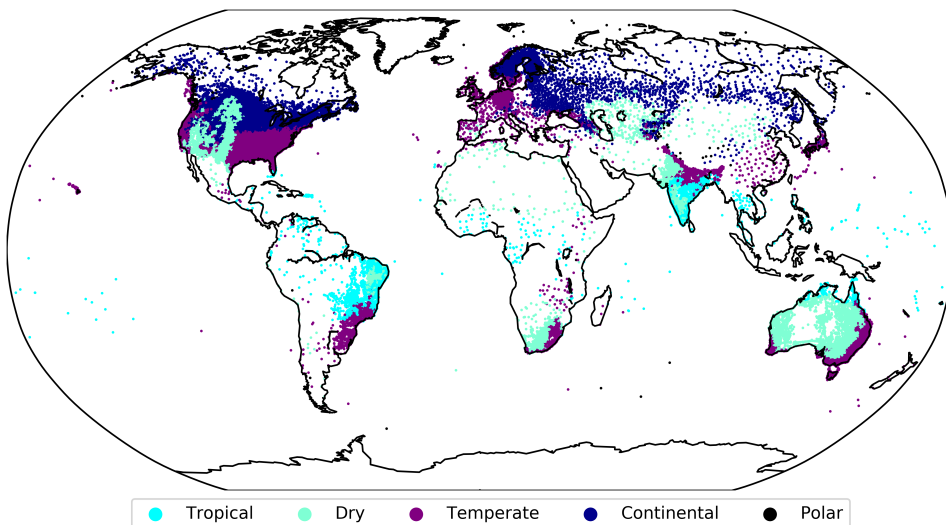

FIGURE 3 All available weather stations meeting the outlined criteria and their associated overarching climate classification [Colour figure can be viewed at wileyonlinelibrary.com]

It is clear from Figure 3 that there is not global coverage, and that some locations, such as the United States and most of Europe, have a significantly higher density of weather stations than others (such as Africa and South America). Nevertheless, even the least sampled class—polar—has over 400 stations with at least 20 years of observed data.

3 | METHODS

3.1 | Markov model fitting

A two-state Markov chain-gamma model was used to simulate the occurrence and amount of precipitation, with two states referring to a day being either wet or dry. A day was described as "wet" if precipitation was above 0.1 mm. This method of stochastic modelling is loosely based on the weather generator WGEN (Richardson and Wright, 1984). The probability of a wet day is conditional on the historical precipitation status. A kth-order Markov-chain refers to the number of conditional k previous days. For the widely used first-order, two-state model, the transition probabilities are:

$$p_{ij} = \frac{n_{ij}}{\sum\limits_{m=0}^{1} n_{im}} \qquad (1)$$

where i and j can represent wet (1) or dry (0) days. For example, p_{01} is the probability of a wet day following a dry day, and n_{01} the number of wet days following dry days (calculated from a historical dataset). This process can be extended to other orders:

$$p_i = \frac{n_i}{\sum\limits_{m=0}^{1} n_m}$$

$$p_{ijq} = \frac{n_{ijq}}{\sum\limits_{m=0}^{1} n_{ijm}}$$

$$p_{ijql} = \frac{n_{ijql}}{\sum\limits_{m=0}^{1} n_{ijqm}} \qquad (2)$$

for orders 0, 2 and 3, respectively. The number of transition probabilities calculated therefore increases exponentially and can be generalized to 2^{k+1} (where k is model order). Transition probabilities were calculated for each month at each weather station, resulting in $12(2^{k+1})$

transition probabilities for each station and each model order. However, the number of independent transition probabilities is half this number, that is, $12(2^k)$, because $p_1 = 1 - p_0$.

It is a common assumption that precipitation amount is conditionally independent of precipitation occurrence (Richardson, 1981). Upon generating a wet day, a random precipitation amount is independently taken from the corresponding month's two-parameter gamma distribution. For each month, shape (α) and scale (β) parameters were calculated from wet-day only data for each station. Thom estimators (Thom, 1958) were used for calculating the shape,

$$\alpha = \frac{1 + \left(1 + \frac{4D}{3}\right)^{\frac{1}{2}}}{4D} \qquad (3)$$

and the scale,

$$\beta = \frac{\bar{x}}{\alpha} \qquad (4)$$

parameters, with sample statistic

$$D = \ln(\bar{x}) - \frac{1}{n} \sum_{i=1}^{n} \ln(x_i) \qquad (5)$$

where x refers to amounts of precipitation and n the number of wet days. This method is used in place of moment estimators, which are considered "inefficient" compared with the Thom estimators that make better use of the information in a dataset (Wilks, 1995).

3.2 | Markov model evaluation

Information criteria are methods that can be used to determine how suitable different order Markov chains are for modelling precipitation occurrence, based on calculations of log-likelihood functions from the transition probabilities (Schoof and Pryor, 2008). Here, we chose the Bayesian information criterion (BIC) over the Akaike information criterion (AIC) because Katz (1981) showed that AIC estimators can be inconsistent, and also BIC is less prone to asymptotic bias and more widely used in recent literature (Harrison and Waylen, 2000; Schoof and Pryor, 2008). This method seeks to find the best model containing the fewest parameters (i.e., minimal k). For a two-state Markov chain model of order k, BIC values are given by
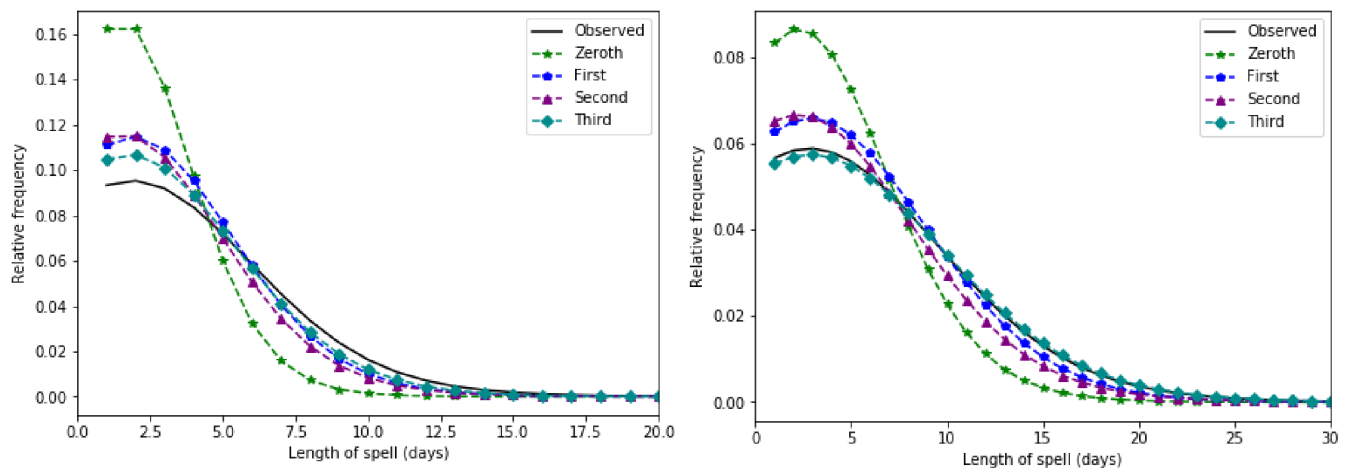
FIGURE 4 Probability density distribution of the length of dry spells (left) and wet spells (right) produced by different model orders over a 50-year period at a weather station in Highcroft, Tasmania (GHCN-D station ID ASN00094028) [Colour figure can be viewed at wileyonlinelibrary.com]

$$BIC(k) = -2L_k + 2^k[\ln(N)] \tag{6}$$

where L_k is the log-likelihood function and N is the number of days in the historical record used to calculate the transition probabilities. L_k is calculated from the estimated transition probabilities using the functions given by Schoof and Pryor (2008) for each model order. The model order that minimizes BIC is chosen as the best order.

As there are 12 sets of transition probabilities for each station, 12 BIC values for each model order were determined and compared enabling an evaluation of annual and seasonal model-order dependence. The mode of the model orders that minimize BIC across each of the 12 months was taken as the annual model order. Seasonal model-order choices were also studied, using the mode of the model orders that minimize BIC across June, July, August (JJA) and December, January, February (DJF).

In addition to BIC, three other assessment criteria are considered, providing additional information on model suitability that cannot be gained by BIC alone. Each order's ability to reproduce the distribution of wet and dry spell lengths was also compared. A wet spell was defined as a period of wet days preceded and followed by a dry day (and vice versa for a dry spell). For each model order, 50 years of precipitation data were generated and the length of every wet and dry spell in the generated data determined.

Kernel density estimation (KDE) can be used to determine a non-parametric probability density function of a random variable (Guidom, 2015). KDE was used here to estimate probability density distributions (Rajagopalan

TABLE 1 Number of weather stations and corresponding grid cells in each climate regime

Regime	No. stations	No. cells
Tropical	4,140	173
Dry	8,555	139
Temperate	20,389	192
Continental	10,552	229
Polar	435	104
TOTAL	44,071	837

et al., 1997) for wet- and dry-spell lengths (see Figure 4 for an example). Upon determining distributions for each model-order, the root-mean-squared-difference between the observed distribution and the distributions produced by each Markov chain was calculated. A bandwidth of 2 days was used for KDE and a test across the climate regimes indicated that the results are not sensitive to this choice. In addition to comparing the full spell length KDE distributions, we also evaluate the performance of each model order at reproducing four percentiles (75, 90, 95 and 99th) in the tail of the distributions. These percentiles are calculated directly from the underlying spell length data, rather than from the KDE distributions to avoid any dependence on the choice of KDE bandwidth.

Finally, each models' ability to reproduce the IAV of precipitation occurrence was tested. Once again, 50 years of daily precipitation occurrence was generated. The total number of wet days in each season or year was recorded, and the SD of theses 50 values (number of wet days per season/year over the 50-year period) was calculated. The model order producing the smallest absolute difference between the generated and observed SD was deemed to perform best.

TABLE 2 Gridded comparison of Markov model-order choices for each of the Köppen climate regimes based on each model order's ability to minimize BIC. Values shown are the % of grid cells within each climate regime where the mode of individual stations' best model-order is equal to 0, 1, 2 or 3

Regime	0			1			2			3		
	Annual	DJF	JJA	Annual	DJF	JJA	Annual	DJF	JJA	Annual	DJF	JJA
Tropical	11.8	17.4	18.5	56.2	52.2	52.8	30.3	27.5	24.7	1.7	2.8	3.9
Dry	18.2	16.1	24.5	63.6	67.8	55.9	17.5	15.4	19.6	0.7	0.7	0.0
Temp.	2.2	4.4	5.5	80.7	80.1	77.9	17.1	15.5	16.6	0.0	0.0	0.0
Cont.	0.4	0.0	0.4	99.2	99.2	98.7	0.4	0.8	0.8	0.0	0.0	0.0
Polar	4.1	3.1	4.1	91.8	84.5	86.6	3.1	11.3	8.2	1.0	1.0	1.0

4 | RESULTS AND DISCUSSION

The methods outlined in Section 3 were applied to 44,071 weather stations that met criteria detailed in Section 2. Results from these weather stations were aggregated into 837 5° latitude by 5° longitude cells (Table 1) to reduce biasing area-average results to locations with dense station coverage (e.g., United States, Figure 3). This also increased the weighting of locations with sparse data coverage, for example, across Africa and South America. The climate regime of a grid-cell was allocated by taking the modal classification from each individual station in the cell. This methodology was also applied to obtain an overall model-order for each grid-cell and each method of analysis used.

4.1 | Bayesian information criterion

The BIC was the first method used to analyse model-order performance. The BIC optimization most often selected a first-order model to minimize BIC across all regimes (Table 2), in agreement with previous literature (Schoof and Pryor, 2008) but extending this finding to more regions, regimes and seasons. Although a first-order model may be allocated overall to each regime, there are clear behavioural differences between the zones. For tropical and dry regimes, a sizable minority of cells show zeroth- or second-order dependence. A sizable minority of temperate cells (17.1%) are minimized by a second-order model; however, unlike tropical and dry regimes, there are very few cells minimized by zeroth-order model. Conversely, first-order models almost universally minimize BIC for continental and polar cells (99.2 and 91.8%, respectively).

These differences in behaviour may arise from spatial variation (Figure 5). The majority of the northern hemisphere extra-tropics are best represented by a first-order model. There is much more variation across the southern hemisphere and the tropics, where zeroth- and second-order models are much more prevalent. For example, Chilean and Peruvian cells do not show first-order dependence, but zeroth or second. Central African cells also show zeroth-order dependence, whereas Mainland Southeast Asia has many cells with second-order dependence.

Central Africa primarily has tropical humid climates (fully humid, monsoon or with dry winters: Af, Am and Aw, respectively) alongside hot dry steppe and desert climates to the north (BSh and BWh, respectively—Figure 2). Cells in tropical humid (Af) regions were found to show both first- and second-order dependence in equal measures (42.4%). Although dry desert (BWh) cells showed overall preference for first-order models (49.0%), a significant minority (32.7%) showed zero-order dependence. These differences in climatic zone account for some of the spatial variation noted in Figure 5. Cold arid dry steppe (BSk) cells are most often found in the northern hemisphere and show differences in behaviour to the overarching regime (dry). BIC was minimized by a first-order model in 91% of BSk cells—almost 30% more than the total dry group. The BSk cells without first-order dependence mostly occur in the South American steppe and instead show mostly second-order dependence. Sub-regimes across temperate, continental and polar regimes performed similarly to their aggregated behaviour.

Several geographical regions have no usable GHCN-D data as indicated by an absence of cells in their location. These regions include Northern and South-West Africa, Maritime Southeast Asia and the Middle East. These locations are often tropical and dry and are located around the equator or southern hemisphere. This is a limitation with using in situ observed data and has the potential to affect overall model-order preferences.

The same overall picture emerges when model orders with the best BIC are considered on a seasonal basis (Table 2), i.e., that first-order models are selected most often in all Köppen regimes but zeroth-order models are
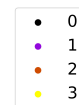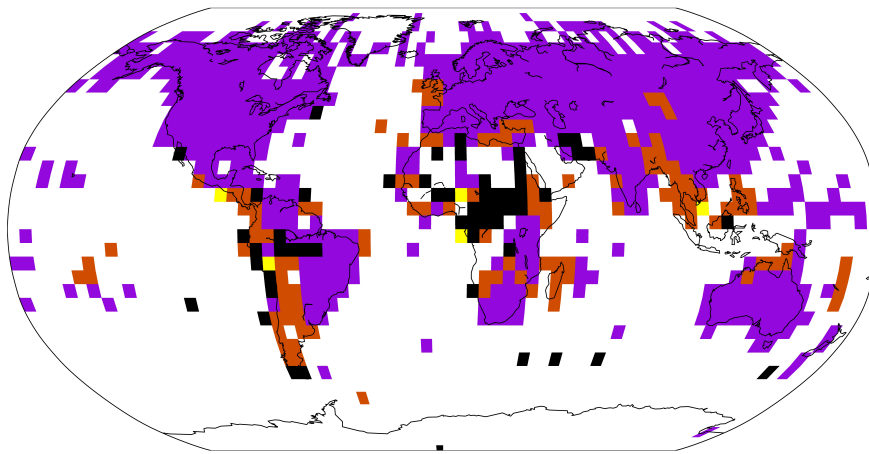
FIGURE 5 Model-order choices for each $5° \times 5°$ grid cell, with colour representing model-order choice as determined by Bayesian information criterion [Colour figure can be viewed at wileyonlinelibrary.com]
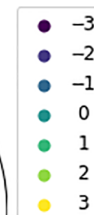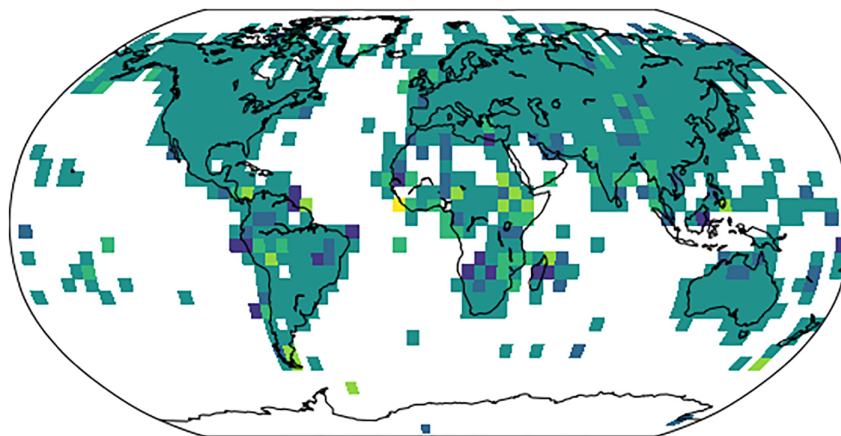

FIGURE 6 Model-order choice for JJA (BIC) minus the model-order choice for DJF (BIC) [Colour figure can be viewed at wileyonlinelibrary.com]

a notable minority in tropical and dry regimes and second-order models are a notable minority in tropical, dry and temperate regimes. Nevertheless, some seasonal differences are apparent and can be seen in the model orders for DJF subtracted from JJA (Figure 6). While the same model order is selected by BIC across most of the northern hemisphere in both seasons, it is once again cells in the tropics and southern hemisphere that experience the most variation. This variation is shown predominantly in tropical and dry regions

once again (Table 2), although these fluctuations are minor. Dry climates exhibit the most noticeable seasonal fluctuation, with first-order minimizing BIC in 67.8% cases in DJF but only 55.9% in JJA. This is reflected in the increased number of zeroth-order cells in JJA.

4.2 | Spell-length analysis

Although first-order models most commonly minimize BIC across all climate regimes, this is not always the case when other metrics are used to assess model-order performance.

Across all regimes except tropical, a first-order model most commonly outperforms others at reproducing the distribution of wet-spell lengths (Table 3). First-order models outperform the other models for wet-spell length at a large majority (89.9%) of continental regime grid cells, whereas sizeable minorities of grid cells are best represented by third-order models across dry, temperate and polar regimes (24.5, 33.7 and 26.8%, respectively). In the tropical regime, by contrast, third-order models most often perform the best, followed by the second- and then first-order models. Across all regimes, zero-order models rarely or never reproduce wet-spell length distributions best.

Though third-order models only outperform other orders for wet-spell lengths in tropical regimes, they are dominant as the best order for reproducing dry-spell length distributions (Table 3) across *all* regimes. This supports the work of Lennartsson *et al.* (2008), who found that higher order models better reproduce the distribution of very long dry spells in Sweden. There is less spatial variation of optimal model order for dry spells than with BIC and wet spells, reiterated by the high percentage (at least 78% across each regime) of cells being best represented by a third-order Markov chain (Table 3).

TABLE 3 Gridded comparison of Markov model-order choices for each of the Köppen climate regimes based on each model's ability to reproduce the IAV of precipitation occurrence (labelled IAV) and distributions of wet-spell length and dry-spell length. Values shown are the % of grid cells within each climate regime where the mode of individual stations' best model-order is equal to 0, 1, 2 or 3

	0			1			2			3		
Regime	Wet	Dry	IAV	Wet	Dry	IAV	Wet	Dry	IAV	Wet	Dry	IAV
Tropical	5.1	0.0	1.7	21.3	6.3	7.9	25.8	8.6	41.0	47.8	85.1	49.4
Dry	0.0	1.4	0.0	67.8	9.2	11.2	7.7	10.6	62.9	24.5	78.7	25.2
Temp.	1.1	1.1	0.0	58.6	6.7	13.3	6.6	3.9	58.6	33.7	88.2	28.2
Cont.	0.0	0.4	0.0	89.9	5.2	25.6	0.0	0.0	52.9	10.1	94.4	21.4
Polar	0.0	2.2	1.0	69.1	4.4	17.5	4.1	1.1	54.6	26.8	92.2	26.8

TABLE 4 The gridded median percentage difference between observed and generated dry-spell length percentiles for each of the Köppen climate regimes and each model order

		Percentile			
	Model order	75th	90th	95th	99th
Tropical	0	−6.5	−20.0	−25.3	−38.5
	1	14.5	2.4	−4.0	−19.9
	2	0.0	0.0	−4.6	−15.2
	3	3.8	8.3	1.2	−8.0
Dry	0	−20.5	−29.3	−34.2	−38.3
	1	5.8	−5.8	−13.1	−24.0
	2	−15.3	−16.3	−20.1	−23.6
	3	5.0	−2.8	−9.3	−17.7
Temperate	0	−26.4	−32.4	−36.3	−39.3
	1	4.8	−1.6	−6.7	−16.2
	2	−25.0	−18.1	−16.5	−18.4
	3	3.5	0.3	−0.8	−6.0
Continental	0	−25.0	−32.5	−35.4	−39.9
	1	3.0	−3.4	−7.9	−13.7
	2	−25.0	−17.9	−16.7	−17.0
	3	3.0	0.4	−1.0	−5.4
Polar	0	−25.0	−33.3	−38.5	−46.7
	1	4.1	−2.1	−8.2	−18.2
	2	−25.0	−16.7	−18.0	−20.7
	3	0.0	0.0	0.0	−7.7

For tropical and dry climates, there is a monotonic increase in preference for a model as the model's order increases (hence with the lowest percentages for zeroth-order models). For the other climate regimes, more than 88% of grid cells have third-order models as the best performing for dry spells, and first-order models are the next most frequent.

The ability of each model order to reproduce percentiles of the dry-spell length distributions was also studied. Percentage differences between the observed and generated dry-spell length 75th, 90th, 95th and 99th percentiles were calculated for each station, with the median percentage difference across grid cells given in Table 4. Medians were taken across regimes as opposed to means to reduce the influence of outliers.

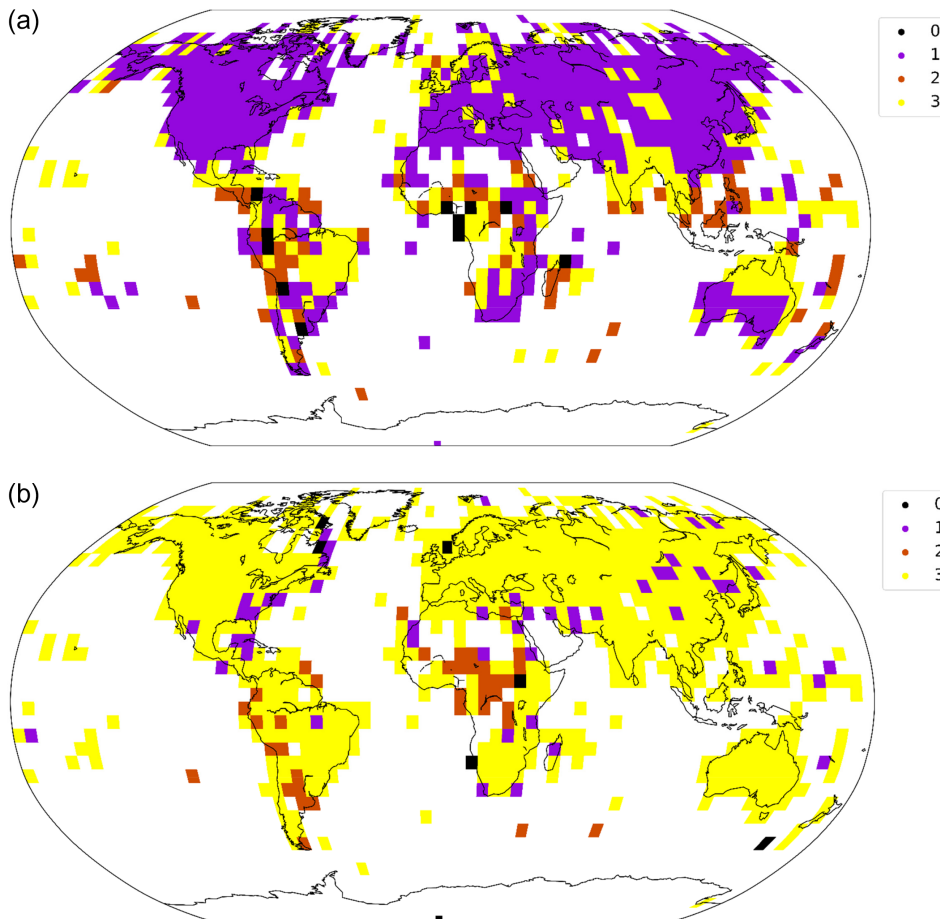While each model order underestimates the 99th percentile of observed dry spells, there is a large

improvement from zeroth order to third. Second- and first-order models tend to underestimate by a similar percentage, with third orders having as little as a 5.4% underestimation in continental regimes. This supports the work of Lennartsson *et al.* (2008), suggesting that lower order models do not accurately reproduce the distribution of very long dry spells. The largest median bias in these extreme (99th percentile) dry spells occurs in the dry Köppen regime, but even here the bias is much less when considering something slightly less extreme (e.g., 95th percentile) providing that a third-order model is used. In all regimes, there is again a large improvement in reproducing the 95th percentile from zeroth to third order, though in this case second-order models tend to have larger median biases than first order. This is particularly noticeable in temperate, continental, and polar regimes. Third-order models continue to have the smallest percentage differences at the 75th and 90th percentiles except for the tropical regime where the second-order model performs best.

The dominant patterns in the results are: (a) at the 75th percentile, first- and third-order models slightly overestimate dry-spell lengths whereas the other two orders underestimate them; (b) the median bias in dry-

spell lengths mostly decreases monotonically as the percentile increases, ending in the underestimation of the 99th percentile spell lengths by all model orders and (c) with few exceptions the biases are smallest for the third-order model, though the first-order model performs nearly as well for the 75th and 90th percentiles in most regimes.

Spatial variations are once again important for model-order performance at reproducing both wet- and dry-spell length distributions (Figure 7), though not as prominently as seen earlier with BIC. The northern hemisphere extratropics are again almost universally best described by the same model-order (first for wet spells, third for dry spells). However, unlike BIC and dry spells, there is notable variation across Europe in model-order performance for wet spells (where third order appears prominent in the northwest, but first order elsewhere). This variation in model order across Europe contributes to third-order models outperforming other orders across a sizable minority of stations in temperate regimes (33.7%). However, it is across the tropics and southern hemisphere where most variation is present, primarily in tropical and dry regimes. Despite third- and first- order models

FIGURE 8   Model-order choice
for dry spells minus the model-order
choice for wet spells [Colour figure
can be viewed at
wileyonlinelibrary.com]
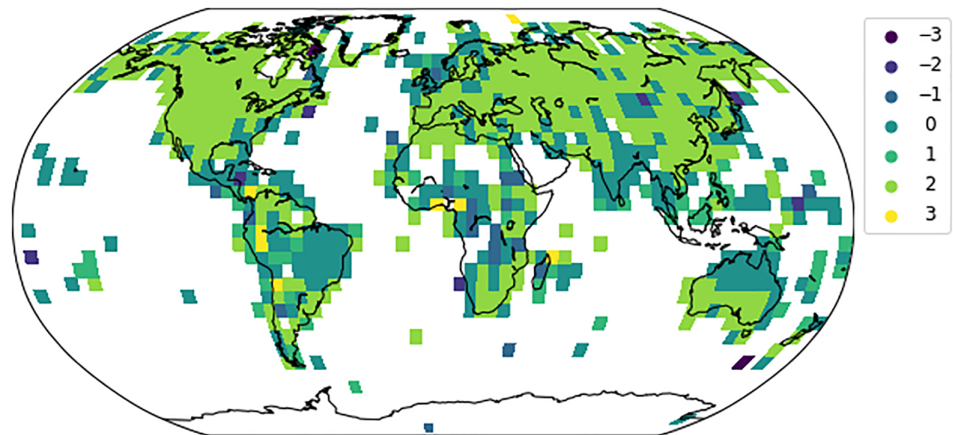
representing wet spells best in tropical and dry regimes, respectively, a majority of tropical grid cells are represented best by an order other than third and a sizable minority (32.2%) of dry regime grid cells are represented by a model order other than first. While third-order models almost entirely outperform others across Brazil, Northern Australia and India, there is much more variation across Central Africa, despite being in the same regime (tropical).

The wet-spell distribution across the Aw (tropical with dry winters) regime (Figure 2), present across much of Brazil and Central Africa, is reproduced best by a third-order model in 61.1% of cells. This is higher than the overall 47.8% of tropical grid cells represented by a third-order model. Dry, hot-arid steppe (BSh) regions, present to the north of Central Africa, and across Australia and India, also often show third-order dependence (51.4%) despite first-order models most commonly outperforming others in dry regime grid cells. Sub-regimes for temperate, continental and polar climates follow similar patterns to their aggregated classifications for both wet- and dry-spell distributions.

For both wet- and dry-spell distributions, there is notably less variation in model-order performance across continental regimes than other classifications. It is important to note that continental climates (category D, termed "Snow," in Figure 2) are almost exclusively present in the northern hemisphere extra-tropics, with each other regime present across the tropics and in the southern hemisphere. This potentially explains the greater variation in the other regimes.

The difference between model-order performance at reproducing wet-spell and dry-spell distributions was studied (Figure 8) by subtracting the model-order choice for wet-spell distributions from the choice for dry-spell distributions. For example, should a grid cell be best represented by a third-order model for dry spells and a first-order model for wet spells, the difference is +2. A

difference of +2 is widespread across most of the northern hemisphere. This is as expected, with first-order models commonly outperforming all others for wet spells and third order for dry spells. It is important to note that for many locations, including Brazil, Central America, North Australia, India, Europe and Central Africa, the same model-order performs best for both wet and dry spells (a difference of 0, Figure 8).

4.3 | Interannual variability of precipitation occurrence

In this section, we evaluate the performance of the models at generating year-to-year variability in the number of wet days (i.e., precipitation occurrence). In all regimes, second-order models most commonly reproduce the IAV of annual precipitation occurrence best, with the exception of tropical regime cells, where third order most commonly outperforms others (Table 3). In all regimes (except tropical), the IAV of precipitation occurrence in at least 50% of cells is represented best by a second-order model. Although third-order models perform best in tropical regimes (49.4%), the number of cells represented best by a second-order models follows closely (41%). As with BIC and spell-length analysis, zeroth-order models rarely (or never) outperform others.

There are some notable differences in model-order performance at reproducing the IAV of precipitation occurrence for individual seasons across the regimes. Much like the IAV of annual occurrence, second-order models most commonly outperform others, with zeroth order rarely performing best (though with larger minorities than the annual occurrence). For all seasons in tropical regimes, third-order models consistently outperform others. While a small minority of cells are best represented by a first-order model annually (25.6% maximum, found in continental regimes), in JJA, first-order models
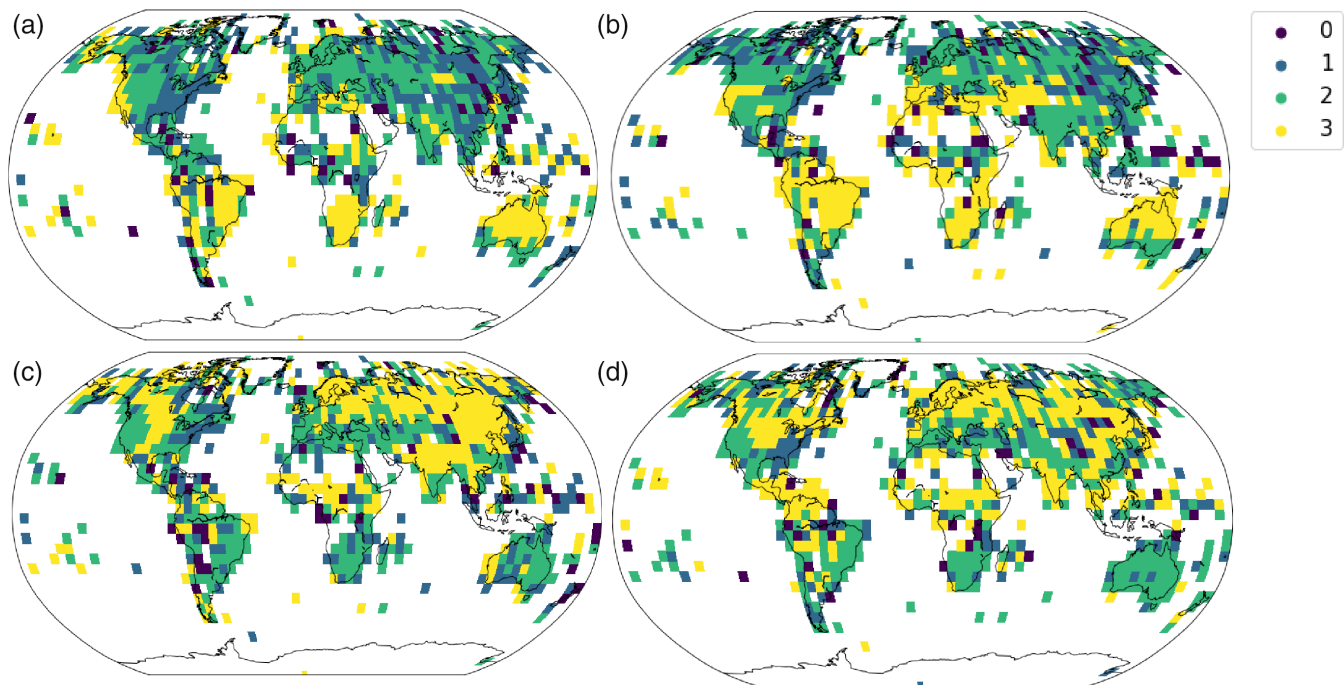
FIGURE 9 Model-order choices for each 5° × 5° grid cell, with colour representing model-order choice as determined by interannual variability of seasonal precipitation occurrence for (a) March, April, May (MAM), (b) June, July, August (JJA), (c) September, October, November (SON) and (d) December, January, February (DJF) [Colour figure can be viewed at wileyonlinelibrary.com]

outperform third-order models in both continental and polar regimes (by 27 and 34.1%, respectively).

There is less coherence in the spatial pattern of model-order choice for reproducing the IAV of annual precipitation occurrence (not shown here) compared with BIC and spell-length analysis (Figures 5 and 7). Nevertheless, the individual seasons do have coherent patterns in the best-performing model order (Figure 9), but these patterns are somewhat opposite between seasons resulting in some cancellation of coherent patterns for the annual results. There is a stark contrast in the performance of third-order models at reproducing the IAV of seasonal occurrence between DJF and JJA across the continental regime. In DJF, third-order models outperform others in 54.4% of continental cells, whereas in JJA this drops to only 6.8%. This is reflected in Figure 9. In JJA and MAM, much of the northern hemisphere is represented best by second- and first-order models, whereas in DJF and SON, many of these cells in are instead represented by a third-order model. These cells are mostly continental and temperate. The reverse is true for South America, where third orders are prevalent in JJA and MAM and second orders in DJF and SON. In JJA, third-order models outperform all others in both tropical and dry regimes (48.3 and 47.1%, respectively), with a sizable minority of second-order cells. However, in DJF, third-order models perform best in only 27.5% of cells, with an increased percentage of second-order cells (49.3%).

TABLE 5 Model orders for each metric with highest percentage of grid cells reproducing the observed data the best

	BIC	Wet spell	Dry spell	IAV
Tropical	1	3	3	3
Dry	1	1	3	2
Temperate	1	1	3	2
Continental	1	1	3	2
Polar	1	1	3	2

Further seasonal differences can be noted across Australia, India and southern Africa.

It is widely known that a common limitation of Markov chain models is in their underestimation of the observed IAV (Wilks, 2010). Here, we find that all model orders tend to underestimate the observed variability, with zeroth-order models consistently underestimating more than higher order models. For each model order, this underestimation is similar across all regimes and tends to decrease as model order increases. However, the improvement stops at second order, and here we find little difference between second and third orders. This agrees with previous findings that lower order models may be more prone to overdispersion (Katz and Parlange, 1998). The IAV of precipitation occurrence tended to be underestimated by a greater percentage by

all model orders in locations that had higher observed variability (China, South America, Pacific Islands) and less so in areas with lower observed variability (Australia, India, West United States).

5 | SUMMARY

It is apparent that although each metric highlights different behaviour across the regimes, with stark variation between tropical/dry and continental regimes, the model order with the highest percentage of grid cells is the same across all regimes except tropical.

The model order with the highest percentage of grid cells in each regime can be said to perform best (Table 5). BIC and dry-spell length analysis each favour a single model order globally, despite behavioural differences in each regime. For wet spells, each regime favours a first-order model, except tropical prescribing third order. The BIC was satisfied by a first-order model across all regimes. A third-order model performs best at reproducing dry spells across all climate regimes (especially for the higher percentiles of the dry-spell length distribution), while second order reproduces the IAV of precipitation occurrence best in all regimes except tropical (once again prescribing third order).

Spatial variation was noted using each method of analysis. Much of the northern hemisphere contains cells with a high percentage following the overall metric's model order prescription, regardless of their climate classification. It is across the tropics and southern hemisphere (South America, Central Africa and Mainland Southeast Asia) that shows the most variation from the model order for each method of analysis. These locations are often categorized into tropical and dry climates.

There are several other points of note. While a third-order model best reproduces observed dry-spell distributions, first-order models are chosen as second best across 76.9% of cells. At the 95th and 99th percentile, while third-order models underestimate the observed percentiles the least, first order performs similarly at the lower percentiles. Third- and first-order models perform second best at reproducing wet-spell distributions across 40.0 and 54.4% of cells, respectively. Third-order models also require the calculation of eight independent transition probabilities (per month) while a first-order model calculates only two, thus the higher order model requires more computing power and, most importantly, increases the possibility of over-fitting a model. Although second-order models most commonly outperform others at reproducing the IAV of precipitation occurrence, first- and third-order models represent a sizable minority of stations in all regimes. It is important to note that while

using a higher order model reduces the underestimation of IAV, all orders underestimate it, with little difference between second- and third-order models.

There is a noticeable dip in how frequently second-order models perform best at reproducing wet- and dry-spell distributions (Table 3). Here, unlike BIC (Table 2), model-order performance is not penalized for the number of parameters used. A potential cause of this dip could be that the percentage of third-order models is boosted by also being representative of the performance of model orders higher than three, which are not considered here. For example, when reproducing wet-spell distributions, there may be decaying performance across higher model orders which, in this study, might be encapsulated by the percentage of third-order grid cells.

However, there are limitations due to data coverage across much of northern and south west Africa, the Middle East, Indonesia and the Philippines (Figures 5–9). Much of this domain falls into tropical and dry regimes, potentially impacting the overall model-order choice of these climate types. Aggregating data into grid cells also loses some of the detail from sub-categorizing the weather stations. In this analysis, continental regimes are almost completely represented by Dfb and Dfc climates (i.e., those with continental humid climates; Figure 2), with 78.6% of continental cells falling into one of these sub-categories. There is no representation of Dsa or Dsd (continental with dry summers) climates using the gridded approach at 5° resolution. Thus, the overall model-order choice for continental climates may not be representative of all sub-categories within the regime. A similar issue is noted across temperate and polar regimes. Only one grid-cell represents Csc and Cwc cells (climates with temperate, cool summers with dry summers and winters, respectively), while over half of the temperate cells represent Cfa and Cfc (temperate, fully humid climates).

6 | CONCLUSION

The ability of four Markov chain weather generator models to generate realistic daily precipitation time series was assessed by four different methods across 44,071 weather stations globally. Each weather station was allocated a regime based on Köppen's climate classification system (Chen and Chen, 2013): tropical, dry, temperate, continental or polar. To extend previous results, the performance of the different order models was assessed using four metrics: the widely used BIC, and the less commonly considered abilities to reproduce observed wet- and dry-spell lengths and the IAV in precipitation occurrence.

Analysis was undertaken for each weather station, and then performance was aggregated on a grid-cell basis to avoid the assessment being dominated by densely observed regions. Model performance measures were aggregated into 837 $5° \times 5°$ cells based on the longitudes and latitudes of the weather stations. Depending on which metric was used, different model orders performed the best. First-order models most frequently minimized BIC and best reproduced observed wet-spell distributions across each regime except tropical. While this agrees with other preceding local or regional studies, including across the United States (Schoof and Pryor, 2008), Nigeria (Jimoh and Webster, 1996) and Costa Rica (Harrison and Waylen, 2000), this finding has now been demonstrated on a near-global scale, and included a range of different climate regimes. Unlike other regimes, third-order models best reproduced observed wet-spell distributions across tropical regime locations. Third-order models also best reproduced dry-spell distributions across all regimes, strengthening previous evidence that a low-order model may not accurately reproduce extreme dry periods (Lennartsson et al., 2008). This is strengthened further by the ability of third-order models to reproduce most successfully the observed 99th percentile (and, in most cases, the 95th percentile) of dry-spell length distributions. Second-order models most commonly reproduced the IAV in precipitation occurrence best in each regime, once again with the exception of tropical regimes where third orders perform best.

Although the most frequently selected model order remained the same across different climate regimes for each metric (except for tropical regimes), interesting variations within and between regimes were found. Tropical and dry regimes showed the most deviation from overall behaviour. Third-order models reproduced both wet- and dry-spell distributions best across Brazil, India, North Australia and Europe, regardless of the climatic region. Central Africa, Mainland Southeast Asia and western South America also showed notable geographical variation of model-order performance for each metric. The behaviour of climate classification sub-categories often reflected the aggregated performance but there were some exceptions primarily across dry and tropical zones. However, many sub-categories were not present or were only represented by a few grid cells in this study, resulting in temperate and continental regimes being more representative of some sub-categories than others. This leads to the notion that analysis of model order selection may be more beneficial on a finer spatial scale, that is, sub-continental, or with greater inclusion of other sub-categories. This may better reflect the noted spatial variance.

This study will inform future work on developing a global stochastic weather generator to synthesize daily time series of a suite of climatic variables, including maximum daily temperature and cloud cover. There are several ways in which these results can be applied to industry and future research. As weather variables produced by a stochastic generator are frequently used as input into hydrological and agricultural models, our results show that it would be beneficial to use different model orders to generate precipitation data depending on the purpose of the model. For example, studies focussing on extended dry periods may favour third-order models due to their superior ability to reproduce the upper tail of the dry-spell distribution, whereas a first-order model acts as a good, computationally efficient "all-rounder" for other studies. Tropical climates are an exception, where third-order models reproduce distributions of both wet and dry spells and the IAV of precipitation occurrence best overall. This study can also inform users on the most appropriate model-order to use based on the location of a specific region. For example, generating realistic wet-spell distributions across Europe requires different model orders to those favoured in the eastern United States, despite both regions falling into the temperate classification (Figures 2 and 7a).

ORCID

Sarah Wilson Kemsley https://orcid.org/0000-0002-7652-5060

REFERENCES

Ailliot, P., Allard, D., Monbet, V. and Naveau, P. (2015) Stochastic weather generators: an overview of weather type models titre: Générateurs stochastiques de condition météorologiques: Une revue des modèles à type de temps. *Journal de la Société Française de Statistique*, 156(1), 101–113.

Belda, M., Holtanová, E., Halenka, T. and Kalvová, J. (2014) Climate classification revisted: from Köppen to Trewartha. *Climate Research*, 59(1), 1–13.

Chen, D. and Chen, H.W. (2013) Using the Köppen classification to quantify climate variation and change: an example for 1901-2010. *Environmental Development*, 6, 69–79.

Chin, E.H. (1977) Modeling daily precipitation occurrence process with Markov chain. *Water Resources Research*, 13(6), 949–956.

Dubrovský, M. (1997) Creating daily weather series with use of the weather generator. *Environment*, 8(5), 409–424.

Eames, M., Kershaw, T. and Coley, D. (2011) On the creation of future probabilistic design weather years from UKCP09. *Building Services Engineering Research and Technology*, 32(2), 127–142.

Foufoula-Georgiou, E. (1987) A Markov renewal model for rainfall occurrences. *Water Resources Research*, 23(5), 875–884.

Gates, P. and Tong, H. (1976) On Markov chain modeling to some weather data. *Journal of Applied Meteorology*, 15(11), 1145–1151.

Guidom, A.C. (2015) *Kernel Density Estimator and Bandwidth Selection for Density and its Derivatives. R package version 1.0.3*. BP 32 El-Alia, U.S.T.H.B, Algeria: University of Science and Technology Houari Boumediene. https://CRAN.R-project.org/package=kedd [Accessed July 8, 2020].

Harrison, M. and Waylen, P. (2000) A note concerning the proper choice for Markov model order for daily precipitation in the humid tropics: a case study in Costa Rica. *International Journal of Climatology*, 20(14), 1861–1872.

Hosseini, R., Le, N. and Zidek, J. (2011) Selecting a binary Markov model for a precipitation process. *Environmental and Ecological Statistics*, 18(4), 295–820.

Jimoh, O.D. and Webster, P. (1996) The optimum order of a Markov chain model for daily rainfall in Nigeria. *Journal of Hydrology*, 185(1–4), 45–68.

Katz, R.W. (1981) On some criteria for estimating the order of a Markov chain. *Technometrics*, 23(3), 243–248.

Katz, R.W. and Parlange, M.B. (1998) Overdispersion phenomenon in stochastic modeling of precipitation. *Journal of Climate*, 11(4), 591–601.

Köppen, W. (1900) Versuch einer Klassifikation der Klimate, vorzugsweise nach ihren Beziehungen zur Pflanzenwelt. *Geographische Zeitschrift*, 6, 657–679.

Larsen, G.A. and Pense, R.B. (1982) Stochastic simulation of daily climatic data for agronomic models. *Agronomy Journal*, 74(3), 510–514.

Lennartsson, J., Bexevani, A. and Chen, D. (2008) Modelling precipitation in Sweden using multiple step markov chains and a composite model. *Journal of Hydrology*, 363(1–4), 42–59.

Liao, Y., Qiang, Z. and Chen, D. (2004) Stochastic modeling of daily precipitation in China. *Journal of Geographical Sciences*, 14(4), 417–426.

Ochola, W. and Kerkides, P. (2003) A Markov chain simulation model for predicting critical wet and dry spells in Kenya: Analysing rainfall events in the Kano plains. *Irrigation and Drainage*, 52(4), 327–342.

Qian, B., Hayhoe, H. and Gameda, S. (2005) Evaluation of the stochastic weather generators LARS-WG and AAFC-WG for climate change impact studies. *Climate Research*, 29(1), 3–21.

Rajagopalan, B., Lall, U. and Tarboton, D. (1997) Evauation of kernel density estimation methods for daily precipitation resampling. *Stochastic Hydrology and Hydraulics*, 11, 523–547.

Richardson, C.W. (1981) Stochastic simulation of daily precipitation, temperature and solar radiation. *Water Resources Research*, 17(1), 182–190.

Richardson, C.W. and Wright, D.A. (1984) *WGEN: A Model for Generating Daily Weather Variables. ARS-8*. Washington, DC: US Department of Agriculture, Agriculural Research Service.

Schoof, J.T. and Pryor, S.C. (2008) On the proper order of Markov chain model for daily precipitation occurrence in the contiguous United States. *Journal of Applied Meteorology and Climatology*, 2000, 2477–2486.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.

Semenov, M. and Barrow, E. (1997) Use of a stochastic weather generator in the development of climate change scenarios. *Climatic Change*, 35(4), 397–414.

Semenov, M. and Barrow, E. (2002). *LARS-WG A Stochastic Weather Generator for Use in Climate Impact Studies Developed by Mikhail A. Semenov Version 3 User Manual*. Harpenden, UK: Rothamsted Research.

Semenov, M., Brooks, R., Barrow, E. and Richardson, C.W. (1998) Comparison of the WGEN and LARS-WG stochastic weather generators for diverse climates. *Climate Research*, 10(2), 95–107.

Soltani, A. and Hoogenboom, G. (2003a) A statistical comparison of the stochastic weather generators WGEN and SIMMETEO. *Climate Research*, 24(3), 215–230.

Soltani, A. and Hoogenboom, G. (2003b) Minimum data requirements for parameter estimation of stochastic weather generators. *Climate Research*, 25(2), 109–119.

Thom, H.C.S. (1958) A note on the gamma distribution. *Monthly Weather Review*, 86(4), 117–122.

Tong, H. (1975) Determination of the order of a Markov chain by Akaike's information criterion. *Journal of Applied Probability*, 12(3), 486–497.

Wilks, D.S. (1995) *Statistical Methods in the Atmospheric Sciences*. London: Academic Press Limited.

Wilks, D.S. (2010) Use of stochastic weather generators for precipitation downscaling. *Wiley Interdisciplinary Reviews: Climate Change*, 1(6), 898–907.