Time Series Project

Time Series Analysis Report

Spatio-Temporal Modelling of Urban and Rural Air Pollution in Occitanie

---

Students group 9:

Baptiste Audrain
Clément Yvernault - Collet
Antoine Montzamir

Tuteur :

Youssef Esstafa

January 16, 2026

Contents

1 Introduction

Air quality has emerged as a critical public health concern, with atmospheric pollutants such as Ozone (O_3), Nitrogen Dioxide (NO_2), and Fine Particles ($PM_{2.5}$) posing significant risks to respiratory and cardiovascular health. Consequently, the ability to not only monitor but effectively *forecast* pollution episodes is paramount for local authorities to implement timely mitigation strategies.

The primary goal of this project is to establish a robust predictive system capable of generating alerts and forecasts regarding air quality across the Occitanie region. This system aims to provide stakeholders with actionable insights based on historical and real-time data.

Modeling air pollution is inherently complex due to the stochastic nature of atmospheric dynamics. Pollutant concentrations exhibit strong seasonality (daily and annual cycles), dependence on meteorological conditions, and phenomenon known as "volatility clustering," where periods of high variance tend to cluster together. To address these challenges, our approach relies on a rigorous Time Series analysis framework:

1. **Typological Segmentation:** Recognizing that pollution dynamics differ significantly by environment, we categorized stations into three distinct typologies: *Urban*, *Peri-urban*, and *Rural*. A "Reference Station" approach was adopted to define the optimal model architecture for each typology/pollutant pair.

2. **Hybrid Modeling (SARIMA-GARCH):** We employ Seasonal Auto-Regressive Integrated Moving Average (SARIMA) models to capture the linear trends and seasonality. Furthermore, to account for heteroskedasticity (varying volatility), we coupled these with Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) models. This hybrid approach allows for dynamic confidence intervals, ensuring that high-risk events are captured within the uncertainty bounds.

3. **Operational Deployment:** The theoretical models are operationalized through an interactive **RShiny dashboard**. This tool functions as a real-time monitoring system, performing 48-hour rolling forecasts and translating statistical probabilities into intelligible alert levels for decision-makers.

The following sections detail the exploratory data analysis, the mathematical derivation of the models, and the final implementation of the forecasting interface.

2 Justification of the Approach

This section justifies the choices made regarding the grouping of station typologies and the selection of only three pollutants among those available.

2.1 Data quality

The data used in this project originates from the official "Atmo Occitanie" open data portal, accessible via the Picto-Occitanie catalogue. We specifically selected the "1-year rolling" (1 an glissant) dataset, which provides validated, high-frequency measurements updated continuously.

The dataset can be downloaded directly via this link: **Download Data (Occitanie)**.

The dataset covers the entire Occitanie region (France). For the purpose of this study, the observation window extends from August 23, 2022, to January 6, 2026. The data is aggregated on an hourly basis, providing the granularity necessary to capture diurnal pollution cycles and train our SARIMA-GARCH models.

The raw dataframe consists of geolocated time series for various pollutants. Table 1 details the key variables used for the analysis.

Table 1: Description of the Variables in the Dataset

Variable	**Description**	**Example**
`nom_station`	Name of the monitoring station	*Toulouse-Berthelot*
`typologie`	Environment type (used for segmentation)	*Urbaine*
`nom_polluant`	Chemical species measured	*NO2, O3, PM2.5*
`valeur`	Concentration level (Target Variable)	*2.6*
`unite`	Measurement unit	*ug.m-3* ($\mu g/m^3$)
`date_fin`	Timestamp of the measurement (hourly)	*2025-03-11 08:00:00*
`x_wgs84 / y_wgs84`	GPS Coordinates (Longitude/Latitude)	*1.444 / 43.587*
`statut_valid`	Data validation status ('t' = true/valid)	*t*

2.2 Station Grouping Strategy

To ensure statistical robustness and simplify the model deployment, we aggregated the 5 initial administrative typologies into 3 functional classes (Urban, Peri-urban, Rural).

Table 2: Methodological grouping of the 5 station typologies

Target Typology (3)	**Initial Typologies (5)**
Urban	Urban
Peri-urban	Peri-urban Rural near Urban Area
Rural	Regional Rural National Rural

2.2.1 Justification for merging the "Rural Regional" and "Rural National" zones

For each pollutant ($PM_{2.5}$, O_3, NO_2), the hourly series measured in the "Rural Regional" and "Rural National" zones were aligned on a common time interval and then compared point by point. Over this interval, ozone (O_3) levels are remarkably similar between the two zones: the means are almost identical (64.8 $\mu g/m^3$ for the regional zone versus 63.9 $\mu g/m^3$ for the national zone), the standard deviations are of the same order (20.4 vs 19.5 $\mu g/m^3$), and medians and quartiles largely overlap, indicating very similar concentration distributions. The correlation between the two O_3 series is high (r = 0.82), highlighting strongly synchronous temporal variations between the two typologies, with ozone peaks and troughs occurring almost at the same time in both zones.

Analogous analyses performed for the other pollutants ($PM_{2.5}$ and NO_2) also reveal mean levels and variabilities of the same order of magnitude in the "Rural Regional" and "Rural National" zones, as well as positive correlations between the time series, which suggests globally coherent temporal profiles across these two typologies. From a time-series modelling perspective, and in line with common practice in the literature that aggregates sites with similar levels and temporal structures, these results support the decision to merge the "Rural Regional" and "Rural National" zones into a single rural analysis zone.

2.2.2 Justification for merging the "Peri-urban" and "Rural Near-Urban" zones

For ozone (O_3), the hourly series from the "Peri-urban" and "Rural Near-Urban" zones were aligned on a common time interval and compared point by point. Over this interval, the mean levels are very close in both zones (61.8 $\mu g/m^3$ for the peri-urban zone versus 59.8 $\mu g/m^3$ for the rural near-urban zone), with standard deviations of the same order (25.2 vs 30.4 $\mu g/m^3$). The medians (63.4 vs 61.2 $\mu g/m^3$) and quartiles (around 45–78 $\mu g/m^3$ for the peri-urban zone and 39–80 $\mu g/m^3$ for the rural near-urban zone) largely overlap, indicating very similar concentration distributions between the two typologies. The correlation between the two O_3 series is moderately high (r = 0.54), reflecting a clear temporal co-variation: major ozone variations (pronounced peaks and drops) tend to occur simultaneously in both zones.

These results show that both the levels and temporal profiles of ozone are broadly comparable between the "Peri-urban" and "Rural Near-Urban" zones. In a time-series modelling perspective, and consistent with approaches in the literature that aggregate sites with similar distributions and dynamics, this evidence justifies merging these two typologies into a single analysis zone.

2.3 Pollutant selection and thresholds

The selection of pollutants for this study focuses on NO_2, O_3, and $PM_{2.5}$ among all available species (NO_2, NO, NO_x, $PM_{2.5}$, PM_{10}, O_3, SO_2, H_2S). According to WHO air quality guidelines and recent European assessments, fine particles $PM_{2.5}$, ozone (O_3), and nitrogen dioxide (NO_2) are key health-relevant pollutants because of their well-documented effects on respiratory and cardiovascular mortality and morbidity [2]. Focusing on these species allows us to target the main public health issues while keeping the model complexity manageable.

In the case of nitrogen pollutants, NO_2 was retained as the representative variable, which led to the exclusion of NO and NO_x. These three indicators are strongly correlated and essentially describe the same emission sources (traffic, combustion), so including them simultaneously would introduce substantial redundancy without providing additional insight. Similarly, $PM_{2.5}$ was preferred over PM_{10}, as it represents the finest particle fraction, generally considered the most toxic, and is closely related to PM_{10} levels. Using $PM_{2.5}$ therefore captures most of the particulate variability with a more health-relevant indicator, while avoiding duplication between two highly similar series.

Finally, SO_2 and H_2S were measured at only a very small number of stations (two sites over the entire dataset). Such a limited spatial coverage makes robust comparisons between

geographical zones difficult and strongly limits the generalisability of the results. Including these series would have introduced a strong imbalance in the database and a high risk of non-generalizable conclusions. For this reason, SO_2 and H_2S were excluded from the analysis in favour of NO_2, O_3, and $PM_{2.5}$, for which the time series are more complete and better suited to a detailed spatio-temporal study.

Air quality standards define, for each pollutant, limit values for ambient concentrations that should not be exceeded in order to protect human health and the environment. These thresholds are based on scientific expertise that combines evidence from epidemiological and toxicological studies, and are set by European and national institutions within a harmonised regulatory framework [1]. The table below summarises the main regulatory concentration limit values for the pollutants considered here, as defined by WHO guidelines and European and French regulations, and serves as a reference for interpreting the levels measured in the different study zones.

Table 3: Regulatory Air Quality Thresholds in France

Pollutant	**Threshold**
Ozone (O_3)	$120\,\mu g/m^3$
Nitrogen Dioxide (NO_2)	$200\,\mu g/m^3$
Fine Particles ($PM_{2.5}$)	$25\,\mu g/m^3$

Source: Airparif - La réglementation en France

3 Modelling

For each geographical cluster and each pollutant, a reference station was selected based on data quality criteria (number of available observations, low proportion of missing values). The time-series model was first calibrated on this reference station and then applied without structural modification to the other stations belonging to the same cluster.

The hourly series from the different stations within a given cluster exhibit similar profiles, with mean levels and variabilities of the same order of magnitude, as well as positive correlations between series, which suggests a common underlying dynamics within each group. In addition, model performance metrics (for example, RMSE and MAE) remain comparable between the reference station and the other stations in the cluster, indicating that the temporal structure learned on the reference site is also suitable for the remaining stations. Taken together, these elements support the assumption of relative intra-group homogeneity and justify using a representative station to calibrate the model and then applying this model to all stations in the corresponding cluster.

The general modeling approach follows a standard Train/Validation/Test split. Performance evaluation is conducted using a **rolling window over 48 hours**, as the ultimate goal is to obtain pollutant concentration forecasts for the next 48 hours.

3.1 Model 1: Rural / $PM_{2.5}$

3.1.1 Construction of mean series and similarity tests

In order to obtain reference series representative of the average behavior of pollutants in rural areas, we constructed, for each pollutant, a mean time series based on the available hourly observations across all rural municipalities.

Definition of mean series Let $y_{i,t}$ denote the concentration observed for municipality i at time t. The mean series associated with a given pollutant is defined as:

$$\bar{y}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} y_{i,t},$$

where N_t denotes the number of rural municipalities with a valid observation at time t. This hourly aggregation smooths local fluctuations while preserving global dynamics, such as seasonality and underlying trends specific to the pollutant under consideration.

Objective of the comparisons The resulting mean series are used as references for the analysis of temporal dynamics and subsequent modeling. To assess their relevance, each local series was compared to the corresponding mean series using several statistical indicators and similarity tests.

Indicators and tests The comparisons are based on the following criteria:

- **Linear correlation**: measures the overall similarity of temporal dynamics.
- **RMSE**: quantifies the average deviation between the local series and the mean series.
- **Kolmogorov–Smirnov test**: assesses the similarity of marginal distributions.
- **Maximum cross-correlation (CCF)**: detects potential temporal shifts.
- **Long-memory parameter** d: compares the temporal dependence properties of the local series (d_{local}) and the mean series (d_{mean}).

The absolute difference $|d_{\text{local}} - d_{\text{mean}}|$ provides a synthetic indicator of the consistency of long-memory properties between the two series.

Comparison results The results of the similarity tests are reported in Table 4.

Table 4: Comparison between local series and mean series by pollutant

Municipality	**Pollutant**	n_{obs}	Corr.	RMSE	KS p-value	CCF max	d_{local}	d_{mean}	$\lvert d_l - d_m \rvert$
BELESTA-EN-LAURAGAIS	O_3	8314	0.928	8.76	1.0×10^{-12}	0.928	0.495	0.497	0.002
BOLQUÈRE	$PM_{2.5}$	8736	0.884	2.88	0	0.884	0.186	0.396	0.210
CORNEILHAN	O_3	8674	0.872	11.5	2.2×10^{-16}	0.872	0.462	0.497	0.035
GAUDONVILLE	O_3	8772	0.947	6.93	7.6×10^{-3}	0.947	0.500	0.496	0.004
PEYRUSSE-VIEILLE	NO_2	8324	1.000	0.00	1	1.000	0.364	0.364	0.000
PEYRUSSE-VIEILLE	O_3	8642	0.881	10.1	5.1×10^{-2}	0.881	0.497	0.496	0.001
PEYRUSSE-VIEILLE	$PM_{2.5}$	8646	0.807	2.89	0	0.807	0.491	7.8×10^{-5}	0.491

Interpretation The results indicate an overall high level of consistency between local series and mean series for all pollutants considered. Correlations are systematically strong, and the deviations measured by the RMSE remain moderate, indicating that the mean series accurately capture the dominant temporal dynamics observed in rural areas.

The Kolmogorov–Smirnov tests and the cross-correlation analysis do not reveal major structural differences nor significant temporal shifts between local and mean series. Furthermore, the long-memory parameters d are, in most cases, close between local and mean series, confirming the consistency of temporal dependence properties.

Although some local differences remain, particularly for $PM_{2.5}$ where greater heterogeneity is observed, these discrepancies do not undermine the use of mean series as global references. They thus represent a relevant compromise between spatial representativeness and statistical stability for all pollutants studied.

Implications for modeling Given these results, the mean series are retained as reference series for the modeling of all pollutants considered. This approach relies on robust and generalizable dynamics, while facilitating comparisons across pollutants and the implementation of common models for forecasting and anomaly detection.

3.1.2 Preliminary analysis

This section presents a first descriptive summary of the $PM_{2.5}$ series in rural areas. Table 5 summarizes the main indicators: sample size, central tendency, dispersion, and higher-order moments.

Table 5: Descriptive statistics of the $PM_{2.5}$ series

n	Min	$Q_{5\%}$	Median	Mean	$Q_{95\%}$	Max	Std. dev.	Variance	Skewness	Kurtosis
17382	0.0	0.7	4.3	5.699	15.1	129.8	5.287	27.947	4.060	47.709

Temporal dependence structure ($PM_{2.5}$) Figure 1 displays the autocorrelation (ACF) and partial autocorrelation (PACF) functions of the $PM_{2.5}$ series in rural areas. The ACF decreases slowly and almost monotonically over a large number of lags, indicating strong temporal persistence and suggesting the presence of an integrated or long-memory component. In contrast, the PACF exhibits a pronounced spike at the first lag, followed by non-significant values at higher lags.

This configuration is typical of a low-order autoregressive process, potentially combined with differencing or a fractional memory parameter. It motivates the use of ARIMA or ARFIMA models to describe the dynamics of $PM_{2.5}$, rather than high-order purely autoregressive models.

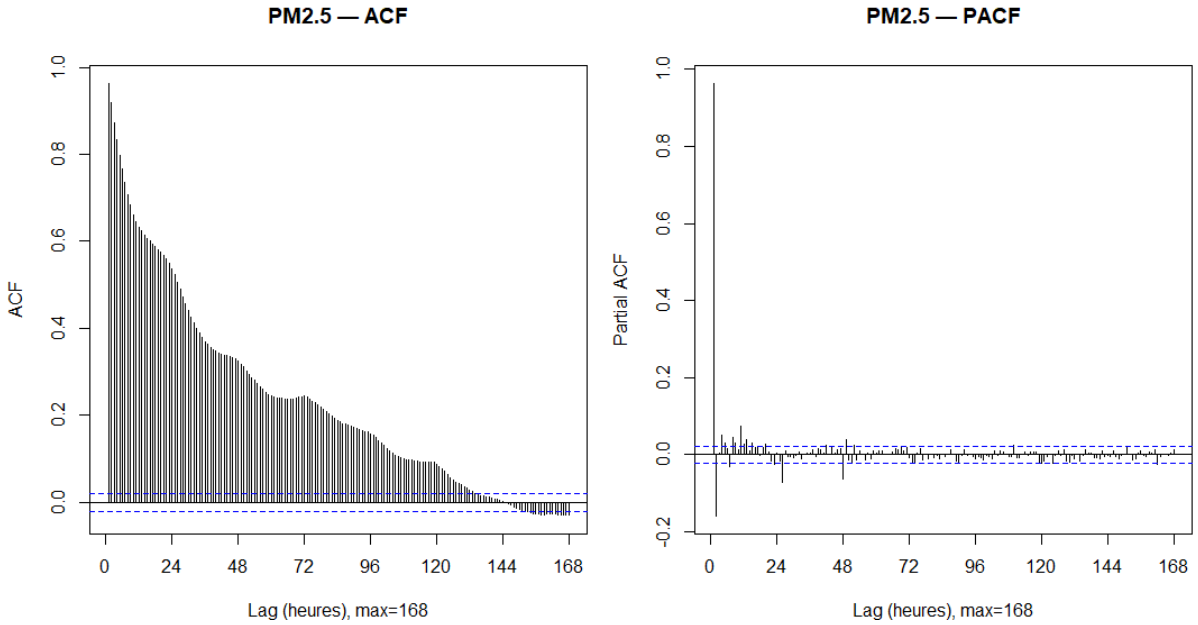

Figure 1: Autocorrelation (ACF) and partial autocorrelation (PACF) functions of the $PM_{2.5}$ series (hourly lags up to 168).

3.1.3 Model selection

Once these preliminary analyses are completed, we proceed to the choice of the model to be specified. Three types of models are considered (SARIMA, FARIMA, and ARMA), motivated by their ability to capture seasonal dynamics, long-memory behavior, and short-run dependence, respectively. Model performance is evaluated using validation metrics such as the RMSE in order to select the most appropriate specification. The corresponding results are reported below.

SARIMA model selection For the $PM_{2.5}$ series, several SARIMA specifications were evaluated on the training sample and compared using the validation RMSE. Although the Akaike Information Criterion (AIC) is also reported as an in-sample goodness-of-fit measure, the final model choice is primarily driven by validation performance, as the objective is anomaly detection on future observations. The retained model therefore corresponds to the lowest validation RMSE, while maintaining parsimony and numerical stability.

For $PM_{2.5}$, models including a first-order non-seasonal differencing ($d = 1$) consistently outperform models without differencing, suggesting a higher degree of persistence in the series. The SARIMA$(1, 1, 2)(1, 0, 1)_{24}$ model yields the lowest validation RMSE and is thus retained.

Table 6: Candidate SARIMA models for $PM_{2.5}$

(p, d, q)	(P, D, Q)	AIC	RMSE_{val}
(1,1,2)	**(1,0,1)**	**18822**	**4.31**
(2,1,1)	(1,0,1)	18821	4.34
(2,0,1)	(1,0,1)	18840	4.42
(1,0,2)	(1,0,1)	18833	4.45
(1,0,1)	(1,0,1)	18848	4.45
(1,0,1)	(1,1,1)	18888	4.46

In addition to SARIMA models, two alternative approaches were considered for modeling

the mean dynamics: (i) ARFIMA models, which allow for long-memory dynamics through fractional differencing, and (ii) ARMA models, which capture short-run dependence without explicit seasonal structure. Both families were evaluated using the validation RMSE, while the AIC is reported as an in-sample criterion. For ARMA models, a small grid search over (p, q) values was performed, and the specification minimizing the validation RMSE was retained (ties were broken using the AIC).

FARIMA Table 7 summarizes the ARFIMA fits obtained using `forecast::arfima` on the training sample, along with the estimated fractional differencing parameter d and the validation RMSE.

Table 7: ARFIMA results (training fit) and validation performance

Pollutant	$\hat{d}$	$\mathbf{RMSE_{val}}$	**Comment**
$PM_{2.5}$	0.1910	4.3437	Close to SARIMA

ARMA Table 8 reports the grid-search results for ARMA models applied to the $PM_{2.5}$ series. The selected ARMA specification is highlighted in bold.

Table 8: Candidate ARMA models for $PM_{2.5}$

p	q	AIC	$RMSE_{val}$
1	**2**	**18847**	**4.44**
1	1	18845	4.45
2	1	18847	4.45
2	2	18847	4.45
2	0	18855	4.46
0	2	25291	4.52
0	1	29245	4.52
0	0	35903	4.53
1	0	19159	5.54

3.1.4 Interpretation

Overall, ARMA models provide a reasonable non-seasonal baseline for $PM_{2.5}$, but they tend to underperform compared to seasonal specifications in terms of out-of-sample accuracy. ARFIMA models capture long-memory behavior through fractional differencing, with the estimated value of d indicating moderate long-range dependence. However, in terms of validation RMSE, the SARIMA model provides the best overall performance and is therefore retained as the reference mean model for trend analysis and anomaly detection.

Moreover, the distribution of $PM_{2.5}$ concentrations exhibits strong right skewness and heavy tails, which is consistent with the occurrence of episodic pollution peaks above the usual background level. These characteristics justify the use of robust exploratory tools and the explicit consideration of conditional volatility in subsequent modeling steps.

3.2 Model 2: Peri-urban / Ozone (O_3)

The modeling strategy for Ozone (O_3) in peri-urban areas relies on a "Reference Station" approach. We selected the Montgiscard station as the champion for this typology due to its data quality and representativeness. The model architecture defined on this station is subsequently applied to all other peri-urban stations, with coefficient re-estimation for each specific location.

3.2.1 Stationarity Analysis and Differentiation

The initial exploratory analysis focused on assessing the stationarity of the Ozone concentration time series, a prerequisite for ARIMA modeling.

First, a Seasonal-Trend Decomposition using Loess (STL) was performed to isolate the signal components. As illustrated in Figure 2, the decomposition reveals a distinct and high-amplitude seasonal component. The strictly repetitive sinusoidal pattern observed in the "Seasonality" panel confirms the presence of a strong diurnal cycle (24-hour periodicity), driven by the photochemical nature of Ozone formation which is highly dependent on solar radiation and traffic patterns.

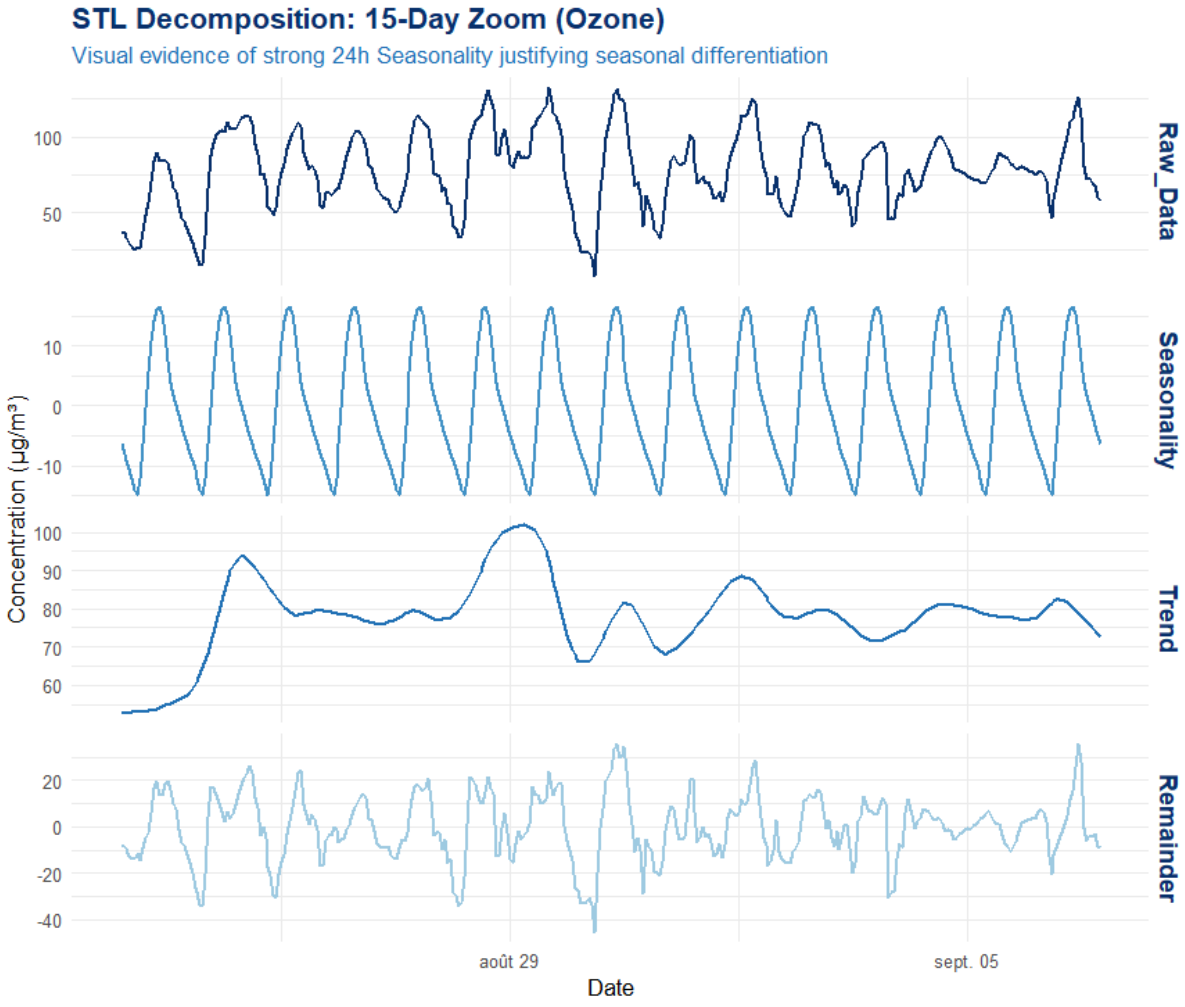

Figure 2: STL Decomposition of Ozone Concentrations (Montgiscard): Evidence of Strong Diurnal Seasonality.

Furthermore, the analysis of the raw time series and its correlograms (Figure 3) corroborates the non-stationarity of the process. The raw series exhibits clear periodic behavior rather than

a constant mean. This is statistically confirmed by the Autocorrelation Function (ACF), which displays a slowly decaying sinusoidal wave pattern rather than a rapid drop to zero. This persistence of significant autocorrelation at high lags is a hallmark of non-stationary seasonal data, necessitating differentiation to stabilize the mean.

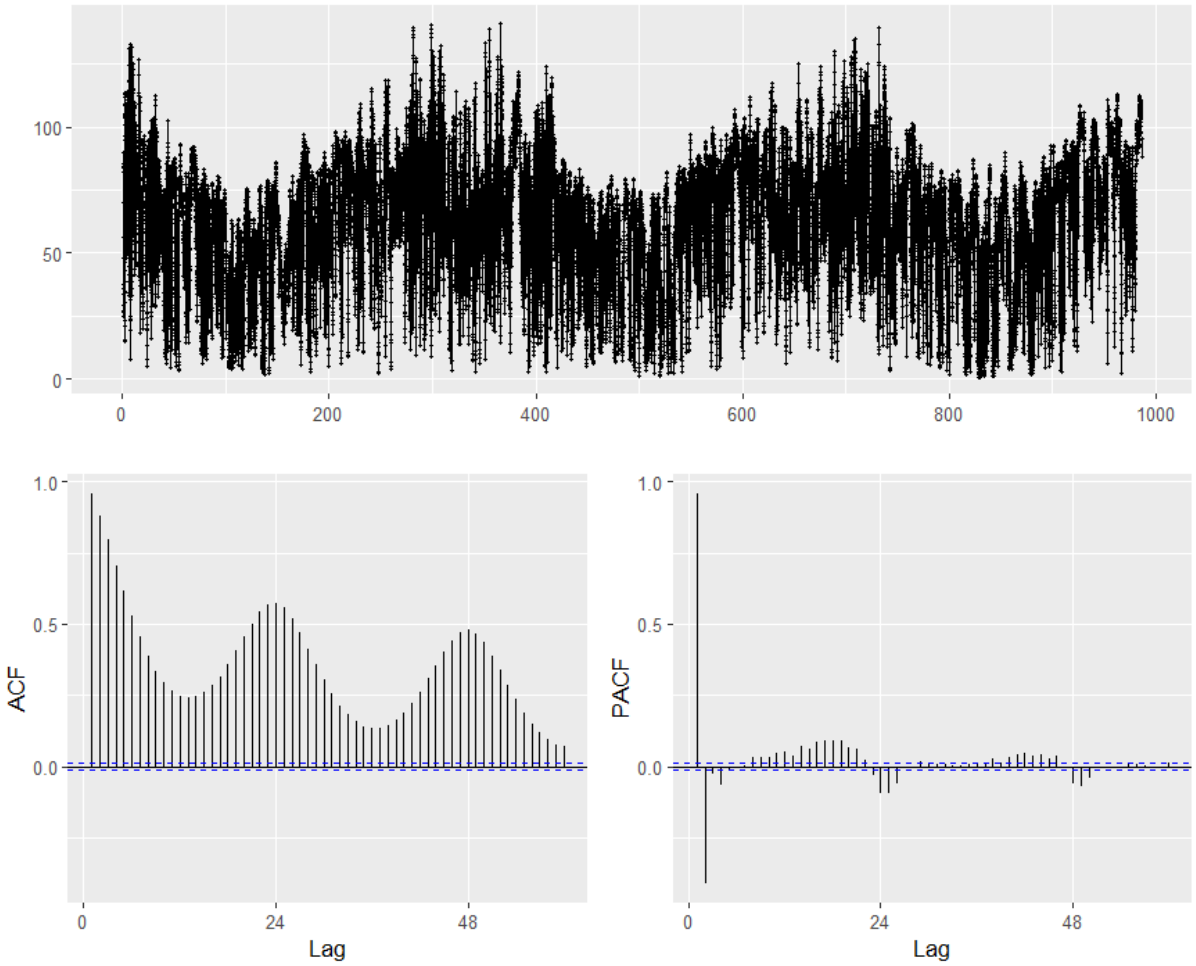

Figure 3: Raw Ozone Time Series Analysis: Signal (Top), ACF (Bottom Left), and PACF (Bottom Right) highlighting Non-Stationarity.

3.2.2 Differentiation Strategy and Stationarity Achievement

Given the non-stationary nature of the raw Ozone series (strong seasonality and local trends), a transformation process was necessary to meet the assumptions of the SARIMA framework.

Step 1: Seasonal Differentiation ($D = 1, s = 24$) The first step aimed to remove the dominant 24-hour cycle identified in the STL decomposition. We applied a seasonal difference operator $\nabla_{24} y_t = y_t - y_{t-24}$.

As shown in Figure 4, while this transformation successfully removed the sinusoidal wave pattern, the resulting series still exhibits non-stationary behavior. The ACF (bottom-left) shows a slow decay rather than a sharp cut-off, and the signal (top) retains visible local trends and variance instability. This indicates that a simple seasonal adjustment is insufficient to achieve stationarity.

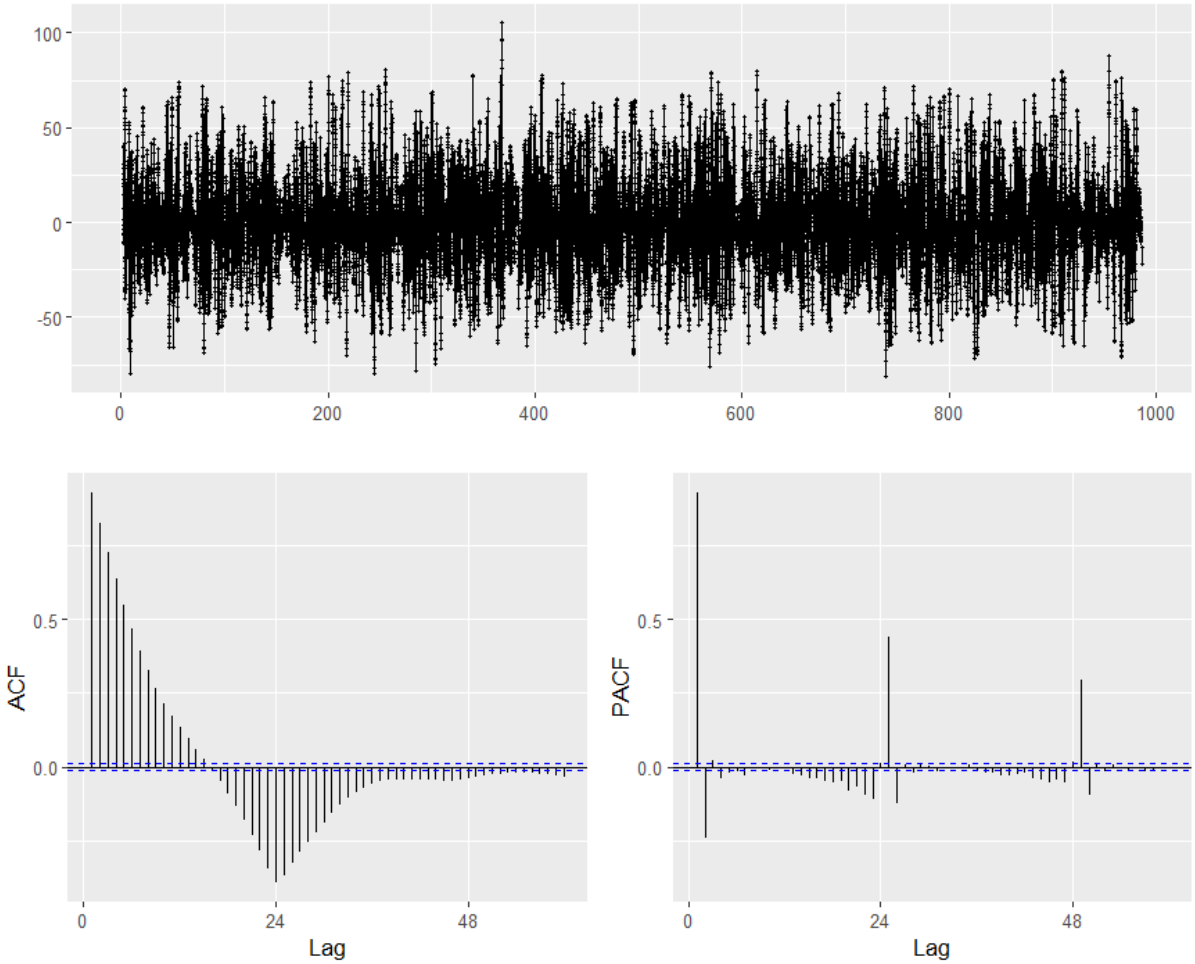

Figure 4: Signal, ACF, and PACF after Seasonal Differentiation ($D = 1$). The slow decay in ACF indicates persistent non-stationarity.

Step 2: Double Differentiation ($D = 1, d = 1$) To address the remaining non-stationarity, we applied a first-order ordinary difference operator to the seasonally adjusted series: $\nabla\nabla_{24} y_t = (y_t - y_{t-24}) - (y_{t-1} - y_{t-25})$. Figure 5 demonstrates the effectiveness of this double differentiation strategy. The time series (top panel) now resembles white noise with a constant mean centered around zero. Crucially, the ACF and PACF plots show a rapid cut-off after lag 1, with no significant recurring patterns at higher lags. This confirms that the series has achieved weak stationarity.

Furthermore, the significant spikes observed at lag 1 in both ACF and PACF suggest the presence of first-order Auto-Regressive (AR) and Moving Average (MA) components, guiding the subsequent selection of candidate SARIMA models.

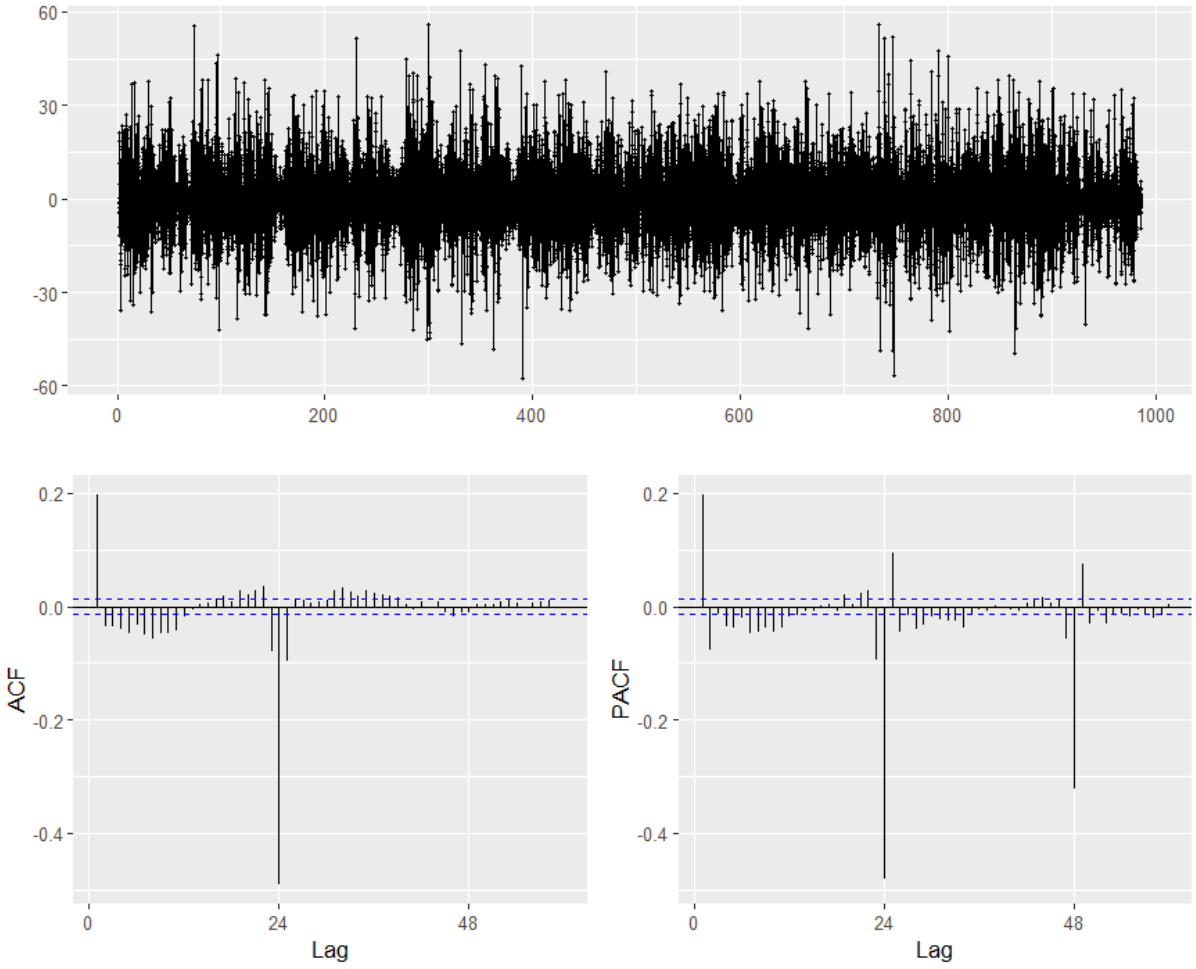

Figure 5: Signal, ACF, and PACF after Double Differentiation ($D = 1, d = 1$). The series is now stationary, ready for model identification.

3.2.3 SARIMA Model Selection

Following the identification of the differentiation order ($d = 1, D = 1$), the analysis of the Partial Autocorrelation Function (PACF) at lag 1 suggested a potential Auto-Regressive process, while the ACF spike suggested a Moving Average process. To rigorously determine the optimal non-seasonal structure (p, q), three candidate models were competed on a held-out validation set (representing 10% of the timeline):

- **MA(1):** $SARIMA(0, 1, 1)(0, 1, 1)_{24}$
- **AR(1):** $SARIMA(1, 1, 0)(0, 1, 1)_{24}$
- **Mixed (ARMA):** $SARIMA(1, 1, 1)(0, 1, 1)_{24}$

The selection process was driven by a "Smart Winner" strategy, which prioritizes predictive accuracy (RMSE) over theoretical parsimony (AIC), as the primary goal of our system is operational forecasting. Table 9 summarizes the performance metrics obtained for each candidate.

Table 9: Performance Comparison of Candidate Models (Validation Set)

Model Architecture	**RMSE** (Validation)	**AIC** (Training)	**Rank (Smart Winner)**
MA(1)	63.45	**152,911.6**	2
AR(1)	**62.17**	153,019.6	**1**
Mixed	63.74	152,912.3	3

Decision Analysis: As observed in Table 9, there is a divergence between the metrics:

- The **AIC criterion** slightly favors the MA(1) model ($152,911$ vs $153,019$), suggesting it offers a slightly better fit-to-complexity ratio on the training data.
- However, the **RMSE metric**, which measures the actual forecast error on unseen data, clearly favors the AR(1) model ($62.17\,\mu g/m^3$ vs $63.45\,\mu g/m^3$).

Given that the drop in RMSE ($> 1.2\,\mu g/m^3$) represents a tangible improvement in forecast reliability, our algorithm selected the **AR(1)** model. Consequently, the final architecture retained for the peri-urban Ozone model is $**SARIMA(1,1,0)(0,1,1)_{24}**$.

3.2.4 Residual Analysis and Volatility Modeling (GARCH)

Although the SARIMA model successfully captures the linear dynamics of the Ozone series (mean equation), a rigorous diagnostic of the residuals (ϵ_t) is required to validate the assumption of constant variance (homoskedasticity).

Visual and Statistical Diagnosis Figure 6 presents the diagnostic plots of the SARIMA residuals. A careful inspection reveals distinct characteristics:

- **ACF (Bottom Left):** Contrary to an ideal white noise process, the autocorrelation function displays several significant spikes. This indicates that the SARIMA architecture, despite its complexity, does not fully capture all linear temporal dependencies. This is a common limitation when modeling high-frequency pollution data with complex exogenous drivers.
- **Histogram (Bottom Right):** The distribution of residuals is clearly non-normal. It exhibits a high peak at zero (leptokurtosis) and thinner tails compared to the Gaussian curve (orange line). This peaked distribution confirms that a standard homoskedastic assumption is ill-suited.
- **Time Series (Top):** The residual plot shows "clusters" of volatility (alternating calm and agitated periods).

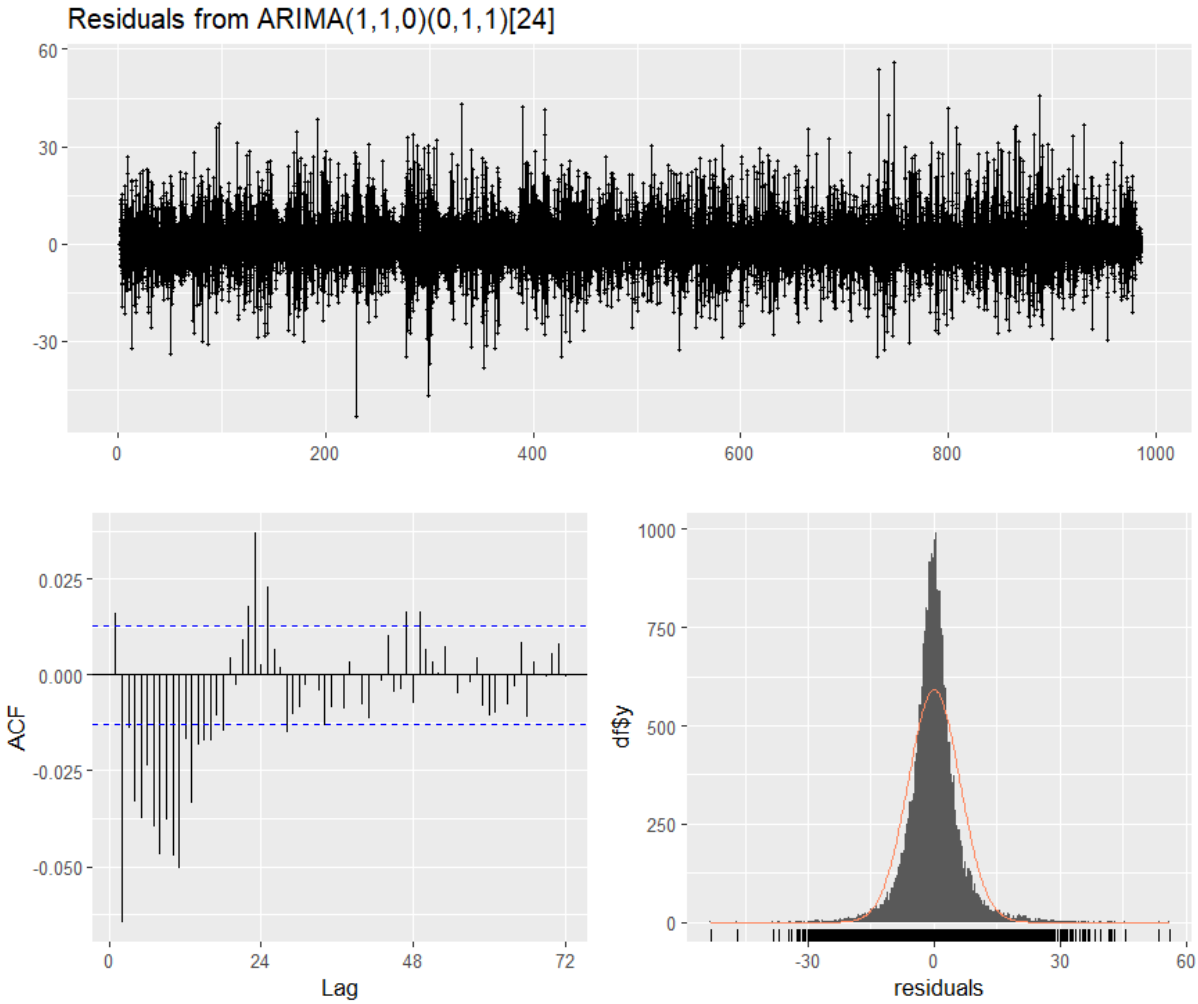

Figure 6: Diagnostic of SARIMA Residuals: Time Series (Top), ACF (Bottom Left), and Histogram vs Normal Distribution (Bottom Right).

It is important to note that the remaining autocorrelation in the ACF (mean equation limitation) does not in itself justify a GARCH model. However, the visual evidence of volatility clustering combined with the specific distribution shape suggests that modeling the *variance* is necessary to improve the reliability of confidence intervals.

To statistically substantiate the need for a conditional variance model—independently of the linear correlation issues discussed above—we relied on the **McLeod-Li** test (Ljung-Box test applied to the squared residuals ϵ_t^2). As detailed in Table 10 and illustrated in Figure **??**, the test yields a p-value of practically zero ($p < 2.2e^{-16}$). This result leads to the unequivocal rejection of the null hypothesis of homoskedasticity, thereby confirming the presence of significant ARCH effects: the volatility of Ozone concentrations is time-dependent, which necessitates the integration of a GARCH component.

Table 10: Ljung-Box Test Results on Squared Residuals

Data Source	Q^* **Stat.**	**Lags**	**D.o.f**	**p-value**	**Conclusion**
SARIMA Residuals	571.57	48	46	$< 2.2 \times 10^{-16}$	Reject H_0

Note: H_0 assumes independent residuals (homoskedasticity). Rejection indicates significant volatility clustering.

GARCH(1,1) Specification To model this conditional variance σ_t^2, we coupled the SARIMA model with a Standard GARCH(1,1) process using a normal distribution. The variance equation is defined as:

$$\sigma_t^2 = \omega + \alpha_1 \epsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \tag{1}$$

Where α_1 represents the reaction to recent shocks (ARCH term) and β_1 represents the persistence of volatility (GARCH term). Table 11 presents the estimated coefficients derived from the model fitting.

Table 11: Estimated Coefficients for the GARCH(1,1) Component

Parameter	**Estimate**	**Std. Error**	**t-value**	**Pr**$(>\|t\|)$
ω (Constant)	7.1717	0.6395	11.21	< 0.001
α_1 (ARCH)	0.2907	0.0165	17.61	< 0.001
β_1 (GARCH)	0.5600	0.0278	20.17	< 0.001

Interpretation: All parameters are highly statistically significant ($p \approx 0$). Furthermore, the sum of the ARCH and GARCH coefficients ($\alpha_1 + \beta_1 = 0.29 + 0.56 = 0.85$) is strictly less than 1. This condition ($\alpha_1 + \beta_1 < 1$) ensures that the estimated variance process is stationary and mean-reverting, confirming the stability and robustness of the chosen GARCH model for operational forecasting.

3.2.5 Final Validation: 48h Rolling Forecast

To rigorously assess the operational viability of the selected SARIMA-GARCH hybrid model, we conducted a Rolling Forecast simulation on the test set, mimicking real-world production conditions.

Simulation Protocol : The testing strategy proceeds iteratively:

1. The model is trained on the historical data up to time t.
2. A multi-step forecast is generated for a horizon of $h = 48$ hours (the target window for public alerts).
3. The observed values for this 48h period are then integrated into the historical dataset.
4. The model parameters are re-estimated (refit), and the process repeats for the next window.

Figure 7 illustrates five consecutive forecast windows (approx. 10 days). The dotted vertical lines indicate the "refit" points where the model updates its knowledge.

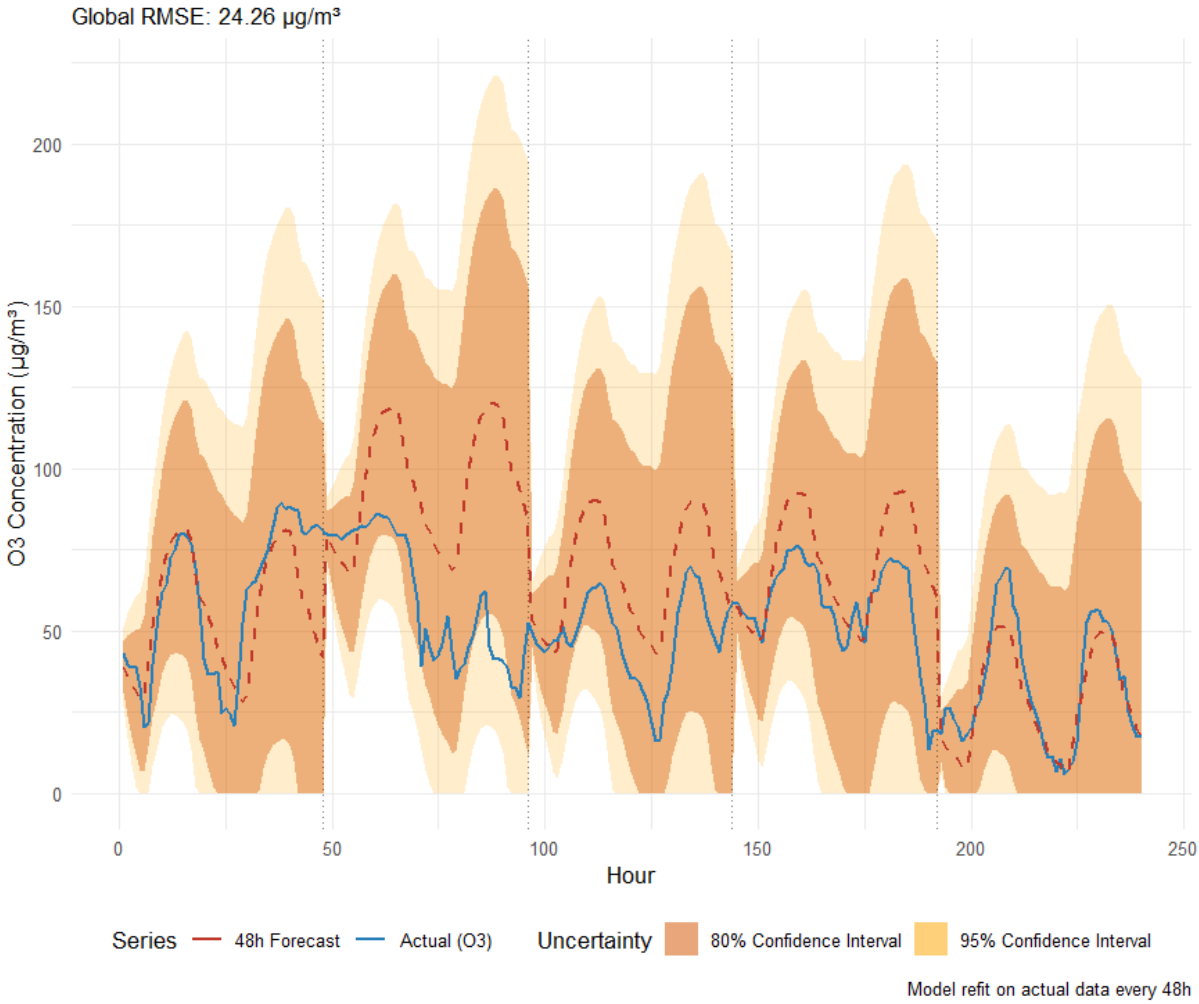

Figure 7: 48h Rolling Forecast Validation. Blue: Actual Observations, Red: SARIMA Forecast, Shaded Areas: Dynamic Confidence Intervals (80% and 95%).

Performance Analysis The quantitative and qualitative evaluation of the results yields the following insights:

- **Global Accuracy (RMSE):** The model achieves a global Root Mean Square Error of $24.26\,\mu g/m^3$. While this error represents a significant portion of the mean concentration, it is largely driven by the difference in signal texture: the model predicts a theoretical, idealized cycle, whereas the actual data contains high-frequency noise.
- **Signal Smoothness vs. Real-World Noise:** A qualitative analysis reveals that the SARIMA component captures the *phase* of the diurnal cycle perfectly (the timing of highs and lows is correct). However, the forecast (red line) is significantly smoother than the observed reality (blue line). The model acts as a filter, projecting a "mean behavior" that ignores the stochastic micro-variations inherent in hourly measurements. In some windows (e.g., hours 50-150), the model even predicts a stronger amplitude than observed, anticipating higher ozone production than what occurred.
- **Reliability and Coverage (Safety):** The most crucial validation for a monitoring system is the coverage of the Confidence Intervals (CI). As visually demonstrated, the actual observations remain entirely contained within the 95% Confidence Interval (light orange

band) throughout the 10-day simulation. This confirms that the GARCH component effectively captures the uncertainty envelope. Even if the point forecast is too smooth or slightly offset in amplitude, the system successfully bounds the risk, which is the primary requirement for reliable safety alerts.

3.3 Model 3: Urban / NO_2

The modelling strategy for nitrogen dioxide (NO_2) in urban areas is also based on a "reference station" approach. An urban station was selected as the champion station for this typology because of the quality of its data (low proportion of missing values, long observation period) and its representativeness of urban background conditions. The model architecture is first defined and calibrated on this reference station, and then the same structure is applied to the other urban stations, with the coefficients re-estimated for each site in order to account for local specificities.

3.3.1 Presentation of the selected model

The model selected for the hourly NO_2 series is a SARIMA(2,1,2)(0,1,1)[24] model coupled with a GARCH(1,1) specification for the conditional variance. The first-order non-seasonal differencing ($d = 1$) and the first-order seasonal differencing ($D = 1$, with a 24 h period) follow directly from the joint analysis of the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the series.

In a first step, only the seasonal differencing of order 1 was applied, leading to the situation illustrated in Figure 8. Visually, the series still exhibits pronounced heteroscedasticity and a strong dependence structure. The ACF decays slowly, with marked spikes at several lags, in particular at lag 24, while the PACF shows significant contributions at the first lags and at some multiples of the daily period. These features indicate that the 24-hour seasonality is not fully removed and that the series is not sufficiently stationary to reliably calibrate a simple ARMA model.

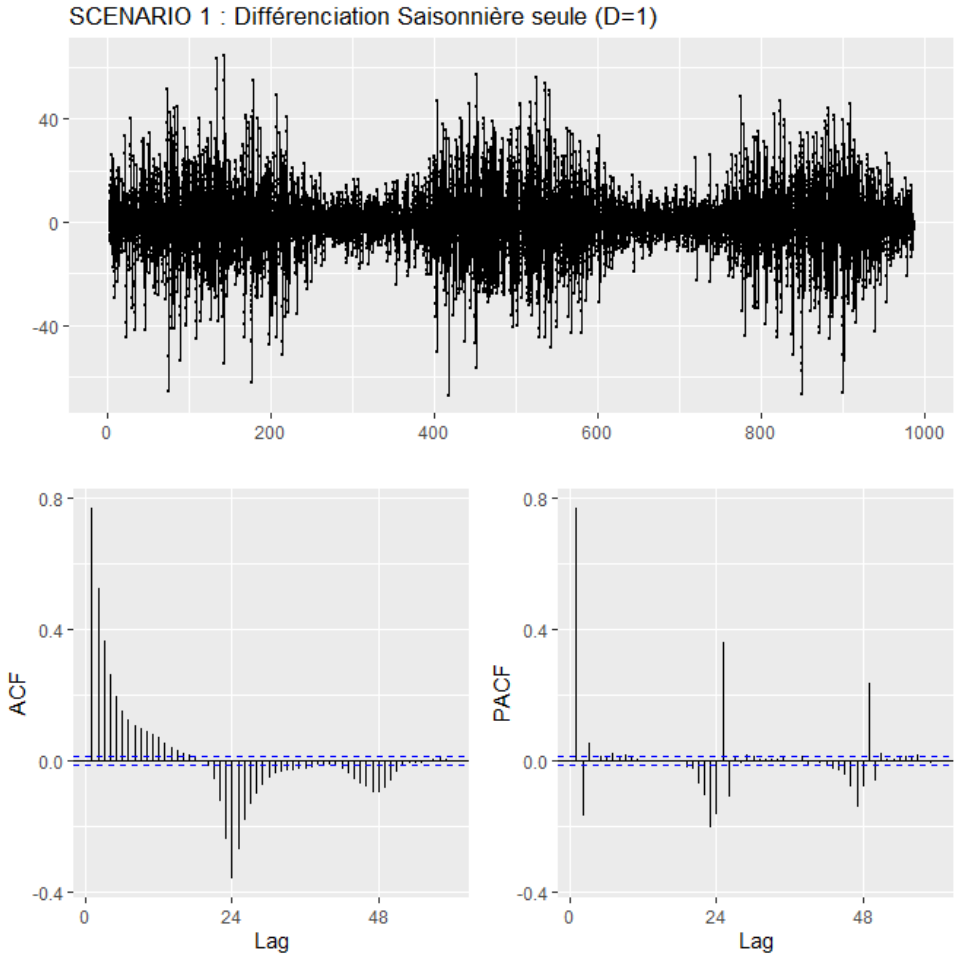

Figure 8: Series of NO_2 after seasonal differencing only (D = 1) and associated ACF/PACF functions.

In a second step, a double differencing was applied, combining a non-seasonal differencing ($d = 1$) and a seasonal differencing ($D = 1$), as illustrated in Figure 9. The resulting series appears more homogeneous, and the dependence structure is substantially simplified. The ACF now displays a rapid decay toward zero, with residual spikes mainly at lags 1–2, while the PACF shows significant contributions restricted to the same lags. This pattern suggests that a low-order ARMA model, combined with a seasonal MA(1) term at period 24, is appropriate to describe the dynamics of the differenced series. On this basis, three candidate structures were evaluated: a MA(2) model, an AR(2) model, and a combined SARIMA(2,1,2)(0,1,1)[24], which were then compared using information criteria (AIC) and validation performance.

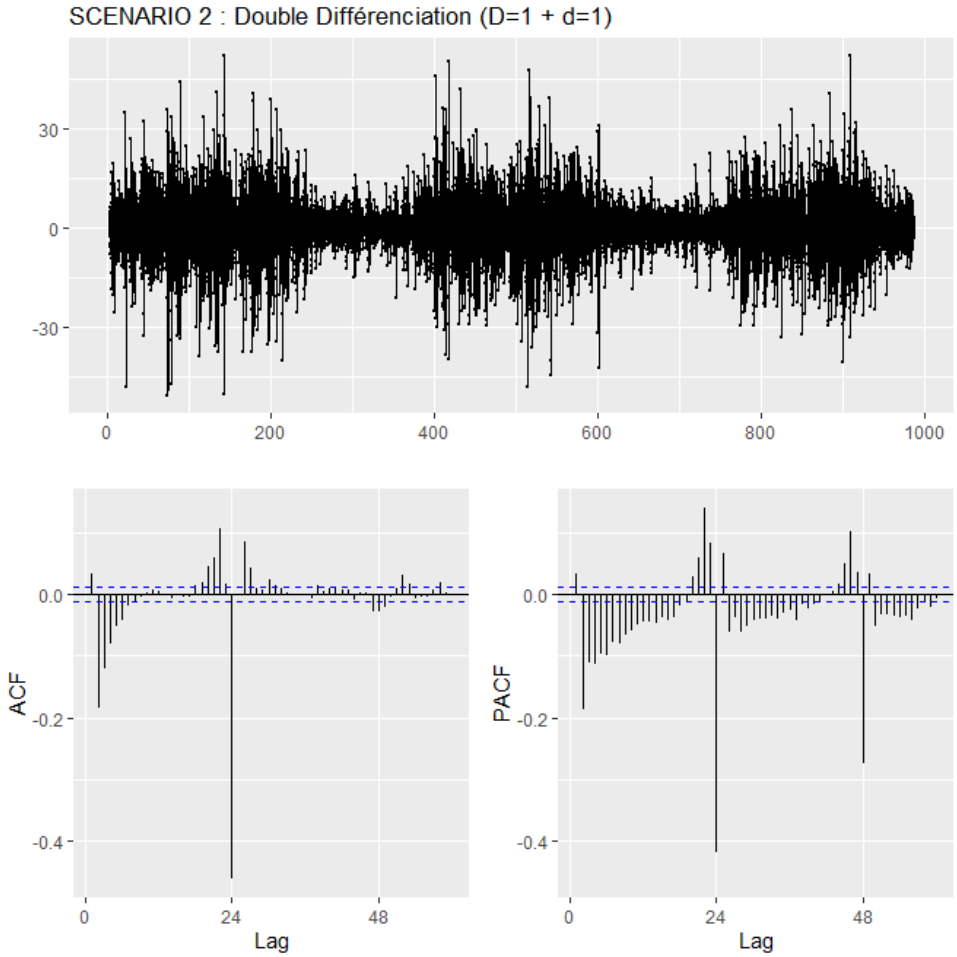

Figure 9: Series of NO_2 after double differencing (d = 1, D = 1) and associated ACF/PACF functions.

3.3.2 SARIMA model fitting

Several model structures were compared using a data split into training, validation, and test sets with a 80/10/10 ratio. On the validation set, the simple MA(2) and AR(2) models lead to root mean squared errors (RMSE) of about 7.24 and 7.23 $\mu g/m^3$, respectively, whereas the combined SARIMA(2,1,2)(0,1,1)[24] model slightly improves the performance with an RMSE of 6.93 $\mu g/m^3$ and a lower Akaike Information Criterion (AIC) value (approximately 135 113 compared with 137 569 and 137 593 for the MA(2) and AR(2) models). On a 48-hour validation window, the model satisfactorily reproduces the hourly dynamics of NO_2: the forecast curve closely follows the observed series, correctly capturing the daily cycles and the main concentration peaks, with an overall RMSE of about 3.3 $\mu g/m^3$ over this period. The 80 % and 95 % prediction intervals contain the vast majority of observations, suggesting that the predictive uncertainty provided by the model is consistent with the variability observed in the data.

Table 12: Selection of the NO_2 model: theory vs. practical performance

Model	**Structure**	**AIC**	**RMSE (val)**
Candidate MA(2)	ARIMA(0,1,2)$(0,1,1)_{24}$	137 568.88	7.24
Candidate AR(2)	ARIMA(2,1,0)$(0,1,1)_{24}$	137 592.54	7.23
Combined candidate	**ARIMA(2,1,2)**$(0,1,1)_{24}$	**135 112.78**	**6.93**

3.3.3 Diagnostics of SARIMA residuals

Before introducing a GARCH component, it is necessary to assess how well the SARIMA model captures the temporal dependence in the series. To this end, the residuals of the SARIMA(2,1,2)(0,1,1)[24] model were analysed using the Ljung–Box test together with the autocorrelation function (ACF) and partial autocorrelation function (PACF). The Ljung–Box test yields a statistic of $Q^* = 387{,}1$ for 43 degrees of freedom, with a p-value lower than $2{,}2 \times 10^{-16}$, indicating that the residuals cannot be treated as independent white noise. Visually, the ACF and PACF of the residuals show a clear reduction of the seasonal structure compared with the original series, but some correlations remain at specific lags, suggesting that the mean component of the process is largely captured while a residual dynamics persists, in particular at the level of the variance.

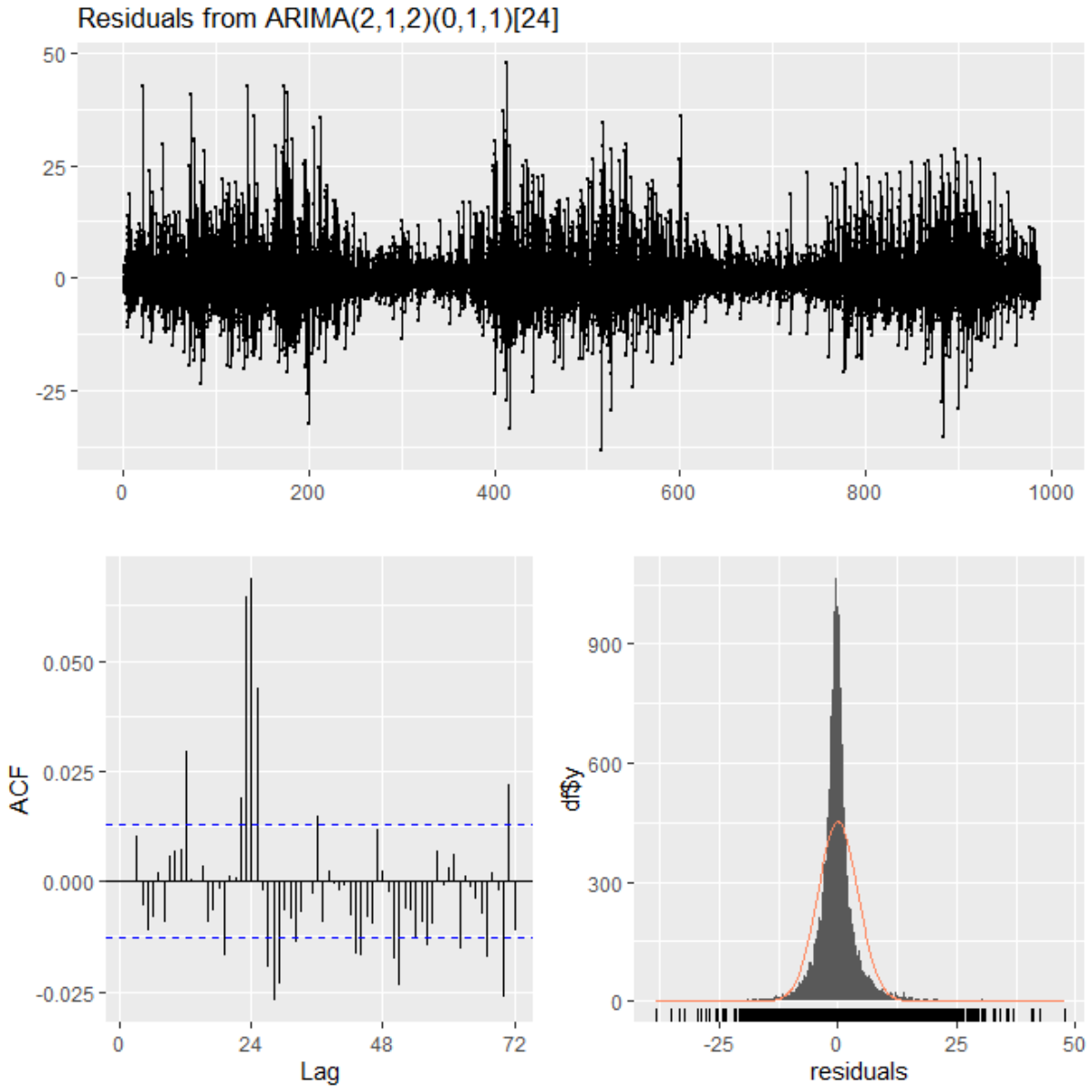

Figure 10: Residuals of the SARIMA(2,1,2)(0,1,1)[24] model for NO_2 and associated diagnostics.

3.3.4 Modélisation de la volatilité : GARCH(1,1)

La structure résiduelle mise en évidence par ces diagnostics suggère la présence d'hétéroscédasticité conditionnelle, avec des périodes de forte et de faible variabilité. Afin de prendre en compte cette volatilité, une composante GARCH(1,1) a été ajustée sur les résidus issus du modèle SARIMA. Les paramètres estimés sont $\omega = 0{,}0584$, $\alpha_1 = 0{,}0437$ et $\beta_1 = 0{,}9553$, tous significatifs au seuil usuel comme en témoignent les valeurs de statistique de test très élevées et les p–values proches de zéro. La somme $\alpha_1 + \beta_1 \approx 0{,}999$ indique une forte persistance de la volatilité, c'est-à-dire que les chocs sur la variance ont des effets durables, cohérents avec l'existence de périodes de forte variabilité des concentrations (épisodes de pollution, conditions météorologiques changeantes). L'introduction de la composante GARCH permet de stabiliser la variance des résidus standardisés et de mieux décrire les clusters de variance observés sur la série, ce qui se traduit par des intervalles de prédiction plus réalistes, notamment lors des épisodes de fluctuations marquées.

Table 13: Estimation of the GARCH(1,1) model parameters for NO_2

Parameter	**Name**	**Estimate**	**Std. Error**	**t value**
ω	omega	0.0584	0.0039	14.88
α_1	alpha1	0.0437	0.0007	59.05
β_1	beta1	0.9553	0.0004	2276.93

3.3.5 Practical interpretation of the results

From an operational perspective, the SARIMA(2,1,2)(0,1,1)[24]–GARCH(1,1) model is able to reproduce the daily seasonality of NO_2 concentrations, with maxima generally associated with traffic periods and nocturnal minima, while also capturing the main episodes of rapid increase or decrease. It provides 48-hour forecasts with a moderate average error (RMSE of about 3 μg/m^3 on the illustrated window and around 7 μg/m^3 on the validation sample), which makes it a relevant tool for analysing NO_2 dynamics and a potential basis for short-term operational scenarios. The presence of persistent volatility, highlighted by the GARCH component, underlines that concentrations remain strongly influenced by external factors (meteorological conditions, local emissions, punctual events) and that prediction intervals should therefore be interpreted as probability bands rather than strict deterministic bounds. Overall, the proposed model offers a satisfactory compromise between statistical realism, predictive performance, and readability for environmental interpretation.

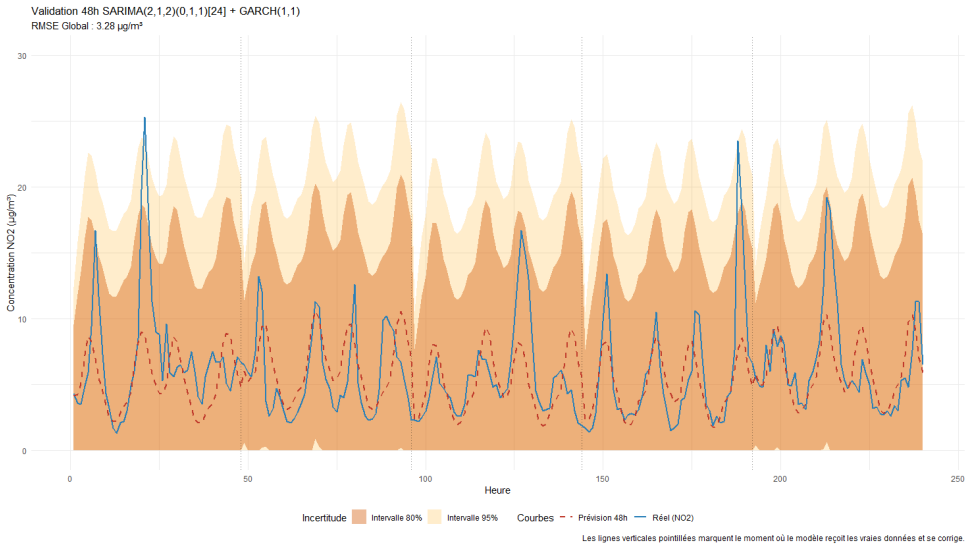

Figure 11: Forecast vs. Actuals for NO2 in Urban areas

3.4 Summary of the 9 models used

Table 14 provides a consolidated overview of the nine predictive models developed, detailing the specific SARIMA and GARCH specifications retained for each Zone/Pollutant pair.

Table 14: Summary of Selected Models for each Pollutant/Zone pair

Zone \Pollutant	$PM_{2.5}$	NO_2	O_3
Rural	SARIMA(1,1,2)$(1,0,1)_{24}$ + sGARCH(1,1)	SARIMA(1,0,2)$(1,0,1)_{24}$ + sGARCH(1,1)	SARIMA(1,0,2)$(1,0,1)_{24}$ + sGARCH(1,1)
Peri-urban	SARIMA(0,1,2)$(0,1,1)_{24}$ + sGARCH(1,1)	SARIMA(2,1,2)$(0,1,1)_{24}$ + sGARCH(1,1)	SARIMA(1,1,0)$(0,1,1)_{24}$ + sGARCH(1,1)
Urban	SARIMA(2,1,2)$(0,1,1)_{24}$ + sGARCH(1,1)	SARIMA(2,1,2)$(0,1,1)_{24}$ + sGARCH(1,1)	SARIMA(0,1,1)$(0,1,1)_{24}$ + sGARCH(1,1)

4 Setting up a User Interface and Interpreting Results

The final stage of this project involves the operational deployment of our predictive models into a dashboard developed with RShiny. This interactive tool bridges the gap between complex time-series analysis (SARIMA-GARCH) and practical decision-making, allowing stakeholders to monitor air quality risks in the Occitanie region in real-time.

4.1 Backend Logic: The "Time Machine" Approach

The application is designed to simulate a real-time production environment. It relies on a "Time Machine" function which automatically calibrates the forecast horizon based on the latest available observation (in this simulation: *January 6, 2026, at 12:00*).

The backend logic follows a strict three-step process:

1. **Dynamic Model Assignment:** The system automatically maps each station to its corresponding predictive model based on the Zone/Pollutant typology defined in the previous section.

2. **Probabilistic Forecasting:** For every station, the system generates a 48-hour forecast. Crucially, it computes not only the Mean Prediction but also the upper bounds of the 80% and 95% Confidence Intervals, leveraging the volatility modeled by the GARCH component.

3. **Threshold Monitoring:** The forecast is compared against regulatory thresholds ($25\mu g/m^3$ for $PM_{2.5}$, $120\mu g/m^3$ for O_3, etc.) to trigger alerts.

4.2 Risk Assessment Strategy

A key innovation of this interface is its management of uncertainty. Instead of a binary alert system, we implemented a hierarchical risk assessment strategy that transforms statistical confidence intervals into operational alert levels (Table 15).

Table 15: Definition of Alert Levels based on GARCH Uncertainty

Risk Level	Statistical Criterion	Operational Interpretation
HIGH ALERT	Mean Prediction $> T$	The exceedance is highly probable. Immediate mitigation actions required.
MEDIUM ALERT	Upper Bound (80% CI) $> T$	There is a significant risk (20% probability) of exceeding the threshold. Enhanced monitoring advised.
POSSIBLE ALERT	Upper Bound (95% CI) $> T$	The exceedance is unlikely but possible in a worst-case volatility scenario (e.g., sudden wind change).

Note: T represents the regulatory threshold for the selected pollutant.

4.3 Dashboard Overview and Smart Synthesis

Figure 12 presents the full interface of the monitoring tool. The layout is divided into three functional areas:

1. **Control Panel (Left):** Allows users to switch pollutants ($PM_{2.5}, NO_2, O_3$), filter specific typologies (e.g., only "Urban" stations), and navigate the 48h horizon.

2. **Geospatial Visualization (Map):** Provides an instant overview of the region. Markers change color based on the pollutant and allow users to access specific station data via popups.

3. **Smart Alert Synthesis (Bottom Table):** To prevent information overload, we developed an aggregation algorithm (using ‘lag‘ and ‘cumsum‘ functions). Instead of listing every hourly violation, the system groups consecutive hours into Time Slots.

For example, as seen in the dashboard, the station *"Montauban - Ramierou Urbain"* triggers a HIGH ALERT for the time slot "From 06/01 18h to 23h". This synthesized view allows operators to instantly identify the duration and intensity of the pollution episode without parsing raw data.

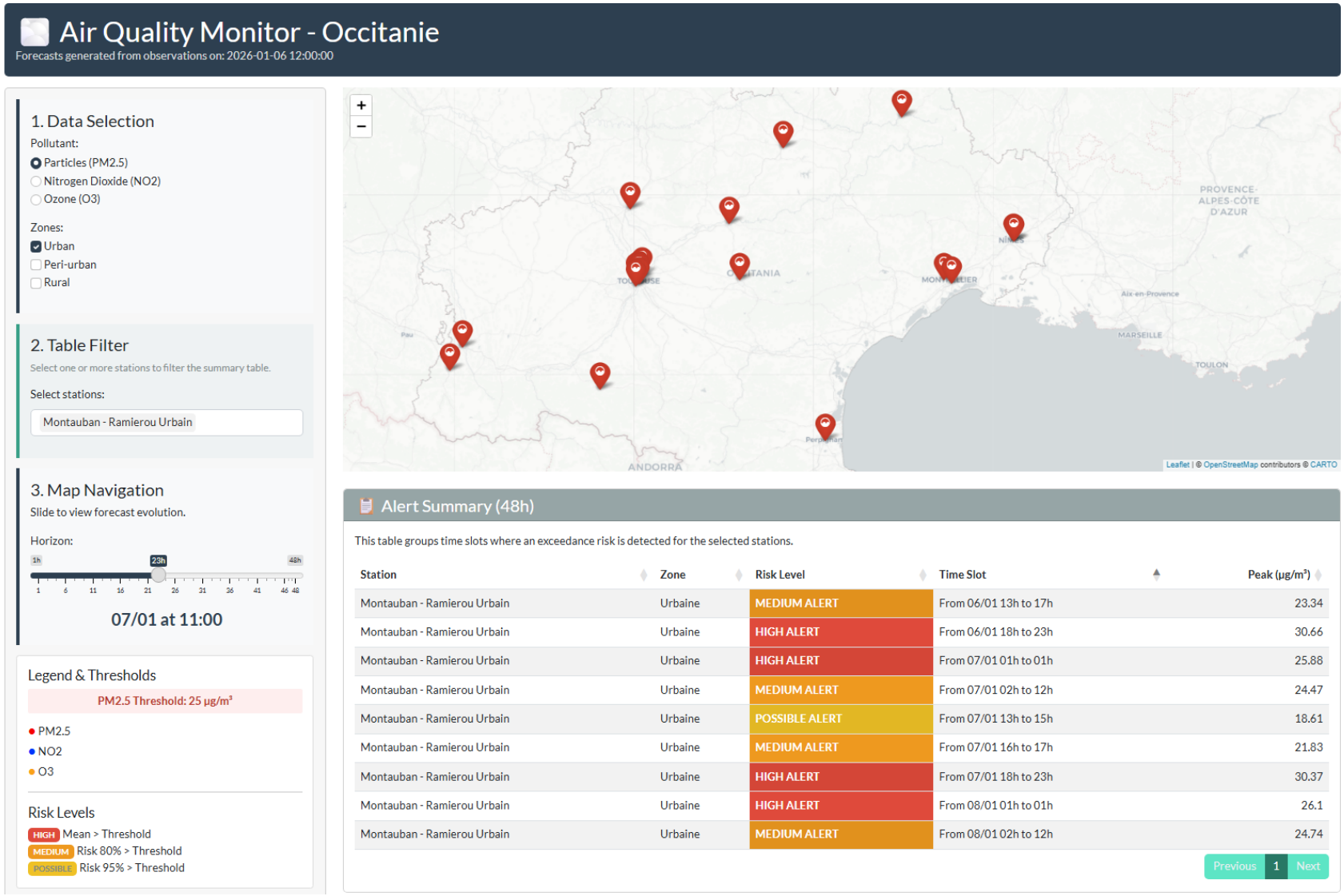

Figure 12: Full Interface of the Air Quality Monitor. The dashboard highlights a High Alert scenario for PM2.5 in Montauban, utilizing the "Smart Synthesis" table to group critical hours.

5 Conclusion and Project Limits

This project successfully demonstrated the potential of Time Series Analysis in building an operational air quality monitoring system for the Occitanie region. By segmenting the analysis into distinct typologies (Urban, Peri-urban, Rural) and employing a "Reference Station" methodology, we were able to deploy tailored predictive models for Ozone (O_3), Nitrogen Dioxide (NO_2), and Fine Particles ($PM_{2.5}$).

The core contribution of this work lies in the hybridization of SARIMA and GARCH models. While the SARIMA component effectively captured the strong diurnal seasonality and linear trends inherent in atmospheric data, the integration of a GARCH component proved decisive for risk management. By modeling the volatility clustering, we transitioned from a simple point forecast to a probabilistic framework, providing dynamic confidence intervals (80% and 95%) that successfully bound the uncertainty.

The final deliverable, an interactive RShiny dashboard, bridges the gap between statistical complexity and decision-making. It translates these mathematical outputs into actionable insights ("High," "Medium," "Possible" alerts) and synthesizes critical time windows, meeting the operational needs of public health monitoring.

Despite the satisfactory performance in capturing general trends and seasonality, the system exhibits certain limitations inherent to the chosen stochastic approach:

1. **Underestimation of Sudden Peaks:** A recurring pattern across our models (particularly for NO_2 and $PM_{2.5}$) is the difficulty in anticipating extreme, sudden spikes in concentration. Autoregressive models (AR) tend to "smooth" the signal and revert to the mean. Consequently, while the *timing* (phase) of the peaks is generally correct, their *amplitude* is often underestimated. This suggests that exogenous factors (such as sudden wind changes or specific local events) are missing from the univariate equation.

2. **Limited Forecast Horizon:** The reliability of the forecasts degrades significantly beyond the 48-hour mark. As the horizon extends, the confidence intervals widen until they become uninformative, and the point forecast flattens out to the series mean. This confirms that this SARIMA-GARCH architecture is strictly suitable for short-term "Nowcasting" (0-48h) and cannot support long-term strategic planning.

3. **Station Heterogeneity:** The "Reference Station" strategy, while computationally efficient, assumes that all stations within a typology behave similarly. In reality, local micro-climates or specific traffic configurations can introduce deviations that the generalized model fails to capture, necessitating a more granular, station-specific tuning in a production environment.

To overcome these limits, future iterations could explore multivariate approaches (ARIMAX) incorporating meteorological covariates (temperature, wind speed) or Machine Learning methods (LSTM, XGBoost), which are better suited for capturing non-linear complexities and extreme deviations.

References

[1] Airparif, « La réglementation en France », https://www.airparif.fr/la-reglementation-en-france

[2] Goshua, Anna and Akdis, Cezmi and Nadeau, Kari C., « World Health Organization Global Air Quality Guideline Recommendations: Executive Summary », https://pmc.ncbi.nlm.nih.gov/articles/PMC12052406/