Regional sub-daily stochastic weather generator based on reanalyses for surface water stress estimation in central Tunisia

Nesrine Farhani[a,b], Julie Carreau[c,d,*], Zeineb Kassouk[a], Bernard Mougenot[b], Michel Le Page[b], Zohra Lili-Chabaane[a], Rim Zitouna-Chebbi[e], Gilles Boulet[b]

[a] *Université de Carthage, Institut National Agronomique de Tunisie, Lr GREEN-TEAM, 43 avenue Charles Nicolle, Tunis, Tunisie*
[b] *Centre d'Études Spatiales de la Biosphère, Univ. de Toulouse, CNRS, CNES, IRD, UPS, INRAE, Toulouse, France*
[c] *HydroSciences Montpellier, Univ. de Montpellier, CNRS, IRD, Montpellier, France*
[d] *Polytechnique Montréal, Montréal, Canada*
[e] *LR VENC, INRGREF, University of Carthage, Rue Hedi Karray 2080, Ariana, Tunisia*

Abstract

We present `MetGen`: a sub-daily multi-variable stochastic weather generator implemented as an `R` library that can be used to perform gap-filling and to extend in time meteorological observation series. `MetGen` is tailored to provide surrogate series of air temperature, relative air humidity, global radiation and wind speed needed for surface water stress estimation that requires sub-daily resolution. Multiple gauged stations can be used to increase the calibration data although spatial dependence is not modeled. The approach relies on Generalized Linear Models that use, among their covariates, large-scale variables derived from ERA5 reanalyses. `MetGen` aims at preserving key features of the meteorological variables along with inter-variable dependencies. We illustrate the abilities of `MetGen` using a case study with three stations in central Tunisia. We consider as alternatives a univariate and a multivariate bias correction techniques along with the un-processed large-scale variables.

Keywords: stochastic weather generator, bias correction, surface water stress estimation, sub-daily resolution, ERA5 reanalyses

1. Introduction

In semi-arid areas, water is a major limitation factor for agricultural production. Indeed, these areas are characterized by short rainy seasons and strong variability of precipitation events in time and space [1]. Natural variations in the water cycle affect the availability of water, leading to irregularities in agricultural production [2] and constitutes the main driver of agricultural droughts. The vegetation health status being generally representative of water availability [3], an important issue concerns the detection of surface water stress and the estimation of evapotranspiration (ET). Water stress may be deduced from ET with energy balance models. At satellite overpass time, energy balance models compute instantaneous ET as the residual term of

[*] Corresponding author
Email address: julie.carreau@ird.fr (Julie Carreau)

the land surface energy balance equation, once net radiation, soil heat flux and sensible heat flux are derived from remotely sensed surface temperature [4, 5, 6]. Such water stress estimates are particularly informative for the detection of incipient plant stress during early stages of drought development compared to estimates derived from other wave lengths (microwave or visible) and allow to launch early drought alerts [7].

Energy balance models use as inputs satellite data (normalized difference vegetation index, albedo and surface temperature) and in-situ meteorological observations (air temperature **AirT**, relative air humidity **Rh**, global radiation **GR** and wind speed **WS**) as provided by gauged networks. ET and water stress estimates computed from the instantaneous surface energy budget constrained by the surface temperature require meteorological observations acquired at the satellite overpass time. To ensure precise timing with satellite information, in-situ meteorological observations must be available at sub-daily resolution. In this work, we use satellite data provided by the latest MODIS collection (`http://earthexplorer.usgs.gov`) that has a 1 km spatial resolution. ET and water stress are estimated over a region covered by several MODIS grid cells. This region is defined so that it can be considered to be homogeneous in terms of climate and weather. As a consequence, a single multi-variable meteorological series representative of the region is needed. Nevertheless, there may be several gauged stations in the region with different observation periods and different gaps in the observation series. Moreover, it is often the case that the observed meteorological series are available over too short periods of time. Therefore, an important task is to develop a rigorous way to obtain a representative sub-daily multi-variable meteorological surrogate series in which gaps are filled and that extends in time the original series by exploiting the information provided by all the gauged stations in the region.

Stochastic Weather Generators (SWGs) are stochastic models based on statistical approaches for simulating, at high spatial resolution, surrogate meteorological series that are similar to observation series in terms of distributional properties, preserving both systematic and random variations [8]. SWGs are thus very useful models to perform coherent gap filling and to generate realistic surrogate series over periods for which no observations are available. In the aforementioned surface water stress application, the sub-daily series of the four meteorological variables (AirT, Rh, GR and WS) display both annual and diurnal cycles. Once these primary systematic variations are accounted for, there will likely remain some random variations. There are two main strategies in SWGs to model systematic and random variability which are applicable in the case of multiple meteorological variables at several gauged stations : the weather type approach that breaks down weather into classes (or types or states) of typical recurring meteorological situations (e.g., clear blue sky, cloudy, rainy, *etc...*) and the stochastic regression approach that captures the variability of weather smoothly by using suitable covariates.

In the weather type approach, the underlying assumption is that each time step belongs to one weather type and all the time steps belonging to the same weather type can be modeled with a relatively simple statistical approach. In other words, the temporal sequence is grouped into blocks, each block being associated to one weather type [8, 9]. One of the earliest weather type SWG proposed by [10] models precipitation as a two-state Markov chain that corresponds to two simple weather types : wet and dry types. [11] builds on the latter model and represents the intensity of precipitation, for the wet weather type, as an exponential probability distribution. The other three variables (maximum/minimum temperature and solar radiation) are modeled with a multivariate normal distribution whose means and standard deviations change according to the wet or dry types. In more recent approaches, more general weather types may be obtained automatically as an unsupervised classification problem. They thus are defined as the classes resulting from the clustering of time steps with each time step characterized by climatic or meteorological features [12]. Weather types may also be defined indirectly as the states of a latent variable using, for instance, hidden markov models [13]. The analog approach in SWGs can be seen as pushing the weather type strategy to the limit where each time step constitutes a weather type [14, 15]. Analog-based SWGs may be entirely non-parametric and they may succeed in reproducing complex patterns between variables and between sites. However, non-parametric analog-based SWGs are essentially resampling schemes that are unable to simulate values and patterns that differ from those present in the observations. This may be a serious drawback when observation periods are not long enough to contain all potential patterns. To account for annual cycles, weather types may be modeled separately for each season, with the difficulty that the definition of seasons might be somewhat arbitrary [12, 16]. Except in the case of rainfall [17], sub-daily weather typing for variables such as temperature, humidity and solar radiation requires a suitable model of the daily cycle. The interpretation of weather types at the sub-daily scale may be less intuitive than at the daily scale. It is not yet clear how to adapt the weather type strategy in order to account for the presence of diurnal cycles [8].

Stochastic regression or, equivalently conditional distribution modeling, is another widely used strategy that can account for both systematic and random variability in SWGs. For instance, [18] links the two parameters of the gamma distribution that models the intensity of rainfall and the probability of rainfall to information on the rainfall pattern on the preceding day, the time of the year, *etc* ...with a one-hidden-layer feed-forward neural network. More generally, instead of focusing on estimating the conditional mean as is the case in conventional regression, stochastic regression seeks to estimate the full conditional distribution from which simulations can be drawn thereby allowing to reproduce the observed variability. Covariates that carry temporal and spatial information can be introduced letting the parameters of the conditional distribution vary in time and space. As an alternative to neural networks, Generalized Linear Models (GLMs) have been used within SWGs for the past 20 years or so, see for instance [19, 20] and the references therein. In GLMs, the conditional distribution belongs to the exponential family that encompasses the gaussian distribution. The link between the parameters and the covariates is established with a potentially transformed (e.g., with a logarithm) linear regression [21]. Routines to implement GLMs are readily available in standard statistical software (e.g., `R` [22]). Spatial dependence may be accounted for by modeling the dependence structure of the residuals, e.g., with gaussian processes [20]. If informative enough covariates are used, GLM-based SWGs can reproduce very accurately both systematic and random variability [23]. In particular, despite that GLM-based SWGs generally operate at the daily resolution, sub-daily resolution modeling may be achieved by introducing covariates carrying sub-daily information. Most GLM-based SWGs simulate only one or two meteorological variables (very often, precipitation and temperature as in [20]). One notable exception is [23] who proposes a simple scheme to model jointly several meteorological variables based on the decomposition of the multivariate density into a product of conditional univariate densities.

In this work, we introduce `MetGen`, an SWG based on GLM, hence relying on stochastic regression, that extends the approach described in [23] to the sub-daily resolution. Its implementation is publicly and freely available as an `R` library (`https://CRAN.R-project.org/package=MetGen`). In `MetGen`, the scheme proposed in [23] to model jointly several meteorological variables is adapted to the four meteorological variables (AirT, Rh, GR and WS) required for surface water stress estimation for which inter-variable dependencies are rather strong. In contrast to [23] who has proposed a way to model spatial dependence, inter-site dependence is not explicitly modeled in `MetGen`. The proposed SWG works in a manner similar to the so-called *regional approach* developed in hydrology [24]. Indeed, several stations within the region of interest may be used to calibrate the SWG to augment the size of the data set. Instead of relying on the homogeneity assumption of the regional approach, spatial variability, when present, is modeled with covariates. In addition to the covariates proposed in [25], special covariates are considered to enable the reproduction of diurnal cycles, based on pairs of sines and cosines, similarly as for annual cycles. An important category of covariates used to carry sub-daily information albeit at a large-scale (horizontal resolution of 31 km) are the meteorological reanalyses provided by ERA5, available at hourly resolution [26].

Since `MetGen` makes use of reanalyses in its covariates and because there are no other, to our knowledge, publicly and freely available multi-variable sub-daily SWG, we resorted to two *statistical downscaling* methods as comparative approaches. Statistical downscaling aims to bridge the gap between low resolution and potentially biased simulations from global climate models and the high resolution series required for impact studies such as observation series from gauged networks [27, 28]. Although often used to obtain climate change scenarios over future periods, statistical downscaling methods may be applied to reanalysis products in order to generate surrogate series over past periods [27]. An active area of research in statistical downscaling concerns the so-called bias correction methods [29]. Bias correction aims at transforming the low resolution series of meteorological variables such as provided by reanalyses so as to match, in terms of distributional properties (e.g., in terms of means), the high resolution series such as measured at gauged stations. In order to assess whether explicitly accounting for inter-variable dependencies is essential, we include as comparative approaches a univariate (`CDF-t` developed in [30]) and a multivariate bias correction method (`MBCn` proposed by [31]). Both methods provide fast, non-parametric (i.e., without strong distributional assumptions) alternatives to `MetGen` and are implemented as publicly available `R` libraries.

The paper is organized as follows. Section 2 presents our study area, the Merguellil plain in central Tunisia together with the meteorological data provided by gauged stations and derived from ERA5 reanalyses. The multi-variable sub-daily GLM-based SWG `MetGen` is described in section 3 along with the two aforementioned bias correction methods and their adaptation to enable their application at the sub-daily resolution. Section 4 is dedicated to the comparison of the statistical methods at generating surrogate meteorological series both in terms of the ability to fill gaps in the observation series and in terms of the ability to extend in time the observation series. Section 5 reports an evaluation of the surrogate meteorological series in terms of surface water stress estimation. In section 6, a discussion is presented followed by conclusions and research perspectives in section 7.

2. Study area and meteorological data

2.1. Study area : the Merguellil plain

The study area is part of the downstream plain of the Merguellil catchment called the Merguellil plain, see Fig 1. Lying in a semi-arid region located in central Tunisia, the catchment is characterized by a relatively mountainous upstream area (1200 km^2) and by a downstream alluvial plain (676 km^2). The upstream area presents a hilly topography (altitude between 200 and 1200 m with a median elevation of 500 m) [32]. In the plain, the landscape is mainly flat, and the vegetation is typical of semi-arid regions : rainfed agriculture (olive tree and cereals) and summer vegetables (melons, peppers and tomatos). Downstream farms are composed mainly of small cultivated areas [33]. The upstream and downstream areas are separated by the El Haouareb dam (Fig 1), which was built in 1989 to protect villages from inundations and to store irrigation water for the plain [34]. The study area is influenced both by the Mediterranean climate (dry subhumid) and the pre-Saharan climate (arid) [1]. It is characterized by the inter-annual irregularity of precipitation, with an average of annual rainfall of about 300 mm per year, and by a high evaporative demand of about 1600 mm per year. Water supply is by far insufficient to meet water demand which is rising steadily. The rise is due to the increase in population and industrial development and, most importantly, to the intensification of agriculture, which is the main water consumer (around 80 %) [35].

2.2. Meteorological observations

Hourly observation series of the four meteorological variables (air temperature **AirT**, relative air humidity **Rh**, global radiation **GR** and wind speed **WS**) needed for surface water stress estimation are collected from the three gauged stations, Ben Salem, Chebika and Barrouta, located in the Merguellil plain (see Fig. 1). The observation period, approximate number of observations and approximate percentage of missing values are given in Table 1. In addition, Fig. 2 illustrates the observation period and the positions of the gaps in the series.

Table 1: *Observation series at the three stations in our study area (see Fig. 1). The number of observations and of missing values may vary slightly depending on the meteorological variable.*

Station	Obs. period	# Obs.	% Miss. values
Chebika	2011 - 2016	54 181	0.01
Ben Salem	2012 - 2016	43 818	0.2 - 1.2
Barrouta	2014 - 2016	18 106	4

The three gauged stations provide similar meteorological information owing to their geographical proximity (7.9 km between Ben Salem and Chebika, 11.7 km between Ben Salem and

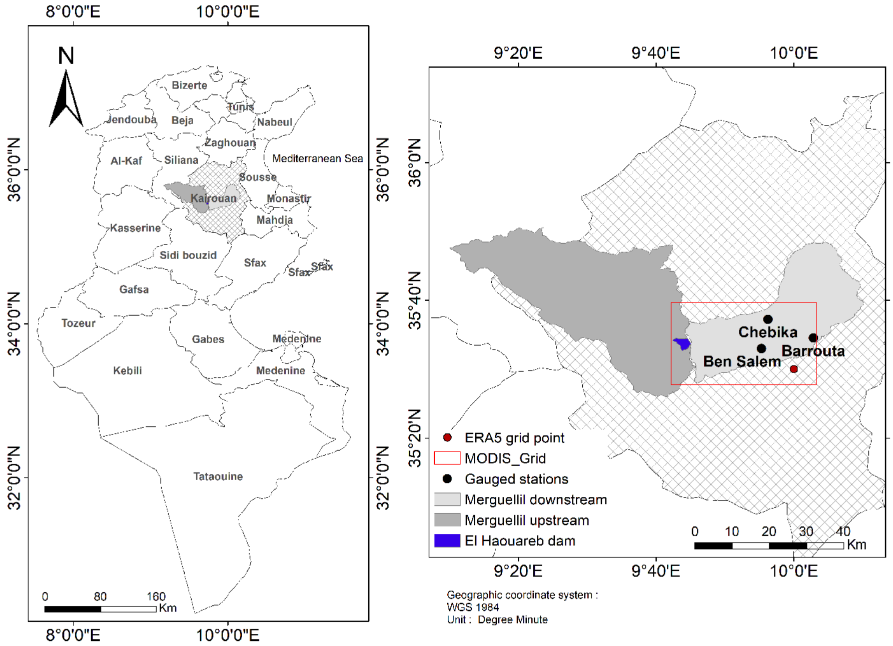


Figure 1: *Localisation of gauged stations : the plain located downstream of the Merguellil catchment in central Tunisia.*

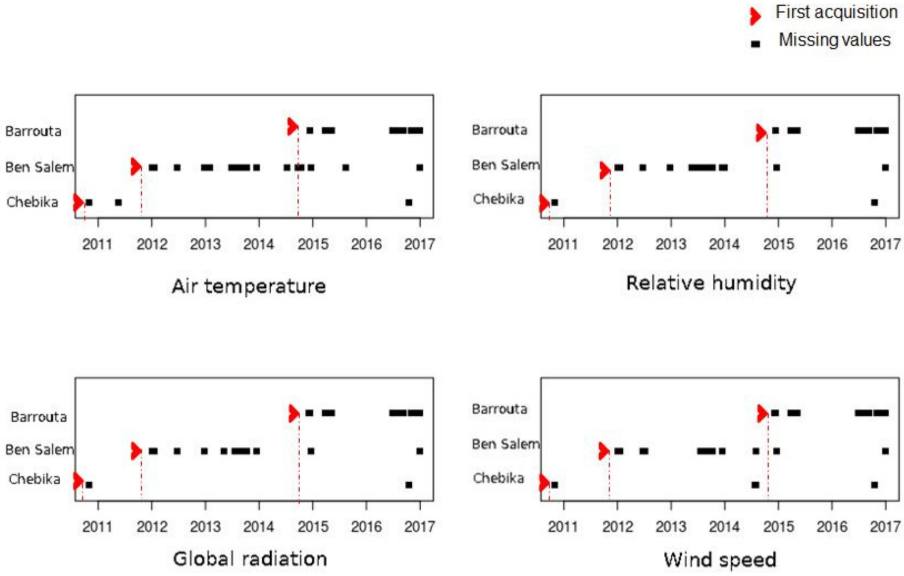
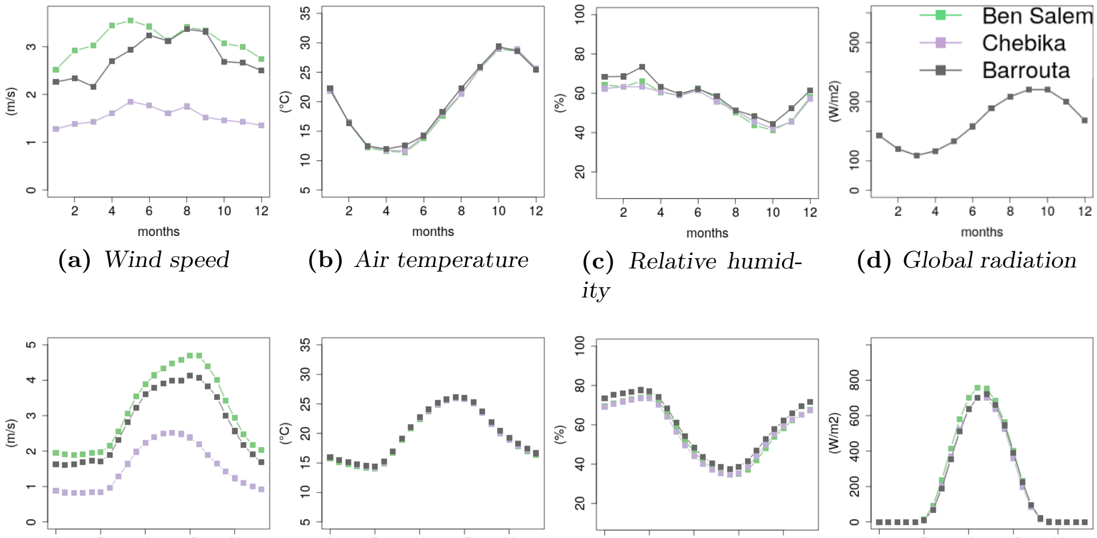


Figure 2: *Observation period and lengths of the gaps for each meteorological variables at the three gauged stations in the Merguellil plain.*

Barrouta and 11 km between Chebika and Barrouta). As shown by the annual and diurnal cycles in Fig. 3, these stations share similar climatic behaviors with the exception of the wind speed observed at Chebika. Lower wind speed values are caused by the presence of a windbreak in the vicinity of the station. In addition, inter-station pair plots (not shown) confirmed the strong relationship in the meteorological information provided by the three stations. The observation series at these three stations enter in the calibration of the stochastic generator `MetGen` proposed in this work. When building the statistical model, spatial covariates are selected to account for differences in the distribution of the meteorological variables at each of the station. In particular, a special covariate is used for the wind to account for the presence of the windbreak (see details in § 3.1). Ben Salem is selected as the *reference station* as it complies best with the meteorological standards according to the WMO guidelines [36]. Therefore, only the simulations of surrogate series corresponding to Ben Salem station are used in the evaluation and comparison of `MetGen`.



(a) *Wind speed* **(b)** *Air temperature* **(c)** *Relative humidity* **(d)** *Global radiation*

(e) *Wind speed* **(f)** *Air temperature* **(g)** *Relative humidity* **(h)** *Global radiation*

Figure 3: *Annual (top row) and diurnal (bottom row) cycles for each observed meteorological variable at the three gauged stations in the Merguellil plain.*

2.3. Meteorological reanalyses (ERA5)

Reanalyses combine forecast models and observations through data assimilation schemes thereby providing a multivariate, spatially complete and coherent record, without gaps, of atmospheric, land and oceanic climate variables [37, 26]. In particular, ERA5 reanalyses are available for a long period in the past, from 1950 till now [26]. Despite being available at hourly resolution, the ERA5 spatial resolution is low (horizontal resolution of 31 km [26]) and thus local-scale variability might not sufficiently be accounted for [38]. Besides the mismatch in spatial resolution, several limitations affected the quality of previous reanalyses such as ERA-Interim which were improved with respect to most aspects for ERA5 [26].

The three gauged stations from the Merguellil plain lie in the same ERA5 grid cell whose center is shown in Fig. 1. ERA5 reanalyses were extracted at this grid cell and were combined to obtain the six large-scale meteorological variables listed in Table 2. In most cases, the large-scale variables correspond to raw reanalysis products. There are two exceptions. The first one concerns the wind speed that was derived by taking the Euclidean norm of the 10 m vertical and horizontal wind components. The second exception concerns the relative humidity that was derived based on 2 m temperature and 2 m dewpoint temperature ERA5 products, according to the procedures defined in [39].

The six large-scale variables from Table 2 serve as covariates in the statistical methods described in section 3 to obtain surrogate meteorological series for the Merguellil plain. Among these, the first four are the large-scale counterpart of the meteorological variables needed for the surface water stress application. To evaluate the quality of the reanalysis products, these four large-scale variables are used without further processing as one of the candidate surrogate meteorological series. It is expected that the statistical methods should be able to correct departures in terms of distributional properties of the large-scale variables. Therefore, other large-scale data, whether reanalyses or remote sensed, could be used instead of ERA5.

Table 2: *Large-scale meteorological variables deduced from ERA5 reanalyses at the grid cell encompassing the three gauged station from the Merguellil plain. The second column indicates when the large-scale variable is considered as the large-scale counterpart of one of the meteorological variable needed in the surface water stress application.*

Large-scale meteo. var.	counterpart for
wind speed - 10 m (derived)	WS
air temperature - 2 m (raw)	AirT
relative humidity - 2 m (derived)	Rh
surface solar radiation downwards (raw)	GR
total cloud cover (raw)	
mean sea level pressure (raw)	

3. Statistical methods

3.1. `MetGen` : a regional multi-variable sub-daily GLM-based SWG

We focus on `MetGen` implementation for the surface water stress application in central Tunisia. The workflow sequence, summarized in Fig. 4, is the main contribution of this work and can be adapted in principle to any study area and to any other meteorological variables. `MetGen` is regional in the sense that the observations from several gauged stations can be used in the calibration to increase the sample size. As discussed in § 2.2, any spatial variability in terms of distribution is captured through dedicated covariates. Once calibrated, `MetGen` simulates series at all the gauged stations. However, for our surface water stress application, a single series, representative of the region, is needed. To this end, we make use of the series corresponding to Ben Salem as it is our reference station, see § 2.2.

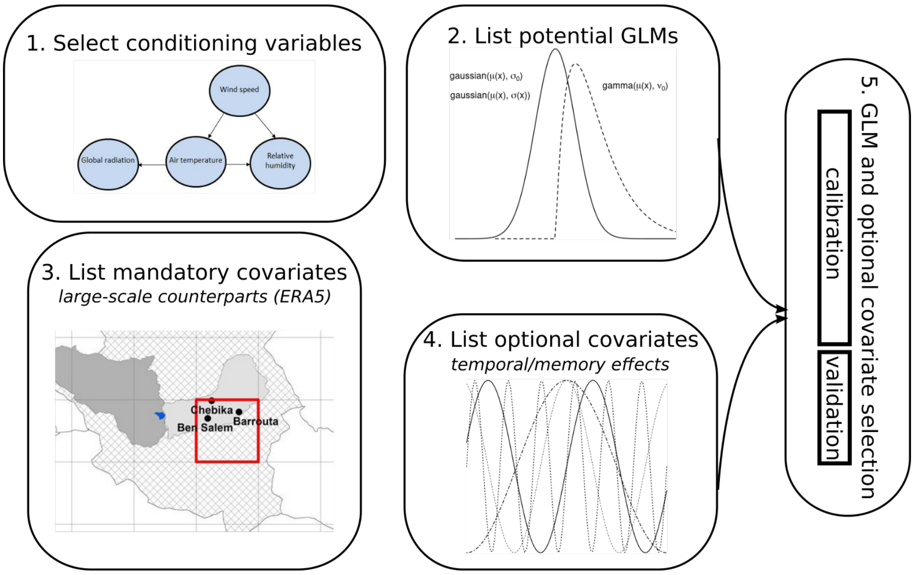


Figure 4: *`MetGen` implementation steps.*

3.1.1. Multi-variable modeling : conditioning variables

The first step to implement `MetGen` consists in modeling the inter-variable dependencies by means of conditioning variables (see Fig. 4). This follows the proposal of `RGlimclim` [25] by which a multivariate distribution can be decomposed with the product rule into conditional univariate distributions. To determine the order of the decomposition in the product rule and to reduce the number of conditioning variables, we rely on the dependence graph shown in Fig. 5. It has been adapted from the one made in the HydEF project (`https://www.imperial.ac.uk/media/imperial-college/research-centres-and-groups/environmental-and-water-resource-engineering/UCL15Feb2012.pdf`) to apply `RGlimclim` in the UK. More precisely, the multivariate distribution of the four meteorological variables needed for the surface water stress application boils down to modeling four conditional univariate distributions (one for each meteorological variable) and including in the covariates the appropriate conditioning variables :

$$\mathbb{P}(WS|\boldsymbol{x}) \tag{1}$$

$$\mathbb{P}(AirT|WS, \boldsymbol{x}) \tag{2}$$

$$\mathbb{P}(Rh|AirT, WS, \boldsymbol{x}) \tag{3}$$

$$\mathbb{P}(Gr|AirT, \boldsymbol{x}). \tag{4}$$

where $\boldsymbol{x}$ are additional covariates to be described in § 3.1.3. The choice of the conditional distribution model, the selection of covariates and the calibration can be performed separately for each meteorological variable. The simulation of the multi-variable surrogate series proceeds following the order dictated by the dependence graph in Fig. 5 : wind speed is simulated first, then air temperature is simulated including among the covariates the series simulated for wind speed, relative humidity is simulated afterwards with the previously simulated series for air temperature and wind speed included in the covariates and finally, global radiation is simulated conditionally on the series simulated for air temperature.

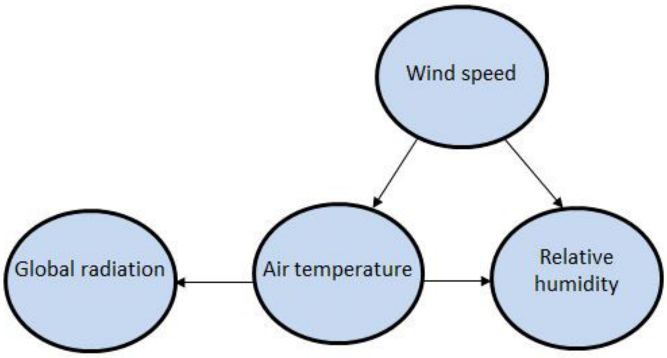


Figure 5: *Inter-variable dependency graph yielding the conditional univariate distributions in* (1)-(4) *which allows to model the dependencies among the four meteolorogical variables in* `MetGen`.

3.1.2. Conditional univariate distribution models : Generalized Linear Models (GLMs)

In the second step of `MetGen`, potential conditional univariate distribution models, which are from the Generalized Linear Model (GLM) family, for each meteorological variable must be defined (see Fig. 4). At present, three possible choices of probability distributions for the GLMs are available in `MetGen` : the gaussian distribution with constant (homoscedastic) or non-constant variance (heteroscedastic) and the gamma distribution. In the GLMs, the parameters of the probability distributions may vary according to covariates. In `MetGen`, we made the following choices to link the covariates to the probability distribution parameters. Let $\boldsymbol{x}_\mu$ and $\boldsymbol{x}_\sigma$ be two covariate vectors, let β_μ, and β_σ be regression coefficient vectors of the same length as $\boldsymbol{x}_\mu$ and $\boldsymbol{x}_\sigma$ respectively and let μ_0, $\sigma_0 > 0$ and $\nu_0 > 0$ be three constants. Then, the parameters of the conditional distributions are provided as follows for each of the three possible choices :

$$\text{homoscedastic gaussian} \quad : \quad \begin{cases} \mu(\boldsymbol{x}_\mu) = \boldsymbol{x}'_\mu \beta_\mu + \mu_0 & \text{Location param.} \\ \sigma_0 & \text{Scale param.} \end{cases} \tag{5}$$

$$\text{heteroscedastic gaussian} \quad : \quad \begin{cases} \mu(\boldsymbol{x}_\mu) = \boldsymbol{x}'_\mu \beta_\mu + \mu_0 & \text{Location param.} \\ \sigma(\boldsymbol{x}_\sigma) = \exp\left(\boldsymbol{x}'_\sigma \beta_\sigma + \sigma_0\right) & \text{Scale param.} \end{cases} \tag{6}$$

$$\text{gamma} \quad : \quad \begin{cases} \mu(\boldsymbol{x}_\mu) = \exp\left(\boldsymbol{x}'_\mu \beta_\mu + \mu_0\right) & \text{Location param.} \\ \nu_0 & \text{Shape param.} \end{cases} \tag{7}$$

Each of these models may be fitted by maximizing the log-likelihood (with the `glm` function in the base package of `R` for (5) and (7) and with the package `lmvar` for (6)).

Some preprocessing is performed on the raw observed series before model fitting. First, time steps for which global radiation is assumed to be zero (i.e., during the night) are determined based on the time of the sunrise and of the sunset at the coordinates of the station and for the given day of the year (see `R` package `insol`). Model fitting and simulation for global radiation is performed only on identified diurnal time steps. Second, preliminary transformations are defined for three meteorological variables (WS, Rh and GR) so as to remove range constraints and make them more likely to be suitably modeled by the gaussian distribution, see Table 3. For each meteorological variable, either two (the homo- or heteroscedastic gaussian, see (5)-(6)) or three choices of probability distributions (including also the gamma, see (7), for WS and GR that take only positive values) with a preliminary transformation when necessary are considered as potential models for each meteorological variable, see the complete list in Table 3.

Table 3: *Potential conditional distribution models, with a preliminary transformation if necessary, considered for each meteorological variable.* $\Phi^{\leftarrow}(\cdot)$ *indicates the quantile function of the standard Normal distribution and 1360.4 W/m^2 is the solar constant.*

Meteo. var.	Range constraint	Transformation	Prob. distr.	Model
WS	$WS > 0$	$\ln(\exp(WS) - 1)$	Homo. gaussian (5)	$\mathcal{M}_1^{WS}$
		$\ln(\exp(WS) - 1)$	Hetero. gaussian (6)	$\mathcal{M}_2^{WS}$
		✗	Gamma (7)	$\mathcal{M}_3^{WS}$
AirT	✗	✗	Homo. gaussian (5)	$\mathcal{M}_1^{AirT}$
			Hetero. gaussian (6)	$\mathcal{M}_2^{AirT}$
Rh	$0 < Rh < 1$	$\Phi^{\leftarrow}(Rh)$	Homo. gaussian (5)	$\mathcal{M}_1^{Rh}$
			Hetero. gaussian (6)	$\mathcal{M}_2^{Rh}$
GR	$0 < GR < 1360.4$	$\Phi^{\leftarrow}(GR/1360.4)$	Homo. gaussian (5)	$\mathcal{M}_1^{GR}$
		$\Phi^{\leftarrow}(GR/1360.4)$	Hetero. gaussian (6)	$\mathcal{M}_2^{GR}$
		✗	Gamma (7)	$\mathcal{M}_3^{GR}$

3.1.3. Mandatory and optional covariates

This corresponds to steps 3 and 4 in Fig. 4. Mandatory covariates, which are always included in the models defined in (5)-(7), are set as follows. For the location parameter $\mu(\boldsymbol{x}_\mu)$, either of the gaussian distributions in (5)-(6) or of the gamma distribution in (7), $\boldsymbol{x}_\mu$ includes the conditioning variables and the large-scale variables listed in the column 2 and 3 respectively of Table 4. To limit model complexity, the mandatory covariates included in $\boldsymbol{x}_\sigma$ for the scale parameter of the heteroscedastic gaussian distribution, $\sigma(\boldsymbol{x}_\sigma)$ in (6), only include the large-scale variables, listed in the 3rd column of Table 4.

Table 4: *Mandatory covariates used in* `MetGen` *for each meteorological variable needed in the surface water stress application : conditioning variables to introduce inter-variable dependencies based on the dependence graph from Fig. 5 and large-scale variables obtained from ERA5 reanalyses (see Table 2).*

Meteo. var.	conditioning var.	large-scale variables
WS	✗	10 m wind speed & mean sea level pressure
AirT	WS	2 m temperature
Rh	AirT, WS	relative humidity
GR	AirT	surface solar radiation downwards & total cloud cover

In addition to these mandatory covariates, other covariates may be optionally included in $\boldsymbol{x}_\mu$ to model systematic temporal variability (annual and diurnal cycles) and to account for temporal persistence (memory effects). Let $Y_{t,s}$ be the meteorological variable of interest (either wind speed, air temperature, relative humidity or global radiation), at time step t and at site $s \in \{1, \ldots, S\}$. The following optional covariates are considered :

- pairs of cosines and sines with annual oscillations :

$$\text{Annual cycle covariates : } \cos\left(\frac{2\pi d}{k_d}\right) \quad \sin\left(\frac{2\pi d}{k_d}\right) \tag{8}$$

  with $1 \leq d \leq 366$, the day of the year associated to time step t and $k_d \in (365, 183, 91, 30)$;

- pairs of cosines and sines with diurnal oscillations :

$$\text{Diurnal cycle covariates : } \cos\left(\frac{2\pi h}{k_h}\right) \quad \sin\left(\frac{2\pi h}{k_h}\right) \tag{9}$$

  with $1 \leq h \leq 24$, the hour of the day associated to time step t and $k_h \in (24, 12, 6)$.

- lagged values of the meteorological variable :

$$\text{Var.lagk : } Y_{t-k,s} \quad k \geq 1 \tag{10}$$

- lagged values of the spatial average of the meteorological variable :

$$\text{SA.lagk : } \frac{1}{S}\sum_{s=1}^{S} Y_{t-k,s} \quad k \geq 1 \tag{11}$$

- lagged values of b-moving averages, with $b > 1$:

$$\text{MAb.lagk : } \frac{1}{b}\sum_{j=1}^{b} Y_{t-j+1-k,s} \quad k \geq 1 \tag{12}$$

- lagged values of spatial b-moving averages, with $b > 1$:

$$\text{SMAb.lagk : } \frac{1}{bS}\sum_{j=1}^{b}\sum_{s=1}^{S} Y_{t-j+1-k,s} \quad k \geq 1. \tag{13}$$

Optional covariates that convey information on the spatial variability can also be considered. Besides the conventional x- and y-coordinates along with elevation, we include a special binary covariate to account for the presence of the windbreak at Chebika station (0 indicates no windbreak while 1 indicates the presence of a windbreak). This binary covariate allows the intercept term in $\boldsymbol{x}'_\mu \beta_\mu + \mu_0$ (see (5)-(7)) to take on a different value according to whether there is a windbreak or not.

3.1.4. Selection of the conditional distribution model and of optional covariates

This is the last step to implement `MetGen`, step 5 in Fig. 4. For each meteorological variable, a conditional distribution model, see Table 3, must be selected and additional optional covariates may be included in the covariate set. Statistical tests that are conventionally used to perform model and covariate selection, e.g., based on p-values or likelihood ratios, rely on the assumption that observations are conditionally independent which is likely not the case in our application. To circumvent this issue, model and covariate selection are performed with a calibration-validation scheme in which the data is split into a calibration period, used for model fitting, and a validation period, used to assess model performance. The performance criteria are detailed in Table 5.

The first stage consists of selecting the conditional distribution model for a given meteorological variable. All the potential models, see Table 3, with the covariate set restricted to the mandatory covariates, see Table 4, are fitted on the calibration period. The choice of conditional distribution model yielding the best performance computed on the validation period is retained for the subsequent stages. In the following stages, the conditional distribution model is thus fixed and optional covariates, among those in (8)-(13) along with spatial covariates, are added to the covariate set gradually. The model with the larger covariate set is fitted on the calibration period. The optional covariates are retained when the performance criteria computed on the validation period are improved.

Table 5: *Performance criteria for the selection of a conditional distribution model and of optional covariates. $y_{(i),s}$ and $\hat{y}_{(i),s}$ with $i = 1, \dots, n$ are resp. the sorted observations and sorted surrogates of a given meteorological variable at site s.*

$$\text{Main criterion:} \qquad \sum_{s} \sum_{i=1}^{n} (y_{(i),s} - \hat{y}_{(i),s})^2 \tag{14}$$

Visual criteria:	annual and diurnal cycles plots temporal auto-correlation plots inter-variable dependence plots

The strategy adopted for performance assessment is based on the minimization of (14), the mean squared error of quantile-quantile plots (qq-plots) that assesses how well the distribution is reproduced, as long as no lacks of fit are detected from the plots (see the visual criteria in Table 5). An example of addition of optional covariates that arose was when a lack of auto-correlation in the surrogate series was noticed that led to the inclusion of memory effects, see (10)-(13). Another example is the presence of the windbreak at Chebika station that generates differences in the diurnal and annual cycles, see Fig. 3a and 3e and led to the design and the inclusion of a special binary covariate indicating the presence of the windbreak as explained in § 3.1.3.

3.2. Sub-daily bias correction techniques

3.2.1. A univariate and a multivariate bias correction techniques

Initial bias correction techniques such as the quantile-matching method [40] are univariate, i.e., they seek to transform a single series (representing a single meteorological variable at a single location). The quantile-matching method relies on a transformation that combines the cumulative distribution functions (CDFs) of the high resolution and the low resolution series estimated over a calibration period. This transformation ensures that the CDF of the corrected series matches the high resolution series' CDF accurately over the calibration period. [30] built on the quantile-matching method to propose a transformation, called `CDF-t`, that incorporates additionally the CDF of the low resolution series over the study (or validation) period. More recent bias correction methods are multivariate, i.e., they correct jointly multiple series (either from several locations or for several meteorological variables or both) seeking to reproduce, in addition to univariate distributional properties, the dependence structures present in the series [31, 41]. One such recent multivariate bias correction method is the N-dimensional probability density function transform (`MBCn`) proposed by [31]. The `MBCn` method looks iteratively for linear combinations of the variables and performs bias correction with a univariate bias correction method such as quantile-matching or `CDF-t` on the linear combinations rather than on each variable separately. These two bias correction techniques, `CDF-t` and `MBCn` are described next.

As previously, let $Y_{t,s}$ be the meteorological variable of interest (either wind speed, air temperature, relative humidity or global radiation), at time step t and at given site s. Let $X_{t,m}$ be its large-scale counterpart provided by the ERA5 reanalyses, as listed in the first four rows of Table 2, at the same time step t and at the grid cell m that contains the site s. In other words, $Y_{t,s}$ is the high resolution meteorological variable from the gauged station and $X_{t,m}$ is its large-scale version obtained from the reanalyses. Let us assume that there is a period used for calibration for which both $Y_{t,s}$ and $X_{t,m}$ are available and a period used for validation for which only $X_{t,m}$ is available. `CDF-t` estimates $Y_{t,s}$, a single meteorological variable at a single site, over the validation period as :

$$\hat{y}_{t,s} = \widetilde{F}^{\leftarrow}_{X_m}(F_{X_m}(F^{\leftarrow}_{Y_s}(\widetilde{F}_{X_m}(x_{t,m})))) \tag{15}$$

where $x_{t,m}$ is the value of the large-scale variable $X_{t,m}$ that actually occurred on time t of the validation period, F_Z and $F^{\leftarrow}_Z$ denote respectively the empirical cumulative distribution function of the random variable Z and its inverse, the quantile function, and F_Z ($\widetilde{F}_Z$) indicates the empirical distribution function estimated over the calibration (validation) period (see [30] for more details).

In contrast to `CDF-t` which needs to be applied separately to each variable meteorological variable, `MBCn` works directly with all the meteorological variables for which bias correction needs to be performed. Let $\mathbf{Y}_{t,s}$ be the vector of four meteorological variables at site s and time step t. Similarly, let $\mathbf{X}_{t,m}$ ($\widetilde{\mathbf{X}}_{t,m}$) be the 4-dimensional vector of large-scale meteorological variables at the grid cell containing the gauged-station for the calibration (validation) period. `MBCn` relies on random orthogonal matrices $\mathbf{R}$. A univariate bias correction technique (in the implementation of `MBCn`, the quantile delta mapping is used, see [42] for detailed explanations) is applied separately by working on each element of the rotated vectors $\mathbf{X}_{t,m}\mathbf{R}$, $\widetilde{\mathbf{X}}_{t,m}\mathbf{R}$ and $\mathbf{Y}_{t,s}\mathbf{R}$. Then the bias corrected large-scale variable vectors are rotated back. These steps are summarized as follows, with $T(\cdot)$ denoting the univariate bias correction operator applied elementwise :

$$\mathbf{X}_{t,m} \leftarrow T\left(\mathbf{X}_{t,m}\mathbf{R}\right)\mathbf{R}^{-1} \tag{16}$$

$$\widetilde{\mathbf{X}}_{t,m} \leftarrow T\left(\widetilde{\mathbf{X}}_{t,m}\mathbf{R}\right)\mathbf{R}^{-1}. \tag{17}$$

This procedure is iterated with new random matrices $\mathbf{R}$ until the multivariate distribution of $\mathbf{X}_{t,m}$ matches the one of $\mathbf{Y}_{t,s}$. Then $\widetilde{\mathbf{X}}_{t,m}$ contains the bias corrected series in the validation period.

3.2.2. Working with anomalies at a single station

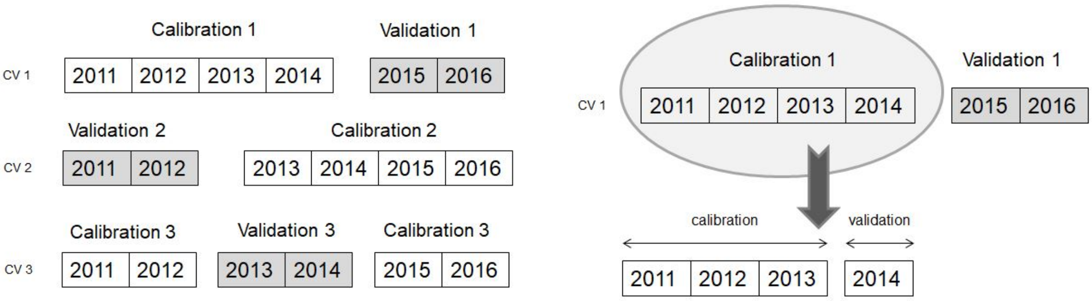
As our goal for the surface water stress application is to obtain a single surrogate series for each of the four meteorological variables that is representative of a homogeneous area, the two bias correction techniques are applied to the observation series at a single station, namely Ben Salem which is the reference station (see § 2.2). In other words, the large-scale variables obtained from ERA5 reanalyses are corrected, either with `CDFt` or with `MBCn`, to fill the gaps and extend in time the observation series at Ben Salem. `CDFt`, being a univariate approach, is applied separately for each meteorological variable while `MBCn` is applied to all four meteorological variables at once.

The following procedure is adopted to apply the two bias correction techniques at the sub-daily resolution. A conventional way to deal with the presence of annual cycles is to split the year into seasons and to apply bias correction separately on each season, see for instance [27]. However, the four meteorological variables used in the energy balance model (WS, AirT, Rh, GR) also display clear diurnal cycles, see Fig. 3. As this strategy (splitting the year into seasons) is not straightforward to extend to deal with diurnal cycles, we propose instead to work on anomalies of diurnal cycles with the diurnal cycle that is allowed to change with the season. More precisely, diurnal cycles are computed for three seasons : summer (June to August), winter (November to March) and inter-season (the remaining months). Observed (large-scale) anomalies are computed by subtracting the observed (large-scale) diurnal cycles from the observation (large-scale) series. Working with anomalies allows to remove systematic fluctuations from the meteorological variables and to focus on random fluctuations around the diurnal cycles. Bias corrected meteorological series are obtained by adding the observed diurnal cycles for each season to the bias corrected anomalies.

4. Evaluation and comparison

4.1. Cross-validation scheme

A cross-validation scheme is used to evaluate the statistical methods described in section 3 in terms of their ability to extend in time the original series, i.e., to simulate on periods for which no observations are available. Indeed, cross-validation is convenient for small data sets and is frequently used to evaluate out-of-sample performance [43, 44]. The cross-validation scheme is made of three temporal partitions of the observations at all the stations, CV1, CV2 and CV3, as presented in Fig. 6a. In each partition, the observation period (from 2011 to 2016) is split into a calibration period made of four years used for model fitting and a validation period consisting of two years for model evaluation and comparison. The statistical models thus simulate out-of-sample surrogate series over the 2 year validation set of each partition. To account for the stochastic aspect of MetGen, the series are replicated 50 times, i.e., for each time step, 50 values are drawn from the conditional models. The replications are limited to 50 to keep the computation times reasonable while still permitting to explore the uncertainty captured by the models. Out-of-sample surrogate series are generated over the complete observation period by putting together the validation periods of the three partitions.



(a) *Cross-validation scheme : three partitions (CV1, CV2 and CV3).*

(b) *Calibration-validation scheme : illustration for the first partition (CV1).*

Figure 6: *Performance evaluation and comparison. A second calibration/validation split within the calibration period of each partition of the cross-validation scheme is introduced to perform model and covariate selection for MetGen.*

For each partition of the cross-validation scheme, the conditional distribution model and optional covariates must be selected for MetGen, as described in § 3.1.4. To this end, the calibration set of the partition, which has four years, is split into a smaller calibration period of three years and a validation period of one year, see Fig. 6b. The selected conditional distribution model with the selected optional covariate set and the mandatory covariates is then calibrated anew over the whole calibration set (four years) of the partition. Note that different selections of model and of optional covariates may occur for each of the three partitions of the cross-validation scheme.

The selections of conditional distribution models and optional covariates for each of the three partitions (CV1, CV2 and CV3) are as indicated in Table 6. For all the partitions, the conditional distribution model selected for all four meteorological variables is the heteroscedastic gaussian distribution, plus a preliminary transformation when needed (models $\mathcal{M}_2^{WS}$, $\mathcal{M}_2^{AirT}$, $\mathcal{M}_2^{Rh}$ and $\mathcal{M}_2^{GR}$ from Table 3). In addition to the mandatory covariates from Table 4, the optional covariates included in the final set of covariates are listed in Table 6. For a given meteorological variable, different covariate sets may be selected for each partition of the cross-validation scheme as the selection was performed separately for each partition. Nevertheless, the covariate sets are very similar in most cases. For the wind speed, for example, the special binary covariate indicating the presence of a windbreak at one of the station was deemed necessary for all partitions. No other covariate conveying spatial information was retained. Besides, no optional covariates related to seasonal or diurnal cycle for the relative humidity and no optional covariates related to memory effects for the global radiation were included for any partition as the improvement in performance was not significant. Some interactions among the covariates were tested but none of them brought significant performance improvements in our application so that none were retained.

Table 6: *Model and covariate selection results for* `MetGen` *: for each meteorological variable conditional distribution models selected (see Table 3) and optional covariates (see* (8)-(13)*). For WS, the special binary covariate indicating the presence of a windbreak is included for each partition.*

Partition	Meteo. var.	Annual	Diurnal	Memory effects
	& model	oscill. (days)	oscill. (hours)	lagk = longest lag
CV1	WS : $\mathcal{M}_2^{WS}$	$30, 365$	$12, 24$	✗
	AirT : $\mathcal{M}_2^{AirT}$	183	$24, 12$	✗
	Rh : $\mathcal{M}_2^{Rh}$	✗	✗	✗
	GR : $\mathcal{M}_2^{GR}$	$365, 183$	$24, 12$	✗
CV2	WS : $\mathcal{M}_2^{WS}$	$30, 365$	$12, 24$	✗
	AirT : $\mathcal{M}_2^{AirT}$	$183, 365$	$6, 12, 24$	SA.lag3, MA.lag8, SMA.lag8, Var.lag3
	Rh : $\mathcal{M}_2^{Rh}$	✗	✗	SA.lag3, MA.lag3, SMA.lag7, Var.lag1
	GR : $\mathcal{M}_2^{GR}$	✗	$24, 12$	✗
CV3	WS : $\mathcal{M}_2^{WS}$	30	$12, 24$	✗
	AirT : $\mathcal{M}_2^{AirT}$	183	$24, 12$	SA.lag3, MA.lag3, SMA.lag1, Var.lag3
	Rh : $\mathcal{M}_2^{Rh}$	✗	✗	✗
	GR : $\mathcal{M}_2^{GR}$	✗	$24, 12$	✗

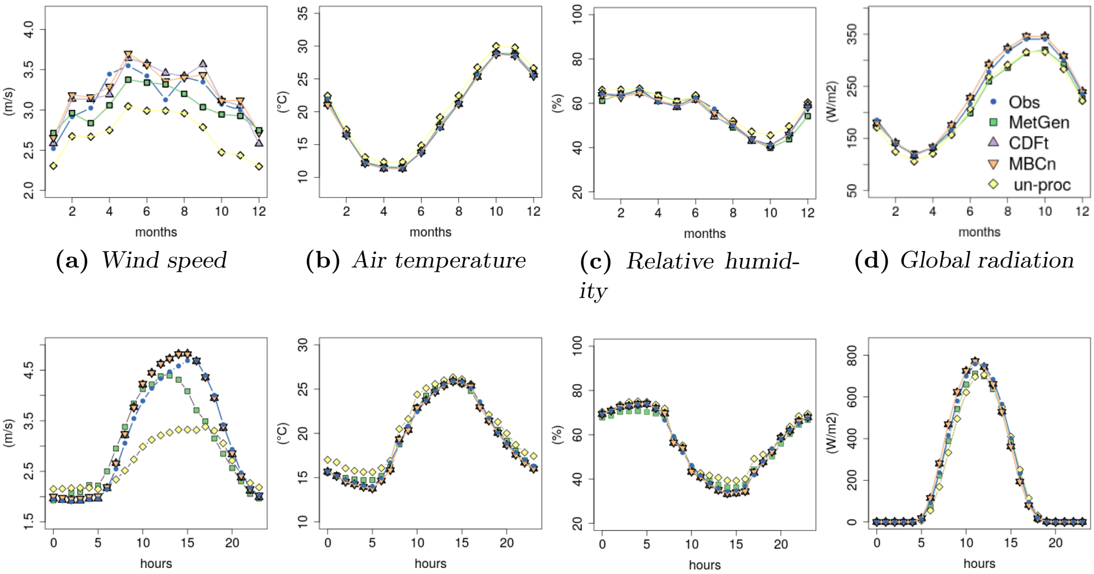
4.2. Cross-validation evaluation

In what follows, we report the evaluation and comparison of the three statistical models, the SWG `MetGen` and the two bias correction techniques `CDFt` and `MBCn` (see § 3), based on the surrogate meteorological series simulated with the cross-validation scheme described in § 4.1. Recall that, for `MetGen`, the surrogate series is replicated 50 times and that only the series related to the Ben Salem station is kept in this evaluation. In addition to the surrogate series from the statistical models, the un-processed large-scale variables obtained from the ERA5 reanalyses, see Table 2, are included in the comparison.

4.2.1. Annual and diurnal cycles

Fig. 7 presents the annual (top row) and diurnal (bottom row) cycles for each of the four meteorological variables. Annual cycles are computed by averaging values in each month over the observation period. Similarly, diurnal cycles are obtained as hourly averages.

The cycles of the un-processed large-scale variables (yellow diamonds in Fig. 7) accurately reproduce the observations' cycles (blue dots in Fig. 7) for most meteorological variables. This is the case for air temperature (Fig. 7f and Fig. 7b), relative humidity (Fig. 7g and Fig. 7c) and global radiation (Fig. 7h and Fig. 7d). However, the wind speed cycles are under-estimated in the un-processed large-scale variables series : the afternoon peak present in the diurnal cycle (Fig. 7e) is too low (about 3 m/s instead of about 4.5 m/s in the observation cycle) and the annual cycle values (Fig. 7a) are consistently below the observed ones. Despite relying on this information through its covariates, `MetGen` is able to correct fairly well the under-estimation



(a) *Wind speed* **(b)** *Air temperature* **(c)** *Relative humidity* **(d)** *Global radiation*

(e) *Wind speed* **(f)** *Air temperature* **(g)** *Relative humidity* **(h)** *Global radiation*

Figure 7: *Annual (top row) and diurnal (bottom row) cycles for each meteorological variable at the Ben Salem station. Comparison between observed (in blue) and surrogate series (see color legend). Note that the surrogate series from the two bias correction techniques (`CDFt` and `MBCn`) are the sum of the observed diurnal cycles plus the corrected anomalies hence the good adequation with the observed cycles (see § 3.2).*

shown in the un-processed large-scale series and to reproduce much more accurately annual and diurnal cycles of the wind speed variable. The series produced by the bias correction techniques are bias corrected anomalies to which observed diurnal cycles are added (see § 3.2). Since the root-mean-squared errors (not reported) of the bias corrected anomalies are low, it follows that the cycles of the corresponding series are well reproduced.

4.2.2. Goodness-of-fit of the whole distribution and of extreme values

Quantile-quantile plots (qq-plots) are used instead of scatter plots to assess whether the distribution of the observation series is well reproduced by the surrogate series. Indeed, all the statistical methods considered to generate surrogate series aim at providing a correction of the distributional properties of the large-scale variables rather than reproducing the chronology of the observed series. Similarly as in (14), let $y_{(i)}$ and $\hat{y}_{(i)}$, with $i = 1, \ldots, n$, be respectively the sorted observations and the sorted surrogates of a given meteorological variable for one of the approach (either one of the three statistical methods or the un-processed large-scale variables) corresponding to the Ben Salem station. Root Mean-Squared Errors (RMSEs) of the qq-plots relative to the standard deviation of the observations are defined as follows :

$$\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{(i)} - \hat{y}_{(i)})^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_{(i)} - \frac{1}{n}\sum_{i=1}^{n} y_{(i)}\right)^2}}. \tag{18}$$

The relative RMSE is near zero when the qq-plot is well aligned on the first bisector which means that the distribution of the observations is well reproduced by the surrogates. It is below (above) one when the RMSE is smaller (greater) than the standard deviation, i.e., when the surrogate series is better (worse) than the empirical average at reproducing the observations.

The relative RMSEs for each meteorological variable and for each type of surrogate series are reported in Table 7. For MetGen, the median of the relative RMSEs of the 50 replications is reported. The relative RMSE is computed either on all the quantiles (indicated as 0-100 %) or the 1 % highest quantiles to focus on how extreme values are reproduced. On the complete distribution (i.e., 0-100 %), all three statistical methods performed quite well (relative RMSEs are all rather close to zero). The multivariate bias correction technique, MBCn, performed best in general although often not by much. The un-processed large-scale variables always have the poorest performance (relative RMSEs are higher) especially for the wind speed. This indicates that all three statistical methods improved upon the distributional properties of the un-processed large-scale variables. Nevertheless, their performance is rather decent as the relative RMSEs are always lower than one. On the extreme values (i.e., the 1 % highest quantiles), the relative RMSEs give a very different picture. MetGen outperforms the two bias correction techniques for the wind speed and the air temperature. The un-processed large-scale variables relative RMSEs may be quite high, higher than one in three instances and sometime by a rather large factor. The extreme values of the relative humidity variable were the most difficult to reproduce in all surrogate series and is the only case in which all statistical methods do worst than the un-processed large-scale variables. This might be caused by the upper bound on the values taken by the relative humidity.

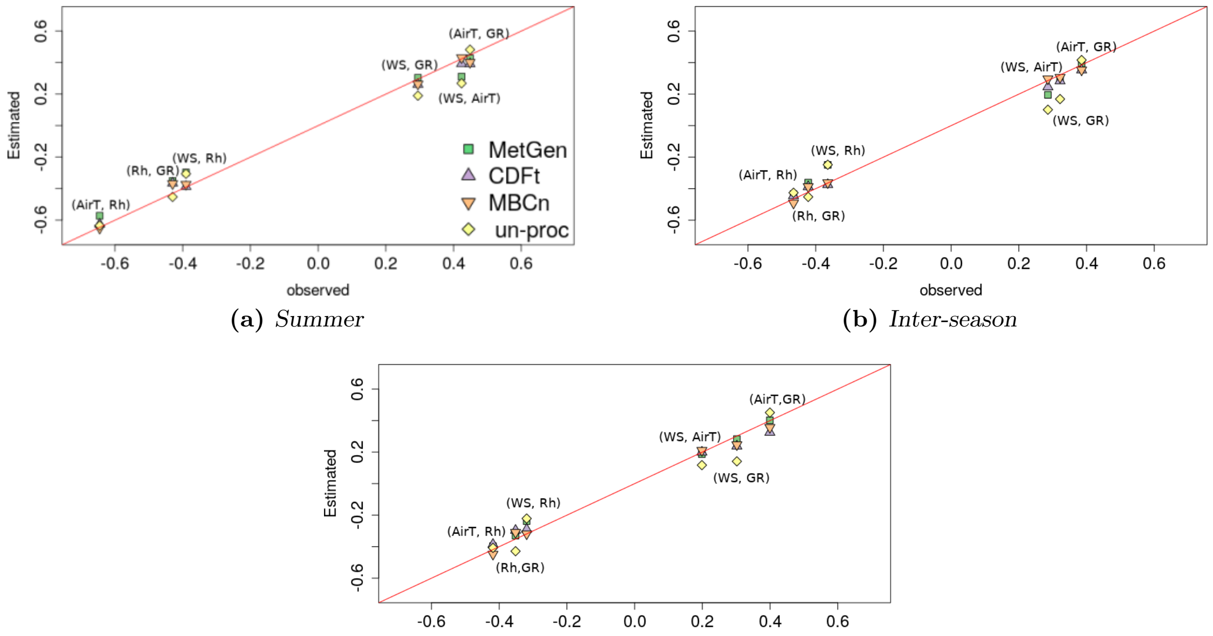
Table 7: *Goodness-of-fit of the whole distribution and of extreme values. The relative RMSEs, see* (18)*, are computed for all the quantiles (0-100%) and for the 1 % highest quantiles. For* MetGen*, the median of the relative RMSE of the 50 replications is provided. The best performance (lowest value) for each meteorological variable is indicated in italic font and values above one are indicated in* ***bold font****.*

Quantiles	Meteo. var.	MetGen	un-proc.	CDFt	MBCn
0-100%	WS	0.12	0.4	0.08	*0.07*
	AirT	0.03	0.12	*0.02*	*0.02*
	GR	0.08	0.11	*0.02*	*0.02*
	Rh	0.05	0.13	0.04	*0.03*
1% highest	WS	*0.71*	**2.3**	**1.34**	0.97
	AirT	*0.16*	0.3	0.41	0.28
	GR	0.94	**3.88**	*0.44*	0.8
	Rh	**2.23**	***1.14***	**1.23**	**1.28**

4.2.3. Inter-variable dependencies

Accurately reproducing inter-variable dependencies is particularly important since surface water stress is generally triggered by a combination of meteorological factors. In this evaluation, inter-variable dependence is summarized by Kendall's τ, a non-parametric correlation coefficient based on ranks, that is suitable for non-gaussian distributions (as opposed to the Pearson correlation coefficient) [45]. Positive values of Kendall's τ indicate that both variables tend to increase or decrease simultaneously while negative values indicate that they tend to vary in an opposite manner. A value near zero signals a lack of dependence. The comparison of correlation coefficients is carried out for each of the three seasons considered for the sub-daily bias correction techniques (see § 3.2) : summer (June to August), winter (November to March) and inter-season (the remaining months).

In Fig. 8, the correlation coefficients computed from the observations are on the x-axis while those derived from the surrogate series are on the y-axis. There is an overall relatively good alignment along the first bisector (the red line) showing that the inter-variable dependence strength is rather well preserved using the different surrogate series and for all three seasons. Nevertheless, Kendall's τ coefficients computed from the large-scale variables series tend to be less tightly aligned, especially when the wind speed (WS) variable is involved. For example, the correlation coefficient in winter (Fig. 8c) between the wind speed (WS) and the air temperature (AirT) in the observation series is about 0.2 whereas it is about 0.1 for the un-processed large-scale variable series. The correlation is always improved with the bias corrected series and, in most cases, with the `MetGen` series. A relatively strong negative dependence between the air temperature (AirT) and the relative humidity (Rh) is preserved in the different surrogate series and in the three seasons, especially in summer as it reaches -0.65 (Fig. 8a).



(a) *Summer*

(b) *Inter-season*

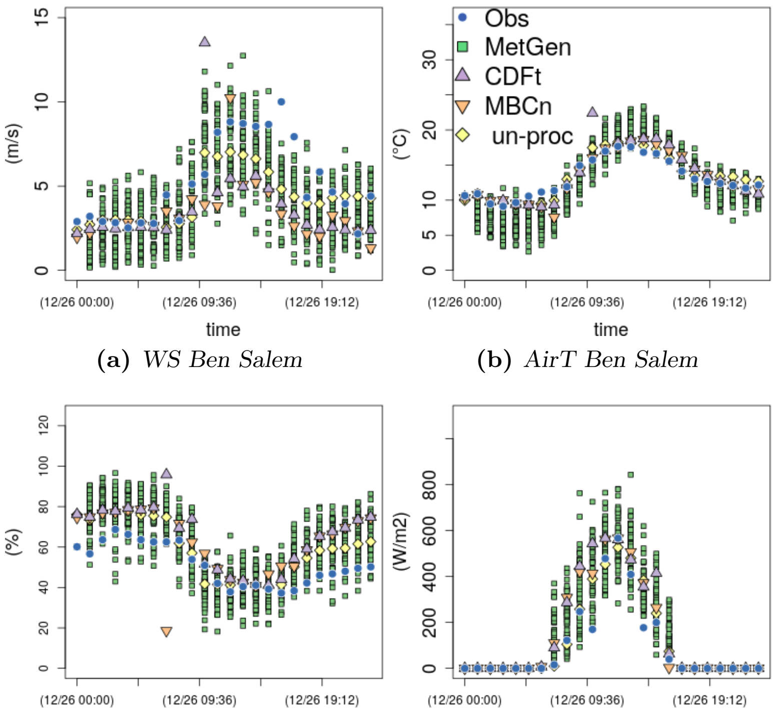
(c) *Winter*

Figure 8: *Inter-variable dependencies : comparison of Kendall's τ according to seasons (summer, winter, inter-season) for each pair of meteorological variables from the observation series on the x-axis and from the surrogate series (see color legend) on the y-axis.*

4.3. Gap-filling exercise

`MetGen` can run in gap-filling mode. As the simulation proceeds step by step, when observations are found missing, surrogate values are simulated by `MetGen` and the covariates introducing memory effects (see (10)-(13)) are updated based on the simulated values. Simulations of `MetGen` in gap-filling mode can be repeated to account for the uncertainty. The two bias correction techniques may also be used to perform gap-filling. However, in contrast to `MetGen`, the surrogate series are produced in the same way as for a validation period. A gap-filling exercise is carried out visually by inspecting the chronological plots of the surrogate series over one day, December 26th 2014, for which the observed values of all four meteorological variables from the Ben Salem station were removed artificially.

Fig. 9 presents the observed series (in blue) at the Ben Salem station over the day selected for the gap-filling exercise. The surrogate series of `MetGen` produced in gap-filling mode (50 replications), of the two bias correction techniques and of the un-processed large-scale variable are superimposed (see color legend). In Fig. 9, we observe that, the simulated values from the three statistical methods reproduce rather well the original values observed at the Ben Salem station. The 50 simulations from `MetGen` are generally centered around the large-scale series and, most importantly, their spread covers the observed series. We also note that `MetGen` is able to rectify values that are too low (e.g., the wind speed in Fig. 9a) or too high (e.g., the relative humidity in Fig. 9c) that are present in the un-processed large-scale variable series.



(a) *WS Ben Salem*

(b) *AirT Ben Salem*

(c) *Rh Ben Salem*

(d) *GR Ben Salem*

Figure 9: *Gap-filling exercise : chronological plots over one day (December 26th 2014) of the four meteorological variables observed at the Ben Salem station. Superimposed are the surrogate series (50 replications) of `MetGen` ran in gap-filling mode along with the surrogate series of the two bias correction techniques (`CDFt` and `MBCn`) ran in validation mode and the un-processed large-scale variable series (see color legend).*

5. Surface water stress application

5.1. SPARSE : a dual-source energy balance model

Surface water stress may be deduced from evapotranspiration (ET) using energy balance models. At satellite overpass time, energy balance models compute instantaneous latent heat flux (LE), expressed in W/m^2, as the residual term of the land surface energy balance equation [4, 5, 6]. In this application of surface water stress estimation, we use the dual-source model Soil Plant Atmosphere and Remote Evapotranspiration (SPARSE) [46] which is based on the same rationale as TSEB (Two-Source Energy Balance model) [5]. SPARSE derives from the remotely sensed surface temperature ($Tsurf$) separate estimates of the instantaneous fluxes of the soil (subscript s) and vegetation (subscript v) components of the energy budget at the satellite overpass time. SPARSE can be run under the two following modes :

- A prescribed mode which simulates evaporation and transpiration rates for known stress levels (for instance, the two extremes of the water status spectrum : fully stressed or maximum moisture, i.e., potential conditions). The prescribed mode provides an estimate of the potential latent flux for the soil and the vegetation ($LEspot$ and $LEvpot$ respectively).

- A retrieval mode which simulates actual evaporation and transpiration : the respective stress levels (between non evaporating/transpiring and potential rates) correspond to two unknown which are solved from the single piece of information ($Tsurf$) [46].

The surface water stress index (SI) is derived from the actual and potential evapotranspiration rates simulated from the retrieval and the prescribed mode respectively at the time of the satellite overpass. SI can be defined so as to describe the water status of a single component : either of the soil or of the vegetation (using LEs or LEv). In this application, we used rather the definition of SI to describe the water status of the soil-vegetation composite :

$$SI = 1 - \frac{LEv + LEs}{LEvpot + LEspot}. \tag{19}$$

The stress index values obtained directly with (19) may contain negative values due to enhanced turbulence in unstable conditions. As negative values cannot theoretically occur, stress index values below -0.5 are replaced by zeros. Besides SI, daily evapotranspiration (ETd) is computed from an extrapolation algorithm in order to reconstruct its sub-daily variations by assuming the self preservation of the evaporative fraction [47]. SPARSE is only ran when remote sensing data are available (i.e., on clear-sky days). The implementation of SPARSE in the Matlab environment is freely available online (`http://tully.ups-tlse.fr/gilles.boulet/sparse`).

5.2. Remote sensing data (MODIS)

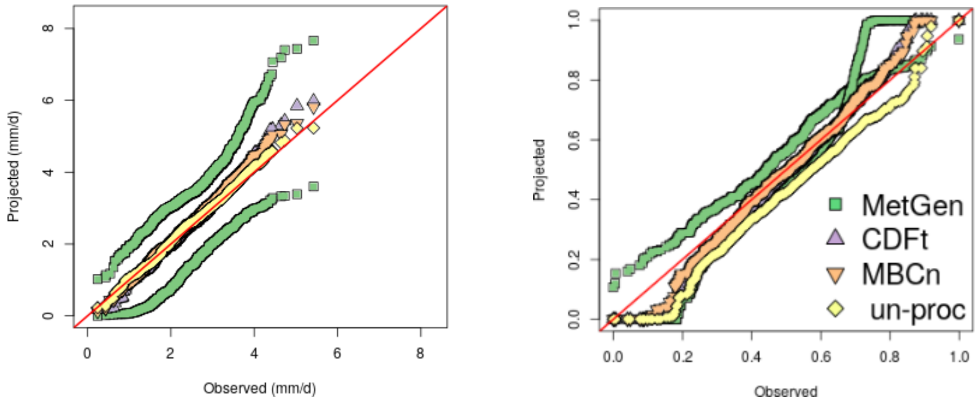
In addition to meteorological information (air temperature, relative air humidity, global radiation and wind speed), SPARSE uses as inputs satellite data (Normalized Difference Vegetation Index (NDVI), albedo and surface temperature) that provide a description of the initial conditions and of the characteristics of the surface cover. To this end, we relied on remotely sensed data from the latest collection 6 of MODIS (`http://earthexplorer.usgs.gov`) that are available from the year 2000. More precisely, we used the temporal 16-day composite series of MODIS NDVI (MOD13A2), daily Land Surface Temperature (LST), surface emissivity and viewing angle from (MOD11A1) and 8-day of albedo series (MCD43A3) having a spatial resolution of 500 m. These data are acquired for the observation period available at the Ben Salem station (2012-2016) at the resolution of the MODIS sensor (1 km). We extracted a sub-image covering the whole Merguellil plain, see Fig. 1. In addition, we performed a temporal interpolation of albedo and NDVI data to have daily information at the time of the satellite overpass. Last, NDVI information is used to compute remotely sensed leaf area index.

5.3. Evaluation of the surrogate series in terms of SI and ETd estimates

We compare various estimates of instantaneous surface water stress index *SI* and of daily evapotranspiration *ETd* by constraining SPARSE, on one hand, with the aforementioned MODIS data and, on the other hand, with different choices of meteorological information. Our ground truth is the estimates of *SI* and *ETd* obtained when the observed meteorological series at Ben Salem station are used. Other estimates of *SI* and *ETd* are obtained when these observed meteorological series are replaced by the surrogate meteorological series produced with the cross-validation scheme (see § 4.1) for the three statistical methods (`MetGen`, `CDFt` and `MBCn`) and by the un-processed large-scale variable series. Out of the 50 replications of `MetGen`, we designed two surrogate series corresponding to meteorological conditions leading to lower or higher surface water stress. The low stress conditions consist of high humidity levels set to the 97.5 % quantile of Rh and low levels of WS, AirT and GR set to 2.5 % quantiles. The quantile levels are reversed to obtain the high stress conditions.

In Fig. 10, the comparison between the various estimates of *SI* and *ETd* is first carried out in terms of distribution with qq-plots. On the x-axis, *ETd* (Fig. 10a) and *SI* (Fig. 10b) are the ground truth estimates (i.e., when the observed meteorological series at Ben Salem station are used to constrain SPARSE). On the y-axis of both panels of Fig. 10, the estimates are obtained by replacing the observed meteorological series with one of the surrogate meteorological series. The distribution of *ETd*, see Fig. 10a, is overall well reproduced by all the estimates computed with the surrogate meteorological series. The low and high stress condition surrogate series from `MetGen` form a sort of confidence band around the first bisector.

More pronounced differences are observed in the comparison of the qq-plots of *SI* in Fig. 10b. As explained in § 5.1, a truncation of *SI* estimates is performed to reduce the importance of negative values which are theoretically not realistic. This creates an atom, i.e., a concentration of values, at zero in the distribution of *SI* estimates which explains the shape of the qq-plot close to zero. The atom is especially important when the un-processed large-scale variable series are used as meteorological information to obtain the *SI* estimates (the atom makes a horizontal line of zero values starting at 0 in the yellow curve of the qq-plot in Fig. 10b). The atom is also present for the high stress condition series from `MetGen` and the two bias corrected series (the plotting symbols are covered partially). The *SI* estimates from the un-processed large-scale variables are globally too low (the yellow curve of the qq-plot is well under the first bisector). The *SI* estimates deduced with the surrogate series from the two bias correction techniques (`CDFt` and `MBCn`) display a much milder under-estimation that concerns mostly the lower values, indicative of high stress levels. The *SI* estimates computed with the low and high stress condition series from `MetGen` can be thought of as forming a sort of confidence band for the lower index values. However, their behavior for the higher *SI* values is much harder to interpret and would require further investigation.



(a) *Daily evapotranspiration (ETd)*

(b) *Instantaneous surface water stress index (SI)*

Figure 10: *Qq-plots of the outputs of the SPARSE energy balance model, see § 5.1, when constrained by the meteorological information from the observed series on the x-axis and from the surrogate series obtained with the cross-validation scheme (see § 4.1) on the y-axis (see color legend). For* `MetGen`*, a low and high stress condition series are computed out of the 50 replicated surrogate series.*

In order to translate differences in distribution as visualized by discrepancies from the first bisector in the qq-plot from Fig. 10b into a more hydrologically interpretable analysis, we propose a comparison based on the probability that the *SI* estimate exceeds a given threshold, so-called *exceedance probability*. In Fig. 11, threshold values are represented on the x-axis while the exceedance probabilities are on the y-axis, both ranging from 0 to 1. Black dots represent the exceedance probabilities computed from the ground truth *SI* estimates along with an empirical 95% confidence band in gray based on binomial proportion confidence intervals :

$$\sqrt{\frac{p(1-p)}{n}} \times 1.96, \tag{20}$$

where p is the probability to have a *SI* estimate value that exceeds the threshold and n is the number of the available time steps. The exceedance probabilities computed from the other *SI* estimates (when relying on the surrogate series for the meteorological information constraining the SPARSE model) are as indicated in the color legend in Fig. 11. With the *SI* estimates obtained when using the un-processed large-scale variable series, the exceedance probability (yellow diamonds in Fig. 11) falls below the grey confidence band for almost all threshold values. This is coherent with the under-estimation detected from the qq-plot in Fig. 10b. With the *SI* estimates obtained when using the surrogate series from the two bias correction techniques, the exceedance probability also tend to fall slightly below the confidence band but only for the lower thresholds. The *SI* estimates based on the low and high stress condition series from `MetGen` form a band that overlaps the empirical confidence band for most threshold values except the larger ones.

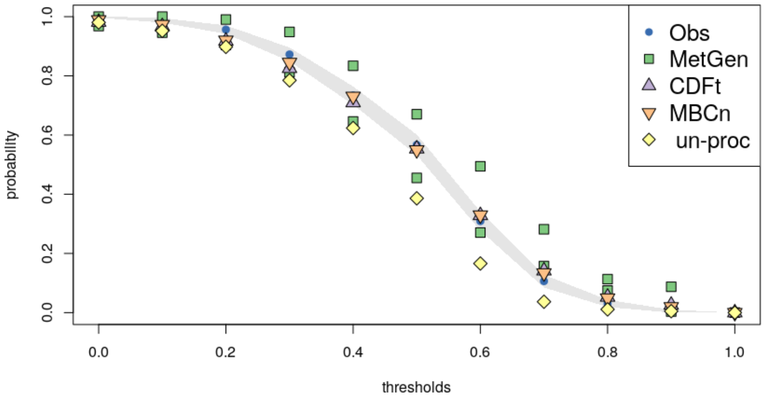


Figure 11: *Exceedance probabilities for increasing threshold values computed from the SI estimates as computed by the SPARSE energy balance model when constrained with the observed meteorological series (in blue) along with a 95% empirical confidence band (in gray) and when constrained with the surrogate series (see the color legend). For* `MetGen`*, a low and high stress condition series are computed out of the 50 replicated surrogate series.*

A final analysis is carried out to illustrate the fluctuations of the *SI* estimates chronologically over a short period. Fig. 12 presents an extract of the *SI* estimates during one month, May 2016. The estimates are derived at the satellite overpass times during this month with the different choices of meteorological information (either the observed series or one of the surrogate series) used to constrain the SPARSE energy balance model. For `MetGen`, the high stress condition series was used as it follows more closely the ground truth estimates. In Fig. 12, we observe that the *SI* estimates derived with the surrogate series generated by the bias correction techniques or the un-processed large-scale variables tend to over- or under-estimate the ground truth estimates.

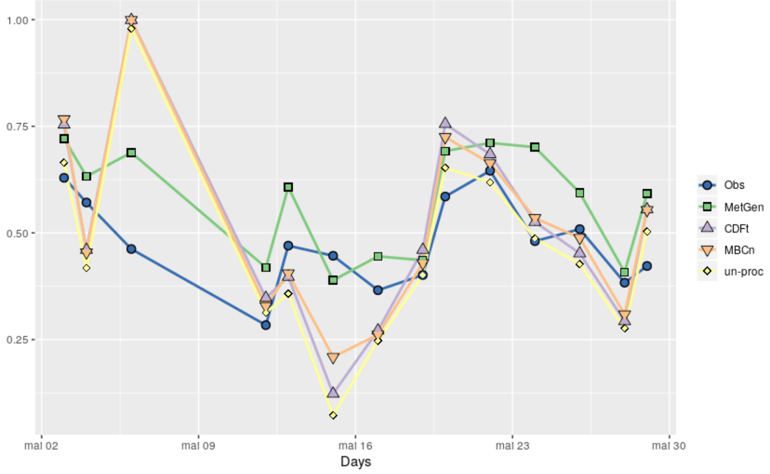


Figure 12: *Chronological fluctuations of SI estimates at the satellite overpass times during May 2016 with the different choices of meteorological information (either the observed series or one of the surrogate series) used to constrain the SPARSE energy balance model (see color legend). For* `MetGen`*, a high stress condition series is computed out of the 50 replicated surrogate series.*

6. Discussion

We proposed an adaptation of the GLM-based SWG developed in [25], called MetGen, to the sub-daily resolution. Indeed, sub-daily resolution is necessary when meteorological observations are used as inputs in energy balance models to estimate surface water stress in order to ensure a precise timing with satellite overpass time. By decomposing the joint distribution into a product of conditional univariate distributions, MetGen can model any number of meteorological variables simultaneously. GLMs are a flexible family to model the conditional distributions of meteorological variables from which surrogate series can be simulated stochastically thereby allowing to take into account uncertainty. Although the inter-station dependence is not modeled, MetGen can exploit the observations of several gauged stations as long as spatial variability can be modeled through covariates. This allows to increase the sample size similarly as in the regional approach frequently used in hydrology. In addition, for the surface water stress application, surrogate series at a single station are needed hence inter-site dependence is not an issue. The framework of MetGen, available freely as an R library (https://CRAN.R-project.org/package=MetGen), can be useful in many cases since consistent gap-filling and the extension in time of a multi-variable sub-daily observation series is a frequent issue [48, 49].

Since observations at successive time-steps and at neighboring gauged stations during the same time-steps are likely to be dependent, standard model selection procedure that relies, for instance, on the bayesian information criterion [50] and covariate selection that relies on p-values of the coefficients cannot be used for MetGen. For this reason, we put forward a model selection procedure that relies entirely on out-of-sample data through a calibration-validation scheme. Model selection proceeds iteratively, by first selecting the type of conditional distribution model using the mandatory covariates only. At the successive stages, optional covariates may be introduced to improve goodness-of-fit criteria computed on the validation set. In addition to the mean-squared error from the qq-plot, we included several visual criteria that help to assess additional important features that may not be reflected in the mean-squared error. One notable example of such primary features are the annual and diurnal cycles that lead to the design and the inclusion of the special covariate used to account for the presence of a windbreak at one of the stations. This covariate proved essential in order to reproduce correctly the different shapes of the annual and diurnal cycles of the wind speed at each station. This shows that the MetGen framework can handle a certain level of spatial variability among the stations used for calibration.

Our analyses showed that relying directly on the un-processed large-variables obtained from ERA5 reanalyses [26] as surrogate series may lead to some considerable biases. This is especially true for the wind speed variable as can be seen in the comparison of the annual and diurnal cycles and in the relative RMSE of the qq-plots. Bias correction techniques are devised to correct this kind of systematic departures in terms of distribution of the large-scale variables. Although based on a completely different kind of statistical approaches than SWGs, bias correction techniques can easily be adapted to the task of gap-filling and temporal extension for which MetGen is designed. These techniques are more straightforward to apply since no model selection is needed.

Nevertheless, in contrast to MetGen, they are not stochastic (a single replication of the surrogate series is produced). In the gap-filling exercise, from the replicated surrogate series simulated by MetGen, confidence bands could be obtained by computing a low and a high quantiles (such as 2.5% and 97.5% for a 95% confidence level) for each time step. In addition, despite the development of multivariate approaches, the bias correction techniques are not devised explicitly to exploit meteorological data from neighboring gauged stations lying in the same ERA5 grid cell. In MetGen, we made the choice to include, among the mandatory covariates in the GLMs, the large-scale variables in order to allow the simulated series to be guided by the non-stationary behavior present in these covariates, e.g., to follow trends and cycles. Owing to the inclusion of these large-scale covariates, MetGen may also be thought of as performing, in some sense, a form of bias correction [51, 52].

The mechanisms that allow the SWG MetGen and the bias correction techniques to reproduce seasonal and diurnal behavior are very different. The surrogate series produced by the two bias correction techniques are constructed by adding the observations' diurnal cycles, computed separately for three seasons, to the bias corrected anomalies. The fact that the cycles, annual and diurnal, computed from these surrogate series are very similar to the observations' cycles for all meteorological variables essentially means that the bias corrected anomalies are almost zero-mean. The construction of these surrogate series also explains why inter-variable correlations can suitably vary from one season to another. This explains also likely why there are no large differences between the univariate and the multivariate bias correction techniques as there is probably not much residual inter-variable dependencies left in the anomalies (in most cases, the plot symbols of CDFt are hidden by those of MBCn) . In contrast, the surrogate series simulated by MetGen are able to mimic seasonal and diurnal behavior by exploiting the information in the covariates included in the GLMs. Even when no covariates dedicated to the cycles are included, such as for the relative humidity variable, information on seasonal and diurnal behavior can be drawn from other covariates such as the large-scale variables. In particular, the reproduction of the annual and diurnal cycles and of the seasonal variation of the inter-variable correlations of the surrogate series simulated by MetGen is only due to the information present in the covariates. This mechanism is, in our opinion, more flexible as, for instance, it allows the starting and ending dates of each season to vary seamlessly from year to year instead of resorting to a specialized model [53].

In our analyses, one of the main differences among the surrogate meteorological series concerns the wind speed variable. This can be explained by the fact that the wind is the most turbulent of the meteorological variables being modeled. The turbulence yields more frequent and important random fluctuations that are challenging for the statistical models. As MetGen relies on stochastic regression, it can better cope with the wind speed fluctuations than the bias correction techniques. These differences in the wind speed surrogate series might be related to the differences in the resulting surface water stress index estimates. Indeed, the SPARSE energy balance model is particularly sensitive to wind speed, especially high wind speed values that can lead to low *SI* estimates. These lower *SI* values are particularly important as they are indicative of incipient water stress. When the low and high stress condition series derived from the replicated series of MetGen are used as inputs to the SPARSE model, a form of confidence band is obtained that seems particularly reliable for the lower *SI* values. More work on the influence of each meteorological variable on the *SI* estimates is required to improve these low and high stress condition series in order to yield proper confidence bands. Besides, the daily evapotranspiration estimates, *ETd*, are not very sensitive to which meteorological variables are used as inputs in the SPARSE model. This lack of sensitivity may be explained by the fact that an extrapolation scheme, that might act as a form of filter of the differences, is applied to obtain daily values from instantaneous ones. However, the sensitivity of the SPARSE model should be explored further.

7. Conclusion

The framework of MetGen provides a solution to the task of obtaining a representative sub-daily multi-variable meteorological series which mimic the observation series but in which gaps are filled and whose simulation period can extend the observation period. It relies on GLMs to model the conditional distributions with large-scale variables used as covariates. These large-scale variables, derived from ERA5 reanalyses in our application, are useful to provide a temporal dynamic of the weather. They may present, in some cases, important biases that can be reduced thanks to the GLMs and the introduction of other covariates. MetGen was compared to two bias correction techniques applied on the anomalies of seasonal diurnal cycles. In contrast to the bias correction techniques, MetGen can exploit the observations series from nearby gauged-stations having a relatively similar climatic behavior and can generate several replications of the same series to allow uncertainty assessment. Nevertheless, MetGen requires a careful selection of the conditional distribution models and of the covariate sets and the simulation can be slow while the bias correction techniques are more straightforward to implement and fast to run.

The analyses performed in this work provide a two-pronged evaluation of the surrogate series generated by MetGen and the two bias correction techniques. Firstly, the evaluation is carried out in terms of the ability to reproduce several statistical properties of the meteorological variables and secondly, in terms of surface water stress estimation when the series serve as inputs in the SPARSE energy balance model. Although the performance of the two bias correction techniques was similar for some criteria to MetGen, the evaluation in terms of the surface water stress application tends to indicate that the surrogate series generated by MetGen may be used to deduce reliable confidence intervals, especially for the lower *SI* values that are important for early drought detection.

Perspectives for this work include the study of the sensitivity of MetGen to the choice of the dependence graph that serves to simplify the decomposition of the multivariate distribution into a product of conditional univariate distributions. In addition, the introduction of a mechanism to explicitly model spatial dependence would certainly be a useful development (see [20]). Also, it would be interesting to evaluate MetGen in diverse climatic conditions (e.g., a coastal area) and for other meteorological variables (e.g., precipitation). The two main challenges for MetGen implementation are to find representative enough large-scale variables and to perform a thorough selection of the GLMs and of the optional covariates. Last, more thorough analyses pertaining to agricultural drought monitoring should be performed.

Acknowledgements:
The authors gratefully acknowledge financial support by the LMI NAILA, the PHC Utique Amande (a French-Tunisian program), ERANETMED CHAAMS and the European PRIMA project ALTOS. We would also wish to thank all the technical teams of the IRD and INAT for their strong collaboration and support.

References

[1] H. Baccour, H. Feki, M. Slimani, C. Cudennec, Interpolation de l'évapotranspiration de référence en Tunisie par la méthode de krigeage ordinaire, Science et changements planétaires/Sécheresse 23 (2) (2012) 121–132.

[2] S. Saadi, G. Boulet, M. Bahir, A. Brut, E. Delogu, P. Fanise, B. Mougenot, V. Simonneaux, Z. Chabaane, Assessment of actual evapotranspiration over a semi arid heterogeneous land surface by means of coupled low-resolution remote sensing data with an energy balance model: comparison to extra-large aperture scintillometer measurements, Hydrology and Earth System Sciences 22 (4) (2018) 2187–2209.

[3] J. Sheffield, E. F. Wood, Drought: past problems and future scenarios, Routledge, 2012.

[4] J. Hoedjes, A. Chehbouni, F. Jacob, J. Ezzahar, G. Boulet, Deriving daily evapotranspiration from remotely sensed instantaneous evaporative fraction over olive orchard in semi-arid Morocco, Journal of Hydrology 354 (1-4) (2008) 53–64.

[5] J. M. Norman, W. P. Kustas, K. S. Humes, Source approach for estimating soil and vegetation energy fluxes in observations of directional radiometric surface temperature, Agricultural and Forest Meteorology 77 (3-4) (1995) 263–293.

[6] W. J. Timmermans, W. P. Kustas, M. C. Anderson, A. N. French, An intercomparison of the surface energy balance algorithm for land (sebal) and the two-source energy balance (tseb) modeling schemes, Remote Sensing of Environment 108 (4) (2007) 369–384.

[7] J. A. Otkin, M. C. Anderson, C. Hain, I. E. Mladenova, J. B. Basara, M. Svoboda, Examining rapid onset drought development using the thermal infrared–based evaporative stress index, Journal of Hydrometeorology 14 (4) (2013) 1057–1074.

[8] P. Ailliot, D. Allard, V. Monbet, P. Naveau, Stochastic weather generators: an overview of weather type models, J. de la Société Francaise de Statistique 156 (1) (2015) 101–113.

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

[10] R. W. Katz, Precipitation as a chain-dependent process, Journal of Applied Meteorology and Climatology 16 (7) (1977) 671 – 676. doi:10.1175/1520-0450(1977)016¡0671:PAACDP¿2.0.CO;2.
URL https://journals.ametsoc.org/view/journals/apme/16/7/1520-0450_1977_016_0671_paacdp_2_0_co_2.xml

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

[12] C. Flecher, P. Naveau, D. Allard, N. Brisson, A stochastic daily weather generator for skewed data, Water Resources Research 46 (7) (2010).

[13] P. Ailliot, C. Thompson, P. Thomson, 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) (2009) 405–426.

[14] F. Oriani, J. Straubhaar, P. Renard, G. Mariethoz, Simulation of rainfall time series from different climatic regions using the direct sampling technique, Hydrology and Earth System Sciences 18 (8) (2014) 3015–3031.

[15] P. Yiou, Anawege: a weather generator based on analogues of atmospheric circulation, Geoscientific Model Development 7 (2) (2014) 531–543.

[16] F. Garavaglia, J. Gailhard, E. Paquet, M. Lang, R. Garçon, P. Bernardara, Introducing a rainfall compound distribution model based on weather patterns sub-sampling, Hydrology and Earth System Sciences 14 (6) (2010) 951–964.

[17] L. Benoit, M. Vrac, G. Mariethoz, Nonstationary stochastic rain type generation: accounting for climate drivers, Hydrology and Earth System Sciences 24 (5) (2020) 2841–2854.

[18] P. M. Williams, Modelling seasonality and trends in daily rainfall data, in: Advances in neural information processing systems, 1998, pp. 985–991.

[19] R. E. Chandler, On the use of generalized linear models for interpreting climate variability, Environmetrics 16 (7) (2005) 699–715.

[20] A. Verdin, B. Rajagopalan, W. Kleiber, G. Podesta, F. Bert, A conditional stochastic weather generator for seasonal to multi-decadal simulations, Journal of Hydrology 556 (2018) 835 – 846.

[21] P. McCullagh, J. A. Nelder, Generalized linear models, Monographs on statistics and applied probability, Chapman and Hall, 1989.

[22] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2020).
URL https://www.R-project.org/

[23] R. E. Chandler, Multisite, multivariate weather generation based on generalised linear models, Environmental Modelling and Software 134 (2020) 104867.

[24] J. R. M. Hosking, J. R. Wallis, Regional frequency analysis: an approach based on L-moments, Cambridge University Press, 2005.

[25] R. E. Chandler, A multisite, multivariate daily weather generator based on generalized linear models, User guide : R package (2015).
URL `https://www.ucl.ac.uk/~ucakarc/work/glimclim.html`

[26] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Munoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume, J.-N. Thépaut, The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society 146 (730) (2020) 1999–2049.

[27] P. V. Ayar, M. Vrac, S. Bastin, J. Carreau, M. Déqué, C. Gallardo, Intercomparison of statistical and dynamical downscaling models under the Euro-and Med-CORDEX initiative framework: present climate evaluations, Climate Dynamics 46 (3-4) (2016) 1301–1329.

[28] D. Maraun, F. Wetterhall, A. M. Ireson, R. E. Chandler, E. J. Kendon, M. Widmann, S. Brienen, H. W. Rust, T. Sauter, M. Themeßl, V. K. Venema, K. P. Chun, C. M. Goodess, R. G. Jones, C. Onof, M. Vrac, I. Thiele-Eich, Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user, Reviews of Geophysics 48 (3) (2010).

[29] B. François, M. Vrac, A. J. Cannon, Y. Robin, D. Allard, Multivariate bias corrections of climate simulations: which benefits for which losses?, Earth System Dynamics 11 (2) (2020) 537–562.

[30] P. A. Michelangeli, M. Vrac, H. Loukos, Probabilistic downscaling approaches: Application to wind cumulative distribution functions, Geophysical Research Letters 36 (11) (2009).

[31] A. J. Cannon, Multivariate quantile mapping bias correction: an n-dimensional probability density function transform for climate model simulations of multiple variables, Climate dynamics 50 (1-2) (2018) 31–49.

[32] C. Leduc, S. Ben Ammar, G. Favreau, R. Beji, R. Virrion, G. Lacombe, J. Tarhouni, C. Aouadi, B. Zenati Chelli, N. Jebnoun, M. Oi, J. L. Michelot, K. Zouari, Impacts of hydrological changes in the Mediterranean zone: environmental modifications and rural development in the Merguellil catchment, central Tunisia - Un exemple d'évolution hydrologique

en Méditerranée: impacts des modifications environnementales et du développement agricole dans le bassin-versant du Merguellil (Tunisie centrale), Hydrological Sciences Journal 52 (6) (2007) 1162–1178.

[33] F. Molle, P. Wester, River basin trajectories: societies, environments and development, Vol. 8, IWMI, 2009.

[34] S. Ben Ammar, K. Zouari, C. Leduc, J. M'barek, Caractérisation isotopique de la relation barrage-nappe dans le bassin du merguellil (plaine de kairouan, tunisie centrale), Hydrological sciences journal 51 (2) (2006) 272–284.

[35] C. Leduc, R. Calvez, R. Beji, Y. Nazoumou, G. Lacombe, C. Aouadi, Evolution de la ressource en eau dans la vallée du Merguellil (Tunisie centrale), in: Séminaire sur la modernisation de l'agriculture irriguée, IAV Hassan II, 2004, pp. 10–p.

[36] OMM, Guide des instruments et des méthodes d'observation météorologiques, OMM 8, Geneve, 2010.

[37] D. P. Dee, S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. C. M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haimberger, S. B. Healy, H. Hersbach, E. V. Hólm, L. Isaksen, P. Kållberg, M. Köhler, M. Matricardi, A. P. McNally, B. M. Monge-Sanz, J. J. Morcrette, B. K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J. N. Thépaut, F. Vitart, The Era-Interim reanalysis: Configuration and performance of the data assimilation system, Quarterly Journal of the royal meteorological society 137 (656) (2011) 553–597.

[38] J. Hooker, G. Duveiller, A. Cescatti, A global data set of air temperature derived from satellite remote sensing and weather stations, Scientific data 5 (2018) 180246.

[39] R. G. Allen, L. S. Pereira, M. Smith, D. Raes, J. Wright, FAO-56 dual crop coefficient method for estimating evaporation from soil and application extensions, Journal of irrigation and drainage engineering 131 (1) (2005) 2–13.

[40] D. Déqué, Frequency of precipitation and temperature extremes over france in an anthropogenic scenario: Model results and statistical correction according to observed values, Global and Planetary Change 57 (1) (2007) 16 – 26, Extreme Climatic Events.

[41] M. Vrac, P. Friederichs, Multivariate-intervariable, spatial, and temporal bias correction, Journal of Climate 28 (1) (2015) 218–237.

[42] A. J. Cannon, Neural networks for probabilistic environmental prediction: Conditional Density Estimation Network Creation and Evaluation (CaDENCE) in R, Computers & Geosciences 41 (2012) 126–135.

[43] P. Bruce, A. Bruce, P. Gedeck, Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python, O'Reilly Media, 2020.

[44] P. Filzmoser, B. Liebmann, K. Varmuza, Repeated double cross validation, Journal of Chemometrics: A Journal of the Chemometrics Society 23 (4) (2009) 160–171.

[45] H. Joe, Multivariate models and multivariate dependence concepts, CRC Press, 1997.

[46] G. Boulet, B. Mougenot, J. Lhomme, P. Fanise, Z. Lili-Chabaane, A. Olioso, M. Bahir, V. Rivalland, L. Jarlan, O. Merlin, et al., The sparse model for the prediction of water stress and evapotranspiration components from thermal infra-red data and its evaluation over irrigated and rainfed wheat, Hydrology and Earth System Sciences 19 (11) (2015) 4653–4672.

[47] E. Delogu, G. Boulet, A. Olioso, B. Coudert, J. Chirouze, E. Ceschia, V. Le Dantec, O. Marloie, G. Chehbouni, J. P. Lagouarde, Reconstruction of temporal variations of evapotranspiration using instantaneous estimates at the time of satellite overpass, Hydrology and earth system sciences 16 (2012) 2995–3010.

[48] S. Er-Raki, A. Chehbouni, S. Khabba, V. Simonneaux, L. Jarlan, A. Ouldbba, J. Rodriguez, R. Allen, Assessment of reference evapotranspiration methods in semi-arid regions: Can weather forecast data be used as alternate of ground meteorological parameters?, Journal of Arid Environments 74 (12) (2010) 1587–1596.

[49] C. Leauthaud, B. Cappelaere, J. Demarty, F. Guichard, C. Velluet, L. Kergoat, T. Vischel, M. Grippa, M. Mouhaimouni, I. Bouzou Moussa, I. Mainassara, B. Sultan, A 60-year reconstructed high-resolution local meteorological data set in central sahel (1950–2009): evaluation, analysis and application to land surface modelling, International Journal of Climatology 37 (5) (2017) 2699–2718.

[50] G. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (2) (1978) 461–464. doi:10.1214/aos/1176344136.

[51] N. Fodor, I. Dobi, J. Mika, L. Szeidl, Applications of the MVWG multivariable stochastic weather generator, The Scientific World Journal 2013 (2013).

[52] J. Chen, F. P. Brissette, R. Leconte, A daily stochastic weather generator for preserving low-frequency of climate variability, Journal of hydrology 388 (3-4) (2010) 480–490.

[53] T. Carey-Smith, J. Sansom, P. Thomson, A hidden seasonal switching model for multisite daily rainfall, Water Resources Research 50 (1) (2014) 257–272.