Water Resources Research

RESEARCH ARTICLE
10.1029/2019WR026416

Special Section:
Big Data & Machine Learning in Water Sciences: Recent Progress and Their Use in Advancing Science

Multisite Weather Generators Using Bayesian Networks: An Illustrative Case Study for Precipitation Occurrence

M. N. Legasa[1] and J. M. Gutiérrez[2]

[1]Meteorology Group, Departamento de Matemática Aplicada y Ciencias de la Computación, Universidad de Cantabria, Santander, Spain, [2]Meteorology Group, Instituto de Física de Cantabria (IFCA, CSIC-UC), Santander, Spain

Abstract Many existing approaches for multisite weather generation try to capture several statistics of the observed data (such as pairwise correlations) in order to generate spatially and temporarily consistent series. In this work, we analyze the application of Bayesian networks to this problem, focusing on precipitation occurrence and considering a simple case study to illustrate the potential of this new approach. We use Bayesian networks to approximate the multivariate (multisite) probability distribution of observed gauge data, which is factorized according to the relevant (marginal and conditional) dependencies. This factorization allows the simulation of synthetic samples from the multivariate distribution, thus providing a sound and promising methodology for multisite precipitation series generation.

Key Points:
- Bayesian networks are introduced as a novel machine learning methodology for multisite precipitation occurrence generation
- Their performance is assessed using several measures in terms of spatial and temporal coherence
- The proposed methodology shows promise, with improvement on several spatiotemporal aspects against existing models

Correspondence to:
M. N. Legasa,
mikel.legasa@unican.es

Citation:
Legasa, M. N., & Gutiérrez, J. M. (2020). Multisite weather generators using Bayesian networks: An illustrative case study for precipitation occurrence. *Water Resources Research*, *56*, e2019WR026416. https://doi.org/10.1029/2019WR026416

Received 26 SEP 2019
Accepted 1 JUN 2020
Accepted article online 8 JUN 2020

1. Introduction

Stochastic weather generators (WGs) produce synthetic time series of weather data of unlimited length for a location based on the statistical characteristics of observed weather at that location. There are a number of reasons why WGs may be required, including the need for long enough time series of daily weather not directly available from observational records and the sparsity and/or missingness of data at some locations. WGs are also needed in climate change studies to avoid the limitations of the raw outputs from global and regional climate models (GCMs and RCMs, respectively), which are typically biased and require some calibration or downscaling for practical applications (Gutiérrez et al., 2019; Manzanas et al., 2019; Maraun et al., 2010). WGs are one of the available (stochastic) downscaling methodologies suitable for some applications, like using "delta" changes obtained from climate change projections (Booij, 2005; Schlabing et al., 2014).

The first approach for developing WGs was proposed by Richardson (1981), in which the generation of precipitation involves itself a two-step process: first modeling the occurrence of wet/dry days using a Markov procedure and then modeling the amount of precipitation falling on wet days. The remaining variables are then computed based on their correlations with each other and with the wet or dry status of each day. These models are often referred to as *Richardson-type,* and produce single-site series with no spatial dependence.

Precipitation has always been a key variable of interest in meteorology and, in particular, for WGs (Ailliot et al., 2015), due to its challenging mixed discrete continuous nature and non-Gaussianity (Duan et al., 2007). Thus, most effort in the construction of WGs has been and is still devoted to precipitation, and most of the currently available WGs still perform separate treatments of the precipitation occurrence (wet/dry) and precipitation amount (Ailliot et al., 2015; Wilks & Wilby, 1999; Zhou et al., 2019). In this work we focus on the discrete aspect: the generation of spatially and temporarily consistent dry/wet days series.

Richardson-type generators have been successfully employed in a wide range of applications in hydrology, agriculture, and environmental management. However, many applications require spatially consistent data, reflecting not only the marginal statistics of the different sites but also intersite (or multisite) statistics such as pairwise correlation (Fiener & Auerswald, 2009; Haile et al., 2009). A number of multisite WG methods have been proposed in the literature, starting with the work of Wilks (1998), in which a multisite WG is presented that prescribes a spatial dependence pattern learnt from the data set. This dependence pattern is simulated by generating uniformly distributed pairwise-correlated series. Wet/dry days are simulated by transforming these series using a threshold that corresponds to the wet-wet and dry-wet transition probabilities. The correlation of the uniform random variables is chosen so that the final binary series have the observed pairwise correlations.

Existing approaches for multisite weather generation build on many different techniques to generate synthetic spatially consistent series, such as empirical orthogonal function analysis (Zhou et al., 2019), weather typing (Ailliot et al., 2015), nearest neighbours (Yates et al., 2003), and Gaussian processes (Kleiber et al., 2012) —which permit the interpolation to locations where there are no direct observations—or more sophisticated approximate Bayesian computation (ABC) methods for thresholded Gaussian processes (Olson & Kleiber, 2017). In general, there is no overall best method, and the Wilks seminal contribution is still a good first choice in terms of complexity and performance, as shown in different intercomparison studies (Keller et al., 2015; Mehrotra et al., 2006). Moreover, this method is typically used as a benchmark to describe new methodologies, thus allowing the (indirect) comparison of methods by contrasting their respective improvement with respect to the Wilks benchmark. Therefore, we use the Wilks method for benchmarking the new weather generation methodology based on Bayesian networks (BNs) introduced in this work.

BNs are a sound and popular machine learning technique which combines graphs and probability theory to build tractable probabilistic models from data, representing the most relevant (pairwise and conditional) dependencies among the variables (Castillo et al., 1997). BNs have gained widespread use in several fields (Niedermayer, 2008; Pourret et al., 2008), boosted by the availability of several commercial and open software packages allowing to efficiently learn them from data, such as the *bnlearn* popular implementation in R used in this work (Scutari, 2010; Scutari & Denis, 2014). However, their application to environmental sciences is still limited (Aguilera et al., 2011; Borunda et al., 2016; Uusitalo, 2007), and only a few applications for water resource management have been described in the literature (Phan et al., 2016; Ropero et al., 2017). The use of BNs in meteorology was first described by Cano et al. (2004), illustrating their potential application for weather prediction and generation. Most of the applications of BNs described so far in this field correspond to probabilistic weather prediction and downscaling (Boneh et al., 2015; Cofiño et al., 2002; Das & Ghosh, 2014, 2017; Hellman et al., 2012; Nandar, 2009; Sharma & Goyal, 2016; Smail, 2018), and there are also some applications for drought/flood forecasting (Garrote et al., 2008; Madadgar & Moradkhani, 2014). However, to our knowledge, a comprehensive application of BNs for stochastic weather generation has not been described yet.

In this work we describe the application of BNs to stochastic weather generation of precipitation occurrence. BNs learn tractable multivariate discrete models from the available historical data which encode the relevant spatial and temporal dependencies among the stations. We illustrate the new methodology using a small case study over Germany and validate the results using the Wilks WG (Wilks, 1998) as benchmark.

In order to facilitate the reproducibility of the results and testing the performance of the proposed methodology, we have prepared an R package with the software used, which builds on the R package bnlearn (Scutari, 2010), and a Jupyter notebook illustrating the creation and use of some of the models presented in this work (both available at https://github.com/MNLR/BNWeatherGen).

The remainder of this article is organized as follows: Section 2 explains the data used throughout this work. Section 3 introduces and explains WGs. Section 4 introduces and explains BNs from a theoretical perspective, and section 5 describes the methodology used to employ BNs as WGs. Since BNs require the choice of a complexity parameter, section 6 is devoted to finding the optimal complexity. All gathered results are described in section 7, and finally, section 8 gives the conclusions and establishes the future lines of work.

2. Area of Study and Data

As an illustrative case study, we consider a subset of 11 stations in southeast Germany extracted from the VALUE spatial validation experiment (Widmann et al., 2019). The data set is provided by the European Climate Assessment & Dataset project (ECA&D), and the daily precipitation records used in this study range from 1979 to 2008 (30 years, corresponding to the VALUE experimental period). We have restricted the experiments to a subset of 11 stations for the sake of illustration and due to the computational cost of the comprehensive analyses performed. Also, we treated observations from summer (JJA) and winter (DJF) separately; that is, models are trained separately for each season using all the available daily data for the particular season. We omit the remaining seasons for brevity reasons, as results are analogous. In total, and having removed observations with missing values (less than 5% of the days of the period considered), we have 2,668 daily observations for the JJA season and 2,614 for the DJF season. Since we require consecutive

data for the WGs, the number of training instances left for the BNs are 2,548 and 2,553 for the JJA and DJF seasons, respectively.

Discretization has been carried out with a day considered *dry* if precipitation amount was lower than 1 mm and *wet* otherwise. Table 1 describes the stations, including geographical information and some basic marginal and temporal statistics which will be used later in the study: wet-day frequencies (WF) and mean of annual maxima dry (DS)/wet (WS) spells. ID is the unique code that identifies the stations in the ECA&D data sets, and Code is the unique code used throughout this work. The orography of the region is shown in Figure 1a.

3. Multisite Weather Generators

Given a set of stations $\{1, ..., p\}$, we consider the random variable $\mathbf{X} = (X_1, ..., X_p)$, where each X_i characterizes rainfall occurrence (binary) at station $i = 1, ..., p$ ($p = 11$ in this work). If $X_i = 1$ codifies a wet day at station i and $X_i = 0$ a dry day at station i, a multivariate WG can be viewed as a model to obtain samples from the distribution $P(\mathbf{X}^t \mid \mathbf{X}^{t-1}, \mathbf{X}^{t-2}, ...)$, where $\mathbf{X}^t$ means $\mathbf{X}$ at time slice t. It is often simplified according to Markov assumption as a Markov-1 process, in which the future is assumed to be dependent on just the day before, that is,

$$P(\mathbf{X}^t \mid \mathbf{X}^{t-1}, \mathbf{X}^{t-2}, \dots) \approx P(\mathbf{X}^t \mid \mathbf{X}^{t-1}) = P((X_1^t, ..., X_p^t) \mid \mathbf{X}^{t-1}). \quad (1)$$

Due to the difficulties in obtaining a tractable multivariate estimation of the above distribution, this process is often achieved by first computing P sampling from the univariate distributions $P(X_i^t | X_i^{t-1})$, $i = 1, ..., p$, with the first step in a WG being computing the probabilities $P_i^{01} = P(X_i^t = 1 | X_i^{t-1} = 0)$ and $P_i^{11} = P(X_i^t = 1 \mid X_i^{t-1} = 1)$ from the data set, often referred to as the transition probabilities.

Then, random number generators are employed to generate weather series that follow these transitions. For multisite WGs, these random numbers are often correlated to mimic the correlations observed between pairs as an ad hoc methodology to impose the desired spatial structure. This is the case for the Wilks WG (see Wilks, 1998 and an evaluation in Mehrotra et al., 2006), which we use in this work as a benchmark for the new proposed methodology.

4. Bayesian Networks

Characterizing a multivariate discrete probability distribution such as (1) involves an intractable number of parameters that grows exponentially with the number of variables, thus hindering practical applications. Bayesian networks (Castillo et al., 1997; Pearl, 1988; Scutari & Denis, 2014) are probabilistic graphical models that combine graph and probability theories to efficiently learn from data the Joint Probability Distribution (JPD) of a set of discrete random variables, while also representing their relationships in an easy-to-interpret graph. For a set of discrete random variables $\{X_1, X_2, ..., X_p$ describing the quantities of interest (rainfall occurrence in a network of p stations in this case), the JPD $P(X_1,...,X_p)$ has 2^p possible categories and thus requires 2^p-1 parameters. This poses a practical problem, since even in the case when this is feasible computationally, we are unlikely to have a large enough data set to be able to adjust that many parameters.

A BN builds on a directed acyclic graph (DAG)—a directed graph with no directed cycles and each node associated with one variable X_i—which encodes the dependence and independence relationships among the variables, so if there is no arc connecting two nodes, the corresponding variables are either independent or conditionally independent given a subset of the remaining variables. In the example from Figure 1b, station 5 is independent of any other station given we know the values of stations 3,7, and 8. These nodes are the so-called Markov Blanket for node 5. These relationships are formalized through the concept of d-separation (see Scutari & Denis, 2014 and Koller & Friedman, 2009 for a formal definition of these concepts).

The independencies defined by d-separation in the graph imply a factorization of the JPD in terms of the probabilities of each of the variables conditioned to its parents (for a configuration in the form $X \rightarrow Y$, X is a parent of Y) (Koller & Friedman, 2009):

Table 1
Description of Meteorological Stations Used in This Study Located in Southeast Germany Showing the Code, ECA&D ID, Altitude (in meters), Longitude, Latitude, Relative Wet-Day Frequency (WF), and Mean of Longest Annual Dry (DS)/Wet (WS) Spells for Summer-JJA (s) and Winter-DJF (w)

Code	ID	Alt	Lon	Lat	WFs	WFw	DSs	DSw	WSs	WSw	Location
1	4007	921	9.94	50.5	0.40	0.48	11.41	11.17	6.76	8.6	WASSERKUPPE
2	4572	415	10.17	49.39	0.33	0.36	13.28	13.20	5.14	5.57	ROTHENBURG OB DER TAUBER
3	4472	435	10.51	48.83	0.34	0.31	12.26	14.73	4.90	5.2	REIMLINGEN
4	4617	937	10.77	50.66	0.41	0.50	10.69	10.43	7.03	8.97	SCHMUCKE
5	52	515	11.54	48.16	0.40	0.34	9.55	13.17	5.59	5.27	MUENCHEN
6	4083	657	11.84	49.98	0.40	0.47	10.62	10.87	7.31	8.53	FICHTELBERG OBERFRANKEN
7	4004	365	12.1	49.04	0.35	0.33	11.55	12.90	5.48	4.8	REGENSBURG
8	4079	472	12.73	48.48	0.37	0.34	11.21	13.10	5.90	5.4	KR.ROTTAL-INN FALKENBERG
9	4954	418	12.87	50.79	0.35	0.32	11.69	14.33	4.76	5.03	CHEMNITZ
10	488	1213	12.96	50.43	0.43	0.49	9.86	9.80	6.52	8.17	FICHTELBERG
11	483	227	13.76	51.13	0.33	0.33	12.59	12.5	5.31	4.47	DRESDEN-KLOTZSCHE

$$P(X_1, ..., X_p) \approx P_{dag}(X_1, ..., X_p) = \prod_{i=1}^{p} P(X_i|\pi(X_i)), \tag{2}$$

where $\pi(X_i)$ is the set of parents of node X_i in the graph. This factorization on the right-hand side requires, for each node, the specification of a conditional probability table (CPT), that is, a probability distribution for the node conditional to each combination of the parents' states. In turn, each of these CPTs require a total of $(s(X_i)-1)*s(\pi(X_i))$ parameters, where s represents the number of states of the variable or combination of variables. This results in a joint probability model that requires only a moderate number of parameters, leading to parsimonious data-driven models.

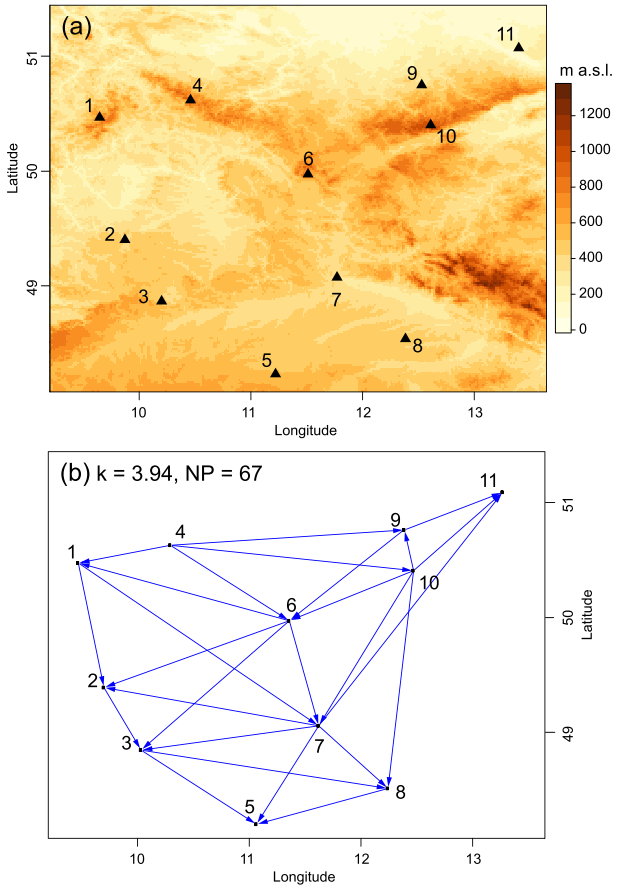

Figure 1. (a) Orography of the regions of study in Southeast of Germany and location of the 11 stations used in this study. (b) Example of a directed acyclic graph showing the dependence structure among the stations.

For instance, Figure 1b shows the DAG approximating the JPD of the set of stations described in the previous section (corresponding to summer-JJA data), implying the following factorization for the JPD:

$$P(X_1, ..., X_{11}) = P(X_4)P(X_{10}|X_4)P(X_9|X_4, X_{10})P(X_6|X_4, X_9, X_6)....$$

Each node's associated CPT depends on the factorization. For example, there is a CPT associated to the term $P(X_9|\pi(X_9))$ for node X_9. It requires the specification of one parameter for the Bernoulli distribution of node X_9 (probability of *rain* at that station) for each condition $(X_{10}=0, X_4=0)$, $(X_{10}=0, X_4=1)$, $(X_{10}=1, X_4=0)$, and $(X_{10}=1, X_4=1)$, thus requiring $(s(X_9)-1)*s(\pi(X_9))=(2-1)*4$ parameters. For this node and this particular data set (JJA), we have the CPT shown in Table 2, as estimated from the data using Bayesian estimation.

Similarly, since X_{10} and X_4 have one and no parents, respectively, their CPTs are those in Table 3.

BNs also allow for a qualitative analysis of the dependencies and independencies codified in the graph, and although this is not the purpose of this work, they can be used for answering questions such as "Is variable X_i independent of X_j given a set of variables $\mathcal{X}_k$?" As we will see later, the density of the graph can vary depending on the complexity required for the particular problem/application, since a DAG with more arcs captures more dependence relationships but requires more parameters.

Table 2
CPT for Node X_9, Station 9 in Figure 1

$s(\pi(X_9))$	$X_{10}=0$, $X_4=0$	$X_{10}=1$, $X_4=0$	$X_{10}=0$, $X_4=1$	$X_{10}=1$, $X_4=1$
$P(X_9=1)$	0.04	0.52	0.27	0.78

4.1. Learning Bayesian Networks from Data

From a practical point of view, the application of Bayesian networks to real-world problems depends on the availability of automatic learning procedures to infer appropriate network structure (DAG) representing the relevant independencies from a given data set. Unfortunately, this problem has been shown to be *NP*-hard (Chickering et al., 2004), since the number of possible DAGs is superexponential on the number of nodes and a graph with p nodes can have up to $\frac{1}{2}p(p-1)$ possible arcs. This results in $2^{O(p^2)}$ possible graphs, and several heuristic methods have been developed for learning the graphical structure (structure learning) and estimating the probabilities (parametric learning) from data in reasonable times.

Structure learning methods can be classified in either *constraint-based* or *score-based* approaches (Scutari & Denis, 2014). Constraint-based algorithms explore the possible conditional dependencies and independencies among sets of variables by applying conditional independence tests, such as χ^2, following the rationale of placing an arc if two nodes are dependent given sets of different nodes. In this case, the graph is formed by the aggregation of the local dependencies. These algorithms are very sensitive to failures in independence tests performed. Scutari et al. (2019) analyzed the performance of different learning methods in climate problems and concluded that score-based methods performed better than constraint-based methods. Indeed, in our experiments, constraint-based algorithms have shown to yield poor results, so we restrict our study to score-based approaches.

Score-based approaches search through the space of possible graphical structures maximizing a score that evaluates how well the graph represents the data set, returning the best model obtained after an iterative process. Therefore, they require

- A score, which should be representative of how well the graph represents the data set, that is, a *quality measure* (Heckerman et al., 1995). The most commonly used score is Bayesian information criterion (BIC), introduced in (Schwarz, 1978) and equivalent to minimal description length (Lam & Bacchus, 1994).
- An optimization algorithm, that is, a heuristic search to maximize the score, from simple greedy approaches like hill-climbing to more elaborate ones like genetic algorithms.

Structure learning can be formalized as finding the DAG G that maximizes $P(G|\mathcal{D})$, where $\mathcal{D}$ is the data set. It can be decomposed using Bayes theorem into

$$P(G|\mathcal{D}) = \frac{P(\mathcal{D}|G)P(G)}{P(\mathcal{D})}.$$

As $P(\mathcal{D})$ is constant, the problem is equivalent to maximizing $P(\mathcal{D}|G)P(G)$, the product of the *prior* distribution over the possible DAGs, $P(G)$; and the probability of obtaining the data from the distribution obtained by the factorization implied by G, $P(\mathcal{D}|G)$ (*likelihood* in the Bayesian setting). The term $P(\mathcal{D}|G)$ is the actual theoretical *score* we are looking to maximize, with $P(G)$ allowing us to introduce informed priors, like linking a station with its *past*. It has been shown that $P(\mathcal{D}|G)$ can be approximated with a bounded error by the so called BIC *score,* defined in the context of BNs as (Scutari & Denis, 2014)

$$\mathrm{BIC}(G, \mathcal{D}) = \sum_{i=1}^{p} (\log(P(X_i|\pi(X_i))) - k|\Theta(X_i)|), \tag{3}$$

which profits again from the factorization 2. $|\Theta(X_i)|$ is the number of parameters (probability values) for each CPT associated to each node X_i, and k is the *regularization parameter*, which takes the value $\log(n)/2$ ($\approx$3.94 for this data set in JJA for building the DAG in Figure 1), where n is the number of observations in the data set. This term penalizes the score of high-density DAGs, thus preventing the number of

Table 3
CPTs for Nodes X_4 and X_{10}, Stations 4 and 10 in Figure 1

$s(\pi(X_4))$	∅	$s(\pi(X_{10}))$	$X_4=0$	$X_4=1$
$P(X_4=1)$	0.41	$P(X_{10}=1)$	0.21	0.74

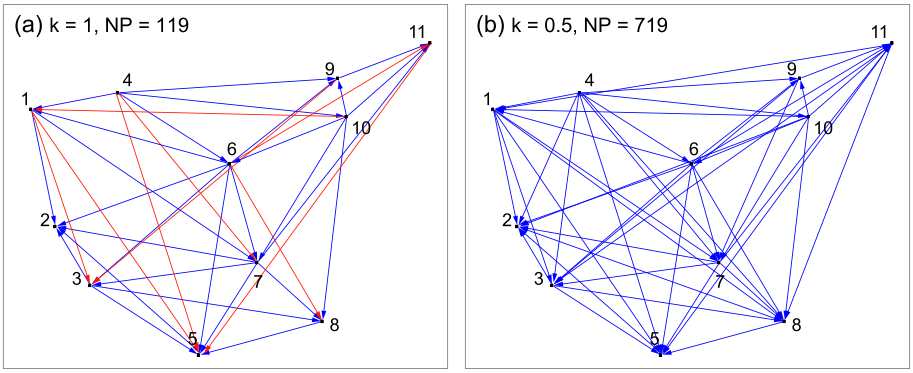

Figure 2. Two graphs learnt from data, as in 1b, but with decreasing regularization parameter (a) $k = 1$ and (b) $k = 0.5$, thus increasing the number of links and parameters (NP). Note that Figure 1b was obtained with the same algorithm but with the default regularization parameter ($k = 3.94$); for the sake of comparison with this figure, panel (a) shows in blue the graph with regularization parameter ($k = 3.94$) and the newly added links corresponding to $k = 1$ in red. Note that these are *descriptive* networks, as they do not have a temporal aspect.

arcs from growing too large. Moreover, taking the logarithm accounts for the fact that, even when there are a lot of observations, the complexity of the model grows exponentially. Although BIC score specifically uses $k = \log(n)/2$, a practical way to adjust the DAG complexity to our needs is by varying this parameter. The lower this value, the more arcs the learning algorithm will place. It should be noted that, when $k = 1$, this coincides with the Akaike information criterion (AIC) (Akaike, 1998).

Once the score is chosen, the optimization is carried out. Ideally, all possible DAGs would be searched for the one that fits best. As we have stated before, the superexponential number of possible DAGs renders this impossible, and heuristic approaches are required. The Tabu search algorithm (Glover, 1990; Koller & Friedman, 2009) is a standard heuristic search procedure for maximizing the score in the discrete space of possible DAGs. In its basic implementation, it is essentially a greedy approach (hill-climbing) with a *memory* that stores a backtrack of already visited DAGs and performs additional searches prohibiting previously visited local maxima, thus yielding generally better results than simple hill-climbing, as described in Scutari et al. (2019). The empirical complexity of greedy search is cubic in the number of nodes but can be reduced to quadratic by constraining the search. The interested reader is referred to Scutari et al. (2019) for a detailed analysis of the empirical complexity of different approaches as a function of both the number of nodes and the size of data available.

Figure 1b shows the graph obtained with Tabu search for the default regularization parameter $k = 3.94$, whereas Figures 2a and 2b show two alternative graphs obtained with $k = 1$ and 0.5, respectively, for JJA (similar results are obtained for DJF, not shown). The number of parameters grows from 67 to 119 and to 719, respectively (with 22, 33, and 48 links). The parameter k is essential to avoid overfitting, and it can be easily checked that with $k = 0$, the result of the algorithm would simply be the complete DAG, with the factorization from Formula 2 being that of the chain rule (Castillo et al., 1997). Dense DAGs yield large parent sets implying large CPTs which have to be estimated from data, in some cases based on few or no samples, thus becoming unstable. For example, node 5 alone in Figure 2b requires a specification of a probability for $2^8 = 256$ (eight parents) different realizations (combinations) of parents, many of which may not even be present in the data set. In this work, we use the $k = 1$ intermediate network, a compromise between fitting data and model simplicity (see section 6) which, as stated before, coincides with AIC (Akaike, 1998).

Once the DAG has been learned, parameter learning becomes straightforward. Probabilities can be estimated from the observed frequencies (*maximum likelihood estimation*, MLE) or in a Bayesian setting, using their posterior distribution (see, e.g., Sivia & Skilling, 2006 for the details of Bayesian estimation), a methodology which prevents the parameters from being exactly 0 in order to avoid sparse tables, with lots of 0 cells. This is required to fulfill regularity conditions of model estimation and inference methods (Koller & Friedman, 2009; Scutari & Denis, 2014).

In particular, the software used in this work (The R package bnlearn Scutari, 2010) computes the *posterior estimates* as a weighted mean of a flat prior and the empirical frequencies, as follows. In the factorization,

suppose we have to estimate $P(X_i = 1 \mid X_j = 0)$ and recall that $P(X_i \mid X_j) = P(X_i, X_j)/P(X_j)$. If we denote as $\widetilde{p}_{X_i,X_j}$ the maximum likelihood estimator of $X_i = 1, X_j = 0$ (i.e., the number of observations for which $X_i = 1, X_j = 0$ divided by the total number of observations) and $\widetilde{p}_{X_j}$ is the number of observations for which $X_j = 0$ divided by the number of observations, then the posterior estimate $\widetilde{P}(X_i = 1 \mid X_j = 0)$ is

$$\widetilde{P}(X_i = 1 \mid X_j = 0) = \frac{\widetilde{P}(X_i = 1, X_j = 0)}{\widetilde{P}(X_j = 0)} = \frac{\frac{n}{n+1}\widetilde{p}_{X_i, X_j} + \frac{1}{n+1}\pi_{X_i, X_j}}{\frac{n}{n+1}\widetilde{p}_{X_j} + \frac{1}{n+1}\pi_{X_j}},$$

where n is the number of observations, $\pi_{X_i, X_j} = 1/4$ and $\pi_{X_j} = 1/2$, since we consider flat priors (uniform distribution) and are working with binary variables. Note that if there are no observations to inform the parameters, then we have $\widetilde{P}(X_i = 1 | X_j = 0) = 0.5$. Other priors based on expert knowledge are possible, but a sensitivity analysis to the choice of different priors is out of the scope of this work.

5. Bayesian Networks as Weather Generators

Bayesian networks allow the factorization of equation 1 according to equation 2 in the following way:

$$P(\mathbf{X}^t \mid \mathbf{X}^{t-1}) = P((X_1^t, ..., X_p^t) \mid \mathbf{X}^{t-1}) \approx \prod_{i = 1, ..., p} P(X_i^t \mid \pi(X_i)^t, \mathbf{X}^{t-1}). \quad (4)$$

This factorization provides a simple form for simulating, one by one, in ancestral ordering—first parents then children—a synthetic value for each of the variables $i = 1, ..., n$.

From now on, we refer to nodes in time slice t as *present* nodes and to nodes in time slice $t-1$ as *past* nodes. Also, *spatial* arcs or dependencies join either two *present* or two *past* nodes, and *temporal*arcs join a *past* and a *present* node (not necessarily in that direction).

Note that if we assume that X_i^t is independent of all X_j^{t-1}, $j \neq i$, given X_i^{t-1} (i.e., all information from the past comes through the node's past value), then 4 can be expressed as

$$P(\mathbf{X}^t \mid \mathbf{X}^{t-1}) \approx \prod_{i = 1, ..., p} P(X_i^t \mid \pi(X_i)^t, X_i^{t-1}), \quad (5)$$

resulting in a very simple factorization where each present node X_i^t can be simulated once the value of their corresponding past node X_i^{t-1} and the node's present parents $\pi(X_i)^t$ are known. We refer to this BN as *Markov* (see Figure 3a).

However, other less restrictive extended approaches can be easily considered to build the temporal dependencies such as the *Unconstrained* approach (see Figure 3b), consisting on leaving the algorithm to automatically learn both spatial and temporal dependencies by using an extended DAG for the extended set of variables $(X_1^{t-1}, ..., X_p^{t-1}, X_1^t, ..., X_p^t)$, where the new data set is built using the values of consecutive days. However, this model may not be optimized for the particular task; for instance, only nodes X_8 and X_{10} are connected to its past counterparts, and all other nodes received information from the past through a different node (building on the data). Therefore, we consider an additional model, obtained by augmenting the Markov BN structure shown in Figure 3a by allowing the training algorithm to include additional links according to the data (see Figure 3c). This model seems to be appropriate since this kind of dependence seems highly likely. Taking into account the dependence structure encoded in the DAG, in practice, this is done by forcing the algorithm to place an arc but letting it choose the direction (Formula 6 considers arcs to go forward for simplicity, from $t-1$ to t). In practice, they can go backwards to build the most accurate model (note that arcs do not represent causal relationships). We call this model the *Augmented model*, for which Equation 4 becomes

$$P(\mathbf{X}^t \mid \mathbf{X}^{t-1}) \approx \prod_{i = 1, ..., p} P(X_i^t \mid \pi(X_i)^t, X_i^{t-1}, \pi(X_i)^{t-1}), \quad (6)$$

with $\pi(X_i)^t$ and $\pi(X_i)^{t-1}$ being decided by the algorithm.

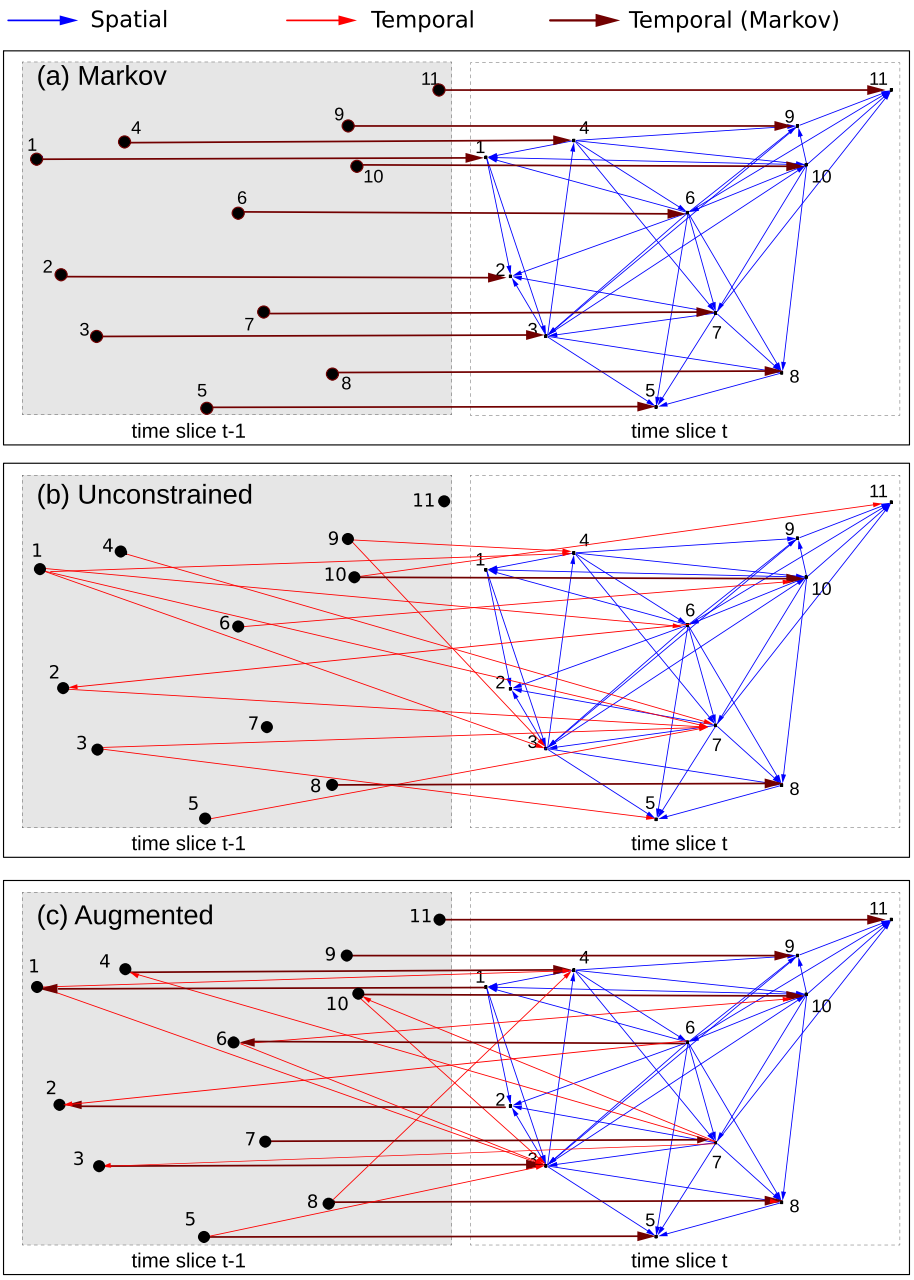

Figure 3. Different Bayesian network models considered for weather generation. (a) Markov, in which only temporal relationships between a node and its past are permitted; (b) Unconstrained, in which the algorithm decides all the arcs; and (c) Augmented, in which the arcs between a node and its past are forced (the direction is chosen by the algorithm) and the algorithm places the rest of the arcs. Note that arcs do not imply a causal relationships, and thus arcs facing backwards are perfectly normal. Their actual direction is decided based on the d-separation criterion (Scutari, 2010), which considers the dependence relationships across all nodes. Also note that spatial arcs in the time slice $t-1$ have been removed for visualization purposes, but they are present in the model with the same spatial structure as the one in time slice t.

Letting the algorithm decide the complete structure is the *pure* machine learning approach, whereas introducing the arcs means using *expert knowledge* (recall the term $P(G)$ in section 4.1) in the assumption that there is a direct dependence between a station on a particular day (X_i^{t-1}) and that same station on the following day (X_i^t). We first learn the spatial structure (learning a DAG for just the stations, without temporary slicing) and then add the temporal structure by adding the past nodes.

Once the network and the resulting probabilities have been learnt, the simulation process to generate weather series is straightforward. A basic property of a DAG is that it has a node with no parents (Pearl, 2013), whose CPT is just its marginal probability distribution. There is also a topological order of the nodes (ancestral ordering) $X_1 \leq ... \leq X_p$ such that either $\pi(X_i) \leq X_i$ or $\pi(X_i) = \emptyset$, that is, parents are ordered before children (note that $\pi(X_1) = \emptyset$). For the DAG in Figure 3b, we have $X_4 \leq X_{10} \leq X_9 \leq X_6 \leq X_1 \leq X_7 \leq X_2 \leq X_3 \leq X_8 \leq X_5 \leq X_{11}$ Finally, as we have explained in section 4, each node has a CPT associated, which comprises a probability function for each combination of fathers' states.

We can generate instances by applying a simple iterative process. We assume that nodes are sorted following an ancestral ordering. We start by X_1, which has no parents, and simulate from its marginal distribution (in Figure 1b, this is Station 4), obtaining $X_1 = x_1$. We then simulate X_2, as X_2 has either no parents or it has X_1 as a parent. In the first case, we simulate as with X_1, and in the second case, we use its CPT $P(X_2 | X_1 = x_1)$, obtaining $X_2 = x_2$ (in Figure 1b, this is Station 10). We follow this process with X_3 up to the last node.

When working with WGs, as with DAGs in Figure 3, we follow this same process by instantiating the *past* nodes X_1^{t-1}, ..., X_p^{t-1} by updating the probabilities of X_i^t, $i = 1, ..., n$ conditioned to this evidence (see chapters 8 and 9 Castillo et al., 1997), and then we follow the process explained above. That is, we start with an observation $\mathbf{X}^0 = (X_1^0 = x_1^0, ..., X_p^0 = x_p^0)$, and we follow with $P(X_1^t | \mathbf{X}^{t-1} = \mathbf{X}^0)$, obtaining x_1^1. We then continue with $P(X_2^t | X_1^t = x_1^1, \mathbf{X}^{t-1} = \mathbf{X}^0)$, and so on. Once we obtain $\mathbf{X}^1 = (X_1^1 = x_1^1, ..., X_p^1 = x_p^1)$, it serves as the next evidence to the iterative process.

6. Choice of the Regularization Parameter *k* Using Cross-Validation

As explained in section 4.1, the regularization parameter k determines the type of model obtained from the learning process, increasing its complexity (larger number of links and parameters) as k decreases toward zero.

In the case of the Bayesian network weather generators, the BIC score is defined with a regularization parameter $k = \log(n)/2 \approx 3.92$, where n is the number of training instances, which penalizes the growth of the number of arcs. However, other alternative values have been also considered in the literature, such as $k = 1$, which corresponds to the AIC, a well-known estimator of the quality of statistical model for a given data set (Akaike, 1998). From a practical point of view, for the end-user, it may be difficult to choose the value of this parameter. By lowering k, the model might be able to better represent the JPD of the data, but care must be taken when doing so, since the number of parameters grows exponentially as the number of arcs increases, leading to overfitting and scarcity of data examples to estimate some of the probability values of the CPT tables.

Therefore, in order to assess the choice of k, we considered the Augmented BN, shown in Figure 3c for k = 1, and performed a cross-validation experiment using a 5-fold cross-validation (see section 7.10 in Hastie et al., 2009) by dividing the data set into five consecutive folds with the observations divided into the years 1979–1984, 1985–1990, 1991–1996, 1997–2002, and 2003–2008. Figure 4 shows the results obtained for each k from 0.25 to 3.75 (mean across all five folds is shown).

The performance of the resulting models was measured considering the log-likelihood of the data given the models, for both the test and training data (Figure 4a). In order to test the complexity of the resulting models, we considered the number of parameters (probabilities of the CPTs) and the percentage of those which cannot be estimated from the data because there are no observations of the particular event; see section 4.1 (Figure 4b). Finally, the predictive capability was measured considering the area under the ROC curve (AUC; see Bradley, 1997) when predicting precipitation occurrence given the *past* values in the test samples (Figure 4c).

Based on these results, we choose $k = 1$ as a compromise between performance and complexity. Higher values, as shown in Figure 5, tend to lose pairwise correlations, whereas panels (a) and (b) from Figure 4 indicate missing parameters and overfitted models for smaller values of k and, even though the potential predictive capability is slightly higher, the log-likelihood in test samples falls substantially for $k = 0.75$. Finally, this choice is supported by the fact that, as stated before, $k = 1$ corresponds to the AIC.

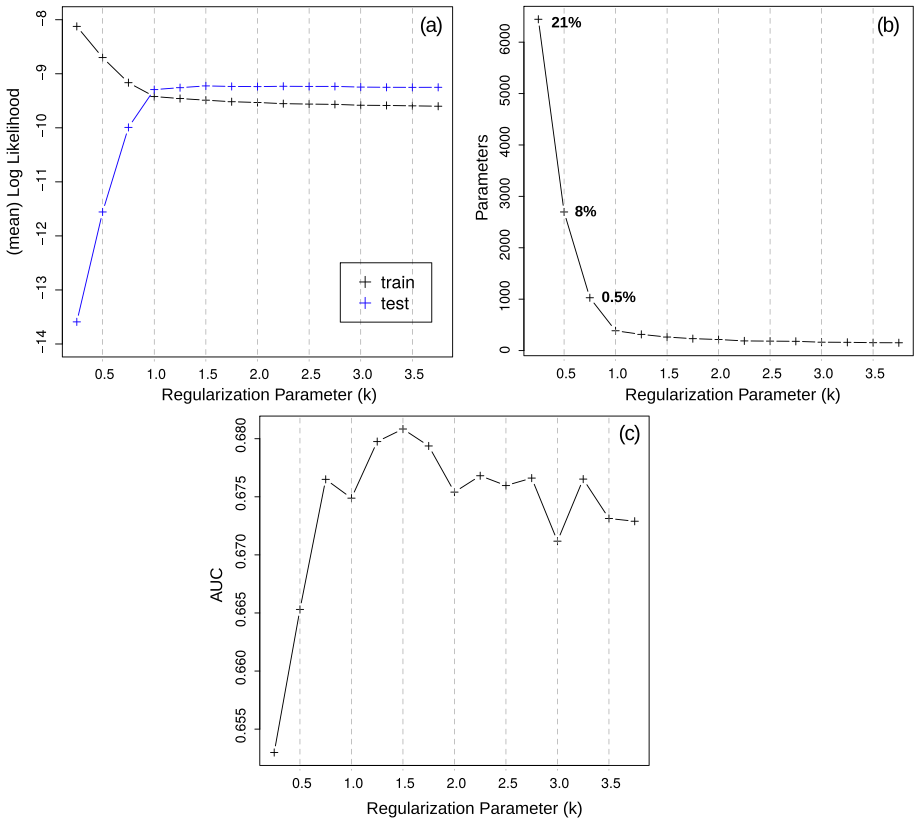

Figure 4. Results of the Bayesian networks for different values of the regularization parameter k. Panels show (a) the log-likelihood of the train and test data, averaged across the data sets' lengths for easy comparison; (b) the number of parameters of the resulting models (numbers inside the panel indicate the proportion of parameters which cannot be learnt from data for the three cases that this happens, corresponding to $k = 0.25,0.5,0.75$); and (c) the predictive performance as measured by the area under the ROC curve (AUC).

7. Results

For a robust comparison of the weather generation models based on both the benchmark, the Wilks method (Wilks, 1998), and the different Bayesian network (BN) models described in the previous sections, we built each model and then generated 250 series of approximately equal length as that of the observational data set ($2{,}668 = 29 \times 92$ for JJA and $2{,}614 \approx 29 \times 90$ for DJF). We replicated series of 92 and 90 continuous days for seasons JJA and DJF, respectively, to account for the discontinuities in the observational data set, thus making spell measures comparable for observed and synthetic series. BNs can generate a synthetic initial observation using (2); however, for the sake of comparison, we used a real observation as initial conditions for both Wilks and Bayesian models (therefore, each generated series has 29 real observations). Throughout this section, all measures are the mean of the 250 diagnostic measures computed for each series, and all results are shown for the JJA season (in general, results for DJF season are similar and not shown) if it is not otherwise specified.

Mehrotra et al. (2006) describes some diagnostics to assess and compare the performance of the Wilks WG. Here we extend this analysis with some additional measures and compare the results against the BN WG. We focus on the representation of both temporal and spatial aspects, including the combination of both (spatiotemporal aspects). The results correspond to the *Unconstrained* and *Augmented* BNs, explained in the previous section.

The common approach to validate the spatial coherence of weather generated series (against observations) is to compute the correlation between pairs of stations (Mehrotra & Sharma, 2009; Mehrotra et al., 2006). However, there are other aspects of the spatial coherence beyond marginal pairwise correlations, such as the conditional dependencies (correlations) or frequency of different realizations. Also, spatiotemporal

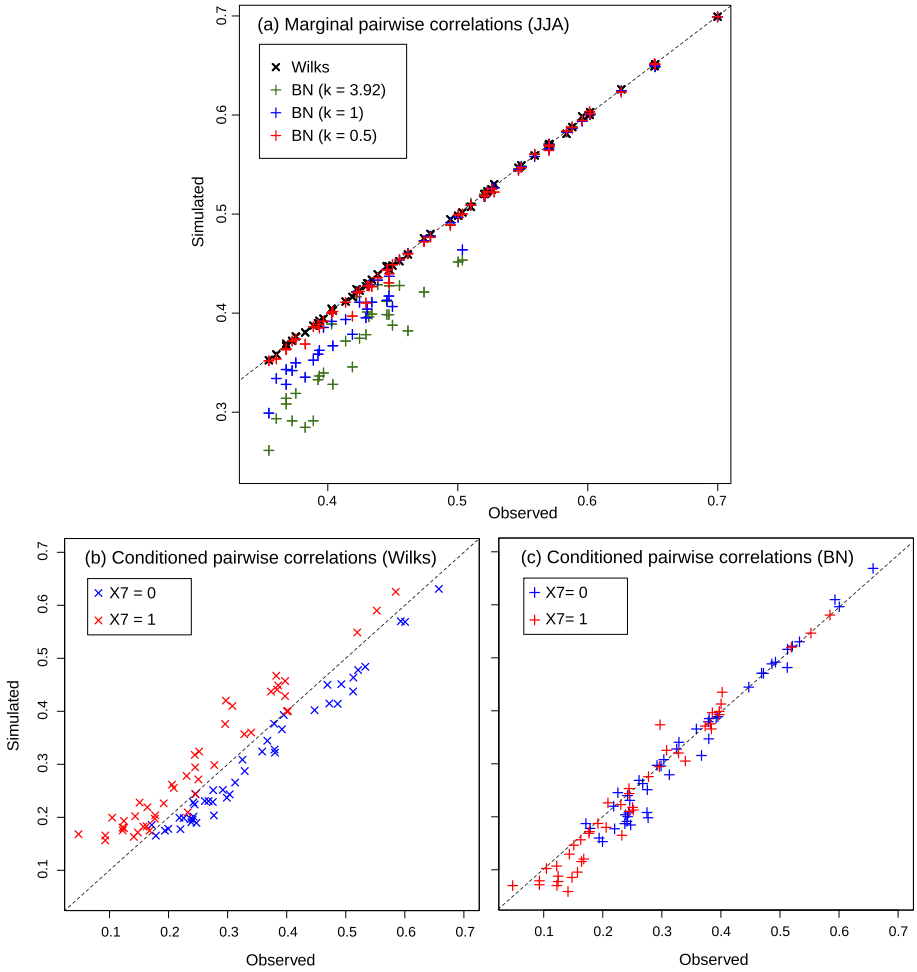

Figure 5. (a) Correlation between pairs of stations plotted against the observed one for Bayesian networks (BN) with different regularization terms (k) and Wilks models. The diagonal shows perfect adjustment. (b, c) Pairwise correlation conditioned to $X_7 = 0$ (blue) and $X_7 = 1$ (red) for the Wilks and Bayesian network ($k = 1$) model, respectively. For marginal pairwise correlations, the mean absolute error for the Wilks model is 0.14. For the BN model with $k = 1$, the mean absolute error is 1.18.

validation measures, which consider both spatial and temporal aspects altogether can be considered: multisite spells and lagged cross-correlations. We first consider the following measures covering the spatial and spatiotemporal aspects of precipitation, which we expand with corresponding figures and some summary error measures based on these (see Table 4):

(a) Pairwise correlations. This is the general performance measure used in previous works, as in Wilks (1998).

(b) Conditional pairwise correlations, which measure the correlation between pairs of stations conditional to a third station.

(c) 1-day lagged cross-correlations, as in Wilks (1998). Note that lagged autocorrelation is included for easy comparison against lagged cross-correlations but is actually a temporal validity measure that captures the same concept as the indices WW and DW in the next section.

(d) Multisite dry/wet observations. We compare the frequency of observations for which there is a dry/wet day for all stations. This further characterizes the spatial coherence of the simulated series. Combined, these observations comprise 37% and 44% of the data set for the JJA and DJF seasons, respectively.

Table 4
Table of Summary Measures Used for Assessing the Spatial Coherence of Weather Simulations

Name	Description
Cor	Mean pairwise correlation
CondCor	Conditional mean pairwise correlation
LaggedCrossCor	Lagged mean pairwise cross-correlation
LaggedAutoCor	Lagged mean autocorrelation
MultisiteDry	Number of multisite dry observations
MultisiteWet	Number of multisite wet observations

Note. Results are shown in Tables 5 and 6.

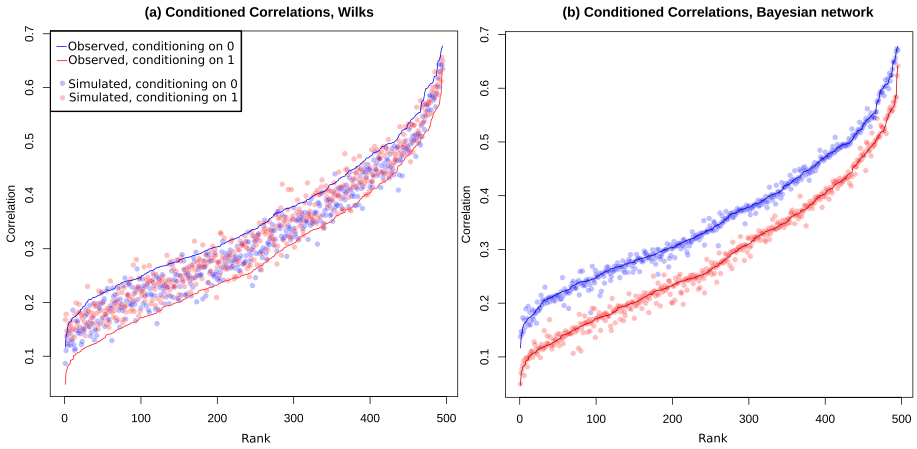

Figure 6. (a,b) Correlations between pairs of stations X_i, X_j given $X_k = 0$ (blue) and $X_k = 1$ (red), for all $i, j, k = \{1,...,11\}$, $i \neq j \neq k$ stations, sorted by their observed value in ascending order (x axis, rank). Observed values are shown as a blue (conditional on $X_k = 0$) and red (conditional on $X_k = 1$) line. Results are shown for the JJA season. Mean absolute error (multiplied by 100) is 4.1 for the Wilks model and 1.9 for both BN models.

(e) Multisite spells. Spells have already been analyzed in previous works, usually for each station individually. Here we also analyze spells occurring simultaneously in more than 90% of the stations, as done for the model introduced in Kleiber et al. (2012). A similar measure has been analyzed in Olson and Kleiber (2017), in which they consider wet spell counts.

Wilks (1998) shows that the Wilks model precisely adjusts probabilities for the events (1,1) and (0,0) between pairs of stations. This is also analyzed in Mehrotra et al. (2006), this time in terms of correlations. This same analysis can be seen in Figure 5a. Clearly, Wilks model is nearly perfect, whereas BNs tend to have more dispersion, particularly for simple models with high regularization parameter (e.g., $k = 3.92$). However, this dispersion errors are mainly concentrated in the lowest observed correlations, as BNs aim to obtain a compromise model for the *whole* JPD by sacrificing *weaker* correlations for the sake of model simplicity. As stated before, the balance between complexity and loss can be achieved by varying the regularization parameter k up to the desired point; for instance, $k = 0.5$ yields very similar results to the Wilks model. The chosen model, with $k = 1$, shows a good compromise between parsimony and loss of correlations for data sets of this size (see section 6).

To further analyze the spatial structure of the generated series, Figures 5b and 5c show an example of correlation between pairs of stations, this time when we condition to Station 7. We observe that the BN is able

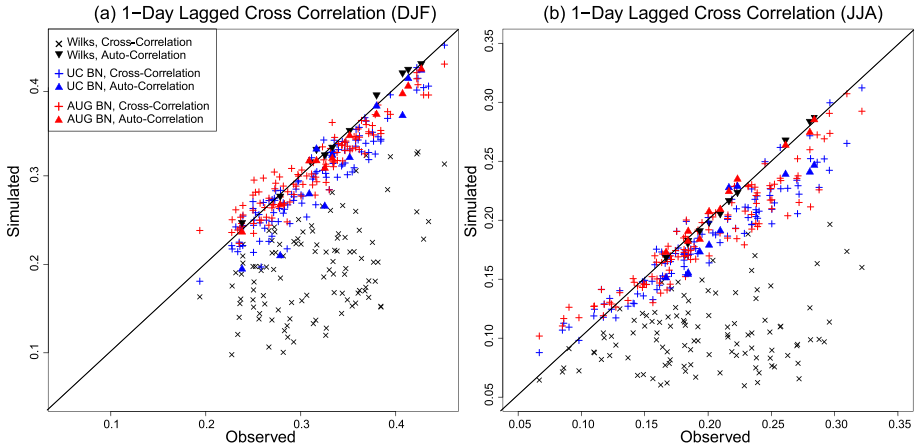

Figure 7. Simulated versus observed 1-day lagged cross-correlations for (a) DJF and (b) JJA for three different models: Wilks, Unconstrained (UC BN), and Augmented (AUG BN) Bayesian network. Triangle-shaped points represent the autocorrelation values between stations. Mean absolute errors (multiplied by 100) are 9.2 (JJA), 12 (DJF) for the Wilks model; 1.7 for both seasons for the Augmented model; and 1.9 (JJA), 2 (DJF) for the Unconstrained BN model.

Table 5
Spatial Validation Summary of Error Results for the Wilks Model, Unconstrained (UN), and Augmented (AUG) Bayesian Network Models for the JJA Season

Measure	Wilks	UN	AUG
Cor	**0.14** (−0.35)	1.18 (−4.13)	1.22 (−4.53)
CondCor	4.10 (12)	**1.90 (-9.9)**	**1.90 (-9.6)**
LaggedCrossCor	9.20 (−21.0)	1.90 (−6.5)	**1.70 (-7.0)**
LaggedAutoCor	**0.31 (-0.56)**	2.27 (−3.9)	0.55 (−0.94)
MultisiteDry	−8.68	**-4.34**	−4.40
MultisiteWet	23.83	−3.71	**-3.63**

Note. For the first four measures, we show the absolute errors multiplied by 100 for easy visualization, that is, $100 \cdot |s-o|$, for the simulated measures (s) and the observed (o) measures. Worst value (multiplied by 100) is shown between parentheses. For the multisite observations (last two rows), we show relative errors in percentage, that is, $100 \cdot \frac{s-o}{o}$. Boldface indicates the best (mean) results for each measure.

to adjust reasonably well to this correlations, whereas Wilks model has more bias. This behavior can be seen very clearly in Figure 6, in which we show all correlation values between pairs of stations conditional to a third one, sorted by their observed value in ascending order. This suggests that there is no real adjustment in the Wilks model for these correlations depending on the condition and that they adopt the mean value between the two: for instance, the correlation of X_1 and X_6, given that variable $X_7 = 0(1)$ obtained directly for data is 0.512(0.297), similar to the values simulated from the BN, 0.512(0.299). However, the Wilks model provides noninformative conditional correlations, 0.437 (0.420); that is, the spatial structure is not consistent for the evidence on the state of station X_7. Not reproducing conditional correlations implies that the model may fail to represent complex orographical or physical (weather states) constraints.

These results raise a warning about the fact that Wilks model simulations are only pairwise coherent and might not be representative of more general (conditional) dependencies encoded in the data set.

Along with spatial coherence, stations are also cross-correlated between time slices due to, for example, the tendency for precipitation systems to move from west to east (spatiotemporal coherence). Wilks (1998) considers the 1-day lagged cross-correlation measure, showing that his model has a considerable bias in this aspect, greatly underestimating these correlation pairs. Figure 7 shows these values for both seasons (as difference in correlations is relevant). BNs clearly outperform the Wilks model and simulate reasonably well lagged cross-correlations. Not surprisingly, the autocorrelations (marked as a triangle in the plot) are well captured by Wilks model, whereas the BNs tends to perform as well as for the cross-correlations (see Tables 5 and 6).

Regarding multisite observations, we show the results in Table 7. We see a clear overestimation of multisite wet days and an underestimation of multisite dry days by the Wilks model. In order to test whether the observed frequencies can be considered a plausible simulation obtained with the Wilks and BN models (among the 250 runs performed for each model), we performed a standard hypothesis test for proportions (Zou et al., 2003), with the null hypothesis $H_0 : p - p_0 = 0$ and alternative hypothesis H_a: $p - p_0 \neq 0$. p corresponds to the mean model proportion of multisite dry/wet days, and p_0 is the corresponding observed frequency.

Table 6
As With Table 5, for DJF Season

Measure	Wilks	UN	AUG
Cor	**0.14** (−0.35)	1.2 (−4.10)	1.2 (−4.60)
CondCor	3.90 (17.0)	**1.7 (9.5)**	**1.7 (9.0)**
LaggedCrossCor	12.0 (−23)	2.0 (−6.8)	**1.7 (-4.6)**
LaggedAutoCor	**0.6 (-0.3)**	2.7 (−6.8)	0.9 (−1.6)
MultisiteDry	−6.26	**-3.14**	−3.97
MultisiteWet	13.99	**-6.54**	−6.64

Table 7
Multisite Dry/Wet Observations Compared to the Observed Values (first row)

Model	Dry JJA	DJF	Wet JJA	DJF
Observed	795	878	193	286
Wilks	726(26.8)	823(32.6)	239(17.9)	326(22.2)
BN	760(30.3)	843(41.5)	186(16.1)	267(18.4)

Note. BN corresponds to the Augmented Bayesian network model, as results are similar for the Unconstrained model. Values are the mean of the 250 simulated series, with the standard deviation shown between parentheses.

The resulting p-values are included in Table 8. For the Wilks model, hypothesis test rejects the null hypothesis with a significance of 99.5% for observations in JJA season and of 97.5% in DJF season. We conclude from the hypothesis test that the BN clearly outperforms the Wilks model in mimicking the proportions of multisite dry and wet observations.

As stated before, we have also considered the multisite spells. Previous studies usually consider spell measures (consecutive sequences of dry/wet days), but they do this for each station, without taking into account multisite spells. In combination with the proportions of multisite dry/wet observations and 1-day lagged cross-correlations, this addresses the spatiotemporal aspect of the simulated series. We compare the percentiles of the distribution of multisite spells, in which a day is considered wet/dry if it is wet/dry in over 90% of the stations. These are shown in Figure 8. Even if there is a clear underestimation of dry spells for both seasons and models, the BN model also outperforms Wilks model. Multisite wet spells are not very long in this small region, and results are very similar for both methods.

Finally, we analyze some measures whose aim is to compare the Wilks WG against the BN models in terms of temporal consistency alone, as is done in, for example, Mehrotra et al. (2006). The measures are computed for each station and the mean value and standard deviation are shown, with each measure explained in Table 9.

Results are shown in Table 10 and for JJA season, with DJF having similar results. A similar performance is found for Wilks and BN models, with largest differences for DW and WW (transition probabilities) which, as explained before, are directly calibrated with the Wilks model (note how this exhibits a similar behavior as that of pairwise correlations). In the case of the BNs, these probabilities are not direct parameters of the models and, therefore, have slightly higher errors (smaller for the case of the Augmented model, which forces these transition probabilities to be considered by the model). The relative errors are 0.011 for DW 0.006 for WW.

In some cases, the models perform poorly in capturing spells, which is a well-known flaw in current WGs. In particular, all models highly underestimate droughts (extreme droughts by 10% for JJA and up to 30% for DJF); this could be attributed to considering a first-order Markov process (Ailliot et al., 2015; Mehrotra et al., 2006; Wilks & Wilby, 1999) and is slightly alleviated by the Augmented model, as opposed to the Unconstrained version.

8. Conclusions and Future Lines of Work

We introduced Bayesian networks as a new methodology with a well-established theoretical background for simulating discrete multisite precipitation series, mimicking spatial and temporal aspects simultaneously. The performance of this method was assessed thorough a spatiotemporal validation analysis using the Wilks model as benchmark.

Table 8
P Values of the Proportions Hypothesis Test with Null Hypothesis $H_0 : p - p_0 = 0$ *and Alternative Hypothesis* $H_a : p - p_0 \neq 0$, *where p is the Model Proportion and* p_0 *is the Corresponding Observed Frequency of Dry/Wet Multisite Observations And Both Seasons*

Model	Dry JJA	DJF	Wet JJA	DJF
Wilks	0.004	0.024	0.001	0.012
BN	0.143	0.152	0.588	0.228

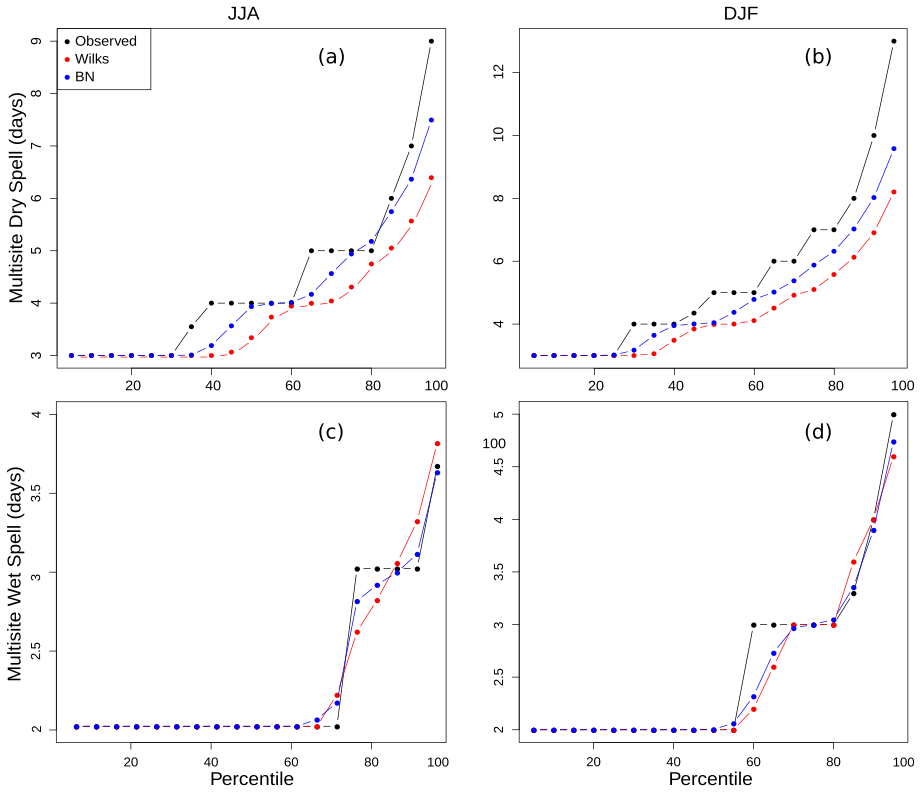

Figure 8. Percentile values (from 5th to 95th) from the sample of multisite dry spell values for (a) JJA and (b) DJF and for multisite wet spells for (c) JJA and (d) DJF seasons. Only spells affecting more than 90% of the territory (10 out of 11 stations) have been considered. For dry spells, spells greater than three days are considered. For wet spells, since they are shorter, we consider spells starting in two consecutive days. Results for the Augmented Bayesian network model are shown. Results for the Unconstrained model are similar.

Instead of focusing on pairwise dependencies among stations, BNs try to represent the whole probability distribution, considering all relevant conditional independencies (including pairwise as a particular case). Thus, transition probabilities and pairwise correlations are expectedly better represented by Wilks model, since this model directly adjusts them from the data set. However, when more complex features like lagged cross-correlations, multisite dry spells, or global proportions are analyzed, BNs perform better than Wilks model. This illustrates the potential of this new approach, although further analysis with more complex examples is necessary to fully test the advantages of this new methodology.

BNs are a sound machine learning technique supported by well-known theory and algorithms and offer some potential advantages compared to other approaches due to

Table 9
Table of Measures Used for the Performance Assessment of Temporal Coherence

Name	Description
DW	Dry-wet transition probability
WW	Wet-wet transition probability
WetFreq	Percentage of wet days
WetSpell (WS)	Mean of wet spells
DrySpell (DS)	Mean of dry spells
WSAnnualMean	Mean of longest annual wet spells
DSAnnualMean	Mean of longest annual dry spells
WSAnnualMax	Maximum of longest annual wet spells
DSAnnualMax	Maximum of longest annual dry spells

Note. All spell measures are in days (≥2) and probabilities are measured in percentage.

Table 10
Results for the Wilks Model and Bayesian Network Models for the JJA Season

Measure	Observed	Wilks	UN	AUG
DW	28.72 (2.18)	**0.06** (0.04)	0.86 (0.55)	0.32 (0.16)
WW	51.89 (4.85)	**0.09** (0.06)	1.33 (0.56)	0.33 (0.29)
WetFreq	37.44 (3.61)	**0.09** (0.06)	0.17 (0.12)	0.1 (0.07)
WetSpell	3.10 (0.24)	**0.05** (0.04)	0.08 (0.07)	**0.05** (0.04)
DrySpell	4.57 (0.26)	0.16 (0.09)	0.15 (0.11)	**0.07** (0.08)
WSAnnualMean	5.88 (0.89)	0.30 (0.18)	0.35 (0.26)	**0.28** (0.19)
DSAnnualMean	11.34 (1.13)	0.80 (0.50)	0.72 (0.51)	**0.51** (0.38)
WSAnnualMax	10.73 (1.90)	0.96 (0.80)	**0.85** (0.87)	1.03 (0.74)
DSAnnualMax	21.18 (4.05)	2.55 (2.58)	2.54 (2.46)	**2.5** (2.18)

Note. "Observed" column shows the mean (and standard deviation) of the observed values of the stations. The three last columns indicate absolute errors, that is, $|s-o|$, between the diagnostics for simulations (s) and observations (o), again in terms of mean and standard deviations. Boldface indicates the best (mean) results for each index.

1. Generalization: There is no need for an ad hoc adjustment of probabilities or correlations. BNs adjust the whole distribution in a robust way, adapting to the local dependencies and characteristics implied by existing records.
2. Flexible complexity: It may not be enough to just adjust transition probabilities for one particular location; it may require data from other stations and/or previous time slices. It may also require information from other locations *combined*.On the other hand, data may be scarce for computing one probability value. A BN will automatically adapt to these situations.
3. Interpretability: A general property of BNs is their interpretability through their DAGs, which can be easily interpreted and the dependence relationships analyzed in a user-friendly way.
4. Expert knowledge: Apart from the information gathered from the data set, probabilistic models allow for expert knowledge to be introduced in the model in a robust way. This feature, combined with interpretability, potentially offers the possibility to further trim the model to user needs or to characteristics not present in the existing data.

On the other hand, a disadvantage of this methodology is the increasing complexity and computation time required to both learn and simulate from the models when the number of variables (and thus the number of arcs) increases. There are a number of studies showing that BNs are tractable for problems from tens to a few hundreds of variables (Scutari et al., 2019). In problems of larger complexity (hundreds or thousands of stations), the complexity of the resulting model can be controlled through the regularization parameter k (see section 6) or using more efficient learning and inference algorithms (Scutari et al., 2019).

This work paves the way for further investigating the potential of BNs in the field of WGs. We focused on the discrete step of the precipitation modeling process, with the continuous counterpart (i.e., precipitation amount on wet days) remaining to be studied in depth. There are several options: Discrete series, either binary or with more bins, could be coupled with an already implemented continuous model for evaluating the added value to the continuous aspect of the problem. The next step should be making use of continuous BNs, either as a separate model for the precipitation amount or even mixing discrete and continuous nodes in the same model (hybrid BNs). It should be noted that this poses a considerable challenge, as BNs with continuous nodes have restrictions (both in the distributions for the continuous nodes and the placement of the arcs) that hinder their application to the problem of modeling the precipitation amount probability distribution, usually considered a gamma distribution.

Another potential is exploring conditional WGs extending the proposed methodology considering nodes with GCM outputs, either at a gridbox level or clustered as weather types. This could potentially improve the overall model performance in certain measures like multisite spells and, on the other hand, provide stochastic climate change projections. Finally, spell reproducibility could also be improved by adding an additional node or set of nodes to track more than one day, making it a higher-order Markov model.

In order to facilitate reproducibility of the results and testing the performance of the proposed methodology, we have prepared an R package, *BNWeatherGen*—which builds on the R package *bnlearn* (Scutari, 2010)—and a Jupyter notebook illustrating the creation and use of some of the models presented in this work

(both available at https://github.com/MNLR/BNWeatherGen). The notebook also contains an additional larger example (44 stations) used to illustrate the complexity and scalability of the proposed methodology, so interested readers can explore different learning alternatives. In particular, for the 44 stations example, learning and simulation can be undertaken in a personal computer in less than 10 min using the default configuration of the package.

Acknowledgments
We acknowledge funding provided by the project MULTI-SDM (CGL2015-66583-R, MINECO/FEDER). Station data were provided by http://www.ecad.eu (Klein Tank et al., 2002). Bayesian networks package for R (bnlearn) was used for all the computations (Scutari, 2010).

References

Aguilera, P. A., Fernández, A., Fernández, R., Rumí, R., & Salmerón, A. (2011). Bayesian networks in environmental modelling. *Environmental Modelling & Software*, *26*(12), 1376–1388. https://doi.org/10.1016/j.envsoft.2011.06.004

Ailliot, P., Allard, D., Monbet, V. A., & Naveau, P. (2015). Stochastic weather generators: An overview of weather type models. *Journal de la Societe Franaise de Statistique*, *156*(1), 101–113.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. *Selected Papers of Hirotugu Akaike* (pp. 199–213). New York: Springer. https://doi.org/10.1007/978-1-4612-1694-0_15

Boneh, T., Weymouth, G. T., Newham, P., Potts, R., Bally, J., Nicholson, A. E., & Korb, K. B. (2015). Fog forecasting for Melbourne airport using a Bayesian decision network. *Weather and Forecasting*, *85*, 1218–1233. https://doi.org/10.1175/WAF-D-15-0005.1

Booij, M. J. (2005). Impact of climate change on river flooding assessed with different spatial model resolutions. *Journal of Hydrology*, *303*(1), 176–198. https://doi.org/10.1016/j.jhydrol.2004.07.013

Borunda, M. A. A., Jaramillo, O. A., Reyes, A., & Ibargengoytia, P. H. (2016). Bayesian networks in renewable energy systems: A bibliographical survey. *Renewable and Sustainable Energy Reviews*, *62*, 32–45. https://doi.org/10.1016/j.rser.2016.04.030

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Cano, R., Sordo, C., & Gutiérrez, J. M. (2004). Applications of Bayesian networks in meteorology. In J. A. Gámez, S. Moral, & A. Salmerón (Eds.), *Advances in Bayesian networks* (pp. 309–328). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-39879-0_17

Castillo, E., Gutirrez, J. A. M., & Hadi, A. S. (1997). *Expert systems and probabilistic network models*. New York, NY: Springer.

Chickering, D. M., Heckerman, D., & Meek, C. (2004). Large-sample learning of Bayesian networks is NP-hard. *Journal of Machine Learning Research*, *5*, 1287–1330.

Cofiño, A. S., Cano, R., Sordo, C., & Gutiérrez, J. M. (2002). Bayesian networks for probabilistic weather prediction. In *Proceedings of the 15th European Conference on Artificial Intelligence (ECAI-2002)* (pp. 695–700). Lyon: IOS Press.

Das, M., & Ghosh, S. K. (2014). A probabilistic approach for weather forecast using spatio-temporal inter-relationships among climate variables. In *2014 9th International Conference on Industrial and Information Systems (ICIIS)* (pp. 1–6). Gwalior: IEEE.

Das, M., & Ghosh, S. K. (2017). semBnet: A semantic Bayesian network for multivariate prediction of meteorological time series data, *93*, 192–201. https://doi.org/10.1016/j.patrec.2017.01.002

Duan, J., Selker, J., & Grant, G. E. (2007). Evaluation of probability density functions in precipitation models for the Pacific Northwest. *JAWRA Journal of the American Water Resources Association*, *34*(3), 617–627. https://doi.org/10.1111/j.1752-1688.1998.tb00959.x

Fiener, P., & Auerswald, K. (2009). Spatial variability of rainfall on a sub-kilometre scale. *Earth Surface Processes and Landforms*, *34*(6), 848–859. https://doi.org/10.1002/esp.1779

Garrote, L., Molina, M., & Mediero, L. (2008). Learning Bayesian networks from deterministic rainfall-runoff models and Monte Carlo simulation. In R. J. Abrahart, L. M. See, & D. P. Solomatine (Eds.), *Practical hydroinformatics: Computational intelligence and technological developments in water applications* (pp. 375–388). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-79881-1_27

Glover, F. (1990). Tabu search: A tutorial. *Interfaces*, *20*(4), 74–94. https://doi.org/10.1287/inte.20.4.74

Gutiérrez, J. M., Maraun, D., Widmann, M., Huth, R., Hertig, E., Benestad, R., et al. (2019). An intercomparison of a large ensemble of statistical downscaling methods over Europe: Results from the VALUE perfect predictor cross-validation experiment. *International Journal of Climatology*, *39*(9), 3750–3785. https://doi.org/10.1002/joc.5462

Haile, A. T., Rientjes, T., Gieske, A., & Gebremichael, M. (2009). Rainfall variability over mountainous and adjacent lake areas: The case of Lake Tana Basin at the source of the Blue Nile River. *Journal of Applied Meteorology and Climatology*, *48*(8), 1696–1717. https://doi.org/10.1175/2009JAMC2092.1

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.), *Springer Series in Statistics*. New York: Springer-Verlag. https://doi.org/10.1007/978-0-387-84858-7

Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, *20*(3), 197–243. https://doi.org/10.1023/A:1022623210503

Hellman, S., McGovern, A., & Xue, M. (2012). Learning ensembles of continuous Bayesian networks: An application to rainfall prediction, *2012 Conference on Intelligent Data Understanding* (pp. 112–117). Boulder, CO: IEEE.

Keller, D. E., Fischer, A. M., Frei, C., Liniger, M. A., Appenzeller, C., & Knutti, R. (2015). Implementation and validation of a Wilks-type multi-site daily precipitation generator over a typical alpine river catchment. *Hydrology and Earth System Sciences*, *19*(5), 2163–2177. https://doi.org/10.5194/hess-19-2163-2015

Kleiber, W., Katz, R. W., & Rajagopalan, B. (2012). Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes. *Water Resources Research*, *48*, W01523. https://doi.org/10.1029/2011WR011105

Klein Tank, A. M. G., Wijngaard, J. B., Knnen, G. P., Bhm, R., Demare, G., Gocheva, A., et al. (2002). Daily dataset of 20th-century surface air temperature and precipitation series for the European climate assessment. *International Journal of Climatology*, *22*(12), 1441–1453. https://doi.org/10.1002/joc.773

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models, principles and techniques*. Cambridge, MA: The MIT Press.

Lam, W., & Bacchus, F. (1994). Learning Bayesian belief networks: An approach based on the Mdl principle. *Computational Intelligence*, *10*(3), 269–293. https://doi.org/10.1111/j.1467-8640.1994.tb00166.x

Madadgar, S., & Moradkhani, H. (2014). Spatio-temporal drought forecasting within Bayesian networks. *Journal of Hydrology*, *512*, 134–146. https://doi.org/10.1016/j.jhydrol.2014.02.039

Manzanas, R., Gutirrez, J. M., Bhend, J., Hemri, S., Doblas-Reyes, F. J., Torralba, V., et al. (2019). Bias adjustment and ensemble recalibration methods for seasonal forecasting: A comprehensive intercomparison using the C3S dataset. *Climate Dynamics*, *53*(3), 1287–1305. https://doi.org/10.1007/s00382-019-04640-4

Maraun, D., Wetterhall, F., Ireson, A. M., Chandler, R. E., Kendon, E. J., Widmann, M., et al. (2010). Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, *48*, RG3003. https://doi.org/10.1029/2009RG000314

Mehrotra, R., & Sharma, A. (2009). Evaluating spatio-temporal representations in daily rainfall sequences from three stochastic multi-site weather generation approaches. *Advances in Water Resources*, *32*(6), 948–962. https://doi.org/10.1016/j.advwatres.2009.03.005

Mehrotra, R., Srikanthan, R., & Sharma, A. (2006). A comparison of three stochastic multi-site precipitation occurrence generators. *Journal of Hydrology*, *331*(1-2), 280–292. https://doi.org/10.1016/j.jhydrol.2006.05.016

Nandar, A. (2009). Bayesian network probability model for weather prediction. In *2009 International Conference on the Current Trends in Information Technology (CTIT)* (pp. 120–124). Dubai: IEEE. https://doi.org/10.1109/CTIT.2009.5423132

Niedermayer, D. (2008). An introduction to Bayesian networks and their contemporary applications. In D. E. Holmes, & L. C. Jain (Eds.), *Innovations in Bayesian networks: Theory and applications* (pp. 117–130). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-85066-3_5

Olson, B., & Kleiber, W. (2017). Approximate Bayesian computation methods for daily spatiotemporal precipitation occurrence simulation. *Water Resources Research*, *53*, 3352–3372. https://doi.org/10.1002/2016WR019741

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA: Elsevier. https://doi.org/10.1016/C2009-0-27609-4

Pearl, J. (2013). *Causality* (2nd ed.). Cambridge, UK: Cambridge University Press.

Phan, T. D., Smart, J. C. R., Capon, S. J., Hadwen, W. L., & Sahin, O. (2016). Applications of Bayesian belief networks in water resource management: A systematic review. *Environmental Modelling & Software*, *85*, 98–111. https://doi.org/10.1016/j.envsoft.2016.08.006

Pourret, O., Naim, P., & Marcot, B. (2008). *Bayesian networks: A practical guide to applications*. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470994559

Richardson, C. W. (1981). Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research*, *17*(1), 182–190. https://doi.org/10.1029/WR017i001p00182

Ropero, R. F., Flores, M. J., Rumí, R., & Aguilera, P. A. (2017). Applications of hybrid dynamic Bayesian networks to water reservoir management. *Environmetrics*, *28*(1), e2432. https://doi.org/10.1002/env.2432

Schlabing, D., Frassl, M., Eder, M. M., Rinke, K., & Brdossy, A. A. (2014). Use of a weather generator for simulating climate change effects on ecosystems: A case study on Lake Constance. 61 https://doi.org/10.1016/j.envsoft.2014.06.028

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

Scutari, M. (2010). Learning Bayesian networks with the bnlearn R package. *Journal of Statistical Software*, *35*(3), 1–22. https://doi.org/10.18637/jss.v035.i03

Scutari, M., & Denis, J.-B. (2014). *Bayesian networks: With examples in R*: Chapman and Hall/CRC.

Scutari, M., Graafland, C. E., & Gutiérrez, J. M. (2019). Who learns better Bayesian network structures: Accuracy and speed of structure learning algorithms. *International Journal of Approximate Reasoning*, *115*, 235–253. https://doi.org/10.1016/j.ijar.2019.10.003

Scutari, M., Vitolo, C., & Tucker, A. (2019). Learning Bayesian networks from big data with greedy search: Computational complexity and efficient implementation. *Statistics and Computing*, *29*(5), 1095–1108. https://doi.org/10.1007/s11222-019-09857-1

Sharma, A., & Goyal, M. K. (2016). Bayesian network for monthly rainfall forecast: A comparison of K2 and MCMC algorithm. *International Journal of Computers and Applications*, *38*(4), 199–206. https://doi.org/10.1080/1206212X.2016.1237131

Sivia, D., & Skilling, J. (2006). *Data Analysis, A Bayesian Tutorial* (2nd ed.). Oxford: Oxford University Press.

Smail, L. (2018). Bayesian network model for temperature forecasting in Dubai. *AIP Conference Proceedings*, *2025*(1), 100,006. https://doi.org/10.1063/1.5064935

Uusitalo, L. (2007). Advantages and challenges of Bayesian networks in environmental modelling. *Ecological Modelling*, *203*(3–4), 312–318. https://doi.org/10.1016/j.ecolmodel.2006.11.033

Widmann, M., Bedia, J., Gutirrez, J. A. M., Bosshard, T., Hertig, E., Maraun, D., et al. (2019). Validation of spatial variability in downscaling results from the VALUE perfect predictor experiment. *International Journal of Climatology*, *39*(3819–3845). https://doi.org/10.1002/joc.6024

Wilks, D. S. (1998). Multisite generalization of a daily stochastic precipitation generation model. *Journal of Hydrology*, *210*(1-4), 178–191. https://doi.org/10.1016/S0022-1694(98)00186-3

Wilks, D. S., & Wilby, R. L. (1999). The weather generation game: A review of stochastic weather models. *Progress in Physical Geography: Earth and Environment*, *23*(3), 329–357. https://doi.org/10.1177/030913339902300302

Yates, D., Gangopadhyay, S., Rajagopalan, B., & Strzepek, K. (2003). A technique for generating regional climate scenarios using a nearest-neighbor algorithm. *Water Resources Research*, *39*(7), 1199. https://doi.org/10.1029/2002WR001769

Zhou, L., Meng, Y., & Abbaspour, K. C. (2019). A new framework for multi-site stochastic rainfall generator based on empirical orthogonal function analysis and Hilbert-Huang transform. *Journal of Hydrology*, *575*, 730–742. https://doi.org/10.1016/j.jhydrol.2019.05.047

Zou, K. H., Fielding, J. R., Silverman, S. G., & Tempany, C. M. (2003). Hypothesis testing I: Proportions. *Radiology*, *226*(3), 609–613 https://doi.org/10.1148/radiol.2263011500