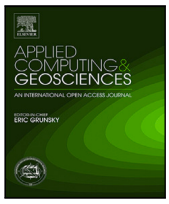
Applied Computing and Geosciences 30 (2026) 100345

Contents lists available at ScienceDirect

Applied Computing and Geosciences

journal homepage: www.sciencedirect.com/journal/applied-computing-and-geosciences



An interpretable machine learning approach for alkalinity reconstruction in the Mediterranean Sea

Teresa Tonelli [a,b,*], Gloria Pietropolli [b], Luigi Rovito [b], Luca Manzoni [a,b], Gianpiero Cossarini [a]

[a] National Institute of Oceanography and Applied Geophysics - OGS, Borgo Grotta Gigante 42/c, Sgonico, 34010, Trieste, Italy
[b] Department of Mathematics, Informatics and Geosciences, University of Trieste, Via Alfonso Valerio 12/1, Trieste, 34127, Trieste, Italy

ARTICLE INFO

Keywords:
Alkalinity
Mediterranean Sea
Machine learning
Symbolic regression
Genetic programming
Artificial intelligence

ABSTRACT

Ocean acidification has significant impacts on marine ecosystems and human activities, and its understanding relies on an accurate characterization of the marine carbonate system, in which alkalinity plays a central role.

We propose a Machine Learning (ML) approach based on Genetic Programming (GP) to model alkalinity and apply this framework to the surface layers of the Mediterranean Sea. Our framework produces interpretable equations that capture alkalinity typical patterns and its finer-scale variability by inferring its relation with key physical and biogeochemical variables.

Results, supported by quantitative metrics and visual analyses, demonstrate that our method reliably reproduces the spatio-temporal variability of alkalinity with a high level of predictive accuracy when compared with in situ observations. Moreover, we use the derived alkalinity equations to produce gap-free 2D surface alkalinity maps using satellite data. The maps correctly capture spatial gradients, seasonal patterns, and riverine contributions, reinforcing the robustness of the proposed approach.

1. Introduction

In recent years, ocean acidification, driven by the uptake of atmospheric carbon dioxide resulting from increased fossil fuel emissions, has been affecting marine ecosystems, with harmful consequences for environment and socio-economic systems (Kapsenberg et al., 2017; Rodrigues et al., 2013). Alkalinity, defined as the seawater's ability to buffer acidification (Zeebe and Wolf-Gladrow, 2001), is a fundamental variable for characterizing the carbonate system and improving our understanding of acidification.

Although its modeling is essential, the accuracy of alkalinity is influenced by the intrinsic complexity of the underlying processes and by the availability of descriptive measurements. In situ observations provide a sparse and discontinuous view of alkalinity spatial and temporal variability (Cossarini et al., 2015; Schneider et al., 2007; Touratier et al., 2012), limiting our capability to accurately capture its dynamics (Gray et al., 2024; Sonnewald et al., 2021; Zhang et al., 2025). To address this, two main modeling approaches have been developed: dynamical models, which simulate alkalinity through transport and biogeochemical processes and require boundary conditions and observational calibration (Baker and Brezonik, 1988; Bergström et al., 1985), and data-driven models, which infer alkalinity from relationships with physical and biogeochemical variables (Beibei et al., 2025; Michałowski and Asuero, 2012). Unlike dynamical models with explicitly defined equations, these approaches rely on algorithms that autonomously learn such relationships from data.

Overall, approaches using salinity to predict alkalinity are the most widely applied (Cossarini et al., 2015; Schneider et al., 2007). Alkalinity and salinity often show similar spatial variability in surface waters (Copin-Montégut, 1993; Wolf-Gladrow et al., 2007), as freshwater addition or removal controls alkalinity by diluting or concentrating its contributing compounds and simultaneously affects salinity (Cossarini et al., 2015; Kapsenberg et al., 2017). As a result, many studies rely on salinity-based numerical models for alkalinity prediction (Cossarini et al., 2015; Schneider et al., 2007). However, salinity cannot capture biological processes that also modify alkalinity, such as nutrient uptake, mineralization, nitrification, and denitrification (Wolf-Gladrow et al., 2007).

For the Mediterranean Sea, several salinity-based regression relationships have been proposed (Copin-Montégut, 1993; Huertas et al.,

* Corresponding author at: Department of Mathematics, Informatics and Geosciences, University of Trieste, Via Alfonso Valerio 12/1, Trieste, 34127, Trieste, Italy.
E-mail addresses: teresa.tonelli@phd.units.it (T. Tonelli), gloria.pietropolli@units.it (G. Pietropolli), luigi.rovito@units.it (L. Rovito), lmanzoni@units.it (L. Manzoni), gcossarini@ogs.it (G. Cossarini).

https://doi.org/10.1016/j.acags.2026.100345
Received 27 November 2025; Received in revised form 25 March 2026; Accepted 25 March 2026
Available online 2 April 2026

2009; Schneider et al., 2007). These models highlight regional differences in the salinity–alkalinity correlation, showing a positive correlation where evaporation–dilution processes prevail (Schneider et al., 2007) and a negative one in areas influenced by freshwater input, such as rivers (Luchetta et al., 2010). Such contrasts make a single unified model ineffective, and the literature therefore recommends partitioning regions with differing correlations and developing separate models for each macro-area. In the Mediterranean Sea, this leads to treating the Adriatic Sea separately from the rest of the basin, as done in several studies (Copin-Montégut, 1993; Cossarini et al., 2015; Giani et al., 2023; Schneider et al., 2007), which consistently show that the Adriatic Sea has distinct characteristics, making a single unified model ineffective.

The growing volume of in situ and satellite ocean observations, and their central role in bringing model predictions closer to reality, have opened the way to data-driven tools. In this context, ML has advanced rapidly in recent decades and has become a powerful approach for handling the complexity and multi-dimensionality of oceanographic data, supporting improved estimates and forecasts in operational oceanography (Amadio et al., 2024; Dong et al., 2022; Lou et al., 2023; Mittal et al., 2022; Pietropolli et al., 2022, 2024, 2025; Sammartino et al., 2020; Tonelli et al., 2026). ML algorithms are now widely used for different applications, including modeling ocean turbulence (Zanna and Bolton, 2021), predicting surface temperature (Lyman and Johnson, 2023), and increasing models resolution (Ducournau and Fablet, 2016).

EAs are a subset of ML which solve optimization problems by mimicking the process of natural evolution. Instead of manually designing a solution, they iteratively improve a population of candidate solutions by evaluating their quality through a fitness function and applying genetic operations to generate increasingly effective solutions. Its interpretability and effectiveness make EA attractive, with applications across engineering, computer science, computer vision, scheduling, and biomedical (Arif et al., 2024; Nakane et al., 2020; Slowik and Kwasnicka, 2020; Zhan et al., 2022). Despite this broad diffusion, only a few studies have applied EAs in oceanography (Álvarez et al., 2002; Fonlupt, 2001; Gaur and Deo, 2008).

This work proposes a new approach to modeling alkalinity based on GP, a type of EA that produces interpretable, human-readable solutions. Recent advances in Symbolic Regression (SR) show that GP-based algorithms often outperform other symbolic regression methods (He et al., 2022; Huynh et al., 2022; Langdon et al., 2008; Radwan et al., 2024).

In this paper, the relationship between alkalinity and a broader set of physical and biogeochemical variables is investigated using an unprecedentedly wide data distribution. We apply this new approach to the Mediterranean Sea, a landlocked and highly dynamic ecosystem characterized by strong coastal anthropic pressure, a wider alkalinity range than the Atlantic, and a high capacity to absorb and buffer anthropogenic CO^2 (Cossarini et al., 2015; Kapsenberg et al., 2017). Following previous studies (Copin-Montégut, 1993; Cossarini et al., 2015; Giani et al., 2023), we divided the basin into two areas, i.e., the Adriatic Sea and the rest of the Mediterranean, to account for their distinct alkalinity–salinity dynamics.

We benchmark our approach against two ML baselines: Multi-Layer Perceptron (MLP) and Linear Regression (LR), allowing us to assess predictive accuracy and interpretability. The MLP achieves the lowest errors in both regions, but its internal mechanisms are not directly accessible. In contrast, GP produces explicit analytical equations that can be inspected from an oceanographic perspective, with resulting models clearly exposing the relationship between input variables and alkalinity.

2. Dataset

Our approach is implemented in the Mediterranean Sea, where a quality-checked in situ dataset for the carbonate system exists and a rich collection of numerical models and satellite observations is available from the Marine Copernicus Service (Cossarini et al., 2019; Salon et al., 2019). Specifically, we use the MedBGCins dataset (Di Biagio et al., 2025), a dataset that collects in situ measurements for the period 1995–2023, integrating the "Mediterranean Sea - Eutrophication and Acidity aggregated datasets" (Hellenic Centre for Marine Research et al., 0000) with other documented oceanographic cruises. The dataset contains more than 5000 alkalinity measurements with other variables such as temperature, salinity and nutrients which are used as explanatory parameters in the model. Measurements number decreases to about 800 when only the surface layer (i.e., 0–20 m) is considered.

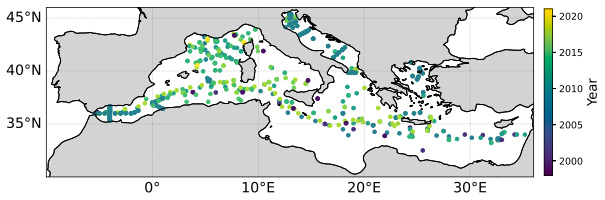


Fig. 1. Distribution of alkalinity data. Points are colored according to the year of observation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Along with MedBGCins, we use datasets of satellite observations from the Marine Copernicus Service. In particular, chlorophyll (OC-CNR-ROMA-IT, 2023), surface temperature (Copernicus Marine Service, 2024), and salinity (Copernicus Marine Service, 2023) are used both as explanatory variables and to reconstruct alkalinity spatial maps using the developed model. Satellite data have daily frequency and a spatial resolution of 1 km (1/8° in the case of salinity) and observations are extracted at the same location (longitude, latitude) and time of the in situ surface measurements to be paired with alkalinity in situ measurements.

Sample spatial distribution is reported in Fig. 1.

3. Background

Symbolic Regression (SR) is a regression approach that searches the space of mathematical expressions to identify models that best describe a dataset, yielding interpretable and human-readable equations (Angelis et al., 2023). Unlike conventional regression methods (Su et al., 2012; Ostertagová, 2012), which rely on fixed functional forms and optimize only coefficients, SR simultaneously discovers both the structure of the equation and its parameters (La Cava et al., 2021). Because identifying the exact expression is often intractable, SR is typically solved using approximate methods, most notably Genetic Programming (GP). Since GP belongs to the broader family of Evolutionary Computation (EC) techniques, we first introduce the core concepts of EC (Section 3.1), then outline the general GP paradigm (Section 3.1.1), and finally describe the specific GP variant employed in this study (Section 3.1.2). The section concludes with the Multi-Layer Perceptron (MLP) and Linear Regression (LR), used as baseline methods for model comparison (Section 3.2).

3.1. Evolutionary computation

Evolutionary Computation (EC) draws inspiration from the principles of Darwinian evolution to solve complex optimization and search problems (Bacardit et al., 2022; Back and Schwefel, 1996; Bäck et al., 1997; Bartz-Beielstein et al., 2014; Leporati et al., 2023). At its core, EC encompasses a family of algorithms, known as Evolutionary Algorithms (EAs), that iteratively improve a population of candidate solutions by simulating natural evolution processes, mimicking mechanisms such as selection, crossover, and mutation. To guide the evolution, the quality of each individual is evaluated using a fitness function f.

Algorithm 1 Generic EA (EA)

1: Initialize a population P of individuals
2: **while** termination criterion not met **do**
3: Evaluate fitness function f on each individual in P
4: Select individuals for the mating pool based on their fitness
5: Apply crossover and mutation to create a new population
6: Replace P with the new population
7: **end while**
8: **return** the best-so-far individual

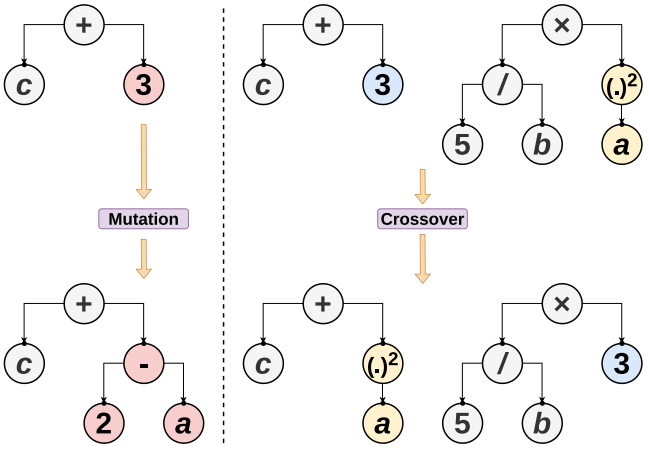


Fig. 2. Examples of GP mutation and crossover.

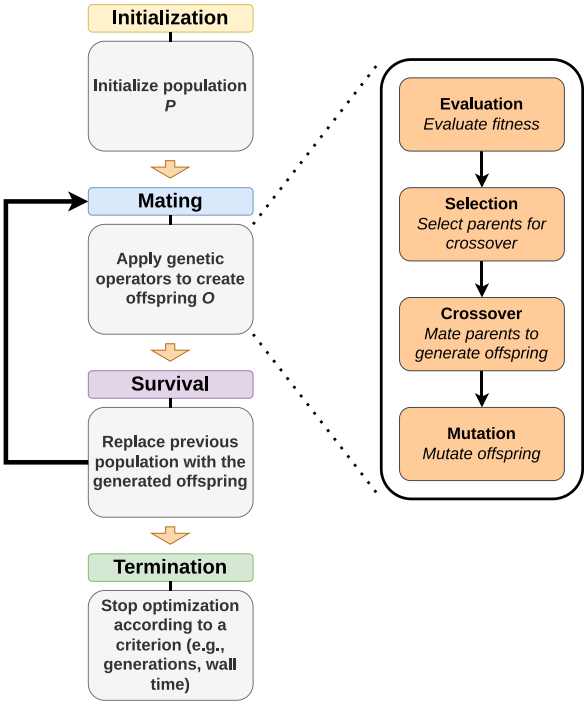


Fig. 3. Example of a generic EA optimization cycle.

An individual represents a solution $s \in S$ within the search space S, and the population $P \subseteq S$ denotes the set of evolved individuals. Selection chooses which individuals undergo variation through crossover and mutation. Crossover recombines parts of selected individuals to produce offspring, while mutation introduces random modifications to preserve diversity and explore new solutions. These operations are applied with probabilities $p_{\texttt{cross}}$ and $p_{\texttt{mut}}$, respectively. A generation corresponds to one complete cycle of evaluation, selection, crossover, and mutation. The evolutionary process continues for multiple generations until predefined termination criteria are met (e.g., fitness evaluations, execution time).

An outline of a generic EA is detailed in Algorithm 1.

Due to its consolidated effectiveness in Symbolic Regression (SR) problems (Angelis et al., 2023; Huynh et al., 2022), we adopt a Genetic Programming (GP) implementation (M. Cranmer, 2023; M.D. Cranmer, 2023; Tonda, 2024).

3.1.1. Genetic programming

GP is an EA where individuals represent computer programs (Banzhaf et al., 2000; Koza, 1994; Langdon et al., 2008; Poli et al., 2007). Thanks to its flexibility and expressive power, GP has been successfully applied across a wide range of domains, including SR, automated design, and modeling complex systems (Brabazon et al., 2020; Ferreira et al., 2020; Marchetti et al., 2024; Pietropolli et al., 2023b; Bonin et al., 2024; Rovito et al., 2025; Tonelli et al., 2024; Zhang et al., 2021)

Here, individuals are encoded as trees, where internal nodes represent functions, and terminal nodes represent inputs or constants (Koza, 1995). During the evolution, GP iteratively transforms individuals by applying tree-specialized genetic operations (Langdon et al., 2008). A GP algorithm shares the structure of Algorithm 1 while it differs in the individuals representation and, consequently, in how operations such as initialization, crossover, and mutation are formalized (Langdon et al., 2008).

Trees are initialized by assigning operations from the function set $\mathcal{F}$ to internal nodes and terminals from $\mathcal{T}$ to leaves. Although $\mathcal{T}$ includes all input variables, the evolutionary process implicitly performs feature selection, and not all variables necessarily appear in the final GP solution. Tree depth, which controls individual complexity, is defined as the length of the longest path from the root to any leaf.

Crossover and mutation are defined to both operate on trees: *sub-tree crossover* (Fig. 2, left panel) switches and recombines sub-trees of different individuals, while *sub-tree mutation* (Fig. 2, right panel) substitutes selected sub-trees with randomly generated ones. The GP evolution process is outlined in Fig. 3.

3.1.2. Island model GP

Island Model GP (Martin, 1997; Whitley et al., 1997, 1999) represents an extension of traditional GP, developed to avoid premature convergence and maintain diversity (Izzo et al., 2012; Duarte et al., 2017). Its core idea is the partition of the population into several subpopulations, called islands. Each island evolves independently, running a standard GP. The islands communicate through migration, a process that periodically transfers a few solutions from one island to another, promoting the interaction between candidate solutions, preventing premature convergence and widely exploring solution space (Izzo et al., 2012). To better lead and control the evolution, the migration process is generally ruled by a policy, which defines the rules for the solution exchanges. Different parameters impact the island models effectiveness,

Algorithm 2 Island Model GP

```
Initialize a population P of individuals
for island i = 1 to M do do in parallel
    for generation = 1 to G do
        Run the genetic programming loop
        if generations then
            Select migrants from island i
            Send migrants to destination islands j based on topology
            Receive migrants from source islands j
            Integrate migrants into island i
        end if
    end for
end for
return best individual(s) from all islands
```

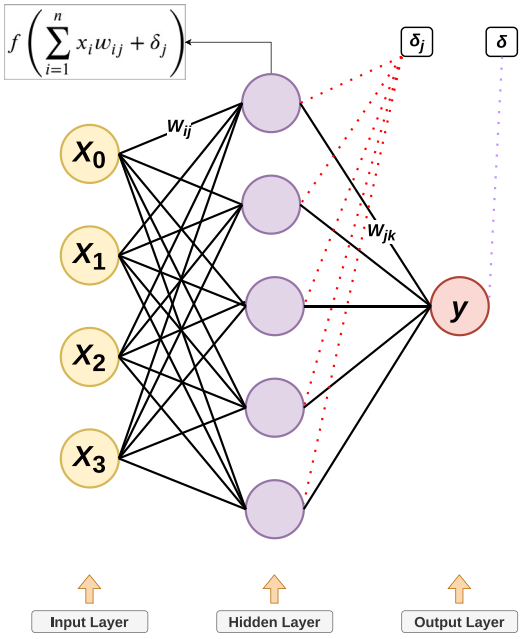


Fig. 4. MLP internal structure.

such as the number of islands N_i, the migration topology $\mathcal{G}$, the migration rate m and the migration frequency f_m.

An outline of the island model GP structure is detailed in Algorithm 2.

3.2. Linear regression and multi-layer perceptron

Linear Regression (LR) (Su et al., 2012) explicitly models the relationship between a dependent variable (target) and one or more independent variables (features) by fitting a hyperplane to the data. Given a features set $x_1, \ldots, x_n$ and a target variable y, the linear regression model estimates the optimal coefficients $\beta_0, \ldots, \beta_n$ that minimize the discrepancy between the predicted and actual values of y, according to the linear model

$$y = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_n \cdot x_n + \epsilon$$

where ϵ represents the error term.

Multi-Layer Perceptron (MLP) is a widely used Artificial Neural Network (ANN) for prediction and forecasting (Dong et al., 2022; Zhao et al., 2024; Sauzède et al., 2017; Bittig et al., 2018; Fourrier et al., 2020; Pietropolli et al., 2023a). Its structure includes an input layer, an output layer, and typically one or more hidden layers composed of neurons, as shown in Fig. 4. Despite its performance, the internal decision process of an MLP cannot be easily inspected, making it a *black-box* model.

Each neuron adjusts its internal parameters, or weights (w_{ij}), during training to approximate complex input–output relationships. Within each layer, a neuron computes a weighted sum I_j of its inputs x_i and a bias term δ_j, followed by a non-linear activation function f:

$$I_j = f\left(\sum_{i=1}^{n} x_i w_{ij} + \delta_j\right).$$

This operation is iteratively applied across layers to produce the output y. The optimal weights w_{ij} are found through continuous optimization combining backpropagation (Rumelhart et al., 1986) and gradient descent, which updates weights as

$$w_{ij}^{\text{new}} = w_{ij}^{\text{old}} - \eta \frac{\partial L}{\partial w_{ij}^{\text{old}}},$$

where η is the learning rate.

4. Proposed method

In this section, we present the GP island model used for SR to model surface alkalinity in the Mediterranean Sea.

4.1. Data pre-processing

The input consists of in situ oceanographic measurements, including physical variables (temperature and salinity), a biogeochemical variable (chlorophyll), and their spatio-temporal coordinates (month, latitude, longitude). These variables are selected based on their availability and their documented relevance to alkalinity (Cossarini et al., 2015; Gemayel et al., 2015; Lee et al., 2006; Millero et al., 1998; Wallace, 1995). In particular, chlorophyll acts as a proxy for biological activity influencing the carbonate system (Champenois et al., 2025; Dash et al., 2022), especially in coastal environments, and captures spatial patterns associated with riverine fertilization, which is a key driver of alkalinity variability in regions such as the Adriatic Sea (Copin-Montégut, 1993).

To reconstruct the surface alkalinity distribution, we restrict the dataset to superficial layers (0–20 m), consistently with previous studies (Copin-Montégut, 1993; Cossarini et al., 2015; Schneider et al., 2007). To incorporate temporal information, the month of the data sampling is encoded through categorical encoders.

The dataset is randomly split into train and test sets, ensuring a balanced distribution across all sub-basins for a fair evaluation of unseen data (Farias et al., 2020; Kratzert et al., 2024).

To address the cardinality and the non-uniform spatial coverage of the Mediterranean basin, we apply data augmentation (Chawla et al., 2002; Shorten and Khoshgoftaar, 2019) to increase the number of samples, particularly in underrepresented areas (see Fig. 1). The procedure generates synthetic observations by interpolating between

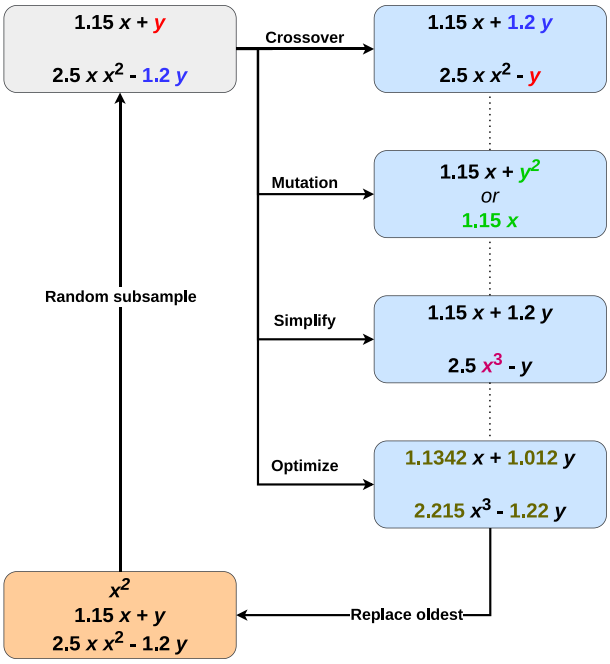


Fig. 5. Structure of the evolve–simplify–optimize cycle.

existing samples in the minority regions. For each minority sample x, the algorithm selects its k nearest neighbors within the same group and creates a new synthetic point along the segment connecting x to one neighbor

$$x_{nn} : x_{\texttt{synthetic}} = x + \delta \cdot (x_{nn} - x), \text{with } \delta \in \{0, 1\},$$

following the formulation introduced in Chawla et al. (2002).

Finally, because the ranges of the input variables differ considerably, we normalize all inputs (Ali et al., 2014) to avoid bias toward variables with larger magnitudes. The mean μ and standard deviation σ are computed on the training set and applied to both training and test data, as follow:

$$x_{\texttt{norm}} = \frac{x - \mu}{\sigma}.$$

4.2. GP for alkalinity prediction

The proposed algorithm follows an island-model structure, with additional components introduced to tailor it to the alkalinity reconstruction task.

The internal search follows an evolve–simplify–optimize loop, illustrated in Fig. 5. After generating new candidate solutions through mutation and crossover, the algorithm simplifies a subset of them into equivalent but less complex expressions. Only part of the population is simplified so that more complex solutions can still emerge during the search. Then, the algorithm adopts the BFGS optimization algorithm (Broyden, 1970) to explicitly optimize constants, making the approach more efficient for real application problems (M. Cranmer, 2023).

Our model adopts a shared global Pareto front (Kang et al., 2024; Smits and Kotanchek, 2005), which all populations can contribute to and access. Instead of selecting solutions based solely on their fitness, the algorithm also accounts for model complexity. This multi-objective approach is adopted because interpretability is a key requirement for our application. The resulting Pareto front collects the best trade-off solutions between effectiveness and complexity, providing the best-performing interpretable expression available at each complexity level.

Table 1
Input variables and their aliases used for each geographical area to predict alkalinity (A). `is_Season` denotes a placeholder naming convention, with Season substituted by the specific season (e.g., `is_Spring`, `is_Summer`, `is_Autumn`, `is_Winter`).

Variables	Aliases
Seasons	`is_Season`
Latitude	ϕ
Longitude	λ
Salinity	S
Temperature	T
Chlorophyll	C

Table 2
PySR hyperparameters used in our experiments.

Hyperparameter	Value
N° Islands N_i	50
Island size P_i	10^3
N° Generations G	10^5
Max equations complexity	20
Max tree depth	8
Tournament size t	15
Crossover probability p_{cross}	0.05
Function set $\mathcal{F}$	$+$, $-$, $*$, $/$, x^2, $\frac{1}{x}$

Table 3
MLP hyperparameters used in our experiments.

Hyperparameter	Value
Epochs	200
Snaperiod	25
Optimizer	Adam
Learning rate	0.001
Layers dimension	[3, 64, 16, 1]
N° hidden layers	3
Normalization	Layer normalization
Activation function	ReLU

Finally, we refined expression complexity (measured by counting the number of nodes in the tree) by assigning different weights to operators, variables, and constants. Instead of treating all nodes as equally costly, complexity is computed as a weighted sum, encouraging expressions built from operators that are more meaningful for the problem. In addition, the algorithm includes constraints that prevent excessively nested structures (for example, long chains of repeated $\frac{1}{x}$ operations), ensuring that the resulting expressions remain interpretable.

5. Experimental settings

This section describes hyperparameter and experimental settings, source code is available at https://github.com/TeresaTonelli/GP-4-Alkalinity.

We implemented the preprocessing pipeline using `scikit-learn` (Pedregosa et al., 2011) and the `imblearn` library (LemaÃŽtre et al., 2017) for data augmentation. The dataset was split into training and test sets using a 70:30 ratio (Vrigazova, 2021). We applied the SMOTE algorithm (Chawla et al., 2002) with an initial $k = 5$, and reduced k when too few samples were available in under-represented geographic regions. One-hot encoding was applied to the time categorical variable. Z-score normalization was then performed using the mean and standard deviation computed from the training set (Patro and Sahu, 2015). We list the input variables in Table 1. The same set of variables, selected a priori based on relevant literature, is used across all methods to ensure a fair comparison.

To run the island-based GP, we used the PySR library (M.D. Cranmer, 2023; Tonda, 2024), which provides a genetic-programming island model for SR with a Python interface and a high-performance

Table 4
Alkalinity models computed by the proposed GP algorithm in the Mediterranean and Adriatic Sea. RMSE values refers to the test set. All variables in the equations, together with their aliases, are listed in Table 1. The variable "is_Spring" holds 1 if the measurement was recorded during spring season, 0 otherwise.

GP alkalinity models			
Mediterranean Sea		Adriatic Sea	
Model equation	RMSE	Model equation	RMSE
$2571.91 + 74.36 \cdot S + 4.45 \cdot \phi$	20.21	$2649.02 + 1.02 \cdot C + 15.55 \cdot \phi$	17.26
$2563.41 + 74.63 \cdot S - 7.48 \cdot T - \lambda^2$	18.23	$2653.13 - 7.25 \cdot S - 7.25 \cdot \lambda - 0.99 \cdot T$	16.58
$2566.02 + 80.04 \cdot S + 7.61 \cdot \phi$	20.34	$2654.07 - 3 \cdot S^2 - 13.76 \cdot S - 2 \cdot T \cdot S + 4 \cdot \phi \cdot S + T^2$	15.98
$2583.31 + 75.09 \cdot S + 1.01 \cdot \texttt{is\_Spring} - \frac{1}{C}$	19.06	$2638.96 - 4.66 \cdot S + 4.66 \cdot T^2 + 4.66 \cdot \phi - 4.66 \cdot \lambda$	16.11

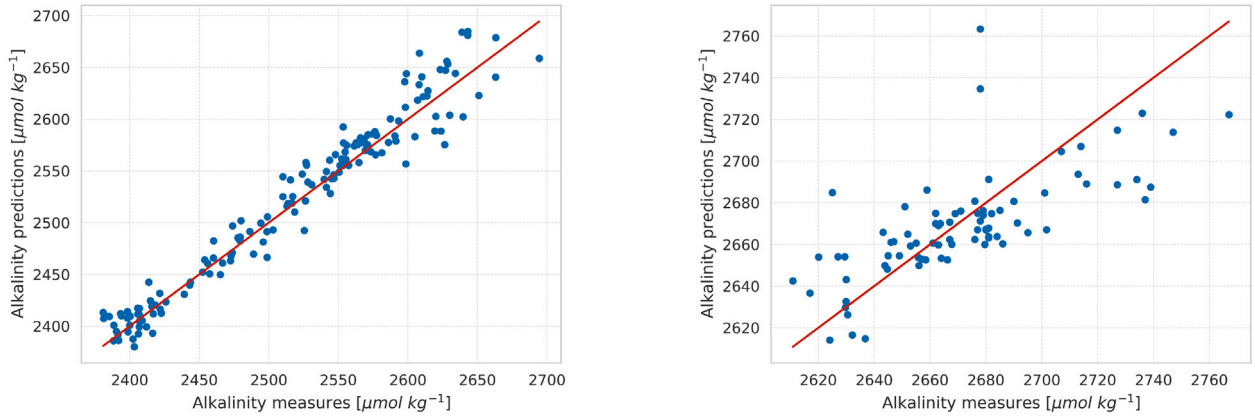


Fig. 6. Scatterplot of measured (x-axis) versus predicted (y-axis) alkalinity for the Mediterranean Sea (left panel) and the Adriatic Sea (right panel). The red line indicates the identity.

Julia backend (M. Cranmer, 2023). We used Root Mean Squared Error (RMSE) as fitness function. GP hyperparameters are summarized in Table 2. We ran 50 populations (N_i) asynchronously, each with 10^3 individuals (P_i), for 10^5 generations (G). To preserve interpretability, we constrained equation complexity to a maximum of 20 and limited tree depth to 8. The unary operators x^2 and $\frac{1}{x}$ were included in the function set to capture relevant nonlinear relationships (e.g., inverse dependencies) between inputs and alkalinity. To further promote simpler expressions, we doubled the complexity cost of unary operators, reducing their probability of being selected when they are not necessary.

Most GP hyperparameters in PySR were left at their default settings. Tree depth and model complexity were limited to prevent overly large symbolic expressions. Population size, number of islands, and number of generations were set according to an initial hyperparameter tuning.

We compared our approach with two baseline Machine Learning (ML) models: MLP (Delashmit et al., 2005; Pinkus, 1999; Zhang, 2002) and LR (Maulud and Abdulazeez, 2020; Su et al., 2012). The MLP used for comparison includes three hidden layers, followed by a linear layer, a normalization layer (Ba et al., 2016), and a ReLU activation function (Agarap, 2018; Dubey and Jain, 2019):

$$\mathrm{ReLU}(x) = \max(0, x) \tag{1}$$

The network is optimized using Adam with a learning rate of 0.001 (Bock and Weiß, 2019). The size of the MLP hidden layers was tuned through brief preliminary tests. All hyperparameters, tuned to maximize predictive performance, are reported in Table 3.

To obtain statistically robust results all methods were run 30 times. We applied a Mann–Whitney U test (Mann and Whitney, 1947) with Holm–Bonferroni correction (Holm, 1979) to compare the error distributions with a significance threshold of $\alpha = 0.05$.

6. Results

This section presents the results of superficial alkalinity modeling.

Table 5
Table detailing 95 % Confidence Intervals (CIs).

Mediterranean Sea		Adriatic Sea	
Method	CI	Method	CI
GP	[20.66–22.30]	GP	[16.72–17.67]
MLP	[16.94–19.56]	MLP	[15.51–16.41]
LR	[21.09–22.85]	LR	[17.47–18.42]

Mediterranean Sea. In the Mediterranean Sea, the best model found is described by the following equation:

$$A = 2565.14 + 74.98 \cdot S + 8.02 \cdot \phi \tag{2}$$

Eq. (2) exhibits the lowest test RMSE (18.06 μmol kg^{-1}) within the subset of models characterized by lower complexity; additional models found are summarized in Table 4. Here, the GP algorithm computes equations that describe alkalinity by adding spatio-temporal, physical, and biogeochemical relations to the traditional salinity–alkalinity one. While physical and biogeochemical variables directly influence alkalinity (Copin-Montégut, 1993; Cossarini et al., 2015; Gemayel et al., 2015), the spatial coordinates act as proxies for underlying processes, such as latitude climate gradient relation, that are not directly measured but are correlated with location-specific features of the basin (Legendre and Legendre, 2012). The presence of the salinity linear term in all alkalinity equations confirms and reinforces its strength as a good alkalinity predictor (Copin-Montégut, 1993; Cossarini et al., 2015; Schneider et al., 2007). Fig. 7 (left) shows the spatial distribution of RMSE, which is relatively uniform across regions, with higher errors mainly concentrated in coastal areas.

Fig. 6 (left) compares the alkalinity predicted by Eq. (2) with measured test samples. Most points align closely with the red bisector, indicating strong agreement between predictions and observations (Fourrier et al., 2020). Similar behavior is observed for all equations reported in Table 4.

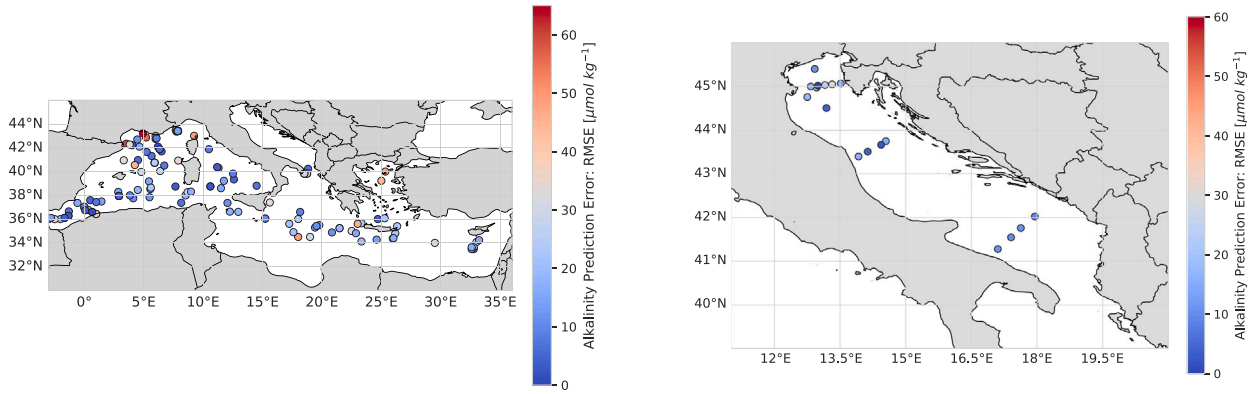


Fig. 7. RMSE distribution on the test set for the Mediterranean Sea (left panel) and the Adriatic Sea (right panel).

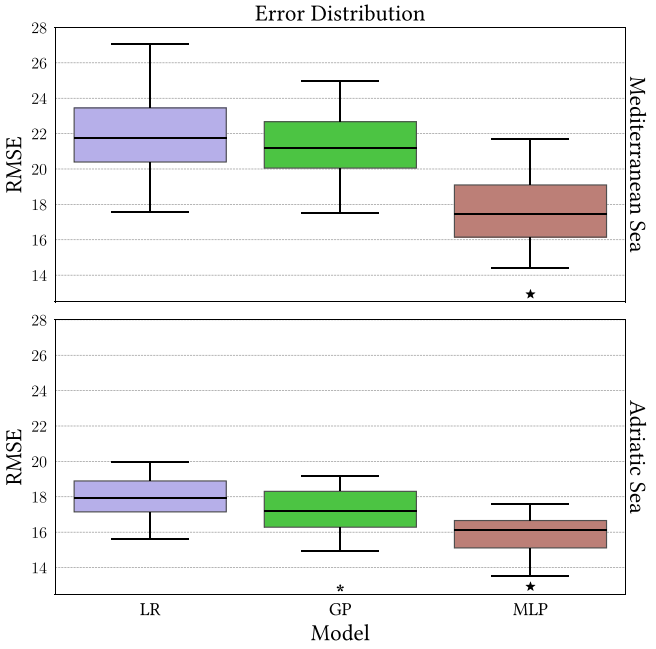


Fig. 8. Boxplots of test RMSE of the discovered models over 30 runs. We mark with an asterisk (*) models that outperform at least another model in the group, and with a star (★) models that outperform all the other ones.

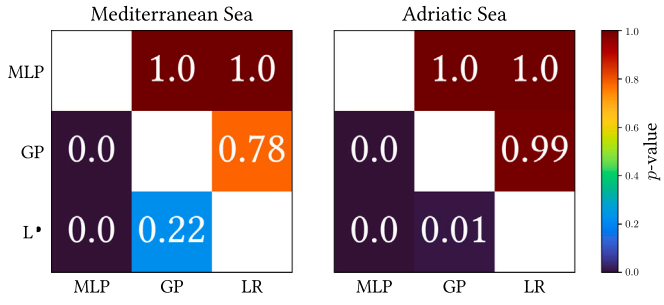


Fig. 9. Results of Mann–Whitney U test (Mann and Whitney, 1947) ($\alpha = 0.05$) over different methods for Mediterranean and Adriatic Sea. Each method on the x-axis is compared against each method on the y-axis to check whether the former produces significantly lower errors than the latter. The comparison is performed by accounting for the RMSE values on the test set.

Adriatic Sea. In the Adriatic Sea, the GP algorithm identifies more complex relationships, revealing an inverse salinity–alkalinity dependence, with a strong influence of temperature and chlorophyll. The most accurate model obtained is:

$$A = 2651.42 - 10.04 \cdot \lambda + 10.04 \cdot \phi + 10.04 \cdot T^2 + 10.04 \cdot C - 7.89 \cdot \texttt{is\_Spring} \quad (3)$$

which achieves the lowest test RMSE (15.81 μmol kg^{-1}). Here, A denotes alkalinity, λ longitude, ϕ latitude, T temperature, C chlorophyll, and `is_Spring` a binary indicator equal to 1 for spring samples and 0 otherwise. The related RMSE spatial distribution is illustrated in the right panel of Fig. 7. Additional models are reported in Table 4. Across all equations, temperature and chlorophyll consistently emerge as key predictors, while salinity (always inversely related to alkalinity) is less influential than in the Mediterranean. As in the Mediterranean Sea, latitude and longitude act as proxies for unobserved regional processes (river mouths, basin orientation, and latitudinal climatic gradients).

The complexity in alkalinity prediction in this region is largely driven by the presence of rivers, which discharge humid acids into coastal waters (Giani et al., 2023). The dilution and degradation dynamics of these compounds overlap with the dynamics of dissolved ionic substances in the Mediterranean waters, increasing reconstruction complexity. As a result, variables such as chlorophyll and temperature become essential to capture the spatial gradients from river-influenced coastal zones to offshore waters.

The scatterplot (Fig. 6, right) compares the alkalinity predicted by Eq. (3) with the measured samples in the test set. This plot shows larger discrepancies between predicted and observed alkalinity than those seen for the Mediterranean Sea, reflecting the higher complexity of this task.

SR and MLP comparison. Fig. 8 compares test-set RMSE distributions across 30 runs for our model, MLP, and LR. The corresponding p-values, obtained using the Mann–Whitney U test (Mann and Whitney, 1947) are shown in Fig. 9. To estimate uncertainty in mean performance, we compute confidence intervals (CI) using the `scipy.statistics` library (Gommers et al., 2024), which derives confidence bounds from the sample mean, standard error, and t-value. Results are reported in Table 5.

The MLP achieves the lowest RMSE, although it provides a non-interpretable, black-box solution. In the Mediterranean Sea, GP slightly outperforms LR. This similarity between these results is expected, as the salinity–alkalinity linear relationship already captures a significant portion of the alkalinity variability in this basin, making linear models naturally competitive. In contrast, in the Adriatic Sea, the GP model outperforms LR, due to its ability to capture nonlinear interactions and proxy-driven variability, which represent key aspects of alkalinity dynamics in this region.

Reconstruction on satellite sensoring data. To extend the analysis beyond the sparse in situ measurements used for training, we applied the best GP models to gap-free satellite observations from the Copernicus Marine Service (Le Traon et al., 2019). The resulting alkalinity fields

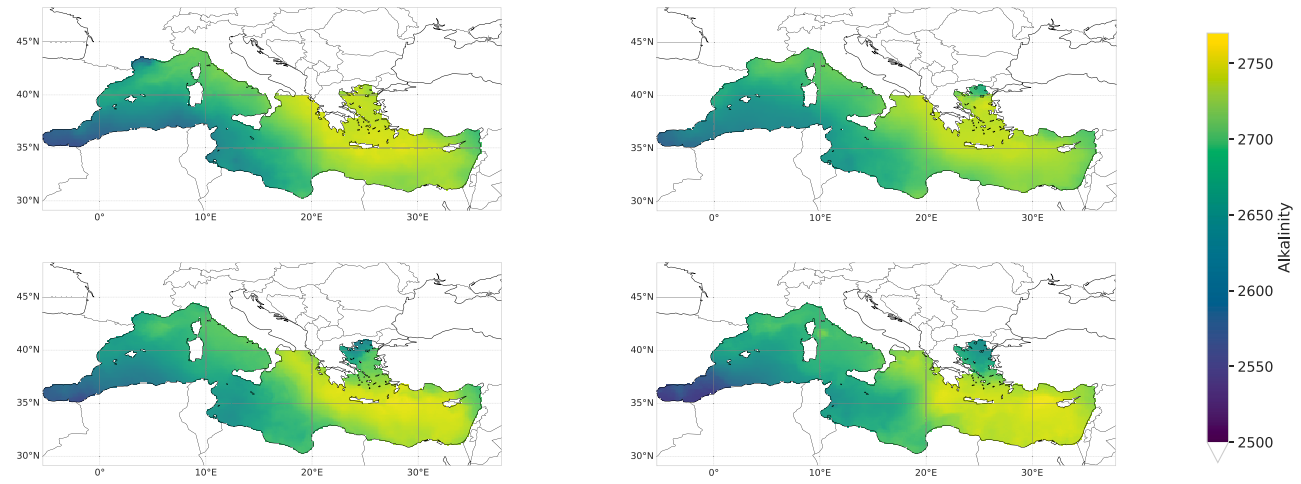


Fig. 10. Maps of alkalinity reconstruction in the Mediterranean Sea. The maps show the monthly mean alkalinity in February (upper left), May (upper right), July (lower left), and October (lower right).

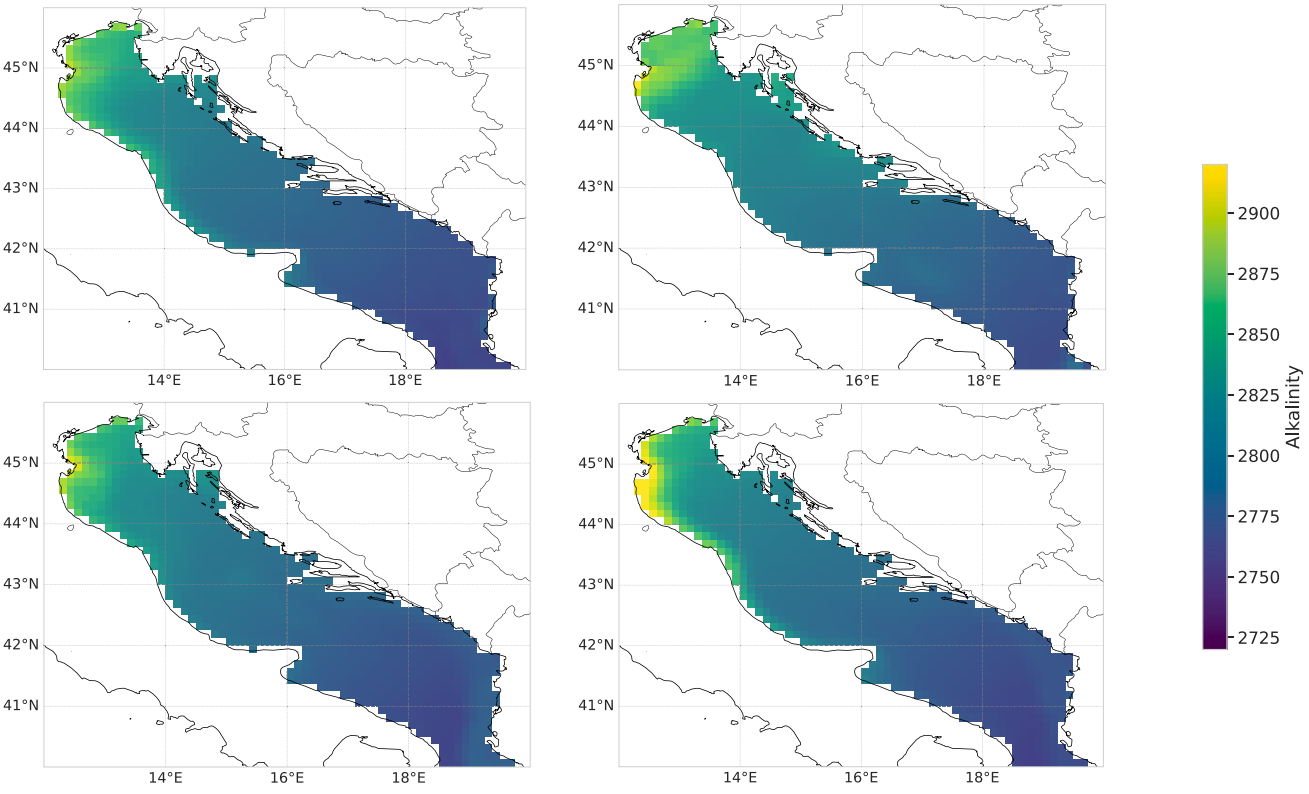


Fig. 11. Maps of alkalinity reconstruction in the Adriatic Sea. The maps show the monthly mean alkalinity in March (upper left), June (upper right), October (lower left), and December (lower right).

for the Mediterranean and Adriatic regions are shown in Figs. 10 and 11, computed using Eqs. (2) and (3), respectively. Because alkalinity exhibits strong seasonal variability, we present one representative map for each season.

Fig. 10 shows that our model is proficient in reconstructing the west-east alkalinity gradient, exhibiting an increase from Gibraltar to the Levantine Sea in all seasons. The seasonal variability of alkalinity shows higher values along the northwestern coasts during winter and spring, whereas the eastern sub-basin exhibits elevated alkalinity during summer and fall, in agreement with previous studies (Cossarini et al., 2015). As reported in the literature, the spatial gradient is smoother in winter and spring but becomes steeper during summer and fall, consistent with these earlier findings. Fig. 11 demonstrates the ability of our model to reasonably reconstruct the spatial gradients and the temporal evolution in the Adriatic region. In particular, alkalinity decreases from north to south, and higher values are predicted near the northern coasts, where river contributions are higher (Copin-Montégut, 1993; Cossarini et al., 2015; Ingrosso et al., 2016). The seasonal variability of alkalinity in this area shows peaks near the northern coastal regions, consistent with the seasonal river discharge cycle, exhibiting steeper gradients in winter and weaker ones in autumn and summer. Different from the Mediterranean case, GP model introduces variables

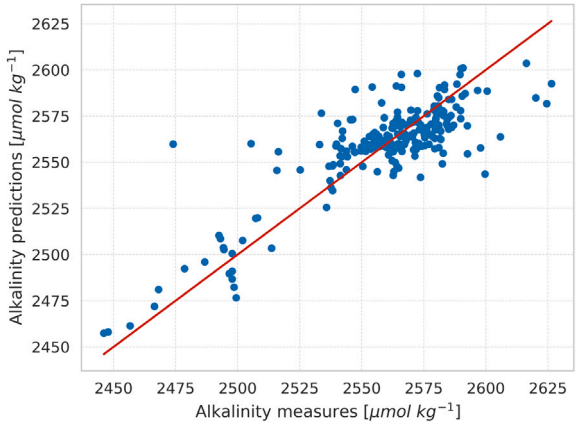


Fig. 12. Scatterplot of measured (x-axis) versus predicted (y-axis) alkalinity for the external validation dataset. The red line indicates the identity.

such as temperature and chlorophyll whose temporal variability proved important to reconstruct the signal of river input and degradation processes in this marginal sea.

External validation. Finally, we validate the model on an external dataset (SNAPO-CO_2-v2, Metzl et al. (2024)), which includes alkalinity measurements collected after 2021 and not used for training or testing. Since this dataset lacks Adriatic Sea samples, the analysis is limited to the Mediterranean Sea. Fig. 12 compares predicted alkalinity values, obtained by applying the Mediterranean model (Eq. (2)), with the corresponding observations. Although the predictions show slightly lower accuracy than in Fig. 6, performance remains comparable (RMSE = 22.23 μmol kg^{-1}; std = 15.99 μmol kg^{-1}), confirming the model's ability to generalize to new data.

7. Discussion

This work presents a framework for modeling alkalinity aimed at identifying explicit, interpretable relationships. The method explores a broad set of candidate equations, enabling the emergence of meaningful patterns without detailed prior knowledge of the underlying physical and biogeochemical processes. The key strength of this approach lies in its interpretability: the method produces transparent, human-readable equations that allow identification of the contribution of each feature to alkalinity variability. At the same time, the physical interpretation of the inferred relationships may become less straightforward in more complex environments. Our results suggest that, moving from the Mediterranean to the Adriatic Sea, alkalinity dynamics become more complex, likely due to coastal circulation, riverine inputs, and stronger biological activity, requiring GP to include additional terms. This does not reduce the transparency of the equations themselves, but it can make their direct mechanistic interpretation more challenging. It is also worth noting that the Adriatic dataset is relatively small, which may further limit the robustness of the inferred relationships.

Dataset size is a common limitation in data-driven approaches, including this study (Caiafa et al., 2021; Kokol et al., 2022). Although more in situ measurements are used than in previous alkalinity studies (Cossarini et al., 2015; Schneider et al., 2007), the dataset remains insufficient to fully capture spatial and temporal variability of the basin (Nwaila et al., 2024; Snieder and Khan, 2025). Data augmentation increases the overall data volume, but the dataset remains temporally limited and, since additional samples are obtained through spatial interpolation, coverage in underrepresented regions is not substantially improved. A strength of our approach is that the choice of variables, operators, and model complexity allows basic domain knowledge to be incorporated directly into the search, helping to mitigate data limitations and improve robustness.

In the Mediterranean Sea, we obtain a RMSE varying between [16 μmol kg^{-1}, 24 μmol kg^{-1}], in line with results from other data-driven methods based on datasets of compatible size (Cossarini et al., 2015). The Adriatic Sea shows RMSE values between [14.5 μmol kg^{-1}, 19 μmol kg^{-1}], consistently with those reported in previous data-driven analyses (Cossarini et al., 2015). Due to the limited number of investigations in this region, direct comparisons remain challenging. Moreover, the absence of an independent dataset for the Adriatic Sea limits this study, preventing direct validation of the GP model in the basin with the most complex alkalinity dynamics.

For the Mediterranean Sea, all models include a linear salinity term, confirming the well-established salinity–alkalinity relationship in this region (Cossarini et al., 2015; Gemayel et al., 2015; Schneider et al., 2007). In contrast, in the Adriatic Sea, salinity exhibits an inverse correlation with alkalinity, consistently with previous studies (Cossarini et al., 2015). This behavior reflects the influence of river inputs, which decrease salinity while increasing alkalinity. Spatial coordinates act as proxies for unresolved regional processes (Meyer et al., 2019; Milà et al., 2024) and are therefore difficult to interpret directly. Although the use of geographic coordinates may limit extrapolation, they are retained to maximize predictive performance for the Mediterranean Sea (Guisan and Zimmermann, 2000). In the Mediterranean Sea, several models include temperature as a linear term, consistent with findings from previous studies in this region (Gemayel et al., 2015), whereas in the Adriatic Sea, it appears within more complex formulations. This reflects the higher regional complexity and the inherent difficulty in isolating the specific influence of temperature (Copin-Montégut, 1993; Cossarini et al., 2015).

To the best of our knowledge, this represents the first attempt to incorporate chlorophyll as input for alkalinity prediction using data-driven approaches. Consequently, a direct comparison of the influence of chlorophyll with results from previous studies is not yet possible. Nonetheless, the inclusion of chlorophyll appears more prominent in the Adriatic Sea, where riverine inputs likely enhance its impact (Cossarini et al., 2015; Giani et al., 2023). An additional consideration concerns transitional regions between the Adriatic and Ionian Seas, which may exhibit mixed dynamics. While not explicitly addressed in the present study, this aspect represents a relevant direction for future work. In particular, future developments will explore strategies to combine predictions from basin-specific models, for instance through weighted approaches based on basin proximity.

Compared to MLP, our approach presents slightly higher RMSE but interpretable solutions. Interpretability has become a central concern in ML, particularly when the objective extends beyond prediction to the understanding of the mechanisms underlying the data (Jobin et al., 2019; Jorgensen et al., 2025; Lipton, 2018; Molnar et al., 2020; Nadizar et al., 2024; Rudin, 2019). In marine applications, transparent models can support the analysis of physical and biogeochemical processes by providing explicit relationships between variables. For alkalinity, this enables the identification of key drivers and the evaluation of their contribution to the overall alkalinity budget, facilitates comparisons across regions and experiments, and allows the inferred relationships to be directly assessed against established domain knowledge. Moreover, given the central role of salinity–alkalinity dynamics in marine ecosystems (Santore et al., 2001; de Paiva Magalhães et al., 2015; Schneider et al., 2007; Luchetta et al., 2010), transparent formulations provide a principled way to analyze how and why specific alkalinity estimates are obtained.

While prediction accuracy is crucial, understanding how models generate solutions is equally important to identify errors, biases, or data issues (Longo et al., 2020; Xu et al., 2019). Several studies investigate, neural network structure, exploring output generation and how feature interactions drive specific activations (Räuker et al., 2023; Sajjad et al., 2022). Unlike post-hoc interpretability methods in deep learning,

GP provides interpretable models by construction, which explicitly describe the relationship between input and target variables.

In our approach, we deliberately restricted the analysis to variables available from satellite observations. This restriction may reduce GP full explanatory power but enables a gap-free reconstruction of alkalinity across the entire basin. This represents a common constraints when a model must interface with broader observing systems (Fourrier et al., 2020; Sauzède et al., 2017). However, our model is easily extensible to include additional variables as they become available at large scale. Coupling data-driven approaches with mechanistic models represents a promising direction for future work, with the potential to improve alkalinity reconstruction and to account for processes not directly captured by observations.

8. Conclusion

This study presents an interpretable, data-driven framework for reconstructing surface alkalinity in the Mediterranean Sea using GP which, alternative to traditional black-box approach, produces explicit and inherently interpretable equations.

In the Mediterranean Sea, the GP model consistently recovers the well-known linear relationship between salinity and alkalinity and complements it with additional variables that improve model accuracy. In the Adriatic Sea our models reveal a different set of dependencies, with temperature, chlorophyll, and spatial indicators reflecting the influence of freshwater discharge and coastal dynamics. The resulting equations generalize well to gap-free satellite fields, enabling the reconstruction of seasonal alkalinity patterns.

CRediT authorship contribution statement

Teresa Tonelli: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Gloria Pietropolli:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Luigi Rovito:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Luca Manzoni:** Writing – review & editing, Supervision, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Gianpiero Cossarini:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Agarap, A.F., 2018. Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375.

Ali, P.J.M., Faraj, R.H., Koya, E., Ali, P.J.M., Faraj, R.H., 2014. Data Normalization and Standardization: A Technical Report. Mach Learn Tech Rep, Vol. 1, pp. 1–6.

Álvarez, A., Vélez, P., Orfila, A., Vizoso, G., Tintoré, J., 2002. Evolutionary computation for climate and ocean forecasting:"El Niño forecasting". In: Elsevier Oceanography Series. vol. 66, Elsevier, pp. 489–494.

Amadio, C., Teruzzi, A., Pietropolli, G., Manzoni, L., Coidessa, G., Cossarini, G., 2024. Combining neural networks and data assimilation to enhance the spatial impact of Argo floats in the Copernicus Mediterranean biogeochemical model. Ocean. Sci. 20 (3), 689–710.

Angelis, D., Sofos, F., Karakasidis, T.E., 2023. Artificial intelligence in physical sciences: Symbolic regression trends and perspectives. Arch. Comput. Methods Eng. 30 (6), 3845–3865.

Arif, M., ur Rehman, F., Sekanina, L., Malik, A.S., 2024. A comprehensive survey of evolutionary algorithms and metaheuristics in brain EEG-based applications. J. Neural Eng. 21 (5), 051002.

Ba, J.L., Kiros, J.R., Hinton, G.E., 2016. Layer normalization. arXiv preprint arXiv:1607.06450.

Bacardit, J., Brownlee, A.E.I., Cagnoni, S., Iacca, G., McCall, J., Walker, D., 2022. The intersection of evolutionary computation and explainable AI. In: Proceedings of the Genetic and Evolutionary Computation Conference Companion. GECCO '22, Association for Computing Machinery, New York, NY, USA, pp. 1757–1762. http://dx.doi.org/10.1145/3520304.3533974.

Bäck, T., Fogel, D.B., Michalewicz, Z., 1997. Handbook of Evolutionary Computation. p. B1, Release 97.

Back, T., Schwefel, H.-P., 1996. Evolutionary computation: An overview. In: Proceedings of IEEE International Conference on Evolutionary Computation. IEEE, pp. 20–29.

Baker, L.A., Brezonik, P.L., 1988. Dynamic model of in-lake alkalinity generation. Water Resour. Res. 24 (1), 65–74.

Banzhaf, W., Koza, J., Ryan, C., Spector, L., Jacob, C., 2000. Genetic programming. IEEE Intell. Syst. their Appl. 15 (3), 74–84. http://dx.doi.org/10.1109/5254.846288.

Bartz-Beielstein, T., Branke, J., Mehnen, J., Mersmann, O., 2014. Evolutionary algorithms. Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 4 (3), 178–195.

Beibei, E., Zhang, S., Carter, E., Meem, T.J., Wen, T., 2025. Predicting salinity and alkalinity fluxes of US freshwater in a changing climate: Integrating anthropogenic and natural influences using data-driven models. Appl. Geochem. 180, 106285.

Bergström, S., Carlsson, B., Sandberg, G., Maxe, L., 1985. Integrated modelling of runoff, alkalinity, and pH on a daily basis. Hydrol. Res. 16 (2), 89–104.

Bittig, H.C., Steinhoff, T., Claustre, H., Fiedler, B., Williams, N.L., Sauzède, R., Körtzinger, A., Gattuso, J.-P., 2018. An alternative to static climatologies: Robust estimation of open ocean CO2 variables and nutrient concentrations from T, S, and O2 data using Bayesian neural networks. Front. Mar. Sci. 5, 328.

Bock, S., Weiß, M., 2019. A proof of local convergence for the adam optimizer. In: 2019 International Joint Conference on Neural Networks. IJCNN, IEEE, pp. 1–8.

Bonin, L., Rovito, L., De Lorenzo, A., Manzoni, L., 2024. Cellular geometric semantic genetic programming. Genet. Program. Evolvable Mach. 25 (1), 1–32. http://dx.doi.org/10.1007/s10710-024-09480-8.

Brabazon, A., Kampouridis, M., O'Neill, M., 2020. Applications of genetic programming to finance and economics: Past, present, future. Genet. Program. Evolvable Mach. 21, 33–53.

Broyden, C.G., 1970. The convergence of a class of double-rank minimization algorithms 1. General considerations. IMA J. Appl. Math. 6 (1), 76–90.

Caiafa, C.F., Sun, Z., Tanaka, T., Marti-Puig, P., Solé-Casals, J., 2021. Machine learning methods with noisy, incomplete or small datasets.

Champenois, B., Bastidas, C., LaBash, B., Sapsis, T., 2025. Data-driven modeling of 4D ocean and coastal acidification in the Massachusetts and Cape Cod Bays from surface measurements. J. Geophys. Res.: Biogeosciences 130 (6), e2024JG008465.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic minority over-sampling technique. J. Artificial Intelligence Res. 16, 321–357.

Copernicus Marine Service, 2023. Multi observation global ocean sea surface salinity and sea surface density. http://dx.doi.org/10.48670/moi-00051.

Copernicus Marine Service, 2024. Mediterranean Sea – high resolution L4 sea surface temperature reprocessed. http://dx.doi.org/10.48670/moi-00173.

Copin-Montégut, C., 1993. Alkalinity and carbon budgets in the Mediterranean Sea. Glob. Biogeochem. Cycles 7 (4), 915–925.

Cossarini, G., Lazzari, P., Solidoro, C., 2015. Spatiotemporal variability of alkalinity in the Mediterranean Sea. Biogeosciences 12 (6), 1647–1658.

Cossarini, G., Mariotti, L., Feudale, L., Mignot, A., Salon, S., Taillandier, V., Teruzzi, A., d'Ortenzio, F., 2019. Towards operational 3D-Var assimilation of chlorophyll Biogeochemical-Argo float data into a biogeochemical model of the Mediterranean Sea. Ocean. Model. 133, 112–128.

Cranmer, M.D., 2023. Interpretable Machine Learning for the Physical Sciences (Ph.D. thesis). Princeton University.

Cranmer, M., 2023. Interpretable machine learning for science with PySR and SymbolicRegression.jl. URL: https://arxiv.org/abs/2305.01582, arXiv:2305.01582.

Dash, P., Devkota, M., Mercer, A.E., Ambinakudige, S., 2022. A geographic weighted regression approach for improved total alkalinity estimates in the Northern Gulf of Mexico. Environ. Model. Softw. 148, 105275.

de Paiva Magalhães, D., da Costa Marques, M.R., Baptista, D.F., Buss, D.F., 2015. Metal bioavailability and toxicity in freshwaters. Environ. Chem. Lett. 13 (1), 69–87. http://dx.doi.org/10.1007/s10311-015-0491-9.

Delashmit, W.H., Manry, M.T., et al., 2005. Recent developments in multilayer perceptron neural networks. In: Proceedings of the Seventh Annual Memphis Area Engineering and Science Conference. MAESC, Vol. 7, p. 33.

Di Biagio, V., Campanella, S., Cossarini, G., 2025. In situ dataset for initialization and validation of the Copernicus Med-MFC biogeochemical model system (MedBGCins). http://dx.doi.org/10.5281/zenodo.15489967.

Dong, C., Xu, G., Han, G., Bethel, B.J., Xie, W., Zhou, S., 2022. Recent developments in artificial intelligence in oceanography. Ocean-Land-Atmosphere Res..

Duarte, G., Lemonge, A., Goliatt, L., 2017. A dynamic migration policy to the island model. In: 2017 IEEE Congress on Evolutionary Computation. CEC, IEEE, pp. 1135–1142.

Dubey, A.K., Jain, V., 2019. Comparative study of convolution neural network's relu and leaky-relu activation functions. In: Applications of Computing, Automation and Wireless Systems in Electrical Engineering: Proceedings of MARC 2018. Springer, pp. 873–880.

Ducournau, A., Fablet, R., 2016. Deep learning for ocean remote sensing: An application of convolutional neural networks for super-resolution on satellite-derived SST data. In: 2016 9th IAPR Workshop on Pattern Recogniton in Remote Sensing. PRRS, IEEE, pp. 1–6.

Farias, F., Ludermir, T., Bastos-Filho, C., 2020. Similarity based stratified splitting: An approach to train better classifiers. arXiv preprint arXiv:2010.06099.

Ferreira, L.A., Guimarães, F.G., Silva, R., 2020. Applying genetic programming to improve interpretability in machine learning models. In: 2020 IEEE Congress on Evolutionary Computation. CEC, pp. 1–8. http://dx.doi.org/10.1109/CEC48606.2020.9185620.

Fonlupt, C., 2001. Solving the ocean color problem using a genetic programming approach. Appl. Soft Comput. 1 (1), 63–72.

Fourrier, M., Coppola, L., Claustre, H., D'Ortenzio, F., Sauzède, R., Gattuso, J.-P., 2020. A regional neural network approach to estimate water-column nutrient concentrations and carbonate system variables in the Mediterranean Sea: CANYON-MED. Front. Mar. Sci. 7, 620.

Gaur, S., Deo, M., 2008. Real-time wave forecasting using genetic programming. Ocean Eng. 35 (11–12), 1166–1172.

Gemayel, E., Hassoun, A.E.R., Benallal, M.A., Goyet, C., Rivaro, P., Abboud-Abi Saab, M., Krasakopoulou, E., Touratier, F., Ziveri, P., 2015. Climatological variations of total alkalinity and total dissolved inorganic carbon in the Mediterranean Sea surface waters. Earth Syst. Dyn. 6 (2), 789–800.

Giani, M., Ogrinc, N., Tamše, S., Cozzi, S., 2023. Elevated river inputs of the total alkalinity and dissolved inorganic carbon in the Northern Adriatic Sea. Water 15 (5), 894.

Gommers, R., Virtanen, P., Haberland, M., Burovski, E., Reddy, T., Weckesser, W., Oliphant, T.E., Cournapeau, D., Nelson, A., Roy, P., et al., 2024. Scipy/scipy: SciPy 1.15. 0. Zenodo.

Gray, P.C., Boss, E., Prochaska, J.X., Kerner, H., Demeaux, C.B., Lehahn, Y., 2024. The promise and pitfalls of machine learning in ocean remote sensing. Oceanography 37 (3), 52–63.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135 (2–3), 147–186.

He, B., Lu, Q., Yang, Q., Luo, J., Wang, Z., 2022. Taylor genetic programming for symbolic regression. In: Proceedings of the Genetic and Evolutionary Computation Conference. pp. 946–954.

Hellenic Centre for Marine Research, Hellenic National Oceanographic Data Centre (HCMR/HNODC), National Institute of Oceanography, Applied Geophysics - OGS, D.o.O., Mediterranean Sea - Eutrophication and Acidity aggregated datasets 1911/2022 v2023 http://dx.doi.org/10.13120/74158cb0-a21f-42ea-8a29-72a96b2a0da2.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 65–70, URL: https://www.jstor.org/stable/4615733.

Huertas, I.E., Ríos, A.F., García-Lafuente, J., Makaoui, A., Rodríguez-Gálvez, S., Sánchez-Román, A., Orbi, A., Ruíz, J., Pérez, F.F., 2009. Anthropogenic and natural CO 2 exchange through the Strait of Gibraltar. Biogeosciences 6 (4), 647–662.

Huynh, Q., Singh, H., Ray, T., Oyama, A., 2022. Improved genetic programming for symbolic regression: Case studies on practical applications. In: 2022 IEEE Symposium Series on Computational Intelligence. SSCI, pp. 1135–1142. http://dx.doi.org/10.1109/SSCI51031.2022.10022279.

Ingrosso, G., Giani, M., Comici, C., Kralj, M., Piacentino, S., De Vittor, C., Del Negro, P., 2016. Drivers of the carbonate system seasonal variations in a Mediterranean Gulf. Estuar. Coast. Shelf Sci. 168, 58–70.

Izzo, D., Ruciński, M., Biscani, F., 2012. The generalized island model. In: Parallel Architectures and Bioinspired Algorithms. Springer, pp. 151–169.

Jobin, A., Ienca, M., Vayena, E., 2019. The global landscape of AI ethics guidelines. Nat. Mach. Intell. 1 (9), 389–399. http://dx.doi.org/10.1038/s42256-019-0088-2.

Jorgensen, S., Nadizar, G., Pietropolli, G., Manzoni, L., Medvet, E., O'Reilly, U.-M., Hemberg, E., 2025. Policy search through genetic programming and LLM-assisted curriculum learning. ACM Trans. Evol. Learn..

Kang, S., Li, K., Wang, R., 2024. A survey on Pareto front learning for multi-objective optimization. J. Membr. Comput. 1–7.

Kapsenberg, L., Alliouane, S., Gazeau, F., Mousseau, L., Gattuso, J.-P., 2017. Coastal ocean acidification and increasing total alkalinity in the northwestern Mediterranean Sea. Ocean. Sci. 13 (3), 411–426.

Kokol, P., Kokol, M., Zagoranski, S., 2022. Machine learning on small size samples: A synthetic knowledge synthesis. Sci. Prog. 105 (1), 00368504211029777.

Koza, J.R., 1994. Genetic programming as a means for programming computers by natural selection. Stat. Comput. 4 (2), 87–112. http://dx.doi.org/10.1007/BF00175355.

Koza, J.R., 1995. Survey of genetic algorithms and genetic programming. In: Wescon Conference Record. Western Periodicals Company, pp. 589–594.

Kratzert, F., Gauch, M., Klotz, D., Nearing, G., 2024. HESS opinions: Never train a long short-term memory (LSTM) network on a single basin. Hydrol. Earth Syst. Sci. 28 (17), 4187–4201.

La Cava, W., Burlacu, B., Virgolin, M., Kommenda, M., Orzechowski, P., de França, F.O., Jin, Y., Moore, J.H., 2021. Contemporary symbolic regression methods and their relative performance. Adv. Neural Inf. Process. Syst. 2021 (DB1), 1, doi:PMC11074949.

Langdon, W.B., Poli, R., McPhee, N.F., Koza, J.R., 2008. Genetic programming: An introduction and tutorial, with a survey of techniques and applications. Comput. Intell.: A Compend. 927–1028. http://dx.doi.org/10.1007/978-3-540-78293-3_22.

Le Traon, P.Y., Reppucci, A., Alvarez Fanjul, E., Aouf, L., Behrens, A., Belmonte, M., Bentamy, A., Bertino, L., Brando, V.E., Kreiner, M.B., et al., 2019. From observation to information and users: The Copernicus Marine Service perspective. Front. Mar. Sci. 6, 234.

Lee, K., Tong, L.T., Millero, F.J., Sabine, C.L., Dickson, A.G., Goyet, C., Park, G.-H., Wanninkhof, R., Feely, R.A., Key, R.M., 2006. Global relationships of total alkalinity with salinity and temperature in surface waters of the world's oceans. Geophys. Res. Lett. 33 (19).

Legendre, P., Legendre, L., 2012. Numerical Ecology, vol. 24, Elsevier.

LemaÃŽtre, G., Nogueira, F., Aridas, C.K., 2017. Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. J. Mach. Learn. Res. 18 (17), 1–5.

Leporati, A., Manzoni, L., Mauri, G., Pietropolli, G., Zandron, C., 2023. Inferring P systems from their computing steps: An evolutionary approach. Swarm Evol. Comput. 76, 101223.

Lipton, Z.C., 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16 (3), 31–57. http://dx.doi.org/10.1145/3236386.3241340.

Longo, L., Goebel, R., Lecue, F., Kieseberg, P., Holzinger, A., 2020. Explainable artificial intelligence: Concepts, applications, research challenges and visions. In: International Cross-Domain Conference for Machine Learning and Knowledge Extraction. Springer, pp. 1–16.

Lou, R., Lv, Z., Dang, S., Su, T., Li, X., 2023. Application of machine learning in ocean data. Multimedia Syst. 29 (3), 1815–1824.

Luchetta, A., Cantoni, C., Catalano, G., 2010. New observations of CO2-induced acidification in the northern Adriatic Sea over the last quarter century. Chem. Ecol. 26 (S1), 1–17.

Lyman, J.M., Johnson, G.C., 2023. Global high-resolution random forest regression maps of ocean heat content anomalies using in situ and satellite data. J. Atmos. Ocean. Technol. 40 (5), 575–586.

Mann, H.B., Whitney, D.R., 1947. On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Stat. 18 (1), 50–60, URL: https://www.jstor.org/stable/2236101.

Marchetti, F., Pietropolli, G., Verdù, F.J.C., Castelli, M., Minisci, E., 2024. Automatic design of interpretable control laws through parametrized genetic programming with adjoint state method gradient evaluation. Appl. Soft Comput. 159, 111654.

Martin, W.N., 1997. Island (migration) models: Evolutionary algorithms based on punctuated equilibria. In: Handbook of Evolutionary Computation. IOP and Oxford University Press.

Maulud, D., Abdulazeez, A.M., 2020. A review on linear regression comprehensive in machine learning. J. Appl. Sci. Technol. Trends 1 (2), 140–147.

Metzl, N., Fin, J., Lo Monaco, C., Mignon, C., Alliouane, S., Bombled, B., Boutin, J., Bozec, Y., Comeau, S., Conan, P., et al., 2024. An updated synthesis of ocean total alkalinity and dissolved inorganic carbon measurements from 1993 to 2023: The SNAPO-CO2-v2 dataset. Earth Syst. Sci. Data Discuss. 2024, 1–39.

Meyer, H., Reudenbach, C., Wöllauer, S., Nauss, T., 2019. Importance of spatial predictor variable selection in machine learning applications–Moving from data reproduction to spatial prediction. Ecol. Model. 411, 108815.

Michałowski, T., Asuero, A.G., 2012. New approaches in modeling carbonate alkalinity and total alkalinity. Crit. Rev. Anal. Chem. 42 (3), 220–244.

Milà, C., Ludwig, M., Pebesma, E., Tonne, C., Meyer, H., 2024. Random forests with spatial proxies for environmental modelling: Opportunities and pitfalls. Geosci. Model. Dev. 17 (15), 6007–6033.

Millero, F.J., Lee, K., Roche, M., 1998. Distribution of alkalinity in the surface waters of the major oceans. Mar. Chem. 60 (1–2), 111–130.

Mittal, S., Srivastava, S., Jayanth, J.P., 2022. A survey of deep learning techniques for underwater image classification. IEEE Trans. Neural Networks Learn. Syst. 34 (10), 6968–6982.

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C.A., Casalicchio, G., Grosse-Wentrup, M., Bischl, B., 2020. Pitfalls to avoid when interpreting machine learning models. In: XXAI: Extending Explainable AI beyond Deep Models and Classifiers, ICML 2020 Workshop. p. 10, URL: http://eprints.cs.univie.ac.at/6427/.

Nadizar, G., Rovito, L., De Lorenzo, A., Medvet, E., Virgolin, M., 2024. An analysis of the ingredients for learning interpretable symbolic regression models with human-in-the-loop and genetic programming. ACM Trans. Evol. Learn. Optim. 4 (1), http://dx.doi.org/10.1145/3643688.

Nakane, T., Bold, N., Sun, H., Lu, X., Akashi, T., Zhang, C., 2020. Application of evolutionary and swarm optimization in computer vision: A literature survey. IPSJ Trans. Comput. Vis. Appl. 12 (1), 3.

Nwaila, G.T., Zhang, S.E., Bourdeau, J.E., Frimmel, H.E., Ghorbani, Y., 2024. Spatial interpolation using machine learning: From patterns and regularities to block models. Nat. Resour. Res. 33 (1), 129–161.

OC-CNR-ROMA-IT, 2023. Mediterranean Sea, bio-geo-chemical, L4, monthly means, daily gap-free and climatology. http://dx.doi.org/10.48670/moi-00300.

Ostertagová, E., 2012. Modelling using polynomial regression. Procedia Eng. 48, 500–506.

Patro, S., Sahu, K.K., 2015. Normalization: A preprocessing stage. arXiv preprint arXiv:1503.06462.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830. http://dx.doi.org/10.5555/1953048.2078195.

Pietropolli, G., Carolina, A., Cossarini, G., Manzoni, L., 2025. GLOBIO: Bridging global and local scales for biogeochemical profiles prediction. In: EGU General Assembly Conference Abstracts. pp. EGU25–3779.

Pietropolli, G., Cossarini, G., Manzoni, L., 2022. GANs for integration of deterministic model and observations in marine ecosystem. In: EPIA Conference on Artificial Intelligence. Springer, pp. 452–463.

Pietropolli, G., Manzoni, L., Cossarini, G., 2023a. Multivariate relationship in big data collection of ocean observing system. Appl. Sci. 13 (9), 5634.

Pietropolli, G., Manzoni, L., Cossarini, G., 2024. PPCon 1.0: Biogeochemical-Argo profile prediction with 1D convolutional networks. Geosci. Model. Dev. 17 (20), 7347–7364.

Pietropolli, G., Manzoni, L., Paoletti, A., Castelli, M., 2023b. On the hybridization of geometric semantic GP with gradient-based optimizers. Genet. Program. Evolvable Mach. 24 (2), 16.

Pinkus, A., 1999. Approximation theory of the MLP model in neural networks. Acta Numer. 8, 143–195.

Poli, R., Langdon, W.B., McPhee, N.F., Koza, J.R., 2007. Genetic programming: An introductory tutorial and a survey of techniques and applications. Univ. Essex School of Computer Science and Eletronic Engineering Technical Report No. CES-475, pp. 1–112.

Radwan, Y.A., Kronberger, G., Winkler, S., 2024. A comparison of recent algorithms for symbolic regression to genetic programming. In: International Conference on Computer Aided Systems Theory. Springer, pp. 157–171.

Räuker, T., Ho, A., Casper, S., Hadfield-Menell, D., 2023. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. In: 2023 IEEE Conference on Secure and Trustworthy Machine Learning. Satml, IEEE, pp. 464–483.

Rodrigues, L.C., van den Bergh, J.C., Ghermandi, A., 2013. Socio-economic impacts of ocean acidification in the Mediterranean Sea. Mar. Policy 38, 447–456.

Rovito, L., Bonin, L., Farinati, D., Vanneschi, L., Manzoni, L., De Lorenzo, A., Pietropolli, G., 2025. Exploring the integration of cellular structures in genetic programming-based methods. In: Xue, B., Manzoni, L., Bakurov, I. (Eds.), Genetic Programming. Springer Nature Switzerland, Cham, pp. 120–138.

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1 (5), 206–215. http://dx.doi.org/10.1038/s42256-019-0048-x.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323 (6088), 533–536. http://dx.doi.org/10.1038/323533a0.

Sajjad, H., Durrani, N., Dalvi, F., 2022. Neuron-level interpretation of deep nlp models: A survey. Trans. Assoc. Comput. Linguist. 10, 1285–1303.

Salon, S., Cossarini, G., Bolzon, G., Feudale, L., Lazzari, P., Teruzzi, A., Solidoro, C., Crise, A., 2019. Novel metrics based on Biogeochemical Argo data to improve the model uncertainty evaluation of the CMEMS Mediterranean marine ecosystem forecasts. Ocean. Sci. 15 (4), 997–1022.

Sammartino, M., Buongiorno Nardelli, B., Marullo, S., Santoleri, R., 2020. An artificial neural network to infer the Mediterranean 3D chlorophyll-a and temperature fields from remote sensing observations. Remote. Sens. 12 (24), 4123.

Santore, R.C., Di Toro, D.M., Paquin, P.R., Allen, H.E., Meyer, J.S., 2001. Biotic ligand model of the acute toxicity of metals. 2. Application to acute copper toxicity in freshwater fish and Daphnia. Environ. Toxicol. Chem. 20 (10), 2397–2402. http://dx.doi.org/10.1002/etc.5620201035.

Sauzède, R., Bittig, H.C., Claustre, H., Pasqueron de Fommervault, O., Gattuso, J.-P., Legendre, L., Johnson, K.S., 2017. Estimates of water-column nutrient concentrations and carbonate system parameters in the global ocean: A novel approach based on neural networks. Front. Mar. Sci. 4, 128.

Schneider, A., Wallace, D.W., Körtzinger, A., 2007. Alkalinity of the Mediterranean sea. Geophys. Res. Lett. 34 (15).

Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. J. Big Data 6 (1), 1–48.

Slowik, A., Kwasnicka, H., 2020. Evolutionary algorithms and their applications to engineering problems. Neural Comput. Appl. 32 (16), 12363–12379.

Smits, G.F., Kotanchek, M., 2005. Pareto-front exploitation in symbolic regression. In: Genetic Programming Theory and Practice II. Springer, pp. 283–299.

Snieder, E., Khan, U.T., 2025. A diversity-centric strategy for the selection of spatio-temporal training data for LSTM-based streamflow forecasting. Hydrol. Earth Syst. Sci. 29 (3), 785–798.

Sonnewald, M., Lguensat, R., Jones, D.C., Dueben, P.D., Brajard, J., Balaji, V., 2021. Bridging observations, theory and numerical simulation of the ocean using machine learning. Environ. Res. Lett. 16 (7), 073008.

Su, X., Yan, X., Tsai, C.-L., 2012. Linear regression. Wiley Interdiscip. Rev.: Comput. Stat. 4 (3), 275–294.

Tonda, A., 2024. Review of PySR: High-performance symbolic regression in Python and Julia. Genet. Program. Evolvable Mach. 26 (1), 7. http://dx.doi.org/10.1007/s10710-024-09503-4.

Tonelli, T., Cossarini, G., Manzoni, L., Pietropolli, G., 2026. Two-phase CNN for model data fusion: Predicting 3D chlorophyll-a in the Mediterranean Sea. Ocean. Model. 102707.

Tonelli, T., Pietropolli, G., Sbaiz, G., Manzoni, L., 2024. Genetic programming for the reconstruction of delay differential equations in economics. In: Proceedings of the Genetic and Evolutionary Computation Conference Companion. pp. 2119–2122.

Touratier, F., Guglielmi, V., Goyet, C., Prieur, L., Pujo-Pay, M., Conan, P., Falco, C., 2012. Distributions of the carbonate system properties, anthropogenic CO 2, and acidification during the 2008 BOUM cruise (Mediterranean Sea). Biogeosciences Discuss. 9 (3), 2709–2753.

Vrigazova, B., 2021. The proportion for splitting data into training and test set for the bootstrap in classification problems. Bus. Syst. Res.: Int. J. Soc. Adv. Innov. Res. Econ. 12 (1), 228–242.

Wallace, D.W., 1995. Monitoring global ocean carbon inventories. Ocean. Obs. Syst. Dev..

Whitley, D., Rana, S., Heckendorn, R.B., 1997. Island model genetic algorithms and linearly separable problems. In: AISB International Workshop on Evolutionary Computing. Springer, pp. 109–125.

Whitley, D., Rana, S., Heckendorn, R.B., 1999. The island model genetic algorithm: On separability, population size and convergence. J. Comput. Inf. Technol. 7 (1), 33–47.

Wolf-Gladrow, D.A., Zeebe, R.E., Klaas, C., Körtzinger, A., Dickson, A.G., 2007. Total alkalinity: The explicit conservative expression and its application to biogeochemical processes. Mar. Chem. 106 (1–2), 287–300.

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., Zhu, J., 2019. Explainable AI: A brief survey on history, research areas, approaches and challenges. In: CCF International Conference on Natural Language Processing and Chinese Computing. Springer, pp. 563–574.

Zanna, L., Bolton, T., 2021. Deep learning of unresolved turbulent ocean processes in climate models. In: Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences. Wiley Online Library, pp. 298–306.

Zeebe, R.E., Wolf-Gladrow, D., 2001. CO2 In Seawater: Equilibrium, Kinetics, Isotopes, vol. 65, Gulf Professional Publishing.

Zhan, Z.-H., Shi, L., Tan, K.C., Zhang, J., 2022. A survey on evolutionary computation for complex continuous optimization. Artif. Intell. Rev. 55 (1), 59–110.

Zhang, G.P., 2002. Neural networks for classification: A survey. IEEE Trans. Syst. Man, Cybern. Part C (Applications Reviews) 30 (4), 451–462.

Zhang, Q., Barri, K., Jiao, P., Salehi, H., Alavi, A.H., 2021. Genetic programming in civil engineering: Advent, applications and future trends. Artif. Intell. Rev. 54 (3), 1863–1885.

Zhang, Z., Chen, P., Zhang, S., Huang, H., Pan, Y., Pan, D., 2025. A review of machine learning applications in ocean color remote sensing. Remote. Sens. 17 (10), 1776.

Zhao, Q., Peng, S., Wang, J., Li, S., Hou, Z., Zhong, G., 2024. Applications of deep learning in physical oceanography: A comprehensive review. Front. Mar. Sci. 11, 1396322.