

**SANTA CATARINA STATE UNIVERSITY - UDESC
COLLEGE OF TECHNOLOGICAL SCIENCES - CCT
GRADUATE PROGRAM IN APPLIED COMPUTING - PPGCAP**

THIAGO BRANDENBURG

**A DATA-DRIVEN APPROACH FOR ATMOSPHERIC CORROSION
PREDICTION**

JOINVILLE

2026

THIAGO BRANDENBURG

**A DATA-DRIVEN APPROACH FOR ATMOSPHERIC CORROSION
PREDICTION**

Thesis submitted to the Graduate Program in Applied Computing at the Center for Technological Sciences of the State University of Santa Catarina, to obtain a Master's degree in Applied Computing.

Advisor: Dr. Rafael Stubs Parpinelli

JOINVILLE

2026

**Ficha catalográfica elaborada pelo programa de geração automática da
Biblioteca Universitária Udesc,
com os dados fornecidos pelo(a) autor(a)**

Brandenburg, Thiago
A Data-driven Approach for Atmospheric Corrosion Prediction
/ Thiago Brandenburg. -- 2026.
115 p.

Orientador: Rafael Stubs Parpinelli
Dissertação (mestrado) -- Universidade do Estado de Santa
Catarina, Centro de Ciências Tecnológicas, Programa de
Pós-Graduação em Computação Aplicada, Joinville, 2026.

1. Machine Learning. 2. Atmospheric Corrosion. 3.
Data-Driven. I. Parpinelli, Rafael Stubs. II. Universidade do Estado
de Santa Catarina, Centro de Ciências Tecnológicas, Programa de
Pós-Graduação em Computação Aplicada. III. Título.

Thiago Brandenburg

A Data-driven Approach for Atmospheric Corrosion Prediction

This dissertation was considered suitable for obtaining the degree of **Master of Science in Applied Computing**, in the concentration area of "Research Methodology and Computer Techniques", and was approved in its final form by the Master's Program in Applied Computing of the Center for Technological Sciences at the Santa Catarina State University.

Examination Committee

Rafael Stubs Parpinelli

Santa Catarina State University (UDESC)
Advisor

Professor Adriano Fiorese

Santa Catarina State University (UDESC)
Examiner 1

Professor Leandro dos Santos Coelho

Federal University of Paraná (UFPR)
Examiner 2

Joinville, 05 de Maio de 2026

I dedicate this work to my family, friends, colleagues, and teachers who have accompanied me and given me strength throughout this magnificent journey.

ACKNOWLEDGEMENTS

First, I would like to thank my advisor, Rafael Stubs Parpinelli, for the opportunity to work on the corrosion project for three years. My time at LABICOM shaped me both as a professional and as a scientist. It was a unique opportunity to experience several “Eureka” moments and see them materialize within an applied research project.

I would also like to thank the other members of the project — Lavínia, Vinícius, José Francisco, Fabiano, Gustavo, Sebastian, and Krigor — for sharing their expertise. I am grateful to my colleagues at LABICOM — Maísa, Draylon, Gonça, Guilherme Razzini, Guilherme Ulissea, Gabriel, and Nicholas — who patiently tolerated me nerding out all day about salt and corrosion.

I would also like to thank my friends. To my homies Dunzer, Manske, Kubaski, Paulo, and Sales, thank you for sticking with me all these years, despite my terrible Dota skills, and also for tolerating me nerding out all day about salt and corrosion.

To my comrades Dan, Pandini, Gilson, Miguel, Rodrigo, Kauã, and Ian, thanks for helping me plug and unplug computers all these years, and definitely for tolerating me nerding out all day about salt and corrosion (yeah I kinda did it a lot). I wanna play some gartic with you folks one day.

I also thank Laís and Danilo for allowing me to experience the fastest speed I’ve achieved on a land vehicle (never imagined it would be a Peugeot). I also thank them and Jota, Schuma, André, Kalyi, Amyr, Chell, Marco, Alexandre, Murilo, and Cyro for allowing me to join the best Telegram server ever created, and ALSO tolerating me nerding out all day about salt and corrosion.

As a programmer, I must point out that there is some intersection between these groups; therefore, I listed only the first occurrence of each name to avoid repetition and redundancy (neeeeerd).

I would also like to thank my family: my father, Silvio, the person I admire most in this world; my mother, Giselli, who is not only my Mother but also my best friend; my brother, Alexandre, whom I love dearly and to whom I demand a rematch in Street Fighter II; and my grandmother, Diva, who has always welcomed me with a smile.

Finally, I thank my girlfriend, Lissa, that I love so much and is also my greatest rival in puzzles (and I am not going to lose next time!).

To everyone, thank you for your patience with me. Yes, folks — it’s corroding!

"The Blades are wise not to trust me. *Onikaan ni ov*. I would not trust another *dovah*. *Dov wahlaan fah rel*. We were made to dominate. The will to power is in our blood. You feel it in yourself, do you not? I can be trusted. I know this. But they do not. *Onikaan ni ov dovah*. It is always wise to mistrust a *dovah*. I have overcome my nature only through meditation and long study of the Way of the Voice. No day goes by where I am not tempted to return to my inborn nature. *Zin krif horvut se suleyk*. What is better - to be born good, or to overcome your evil nature through great effort?"

Paarthurnax

ABSTRACT

Atmospheric corrosion is the degradation of metals caused by interaction with the surrounding environment. Corrosion maintenance and protection costs are estimated at 3% to 5% of the GDP of industrialized countries; therefore, estimating atmospheric corrosion rates is essential for assessing the costs and service lifetime of metallic structures. To estimate the atmospheric corrosion rate of a metal, ISO 9223 proposes Dose-Response Functions (DRFs) as a standardized model.

DRFs have limited applicability because their input variables require long-duration field experiments. Consequently, machine learning-based alternatives have been proposed. In the literature, two approaches are distinguishable: corrosion monitoring models, which are site-specific, and corrosion exposure models, which are generalist multi-climate models. For exposure models, a novel approach is to use remote sensing atmospheric and geospatial data as input variables, exposure data as training entries, and machine learning (ML) algorithms as predictive models.

This work proposes an exposure prediction model developed and deployed through a data-driven atmospheric corrosion prediction system structured in three layers: a data layer, an ML modeling layer, and a user application layer for running case studies. Models for six target variables: two intermediary pollutant deposition variables and four corrosion rate variables were developed. Each model was selected through a comparison of traditional ML algorithms using Random Subsampling validation and the Dunn test.

The selected models for pollutants are Random Forest for predicting sulfur dioxide (SO_2) deposition ($r^2 = 0.69$), a symbolic regression equation for predicting chloride (Cl^-) deposition ($r^2 = 0.71$). For corrosion rate, a Support Vector Machine regressor for predicting zinc (Zn) corrosion rate ($r^2 = 0.69$), and Random Forests for predicting Carbon Steel (Fe), Aluminum (Al), and Copper (Cu), with r^2 values of 0.63, 0.60, and 0.68, respectively.

Key-words: machine learning, atmospheric corrosion, data-driven.

RESUMO

A corrosão atmosférica é a degradação de metais causada pela interação com o ambiente ao redor. Os custos de manutenção e proteção contra corrosão são estimados entre 3% e 5% do PIB dos países industrializados; portanto, estimar as taxas de corrosão atmosférica é essencial para avaliar os custos e a vida útil de estruturas metálicas. Para estimar a taxa de corrosão atmosférica de um metal, a norma ISO 9223 propõe as Funções Dose-Resposta (FDRs) como modelo padronizado.

As FDRs possuem aplicabilidade limitada, pois suas variáveis de entrada requerem experimentos de campo de longa duração. Consequentemente, alternativas baseadas em aprendizado de máquina foram propostas. Na literatura, duas abordagens são distinguíveis: modelos de monitoramento de corrosão, que são específicos para cada local, e modelos de exposição à corrosão, que são modelos generalistas para múltiplos climas. Para modelos de exposição, uma abordagem inovadora consiste em utilizar dados atmosféricos e geoespaciais de sensoriamento remoto como variáveis de entrada, dados de exposição como entradas de treinamento, e algoritmos de aprendizado de máquina (AM) como modelos preditivos.

Este trabalho propõe um modelo de predição de exposição desenvolvido e implantado por meio de um sistema de predição de corrosão atmosférica orientado a dados, estruturado em três camadas: uma camada de dados, uma camada de modelagem AM e uma camada de aplicação ao usuário para execução de estudos de caso. Foram desenvolvidos modelos para seis variáveis-alvo: duas variáveis intermediárias de deposição de poluentes e quatro variáveis de taxa de corrosão. Cada modelo foi selecionado por meio da comparação de algoritmos tradicionais de AM utilizando validação por Subamostragem Aleatória e o teste de Dunn.

Os modelos selecionados para poluentes são: Random Forest para predição da deposição de dióxido de enxofre (SO_2) ($r^2 = 0,69$), e uma equação de regressão simbólica para predição da deposição de cloreto (Cl^-) ($r^2 = 0,71$). Para taxa de corrosão, uma Máquina de Vetores de Suporte (SVR) para predição da taxa de corrosão do zinco (Zn) ($r^2 = 0,69$), e Random Forests para predição do Aço Carbono (Fe), Alumínio (Al) e Cobre (Cu), com valores de r^2 de 0,63, 0,60 e 0,68, respectivamente.

Palavras-chaves: Aprendizado de máquina, corrosão atmosférica, orientado a dados.

LIST OF FIGURES

Figure 1 – Effects of corrosion	16
Figure 2 – Concept of an intelligent atmospheric corrosion prediction system .	17
Figure 3 – Diagram of the three major exposure programs	23
Figure 4 – An example of an atmospheric corrosion station	24
Figure 5 – Pollutant Candles from ISO 9223	27
Figure 6 – Different exposure programs by variables collected	28
Figure 7 – Workflow used to create data-driven atmospheric corrosion maps . .	30
Figure 8 – Map of ISOCORRAG stations	31
Figure 9 – Map of MICAT project stations	32
Figure 10 – Corrosivity map of Chile	33
Figure 11 – Locations of monitoring stations in the Canary Islands	34
Figure 12 – Atmospheric Corrosion Maps of Brazil	35
Figure 13 – ERA-5 map of 2-meter temperature	37
Figure 14 – MERRA-2 map of surface sea salt mass concentration	38
Figure 15 – Distance to coastline dataset	39
Figure 16 – STRM30+ digital elevation model	40
Figure 17 – ESA CCI Land Cover for 2015	41
Figure 18 – Images for an Industrial site in Cubatão, Brazil	43
Figure 19 – SVR fits using linear, polynomial, and RBF kernels on a sine curve .	46
Figure 20 – Caption	48
Figure 21 – Genetic operations in expression trees	49
Figure 22 – Example of a random subsampling with n executions	51
Figure 23 – Corrosion station with ISO specimens and ACM sensors.	56
Figure 24 – Data distribution by classification	56
Figure 25 – Number of publications per year	59
Figure 26 – Datasets used by the publications	60
Figure 27 – Country of origin of the data used	60
Figure 28 – Variables used by the models	61
Figure 29 – Variable correlation analysis	62
Figure 30 – ML models used in the publications	63
Figure 31 – Error metrics used in the publications	64
Figure 32 – LABICORR system workflow	66
Figure 33 – Extended representation of the data layer	68
Figure 34 – Entity-Relationship diagram of the Corrosion DB	69
Figure 35 – Extended diagram of the modeling layer	75
Figure 36 – Extended diagram of the application layer.	79

Figure 37 – Sequence diagram of a case study execution.	80
Figure 38 – Component diagram of the system versions.	81
Figure 39 – Map of the Corrosion Database stations	82
Figure 40 – Distribution of SO_2 entries across pollution levels. Source: Author. . .	83
Figure 41 – SO_2 model variables Spearman correlation.	84
Figure 42 – Results for the SO_2 model	85
Figure 43 – Distribution of Cl^- data by ISO 9223 pollution levels. Source: Author. .	86
Figure 44 – Cl^- model variables Spearman correlation.	87
Figure 45 – Results of the Cl^- model validation	88
Figure 46 – Distribution of Zn data by ISO 9223 corrosivity levels. Source: Author. .	89
Figure 47 – Zn model variables Spearman correlation.	90
Figure 48 – Results of the Zn model validation.	91
Figure 49 – Distribution of Fe data by ISO 9223 corrosivity levels. Source: Author. .	92
Figure 50 – Fe model variables Spearman correlation.	93
Figure 51 – Results of the Fe model validation.	94
Figure 52 – Comparison between ISO 9223 Fe Dose Response Function and adjusted version in the validation.	94
Figure 53 – Distribution of Al data by ISO 9223 corrosivity levels. Source: Author. .	95
Figure 54 – Al model variables Spearman correlation.	96
Figure 55 – Results of the Al model validation.	96
Figure 56 – Distribution of Al data by ISO 9223 corrosivity levels. Source: Author. .	97
Figure 57 – Cu model variables Spearman correlation.	98
Figure 58 – Results of the Cu model validation.	98
Figure 59 – Case Study Page. Source: Author.	99

LIST OF TABLES

Table 1 – ISO 9223 corrosivity classification intervals.	25
Table 2 – Reference table of roughness length (z_0)	42
Table 3 – Results of applying inclusion and exclusion criteria	54
Table 4 – Summary of literature review results	64
Table 5 – Coordination refinement of Argentinian stations of the MICAT project	71
Table 6 – Data Sources	72
Table 7 – Variables of the system	72
Table 8 – Hyperparameter search space for each machine learning algorithm .	77
Table 9 – Hyperparameters of the PySR algorithm.	78
Table 10 – Distribution of SO_2 data by ISO 9223 pollution levels.	83
Table 11 – Distribution of Cl^- data by ISO 9223 pollution levels	86
Table 12 – Distribution of Zn data by ISO 9223 corrosivity levels.	89
Table 13 – Distribution of Fe data by ISO 9223 corrosivity levels.	92
Table 14 – Distribution of Al data adapted from ISO 9223 corrosivity levels. . . .	95
Table 15 – Distribution of Cu from ISO 9223 corrosivity levels.	97

LIST OF ABBREVIATIONS AND ACRONYMS

ANN	Artificial Neural Network
ABNT	Brazilian Association of Technical Standards
ABRACO	Brazilian Association of Corrosion
GP	Genetic Programming
ISO	International Organization for Standardization
ISO/TC	Technical Committee of ISO
ISOCORRAG	International Program of Atmospheric Exposure
ECMWF	European Centre for Medium-Range Weather Forecasts
ET	ExtraTrees
MERRA-2	Modern-Era Retrospective Analysis for Research and Applications, Version 2
MICAT	Ibero-American Maps of Atmospheric Corrosivity
MLP	Multi-Layer Perceptron
NACE	National Association of Corrosion Engineers
NASA	National Aeronautics and Space Administration
NBR	Brazilian Technical Standards
OSM	OpenStreetMap
RF	Random Forest
RMSE	Rooted Mean Squared Error
SQL	Structured Query Language
SVM	Support Vector Machine
SVR	Support Vector Machine Regressor
UDESC	Santa Catarina State University
WCO	World Customs Organization

LIST OF SYMBOLS

%	Percent
μm	Micrometers
<i>Al</i>	Aluminum
C°	Degrees Celsius
<i>Cl</i> —	Chloride ions
<i>Cu</i>	Copper
<i>Fe</i>	Iron, used to represent Carbon Steel
mg/m^2d	Milligrams per square meter per day
O_3	Ozone
PbO_2	Lead Dioxide
r_{corr}	Corrosion rate for 1 year
SO_2	Sulfur dioxide
<i>Zn</i>	Zinc

CONTENTS

1	INTRODUCTION	16
1.1	Motivation	18
1.2	Objective	18
1.3	Previous Works	19
1.4	Structure of the Work	19
2	BACKGROUND	21
2.1	Atmospheric Corrosion	21
2.1.1	ISO/TC 156 International Standards	24
2.1.2	Atmospheric Corrosion Estimation	28
2.1.3	Atmospheric Exposure Data	31
2.1.3.1	International Exposure Program (ISOCORRAG)	31
2.1.3.2	Ibero-American Map of Atmospheric Corrosiveness (MICAT)	31
2.1.3.3	Atmospheric Corrosivity Map of Chile	32
2.1.3.4	Atmospheric Corrosivity Maps of the Canary Islands	33
2.1.3.5	CSIRO Asia Exposures	34
2.1.3.6	Other Atmospheric Exposures in Brazil	34
2.2	Climate and Geospatial Data Products	35
2.2.1	ERA-5 and MERRA-2 Atmospheric Reanalysis	36
2.2.2	Emissions Database for Global Atmospheric Research (EDGAR)	38
2.2.3	Distance to the Sea	38
2.2.4	Elevation	39
2.2.5	Land Cover and Surface Roughness	40
2.2.6	OpenStreetMap	42
2.3	Machine Learning	43
2.3.1	Support Vector Regressors (SVR)	44
2.3.2	Random Forest Regressor	46
2.3.3	Genetic Programming and the PySR Algorithm	48
2.3.4	Model Evaluation	50
3	RELATED WORKS	52
3.0.1	Methodology	52
3.0.2	Protocol Application	54
4	PROPOSED APPROACH	66

4.1	Data Layer	68
4.2	Modeling Layer	75
4.3	Application Layer	79
5	RESULTS AND ANALISYS	82
5.1	Sulphur Dioxide deposition model (SO ₂)	83
5.2	Chloride deposition model (Cl ⁻)	86
5.3	Zinc corrosion model (Zn)	89
5.4	Carbon Steel corrosion model (Fe)	92
5.5	Aluminum corrosion model (Al)	95
5.6	Copper corrosion model (Cu)	97
5.7	Application Layer results	99
6	CONCLUSION	100
	REFERENCES	104
	APPENDIX A – ARTIFICIAL INTELLIGENCE USE DECLARATION	114

1 INTRODUCTION

Corrosion is the degradation of a material's properties or the loss of mass over time due to interaction with the environment. This deterioration reflects the natural tendency of material compounds to return to their most thermodynamically stable state. For most metallic materials, corrosion involves the formation of oxides and sulfides on the material's surface (SCHWEITZER, 2006). As an example, Figure 1 shows the effects of corrosion on a steel pole.



Figure 1 – Effects of corrosion. Source: Author

Due to the high costs associated with corrosion (KOCH et al., 2002), a set of international technical standards (ISO) was developed to assist in defining corrosion for outdoor environments (atmospheric corrosion) for protection coatings. The main standard, ISO 9223, defines the one-year corrosion rate (r_{corr}) as the base variable for the standard's procedures, using it to classify atmospheres according to levels of corrosivity (ISO/9223, 2012) and to estimate long-term corrosion (ISO/9224, 2012).

ISO 9223 allows two ways of calculating r_{corr} : either by exposing the target metal for one year using an exposure rack, or by using dose–response functions (MIKHAILOV; TIDBLAD; KUCERA, 2004), available for the four main metallic alloys used in industry (carbon steel, zinc, copper, and aluminum). These functions have pre-defined input variables: temperature, relative humidity, dry deposition of sulfur dioxide (SO_2), and dry deposition of chloride ions (Cl^-).

However, both workflows have limitations. Exposing the metal is an expensive experimental method that requires long-term testing, ranging from months to years. While dose–response functions can theoretically be applied, in practice they face the same limitation: two of the four variables—chloride and sulfur dioxide pollutants—are also obtained through long-term field experiments.

To overcome these experimental limitations, the solution is to use climate variables with historical data available over large coverage areas. In this regard, meteorological and space agencies provide catalogs of climate and geospatial data. Examples include the Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2) (GELARO et al., 2017), an atmospheric reanalysis of the modern satellite era that offers global data for decades of aerosol observations, and ERA5 (HERSBACH et al., 2020), which provides global coverage of meteorological variables such as temperature and precipitation at high resolution from 1950 to the present.

To connect geospatial and atmospheric variables to corrosion prediction, machine learning models can be used. Given a precise location and exposure period, data from these programs can be used to collect the geospatial and atmospheric variables corresponding to the metal exposure measurements, and then train a statistical model. The result is an intelligent system that can be applied remotely, without the need for experimental methods. Figure 2 shows a diagram illustrating the conceptual idea of an intelligent atmospheric corrosion prediction system.

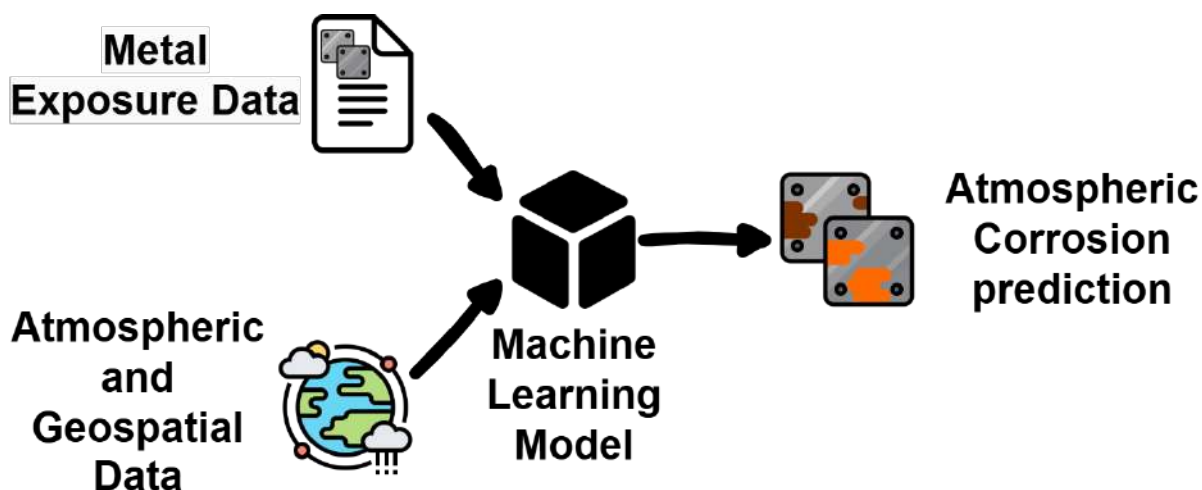


Figure 2 – Conceptual idea of an intelligent atmospheric corrosion prediction system. Source: Author

The concept of an intelligent atmospheric corrosion system is novel in the literature at the time of this work, given that most atmospheric datasets with usable resolutions are recent, with the best high-resolution pollutant data available from 2016 (GELARO et al., 2017) and meteorological data from 2020 (HERSBACH et al., 2020).

To date, most intelligent corrosion solutions are based either on experimental

data or on the use of corrosivity maps developed from the interpolation of previously conducted experiments (SANTA et al., 2022). These maps do not account for microclimates nor are they time-sensitive, which justifies the study of solutions that use recent climate data capable of capturing microclimatic variables.

This work proposes a case study approach, in which data for a location are provided by the user, and the system collects the necessary variables, runs the models, and returns a prepared answer through an interface. To achieve this, two intermediate models for Cl^- and SO_2 deposition prediction and a final target model for zinc (Zn) corrosion rate prediction were developed and integrated into a case study application.

1.1 MOTIVATION

The study of corrosive processes, especially in metallic materials, is of great interest to society because corrosion problems are frequent and occur across a wide range of activities: industry, civil construction, medicine, preservation of artworks, and others (GENTIL, 2011). The costs associated with corrosion are high: in the United States, they have been estimated at 3.1% of the national Gross Domestic Product (GDP) (KOCH et al., 2002), and between 3% and 4% of the GDP of industrialized countries (SCHMITT et al., 2009).

These costs can be direct (replacement of parts and equipment, maintenance of corrosion protection systems) or indirect (losses due to malfunctioning corroded equipment, contamination, and oversizing of projects) (GENTIL, 2011). More than half of the considerable damage due to corrosion is a result of atmospheric impacts on materials, which is logical considering that most metallic equipment and structures operate in the atmospheric environment. For this reason, the action of the atmosphere on metals is one of the major issues in corrosion science (CHICO et al., 2017).

1.2 OBJECTIVE

The main objective of this work is to develop and apply atmospheric corrosion prediction models for each of the ISO 9223 standard metals, along with coupled pollutant models to produce input variables, using machine learning and atmospheric and geospatial data. To achieve this, three specific objectives were established:

- Conduct a systematic literature review on the use of machine learning for atmospheric corrosion prediction, in order to identify which data, models, metrics, and validation methods have been used to develop such models.
- Develop and improve pollutant models for sulfur dioxide (SO_2) and chloride ion (Cl^-) deposition and atmospheric corrosion models for zinc or galvanized steel

(*Zn*), carbon steel (*Fe*), aluminum (*Al*), and copper (*Cu*).

- Develop an intelligent system that integrates atmospheric and geospatial data, target metal models, and intermediate pollutant models into a single application, allowing users to perform case studies of atmospheric corrosion remotely from the studied location, mainly for Brazilian environments.

1.3 PREVIOUS WORKS

This work compiles and extends a series of publications associated with the project “Machine Learning for Predicting the Corrosion Rate in Steels” (PARPINELLI et al., 2023). The project is a partnership between the Laboratory of Computational Intelligence (LABICOM) at the State University of Santa Catarina (UDESC) and ArcelorMittal Brazil.

Initially, an atmospheric corrosion model based on Random Forests using the same variables as the Dose–Response Functions was proposed (GEREMIAS et al., 2024; GEREMIAS et al., 2025a). As the variable-oriented nature of the atmospheric corrosion problem became evident, climate and geospatial data were then used to develop pollutant models for the deposition of sulfur dioxide (GEREMIAS et al., 2025b) and chloride ions (BRANDENBURG, 2024; BRANDENBURG et al., 2024). Later, these atmospheric and geospatial data were applied to corrosion models, coupled with the pollutant models, and used to produce atmospheric corrosivity maps (GEREMIAS, 2024) of Latin America.

Atmospheric corrosivity maps are useful tools, but as the variables became more microclimate-oriented, it became computationally expensive to calculate the variable grids required to generate high-resolution corrosivity maps. The salinity and pollution of environments can vary drastically within just a few meters, so even maps with resolutions of hundreds of meters (which are very high for climate data) are subject to significant imprecision. Therefore, the project focus shifted to location-oriented case studies. The contribution of this work lies in the improvement of the models and in the development of the corrosion system to which they are coupled, along with components necessary for data collection, modeling, case study creation, and user application.

1.4 STRUCTURE OF THE WORK

The proposed development for this project was divided into five chapters. Chapter 2 explains the theoretical background concepts required to understand this work. Chapter 3 presents a systematic literature review on the use of machine learning for atmospheric corrosion. Chapter 4 describes the proposed approach, including the sys-

tem description and analyses. Chapter 5 presents the results of the experiments. Chapter 6 discusses the results and the project development schedule.

2 BACKGROUND

This chapter provides the theoretical background necessary to understand the scope of the project. The first section discusses the science of atmospheric corrosion, its technical standards, and the corrosion data available in the literature, as well as the available atmospheric corrosion systems. The second section discusses the available atmospheric data, the concept of reanalysis data, and the main data catalogs. The third section discusses geospatial data, such as distance from the coast, elevation, and land cover, along with their sources. The fourth section introduces the concept of machine learning, the models used in this project, and the validation techniques and evaluation metrics.

2.1 ATMOSPHERIC CORROSION

Corrosion is the degradation of a material's properties or the loss of mass over time due to interaction with the environment. This deterioration reflects the natural tendency of material compounds to return to their most thermodynamically stable state. For most metallic materials, corrosion involves the formation of oxides and sulfides on the material's surface. Under normal circumstances, iron and steel corrode in the presence of both oxygen and water (SCHWEITZER, 2006). The most common form of corrosion, rusting, refers to the corrosion of iron and plain carbon steel. Rust is ferric oxide, which appears in the familiar red or dark brown color. Other non-ferrous metals, such as aluminum, copper, and zinc, also corrode, but they do not rust; each produces its own corrosion products (MCCAFFERTY, 2010).

The corrosion of metallic materials is an electrochemical process. For any metal M , the electrochemical corrosion process consists of anodic regions (where oxidation occurs) and cathodic regions (where reduction occurs) distributed over the metal surface. In the anodic process, the metal dissolves and is transferred to the solution as M^{+2} ions (cations of metal M). The process resembles an electrical circuit in which there is no accumulation of charges (PANNONI, 2017).

Corrosion in outdoor environments, or simply atmospheric corrosion, has received considerable attention because of the high costs associated with the large number of outdoor structures, such as buildings, fences, bridges, towers, automobiles, ships, and many other applications exposed to the atmosphere. The electrochemical mechanism of atmospheric corrosion involves the base metal, the condensation of moisture as the electrolyte, and the surrounding atmosphere. Therefore, when relative humidity is below 30%, corrosion is negligible. Many variables, such as relative hu-

midity, temperature, pollutants, amount of rainfall, geographical location, and even the position of the metal, influence atmospheric corrosion behavior (SCHWEITZER, 2006).

The science of atmospheric corrosion began in the 1920s and 1930s, when W. H. J. Vernon started systematic field exposure of metals to the environment (VERNON, 1924). Except for some increased sophistication in instrumentation, his experiments were very similar to modern ones: he cleaned metal samples, exposed them to specific concentrations of gases such as sulfur dioxide (SO_2) and carbon dioxide (CO_2), or to natural outdoor environments, and then determined the corrosion rates. It became evident that commonly used metals, particularly steel, copper, zinc, and aluminum, suffered corrosion when exposed in heavily polluted environments. The environments were categorized into rural, marine, urban, and industrial, and it was recognized that the metals exhibited different corrosion behaviors in each of these environments (LEYGRAF et al., 2016).

In urban and industrial atmospheres, all types of sulfur dioxide contamination are significant. The burning of fossil fuels generates large amounts of sulfur dioxide, which is converted into sulfuric and sulfurous acids in the presence of moisture. In marine environments, chloride attack results from the deposition of fine droplets or crystals formed by the evaporation of salt spray carried inland by the wind from the sea (SCHWEITZER, 2006).

Due its enormously complex and interdisciplinary nature, research on atmospheric corrosion stagnated for several decades following Vernon's experiments (LEYGRAF et al., 2016). It became clear that isolated experiments were not sufficient, and that standardization and systematic exposure of metals around the world were necessary. Following the formation of the Technical Committee on Corrosion of Metals and Alloys (ISO/TC 156) in 1974, three ambitious metal exposure programs were carried out during the 1980s and 1990s: the International Exposure Program (ISOCORRAG), the Maps of Corrosivity for Ibero-America (MICAT), and the International Cooperative Programme within the United Nations Economic Commission for Europe (ICP/UNECE).

Figure 3 shows a venn diagram of the countries that participated in these three exposure programs. ISOCORRAG had stations in America, Europe, Asia and Oceania, MICAT focused on Ibero America (Latin America, Portugal and Spain) and ICP Material focused on the European countries. Both MICAT and ISOCORRAG had Antarctic stations. Together, the programs combined had more than 100 stations in 38 countries.

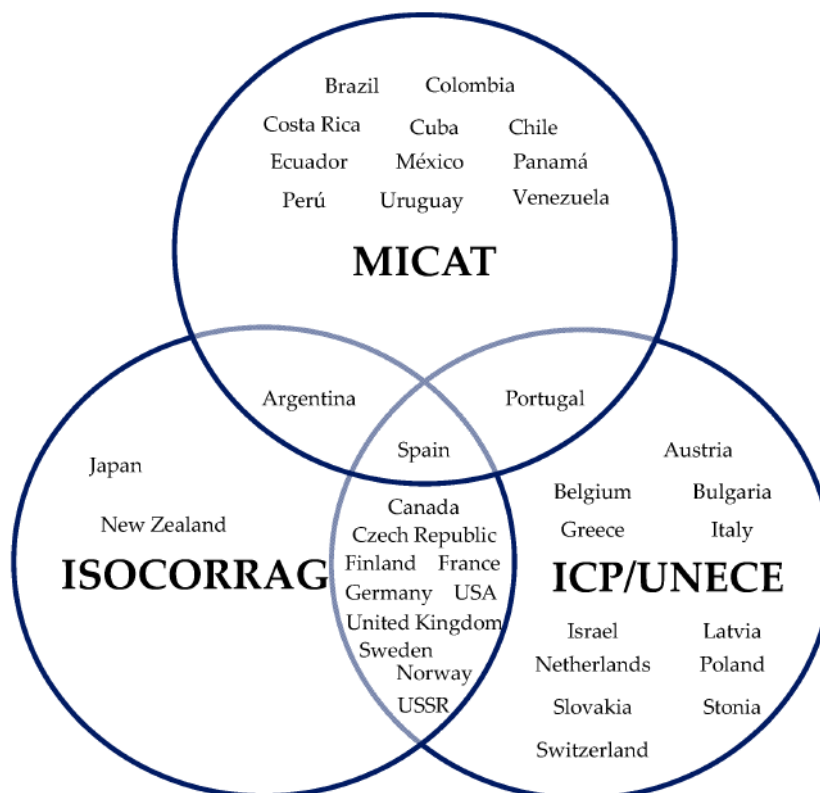


Figure 3 – Diagram of the three major exposure programs and the countries involved in each. Source: (CHICO et al., 2017)

Although each program had its own particular characteristics, they nevertheless shared several common objectives and methodology, such as atmosphere classification of atmospheres (rural, coastal, urban and industrial) and one year expositions of zinc, carbon steel, aluminum and copper. These methodological similarities provided a meeting point between the three programs, allowing, for the first time, a worldwide perspective on the problem of atmospheric corrosion, covering a broad spectrum of climatological and atmospheric pollution conditions never before considered in the extensive published literature on atmospheric corrosion (CHICO et al., 2017).

The exposure programs on metals enabled the improvement of international technical standards. Current versions of the standard include detailed protocols on how to construct racks and material exposure test specimens (ISO/8565, 2011), procedures for collecting pollutants (ISO/9225, 2012), and the classification of environmental corrosivity and corrosion estimation (ISO/9223, 2012), including long-term exposures (ISO/9224, 2012). The data from these programs allowed the creation of equations for atmospheric corrosion prediction called Dose Response Functions (DRF), ICP UNECE data was used to create multi-pollutant DRF (KUCERA et al., 2007) and MICAT and ISOCORRAG were used to create the DRFs which were incorporated to the standards (MIKHAILOV; TIDBLAD; KUCERA, 2004).

2.1.1 ISO/TC 156 International Standards

Within the field of exposure and evaluation of specimens under atmospheric conditions, as well as the classification of atmospheric corrosivity, the standards developed by the Technical Committee on Corrosion of Metals and Alloys of the International Organization for Standardization (ISO/TC 156) should be primarily mentioned (WATT et al., 2009). The atmospheric corrosion standards developed by ISO/TC 156 define procedures for metal exposure (ISO/8565, 2011), calculation of mass or thickness loss in exposed metals (ISO/9226, 2012), collection of environmental parameters that affect corrosion (ISO/9225, 2012), definition of corrosivity categories for atmospheres, estimation of corrosivity through normative models (ISO/9223, 2012), and estimation of corrosion over long exposure periods (ISO/9224, 2012).

For the exposure of metals, ISO 8565 provides specifications for test specimens and corrosion stations. The standard recommends the use of flat sheet specimens measuring 15 cm × 10 cm, exposed on racks inclined at a 45° degree angle (ISO/8565, 2011). Corrosion stations typically follow the structure of the station at figure 4: the station is equipped with an exposure rack designed to hold multiple specimens, allowing several exposures to be conducted simultaneously. A complete atmospheric corrosion station also includes collectors for environmental parameters that influence corrosion, which can be seen on the left side of the station in Figure 4:



Figure 4 – An example of an atmospheric corrosion station: from left to right, one wet candle for chloride pollutants, three lead dioxide sulfation cylinders inside shelters, and six metal exposure racks.

To calculate corrosion in a specimen, it is weighed before placement in the

exposure rack. The specimen is then exposed for a period of time, after which it is removed, chemically cleaned using a standardized process (156, 2021), and weighed again. The difference in weight over the exposure period is calculated. Knowing the dimensions of the specimen, it is possible to convert mass loss into thickness loss, expressed in micrometers per year ($\mu\text{m/a}$) (ISO/9226, 2012).

The most commonly used exposure period is one year, since the corrosion rate for the first year of exposure (r_{corr}) is the main variable used by ISO 9223 tools. For the four main metallic materials used in the industry to produce alloys and metallic anti-corrosion coatings — carbon steel (iron, Fe), zinc (Zn), aluminum (Al), and copper (Cu) — it is possible to classify an atmosphere by its corrosivity using r_{corr} . Categories range from C1 (very low corrosivity) to C5 (very high corrosivity), with an additional CX (extreme corrosivity) category. Table 1 shows the r_{corr} ranges and their corresponding category values:

Corrosivity category	Corrosion rates of metals				
	Unit	Carbon steel	Zinc	Copper	Aluminium
C1	$\text{g}/(\text{m}^2\cdot\text{a})$	$r_{\text{corr}} \leq 10$	$r_{\text{corr}} \leq 0,7$	$r_{\text{corr}} \leq 0,9$	negligible
	$\mu\text{m/a}$	$r_{\text{corr}} \leq 1,3$	$r_{\text{corr}} \leq 0,1$	$r_{\text{corr}} \leq 0,1$	—
C2	$\text{g}/(\text{m}^2\cdot\text{a})$	$10 < r_{\text{corr}} \leq 200$	$0,7 < r_{\text{corr}} \leq 5$	$0,9 < r_{\text{corr}} \leq 5$	$r_{\text{corr}} \leq 0,6$
	$\mu\text{m/a}$	$1,3 < r_{\text{corr}} \leq 25$	$0,1 < r_{\text{corr}} \leq 0,7$	$0,1 < r_{\text{corr}} \leq 0,6$	—
C3	$\text{g}/(\text{m}^2\cdot\text{a})$	$200 < r_{\text{corr}} \leq 400$	$5 < r_{\text{corr}} \leq 15$	$5 < r_{\text{corr}} \leq 12$	$0,6 < r_{\text{corr}} \leq 2$
	$\mu\text{m/a}$	$25 < r_{\text{corr}} \leq 50$	$0,7 < r_{\text{corr}} \leq 2,1$	$0,6 < r_{\text{corr}} \leq 1,3$	—
C4	$\text{g}/(\text{m}^2\cdot\text{a})$	$400 < r_{\text{corr}} \leq 650$	$15 < r_{\text{corr}} \leq 30$	$12 < r_{\text{corr}} \leq 25$	$2 < r_{\text{corr}} \leq 5$
	$\mu\text{m/a}$	$50 < r_{\text{corr}} \leq 80$	$2,1 < r_{\text{corr}} \leq 4,2$	$1,3 < r_{\text{corr}} \leq 2,8$	—
C5	$\text{g}/(\text{m}^2\cdot\text{a})$	$650 < r_{\text{corr}} \leq 1\,500$	$30 < r_{\text{corr}} \leq 60$	$25 < r_{\text{corr}} \leq 50$	$5 < r_{\text{corr}} \leq 10$
	$\mu\text{m/a}$	$80 < r_{\text{corr}} \leq 200$	$4,2 < r_{\text{corr}} \leq 8,4$	$2,8 < r_{\text{corr}} \leq 5,6$	—
CX	$\text{g}/(\text{m}^2\cdot\text{a})$	$1\,500 < r_{\text{corr}} \leq 5\,500$	$60 < r_{\text{corr}} \leq 180$	$50 < r_{\text{corr}} \leq 90$	$r_{\text{corr}} > 10$
	$\mu\text{m/a}$	$200 < r_{\text{corr}} \leq 700$	$8,4 < r_{\text{corr}} \leq 25$	$5,6 < r_{\text{corr}} \leq 10$	—

Table 1 – ISO 9223 corrosivity classification intervals by r_{corr} values. Source: (ISO/9223, 2012).

If it is not possible to determine the corrosivity categories by exposing standard specimens, corrosivity may be estimated from corrosion loss calculated using environmental data or information about environmental conditions and exposure situations, through dose-response functions (DRFs) (ISO/9223, 2012). ISO 9223 provides DRFs for four standard metals, describing corrosion attack after the first year of open-air exposure as a function of sulfur dioxide (SO_2) dry deposition, chloride ions (Cl^-) dry deposition, temperature, and relative humidity. DRFs describe corrosion as a function of these parameters, with coefficients adjusted using data from two exposure programs (ISOCORRAG and MICAT), in addition to Soviet exposure data (PANCHENKO; SHU-

VAKHINA, 1982). Equations 2.1, 2.2, 2.3, and 2.4 show the DRFs for carbon steel, zinc, aluminum, and copper, respectively.

Carbon Steel

$$r_{corr} = 1,77.P_d^{0,52}.exp(0,020.RH + f_{st}) + 0,102.S_d^{0,62}.exp(0,033.RH + 0,04T) \quad (2.1)$$

$$f_{st} = 0,150.(T - 10) \text{ when } T \leq 10^\circ C, \text{ otherwise } -0,054.(T - 10)$$

Zinc

$$r_{corr} = 0,0129.P_d^{0,44}.exp(0,046.RH + f_{zn}) + 0,0017.S_d^{0,57}.exp(0,008.RH + 0,085T) \quad (2.2)$$

$$f_{zn} = 0,038.(T - 10) \text{ when } T \leq 10^\circ C, \text{ otherwise } -0,071.(T - 10)$$

Copper

$$r_{corr} = 0,0050.P_d^{0,26}.exp(0,059.RH + f_{Cu}) + 0,010.S_d^{0,27}.exp(0,036.RH + 0,049.T) \quad (2.3)$$

$$f_{Cu} = 0,126.(T - 10) \text{ when } T \leq 10^\circ C, \text{ otherwise } -0,080.(T - 10)$$

Aluminum

$$r_{corr} = 0,0042.P_d^{0,73}.exp(0,025.RH + f_{Al}) + 0,0018.S_d^{0,60}.exp(0,020.RH + 0,094.T) \quad (2.4)$$

$$f_{Al} = 0,009.(T - 10) \text{ when } T \leq 10^\circ C, \text{ otherwise } -0,043.(T - 10)$$

Where r_{corr} is the corrosion rate for one year of exposure, T is the annual average temperature in Celsius (C), RH is annual average relative humidity as a percentage (%), P_d annual average SO_2 deposition and S_d is the annual average Cl^- deposition. The collection method for each of the four variables is defined by ISO 9225. Temperature and relative humidity can often be obtained from meteorological authorities in the country. SO_2 and Cl^- deposition are airborne contaminants that must be collected on the site, using proper measuring equipment, instruments and samples specified in the standard. For SO_2 is main collection method is the lead dioxide sulfation cylinder and for Cl^- is the wet candle, other methods are specified in the standard, but need proper conversions before being applied in the DRFs (ISO/9225, 2012). Figures 5a and 5b shows a wet candle and a lead dioxide sulfation cylinder respectively:



(a) An Chloride Wet Candle. Source: (BRANDENBURG, 2024)



(b) A lead dioxide sulfation cylinder. Source: (GIRÃO, 2007)

Figure 5 – Pollutant Candles from ISO 9223.

The wet candle consists of a single or double layer of surgical gauze kept moist by submerging its ends in a container of distilled water, as shown in Figure 5a. The gauze is attached to a wick, and its exposed surface area is known and used to calculate deposition (ISO/9225, 2012). Wet candle measurements also allow the classification of locations by salt pollution levels, ranging from S_0 to S_3 , with deposition intervals from 0 to 1500 mg/m²·day (ISO/9225, 2012).

The lead dioxide sulfation cylinder consists of a ceramic or plastic cylinder covered with cloth coated in lead dioxide (PbO_2) paste. The cylinder must be protected from rain and snow without obstructing airflow, so shelters are used, as shown in Figure 5b. Lead dioxide cylinder measurements also allow classification of locations by sulfur compound levels, ranging from P_0 to P_3 , with deposition intervals from 0 to 250 mg/m²·day (ISO/9223, 2012).

2.1.2 Atmospheric Corrosion Estimation

Although DRFs are the models defined by the standard, they are mainly limited by their input variables. First, to create a unified database to validate these models, it was necessary to find common ground among the material exposure projects. For example, ISOCORRAG did not include temperature and relative humidity measurements, so a heuristic was applied to estimate these parameters. The Russian Exposure Program and ICP Materials did not collect salt data (MIKHAILOV; TIDBLAD; KUCERA, 2004). Similar limitations apply to other exposure projects. Table 6 illustrates the variable availability problem in the six exposure programs and the pollutants collected. Besides SO_2 deposition, no other pollutant was collected by all projects.

Dry deposition							Wet deposition
Program	SO_2	NO_2	O_3	HNO_3	Cl	Particles	Precipitation (amount, composition)
NATO - CCMS	X	X			X		
ISO CORRAG	X				X		
MICAT	X				X		
ICP Materials	X	X	X				X
MULTI-ASSESS	X	X	X	X		X	X
RAPIDC	X	X	X	X		X	X

Figure 6 – Different exposure programs by variables collected. Source: (WATT et al., 2009)

Due to differences in the variables, different models have been proposed for each exposure program. For ICP Materials, which mainly includes continental European stations, a multi-pollutant dose–response function was developed (KUCERA et al., 2007). For MICAT, an artificial neural network was trained using the variables collected in the project (PINTOS et al., 2000). Dose–response functions based on Cl^- , SO_2 , and time of wetness (TOW) were also proposed but later discarded (MIKHAILOV; TIDBLAD; KUCERA, 2004). A large number of statistical models based on machine learning techniques have been proposed. Chapter 3 presents a systematic literature review on the use of machine learning models for atmospheric corrosion prediction.

Another problem related to the variables is data collection. DRFs are presented as an alternative to exposing test specimens, but they face a similar issue, since SO_2 and Cl^- collectors also require long-term field experiments. In practice, defining atmospheric corrosivity through experimentation is a lengthy process that takes months or even years (SEBRAO; ORDINE, 2010). Therefore, alternatives have been proposed to

determine environmental corrosivity while minimizing or eliminating the need for field experimentation.

The most common approach to circumvent the experimentation problem is to use an atmospheric corrosion map. The idea behind a corrosion map is simple: for a selected region, samples are exposed at different points within that region, and criteria are applied to extrapolate the experimental data (interpolation and categorization are common methods) to the entire area, generating a map that can be used for future classification. (SANTA et al., 2022) conducted a literature review on corrosivity maps developed for different countries and regions worldwide. Interactive corrosivity maps for Chile (VERA et al., 2012) and the United Kingdom (CORROSION. . . , 2024) have been developed following this methodology.

Another approach made possible by advances in meteorology is the use of globally available atmospheric and geospatial data as input variables for models. In this approach, it is possible to collect environmental data remotely, without the need for field experiments. Sources include meteorological networks, satellite observations, map databases, reanalysis datasets, land-cover information, socioeconomic indicators, and others. An example of this approach is the development of dose–response functions using satellite data (KOUREMADAS et al., 2023). Corrosivity maps can also be integrated with this methodology: for instance, European corrosivity maps have been generated using pollution and wind data derived from climate models (TIDBLAD, 2012), and machine learning models trained on reanalysis data were employed to create South American corrosivity maps (GEREMIAS, 2024). Figure 7 shows the workflow used by (GEREMIAS, 2024) to generate such maps.

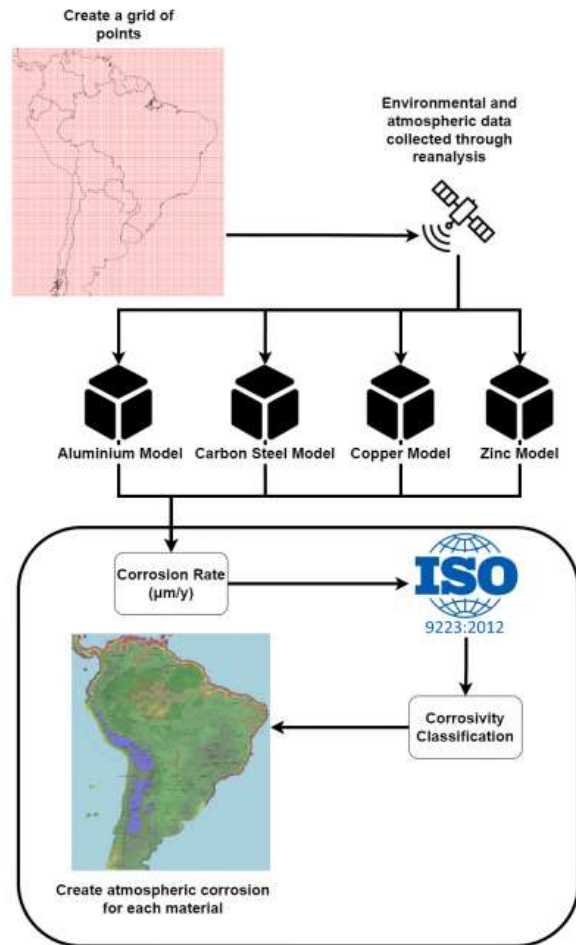


Figure 7 – Workflow used to create data-driven atmospheric corrosion maps. Source: (GEREMIAS, 2024)

This remote data-driven approach relies on three key elements:

- Atmospheric corrosion and pollution data as target variables, including precise location and exposure period of experiments as metadata.
- Atmospheric and geospatial datasets, providing global spatial coverage and long-term temporal consistency to construct input variables.
- Statistical learning models capable of assimilating atmospheric and geospatial variables to predict atmospheric corrosion values.

Therefore, the following background sections focus on exposure projects, atmospheric and geospatial data, and machine learning models, in order to define the necessary requirements for developing a data-driven approach to atmospheric corrosion prediction.

2.1.3 Atmospheric Exposure Data

There are two main sources of atmospheric corrosion exposure data: metal exposure programs and corrosivity maps. Exposure programs are the primary source, as they are generally large scale international initiatives. Some corrosivity map studies include a large number of corrosion stations and can reach the scale of exposure programs; therefore, they also represent important sources of corrosion data.

2.1.3.1 International Exposure Program (ISOCORRAG)

To establish a reliable reference dataset for evaluating atmospheric corrosion, minimize experimental inconsistencies that had undermined earlier studies, and support the revision of international standards, ISO/TC 156 launched the ISOCORRAG project. The initiative involved exposing metal plates and propellers to a wide range of atmospheric environments worldwide. In total, ISOCORRAG operated 53 monitoring stations across 14 countries, covering industrial, urban, rural, and marine settings located in temperate, tropical, and arctic coastal zones. Over its 8-year duration, the project recorded corrosion rates for four metals (carbon steel, zinc, copper, and aluminum), along with temperature, humidity, and pollutant concentrations (KNOTKOVA; KREISLOVA; DEAN, 2010). Figure 8 shows the map of ISOCORRAG corrosion stations.



Figure 8 – Map of ISOCORRAG stations. Source: (KNOTKOVA; KREISLOVA; DEAN, 2010)

2.1.3.2 Ibero-American Map of Atmospheric Corrosiveness (MICAT)

The MICAT Project (MORCILLO, 1995)(MORCILLO et al., 1998) stands out as one of the largest collaborative initiatives ever undertaken to study the atmospheric

corrosion of metallic materials. Conducted between 1988 and 1999, it brought together 14 Ibero-American nations (Argentina, Brazil, Colombia, Costa Rica, Cuba, Chile, Ecuador, Mexico, Panama, Peru, Portugal, Spain, Uruguay, and Venezuela), involving 72 corrosion monitoring stations and the participation of 143 researchers (MORCILLO et al., 2012). During the project, samples of carbon steel, zinc, copper, and aluminum were exposed at these sites for periods of 1, 2, 3, and 4 years. The dataset collected included environmental variables such as chloride and sulfate deposition, temperature, relative humidity, time of wetness, and rainfall. In Brazil, 12 corrosion stations were installed across different regions of the country. Figure 9 presents the geographical distribution of the monitoring sites.

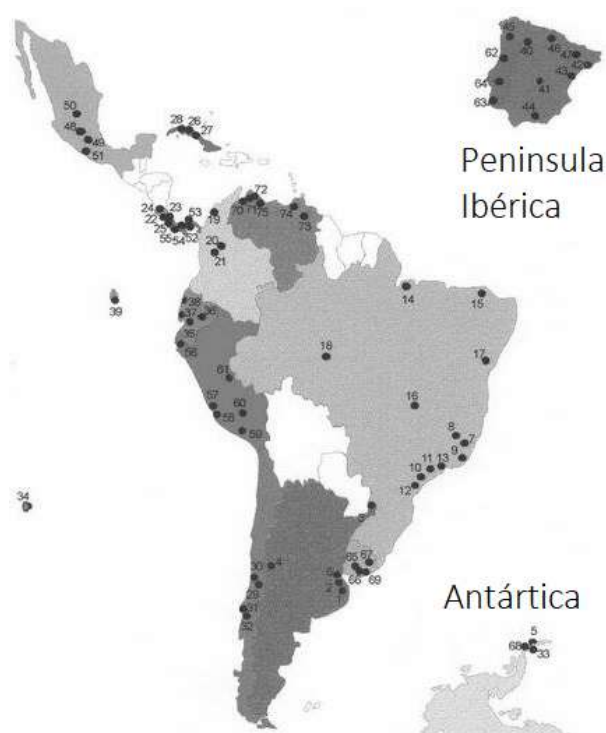


Figure 9 – Map of MICAT project stations. Source: (MORCILLO et al., 2012)

2.1.3.3 Atmospheric Corrosivity Map of Chile

The Chilean Corrosivity Map was coordinated by the University of Chile. According to its description, the initiative aimed at the “construction of atmospheric corrosivity maps of Chile for metals and alloys of greatest technological interest, enabling the optimal selection of materials for use in different environmental zones of the country” (VERA et al., 2012). A total of 31 monitoring stations were distributed across Chile’s diverse climatic regions. The project spanned three years, during which environmental and pollutant data were collected monthly, while samples of the four metallic materials defined by ISO 9223 were collected quarterly. Chile’s dataset is especially valuable due to both the extensive volume of records and the remarkable climatic variability of

the country, with stations established in coastal, industrial, desert, and even Antarctic environments. Figure 10 presents the geographical distribution of the monitoring sites.

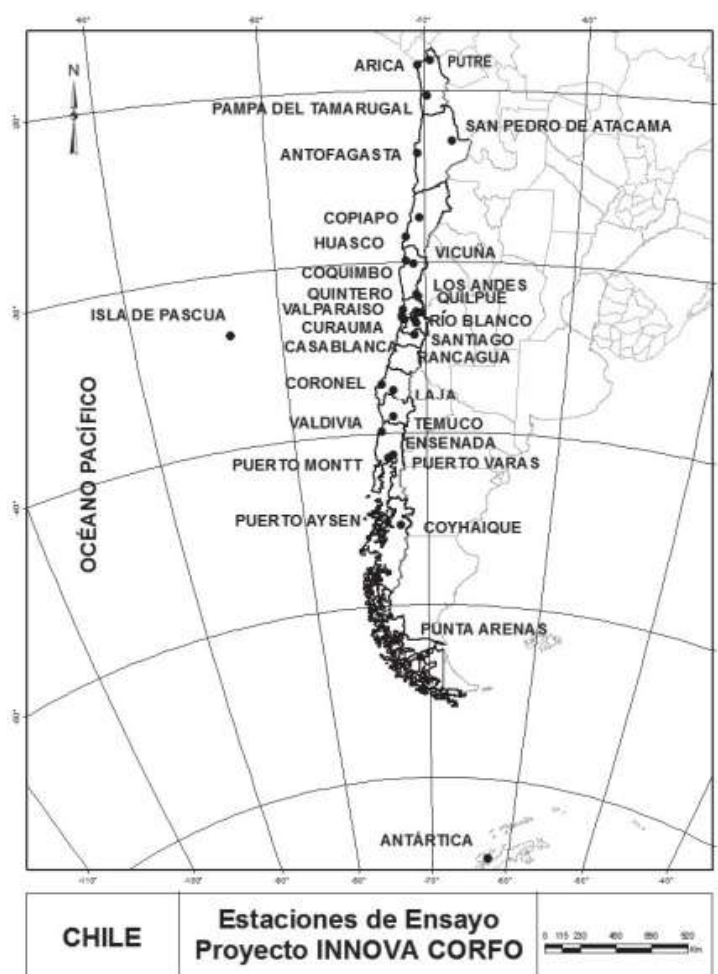


Figure 10 – Corrosivity map of Chile. Source: (VERA et al., 2012)

2.1.3.4 Atmospheric Corrosivity Maps of the Canary Islands

The research project titled “Development of Atmospheric Corrosion Maps for the Canary Islands” aimed to assess the corrosive behavior across the seven islands (SANTANA et al., 2019a). Over a period of three years, the study included 74 monitoring stations. Test samples of carbon steel, zinc, and copper were exposed, while chloride and sulfate pollutants were systematically measured.

The Canary Islands dataset is particularly noteworthy due to the extensive number of stations—comparable to major international initiatives—and the dominant influence of chloride on the islands’ corrosivity. The interplay between the islands’ varied climates and complex topography generates multiple microclimates within a single island, earning the archipelago the description of a “miniature continent” (SANTANA et al., 2019b). Figure 11 illustrates the locations of the monitoring sites.

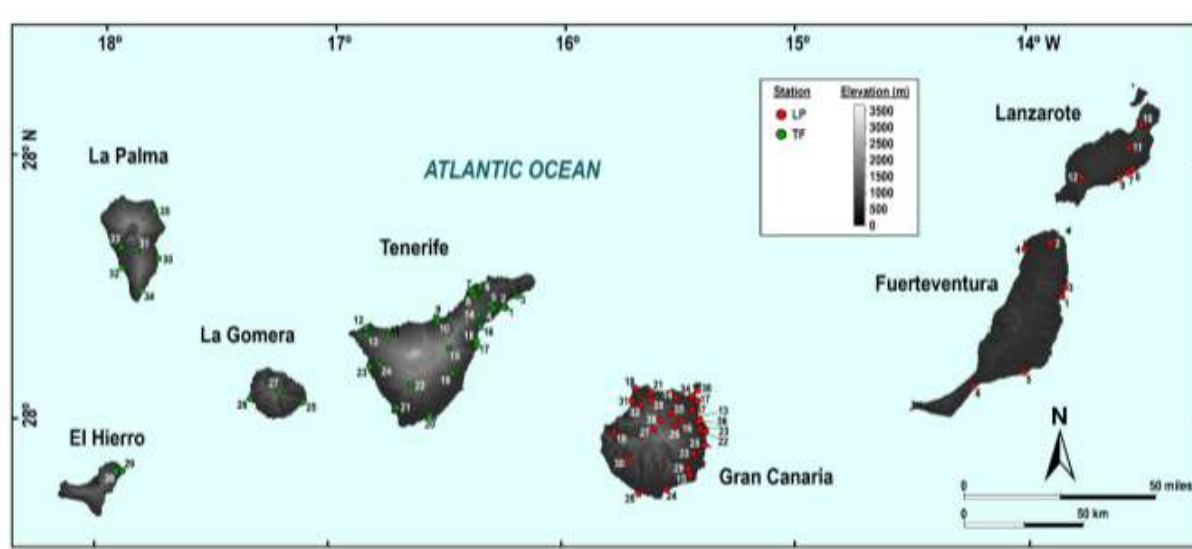


Figure 11 – Locations of monitoring stations in the Canary Islands. Source: (SANTANA et al., 2019b)

2.1.3.5 CSIRO Asia Exposures

Regarding metal exposure in Asia, multiple studies have been coordinated by the Commonwealth Scientific and Industrial Research Organization (CSIRO). One exposure project across five countries (Australia, the Philippines, Thailand, Indonesia, and China), comprising 16 stations, exposed specimens of zinc and carbon steel (COLE et al., 2001). Another study evaluated salinity using wet candles at 34 coastal stations in Australia (COLE; PATERSON; GANTHER, 2003).

2.1.3.6 Other Atmospheric Exposures in Brazil

The corrosivity map for the state of Ceará was produced in 2003 by COELCE (currently Enel Ceará) through the Ceará Industrial Technology and Quality Center (NUTEC), in collaboration with the Technological Education Center Institute (CENTEC) and the Ceará Foundation for Meteorology and Water Resources (FUNCEME), as part of the project “Corrosion and Atmospheric Degradation of Electrical Materials – Cycle II” (LUNA et al., 2008). A total of 19 monitoring stations were installed with chloride and sulfate collectors, and six of these stations also exposed carbon steel and copper test specimens. Contaminant measurements were conducted throughout the four-year duration of the project. Figure 12a shows the locations of these stations across Ceará.

In São Luís (MA), the corrosivity map was developed as part of two projects coordinated by Companhia Energética do Maranhão (CEMAR) and Centrais Elétricas do Norte do Brasil (ELETRONORTE) (SICA, 2006; SICA et al., 2007). Fifteen stations were deployed on power poles and transmission towers between March 2002 and May 2005, at sites defined by the two companies. Environmental monitoring included chloride and sulfate deposition, temperature, relative humidity, global radiation, wind speed,

and wind direction, while test specimens of the four ISO 9223 materials were exposed. Figure 12b presents the resulting corrosivity map along with the station locations.

For Salvador, data were collected from eight stations located at COELBA substations in the city and surrounding areas. Exposures began in September 2009 and lasted one year. Chloride and sulfur deposition measurements were taken, and test specimens of the four ISO 9223 materials were deployed (PORTELLA et al., 2012). Figure 12c displays the corrosivity map for the monitored regions, as well as the locations of the substations. Notably, the map in the publication includes 19 substations—eleven more than those directly used in the study—indicating that additional data, possibly from previous research, were incorporated.

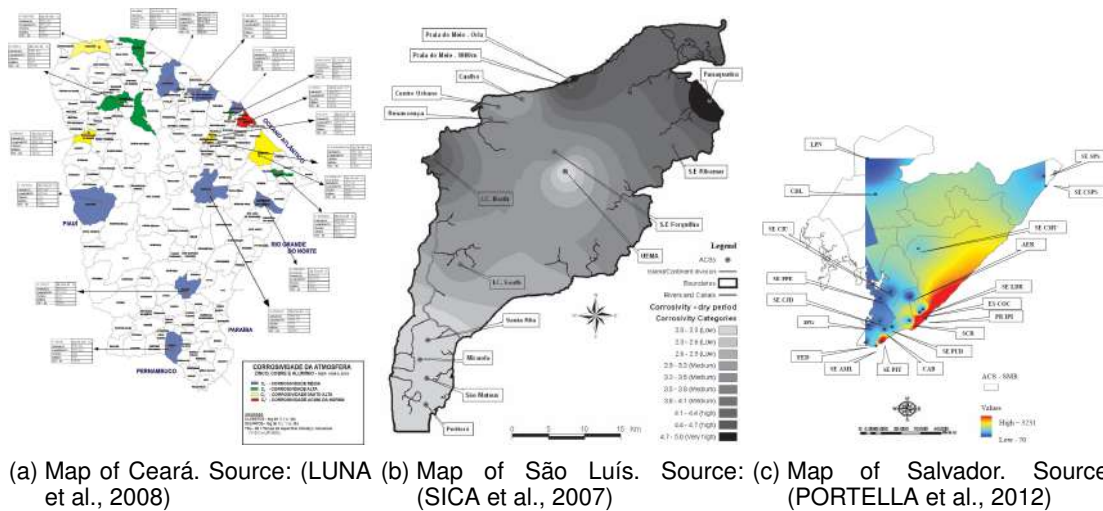


Figure 12 – Atmospheric Corrosion Maps of Brazil

In addition to corrosivity maps, other studies were coordinated in Brazil. Studies measuring salinity varying with distance from the sea were conducted in Recife (BEZERRA-PONTES et al., 2011) and Santa Catarina (VITALI, 2013).

2.2 CLIMATE AND GEOSPATIAL DATA PRODUCTS

Currently, meteorological and space agencies provide large volumes of weather and climate data. This data is generated through observations and climate models, and made available as datasets stored in data catalogs. Among these catalogs, notable examples include the Copernicus Climate Data Store from the European Union's Earth Observation Program (Copernicus) and the Goddard Earth Sciences Data and Information Services Center (GES DISC) from the North American National Aeronautics and Space Administration (NASA).

Among geospatial datasets, there are land cover (ESA, 2017), elevation (SANDWELL; SMITH; AND, 2014), distance to coastline datasets (OBPG, 2023), among others.

For atmospheric datasets, reanalysis products are generally used, for both meteorological (HERSBACH et al., 2020) and pollutant variables (GELARO et al., 2017). Emissions maps should also be mentioned (CRIPPA et al., 2018). Reanalysis data, like many scientific datasets, are commonly distributed in the Network Common Data Form (NetCDF) format. NetCDF provides a set of software libraries and a machine-independent data format for creating, accessing, and sharing array-oriented scientific data, and it has become a community standard for exchanging scientific datasets (UNIDATA, 2025).

2.2.1 ERA-5 and MERRA-2 Atmospheric Reanalysis

Although experimental methods remain the most common approach for determining contaminant levels in a given microclimate, alternative sources exist at the mesoscale (resolutions on the order of kilometers), such as atmospheric reanalyses. Atmospheric reanalysis is a systematic method for generating consistent datasets for climate monitoring and research. These datasets are produced through data assimilation of observations and numerical climate models, providing a dynamically consistent estimate of the atmospheric state at each time step (FASULLO et al., 2016). Data sources include radiosondes, satellites, buoys, aircraft, ships, and other observational platforms. Reanalysis datasets are multivariate, spatially and temporally complete, and gridded.

Continuous reanalysis grids also mitigate the spatial imbalance typical of meteorological station networks. In Brazil, for instance, the national meteorological authority, the *Instituto Nacional de Meteorologia* (INMET), operates more than 700 stations (INMET, 2023a; INMET, 2023b), but most are concentrated in densely populated areas, leaving sparsely populated regions like the Amazon rainforest underrepresented. Two prominent reanalyses stand out: ERA-5, containing high-resolution meteorological data (HERSBACH et al., 2020), and the *Modern-Era Retrospective Analysis for Research and Applications, version 2* (MERRA-2), containing atmospheric pollutant data (GELARO et al., 2017).

ERA-5 is the fifth generation of atmospheric reanalyses produced by the *European Centre for Medium-Range Weather Forecasts* (ECMWF). It is generated using hybrid incremental 4D-Var data assimilation together with the ECMWF *Integrated Forecast System* (ECMWF IFS), version Cy41r2 (HERSBACH et al., 2020).

ERA-5 provides data from 1940 to the present, with a horizontal resolution of 31 km (distributed at $0.25^\circ \times 0.25^\circ$ in geographical degrees for atmospheric variables) (HERSBACH et al., 2023), and with 37 available pressure levels, in addition to single-level and surface pressure products (ECMWF, 2023). Its spatial resolution is higher than that of MERRA-2; however, unlike MERRA-2, it does not include aerosol data.

Notable ERA-5 variables include near-surface temperature (2 m), wind speed (zonal, meridional, and mean), temperature profiles, precipitation, evaporation, UV radiation, cloud data, ocean wave parameters, atmospheric pressure, and albedo (ECMWF, 2023). Figure 13 illustrates ERA-5 data, showing a map of 2-meter temperature for January 2023.

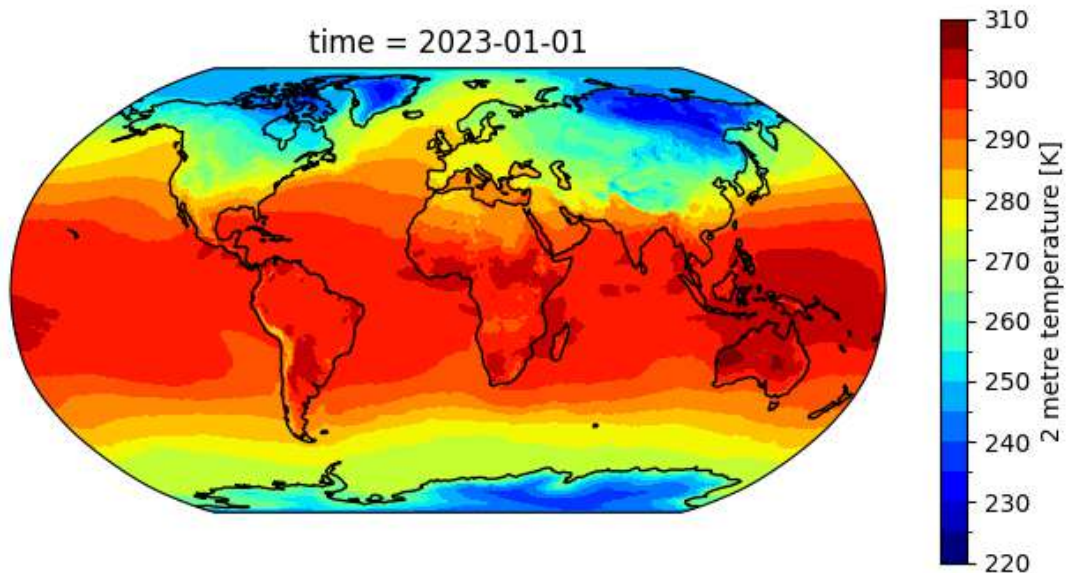


Figure 13 – ERA-5 map of 2-meter temperature for January 2023. Source: (HERSBACH et al., 2020)

MERRA-2 (GELARO et al., 2017) is the current version of the *Modern-Era Retrospective Analysis for Research and Applications* developed by NASA's *Global Modeling and Assimilation Office*. It is built using the *Goddard Earth Observing System Model, version 5 atmospheric data assimilation system* (GEOS ADAS) (RANGLES et al., 2017). MERRA-2 provides data from 1980 to the present and employs a cubed-sphere horizontal discretization (i.e., the Earth is represented as a cubic grid mapped onto a sphere) with an approximate resolution of 50 km × 70 km and 72 hybrid-eta levels from the surface up to 0.01 hPa. In addition to standard meteorological analyses, MERRA-2 also includes aerosol analysis (GELARO et al., 2017).

Aerosols in MERRA-2 are simulated using the radiatively coupled version of the *Goddard Chemistry, Aerosol, Radiation, and Transport model* (GOCART) (RANGLES et al., 2017; CHIN et al., 2002). GOCART represents the sources, sinks, and chemistry of dust, sulfate, sea salt, elemental carbon, and organic carbon. Dust and sea salt emissions are parameterized as functions of wind speed (RANGLES et al., 2017). Sea salt emissions in the model follow the parameterization of (MONAHAN; SPIEL; DAVIDSON, 1986). Figure 14 shows an example of MERRA-2 output: a map representing the surface mass concentration of sea salt for January 2023.

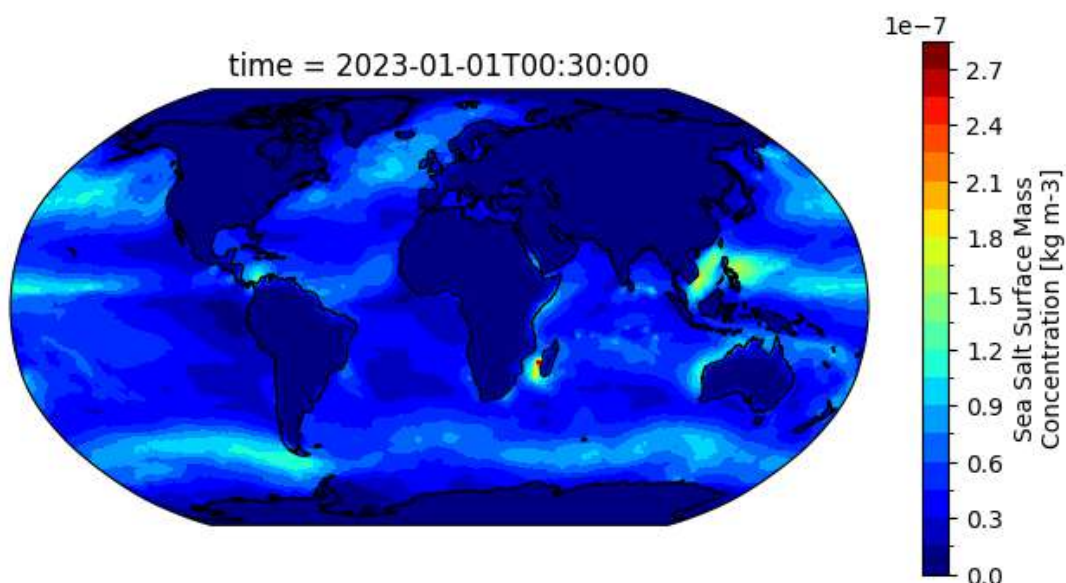


Figure 14 – MERRA-2 map of surface sea salt mass concentration for January 2023. Data source: (GELARO et al., 2017)

2.2.2 Emissions Database for Global Atmospheric Research (EDGAR)

The Emissions Database for Global Atmospheric Research (EDGAR) compiles gaseous and particulate air pollutant emissions, using standardized anthropogenic sectors (CRIPPA et al., 2018). EDGAR has a spatial resolution of 0.1° (approximately 10 km) and provides annual data from 1970 to the present, organized by sector, country, or gridded format. It includes emissions of greenhouse gases (CO_2 , CH_4 , N_2O) and air pollutants (SO_2 , NO_x , CO , $NMVOC$, NH_3 , PM_{10} , $PM_{2.5}$, BC , OC).

Although EDGAR contains emissions rather than deposition data, its importance for atmospheric corrosion studies should not be overlooked. EDGAR emissions have a higher spatial resolution than pollutant deposition data from reanalyses, providing insights into metal corrosivity near emission sources.

2.2.3 Distance to the Sea

The dataset of distances to the nearest coastline was first developed in 2009 by NASA's Ocean Biology Processing Group (OBPG) for oceanographic applications (BAILEY, 2024). The original data were produced at a spatial resolution of 0.04° , and later bilinearly interpolated to 0.01° . The estimated positional uncertainty across the grid is approximately 1 km (OBPG, 2023).

Although widely used, this dataset has limitations for the present study. A horizontal accuracy of 1 km is too coarse for direct applications, as marine aerosol concentrations can vary sharply within just a few meters from the shoreline. Nonetheless, the dataset remains valuable for identifying the closest coastal point to a given loca-

tion—an essential feature for models that depend on in-situ measurements near the shore. Figure 15 illustrates the global map derived from this dataset.

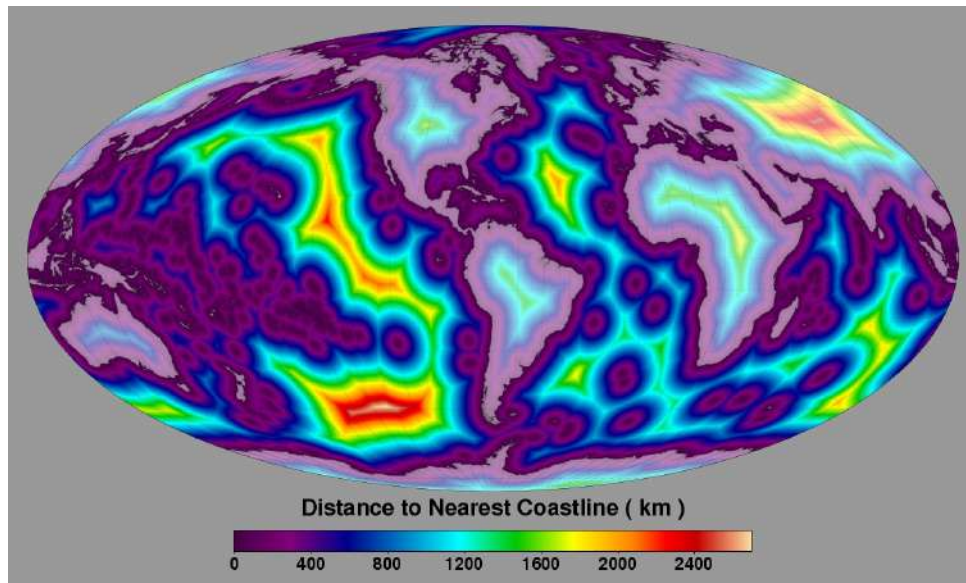


Figure 15 – Distance to coastline dataset. Data source: (OBPG, 2023)

2.2.4 Elevation

The STRM30+ is a digital elevation model (DEM) with global resolution of 1 km, derived from data collected by NASA's Shuttle Radar Topography Mission (SRTM) and provided by the United States Geological Survey (USGS) (SANDWELL; SMITH; AND, 2014). Digital elevation models provide elevation information for any point on Earth, on land or in the oceans. Elevation is particularly useful in the context of salinity, as it represents the topography of a site and is associated with other factors such as humidity, pressure, and wind flow patterns. Figure 16 shows the global map generated from the 1 km STRM30+ DEM.

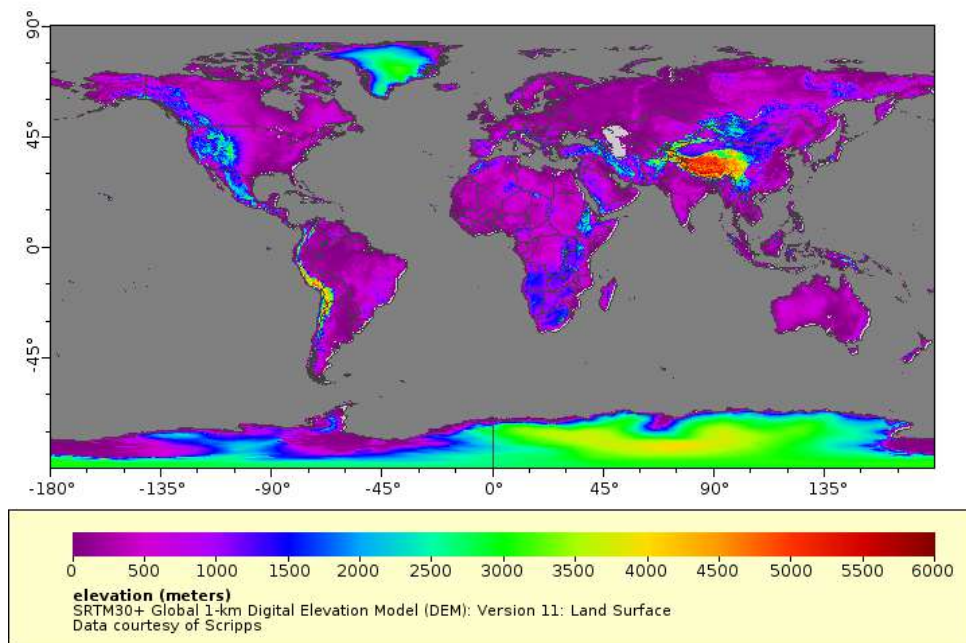


Figure 16 – SRTM30+ digital elevation model. Source: (SANDWELL; SMITH; AND, 2014)

2.2.5 Land Cover and Surface Roughness

Land cover data describe the extent to which a region is covered by forests, wetlands, impervious surfaces, agricultural areas, and other land or water types, typically derived from satellite observations (NOAA, 2024). A global land cover dataset represents land cover patterns at a given spatial resolution and classification system.

One widely used dataset is the European Space Agency's Climate Change Initiative Land Cover dataset (ESA CCI LC), which provides annual maps from 1992 to 2015 at 300 m resolution (ESA, 2017). This dataset uses the *Land Cover Classification System* (LCCS), dividing global land cover into 22 categories. Complementing ESA CCI LC, the Copernicus Climate Change Service developed the C3S Global Land Cover dataset (C3S LC) for the years 2016 to present, ensuring continuity with ESA CCI LC (DEFOURNY et al., 2021).

Although higher-resolution land cover datasets exist, the advantage of ESA CCI LC and C3S LC lies in their temporal coverage: higher-resolution products are generally available only after 2010, whereas ESA CCI LC spans from 1992, which aligns with the exposure periods of major corrosion databases. Figure 17 shows the global land cover map from ESA CCI for 2015.

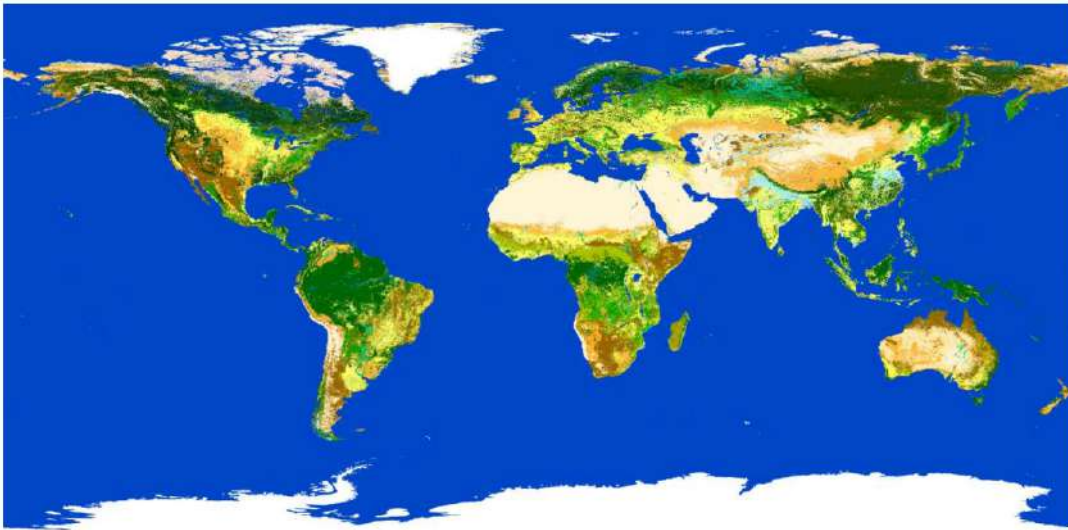


Figure 17 – ESA CCI Land Cover for 2015. Source: (ESA, 2017)

Land cover is relevant for corrosion and environmental salinity due to its influence on aerosol transport by wind, with surface roughness (e.g., trees, buildings) being the most significant factor. Small particles can follow wind flows and pass over obstacles, whereas heavier aerosols are affected by surface structures. Thus, forests and urban areas can strongly reduce airborne salt concentrations up to the height of obstacles (COLE; PATERSON; GANTHER, 2003). This drag effect caused by surface structures is quantified as surface roughness.

The surface roughness length (z_0) is a parameter used in numerical models to represent surface resistance to wind (LINACRE; GEERTS, 1997). It is a key parameter in the logarithmic wind profile, which calculates wind speeds in the lowest atmospheric layers (AMS, 2019). Baseline z_0 values are usually obtained experimentally, though in field studies, they can be estimated visually. The determination of z_0 can also be automated using land cover data, assigning each LCCS class a corresponding z_0 value (FLOORS et al., 2021). Table 2 presents the roughness scheme adopted by the Wind-Pro model (WINDPRO, 2014) for the LCCS classification system.

ID	Land cover type	EMD roughness class
11	Post-flooding or irrigated croplands (or aquatic)	$z_0 = 0.100$
14	Rainfed croplands	$z_0 = 0.100$
20	Mosaic cropland (50–70%) / vegetation (grassland/shrubland/forest) (20–50%)	$z_0 = 0.07$
30	Mosaic vegetation (grassland/shrubland/forest) (50–70%) / cropland (20–50%)	$z_0 = 0.07$
40	Closed to open (>15%) broadleaved evergreen or semi-deciduous forest (>5 m)	$z_0 = 0.5$
50	Closed (>40%) broadleaved deciduous forest (>5 m)	$z_0 = 0.4$
60	Open (15–40%) broadleaved deciduous forest/woodland (>5 m)	$z_0 = 0.4$
70	Closed (>40%) needleleaved evergreen forest (>5 m)	$z_0 = 0.5$
90	Open (15–40%) needleleaved deciduous or evergreen forest (>5 m)	$z_0 = 0.4$
100	Closed to open (>15%) mixed broadleaved and needleleaved forest (>5 m)	$z_0 = 0.4$
110	Mosaic forest or shrubland (50–70%) / grassland (20–50%)	$z_0 = 0.07$
120	Mosaic grassland (50–70%) / forest or shrubland (20–50%)	$z_0 = 0.07$
130	Closed to open (>15%) (broadleaved or needleleaved, evergreen or deciduous) shrubland (<5 m)	$z_0 = 0.07$
140	Closed to open (>15%) herbaceous vegetation (grassland, savannas or lichens/mosses)	$z_0 = 0.05$
150	Sparse (<15%) vegetation	$z_0 = 0.07$
160	Closed to open (>15%) broadleaved forest regularly flooded (semi-permanently or temporarily) – fresh or brackish water	$z_0 = 0.1$
170	Closed (>40%) broadleaved forest or shrubland permanently flooded – saline or brackish water	$z_0 = 0.1$
180	Closed to open (>15%) grassland or woody vegetation on regularly flooded or waterlogged soil – fresh, brackish or saline water	$z_0 = 0.03$
190	Artificial surfaces and associated areas (urban areas >50%)	$z_0 = 0.4$
200	Bare areas	$z_0 = 0.02$
210	Water bodies	$z_0 = 0.0002$
220	Permanent snow and ice	$z_0 = 0.001$
230	No data (burnt areas, clouds, . . .)	$z_0 = \text{N/A}$

Table 2 – Reference table of roughness length (z_0) for the LCCS land cover scheme. Adapted from (WINDPRO, 2014)

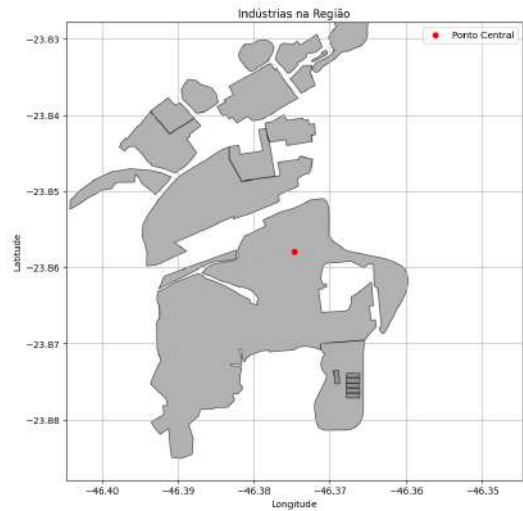
2.2.6 OpenStreetMap

OpenStreetMap (OSM) is an editable map database built and maintained by volunteers, and distributed under the Open Data Commons Open Database License. Started in 2004, the mapping infrastructure, along with many associated OSM projects, is developed and maintained by volunteers. Computing and network resources are donated by universities and organizations. For an introduction to the OSM project, see the official about page. Prior research projects using OSM data are also available for reference (OpenStreetMap contributors, 2017).

OSM contains more than 1.1 billion polygons and linear features, offering a rich database of urban and industrial sites, which can be used to study pollution. While Land Cover (LC) maps generally have better temporal consistency than OSM, they do not differentiate industrial from urban areas, which is essential for defining pollution in micro-climates. Figures 18a and 18 illustrate an industrial site in Cubatão, Brazil, as seen from satellite imagery and OSM polygons, respectively:



(a) Satellite image from Google Earth



(b) Polygons from OpenStreetMap.

Figure 18 – Images for an Industrial site in Cubatão, Brazil. Sources: (OpenStreetMap contributors, 2017), (Google, 2025)

2.3 MACHINE LEARNING

Machine learning is a subfield of computer science concerned with building algorithms that, to be useful, rely on a collection of examples of some phenomenon. These examples can come from nature, be handcrafted by humans, or be generated by another algorithm. Machine learning can also be defined as the process of solving a practical problem by first gathering a dataset and then algorithmically building a statistical model based on that dataset. That statistical model is assumed to be used somehow to solve the practical problem (BURKOV, 2019).

Machine learning can be classified according to the type of learning, which can be supervised, semi-supervised, unsupervised, or reinforcement learning. Broadly speaking, supervised learning involves building a model to predict an output based on one or more inputs. Problems of this nature occur in fields as diverse as business, medicine, astrophysics, and public policy. In unsupervised statistical learning, there are inputs but no supervising output; nevertheless, we can learn relationships and structures from such data (MCCAFFERTY, 2010).

It is possible to formulate a supervised learning problem as a function. A set of

inputs X is readily available, but the output Y cannot be easily obtained. In this setting, since the error term averages to zero, we can predict Y using Equation 2.5:

$$\hat{y} = \hat{f}(x) \quad (2.5)$$

where $\hat{y}$ represents our estimate for f , and $\hat{Y}$ represents the resulting prediction for Y . In this setting, $\hat{y}$ is often treated as a black box, in the sense that one is not typically concerned with the exact form of $\hat{y}$, provided that it yields accurate predictions for Y (MCCAFFERTY, 2010). If y takes its value from a discrete set of categories, we are dealing with a classification problem; otherwise, if y consists of one or more continuous variables, we are dealing with a regression problem.

2.3.1 Support Vector Regressors (SVR)

Support Vector Machines (SVM) are a popular machine learning algorithm widely regarded as one of the best “out-of-the-box” models (MCCAFFERTY, 2010). Originally conceived as classifiers, SVMs identify an optimal separating hyperplane, which is an $(n - 1)$ -dimensional subspace that maximises the margin between two classes. For a 3-dimensional feature space, this hyperplane is a plane; for a 2-dimensional space, it reduces to a line.

Although SVMs were originally designed for classification, they have a regression variant known as Support Vector Regression (SVR). Instead of separating classes, SVR fits a function $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ to the data within an adjustable ε -insensitive tube: predictions that fall within a distance ε of the true value incur no penalty, and only points outside the tube contribute to the loss (BRERETON; LLOYD, 2010). This tolerance zone is what gives SVR its robustness to small noise. Formally, the SVR primal optimisation problem is:

$$\min_{\mathbf{w}, b, \xi, \xi^*} \quad \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N (\xi_i + \xi_i^*) \quad (2.6)$$

subject to:

$$y_i - \mathbf{w}^\top \mathbf{x}_i - b \leq \varepsilon + \xi_i \quad (2.7)$$

$$\mathbf{w}^\top \mathbf{x}_i + b - y_i \leq \varepsilon + \xi_i^* \quad (2.8)$$

$$\xi_i, \xi_i^* \geq 0, \quad \forall i \quad (2.9)$$

where ξ_i and ξ_i^* are slack variables that allow individual points to lie outside the tube, and the regularisation parameter $C > 0$ controls the trade-off between tube width and

the total penalty incurred by those violations. By default, both SVM and SVR are formulated for linearly separable problems. To handle non-linear data, kernel functions can be used to implicitly map the inputs into a higher-dimensional feature space where a linear model becomes sufficient, a technique known as the *kernel trick* (DING et al., 2021). Given a feature map $\phi : \mathbb{R}^d \rightarrow \mathcal{H}$, the kernel function computes inner products in the transformed space without explicitly evaluating ϕ :

$$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{H}} \quad (2.10)$$

This substitution allows the SVR dual problem to be solved entirely in terms of pairwise kernel evaluations. The three most common kernel functions are:

- **Linear:**

$$k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j \quad (2.11)$$

- **Polynomial:**

$$k(\mathbf{x}_i, \mathbf{x}_j) = (\gamma \mathbf{x}_i^\top \mathbf{x}_j + r)^p \quad (2.12)$$

- **Radial Basis Function (RBF):**

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2) \quad (2.13)$$

where $\gamma > 0$, r , and p are kernel hyperparameters. The RBF kernel is the most widely used, as it can model complex non-linear relationships with a single tunable parameter γ .

The practical effect of the kernel choice is illustrated in Figure 19, which shows SVR fits to a sine curve using each of the three kernels. The solid line represents the model prediction and the shaded region the corresponding ε -tube. Circled points lying on the tube boundary are the support vectors the only training examples that influence the fitted model. The linear kernel fails to capture the non-linear behaviour of the data, whereas the polynomial and RBF kernels fit the curve closely, demonstrating the expressiveness that the kernel trick provides.

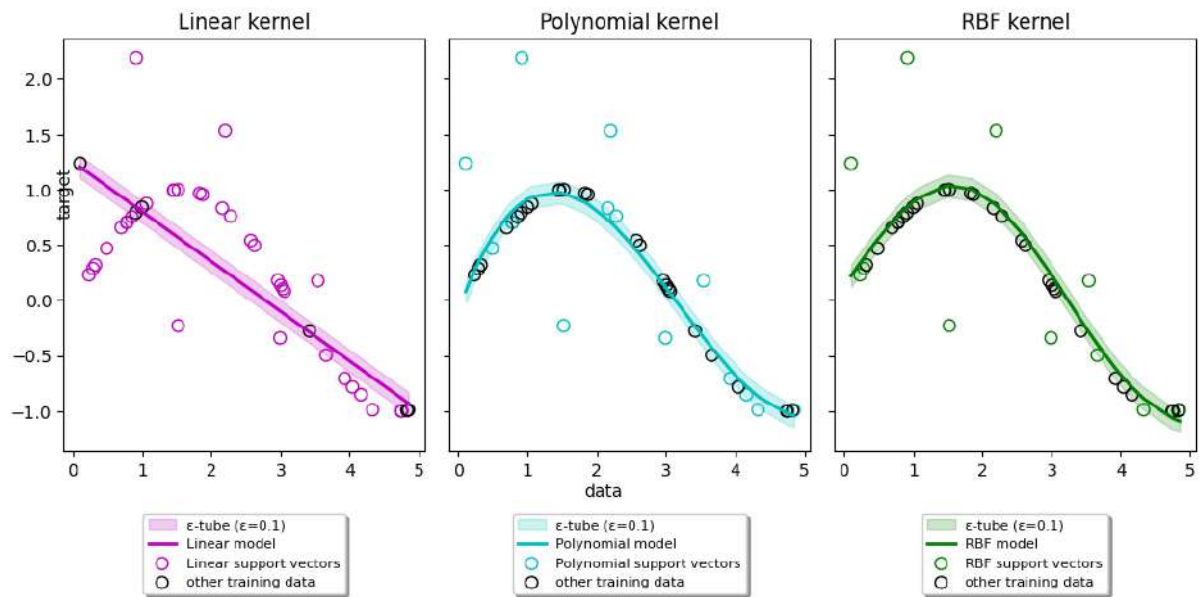


Figure 19 – SVR fits using linear, polynomial, and RBF kernels on a sine curve. Solid lines are model predictions; shaded regions are the ϵ -insensitive tubes; circled points are support vectors. Source: (SCIKIT-LEARN, 2026).

2.3.2 Random Forest Regressor

A decision tree classifies items by asking a series of tests or questions about their characteristics. Each question is contained in a node, and each node points to a child node for each possible answer to that question. The questions therefore form a hierarchy encoded as a tree (KINGSFORD; SALZBERG, 2008). Decision trees are easy to explain and are very similar to the methods humans use to make decisions. However, they are not very robust, as small deviations in the data can generate large deviations in the predicted values. By grouping multiple decision trees into a single model, we can substantially improve their performance (JAMES et al., 2013).

Random forests (BREIMAN, 2001) consist of training multiple trees and using the average of their predictions as the final result. The main premise of this approach is that, by combining several models, the errors of a single model are likely to be compensated for by others and, consequently, the overall performance will be better than that of a single predictor. This principle is called ensemble learning (SAGI; ROKACH, 2018). Additional operations, such as using only a subsample of the data for each tree (bootstrapping) or for each node (bagging), can also be applied.

Formally, given a training set $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, a random forest builds M randomized regression trees and aggregates their predictions by averaging. Each tree is grown on a bootstrap (or subsampled) replica of the data, and at every node, instead of evaluating all p available features, only a random subset of `mtry` features is considered for splitting. This double layer of randomness — resampling the data and restricting

the candidate features at each split — decorrelates the individual trees and is the key mechanism behind the variance reduction achieved by the ensemble (WAGER, 2016). Algorithm 1 describes, step by step, how a single prediction is obtained from a trained forest, following the formulation given by Biau and Scornet (WAGER, 2016).

Algorithm 1 Breiman’s random forest predicted value at $\mathbf{x}$ (WAGER, 2016)

Require: Training set $\mathcal{D}_n$, number of trees $M > 0$, $a_n \in \{1, \dots, n\}$, $\text{mtry} \in \{1, \dots, p\}$, $\text{nodesize} \in \{1, \dots, a_n\}$, and query point $\mathbf{x} \in \mathcal{X}$.

Ensure: Prediction of the random forest at $\mathbf{x}$.

```

1: for  $j = 1, \dots, M$  do
2:   Select  $a_n$  points, with (or without) replacement, uniformly from  $\mathcal{D}_n$ . Only these
   observations are used in the following steps.
3:   Set  $P \leftarrow \{\mathcal{X}\}$ , the list containing the cell associated with the root of the tree.
4:   Set  $P_{\text{final}} \leftarrow \emptyset$ .
5:   while  $P \neq \emptyset$  do
6:     Let  $A$  be the first element of  $P$ .
7:     if  $A$  contains fewer than  $\text{nodesize}$  points, or all  $\mathbf{X}_i \in A$  are equal then
8:       Remove  $A$  from  $P$  and append it to  $P_{\text{final}}$ .
9:     else
10:      Select, uniformly without replacement, a subset  $\mathcal{M}_{\text{try}} \subset \{1, \dots, p\}$  of cardinality  $\text{mtry}$ .
11:      Find the best split in  $A$  by maximising the CART-split criterion (Eq. 2.14) over the directions in  $\mathcal{M}_{\text{try}}$ .
12:      Cut  $A$  according to the best split, producing child cells  $A_L$  and  $A_R$ .
13:      Remove  $A$  from  $P$  and append  $A_L, A_R$  to  $P$ .
14:    end if
15:  end while
16:  Compute the predicted value  $m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n)$  as the average of the  $Y_i$  falling in the cell of  $\mathbf{x}$  in the partition  $P_{\text{final}}$ .
17: end for
18: Compute the forest estimate  $m_{M,n}(\mathbf{x}) = \frac{1}{M} \sum_{j=1}^M m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n)$ .

```

At each node, the best split (j^*, z^*) is the one that maximises the CART-split criterion (WAGER, 2016):

$$L_{\text{reg},n}(j, z) = \frac{1}{N_n(A)} \sum_{i=1}^n (Y_i - \bar{Y}_A)^2 \mathbf{1}_{\mathbf{x}_i \in A} - \frac{1}{N_n(A)} \sum_{i=1}^n \left(Y_i - \bar{Y}_{A_L} \mathbf{1}_{X_i^{(j)} < z} - \bar{Y}_{A_R} \mathbf{1}_{X_i^{(j)} \geq z} \right)^2 \mathbf{1}_{\mathbf{x}_i \in A} \quad (2.14)$$

where A is the cell being split, A_L and A_R are the candidate left and right child cells resulting from cut (j, z) , $N_n(A)$ is the number of training points falling in A , and $\bar{Y}_A$, $\bar{Y}_{A_L}$, $\bar{Y}_{A_R}$ denote the average response value within each respective cell. Intuitively, this criterion measures the reduction in the variance of Y achieved by splitting A into A_L and A_R : the larger the reduction, the better the split.

Figure 20 illustrates how a single trained tree produces a prediction for a given input. Starting at the root node, the input traverses the tree by following, at each internal node, the branch corresponding to the outcome of the associated test, until a leaf is reached. The path actually taken by the input — that is, the sequence of nodes visited from the root to the predicted leaf — is highlighted in blue, while the remaining (unvisited) nodes and branches are shown in their default colour. The predicted value returned by the tree corresponds to the average of the training responses contained in the highlighted leaf node, as described in Algorithm 1. In a random forest, this same procedure is repeated independently for each of the M trees, and the final prediction is obtained by averaging the individual leaf values across all trees.

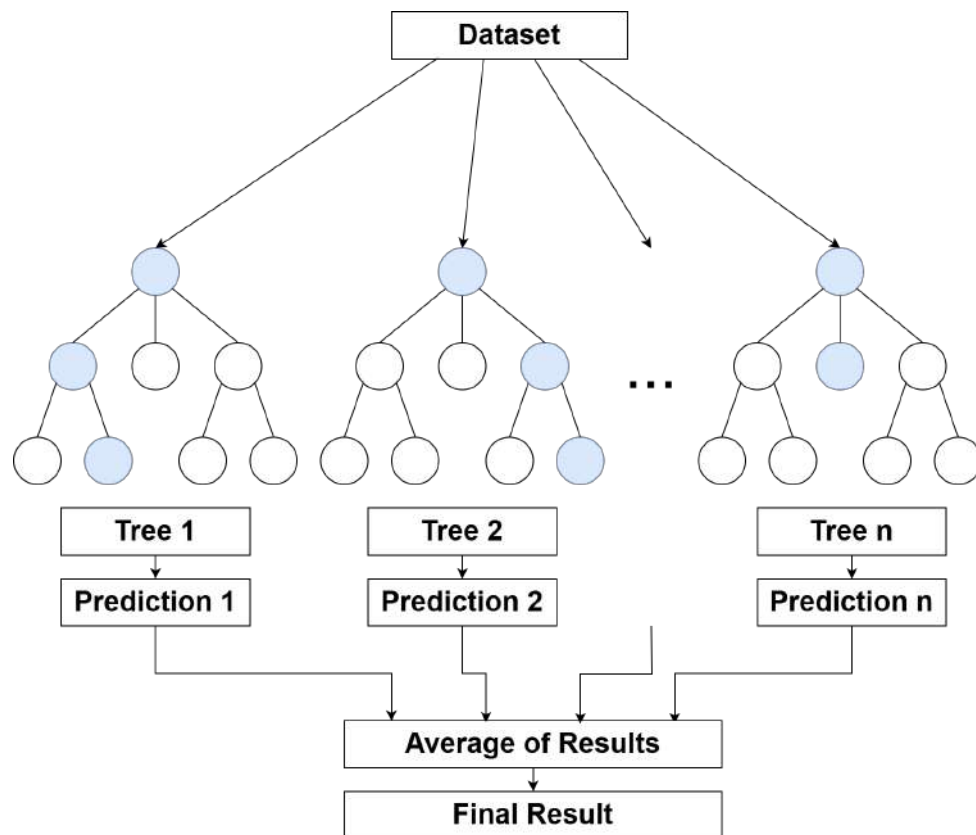


Figure 20 – Caption

2.3.3 Genetic Programming and the PySR Algorithm

The search for an equation that best describes a problem, both in terms of accuracy and simplicity, is called symbolic regression, for which genetic programming algorithms are typically applied. Genetic Programming (GP) is a systematic method for getting computers to automatically solve a problem starting from a high-level statement of what needs to be done. It is a domain-independent method that genetically breeds a population of computer programs to solve a problem. Specifically, GP iteratively transforms a population of programs into a new generation by applying analogs of naturally occurring genetic operations (KOZA; POLI, 2005).

Genetic programming is an extension of the genetic algorithm, but instead of having a population of strings that encode candidate solutions to a problem, it uses programs encoded as expression trees. In such trees, intermediate nodes are mathematical operations called functions, and leaf nodes are constants or variables called terminals. The genetic operations applied between expression trees include selection, crossover (sexual recombination), mutation, reproduction, gene duplication, and gene deletion (KOZA; POLI, 2005). Figure 21 shows examples of mutation and crossover operations, respectively. A mutation randomly changes the value of a node in a tree, and crossover exchanges branches between two trees.

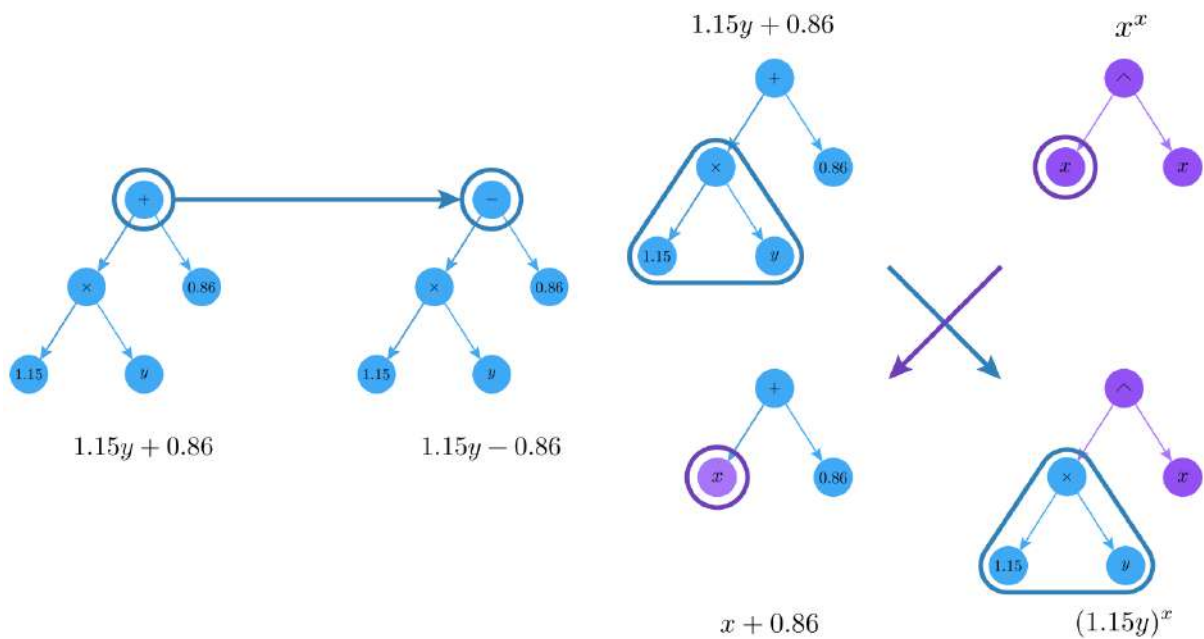


Figure 21 – Genetic operations in expression trees. Source: (CRANMER, 2023)

Among modern variations of Genetic Programming algorithms, the PySR algorithm evolves equations through an evolutionary system that operates with multiple populations of candidate expressions—internally represented as expression trees. These populations evolve in parallel and independently, typically distributed across multiple processing cores or computational clusters. A migration process is periodically performed, in which the best expressions found in one population are shared with others, improving the efficiency of the search for good solutions. PySR also incorporates simulated annealing into the selection process. This means that a “temperature” parameter controls the probability of accepting mutations that may worsen the fitness of an expression. At high temperatures, for example, the algorithm may accept worse solutions more frequently, allowing broader exploration and avoiding local optima. Conversely, as the temperature decreases over the course of the search, the algorithm becomes more selective, focusing on optimizing the best solutions found (CRANMER, 2023).

2.3.4 Model Evaluation

To measure a models performance, its necessary both an evaluation metric and a validation sampling scheme to avoid bias. Model performance in regression problems is often measured using statistical metrics that quantify the discrepancy between predicted and observed values. Mean Squared Error (MSE) is one of the most widely used metrics, defined as the average of the squared differences between predicted ($\hat{y}_i$) and observed (y_i) values, as shown in Equation 2.15. MSE advantage is that errors of high magnitude are more important than small errors (JAMES et al., 2013).

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2.15)$$

Another complementary metric is the *Coefficient of Determination* (R^2), which measures the proportion of variance in the dependent variable that is predictable from the independent variables. It is defined as shown in Equation 2.16. An R^2 value close to 1 indicates that the model explains most of the variability in the target data, while values near 0 suggest that the model fails to capture the underlying relationships (CHICCO; WARRENS; JURMAN, 2021).

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (2.16)$$

To assess the performance and generalization ability of a machine learning model, it is essential to evaluate it on data not used during the training phase. This procedure aims to estimate how well the model performs on unseen samples, mitigating the risk of overfitting, a phenomenon in which a model fits the training data too closely and fails to generalize to new instances (JAMES et al., 2013).

Among the various resampling-based validation methods, random subsampling (also known as Monte Carlo cross-validation) is one of the most flexible approaches. In this method, the dataset is repeatedly divided into two independent subsets: one for training and another for testing. For each iteration, a model is trained and evaluated, and the resulting performance metrics are averaged to obtain a robust estimate of predictive accuracy. Unlike other cross-validations, random subsampling does not ensure that every observation is used in both training and validation, but it allows full control over the number of iterations and the size of each split, providing an adaptable framework for model evaluation (REFAEILZADEH; TANG; LIU, 2009). Figure 22 illustrates how a random subsampling separates data:

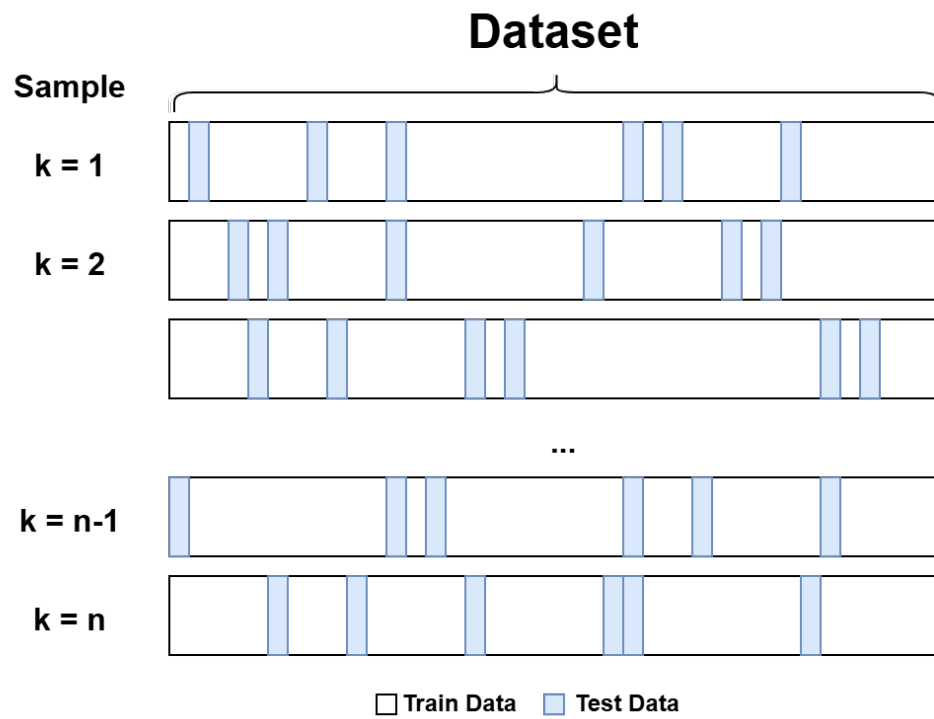


Figure 22 – Example of a random subsampling with n executions. Adapted from (BRANDENBURG, 2024)

3 RELATED WORKS

Multiple literature reviews have been conducted on the prediction of corrosion. The first notable work addressed the earliest models of the decades of 1980 and 1990, based on the three major exposure programs: ISOCORRAG, MICAT, and ICP Materials (ROBERGE; KLASSEN; HABERECHT, 2002). At that time, the approaches used included a regression decision tree called the PACER-LIME model (SUMMITT; FINK, 1980). This early version also included models in the form of equations, although more primitive than current dose-response equations, often taking the form of linear regressions.

In the decade of 2010, dose response functions appeared, such as the multi-pollutant dose response function (KUCERA et al., 2007) and the ISOCORRAG-MICAT dose response functions (MIKHAILOV; TIDBLAD; KUCERA, 2004), which were incorporated into the ISO 9223 standard in 2012, therefore becoming the standardized method for predicting r_{corr} . Subsequently, the work of (CAI et al., 2020) presented a traditional review, proposing a conceptual separation of atmospheric corrosion models for different environments into four categories: dose-response equations, multi-factor combination models, response surface models, and statistical learning models. Following that, a review focused specifically on statistical learning models was conducted (COELHO et al., 2022), which was in fact the first study to adopt a systematic approach. However, its scope included corrosion modeling in general, without a specific focus on atmospheric corrosion and its particularities.

This chapter aims to complement the work by providing a systematic literature review specifically focused on machine learning applied to atmospheric corrosion, with objective of defining methodology, models and metrics used in this work. Due to its recent publication, systematic approach, and close relation to the objective of this research, the work of (COELHO et al., 2022) served as a foundation for the methodology and processes developed in this study. The main difference of this work lies in its exclusive focus on atmospheric corrosion models, offering detailed analyses of aspects such as the atmospheric exposure datasets used for training, model error metrics, data splitting strategies, and the types of data employed.

3.0.1 Methodology

This review follows the guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA Statement). The PRISMA Statement consists of a 27-item checklist and a four-phase flow diagram designed to help authors

improve the reporting of systematic reviews and meta-analyses. Although its primary focus is on randomized studies, it can also serve as a basis for reporting systematic reviews of other types of research, especially intervention assessments (LIBERATI et al., 2009).

Initially, it is necessary to define the objectives of the work in terms of primary (PQ) and secondary (SQ) research questions. The primary research question is:

- **PQ:** Which machine learning models are being applied to predict atmospheric corrosion?

The secondary research questions are:

- **SQ01:** What datasets are being used to train the models?
- **SQ02:** What variables are being used?
- **SQ03:** What evaluation metrics are most frequently used?
- **SQ04:** Are the models being generated aligned with the standardized dose-response equations?

For the identification stage, the search phrase was adapted from (COELHO et al., 2022) ("machine learning" AND "corrosion") to fit the context of atmospheric corrosion, using "machine learning" AND "atmospheric corrosion".

For the screening and eligibility phases, inclusion criteria (IC) and exclusion criteria (EC) were defined as follows:

- **IC01:** The study must use machine learning to predict atmospheric corrosion.
- **EC01:** Articles focused on corrosion of non-metallic materials.
- **EC02:** Articles without full access or only partially available.
- **EC03:** Incomplete works.
- **EC04:** Duplicate materials.
- **EC05:** Secondary and tertiary studies.

The EC and IC criteria are applied in this order, initially based on the title, abstract, and keywords. If it is unclear whether the criteria are met, the introduction and conclusion will be consulted, respectively. Inclusion of articles by snowballing (searching references of selected studies) is outside the scope of this work.

The time period analyzed extends from the publication of the preliminary results of the ISOCORRAG project—the first multi-climate dataset developed for this topic, which led to the creation of the earliest models—until the present, i.e., from 1987 to early 2025. Within the scope of this research are journal articles, conference publications, books, and scientific reports available in academic search databases. The language of interest is English. The databases used were Scopus and Web of Science (WoS), both accessed via the CAPES platform. These two databases were selected due to their relevance, accessibility in the CAPES platform in Brazil, and compatibility with the study area. WoS has previously been used in systematic literature reviews on this topic (COELHO et al., 2022).

The data extracted from the selected studies include the models used, error metrics, datasets, data sources, input variables, machine learning models and any observable classifications related to the models. Regarding data analysis, the work of (COELHO et al., 2022) employed a quantitative analysis of model error metrics. However, this study found such an approach inadequate, as models trained on different datasets have distinct scopes of application and are therefore not directly comparable. The primary focus of this study is a quantitative analysis of the data and processes used in the development of the models, followed by a qualitative analysis of the findings.

The results will be organized in virtual spreadsheets using Google Sheets. Information will be separated into columns for models, datasets, metrics, and comparison with the standard. The data will then be analyzed in a Jupyter Notebook environment using the Python programming language. The Pandas library will be used for data manipulation, while the Matplotlib, Seaborn, and Plotly libraries will be used for data visualization. The research was conducted entirely by a single author, who developed the protocol, carried out the review, performed the analyses, and wrote the full text.

3.0.2 Protocol Application

Table 3 presents the results from applying the inclusion and exclusion criteria. The search phrase was applied to both databases, resulting in 73 hits—50 from Scopus and 23 from WoS. Most of the full-text publications found in WoS overlapped with those from Scopus, resulting in only one additional unique publication from the WoS database. After applying the inclusion and exclusion criteria (IC and EC), the resulting dataset consisted of 24 full-text studies.

	Initial	EC01	EC02	EC03	EC04	EC05	IC01	Result
Scopus	50	44	38	37	37	36	20	20
WoS (CAPES)	23	23	23	22	5	5	1	1
Total	73							21

Table 3 – Results of applying inclusion and exclusion criteria

The first observation was that in addition to the models predicting one-year corrosion rates, which were the initial focus of this study, the execution of the methodology revealed a significant number of works fitting the search phrase that instead used a regression models based of time series. Such models used a different method to measure corrosion, using sensors that measures corrosion using a small sample of metal and current, typically called Atmospheric Corrosion Monitor (ACM). Such works were included in the study and was necessary to categorize models in two separated categories:

- **Atmospheric corrosion exposure models (ACE models):** Models that use a small amount of data per station but cover a large number of stations. These models take climatic and pollutant data as inputs and return the corrosion rate over a period of metal exposure (generally one year). Typically trained on established datasets obtained from exposure stations following ISO 8565, such as the ones that cover most of the corrosion station from Figure 23. Specimens in ISO 8565 format are destroyed when cleaned, therefore such methods are not totally appropriate for monitoring corrosion continuously. These models usually aim to understand corrosion behavior over large regions, such as states, countries, or even global scale.
- **Atmospheric corrosion monitoring models (ACM models):** Models trained with large amounts of data from few or even a single location. Corrosion data are usually collected by sensors or other techniques capable of monitoring corrosion more frequently than the ISO 8565 method, typically using an ACM sensor such as shown at the top-right of the corrosion station of Figure 23. These models often take the form of time series, also using climatic and pollutant data together with the last corrosion measurement as inputs. Although they tend to show better error metrics than standard prediction models, they are limited in scope to the microclimate of the experimental location.

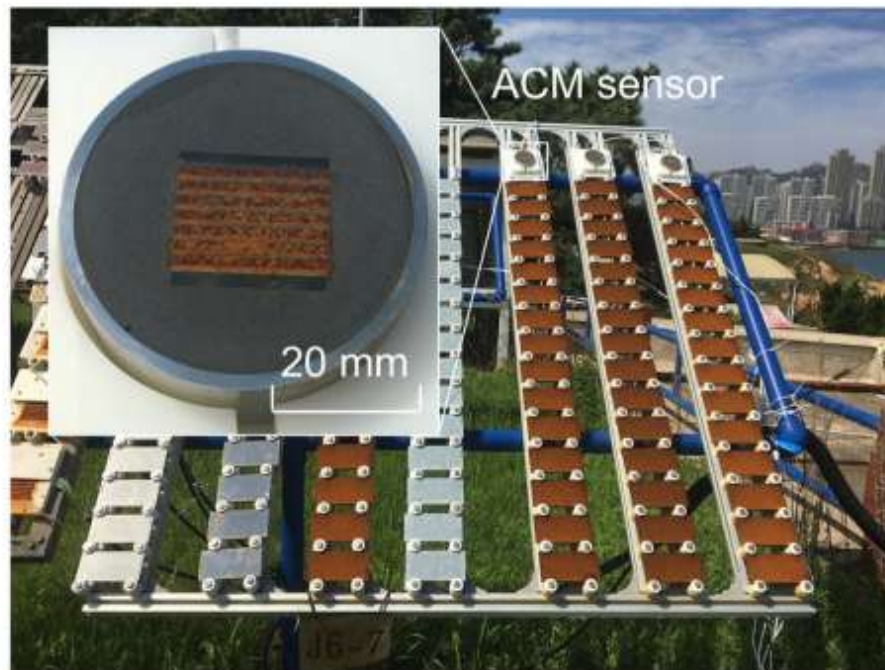


Figure 23 – Corrosion station with ISO specimens and ACM sensors. Source: (PEI et al., 2020)

Figure 24 presents the number of studies by category. Two-thirds of the publications used the traditional regression approach, while one-third used monitoring-based approaches. While the regression approach typically follows ISO 8565 with standardized exposure samples, there is no standardized process for corrosion monitoring. Nevertheless, the vast majority of monitoring publications used Atmospheric Corrosion Monitoring (ACM) sensors, which are designed to monitor real-time corrosion behavior of materials in dynamic outdoor environments (MENG et al., 2021).

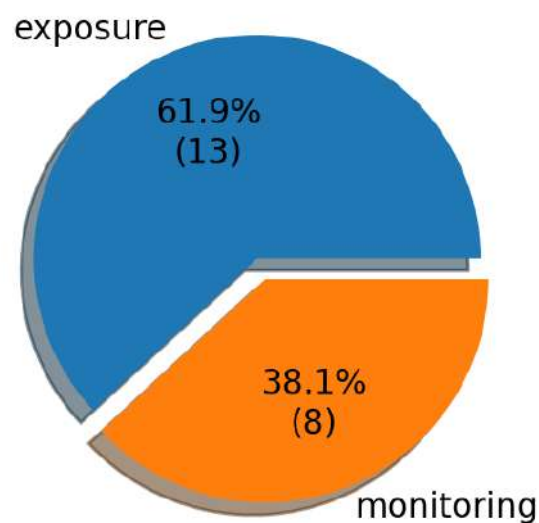


Figure 24 – Data distribution by classification

Regarding the use of ACMs, (KUCHAŘ et al., 2024) used the AirCorr ACM sensor to monitor corrosion of aluminum and stainless steel components from aircraft, both indoors and outdoors at the Kbely Aviation Museum in the Czech Republic. The study used climatic and pollutant data and compared different ML models to a Gaussian process approach, using R^2 and MAPE as evaluation metrics with an 80-20 hold-out split over one year of data collection. ECORR ACM sensors were used to monitor steel specimens exposed in Qingdao, China—an experiment that resulted in three publications found using the search query. The first study analyzed the influence of climatic and pollutant variables using multiple correlation metrics and proposed a custom ML model that fused variable relationships, comparing it to SVM and ANN models (MENG et al., 2021). The second study applied a Random Forest (RF) model to investigate whether corrosion behavior could be monitored in real time (YANG et al., 2022). The third publication compared the performance of Artificial Neural Networks (ANNs), Support Vector Machine Regressors (SVRs), and RF models on the collected data. ACMs were also used alongside saline treatments to create a corrosion dataset sensitive to salinity, using sensor-based data collection (OKURA et al., 2022).

Other monitoring-based studies included one that assessed whether accelerated corrosion test data could be used to predict atmospheric corrosion in uncontrolled outdoor environments, using time-series data from a Chinese station (VANGRUNDER-BEEK et al., 2023). Sensors were also used to build a dataset for testing various ML models to monitor corrosion in high-temperature atmospheric environments, such as nuclear, solar, and geothermal power plants (MUTHUKRISHNAN et al., 2024). Another study used data from metal specimens mounted on a test vehicle over a period of three years and developed a prediction model using SVR. It also compared the performance of different metaheuristics for hyperparameter optimization—this was the only monitoring-focused study found that used test specimens instead of sensors (SONG et al., 2024).

Machine Learning was also applied to evaluate the impact of the variables collected using an ACM sensor in the corrosion prediction. One study measured the impact of the variables in the prediction using SHapley Additive exPlanations (SHAP) (WANG et al., 2025). Another study used data from ACM sensors combined with clustering to classify atmospheric corrosivity (YANG et al., 2025).

Among the prediction-focused studies, most used proprietary datasets for model training, as shown in Figure 26. Among the more established datasets, such as ISO-CORRAG and the MICAT project, ISOCORRAG data were used to train a machine learning model divided into two stages: data preprocessing using the Multivariate Adaptive Regression Splines (MARS) model and prediction using an ANN model known as Self Organizing Maps (SOM) (TERRADOS-CRISTOS et al., 2021). Another study

conducted a direct comparison with the standard: it used aluminum data from three exposure programs—MICAT, ISOCORRAG, and the Russian/Soviet Corrosion Program—and trained a Random Forest model whose performance was compared with dose-response functions (GAVRYUSHINA; MARSHAKOV; PANCHENKO, 2023). Other interesting approach was the use of chloride expositions from the MICAT dataset to predict not the corrosion rate, but the classification of airborne salinity (TERRADOS-CRISTOS et al., 2025), a similar approach to this work.

A frequently used dataset is Japan's National Institute for Materials Science (NIMS) Materials Database (MatNavi). Data from three Japanese stations were used to train ML models, perform variable selection, conduct data mining using SHapley Additive Explanations (SHAP) analysis, and compare ensemble tree models, SVRs, and linear regression (YAN et al., 2020). Another publication using the same dataset evaluated the performance of the RF algorithm (YAN; DIAO; GAO, 2020). Yet another study used the same dataset to compare RF, SVR, and linear regression models (THANUSH et al., 2022).

A significant portion of the prediction models were based on research using Chinese data. One dataset, from the China Gateway to Corrosion and Protection, contains data from 1, 2, 4, 8, and 16 years of exposure across six corrosion stations in China, totaling 409 data points. A model integrating KNN and RFS, called Random Forest Weighted K Nearest Neighbors, was developed and its performance compared to ANN, SVRs, and cForest algorithms using this dataset (ZHI; YANG; FU, 2020). Another study used the same dataset to create an ensemble model combining a neural network and the output of RF, XGBoost, LightGBM, and a Gradient Boosted Decision Tree (GBDT) regressor (CHEN et al., 2024). Yet another study using the same dataset compared the performance of SVM, RF, ANN, and linear regression (ZHI et al., 2019).

Other publications using Chinese data include one in which exposure data was used to train a two-stage model—first reducing dimensionality using RF and Spearman correlation, followed by training an SVR model (ZHI et al., 2021). RF, SVR, and linear regression models were also used to generate a corrosion map of the Hunan region in China, using one-year exposure data from 50 stations across the region (ZHANG et al., 2024).

Within the international scope, one dataset composed of Chinese, Indian, and Iranian data was used to train and compare ANN, LASSO, RF, SVR, GBDT, and XGBoost models (KUANG; LONG, 2024). Vietnamese exposure data were used to train an ANN and compare its performance with dose-response functions fitted to different datasets (TRAN et al., 2021). Another study used data from 12 corrosion stations in the metropolitan area of Mexico City, comparing ANN, regression trees, tree ensembles, SVMs, and linear regression models (RODRÍGUEZ, 2024).

Regarding the publication dates, Figure 25 shows a histogram of the studies by year. It is evident that these publications are recent: the first paper dates from 2019, and the publication rate remained stable from 2020 to 2023. The research year 2024 had the highest number of publications, with seven. Therefore, it can be concluded that the use of the term "machine learning" in the context of atmospheric corrosion is a recent trend. It is possible that earlier studies were conducted under different terminology before "machine learning" became widely adopted.

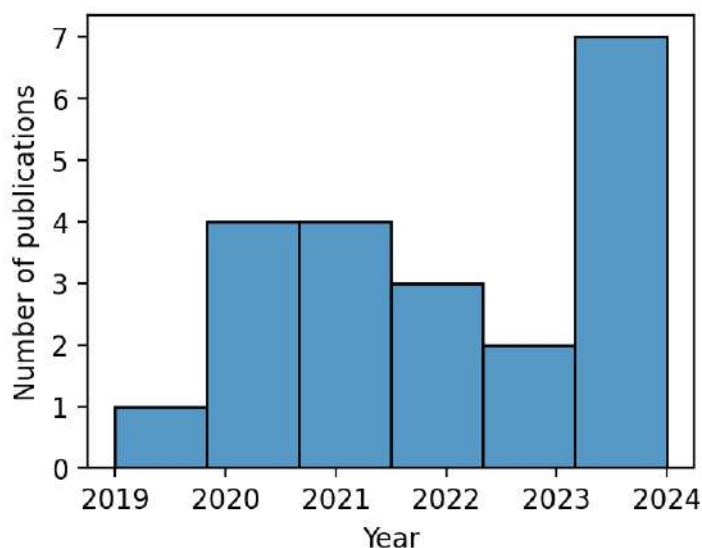


Figure 25 – Number of publications per year

Regarding the datasets used in each publication, Figure 26 shows the distribution across sources. Most studies used proprietary data for model training, which is expected since monitoring models typically rely on custom datasets—although this is also common in standard prediction models. The National Environmental Corrosion Platform of China dataset was used in four publications, followed by the China Gateway to Corrosion and Protection, highlighting the significance of Chinese research in this field. The Japanese MatNavi dataset was used in three studies. Although MatNavi is an international database, only Japanese stations were used in the referenced works. Among the more established datasets, ISOCORRAG appeared in only two studies—one in combination with MICAT and the Russian (RUS) exposure program.

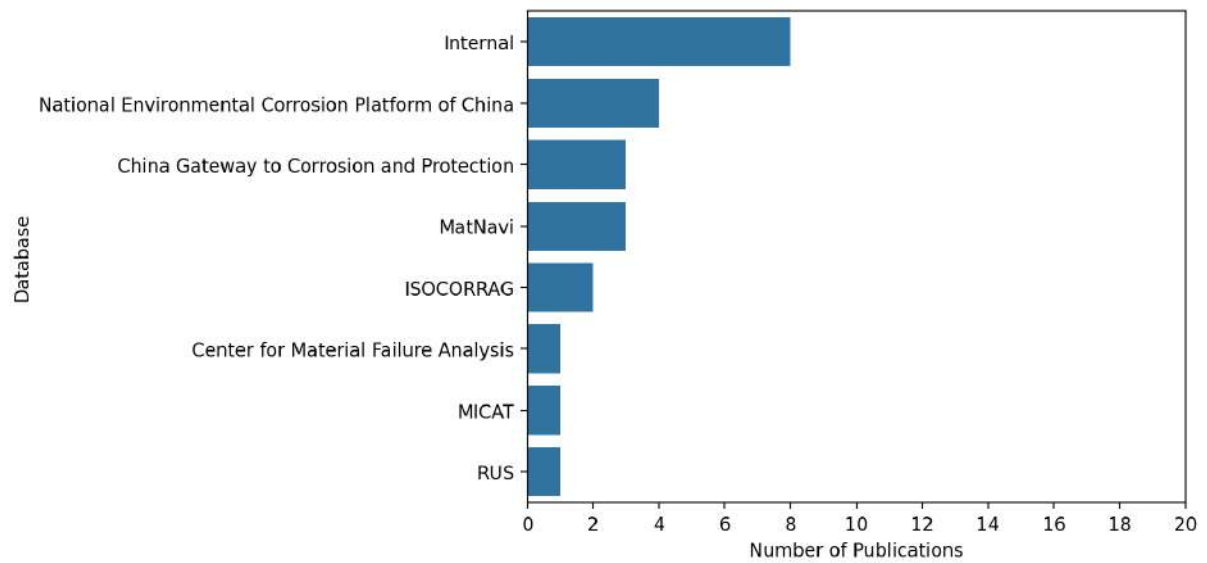


Figure 26 – Datasets used by the publications

Given the relevance of Chinese publications within the datasets, Figure 27 also presents the origin countries of the datasets. It is notable that Chinese data accounts for half of the datasets used in the reviewed publications, while international datasets corresponded to only two of the identified sources. Most datasets also have Asian origins, as Japan appears with two datasets and Vietnam with one. This reveals a gap in comparing ML models based on the well-established MICAT, ISOCORRAG, and ICP Materials datasets, highlighting an avenue yet to be explored in the use of ML models.

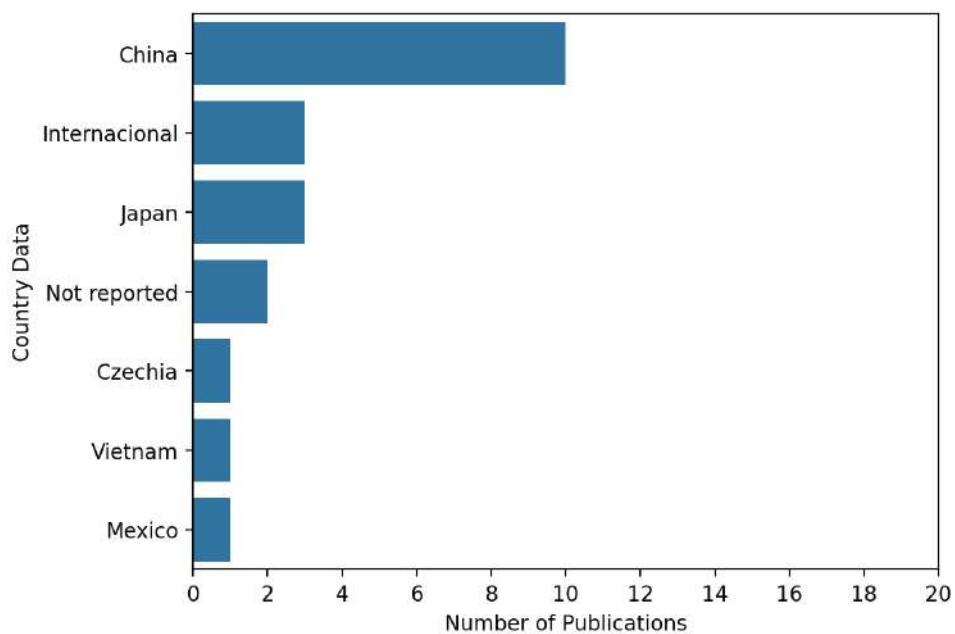


Figure 27 – Country of origin of the data used

Regarding the variables, Figure 28 shows those used in the models. Only Temperature (T) and Relative Humidity (RH) were present in all studies, likely due to their

high importance, being standardized variables for atmospheric corrosion and easy to collect. The most used pollutant was sulfur dioxide (SO_2), another standardized variable, which explains its frequent presence in the literature. Among the four standardized variables for dose-response functions, dry deposition of chloride ions (Cl^-) appeared in just over half of the publications. Although theoretically mandatory, in practice chloride contamination usually only occurs in marine environments, so this variable is often excluded in continental environments. Composition is also a common variable, representing the inclusion of the alloy's chemical composition as a model input. Among other climatic variables, noteworthy ones include Precipitation (Pr), Rain pH (pH), Time of Wetness (TOW), Solar Radiation (Solar), Hours of Sunshine (HoS), and Wind Speed (Wind), all used in multiple studies. For pollutants, Nitrogen Dioxide (NO_2), Ozone (O_3), Nitric Oxide (NO), and Carbon Monoxide (CO) appeared in several publications, as did air quality metrics such as PM 2.5 and PM 10 particulates.

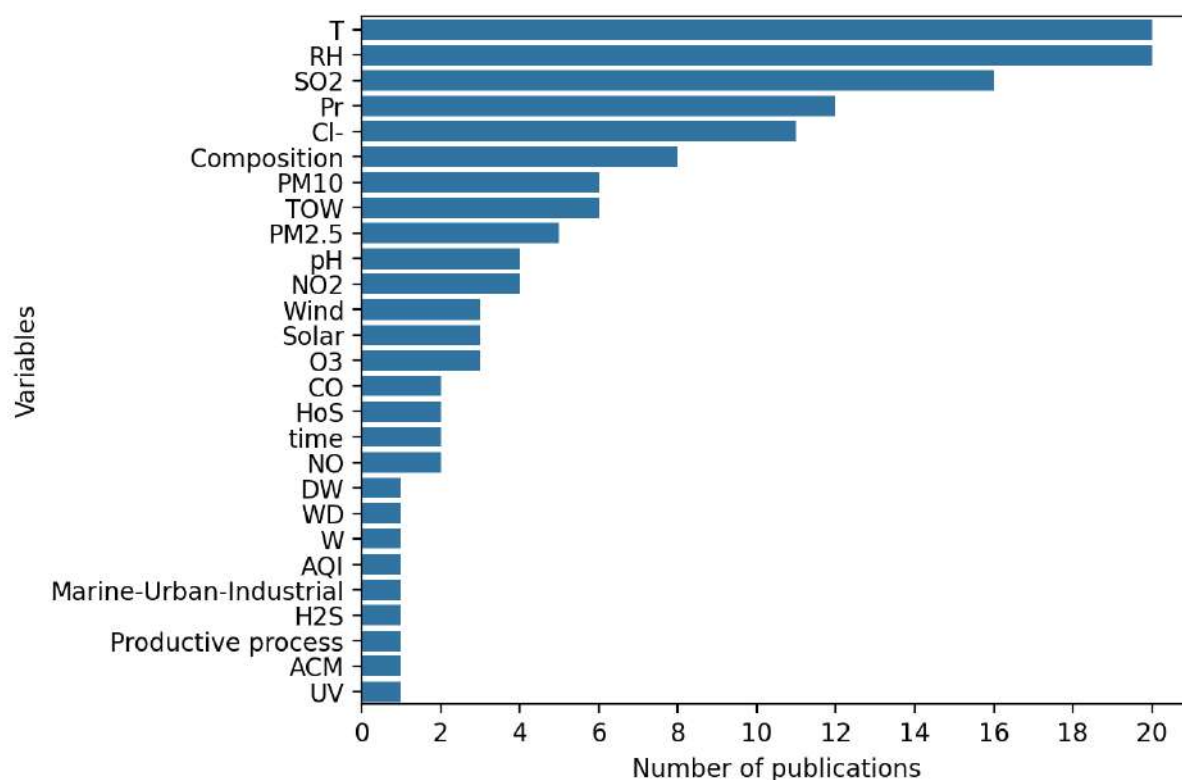


Figure 28 – Variables used by the models

Regarding whether correlation analysis was performed, Figure 29 displays the types of correlations used in the studies. Most studies did not present correlation analysis or variable selection. When used, Pearson and Spearman correlations were the most common.

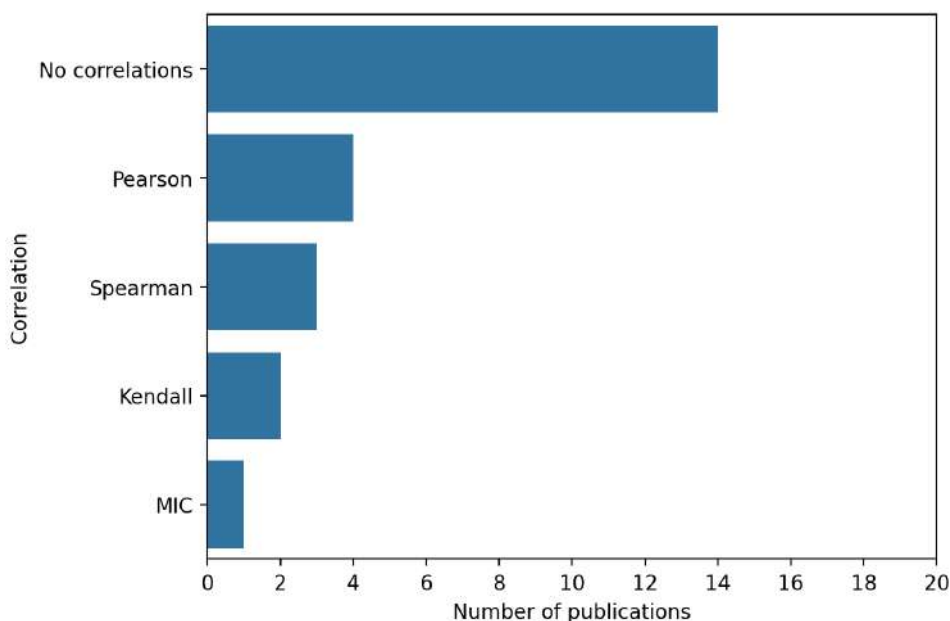


Figure 29 – Variable correlation analysis

As for the models used, Figure 30 shows the machine learning models applied in the studies. Support Vector Machine Regressor (SVR) and Random Forest (RF) were the most commonly used, appearing in the majority of the studies. Overall, tree-based and boosting ensemble models such as RF, GBRT, XGBoost, Light Gradient Boosting (LightGBM), and Adaptive Boosting (AdaBoost) were widely applied. Artificial Neural Networks (ANNs) appeared in half of the publications. The preference for RF and SVR over ANN in predictive tasks is largely due to the difficulty in obtaining datasets large enough to train a neural network—more versatile models such as tree ensembles tend to be more attractive options. Linear models also appeared, with Linear Regression (LinReg), Lasso, and Ridge Regression used in multiple publications. Linear regressions are not well suited for time series or multivariate regression in corrosion studies, so they should always be compared to other approaches to ensure modeling performance.

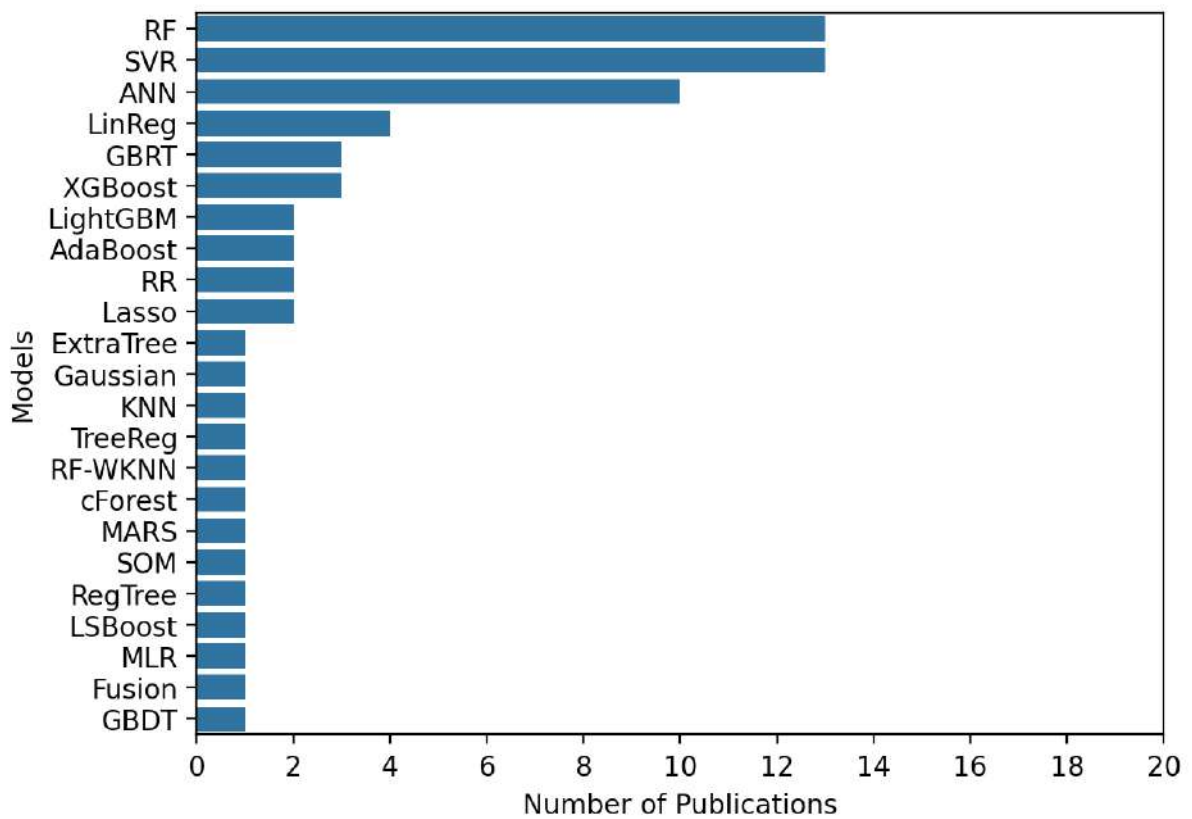


Figure 30 – ML models used in the publications

Regarding metrics, Figure 31 displays the error metrics used in the reviewed works. The coefficient of determination (R^2) is the most used, appearing in most publications. This is likely due to its intuitive interpretation, widespread use in engineering, and the fact that it is the metric used in atmospheric corrosion standards. Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE) were also widely applied. MAPE is ideal for datasets with a large range of values, while RMSE gives greater importance to larger errors. Other metrics such as Mean Absolute Error (MAE) and Accuracy, Percentage Error (PE), Absolute Error (AE), Relative Error (RE), Mean Squared Error (MSE), and Symmetric Mean Absolute Percentage Error (SMAPE) were only used occasionally.

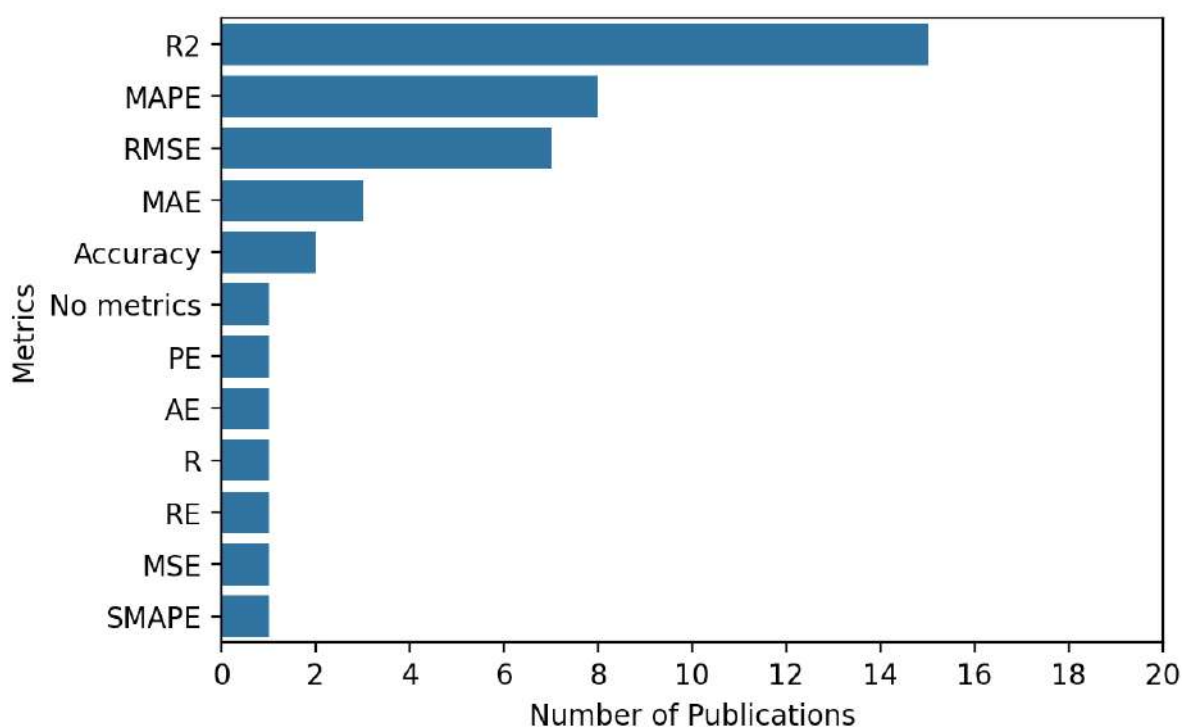


Figure 31 – Error metrics used in the publications

A summary of the found works are shown in Table 4. Most of the selected publications aligned with the initial focus on *rcorr* prediction (ACE models), but the study also revealed the rise of a new approach based on corrosion monitoring through sensors (ACM models). For the datasets, half of the selected studies used Chinese datasets instead of well-established datasets such as MICAT and ISOCORRAG. As for the variables, temperature and relative humidity were used in all applications. The most commonly used correlations coefficients were Spearman and Pearson. Regarding the models, Support Vector Machines and Random Forests were the most frequently used, followed by Artificial Neural Networks and other tree-based ensemble regression models. The most common metric was the coefficient of determination (R^2).

Characteristics	Most popular approaches
Type of prediction	Atmospheric corrosion exposure (ACE) prediction, atmospheric corrosion monitoring (ACM) prediction
Machine Learning Model	Random Forest and other tree ensemble algorithms, Support Vector Regressors, Artificial Neural Networks
Dataset	Chinese National Datasets, MatNavi, Russian Exposure data, MICAT and ISOCORRAG.
Variables	Temperature, Relative Humidity, Precipitation, SO ₂ , Cl-
Metrics	Coefficient of Determination (R^2), Rooted Mean Squared Error (RMSE), Absolute Percentual Error (MAPE),

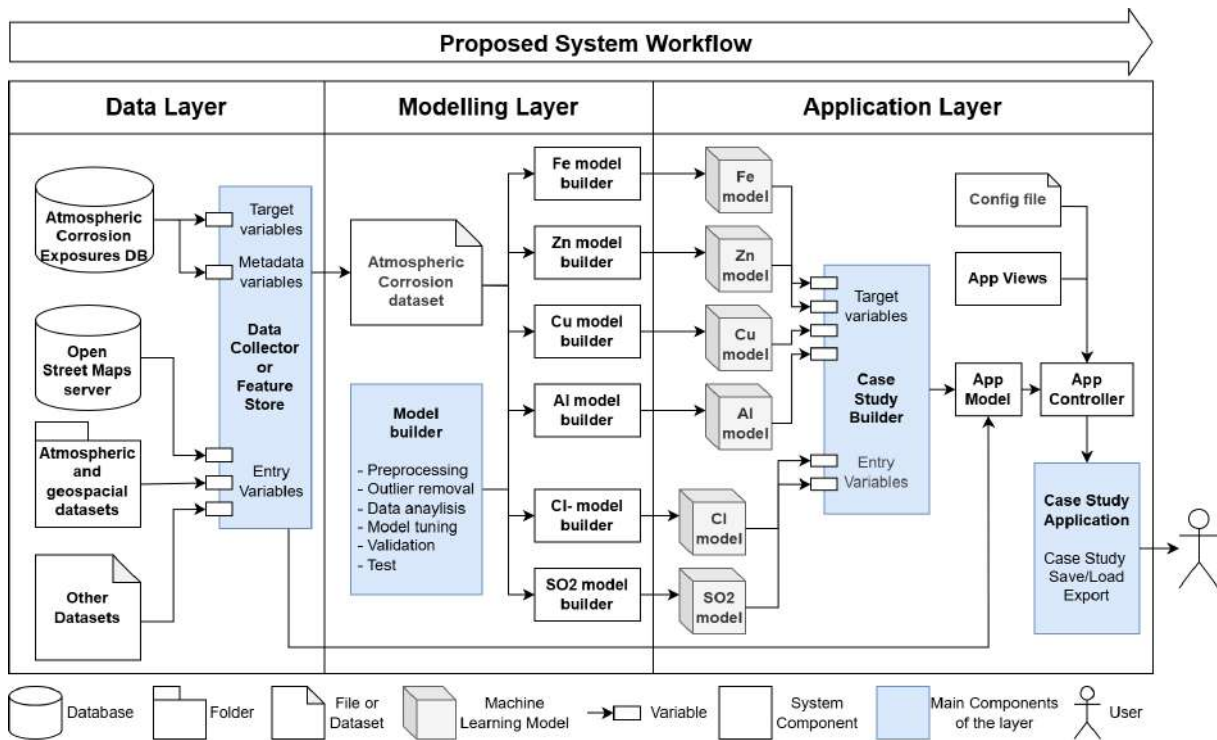
Table 4 – Summary of literature review results

The proposed approach of this work relies on data from exposure projects; therefore, the developed system will adopt an ACE (Atmospheric Corrosion Estimation) model. Data will be sourced from exposure projects such as MICAT and ISOCORRAG, while Chinese datasets remain outside the scope of this study. For model development, a comparison between Support Vector Regressors (SVRs), Random Forests (RFs), and Artificial Neural Networks (ANNs) will be conducted, using the R^2 metric as the primary performance indicator. Common meteorological variables, such as temperature, relative humidity, and precipitation, can be obtained from reanalysis datasets like ERA5. However, SO_2 and Cl^- concentrations are not directly available, and therefore, intermediate pollutant models will be employed to estimate these input variables. The next chapter provides a detailed explanation of the proposed methodology.

4 PROPOSED APPROACH

To address two of the specific objectives of this work — the improvement of predictive models and the execution of case studies in Brazilian environments — this study proposes the development of a data-driven atmospheric corrosion prediction system, referred to as LABICORR. The system encompasses the entire workflow, from the collection of atmospheric, geospatial, and corrosion data and their organization within a structured data storage framework, to the prediction of metal corrosivity based on user input through an interactive interface.

Figure 32 illustrates the workflow of the proposed system, which is divided into three layers, represented by lanes: the Data Layer, the Modeling Layer, and the Application Layer. Each layer is briefly introduced in this section and further detailed in the corresponding sections that follow.



The Data Layer centralizes the collection and data engineering of all project variables through the Data Collector component. This layer outputs all variables in the form of unified datasets, which are used for training models or for case studies.

The Modeling Layer is where the machine learning models are trained using the datasets provided by the Data Layer. There are six models in the project: four target r_{corr} models for Zn , Fe , Al , and Cu , and two intermediary pollutant deposition models

for Cl^- and SO_2 . Each model is created through a builder that follows the typical steps in the development of a statistical model: preprocessing data, removing outliers, data analysis, model tuning, model comparison and validation.

The Application Layer integrates the models and components of the project into an application that allows users to run case studies for selected sites. The application follows a model–view–controller architecture, where the components of the previous layers are grouped within the model part.

The system was developed using the Python 3.10.12 programming language. The following packages and libraries were used to develop the system, along with their respective functions:

- **Xarray**: Manipulation of atmospheric and geospatial NetCDF4 datasets.
- **Pandas**: Manipulation of CSV, SQL and Excel datasets, mainly of atmospheric corrosion data.
- **Matplotlib**: Creation of statistical plots.
- **Seaborn**: Creation of statistical plots.
- **Plotly**: Creation of interactive plots, mainly for identifying outliers.
- **Cartopy**: Map plotting.
- **Scikit-learn**: Implementation of machine learning models.
- **PySR**: Implementation of the PySR genetic programming algorithm.
- **SciPy**: Implementation of the least-squares method to compare analytical functions to machine learning models.
- **Streamlit**: Front-end framework for the application.
- **Jupyter Notebook**: Interactive environment used to analyze and train the models.
- **PyInstaller**: Used to create an executable application for Windows and Linux.

The user application was developed for Windows and Linux operating systems, requiring a minimum of 10 GB of storage and internet access. The development version of the system was hosted on a server, whose specifications are described below:

- **Workstation**: Dell PowerEdge R650

- **Operating System:** Ubuntu 22.04.5 LTS x86_64
- **CPU:** Intel(R) Xeon(R) Gold 6330N, 3.4 GHz, 56 cores
- **Main Memory:** 64 GB RAM
- **Secondary Storage:** 5 TB SSD

4.1 DATA LAYER

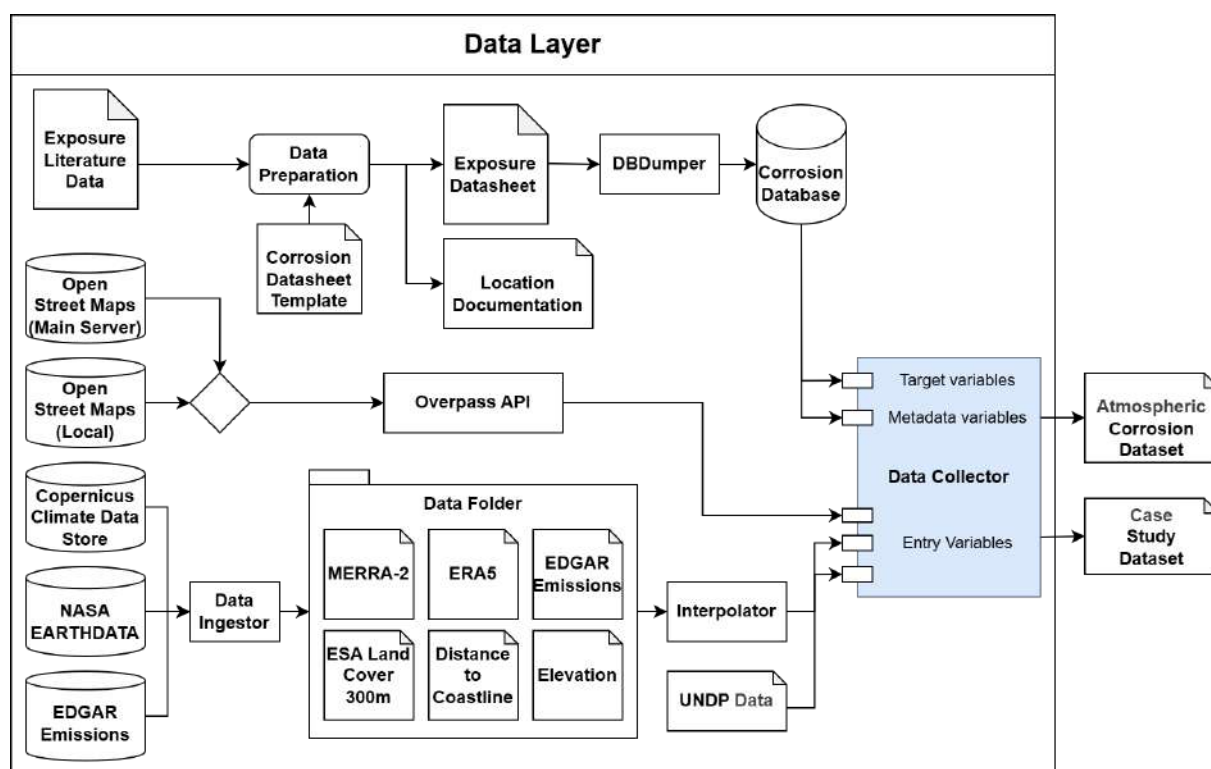


Figure 33 – Extended representation of the data layer

The data-oriented nature of atmospheric corrosion prediction demands that the Data Layer be a robust architecture for data ingestion, organization, and engineering. The main function of the Data Layer is to create on-demand datasets containing all variables needed to train and run the models in the system. For that, the Data Collector component was created: it uses coordinates and exposure periods to retrieve variables from datasets and databases located on the server or workstation running the system. Figure 33 shows an extended version of the Data Layer, with all components connected to the Data Collector.

The data structure manipulated by the Data Layer is a two-dimensional dataset. Each row of this dataset represents a group of experiments called exposures, and each column represents a variable. An exposure has at most one measurement for each

ISO 9223 metal and/or SO_2 and Cl^- deposition, at the same place for the same period. Variables are collected from the site of exposure or remotely from atmospheric and geospatial datasets. Variables are divided into three categories:

- **Metadata Variables:** Information about the exposure other than experimental results, such as station name, publication, location, exposure start date, etc.
- **Input Variables:** Independent variables, i.e., model inputs that are relevant for prediction, typically atmospheric or geospatial variables. Examples include temperature, relative humidity, precipitation (from site or ERA5), surface mass concentrations of SO_2 and Cl^- (from MERRA-2), elevation, distance to coastline, and SO_2 and Cl^- deposition (for r_{corr} models), among others.
- **Target Variables:** Dependent variables, i.e., model outputs. There are six output variables: SO_2 , Cl^- , Zn , Fe , Al , and Cu . SO_2 and Cl^- act as both input and output variables, depending on whether we are referring to a metal model or an intermediary pollutant model, respectively.

Before generating the final version of the datasets, metadata variables and output variables from corrosion data are stored in one component: the Corrosion Database (Corrosion DB). The Corrosion DB is a relational Structured Query Language (SQL) database structured as in the Entity-Relationship diagram of Figure 34:

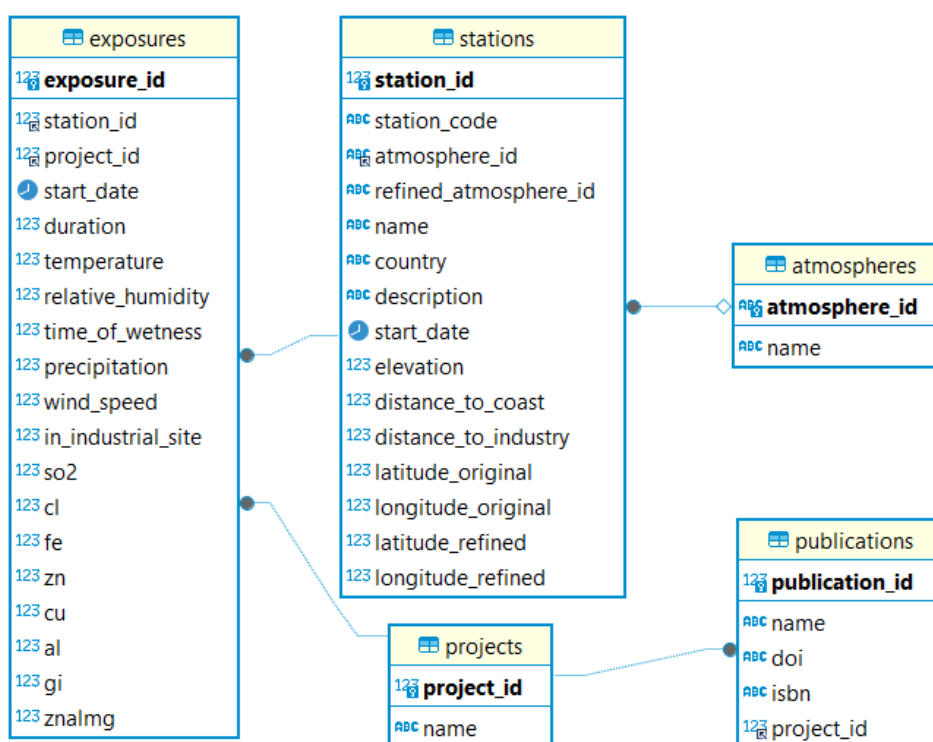


Figure 34 – Entity-Relationship diagram of the Corrosion DB

Exposures are a set of observations for the same station and the same period, contains output variables and some input variables collected on site. Stations is where exposures take place, each exposure is related to only one station and aren't directly related to projects, since stations can be permanent sites used for multiple exposure projects. A project is set of exposures typically associated with an exposure program, usually published in one or more publications. Publications are data sources, including articles, books, theses, technical reports, and other academic works. Atmospheres are used for classifying pollution types, generally including 'coastal,' 'rural,' 'urban,' and 'industrial' categories.

Data in the Corrosion DB comes from ISOCORRAG (KNOTKOVA; KREISLOVA; DEAN, 2010), MICAT (MORCILLO et al., 1998), Asian exposures (COLE et al., 2001), Corrosivity Map of Chile (VERA et al., 2012) and Canary Islands (MORALES et al., 2005) and exposures in Brazil (LUNA et al., 2008; SICA, 2006; PORTELLA et al., 2012; BEZERRA-PONTES et al., 2011; VITALI, 2013), with continuous expansions to other works. Since most data come from publications, two steps are necessary to convert exposure data from publications into the Corrosion DB: the data preparation that converts exposure data in the corrosion template format and the dumping of the dataset in the Corrosion DB using the DB Dumper module, as shown in Figure 33

The Data Preparation step is to collect the data from an exposure project and organize it into a dataset with a structure similar to that of the Corrosion DB. The first part of data preparation is to locate and document the exact positions of corrosion stations that participate in a project, process called coordinate refinement. This is necessary because many atmospheric corrosion studies are decades old and were not designed for integration with external variables from modern climate datasets. Therefore, several experimental locations were imprecise: the MICAT project provides coordinates only in degrees and minutes (with an inaccuracy of about 1 km), ISOCORRAG does not provide coordinates at all—only country and station name—and similar issues occur in other publications.

Table 5 shows the template for the coordinate refinement. The original coordinate is recorded, and a refined coordinate is proposed based on a documentation process using descriptions, names, photos, maps, related publications, and author testimonials. Additionally, the exposure period and distance to coastline are also noted in this step. This process is mandatory for all stations, since a distortion of even a few hundred meters can cause large errors in the collection of fine-resolution variables. The location documentation procedure was previously presented in (BRANDENBURG, 2024) and is a continuous process that iteratively improves the Corrosion DB metadata, the precision of input variables, and the quality of the models.

Station	Description	Given Co-ord.	Refined Coord.	Images
Camet	Exact location found. Identified by Gmaps.	-37.833333, -57.583333	-37.889949, -57.521831	
Vila Martelli	Exact location found. Located on CITEFA, According to description.	-34.583333, -58.483333	-34.557891, -58.506272	
Iguazu	Moved to Iguazu Airport, according to description	-25.733333, -54.400000	-25.731303, -54.477952	
San Juan	Allocated at the Félix Aguiar Observatory, according to description.	-31.600000, -68.550000	-31.509417, -68.623889	
La Plata	Exact location found. Located on CIDEPINT 8 km of given coord.	-34.833333, -57.883333	-34.902686, -57.935166	
Jubany	Moved south, 100m from coast. Visually confirmed.	-62.233333, -58.666667	-62.238576, -58.667600	

Table 5 – Coordination refinement of Argentinian stations of the MICAT project. Updated from (BRANDENBURG, 2024)

After documenting the stations, an atmospheric corrosion dataset is created for each project and saved in a spreadsheet file with the same structure as the Corrosion DB. The datasets are stored in an external drive, and are inserted in Corrosion DB using the DB Dumper component. The Dumper manages the index of exposures and ensures no duplicated datasets are inserted in the DB. The final result of the Corrosion DB is shown in Table 6, where the number of stations and exposures per project are shown:

Project	N. Stations	SO ₂	Cl-	Fe	Zn	Al	Cu
MICAT	79	184	181	196	176	171	191
Canary Islands	74	74	74	73	74		73
ISOCORRAG	52	243	194	287	298	293	294
Chile	30	30	30	30		30	30
Australia	28		28				
CSIRO Asia	18	12	12	12	12		
Brazil - Ceará	19	72	72	6			5
Brazil - Salvador	7	7	7	7	7	7	7
Brazil - S. Francisco do Sul	6		6				
Brazil - Recife	5		5				
Brazil - Sergipe	2			6	6		
Total	320	622	609	617	573	501	600

Table 6 – Data Sources

While exposures with metadata and output variables are stored in the Corrosion DB, input variables are provided by other sources. Each variable has a different source and pre-processing method to make it compatible with the Data Collector. The data sources must provide both spatial and temporal coverage for exposure projects and contemporary case studies, meaning they require global spatial coverage and temporal continuity from the 1980s to the present. Table 7 lists all input variables implemented in the system, their respective sources and target models:

Variable	Unit	Source	SO ₂	Cl-	Zn	Fe	Al	Cu
Distance from the sea	km	Map Ruler Tool		X				
Elevation	m	STRM30+		X	X	X	X	X
Trajectory roughness	m	ESA CCI LC, dist2coast		X				
Weighted Urbanization	No Unit	ESA CCI LC	X					
Weighted Industrialization	No Unit	OpenStreetMap	X					
SO ₂ Emissions	mg/m ² d	EDGAR	X					
Wind speed	m/s	ERA5		X	X	X	X	X
Temperature	°C	ERA5			X	X	X	X
Relative humidity	%	ERA5			X	X	X	X
Precipitation	mm	ERA5			X	X	X	X
Sea salt mass concentration	kg/m ³	MERRA-2		X				
Sulfur Dioxide mass concentration	kg/m ³	MERRA-2	X					
Cl ⁻ dry deposition	mg/m ² d	Corrosion DB			X	X	X	X
SO ₂ dry deposition	mg/m ² d	Corrosion DB			X	X	X	X

Table 7 – Variables of the system. Each model variables are marked with an "X". Source: Author

Datasets such as dis2coast, STRM30+ DEM, EDGAR, MERRA-2, and ERA5 are global 2D NetCDF grid maps, with latitude as the X-axis and longitude as the Y-axis, at a specific lat × lon resolution. To obtain values for a specific location, it is necessary to interpolate the discrete grid cells to produce continuous spatial data. The Interpolator module performs this process, offering two methods: *closest*, which assigns the value of the nearest grid cell, and *bilinear*, which computes a weighted average of the four surrounding cells.

Distance to the coastline is the main variable for the *Cl*– deposition model. Chlorides are emitted from the sea and have maximum concentrations near the coast, decreasing exponentially as distance increases, as proposed in salinity models (FELIU; MORCILLO; CHICO, 1999). This variable is collected manually using satellite imagery and ruler tools.

Elevation affects several environmental factors, such as salt pollutants, air pressure, and humidity. Therefore, elevation is included in the *Cl*– and metal models. Elevation data are obtained from the STRM30+ Digital Elevation Model (DEM), with a resolution of approximately 800 meters, interpolated bilinearly.

Trajectory roughness is an engineered variable that quantifies the resistance imposed by terrain to wind flow (surface roughness). In areas with large surface roughness, such as forests and cities, the decay of salt aerosol concentration with distance is faster; therefore, it is included in the *Cl*– model. A greedy search algorithm is applied to dis2coast to find the nearest coastal coordinate with an inaccuracy of 1 km, then refined using land-cover data to locate the nearest water body (reducing inaccuracy to 150 m). The Bresenham line-drawing algorithm is used to draw a straight line of land-cover tiles between the coastal point and the target location. The WindPro conversion scheme (Table 2) converts land-cover values to roughness values. The resulting roughness path is averaged to obtain the trajectory roughness.

Weighted urbanization and industrialization are engineered variables proposed by this work to quantify SO_2 emissions within a microclimate pondered by its distance. SO_2 emissions mainly result from fossil fuel combustion by vehicles (urban) and factories (industrial). Urban areas are obtained from Land Cover cells. Industrial areas are obtained from OpenStreetMap (OSM) polygons. Equation 4.1 shows the formula used:

$$\text{Weighted Area} = \sum_{\substack{k=1 \\ \text{distance}_k < R}}^n \frac{\text{area}_k}{\max(\text{distance}_k, 50 \text{ m})} \quad (4.1)$$

Were R is the radius of observation and n the total number of urban cells or industrial polygons within the radius. For a k cell or polygon, area_k is the total area and distance_k is distance to center to the location analyzed (center of the circle). The

minimum distance of 50 meters is applied to avoid division by zero.

SO_2 emissions are provided by EDGAR emission maps. Although emissions and deposition of SO_2 do not necessarily occur in the same location, areas with high emissions tend to have higher depositions. EDGAR has a spatial resolution of $10 \text{ km} \times 10 \text{ km}$, which is coarser than the SO_2 mass concentration data from MERRA-2. SO_2 emissions are collected using closest interpolation since they are spatially localized and smoothing interpolations such as bilinear tend to distort their distribution.

Meteorological variables such as wind speed, temperature, relative humidity, and precipitation are obtained from the ERA5 reanalysis, at an approximate resolution of $31 \text{ km} \times 31 \text{ km}$. These variables influence the formation of the thin condensed water film on metal surfaces, which acts as an electrolyte. Thus, meteorological variables are part of the metal models. Wind speed also influences the emission and transport of sea salt aerosols, hence it is a variable of the $Cl-$ model. Relative humidity (RH) is not directly available in ERA5 and is calculated using Equation 4.2 (WMO, 2024). ERA5 variables are interpolated bilinearly, as smoothing tends to increase precision for these variables.

$$RH = 100 \cdot \frac{e_s(T_d)}{e_s(T)} \quad (4.2)$$

$$e_s(T) = 6.112 \exp\left(\frac{17.67 T}{T + 243.5}\right)$$

Where T is temperature, T_d is dew-point temperature. $e_s(T)$ and $e_s(T_d)$ are the saturation for water vapor pressure for temperature and dew-point temperature, respectively.

Pollutant variables of Atmospheric Surface Mass Concentration for both SO_2 (SO2MASS) and $Cl-$ (SSSMASS) are provided by the MERRA-2 reanalysis. Although MERRA-2 also provides dry deposition data for both pollutants, these are separated into five particle-size bins (up to $10 \mu\text{m}$), while SO2MASS and SSSMASS represent total concentrations, which show stronger correlations with peroxide cylinder and wet candle measurements. SO2MASS is a variable of the SO_2 model and is obtained using closest interpolation, as smoothing interpolations generate artificial values for small-area emitters such as cities and factories. SSSMASS is a variable of the $Cl-$ model and is interpolated bilinearly, as smoothing reduces abrupt concentration transitions between continental and coastal grid cells.

4.2 MODELING LAYER

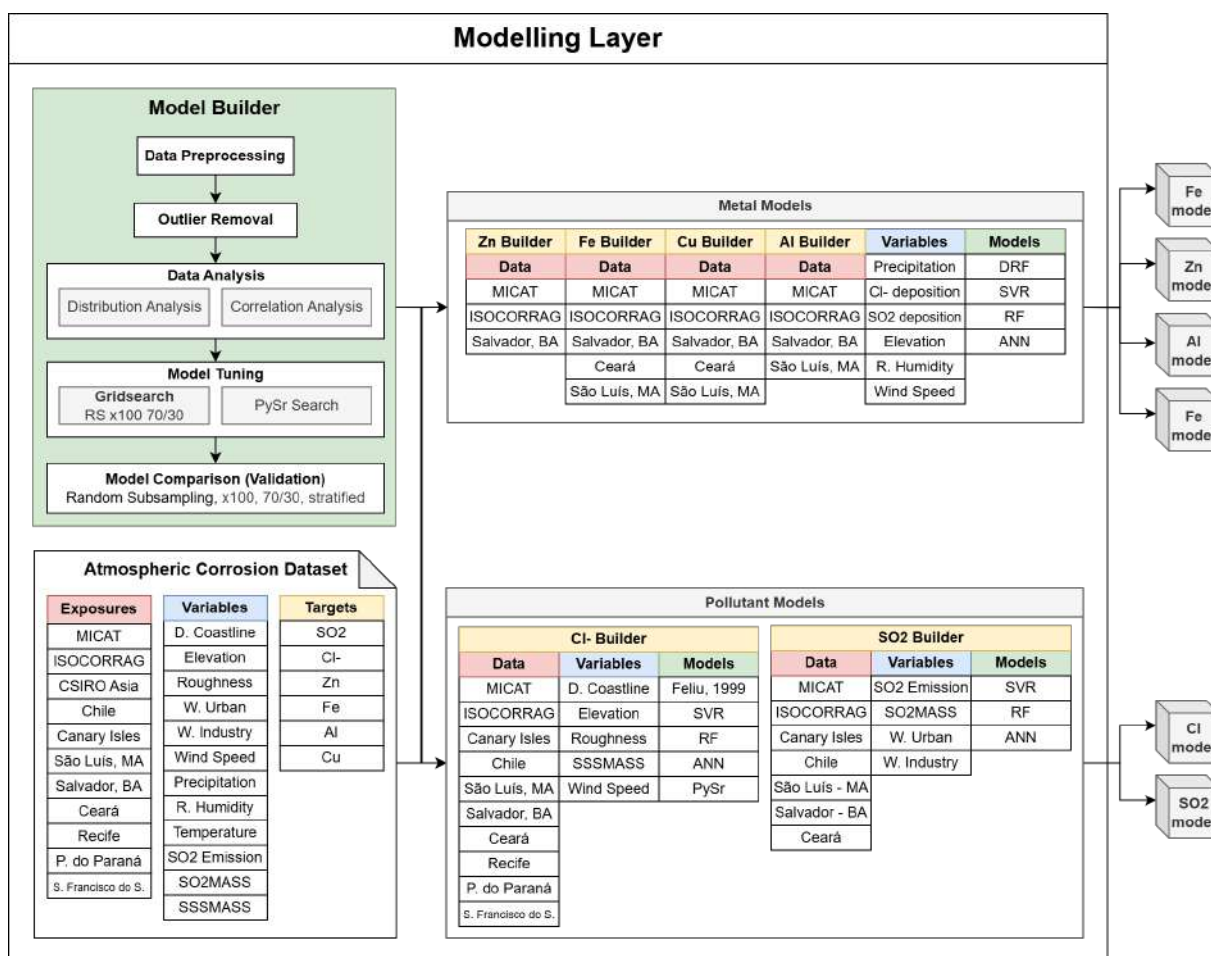


Figure 35 – Extended diagram of the modeling layer. Source: Author

Once the atmospheric corrosion dataset has been constructed in the data layer, it is necessary to train, validate, and compare corrosion and pollutant models, for which the modeling layer was developed. Figure 35 shows an extended diagram of the components in this layer. The first component is the atmospheric corrosion dataset, containing data from five international exposure projects and seven Brazilian research projects (Table 6), twelve input variables from multiple sources, and six target variables (Table 7). A specific model is chosen for each target and then forwarded to the output layer.

The second component is the Model Builder, an interface that defines the steps each of the six target models must follow to prepare data, analyze, train, compare, and select models. Each model has a separate instance of the Model Builder, which manages which projects, variables, and algorithms will be used for each target. The *Cl*⁻ model primarily works with terrain-, salt-, and wind-related variables, with data available in almost all data sources. The *SO*₂ model focuses on urban and industrial variables, such as area, emissions, and depositions. Metal models mainly rely on meteorological

variables, and pollutant data are limited to SO_2 and Cl^- deposition; therefore, other variables are indirectly available through these two pollutant models. Metal models have an almost identical structure, with only minor differences in training data due to availability.

The first step in building a model is preprocessing, which consists of performing the necessary operations to convert the corrosion dataset into training datasets. This includes variable conversions, removal of entries with missing variables, creation of stratification values based on levels of corrosivity ($C1$ to CX) or pollution levels (P_0 to P_3 for SO_2 , S_0 to S_3), and filling gaps in the Zn model with zinc-galvanized steel (GI). It is worth noting that the datasets created for metal models only include entries with on-site Cl^- and SO_2 deposition. Although filling the gaps using the SO_2 and Cl^- models is possible, it would propagate the errors of the pollutant models to the metal models. Therefore, all models will be used only in the output layer.

The second step in building a model consists of outlier removal. It involves selecting data to be removed from the training dataset. Data can be removed either by exposure or by station. The first criterion is that data points outside the interval for corrosivity and pollution levels defined by ISO 9223 are removed. Since the station location is a documented process, stations without precise coordinates are susceptible to collection errors. When multiple measurements from one station have consistently poor metric values, that station is temporarily removed and returned to the data documentation process. Additionally, for the Cl^- model, outliers include stations with non-marine chloride deposition, such as roads where snow is melted with salt or continental salt ponds, for which the input variables cannot explain the deposition and are, therefore, out of scope.

The third step of model building is data analysis. In this stage, plots are created from the training data, focusing mainly on distributions and correlations. Distribution analysis involves checking the model data by category in terms of corrosivity or pollution levels to verify whether some categories have too little data and must be merged. This is important because these levels are used to stratify the data during validation, and too few entries per category can compromise the results. Correlation analysis consists of plotting a heatmap of the binary correlations between variables. Since pollutant and metal corrosion typically exhibit non-linear relationships with independent variables, the Spearman correlation was used to analyze them.

After the removal of outliers, the builders have the final version of the datasets, and only model training remains. For the six models, Support Vector Machine, Random Forest, and Multilayer Perceptron regressors are applied—these are the main non-linear regression models used for atmospheric corrosion prediction, as shown in Figure 30. For the salt model, two additional models are compared: an equation describing salt

decay as a function of distance from the sea (FELIU; MORCILLO; CHICO, 1999), which serves as a baseline, and a PySR regressor, which is a genetic programming algorithm. For the metal models, the DRFs are used as baselines, with constants provided by ISO 9223 as initials values which are optimized by least-squares method during training, which allows for a fair comparison with other models..

The first step of the builder to prepare the models is model tuning, which consists of finding the best configuration for each machine learning model. For most models, a hyperparameter grid search is used, which involves testing every possible combination within a discrete search space. Each combination was evaluated using stratified random subsampling, with 100 executions and a 70/30 train/validation split. The metric used to compare the solutions is the Mean Squared Error (MSE), which is computationally efficient for grid searches and is typically used in atmospheric corrosion models (see Table 28). Table 8 shows the search space for the grid search for each of the machine learning algorithms tested.

Model	Search Space
Support Vector Machine Regressor (SVR)	kernel: [linear, rbf, sigmoid] C: [0.1, 1, 10, 100] degree: [3, 8] coef0: [0.0, 0.01, 0.5, 10] gamma: [auto, scale, 0.1, 0.01, 0.001, 0.0001]
Random Forest (RF), ExtraTrees (ET)	n_estimators: [100, 200, 300, 400] bootstrap: [True, False] max_depth: [7, 9, 11, 15, 20] max_features: [1.0, sqrt] min_samples_leaf: [2, 4, 8] min_samples_split: [2, 5, 10]
Multi Layer Perceptron (MLP)	Hidden layers architecture: [(16,16), (32,32), (64,64), (128,128), (32,16), (64,32), (128,64), (64,32,16)] Epochs: [100, 200, 300], batch_size: [16, 32, 64] optimizer: [adam, rmsprop] validation_size: 0.2 patience: [30, 50] min_delta: [0.001, 0.0001]

Table 8 – Hyperparameter search space for each machine learning algorithm

While grid search is an effective way of defining hyperparameters for most

models, it is not a feasible tuning solution for the PySR algorithm. Genetic programming algorithms have a large number of parameters and long training periods. Therefore, the parameters were defined based on necessities of the problem, every operator present in the literature models was used, constraints were applied to avoid recursive exponential operations, and other parameters were set to default values recommended in the PySR article citeCranmer2023pysr. Training was executed ten times; the best solution among the executions was selected. Table 9 shows the hyperparameters used for the PySR executions.

Number of executions	10
Hyperparameter	Value
Loss	MSE
Max complexity	15
Iterations	10000
Number of populations	100
Individuals per population	27
Simulated annealing	True
Operators	{+, −, ×, ÷, pow, exp}
Constraints pow exp	(1,6) 6
Nested constraints pow exp	exp:0, pow:0 pow:0, exp:0
Number of elites	1
Mutation add, insert, delete nothing, constant, operator swap, rotate, randomize simplify, optimize	2.47, 0.0112, 0.870 0.273, 0.0346, 0.293 0.198, 4.26, 0.000502 0.00209, 0.0
Crossover probability	0.0259

Table 9 – Hyperparameters of the PySR algorithm.

The final step for each builder is model comparison. A random subsampling with 100 repetitions and a 70/30 split is applied, using the determination coefficient (r^2) as the metric. Additionally, a Dunn test with a 5% significance level is applied to compare the statistical equivalence between the models. The chosen model is the one with the highest r^2 value that is not statistically equivalent to other models. If there is a set of statistically equivalent best models, the simplest model to deploy is selected, fitted with the entire training data, and send to production in the application layer.

4.3 APPLICATION LAYER

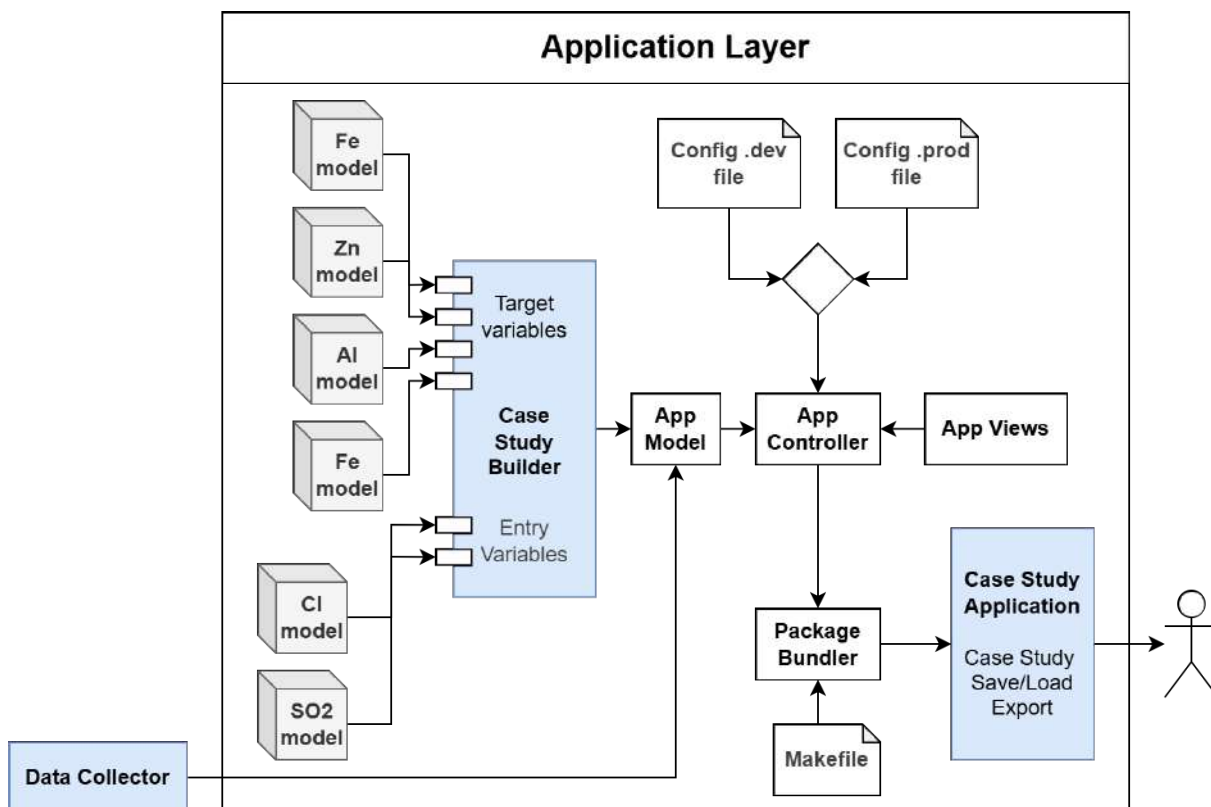


Figure 36 – Extended diagram of the application layer.

The function of the application layer is to generate case studies, that is, the calculation of r_{corr} and corrosivity for an arbitrary point through the models, and to provide a user interface for inputting the necessary metadata variables to run such models and obtain the results. Figure 36 shows the extended diagram of the application layer. This layer adopts a Model–View–Controller architecture, where the machine learning models, the Case Study Builder, and the Data Collector are contained in the model component. The views contain the interface elements developed using Streamlit, and the controller wraps the models, views, and configurations. Using a package bundler, the project is converted into a desktop application, which is the final version of the system, intended for use by atmospheric corrosion specialists.

The first components of the application layer are the six target models produced by the modeling layer. The models are wrapped by the Case Study Builder, the component responsible for structuring and executing case studies using the models. Figure 37 shows a sequence diagram of the case study execution process. Through the application interface, the user provides all necessary metadata variables for the studied location. The controller passes this information to the Case Study Builder, which creates a dataset and, together with the Data Collector, loads the input variables from the datasets and the models from the project data folder. The SO_2 and Cl^- deposition

models are executed, and given the final input variables, the r_{corr} is calculated using the metal models.

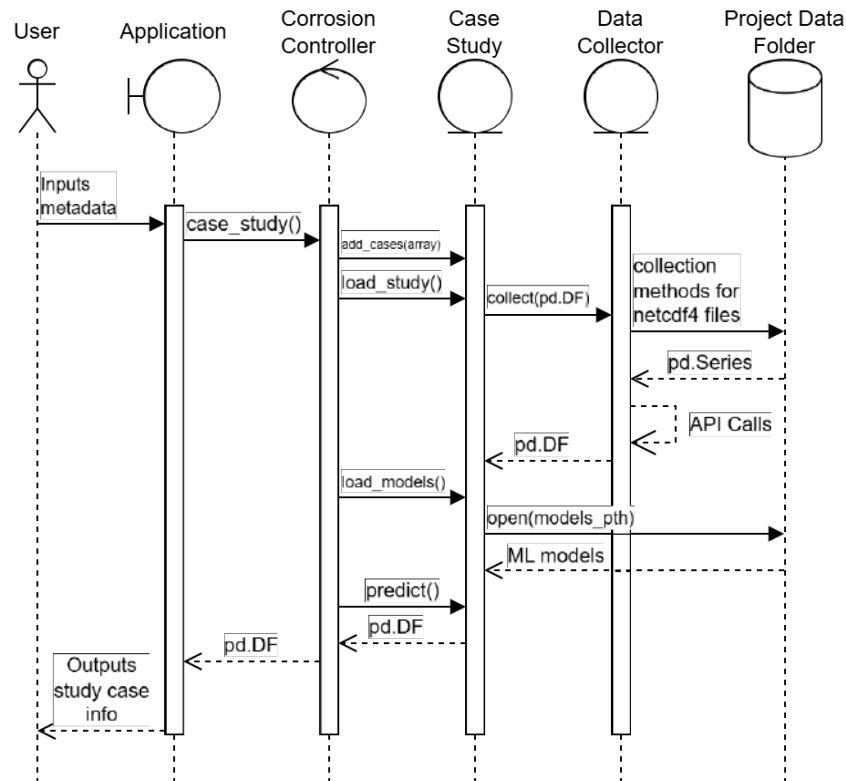


Figure 37 – Sequence diagram of a case study execution.

A challenge with this approach is the reliance on the project data folder. To train the models, the folder contains atmospheric and geospatial datasets from 1980 to the present, with global coverage, monthly resolution, and dozens of variables. Together with a local clone of the OpenStreetMap server, this folder occupies up to 800 GB of server storage, making it unfeasible to encapsulate the data folder within the application.

The solution to this data challenge is to use two versions of the system: a development version and a production version. Figure 38 shows a component diagram illustrating this division. The development version has no interface; its controllers and models are manipulated through Jupyter notebooks, mainly to train models and run experiments. The production version includes the application interface and, while maintaining the same components as the development version, is coupled with a subset of the data directory. This subset contains only data from 2023 and relies on online API calls when possible. This reduces the 800 GB dataset to only 10 GB, greatly increasing the system's applicability.

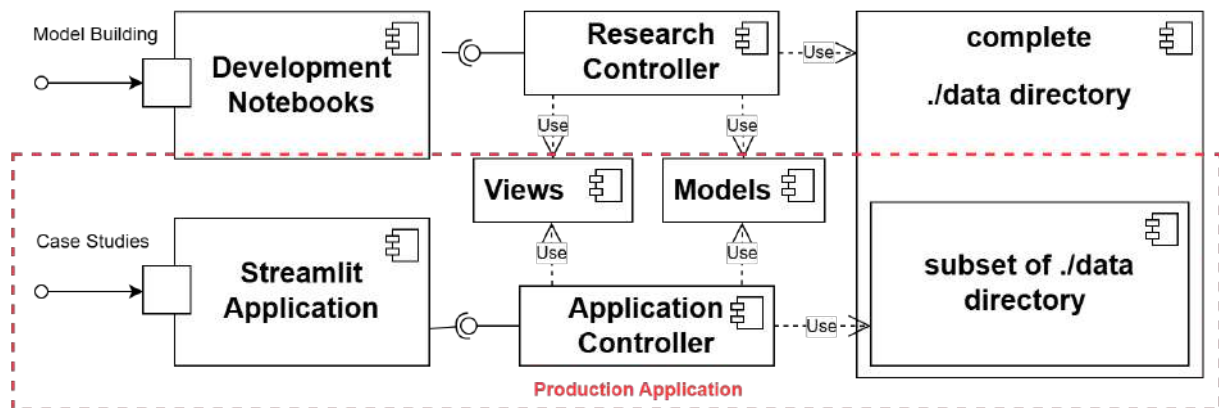


Figure 38 – Component diagram of the system versions.

The application interface must allow the execution of case studies. For clarity, each entry in a case study is referred to as a “case,” and a set of cases is referred to as the “study.” The application must meet the following requirements:

- A form for entering metadata information for each case individually, with an option to insert the case into the study.
- An option to run the study, which uses the Case Study Builder and the Data Collector to gather the variables and run the models.
- A map interface to add coordinates using markers and measure the distance to the coastline with ruler tools.
- Options to save, load, and download case studies.

The final result is an web application which allows for execution of case studies for anywhere in the world for the specified year of the subset. Images and descriptions of the final application are shown in the Application section of the Results chapter.

5 RESULTS AND ANALISYS

The preliminary results of this work were separated in seven sections: one for each of the six models (SO_2 , Cl^- , Zn , Fe , Al and Cu) and one for the application. In each model section, a distribution,

The dataset used for the six models obtained by the Data Layer unifies data from MICAT, ISOCORRAG, Asia, Maps of Chile and Canary Island and Brazilian exposures, the final Corrosion DB has a total of 320 stations, located in all continents. Figure 39 shows the map of the stations present in the Corrosion DB. Most of data in concentrated in South America, Europe and Southeast Asia. A reasonable amount of data is from tropical and subtropical regions, which should allow a better adjustment of the model for Brazilian climates.



Figure 39 – Map of the Corrosion Database stations

The number of entries varies according to the target, depending on availability and taking into account the presence of outliers. The data for each target is stratified into categories provided by ISO 9223, which will be presented in the sessions for each model/target.

5.1 SULPHUR DIOXIDE DEPOSITION MODEL (SO_2)

For the SO_2 model, the resulting training dataset has 506 entries from 123 stations. Data from Ceará showed values approximately ten times higher than the rest of the dataset and were removed as outliers. The dataset was stratified according to the pollution levels given by ISO 9223, which are presented in Table 10:

SO_2 (mg/m ² .d)	Level
$P_d \leq 4$	P_0 Rural atmosphere
$4 < P_d \leq 24$	P_1 Urban atmosphere
$24 < P_d \leq 80$	P_2 Industrial atmosphere
$80 < P_d \leq 200$	P_3 Highly polluted industrial atmosphere

Table 10 – Distribution of SO_2 data by ISO 9223 pollution levels. Adapted from (ISO/9223, 2012).

Figure 40 shows a histogram of the exposure distribution across pollution levels. The dataset shows imbalance only in P_1 , with approximately twice as many entries as the other levels, which demonstrates our dataset has a slight bias toward urban environments, though it is worth noticing that both P_2 and P_3 represent industrial atmospheres, which implies the truly underrepresented category in the dataset is rural atmospheres. Therefore, stratification is important for maintaining proportions of the categories during sampling, even though the impact is probably smaller compared to other datasets presented in this work.

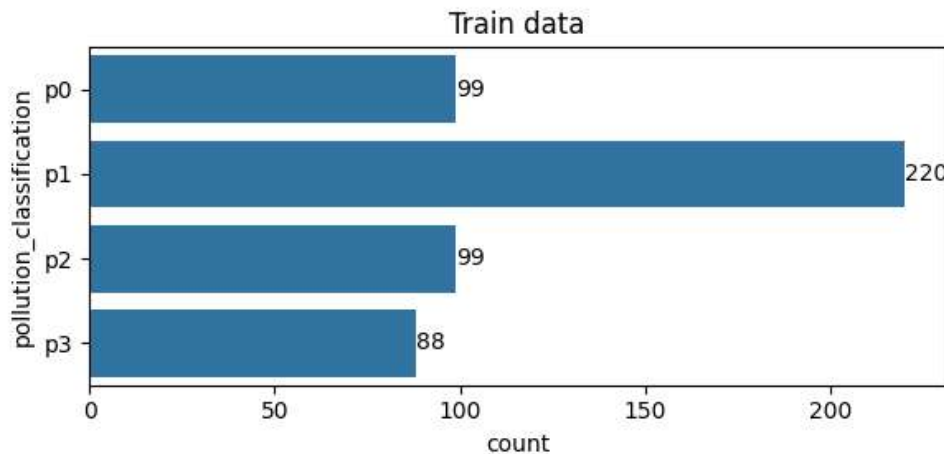


Figure 40 – Distribution of SO_2 entries across pollution levels. Source: Author.

The Spearman correlations of the variables are shown in Figure 41. The target variable SO_2 is displayed in the last row, with correlation values ranging from -1 (blue) to 1 (red). The strongest correlation is with weighted industrial areas with a correlation of 0.65, confirming industrial emissions as the main source of SO_2 . This result is impressive considering we don't have a temporal resolution for this variable, implying that maps of industrial areas, even decades apart, can still be used to estimate pollution.

As for atmospheric variables, it is worth noticing that EDGAR emissions show higher correlations than MERRA-2 depositions, even though deposition is more closely related to the lead candle measurement, this is probably due to the higher resolution of EDGAR compared to MERRA-2 (10km to 30km), reinforcing the microclimate nature of SO_2 deposition. While CO_2 production per capita presents low correlation values, its effects are assumed to be interpretable only in conjunction with weighted urban areas.

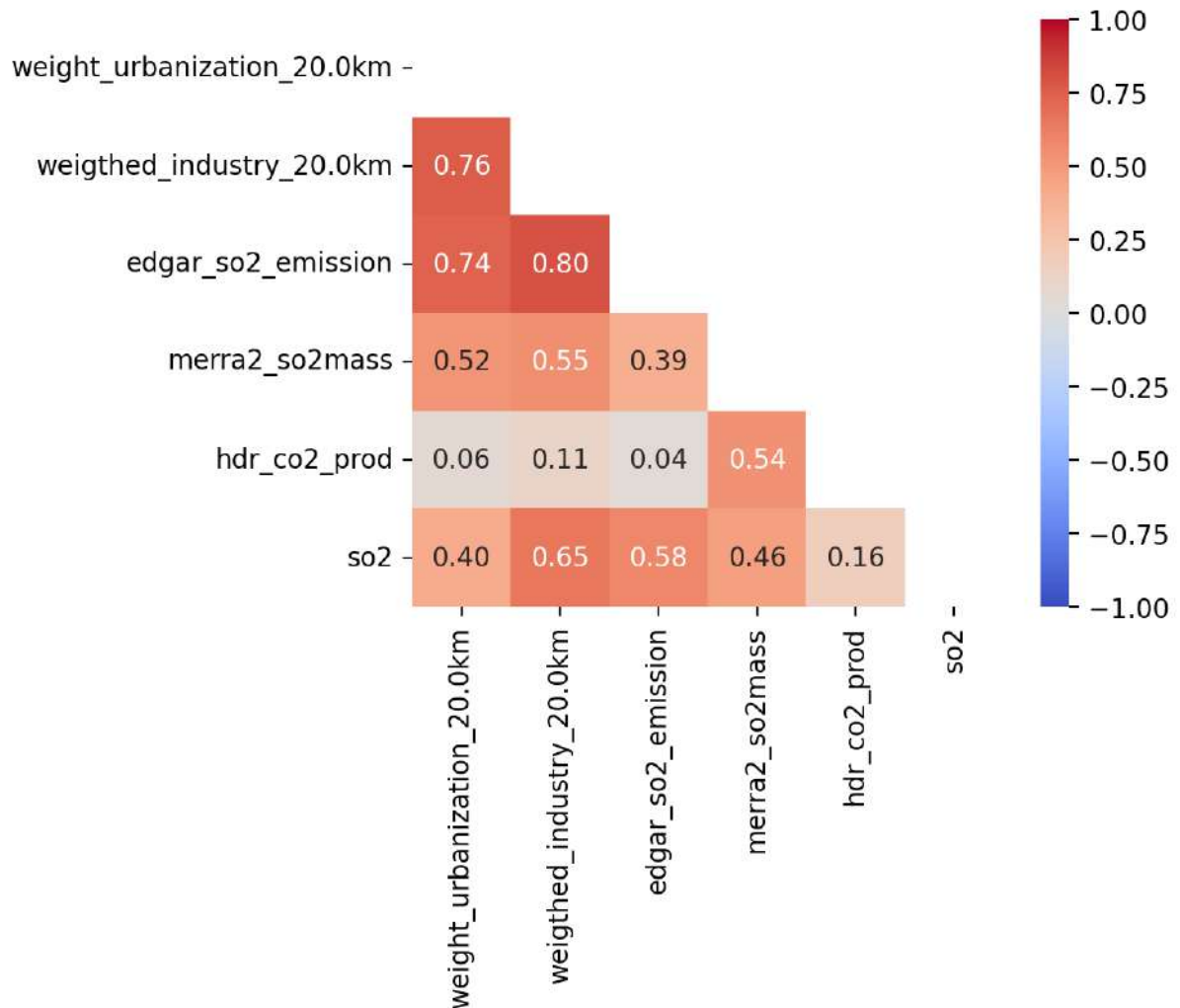
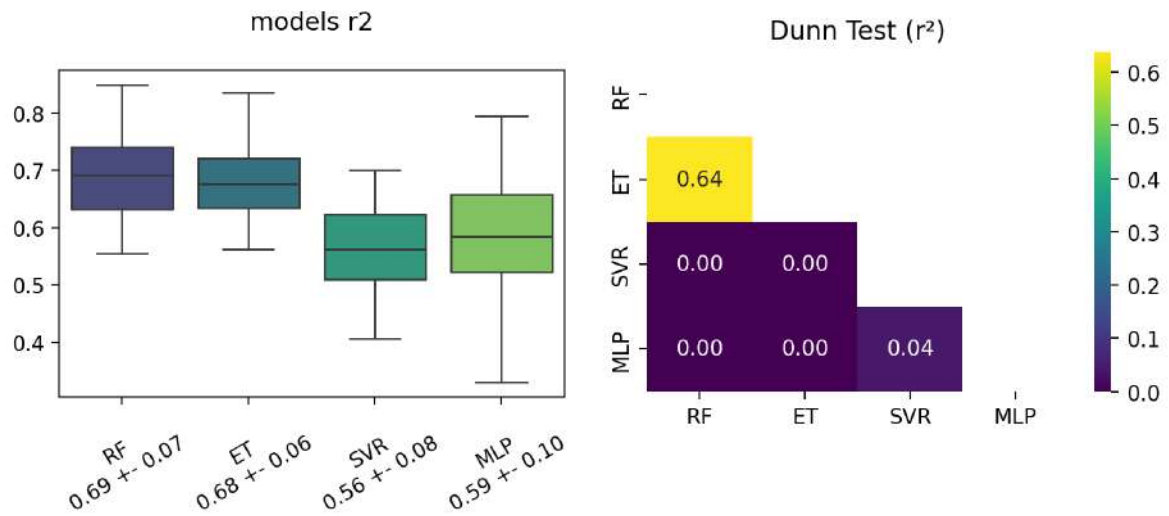


Figure 41 – SO_2 model variables Spearman correlation.

The results of the cross-validation are shown in Figure 42a, which contains a boxplot of validation scores for the 100 random 70/30 subsamplings, and Figure 42b shows a confusion matrix of the Dunn test p-values. The tree ensemble models achieved the best results of validation, with the Random Forest (RF) and Extra Trees (ET) models being statistically equivalent at the 5% confidence level, with r^2 values of 0.69 and 0.68, respectively. The Multi-Layer Perceptron (MLP) and the Support Vector Regressor (SVR) showed similar results, with r^2 values of 0.56 and 0.59, respectively, with worse results compared to the RF and ET.



(a) Boxplot for the random subsampling residuals (b) Dunn Test between residuals distribution

Figure 42 – Results for the SO_2 model. Source: Author.

Therefore, for SO_2 deposition prediction, ensemble models demonstrated the best performance and therefore are used in the system application. These results are consistent with previous works (GEREMIAS et al., 2025b), and show it is possible to predict ISO 9223 pollution even with temporal limitations of the corrosion datasets.

However, it is important to note that using data from the 1980s and 1990s to make predictions for the present is far from the best scenario. Unlike chloride, pollutant emissions are primarily caused by human activity and have shown significant regional variations over the years. This change will allow for both greater confidence in the use of the model and the use of more modern variables.

5.2 CHLORIDE DEPOSITION MODEL (Cl^-)

For chloride dry deposition prediction (Cl^-), the final version of the Cl^- dataset has 419 entries from 213 stations. Continental stations with non-marine high chloride concentrations, such as salt desert stations (San Pedro de Atacama, Pampa del Tamarugal) and stations near roads in cold climates (Boucherville, Vladivostok, Batumi), were removed. The final dataset was stratified according to the pollution classification from ISO 9223, from S_0 to S_3 , as shown in Table 11:

Cl^- (mg/m ² .d)	Level
$S_d \leq 3$	S_0
$3 < S_d \leq 60$	S_1
$60 < S_d \leq 300$	S_2
$300 < S_d \leq 1500$	S_3

Table 11 – Distribution of Cl^- data by ISO 9223 pollution levels. Adapted from (ISO/9223, 2012).

The results of this classification are shown in the histogram in Figure 43. Unlike the SO_2 model, the Cl^- dataset is highly imbalanced, with most entries in level S_1 (3 to 60), characteristic of continental regions. This highlights the importance of data stratification, since the lack of high Cl^- deposition values means that a random subsample could yield an extremely unbalanced split between training and validation datasets, undermining the validity of the results.

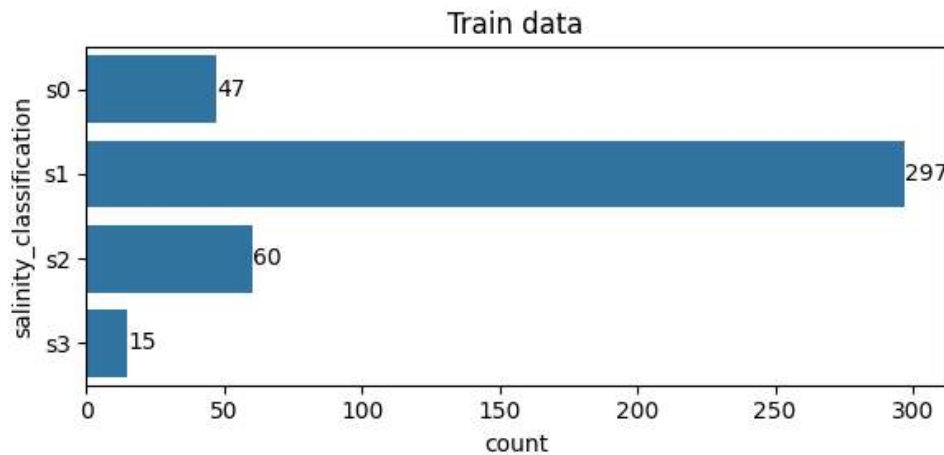


Figure 43 – Distribution of Cl^- data by ISO 9223 pollution levels. Source: Author.

The Spearman correlations for the Cl^- dataset are shown in the heatmap in Figure 44. The variable with the highest correlation magnitude is the distance to the coastline (-0.68), consistent with the literature, as it governs the decay of airborne salt concentration in all proposed models (GUSTAFSSON, 1997; FELIU; MORCILLO; CHICO, 1999; MEIRA et al., 2008). The other two geospatial variables also showed

significant correlation, with trajectory roughness showing a correlation of -0.50 and elevation -0.42, demonstrating the importance of micro-climate features.

As for atmospheric variables, ERA5 Wind Speed follows has the second-highest correlation with chloride (0.57), and while the observation agrees is according to the literature, it introduces a new aspect of since reanalysis data has a much coarse precision compared to on-site collections, yet it still maintains high correlations. MERRA-2 surface mass concentration (SSSMAS), which has an even coarser resolution (30km, compared to 10km on ERA5), and yet achieves a correlation of 0.49. Observing the variables, we can affirm that both macro-climate and micro-climate variables influence in the dry deposition of chlorides.

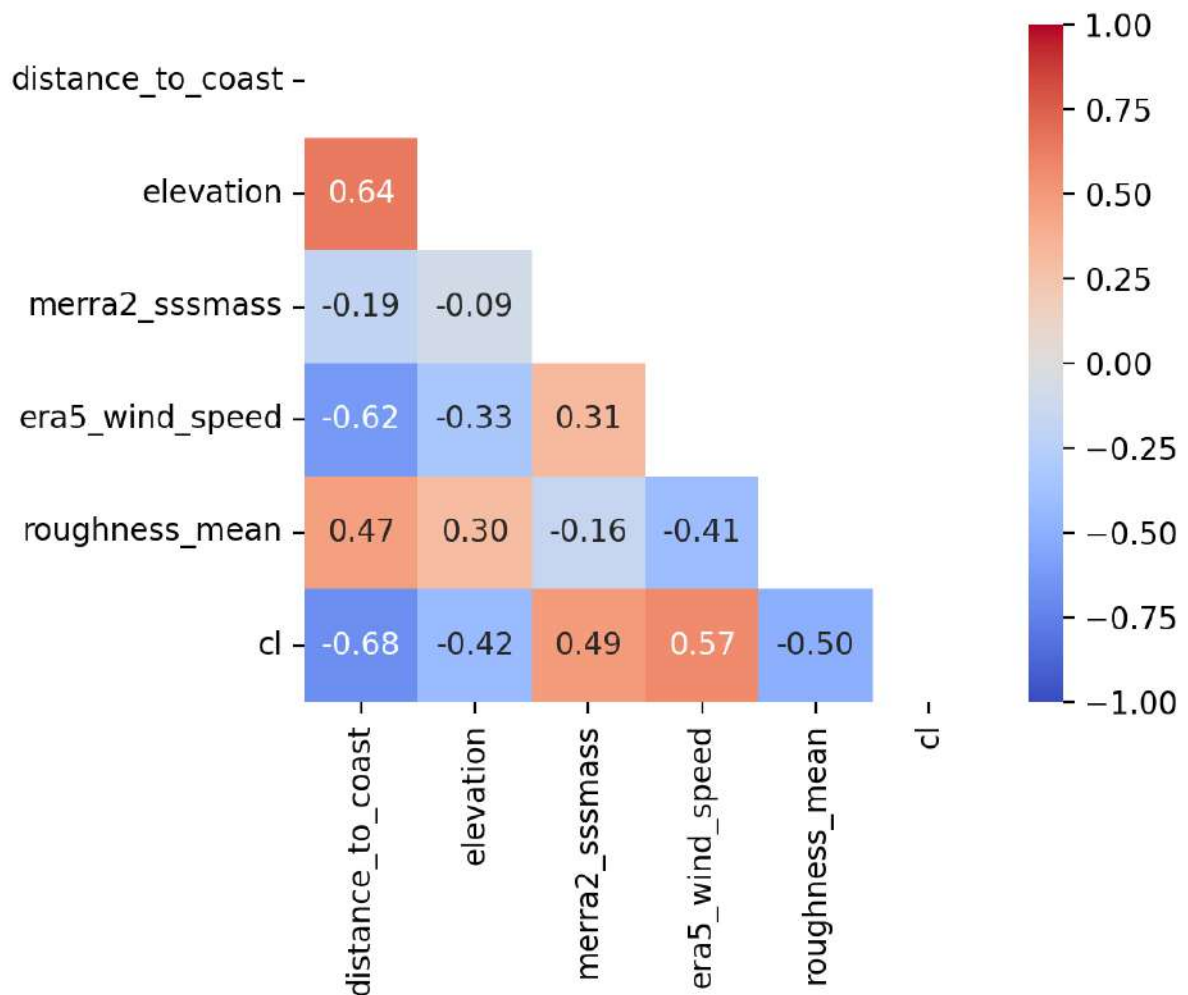


Figure 44 – Cl^- model variables Spearman correlation.

The results of the Cl^- model validation are shown in Figure 45. Figure 45a shows a boxplot of the r^2 residuals from the random subsampling, and Figure 45b shows a heatmap of the Dunn test p-values between models residuals. The best r^2 values are from the Random Forest (0.72) and the PySR (0.71), which are statistically equivalent at a 5% confidence level. The literature's decay model (FELIU; MORCILLO;

CHICO, 1999) (0.53) performs slightly worse than the Multi-Layer Perceptron (0.58), and the Support Vector Machine Regressor shows the worst result (0.33).

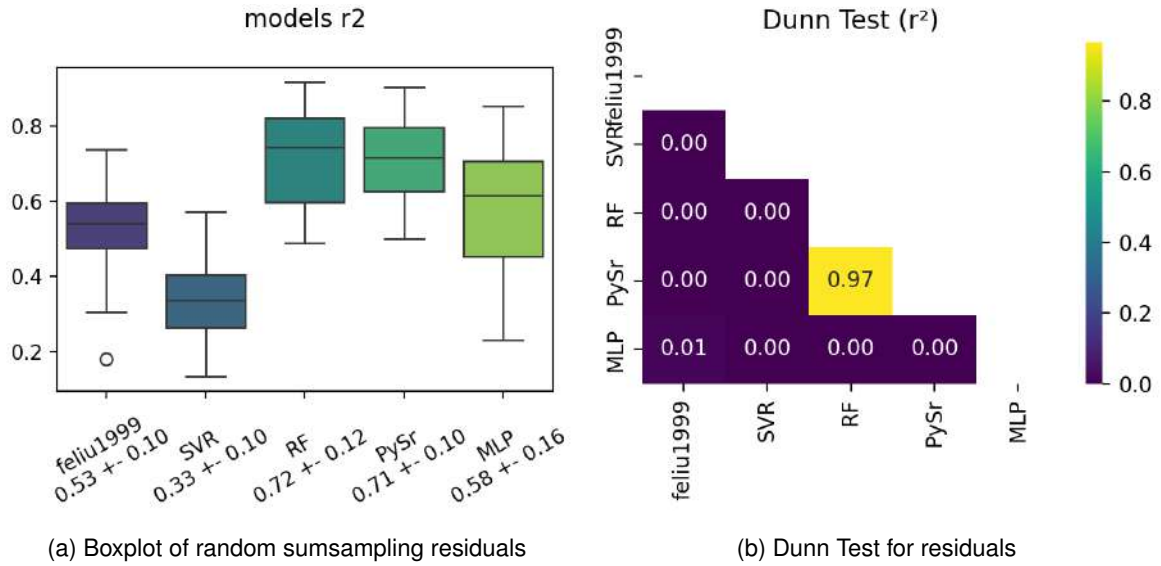


Figure 45 – Results of the Cl^- model validation. Source: Author.

Between the Random Forest and the PySR, the latter was chosen since its final result is an explicit equation (Equation 5.1), which is much more interpretable than a black-box approach such as Random Forest. The obtained equation shows a decay with distance to the coastline similar to that found in the literature, but with two additional terms guiding the concentration: MERRA-2 surface mass concentration and wind speed by elevation. An empirical interpretation is that low-elevation regions with high wind speed, high salt concentration, and proximity to the coast tend to have higher deposition values, consistent with the observed Spearman correlations.

$$Cl^- = \frac{\frac{0.001 \cdot 4.172 \cdot \text{wind speed}}{\text{elevation} - 7.296} + \text{ssmass}}{\text{dist}^{0.476}} \quad (5.1)$$

Where *wind speed* is the ERA5 Windspeed, *ssmass* is the MERRA-2 Sea Salt Surface Concentration, *elevation* is the SRTM30+ elevation and *dist* is the distance to coastline collected by gmaps ruler tool.

5.3 ZINC CORROSION MODEL (ZN)

For Zinc corrosion rate prediction, the Zn dataset contains 259 entries from 88 stations worldwide. Outliers were mainly stations with imprecise location data. The dataset was stratified using the atmospheric corrosivity scheme from ISO 9223, shown in Table 12:

Range ($\mu\text{m}/\text{year}$)	Corrosivity (Zn)
$r_{corr} \leq 0.1$	C1
$0.1 < r_{corr} \leq 0.7$	C2
$0.7 < r_{corr} \leq 2.1$	C3
$2.1 < r_{corr} \leq 4.2$	C4
$4.2 < r_{corr} \leq 8.4$	C5
$8.4 < r_{corr} \leq 25$	CX

Table 12 – Distribution of Zn data by ISO 9223 corrosivity levels. Adapted from (ISO/9223, 2012).

The data distribution across corrosivity categories is shown in the histogram in Figure 46. Since the C1 and CX categories had insufficient entries for subsampling, they were merged into the C1-C2 and C5-CX categories, respectively. The dataset is heavily imbalanced, with most data falling into the C3 category. Therefore, stratification is necessary to maintain class balance during subsampling, especially given the scarcity of high-corrosivity zinc environments. This measure provides a more reliable results, reducing extreme scenarios in sampling process.

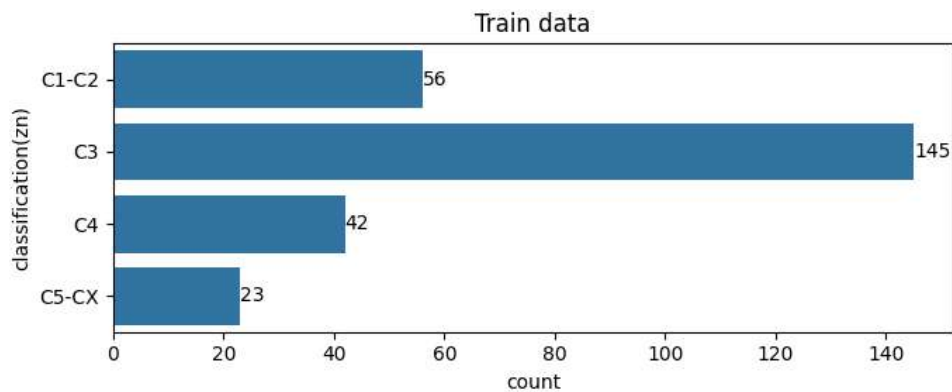


Figure 46 – Distribution of Zn data by ISO 9223 corrosivity levels. Source: Author.

The Spearman correlations of the Zn model variables are shown in Figure 47, with the last row representing the target variable and other rows the input variables. The strongest absolute correlation is with Cl^- dry deposition (0.57), much higher than for SO_2 deposition, consistent with zinc and galvanized steel being more susceptible to chloride-induced corrosion than urban/industrial pollutants. Elevation and distance to the coastline are the next most correlated variables (-0.48 for both), indicating that

chloride deposition alone does not fully capture the micro-climate effects of coastal regions. Experimental removal of either variable significantly decreases the model's r^2 .

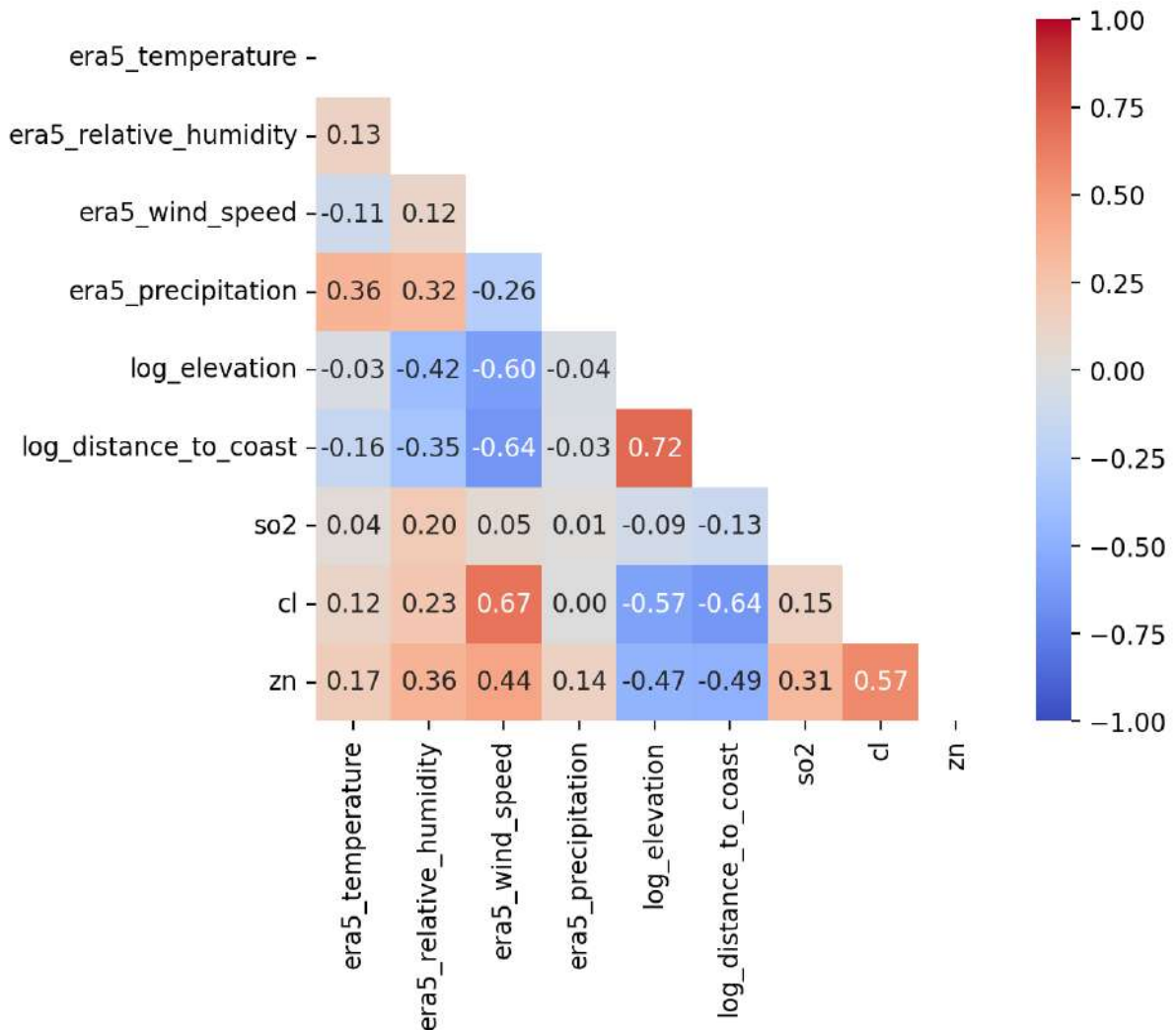
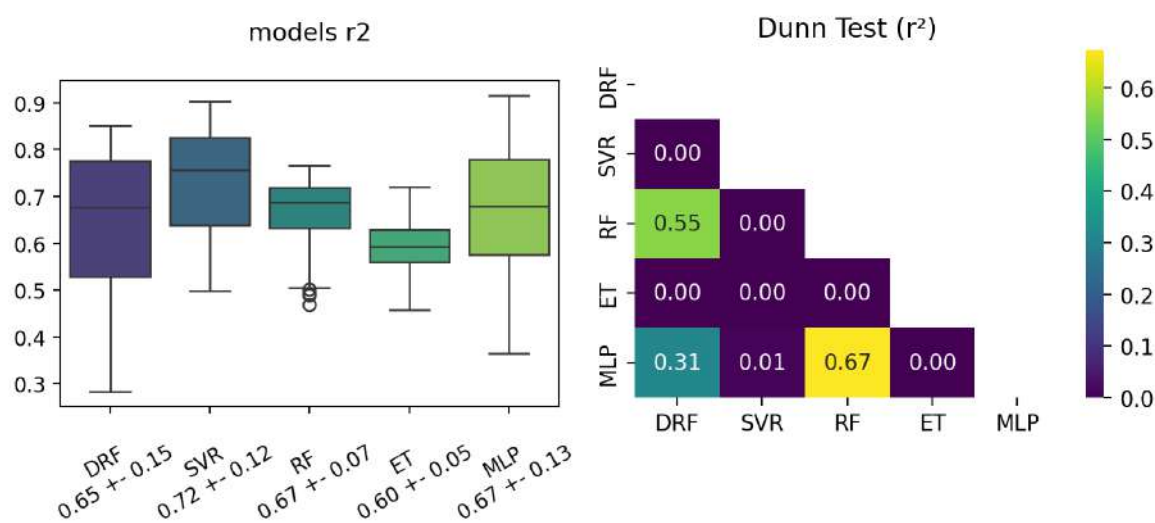


Figure 47 – Zn model variables Spearman correlation.

The results of validation using random subsampling are shown in Figure 48. Figure 48a shows a boxplot of validation residuals, and Figure 48b is a heatmap of Dunn test p-values between models residuals. The Support Vector Machine Regressor (SVR) achieved the highest score, with an r^2 value of 0.72, statistically different from other models at a 5% confidence level according to the Dunn test. The baseline Dose-Response Function, Random Forest, and Multi-Layer Perceptron ANN models performed slightly worse, with r^2 values ranging from 0.60 to 0.67, and are statistically equivalent to each other at a 5% confidence level.



(a) Boxplot of random subsampling residuals.

(b) Dunn test of the models residuals.

Figure 48 – Results of the Zn model validation. Source: Author.

Based on its performance compared to other models, the SVR was selected as the model to be implemented within the system. It is worth noting that, among the metal models, zinc was the only one in which SVR outperformed the ensemble tree models. Considering that the SVR model works is limited to curves in comparison to a more versatile model like tree ensembles, the hypothesis arises that the zinc dataset less sparse and better consolidated, and the curves that represent the relations between variables are more solid.

This observation is important because it suggests that it may be possible to obtain symbolic regression equations for the metal models upon a better consolidation of the training dataset. For this study, it was only possible to obtain consistent results from symbolic regressions for the chloride model; however, with some adjustments, it may be feasible to apply the same method to the metal models.

5.4 CARBON STEEL CORROSION MODEL (FE)

The Fe dataset contains 332 entries from 116 stations worldwide. Outliers were mainly stations with imprecise location data. The dataset was stratified using the Fe atmospheric corrosivity scheme from ISO 9223, shown in Table 15. It is noticeable that corrosion values for Carbon steel are much higher than other metals from ISO 9223, generally measured around 10 times higher than Zn corrosion, reason the other metals of the standard are often used for coating steel, acting as both physical and electrochemical protection.

Range ($\mu m/year$)	Corrosivity (Fe)
$r_{corr} \leq 1.3$	C1
$1.3 < r_{corr} \leq 25$	C2
$25 < r_{corr} \leq 50$	C3
$50 < r_{corr} \leq 80$	C4
$80 < r_{corr} \leq 200$	C5
$200 < r_{corr} \leq 700$	CX

Table 13 – Distribution of Fe data by ISO 9223 corrosivity levels. Adapted from (ISO/9223, 2012).

The data distribution across corrosivity categories is shown in the histogram in Figure 49. Due to Carbon Steel high sensibility to corrosion, there's almost no samples for the C1 corrosivity in the dataset, therefore it was merged onto C1-C2 category. Due to the dataset heavily imbalance, with most data falling into C2 and C3 categories, Stratification is necessary to maintain class balance during subsampling.

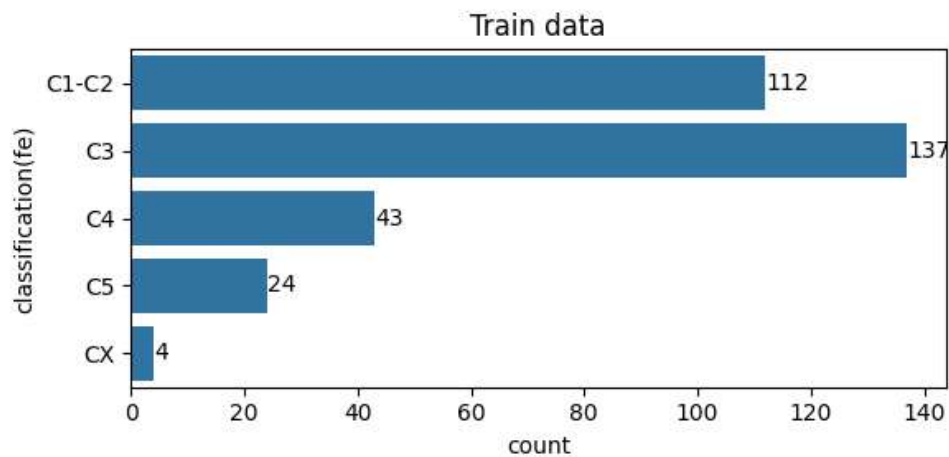


Figure 49 – Distribution of Fe data by ISO 9223 corrosivity levels. Source: Author.

The Spearman correlations of the Fe model variables are shown in Figure 57, with the last row representing the target variable and other rows the input variables. Correlation for pollutants were 0.65 for Cl^- deposition and 0.45 for SO_2 deposition, slightly higher than for other metals due to the sensibility of steel to corrosion. Eleva-

tion and wind speed also showed good correlations, with -0.59 and 0.52 respectively. Precipitation and Temperatures had almost irrelevant correlations.

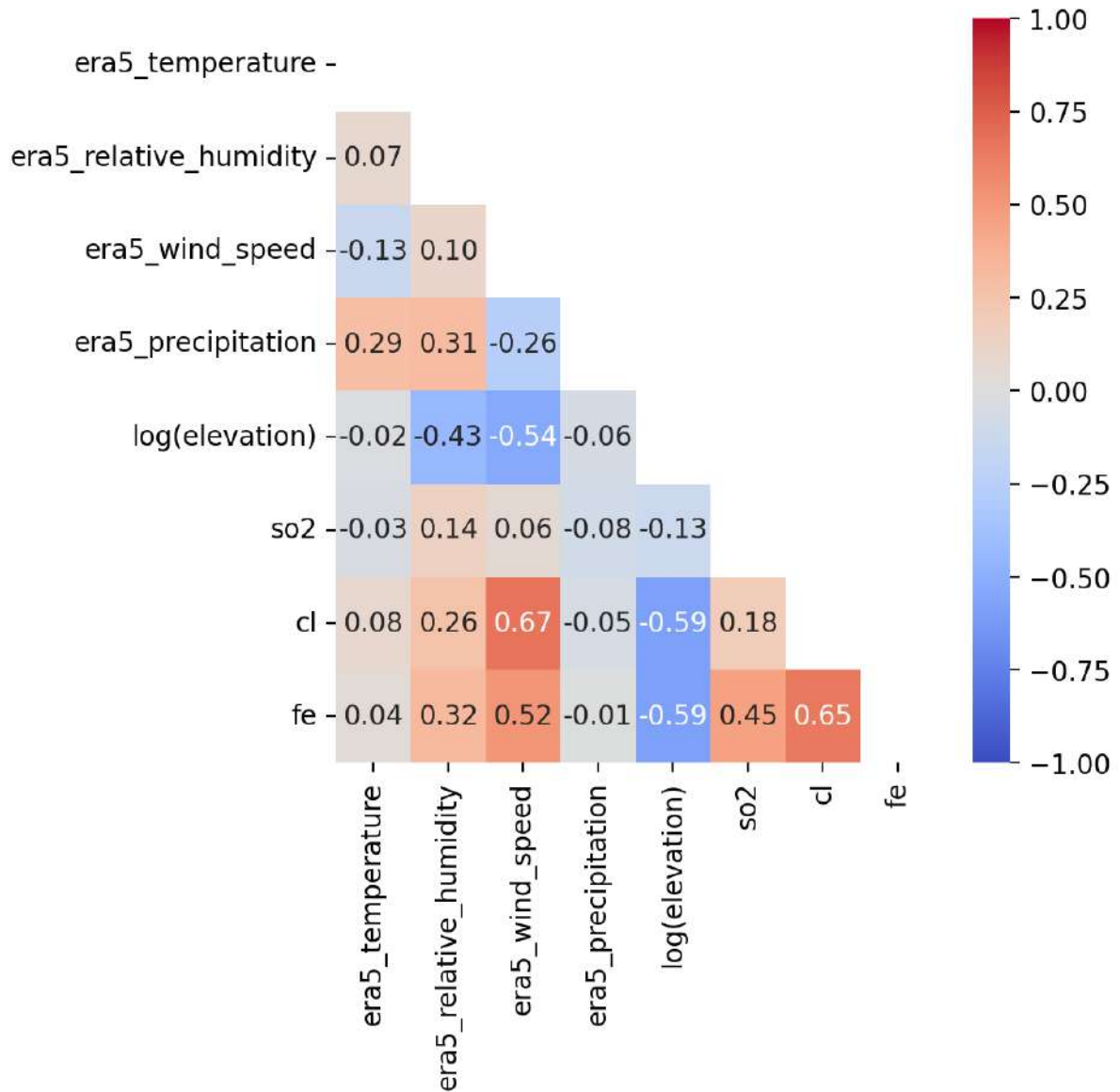
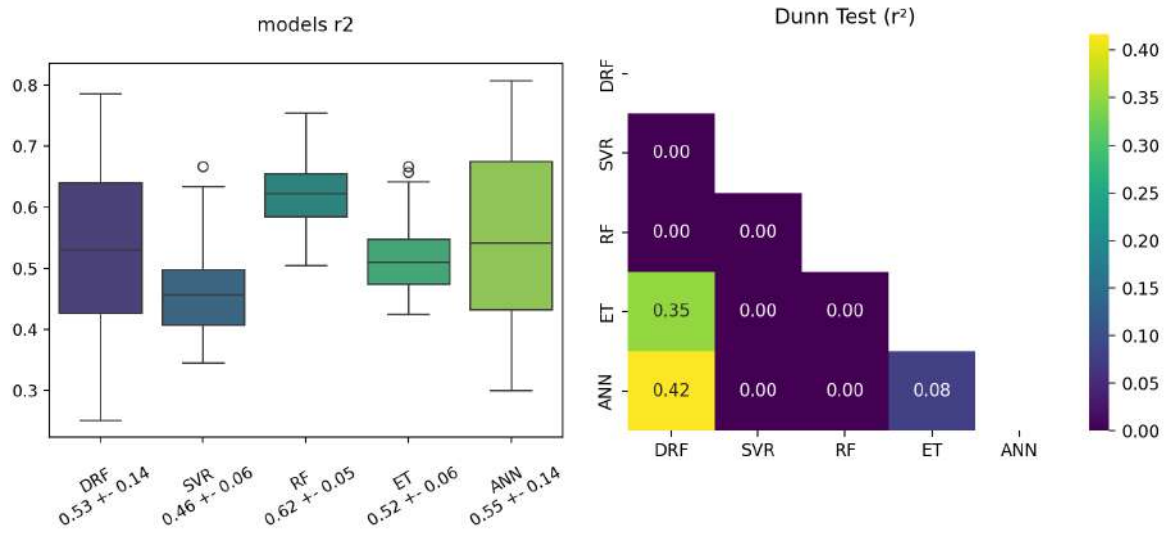


Figure 50 – *Fe* model variables Spearman correlation.

The results of validation using random subsampling are shown in Figure 51. Figure 51a shows a boxplot of validation residuals, and Figure 51b is a heatmap of Dunn test p-values between models residuals. The Random Forest (RF) obtained the best results, with an r^2 of 0.62, being statistically different than the other models. The Artificial Neural Network (ANN) model achieved equivalent results with the Dose Response Function (DRF), with r^2 values 0.55 and 0.53 respectively. Differently from the *Zn* model, the Support Vector Machine (SVR) performed poorly ($r^2 = 0.46$) than other models.

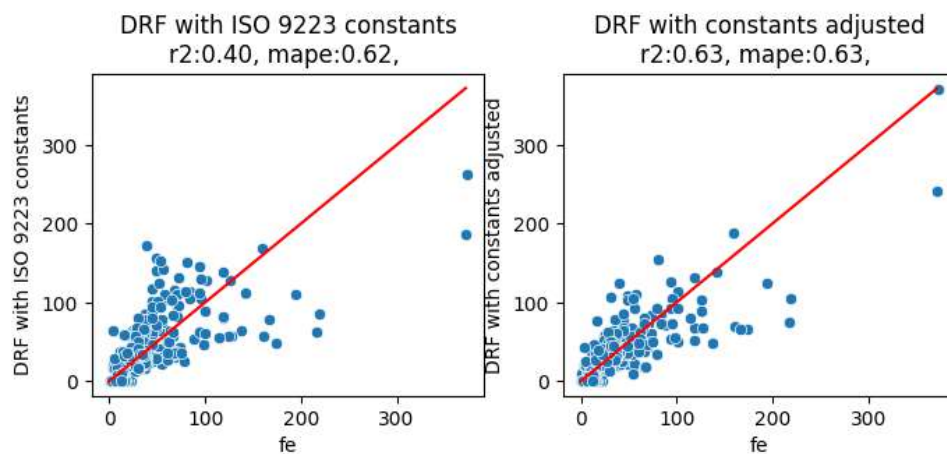


(a) Boxplot of random subsampling residuals.

(b) Dunn test of the models residuals.

Figure 51 – Results of the Fe model validation. Source: Author.

One observation made during the experiments is that the coefficients obtained for iron during validation differ significantly from those provided by ISO 9223, and there is a much higher difference in r^2 values between the equations when taking into account the entire dataset, as shown in Figure 52. This was not the case for other metals, where adjusted DRFs stayed similar. This, combined with the fact that the R^2 values for iron are lower than those for zinc, suggests that the corrosion data for iron are significantly more dispersed than those for zinc,

Figure 52 – Comparison between ISO 9223 Fe Dose Response Function and adjusted version in the validation.

5.5 ALUMINUM CORROSION MODEL (AL)

For aluminum corrosion prediction, the *Al* dataset contains 259 entries from 88 stations worldwide. The dataset was stratified according to the atmospheric corrosivity scheme adapted from ISO 9223, shown in Table 14. Since the standard does not explicitly define corrosivity categories for *Al*, the categories were derived from mass loss values.

Range ($g/(m^2a)$)	Range ($\mu m/year$)	Corrosivity (<i>Al</i>)
negligible	negligible	<i>C1</i>
$r_{corr} \leq 0.6$	$r_{corr} \leq 0.22$	<i>C2</i>
$0.6 < r_{corr} \leq 2.0$	$0.22 < r_{corr} \leq 0.74$	<i>C3</i>
$2.0 < r_{corr} \leq 4.2$	$0.74 < r_{corr} \leq 1.56$	<i>C4</i>
$4.2 < r_{corr} \leq 8.4$	$1.56 < r_{corr} \leq 3.11$	<i>C5</i>
$8.4 < r_{corr} \leq 25$	$3.11 < r_{corr} \leq 9.26$	<i>CX</i>

Table 14 – Distribution of *Al* data adapted from ISO 9223 corrosivity levels. Adapted from (ISO/9223, 2012).

The data distribution across corrosivity categories is shown in the histogram in Figure 53. The dataset is highly imbalanced, with most data concentrated in the *C3* category; therefore, stratification plays an important role in maintaining stability during validation.

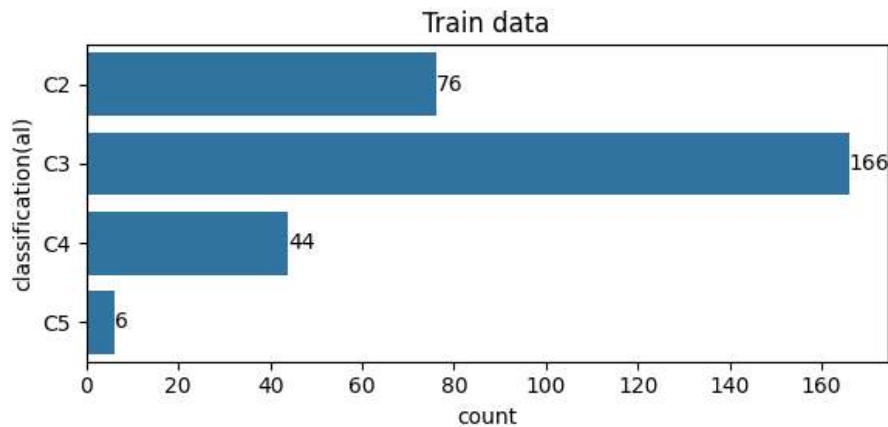


Figure 53 – Distribution of *Al* data by ISO 9223 corrosivity levels. Source: Author.

The Spearman correlations of the *Al* model variables are shown in Figure 54, with the last row representing the target variable and the other rows the input variables. The *Al* variables show lower correlations compared to other metals, with a possible explanation being that aluminum is primarily affected by pitting corrosion rather than uniform thickness loss, which may not be fully captured by r_{corr} . The strongest absolute correlation is observed for Cl^- dry deposition (0.47), followed by temperature and wind speed, indicating a strong influence of coastal environments on aluminum corrosion.

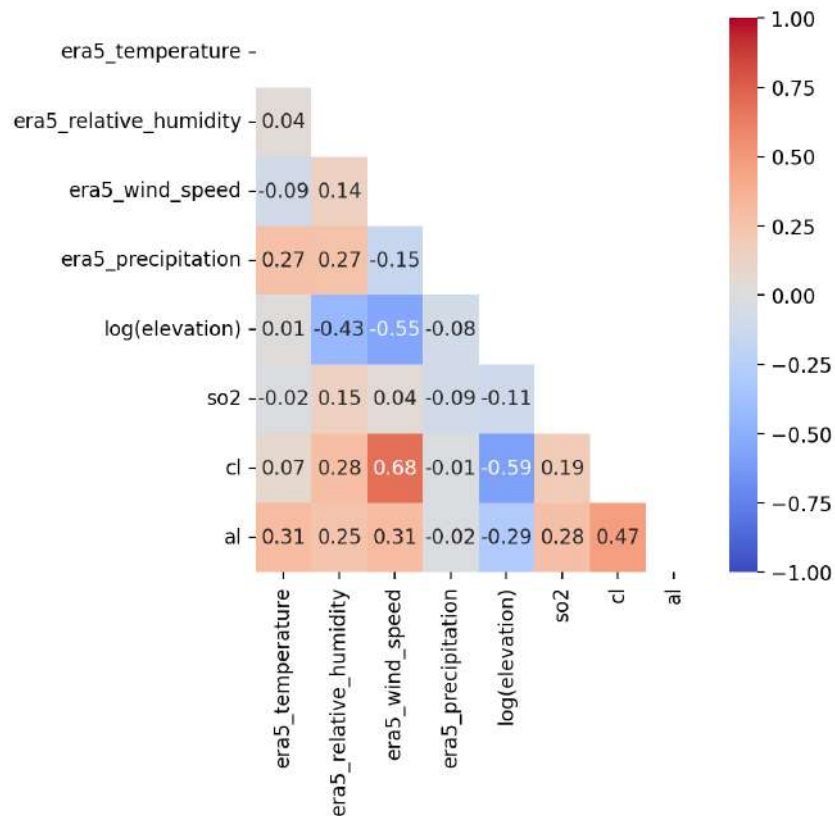
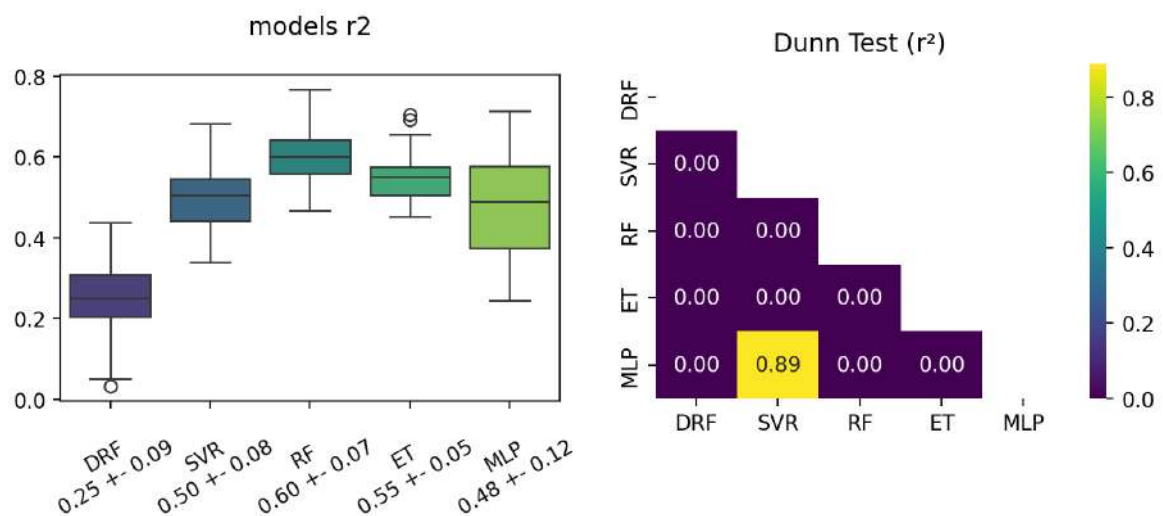


Figure 54 – AI model variables Spearman correlation.

The results of the validation using random subsampling are shown in Figure 55. Figure 55a presents the boxplot of validation residuals, while Figure 55b shows the heatmap of Dunn test p-values between model residuals. Tree-based ensemble models achieved the best results for the AI model, with r^2 values of 0.60 for the Random Forest and 0.55 for the Extra Trees model. Overall, the r^2 values for aluminum are lower than those of other models, which is consistent with ISO 9223 explanations.



(a) Boxplot of random subsampling residuals.

(b) Dunn test of the model residuals.

Figure 55 – Results of the AI model validation. Source: Author.

5.6 COPPER CORROSION MODEL(CU)

For Copper, the Cu dataset contains 339 entries from 118 stations worldwide. Outliers were mainly stations with imprecise location data, which performed poorly in analysis. Originally, ISO 9223 The Cu dataset was stratified using the atmospheric corrosivity scheme from ISO 9223, shown in Table 15:

Range ($\mu m/year$)	Corrosivity (Cu)
$r_{corr} \leq 0.1$	$C1$
$0.1 < r_{corr} \leq 0.6$	$C2$
$0.6 < r_{corr} \leq 1.3$	$C3$
$1.3 < r_{corr} \leq 2.8$	$C4$
$2.8 < r_{corr} \leq 5.6$	$C5$
$5.6 < r_{corr} \leq 10$	CX

Table 15 – Distribution of Cu from ISO 9223 corrosivity levels. Adapted from (ISO/9223, 2012).

The data distribution across corrosivity categories is shown in the histogram in Figure 53. The Cu datasets has almost no entries within the C1 and CX categories. Therefore, for the purpose of stratification, such classes were unified in C1-C2 and C5-CX respectively. The results still is a dis-balanced dataset, with most entries concentrated in the C3 and C4 categories.

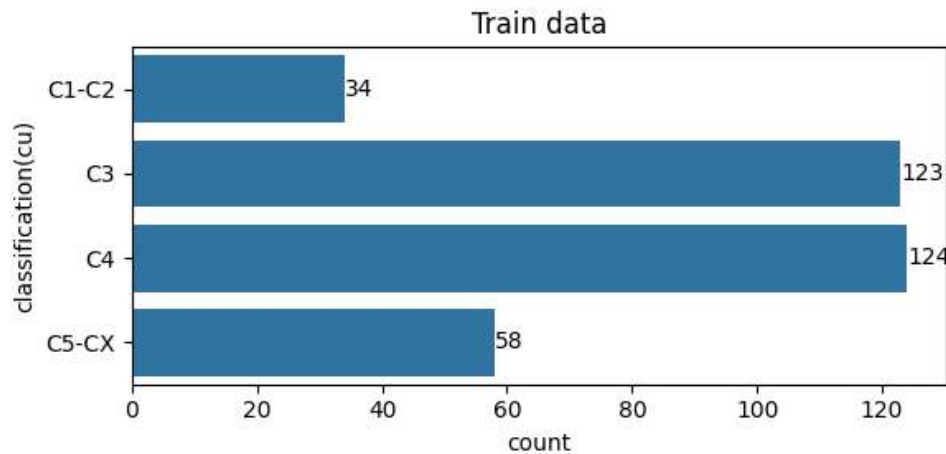
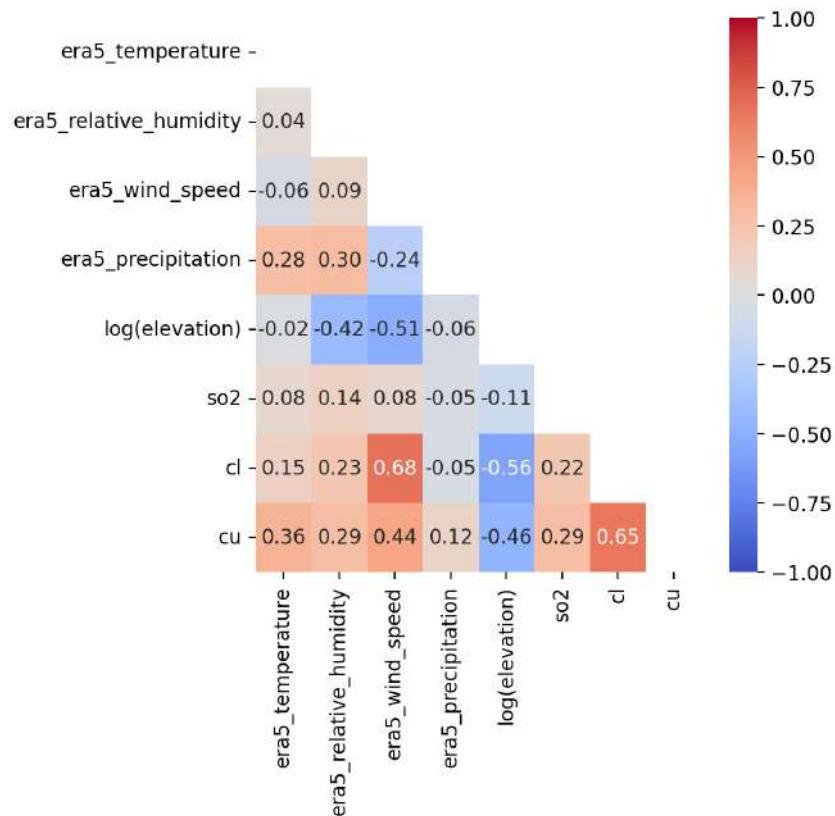
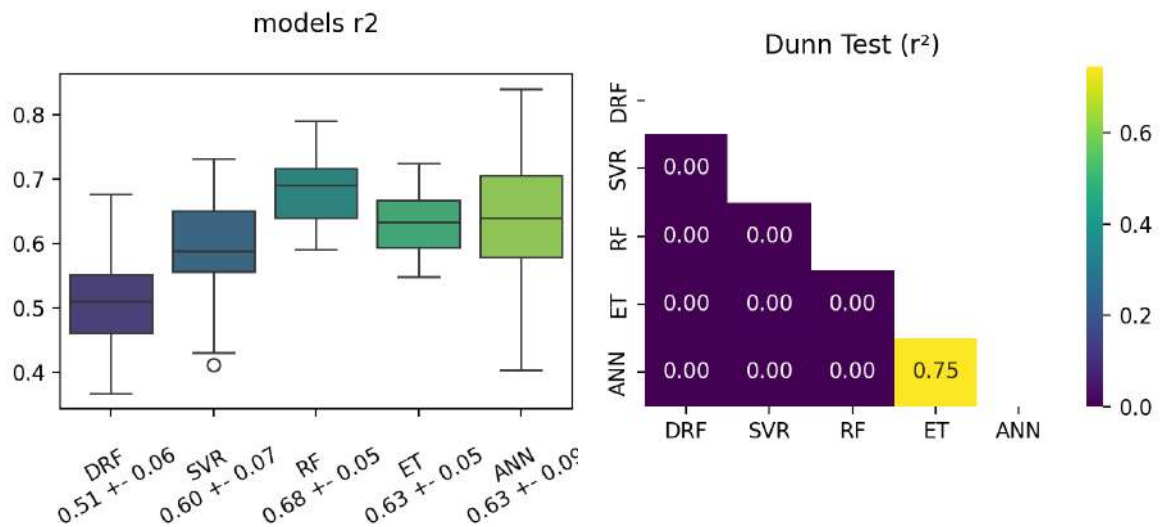


Figure 56 – Distribution of Cu data by ISO 9223 corrosivity levels. Source: Author.

The Spearman correlations of the Cu model variables are shown in Figure 57, with the last row representing the target variable and other rows the input variables. The corrosion rate of copper shows a strong correlation with chloride deposition (spearman=0.65), and the low correlation with sulfur dioxide, much like the aluminum model.

Figure 57 – *Cu* model variables Spearman correlation.

The results of validation using random subsampling are shown in Figure 58. Figure 58a shows a boxplot of validation residuals, and Figure 58b is a heatmap of Dunn test p-values between models residuals. The Random Forest model achieved the highest score, with an r^2 value of 0.68, statistically different from other models at a 5% confidence level according to the Dunn test.



(a) Boxplot of random subsampling residuals.

(b) Dunn test of the models residuals.

Figure 58 – Results of the *cu* model validation. Source: Author.

5.7 APPLICATION LAYER RESULTS

Figure 59 shows the LABICORRs case study page. The interface contains four main elements: the form, the map widget, the data editor, and the save/load/download menu. Each element includes information bullets to guide the user.

Formulary

Name:

Latitude: Longitude:

Distance to coast (km): Reference year:

Metals: ☒ Zn ☒ Fe ☐ Al ☐ Cu

Map

Map widget showing a satellite view of a coastal area with a blue location pin and a green line indicating the distance to the coast. Map controls (zoom in, zoom out, full screen) are visible on the left.

Table Editor

	Lat	Lon	Dist. Coast (km)	Wind Speed (m/s)	Temperature (C)	Humidity (%)	SO2 (mg/m ² d)	Cl- (m)

Save Data **Load Data** **Download**

Figure 59 – Case Study Page. Source: Author.

The user inputs metadata for each case individually using the form on the left. Two options are available: add the case to the study or run the study, which then collects the variables and executes the models. On the right, a map widget allows the user to place markers to update locations and use ruler tools to calculate the distance to the coastline.

Below the form and map widget is a table editor, which allows manual editing of case studies, similar to a spreadsheet editor. Beneath the table are options to save, load, and download studies. Files can be downloaded in CSV format from the table editor or XLSX format from the download button.

6 CONCLUSION

Corrosion is the deterioration of metals through interaction with the environment. The costs associated with maintenance and prevention of corrosion correspond to approximately 3% to 4% of the GDP of industrialized countries (SCHMITT et al., 2009), which justifies the study of corrosion across different environments to ensure effective material protection. For atmospheric conditions, standards developed by ISO/TC 156 provide technical methods to measure and estimate corrosion rates and atmospheric corrosivity, mainly through dose-response functions and classification tables.

Although widely used, the ISO 9223 dose-response functions present limitations in applicability, as their input variables must be collected through on-site field experiments, which require long exposure periods—typically up to one year—and specialized technical expertise. To overcome these limitations, alternative approaches have been proposed, such as corrosivity mapping using interpolation techniques and machine learning models that rely on readily available environmental variables (CRISTOS et al., 2025; GEREMIAS, 2024).

In this context, machine learning has been increasingly adopted to estimate corrosion rates based on available data. Two main approaches can be identified: Atmospheric Corrosion Exposure (ACE) prediction, a generalist approach that combines ISO 9223 exposure data from multiple locations with environmental variables, and Atmospheric Corrosion Monitoring (ACM), which focuses on localized, time-series data collected through sensors.

More recently, the availability of global-scale geospatial and atmospheric datasets has enabled the enhancement of the map-based approach through the integration of machine learning, allowing the generation of corrosivity maps for large regions such as South America (GEREMIAS, 2024). Despite its effectiveness, this approach is inherently limited by spatial resolution. Atmospheric corrosion is strongly influenced by microclimatic effects—for instance, variations of only a few meters in distance to the coastline can significantly affect corrosivity. Even high-resolution maps, on the order of hundreds of meters, are often insufficient to capture such fine-scale variability, highlighting the need for alternative approaches.

To address this limitation, this work proposes a case study approach, in which user-provided information about a specific location is used to collect relevant variables, perform predictions using machine learning models, and return both corrosion rates and corrosivity classifications. The proposed system, LABICORR, is structured into three layers: the data layer, the modeling layer, and the application layer. The data layer

is responsible for managing data sources and organizing variables and exposure data into structured datasets. The modeling layer focuses on the development of machine learning models for six target variables: deposition of Cl^- and SO_2 , and corrosion rates of Zn , Fe , Al , and Cu . The application layer enables users to execute case studies by providing input data and running the trained models.

This work was developed with the explicit aim of supporting future research in the advancement of atmospheric corrosion prediction systems. Its primary contribution lies in the design and comprehensive documentation of an end-to-end intelligent system for corrosion prediction, capable of making predictions instantly, without field experimentation, only with location coordinates. With the exception of exposure of metals themselves, all stages of the development pipeline are covered, including data acquisition, preprocessing, database construction, model development and validation, as well as the implementation of both development and production environments and a final user-oriented application.

The proposed models extend previous works (BRANDENBURG et al., 2024; GEREMIAS, 2024) by incorporating a case study perspective, new datasets, additional variables, and a broader comparison of machine learning algorithms, including Random Forests, Extra Trees, Support Vector Machine Regressors, Genetic Programming, and Artificial Neural Networks, among others depending on the target variable. All six models (SO_2 , Cl^- , Zn , Fe , Al , and Cu) are developed within a unified modeling framework, ensuring that improvements in preprocessing and analysis are consistently shared across models.

For each target variable, one model was chosen based on r^2 value and significant statistical difference to other models. For the SO_2 model, industrial variables derived from OpenStreetMap data were incorporated, with the Random Forest model achieving the best performance ($r^2 = 0.69$). In the Cl^- model, symbolic regression was introduced, resulting in an interpretable empirical salt decay equation with a high performance ($r^2 = 0.71$), which outperformed consolidated literature models.

For the Zn model, the Support Vector Machine Regressor outperformed ensemble methods, achieving an r^2 of 0.69. For the Fe , Al , and Cu models, Random Forests provided the best results, with r^2 values of 0.62, 0.60, and 0.68, respectively. The higher values compared to the baseline Dose-Response Functions from the literature indicates that using geospatial and atmospheric data through a data-driven approach significantly contribute to improving prediction performance. This suggests that the future of atmospheric corrosion research is moving toward a data-driven approach, utilizing external sources of variables rather than on-site measurements.

It is important to note that the obtained r^2 values, ranging from 0.6 to 0.7,

are moderate. This is expected given the nature of ACE models, which operate in uncontrolled environments influenced by numerous variables and unpredictable factors. Additionally, the use of a robust validation strategy—70/30 random subsampling repeated 100 times—reduces the likelihood of overfitting, resulting in more conservative but reliable performance estimates.

The final application has an interface that allows users to enter data about a location through a formulary, a map component and a ruler tool for distance to coastline. Case studies can be performed multiple locations. The separation between the development and production system allowed for the existence of a robust server-based development version and a portable production version to be run on desktops.

For future work, it was observed that significantly lower r^2 values are obtained for the *Fe* and *Al* models. In contrast, in cases where higher r^2 values were achieved, curve-based models such as Support Vector Regression (SVR) and symbolic regression tended to perform better. This suggests the possible presence of outliers that were not manually identified. Therefore, the adoption of more robust and systematic outlier detection methods is recommended.

Another important recommendation is the development of an updated dataset for SO_2 prediction. The current dataset is largely composed of data from the 1980s and 1990s, which limits its applicability to present-day environmental conditions.

It is also recommended to explore additional reanalysis data sources. For the chloride model, ocean wave reanalysis datasets—such as the Global Ocean Waves Reanalysis (European Union-Copernicus Marine Service, 2019)—have not yet been considered. A comprehensive survey of relevant variables in other atmospheric reanalysis datasets, including ERA5 and MERRA-2, is also recommended.

Another potential line of development is the integration of the map-based approach with the case study approach. As the number of microclimate variables increases, these approaches tend to become incompatible. However, this limitation could be mitigated through the development of high-resolution maps of key microclimate variables, such as elevation, distance to the coastline, surface roughness, and weighted land-use areas. While high-resolution elevation datasets are already available, the remaining variables would need to be constructed—likely using greedy algorithms—requiring further optimization to ensure computational feasibility. In particular, the development of a high-resolution distance-to-coastline dataset would represent a valuable contribution, as current datasets (e.g., *dist2coast*) developed by NASA and NOAA have a relatively coarse spatial resolution of approximately 1 km.

Finally, another unexplored direction is the incorporation of long-term corrosion modeling. At present, the proposed models provide estimates of r_{corr} and corrosivity

classification, leaving the determination of total protective coating thickness to the specialist. However, ISO 9224 provides formulations for cumulative corrosion (total attack), which could be integrated as a baseline or extended modeling component.

REFERENCES

156, I. *ISO 8407:2021 - Corrosion of metals and alloys — Removal of corrosion products from corrosion test specimens*. 2021.

AMS. **Aerodynamic roughness length**. 2019. <https://glossary.ametsoc.org/wiki/Aerodynamic_roughness_length>. Acesso em 2024-05-06.

BAILEY, S. **DISTANCE TO THE NEAREST COAST**. 2024. <<https://oceancolor.gsfc.nasa.gov/resources/docs/distfromcoast/>>. Acesso em: 2024-05-19.

BEZERRA-PONTES, R. et al. Clasificación de la agresividad costera de recife en función del nivel de contaminación por cloruros. **Ingeniería**, Universidad Autónoma de Yucatán, v. 15, n. 1, p. 19–26, 2011.

BRANDENBURG, T. **Aplicação de Dados de Reanálise Climática e Aprendizado de Máquina para Predição de Deposição Seca de Íons de Cloreto**. Dissertação (Graduation's thesis) — Universidade do Estado de Santa Catarina, Joinville, SC-Brazil, June 2024. Available at <<https://example.com/thesis.pdf>>.

BRANDENBURG, T. et al. A data-driven approach with reanalysis and geospatial data for chloride deposition prediction. **Earth Science Informatics**, Springer Science and Business Media LLC, v. 18, n. 1, dez. 2024. ISSN 1865-0481. Disponível em: <<http://dx.doi.org/10.1007/s12145-024-01660-5>>.

BREIMAN, L. Random forests. **Machine Learning**, Springer Science and Business Media LLC, v. 45, n. 1, p. 5–32, 2001. Disponível em: <<https://doi.org/10.1023/a:1010933404324>>.

BRERETON, R. G.; LLOYD, G. R. Support vector machines for classification and regression. **The Analyst**, Royal Society of Chemistry (RSC), v. 135, n. 2, p. 230–267, 2010. ISSN 1364-5528. Disponível em: <<http://dx.doi.org/10.1039/b918972f>>.

BURKOV, A. **The hundred-page machine learning book**. Quebec City, QC, Canada: Andriy Burkov, 2019. v. 1.

CAI, Y. et al. Atmospheric corrosion prediction: a review. **Corrosion Reviews**, Walter de Gruyter GmbH, v. 38, n. 4, p. 299–321, jul. 2020. ISSN 2191-0316. Disponível em: <<http://dx.doi.org/10.1515/corrrev-2019-0100>>.

CHEN, Q. et al. Data-driven atmospheric corrosion prediction model for alloys based on a two-stage machine learning approach. **Process Safety and Environmental Protection**, Elsevier BV, v. 188, p. 1093–1105, ago. 2024. ISSN 0957-5820. Disponível em: <<http://dx.doi.org/10.1016/j.psep.2024.06.028>>.

CHICCO, D.; WARRENS, M. J.; JURMAN, G. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. **PeerJ Computer Science**, v. 7, p. e623, 2021.

CHICO, B. et al. Annual atmospheric corrosion of carbon steel worldwide. an integration of ISOCORRAG, ICP/UNECE and MICAT databases. **Materials**, MDPI AG, v. 10, n. 6, p. 601, maio 2017. Disponível em: <<https://doi.org/10.3390/ma10060601>>.

CHIN, M. et al. Tropospheric aerosol optical thickness from the GOCART model and comparisons with satellite and sun photometer measurements. **Journal of the Atmospheric Sciences**, American Meteorological Society, v. 59, n. 3, p. 461–483, fev. 2002. Disponível em: <[https://doi.org/10.1175/1520-0469\(2002\)059<0461:taotft>2.0.co;2](https://doi.org/10.1175/1520-0469(2002)059<0461:taotft>2.0.co;2)>.

COELHO, L. B. et al. Reviewing machine learning of corrosion prediction in a data-oriented perspective. **npj Materials Degradation**, Nature Publishing Group UK London, v. 6, n. 1, p. 8, 2022.

COLE, I. et al. A mechanistic study of the effect of environmental conditions in china, south-east asia and australia on the corrosion of zinc and steel. **International Symposium on Corrosion and Corrosion Protection**, p. 749–760, 01 2001.

COLE, I. S.; PATERSON, D. A.; GANTHER, W. D. Holistic model for atmospheric corrosion part 1 - theoretical framework for production, transportation and deposition of marine salts. **Corrosion Engineering, Science and Technology**, Informa UK Limited, v. 38, n. 2, p. 129–134, jun. 2003. Disponível em: <<https://doi.org/10.1179/147842203767789203>>.

CORROSION Map. Victoria Rd, The Parade, Birmingham, Sutton Coldfield B72 1SX, United Kingdom: Galvanizers Association, 2024. <<https://galvanizing.org.uk/corrosion/corrosion-map/>>.

CRANMER, M. **Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl**. arXiv, 2023. Disponível em: <<https://arxiv.org/abs/2305.01582>>.

CRIPPA, M. et al. Gridded emissions of air pollutants for the period 1970–2012 within edgar v4.3.2. Copernicus GmbH, 2018. Disponível em: <<http://dx.doi.org/10.5194/essd-2018-31>>.

CRISTOS, M. T. et al. 03-033 – prediction of chloride deposition using machine learning for corrosion assessment; [predicción de la deposición de cloruros con aprendizaje automático para la evaluación de la corrosión]. In: . Asociacion Espanola de Direccion e Ingenieria de Proyectos (AEIPRO), 2025. p. 1155 – 1167. Cited by: 0. Disponível em: <<https://www.scopus.com/pages/publications/105022845501?origin=resultslist>>.

DEFOURNY, P. et al. **Product User Guide and Specification: ICDR Land Cover 2016-2020**. Copenhagen, Denmark, 2021.

DING, X. et al. Random radial basis function kernel-based support vector machine. **Journal of the Franklin Institute**, Elsevier BV, v. 358, n. 18, p. 10121–10140, dez. 2021. ISSN 0016-0032. Disponível em: <<http://dx.doi.org/10.1016/j.jfranklin.2021.10.005>>.

ECMWF. **ERA5: data documentation**. Reading, Reino Unido, 2023.

ESA. **Land Cover CCI Product User Guide Version 2**. Croix du S 2, 1348 Ottignies-Louvain-la-Neuve, Belgium: European Space Agency, 2017.

European Union-Copernicus Marine Service. **Global Ocean Waves Reanalysis WAVERYS**. Mercator Ocean International, 2019. Disponível em: <https://resources.marine.copernicus.eu/product-detail/GLOBAL_MULTIYEAR_WAV_001_032/INFORMATION>.

FASULLO, J. et al. Atmospheric reanalysis: Overview & comparison tables. **NATIONAL CENTER FOR ATMOSPHERIC RESEARCH**, 2016.

FELIU, S.; MORCILLO, M.; CHICO, B. Effect of distance from sea on atmospheric corrosion rate. **CORROSION**, Association for Materials Protection and Performance (AMPP), v. 55, n. 9, p. 883–891, set. 1999. Disponível em: <<https://doi.org/10.5006/1.3284045>>.

FLOORS, R. et al. Satellite-based estimation of roughness lengths and displacement heights for wind resource modelling. **Wind Energy Science**, Copernicus GmbH, v. 6, n. 6, p. 1379–1400, nov. 2021. ISSN 2366-7451. Disponível em: <<http://dx.doi.org/10.5194/wes-6-1379-2021>>.

GAVRYUSHINA, M. A.; MARSHAKOV, A. I.; PANCHENKO, Y. M. Application of the random forest algorithm of corrosion losses of aluminum for the first year of exposure in various regions of the world. **Protection of Metals and Physical Chemistry of Surfaces**, Pleiades Publishing Ltd, v. 59, n. 1, p. 85–95, fev. 2023. ISSN 2070-206X. Disponível em: <<http://dx.doi.org/10.1134/S2070205123700259>>.

GELARO, R. et al. The modern-era retrospective analysis for research and applications, version 2 (merra-2). **Journal of climate**, v. 30, n. 14, p. 5419–5454, 2017.

GENTIL, V. **Corrosão. 6ª edição**. Rio de Janeiro, Brasil: LTC, 2011.

GEREMIAS, V. M. **Machine Learning for predicting atmospheric corrosion rate in metallic materials using reanalysis data**. Dissertação (Mestrado) — Universidade do Estado de Santa Catarina, 2024.

GEREMIAS, V. M. et al. Predição de corrosão atmosférica em materiais metálicos utilizando aprendizado de máquina. In: **Anais do XVI Congresso Brasileiro de Inteligência Computacional**. SBIC, 2024. (CBIC 2023). Disponível em: <<http://dx.doi.org/10.21528/CBIC2023-101>>.

GEREMIAS, V. M. et al. Atmospheric corrosion prediction in metallic materials using machine learning. In: _____. **Hybrid Intelligent Systems**. Springer Nature Switzerland, 2025. p. 42–52. ISBN 9783031789342. Disponível em: <http://dx.doi.org/10.1007/978-3-031-78934-2_5>.

GEREMIAS, V. M. et al. Predicting sulfur dioxide deposition in microclimates using machine learning and reanalysis data. **Modeling Earth Systems and Environment**, Springer Science and Business Media LLC, v. 11, n. 2, fev. 2025. ISSN 2363-6211. Disponível em: <<http://dx.doi.org/10.1007/s40808-025-02327-w>>.

GIRÃO, S. M. B. **Estimativa de sulfatação total na atmosfera em diferentes regiões do Ceará**. Dissertação (Mestrado) — Universidade Federal do Ceará, 2007.

Google. **Industrial site of Cubatão through Google Earth**. 2025. Disponível em: <<https://earth.google.com/web/>>.

GUSTAFSSON, M. E. Raised levels of marine aerosol deposition owing to increased storm frequency; a cause of forest decline in southern sweden? **Agricultural and Forest Meteorology**, Elsevier BV, v. 84, n. 1–2, p. 169–177, mar. 1997. ISSN 0168-1923. Disponível em: <[http://dx.doi.org/10.1016/S0168-1923\(96\)02384-2](http://dx.doi.org/10.1016/S0168-1923(96)02384-2)>.

HERSBACH, H. et al. The ERA5 global reanalysis. **Quarterly Journal of the Royal Meteorological Society**, Wiley, v. 146, n. 730, p. 1999–2049, jun. 2020. Disponível em: <<https://doi.org/10.1002/qj.3803>>.

HERSBACH, H. et al. Era5 hourly data on single levels from 1940 to present. **Copernicus climate change service (c3s) climate data store (cds)**, ECMWF Reading, UK, 2023.

INMET. **Catálogo de Estações Automáticas**. Brasília, Brasil, 2023.

INMET. **Catálogo de Estações Convencionais**. Brasília, Brasil, 2023.

ISO/8565. **Metals and alloys — Atmospheric corrosion testing — General requirements**. 2011.

ISO/9223. **Corrosion of metals and alloys - Corrosivity of atmospheres - Classification, determination and estimation**. 2012.

ISO/9224. **Corrosion of metals and alloys — Corrosivity of atmospheres — Guiding values for the corrosivity categories**. 2012.

ISO/9225. **Corrosion of metals and alloys — Corrosivity of atmospheres — Measurement of environmental parameters affecting corrosivity of atmospheres**. 2012.

ISO/9226. **Corrosion of metals and alloys — Corrosivity of atmospheres — Determination of corrosion rate of standard specimens for the evaluation of corrosivity**. 2012.

JAMES, G. et al. **An introduction to statistical learning**. Nova Iorque, Estados Unidos: Springer, 2013. v. 112.

KINGSFORD, C.; SALZBERG, S. L. What are decision trees? **Nature Biotechnology**, Springer Science and Business Media LLC, v. 26, n. 9, p. 1011–1013, set. 2008. Disponível em: <<https://doi.org/10.1038/nbt0908-1011>>.

KNOTKOVA, D.; KREISLOVA, K.; DEAN, S. W. (Ed.). **ISOCORRAG International Atmospheric Exposure Program: Summary of Results**. West Conshohocken, Estados Unidos: ASTM International, 2010. Disponível em: <<https://doi.org/10.1520/ds71-eb>>.

KOCH, G. H. et al. **Corrosion cost and preventive strategies in the United States**. McLean, Estados Unidos, 2002.

KOUREMADAS, G. et al. Satellite sensed data-dose response functions: A totally new approach for estimating materials' deterioration from space. **Remote Sensing**, MDPI AG, v. 15, n. 12, p. 3194, jun. 2023. Disponível em: <<https://doi.org/10.3390/rs15123194>>.

KOZA, J. R.; POLI, R. Genetic programming. In: _____. **Search Methodologies**. Springer US, 2005. p. 127–164. ISBN 9780387234601. Disponível em: <http://dx.doi.org/10.1007/0-387-28356-0_5>.

KUANG, J.; LONG, Z. Prediction model for corrosion rate of low-alloy steels under atmospheric conditions using machine learning algorithms. **International Journal of Minerals, Metallurgy and Materials**, Springer Science and Business Media LLC, v. 31, n. 2, p. 337–350, fev. 2024. ISSN 1869-103X. Disponível em: <<http://dx.doi.org/10.1007/s12613-023-2679-5>>.

KUCERA, V. et al. Un/ece icp materials dose-response functions for the multi-pollutant situation. **Water, Air, & Soil Pollution: Focus**, Springer Science and Business Media LLC, v. 7, n. 1–3, p. 249–258, jan. 2007. ISSN 1573-2940. Disponível em: <<http://dx.doi.org/10.1007/s11267-006-9080-z>>.

KUCHAŘ, M. et al. Corrosion prediction for preventive protection of aircraft heritage. **Heritage Science**, Springer Science and Business Media LLC, v. 12, n. 1, mar. 2024. ISSN 2050-7445. Disponível em: <<http://dx.doi.org/10.1186/s40494-024-01212-6>>.

LEYGRAF, C. et al. **Atmospheric corrosion**. Hoboken, Estados Unidos: John Wiley & Sons, 2016.

LIBERATI, A. et al. The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. **Annals of internal medicine**, American College of Physicians, v. 151, n. 4, p. W–65, 2009.

LINACRE, E.; GEERTS, B. **Resources in Atmospheric Sciences: Roughness Length**. 1997. <<http://www-das.uwyo.edu/~geerts/cwx/notes/chap14/roughness.html>>. Acesso em: 2024-05-06.

LUNA, A. M. de et al. Mapeamento da corrosividade atmosférica do estado do ceará. **Intercorr**, ABRACO, 2008.

MCCAFFERTY, E. **Introduction to corrosion science**. Nova Íorque, Estados Unidos: Springer Science & Business Media, 2010.

MEIRA, G. et al. Modelling sea-salt transport and deposition in marine atmosphere zone – a tool for corrosion studies. **Corrosion Science**, Elsevier BV, v. 50, n. 9, p. 2724–2731, set. 2008. Disponível em: <<https://doi.org/10.1016/j.corsci.2008.06.028>>.

MENG, J. et al. Data mining to atmospheric corrosion process based on evidence fusion. **Materials**, MDPI AG, v. 14, n. 22, p. 6954, nov. 2021. ISSN 1996-1944. Disponível em: <<http://dx.doi.org/10.3390/ma14226954>>.

MIKHAILOV, A. A.; TIDBLAD, J.; KUCERA, V. The classification system of ISO 9223 standard and the dose–response functions assessing the corrosivity of outdoor atmospheres. **Protection of Metals**, Springer Science and Business Media LLC, v. 40, n. 6, p. 541–550, nov. 2004. Disponível em: <<https://doi.org/10.1023/b:prom.0000049517.14101.68>>.

MONAHAN, E. C.; SPIEL, D. E.; DAVIDSON, K. L. A model of marine aerosol generation via whitecaps and wave disruption. In: **Oceanographic Sciences Library**. Springer Netherlands, 1986. p. 167–174. Disponível em: <https://doi.org/10.1007/978-94-009-4668-2_16>.

MORALES, J. et al. Atmospheric corrosion in subtropical areas: influences of time of wetness and deficiency of the iso 9223 norm. **Corrosion Science**, Elsevier BV, v. 47, n. 8, p. 2005–2019, ago. 2005. ISSN 0010-938X. Disponível em: <<http://dx.doi.org/10.1016/j.corsci.2004.09.005>>.

MORCILLO, M. Atmospheric corrosion in ibero-america: The MICAT project. In: **Atmospheric Corrosion**. ASTM International, 1995. p. 257–257–19. Disponível em: <<https://doi.org/10.1520/stp14924s>>.

MORCILLO, M. et al. **Corrosion y Protección de Metales en las Atmósferas de Iberoamérica. Parte I — Mapas de Iberoamérica de Corrosividad Atmosférica (Proyecto MICAT, XV.1/CYTED)**. Madrid, Espanha: CYTED, 1998. ISBN 84-930448-1-4.

MORCILLO, M. et al. Looking back on contributions in the field of atmospheric corrosion offered by the MICAT ibero-american testing network. **International Journal of Corrosion**, Hindawi Limited, v. 2012, p. 1–24, 2012. Disponível em: <<https://doi.org/10.1155/2012/824365>>.

MUTHUKRISHNAN, R. et al. Machine learning approach to investigate high temperature corrosion of critical infrastructure materials. **High Temperature Corrosion of Materials**, Springer Science and Business Media LLC, v. 101, n. S1, p. 309–331, set. 2024. ISSN 2731-8400. Disponível em: <<http://dx.doi.org/10.1007/s11085-024-10312-4>>.

NOAA. **What is the difference between land cover and land use?** [S.l.]: National Ocean Service website, 2024. <<https://oceanservice.noaa.gov/facts/lclu.html>>. Data de Acesso: 30-04-2024.

OBPG, N. O. B. P. G. **Dataset: Distance to Nearest Coastline/0.01-Degree Grid/Distance to Nearest Coastline: 0.01-Degree Grid**. Honolulu, Estados Unidos: Pacific Islands Ocean Observing System (PacIOOS), 2023.

OKURA, T. et al. Application of machine learning for data with an atmospheric corrosion monitoring sensor based on strain measurements. **Metals**, MDPI AG, v. 12, n. 7, p. 1179, jul. 2022. ISSN 2075-4701. Disponível em: <<http://dx.doi.org/10.3390/met12071179>>.

OpenStreetMap contributors. **Planet dump retrieved from <https://planet.osm.org>**. 2017. <<https://www.openstreetmap.org>>.

PANCHENKO, Y. M.; SHUVAKHINA, L. Atmospheric corrosion of metals in the far east. **Zashch. Met.**, v. 18, n. 4, p. 575–582, 1982.

PANNONI, F. D. **Manual De Construção do Aço: Projeto e Durabilidade**. Rio de Janeiro, Brasil: Instituto Aço Brasil, 2017.

PARPINELLI, R. S. et al. Aprendizado de máquina para predição da taxa de corrosão em aços. **ArcelorMittal Brasil**, 2023.

PEI, Z. et al. Towards understanding and prediction of atmospheric corrosion of an fe/cu corrosion sensor via machine learning. **Corrosion Science**, Elsevier BV, v. 170, p. 108697, jul. 2020. Disponível em: <<https://doi.org/10.1016/j.corsci.2020.108697>>.

PINTOS, S. et al. Artificial neural network modeling of atmospheric corrosion in the micat project. **Corrosion science**, Elsevier, v. 42, n. 1, p. 35–52, 2000.

PORTELLA, M. et al. Atmospheric corrosion rates of copper, galvanized steel, carbon steel and aluminum in the metropolitan region of salvador, BA, northeast brazil. **Procedia Engineering**, Elsevier BV, v. 42, p. 171–185, 2012. Disponível em: <<https://doi.org/10.1016/j.proeng.2012.07.408>>.

RANDLES, C. A. et al. The MERRA-2 aerosol reanalysis, 1980 onward. part i: System description and data assimilation evaluation. **Journal of Climate**, American Meteorological Society, v. 30, n. 17, p. 6823–6850, set. 2017. Disponível em: <<https://doi.org/10.1175/jcli-d-16-0609.1>>.

REFAEILZADEH, P.; TANG, L.; LIU, H. Cross-validation. In: **Encyclopedia of Database Systems**. [S.l.]: Springer, 2009. p. 532–538.

ROBERGE, P.; KLASSEN, R.; HABERECHE, P. Atmospheric corrosivity modeling—a review. **Materials & Design**, Elsevier, v. 23, n. 3, p. 321–330, 2002.

RODRÍGUEZ, E. B. Application of machine learning for the prediction of atmospheric corrosion in the metropolitan area of mexico city. **ITEGAM- Journal of Engineering and Technology for Industrial Applications (ITEGAM-JETIA)**, GN1 Sistemas e Publicacoes Ltd., v. 10, n. 48, 2024. ISSN 2447-0228. Disponível em: <<http://dx.doi.org/10.5935/jetia.v10i48.1089>>.

SAGI, O.; ROKACH, L. Ensemble learning: A survey. **WIREs Data Mining and Knowledge Discovery**, Wiley, v. 8, n. 4, fev. 2018. Disponível em: <<https://doi.org/10.1002/widm.1249>>.

SANDWELL, D.; SMITH, W. H.; AND, J. J. B. **SRTM30+ Global 1-km Digital Elevation Model (DEM): Version 11: Land Surface**. Honolulu, Estados Unidos: Pacific Islands Ocean Observing System (PacIOOS), 2014.

SANTA, A. C. et al. Atmospheric corrosion maps as a tool for designing and maintaining building materials: A review. **Heliyon**, Elsevier BV, v. 8, n. 9, p. e10438, set. 2022. ISSN 2405-8440. Disponível em: <<http://dx.doi.org/10.1016/j.heliyon.2022.e10438>>.

SANTANA, J. J. et al. Atmosferic corrosion in the canary islands. **Repositório institucional de la Universidad de La Laguna**, Universidad de La Laguna, 2019.

SANTANA, J. J. et al. Shortcomings of international standard iso 9223 for the classification, determination, and estimation of atmosphere corrosivities in subtropical archipelagic conditions—the case of the canary islands (spain). **Metals**, MDPI AG, v. 9, n. 10, p. 1105, out. 2019. ISSN 2075-4701. Disponível em: <<http://dx.doi.org/10.3390/met9101105>>.

SCHMITT, G. et al. Global needs for knowledge dissemination, research, and development in materials deterioration and corrosion control. **World Corrosion Organization**, Citeseer, v. 38, p. 14, 2009.

SCHWEITZER, P. A. **Fundamentals of Metallic Corrosion: Atmospheric and Media Corrosion of Metals**. Boca Raton, Estados Unidos: CRC Press, 2006.

SCIKIT-LEARN. **Support Vector Regression (SVR) using linear and non-linear kernels**. 2026. Disponível em: <scikit-learn.org/stable/auto_examples/svm/plot_svm_regression.html>.

SEBRAO, M. Z.; ORDINE, A. P. Uma metodologia alternativa para determinar a agressividade atmosférica. **ABRACO**, 2010.

SICA, Y. C. **Mapeamento da Corrosividade Atmosférica de São Luís-MA e a Correlação das variáveis ambientais que influenciam na degradação dos materiais metálicos**. Dissertação (Mestrado) — Universidade Federal do Paraná, Curitiba, 2006.

SICA, Y. C. et al. Atmospheric corrosion performance of carbon steel, galvanized steel, aluminum and copper in the north brazilian coast. **Journal of the Brazilian Chemical Society**, FapUNIFESP (SciELO), v. 18, n. 1, p. 153–166, 2007. Disponível em: <<https://doi.org/10.1590/s0103-50532007000100017>>.

SONG, X. et al. Corrosion behavior and corrosion prediction of carbon steel under dynamic atmospheric corrosion environment in harbin. **Journal of Materials Engineering and Performance**, Springer Science and Business Media LLC, jun. 2024. ISSN 1544-1024. Disponível em: <<http://dx.doi.org/10.1007/s11665-024-09563-8>>.

SUMMITT, R.; FINK, F. T. Pacer lime: An environmental corrosion severity classification system. part i. **University of Michigan State**, v. 34, p. 35, 1980.

TERRADOS-CRISTOS, M. et al. Corrosion risk assessment in coastal environments using machine learning-based predictive models. **Sensors**, Multidisciplinary Digital Publishing Institute (MDPI), v. 25, n. 13, 2025. Cited by: 2; All Open Access; Gold Open Access; Green Open Access. Disponível em: <<https://www.scopus.com/pages/publications/105010319239?origin=resultslist>>.

TERRADOS-CRISTOS, M. et al. Corrosion prediction of weathered galvanised structures using machine learning techniques. **Materials**, MDPI AG, v. 14, n. 14, p. 3906, jul. 2021. ISSN 1996-1944. Disponível em: <<http://dx.doi.org/10.3390/ma14143906>>.

THANUSH, A. A. et al. Atmospheric corrosion rate prediction of low-alloy steel using machine learning models. **IOP Conference Series: Materials Science and Engineering**, IOP Publishing, v. 1248, n. 1, p. 012050, jul. 2022. ISSN 1757-899X. Disponível em: <<http://dx.doi.org/10.1088/1757-899X/1248/1/012050>>.

TIDBLAD, J. Atmospheric corrosion of metals in 2010–2039 and 2070–2099. **Atmospheric Environment**, Elsevier BV, v. 55, p. 1–6, ago. 2012. ISSN 1352-2310. Disponível em: <<http://dx.doi.org/10.1016/j.atmosenv.2012.02.081>>.

TRAN, N.-L. et al. A machine learning-based model for predicting atmospheric corrosion rate of carbon steel. **Advances in Materials Science and Engineering**, Wiley,

v. 2021, n. 1, jan. 2021. ISSN 1687-8442. Disponível em: <<http://dx.doi.org/10.1155/2021/6967550>>.

UNIDATA. **NetCDF**. 2025. <https://www.unidata.ucar.edu/software/netcdf>.

VANGRUNDERBEEK, V. et al. Exploring the potential of transfer learning in extrapolating accelerated corrosion test data for long-term atmospheric corrosion forecasting. **Corrosion Science**, Elsevier BV, v. 225, p. 111619, dez. 2023. ISSN 0010-938X. Disponível em: <<http://dx.doi.org/10.1016/j.corsci.2023.111619>>.

VERA, R. et al. Mapa de corrosión atmosférica de Chile: resultados después de un año de exposición. **Revista de la construcción**, Pontificia Universidad Católica de Chile, v. 11, n. 2, p. 61–72, ago. 2012. ISSN 0718-915X. Disponível em: <<http://dx.doi.org/10.4067/S0718-915X2012000200007>>.

VERNON, W. First (experimental) report to the atmospheric corrosion research committee (of the British non-ferrous metals research association). **Transactions of the Faraday Society**, Royal Society of Chemistry, v. 19, n. March, p. 839–845, 1924.

VITALI, M. Efeito do distanciamento ao mar da contaminação do concreto por cloretos. **Universidade Federal de Santa Catarina**, 2013.

WAGER, S. Comments on: A random forest guided tour. **TEST**, Springer Science and Business Media LLC, v. 25, n. 2, p. 261–263, abr. 2016. ISSN 1863-8260. Disponível em: <<http://dx.doi.org/10.1007/s11749-016-0482-6>>.

WANG, B. et al. Corrosion big-data driven continuous observation of low alloy steel rust layer evolution and mining of influence rules of atmospheric environment interaction. **Corrosion Science**, Elsevier Ltd, v. 255, 2025. Cited by: 5. Disponível em: <<https://www.scopus.com/pages/publications/105008685138?origin=resultslist>>.

WATT, J. et al. **The Effects of Air Pollution on Cultural Heritage**. Springer US, 2009. ISBN 9780387848938. Disponível em: <<http://dx.doi.org/10.1007/978-0-387-84893-8>>.

WINDPRO. **Glob Cover**. 2014. Acesso em 2024-08-05.

WMO. **Guide to Instruments and Methods of Observation (WMO-No. 8), Volume I – Measurement of Meteorological Variables**. World Meteorological Organization (WMO), 2024. Disponível em: <<http://dx.doi.org/10.59327/WMO/CIMO/1>>.

YAN, L.; DIAO, Y.; GAO, K. Analysis of environmental factors affecting the atmospheric corrosion rate of low-alloy steel using random forest-based models. **Materials**, MDPI AG, v. 13, n. 15, p. 3266, jul. 2020. ISSN 1996-1944. Disponível em: <<http://dx.doi.org/10.3390/ma13153266>>.

YAN, L. et al. Corrosion rate prediction and influencing factors evaluation of low-alloy steels in marine atmosphere using machine learning approach. **Science and Technology of Advanced Materials**, Informa UK Limited, v. 21, n. 1, p. 359–370, jan. 2020. ISSN 1878-5514. Disponível em: <<http://dx.doi.org/10.1080/14686996.2020.1746196>>.

YANG, G. et al. Innovative dynamic evaluation and classification method for marine atmospheric corrosion based on corrosion sensors and machine learning. **Materials Today Communications**, Elsevier Ltd, v. 42, 2025. Cited by: 8. Disponível em: <<https://www.scopus.com/pages/publications/85214577813?origin=resultslist>>.

YANG, X. et al. Data-mining and atmospheric corrosion resistance evaluation of sn- and sb-additional low alloy steel based on big data technology. **International Journal of Minerals, Metallurgy and Materials**, Springer Science and Business Media LLC, v. 29, n. 4, p. 825–835, abr. 2022. ISSN 1869-103X. Disponível em: <<http://dx.doi.org/10.1007/s12613-022-2457-9>>.

ZHANG, H. et al. Comparative study on optimization algorithms for atmospheric corrosion map of q235 steel in hunan, china. **Process Safety and Environmental Protection**, Elsevier BV, v. 184, p. 105–115, abr. 2024. ISSN 0957-5820. Disponível em: <<http://dx.doi.org/10.1016/j.psep.2024.01.107>>.

ZHI, Y. et al. Prediction and knowledge mining of outdoor atmospheric corrosion rates of low alloy steels based on the random forests approach. **Metals**, MDPI AG, v. 9, n. 3, p. 383, mar. 2019. ISSN 2075-4701. Disponível em: <<http://dx.doi.org/10.3390/met9030383>>.

ZHI, Y. et al. Improving atmospheric corrosion prediction through key environmental factor identification by random forest-based model. **Corrosion Science**, Elsevier BV, v. 178, p. 109084, jan. 2021. ISSN 0010-938X. Disponível em: <<http://dx.doi.org/10.1016/j.corsci.2020.109084>>.

ZHI, Y.; YANG, T.; FU, D. An improved deep forest model for forecast the outdoor atmospheric corrosion rate of low-alloy steels. **Journal of Materials Science & Technology**, Elsevier BV, v. 49, p. 202–210, jul. 2020. ISSN 1005-0302. Disponível em: <<http://dx.doi.org/10.1016/j.jmst.2020.01.044>>.

APPENDIX A – ARTIFICIAL INTELLIGENCE USE DECLARATION

I, Thiago Brandenburg, declare that during the elaboration of this work, I used the Artificial Intelligence tools described below with the purpose of reviewing and improving the writing, correcting grammatical errors, and enhancing the fluency of the text. After use, all content was reviewed and edited by me, and I assume full responsibility for the final content of this publication.

- **ChatGPT** (OpenAI, version GPT-4o): used for grammatical review, rewriting suggestions, and verification of textual coherence.
- **Claude** (Anthropic, version Claude Sonnet 4.6): used for text revision, improvement of fluency and clarity of writing.
- **Grammarly** (Grammarly Inc.): used for automated spell and grammar checking throughout the writing of this work.

The use of these tools was restricted to assistance in the linguistic revision of the text. The ideas, arguments, analyses, and conclusions presented are my sole responsibility.



JOINVILLE
CENTRO DE CIÊNCIAS
TECNOLÓGICAS

UNIVERSIDADE DO ESTADO DE SANTA CATARINA – UDESC
BIBLIOTECA UNIVERSITÁRIA
REPOSITÓRIO INSTITUCIONAL

CENTRO DE CIÊNCIAS TECNOLÓGICAS – CCT

ATESTADO DE VERSÃO FINAL

Eu, Rafael Stubs Parpinelli, professor do curso de Mestrado em Computação Aplicada, declaro que esta é a versão final aprovada pela comissão julgadora da dissertação intitulada: **“A Data-driven Approach for Atmospheric Corrosion Prediction”** de autoria do acadêmico Thiago Brandenburg.

Joinville, 25 de Junho de 2026.

Assinatura digital do orientador:

Documento assinado digitalmente
 **RAFAEL STUBS PARPINELLI**
Data: 25/06/2026 13:33:31-0300
Verifique em <https://validar.iti.gov.br>

RAFAEL STUBS PARPINELLI

***PRAZO MÁXIMO PARA SUBMISSÃO DA VERSÃO FINAL NO RI:
ATÉ 60 DIAS APÓS A DEFESA**