Spatio-temporal weather generator for the temperature over France

Caroline Cognot [1&2] & Liliane Bel [2] & Sylvie Parey [1] & David Métivier [3]

[1] *EDF R&D, 91120, Palaiseau, France - caroline.cognot@edf.fr*
[2] *Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120, Palaiseau, France*
[3] *UMR MISTEA, 34000, Montpellier, France*

Résumé. Les générateurs de temps sont des simulateurs qui permettent de reproduire la variabilité climatique et générer un grand nombre de situations pour des variables météorologiques selon un modèle statistique ajusté sur les observations. La plupart des générateurs de temps fournissent des simulations pour une ou plusieurs variables météorologiques site par site. Dans ce travail, nous nous proposons de concevoir un générateur spatialisé pour une seule variable météorologique. Nous nous focalisons sur la température. La moyenne et la variance sont décomposées chacune en une tendance et une saisonnalité. Les tendances sont modélisées par des fonctions non paramétriques, les saisonnalités par des polynômes trigonométriques. Chaque fonction est ajustée site par site, puis étendue à tout l'espace par krigeage pour les saisonnalités et par distance inverse pour la partie tendance. La partie stochastique est modélisée par un champ gaussien avec une fonction de corrélation spatio-temporelle non séparable. Ces deux étapes permettent de simuler sur une grille, sur n'importe quelle période de temps. La validation du modèle a été réalisée à la fois sur le jeu de données des stations et sur la grille. En calculant plusieurs indicateurs liés à la structure spatiale, à la structure temporelle ou aux extrêmes, on constate que le générateur fournit des simulations adéquates et permet la génération de séries de températures spatialement cohérentes, à l'exception toutefois des extrêmes élevés, ce qui était attendu, les processus gaussiens étant connus pour être peu performants sur cet aspect.

Mots-clés. Générateur de temps, processus spatio-temporel, processus Gaussiens

Abstract.

Weather generators are simulators useful to reproduce climate variability and generate a great number of situations for meteorological variables according to a statistical model fitted on observations. Most weather generators provide simulations for one or several meteorological variables site by site, we aim in this work to design a spatial weather generator for one meteorological variable. We focus on temperature. The mean and variance are each modeled as a trend and a seasonality function. Trends are non-parametric functions while seasonality parts are written as trigonometric polynomials. They are fitted separately on each site and then extended to the whole domain by kriging for the seasonality coefficient and by inverse distance weighting for the trends. The stochastic part is modeled as a Gaussian field with a non-separable spatio-temporal correlation function. These two steps allow for simulation on a grid, over any time period. The validation of the model was conducted both on the stations dataset and the grid. By computing several indicators related to the spatial structure, the

temporal structure or extremes, it is found that the generator provides adequate simulations and allows the generation of spatially coherent temperature series, except for high extremes, which was to be expected as Gaussian processes are known to perform poorly in this aspect.

Keywords. Weather generator, Spatio-temporal process, Gaussian Processes

1 Introduction

Weather generators consider the observations of a meteorological variable to be one of the infinite possible trajectories of a stochastic model. The main objective is to allow the simulation of many plausible sequences of a meteorological variable, that share common statistical properties with the observations. The parameters are estimated on the observations. Simulations are fast and allow for a more complete sampling of the climate variables than physical models. Since the first weather generator of Richardson [1981] many others have been proposed using various statistical tools, for univariate and multivariate series. In the case of multisite temperature-focused models, Dubrovsky et al. [2020] propose an auto-regressive model with a high number of matrix parameters to estimate, and Sparks et al. [2018] use Empirical Orthogonal Functions for dimension reduction. These approaches however do not take advantage of the spatial structure of the data. A good option is to turn to geostatistics. The main improvement we can get from these tools is the use of spatial information in the dependencies, with correlation functions that involve spatial distances. Many models use this concept using one or several Gaussian fields with a chosen covariance function. They are combined if necessary with an anamorphosis function, as in Kleiber et al. [2012], Verdin et al. [2015], or Sparks et al. [2018], in order to be used on non-Gaussian data, mainly precipitations. These Gaussian fields can have varying complexity, with spatio-temporal covariance function ranging from the simple $C(h,t) = \exp(-||h||/A(t))$ with h the distance between spatial points and A a scale function in Kleiber et al. [2012] to non-separable covariance functions in Bourotte et al. [2016],Allard et al. [2022].

In this work, we focused on building a spatio-temporal Stochastic Weather Generator (SWG) for the temperature. The model is built in two steps: first, the mean and variance are represented as a sum of trend and seasonality functions. Trends are depicted as non-parametric functions, whereas the seasonality components are expressed using trigonometric polynomials. Then, a Gaussian field model is fitted on the residuals.

2 Methods

2.1 Preprocessing of the temperature

Single-site decomposition : Let $X(s,t)$ be the temperature at site s and time t, with $s \in \{s_1, ..., s_{n_S}\} \subset D$, and $t \in \{1, .., n_T\}$. At each station, we use the decomposition of Hoang [2010] to separate the mean and variance into trend and seasonality components.

The decomposition writes

$$X(s,t) = T_m(s,t) + S_m(s,t) + T_\sigma(s,t)S_\sigma(s,t)Z(s,t), \tag{1}$$

where T_m is the trend in mean, S_m is the seasonality in mean, T_σ is the trend in standard-deviation and S_σ is the seasonality in standard-deviation. Z is the residual signal, carrying the random information.

The seasonality terms are modeled by trigonometric polynomials of the following form:

$$S(s,t) = \beta_1(s) + \sum_{i=1}^{d} \left[\beta_{2i}(s) \cos\left(\frac{2i\pi t}{365}\right) + \beta_{2i+1}(s) \sin\left(\frac{2i\pi t}{365}\right) \right]. \tag{2}$$

For each site s, the different components in Eq. (1) are obtained iteratively using LOESS (LOcally Estimated Scatterplot Smoothing) regression for the trend terms T_m, T_σ and linear regression for the trigonometric polynomial coefficients $\boldsymbol{\beta}_m(s), \boldsymbol{\beta}_\sigma(s)$ involved in S_m, S_σ. The LOESS regression gives a non-parametric form to the trends, avoiding the over-simplification of linear trends. This model takes into account the seasonal cycle of the variance, which is often neglected in most weather generators.

Spatial extension: In order to use the model anywhere in space, the trends T_m, T_σ and the coefficients of the seasonality $\boldsymbol{\beta}_m(s), \boldsymbol{\beta}_\sigma(s)$ have to be extended. The trends are assumed to slowly vary in time and are interpolated using inverse distance weighting. For the seasonality coefficients, we use ordinary kriging.

2.2 A geostatistics model

Once the coefficients of the decomposition are estimated, we are left with centered residuals Z of variance 1 when averaged over a year. They are spatially and temporally correlated, and in a sense, represent how the weather at time t, site s deviates from a generic value. Our approach for the residuals $Z(s,t)$ makes use of geostatistics to include geographical proximity in the dependence structure, and allow simulation at any new point in space.

We suppose $Z(.)$ to be a Gaussian stationary process, with mean 0. Under these two assumptions, it is entirely described by its covariance function $C(h,u)$ where h is a space lag and u is a time lag. We choose to use the Gneiting-Matern covariance model (16) of Gneiting [2002], a non-separable model that allows interaction between space and time,

$$C(h,u) = \frac{\sigma^2}{(\alpha u^{2a}+1)^\tau} \mathcal{M}\left(\frac{h}{\sqrt{\alpha u^{2a}+1}^b}; r; \nu \right), \tag{3}$$

where $\mathcal{M}(d,r,\nu)$ is the value of the Matern covariance function at distance d, with smoothness parameter ν and scale parameter r.

Simulation When working with spatiotemporal grids of limited size, simulating a Gaussian process with the covariance model in Eq.(3) is straightforward, using standard simulation procedures such as Cholesky decomposition for instance. When working with finer grids and long periods of time, this is no longer computationally feasible. We instead choose to simulate sequentially in time making the values at time t only depending on the l previous time steps, with l coherent with the temporal range parameter that is estimated. This is made possible by the Gaussian properties of the model allowing easy simulation conditionally to the previous time steps.

2.3 Validation

To validate such a model, we choose to look at various indicators, such as the annual cycle in mean and variance at each point or the pairwise correlations between pairs of sites. The main idea is to compute many simulations, obtain an empirical distribution of these indicators, and compare them to the observed values. The model is adequate if the observed values are in the range of the simulations.

Pairwise correlations between sites For each pair of sites i, j, we compute for the observations and 100 simulations the pairwise empirical correlations of the temperature X.

Pairwise conditional threshold exceedance We are interested in how extremes in one location are related to the same extreme in another location, and we look at pairwise conditional threshold exceedances for high (resp. low) quantiles.

For each station i, define the probability α and quantile $q_\alpha(i)$ such that $\alpha = P(X_i > q_\alpha(i))$ (resp. $x = P(X_i < q_\alpha(i))$) in the observations. Then, for every pair of stations i, j, define

$$p_{i,j}^{\alpha} = P(X_i > q_\alpha(i) | X_j > q_\alpha(j)) \tag{4}$$

$$\hat{p}_{i,j}^{\alpha} = \frac{\sum_{t=1}^{N_t} \mathbb{1}_{X_i > q_\alpha(i) \cap X_j > q_\alpha(j)}}{\sum_{t=1}^{N_t} \mathbb{1}_{X_j > q_\alpha(j)}}. \tag{5}$$

With inverted signs in the case of low quantiles. The values of $\hat{p}_{i,j}^{x}$ for all i, j in the observations are then compared with the values from the simulations.

Temporal correlation As we wish to be able to represent well the temporal structure of the data, we also look at the temporal correlation function. In particular, we compare the lagged correlation in the full covariance simulation and in the simulations where we use only 10 previous days to simulate time t.

3 Results

3.1 Parameter estimation

The data used is a set of 41 ECA&D (Klein Tank [2002]) stations in France. The trends and seasonality are estimated on 61 years (1960-2020) and the covariance parameters are estimated on 31 years (1985-2015). The covariance parameters for (3) are computed with a composite pairwise likelihood estimation using R package GeoModels (Bevilacqua et al. [2018]) for each extended season each year (extended winter: September- October - November - December - January- February; extended summer: March - April -May-June-July-August). This separation is both for the sake of computational efficiency and to account for differences between the seasons observed during an initial study. This leads to 30 sets for winter and 31 for summer (the first and last winter are not complete when working on data that start from 1st January), see table 1 for the median and standard deviation of the found parameters. The final set of parameters is chosen as the median over all years for each of the parameters.

Parameter	Winter		Summer	
	median	sd	median	sd
σ^2	1.0199e+00	1.252e-01	9.713e-01	1.474e-01
τ	1.006e+00	8.715e-01	1.142e+00	2.606e-01
ν	8.662e-01	1.750e-01	6.834e-01	1.282e-01
$2a$	1.507e+00	2.405e-01	1.619e+00	2.085e-01
$\frac{r}{1000}$	1.000e+00	1.173e-03	9.997e-01	1.026e-03
$\frac{1}{\alpha}$	3.118e+00	1.082e+00	2.871e+00	4.072e-01
nugget	9.631e-02	2.297e-02	1.009e-01	2.285e-02

Table 1: Estimated covariance parameters

We find some of the parameters to be quite different between the two seasons. Notably, the smoothing Matérn parameter and the variance are higher in winter than in summer. The temporal scale parameter tends to be higher in the winter but the results are also a lot more dispersed than for the summer.

We also find that seasonal coefficients vary with the geographic coordinates. Compared to a simple longitude/latitude regression, kriging the values of the seasonal coefficients allows to highlight regional characteristics of the French climate. As seen in Figure 1 showing the kriging for the mean coefficients, we can retrieve the regions with oceanic, continental, and Mediterranean climates.

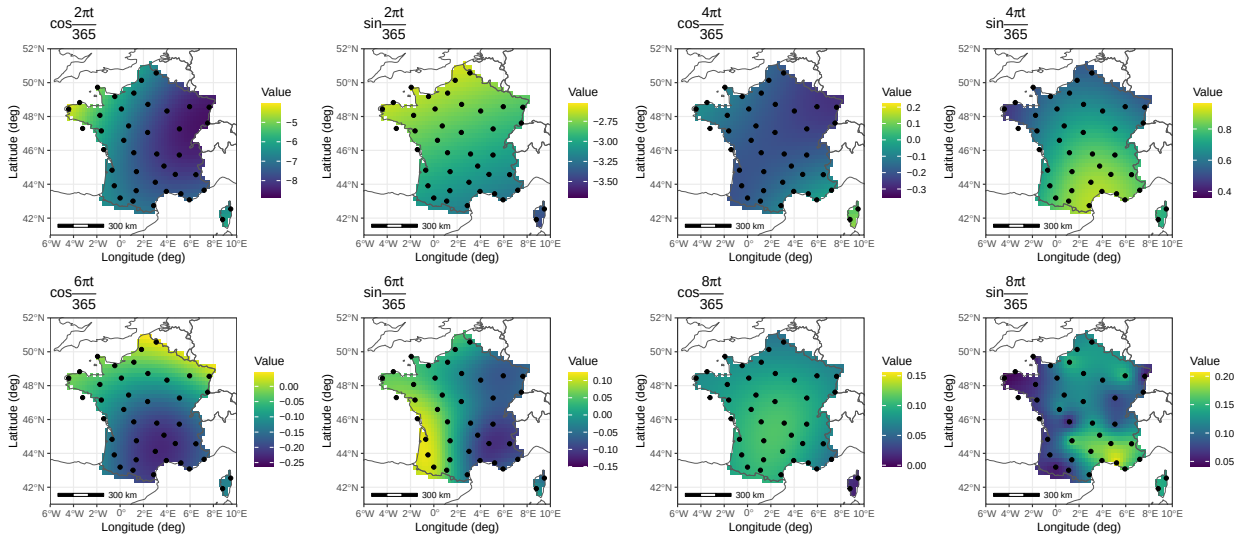

Figure 1: Kriging the coefficients of S_m.

3.2 Simulation at the fitting points

Temperature series are simulated 100 times at each of the 41 fitting data points in three ways. First, a 31 years simulation is obtained by first simulating 31 summers and 31 winters from the fitted covariance model in Eq. (3) and then adding the deterministic parts in Eq. (1). Then, we simulate from Eq. (3) using 10 days conditional. Finally, we simulate using only a temporal covariance function, to evaluate the gain of the spatial part.

We find that although the spatial correlation is necessarily embedded in the removed deterministic parts, the spatial structure of the residuals is still important to correctly reproduce the spatial correlation of temperature. In figure 2 (left), we plot all the pairwise correlations between stations and compare them to the values computed from the observations. We expect a good model to give points close to the 1:1 line in black. Compared to simulating from a non-spatial model (blue dots), the pairwise correlations between stations are indeed improved by our model (red dots for simulations with full covariance, almost entirely covered by the green dots from the 10 days conditional simulation). Figure 2 (right) shows, for every pair of station i, j, the probability of station i exceeding a quantile, given station j does. Again, we expect points close to the 1:1 line, which is not entirely true, especially for very low (q1) or very high (q99) quantiles. This probably comes from the fact that Gaussian fields are not able to produce extremal dependence. However, there is definitely an improvement in using a spatiotemporal model instead of a purely temporal one.

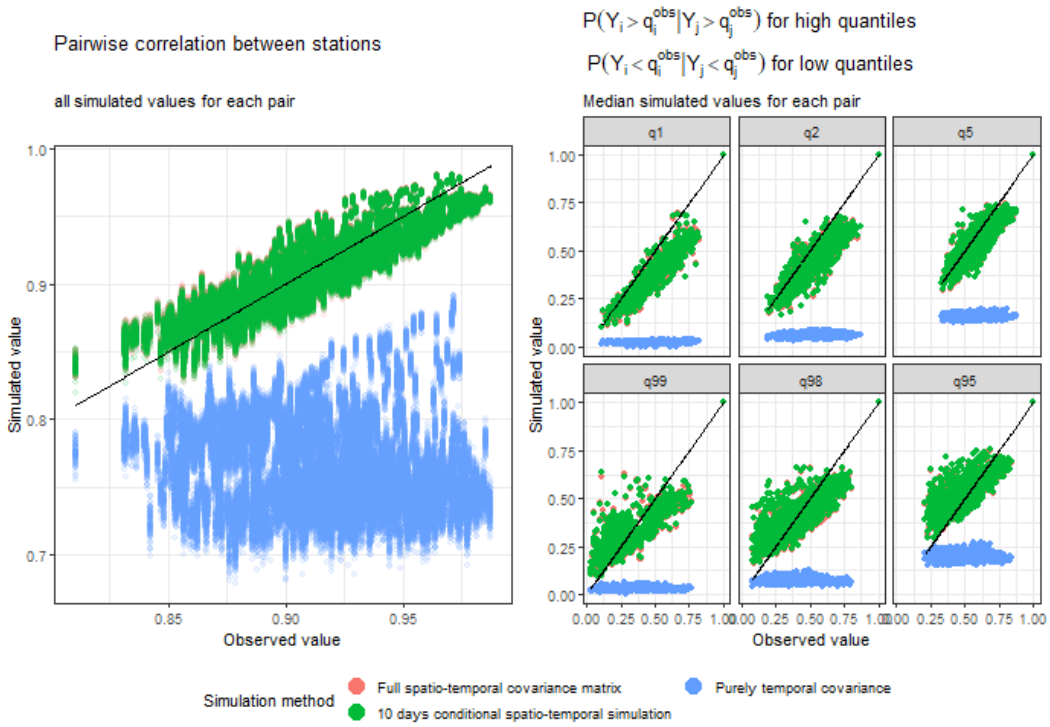

Figure 2: Pairwise correlation between stations (left) and extremal dependence (right).

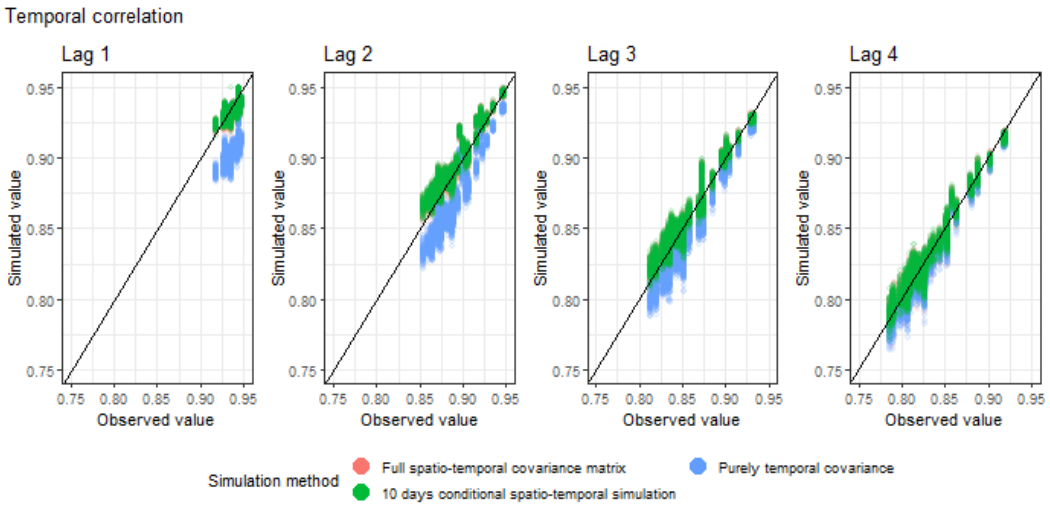

Figure 3: Temporal correlations at each of the 41 stations

The lagged temporal correlations, in figure 3, are also quite well represented in the simulations.

3.3 Simulation on a regular grid

We make simulations on a grid using the kriged coefficients in 1, the interpolated trends and residuals simulated from Eq. (3) using the 10 days conditional. Two simulations for the same period in June 2005 are represented in figure 4 .

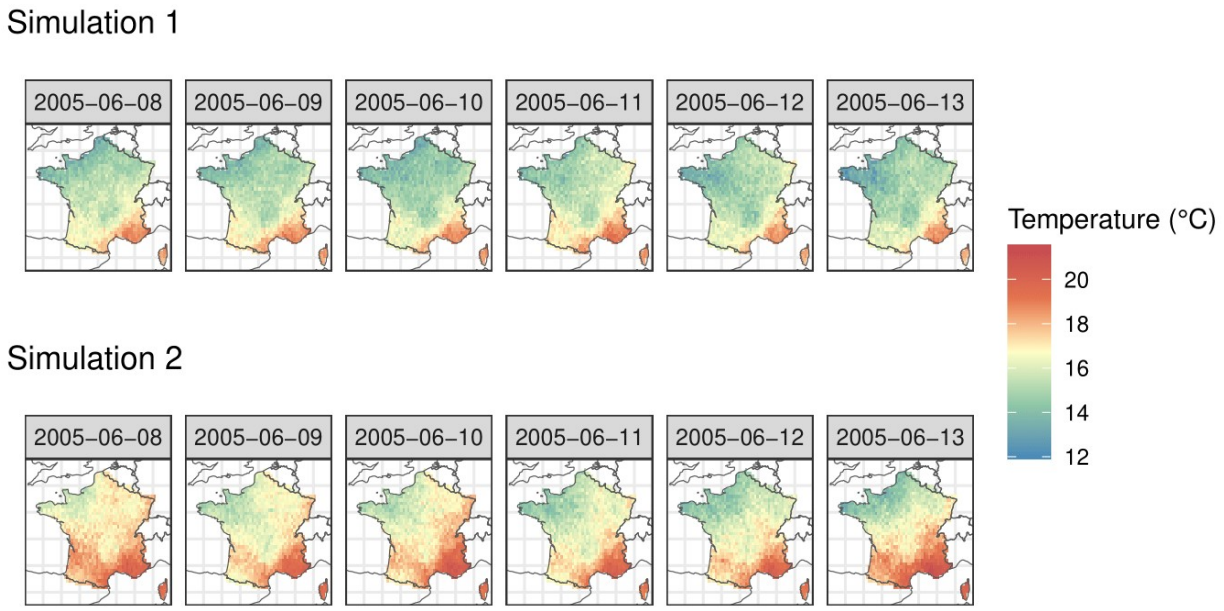

Figure 4: Two simulations, equivalent to the same period in June 2005

The simulated temperature values depend both on the deterministic parts (trends and seasonality) and the stochastic part. We see on both of the 6-day sequences that the two simulations give different results: the simulated temperature is a lot warmer in the southwest in the second simulation than in the first.

4 Conclusion and perspectives

In this work, we construct a spatio-temporal weather generator for the temperature. It can be used to simulate on grids over any period of time, provided trends are available. Trends in mean and standard deviation can be derived from climatic studies while supposing the stochastic structure to be stationary.

Our simulations will be compared with the E-OBS dataset. The E-OBS dataset is a gridded set of observations, obtained from interpolating from the ECA&D dataset, the same we used in this study. Because of that, it can be interesting to see whether our model is consistent with it.

We will investigate other simulation methods for Gaussian processes in order to be able to handle large datasets and several meteorological variables.

References

D. Allard, L. Clarotto, and X. Emery. Fully nonseparable gneiting covariance functions for multivariate space–time data. Spatial Statistics, 52:100706, 2022.

M. Bevilacqua, V. Morales-Oñate, and C. Caamaño-Carrillo. GeoModels: Procedures for Gaussian and Non Gaussian Geostatistical (Large) Data Analysis, 2018. URL https://vmoprojs.github.io/GeoModels-page/. R package version 1.0.0.

M. Bourotte, D. Allard, and E. Porcu. A flexible class of non-separable cross-covariance functions for multivariate space–time data. Spatial Statistics, 18:125–146, 2016.

M. Dubrovsky, R. Huth, H. Dabhi, and M. W. Rotach. Parametric gridded weather generator for use in present and future climates: focus on spatial temperature characteristics. Theoretical and Applied Climatology, 139:1031–1044, 2020.

T. Gneiting. Nonseparable, stationary covariance functions for space–time data. Journal of the American Statistical Association, 97(458):590–600, 2002.

T. T. H. Hoang. Modeling of non-stationnary, non-linear time series - Application to the definition of trends on mean, variability and extremes of air temperatures in Europe. PhD thesis, Université Paris Sud-Paris XI, 2010.

W. Kleiber, R. W. Katz, and B. Rajagopalan. Daily spatiotemporal precipitation simulation using latent and transformed gaussian processes. Water Resources Research, 48(1), 2012.

A. a. C. Klein Tank. Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. Int. J. of Climatol., 22, 1441-1453, 2002.

C. W. Richardson. Stochastic simulation of daily precipitation, temperature, and solar radiation. Water resources research, 17(1):182–190, 1981.

N. J. Sparks, S. R. Hardwick, M. Schmid, and R. Toumi. Image: a multivariate multi-site stochastic weather generator for european weather and climate. Stochastic environmental research and risk assessment, 32:771–784, 2018.

A. Verdin, B. Rajagopalan, W. Kleiber, and R. W. Katz. Coupled stochastic weather generation using spatial and generalized linear models. Stochastic Environmental Research and Risk Assessment, 29:347–356, 2015.