Impact of Weather on Energy Demand: Cross-Section Prediction and Time-Series Forecasting

Written and presented by

R. AL AFIA, S. AUGUSTE, N. BOIMOND

D. Mamadou, P. OUEDRAOGO

Under the supervision of

Professor Christophe MULLER

Predictive Methods Course

January 20, 2026

Contents

1 Introduction

1.1 General context

Energy demand is a critical factor in modern infrastructure, influencing everything from economic stability to environmental sustainability. As global energy systems evolve, understanding the variables that drive consumption becomes increasingly important. Among these variables, weather stands out as a powerful and dynamic influence. Temperature fluctuations, precipitation, wind patterns, and seasonal changes directly impact heating, cooling, and overall electricity usage both at the household and industrial levels.

1.2 Motivation and research questions

This report explores the impact of weather on energy demand, employing two complementary analytical approaches: cross-section prediction and time-series forecasting. By examining how weather variables correlate with energy consumption across different regions and over time, we aim to answer key questions: How do specific weather conditions alter energy demand patterns? Can we accurately predict future demand based on weather forecasts? What are the implications for energy providers, policymakers, and consumers ?

The significance of this topic lies in its potential to enhance energy efficiency, optimize resource allocation, and mitigate the risks of supply shortages or surpluses. As climate change intensifies weather variability, the ability to forecast demand with precision becomes not only valuable but essential. Through this analysis, we seek to provide actionable insights for stakeholders in the energy sector, contributing to more resilient and sustainable energy systems.

2 Data Description

2.1 Data Sources

The data used in this study come from the Météo France[1] database, accessed through the official French open data platform data.gouv.fr. Météo France is the national meteorological service of France and provides The data used in this project come from two main open data sources: Météo France and Open Data Réseaux Énergies (ODRÉ)[2]. Both sources are publicly available through official French open data platforms and provide reliable, high-quality information relevant to the study of energy demand and its relationship with weather conditions.

Meteorological data were obtained from the Météo France database, hosted by the French official open data portal. This dataset contains daily observations collected by

weather stations located across metropolitan France over the period from 2013 to 2023. The data include several weather indicators such as temperature, precipitation, and other atmospheric variables. In this study, particular attention is given to daily mean temperature, as it is a key determinant of heating and cooling needs and, consequently, energy demand.

In addition to meteorological data, energy-related data were sourced from the Open Data Réseaux Énergies (ODRÉ) platform. ODRÉ is a public initiative that provides open-access data related to energy production, consumption, infrastructure, markets, and territories in France.

By combining meteorological data from Météo France with energy-related data from ODRÉ, this project relies on complementary sources that allow for a more comprehensive analysis of the interaction between weather conditions and energy demand. Both datasets are collected on a regular basis, cover several years, and are suitable for quantitative analysis, which makes them well adapted to the objectives of this study.

2.2 Data description and preprocessing

2.2.1 Weather

The dataset from météo France can be characterized as panel data, combining observations across time and across spatial units (weather stations and departments). Each observation corresponds to a specific station on a given day, which results in a large number of observations but also introduces heterogeneity in data availability across stations.

The raw dataset includes several meteorological indicators such as temperature, precipitation, wind speed, global radiation, sunshine duration, and humidity. Due to differences in station activity and reporting practices, some stations exhibit substantial missing values. To ensure data quality and temporal consistency, a station-level completion rate was computed for each variable. Only stations with a completion rate of at least 80% were retained for further analysis, which limits the influence of inactive or unreliable stations.

For the purpose of this study, two different data structures were constructed depending on the modeling objective. For time-series analysis, the data were reshaped to obtain daily temperature series at the department level for the ProvenceAlpesCôte dAzur (PACA) region. After filtering active stations, daily temperatures were aggregated by department using the mean across stations. In this setting, the data take the form of multivariate time-series data, where each department is observed repeatedly over time and each observation represents the daily average temperature of a given department on a specific date.

For cross-sectional prediction models, the data were organized differently. The full metropolitan area was retained in order to capture spatial variability across departments.

Multiple meteorological variables were selected as explanatory features, while temperature was used as the target variable. In this case, each observation corresponds to a department-day combination, which allows the analysis of relationships between temperature and weather-related variables across space rather than over time.

After aggregation, a missing value analysis revealed only a limited number of remaining missing observations, mainly affecting global radiation and sunshine duration. Given their small proportion, these values were imputed using the mean of the corresponding variables. The final meteorological dataset was then sorted chronologically and stored in a processed format.

Overall, the dataset combines temporal, spatial, and quantitative dimensions, which makes it particularly rich but also requires careful preprocessing. The distinction between time-series data and cross-sectional data allows the project to address complementary research questions, while ensuring that the data structure is well adapted to each modeling approach.

Date	Department name	Temperature
2013-01-01	Alpes-de-Haute-Provence	0.515
2013-01-01	Alpes-Maritimes	3.134
⋮	⋮	⋮
2013-01-02	Alpes-de-Haute-Provence	1.084
2013-01-02	Alpes-Maritimes	4.848
⋮	⋮	⋮

Table 1: Panel data : Daily temperatures by department of the P.A.C.A. region

2.2.2 Energy

The energy data used in this study take the form of aggregated univariate time-series data. Each observation represents total daily energy consumption in the ProvenceAlpesCôte dAzur (PACA) region on a given date. The data are ordered over time and recorded at a daily frequency, making them well suited for temporal analysis and forecasting. In their raw form, the energy data include daily observations of electricity and gas consumption reported in megawatts (MW) at the regional level. Together, these variables provide a comprehensive measure of energy usage across different energy carriers. As an initial preprocessing step, the data were cleaned and the date variable was converted into a datetime format. The analysis was then restricted to the PACA region to ensure geographical consistency with the meteorological data.

Total daily energy consumption was constructed by summing gross electricity consumption and gross gas consumption, resulting in a single indicator of overall energy demand. The data were subsequently aggregated at the daily level and filtered to match the study period starting in 2013.

Finally, the processed energy dataset was merged with the meteorological dataset using the date variable, producing a temporally aligned and consistent dataset suitable for analyzing the relationship between weather conditions and energy demand.

Y: Energy_cons_MW	β_1: Temperature	$\cdots$	β_6: Humidity
185715.0	4.795	$\cdots$	4.882
267200.0	4.476	$\cdots$	0.038
281535.0	4.670	$\cdots$	0.0201
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 2: Prediction datasets shape

3   Explanatory data analysis

XXXXXXXXXXXXXXXXXXXXXXXX

4   Variable selection and empirical strategy

4.1   Energy : cross-section analysis

For the cross section part (energy) of this project, the dependent variable in this study is total energy consumption, denoted Y_i , measured in megawatts (MW). It corresponds to the variable *total_energy_consumption_MW*' in the final dataset. It represents the sum of gross electricity consumption and gross gas consumption, aggregated at a daily frequency for the PACA region. Energy consumption is a key economic indicator, as it reflects households and firms demand for heating, cooling, lighting, and other energy-intensive activities. From a statistical perspective, the variable exhibits substantial temporal variability and pronounced seasonal patterns, making it well suited for econometric analysis and prediction.

The explanatory variables are derived from daily meteorological observations provided by Météo France. They include average daily temperature (TM, in řC), average humidity (UM, in %), precipitation (RR, in mm), wind speed (FFM, in m/s), global solar radiation (GLOT, in Wh/mš), and sunshine duration (INST, in hours)[3]. These variables are expected to influence energy consumption through several channels: temperature directly affects heating and cooling demand, humidity modifies thermal comfort, while solar radiation and sunshine duration impact lighting needs and energy production conditions. Wind and precipitation may also indirectly affect energy demand by influencing perceived temperature and usage behavior. To account for potential non-linear relationships, the empirical specification includes a quadratic term in temperature and an interaction term

between temperature and humidity, allowing the effect of temperature on energy consumption to vary with humidity levels. All continuous variables involved in non-linear terms are centered to improve coefficient interpretability and reduce multicollinearity. The empirical strategy consists in modeling the conditional mean of energy consumption as a function of meteorological variables and comparing several predictive models based on their out-of-sample performance, using RMSE and MAE as evaluation criteria.

4.2   Time-series analysis

For the times series part (weather) of this project, the variable of interest is the average daily temperature, denoted TM_t, measured in degrees Celsius (řC) and observed at a regular daily frequency. The temperature series is obtained by averaging department-level observations across France, resulting in a single national temperature time series. The objective is to predict future temperature values TM_{t+h} using only past information contained in the series itself, in line with standard univariate time-series forecasting frameworks.

From a statistical perspective, this variable exhibits strong seasonal patterns and serial dependence, which are characteristic features of meteorological time series. From an applied standpoint, temperature forecasting is a central task in meteorology and is highly relevant for downstream applications such as energy demand forecasting, agriculture, and climate-related decision-making.

About the explanatory variables of weather, they consist exclusively of past realizations of the dependent variable itself. Specifically, lagged values $\{TM_{t-1}, TM_{t-2}, \cdots\}$ are used to capture short-term temporal dependence, reflecting the persistence of weather conditions over consecutive days. Moving-average components are introduced to model the dependence structure of forecast errors, allowing the model to account for shocks that affect temperature temporarily.

Given the strong annual seasonality of temperature data, seasonal autoregressive and moving-average terms are also included, associated with a yearly periodicity. These components capture recurring patterns linked to the calendar, such as warmer summers and colder winters. Preliminary graphical analysis, including the raw series and moving averages, suggests that seasonality is stable over time and that no dominant deterministic long-term trend is present.

The empirical strategy consists of identifying the appropriate temporal structure of the series before specifying a forecasting model. This includes assessing stationarity, serial correlation, and seasonal patterns. Based on these properties, alternative specifications such as ARMA, ARIMA, or SARIMA models are considered. Model performance is evaluated out of sample using standard forecasting accuracy criteria, including the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE). This approach

ensures that the selected model provides a statistically sound and empirically reliable representation of temperature dynamics over time.

5 Econometric Modeling

5.1 Energy consumption

5.1.1 Model specification

Several predictive models are considered to explain total energy consumption using meteorological variables. The baseline specification relies on a linear regression model estimated by Ordinary Least Squares (OLS), which serves as a reference framework for both interpretation and comparison with more flexible approaches.

Baseline Model: Linear Regression (OLS)

The reference model assumes a linear relationship between energy consumption and weather conditions. The econometric specification is given by :

$$Y_i = \beta_0 + \beta_1 TM_i + \beta_2 TM_i^2 + \beta_3 UM_i + \beta_4 RR_i + \beta_5 FFM_i + \beta_6 GLOT_i + \beta_7 INST_i + \beta_8 (TM_i \times UM_i) + u_i \tag{1}$$

Where Y_i denotes total daily energy consumption measured in megawatts (MW), and u_i is an error term capturing unobserved factors. The model is estimated under the standard exogeneity assumption $\mathbf{E}(u_i \mid X_i) = 0$, implying that the linear specification aims to approximate the conditional mean of energy consumption given the meteorological variables.

The inclusion of a quadratic temperature term allows for a non-linear relationship between temperature and energy demand, reflecting increased consumption during both cold and hot extremes due to heating and cooling needs. The interaction term between temperature and humidity captures the idea that thermal discomfortand thus energy demandmay increase more strongly when high temperatures coincide with high humidity levels. Variables related to precipitation, wind speed, solar radiation, and sunshine duration account for additional channels through which weather conditions can affect energy usage, such as heat losses, lighting needs, and solar gains.

Penalized Regression Models: Ridge and Lasso

Meteorological variables are often highly correlated, particularly temperature, solar radiation, and sunshine duration. To address potential multicollinearity and improve out-of-sample predictive performance, penalized regression models are also considered. Ridge regression introduces an 2 penalty that shrinks coefficients toward zero while retaining all regressors, thereby stabilizing estimation in the presence of strong correlations. Lasso regression relies on an penalty, which1 can set some coefficients exactly to zero and thus performs variable selection in addition to regularization.

All models are evaluated based on their predictive performance using out-of-sample criteria. The primary loss functions considered are the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE), in accordance with the project guidelines. Comparing OLS with Ridge and Lasso allows assessing the trade-off between interpretability and predictive accuracy in explaining energy consumption from meteorological conditions.

5.1.2 Estimation method

In the cross-sectional framework devoted to energy consumption forecasting, several estimation methods are considered in order to compare their predictive performance and to account for the statistical properties of the meteorological explanatory variables. The benchmark estimation method is Ordinary Least Squares (OLS). Under the standard assumption of conditional exogeneity, $\mathbf{E}(u_i \mid X_i) = 0$, the OLS estimator is unbiased and consistent. It provides a natural reference model, as it allows a direct economic interpretation of the estimated coefficients and makes it possible to assess the relevance of nonlinear effects, such as the quadratic term in temperature and the interaction between temperature and humidity. In particular, OLS identifies average marginal effects of meteorological variables on energy consumption. However, several explanatory variables are potentially highly correlated, notably temperature, its squared term, sunshine duration, and global radiation. This multicollinearity may inflate the variance of OLS estimators and weaken out-of-sample predictive performance. To address this issue, penalized regression methods are also implemented.

Ridge regression introduces an L_2 penalty on the coefficients and is defined as :
$$\hat{\beta}^R = argmin_\beta \sum_i (Y_i - X_i\beta)^2 + \lambda \sum_{j \geq 1} \beta_j^2. \tag{2}$$
This penalization reduces the variance of the estimators in the presence of multicollinearity, at the cost of a controlled bias, and improves the stability of the estimated coefficients.

Lasso regression relies on an L_1 penalty and is defined as:
$$\hat{\beta}^L = argmin_\beta \sum_i (Y_i - X_i\beta)^2 + \lambda \sum_{j \geq 1} \mid \beta_j \mid. \tag{3}$$
Unlike Ridge regression, Lasso performs automatic variable selection by allowing some coefficients to be exactly zero. This feature is particularly useful when not all meteorological variables are equally relevant for predicting energy consumption.

Finally, model selection is based on out-of-sample predictive performance, in accordance with the project guidelines. The OLS, Ridge, and Lasso models are compared using standard loss criteria such as the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE). The final choice relies on a biasvariance trade-off and on the ability of the model to generalize beyond the estimation sample.

5.2 Weather

5.2.1 Stationarity assessment

Before estimating a time-series model, it is necessary to assess the stationarity properties of the temperature series. Time-series models such as ARMA, ARIMA, and SARIMA rely on the assumption of stationarity, at least after possible transformations. The identification of unit roots and seasonal patterns at this stage therefore directly guides the choice of the appropriate modeling framework. A preliminary visual inspection of the series suggests the presence of a pronounced and stable annual seasonality. To formally test for stationarity, two complementary unit root tests are employed: the Augmented Dickey-Fuller (ADF) test and the KPSS test. These tests are used jointly because they rely on opposite null hypotheses, which strengthens the robustness of the inference. For the ADF test, the null hypothesis is that the temperature series contains a unit root and is therefore non-stationary :

$$H_0^{ADF} : \text{the series is non-stationary.} \tag{H}$$

The alternative hypothesis is :

$$H_1^{ADF} : \text{the series is stationary.} \tag{H}$$

The Augmented Dickey-Fuller (ADF) test yields a test statistic of 4.29 with a p-value of 0.00047, which is far below standard significance levels. Consequently, the null hypothesis of a unit root is rejected, indicating that the temperature series is stationary in level.

For the KPSS test, the hypotheses are reversed. The null hypothesis assumes stationarity of the series :

$$H_0^{KPSS} : \text{the series is stationary.} \tag{H}$$

The alternative hypothesis is :

$$H_1^{KPSS} : \text{the series is non-stationary.} \tag{H}$$

The KPSS test produces a test statistic of 0.077 with a p-value of 0.10. As a result, the null hypothesis of stationarity cannot be rejected. The warning message indicates that the test statistic is very small and that the true p-value is even larger than the reported value, which further supports stationarity.

Taken together, the ADF and KPSS tests provide consistent evidence that the temperature series is stationary, despite the presence of a stable annual seasonal pattern. Therefore, no additional differencing is required before estimating time-series models, and seasonality can be directly modeled within a SARIMA framework.

The figure illustrates the daily temperature series used for estimating the SARIMA model, along with a 12-month moving average. The series is considered in levels, in line with the stationarity test results. A pronounced and stable annual seasonality clearly emerges, while no strong long-term deterministic trend is observed. This visual evidence supports the inclusion of seasonal components in the SARIMA specification to capture the yearly temperature cycle.
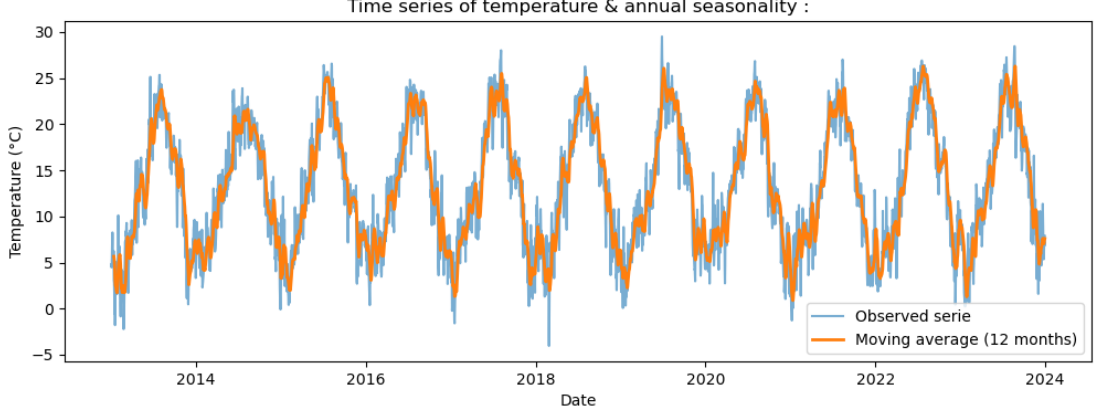
Figure 1: **SARIMA** illustration

5.2.2 Model specification

We consider a time-series forecasting framework for the monthly average temperature, denoted TM_t. As a baseline, a seasonal naïve forecasting model is used, defined by :

$$\widehat{TM}_t = TM_{t-12} \tag{4}$$

This approach predicts the temperature of a given month using the observed value from the same month of the previous year. It provides a minimal benchmark against which the performance gains of more sophisticated models can be evaluated.

SARIMA model

To jointly capture short-term temporal dependence and the pronounced annual seasonality observed in the data, a SARIMA model is considered. The specification is :

$$\mathbf{SARIMA}(1, 0, 1) \times (1, 1, 1)_{12} \tag{5}$$

The non-seasonal components (1) model short-run dynamics, while the seasonal components (1) capture dependencies between observations separated by 12 months. The seasonal order $= 12$ is imposed by the monthly frequency of the data.

Stationarity tests (ADF and KPSS) indicate that the series is stationary in levels, which justifies the absence of non-seasonal differencing ($d = 0$). However, given the strong and stable annual seasonality, a first-order seasonal differencing ($D = 1$) is introduced to stabilize the seasonal pattern.

Overall, the inclusion of first-order autoregressive and moving-average components at both the non-seasonal and seasonal levels allows the model to capture temporal dependence while maintaining a parsimonious specification. The relevance of the SARIMA model is assessed by comparing its out-of-sample forecasting performance to that of the seasonal naïve benchmark using RMSE and MAE criteria.

11

5.2.3 Estimation methods

In the time-series framework, the SARIMA model is estimated using **Maximum Likelihood Estimation (MLE)**. This method consists in selecting the model parameters that maximize the conditional likelihood of the observed series, given the information available up to time $t - 1$. Let F_{t-1} denote the information set generated by past observations, and let ϵ_t be the innovation term. The central assumption of the model is :

$$\mathbf{E}(\epsilon_t \mid F_{t-1}) = 0 \tag{6}$$

which implies that forecast errors are not predictable using past information. Under the standard SARIMA assumptions, the innovations ϵ_t are uncorrelated, have zero mean and constant variance, and are generally assumed to follow a Gaussian distribution for inference purposes. Under these conditions, Maximum Likelihood Estimation provides efficient estimators of the model parameters.

MLE estimation also allows the direct derivation of **point forecasts** for the temperature series TM_t, as well as prediction intervals, which quantify the uncertainty associated with the forecasts. In this project, the model is estimated on a training sample, and its predictive performance is evaluated out-of-sample and compared to that of the seasonal naïve benchmark using RMSE and MAE criteria.

Why Maximum Likelihood Estimation ?

In time-series forecasting, the objective is not only to explain the variable of interest but primarily to produce **optimal forecasts conditional on past information**. SARIMA models belong to the class of **stochastic dynamic models**, in which the observed variable depends on its past values, unobserved random innovations, and, when relevant, seasonal components.

In this context, Maximum Likelihood Estimation is the natural estimation method, as it relies directly on the **conditional distribution of the series given past information**. The likelihood function measures the probability of observing the realized trajectory of the series conditional on the model parameters. Maximizing this likelihood therefore amounts to choosing the parameters that make the observed data the most plausible, given the temporal dynamics imposed by the model.

This estimation method presents several key advantages in the context of this project. First, it yields efficient estimators when the model is correctly specified. Second, it is well suited to dynamic models featuring temporal dependence and seasonality. Finally, it allows the direct construction of point forecasts and prediction intervals, which are essential for assessing forecast uncertainty.

Overall, the use of Maximum Likelihood Estimation is fully consistent with the forecasting objective of weather prediction and with the probabilistic structure of the SARIMA model.

6 Estimation Results

6.1 Energy estimation output

XXXXXXXXXXXXXXXXXXXXXX

6.2 Weather estimation output

XXXXXXXXXXXXXXXXXXXXXX

7 Comparison of Models and Choice of the Forecasting Method

7.1 Energy consumption prediction (Cross-Section)

Several models were estimated to predict total energy consumption using meteorological variables. The approaches considered include a linear regression estimated by Ordinary Least Squares (OLS), as well as two penalized regression models, Ridge and Lasso.

Model selection is based exclusively on out-of-sample predictive performance, assessed using forecasting loss functions adapted to prediction problems, namely the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE), in accordance with the project guidelines.

The results indicate that the predictive performances of the three models are overall very close. The OLS model provides competitive forecasts, while the Ridge and Lasso models do not lead to a substantial improvement in terms of RMSE or MAE on the test sample.

In this context, the OLS model is retained as the reference model, as it offers a good compromise between predictive accuracy, simplicity, and economic interpretability. The penalized models nevertheless remain informative: Ridge regression helps stabilize coefficient estimates in the presence of multicollinearity, while Lasso highlights that some meteorological variables have a limited marginal contribution to prediction.

Consistent with the objective of the project, this choice is driven by predictive performance rather than by strict causal interpretation of the estimated coefficients.

7.2 Temperature forecasting (Time series)

The second forecasting task focuses on predicting the monthly average temperature based solely on its past temporal dynamics. Two approaches are compared :

1. A seasonal naïve model, used as a benchmark.

2. A SARIMA model, designed to capture both short-term temporal dependence and annual seasonality observed in the data.

The seasonal naïve model assumes that the temperature in a given month is equal to the temperature observed in the same month of the previous year. It provides a simple but demanding baseline against which more sophisticated models are evaluated.

The SARIMA model, estimated by maximum likelihood, exploits the information contained in both short-run dynamics and the seasonal structure of the temperature series. Its specification is guided by graphical analysis and stationarity tests, in line with the theoretical framework developed in the course.

The comparison of out-of-sample forecasting performance shows that the SARIMA model improves prediction accuracy relative to the naïve benchmark, particularly in terms of RMSE and MAE. This result indicates that temperature dynamics cannot be reduced to a simple annual repetition and that intra-annual temporal dependence contains valuable predictive information.

As a result, the SARIMA model is selected as the preferred forecasting method, as it provides a better balance between predictive accuracy and a rigorous representation of the underlying temporal structure of the data.

7.3 Methodological remark

In both forecasting exercises, the final model selection relies exclusively on out-of-sample predictive performance criteria, rather than on in-sample goodness-of-fit measures or the individual statistical significance of estimated parameters.

This methodological choice is fully consistent with the project instructions and with the spirit of the course, which emphasize a rigorous evaluation of forecasting performance over causal interpretation of model coefficients.

8 Final forecasts and comparative analysis

XXXXXXXXXXXXXXXXXXXXXXXX

9 Robustness analysis and model limitations

XXXXXXXXXXXXXXXXXXXXXXXX

10 Final conclusion of the project

XXXXXXXXXXXXXXXXXXXXXXXX

References

[1] Météo France. *Données climatologiques de base - quotidiennes* Licence Ouverte / Open Licence version 2.0.

[2] ODRE. *Consommation quotidienne brute régionale* Licence Ouverte / Open Licence version 2.0.

[3]
- **RR**: Quantité de précipitation tombée en 24 heures, mesurée en millimètres et 1/10.
- **TM**: Moyenne quotidienne des températures horaires, exprimée en degrés Celsius avec une précision de 1/10.
- **FFM**: Moyenne quotidienne de la force du vent, mesurée en mètres par seconde (m/s) et moyenne sur 10 minutes, à 10 m de hauteur.
- **INST**: Durée d'insolation quotidienne mesurée en minutes.
- **GLOT**: Rayonnement global quotidien, exprimé en joules par centimètre carré (J/cm^2).
- **UM**: Moyenne quotidienne des humidités relatives horaires, exprimée en pourcentage (%).