A multivariate and space-time stochastic weather generator using a latent Gaussian framework

Said Obakrim, Lionel Benoit, Denis Allard

A multivariate and space-time stochastic weather generator using a latent Gaussian framework

Said Obakrim[1,2*], Lionel Benoit[1†] and Denis Allard[1†]

[1*]Biostatistique et processus SPatiaux (BioSP), INRAE, Avignon, 84914, France.
[2*]Institute of Earth Surface Dynamics, University of Lausanne, Lausanne, Switzerland.

*Corresponding author(s). E-mail(s): said.obakrim@unil.ch;
Contributing authors: lionel.benoit@inrae.fr; denis.allard@inrae.fr;
†These authors contributed equally to this work.

Abstract

Stochastic weather generators are probabilistic tools used to simulate synthetic weather time series whose statistics resemble those observed. These tools face difficulties when it comes to accurately simulating multiple meteorological variables in space and time, because they necessitate models that can capture the complex inter-variable and space-time dependencies. We propose a new multivariate space-time weather generator, called MSTWeatherGen, which takes advantage of the recent development of multivariate space-time covariance functions to model and simulate different weather variables, including temperature, precipitation, wind speed, humidity, and solar radiation, across space and time. Specifically, we employ an approach that involves a non-linear and non-stationary marginal transformation of a multivariate Gaussian random field, characterized by a stationary and non-separable spatio-temporal multivariate cross-covariance function. To further address the time-varying nature of the weather variables, we split the time domain into states called weather types. The method is assessed on the Provence-Alpes-Côte-d'Azur region in France, which is characterized by heterogeneous topography and meteorological conditions. Evaluation results demonstrate the effectiveness of this new stochastic weather generator in reproducing a wide range of weather statistics, including highly non linear indicators such as heat wave or fire weather index.

Keywords: Weather types, Gaussian random field, Multivariate covariance function, space-time covariance function, heat waves, Fire Weather Index

1 Introduction

In environmental sciences, long-term time series of meteorological variables are essential inputs for a wide range of deterministic models aiming to explore the impact of weather on environmental processes such as hydrology (Milly et al, 2005), agronomy (Lobell and Field, 2007), or wildfires (Westerling et al, 2006). However, observational or reanalysis data may be insufficient for a thorough uncertainty analysis of the environmental response to weather forcing, in particular because they provide only a single scenario of meteorological conditions (Semenov et al, 1998). For impact studies focusing on uncertainty assessment, large ensembles of weather data can therefore be beneficial, especially when focusing on rare weather events with low probability of occurrence but high impact (Leutbecher, 2019). While numerical weather models (NWMs) can simulate realistic ensembles of weather data by varying the initial conditions, these simulations are computationally intensive and require considerable expertise to set-up and run (Mearns et al, 2001). As an alternative, a class of statistical models known as stochastic weather generators (SWGs) (Wilks and Wilby, 1999) has been developed to produce realistic ensembles of meteorological time series that match the statistical properties of the observed data. SWGs are less computationally expensive than NWMs and have been successfully applied to the generation of various variables such as precipitation, temperature, or wind (Richardson, 1981; Ailliot et al, 2020; Peleg et al, 2017).

The seminal work of Richardson (1981) proposed to first model precipitation as a Markov chain-exponential model, and to subsequently model the three other variables (minimum temperature, maximum temperature and solar radiation) conditionally on the wet or dry status determined by the precipitation model. This binary classification of days into wet and dry was the precursor of rain typing, that is, the clustering of the timeline into periods with homogeneous precipitation properties (Ailliot et al, 2015). The idea of clustering days with a similar statistical signature has later been extended to the full set of meteorological variables of interest, leading to the concept of weather types (Ailliot et al, 2015). Several methods for constructing weather types are documented in the literature, including empirical orthogonal functions (Wanner et al, 2001), hidden Markov models (Varin et al, 2011), and clustering algorithms (Boé et al, 2006).

Past efforts in improving the realism of environmental modeling led to the emergence of distributed impact models, which enable the simulation of the process of interest at many geographical locations, often distributed on a regular grid (see e.g., Fatichi et al (2016) for application in hydrology, and Han et al (2019) for application

in agronomy). This in turn requires the simulation of spatially explicit meteorological variables in order to provide inputs for the distributed environmental models. Gaussian random fields (GRFs) are a common keystone for the statistical modeling of spatially explicit variables, in particular in the framework of Geostatistics (Cressie, 2015; Chilès and Delfiner, 2012) and Gaussian Processes (Williams and Rasmussen, 2006). In the context of meteorological variables, SWGs based of GRFs have for instance been applied to the simulation of wind (Baxevani and Lenzi, 2018) or precipitation (Allcroft and Glasbey, 2003). GRFs are primarily restricted to the modeling of normally distributed variables, but the application of parametric transform functions to latent GRFs has proved to be an efficient way to model non-Gaussian meteorological variables, in particular precipitation (Allard and Bourotte, 2015; Benoit and Mariethoz, 2017; Paschalis et al, 2013). Latent GRFs also enable the modeling of zero inflated data using truncated Gaussian processes, which is often leveraged for the joint modeling of precipitation occurrence and intensity (Ailliot et al, 2009; Paschalis et al, 2013; Benoit et al, 2018). A wide range of transform functions have been proposed to link latent GRFs to the actual meteorological variable to model, for instance the power (Ailliot et al, 2009) or power-exponential (Allard and Bourotte, 2015) transformation for precipitation, or more flexible Box-Cox transformation (Box and Cox, 1964) or semiparametric Ordered Quantile Normalization (OQN) (Bartlett, 1947; Peterson and Cavanaugh, 2019) in cases where few prior information is available about the distribution of the target meteorological variable (Sparks et al, 2018).

In the framework of (possibly truncated and transformed) Gaussian processes, the spatial dependence structure of the meteorological variables is modeled by the covariance function of the latent Gaussian field. A natural extension of this framework is to use a space-time GRF to build a SWG able to not only capture and reproduce the spatial distribution of meteorological variables, but also their dynamics. This has been made possible by the development of non-separable space-time covariance functions which are sufficiently flexible to model the space-time dependence structure of most meteorological variables (Gneiting, 2002; Gneiting et al, 2006; Chen et al, 2021; Porcu et al, 2021). This approach has been applied in particular to the modeling of precipitation fields at high resolution (Baxevani and Lennartsson, 2015; Benoit et al, 2018; Boutigny et al, 2023).

Going one step further, it is appealing to model some dependencies between variables in addition to the space-time dependencies within a single meteorological variable. This allows SWGs to simulate meteorological variables that are consistent

with each other, which is crucial when investigating the impact of weather on processes that are simultaneously impacted by several meteorological variables (e.g., evapo-transpiration (Kimball et al, 2023) or wildfires (Van Wagner, 1987)) and also when examining compounds extreme events (Bevacqua et al, 2023; Dabhi et al, 2021). Despite the considerable interest of designing multivariate space-time SWGs, few studies have yet explored this approach. These include the SWG of Verdin et al (2015, 2019) that considers precipitation occurrence as well as minimum and maximum temperature with spatial correlations modeled by Gaussian processes. However, as in Richardson's model (Richardson, 1981), the intensity of precipitation is modeled separately from its occurrence, and the minimum and maximum temperatures are modeled conditionally on precipitation occurrence and not jointly with precipitation. In a slightly different approach, Sparks et al (2018) proposed the IMAGE model which is based on latent Gaussian variables with temporal correlations modeled by an autoregressive model (AR) and spatial correlations modeled by empirical orthogonal functions (EOFs) applied to the AR parameters. In general, SWGs based on latent Gaussian processes require valid covariance functions to model the structure of dependence in space, in time, and between variables. However, fully multivariate and space-time covariance models are complex to design. To simplify the problem the aforementioned SWGs chose to disregard some parts of the dependence structure at the cost of over-simplistic dependencies within or between the simulated meteorological variables.

To overcome this limitation, we build on recent progress in the field of multivariate space-time covariance models (De Iaco et al, 2019; Apanasovich and Genton, 2010; Bourotte et al, 2016; Genton and Kleiber, 2015; Allard et al, 2022) to design a fully multivariate space-time SWG. We start from the work of Bourotte (2016) and replace the original covariance of the latent Gaussian fields (Bourotte et al, 2016) by the more flexible class of multivariate space-time covariance proposed by Allard et al (2022). The resulting space-time stochastic weather generator, called MSTWeatherGen, is therefore especially designed to simulate jointly several meteorological variables, for instance precipitation, humidity, wind, solar radiation, and maximum and minimum temperature, over space and time. In addition, we allow the marginal distribution of the meteorological variables to be flexible thanks to an extension of the Ordered Quantile Normalization (Peterson and Cavanaugh, 2019) to truncated distributions.

The remaining of the paper is organized as follows. The proposed model with all its components is first presented in Section 2. Next, Section 3 details how all parameters can be estimated. The simulation algorithm is subsequently presented in Section 4. Then, Section 5 shows how seasonal effects are taken into account. Section 6

evaluates the performance of the SWG on a case study focusing on the Provence-Alpes-Côte d'Azur region in southern France. Finally, Section 7 provides some elements of discussion and summarizes the study.

2 Multivariate and space-time stochastic weather generator

2.1 General framework

For the sake of simplicity, we first present a framework for weather variables that do not exhibit a seasonal cycle. Then in Section 5, we will show how seasonality can be taken into account.

The p meteorological variables studied over a space-time domain $\mathcal{D} \times \mathcal{T} \subset \mathbb{R}^2 \times \mathbb{R}$ are represented by a p-dimensional stochastic process $\mathbf{Y}(\mathbf{s}, t) = [Y_i(\mathbf{s}, t)]_{i=1}^p$, with $(\mathbf{s}, t) \in \mathcal{D} \times \mathcal{T}$. We assume that the time domain can be split into K sub-domains, referred to hereafter as weather types, in which all p meteorological variables have homogeneous statistics (i.e., are stationary in time). The weather type at any time $t \in \mathcal{T}$ is represented by a discrete one-dimensional stochastic process $X(t)$ taking values in $\mathcal{S} = \{1, \ldots, K\}$. We also assume the existence of K p-dimensional latent Gaussian random fields $\mathbf{Z}_k(\mathbf{s}, t) = [Z_{k,i}(\mathbf{s}, t)]_{i=1}^p$, for $k = 1, \ldots, K$, such as within each weather state k each component of the stochastic vector $\mathbf{Y}(\mathbf{s}, t)$ is related to the corresponding component of $\mathbf{Z}_k(\mathbf{s}, t)$ through non-linear transformation functions $\Psi_{k,i,\mathbf{s}}$, $i = 1, \ldots, p$:

$$Y_i(\mathbf{s}, t) = \Psi_{k,i,\mathbf{s}}(Z_{k,i}(\mathbf{s}, t)), \quad \text{with } k = X(t). \tag{1}$$

We also suppose without loss of generality that every component of each latent random field $\mathbf{Z}_k(\mathbf{s}, t)$ has zero mean and unit variance, i.e., $\mathbb{E}(Z_{k,i}(\mathbf{s}, t)) = 0$ and $\mathbb{E}(Z_{k,i}^2(\mathbf{s}, t)) = 1$, $\forall (\mathbf{s}, t) \in \mathbb{R}^2 \times \mathbb{R}$, $\forall i = 1, \ldots, p$ and $k \in \mathcal{S}$. In addition, we assume that each random field $\mathbf{Z}_k(\mathbf{s}, t)$ is second-order stationary in space and time, i.e., its covariance function depends only on the space-time lag $(\mathbf{h}, u)$:

$$\mathrm{Cov}(Z_{k,i}(\mathbf{s}, t), Z_{k,j}(\mathbf{s} + \mathbf{h}, t + u)) = C_{k,ij}(\mathbf{h}, u), \tag{2}$$

$\forall i, j = 1, \ldots, p$, $\forall (\mathbf{s}, t) \in \mathbb{R}^2 \times \mathbb{R}$ and $\forall (\mathbf{h}, u) \in \mathbb{R}^2 \times \mathbb{R}$.

The model (1) requires the parameterization of three elements: the distribution of the weather type process $X(t)$, all non-linear transformation functions $\Psi_{k,i,\mathbf{s}}$, and the multivariate covariance of each p-dimensional Gaussian random field $\mathbf{Z}_k(\mathbf{s}, t)$. The

subsequent subsections will detail the proposed parameterization for each of these elements.

2.2 Weather types

Following Flecher et al (2010) and Ailliot et al (2015), we suppose that the temporal domain is discretized into weather types prior to the statistical modeling of $\mathbf{Y}$. To simplify the modeling we also assume that the weather types are common for the whole spatial domain. This assumption, which may not be realistic at large scales, is reasonable at the regional scale as shown in the case study presented in Section 6.

Weather types are estimated using a clustering algorithm based on the characteristics of the multivariate weather process over the whole region. This task requires the time t to be discrete. From now on, we shall thus assume $t \in \mathbb{N}$. For estimating the weather types, we first reduce the dimensionality of the data using Principal Tensor Analysis (PTA) (Leibovici and Sabatier, 1998). Once the data is compressed, we employ a Gaussian mixture model estimated by the expectation maximization (EM) algorithm to identify the weather types. The optimal number of weather types is determined by Bayesian Information Criterion (BIC) minimization. For this task, we use the `R` package `mclust` (Scrucca et al, 2023). After estimating the weather types, the process $X(t)$ is modeled as a first-order Markov chain, so that:

$$\mathbb{P}(X(t) \mid X(t-1), \ldots, X(0)) = \mathbb{P}(X(t) \mid X(t-1)), \quad t \in \mathbb{N}^*. \tag{3}$$

A higher order Markov chain could also be used, but for the simplicity of the exposition we stick to a first order Markov chain. To account for non-stationary transitions between weather types that can vary along time, the Markov chain is supposed to be non-homogeneous. Hence, $X(t)$ is characterized by transition probabilities that are functions of time:

$$\pi_{kl}(t) = \mathbb{P}(X(t) = l \mid X(t-1) = k), \;\; k, l \in \mathcal{S}. \tag{4}$$

2.3 Transformation functions

Given a weather state $X(t) = k \in \mathcal{S}$, each transformation function $\Psi_{k,i,\mathbf{s}}$ is a non-linear function mapping from the latent Gaussian random field $Z_{k,i}(\mathbf{s}, t)$ to the meteorological variable $Y_i(s, t)$. When the mapping is monotonic, the inverse of the transformation function can be seen as the normalization transformation function that transforms each meteorological variable at each spatial location into a standard Gaussian distribution. In this study, we consider the Ordered Quantile Normalization (OQN) method

developed in Peterson and Cavanaugh (2019) to parameterize our transform functions because it allows modeling beyond the observed range of meteorological variables. In addition, we extend this framework to truncated Gaussian variables. Given a meteorological variable $Y_i(\mathbf{s},t)$ at space-time coordinates $(\mathbf{s},t)$ with $X(t)$ in state $k \in \mathcal{S}$, the inverse OQN transformation takes the form:

$$Y_i(\mathbf{s},t) = \Psi_{k,i,\mathbf{s}}\left(Z_{k,i}(\mathbf{s},t)\right) = \begin{cases} F^{-1}_{k,i,\mathbf{s}}\left(\Phi_{T_{k,i,\mathbf{s}}}\left(Z_{k,i}(\mathbf{s},t)\right)\right) & \text{if } Z_{k,i}(\mathbf{s},t) > T_{k,i,\mathbf{s}} \\ 0 & \text{if } Z_{k,i}(\mathbf{s},t) \leq T_{k,i,\mathbf{s}}, \end{cases} \tag{5}$$

where $F_{k,i,\mathbf{s}}$ is the Cumulative Distribution Function (CDF) of the meteorological variable $Y_i(\mathbf{s},t)$ and $\Phi_{T_{k,i,\mathbf{s}}}$ is the left truncated Gaussian CDF (Burkardt, 2014):

$$\Phi_{T_{k,i,\mathbf{s}}}\left(Z_{k,i}(\mathbf{s},t)\right) = \begin{cases} \left(\Phi\left(Z_{k,i}(\mathbf{s},t)\right) - \Phi(T_{k,i,\mathbf{s}})\right) / \left(1 - \Phi(T_{k,i,\mathbf{s}})\right) & \text{if } Z_{k,i}(\mathbf{s},t) > T_{k,i,\mathbf{s}} \\ 0 & \text{if } Z_{k,i}(\mathbf{s},t) \leq T_{k,i,\mathbf{s}}, \end{cases}$$

where Φ is the $(0,1)$ Gaussian CDF.

2.4 Covariance function

Given a weather state $X(t) = k \in \mathcal{S}$, the multivariate random field $\mathbf{Z}_k(\mathbf{s},t)$ controls the intra- and inter-variable space-time correlations of $\mathbf{Y}(\mathbf{s},t)$. Since $\mathbf{Z}(\mathbf{s},t)$ is assumed to be Gaussian, it is fully characterized by its space-time cross-covariance function $C_{k,ij}(\mathbf{h},u)$. To take into account the temporal variations of the meteorological variables that affect the whole region, we follow Bourotte et al (2016) and decompose the cross-covariance function into:

$$C_{k,ij}(\mathbf{h},u) = C^{(1)}_{k,ij}(u) + C^{(2)}_{k,ij}(\mathbf{h},u), \tag{6}$$

where $C^{(1)}_{k,ij}(u)$ is a fully temporal cross-covariance function modeling the temporal fluctuations of the meteorological variables at the regional scale and $C^{(2)}_{k,ij}(\mathbf{h},u)$ is a space-time cross-covariance function modeling the fluctuations of the meteorological variables within the area of interest. For the temporal cross-covariance function $C^{(1)}_{k,ij}(u)$, we consider a multivariate exponential covariance (Gneiting et al, 2010; Genton and Kleiber, 2015):

$$C^{(1)}_{k,ij}(u) = \beta^{(1)}_{k,ij} \exp\left(-r^{(1)}_{k,ij}|u|\right) \tag{7}$$

with $2\left(r^{(1)}_{k,ij}\right)^2 = \left(r^{(1)}_{k,ii}\right)^2 + \left(r^{(1)}_{k,jj}\right)^2$, $\beta^{(1)}_{k,ij} = \rho^{(1)}_{k,ij}\left(r^{(1)}_{k,ii} r^{(1)}_{k,jj}\right)^{1/2}\left(r^{(1)}_{k,ij}\right)^{-1}$, and $\left[\rho^{(1)}_{k,ij}\right]_{i,j=1}^{p}$ is a correlation matrix.

To model $C^{(2)}_{k,ij}$ we use the class of Gneiting space-time non-separable cross-covariance functions proposed in Allard et al (2022). Specifically, we consider the multivariate Gneiting-Matérn model defined as:

$$C^{(2)}_{k,ij}(\mathbf{h},u)=\beta^{(2)}_{k,ij}\frac{\exp\left(-r^{(2)}_{k,ij}|u|\right)}{(\eta_{k,ij}(u)+1)^{d/2}}C_{\mathcal{M}}\left(\mathbf{h};\sqrt{\frac{a^2_{k,ij}}{(\eta_{k,ij}(u)+1)}},\nu_{k,ij}\right), \tag{8}$$

with $i,j=1,\ldots,p$ and where $C_{\mathcal{M}}(\mathbf{h};r,\nu)$ is the Matérn covariance function with scale parameter $r>0$ and smoothness parameter ν:

$$C_{\mathcal{M}}(\mathbf{h};r,\nu)=\frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(r||\mathbf{h}||)^{\nu}K_{\nu}(r||\mathbf{h}||), \tag{9}$$

where K_ν is the Bessel function of second kind with parameter $\nu>0$. In (8), the following conditions must hold: $2\left(r^{(2)}_{k,ij}\right)^2=\left(r^{(2)}_{k,ii}\right)^2+\left(r^{(2)}_{k,jj}\right)^2$, and $\boldsymbol{\eta}(u)=[\eta_{k,ij}(u)]^p_{i,j=1}$ is a $p\times p$ unbounded pseudo-variogram on $\mathbb{R}$. Pseudo-variograms on $\mathbb{R}$ are defined in the following way. Let $W_i(\cdot)_{i=1,\ldots,p}$ be p stochastic processes. Then,

$$\eta_{k,ij}(u)=0.5\mathrm{Var}\big(W_i(t)-W_j(t+u)\big)$$

provided it exists for all $t,u\in\mathbb{R}$ and for all $i,j=1,\ldots,p$. The pseudo-variogram has nonnegative entries and is not necessarily an even function. For any $i=1,\ldots,p$, the function η_{ii}, is a usual variogram, i.e., a conditionally negative semidefinite function (Chilès and Delfiner, 2012). The pseudo-variogram must be unbounded for the direct and cross-covariances $C_{ij}(\mathbf{h},u)$ to vanish as $|u|\to\infty$. Necessary and sufficient conditions for a matrix-valued function to be a pseudo-variogram have been provided in Dörr and Schlather (2023). Despite these recent results, building valid unbounded matrix-valued pseudo-variograms with different diagonal entries (direct variograms) is still an open question. As a simple model for this, we will follow Allard et al (2022) and set:

$$\eta_{k,ij}(u)=\left((a_k|u|)^{2b_k}+1\right)^{c_k}-A_{k,i}A_{k,j}\left((d_k|u|)^{2e_k}+1\right)^{-c_k}, \tag{10}$$

where $a_k>0$, $0<b_k\leq 1$, $0\leq c_k\leq 1$, $0\leq A_{k,i}<1$, for all $i,j=1,\ldots,p$, $d_k>0$, and $0<e_k\leq 1$. The sufficient conditions for the positive definiteness of $C^{(2)}_{k,ij}(h,u)$ are that, for all $i,j=1,\ldots,p$, $2a^2_{k,ij}=a^2_{k,ii}+a^2_{k,jj}$, $2\nu_{k,ij}=\nu_{k,ii}+\nu_{k,jj}$ and

$$\beta^{(2)}_{k,ij}=\rho^{(2)}_{k,ij}\frac{a^{v_{k,ii}}_{k,ii}a^{v_{k,jj}}_{k,jj}}{a^{v_{k,ij}}_{k,ij}}\frac{\Gamma(\nu_{k,ij})}{\Gamma(\nu_{k,ii})^{1/2}\Gamma(\nu_{k,jj})^{1/2}}\frac{\left(r^{(2)}_{k,ii}r^{(2)}_{k,jj}\right)^{1/2}}{r^{(2)}_{k,ij}}\sqrt{(2-{A_{k,i}}^2)(2-{A_{k,j}}^2)},$$

where $\left[\rho^{(2)}_{k,ij}\right]_{i,j=1}^{p}$ is a correlation matrix.

Finally, for a weather state $k \in \mathcal{S}$, the parameters of the random field $\mathbf{Z}_k(\mathbf{s}, t)$ are:

$$\theta_k = \left\{ a_k, b_k, c_k, d_k, e_k, A_{k,i}, r^{(1)}_{k,ii}, r^{(2)}_{k,ii}, \rho^{(1)}_{k,ij}, \rho^{(2)}_{k,ij}, a_{k,ii}, \nu_{k,ii},\ \ i,j = 1,\dots,p \right\}, \quad (11)$$

which makes $5 + 5p + p(p-1)$ parameters per weather state.

3 Estimation of the parameters

3.1 Estimation of weather type transitions

Let $\mathbf{X} = \{X(t_1), \dots, X(t_{n_t})\}$ be the observed weather type process. The non-homogeneous transition probabilities of the first order Markov chain are estimated by their empirical frequencies in a sliding window of length $2L + 1$ with:

$$\hat{\pi}_{kl}(t_\gamma) = \frac{\sum_{\delta \in V(t_\gamma, L)} \mathbb{I}(X(t_\delta - 1) = k, X(t_\delta) = l)}{\sum_r \sum_{\delta \in V(t_\gamma, L)} \mathbb{I}(X(t_\delta - 1) = k, X(t_\delta) = r)}, \quad (12)$$

where $V(t_\gamma, L) = \{t_\delta, |t_\delta - t_\gamma| \leq L\}$ is a time window of length $2L + 1$ centered at t_γ, and $\mathbb{I}(A)$ is the indicator function equal to 1 when A holds and equal to 0 otherwise.

The parameter L is a hyperparameter that influences the temporal smoothness of the transition probabilities. If L is too small it may lead to overfitting, making the simulated weather types closely resemble the observed ones, while if L is too large it can cause the simulated weather types to be overly homogeneous over time. The appropriate choice of L depends on the specific application and on the size of the time domain being considered.

3.2 Estimation of transformation functions

We now present the method for estimating the transformation function applied to each variable, weather type, and spatial location. To simplify the notation in this subsection, indices for spatial locations, weather types, and variables will not be used. Instead, we only use the index t, which represents a time-specific observation of a variable at a particular spatial location.

The OQN transformation in equation (5) necessitates the extrapolation of the function $\Psi^{-1}_{k,i,\mathbf{s}}$ (hereafter noted Ψ^{-1}) beyond the observed data range. We utilize the method developed by Peterson and Cavanaugh (2019), which employs a logit approximation for data outside the observation range. Consider a data vector $\mathbf{y} =$

$(y_1, \ldots, y_n)$. For a given value $y_t \in \mathbf{y}$, we define the function g as follows:

$$\begin{aligned} g(y_t) &= \Phi_T^{-1}(F(y_t)) \\ &= \Phi_T^{-1}\left\{\frac{r_t - 1/2}{n} + \Phi(T)\left(1 - \frac{r_t - 1/2}{n}\right)\right\}, \end{aligned} \tag{13}$$

where r_t is the rank of y_t and T represents the truncation threshold as in (5). The functions F, Φ_T, and Φ denote the CDF of y, the standard truncated Gaussian CDF (T being the truncation threshold), and the standard Gaussian CDF, respectively. Note that the first part of the equation is simply the inverse of equation (5). For an arbitrary point y^*, if $y^* \notin \mathbf{y}$, let y_l and y_r be the closest points to y^* in the original data, with $y_l < y^*$ and $y_r > y^*$, and define $a_{y^*} = 1/(y_r - y_l)$. Following Peterson and Cavanaugh (2019), the truncated OQN $\Psi^{-1}(y^*)$ is defined as follows:

$$\Psi^{-1}(y^*) = \begin{cases} g(y^*) & \text{if } y^* \in \{\mathbf{y}\}, \\ a_{y^*}\big(g(y_r) - g(y_l)\big) & \text{if } y^* \notin \{\mathbf{y}\} \text{ and } \min(\mathbf{y}) < y^* < \max(\mathbf{y}), \\ \ell(y^*) + \min_i\big(g(y_i)\big) - \min_i\big(\ell(y_i)\big) & \text{if } y^* < \min(\mathbf{y}), \\ \ell(y^*) + \max_i\big(g(y_i)\big) - \max_i\big(\ell(y_i)\big) & \text{if } y^* > \max(\mathbf{y}), \end{cases} \tag{14}$$

where

$$\ell(y) = \Phi_T^{-1}\left(\frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 y)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 y)}\right), \tag{15}$$

and $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimated parameters of the following generalized linear model, determined by likelihood maximization:

$$\text{logit}\left(\Phi_T(\Psi^{-1}(y_i))\right) = \beta_0 + \beta_1 y_i, \quad i = 1, \ldots, n. \tag{16}$$

The threshold value T varies depending on the meteorological variable. For continuous variables like temperature, the threshold is negative infinity. For zero-inflated variables such as precipitation, the threshold equals $\Phi^{-1}(f_0)$, where f_0 is the frequency of zero values.

3.3 Estimation of the covariance function parameters

The parameters of the covariance function are estimated independently for each weather type successively. For ease of notations, we drop the reference to the weather type k for the rest of this section. Let us denote $\mathbf{Z} = (Z_1(\mathbf{s}_1, t_1), \ldots, Z_p(\mathbf{s}_1, t_1), \ldots, Z_p(\mathbf{s}_{n_S}, t_{n_T}))^T$ the vector of transformed observations

corresponding to the p meteorological variables measured at n_S spatial locations and n_T time steps. Because of the preliminary OQN transformation, the values in $\mathbf{Z}$ are supposed to follow a Gaussian or truncated Gaussian distribution.

When dealing with multivariate space-time Gaussian random fields, the maximum likelihood approach requires computing the determinant and the inverse of large matrices, which is impractical for domains of even moderate size. As an example, a moderate spatio-temporal domain with $n_S = 100$, $n_T = 30$ and three variables would lead to matrices of size 9000×9000. To overcome this limitation, approximations of the likelihood can be considered. Here we use the pairwise likelihood (PL), a special case of composite likelihood (Varin et al, 2011). PL is the product of marginal likelihoods computed on a well-chosen selection of pairs. It has been successfully considered in spatio-temporal contexts in Bourotte et al (2016) and Allard et al (2022) where it was shown that PL provides estimates of the covariance function parameters with only a small loss in efficiency compared to a full likelihood approach (when it is possible), with a significant gain in terms of computation.

The computational cost induced by PL is of the order of $\mathcal{O}\left((n_S n_T p)^2\right)$ if all pairs are considered. It can nevertheless be significantly reduced if most pairs of observations whose distance is beyond the correlation range are discarded, although some distant pairs should be included in the PL to improve parameter estimation (Allard et al, 2021). Hence, we first select the set A of n_A optimal locations (centroids) that partition the spatial domain D into n_A subgroups of minimal area. Then, for each $\alpha \in A$, we build the sets B_α containing n_B locations that have been randomly selected among all possible locations with probabilities inversely proportional to their distance from $\mathbf{s}_\alpha$:

$$B_\alpha = \left\{ \mathbf{s}_\beta \in D \mid \mathbb{I}\left(U_\beta \leq \frac{b_\alpha}{\|\mathbf{s}_\alpha - \mathbf{s}_\beta\|^2 + 1} \right) \right\}, \tag{17}$$

where U_β are independent copies of a uniform random variable on $[0, 1]$ and where b_α has been adjusted so that $\mathbb{E}[|B_\alpha|] = n_B$.

The set of pairs Λ to be used in PL is then defined as

$$\Lambda = \{(\alpha, \beta, \gamma, \delta) \ : \ \alpha \in A;\ \beta \in B_\alpha;\ |t_\gamma - t_\delta| \leq t_{max},\} \tag{18}$$

where t_{max} denotes the maximum time lag. Compared to the original PL, the number of pairs has been reduced by a factor approximately equal to $n_S^2/(n_A n_B) \times n_T/t_{max}$, which potentially reach several orders of magnitude and has significant impact on the optimization efficiency. The hyper-parameters t_{max}, n_A and n_B depend on the size of

the space-time domain under consideration, and should therefore be defined by the user.

For the set of pairs Λ, the pairwise log-likelihood is

$$\mathrm{pl}(\theta) = \sum_{i,j=1}^{p} \sum_{(\alpha,\beta,\gamma,\delta)\in\Lambda} \ell(i,j,\alpha,\beta,\gamma,\delta;\theta), \tag{19}$$

where θ represents the parameters of the covariance function and $\ell(i,j,\alpha,\beta,\gamma,\delta;\theta)$ is the bivariate log-likelihood for a pair $(Z_i(\mathbf{s}_\alpha,t_\gamma), Z_j(\mathbf{s}_\beta,t_\delta))$. Depending on the values of $Z_i(\mathbf{s}_\alpha,t_\gamma)$ and $Z_j(\mathbf{s}_\beta,t_\delta)$, three cases must be considered:

- If $Z_i(\mathbf{s}_\alpha,t_\gamma) > T_{i,\mathbf{s}_\alpha}$ and $Z_j(\mathbf{s}_\beta,t_\delta) > T_{j,\mathbf{s}_\beta}$,

$$\ell(i,j,\alpha,\beta,\gamma,\delta;\theta) = -\frac{1}{2}\Bigg(\log\left(1-C_{ij}(\mathbf{h},u)^2\right) + \frac{Z_i(\mathbf{s}_\alpha,t_\gamma)^2 - 2C_{ij}(\mathbf{h},u)Z_i(\mathbf{s}_\alpha,t_\gamma)Z_j(\mathbf{s}_\beta,t_\delta) + Z_j(\mathbf{s}_\beta,t_\delta)^2}{1-C_{ij}(\mathbf{h},u)^2}\Bigg).$$

- If $Z_i(\mathbf{s}_\alpha,t_\gamma) \leq T_{i,\mathbf{s}_\alpha}$ and $Z_j(\mathbf{s}_\beta,t_\delta) > T_{j,\mathbf{s}_\beta}$;

$$\ell(i,j,\alpha,\beta,\gamma,\delta;\theta) = \log\Phi\left(\frac{T_{i,\mathbf{s}_\alpha} - C_{ij}(\mathbf{h},u)Z_j(\mathbf{s}_\beta,t_\delta)}{\sqrt{1-C_{ij}(\mathbf{h},u)^2}}\right).$$

- If $Z_i(\mathbf{s}_\alpha,t_\gamma) \leq T_{i,\mathbf{s}_\alpha}$ and $Z_j(\mathbf{s}_\beta,t_\delta) \leq T_{j,\mathbf{s}_\beta}$,

$$\ell(i,j,\alpha,\beta,\gamma,\delta;\theta) = \log\Phi_2\left(T_{i,\mathbf{s}_\alpha}, T_{j,\mathbf{s}_\beta}; C_{ij}(\mathbf{h},u)\right),$$

where $\mathbf{h} = ||\mathbf{s}_\alpha - \mathbf{s}_\beta||$, $u = |t_\gamma - t_\delta|$, and Φ_2 is the bivariate Gaussian CDF.

4 Simulation of space-time and multivariate synthetic weather data

The meteorological variables are simulated using a simulation algorithm sequential in time. Consider $\mathbf{Z}(t) = \mathrm{vec}\left([[Z_i(\mathbf{s}_\alpha,t)]_{i=1}^{p}]_{\alpha=1}^{n_s}\right)$, where $\mathrm{vec}(\cdot)$ is the operator that stacks the columns of a matrix into a single column vector. In principle, at a given time t, the simulation of $\mathbf{Z}(t)$ should be conditional on the whole sequence $(X(1),\ldots,X(t-1),\mathbf{Z}(1),\ldots,\mathbf{Z}(t-1))$. This would require inverting matrices of size up to $n_S \times (n_T - 1) \times p$, which is in most cases not possible. To keep the conditioning computationally

efficient, we make the following Markov-type assumption:

$$p(\mathbf{Z}(t) \mid X(0), \ldots, X(t), \mathbf{Z}(0), \ldots, \mathbf{Z}(t-1)) = p(\mathbf{Z}(t) \mid X(t), \mathbf{Z}(t-M), \ldots, \mathbf{Z}(t-1)) \tag{20}$$

for a given M, where p is the probability distribution function.

The pseudo-code of the simulation algorithm is shown in Algorithm 1. Step 2 consists of simulating $\mathbf{Z}(t)$ conditionally on $X(t) = k$ and on $(\mathbf{Z}(t-1), \ldots, \mathbf{Z}(t-M))$. Since the vectors $(\mathbf{Z}(t), \ldots, \mathbf{Z}(t-M))$ are jointly Gaussian, the conditional distribution of $\mathbf{Z}(t)$ given $X(t)$ and $(\mathbf{Z}(t), \ldots, \mathbf{Z}(t-M))$ is

$$\mathbf{Z}(t) = \mathbf{B}_{k,1}\mathbf{Z}(t-1) + \cdots + \mathbf{B}_{k,M}\mathbf{Z}(t-M) + \mathbf{L}_{k,0}\boldsymbol{\epsilon}(t), \tag{21}$$

where $\boldsymbol{\epsilon}(t)$ is a vector of appropriate size of i.i.d standard Gaussian random variables. The matrices $\mathbf{B}_{k,m}$ with $m = 1, \ldots, M$ are solution of the linear system

$$\begin{pmatrix} \mathbf{B}_{k,1} \\ \mathbf{B}_{k,2} \\ \vdots \\ \mathbf{B}_{k,M} \end{pmatrix} = \begin{pmatrix} \mathbf{C}_k(0) & \mathbf{C}_k(1) & \cdots & \mathbf{C}_k(M-1) \\ \mathbf{C}_k(1) & \mathbf{C}_k(0) & \cdots & \mathbf{C}_k(M-2) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_k(M-1) & \mathbf{C}_k(M-2) & \cdots & \mathbf{C}_k(0) \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{C}_k(1) \\ \mathbf{C}_k(2) \\ \vdots \\ \mathbf{C}_k(M) \end{pmatrix} \tag{22}$$

and

$$\mathbf{L}_{k,0}^T\mathbf{L}_{k,0} = \mathbf{C}_k(0) - \mathbf{B}_{k,1}\mathbf{C}_k(0) - \cdots - \mathbf{B}_{k,M}\mathbf{C}_k(M).$$

with

$$\mathbf{C}_k(m) = \operatorname{cov}\left(\mathbf{Z}(t), \mathbf{Z}(t-m) \mid X(t) = k\right), \ m = 0, \ldots, M.$$

Each matrix $\mathbf{C}_k(m)$ is the Gram matrix computed using the spatio-temporal covariance defined in equation (6) at the temporal lag $u = m$ and the parameters estimated in weather type k. The simulation algorithm requires inverting and storing the matrix of size $n_S \times M \times p$ in equation (22) which might be computationally expensive depending on the space-time domain size. The choice of an appropriate M permits to reduce the computational burden.

5 Dealing with seasonality

In most cases, some seasonality is present in the meteorological data. Seasonality impacts the mean value and the variance of each variable, but also the correlation between variables. A well known example is the correlation between solar radiation and temperature, which in temperate climates is positive in the summer and negative

Algorithm 1 Multivariate Space-Time Stochastic Weather Generator

1: **Input:** Initial values $X(0)$ and $\mathbf{Z}(0), \ldots, \mathbf{Z}(M)$
2: **Output:** Simulated weather variables
3: **for** $t > M$ **do**
4: Sample a weather state $X(t)$ conditionally on $X(t-1)$ using the transition probabilities in (12)
5: Sample the latent random field $\mathbf{Z}(t)$ conditionally on $X(t)$ and $(\mathbf{Z}(t-1), \ldots, \mathbf{Z}(t-M))$ as per (21).
6: Transform the random field $\mathbf{Z}(t)$ into meteorological variables using the transform function $\Psi_{X(t),i,\mathbf{s}}$.
7: **end for**

in the winter. In our framework, seasonality is accounted for in two ways: first, any seasonal cycle in each variable is removed; then, seasons during which the dependencies between the weather variables are considered constant are defined. Seasons must be provided by the users as input parameters because their definition depend on the precise location of the studied area. Except for precipitation, the central tendency and standard deviation are calculated for each variable and each day of the year. The moving average method is then used to smooth the mean and standard deviation. The observed variables are then standardised by subtracting the calculated central tendency and dividing the residuals by the smoothed standard deviation. Finally, all parameters are estimated separately for each season, including transition probabilities between weather types, transformation functions, and covariance function parameters.

For the simulation, the weather variables are generated for an entire season using the Algorithm 1 and the corresponding estimated parameters. When changing from one season to the next, a weather state is required for the first day of the new season. Since the probability transitions between the states of two different seasons have not been estimated (due to insufficient data), we need a special procedure. Let us denote $\mathbf{Z}(t_{\text{last}})$ the variables simulated for the last day of the season. The likelihood of $\mathbf{Z}(t_{\text{last}})$ is then computed using the parameters of the new season, and the most likely state, denoted $\hat{X}_{\text{last}}$, is selected. Then, weather states within the new season are simulated according to the new transition probabilities with initial state $X(0) = \hat{X}_{\text{last}}$. At the end of the year, all simulated variables are multiplied by the smoothed standard deviation. Finally, the smoothed central tendency is added.

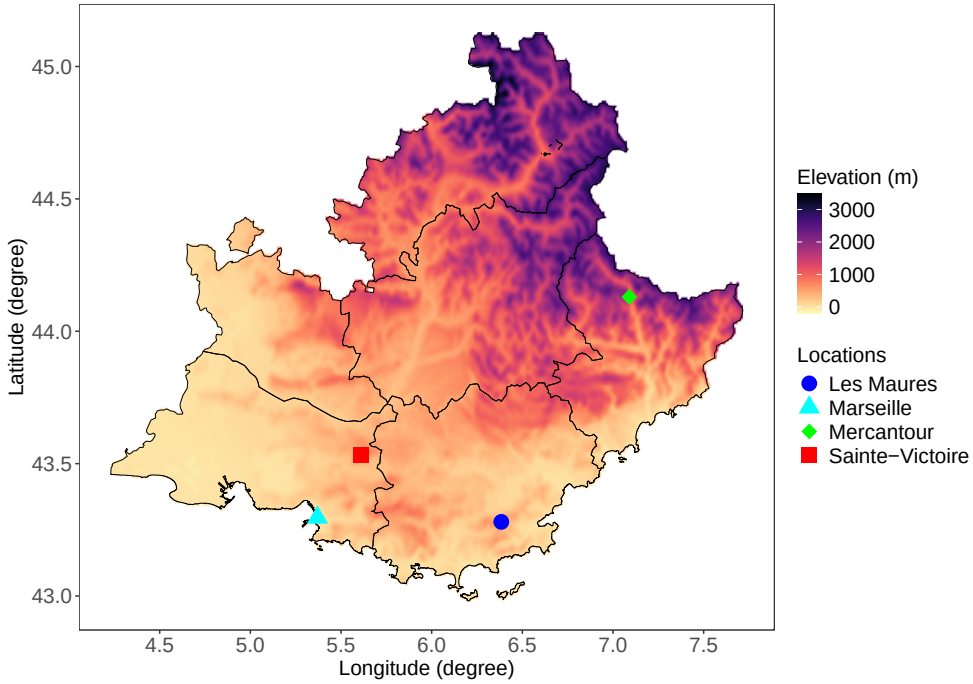

Fig. 1: Elevation of the study area (PACA region) along with the four key locations analyzed in this study: Marseille, Les Maures, Mercantour, and Sainte-Victoire

6 Application to weather generation over South-East France

6.1 Study area and data

The performance of MSTWeatherGen is assessed for the period 2012-2021 using a daily gridded dataset encompassing six meteorological variables: absolute humidity, precipitation, radiation, minimum temperature, maximum temperature, and wind speed. This dataset derives from the SAFRAN reanalysis (Quintana-Segui et al, 2008), which covers mainland France at a 8 km $\times$ 8 km resolution. Our analysis focuses on the Provence-Alpes-Côte d'Azur (PACA) region, covering a total of 498 pixels. This region is chosen due to its diverse climate patterns: the southern part is characterized by a Mediterranean climate while the North-East of the region experiences a mountain climate (see Figure 1).

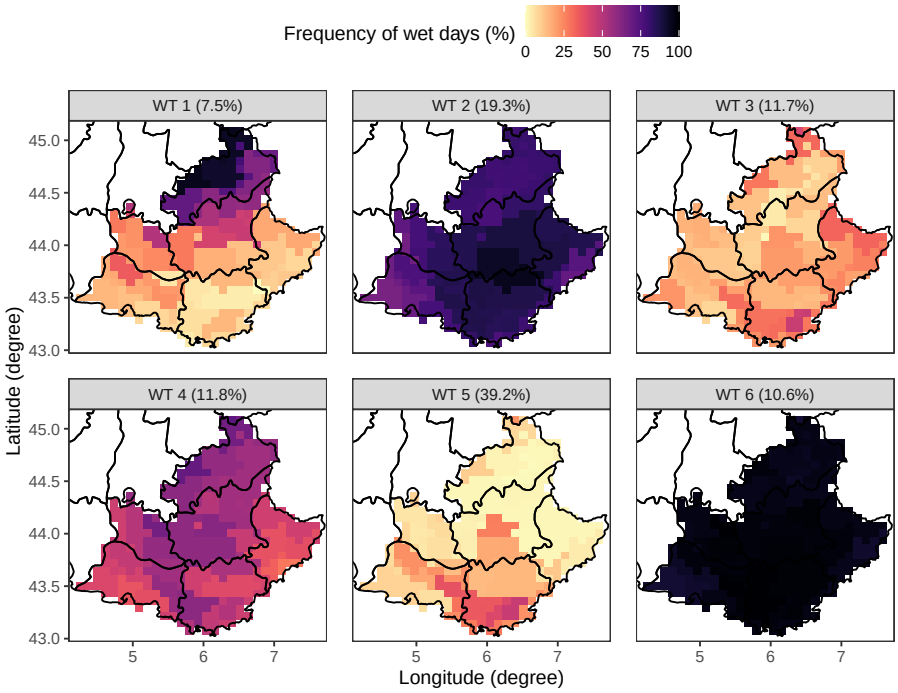

Fig. 2: Frequency of wet days for each weather type at each grid point during the winter season (DJF). The percentage in brackets corresponds to the percentage of each weather type

6.2 Parameter estimation

Since the weather variables exhibit seasonality in PACA, we adopt the method presented in Section 5 and remove the annual cycle before estimating the parameters of MSTWeatherGen for each season separately. We follow the common practice of defining four seasons for a mid-latitude climate: winter (DJF), spring (MAM), summer (JJA) and autumn (SON). After removing the seasonal signal, we estimate the weather types using the method described in Section 3.1. Figure 2 illustrates the frequency of wet days for each weather type of the winter season, and shows that WT 2 and 6 predominantly experience wet conditions across the entire region while WT 3 and 5 are characterized by mostly dry conditions, and WT 1 exhibits intermediate conditions

with strong spatial gradients. After weather typing, the transitions between weather types are modeled using a first-order non-homogeneous Markov chain (cf. Section 3.1) whose parameters are estimated using a sliding window of length $L = 30$ days.

All detrended variables are subsequently transformed into Gaussian scores using the ordered normalization method (cf. Section 3.2), and the parameters of the covariance function are estimated for each weather type separately (cf. Subsection 2.4) with the hyperparameters n_A, n_B, and t_{max} set to respectively 15, 15, and 3. Figure 3 displays the empirical and modeled covariances for winter WT 1, and shows that the model effectively captures the intra-variable space-time covariances as well as most of inter-variable covariances. The most complex relationships (e.g., between maximum temperature and minimum temperature for which the one-day-lag spatial correlation is higher than the zero-day correlation) are nevertheless only imperfectly captured because the model of covariance lacks flexibility to accommodate the intricate behavior of the empirical covariance.

6.3 Stochastic weather generation

Stochastic weather generation is performed using Algorithm 1 in which we opt for a third-order autoregressive model to maintain computational efficiency (we were therefore able to perform all computations on a usual laptop computer). We set $M = 3$ in Equation (21), which results in $M \times K = 18$ matrices $\mathbf{B}_{k,m}$ of size $n_S p \times n_S p = 2988 \times 2988$.

Simulation of marginal distributions

We assess the marginal distributions simulated by MSTWeatherGen for all target variables at one location, namely the grid-point encompassing the city of Marseille which is the main city of PACA. Figure 4 compares the observed and simulated empirical probability density functions for all meteorological variables, apart from precipitation whose simulation is assessed in Figure 5. Figure 4 shows that MSTWeatherGen effectively reproduces the marginal distributions of all target variables, and in particular their multimodality (see e.g., radiation or minimum temperature in Fig. 4) thanks to the use of the ordered normalization method to transform the data (see Section 3.2). Regarding precipitation, the quantile-quantile plots in Figure 5 show that the simulated intensities closely match the observed ones, the observed values being well within the simulation envelopes.

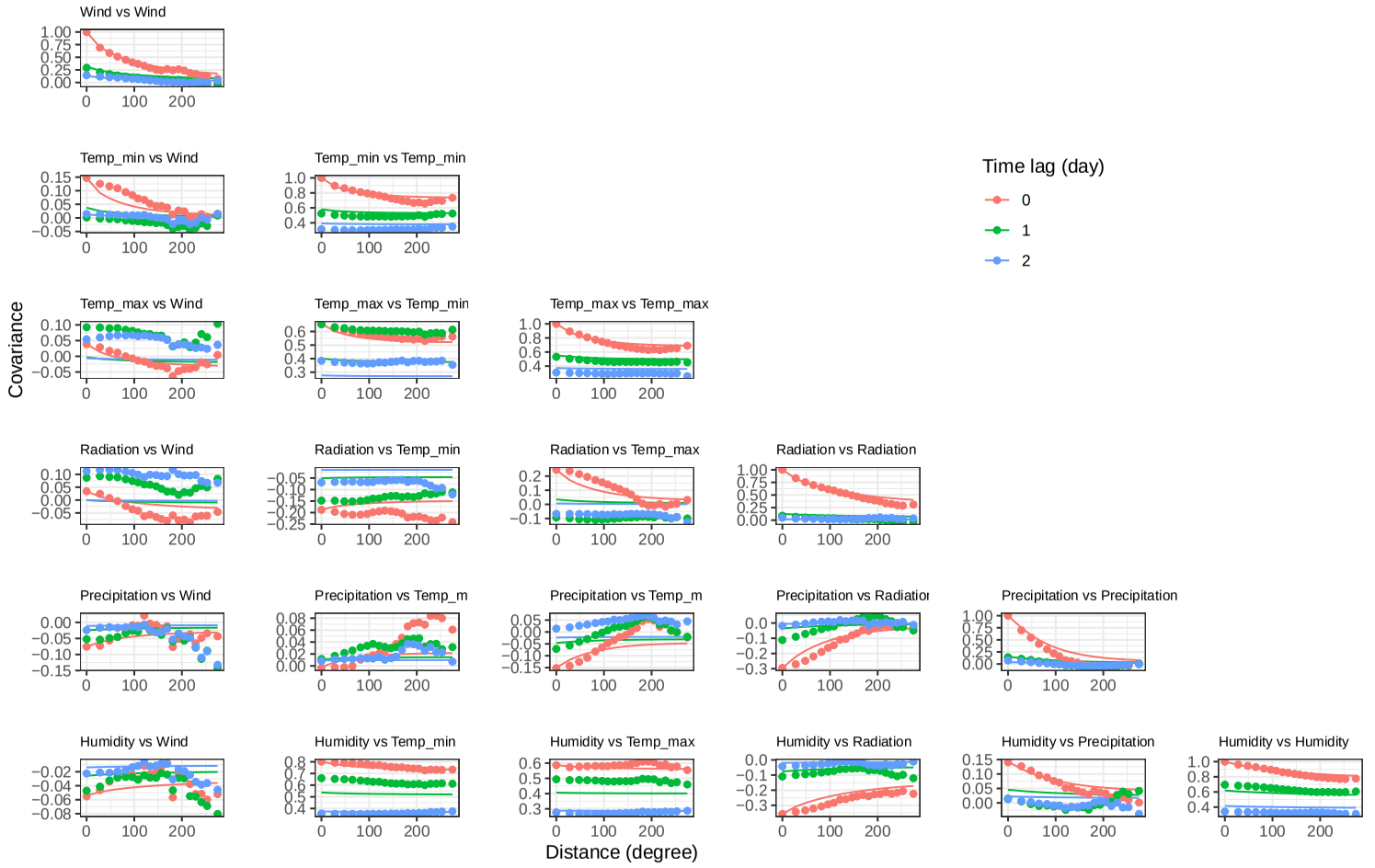

Fig. 3: Spatial covariance for normalized observations (dots) and model (solid line) at lag zero, one, and two days (respectively in red, green, and blue) between all meteorological variables and for the weather type WT 1 - winter season

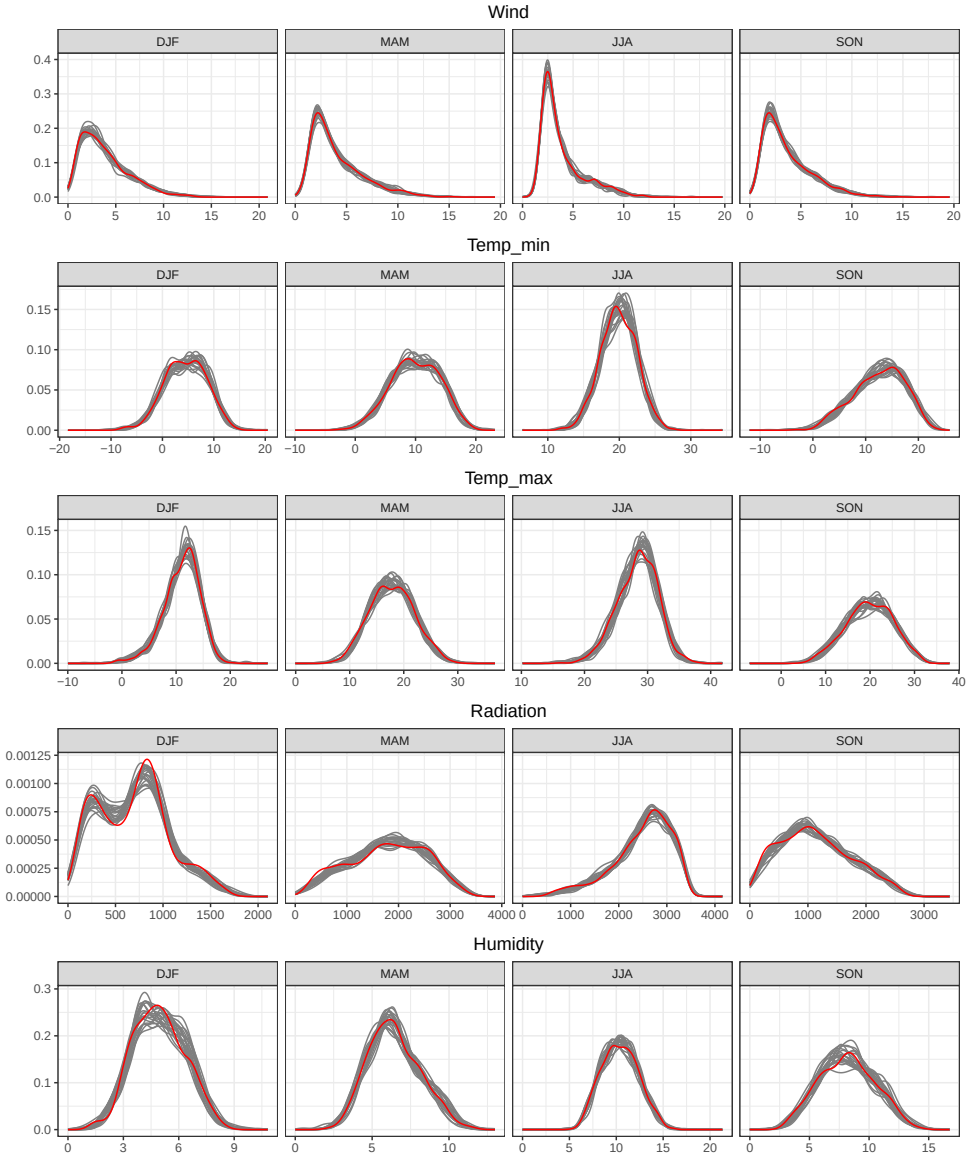

Fig. 4: Empirical probability density function of observations (red curve) and 20 simulated realizations (gray curves) in the city of Marseille for the period 2012-2021. From top to bottom: wind speed (m/s), minimum temperature (°C), maximum temperature (°C), radiation (W/m^2), and absolute humidity (g/m^3)

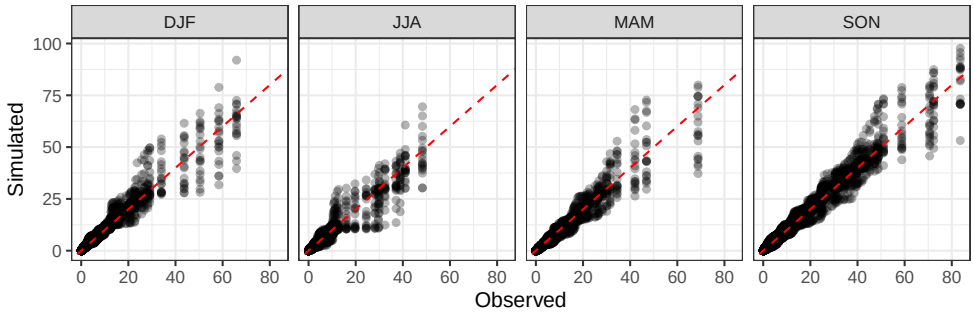

Fig. 5: Simulated precipitation quantiles vs. observed precipitation quantiles for 20 simulated 10 year long realizations. The red dashed line indicates perfect agreement. City of Marseille, period 2012-2021

Simulation of bivariate statistics

The stochastic generation of multiple variables over a spatial domain requires the accurate reproduction of the bivariate statistics between pairs of variables and locations. To assess the simulation of such bivariate statistics by MSTWeatherGen, a total of 10 locations were randomly selected within the PACA region, and the correlations between pairs of locations and variables were calculated for each season. The results are displayed in Figure 6, which shows that overall the model successfully reproduces both inter-variable and cross-variable correlations. However, MSTWeatherGen tends to overestimate the correlation between some pairs of variables, for instance in summer between precipitation and radiation as well as between precipitation and maximum temperature. This overestimation may be due to the decreased rainfall observed in the summer months. In addition, the inter-sites correlation of precipitation is imperfectly simulated for some pairs of locations in winter, which may be due to the non-stationarity of the spatial covariance of this variable at that time of the year. On a side note, one can notice that the correlations reported in Figure 6 vary significantly between seasons, which supports the choice of fitting distinct models for each season.

Simulation of temporal persistence

The simulation of realistic temporal persistence patterns within the weather system is crucial for impact studies because many environmental processes integrate weather variables through time, and are therefore sensitive to the duration spent in a given weather condition. For instance, dry and wet spells (defined as sequences of consecutive dry and wet days) are critical in hydrology because they significantly affect

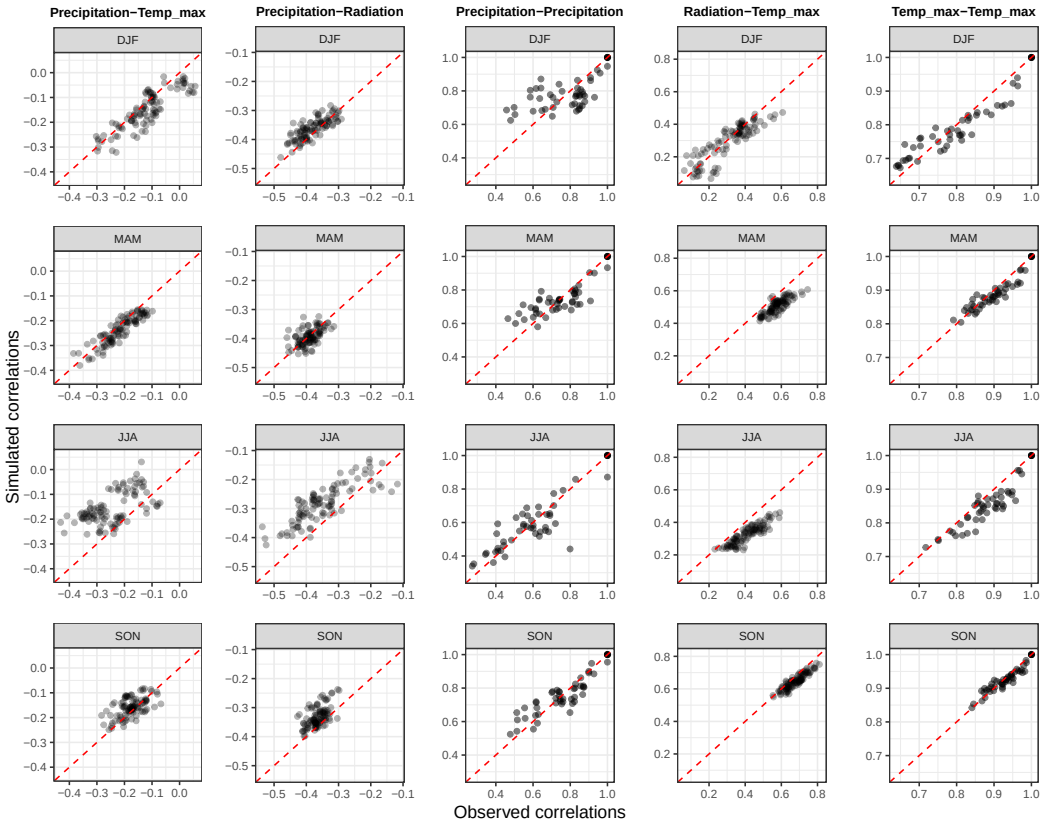

Fig. 6: Bivariate correlations of observed and simulated variables evaluated for all possible pairs derived from a set of 10 randomly selected locations. From top to bottom: winter, spring, summer, and autumn. The red dashed line indicates a perfect correlation

the occurrence of droughts (long dry spells) and floods (long wet spells) (Mathlouthi and Lebdi, 2021). Figure 7 compares the observed and simulated wet spells duration in winter and dry spells duration in summer. Results show that MSTWeatherGen effectively reproduces the duration of dry and wet spells, except for an overestimation of short winter wet spells in mountain areas (North of the target area). One can also notice that the simulation is less regular than the observations, which is due to the artificial patchiness of precipitation in the original SAFRAN reanalysis dataset (Quintana-Segui et al, 2008). To complement this evaluation of precipitation persistence, Figure 8 compares the persistence of observed and simulated cold winter and

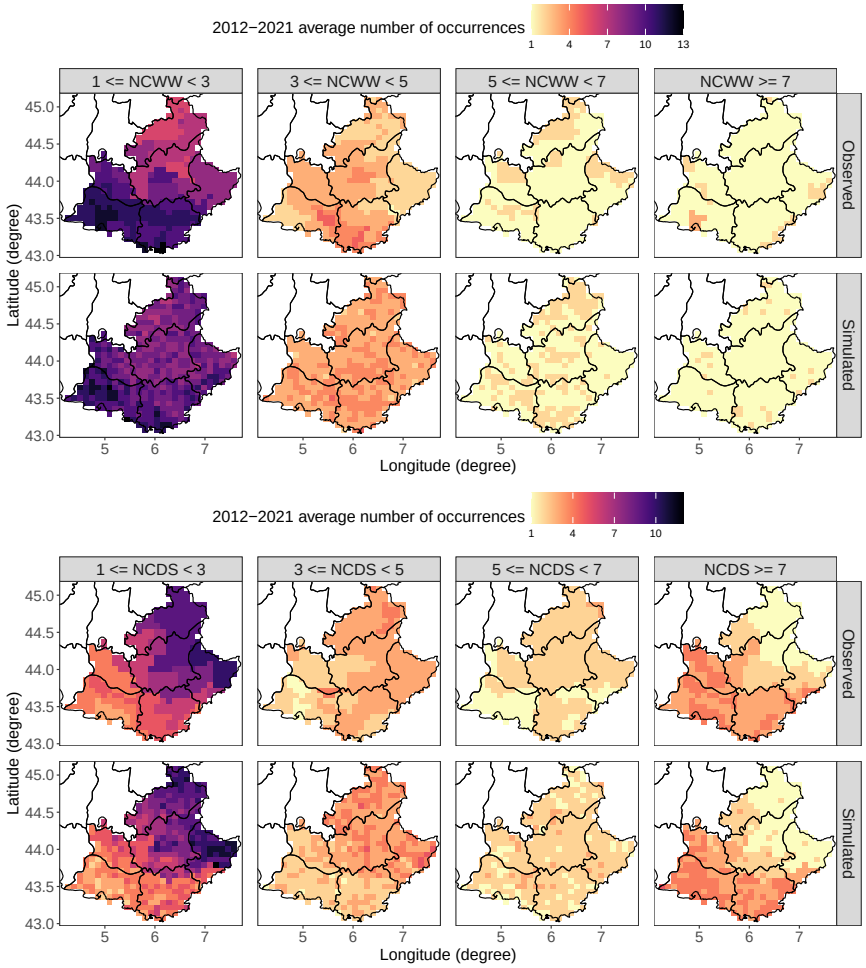

Fig. 7: Observed and simulated average wet and dry spells for the period 2012-2021. Upper panel: Number of consecutive wet winter days (NCWW). Bottom panel: Number of consecutive dry summer days (NCDS)

hot summer spells, and the results show that the model performs well also with regard to the persistence of temperature threshold passing.

Simulation of compound events: the example of heatwaves

Heatwaves are significant compound extreme events with substantial impacts on human health and the economy, and attract a growing interest as their frequency and intensity have been increasing in recent years due to climate change (Ouzeau

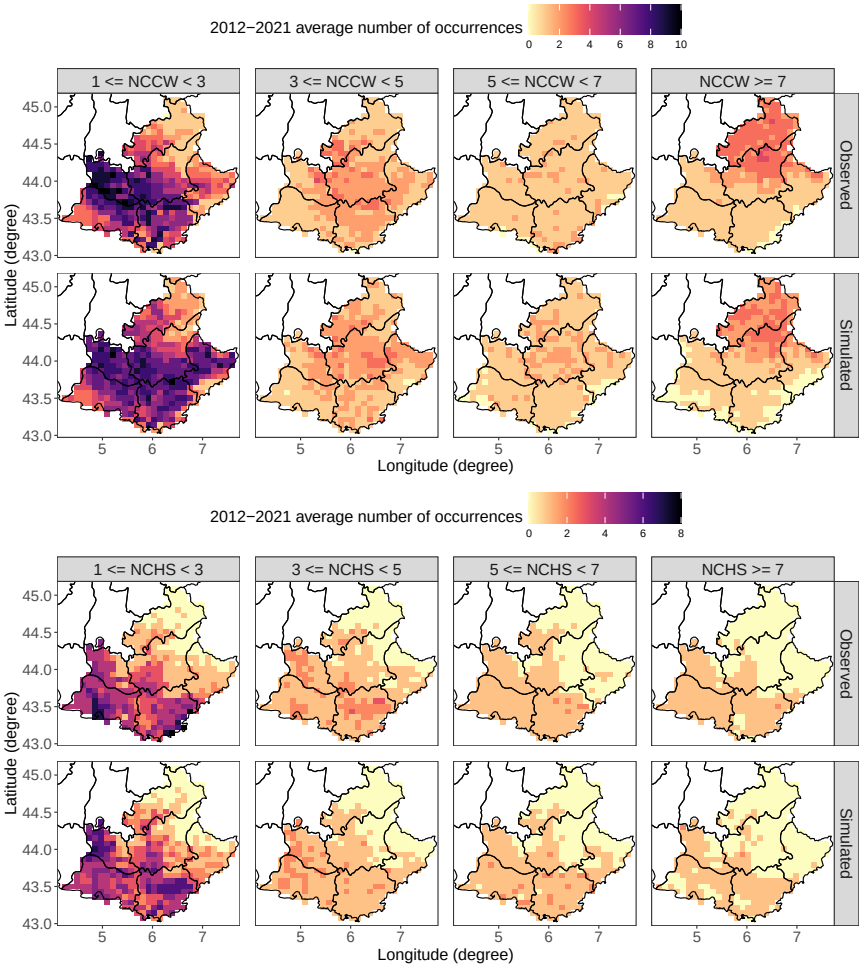

Fig. 8: Observed and simulated average cold and hot spells for the period 2012-2021. Upper panel: Number of consecutive cold winter days (NCCW). Bottom panel: Number of consecutive hot summer days (NCHS). Cold days are defined as days with a minimum temperature below 0°C and hot summer days as days with a maximum temperature above 31.8°C, which represents the 90th percentile for the region

et al, 2016). Accurately reproducing these events using stochastic weather generators is essential for conducting impact studies (Yiou and Jézéquel, 2020). The criteria used to define heatwaves vary depending on the context and location. For example the French weather agency Météo-France issues a public warning (Orange Vigilance

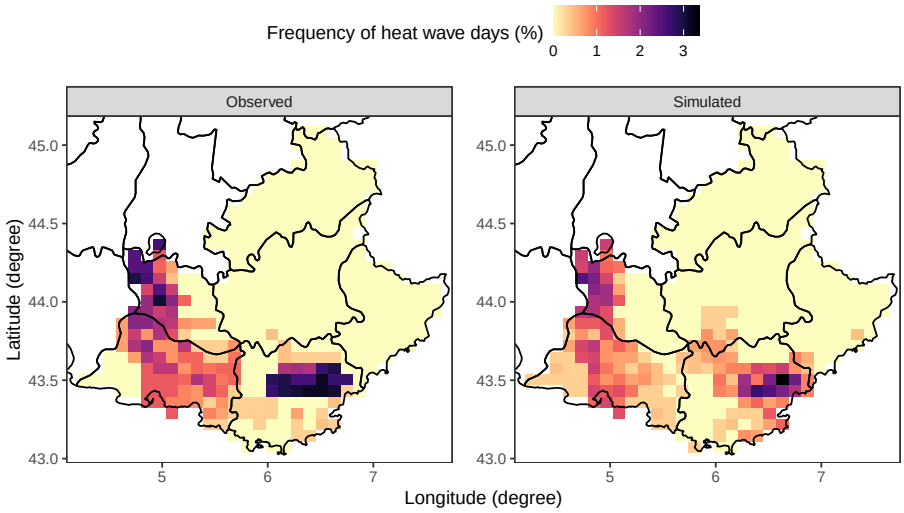

Fig. 9: Observed and simulated frequency of heatwaves during summer from 2012 to 2021. Heatwaves are defined as periods during which both maximum and minimum temperatures exceed high thresholds for at least three consecutive days. The thresholds used here are 21.5°C for the minimum temperature and 34.5°C for the maximum temperature

alert) when both the minimum and maximum temperatures exceed certain thresholds for at least three consecutive days. For illustrative purposes, we choose to assess the ability of MSTWeatherGen to simulate heatwaves with thresholds defined as the mean of the Météo-France Orange Vigilance alert thresholds over PACA (i.e., 21.5°C for the minimum temperature and 34.5°C for the maximum temperature). Figure 9 compares the observed and simulated frequency of heat waves during summer from 2012 to 2021, and shows that MSTWeatherGen effectively reproduces the frequency of this compound event, including its spatial patterns.

Simulation of multivariate environmental indices: the example of the Fire Weather Index

MSTWeatherGen has been designed with the objective of simulating weather data with realistic inter-variable dependencies. We therefore assess the simulation of the co-fluctuation of many variables by MSTWeatherGen, and illustrate such multivariate environmental process through the example of the risk of fire evaluated hereafter by the Canadian Forest Fire Weather Index (FWI, (Wang et al, 2017)).

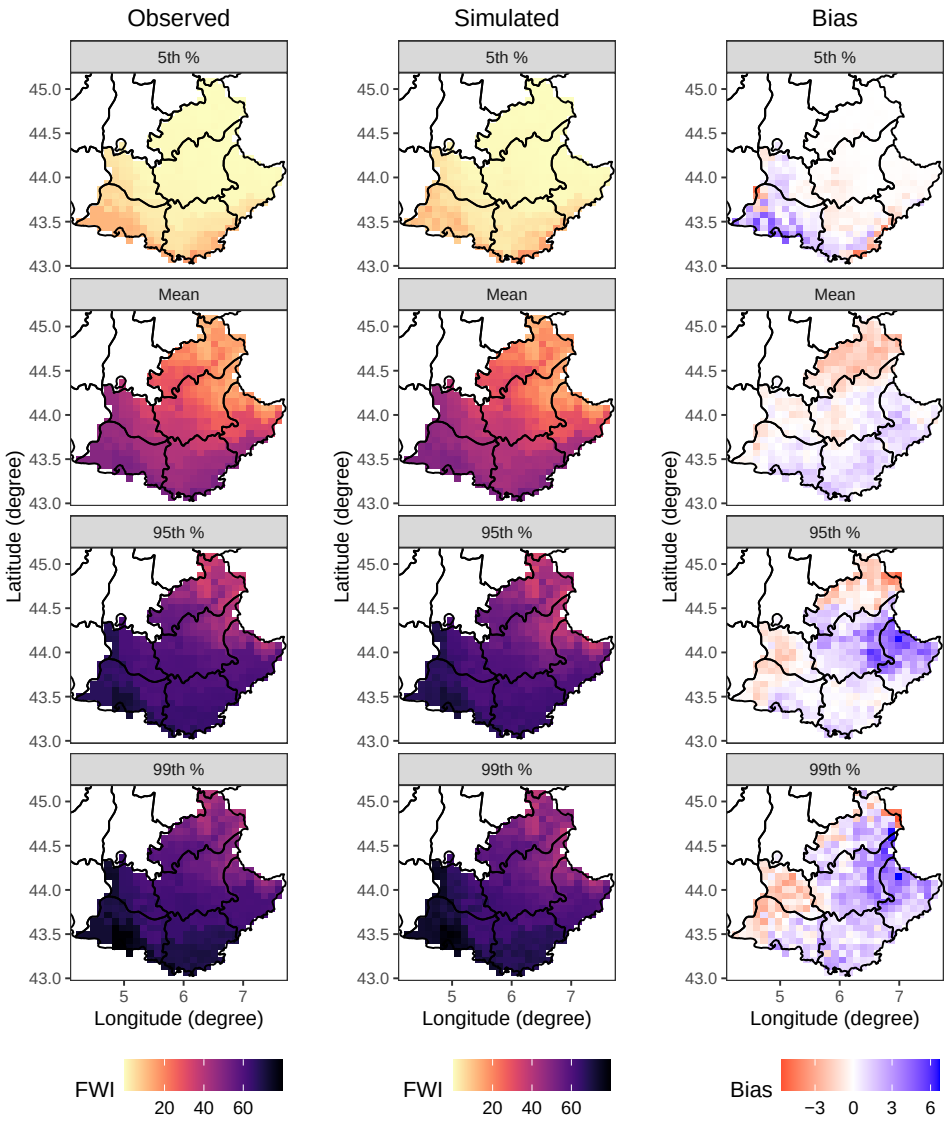

Fig. 10: Spatial distribution of key FWI statistics. Left and middle panels: spatial distribution of observed and simulated mean, 5th, 95th, and 99th percentile of the Fire Weather Index (FWI) during the summer periods from 2012 to 2021. Right panel: bias (observed minus simulated) of these statistics

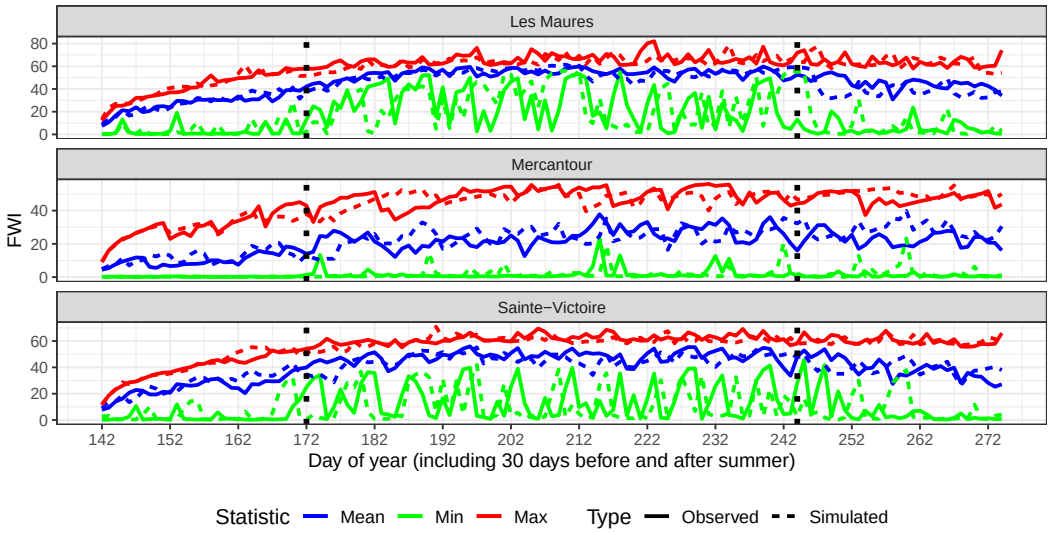

Fig. 11: Observed (solid line) and simulated (dashed line) daily FWI statistics over 10 years (2012 to 2021) in summer (including 30 days before and after summer) at three locations: Les Maures, Mercantour, and Sainte-Victoire. The mean is represented in blue, minimum in green, and maximum in red. Vertical dashed lines determine the summer period

The FWI is a meteorologically-based metric used to assess fire danger by integrating the joint influence and dynamics of several key weather variables: precipitation, temperature, relative humidity, and wind speed. We use the `cffdrs` R package (Wang et al, 2017) to calculate the summer FWI from 2012 to 2021 across PACA, for both observed and simulated data, and with default initialization during spring time. Figure 10 displays the spatial distribution of key FWI statistics computed on each series of 920 values (3 summer months $\times$ 10 years) across the studied area and shows that MSTWeatherGen effectively reproduces the overall spatial patterns of the FWI. The mean, the lower and upper percentiles of observations and simulations are in very good agreement. Indeed, the first row of Figure 10 shows an almost perfect simulation of the 5th percentile, except for a minor underestimation in the south-west of the area. The bias of the mean is also very small (magnitude less than 3), in particular when compared to the actual mean FWI whose spatial average is around 34. Finally, the bias of the high percentiles (95th and 99th) is slightly higher in FWI units, in particular in the north-east of the region where the underestimation reaches 6 for the 99th percentile, but once again it is very limited when compared to the high FWI values

reached in this configuration (for the 99th percentile the FWI ranges from 37 to 79 across PACA).

The dynamics of the Fire Weather Index (FWI) is also crucial for understanding the risk of fire over time. We examine this dynamics at three locations corresponding to landmark forest landscapes in Provence: the small mountains of Sainte-Victoire (43.532°N, 5.6128°E) and Les Maures (43.28°N, 6.384°E), and the Mercantour massif (44.13°N, 7.09°E), both densely vegetated with evergreen forests. The Sainte-Victoire and Les Maures areas experience hot and dry summers typical of the Mediterranean climate which lead to a high fire risk during the dry season. In contrast, the Mercantour massif experiences a Southern alpine climate with cooler and wetter summers, and hence a lower fire risk. For each of the three locations, we calculate the daily FWI statistics (mean, minimum, and maximum) during summer and over 10 years (2012 to 2021). Note that the difference between this approach and the one displayed in Figure 10 is that the statistics in Figure 11 are calculated on a daily basis, whereas those in Figure 10 are calculated for the all summer season. Figure 11 compares the observed and simulated FWI statistics in these locations during summer (i.e., the peak fire season), and demonstrates that MSTWeatherGen accurately captures the temporal dynamics of FWI at these sites. In all three locations, the FWI starts below 20 at the end of the winter, then gradually increases to reach its peak between July and August, and finally declines at the end of the summer. One can notice that at Les Maures and Sainte-Victoire locations the FWI often exceeds 38 on average (corresponding to a very high fire risk according to the European Forest Fire Information System (San-Miguel-Ayanz et al, 2012)), and that these high values are properly captured in MSTWeatherGen simulations. All in all, Figures 10 and 11 show that MSTWeatherGen effectively reproduces both the spatial and the temporal observed patterns and extremes of FWI, thereby providing reliable simulations for daily fire risk assessment. This is a very encouraging result considering that MSTWeatherGen was not designed to reproduce extreme multivariate combinations of the weather variables.

7 Concluding remarks

Splitting the modeling effort to embrace the complexity of the weather system

In the realm of stochastic weather generation, the simulation of realistic meteorological scenarios involves the joint simulation of multiple variables with complex interactions and space-time dynamics. This study tackles this challenge by splitting the stochastic

weather model into three components: a regional weather type process, a set of site-specific transformation functions, and a multivariate space-time Gaussian process. We discuss hereafter the influence of each component and their underlying hypotheses on the statistical and meteorological properties of the resulting synthetic weather data.

The regional weather types process has been designed to identify states within which the weather is expected to be homogeneous when considered at the regional scale. For example, the six weather types identified for the PACA region during winter and displayed in Figure 2 characterize the main weather patterns observed in this region during the cold season. By replicating such patterns, the SWG effectively reproduces the temporal variability of the regional weather system within each season, which in the present case is achieved through a non-homogeneous Markov chain as outlined in Section 3.1. One limitation of the current approach is the assumption of a first-order Markov chain, which may impact the temporal persistence of the weather types (Ailliot et al, 2015). If this problem emerges, a potential solution would be to condition the transition probabilities of the weather types on large-scale covariates such as the continental pattern of geopotential height or the phase of the North Atlantic Oscillation (Furrer and Katz, 2007).

MSTWeatherGen adopts a trans-Gaussian framework to model the multivariate and space-time behavior of the weather system within each weather type. Nonlinear transformation functions are central in this framework, as they link the standard Gaussian variables used for modeling with the target meteorological variables. The inverse of these transformation functions can be seen as normalization functions. In this study, we adopted the OQN method to design the transform functions (Peterson and Cavanaugh, 2019). This method effectively transforms complex continuous distributions, including multimodal ones, into standard Gaussian distributions. In practice, the OQN allows MSTWeatherGen to accurately reproduce the marginal distributions of all meteorological variables of interest, as shown in Figure 4. Furthermore, we successfully extended this method to accommodate truncated distributions, which is particularly useful for zero-inflated variables such as precipitation (Figure 5). Another advantage of the OQN is its ability to extrapolate the distribution of each meteorological variable beyond the range of observed values using a generalized linear model (Peterson and Cavanaugh, 2019). This is crucial in stochastic weather generation, and in particular when studying extreme events.

The dependence structure between variables is modeled by a latent Gaussian field with the non-separable multivariate and space-time covariance function proposed in Allard et al (2022). Our case study demonstrates that this covariance model aligns

well with the empirical covariances derived from a high-resolution reanalysis dataset over South-East France, and that this model captures the majority of inter-variable and intra-variable dependencies of the regional weather system (Figure 3). However, a few complex dependence structures challenge the current model, in particular when the second order non-stationarity of the latent field becomes substantial, for example due to the presence of strong elevation differences, despite our efforts to restrict it through seasonal weather typing and data transformation. To improve this aspect, new non-stationary, multivariate and space-time covariance models need to be designed.

MSTWeatherGen - a ready to use multivariate and space-time SWG

This paper introduces MSTWeatherGen, a stochastic weather generator designed to simulate a broad range of meteorological variables across space and time. The targeted resolution is daily in time, and 8 km x 8 km in space (typical of a regional-scale reanalysis dataset) with a footprint of the order of 500 pixels (corresponding to a French administrative Region) but MSTWeatherGen can be used at other scales and resolutions. The requirement is that there are enough pixels for a good estimation of the covariance function.

By integrating advancements in weather typing, Markov chain modeling, normalization functions, and multivariate space-time covariance functions, MSTWeatherGen offers a flexible and robust framework for generating realistic weather scenarios. Applications range from hydrology- or agriculture-focused impact studies to risk assessment, and in this field an illustrative case study showed that MSTWeatherGen has very good skills at simulating realistic Forest Fire Weather Index patterns and dynamics across a gradient of Mediterranean to mountain climates.

An R package has been developed to facilitate the use of MSTWeatherGen. It is available at https://github.com/sobakrim/MSTWeatherGen. This package allows users to calibrate MSTWeatherGen from any reanalysis or RCM climate projection dataset, and to swiftly emulate long (i.e., hundred of years) and realistic (i.e., faithfully reproducing the statistics of the training dataset) weather series. The resulting simulations preserve the space-time and multivariate dependencies of the target weather system, which is deemed essential for high resolution and multivariate impact studies and risk assessment.

Acknowledgments. The authors acknowledge the support of funding from the French National Agency (ANR) for the BEYOND project (contract 20-PCPA-0002)

that supported the work of all authors. The authors also acknowledge the support of the Chaire Geolearning.

Declarations

- Funding
  This research was funded by the French National Agency (ANR) as part of the BEYOND project (contract 20-PCPA-0002).
- Conflict of interest
  The authors declare no conflict of interest.
- Availability of data and materials
  The SAFRAN data can be accessed publicly at: https://meteo.data.gouv.fr/.
- Code availability
  MSTWeatherGen R package is available at https://github.com/sobakrim/MSTWeatherGen.
- Authors' contributions
  All authors contributed to the study conception and design. SO designed and implemented the model, performed the numerical experiments of the case study, and created the R package. SO wrote the paper with input and corrections from all co-authors. All authors read and approved the final manuscript.

References

Ailliot P, Thompson C, Thomson P (2009) Space-time modelling of precipitation by using a hidden Markov model and censored Gaussian distributions. Journal of the Royal Statistical Society Series C: Applied Statistics 58(3):405–426. https://doi.org/10.1111/j.1467-9876.2008.00654.x

Ailliot P, Allard D, Monbet V, et al (2015) Stochastic weather generators: an overview of weather type models. Journal de la société française de statistique 156(1):101–113

Ailliot P, Boutigny M, Koutroulis E, et al (2020) Stochastic weather generator for the design and reliability evaluation of desalination systems with Renewable Energy Sources. Renewable Energy 158:541–553. https://doi.org/10.1016/j.renene.2020.05.076

Allard D, Bourotte M (2015) Disaggregating daily precipitations into hourly values with a transformed censored latent Gaussian process. Stochastic Environmental Research and Risk Assessment 29:453–462. https://doi.org/10.1007/

s00477-014-0913-4

Allard D, Clarotto L, Opitz T, et al (2021) Discussion on "Competition on Spatial Statistics for Large Datasets". Journal of Agricultural, Biological and Environmental Statistics 26(4):604–611. https://doi.org/10.1007/s13253-021-00462-2

Allard D, Clarotto L, Emery X (2022) Fully nonseparable Gneiting covariance functions for multivariate space-time data. Spatial Statistics 52:100706. https://doi.org/10.1016/j.spasta.2022.100706

Allcroft DJ, Glasbey CA (2003) A latent Gaussian Markov random-field model for spatiotemporal rainfall disaggregation. Journal of the Royal Statistical Society Series C: Applied Statistics 52(4):487–498. https://doi.org/10.1111/1467-9876.00419

Apanasovich TV, Genton MG (2010) Cross-covariance functions for multivariate random fields based on latent dimensions. Biometrika 97(1):15–30. https://doi.org/10.1093/biomet/asp078

Bartlett MS (1947) The use of transformations. Biometrics 3(1):39–52. https://doi.org/10.2307/3001536

Baxevani A, Lennartsson J (2015) A spatiotemporal precipitation generator based on a censored latent Gaussian field. Water Resources Research 51(6):4338–4358. https://doi.org/10.1002/2014WR016455

Baxevani A, Lenzi A (2018) Very short-term spatio-temporal wind power prediction using a censored Gaussian field. Stochastic Environmental Research and Risk Assessment 32:931–948. https://doi.org/10.1007/s00477-017-1435-7

Benoit L, Mariethoz G (2017) Generating synthetic rainfall with geostatistical simulations. Wiley Interdisciplinary Reviews: Water 4(2):e1199. https://doi.org/10.1002/wat2.1199

Benoit L, Allard D, Mariethoz G (2018) Stochastic rainfall modeling at sub-kilometer scale. Water Resources Research 54(6):4108–4130. https://doi.org/10.1029/2018WR022817

Bevacqua E, Suarez-Gutierrez L, Jézéquel A, et al (2023) Advancing research on compound weather and climate events via large ensemble model simulations. Nature Communications 14(1):2145. https://doi.org/10.1038/s41467-023-37847-5

Boé J, Terray L, Habets F, et al (2006) A simple statistical-dynamical downscaling scheme based on weather types and conditional resampling. Journal of Geophysical Research: Atmospheres 111(D23). https://doi.org/10.1029/2005JD006889

Bourotte M (2016) Générateur stochastique de temps multisite basé sur un champ gaussien multivarié. PhD thesis, Université d'Avignon

Bourotte M, Allard D, Porcu E (2016) A flexible class of non-separable cross-covariance functions for multivariate space–time data. Spatial Statistics 18:125–146. https://doi.org/10.1016/j.spasta.2016.02.004

Boutigny M, Ailliot P, Chaubet A, et al (2023) A meta-Gaussian distribution for sub-hourly rainfall. Stochastic Environmental Research and Risk Assessment pp 1–13. https://doi.org/10.1007/s00477-023-02487-0

Box GE, Cox DR (1964) An analysis of transformations. Journal of the Royal Statistical Society Series B: Statistical Methodology 26(2):211–243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Burkardt J (2014) The truncated normal distribution. Department of Scientific Computing Website, Florida State University 1:35

Chen W, Genton MG, Sun Y (2021) Space-time covariance structures and models. Annual Review of Statistics and Its Application 8:191–215. https://doi.org/10.1146/annurev-statistics-042720-115603

Chilès JP, Delfiner P (2012) Geostatistics: Modeling Spatial Uncertainty, Second Edition. John Wiley & Sons, https://doi.org/10.1002/9781118136188

Cressie N (2015) Statistics for spatial data. John Wiley & Sons, https://doi.org/10.1002/9781119115151

Dabhi H, Rotach MW, Dubrovskỳ M, et al (2021) Evaluation of a stochastic weather generator in simulating univariate and multivariate climate extremes in different climate zones across Europe. Meteorologische Zeitschrift 30(2):127–151. https://doi.org/10.1127/metz/2020/1021

De Iaco S, Palma M, Posa D (2019) Choosing suitable linear coregionalization models for spatio-temporal data. Stochastic Environmental Research and Risk Assessment 33:1419–1434. https://doi.org/10.1007/s00477-019-01701-2

Dörr C, Schlather M (2023) Characterization theorems for pseudo cross-variograms. Journal of Applied Probability 60(4):1219–1231. https://doi.org/10.1017/jpr.2022.133

Fatichi S, Vivoni ER, Ogden FL, et al (2016) An overview of current applications, challenges, and future trends in distributed process-based models in hydrology. Journal of Hydrology 537:45–60. https://doi.org/10.1016/j.jhydrol.2016.03.026

Flecher C, Naveau P, Allard D, et al (2010) A stochastic daily weather generator for skewed data. Water Resources Research 46(7). https://doi.org/10.1029/2009WR008098

Furrer EM, Katz RW (2007) Generalized linear modeling approach to stochastic weather generators. Climate Research 34(2):129–144. https://doi.org/10.3354/cr034129

Genton MG, Kleiber W (2015) Cross-covariance functions for multivariate geostatistics. Statistical Science 30(2):147–163. https://doi.org/10.1214/14-STS487

Gneiting T (2002) Nonseparable, stationary covariance functions for space-time data. Journal of the American Statistical Association 97(458):590–600. https://doi.org/10.1198/016214502760047113

Gneiting T, Genton MG, Guttorp P (2006) Geostatistical space-time models, stationarity, separability, and full symmetry. Monographs On Statistics and Applied Probability 107:151

Gneiting T, Kleiber W, Schlather M (2010) Matérn cross-covariance functions for multivariate random fields. Journal of the American Statistical Association 105(491):1167–1177. https://doi.org/doi.org/10.1198/jasa.2010.tm09420

Han C, Zhang B, Chen H, et al (2019) Spatially distributed crop model based on remote sensing. Agricultural Water Management 218:165–173. https://doi.org/10.1016/j.agwat.2019.03.035

Kimball BA, Thorp KR, Boote KJ, et al (2023) Simulation of evapotranspiration and yield of maize: An Inter-comparison among 41 maize models. Agricultural and Forest Meteorology 333:109396. https://doi.org/10.1016/j.agrformet.2023.109396

Leibovici D, Sabatier R (1998) A singular value decomposition of a k-way array for a principal component analysis of multiway data, PTA-k. Linear Algebra and its Applications 269(1-3):307–329. https://doi.org/10.1016/S0024-3795(97)81516-9

Leutbecher M (2019) Ensemble size: How suboptimal is less than infinity? Quarterly Journal of the Royal Meteorological Society 145:107–128

Lobell DB, Field CB (2007) Global scale climate–crop yield relationships and the impacts of recent warming. Environmental Research Letters 2(1):014002. https://doi.org/10.1088/1748-9326/2/1/014002

Mathlouthi M, Lebdi F (2021) Comprehensive study of the wet and dry spells and their extremes in the Mediterranean climate basin northern tunisia. SN Applied Sciences 3:1–17. https://doi.org/10.1007/s42452-021-04834-8

Mearns L, Easterling W, Hays C, et al (2001) Comparison of agricultural impacts of climate change calculated from high and low resolution climate change scenarios: Part i. the uncertainty due to spatial scale. Climatic Change 51:131–172. https://doi.org/10.1023/A:1012297314857

Milly PC, Dunne KA, Vecchia AV (2005) Global pattern of trends in streamflow and water availability in a changing climate. Nature 438(7066):347–350. https://doi.org/10.1038/nature04312

Ouzeau G, Soubeyroux JM, Schneider M, et al (2016) Heat waves analysis over France in present and future climate: Application of a new method on the EURO-CORDEX ensemble. Climate Services 4:1–12. https://doi.org/10.1016/j.cliser.2016.09.002

Paschalis A, Molnar P, Fatichi S, et al (2013) A stochastic model for high-resolution space-time precipitation simulation. Water Resources Research 49(12):8400–8417. https://doi.org/10.1002/2013WR014437

Peleg N, Fatichi S, Paschalis A, et al (2017) An advanced stochastic weather generator for simulating 2-d high-resolution climate variables. Journal of Advances in Modeling Earth Systems 9(3):1595–1627. https://doi.org/10.1002/2016MS000854

Peterson RA, Cavanaugh JE (2019) Ordered quantile normalization: a semiparametric transformation built for the cross-validation era. Journal of Applied Statistics https://doi.org/10.1080/02664763.2019.1630372

Porcu E, Furrer R, Nychka D (2021) 30 years of space-time covariance functions. Wiley Interdisciplinary Reviews: Computational Statistics 13(2):e1512. https://doi.org/10.1002/wics.1512

Quintana-Segui P, Le Moigne P, Durand Y, et al (2008) Analysis of near-surface atmospheric variables: Validation of the SAFRAN analysis over France. Journal of Applied Meteorology and Climatology 47(1):92–107. https://doi.org/10.1175/2007JAMC1636.1

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

San-Miguel-Ayanz J, Schulte E, Schmuck G, et al (2012) Comprehensive monitoring of wildfires in Europe: the European forest fire information system (EFFIS). In: Approaches to managing disaster-Assessing hazards, emergencies and disaster impacts. IntechOpen, https://doi.org/10.5772/28441

Scrucca L, Fraley C, Murphy TB, et al (2023) Model-Based Clustering, Classification, and Density Estimation Using mclust in R. Chapman and Hall/CRC, https://doi.org/10.1201/9781003277965, URL https://mclust-org.github.io/book/

Semenov MA, Brooks RJ, Barrow EM, et al (1998) Comparison of the WGEN and LARS-WG stochastic weather generators for diverse climates. Climate Research 10(2):95–107. https://doi.org/10.3354/cr010095

Sparks NJ, Hardwick SR, Schmid M, et al (2018) Image: a multivariate multisite stochastic weather generator for European weather and climate. Stochastic Environmental Research and Risk Assessment 32:771–784. https://doi.org/10.1007/s00477-017-1433-9

Van Wagner C (1987) Development and structure of the Canadian forest fire weather index system. Tech. rep., Canadian Forestry Service

Varin C, Reid N, Firth D (2011) An overview of composite likelihood methods. Statistica Sinica pp 5–42

Verdin A, Rajagopalan B, Kleiber W, et al (2015) Coupled stochastic weather generation using spatial and generalized linear models. Stochastic Environmental Research

and Risk Assessment 29:347–356. https://doi.org/10.1007/s00477-014-0911-6

Verdin A, Rajagopalan B, Kleiber W, et al (2019) BayGEN: A Bayesian space-time stochastic weather generator. Water Resources Research 55(4):2900–2915. https://doi.org/10.1029/2017WR022473

Wang X, Wotton BM, Cantin AS, et al (2017) cffdrs: an R package for the canadian forest fire danger rating system. Ecological Processes 6:1–11. https://doi.org/10.1186/s13717-017-0070-z

Wanner H, Brönnimann S, Casty C, et al (2001) North Atlantic Oscillation – concepts and studies. Surveys in Geophysics 22:321–381. https://doi.org/10.1023/A:1014217317898

Westerling AL, Hidalgo HG, Cayan DR, et al (2006) Warming and earlier spring increase western US forest wildfire activity. Science 313(5789):940–943. https://doi.org/10.1126/science.1128834

Wilks DS, Wilby RL (1999) The weather generation game: a review of stochastic weather models. Progress in Physical Geography 23(3):329–357. https://doi.org/10.1002/qj.3387

Williams CK, Rasmussen CE (2006) Gaussian processes for machine learning. MIT press Cambridge, MA, https://doi.org/10.1142/S0129065704001899

Yiou P, Jézéquel A (2020) Simulation of extreme heat waves with empirical importance sampling. Geoscientific Model Development 13(2):763–781. https://doi.org/10.5194/gmd-13-763-2020