environments

MDPI

Article

Modeling Hydrological and Atmospheric Interactions in the Mar Menor Lagoon Through Machine Learning: Feature Selection, Time Association, and Predictive Modeling

Juan Fernando Lima [1,*], Jose-María Cecília [2], Carlos T. Calafate [2] and Marcos Orellana [1]

[1] Computer Science Research and Development Laboratory (LIDI), Universidad del Azuay (UDA), Cuenca 010204, Ecuador; marore@uazuay.edu.ec
[2] Department of Computer Engineering (DISCA), Universitat Politècnica de València (UPV), 46022 Valencia, Spain; jmcecilia@disca.upv.es (J.-M.C.); calafate@disca.upv.es (C.T.C.)
* Correspondence: flima@uazuay.edu.ec

Abstract

The Mar Menor Lagoon (Spain), recently granted legal personality after severe eutrophication crises, is monitored by sensors that generate multi-source, high-frequency data, yet hydrological and atmospheric variables are usually analyzed in isolation. This study examines how these variables interact over time. Using 2023 data from the lagoon buoy, AEMET meteorological stations (San Javier Airport, Cabo Palos), the La Pedrera reservoir and the MURC2 tide gauge. We follow the CRISP-DM methodology and combine Lasso and Elastic Net feature selection, K-means and Gaussian Mixture clustering, time-rolling-window Spearman correlation, and Random Forest prediction of chlorophyll-a and turbidity. Results indicate that (1) air and lagoon water temperature are strongly coupled, and the disruption of this coupling flagged a sensor anomaly in June 2023; (2) the positive association between conductivity and temperature holds only during the first six months, suggesting a shift in the dominant drivers of ion mobility; and (3) wind power is negatively associated with temperature and conductivity, consistent with evaporative cooling and surface mixing. In this single-year record, associations are therefore time-varying, which challenges static, compartmentalized monitoring. The workflow is reproducible and can support sensor quality control and evidence-based management of legally protected water bodies.

Keywords: environmental monitoring; feature selection; rolling correlation; clutsering; forecasting; Mar Menor; Spain

Check for updates

Academic Editors: William A. Anderson and Ching-Yuan Chang

Received: 25 September 2026
Revised: 18 October 2026
Accepted: 25 October 2026
Published: 29 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

In recent years, concerns regarding climate change have attained heightened importance within policy agendas, prompting governments to implement measures aimed at mitigating air and water pollution. In fact, water protection in Spain is governed by a multilevel legal framework in which EU law sets the objectives and national and regional law implements them [1]. It requires Member States to reach good ecological and chemical status in surface and coastal waters, and to run monitoring programmes based on biological and physico-chemical quality elements [2]. This fundamental legislation has been subsequently amended to regulate both point-source and nonpoint-source pollution. Climate-related pressures are addressed by the European Climate Law (Regulation (EU) 2021/1119) [3] and Spain's Law 7/2021 on climate change [4].

The Mar Menor Lagoon in Spain, a substantial coastal water body, has been the subject of specific legislation and regulations aimed at safeguarding its waters for agricultural purposes. This legal response follows a series of environmental crises that have heightened concerns for the ecological integrity of the region. These conservation efforts culminated in the landmark recognition of the lagoon as a legal entity with rights [5,6]. Concurrently, a comprehensive monitoring program has been implemented as a core component of its stewardship. This program involves the ongoing monitoring of essential water quality parameters via an array of sensors installed at various surface locations and depths. Parameters subjected to ongoing monitoring encompass water temperature, dissolved oxygen levels, oxygen saturation, electrical conductivity, chlorophyll-a, and turbidity. This long-term, high-resolution data collection provides the empirical foundation for informed management decisions aimed at the lagoon's preservation [7,8].

In fact, decisions regarding local policies are founded upon the data of the aforementioned variables; consequently, various fields of knowledge have contributed to the formulation of these policies. Computer science is among the disciplines that have proposed a range of instruments, such as a model to simulate several Best Management Practices (BMP) [9], an approach for selecting cost-effective measures [10], and the assessment of the impact of climate risk manifestation and ecological deterioration on house prices [11]. Data serves as the foundation for specialized monitoring algorithms, including deep learning techniques for efficient monitoring [12] and predicting streamflow behavior [13], or forecasting warning indicators such as chlorophyll-a concentrations, which are essential for the development of marine ecosystems [14,15].

Environmental analysis of coastal lagoons is typically conducted within discrete compartments, with variables categorised as hydrological, atmospheric or chemical. Yet the physical processes that control lagoon water quality cross these boundaries, for the Mar Menor, whose management now carries a legal mandate to demonstrate the protection of the lagoon, this creates a concrete problem. Monitoring networks produce large volumes of heterogeneous, multi-source, high-frequency data, but it is unclear which of these variables carry the information that matters, how the lagoon–atmosphere relationships evolve over time rather than remaining constant, and whether the resulting knowledge can support anomaly detection and short-term prediction of key eutrophication indicators.

Accordingly, the objective of this paper is to characterize the temporal dynamics of hydrological–atmospheric interactions in the Mar Menor lagoon and to assess their usefulness for monitoring and prediction. To this end, we apply machine-learning techniques as analytical instruments: regularized regression (Lasso and Elastic Net) for variable selection, clustering for exploratory structure analysis, time-rolling-window Spearman correlation for non-stationary associations, and Random Forest for non-linear prediction. We use one year (2023) of data from five sources within a CRISP-DM workflow. The contributions are: (i) a parsimonious set of drivers selected from a multi-source dataset; (ii) evidence of associations across the year, which challenges static management approaches; (iii) a correlation-based procedure that flagged a sensor anomaly in June 2023; and (iv) a reproducible pipeline that can be re-applied as new data become available.

2. Related Work

Currently, marine pollution in the Mar Menor Lagoon represents a critical environmental challenge that has escalated since the 1980s. Over the decades, intensive irrigated agriculture has discharged elevated concentrations of nitrates and phosphates into the lagoon, triggering massive mortality events among marine organisms. The economic repercussions extend beyond ecological damage; empirical evidence indicates that even a marginal increase of 1 mg/m^3 in chlorophyll-a concentration, a standard proxy for eu-

trophication, significantly raises the probability of business failure for firms operating in adjacent areas. Paradoxically, agricultural and transport sectors continue to derive economic benefits from coastal proximity despite progressive environmental degradation, thereby diminishing their incentive to adopt conservation measures [16,17]. Consequently, these compounding pressures have necessitated systematic efforts to monitor the lagoon's environmental variables through continuous observation protocols.

To address the analytical demands posed by such high-dimensional environmental monitoring, the academic community has increasingly integrated machine learning algorithms into marine data analysis frameworks [18]. These computational methodologies enable data administrators and environmental strategists to synthesize heterogeneous sensor outputs and comprehend complex temporal patterns in near real time. Specifically, algorithms oriented toward data description, classification, and predictive modeling have become standard tools for interpreting environmental datasets [19]. The operational pipeline typically involves the systematic collection, validation, and archival of numerical measurements corresponding to multiple biogeochemical context variables. In aquatic ecosystems such as lagoons, seas, and rivers, key monitoring parameters include water temperature, turbidity, Secchi disk depth, salinity, dissolved oxygen, nitrogen compounds, and total phosphorus, frequently sampled at multiple depths to capture vertical stratification [12,20].

Conversely, not only water variables are prone to analysis; studies have also confirmed the relationships between water and meteorological variables, thereby achieving a harmonious balance in environmental synergy [21]. Furthermore, it has been demonstrated that demographic, landscape, and meteorological variables are factors that influence the microbial pollution of coastal waters [22].

Novel computational algorithms, including clustering, regression, and neural networks, are now frequently applied to monitor and analyze environmental systems. Specifically, algorithms such as Recursive Feature Elimination (RFE), Genetic Algorithm (GA) and Simulated Annealing (SA), combined with Random Forest (RF) or Support Vector Regression (SVR), are new approaches to predicting the chlorophyll-a concentration, yielding successful results such as R^2 values greater than 0.7 [20]. This work provides an overview of the power of data mining techniques, opening up new perspectives for lagoon monitoring. It not only applies to this lagoon, but also seeks to generalize the techniques to other environmental scenarios.

The widespread deployment of Internet of Things (IoT) devices has enabled the retrieval of environmental data from multiple heterogeneous sources, thereby introducing significant preprocessing challenges prior to analysis. For instance, López-Andreu et al. [23] integrate biophysical, nutrient, meteorological, hydrological, and satellite datasets within a unified analytical framework. Beyond source heterogeneity, the geographical dispersion of monitoring stations introduces spatial misalignment that must be reconciled during data fusion. Furthermore, temporal periodicity constitutes an additional complicating dimension; even when datasets span comparable temporal windows, their sampling frequencies frequently differ across sources. These compounding heterogeneities, spanning data origin, spatial coordinates, and temporal resolution, substantially hinder the structuring, integration, and archival of the consolidated dataset.

Dealing with high-dimensional data is a challenging task that requires a framework to be executed properly; computational techniques are used to address this challenge. The framework implies a flow of activities that includes: a full dataset file, a subset generation, a subset evaluation (using metrics such as Mean Absolute Error (MAE) or similar), and the stop criteria [24]. However, newer algorithms seek to use this flow and new metrics to evaluate the best scores using cross-validation, such as the case of the Lasso and Elastic Net algorithms, which have been proposed as methods to create a predictor while excluding variables that do not satisfy a specific criterion [25].

Despite the advances reviewed above, existing studies addressing the Mar Menor Lagoon, and comparable coastal ecosystems, have largely treated hydrological and atmospheric variables as separate analytical compartments [20,23], or have relied on static predictive models that do not capture the temporal evolution of cross-domain associations. Although feature selection frameworks such as RFE and GA have improved chlorophyll-a prediction accuracy [20], the joint application of regularized regression, specifically Lasso and Elastic Net, with time-rolling correlation windows remains unexplored for modeling nonstationary interactions in this ecosystem. This paper addresses that gap by presenting a unified Cross-Industry Standard Process for Data Mining (CRISP-DM) pipeline that integrates multi-source IoT data with regularized feature selection, exploratory clustering, and time-rolling-window correlation to quantify temporally evolving hydrological–atmospheric couplings in the Mar Menor, thereby advancing beyond compartmentalized or purely static analyses.

3. Method and Materials

To achieve proper data analysis, several approaches have been evaluated. In this context, the work of Schöer et al. [26] reveals that the CRISP-DM is a robust and well-proven methodology for data analysis [27]. It includes a set of activities to understand and manage data, from the collection stage to model deployment. Based on the different data sources shown in Table 1, data for 2023 are used in this study to ensure uniformity; however, the analysis can be reproduced when new data become available.

Table 1. Summary statistics for Mar Menor Lagoon environmental variables, 2023.

Variable	Abbreviation	M	SD	Min	Q1	Q2	Q3	Max
Mar Menor Buoy (3 m)								
Air relative humidity (%)	AIR_RELHUMIDITY	74.47	9.66	42.49	69.45	76.57	81.62	92.16
Air temperature (°C)	AIR_TEMP_HS	19.48	6.02	8.10	14.63	19.30	25.05	30.06
Vapor pressure (hPa)	AIR_VAPOR_ PRESSURE	1.80	0.72	0.52	1.22	1.62	2.40	3.38
Wind power (W/m^2)	AIR_WS3	143.01	177.87	9.88	49.55	80.34	154.10	1333.42
Wind speed (m/s)	AIR_WS3_MS	4.13	1.51	1.58	3.07	3.79	4.81	10.38
Chlorophyll-a (µg/L)	SEA_CHL	0.96	0.50	0.38	0.67	0.82	1.02	3.08
O_2 concentration (mg/L)	SEA_O2_conc	9.97	2.03	6.21	8.34	9.86	11.22	15.55
O_2 saturation (%)	SEA_O2_sat	99.43	11.85	74.72	89.74	101.85	108.38	127.89
Conductivity (µS/cm)	SEA_SDI_Cond	57,516.74	7252.14	44,390.08	50,813.52	59,452.26	63,960.35	67,963.44
Water temperature (°C)	SEA_SDI_Temp	20.97	6.66	10.18	14.39	20.42	27.22	31.31
Temp.-corrected conductivity (µS/cm)	SEA_SDI_ TempCorrCond	61,339.79	2263.60	54,707.26	60,458.94	61,958.81	62,895.65	64,880.45
Turbidity (NTU)	SEA_TURB	1.63	0.45	1.21	1.36	1.55	1.72	4.67
Cabo Palos (S7019X)								
Wind direction (°)	S7019X_DIR	18.94	19.65	1.00	6.00	19.00	21.00	99.00
Relative humidity (%)	S7019X_HRMEDIA	74.99	11.90	33.00	68.00	77.00	82.25	100.00
Precipitation (mm)	S7019X_PREC	0.65	6.01	0.00	0.00	0.00	0.00	107.80
Wind gust (m/s)	S7019X_RACHA	9.90	3.43	3.90	7.20	9.40	11.90	24.20
Mean temperature (°C)	S7019X_TMED	19.26	5.95	7.90	14.20	19.15	24.65	30.20
Mean wind speed (m/s)	S7019X_VELMEDIA	4.15	1.65	1.40	3.10	3.90	5.00	11.70
San Javier Airport (S7031)								
Precipitation (mm)	S7031_PREC	−32.47	178.46	0.00	0.00	0.00	0.00	28.00
Insolation (h)	S7031_SOL	8.62	3.01	0.00	7.20	9.20	10.48	13.70
Mean temperature (°C)	S7031_TMED	18.78	6.28	6.00	13.80	18.80	24.20	30.30
Embalse de La Pedrera (EM085)								
Evaporation (mm)	EM085_EVAP_MM	76.76	44.39	13.00	45.00	68.00	94.00	300.00
Precipitation (mm)	EM085_PRECIP_MM	4.79	27.25	0.00	0.00	0.00	0.00	342.00
Tide Gauge (MURC2)								
Water level (mm)	MURC2_LEVEL_MM	−2333.72	997.05	−2231.00	−2278.00	−2195.00	−2135.00	−2004.00

Note. M = mean; SD = standard deviation; Q1 = first quartile; Q2 = median; Q3 = third quartile. Abbreviations correspond to the variable names used in Figures 3 and 5. The negative mean precipitation at S7031 reflects a data-encoding artifact in the source repository rather than a physical measurement; consequently, this variable was excluded from correlation analyses.

3.1. Data Understanding and Data Sources

Before describing the data, this subsection presents the data sources, the stations and the variables. Figure 1 shows the geographic locations of the five data sources. (1) The primary source is the Mar Menor monitoring buoy [28], which records meteorological variables above the water surface and water-quality variables at fixed depths in the water column. Water temperature (°C) is measured by different sensors covering ten levels (0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 4.0, 5.0, 6.0 and 6.5 m); oxygen saturation (%) and dissolved oxygen concentration (mg/L) at 1.0, 3.0 and 6.5 m; electrical conductivity (µS/cm), temperature-corrected conductivity (µS/cm at 25 °C) and salinity at 1.0, 3.0 and 6.0 m; and chlorophyll-a (µg/L) and turbidity (NTU) only at 3.0 m. Since chlorophyll-a and turbidity, the target variables of this study, are available only at 3.0 m, all water-column measurements used here correspond to that depth, which gives a consistent vertical reference across parameters providing the variability needed to analyze the dynamics of the lagoon.



Figure 1. Geographic location of data sources. Mar Menor is enclosed by La Manga, a narrow strip of sand approximately 22 km long and ranging in width from 100 to 1200 m. This natural barrier prevents the Mar Menor from being fully open to the open sea.

The two subsequent stations are included in the open data provided by the AEMET of the Spanish Government, which reports measurements from various locations across Spain. In this instance, only stations in the vicinity of the Mar Menor were considered. (2) Cartagena/Salinas Cabo de Palos (7019X), situated near "La Manga" in the middle of the coastal spit (or sandbar) that separates the Mar Menor lagoon from the Mediterranean Sea. The prefix 7019 was designated for each variable from this station (indicated in red). (3) Similarly, San Javier Airport (7031) is a station located centrally within the airport premises, and it is employed to monitor air quality variables.

Variables associated with external environmental behavior were also incorporated, focusing on modeling the relationships underlying the ecosystem of the Mar Menor. However, the objective variables were not reported at the aforementioned stations; consequently, the nearest stations were utilized. The tide sensor, also referred to as a tide gauge, situated in Puerto Cartagena, was included to measure tidal height. Specifically, it is designated as tide gauge station MURC2, indicated in purple. Lastly, the reservoir located at the Embalse de La Pedrera (EM-085) reports variables such as evaporation and precipitation. These variables will be employed for cross-validation purposes.

3.2. Data Preparation

A relational database was prepared to store data related to the Mar Menor. This Database Management System (DBMS) was used for its features in data joins, aggregations, filters, and performing separate calculations within each logic group.

3.2.1. Variables Collection

Data profiling is focused on examining, analyzing, and summarizing data from an existing source to gain insights into its structure, content, quality, and relationships. Think of it as taking a "data census" to get a comprehensive report card on the data. In the case of Mar Menor variables, it is done by means of governmental policies; for some variables, it has allowed the creation of Application Programming Interface (API)s and data is shared publicly [28]. Data auditing is focused on ensuring that collected data can represent the reality of the environmental behavior due to issues that can alter the measurements; these include device calibration, services not available, or data manipulation. Data auditing is in charge of government organizations, which validate data according to the Spanish government [29].

3.2.2. Data Cleaning

Data cleaning encompasses a series of preparatory procedures designed to convert raw data into a structured format suitable for subsequent analytical modeling. Within this stage, two fundamental operations are performed: transforming the dataset's architectural organization to ensure compatibility with analytical frameworks, and standardizing measurement scales to prevent algorithmic bias.

Data Transformation: This step involves deriving new attributes from existing variables and restructuring the dataset to facilitate downstream analysis. The raw dataset is characterized by hierarchical headers representing variables at varying levels of granularity, thereby forming distinct measurement subgroups. To render this structure compatible with data management frameworks such as Pandas, restructuring operations are required. Specifically, the `melt` function reconfigures the DataFrame by designating one or more columns as identifiers (`id_vars`), while all remaining columns, treated as measured variables (`value_vars`), are unpivoted along the row axis, yielding a normalized format comprising `variable` and `value` columns [30].

Data Standardization: Standardization is essential to ensure algorithmic validity, particularly for distance-based methods such as clustering, which are inherently sensitive to disparities in feature scales. Without normalization, variables with larger ranges may disproportionately influence model outcomes, leading to biased results. Consequently, statistical normalization techniques are systematically applied to rescale features to a common distribution, thereby preserving the relative importance of each variable across the dataset.

3.3. Feature Selection

Feature selection is a method frequently employed in machine learning whereby an algorithm autonomously optimizes the selection of the most pertinent features (variables) from the initial dataset. Its primary aims include enhancing the model's performance, minimizing computational expenses, and simplifying the model to facilitate easier interpretation by eliminating irrelevant or redundant features.

The application of this process is aimed at achieving: (a) improved model performance; removing noisy or irrelevant features can prevent overfitting, which occurs when a model learns the training data excessively well, and subsequently underperforms on new, unseen data [31]; (b) reduced training duration; with fewer features, the model processes less data,

thereby significantly accelerating the training process [32]; and (c) enhanced interpretability; a model incorporating fewer features proves simpler to explain and comprehend. This aspect is particularly critical in domains such as healthcare or environmental policy, where one needs to justify a model's decisions based on the input variables [31].

This paper primarily focuses on the interpretability objective described above. The preceding subsections provide multiple sources that contribute several variables, some of which are duplicated across stations. In this context, the feature selection process enables us to identify the most pertinent features. To obtain interpretable models and prevent overfitting, the regularization approach incorporates a penalty into the model's loss function, thereby constraining the magnitude of the coefficients. Consequently, several candidate algorithms based on regularization are appropriate for consideration at this stage. Specifically, the Lasso, Ridge, and Elastic Net methods are employed to shrink coefficients in order to prevent overfitting [33], while Lasso and Elastic Net additionally facilitate feature selection by reducing the influence of certain variables to exactly zero [34].

Lasso regression's L1 penalty automatically shrinks the coefficients of less important features to zero. This makes it a powerful tool for feature selection because any feature with a zero coefficient has been effectively removed from the model. On the other hand, Elastic Net combines the L1 penalty of Lasso and the L2 penalty of Ridge; it also performs feature selection by shrinking coefficients to zero, but it is particularly useful when one has a group of highly correlated features. Unlike Lasso, which tends to arbitrarily select a single member from a correlated group, Elastic Net distributes coefficients across correlated predictors [33].

To obtain the best parameters and, therefore, an accurate feature selection, `GridSearchCV` from scikit-learn was used; it performs an exhaustive search over specified parameter values for an estimator. In the case of both Lasso and Elastic Net, the list of alphas was: `logspace(start=-3, stop=2, num=50)`. The `logspace` function generates numbers spaced evenly on a log scale; the number of cross-validation folds was 5. Specifically, for Elastic Net, the list of L1 ratios was [0.01, 0.05, 0.10, 0.50, 0.95].

Lasso and Elastic Net are regarded as regression techniques. This characteristic serves as the basis for training and assessing the current variables during the feature selection phase. A prior study aimed to predict chlorophyll and turbidity levels to reduce sensor operational time, thereby conserving energy. The researchers applied 10-fold cross-validation to the training data, foregoing a separate validation set. Within each fold of the cross-validation, the data were standardized through scaling, subtracting the mean, and dividing by the standard deviation, prior to the evaluation of algorithms and the selection of the most promising models for final testing on the hold-out test set [15]. The performance of the models was evaluated using two primary metrics: MAE and the Pearson Correlation Coefficient (PCC) [35]. In the present study, to facilitate feature selection, chlorophyll and turbidity were used as target variables in the regression, and MAE was used to measure accuracy by quantifying the average absolute difference between the true and predicted values.

3.4. Exploratory Clustering Analysis

Clustering, irrespective of the chosen technique, is an unsupervised machine learning approach that categorizes data points according to their similarities. It is employed to identify and systematically organize data into natural groupings, whereby data points within a given cluster exhibit greater similarity to one another than to those in other groups. Nonetheless, selecting a clustering algorithm that is universally applicable to all datasets is difficult [36].

In this context, various techniques have been proposed based on specific characteristics of the data. Consequently, k-means, recognized as the traditional clustering technique,

utilizes metrics such as Euclidean, Manhattan, or Bregman divergences. Nevertheless, while these methods offer certain advantages, they also present limitations regarding their use in high-dimensional spaces [37]. The scikit-learn library was used to train a K-means model with variables of the Mar Menor Lagoon; it was trained with the following parameters: `n_clusters: 2-10`, `max_iter=250`, `random_state=42`, `init=k-means++`, `algorithm=lloyd`.

On the other hand, Gaussian Mixture Models (GMM) offers a sophisticated "soft" clustering approach. Instead of just calculating distances, GMM assumes the data is composed of multiple Gaussian distributions, each with its own mean and covariance. This allows GMM to capture elliptical shapes and provide a probability score for a point belonging to various clusters, rather than a binary "yes" or "no" [38]. The training parameters were: `n_components=10`, `covariance_type=''full''`, `weight_concentration_prior_type=''dirichlet_process''`, `weight_concentration_prior=0.1`, `random_state=42`.

In standard practice, feature selection is often recommended prior to clustering to mitigate the curse of dimensionality and collinearity effects [39]. However, in this study, clustering was initially conducted as an exploratory analysis on the full standardized dataset to assess whether natural groupings emerged without prior variable pruning. The poor results obtained (Section 4.2) confirmed the necessity of dimensionality reduction and justified the subsequent application of Lasso and Elastic Net feature selection before proceeding to predictive modeling and association analysis.

3.5. Modeling Associations by Time Rolling Windows

Rolling correlation is a statistical method employed to assess the relationship between time series variables across a designated, moving temporal window. In contrast to a singular, static correlation computed for the entire dataset, a rolling correlation yields a series of correlation coefficients. This technique is utilized to illustrate the evolution of the relationship between two variables over time [40]. The fundamental aspect of this technique involves determining a window size within the time series, which delineates the relationship between variables. It offers benefits such as providing a dynamic perspective on the evolution of the relationship between two variables, and being valuable for forecasting purposes by demonstrating how prior relationships have changed [41].

Initially, it is necessary to select the appropriate correlation coefficient, considering the existence of parametric and non-parametric approaches. Based on Shapiro–Wilk normality tests applied to the collected variables, the data do not originate from a normal distribution; therefore, the Spearman non-parametric rank correlation coefficient was employed. Subsequently, defining the size of the rolling window becomes essential; in this context, multiple tests were conducted to determine the optimal window size and rolling step. It was identified that a window size of fifteen days with a rolling step of seven days effectively models the short-term relationships. Additionally, a window size of thirty days with a rolling step of ten days captures the overall long-term behavior.

The rolling correlation method generated a new time series that illustrates the behavior of relationships between the variables under analysis. It can be interpreted as a line graph where each point represents the correlation value between two variables within a specified time window. Figure 2 demonstrates an example of the rolling window correlation between the air temperature recorded in the lagoon and the temperature recorded at San Javier Airport. The correlation values are notably high, which aligns with expected physical patterns; however, in June, an anomaly in the correlation values raised concerns that one of the sensors may have been malfunctioning. Administrative decisions, such as the revision or removal of sensors, may be informed by these findings, or technical interventions such

as data imputation might be necessary. The assessment of variable associations is based on this methodology, resulting in multiple correlation values throughout the time series.

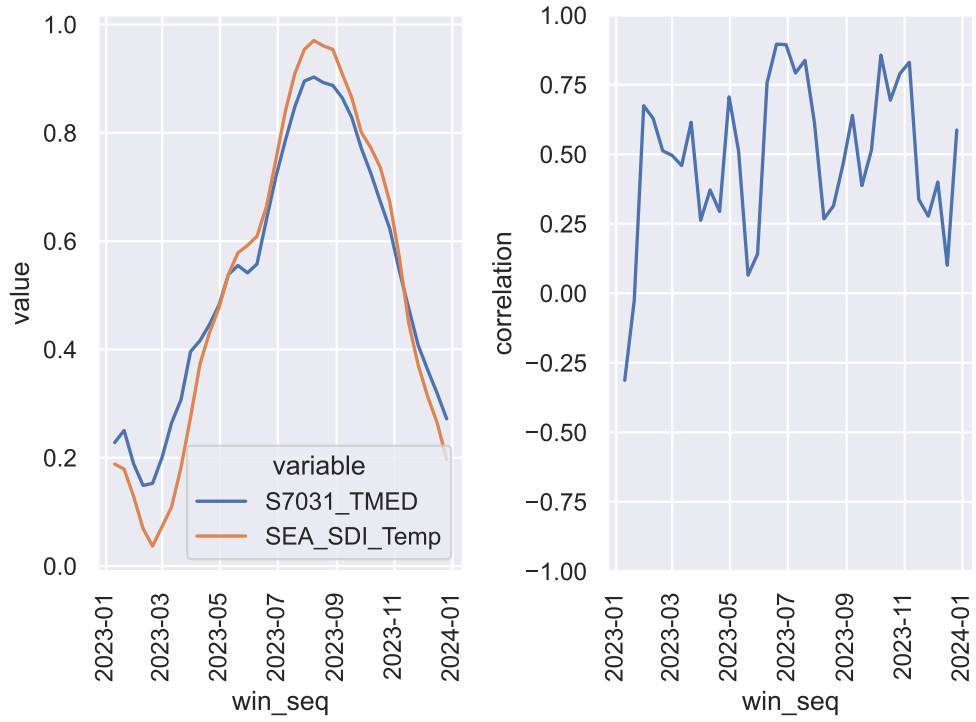


Figure 2. Associations between two variables by means of correlation windows: normalized temperature series (**left**) and rolling Spearman correlation with a 15-day window and 7-day step (**right**). The June anomaly is highlighted by the sudden correlation drop.

4. Results and Discussion

This section reports the findings obtained from the sequential analytical pipeline. The results are organized to reflect the logical progression of the methodology: supervised feature selection to reduce dimensionality, an exploratory evaluation of data structure via clustering, a time-resolved correlation analysis, and culminating in the development of predictive models for chlorophyll and turbidity. The following subsections examine each stage in turn.

4.1. Feature Selection Results

The coefficients obtained from applying Lasso and Elastic Net regression are visualized in Figure 3. Both algorithms consistently shrink several coefficients to zero, thereby eliminating redundant or non-informative variables. Notably, variables such as `AIR_TEMP_HS`, `EM085_PRECIP_MM`, and most directional wind measurements from auxiliary stations receive zero coefficients, indicating that they do not contribute significantly to predicting chlorophyll or turbidity when regularization is applied.

Table 2 reports the accuracy of the linear regularized models when predicting turbidity and chlorophyll. For turbidity, both Lasso and Elastic Net achieve comparable performance, with R^2 values around 0.10, suggesting that linear models capture only a modest proportion of the variance in turbidity. For chlorophyll, LassoCV yields an R^2 of approximately 0.00, indicating a complete failure of the linear model to explain the variance in chlorophyll-a concentration. ElasticNetCV performs marginally better ($R^2 = 0.233$), yet still explains less than a quarter of the variance.

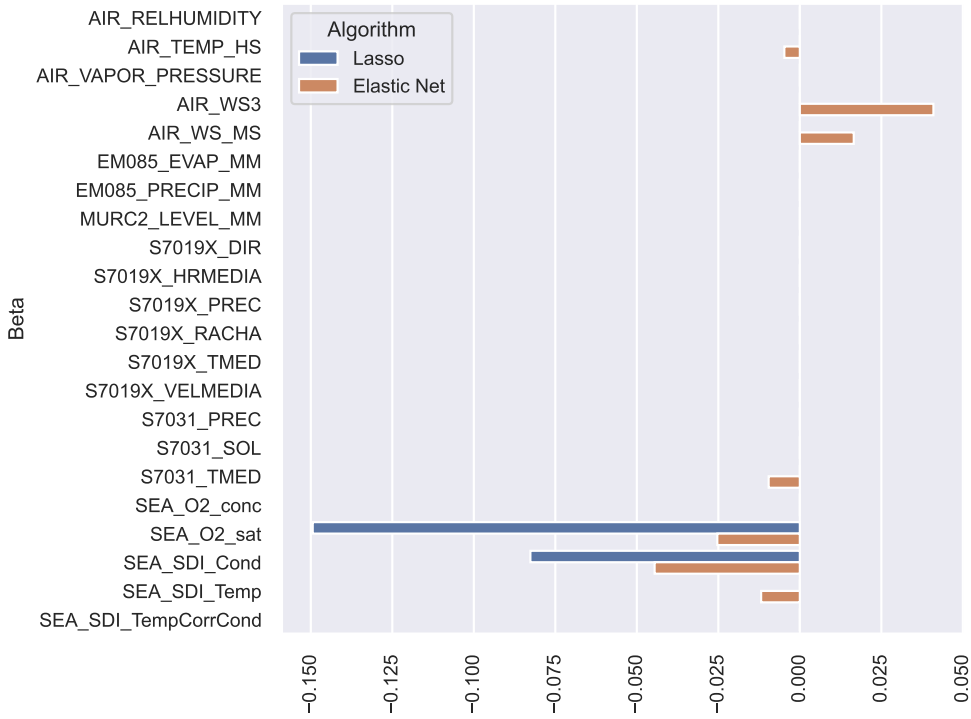


Figure 3. Coefficients of the variables calculated by Lasso and Elastic Net regularization. Features with zero coefficients are excluded from subsequent modeling stages.

Table 2. Accuracy of linear regression models in the feature selection task.

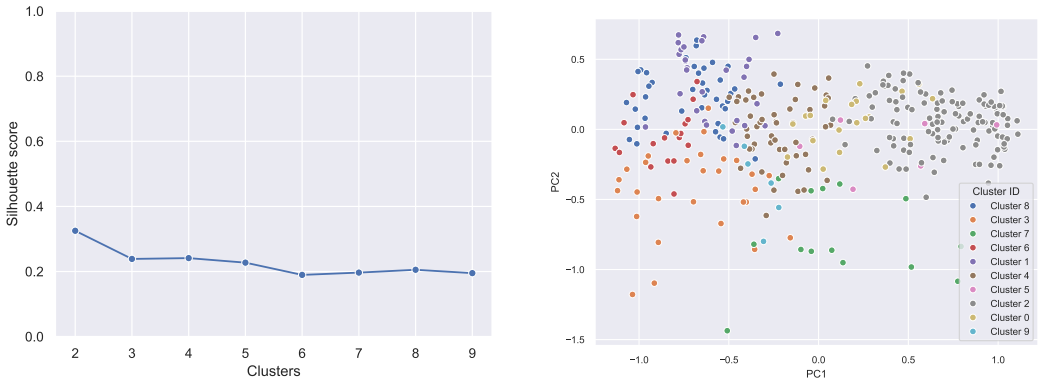
Target	Model	RMSE	MAE	R^2
Turbidity	LassoCV	0.082	0.033	0.102
	ElasticNetCV	0.082	0.033	0.101
Chlorophyll	LassoCV	0.128	0.067	0.000
	ElasticNetCV	0.112	0.063	0.233

The poor predictive performance of linear regularized models for chlorophyll suggests that the relationship between this target and the predictor set is inherently non-linear. Chlorophyll-a dynamics in coastal lagoons are governed by complex biogeochemical processes, including nutrient loading, light availability, grazing pressure, and thermal stratification, that may not be adequately captured by linear combinations of meteorological and hydrological predictors. Consequently, the feature selection stage should be interpreted primarily as a dimensionality reduction and interpretability tool rather than a predictive endpoint.

The weak performance of the linear models (R2 = 0.10 for turbidity; 0.00 and 0.233 for chlorophyll-a) agrees with earlier Mar Menor work, where the correlations between chlorophyll-a and water-quality variables were weak and motivated non-linear models [18]. A comparable gap appears in Lake Taihu, where linear regression reached R2 = 0.30 against 0.78 for XGBoost. In that study, electrical conductivity and water temperature were the most influential features, consistent with their retention in our selection [42]. In other systems, nutrients and water temperature rank highest [43], but our buoy data contain no nutrient, light or grazing information. Their absence probably caps what linear combinations of the available variables can explain.

4.2. Exploratory Clustering Results

Figure 4 presents the clustering results obtained from both algorithms applied to the full standardized dataset. Figure 4a reports the silhouette scores for K-means, which range from -1 to 1, where 1 indicates perfectly separated clusters and -1 indicates substantial overlap [44]. The observed scores, constrained between 0.2 and 0.4, fall well below the threshold for reliable partitioning, thereby confirming that K-means yields deficient groupings under the current high-dimensional setting.



(**a**) K-means clustering (**b**) Gaussian Mixture clustering

Figure 4. A comparison between clustering techniques applied to the full standardized environmental dataset.

Similarly, GMM identifies multiple clusters despite its theoretical capacity to model elliptical distributions through covariance heterogeneity. However, the resulting partition exhibits considerable fragmentation, suggesting that the pronounced collinearity among environmental variables undermines the statistical assumptions underlying both distance-based and distribution-based clustering. These findings confirm that clustering techniques must be applied with caution when variables exhibit strong mutual associations; this aligns with established recommendations advocating variable selection prior to the clustering stage [39]. The failure of exploratory clustering reinforced the decision to proceed with regularized feature selection before predictive modeling, thereby avoiding the propagation of collinearity-induced artifacts into downstream analyses.

4.3. Data Associations by Time Rolling Windows

Following feature selection, the time-rolling-window correlation was employed to reveal the dynamic correlation between lagoon variables and air quality variables. Beginning on the first day of the year, the Spearman correlation was computed within each window. Figure 5 illustrates the temporal evolution of these associations.

The heatmap reveals several physically meaningful patterns:

- Air-lagoon temperature coupling. A consistently positive association is observed between air temperature (S7031_TMED) and lagoon water temperature (SEA_SDI_Temp). This is a fundamental consequence of heat transfer, where the shallow lagoon readily absorbs heat from the surrounding atmosphere. This correlation is crucial as it serves as a process validation, confirming that the sensor readings accurately reflect standard environmental physics. The transient disruption of this correlation in June 2023 (see Figure 2) was instrumental in flagging a potential sensor calibration drift.
- Seasonal decoupling of conductivity and temperature. A dynamic association is noted between Electrical Conductivity (EC) (SEA_SDI_Cond) and water temperature, which exhibits a strong positive correlation only during the initial six months of observation. This initial association reflects the basic chemical principle that rising water

temperature increases the mobility of dissolved ions, thus increasing conductivity. However, the subsequent disappearance of this correlation strongly suggests that other environmental drivers, such as seasonal runoff, changes in evaporation rates, or biological cycles, become the dominant factors influencing salinity and EC over the long term. This seasonal decoupling challenges the assumption of stationary hydrological–atmospheric relationships and necessitates temporally adaptive management strategies.

- Wind-driven evaporative cooling. A negative association is found between Wind Power (WS3) (AIR_WS3) and both temperature and conductivity. This relationship suggests that stronger winds promote evaporative cooling at the water surface, thereby reducing the water temperature. This rapid surface cooling, potentially combined with increased vertical mixing, may also lead to a minor decrease in surface ion concentration, resulting in lower conductivity readings.
- Minor thermal influence of wind. The relatively low magnitude of the association between wind power and water temperature implies that, despite wind's measurable effect on cooling and conductivity, its overall influence on the lagoon's thermal budget is minor compared to more significant factors such as solar radiation and ambient air temperature.
- Oxygen saturation dynamics. The low negative association between water temperature and oxygen saturation (SEA_O2_sat) indicates a classic physicochemical relationship: the solubility of dissolved gases decreases as water temperature rises. This behavior, consistent with Henry's law, confirms that oxygen saturation acts as a reliable inverse proxy for thermal stress in the lagoon ecosystem.

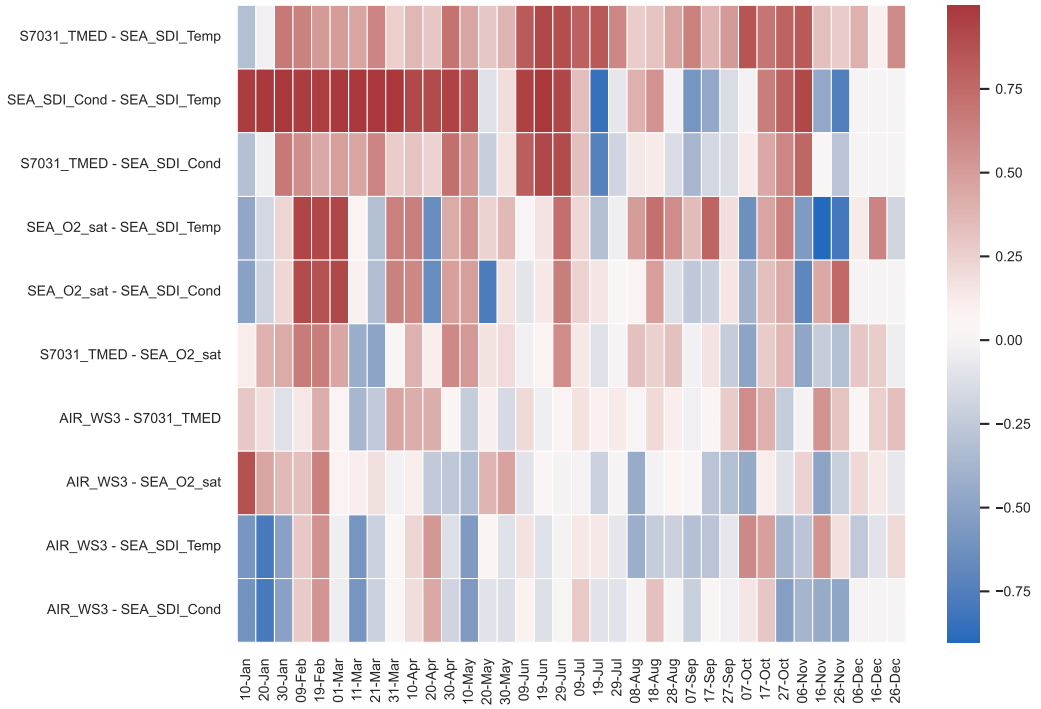


Figure 5. Time-resolved Spearman correlation heatmap: window width of 15 days and rolling step of 7 days. Red indicates positive association, blue indicates negative association, and white indicates near-zero correlation.

The strong coupling between air and lagoon temperature matches the finding that local air temperature is among the strongest correlates of Mar Menor surface temperature, and that local drivers dominate [45]. The same study points to high evaporation as a driver

of hypersalinity, which fits our interpretation that evaporation, rather than temperature, controls conductivity in the second half of 2023 [45]. The negative wind association is compatible with evaporative cooling and mixing. Wind also resuspends organic sediment and lowers dissolved oxygen in the lagoon [46], so wind effects on oxygen deserve testing in future work. The June 2023 correlation drop should be called a potential sensor fault. Flagged anomalies in buoy data can reflect real water-mass changes and need expert review [47].

4.4. Prediction of Chlorophyll and Turbidity

Based on the features selected via Lasso and Elastic Net, non-linear predictive models for chlorophyll and turbidity were constructed using Random Forest (RF), which is capable of capturing complex interactions and non-additive effects that linear regularized models failed to encode. To rigorously evaluate the benefit of feature selection, a baseline model trained on the full set of standardized features (without selection) was included for comparison. All models were evaluated using a hold-out test set comprising 30% of the data, with hyperparameters optimized via 5-fold cross-validation on the training partition.

Table 3 reports the predictive accuracy for each configuration. The baseline RF using all 24 standardized features achieves moderate performance for turbidity ($R^2 = 0.712$) but poor performance for chlorophyll ($R^2 = 0.341$), confirming that raw dimensionality introduces noise for the more biologically complex target. When feature selection is applied, predictive performance improves substantially. Using the single feature retained by Lasso, RF achieves $R^2 = 0.733$ for turbidity. Using the six features retained by Elastic Net, RF achieves $R^2 = 0.929$ for turbidity and $R^2 = 0.977$ for chlorophyll. These results demonstrate that (1) feature selection is essential for this high-dimensional environmental dataset, and (2) the non-linear nature of chlorophyll-a dynamics requires both dimensionality reduction and non-linear modeling capacity.

Table 3. Random Forest accuracy by means of selected features compared against a baseline using all features.

Target	**All Features**		**Lasso**		**ElasticNetCV**	
	Turb.	**Chlorop.**	**Turb.**	**Chlorop.**	**Turb.**	**Chlorop.**
RF Parameters						
max_depth:	10	15	10	20	10	None
min_samples_split:	2	2	5	2	2	2
n_estimators:	100	100	50	200	200	50
RMSE	0.048	0.095	0.044	0.020	0.023	0.019
MAE	0.021	0.071	0.018	0.009	0.011	0.010
$\boldsymbol{R^2}$	0.712	0.341	0.733	0.974	0.929	0.977

The Elastic Net random forest reached R^2 = 0.929 for turbidity and 0.977 for chlorophyll-a. This is above reported chlorophyll-a values for the Mar Menor, from R^2 = 0.68 [18] to 0.76 at the surface and 0.60 at 3–4 m [48], and for lakes (0.66–0.78 [49]); only one study with XGBoost approaches ours [50]. The jump from R^2 = 0.233 (linear) to 0.977 using the same six variables is large enough to warrant caution. Two features of our design could inflate it: features were selected on the full dataset before the 30% hold-out split, a recognized source of leakage [43,49], and the random split places temporally adjacent, autocorrelated samples in both training and test sets. Our rolling correlations also indicate non-stationarity, under which random cross-validation gave worse estimates than order-preserving hold-out [51]. The scores should therefore be seen as optimistic upper bounds, to be re-tested with chronological or blocked validation [7].

4.5. Synthesis of Hydrological and Atmospheric Interactions

The integrated analysis reveals that the Mar Menor Lagoon cannot be understood through static or compartmentalized models. The temporal evolution of associations, particularly the six-month decoupling of conductivity and temperature, demonstrates that atmospheric forcing modulates lagoon water quality in nonstationary ways. The feature selection stage identified a parsimonious subset of drivers (air temperature, electrical conductivity, wind power, and oxygen saturation) that carry the majority of predictive information, while the rolling correlation analysis resolved the timescales at which these couplings operate. Together, these results suggest that effective environmental stewardship of the lagoon requires adaptive monitoring frameworks that update their internal models seasonally, rather than relying on annually static correlation structures.

5. Conclusions

This study set out to characterize how hydrological and atmospheric variables interact over time in the Mar Menor lagoon and to assess their usefulness for monitoring and prediction. Using one year (2023) of multi-source data and variables measured at a depth of 3.0 m. First, linear regularization explained little of the variance in the eutrophication indicators (turbidity R^2 = 0.10; chlorophyll-a 0.00 with Lasso and 0.233 with Elastic Net), in line with earlier Mar Menor work, where weak linear relationships motivated non-linear models [18], and with lake studies where linear models trailed tree-based ones [42]. Lasso and Elastic Net were therefore most useful for reducing 24 variables to a small interpretable subset that included conductivity, a top-ranked feature in comparable studies [42]. Second, clustering the full standardized dataset gave weak partitions (silhouette 0.2–0.4), so redundant variables must be reduced before grouping [37]. Third, the relationships between variables are not constant: air and water temperature were strongly coupled, consistent with the dominance of local drivers reported for the lagoon [45]; conductivity was positively tied to temperature only in the first six months; and wind power was negatively associated with temperature and conductivity, which fits evaporative cooling and mixing. A static correlation structure would have hidden these patterns. Fourth, a drop in the air–water temperature correlation in June 2023 flagged a possible sensor fault, which still requires expert review because flagged anomalies can also reflect real water-mass events [47]. Fifth, Random Forest on the selected features reached R^2 = 0.929 for turbidity and 0.977 for chlorophyll-a, against 0.712 and 0.341 with all 24 variables. These scores exceed those reported for the Mar Menor and for lakes [18,42,43,48], and the random hold-out split and the selection of features before splitting may inflate them [51], so they are best read as upper bounds.

In practical terms, the main users of this approach are those who manage the lagoon's monitoring and decisions: the operators of the buoy and sensor network, such as its Monitoring Commission and Scientific Committee. The variables analyzed overlap with the quality elements used in WFD-type monitoring [1]. At present, the results can support two uses on retrospective data: a quality-control aid for buoy records, by watching rolling correlations between related variables, and a seasonal reading of the conductivity–temperature and wind–temperature relationships when interpreting records. They are not yet suited to an operational forecast of chlorophyll-a or turbidity, because validation was random rather than chronological, the data come from a single point at 3.0 m in one year, and nutrient and light measurements, which tree-based studies rank among the main drivers [43], were not available. Nor can they be used for the formal classification of water-body status. The 15-day window with a 7-day step also means that a drift flag may arrive up to two weeks after a problem begins, so the method suits periodic review, not real-time alarms.

Predictive use should wait for a multi-year record, chronological validation and seasonal re-estimation, since the associations changed within 2023 itself.

A further limitation concerns freshwater and nutrient inflows. Although the Mar Menor has no permanent river, its catchment (about 1244 Km^2, roughly 75% agricultural) delivers water through the Rambla del Albujón, which carries a small continuous flow, and through ephemeral streams that discharge only after rainfall. Floods are the main phosphorus input, with loads estimated at 10–20 times the regular ones, while nitrate arrives mainly through the aquifer and the Albujón [52]. Official reports describe 2023 as a dry start: rainfall was scarce in early 2023, salinity fell despite the lack of rain because of a high water table, nitrate rose from January, and chlorophyll-a stayed mostly below 2 µg/L; during the summer, nitrate entered mainly near the Albujón and phosphorus along the inner shore, and salinity rose slightly over the year [53]. Our dataset contains no direct measure of these inflows, so their effect on the results can only be inferred. Dilution or nutrient pulses could contribute to the decoupling of conductivity and temperature after six months, which we attributed to evaporation and biogeochemical processes. Such unmeasured nutrient inputs may also explain why linear models explained little of the chlorophyll-a variance and why the nonlinear gain was so large. The rainfall and evaporation records of the La Pedrera reservoir and Cabo de Palos therefore act only as regional proxies, not as inflow measurements. A year with a stronger flood-pulse regime might change the associations, so future work will add rambla discharge and nutrient-load data and will relate the dates of the highest rainfall records (Table 1) to the timing of the changes in correlation.

As a further work, Granger Causality could explicitly establish the chronological order of events, such as verifying whether the rise in wind power occurs prior to the decrease in temperature, or vice versa, thereby reinforcing the mechanistic conclusion of evaporative processes cooling. Applying this technique to the time series data would generate comprehensive evidence concerning the lead–lag relationships among meteorological, hydrological, and water quality parameters. The resulting model could subsequently forecast future water quality variations by considering not only current conditions, but also the anticipated sequence of significant environmental events.

Author Contributions: Methodology, J.F.L. and M.O.; Validation, J.-M.C.; Formal analysis, J.F.L.; Investigation, J.F.L. and M.O.; Resources, J.-M.C., C.T.C. and M.O.; Data curation, J.F.L., J.-M.C. and C.T.C.; Writing—original draft, J.F.L.; Writing—review & editing, J.F.L. and M.O.; Project administration, C.T.C. and M.O.; Funding acquisition, C.T.C. and M.O. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by research project MORELLINO (CIPROM/2023/29), funded by "Direcció General de Ciència i Investigació" Generalitat Valenciana—SPAIN, through the AIGUALERT, funded by the European Union under the European Regional Development Fund (ERDF) Programme Comunitat Valenciana 2021–2027, by IVACE+i Innovation calls for proposals, 'Strategic Cooperation Projects' (Reference INNEST/2025/486), and by the European Union's Horizon Europe project REMARKABLE under the MSCA-SE grant ID: 101086387. In addition, the Universidad del Azuay supported this research by means of the research project titled "Aplicación de la ciencia de datos en la predicción del impacto ambiental de la laguna del Mar Menor en España", reference 2025-0036.

Informed Consent Statement: Not applicable.

Data Availability Statement: The environmental datasets analyzed in this study are publicly available through the following open-data portals: the Mar Menor buoy data are provided by the Spanish Ministry for the Ecological Transition (https://www.miteco.gob.es/, accessed on 24 October 2026); AEMET meteorological data are accessible via the Open Data portal of the Spanish State Meteorological Agency (https://opendata.aemet.es/, accessed on 24 October 2026); tide gauge records (MURC2) are available from the Intergovernmental Oceanographic Commission

(https://www.ioc-sealevelmonitoring.org/, accessed on 24 October 2026); and reservoir data (Embalse de La Pedrera) are published by the Spanish Ministry of Agriculture, Fisheries and Food. Processed data and analysis scripts supporting the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: This work was partially supported by research project MORELLINO (CIPROM/2023/29), funded by "Direcció General de Ciència i Investigació" Generalitat Valenciana—SPAIN, through the AIGUALERT, funded by the European Union under the European Regional Development Fund (ERDF) Programme Comunitat Valenciana 2021–2027, by IVACE+i Innovation calls for proposals, 'Strategic Cooperation Projects' (Reference INNEST/2025/486), and by the European Union's Horizon Europe project REMARKABLE under the MSCA-SE grant ID: 101086387. The authors wish to thank the Vice-Rector for Research of the Universidad del Azuay and academic support, and the entire staff at the Computer Science Research & Development Laboratory (LIDI).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
API	Application Programming Interface
PFAS	Per- and Polyfluoroalkyl Substances
BMP	Best Management Practices
DBMS	Database Management System
RFE	Recursive Feature Elimination
GA	Genetic Algorithm
EC	Electrical Conductivity
LASSO	Least Absolute Shrinkage and Selection Operator
GMM	Gaussian Mixture Models
SA	Simulated Annealing
RF	Random Forest
SVR	Support Vector Regression
IoT	Internet of Things
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
PCC	Pearson Correlation Coefficient
CRISP-DM	Cross-Industry Standard Process for Data Mining
AEMET	State Meteorological Agency
WS3	Wind Power
O_2S	Oxygen Saturation

References

1. European Parliament and Council of the European Union. *Directive 2000/60/EC of 23 October 2000 Establishing a Framework for Community Action in the Field of Water Policy*; Official Journal of the European Communities: Luxembourg, 2000.
2. Ministerio de Medio Ambiente. *BOE-A-2001-14276*; Real Decreto Legislativo 1/2001, de 20 de Julio, por el que se Aprueba el Texto Refundido de la Ley de Aguas. Boletín Oficial del Estado: Madrid, Spain, 2001.
3. European Parliament and Council of the European Union. *Regulation (EU) 2021/1119 of 30 June 2021 Establishing the Framework for Achieving Climate Neutrality and Amending Regulations (EC) No 401/2009 and (EU) 2018/1999 ('European Climate Law')*; Official Journal of the European Union: Luxembourg, 2021.
4. Jefatura del Estado. *BOE-A-2021-8447*; Ley 7/2021, de 20 de Mayo, de Cambio Climático y Transición energética. Boletín Oficial del Estado: Madrid, Spain, 2021.
5. Kramer, L. Rights of Nature in Europe: The Spanish Lagoon Mar Menor Becomes a Legal Person. *J. Eur. Environ. Plan. Law* **2023**, *20*, 5–23. [CrossRef]

6. Juste Ruiz, J. The Spanish Law for the recognition of legal personality to the Mar Menor lagoon and its basin in the international context: Another step towards the implementation of the rights of nature. *Riv. Giuridica Dell'Ambiente* **2024**, 165–188. [CrossRef]
7. Erena, M.; Domínguez, J.A.; Aguado-Giménez, F.; Soria, J.; García-Galiano, S. Monitoring Coastal Lagoon Water Quality through Remote Sensing: The Mar Menor as a Case Study. *Water* **2019**, *11*, 1468. [CrossRef]
8. Giménez, J.G.; Granero, A.; Senent-Aparicio, J.; Gómez-Jakobsen, F.; Mercado, J.M.; Blanco-Gómez, P.; Ruiz, J.M.; Cecilia, J.M. Assessment of oceanographic services for the monitoring of highly anthropised coastal lagoons: The Mar Menor case study. *Ecol. Inform.* **2024**, *81*, 102554. [CrossRef]
9. López-Ballesteros, A.; Trolle, D.; Srinivasan, R.; Senent-Aparicio, J. Assessing the effectiveness of potential best management practices for science-informed decision support at the watershed scale: The case of the Mar Menor coastal lagoon, Spain. *Sci. Total Environ.* **2023**, *859*, 160144. [CrossRef] [PubMed]
10. Perni, A.; Martínez-Paz, J.M. A participatory approach for selecting cost-effective measures in the WFD context: The Mar Menor (SE Spain). *Sci. Total Environ.* **2013**, *458–460*, 303–311. [CrossRef] [PubMed]
11. Lamas Rodríguez, M.; Garcia Lorenzo, M.L.; Medina Magro, M.; Perez Quiros, G. Impact of climate risk materialization and ecological deterioration on house prices in Mar Menor, Spain. *Sci. Rep.* **2023**, *13*, 11772. [CrossRef] [PubMed]
12. Giménez, J.G.; González, M.; Martínez-España, R.; Cecilia, J.M.; López-Espín, J.J. Enhancing shallow water quality monitoring efficiency with deep learning and remote sensing: A case study in Mar Menor. *J. Ambient. Intell. Smart Environ.* **2025**, *17*, 182–197. [CrossRef]
13. Cisterna-García, A.; González-Vidal, A.; Martínez-Ibarra, A.; Ye, Y.; Guillén-Teruel, A.; Bernal-Escobedo, L.; Skarmeta, A.F. Artificial intelligence for streamflow prediction in river basins: A use case in Mar Menor. *Sci. Rep.* **2025**, *15*, 19481. [CrossRef] [PubMed]
14. González-Enrique, J.; RodrÍguez-GarcÍa, M.I.; Ruiz-Aguilar, J.J.; Carrasco-GarcÍa, M.G.; Enguix, I.F.; Turias, I.J. Chlorophyll-a forecasting using LSTM, bidirectional LSTM and GRU networks in El Mar Menor (Spain). *Log. J. IGPL* **2025**, *33*, jzae046. [CrossRef]
15. Delnevo, G.; Tumedei, G.; Ghini, V.; Prandi, C. Evaluating the use of machine learning algorithms in environmental sensing for energy saving. In *Proceedings of the 2nd Workshop on Networked Sensing Systems for a Sustainable Society*; ACM: New York, NY, USA, 2023; NET4us '23; pp. 201–206. [CrossRef]
16. Maté-Sánchez-Val, M.; Aparicio-Serrano, G. The impact of marine pollution on the probability of business failure: A case study of the Mar Menor lagoon. *J. Environ. Manag.* **2023**, *332*, 117381. [CrossRef] [PubMed]
17. Conesa, H.M.; Jiménez-Cárceles, F.J. The Mar Menor lagoon (SE Spain): A singular natural ecosystem threatened by human activities. *Mar. Pollut. Bull.* **2007**, *54*, 839–849. [CrossRef] [PubMed]
18. Tuia, D.; Kellenberger, B.; Beery, S.; Costelloe, B.R.; Zuffi, S.; Risse, B.; Mathis, A.; Mathis, M.W.; Van Langevelde, F.; Burghardt, T.; et al. Perspectives in machine learning for wildlife conservation. *Nat. Commun.* **2022**, *13*, 792. [CrossRef] [PubMed]
19. Zhong, S.; Zhang, K.; Bagheri, M.; Burken, J.G.; Gu, A.; Li, B.; Ma, X.; Marrone, B.L.; Ren, Z.J.; Schrier, J.; et al. Machine Learning: New Ideas and Tools in Environmental Science and Engineering. *Environ. Sci. Technol.* **2021**, *55*, 12741–12754. [CrossRef] [PubMed]
20. Jimeno-Sáez, P.; Senent-Aparicio, J.; Cecilia, J.M.; Pérez-Sánchez, J. Using Machine-Learning Algorithms for Eutrophication Modeling: Case Study of Mar Menor Lagoon (Spain). *Int. J. Environ. Res. Public Health* **2020**, *17*, 1189. [CrossRef] [PubMed]
21. Kayes, I.; Shahriar, S.A.; Hasan, K.; Akhter, M.; Kabir, M.; Salam, M. The relationships between meteorological parameters and air pollutants in an urban environment. *Glob. J. Environ. Sci. Manag.* **2019**, *5*, 265–278. [CrossRef]
22. Mallin, M.A.; Ensign, S.H.; McIver, M.R.; Shank, G.C.; Fowler, P.K. Demographic, landscape, and meteorological factors controlling the microbial pollution of coastal waters. *Hydrobiologia* **2001**, *460*, 185–193. [CrossRef]
23. López-Andreu, F.J.; López-Morales, J.A.; Atenza Juárez, J.F.; Alcaraz, R.; Hernández, M.D.; Erena, M.; Domínguez-Gómez, J.A.; García Galiano, S. Monitoring System of the Mar Menor Coastal Lagoon (Spain) and Its Watershed Basin Using the Integration of Massive Heterogeneous Data. *Sensors* **2022**, *22*, 6507. [CrossRef] [PubMed]
24. Gnana, D.A.A.; Balamurugan, S.A.A.; Leavline, E.J. Literature review on feature selection methods for high-dimensional data. *Int. J. Comput. Appl.* **2016**, *136*, 9–17. [CrossRef]
25. Araveeporn, A. The Higher-Order of Adaptive Lasso and Elastic Net Methods for Classification on High Dimensional Data. *Mathematics* **2021**, *9*, 1091. [CrossRef]
26. Schröer, C.; Kruse, F.; Gómez, J.M. A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Comput. Sci.* **2021**, *181*, 526–534. [CrossRef]
27. Wirth, R.; Hipp, J. CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, Manchester, UK, 11–13 April 2000; Volume 1, pp. 29–39.

28. Prandi, C.; Cecilia, J.M.; Manzoni, P.; Peña Haro, S.; Pierson, D.; Colom, W.; Blanco, P.; Garcìa, C.A.; Navarro, I.J.; Senent, J. On integrating intelligent infrastructure and participatory monitoring for environmental modelling: The SMARTLAGOON approach. In *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*; ACM: New York, NY, USA, 2022; GoodIT '22; pp. 236–243. [CrossRef]
29. Chen, Y.C.; Ahn, M.J.; Wang, Y.F. Artificial Intelligence and Public Values: Value Impacts and Governance in the Public Sector. *Sustainability* **2023**, *15*, 4796. [CrossRef]
30. The Pandas Development Team. *Pandas-Dev/Pandas: Pandas*, Version v3.1.0rc0, Pandas, 2020, United States. Available online: https://zenodo.org/records/23565656 (accessed on 24 October 2026).
31. El-Hasnony, I.M.; Barakat, S.I.; Elhoseny, M.; Mostafa, R.R. Improved Feature Selection Model for Big Data Analytics. *IEEE Access* **2020**, *8*, 66989–67004. [CrossRef]
32. Bolón-Canedo, V.; Alonso-Betanzos, A. Feature Selection. In *Recent Advances in Ensembles for Feature Selection*; Springer International Publishing: Cham, Switzerland, 2018; pp. 13–37. [CrossRef]
33. Obi, J.C.; Jecinta, I.C. A review of techniques for regularization. *Int. J. Res. Eng. Sci.* **2023**, *11*, 360–367. [CrossRef]
34. Khan, M.H.R.; Bhadra, A.; Howlader, T. Stability selection for lasso, ridge and elastic net implemented with AFT models. *Stat. Appl. Genet. Mol. Biol.* **2019**, *18*, 20170001. [CrossRef] [PubMed]
35. Bosson-Amedenu, S.; Ayitey, E.; Ayiah-Mensah, F.; Asare, L. Evaluating key predictors of breast cancer through survival: A comparison of AFT frailty models with LASSO, ridge, and elastic net regularization. *BMC Cancer* **2025**, *25*, 665. [CrossRef] [PubMed]
36. Golalipour, K.; Akbari, E.; Hamidi, S.S.; Lee, M.; Enayatifar, R. From clustering to clustering ensemble selection: A review. *Eng. Appl. Artif. Intell.* **2021**, *104*, 104388. [CrossRef]
37. Chormunge, S.; Jena, S. Evaluation of Clustering Algorithms for High Dimensional Data Based on Distance Functions. In *Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies*; ACM: New York, NY, USA, 2014; ICTCS '14. [CrossRef]
38. Zhuang, X.; Huang, Y.; Palaniappan, K.; Zhao, Y. Gaussian mixture density modeling, decomposition, and applications. *IEEE Trans. Image Process.* **1996**, *5*, 1293–1302. [CrossRef] [PubMed]
39. Maugis, C.; Celeux, G.; Martin-Magniette, M.L. Variable Selection for Clustering with Gaussian Mixture Models. *Biometrics* **2009**, *65*, 701–709. [CrossRef] [PubMed]
40. Zivot, E.; Wang, J. Rolling Analysis of Time Series. In *Modeling Financial Time Series with S-Plus®*; Springer: New York, NY, USA, 2003; pp. 299–346. [CrossRef]
41. Xu, X. Price dynamics in corn cash and futures markets: Cointegration, causality, and forecasting through a rolling window approach. *Financ. Mark. Portf. Manag.* **2019**, *33*, 155–181. [CrossRef]
42. Sun, G.; Zhu, W.; Qian, X.; Wei, C.; Xie, P.; Shi, Y.; Cao, X.; He, Y. Machine Learning Models for Chlorophyll-a Forecasting in a Freshwater Lake: Case Study of Lake Taihu. *Water* **2025**, *17*, 1219. [CrossRef]
43. Yun, Y.J.; Jargal, N.; An, K.G. Developing practical models for predicting chlorophyll levels in the Andong Reservoir using empirical and machine learning approaches based on integrated analysis of physical, chemical, and hydrological indicators. *Ecol. Indic.* **2024**, *167*, 112659. [CrossRef]
44. Shutaywi, M.; Kachouie, N.N. Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering. *Entropy* **2021**, *23*, 759. [CrossRef] [PubMed]
45. Amos, C.L.; Kassem, H.; Martínez-Alvarez, V.; Al Rashidi, T. A Heat Budget of the Mar Menor Lagoon, Spain. *Water* **2026**, *18*, 533. [CrossRef]
46. Morote-Sánchez, B.; Colomer, J.; Soler, M.; Gilabert, J.; Rodellas, V.; Garcia-Orellana, J.; Serra, T. Aquatic deoxygenation associated with resuspension of anthropogenic organic matter. *Water Res.* **2025**, *278*, 123327. [CrossRef] [PubMed]
47. Shao, W.; Ning, C.; Ma, B.; Li, C.; Li, H.; Yao, Z.; Zeng, L. Intelligent quality control of ocean buoy profile data using a GRU-mean teacher framework. *Front. Mar. Sci.* **2025**, *12*, 1661373. [CrossRef]
48. Martínez-Ibarra, A.; González-Vidal, A.; Cánovas-Rodríguez, A.; Skarmeta, A.F. Chlorophyll-a Mapping and Prediction in the Mar Menor Lagoon Using C2RCC-Processed Sentinel 2 Imagery. *arXiv* **2026**, arXiv:2510.09736.
49. Lones, M.A. Avoiding common machine learning pitfalls. *Patterns* **2024**, *5*, 101046. [CrossRef] [PubMed]
50. Han, J.W.; Cho, Y.J.; Ryu, I.G.; Park, J.H.; Kang, T.; Koo, J. Explainable AI-based modeling of chlorophyll-a using water quality and DOM characteristics: Toward interpretable prediction and actionable management strategies. *Ecol. Inform.* **2026**, *94*, 103650. [CrossRef]
51. Cerqueira, V.; Torgo, L.; Mozetič, I. Evaluating time series forecasting models: An empirical study on performance estimation methods. *Mach. Learn.* **2020**, *109*, 1997–2028. [CrossRef]

52. Pérez-Martín, M.A. Understanding Nutrient Loads from Catchment and Eutrophication in a Salt Lagoon: The Mar Menor Case. *Water* **2023**, *15*, 3569. [CrossRef]
53. Comunidad Autónoma de la Región de Murcia (CARM). *Informe de Situación del Mar Menor, Primer Trimestre de 2024*; Technical report; Canal Mar Menor, Comunidad Autónoma de la Región de Murcia: Murcia, Spain, 2024.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.