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 21, 2026

Contents

1 Introduction

1.1 General context

Energy demand plays a central role in modern economies, as it conditions economic activity, infrastructure planning, and environmental sustainability. Accurate forecasting of energy consumption is therefore a key concern for energy providers, policymakers, and system operators, particularly in a context of increasing climate variability and energy transition. Among the many determinants of energy demand, weather conditions stand out as a major and highly volatile factor. Temperature, humidity, precipitation, wind, and solar exposure directly affect heating, cooling, lighting needs, and indirectly influence industrial and household energy use.

At the same time, weather variables themselvesespecially temperatureexhibit strong temporal structures, characterized by persistence and pronounced seasonal patterns. Understanding and forecasting these dynamics is essential, both as an object of interest in its own right and as an input for energy demand forecasting. As a result, weather and energy forecasting naturally call for complementary empirical approaches, combining cross-sectional prediction models and time series methods.

1.2 Motivation and research questions

This project investigates the relationship between weather conditions and energy consumption using two complementary forecasting frameworks. First, energy consumption is predicted in a cross- sectional setting using meteorological variables as exogenous predictors. Second, temperature is modeled and forecasted using univariate time series techniques, relying exclusively on its past dynamics.

The analysis is guided by the following research questions :

1. *How well can energy consumption be predicted using meteorological information in a cross- sectional framework ?*

2. *How effectively do time series models capture and forecast the seasonal dynamics of temperature ?*

3. *How do cross-sectional and time series approaches complement each other in forecasting applications ?*

Rather than aiming at causal inference, the focus of this project is strictly predictive. Models are evaluated based on their out-of-sample forecasting performance, in line with modern forecasting practice and course guidelines.

1.3 Related literature

A large body of empirical literature documents the strong link between weather conditions and energy consumption. Early studies such as **Bessec and Fouquau (2008)** show that temperatureenergy relationships are inherently non-linear, reflecting asymmetric heating and cooling needs. More recent contributions, including **Auffhammer et al. (2017)** and **Deschênes and Greenstone (2011)**, emphasize the growing sensitivity of energy demand to extreme temperatures in the context of climate change.

From a methodological perspective, cross-sectional and panel regression models using weather variables have been widely employed for short-term energy demand forecasting (see **Hong et al., 2016**). These models are valued for their interpretability and their ability to incorporate rich meteorological information, but they may suffer from multicollinearity and limited ability to capture temporal dependence.

In parallel, time series models such as ARIMA and SARIMA have long been used to forecast meteorological variables, particularly temperature, due to their strong seasonal structure (**Box et al., 2015**). Seasonal naïve models provide a natural benchmark, but more flexible stochastic models often yield superior predictive performance by exploiting short-term persistence in the data.

1.4 Contribution and structure of the report

The contribution of this project is twofold. First, it provides an empirical assessment of how far meteorological variables alone can explain and predict daily energy consumption when non-linearities and interactions are explicitly modeled. Second, it illustrates how univariate time series models capture the seasonal and dynamic structure of temperature, and how this affects forecast accuracy relative to simple benchmarks. The report is organized as follows :

- **Part 1** : focuses on weather data and temperature forecasting using time series methods. It presents the data, exploratory analysis, stationarity tests, model specification, estimation, and forecast evaluation.
- **Part 2** : examines energy consumption in a cross-sectional framework, including data description, explanatory analysis, econometric modeling, estimation results, and model comparison.
- **Finally**, we will compare forecasting approaches, discuss robustness and limitations, and conclude on the implications of the results for predictive modeling.

Part I

2 Weather data & forecasting

2.1 Data description

2.1.1 Data sources

The data used in this study come from the Météo France[?] database, accessed through the official French open data platform data.gouv.fr. Météo France is the national meteorological service of France and provides Meteorological data were obtained from the Météo France database, hosted on data.gouv.fr, 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.

2.1.2 Data description and pre-processing

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 Exploratory Data Analysis

With the time series forecast exercise, we will try to check if there is something to predict (seasonality, trend or cyclicity) :

These figures (1 & 2) reveal 2 things :

- Identical shape and patern (Up and down)
- Fixed period of repetition (12 months)

This is exactly the definition of a seasonality ! Thus this will be our target for the forecast exercise. We can further add that there is a trend (although almost invisible to the naked eye) of rise in temperature, this could illustrate perfectly global warming.

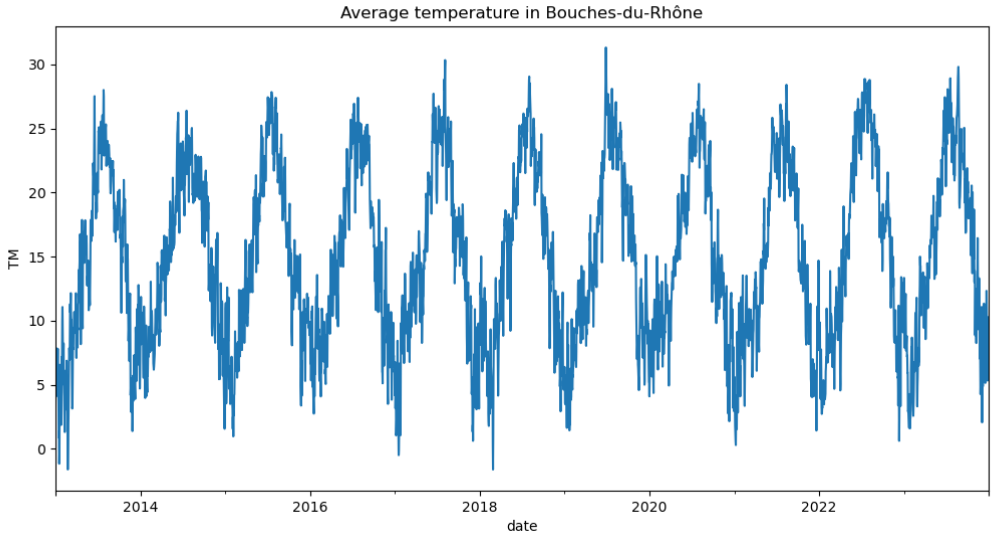

Figure 1: Time series of the Bouche du Rhône average temperature

2.3 Variable selection and empirical strategy

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

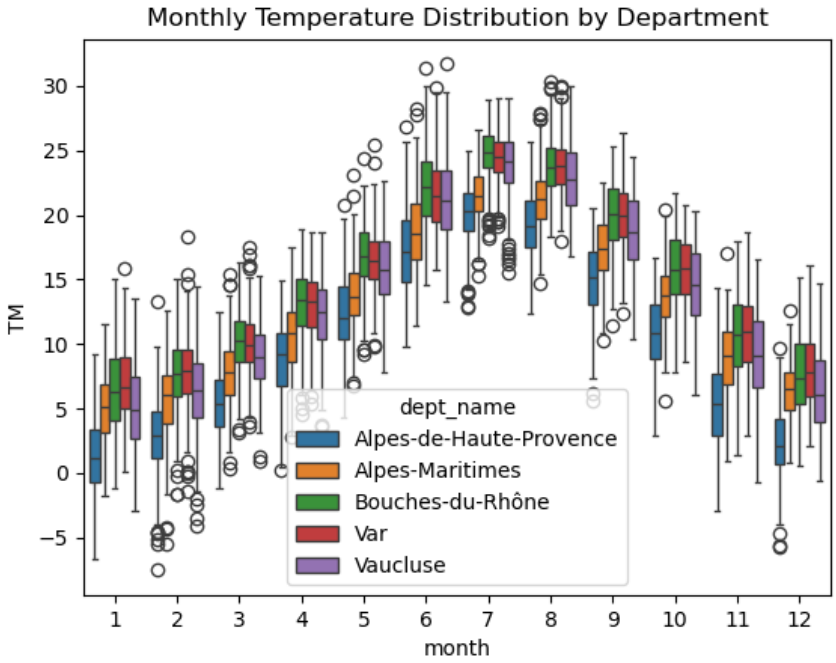

Figure 2: Seasonal patern

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.

2.4 Econometric modeling

2.4.1 Stationnary assessement

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 :

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

The alternative hypothesis is :

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

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 :

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

The alternative hypothesis is :

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

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.

2.4.2 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 :

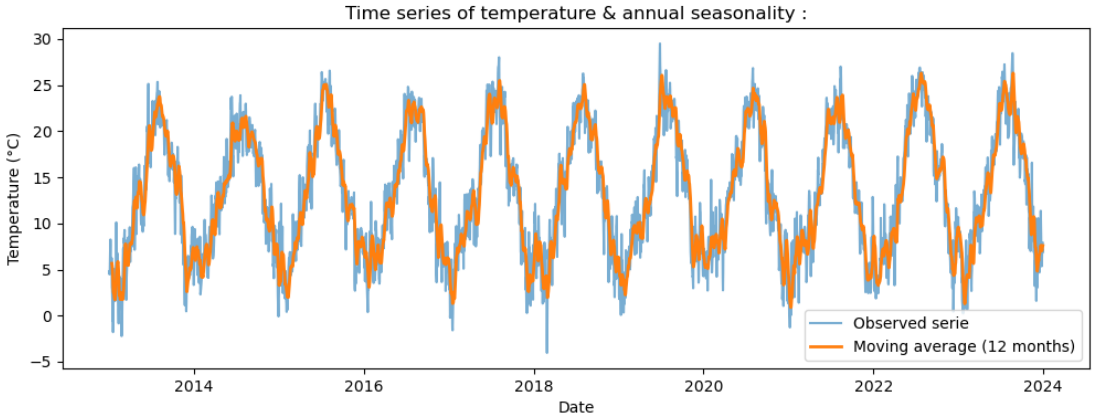

Figure 3: **SARIMA** illustration

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

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 :

$$\textbf{SARIMA}(1,0,1) \times (1,1,1)_{12}$$

The non-seasonal components (1,0,1) model short-run dynamics, while the seasonal components (1,1,1) 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.

2.4.3 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.

2.5 Estimation results

The second part of the analysis focuses on modeling and forecasting average monthly temperature. The variable of interest is the average temperature (TM), observed at a

monthly frequency from January 2013 to September 2021. The objective is to generate forecasts using only the information contained in the past realizations of the series, within a univariate time series framework.

Exploratory analysis reveals a strong and stable annual seasonality, with systematic temperature peaks during summer months and troughs during winter months, and no clear evidence of a deterministic long-term trend. These visual findings are confirmed by formal stationarity tests: the augmented DickeyFuller test rejects the null hypothesis of a unit root, while the KPSS test does not reject the null hypothesis of stationarity. Taken together, these results indicate that the series can be modeled in levels, provided that seasonality is explicitly incorporated.

On this basis, a SARIMA model is adopted to capture both short-term temporal dependence and annual seasonal patterns. The selected specification is a SARIMA(1,0,1) $\times$ $(1,1,1)_{12}$, reflecting monthly data with a twelve-month seasonal cycle. The model is estimated by maximum likelihood.

The estimation results highlight the importance of seasonal dynamics. Seasonal autoregressive and moving average parameters are statistically significant, confirming strong dependence between observations separated by one year. The non-seasonal moving average term is also significant, indicating that temperature shocks have a short-lived impact on the series. The residual variance is stable and consistent with observed fluctuations, suggesting that the model captures a substantial share of temperature variability.

Model diagnostics support the validity of the specification. Residual autocorrelation tests do not indicate remaining serial dependence, and normality tests do not reject the Gaussian assumption. No evidence of heteroskedasticity is detected, reinforcing the appropriateness of maximum likelihood estimation.

Forecasting performance is evaluated by comparing the SARIMA model to a seasonal naïve benchmark, which assumes that temperature in a given month equals that of the same month in the previous year. The SARIMA model consistently outperforms the benchmark in terms of RMSE and MAE on the test sample. This improvement demonstrates that temperature dynamics cannot be reduced to simple annual repetition and that short-term information significantly enhances predictive accuracy.

As in the cross-sectional analysis, the interpretation of the estimated parameters remains strictly predictive. The results show that temperature follows a well-structured temporal process dominated by stable seasonality and short-term dependence, and that SARIMA models provide an effective framework for medium-term temperature forecasting.

2.6 Comparison of Models and Choice of the Forecasting Method

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.

Methodological Remark

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.

2.7 Final forecasts and comparative analysis

Final temperature forecasts are produced using the time series model retained at the end of the selection phase. Forecasts are generated over an out-of-sample horizon and constitute the final predictive output of the temporal approach.

The predicted temperature trajectory extends the dynamics observed over the estimation period in a coherent manner. Forecasts reproduce the strong annual seasonality

of the series and follow the general trajectory of observed values, while smoothing short-term fluctuations. This behavior is characteristic of univariate time series models, which prioritize stability of the predicted trajectory over exact reproduction of transitory shocks.

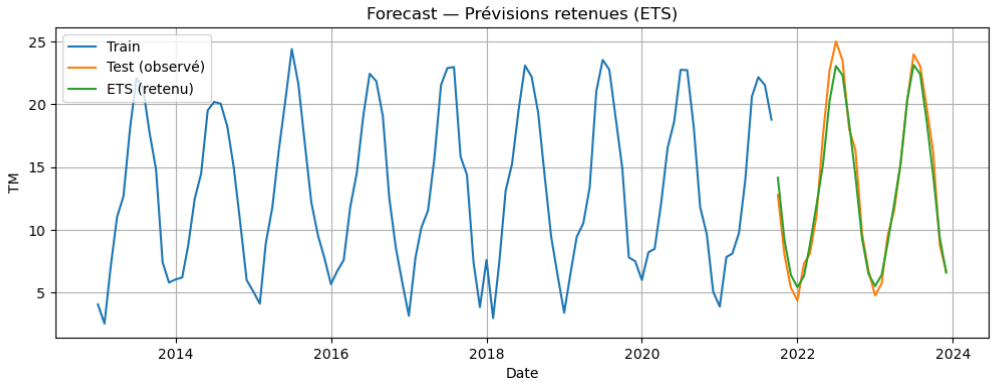

Figure 4: Observed and predicted values

Beyond point forecasts, uncertainty associated with the forecasts is explicitly taken into account through the construction of prediction intervals.

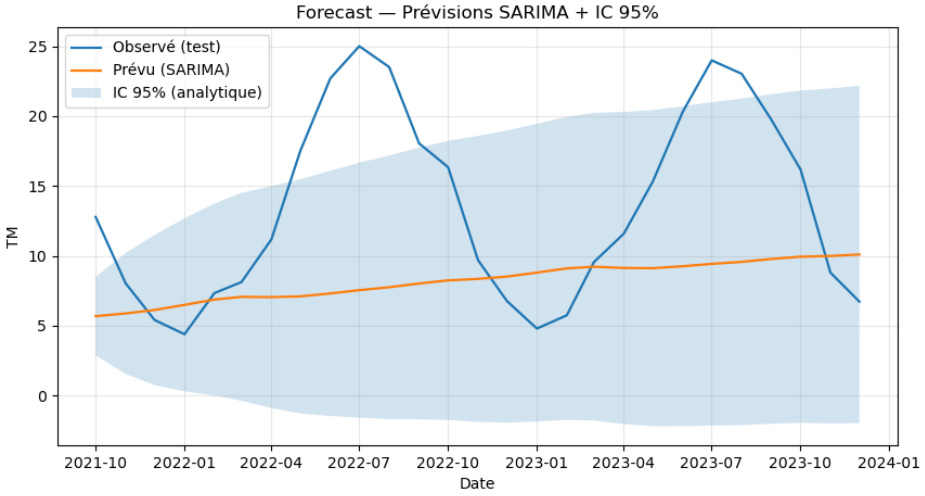

Figure 5: Temperature forecast with 95% prediction intervals

The prediction intervals show that most out-of-sample observations fall within the confidence bands, while their progressive widening as the forecast horizon increases reflects the natural accumulation of uncertainty in a time series framework.

Comparative Analysis

The temperature forecasting exercise highlights the effectiveness of a univariate time-series approach in capturing structured temporal dynamics. By relying exclusively on past realizations of the series, the SARIMA model exploits strong seasonality and short-term dependence to generate forecasts that are temporally coherent and stable. Compared

to the seasonal naïve benchmark, this approach delivers improved predictive accuracy, particularly at medium horizons. Nevertheless, the reliance on historical patterns implies a smoothing of short-term fluctuations, limiting the models ability to anticipate sudden or irregular temperature shocks. As a result, the time-series approach prioritizes structural regularity over responsiveness to unexpected variations

2.8 Robustness analysis and model limitations

Robustness of the results

The robustness of temperature forecasts is evaluated by comparing alternative time-series specifications and by examining the behavior of out-of-sample forecast errors. Models that explicitly incorporate annual seasonality consistently generate stable and coherent predictive trajectories, with no evidence of systematic bias. The SARIMA specification performs robustly relative to simpler benchmarks, indicating that the main temporal dynamics of the temperature series are adequately captured.

Model and data limitations

Nevertheless, several limitations apply. Temperature forecasting relies on univariate time-series models, which assume that all relevant predictive information is contained in past realizations of the series. As a result, unexpected shocks or structural changes cannot be anticipated. Moreover, forecast uncertainty increases with the prediction horizon, as reflected by widening prediction intervals, which limits precision at medium and long horizons.

Implications for interpretation

Temperature forecasts should therefore be interpreted as probabilistic projections of expected seasonal patterns rather than precise point predictions. They are particularly informative for identifying medium-term trends and seasonal behavior, while short-term irregular variations remain inherently difficult to predict.

Part II

3 Energy data & Prediction

3.1 Data description

3.1.1 Data sources

The data used in this part of the project come from Open Data Réseaux Énergies (ODRÉ) [**?**]. This source is 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.

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.

3.1.2 Data description and pre-processing

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.

3.2 Exploratory Data Analysis

For the prediction part, we can observe the following table :

As we had to restrict ourselve from using a lot of different features (due to extensively high missing value rate in the weather dataset), we only retained 6 defaults β. Here, a

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

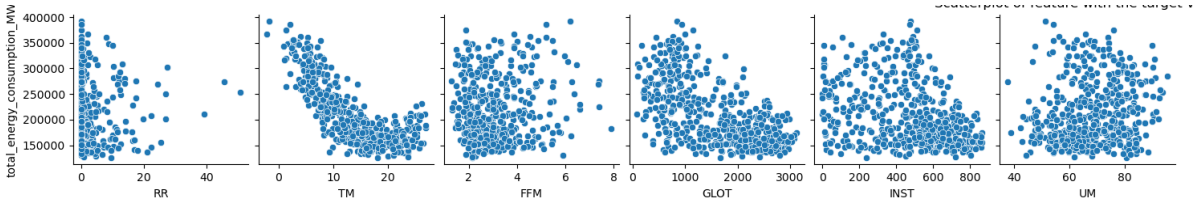

Figure 6: Features scatterplot with the target

simple glance is enough to draw a few conclusion :

- Some features are cleary uncorrelated to the target (a more in-depth analysis [**?**])
- We can expect to find non-linearity in the relation (J-shaped curve with β_{TM})

Meaning, that coming up with prediction with this few information, might be challenging in both a methodological and theoretical point of view (lack of meaning).

3.3 Variable selection and empirical strategy

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)[**?**]. 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.

3.4 Econometric modeling

3.4.1 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 :

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

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.

3.4.2 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 :

$$\hat{\beta}^R = argmin_\beta \sum_i (Y_i - X_i\beta)^2 + \lambda \sum_{j\geq 1} \beta_j^2.$$

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:

$$\hat{\beta}^L = argmin_\beta \sum_i (Y_i - X_i\beta)^2 + \lambda \sum_{j\geq 1} \mid \beta_j \mid.$$

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.

3.5 Estimation results

This subsection presents the estimation results for daily energy consumption using a cross-sectional framework. The dependent variable is total daily energy consumption, measured in megawatts, constructed as the sum of gross electricity and gas consumption. It provides a comprehensive indicator of overall energy demand.

The explanatory variables are exclusively meteorological and include average temperature, average humidity, precipitation, wind speed, global solar radiation, and sunshine duration. These variables are selected based on their expected influence on heating, cooling, lighting, and other weather-sensitive energy uses. This choice is fully consistent with the exploratory analysis, which highlights strong correlations between weather conditions and energy consumption.

Given evidence of non-linearity in the temperatureenergy relationship, the empirical specification incorporates a quadratic temperature term as well as an interaction between temperature and humidity. These terms allow the model to capture increased energy demand at temperature extremes and to account for the fact that temperature effects may vary with humidity levels. All continuous variables involved in non-linear transformations are centered to improve numerical stability and facilitate interpretation in the presence of multicollinearity.

The baseline model is estimated using ordinary least squares (OLS). In addition, Ridge and Lasso regressions are implemented to assess robustness to multicollinearity and to compare out-of-sample predictive performance.

The OLS estimation results show that meteorological variables explain a substantial share of the variation in energy consumption. The adjusted Rš is high, and the global F-test strongly rejects the null hypothesis of no joint explanatory power. Temperature

emerges as the dominant determinant of energy demand: the linear temperature coefficient is negative, while the quadratic term is positive and highly significant, revealing a clear U-shaped relationship. Energy consumption increases at low temperatures due to heating needs and at high temperatures due to cooling demand, a result that aligns closely with economic intuition.

Humidity also plays a significant role. The interaction between temperature and humidity is statistically significant, indicating that high humidity amplifies the effect of temperature on energy demand, particularly during hot periods. Variables related to solar exposure exhibit contrasting effects: sunshine duration is positively associated with energy consumption, while global solar radiation has a negative coefficient, reflecting potentially different channels such as lighting needs or indirect temperature effects. By contrast, precipitation and wind speed do not appear statistically significant once other meteorological variables are controlled for.

Diagnostic analysis reveals substantial multicollinearity among regressors, especially those related to temperature and solar exposure, as indicated by a high condition number. While this does not invalidate the model from a predictive standpoint, it motivates the use of penalized regressions. Ridge and Lasso models are therefore estimated and evaluated on a test sample using RMSE and MAE. The results show that all three models deliver very similar out-of-sample performance, with no meaningful gain from regularization. Consequently, the OLS specification is retained as the preferred forecasting model, as it combines strong predictive performance with simplicity and economic interpretability.

It is important to emphasize that these results are interpreted strictly in a predictive sense. The estimated coefficients capture statistical associations useful for forecasting and do not represent causal effects. Overall, the analysis confirms that meteorological variablesespecially temperature and humidityare key drivers of energy consumption and that relatively parsimonious models can yield reliable forecasts when properly specified and evaluated out of sample.

3.6 Comparison of Models and Choice of the Forecasting Method

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.

Methodological Remark

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.

3.7 Final forecasts and comparative analysis

Final forecasts of energy consumption are produced using the model retained at the end of the selection phase. Predictions are generated on the test sample using observed meteorological variables and constitute the final predictive output of the cross-sectional approach.

The predicted trajectory of energy consumption is globally consistent with observed levels in the out- of-sample period. Forecasts correctly reproduce the order of magnitude of consumption as well as its main variations, while exhibiting greater dispersion during episodes of high demand. This behavior suggests that, although meteorological conditions contain substantial predictive information, certain extreme demand episodes remain difficult to anticipate using these variables alone.

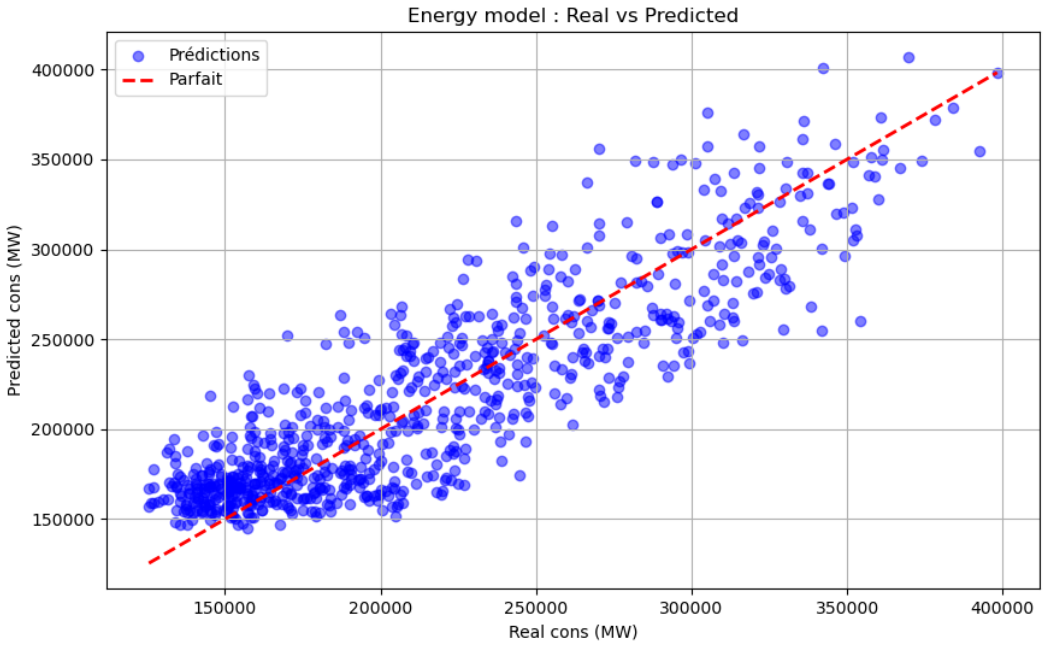

Figure 7: Forecast of energy consumption: observed Vs predicted values (test sample)

This figure illustrates the relationship between observed and predicted energy consumption values on the test sample. Most observations lie close to the line of perfect prediction, indicating overall coherence between forecasts and realizations. The largest deviations appear at the highest consumption levels, highlighting model limitations during periods of exceptional demand. Overall, these forecasts show that the retained model provides a plausible representation of out-of- sample energy consumption and can be used to anticipate demand levels based on observed meteorological conditions, within a strictly predictive framework.

Comparative Analysis

The energy consumption forecasting exercise illustrates the strengths and limitations of a cross- sectional approach based on exogenous meteorological information. By incorporating contemporaneous weather variables, the model provides a coherent and economically interpretable anticipation of average demand levels. This framework performs well under typical conditions, capturing the main drivers of energy consumption related to temperature and humidity. However, its predictive accuracy weakens during episodes of exceptional consumption, which may be driven by extreme weather events, behavioral responses, or structural factors not fully captured by meteorological variables. These results underline the limits of static cross-sectional models when applied to rare or atypical situations.

3.8 Robustness analysis and model limitations

Robustness of the results

The robustness of energy consumption forecasts is assessed by estimating several alternative specifications, including a linear OLS model, non-linear extensions incorporating quadratic and interaction terms, and regularized models such as Ridge and Lasso. Across these specifications, out- of-sample predictive performance remains broadly similar, as measured by RMSE and MAE. This stability indicates that the forecasts are not overly sensitive to the specific estimation method, provided that the model structure is consistent with the underlying data-generating process. In particular, the inclusion of non-linear temperature effects appears more important for predictive accuracy than the choice between OLS and penalized estimators.

Model and data limitations

Despite these encouraging results, several limitations must be acknowledged. First, energy consumption is explained exclusively by meteorological variables, whereas demand is also influenced by economic activity, institutional settings, prices, and behavioral factors that are not observed in the dataset. This omission may reduce predictive accuracy during episodes of exceptional consumption. Second, the cross-sectional framework does not explicitly model the temporal dynamics of energy demand, thereby excluding potential intertemporal dependence that could improve forecasts in some contexts. Finally, results depend on data quality and aggregation choices, which may mask local heterogeneity or short-term shocks.

Implications for interpretation

As a consequence, energy consumption forecasts should be interpreted as indicative predictions capturing average demand responses to weather conditions rather than exact forecasts. They are best suited for anticipating general demand levels and relative variations, particularly under typical meteorological conditions.

4 Conclusion

This project aimed to explore and compare predictive methods applied to two distinct settings: forecasting energy consumption based on meteorological variables and forecasting temperature using time series models. By combining a cross-sectional and a temporal approach, the analysis highlights the importance of aligning data structure, methodological choices, and predictive objectives.

The results show that meteorological variables provide relevant information for anticipating energy consumption. A relatively simple regression model enriched with non-

linearities generates coherent out-of- sample forecasts and captures the main variations in energy demand. However, some extreme situations remain difficult to predict, underscoring the limitations inherent in models relying solely on observable climatic factors.

In parallel, temperature analysis reveals a strongly structured temporal dynamics dominated by stable annual seasonality. The retained time series models effectively reproduce this structure and provide plausible short- and medium-term forecasts. Explicit treatment of uncertainty through prediction intervals nonetheless highlights the declining precision of forecasts as the horizon increases.

Beyond specific results, this project emphasizes the importance of a rigorous approach based on out-of-sample evaluation, a clear separation between estimation, model selection, and forecasting, and explicit acknowledgment of model limitations. It also shows that relatively simple models, when properly specified and used within a predictive framework, can yield informative and operational results.

Finally, the project highlights the complementary nature of cross-sectional and time series approaches. Their joint use allows forecasting strategies to be adapted to the nature of the phenomenon under study, while reminding that all forecasts must be interpreted with caution and within the assumptions underlying the models employed.

References

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

[2]
- **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 (%).

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

[4]

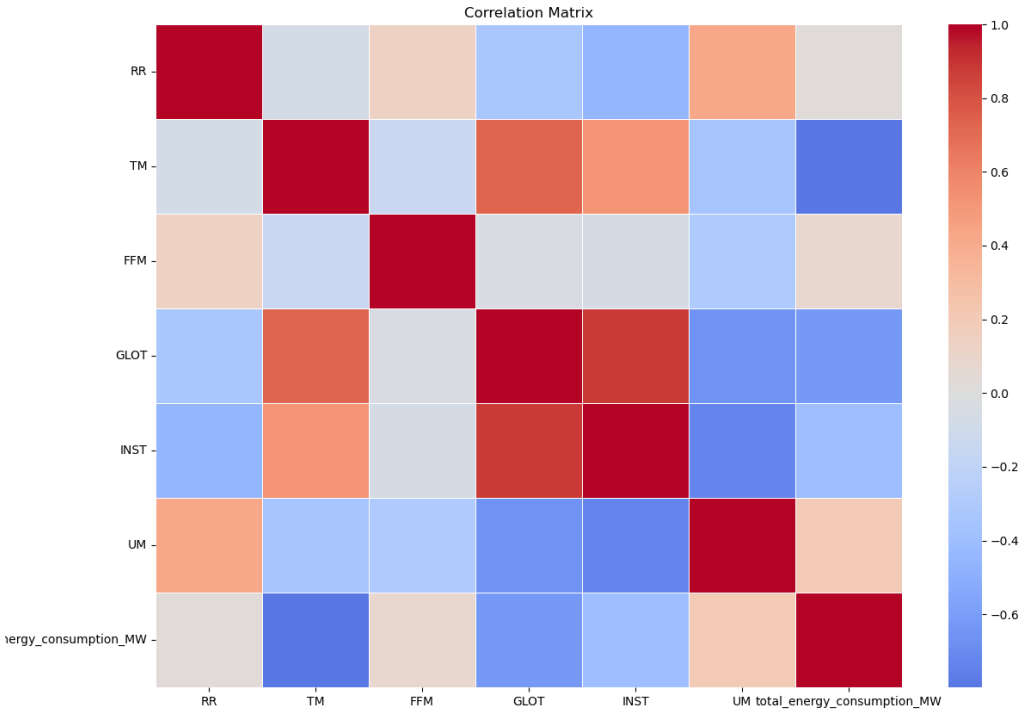

Figure 8: Correlation Matrix

[5] SDES. *Bilan énergétique de la France | Chiffres clés de l'énergie - Édition 2025.*

[6] Auffhammer, M., Baylis, P., & Hausman, C. (2017). *Climate change is projected to have severe impacts on the frequency and intensity of peak electricity demand across the United States. Proceedings of the National Academy of Sciences.*

[7] Bessec, M., & Fouquau, J. (2008). *The non-linear link between electricity consumption and temperature in Europe.* Energy Economics.

[8] Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time Series Analysis: Forecasting and Control. Wiley.

[9] Deschênes, O., & Greenstone, M. (2011). *Climate change, mortality, and adaptation: Evidence from annual fluctuations in weather in the US.* American Economic Journal: Applied Economics.

[10] Hong, T., Pinson, P., & Fan, S. (2016). *Global energy forecasting competition 2012. International Journal of Forecasting.*