POLITECNICO MILANO 1863

School of Civil and Environmental Engineering
MSc in Environmental and Land Planning Engineering

Improving estimates of Greenhouse gas emissions from large dams through Machine Learning

Giorgia Micacchi

Advisor: Prof. Andrea Castelletti
Co-advisor: Prof. Rafael Schmitt - Ing. Bruno Invernizzi
Academic Year: 2024 – 2025

Abstract

Hydropower currently represents the main source of renewable energy worldwide and, driven by the growing energy demand and the need to accelerate the transition towards clean energy sources, is expected to expand further, with more than 3,700 new projects currently under development. For years it has been considered a near zero-emission technology, but it is now evident that hydropower reservoirs can constitute a significant source of biogenic greenhouse gases. The primary cause of these emissions is the high load of organic matter that, following impoundment, accumulates in the reservoir waters and, through decomposition, is transformed into CO_2 and CH_4.

Due to the complexity of the mechanisms regulating the emission processes and the multitude of climatic, biochemical, and environmental factors involved—combined with the limited availability of flux measurements—it is still not possible to provide reliable global estimates or sufficiently accurate predictive models.

Our study proposes a Machine Learning approach to the modeling of emissions, with the aim of developing a flexible and globally applicable tool. The reference model is G-res, which is currently considered the most advanced standard but, without simplifications, is hardly applicable in large-scale analyses. To achieve our goal, we developed two models: the first employs the G-res predictors, in order to compare the performance of an ML algorithm against a linear model, while the second relies on a new set of variables, consisting exclusively of information directly or indirectly related to emission processes and available in global datasets.

If the ML approach proves comparable to the G-res model in terms of accuracy, the results highlight that the use of easily accessible variables significantly improves the inclusiveness of the model, providing it with greater robustness in large-scale applications.

We subsequently applied our model to the analysis of emissions from European reservoirs, with particular focus on hydropower facilities. The study first demonstrates the ability of our models to include almost the entirety of artificial reservoirs in the analysis, substantially expanding the scope of application compared to existing approaches and providing a more comprehensive assessment of emissions. The results show that biogenic

emissions contribute significantly to the carbon intensity of hydropower: when added to those generated throughout the entire dam life cycle, our model estimates an intensity of 46 gCO_{2eq} kWh^{-1}, about twice the expected value.

This study represents a first step towards the development of an innovative tool for energy planning, aimed at ensuring that the transition to new energy sources takes place along a truly sustainable pathway.

Keywords: Hydropower reservoirs, GHG Emission, Machine Learning

Abstract in lingua italiana

L'idroelettrico costituisce oggi la principale fonte di energia rinnovabile a livello globale e, spinto dalla crescente domanda energetica e dalla necessità di accelerare la transizione verso fonti pulite, è destinato a una ulteriore espansione, con oltre 3.700 nuovi progetti attualmente in fase di sviluppo. Se per anni è stato considerato una tecnologia a emissioni quasi nulle, è ormai evidente che i bacini idroelettrici possono costituire una fonte significativa di gas serra di origine biogenica. La principale causa di tali emissioni è l'elevato carico di sostanza organica che, a seguito dell'invaso, si accumula nelle acque del serbatoio e, degradandosi, viene trasformata in CO_2 e CH_4.

A causa della complessità dei meccanismi che regolano i processi emissivi e della molteplicità dei fattori climatici, biochimici e ambientali coinvolti, complice anche la limitata disponibilità di misurazioni dei flussi, non è ancora possibile disporre di stime globali affidabili, né di modelli predittivi sufficientemente accurati.

Il nostro studio propone un approccio di tipo Machine Learning alla modellizzazione delle emissioni, con l'obiettivo di sviluppare uno strumento flessibile e globalmente applicabile. Il modello di riferimento è G-res, che ad oggi è considerato lo standard più avanzato, ma che, senza approssimazioni, è poco applicabile in analisi ad ampia scala. Per raggiungere il nostro obiettivo abbiamo sviluppato due modelli: il primo utilizza i predittori di G-res, in modo da confrontare le prestazioni di un algoritmo ML rispetto ad un modello lineare, mentre il secondo si basa su un nuovo insieme di variabili, costituito esclusivamente da informazioni direttamente o indirettamente legate ai processi emissivi contenute in dataset globali.

Se l'approccio ML si dimostra comparabile al modello G-res a livello di accuratezza, i risultati evidenziano che l'uso di variabili facilmente accessibili migliora sensibilmente l'inclusività del modello, conferendogli maggiore robustezza nelle applicazioni ad ampia scala.

Abbiamo successivamente applicato il nostro modello all'analisi delle emissioni dei serbatoi europei, con particolare attenzione a quelli idroelettrici. Lo studio evidenzia innanzitutto la capacità dei nostri modelli di includere nell'analisi la quasi totalità dei bacini artificiali,

ampliando in modo sostanziale il campo di applicazione rispetto agli approcci esistenti e fornendo una valutazione più completa delle emissioni.
I risultati mostrano che le emissioni biogeniche contribuiscono in misura significativa all'intensità carbonica dell'idroelettrico: sommate a quelle generate lungo l'intero ciclo di vita della diga, il nostro modello stima un'intensità pari a 46 gCO_{2eq} kWh^{-1}, circa il doppio rispetto a quanto atteso.

Questo studio rappresenta un primo passo nella realizzazione di uno strumento innovativo per la pianificazione energetica, volto ad assicurare che la transizione verso nuove fonti di energia si realizzi lungo un percorso autenticamente sostenibile.

Parole chiave: Bacino Idroelettrico, Emissioni, Machine Learning

Contents

1 | Introduction

In recent decades, the steadily increasing demand for energy, together with the ongoing climate change, has placed the energy transition towards renewable sources at the centre of the global debate. In 2024, more than 40% of global electricity generation was derived from renewable sources, with a relevant contribution coming from hydroelectric, 14.3% (Ember, 2025).
Among clean energy sources, in fact, hydropower power plays a pivotal role. Thanks to its flexibility and dispatchability, it has historically been the renewable technology with the largest installed capacity (IRENA), a landscape that is expected to keep growing. In fact, according to the International Hydropower Association (IHA), global hydropower capacity will need to double to align with net-zero targets by 2050. This central role, today and in the decades ahead, highlights the urgency of assessing hydropower's environmental footprint, and in particular its actual carbon intensity. As government design policies to decarbonize their energy sectors, every relevant source of emissions must be accounted for. Only by doing so can policymakers develop reliable strategies to reach carbon neutrality and avoid serious setbacks in the future.

In the recent past, hydropower generation was considered near carbon neutrality. The energy is produced by converting the kinetic energy of flowing water, without any direct release of greenhouse gases (Li and He, 2022).
However, hydropower reservoirs, as all inland aquatic systems, represent dynamic environments in which carbon input, received from surrounding lands, is actively degraded and converted. A portion of this carbon is buried in sediments, and another fraction is emitted to the atmosphere as greenhouse gases (Prairie et al., 2018).

Every time a large dam is built and a reservoir is created—regardless of its intended purpose—the carbon dynamics of the watershed are altered (Forsberg and Dunne, 2017). Compared to natural lakes, artificial reservoir waters are characterized by a higher carbon load. Vegetation and organic matter stored in the soil prior to impoundment becomes the main source of biodegradable carbon, and consequently, of GHG. CO_2 is primarily produced by the aerobic respiration of available organic matter, whereas CH_4 is produced

under anoxic conditions through methanogenesis. Once supersaturation is reached, the excess is released into the atmosphere through multiple pathways, primarily CO_2 and CH_4 diffusion and CH_4 ebullition (Hertwich, 2013).

Additionally, artificial reservoirs deeply affect the hydrology, climate, ecology, and biodiversity of local and downstream regions, causing several environmental and social concerns. Damming a river alters both the natural flow and the sediment cycle (Li and He, 2022). This affects the ecosystem of reservoirs, favoring stratification and eutrophication. The system alterations resulting from the impoundment, together with dam operations, favor the GHG production and release (Yang and Wang, 2014). In particular, the longer water residence times and stratification promote the development of anoxic, sediment-rich environments where methanogenesis occurs (Barros et al., 2011). Furthermore, hydropower reservoirs experience frequent depth variations due to dam operations, which favors the transport of CH_4 bubbles through the water column. In contrast to natural lakes and other artificial reservoirs, hydropower dam operations also cause CH_4 degassing emissions, which occur when water passes through the turbines.

Recent studies highlight the magnitude of biogenic emissions from large dams (Barros et al., 2011; Hertwich, 2010; Scherer and Pfister, 2016). When these emissions are taken into account, the carbon intensity of the hydroelectric power becomes comparable to that of other energy sources, such as biomass combustion, and in some cases even exceeds that of thermal power plants (Scherer and Pfister, 2016).
Estimates suggest that, on a global scale, the GHG emissions from all reservoirs range from 0.61–2.49 Pg CO_{2eq} yr^{-1} as CH_4 and from 0.135–0.99 Pg CO_{2eq} yr^{-1} as CO_2 (Yan et al.). Considering only the emissions originating from hydropower reservoirs, they account for more of 1% of the total anthropogenic carbon footprint (Li and He, 2022).
These findings highlight the urgent need to quantify greenhouse gas emissions from artificial reservoirs created by large dams, regardless of their primary function. Within this broader context, the hydropower sector deserves particular attention: given its central role in today's energy mix and its expected expansion to meet net-zero goals, understanding the true global carbon footprint of hydropower reservoirs is essential.

Firsts global estimates of GHG emissions from reservoirs adopted a statistical method, multiplying the average regional carbon fluxes by the total surface area of reservoirs (Li and He, 2022). While useful as a starting point, this method soon revealed its limitations. Greenhouse gas emissions from reservoirs are not uniform but arise from a complex interplay of physicochemical, meteorological, and reservoir-specific characteristics. As a result, fluxes vary widely—not only across latitudinal gradients but also at much finer, local scales (Almeida and Figueiredo, 2016).

To date, reservoir carbon fluxes is modeled through empirical approaches (Scherer and Pfister (2016), Barros et al. (2011)). These methods aim to reproduce the complex processes controlling CO_2 and CH_4 emissions, considering the reservoir characteristics and the environmental, climatic, and biological variables involved.
Since interest in reservoirs emissions has arisen, research has focused on identifying their key drivers, although discussion remains ongoing. Among reservoir features, studies agree on the importance of area, depth, and age (Barros et al. (2011), Yang et al. (2014)). Temperature, precipitation, wind speed, and cumulative solar radiation also influence both the GHG production and emission processes (Yang et al., 2014). Particularly important is the role of temperature, which enhances aerobic and anaerobic decomposition and, consequently, high GHG fluxes (Hou et al., 2013).
The availability of organic matter remains the main cause of GHG fluxes. As a consequence, including the organic carbon content of the reservoir waters, as well as their level of eutrophication and trophic status, significantly improves the accuracy of emission estimates (Deemer et al., 2016). Over the years, these conditions have been represented by concentrations of chlorophyll, phosphorus, and dissolved organic carbon (Ion and Ene, 2021). Despite their importance, information on these variables remains scarce and, consequently, their use is unfeasible in global applications.

One of the major challenges in modeling reservoir GHG fluxes lies in the scarcity of reliable data, concerning not only the key drivers of emissions but also the direct measurements themselves (Deemer et al., 2016). This limitation makes it difficult to capture the underlying processes with accuracy and, at the same time, to design models that can be easily scaled up for global assessments.
At present, available GHG flux observations remain limited. The most comprehensive dataset is the G-res Tool set, which provides information on 223 globally distributed reservoirs. Nonetheless, like all available datasets, it is affected by important limitations. In general, the coverage of emission pathways across the available datasets is not uniform, with the lack of information more severely affecting CH_4 fluxes (Johnson et al., 2021). This scarcity, in combination with the complexity of emission processes, impedes a clear identification of the relationships among the factors involved and prevents the definition of unambiguous correlations between emission drivers and GHG fluxes. Moreover, the spatial distribution of these datasets is unbalanced, with particularly limited information from tropical areas, where GHG emissions from reservoirs are more pronounced. Globally, GHG fluxes from large dam reservoirs exhibit wide latitudinal gradients (Almeida and Figueiredo, 2016), which are difficult to capture with the currently limited data availability. As a result, most models developed to date tend to underestimate emissions, largely

because the regions with the highest fluxes, namely tropical areas, are underrepresented in existing datasets (Karambelkar and Ames, 2025). Considering the expected expansion of hydropower in the coming years, especially in tropical regions, underestimation is one of the most critical concerns of current modeling efforts.

G-res Tool is the most advanced approach to estimate GHG emissions from reservoirs. G-res includes the specific environmental conditions of the reservoir to predict its associated carbon intensity (Prairie et al., 2017). It characterizes CO_2 and CH_4 fluxes for each emission pathway, namely CO_2 diffusion, CH_4 diffusion, bubbling, and degassing—and further simulates their temporal evolution to estimate the net carbon footprint of a reservoir over its expected lifetime.
The G-res framework consists in several modules. Among them, the most important is the post-impoundment emission module, containing the empirical models for the four emission pathways. G-res adopts linear regression models with logarithmic transformation of predictors. Compared to previous approaches, they were developed by selecting covariates from a larger set of possible predictors. In fact, G-res considered in the selection procedure most of the factors identified in the literature as key drivers, including organic variables such as soil phosphorus and organic carbon content prior to impoundment, as well as reservoir characteristics such as water residence time, thermocline depth, and the proportion of shallow areas. Accordingly with the previously mentioned scarcity of information, these factors are often unavailable, especially for global applications. For this reason, G-res also contains a series of submodules that use empirical models from the literature to estimate the predictors of the models. While the inclusion of these aspects allows for a thorough understanding of emission processes, on the other hand, it reduces the longitudinal applicability of the model and introduces additional sources of uncertainty, which propagate from the empirically derived covariates into the final estimates.

Although the need to better model GHG emissions from reservoirs is becoming increasingly relevant, existing approaches still suffer from severe limitations, and, remarkably, Machine Learning (ML) has not been systematically explored as a potential solution. Its flexibility and ability to capture complex, non-linear interactions makes it, in fact, a promising tool for improving both the accuracy and the scalability of emission estimates. This thesis addresses this gap by applying ML techniques to model GHG emissions from large dam reservoirs, with the aim of overcoming the main limitations of previous approaches.
Building on the G-res framework, and using it as a benchmark, the work develops along two complementary directions.

The first one focuses on the enhancement of the quality and reliability of estimates of bio-

genic reservoirs emissions. As previously discussed, the underlying relationships between GHG fluxes and their drivers are complex and difficult to detect. The use of ML algorithms is intended to achieve a better understanding of these relationships compared to linear models. Indeed, ML algorithms provide greater flexibility than linear approaches, enabling the capture of complex and non-linear interactions (Bragadeesh, 2020). By leveraging the extended G-res dataset, this study applies Support Vector Regression (SVR) to model CO_2 and CH_4 diffusive fluxes, providing an initial analysis that compares the estimates obtained through ML with those from the original G-res models. These newly developed SVR models are expected to better capture the relationships among the variables included in the G-res dataset, thereby providing a more accurate representation of emission processes compared to the empirical models currently employed.

Secondly, the study seeks to overcome the limitations posed by scarce predictor availability, which constrain the global applicability of models like G-res, by exploiting the flexibility of ML. In this phase, the ML approach replaces the empirical submodels employed in G-res to estimate certain covariates. ML models, being more flexible than linear regressions, can capture the underlying relationships directly from raw data without the need for intermediate empirical estimations. Moreover, their adaptability enables a more effective use of proxy variables, thereby allowing the inclusion of information that cannot be directly observed.
New ML models are developed on a newly compiled dataset that integrates raw information on reservoir characteristics, climatic variables, soil profile descriptors, and catchment composition. All included information is identified in the literature as directly or indirectly involved in GHG fluxes from reservoirs and can be easily extracted from globally accessible datasets.
This approach aims to enhance the applicability of the models in global applications, while removing the uncertainty originated by the empirical covariates used by G-res.

Nevertheless, the Machine Learning approach is also affected by well known limitations. While it has the potential to improve the accuracy of estimates and amplify the applicability of models to the global scale, its performance strongly depends on data quality and quantity. As ML models learn from data, the fewer the information provided, the lower their ability to identify meaningful relationships among variables. In particular, skewed and limited datasets reduce model robustness and generalizability, often leading to overfitting on the training data.

Recent studies on biogenic emissions from reservoirs have revealed the extent to which they can reach. Consequently, realistically quantifying those from hydroelectric reservoirs appears essential to assess the actual carbon intensity of hydroelectric power, considering

its importance and development. Currently, around 3,700 dams with an installed capacity greater than 1 MW are planned or under construction (Zarl and Tydecks, 2014) and most of them are located in Africa, China and South America, that have the highest remaining hydropower potential. On the other hand, these areas are located in the tropical climatic zone, which is well known characterized by the higher GHG fluxes from reservoirs.

The application of emission models to reservoirs underscores the relevance of the two issues addressed in this thesis, namely the reliability of estimates and the applicability of models. Particularly, the work analyses the biogenic emissions from European hydropower reservoirs. In Europe, hydropower is an essential player in clean energy production, generating the 34% of renewable energy. Currently, the reservoirs dedicated to hydropower are almost the 50% of all artificial reservoirs, and their number is expected to increase, with more than 600 planned dams (VGBE).
In this context, our models are applied to estimate current emission levels from existing artificial reservoirs. This step aims to evaluate the reservoirs behavior under the climatic and environmental conditions typical of Europe, and also offers an effective tool for projecting the impacts of future dams.
Focusing on hydropower reservoirs, the estimated GHG fluxes are added to those originating from the dam life cycle to assess the actual carbon intensity of hydropower in Europe. In this step, model applicability highlights its primary importance, as the reliability of the overall assessment increases with the number of reservoirs included.

1.1. Objectives and Research Questions

This thesis develops a new approach to modeling GHG emissions from reservoirs, with the aim of overcoming the main limitations of previous methods. The first objective is to improve model robustness and accuracy by implementing ML model algorithm on the variables included in the G-res dataset, with the purpose of enhancing performance compared to the linear approach originally adopted.
The second step involves applying the algorithm to a newly compiled dataset composed exclusively of raw and globally available variables. This step aims not only to ensure model accuracy, but also to test its applicability and inclusiveness on a global scale
Finally, the third objective concerns the application of the proposed approach to a real case. The contribution of biogenic emissions from reservoirs to the carbon intensity of European hydropower is assessed, providing a more realistic evaluation of the actual climate impact of this renewable source.

In conclusion, the research questions that this work seeks to address are

1. Can Machine Learning provide more accurate and robust estimates of reservoir GHG emissions compared to the G-res model?

2. Is it possible to develop globally applicable emission models through Machine Learning using only raw and globally easily available variables?

3. How can our study improve the estimation of current GHG emissions from existing artificial reservoirs in Europe, in terms of both the number of reservoirs included and the total emissions compared to previous approaches like G-res?

4. What is the contribution of biogenic emissions from reservoirs to the carbon intensity of hydropower in Europe?

1.2. Outline of the thesis

This thesis is organized as follows. Chapter 2 describes the main physical and biochemical processes underlying CO_2 and CH_4 emissions from reservoirs, providing an overview of the key drivers. Chapter 3 reviews the evolution of reservoir emission modeling, from its origins to G-res, which is regarded as the most advanced framework to date. Particular attention is given to G-res, with a detailed discussion of its underlying dataset and results. Chapter 4 shows how final models are developed. The first Sections describe the treatment of the G-res dataset, which constitutes our training dataset, and the procedure used to derive the G-res estimates, which serve as the benchmark for this study. Then, the pipeline of the ML modeling procedure is detailed (Section 4.4.2) and described for each emission pathways (Sections 4.5.1 and 4.5.2). Finally Section 4.6 presents the new dataset, examining the extraction procedure for each variable included (Sections 4.6.1, 4.6.4, 4.6.2, 4.6.1, 4.6.5). Chapter 5 details the performance of the ML models. First, the ML models using G-res variables are evaluated, followed by the ML model based on the new variables. Moreover, Sections 5.3 and 5.4 show the comparison between ML and G-res models, underlining their accuracy and applicability. In Section 5.5, the models are applied to the European case study, with a focus on the assessment of biogenic emissions from large dams (5.5.1) and on the evaluation of the contribution of GHG fluxes to the hydropower carbon footprint (5.5.2). Finally, Chapter 6 summarizes the main findings and outlines potential directions for future research.

2 | Physical processes

This chapter describes the main physical processes governing greenhouse gas emissions from reservoirs. It explains the factors involved in GHG production and release, and how these govern the emission dynamics. Furthermore, it highlights the mechanisms that distinguish gas production in artificial reservoirs from those in natural lakes, thereby clarifying the extent of the anthropogenic impact.

2.1. GHG in artificial reservoirs

Freshwater impoundments are sites of active carbon processing and transport. Both natural and artificial reservoirs receive organic and inorganic carbon from surrounding terrestrial landscapes and transform it into different species. Part of this carbon is emitted into the atmosphere, part is permanently buried in sediments, and part is transported downstream. Assessing the actual anthropogenic contribution to GHG emissions from artificial reservoirs requires considering only the fraction released as a direct consequence of transforming a river into a reservoir, while excluding the so-called *pre-impoundment* fluxes, namely the GHG emissions that would have occurred from the mosaic of terrestrial and aquatic components of the landscape even in the absence of damming (Prairie et al., 2018). Depending on soil types and ecosystem characteristics, the net balance can be either positive or negative. The loss of a terrestrial carbon sink should be considered an additional source of emissions, whereas pre-impoundment positive fluxes should, conversely, be subtracted from the anthropogenic contribution to GHG release (Prairie et al., 2018).

In addition to this baseline, impoundments induce further landscape, environmental, and hydrological transformations, altering flow regimes and sediment dynamics, and promoting thermal stratification, water-level fluctuations, and eutrophication (Li and He, 2022).

Post-impoundment emissions arise from the in situ degradation of both dissolved and particulate organic carbon of allochthonous origin. These transformations are mediated by aquatic microbes and photochemical processes, primarily producing CO_2 and CH_4

(St. Louis et al., 2000). Although such processes occur in both natural and artificial water bodies , the altered physical conditions in human-made reservoirs tend to enhance GHG emissions, which should therefore be regarded as anthropogenic.

However, distinguishing between pre- and post-impoundment emissions is still challenging, particularly because it is difficult to partition fluxes into those that would have occurred naturally and those attributable to anthropogenic alteration (Prairie et al., 2018).

The following sections focus on CO_2 and CH_4 production and release processes. Among GHG species, they have been identified as the most significant contributors to the global carbon footprint of reservoirs (Wang et al., 2024).

2.2. Carbon Dioxide

Carbon dioxide is by far the most abundant dissolved greenhouse gas in aquatic ecosystems (Prairie et al., 2018). CO_2 fluxes are largely correlated to allochthonous organic matter degradation, occurring primarily through aerobic respiration and photochemical oxidation. When the concentration of CO_2 exceeds saturation levels, it is emitted into the atmosphere (Schneider et al., 2020). Consequently, the higher the amount of organic carbon inputs received by waterbodies, the more elevated the emissions.
The conversion of organic carbon to CO_2 is also enhanced by several factors, including longer exposure to ultraviolet radiation (Prairie et al., 2018), smaller reservoir surface area (DelSontro et al., 2018), warmer temperatures and higher dissolved O_2 concentration (Paranaíba et al., 2018).
The main emission pathway for carbon dioxide, accounting for approximately 90% of the total (Paranaíba et al., 2018), is diffusive flux across the air–water interface. Gas diffusion rates are governed by the product of the CO_2 concentration gradient between water and air and the gas exchange velocity (k). Wind speed at the reservoir site is therefore a key factor influencing CO_2 diffusion, as it directly affects k (Roland et al., 2010). The remaining fraction of CO_2 is released through ebullition; however, this pathway is often negligible compared to diffusion (Diem et al., 2012).

2.2.1. Reservoirs as carbon sink

In principle, reservoirs could act either as carbon sinks or sources, depending on their age, location, climate and productivity of the given reservoir (Phyoe and Wang, 2019). Not all the organic carbon present in reservoir waters is released into the atmosphere as CO_2; a portion settles and is buried in the sediments, making the reservoir a potential long-term

carbon sink.

Artificial reservoirs, as well as natural ones, can therefore function as carbon sinks. Carbon burial is proven to be more efficient in anoxic environments, which prevents carbon oxidation. Such low-oxygen conditions are commonly found in artificial reservoirs due to slowed water flow and increased primary production (Li et al., 2024).

Additionally, the high rates of primary productivity may offset the function of reservoir as GHG source through algal atmospheric CO_2 fixation via photosynthesis, especially during summer (Phyoe and Wang, 2019).

In summary, negative CO_2 fluxes can occur as a result of carbon burial and atmospheric CO_2 uptake. These processes, like emission dynamics, are influenced by various hydrological and environmental conditions. High temperature enhances organic carbon mineralization in reservoir sediments (Gudasz et al., 2010). Meanwhile, the age of reservoirs is negatively correlated with carbon burial as their capacity to function as carbon sinks tends to decline over time with the evolution of local ecosystems.

At the same time, atmospheric CO_2 uptake is observed especially during the summer, when productivity levels are higher (Li et al., 2024).

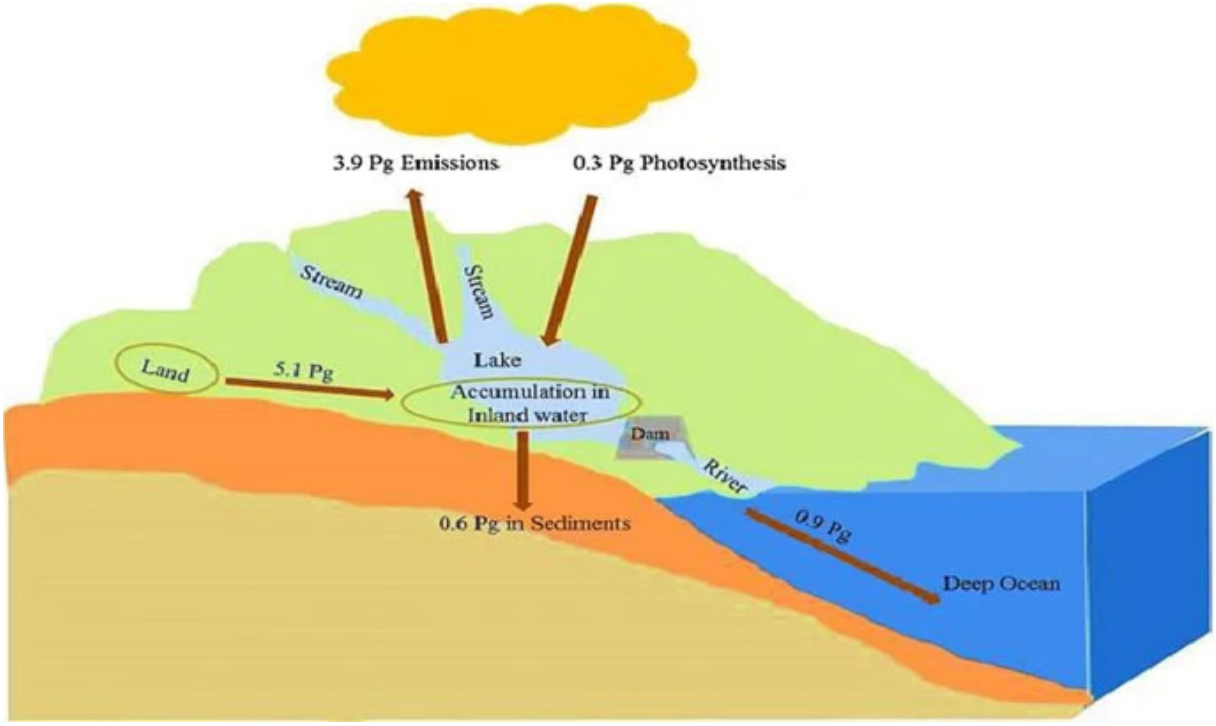

Figure 2.1: Carbon cycle in inland waters with estimates (Phyoe and Wang, 2019)

2.3. Methane

The biochemical process responsible for the conversion of organic carbon into methane is known as *methanogenesis*. The main agents of this process are methanogens, a group of archaea that thrive in anoxic environments. Through anaerobic respiration, they convert organic matter and other substances such as hydrogen and carbon dioxide into CH_4 (Bambace et al., 2007). The formation of anoxic zones is intrinsically associated with river damming. As previously mentioned, impoundment reduces water flow and increases sedimentation rates. Once deposited, carbon-rich sediments promote the formation of an anoxic layer just below the sediment–water interface, where methanogenesis can occur (Maeck et al., 2013). Furthermore, impoundment causes a longer residence time of the water at the reservoir site, which favors thermal stratification of the water column. Depending on local environmental conditions, the hypolimnetic zone may also become anoxic, thereby representing an additional source of methane (Kang et al., 2023). As a consequence, CH_4 concentrations are typically higher in the deeper layers of the reservoir.

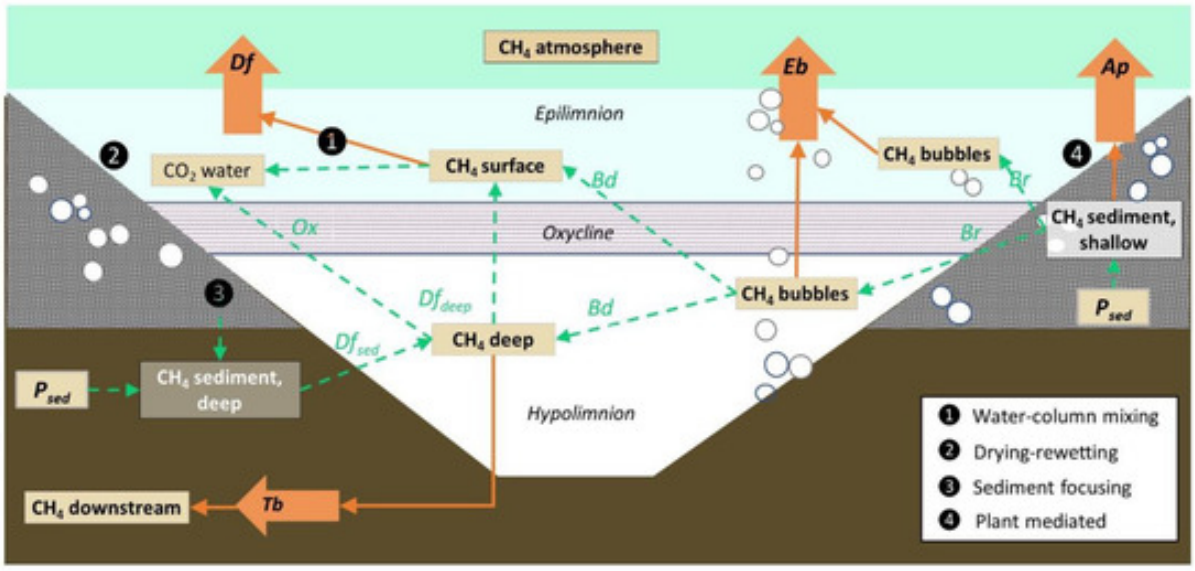

Figure 2.2: Pathways of methane production, transport, and emission in a stratified reservoir (Jager et al., 2023)

Methane produced in sediments can be transported to or through the water column via two main pathways, diffusion and ebullition.
If sediments are overlain by oxic water, a large portion of diffusing methane is rapidly oxidized, representing an additional source of CO_2 for reservoirs (Prairie et al., 2018).

On the other hand, dissolved CH_4 near the water surface diffuse through the air-water interface (DelSontro, 2011).
Methane transported via ebullition avoids oxidation by remaining trapped in gas bubbles during its ascent through the water column. Bubbles form when the methane production rate in sediments exceeds the rate at which it can diffuse. Their release depends on the hydrostatic pressure fluctuations and is influenced by temperature. The lower the water level, the lower the hydrostatic pressure on sediments and the higher the rate of bubbles release (Jager et al., 2023). Additionally, ebullition is proven to be strongly positively correlated to water temperature and wind speed (McClure et al., 2020). Due to its dependence on water level, also, ebullition is often considered the dominant CH_4 emission pathway in the shallow area of reservoirs (Yang et al., 2023).
The bubbling methane fluxes have been shown to be generally higher in hydroelectric reservoirs than in natural lakes. This trend is largely due to the greater fluctuations in water level caused by reservoir operations. As consequence, reservoir management represents a relevant factor influencing bubbling emissions.

2.3.1. Additional emission pathways in hydropower reservoirs

Furthermore, hydropower reservoirs account for two additional emission pathways that are negligible in natural lakes and are particularly relevant for methane compared to carbon dioxide. They are the degassing and downstream emissions.
Degassing emission occurs when water passes through turbines and spillways and is subjected to depressurization and aeration (Levasseur et al., 2021). Downstream emissions refer to the amount of CH_4 that remains dissolved in water after passing through the dam and is released downstream. The magnitude of both types of emission depend on the methane concentration in reservoir water. These fluxes are particularly relevant when the turbined water is drawn from the deep layers of the reservoir, where the methane concentrations are higher (Mercier-Blais, 2024). The depth of the water intake is therefore a key decision variable for limiting these emissions.

Deemer et al. (2016) emphasized the significant role that each emission pathway has in the total CH_4 flux from reservoir, underscoring the importance of accounting for all of them to assess the realistic GHG reservoir budget (Figure 2.3).

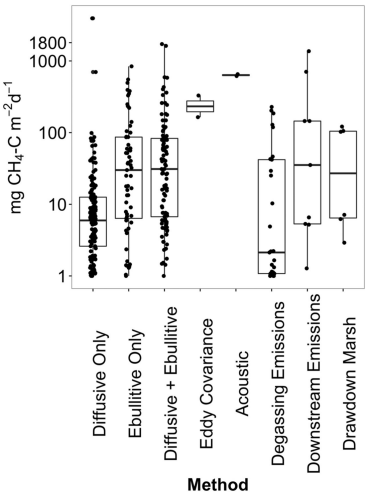

Figure 2.3: Areal CH_4 fluxes associated with reservoir (Deemer et al., 2016)

3 | State of the Art

Given the relevance of greenhouse gas (GHG) emissions from hydropower reservoirs, their estimation has become crucial. This chapter outlines the evolution of emission modeling, starting from the earliest approaches to G-res, the most advanced and widely used model currently available.
The final section provides a brief introduction to Machine Learning algorithms, with a particular focus on the Support Vector Regressor algorithm (SVR), implemented in this study.

3.1. Emission modeling

Nowadays, hydropower plays a central role in renewable energy production. It is considered a reliable, dispatchable, and, in principle, carbon-free source of energy (Li and He, 2022). However, hydropower reservoirs also constitute an important source of biogenic greenhouse gases, primarily due to the decomposition of organic matter submerged after impoundment (Scherer and Pfister, 2016). Recent studies (Barros et al., 2011; Hertwich, 2010; Scherer and Pfister, 2016), which highlight the magnitude of these biogenic fluxes, emphasize the need for their reliable quantification in order to accurately assess both the current and future impact of hydropower on global carbon emissions and its role as a renewable energy source.

Emission modeling can be approached in two main ways: through zonal statistics or empirical models. The zonal statistics method estimates emissions by multiplying the regional average flux by the corresponding reservoir area. In most of the cases, this method proved to be overly simplistic, due to the complexity and spatial and temporal variability that govern the emission processes (Li and He, 2022). In fact, the average regional flux is often not representative of all reservoirs within the region, particularly when measurement data are limited.

Empirical models, by contrast, are designed to account for reservoir specific characteristics and environmental conditions, and provide a more accurate representation of emission dynamics and more reliable estimates (Li and He, 2022).
This work focuses on this category, building on the pioneering study by St. Louis et al. (2000) on the identification of the main forcing factors influencing emissions.
The following paragraphs provide an overview of the key drivers of GHG fluxes and of the models developed to simulate them, with particular emphasis on G-res Tool.

3.1.1. Key drivers

The load of organic carbon in the water plays a central role in the release of GHG from reservoirs, representing the primary source of emissions. Its decomposition is driven by a combination of physical and biochemical processes, including both aerobic and anaerobic respiration, which converts organic matter into CO_2, CH_4, and N_2O (St. Louis et al., 2000). The gases, through different pathways, reach the water surface and leave the reservoirs, causing emissions (Chapter 2).

Multiple factors influence these processes and promote gases releases.
Most studies report a significant negative correlation between areal emissions of both CO_2 and CH_4 and reservoir age, which is currently identified as a key driver of emissions (Barros et al., 2011). Newly flooded organic carbon, such as litter and leaves, decomposes much more rapidly than soil organic carbon, causing larger releases in young reservoirs (Kelly et al., 1997). The depth and surface area of the reservoirs are also among the factors of interest in the emissions from reservoir studies. Both influence the physical dynamics associated with gas release. Reservoir depth, specifically, plays a main role in CH_4 emissions, that proved to be larger in shallow water zones compared to deep water ones (Yang et al., 2014). In addition to physical characteristics, climatic factors such as temperature, wind speed, and precipitation are positively correlated with GHG emissions. Emissions in tropical systems are typically higher (Yang et al., 2024), proving that the water temperature influences the carbon cycle and emission rates in reservoirs (Hou et al., 2013). Strong winds and precipitation affect sediment flow and gas transport coefficients, reduce deeply photosynthesis and water oxygen concentration, and promote GHG surface fluxes (Yang et al., 2014).
The most recent studies have brought to light the importance of the biological profile of flooded soil and reservoir water, showing that the type of ecosystem may influence the observed relationship. Factors such as dissolved organic carbon (DOC), total phosphorus, and chlorophyll content were found to be key factors in the GHG production processes. Chlorophyll-a and the oxygen-poor and C-rich environments favors CH_4 production while

phosphorus tends to covary with DOC in waters, driving CO_2 concentration in reservoirs (DelSontro and del Giorgio, 2018).

3.1.2. Models

To date, numerous studies have estimated the contribution of hydropower reservoirs to the anthropogenic carbon footprint (Barros et al., 2011; Hertwich, 2013; Deemer et al., 2016; Rosalina et al., 2016).
The global estimates produced by these models exhibit notable differences. The scarcity and variability of the available data, combined with the complexity of the underlying biogeochemical processes, have impeded a clear understanding of the relationships between emissions and their driving factors. Various combinations of the factors mentioned above have been fitted to linear regression models to predict global emission from reservoirs. As a consequence, the approach to the GHG emission modeling is not consistent; the developed models rely on distinct sets of covariates, even reporting, in some cases, contradictory correlation patterns among variables and fluxes.

Data scarcity affects both the direct measurement of greenhouse gas fluxes and the availability of information on key environmental drivers. Relatively few measurements are available for GHG emissions from reservoirs, especially of methane. Moreover, fluxes are measured using diverse tracing techniques, with sampling campaigns differing in duration, spatial coverage, and whether or not they account for the various emission pathways.
Although recent advances in satellite imagery are beginning to address the issue of the limited availability of spatially explicit data sets on lake characteristics on a global scale (DelSontro and del Giorgio, 2018), this remains a significant challenge for emissions modeling. Currently, information on water quality, especially water temperature and trophic status, often needs to be estimated or represented using proxy variables. The high water temperature in tropical reservoirs is often represented by latitude (Barros et al., 2011). With regard to the level of eutrophication in the basins, several proxy variables have been tested to represent it. However, the scientific debate on the most robust and reliable indicators remains open.

To illustrate the diversity of approaches adopted in reservoir GHG emission modeling, two representative studies are presented below, each highlighting specific key drivers for representing emissions.

Barros et al. (2011) implement multiple-regression linear models (MRL) to estimate global CO_2 and CH_4 emissions from hydropower reservoirs. They rely on a training dataset comprising 85 globally distributed reservoirs, including 141 observations for CO_2 and 85 for

CH_4. The selected predictors include reservoir age, biome, morphometric characteristics, and chemical properties.

Areal emissions of both gases were found to be negatively correlated with age and latitude. The empirical evidence of the latitude and emissions relationship is particularly relevant, as it had only been hypothesized prior to this study (Figure 3.1). In line with this finding, the authors suggest that latitude can act as a proxy for water temperature, given its positive correlation with GHG emissions. Accordingly, their models employ latitude as an indicator of the thermal gradient across reservoirs.

The most robust MRL models for both CO_2 and CH_4 include reservoir age, latitude and carbon inputs, represented through dissolved organic carbon (DOC) in water. Additionally, the CH_4 model uses as a predictor reservoir depth, reflecting the demonstrated negative relationship between depth and CH_4 fluxes (Section 2.3).

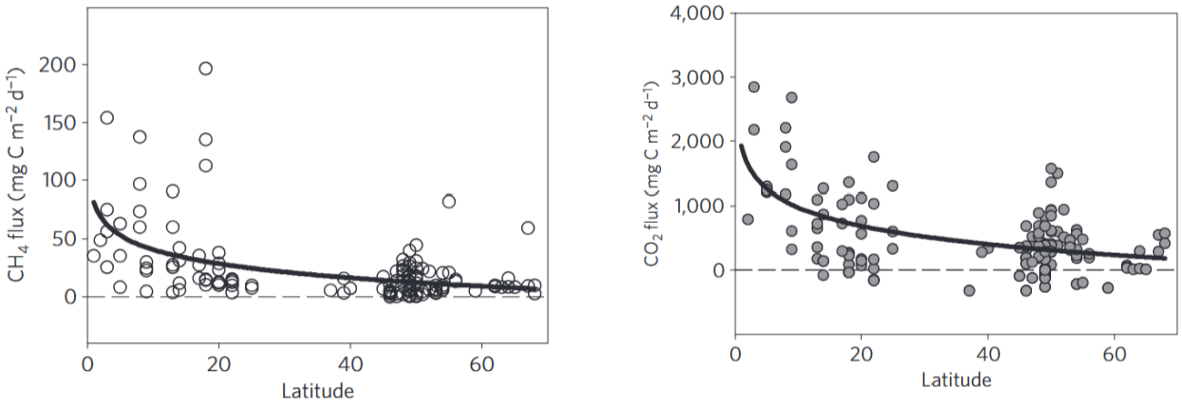

Figure 3.1: Scatter plot and exponential decline for the relationships between CH_4 and CO_2 emissions and latitude (Barros et al., 2011).

The work by Barros et al. (2011) served as the foundation for the study by Scherer and Pfister (2016), which aim to assess the global carbon impact of hydropower reservoirs under the assumption that post-impoundment emissions are representative of net fluxes. Their models build upon the same training dataset used by Barros et al. (2011), enriched with information on electricity production, and are subsequently applied to estimate emissions from 1473 globally distributed reservoirs.

A key point emphasized by the authors is that DOC, used as a predictor in the models by Barros et al. (2011), is not available in global datasets. To address this limitation, they introduce the erosion rate (ERR) as a proxy for organic matter input. The primary focus of Scherer and Pfister (2016) is, in fact, the global applicability of emission models. Therefore, only predictors available in global datasets are considered. In addition to ERR, the potential predictor variable set incorporates reservoir area, area-to-electricity ratio,

age, volume, volume-to-area ratio (as a proxy for depth), mean, minimum, and maximum air temperature, net primary productivity, and topsoil organic carbon content.
Carbon emissions per energy unit (kg MWh^{-1}) and areal fluxes (mgC m^{-2} d^{-1}) are estimated using Generalized Linear Models (GLMs). The two models are then averaged to produce a combined emission estimate. Aware of the likely underestimation of CH_4 flux measurements and CO_2 overestimation, correction factors are applied to both gases final models.
The CO_2 emissions per energy unit model includes the area-to-electricity ratio and reservoir area, while the corresponding flux model incorporates reservoir age and erosion rate (ERR). The CH_4 energy-based model includes age, area-to-electricity ratio, and maximum temperature; the flux-based model, on the other hand, includes age, ERR, area, and maximum temperature. Interestingly, a negative correlation is found between CH_4 emissions and ERR, in contrast to the strong positive correlation observed between ERR and CO_2 emissions.

Although the modeling strategies and selected covariates differ, the studies exhibit similar limitations which additionally are consistent with those commonly observed in other linear models reported in the literature. A common concern across these studies is the potential underestimation of emissions, primarily due to the lack of accurate data on methane fluxes, and the scarcity of measurements from reservoirs in the tropical region, to which much of the observed variability is attributed. Aquatic methane fluxes are difficult to accurately measure because measuring all different emission pathways is complicated and time consuming, particularly due to the high degree of spatial and temporal variability (Hertwich, 2010). Ebullition and degassing fluxes, both poorly represented in the data set, appear to have a more significant impact than diffusive fluxes, which are, by contrast, more easily and frequently sampled (Raymond et al., 2013). As a result, global estimates of emissions from reservoirs are likely still underestimated.
Moreover, improving the accuracy of the trophic state representation appears essential to refine emission estimates, especially for the highest emission values that remain unexplained by the only environmental descriptors.
Nonetheless, these early approaches provided the conceptual basis for the development of G-res, to date the most advanced model to predict the GHG emissions from reservoirs. The G-res framework synthesizes the published literature on reservoir emissions into a series of empirical models, proposing a novel, integrative approach to the estimation of Net GHG footprint of dams.

3.2. G-res Tool

G-res Tool was developed in a decade by a research project led by the International Hydropower Association (iha) and the UNESCO Chair in Global Environmental Change, involving scientists from the University of Quebec at Montreal (UQAM), the Norwegian Foundation for Scientific and Industrial Research (SINTEF) and the Natural Resources Institute of Finland (LUKE). The result was the development of an online modeling platform, globally applicable, capable of predicting emissions of both CH_4 and CO_2 over the expected lifetime of the reservoirs. Whereas previous studies focused on modeling specific emission pathways or isolated aspects of carbon dynamics, the G-res framework is capable of accounting for all components that contribute to the net carbon footprint of a reservoir. Starting from the specific environmental conditions of the reservoirs and on their and their catchments characteristics, G-res dynamically predicts the carbon fluxes, partitioned into four emission pathways, CO_2 diffusion and CH_4 diffusion, ebullition, and degassing. Additionally, accounting for the temporal dynamics of emissions, it enables a more accurate and complete assessment of the carbon footprint of reservoirs throughout their operational lifetime (100 years).

G-res framework aims at predicting the reservoir induced change in the GHG fluxes to the atmosphere of the flooded landscape. Pre-impoundment emissions are subtracted to post-impoundment emissions to obtain the 100 yr-average daily net emission rate, expressed per unit of area or per reservoir area. The model output is the result of a combination of empirical models contained in the series of modules constituting G-res framework, summarized in the Figure 3.2. The tool is globally applicable and easily accessible through an online platform (G-res Tool). Moreover, it can be applied both for existing and new reservoirs, making it a useful resource to support the decision-making process during the reservoir planning and development phases.

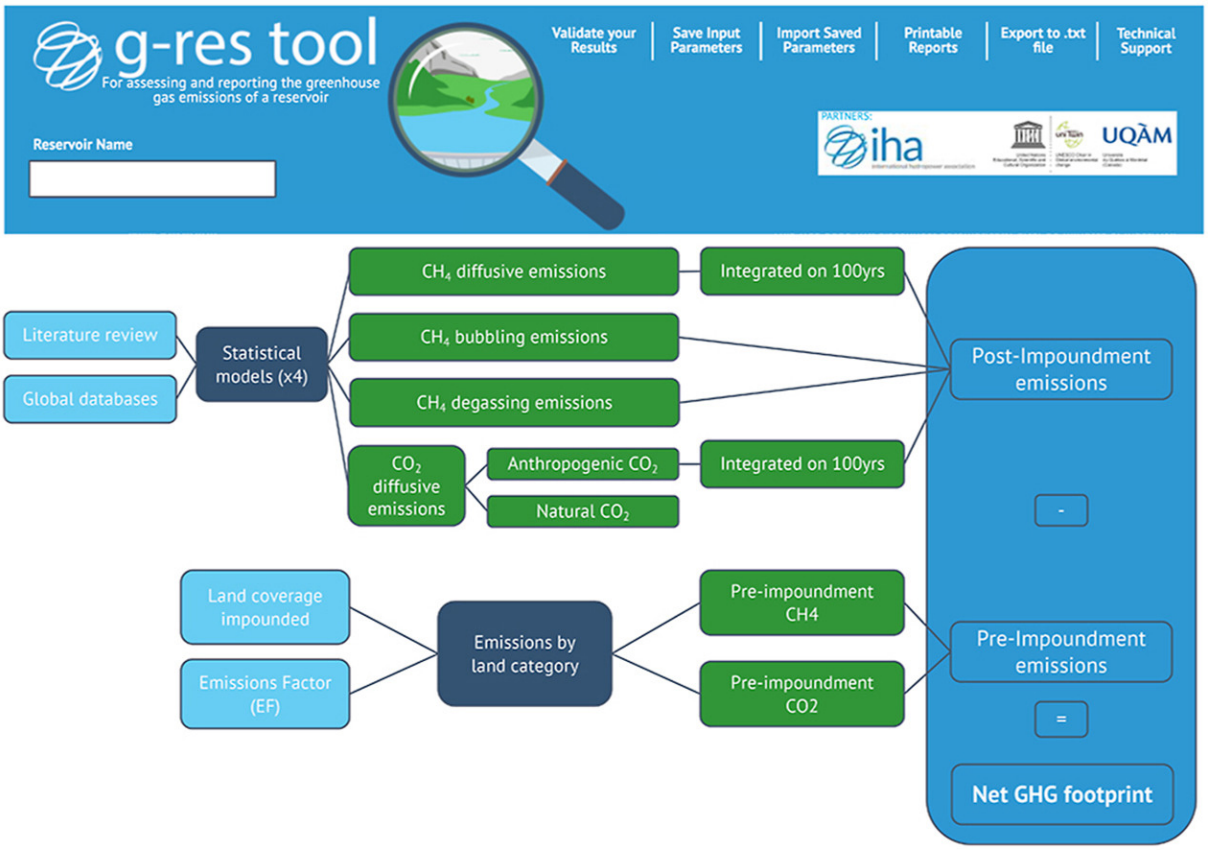

Figure 3.2: G-res conceptual framework for estimating GHG emissions from reservoirs (Xu et al., 2021).

3.2.1. G-res inputs

G-res training dataset was compiled through an extensive review of the available literature up to 2016. It contains information for 223 globally distributed basins, classified into four climate zones: temperate, boreal, tropical, and subtropical. The four emission pathways are not uniformly covered in the dataset, which includes 279 field assessments of diffusive CO_2, 205 of diffusive CH_4 emissions, 59 of bubbling CH_4 emissions and 52 of degassing CH_4 emissions. To address the issue of an inconsistent time scale among emission measurements, the G-res developers implemented an annualization procedure. For the unsampled months, this method adjusts measured fluxes upward or downward based on their relationship with the local water temperature, applying specific correction metrics for CO_2 and CH_4. The resulting monthly estimates were then aggregated to obtain annual values.

In addition to GHG flux data, information on climatic, geographic, edaphic and hydrologic conditions of each reservoir and its catchment are included. Potential predictor variables were collected from a variety of sources including literature, worldwide GIS lay-

ers and information contained in the GRanD database, a product of the Global Water System Project, which provides a geographically explicit and reliable database for existing dams Global Dam Watch. Zonal statistics tools were applied to estimate the climatic and environmental characteristics of each reservoir and its basin, which is delineated using Hydrobasins GIS product (Prairie et al., 2017).

Not all plausible input factors are readily available. A notable feature of the G-res dataset is that many of the predictors used are not directly observed, but estimated through empirical models described in the literature. Among the most relevant ones, there are the thermocline depth, the percentage of littoral area, the water residence time and the nutrient status (G-res Team, 2022). The equations of the empirical models for the covariate estimation are available in Appendix A of *G-res Tool: A comprehensive framework for estimating greenhouse gas emissions from reservoirs* (Xu et al., 2021). Table 3.1 lists the predictor variables contained in G-res dataset and reports their source.

Table 3.1: List of predictor variables included in G-res training dataset, including units and data sources.

Variable	**Unit**	**Source**
Reservoir variables		
Country	–	Literature, GRanD
Climate zone	–	Rubel & Kottek
Dam coordinates	decimal degrees	Literature, GRanD
Impoundment year	year	Literature, GRanD
Reservoir area	km^2	Literature, GRanD, GIS
Reservoir volume	km^3	Literature, GRanD
Maximum depth	m	Literature, GRanD, estimated
Mean depth	m	Literature, GRanD, estimated
Thermocline depth	m	Estimated from literature
Littoral area	%	Estimated from literature
Water residence time	year	Estimated from literature
Mean monthly/annual air temperature	°C	Hijmans et al. (2005)
Annual precipitation	mm/yr	Hijmans et al. (2005)
Mean monthly/annual wind speed	m/s	Hastings et al. (1999), NOAA GLOBE
Mean global horizontal irradiance	$kWh/m^2/day$	NASA SSE (2008)
Phosphorus concentration	µg/L	Literature, estimated
Catchment variables		
Soil carbon content (inundated area)	kgC/m^2	SoilGrids (Hengl et al., 2017)
Catchment area	km^2	Literature, GRanD, GIS
Mean annual runoff	mm/yr	Fekete et al. (2000)
Population density	$persons/km^2$	CIESIN (2005)
Annual discharge	m^3/s	Estimated from literature
Land cover	%	ESA CCI (2014)

3.2.2. G-res models

The main modules of the G-res framework contains the models to estimate emissions.

The estimation of pre-impoundment emissions is based on a zonal statistics approach. Since soil carbon fluxes depend on multiple factors, including land cover, soil type (mineral or organic), and climatic conditions, emissions vary across the diverse ecosystems that compose large landscapes.
The pre-impoundment GHG footprint (F^{pre}) is therefore quantified through the weighted sum of the GHG balance of each landscape component. Emissions from each individual ecosystem type are represented by default values derived from IPCC (Intergovernal Pannel of Climate Change) through specific emission factors for both CO_2 and CH_4. Formally, it is defined as:

$$F^{\text{pre}} = \frac{1}{A_{\text{res}}} \sum_{g \in GHG} \sum_{i \in LC} EF_{i,g}\, A_i \tag{3.1}$$

where A_{res} is the reservoir area, A_i the area of land-cover class i, and $EF_{i,g}$ the emission factor of gas g associated with land-cover class i.

The approach to the post-impoundment GHG emissions estimation is empirical modeling. Since the four emission pathways are controlled by different processes and drivers, they are analyzed and modeled separately. G-res Tool relies on a series of multivariate statistical models, using both reservoir and catchment predictor variables.
Forcing factors selection was carried out using the elastic net regression procedure, that is particularly suitable to modeling cases with many potential predicting variables and low sample size (G-res Team, 2022). This method imposes a penalty on large coefficients in order to face the model instability caused by highly correlated variables. According to the penalty parameter, they can be reduced to zero allowing an objective selection of the variables (Xu et al., 2021). The logarithmic transformation is applied to both model target and inputs, to fulfill the requirement of the procedure of normal residuals.
The resulting model equations are:

$$CH_4^{\text{diff}} = 10^{\left(0.8032 - 0.01419 \cdot \text{Age} + 0.4594 \cdot \log_{10}\left(\frac{A_{\text{lit}}}{100}\right) + 0.04819 \cdot \text{T}_{\text{CH}_4}\right)} \tag{3.2}$$

$$CH_4^{\text{bub}} = 10^{\left(-1.3104 + 0.8515 \cdot \log_{10}\left(\frac{A_{\text{lit}}}{100}\right) + 0.05198 \cdot \text{GHI}_{\text{res}}\right)} \tag{3.3}$$

$$CH_4^{\text{deg}} = 10^{\left(-6.9106+2.950\cdot\log_{10}(\text{CH}_4^{100\text{yrs}})+0.6017\cdot\log_{10}(\text{WRT})\right)} \cdot \frac{A_{\text{cat}} \cdot 10^6 \cdot R_{\text{ann}}}{10^9 \cdot 1000} \cdot 0.9 \tag{3.4}$$

$$\text{CO}_2^{\text{diff}} = 10^{\left(1.860-0.330\cdot\log_{10}(Age)+0.0332\cdot\text{T}_{\text{CO}_2}+0.0799\cdot\log_{10}(A_{\text{res}})+0.0155\cdot\text{C}_{\text{soil}}+0.2263\cdot\log_{10}(\text{TP})\right)} \tag{3.5}$$

For both CH_4 and CO_2 diffusive emissions, the age of the reservoirs was selected by the elastic net procedure as one of the strongest predictors (Prairie et al., 2017). As consequence, emissions are estimated at each reservoir specific age, and then integrated over reservoir lifetime to yield the total net carbon footprint.

For the prediction of the diffusive CH_4 emissions (Eq. 3.2), the age of the reservoirs has the strongest influence. The only other predicting variables selected are the mean effective CH_4 annual temperature (T_{CH_4}) and the percent of the littoral area (A_{lit}). Both of them are computed through empirical models contained in G-res sub-modules.
The effective temperature is derived from the mean annual air temperature. Monthly temperatures are corrected by a coefficient (0.052) that accounts for methane fluxes and air temperature correlation and then averaged to obtain the mean effective temperature T_{CH_4}. The estimation of the littoral area involves the maximum (D_{max}) and mean depth (D_{mean}) of the reservoirs.

$$T_{\text{corr},CH_4} = 10^{(T_{\text{month}}\cdot 0.052)} \tag{3.6}$$

$$T_{CH_4} = \frac{\log_{10}\left(\text{Average}\left(T_{\text{corr},CH_4}^{12\text{ months}}\right)\right)}{0.052} \tag{3.7}$$

$$q = \frac{D_{\text{max}}}{D_{\text{mean}}} - 1 \tag{3.8}$$

$$\%A_{\text{lit}} = \left(1 - \left(1 - \frac{3}{D_{\text{max}}}\right)^q\right) \cdot 100 \tag{3.9}$$

The CH_4 bubbling emissions model uses cumulative global horizontal irradiance (GHI_{res}) and the percentage of littoral area, estimated through Equation 3.9, as predictors. This choice reflects the established negative correlation between bubbling emissions and water depth, as well as their connection with zones of elevated sedimentation (Xu et al., 2021).

To develop the degassing emissions model, degassing fluxes observations were derived as the difference of methane concentrations upstream and downstream the dams. These differences were then multiplied by the mean annual flow through the turbines, estimated as 90% of the annual runoff.
The best model to predict CH_4 concentration differences uses the water residence time (WRT) and post-impoundment annual estimated CH_4 diffusive emission ($CH_4^{100\text{yrs}}$), as a proxy of CH_4 production. WRT is estimated by Equation 3.10, where A_{cat} is the catchment area and R_{ann} the annual runoff.
Since this pathway is relevant when the turbined water comes from the hypoliminium, where methane concentrations are higher, the G-res Tool estimates degassing emissions only when the water intake is located below the thermocline.

$$WRT = \frac{D_{\text{mean}} \cdot A_{\text{res}}}{A_{\text{cat}} \cdot R_{\text{ann}}} \cdot 1000 \tag{3.10}$$

Notably, degassing and bubbling estimates are associated with a higher uncertainty compared to those derived from the diffusive pathway, as the models and their analysis are based on an extremely small dataset.

The best model for diffusive CO_2 flux includes, in addition to reservoir age, the mean effective annual temperature (Eq. 3.12), phosphorus load, reservoir area and pre-impoundment reservoir surface soil carbon content (C_{soil}). Unlike the other models, this is the only one that includes factors related to the organic input to reservoirs, namely total phosphorus and pre-impoundment soil carbon content. However, phosphorus load is not directly available and is instead estimated through a set of empirical models, specific to the land cover types that characterize the landscape. A different phosphorus load is associated with each basin land-cover type, represented through specific coefficients.
A strong positive correlation is found between CO_2 emissions and temperature. This is particularly important since it is suggesting that CO_2 emissions could increase significantly with increasing temperatures due to climate change.
Below, we present the equation used to estimate the effective temperature, along with an example of phosphorus load calculation, where $\%LC$ represents the percentage of land cover:

$$T_{\text{corr},CO_2} = 10^{(T_{\text{month}} \cdot 0.05)} \tag{3.11}$$

$$T_{CO_2} = \frac{\log_{10}\left(\text{Average}\left(T_{\text{corr},CO_2}^{12\,\text{months}}\right)\right)}{0.05} \tag{3.12}$$

$$PL_{\text{forest}} = \frac{10^{\left(0.914 - 0.014 \cdot \log_{10}\left(\frac{\%LC_{\text{forest}}}{100} \cdot A_{\text{cat}}\right)\right)}}{100} \tag{3.13}$$

3.2.3. G-res results and limits

To date, the G-res Tool represents the most comprehensive framework for predicting the global GHG footprint of reservoirs, as it accounts for the relative contribution of the four CH_4 and CO_2 emission pathways and quantifies their evolution over time.
However, it presents some limitations.
Looking at the R^2 of the G-res models (Table 3.2), the results highlight the high uncertainty associated with bubbling emissions. However, recent studies underscore their significant contribution to total carbon emissions, making accurate estimation essential to avoid underestimation (Hou et al., 2013).
The developers also pointed out that a global enhancement will be obtained including a direct link with the trophic status of the reservoirs for all the emission pathways. Until now, only the CO_2 diffusion model accounts for the soil carbon content and the phosphorus load, that are also derived in turn, by empirical models (Section 3.2.2).
As previously described, in addition to phosphorus load, several other predictor variables must be indirectly computed through empirical sub-modules. Whereas the use of modeled variables enables the inclusion of a wider range of explanatory predictors, it also introduces an additional layer of uncertainty into the modeling process. This uncertainty, associated with the predictors themselves, propagates through the model and ultimately affects the reliability of the emission estimates.

Emission pathway	n	R^2	RMSE
CH_4 diffusion	160	0.51	0.52
CH_4 bubbling	46	0.26	0.80
CH_4 degassing	38	0.68	0.81
CO_2 diffusion	169	0.36	0.39

Table 3.2: G-res model performance statistics for each emission pathway.

3.3. Machine Learning overview

In our work, we rely on Machine Learning to improve GHG emission estimates and better represent the complex relationships between emissions and their drivers.

Machine Learning (ML) can be defined as the study of computer algorithms that are automatically enhanced through experience and by the advent of data. The goal of these algorithms is to build models based on training data that are able to make predictions without any explicit external programming rules (Savale, 2021).
The development of these models can be divided into two main phases: training and validation. During the training phase, the model learns and generalizes the relationships within the provided data in order to reproduce the outputs. Once trained, it is validated on previously unseen data to assess its accuracy. If the model proves to be reliable, it can then be used to predict or classify new information.
ML models are classified into four categories: supervised, semi-supervised, unsupervised, and reinforcement learning (Figure 3.2).

Of particular interest to us are the supervised algorithms.
Supervised learning models learn from labeled training data, where input-output pairs are provided. They rely on a supervisor to guide the learning process and correct predictions (Talaei Khoei and Kaabouch, 2023). The human developer plays an important role in processing the training data and searching of model parameters. During this step, the optimal parameters that best represent the relationship between the inputs and the target are identified, minimizing a loss function. The equation that defines the optimal parameter $p*$ is:

$$p* = \arg\min_{p} \left(\sum_{i} \mathcal{L}(y_i, f_p(\vec{x}_i)) + \Omega(p) \right) \tag{3.14}$$

where p is a non-optimal parameter and Ω a term to limit the complexity of the model and

avoid overfitting. After the optimal parameters are set, the model generates the output y as a function of the input $\vec{x}$.

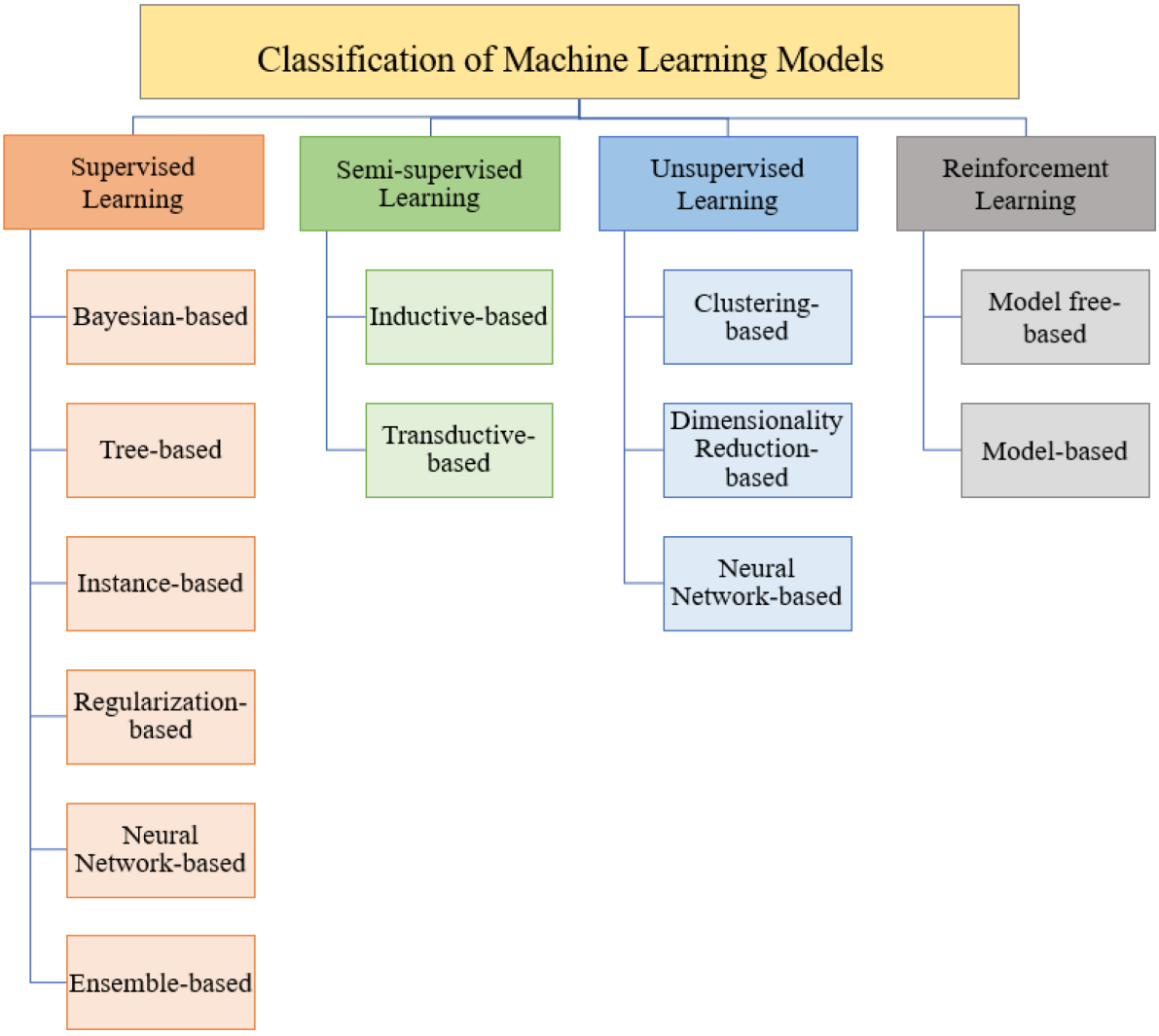

Figure 3.3: ML models classification.

3.3.1. Support Vector Regressor

This sub-section will briefly explain the functioning of the Support Machine Regressor (SVR), which is the algorithm used to build the final model in this project. The choice is driven by the ability of SVRs to afford balanced predictive performance, even in studies where sample sizes may be limited (Pisner and Schnyer, 2020).

SVR is a type of Support Vector Machine (SVMs). These are classification algorithms that map the data in a high-dimensional space and try to find the hyperplane that maximally separates different classes or output values (GeeksforGeeks).

SVR algorithms instead work in regression, aiming at define a function *f(x)* that has at most the ϵ deviation from the actually obtained targets y_i for all training data, and at the same time is as flat as possible. The errors are irrelevant as long as they are lower than ϵ, while they are not accepted if they are larger than ϵ (Smola and Schölkopf, 2004). The model flexibility can be increased by allowing a controlled degree of tolerance for errors exceeding ϵ, regulated by a penalty parameter.

The optimization problem in flexible SVR is expressed as:

$$\min \left(\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{\ell} (\xi_i + \xi_i^*) \right) \tag{3.15}$$

where $\|w\|^2$ represents the flatness of the function $f(x)$, ξ_i and ξ_i^* are variables that relax the constraints for errors beyond the ϵ margin, and C is a regularization parameter that defines the trade-off between model complexity and flexibility to account for the deviations. A function $\phi(x)$ maps the training data into the higher-dimensional feature space where SVR learns a function that approximates the target values while maintaining a margin of tolerance ϵ. This allows the model to capture non-linear relationships in the original input space. The use of a kernel function, $K(x, z) =< \phi(x), \phi(z) >$, allows the definition of the function without explicitly carrying out the mapping into feature space. Kernels in SVR can range from simple linear functions, suitable for modeling linearly correlated data, to more complex functions such as radial basis (RBF) or sigmoid kernels, which enable the model to capture non-linear and intricate relationships between variables. Figure 3.4 provides a graphical representation of the functioning of SVR algorithm (Suykens and Vandewalle, 1999).

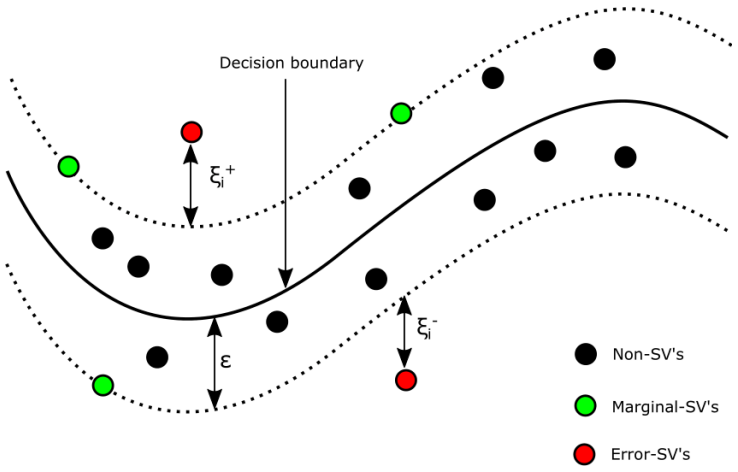

Figure 3.4: Schematic SVR algorithm representation (Mustafa et al., 2024)

4 | Materials and Methods

This chapter provides an overview of the data and methodologies employed in the study. The first section centers on the analysis and management of the G-res dataset, and includes the implementation of the G-res models, which serve as a benchmark for evaluating the performance of the new predictive models. The core of the chapter is structured around two main sections, each addressing one of the research questions guiding this study. The first part outlines the methodology adopted for training and validating new Machine Learning models using G-res original data; the second section introduces new potential emissions drivers and shows how they have been collected to create a new enreached dataset. It describes the procedure for extracting these covariates from global datasets and their use in the construction of the new Machine Learning models, detailed in the final part of the chapter.

4.1. G-res dataset

G-res training dataset is described in detail in Section 3.2.1. It accounts for 223 globally distributed reservoirs, containing a total of 279 field assessments of diffusive CO_2, 205 of diffusive CH_4, 59 of bubbling CH_4, and 52 of degassing CH_4 emissions. These are reported in (mgC m^{-2} d^{-1}) for diffusive and bubbling emissions, and in (mgC L^{-1}) for degassing emissions. For some of the 223 reservoirs, information covers several years, while for others, only single year observations are available. Even within the same reservoir, data may originate from different sources, meaning that multiple measurements for the same basin and year coexist in the dataset. The origin sources are mostly reported, but no information is provided regarding their reliability.
With regard to input variables, Table 4.2 lists those contained in the dataset, downloadable from *"A new modelling framework to assess the biogenic emissions of greenhouse gases from reservoirs"* (Prairie et al., 2017). Among the potential predictors considered by G-res developers, these covariates are the ones selected for their final emission models. As described in Section 3.2.1 and further detailed in Section 3.2.2, these variables are often not directly available from global, or even local, datasets. Empirical models were

therefore used to estimate them (see Equations 3.12, 3.9, 3.10, 3.8, and 3.7). Notably, the original raw input data used to feed these empirical models are not publicly available, since G-res developers decided to directly publish only their outputs. However, despite the use of empirical models to estimate these variables, missing data remain widespread, indicating that even the raw values required for covariate estimation through empirical models are often unavailable.
As highlighted in the overview of the G-res platform, flux measurements were collected through sampling campaigns that differed in duration and timing. To ensure consistency, the developers applied an annualization procedure (Section 3.2.1), and the dataset reports only the standardized annual values resulting from that process.

4.1.1. G-res dataset criticality

During the data exploration phase, we identify two main issues in the original dataset that require specific handling. The first concerns the previously mentioned presence of multiple observations for the same reservoir (and sometimes for the same year); the second involves negative emission values, especially for carbon dioxide. Firstly, biases potentially linked to the data sources have been investigated. However, no systematic patterns have emerged.
The first issue is addressed by preserving the annual variability of emissions in order to analyze temporal changes in the underlying processes and to capture the correlation between fluxes and reservoir age. To retain the information associated with different reservoir ages, we keep dataset entries, namely set rows, referring to different ages unchanged and performed arithmetic averaging only among flux observations corresponding to the same reservoir and age. The corresponding environmental characteristics are averaged as well, while missing or non-unique reservoir-specific attributes are completed by consulting the GRanD dataset, whenever possible.
This procedure is motivated by the lack of detailed information on the reliability and accuracy of the individual measurements and data, which are therefore treated as equally valid.
The second matter is dealt with the analysis of the processes that lead to negative fluxes. In subsection 2.2.1 is explained that reservoirs can exhibit a net negative carbon footprint due to the processes of carbon burial and atmospheric CO_2 uptake. The first refers to the permanent sedimentation of the organic and inorganic carbon at the bottom of the reservoirs. It is therefore a fundamentally different process from diffusive gas exchange, which the model aims to quantify. Measuring the magnitude of stored carbon through burial requires different tracing methods than those used to assess gas exchange at the

water–air interface. As a consequence, we assume that negative emission values in the dataset refer to CO_2 uptake from the atmosphere.
CO_2 uptake process is driven by the abundance of aquatic plants that fix carbon dioxide through photosynthesis. Although it is generally associated with eutrophic conditions and elevated water temperatures, the dataset includes several negative emission values for oligotrophic reservoirs located in boreal regions. Given that this pattern is inconsistent with the expected theoretical framework and is not supported by the literature for such ecosystems, we excluded these values from the analysis owing to their questionable reliability.

After performing these adjustments the dataset contains 202 reservoirs, including 222 field-assessments for CO_2, 183 for CH_4 diffusion, 55 for CH_4 bubbling and 47 for CH_4 degassing.

4.2. Re-Emission

The G-res emission estimates serve as a benchmark for evaluating the performance of the new models. To this end, the Python library *Re-Emission*, available on GitHub (Janus, 2023) is used to run the G-res models on the reservoirs included in the G-res dataset. It calculates the net full life-cycle emissions as well as emission profiles over time for three greenhouse gases, CO_2, CH_4 and N_2O (Janus, 2023).

The repository is structured into several modules, each handling a different aspect of the model pipeline.
The most important is the *Emission* module, which contains the equations for calculating net, total and life-cycle integrated emissions for the three greenhouse gases. The necessary inputs and coefficients of this module directly come from the other modules. The *Temperature* module computes the CO_2, CH_4 and N_2O effective temperature based on monthly air temperature (Eq. 3.12, 3.7). The characteristics of the reservoir and its basins are defined in the corresponding scripts. They contain both the data types of the input variables and the empirical models used to compute them, some of which presented in the Section 3.2.2 (Eq. 3.9, 3.8, 3.10). The input data for calculating both models' covariates and emissions must be structured according to a specific configuration, described in a file contained in the library.

As mentioned, the downloadable G-res dataset contains the estimated inputs, ready to be used by the models. Accordingly, the scripts must be adapted to match the available input file. Additionally, the *Emission* module is modified to provide only the estimates of our interest, namely CO_2 and CH_4 yearly post-impoundment values.

4.2.1. Re-Emission evaluation

The metric used by G-res developers to assess model performance is the coefficient of determination (R^2). Accordingly, the estimates obtained in this study are evaluated using the same criterion.
Table 4.1 reports the corresponding values. These results differ from those published by Prairie et al. (2017), shown in Table 3.2. This discrepancy can be attributed to differences in the sample considered in our analysis and in G-res, as well as to the data treatment method we applied, described in Section 4.1.1. Moreover, the G-res developers applied the Cook's Distance $> 3\mu$ criterion to detect outliers, whereas in our analysis no values were excluded at this stage, except for negative fluxes.

Figure 4.1 compares G-res estimates to the observations, providing a visual interpretation of the R^2 value and its implications in terms of prediction errors.

Emission pathway	**n**	**R^2**
CH_4 diffusion	157	-0.04
CH_4 bubbling	48	-0.17
CH_4 degassing	44	0.35
CO_2 diffusion	151	0.21

Table 4.1: R^2 values obtained using the Re-Emission library and the number of samples included in the evaluation analysis (n) for different emission types.

Reservoir name	Dam or reservoir name
GRanD ID	ID from the GRanD DB
Impoundment year	Year when the reservoir impoundment was completed
Longitude (DD)	In degree decimal
Latitude (DD)	In degree decimal
Climate zone	Climate classification: boreal, temperate, tropical, subtropical
Reservoir area (km^2)	In km^2
Littoral area	Area shallower than 3m [%]
River area before impoundment	Percentage of the catchment area occupied by the river channel prior to impoundment [%]
Catchment area	[km^2]
Maximum dept	[m]
Mean depth	[m]
Thermocline depth	[m]
Discharge	Dam discharge [m^3 s^{-1}]
WRT	Water residence time [yrs]
Total Phosphorus	Phosphorus input from the catchment [µg L^{-1}]
Niveau trophique	From broad trophic state categories
Soil Carbon Content	Soil content in the surface layer (0-30 cm) [kgC m^{-2}]
Cumulative radiance	Calculated for ice-free period and number of months over 0°C [kWh m^{-2} period]
Effective temperature CH_4	Annual mean temperature corrected with temperature correction coefficient representing the temperature dependence of GHG production [°C]
Effective temperature CO_2	Annual mean temperature corrected with temperature correction coefficient representing the temperature dependence of GHG production [°C]
References	Source of emissions field measurement
Year of sampling	Year when field measurements were sampled
Period of sampling	Annual, Ice-free or month of sampling

Table 4.2: Description of reservoir and environmental variables available G-res database.

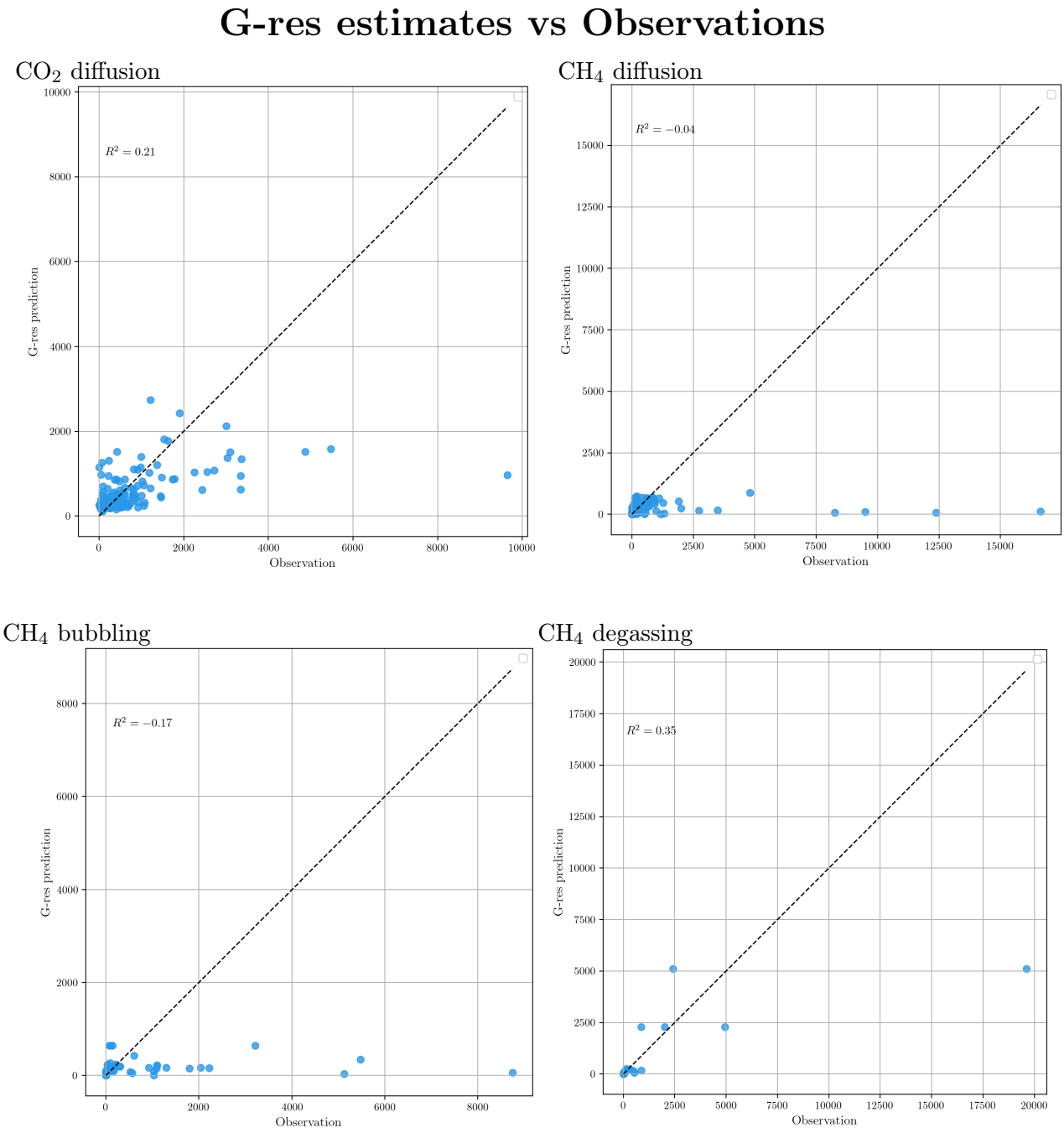

Figure 4.1: Comparison between G-res model estimates and observed values for the four emission components (gCO_{2eq} m^{-2} yr^{-1})

4.3. Fluxes observations analysis

This section briefly describes an additional analysis of emission field measurements, conducted to detect potential outliers, identify the most relevant emission pathways, and support the selection of the most suitable Machine Learning algorithm.

4.3.1. Outliers detection

The analysis of the observations for the four emission pathways, as well as the comparison between observations and G-res estimates (Figure 4.1), demonstrates the high variability of fluxes measurements, characterized by limited high emission values. While recent studies emphasize the possibility of substantial reservoir GHG releases, particularly in tropical areas and eutrophic reservoirs (De Faria and Barros, 2015), the limited number of available measurements, along with their high uncertainty and inconsistency with other observations, suggests that these values should be treated as outliers.
In order to account for the potential occurrence of high emissions and their variability within climatic zones, outlier detection is performed separately for each zone using the *Median Absolute Deviation* (MAD) technique. This method is particularly well suited to our analysis, being insensitive to the presence of outliers and immune to the sample size (Leys et al., 2013).
The MAD is defined as:

$$MAD = b\ M_i\left(|x_i - M_j(x_j)|\right) \tag{4.1}$$

where x_i is the original observation, $M_j(x_j)$ is the median of the series and b is a correction factor to account for skewed and heavily-tailed distributions.
An observation x_i is labeled as an outlier according to the decision criterion:

$$\frac{x_i - M}{MAD} > |\pm threshold| \tag{4.2}$$

The larger the threshold, the more lax the algorithm.
A common value for the threshold is 3. However, in this study, we adopt more permissive threshold values to account for the few but still realistic high emission cases. Table 4.3 shows the number of outliers identified in each climatic zone, along with the threshold values applied to the different gas emission pathways. For diffusive fluxes, values exceeding approximately one order of magnitude relative to the MAD are considered outliers. For ebullitive and degassing fluxes, the threshold is increased to roughly two orders of mag-

nitude relative to the MAD, in order to account for the high variability and uncertainty observed in the very limited sample size.

	Boreal	**Temperate**	**Tropical**	**Subtropical**
CO_2 diffusion				
Threshold	6.745	6.745	6.745	6.745
N outliers	1	1	0	0
CH_4 diffusion				
Threshold	6.745	6.745	6.745	6.745
N outliers	1	9	2	2
CH_4 bubbling				
Threshold	21.3	21.3	21.3	21.3
N outliers	1	3	0	1
CH_4 degassing				
Threshold	21.3	21.3	21.3	21.3
N outliers	0	4	0	1

Table 4.3: Outlier thresholds and number of outliers by emission pathway and climate zone.

A total of 26 observations are excluded: 2 of CO_2 diffusion, 14 of CH_4 diffusion, 5 of CH_4 bubbling, and, finally, 5 of degassing. Following the outliers' detection and the exclusion of the negative emissions, the observations for the four emission pathways are, respectively, 220, 169, 50 and 42.
From this stage onward , all the subsequent analysis and computations are performed using the refined version of the dataset.

4.3.2. Emission pathways relevance

As previously mentioned, the number of available observations is relatively small, with the scarcity being particularly severe for CH_4 bubbling and degassing. Figure 4.2 clearly highlights the discrepancy between the pathways, comparing the reservoirs for which emission data are available with the total number of reservoirs in the dataset.

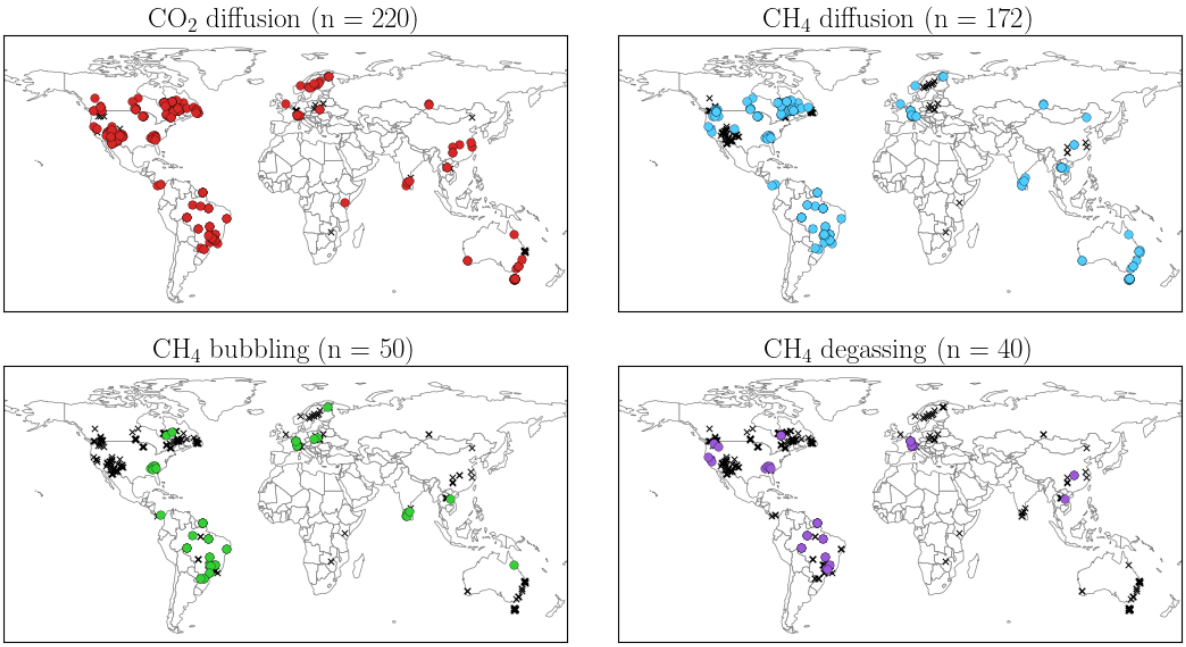

Figure 4.2: G-res training dataset coverage for the four emission pathways. n counts the number of observations for each type of flux.

Since Machine Learning algorithms rely heavily on the amount and quality of input data, a limited dataset may lead to overfitting and model instability, thereby reducing model reliability and its ability to generalize.
To address this limitation, an analysis of emission pathway relevance is conducted to evaluate the relative importance of all four emission components. The goal is to assess whether excluding bubbling and degassing from the modeling effort is justifiable, given the potential uncertainty their limited data may introduce. The analysis of the magnitude of the four emission paths starts with the study of their ranges, shown in the boxplots of Figure 4.3.

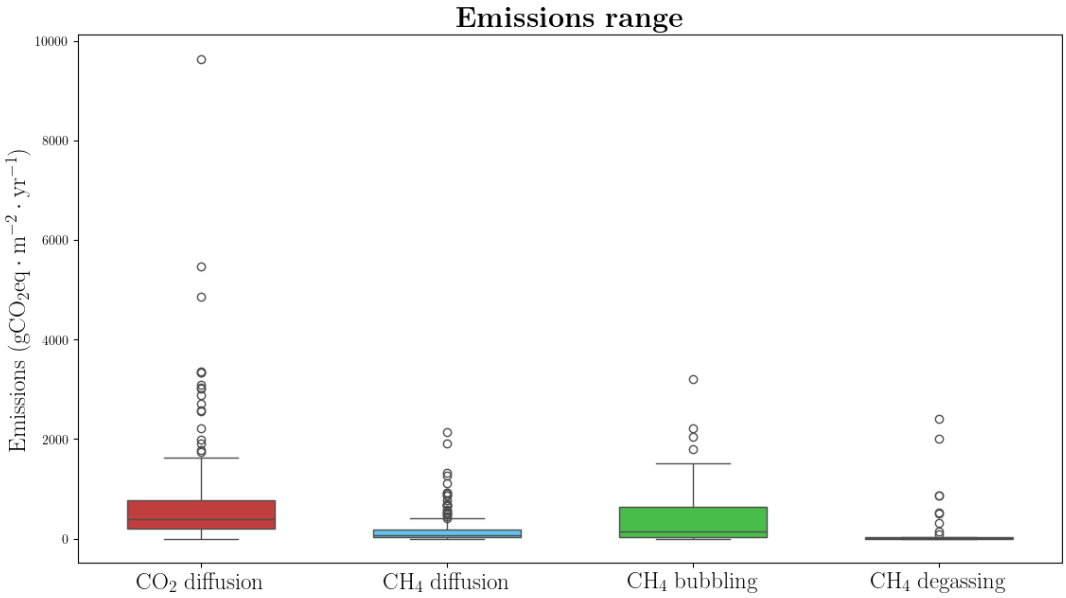

Figure 4.3: Range of values for the four greenhouse gas emission pathways observed in field measurements in gCO_{2eq} m^{-2} yr^{-1}.

The highest values are observed for diffusive CO_2, which also shows the highest mean and median among all distributions. However, its distribution is also the most widespread, with several extreme values lying far outside the interquartile range.
The range of CH_4 emissions appears to be narrower, with CH_4 bubbling emerging as the second most impactful emission pathway. However, these conclusions may be biased by the limited number of available observations, which may not accurately capture the actual distribution of emissions. Moreover, Figure 4.2 shows that data on CH_4 bubbling and degassing emissions are not globally distributed, but are mainly clustered in South America. This spatial bias, combined with the scarcity of measurements, further reduces the representativeness of the dataset and may compromise the robustness of the findings.
In order to extend the sample size of the analysis, we additionally include and compare the emission estimates generated by the G-res models. For consistency, this comparison considers only the estimates obtained for the four emission pathway. The resulting plots confirm the patterns already observed in field measurement comparison: the most relevant emission pathway appears to be CO_2 diffusion, followed by CH_4 bubbling and diffusing emissions. However, it is important to consider the uncertainty associated with CH_4 bubbling, also highlighted by the G-res developers (Prairie et al., 2017).
Given the predominance of diffusive processes and the relevance of data in ML training, we chose to focus on CO_2 and CH_4 diffusive fluxes for the new models development.

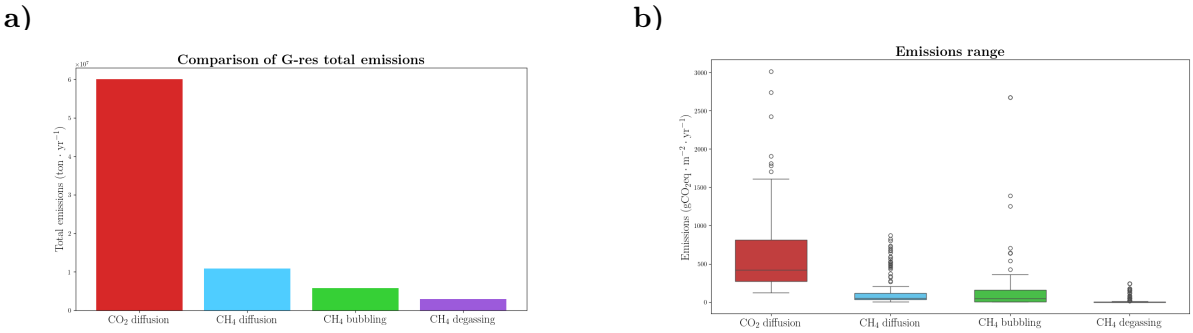

Figure 4.4: Cumulative values (panel *a*) and statistical distribution (panel *b*) of reservoir GHG emissions estimated by G-res by gas and pathway.

4.4. Machine Learning models

After analyzing and filtering the G-res dataset, the process proceeds to the development of new Machine Learning models.
As an initial step, these models are trained using the predictors included in the G-res dataset.
Although a greater number of observations is available for diffusive emissions compared to bubbling and degassing pathways, the overall sample size remains relatively limited and highly variable. This constraint substantially affects the entire model-building process, shaping both the choice of the Machine Learning algorithm and the design of the implementation pipeline.

4.4.1. ML Algorithm

The selection of ML algorithms to be tested is based on a comprehensive literature review, aimed at identifying those considered most suitable for small datasets characterized by a wide range of values. In particular, Support Vector Regression (SVR), Adaptive Boosting (AdaBoost), Ridge Regression (RR), Random Forest (RF), and Extreme Gradient Boosting (XGBoost) have proven to be especially effective for this task (Almeida and Coelho, 2023). All of these algorithms are evaluated through a careful hyperparameter optimization process, designed to balance model flexibility against the risk of overfitting.

Among the tested algorithms, the one that provides the best performance is the SVR, whose theoretical principles are briefly described in Section 3.3.1.
Model training and optimization are carried out using *scikit-learn*, a widely adopted Python library for Machine Learning. It allows three different implementations of Support Vector Regression: SVR, νSVR and LinearSVR. LinearSVR provides a faster implemen-

tation than SVR but only considers the linear kernel, while νSVR implements a slightly different formulation than SVR. In this study, we adopt the standard SVR implementation, which allows us to control both the kernel flexibility and the error margin.
This implementation requires specifying the kernel type among linear, polynomial, radial basis function (RBF), and sigmoid and its specific parameters that define its complexity. Additionally, the hyperparameters C and ϵ must be tuned: C regulates the trade-off between model complexity and training error, while ϵ defines the width of the margin within which no penalty is applied in the training loss function.
The following section describes the implementation of the SVR algorithm, detailing the steps from hyperparameter tuning to the development of the final models.

4.4.2. Pipeline for SVR Model Tuning, Training and Validation

The kernel-shape and hyperparameters tuning is automatically selected through the nested cross-validation. Based on the k-fold cross-validation, it attempts to overcome the problem of overfitting the training dataset treating model hyperparameter tuning as part of the model itself (Brownlee, 2021). The nested cross-validation procedure involves two cross-validation loops: an outer loop and an inner loop. At each iteration, the outer loop splits the dataset into k folds, using k - 1 folds for training and the remaining one for testing. The inner loop further splits the outer training set into k folds and is used to select the best hyperparameters through cross-validation. The model trained with the best hyperparameters is then evaluated on the outer test fold.

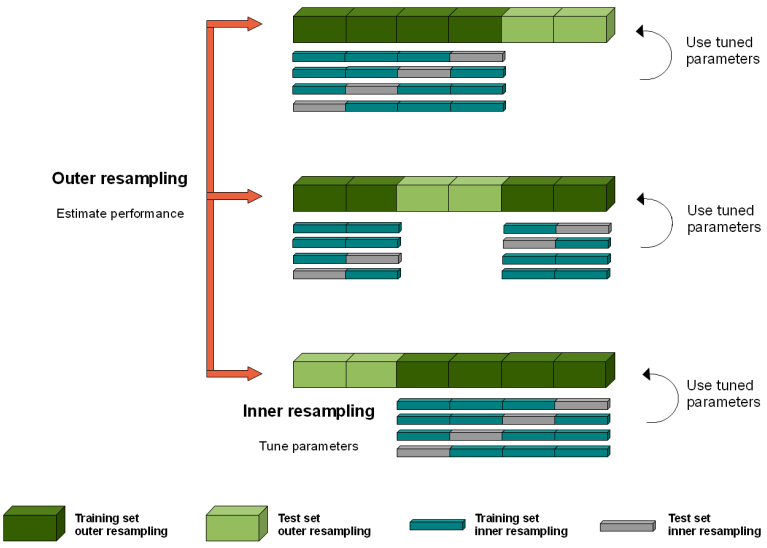

Figure 4.5: Representation of nested cross-validation workflow (Jin, 2018).

The procedure of data splitting and model initialization is inherently stochastic; the starting points, known as seeds, are selected randomly and can lead to variations in model performance. This aspect is particularly relevant to our study. The observed wide variability in the range of target values strongly influences both the hyperparameter tuning phase and the evaluation of model performance. This effect is amplified by the small size of the dataset, making the outcome highly sensitive to how the data are split into training and test sets.
To overcome this issue, we adopt two-step strategy. First, an outer loop is defined to partition the dataset into a training set (80% of the total data) and a test set (20% of the total data). The training set is used for hyperparameter tuning and training. Tuned hyperparameters are selected through nested cross-validation, where the inner loop is repeated using 10 different random seeds to further strengthen the robustness and reliability of the parameter selection process. For each hyperparameter combination, the mean R^2 score obtained across all folds of the inner cross-validation is computed, and the combination yielding the highest average performance is selected.
Once determined, the entire training set is used to train the final model in order to exploit the full training sample. The evaluation is then carried out on the unseen test set. This procedure is repeated 25 times, generating 25 different train-test splits.
Using the hyperparameters and predictors that achieve the best overall performance across the 25 repetitions, a single final model is then trained on a new training–test split and evaluated on the unseen test set.

Repeating the training process across 25 different training–test splits allows for the identification of both hyperparameters and covariates sets that aim to achieve stable and consistent performance, regardless of how the data is partitioned. This step is essential considering the limited data availability. Flux observations cover a wide range of values, and the resulting data splits can differ substantially. Moreover, the dataset does not uniformly represent the spatial domain (Figure 4.2), with splits being particularly influenced by the limited number of observations in tropical areas. The combination of measurement uncertainty and scarce information prevents a precise assessment of the frequency of high and low fluxes, as well as of their actual magnitude. For this reason, establishing a model configuration that ensures consistently good performance across different data splits is essential to address data uncertainty and to obtain reliable models regardless of the underlying distribution. Once this stability is assessed, the final model is trained using the selected configuration on a new single training–test split. Given the consistency observed across the 25 iterations, this final model is expected to exhibit similar performance and can be considered suitable for further applications.

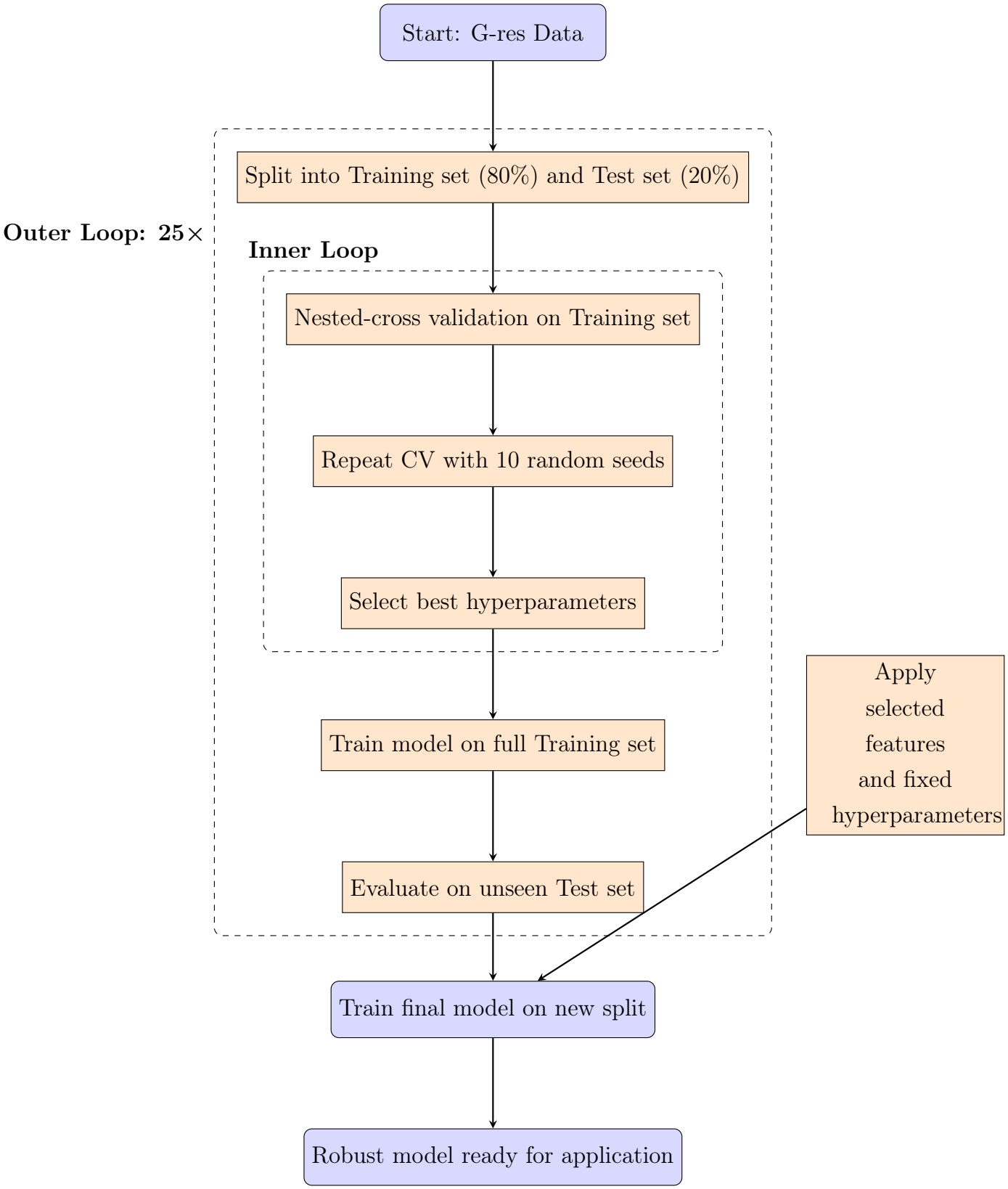

Figure 4.6: Pipeline for SVR model tuning, training and validation.

Feature selection is manually performed, considering both the redundancy of the variables and their relative importance in the models. Since multicollinearity among variables can negatively affect model performance by introducing overfitting and instability, variables correlation is investigated to exclude redundant variables from the pool of input covariates. It is defined using Spearman rank correlation coefficient, which measures the strength and direction of a relationship between two variables without assuming linearity. Cumulative radiance and latitude, littoral area and mean depth, and temperature and latitude show strong negative correlation. In contrast, the highest positive correlations are observed for the pairs discharge and catchment area and soil carbon and latitude (Figure 4.7).
These findings are integrated with results from permutation importance analysis, which quantifies the impact of each input variable on the model output. The metric is computed across the 25 loops to assess which variables consistently influence the flux predictions, regardless of the specific split.
The final set of variables is selected based on the combination that provides the most stable and efficient model performance, identified by repeatedly applying the procedure described above with different covariates combinations.

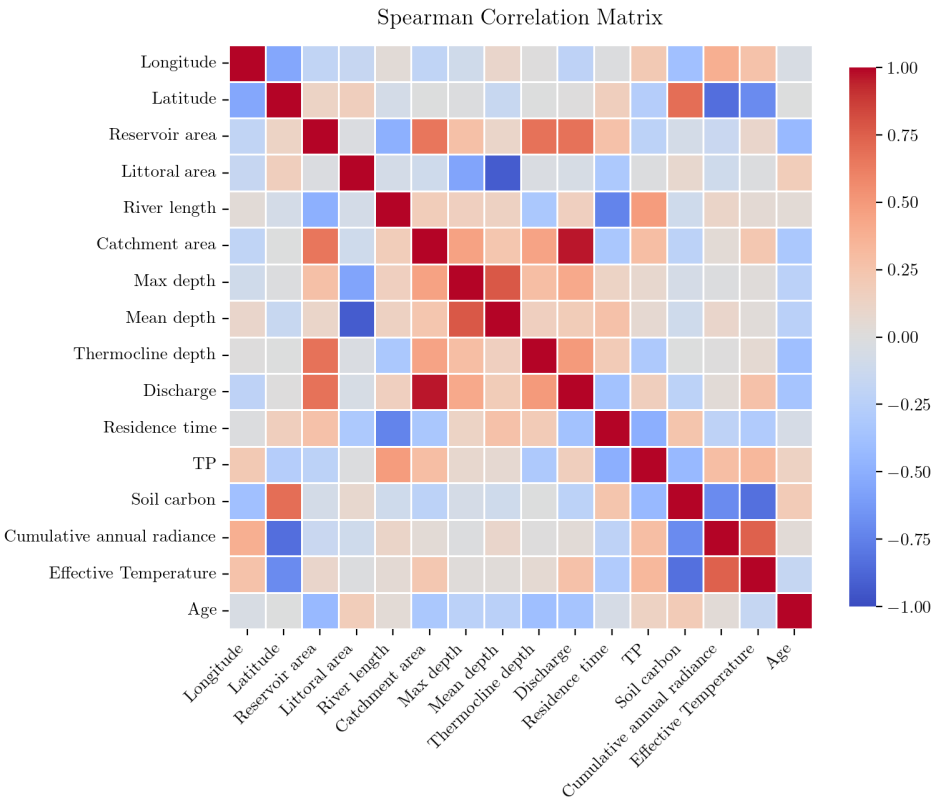

Figure 4.7: Spearman correlation matrix among the predictors included in G-res dataset.

4.5. Algorithm design

The implementation of the algorithm is carried out separately for the SVR models of CO_2 and CH_4. Treating each model individually allows for tailored customization, ensuring a more accurate representation of the specific flux dynamics of the two gases. Gas-specific customization is defined by testing different settings in the first loop of the procedure described in Section 4.4.2. The alternative SVR implementations are performed using the G-res dataset and subsequently compared with each other.
Once the SVR implementation settings were established, the procedure described in Section 4.4.2 was applied to both the G-res dataset and the newly compiled dataset to build the models.

The following sections describe the specific design adopted for CO_2 and CH_4 modeling, outlining the rationale behind each methodological choice.

4.5.1. CO_2 diffusion Model: SVR with Weights

CO_2 emissions exhibit a wide range, with relatively few but realistic high-emission values (Figure 4.3). The underestimation of these values represents one of the main limitations of G-res and other linear models, ultimately leading to an underestimation of the overall carbon footprint of reservoirs. Improving the representation of these high-emission cases has been identified by the model developers themselves as a necessary advancement in emission modeling.

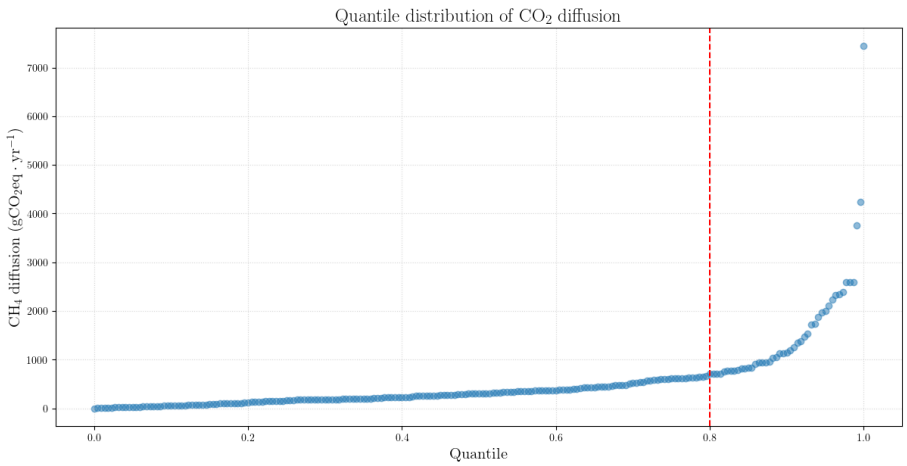

Figure 4.8: CO_2 diffusion quantile distribution

To better understand the distribution of these values, we first plot the quantile distribution to identify where high values lie.
As shown in Figure 4.8, the distribution slope increases exponentially after the 0.8^{th} quantile, which can therefore be defined as a threshold.

To cope with the skewed distribution of the observations and better describe high emission values, we try to assign a greater importance to the samples above the 80^{th} percentile in the model training phase. Hence, sample weighting is introduced in the SVR objective function. A sample-specific weight vector assigns a weight of 1.5 to all observations at or above the 80^{th} precentile of y, while all other observations are assigned a weight of 1.0. Formally, the weight vector w is defined as:

$$w_i = \begin{cases} 1.5 & \text{if } y_i \geq \text{Percentile}_{80}(y) \\ 1.0 & \text{otherwise} \end{cases} \tag{4.3}$$

This approach effectively modifies the empirical risk minimization objective by placing greater emphasis on the prediction errors associated with high-emission values.
The original SVR formulation (Eq. 3.15) is therefore modified as follows:

$$\min \left(\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{\ell} w_i(\xi_i + \xi_i^*) \right) \tag{4.4}$$

where w_i is the weight assigned to the i^{th} observation.
Assigning higher weights to observations in the upper tail of the distribution helps partially correct the skewness of the dataset. Although this strategy does not alter the composition of the data, the differentiated weighting of high-emission and mid-range samples in the loss function effectively forces the model to consider the former relevantly, despite their relative scarcity.
On one hand, this approach enhances the model's sensitivity to the upper tail; however, it may also affect performance on lower values, potentially resulting in their overestimation. For this reason, the weighting scheme is calibrated to achieve a reasonable trade-off between reducing the underestimation of high-emission cases and preserving overall predictive accuracy.
Figure 4.9 illustrates the impact of weighting by comparing the predictions of the standard and the weighted SVR, for one of the 25 data splits. The results suggest that weighting not only improves the prediction of the high emissions, but in some cases can also enhance the estimates of the middle and low emission values.

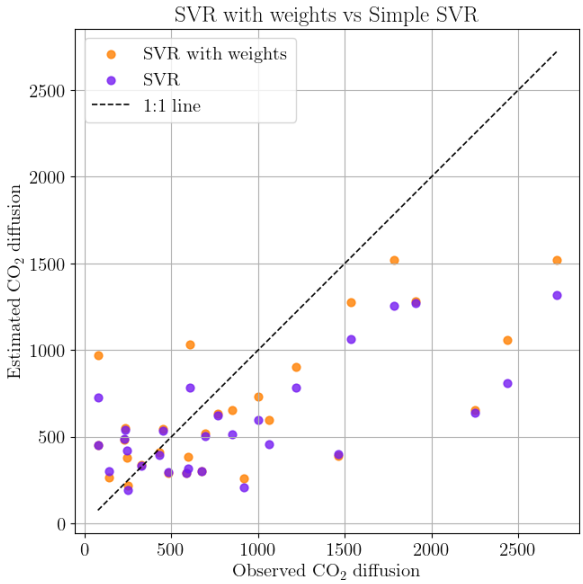

Figure 4.9: Predicted emissions (gCO_{2eq} m^{-2} yr^{-1}) from SVR and weighted SVR on one randomly selected test set among the 25 cross-validation splits.

4.5.2. CH_4 diffusion Model: SVR with logaritmic transformation

CH_4 fluxes span a much narrower range compared to CO_2. While CO_2 emissions can reach up to 7000 mgC m^{-2} d^{-1}, CH_4 values typically remain below 50 mgC m^{-2} d^{-1}. However, the distribution of CH_4 fluxes is still skewed (Figure 4.10), with relatively few but impactful high values.

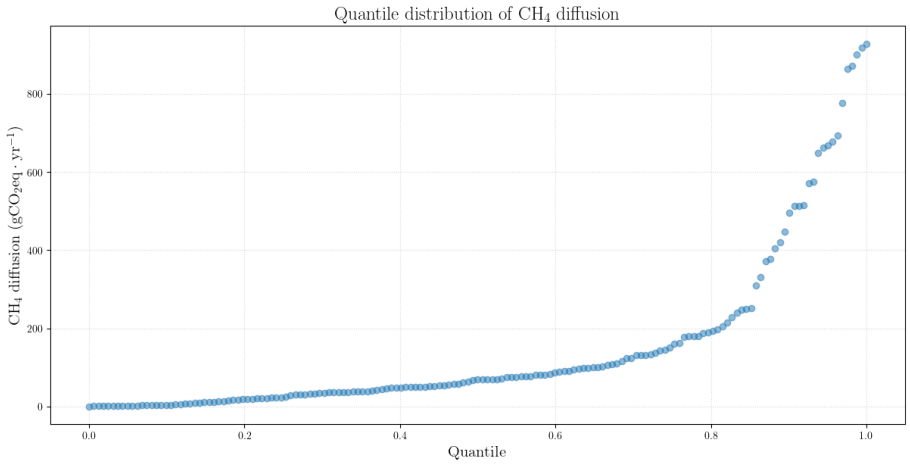

Figure 4.10: CH_4 diffusion quantile distribution

This asymmetry affects the error minimization process, pushing the model to reproduce an average behavior that fails to represent the extremes of the distribution, leading to an overestimation of the lowest values and an underestimation of the highest ones.
To reduce the model's bias towards the center of the distribution a logarithmic transformation is applied to the target variable before training.
Formally, the SVR algorithm solves:

$$\min\left(\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{\ell}(\xi_i + \xi_i^*)\right) \quad \text{subject to:} \quad \begin{cases} \log(y_i) - f(\mathbf{x}_i) \leq \varepsilon + \xi_i \\ f(\mathbf{x}_i) - \log(y_i) \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \tag{4.5}$$

and then:

$$\hat{y} = \exp(f(\mathbf{x})) \tag{4.6}$$

Figure 4.11 compares the performance of the standard SVR and the SVR trained on log-transformed targets, illustrating the general prediction trends and the enhancement achieved through logarithmic transformation.

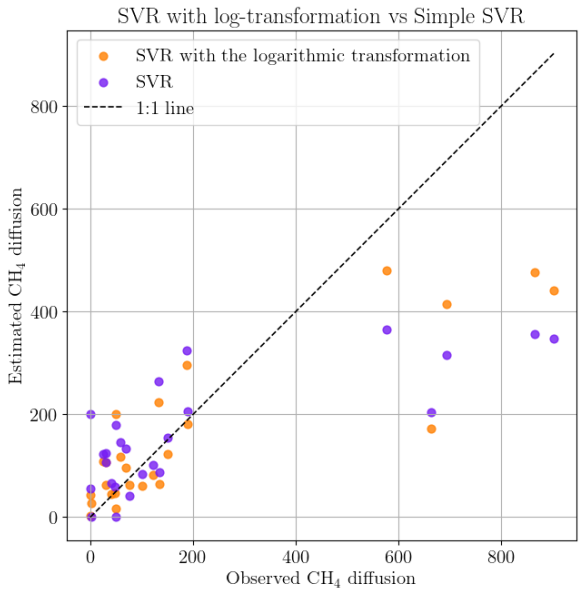

Figure 4.11: Predicted emissions (gCO_{2eq} m^{-2} yr^{-1}) from SVR and SVR trained on log-transformed target on one randomly selected test set among the 25 cross-validation splits.

4.6. New dataset

While the first part of our work tested the capability of ML algorithms to reproduce emission processes, the second part evaluated their potential to estimate emissions using an alternative dataset to G-res, based on easily accessible variables.
To this end, we generated a new dataset integrating several promising covariates. It contains four types of information which can be classified as reservoir characteristics, climatic variables, basin descriptors, and soil biogeochemical properties. These features are extracted from publicly available global datasets and can be directly used as potential covariates in the SVR modeling framework, marking the key distinction from the G-res variables (Sections 3.2.1 and 3.2.2).

Reservoir information includes geographic coordinates (Longitude and Latitude), year of impoundment, surface area (km^2), and mean depth (m) and are collected from the G-res GRanD dataset (Global Dam Watch).
Hydrological basins are delineated for each reservoir using GIS tools and global spatial datasets. Then, for each basin, the total area (km^2), population, and land cover composition are defined. The inclusion of the latter two is motivated by their influence on the organic inputs received by reservoirs. Particularly, croplands and forests are associated with high biological activity, which can contribute significant amounts of organic matter to downstream water bodies. Similarly, urban areas and high population density often indicate substantial wastewater production, typically characterized by elevated organic content.
Climatic variables include temperature, wind speed, precipitation, cumulative solar radiation, and surface runoff. Their selection reflects their involvement in the physical processes described in Chapter 2 and is consistent with findings reported in the scientific literature (Section 3.1.1). Temperature has been shown to favour GHG fluxes (Paranaíba et al., 2018) while wind speed and precipitation play a key role in gas exchange at the water–atmosphere interface (Yang et al., 2024). Solar radiation influences biological processes (Prairie et al., 2018), increases water temperature, and affects reservoir stratification; runoff, along with precipitation, serves as a proxy for the rate of organic matter input into reservoir waters. The basin soil biogeochemical profile is described by the content of organic carbon and total nitrogen, the availability of phosphorus, the input of nitrogen and phosphorus fertilizers, and the pH of soil water. These variables serve as proxies for the potential availability of organic and nutrient-rich inputs that support microbial activity contributing, consequently, to greenhouse gas production.
Table 4.4 provides an overview of the variables included in the generated dataset along

with corresponding units of measurement and source; details on their extraction procedures are presented in the subsequent paragraphs.

4.6.1. Catchment

The first step in the construction of the new dataset is the definition of hydrological basins. Their delineation is crucial for extracting all other variables included in the dataset, excluding the reservoir features.

The extraction procedure is divided into three main steps: flow accumulation and direction maps creation, basin extraction, and data validation.
Flow accumulation and flow direction maps are generated by *r.watershed* module, a specific GIS tool within GRASS GIS. Using as input a *Digital Elevation Model* (DEM) it implements the D8 (Deterministic 8) method for calculating flow direction. The algorithm assumes that each cell in the DEM drains into one of its eight adjacent cell and codifies the flow direction into the 8 cardinal direction. Once flow direction is defined, flow accumulation is calculated by counting, for each cell the number of upstream cells that drain into it.
In our analysis, we supply to *r.watershed* the DEM produced by HydroSHEDS, a global database that offers digital data layers suitable for hydro-ecological application worldwide (HydroSHEDS Team). Prior to processing, artificial depressions in the elevation model are removed using a GIS tool for sink filling to avoid creation of artificial flow accumulation areas. Then, *r.watershed* is applied. We produce flow accumulation and direction maps separately for each continent, in order to reduce the computational load and further improve processing performance.
The corresponding basins of the reservoirs included in the G-res dataset are extracted through *r.water.outlet*, a powerful GRASS GIS module used for delineation of individual catchment draining to one outlet point. For accurate basin outline, it is essential that the outlet points, corresponding in our study to the G-res dams, are located on the cells of maximum flow accumulation within their respective reservoir areas. These points are automatically identified through a Python script that uses as input the flow accumulation map and the dams coordinates, and searches for the cells with the highest accumulation within a 0.05° window centered around each dam. Figure 4.12 shows a zoomed-in view over an area in North America, illustrating the improved alignment of dam locations with the drainage network resulting from the maximum flow accumulation search.

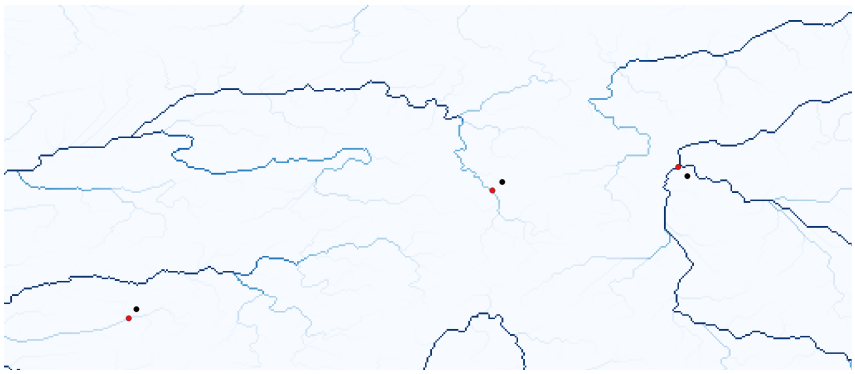

Figure 4.12: Repositioning of maximum accumulation points onto the correct locations of the drainage network (G-res dams coordinate in black and max accumulation points in red).

Once corrected the outlet points, the module *r.water.outlet* extracts the catchments. The module is executed via the Python console, which allows both the extraction of catchments and the calculation of their areas (km^2). The resulting basins are then compared with the G-res values and validated through literature review and visual inspection using Google Earth map. If inconsistencies are found, basin corrections are performed manually by repositioning the outlet points in their correct locations.
Figure 4.13 shows the post-correction comparison between the extracted and G-res basin areas. Some differences still remain and correspond to G-res values that are inconsistent with those reported in the literature.

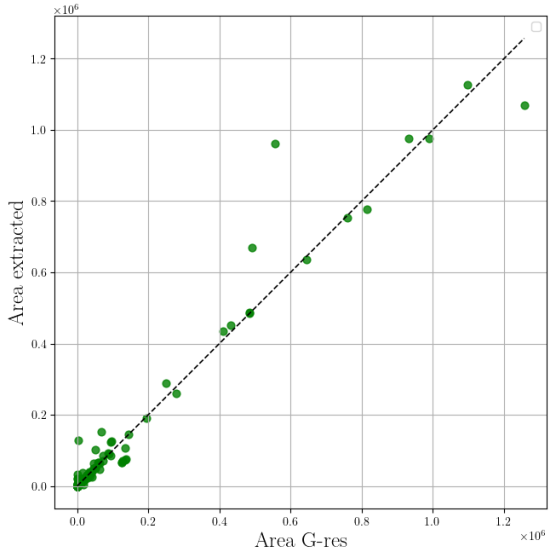

Figure 4.13: Comparison between the extracted basins area and the G-res published values in km^2 (Extracted area in y-axis, G-res in x-axis).

4.6.2. Reservoir area and depth

The spatial footprint of global reservoirs is defined in the GRanD database, which also provides a shapefile for their visualization. This shapefile is used to extract the reservoirs included in the G-res dataset. Starting from the dams coordinate in G-res dataset, an automatic procedure implemented in Python couples G-res dam coordinates to the corresponding GRanD reservoir. The algorithm assigns each dam to the reservoir whose centroid is closest to the dam location, and discards the match if their distance exceeds a threshold of 50 km. The implementation of this procedure is necessary due to the lack of a one-to-one correspondence between dam coordinates and names in the GRanD and G-res datasets, that makes a direct match based only on these attributes unfeasible. Reservoirs of mismatched or unmatched dams are manually reassigned or identified using satellite maps from Google Earth and data from literature. Finally, results are validated using the same approach described for the catchment areas (Section 4.6.1).
Reservoir depth is extracted from G-res dataset; missing values are completed using the GRanD database, information from literature, or estimated using Google Earth tools.

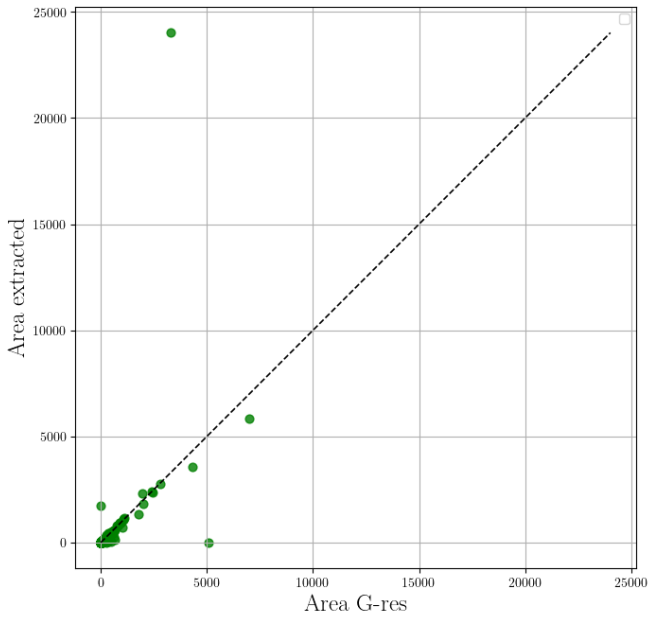

Figure 4.14: Comparison between the extracted reservoirs area and the G-res published values in km^2 (Extracted area in y-axis, G-res in x-axis).

4.6.3. Land cover and Population

Land cover composition and population are, in addition to the area, the features of interest for upstream basins; as mentioned above, they could be representative of the organic

inputs into the reservoirs.

The land cover map is downloaded from NASA EARTHDATA (NASA Earthdata, 2024), which provides global land cover maps from 2001 to the present with a spatial resolution of 500 m. Considering the relatively slow variation of the landscape over time and that fluxes observations span in an interval of only 30 years, from 1988 to 2016, a constant land cover is assumed. Specifically, we use the 2003 map, which is the median year of the observation period. The map has a multi-layer structure, with each one corresponding to a different land cover classification scheme. In line with our analysis, we use the IGBP (International Geosphere-Biosphere Programme) classification, which describes landscape composition using 17 classes: Evergreen Needleleaf and Broadleaf Forest, Deciduous Needleleaf and Broadleaf Forest, Mixed Forests, Closed and Open Shrublands, Woody and Herbaceous Savannas, Grasslands, Permanent Wetlands, Croplands, Urban and Built-Up Lands, Natural Vegetation Mosaics, Snow and Ice, Barren or Sparsely Vegetated Lands, and finally Water Bodies. The global map is clipped with the catchment shapefile (Section 4.6.1), and the area of each land cover class within the basin is then extracted and normalized by the total basin area to compute land cover fractions(%).

Population data are obtained from the Copernicus Global Human Settlement Layer – Download Portal (European Commission, Joint Research Centre (JRC)), an open-access data source for assessing human presence on the planet. The spatial raster dataset, which includes the distribution of the residential population is called GHS-POP. It is expressed as the number of people per cell and has a resolution of 100 m. Population information is available from 1975 to the present, with a temporal resolution of 5 years. Considering the recent global population growth, especially in emerging countries, the temporal variability of the population is maintained in the new dataset. For each reservoir, the closest GHS-POP raster year to the sampling campaign year is selected, choosing the most recent one in case of equal distance. The information included in the compiled dataset is the population density, obtained dividing the total population within the basin by the basin area.

4.6.4. Biogeochemical soil profile

The biogeochemical characteristics of the catchment soil play a key role in driving GHG emissions from reservoirs, as they contribute as organic matter that may be directly converted into CH_4 and CO_2 or participate in microbial and chemical transformation processes.

Organic carbon content (gC kg^{-1}), total nitrogen (gN kg^{-1}) and pH (-) of groundwater are

provided by ISRIC (ISRIC – World Soil Information, 2024). Using SoilGrids (Global Gridded Soil Information), it produces global maps with a 30-arcsecond resolution ($\approx$ 1 km), each containing a specific soil property. Information is available for both TOP (0 cm - 20 cm) and SUB (20 cm - 1 m) soil layers. In our dataset, both layers are included in order to explore all the potential relationship between soil features and GHG emissions. Median and average values of the three characteristics are collected for each catchment.

Top-soil Olsen phosphorus, which represents the fraction available to plants, is included as indicator of soil fertility and biological activity within the basin. Data comes from Global Available Soil Phosphorus Database (McDowell, 2023). It is produced through a predictive model ($R^2 \approx 0.54$) which combines in-situ soil samples, soil bulk density and other soil properties to produce continuous global map of Olsen-P with a 30 arcsecond resolution.
The map is used to extract the data for each reservoir catchment; the dataset provides the average, median and total phosphorus values (mg kg^{-1}) computed at the basin-level.

Global nitrogen and phosphorus fertilizer application is a proxy for agricultural productivity in croplands. Lu and Tian (2017) produced global maps of annual fertilizer use combining crop-specific fertilizer use rates and harvested area maps to produce global maps with a 0.5° ($\approx$ 25–55 km depending on the location on the globe) resolution from 1961 to 2013.
Our dataset assigns to each reservoir basin-specific fertilizer data extracted from the annual map corresponding to the fluxes sampling campaign year. As representative statistics at the basin level, both the average and median are reported, expressed in units of gN m^{-2} yr^{-1} and gP m^{-2} yr^{-1}, respectively.

4.6.5. Climatic variables

Climatic data are downloaded from ERA5 (Copernicus Climate Change Service (C3S), 2017), the latest global climate reanalysis produced by the European Centre for Medium-Range Weather Forecasts (ECMWF). Reanalysis combines model data with observations from across the world using the laws of physics into a globally complete and consistent dataset. The principle of reanalysis is called assimilation. Observations are optimally combined with model estimates to produce the best possible representation of the state of the atmosphere, land and oceans. This is the same method used for weather predictions but, compared to this, the lower resolution of reanalysis products allows them to extend further back in time, incorporating consistent historical observations. ERA5 provides hourly and monthly estimates for a large number of atmospheric, ocean-wave and land-

surface quantities with a resolution of 0.25° (≈ 14–28 km).

Among the available variables, we downloaded monthly estimates of 2m-temperature, 10m-wind speed, surface downward solar radiation, runoff, and total precipitation for the period corresponding to the time span of the emission observations. To ensure the temporal consistency between environmental conditions and observed emissions, variable values are extracted from the annual dataset corresponding to the year of the fluxes sampling campaign.

2m-temperature represents the temperature of air at 2 m above the surface of land, sea or inland waters expressed in Kelvin (K); 10m-wind speed is the horizontal speed of the wind, or movement of air, at a height of ten meters above the surface of the Earth (m s^{-1}). These two variables are directly involved in GHG release from reservoirs, as they influence the physical processes underlying gas diffusion across the water–air interface (Chapter 2). Consequently, data are extracted at the spatial extent of each reservoir.

In contrast, the variables of runoff, precipitation, and solar radiation mostly affect emissions indirectly, by altering organic inputs and nutrient loads. These factors affect the production and transport of organic matter within the basin into the reservoir. Accordingly, data for these variables are extracted at the catchment level.

Table 4.4: List of variables used in the new dataset, including units and data sources.

Variable	Unit	Source
Name	–	G-res
Impoundment year	year	G-res
Longitude	decimal degrees	G-res
Latitude	decimal degrees	G-res
Catchment area	km^2	Extracted
Reservoir area	km^2	GRanD / Literature
Mean depth	m	G-res / Literature
Sampling year	year	G-res
Age	year	Sampling yr – Impoundment yr
Olsen P (mean-median-sum)	mg/kg	ISRIC SoilGrids
Fertilizer N (mean-median)	g N/m^2	Lu & Tian (2017)
Fertilizer P (mean-median)	g P/m^2	Lu & Tian (2017)
Top and Sub-soil Organic carbon (mean-median)	g C/kg soil	ISRIC SoilGrids
Top and Sub-soil Total nitrogen (mean-median)	g N/kg soil	ISRIC SoilGrids
Top and Sub-soil Soil water pH (mean-median)	pH	ISRIC SoilGrids
Annual mean precipitation	mm	ERA5
Annual mean runoff	mm/s	ERA5
Annual cumulative SSRD	J/m^2	ERA5
Maximum wind speed	m/s	ERA5
Minimum wind speed	m/s	ERA5
Mean wind speed	m/s	ERA5
Annual mean temperature	°C	ERA5
Annual max temperature	°C	ERA5
Annual min temperature	°C	ERA5
Months over 0°C	months	Computed from ERA5
Total population	persons	GHSL (GHS-POP)
Land cover classes	%	MODIS MCD12Q1 (IGBP scheme)

4.7. SVR on the newly developed dataset

Once collected the data, the specific implementation of the SVR algorithm is applied to train CO_2 (Section 4.5.1) and CH_4 (Section 4.5.2) models using new covariates.
The model building process follows the pipeline described in Section 4.4.2 including the double-loop structure used to assess the model stability, followed by the training of single final model with the optimal hyperparameters and features set.
Compared to G-res, the new set includes a greater number of covariates and exhibits higher mutual correlations (Figure 4.15). The feature selection plays therefore a key role. It is manually performed considering the same metrics used to prune the G-res variables, namely the Spearman rank correlation coefficient and the permutation importance. Several combinations are tested, varying not only the selected variables but also the descriptive statistics used, precisely the mean, median, sum per basin.

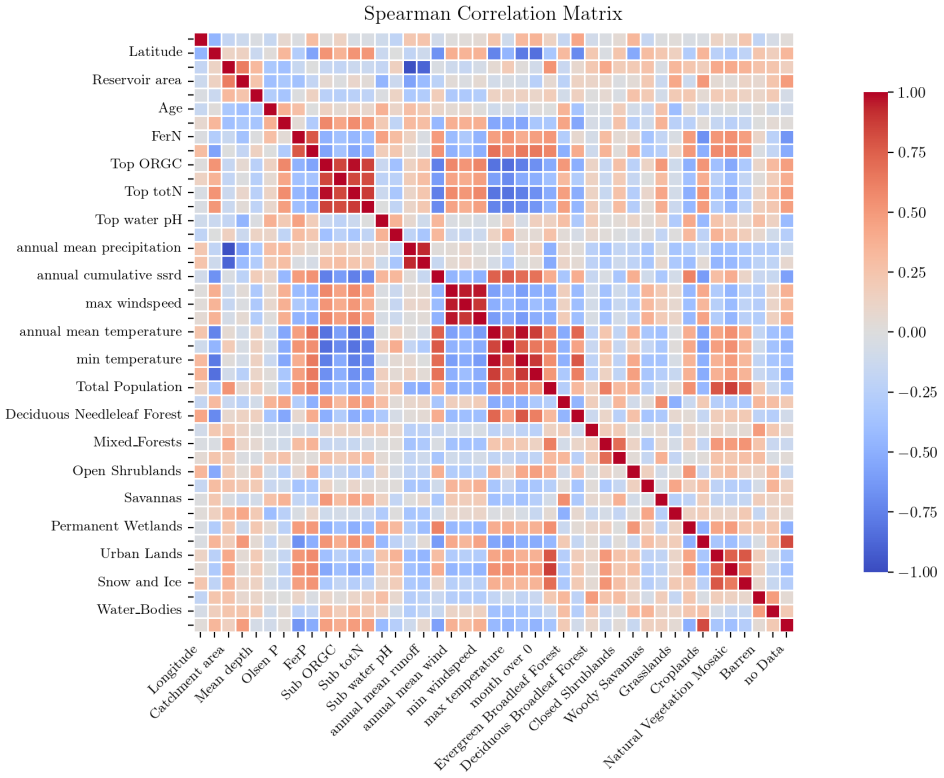

Figure 4.15: Spearman correlation matrix among the predictors included in the new dataset.

5 | Results

In this chapter, the results of the thesis are exhaustively presented.
The structure of the chapter aims to answer the key research questions of the work, namely evaluating the contribution of Machine Learning in modeling GHG fluxes from reservoirs and assessing the potential use of easily accessible variables. Accordingly, the predictive capability of Machine Learning models is investigated both using original G-res variables (Table 4.2) and the new identifed predictor factors (Table 4.4).
Following the modeling pipeline described in Section 4.4.2, the SVR algorithm is firstly evaluated on different subsets in order to assess models generability. Subsequently, the performance of the final model is examined. Moreover, the results are directly compared with the G-res performance, employed as a benchmark across all stages of the ML modeling process.
As a conclusion of our work, we use our models in a real-world application; particularly, they are employed to assess the carbon footprint of European hydropower reservoirs. The analysis provides a quantitative estimate of reservoirs' emissions, demonstrating the practical applicability of the models.

5.1. SVR algorithm applied on G-res dataset

In line with the development of our work, we first evaluate the results obtained by applying the SVR algorithm to the G-res dataset. We repeat the analysis twice, attempting to describe the two processes for which sufficient data is available, thus producing two different final models, one to describe the diffusion process of CO_2 and one for that of CH_4. The outcomes are presented in the following sections for both the CO_2 and CH_4 applications.

5.1.1. SVR algorithm performance

This section presents the performance of the SVR configuration with particular emphasis on its stability.

To select the SVR model features, each combination of hyperparameters and covariates set was evaluated over 25 different data splits. The results obtained for different configurations were then compared on an R^2 basis across the training, test, and full datasets. The selected configuration corresponds to the one that ensures the best R^2 sets, striking a balance between stability and accuracy, and limiting, at the same time, the overfitting across the three datasets.
In the subsequent parts, we report the final set of identified covariates to model the two processes and the respective performances of the SVR models.

CO_2 application

The set of covariates used for modeling CO_2 diffusion, along with their permutation importance, is presented in Figure 5.1.

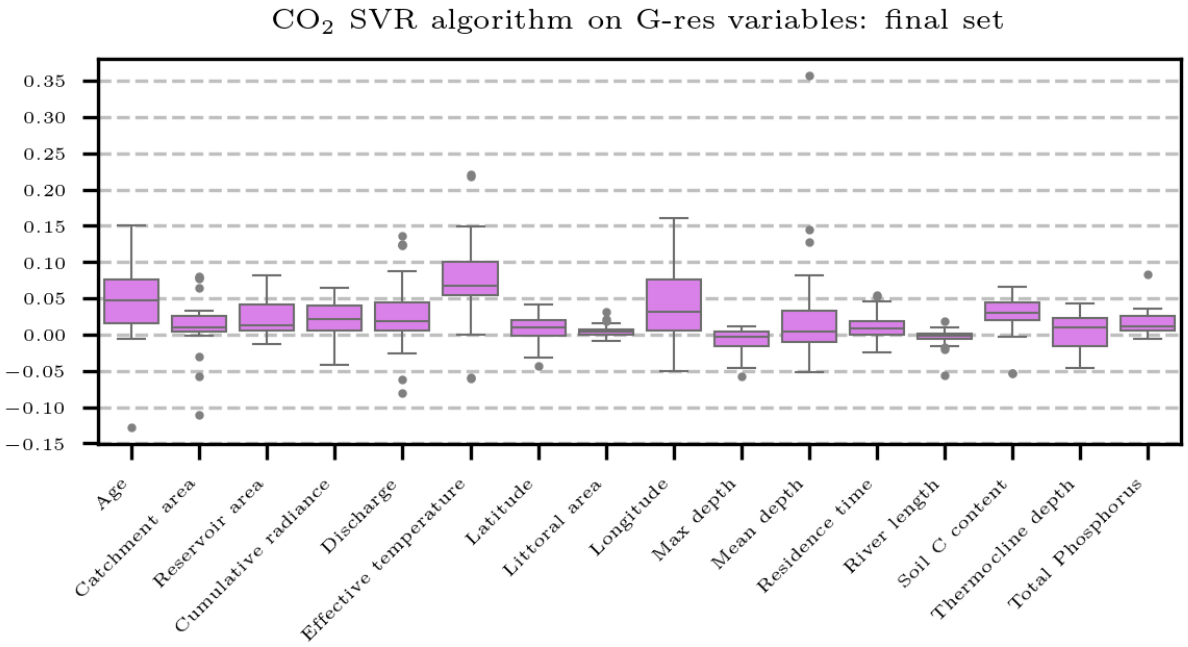

Figure 5.1: Permutation importance of the final set of covariates selected from G-res variables for reproducing CO_2 diffusion.

Among the selected predictors, age, mean depth, and especially effective temperature and soil carbon content show the highest and most consistent permutation importance across the 25 models. These results are in line with the findings of the G-res study, which includes them in the CO_2 model (Eq. 3.5). In particular, the key role played by temperature in CO_2 diffusive flux from reservoirs is in accordance with the conclusions of the G-res developers, who explicitly highlighted its central importance (Prairie et al.,

2017).
Some variables such as latitude and river length exhibit low or even negative importance values in some iterations, suggesting a limited or unstable contribution to the algorithm performance. However, they are retained since their inclusion tends to improve the overall stability of the SVR configuration. They appear to play a relevant role in those data splits that are more challenging to predict, enabling the training of relatively reliable models regardless of the actual data distribution.

The corresponding performance suggests that the SVR algorithm does not show excessively high signs of overfitting on the training data, excluding few samples, and leads to fairly stable performance in the training and entire sets. By contrast, a wider range of R^2 is observed in the test sets, where the lowest R^2 values are achieved. The metric range between [0.54; 0.32] for the training set, [0.58; 0.08] for the test set, and [0.40; 0.28] for the entire set, with average values of 0.40, 0.28 and 0.36, respectively.

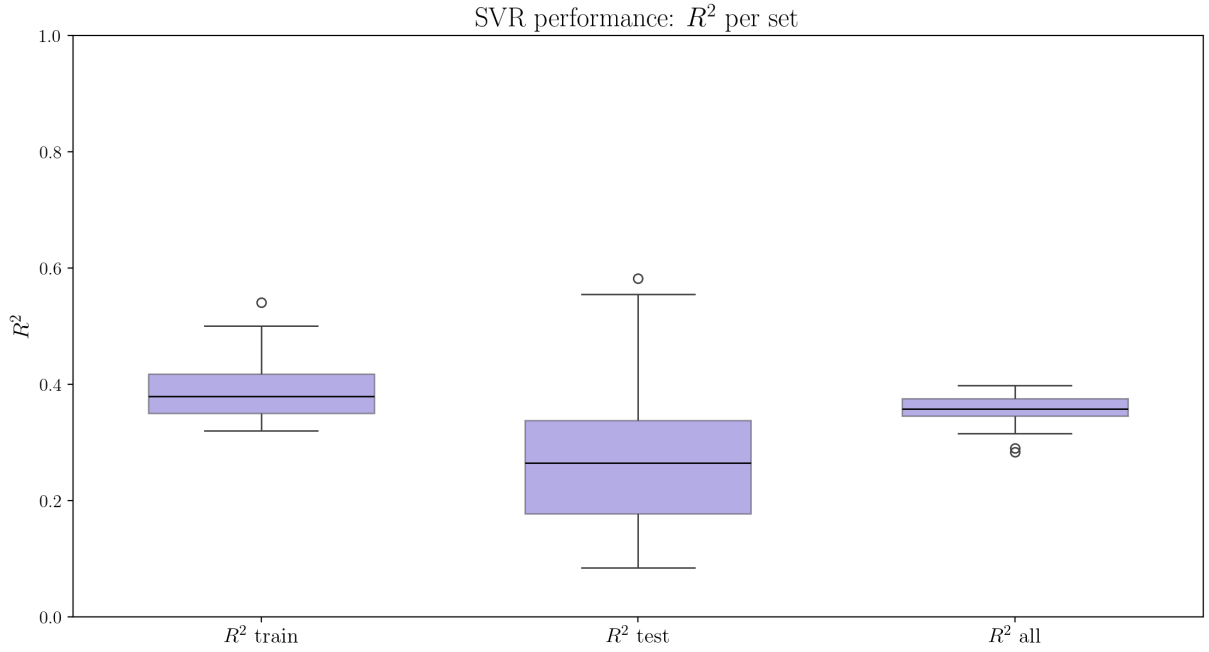

Figure 5.2: R^2 values computed on the training, test and the entire set for the SVR application using G-res covariates to reproduce CO_2 fluxes.

Comparing the R^2 values across the 25 iterations, we identified only four cases that significantly deviate from the others, showing markedly poorer performance on the test set. They are further investigated to identify possible causes; Figure 5.3 provides a visual assessment of SVR predictions compared to observations in these four test sets. Remarkably, the corresponding sets displayed in panel a), b) and c) include an emission value

that is extremely higher than all others (9576 gCO_{2eq} m^{-2} yr^{-1}), which is severely underestimated. Its acute underestimation has a strong negative effect on R^2 computation, reducing the ability of the criteria to represent the actual quality of the models. Aside from the extreme case, in fact, the SVR configuration appears to predict reasonably well the CO_2 emission process from the reservoirs, capturing the overall pattern.
The low R^2 values in these cases are not attributable to a lack of robustness of the SVR configuration. Instead, it is likely driven by the complexity of modeling high emissions.

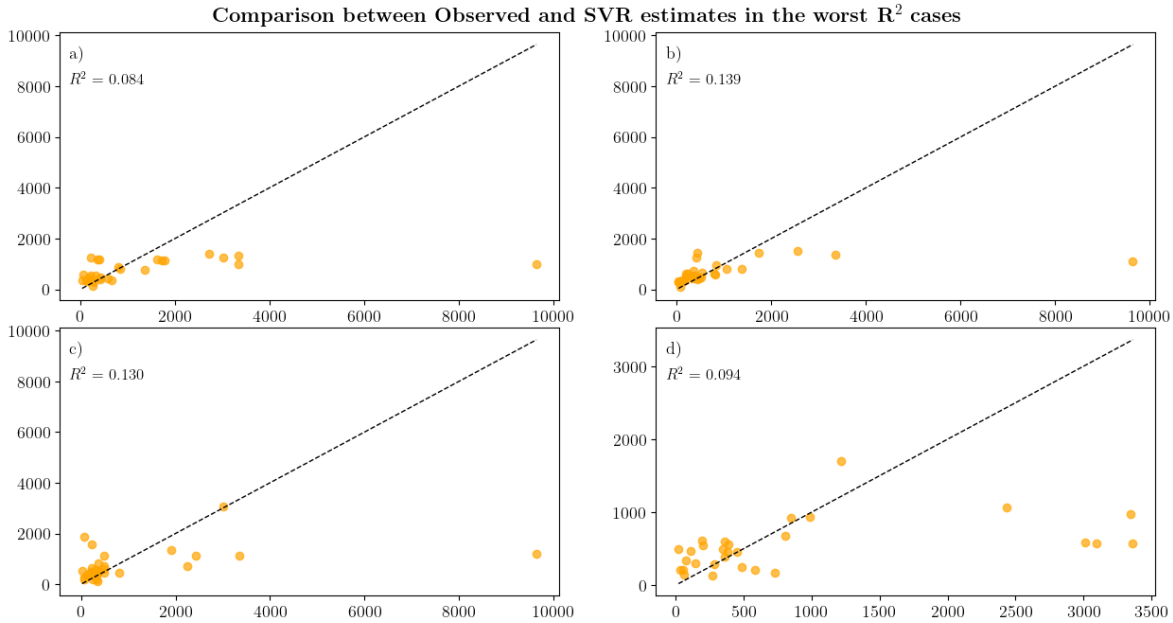

Figure 5.3: Comparison between CO_2 emissions (gCO_{2eq} m^{-2} yr^{-1}) estimated by the SVR using G-res covariates and the observed values in the test set where the SVR algorithm performs worst. The y-axis represents the predicted values, while the x-axis represents the observations.

CH_4 application

The permutation importance of the predictors selected for CH_4 ML model is shown in Figure 5.4.
Our findings are in accordance with the G-res study. Analogously to G-res and CO_2 case, temperature and reservoir age show the highest importance, demonstrating their central role in GHG emission processes. Moreover, both our and G-res analyses do not detect any correlation between CH_4 diffusion and biological variables. The contribution of total phosphorus is almost negligible and soil carbon content shows a relatively low importance. In our analysis, catchment area and discharge also appear relevant. However, the magni-

tude of their contribution is not constant across iterations, indicating that their relevance is significantly sensitive to the data split. This can be observed, albeit to a vary degree, for all factors in the set, especially compared with CO_2 covariate set (Figure 5.1). Such instability is likely related to the smaller size of the CH_4 dataset—which includes 44 fewer samples than the CO_2 training dataset—thereby increasing the sensitivity of ML algorithms to data splits.

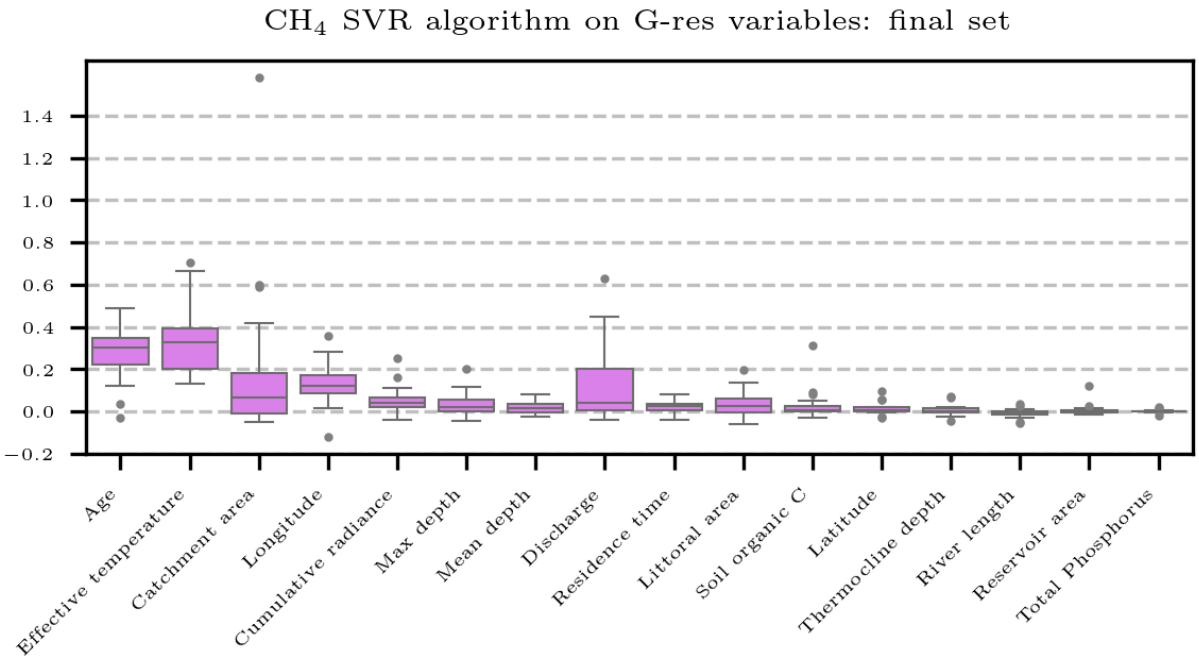

Figure 5.4: Permutation importance of the final set of covariates in the optimal SVR configuration on G-res variables reproducing CH_4 diffusion.

This sensitivity to data splits is reflected in R^2 values in training, test and entire set which show relatively wide ranges (Figure 5.5). Particularly, they vary between 0.45 and 0.62 in the training sets, between 0 and 0.72 in the test sets and between 0.45 and 0.56 in the full sets.

The limited size of the dataset also tends to drive the SVR models toward overfitting the training data. However, this effect remains relatively modest in most cases, with a high difference in training and test performance observed for only two splits (best and worst case).

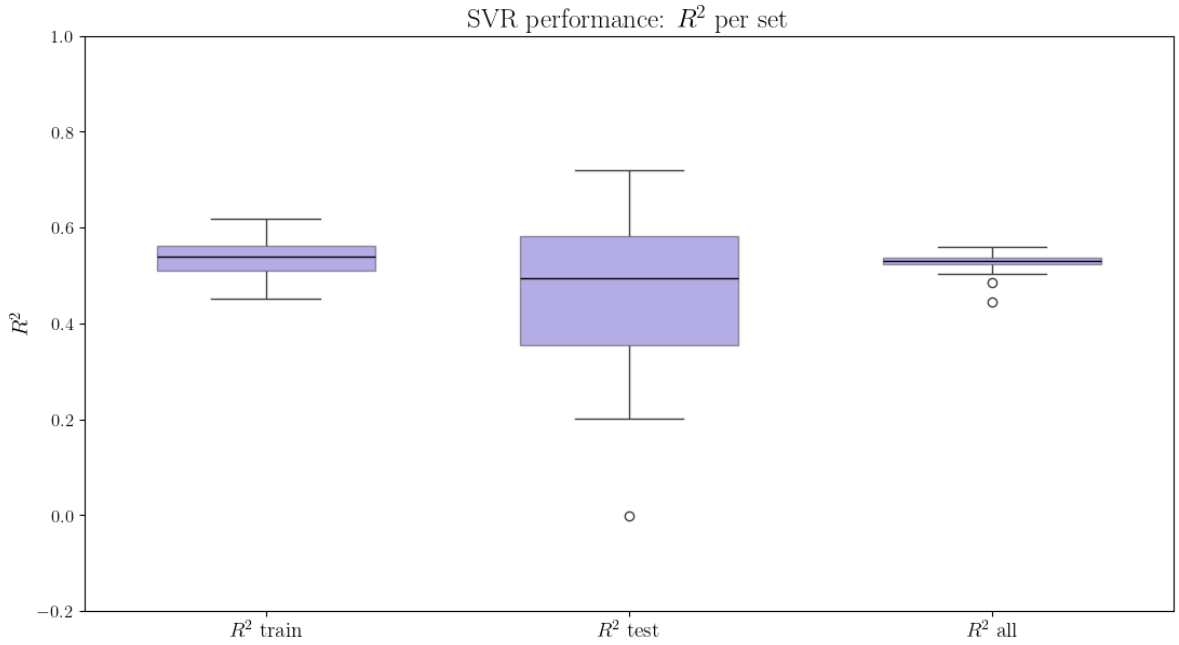

Figure 5.5: R^2 values computed on the training, test and the entire set for the SVR application using G-res covariates to reproduce CH_4 fluxes.

5.1.2. SVR algorithm and G-res: performance comparison

As shown in the previous section, the SVR model has reasonably good predictive ability, with median R^2 values in test sets of 0.38 and 0.50 for CO_2 and CH_4, respectively. Nevertheless, the performance still shows a notably variability with data splits.
In line with our objective of improving the G-res assessment, we compared their performances within the same subsets. This strategy allowed us to evaluate whether the predictions generated by ML models outperform those provided by G-res, while accounting for the uncertainty in the actual data distribution.

The R^2 analysis suggests that, overall, the SVR algorithm achieves more stable performance across the different subsets, as reflected in the narrower box plots (panel *a* and panel *c*, Figure 5.6).

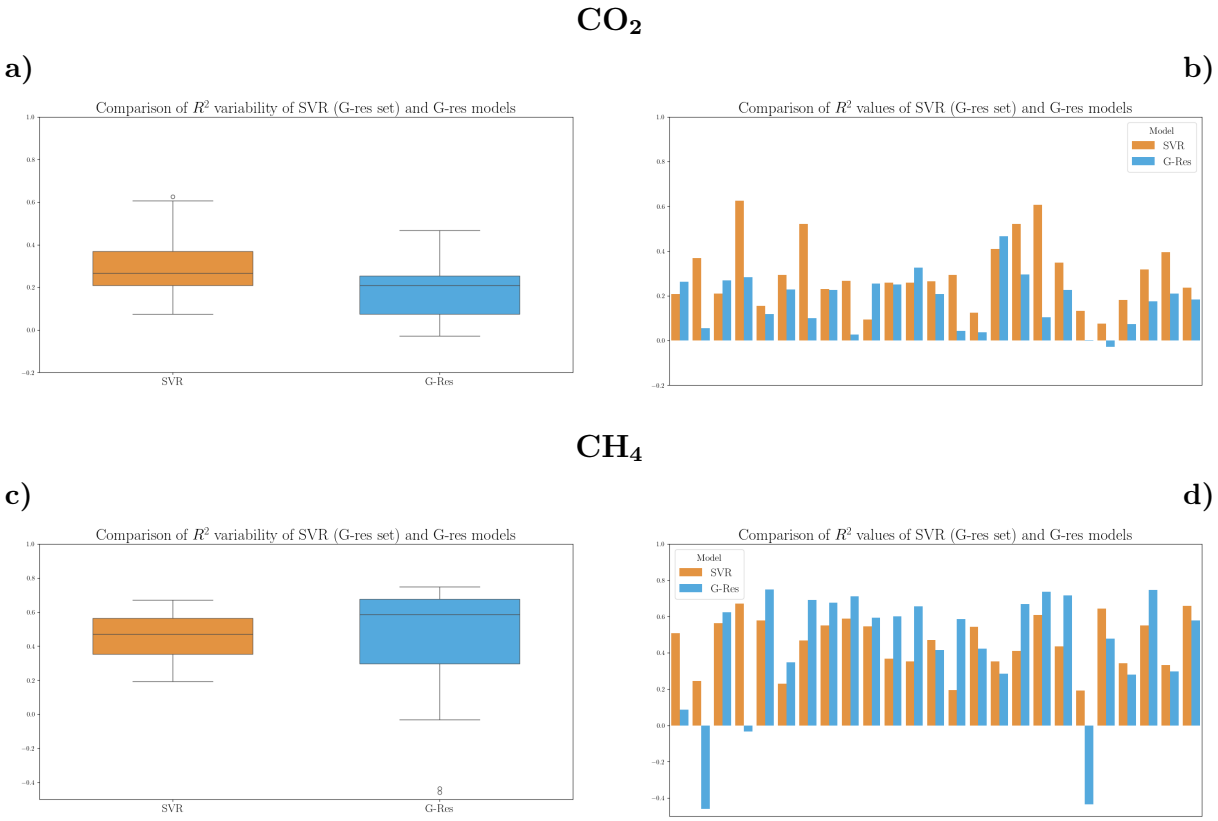

Figure 5.6: Comparison of R^2 values achieved by SVR and G-res models in corresponding subset. Panel *a* and *c* reveals the metric variability while panel *b* and *d* display subset-by-subset comparison.

Modeling CO_2 diffusion, the SVR algorithm outperforms regularly the G-res original model. The SVR R^2 values distribution over the different 25 data splits is consistently higher than that of G-res. Moreover, comparing directly performances across subsets (panel *b*, Figure 5.6), the SVR demonstrates higher values in all cases except for five subsets. Notably, even in those subsets where G-res obtains a higher R^2 value, the advantage over SVR remains relatively limited.

By contrast, for CH_4 modeling, the ML approach provides a marginal enhancement. The SVR algorithm and the original G-res model perform equivalently in almost all subsets. Only 3 cases deviate this trend, where the G-res model yields markedly negative R^2 values.

SVR and G-res comparison also includes the analysis of the bias. This performance criteria is introduce to evaluate the models' tendency towards underestimation. Indeed, accurately reproducing extreme fluxes has been reported as one of the major challenges in reservoir emission modeling (Abbasi and Abbasi, 2020).

In the CO_2 comparison assessment, considering the influence of extreme values on statistical accuracy metrics (Section 5.1.1), we performed a bias analysis on the entire dataset and further partitioned it into low- and high-emission subsets. In line with the considerations presented in Section 4.5.1, the 80^{th} percentile is used as the threshold for high emissions, with all lower values included in the low-emissions subset. This approach allows evaluation of model performance on the highest values as well as on the rest of the dataset, providing an unbiased measure of over- and underestimation.

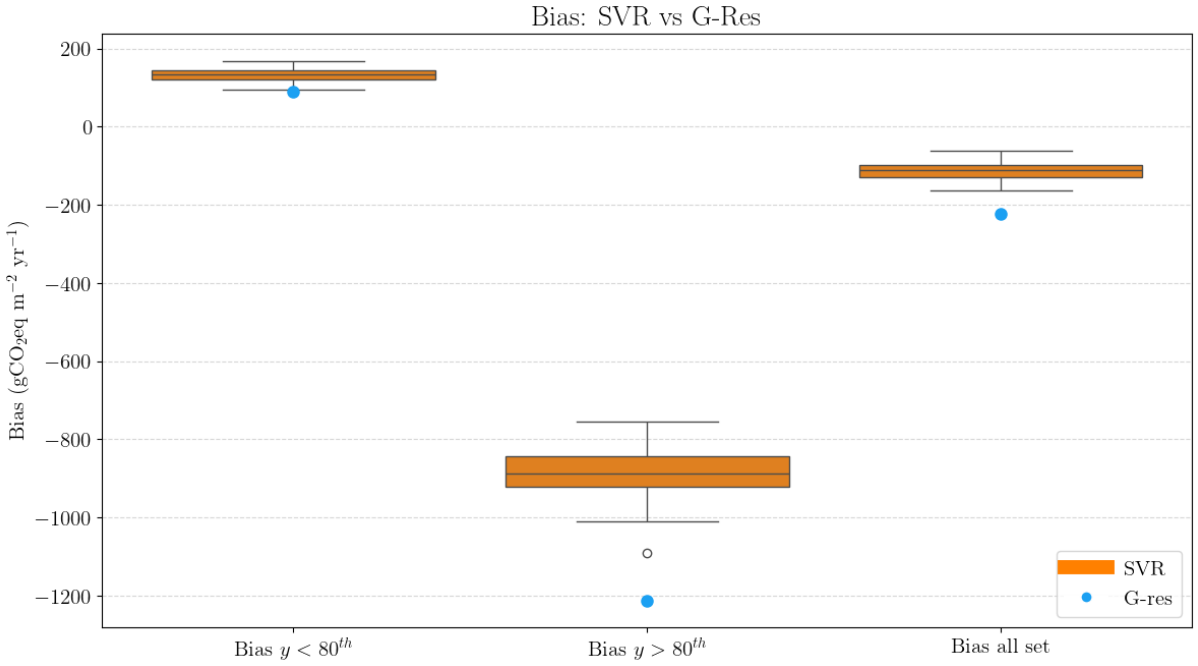

Figure 5.7: Comparison of bias values between SVR using G-res covariates and G-res models for three data groups: emissions below the 80^{th} percentile (*Bias* $y < 80^{th}$), emissions above the 80^{th} percentile (*Bias* $y > 80^{th}$), and the entire dataset (*Bias all set*).

Both SVR and G-res models exhibit a severe underestimation of the highest emission values. This pattern is also reflected in the bias calculated over the full dataset, with the underestimation being much more pronounced for G-res. Overestimation is limited in both cases, although the SVR models show a slightly greater tendency towards positive bias in the low values set. This may be an effect of the weighting strategy during the training phase (Section 4.5.1). However, when comparing the reduction in underestima-

tion, the resulting increase in positive bias is negligible.
Quantitatively, the mean bias for SVR is –114 gCO_{2eq} m^{-2} yr^{-1}, compared to –223 gCO_{2eq} m^{-2} yr^{-1} for G-res in the entire set. For the high values subset, the average SVR bias is -890 gCO_{2eq} m^{-2} yr^{-1}, against -1208 gCO_{2eq} m^{-2} yr^{-1} for G-res, while for the low values subset the values are 113 and 90 gCO_{2eq} m^{-2} yr^{-1}, respectively.
Thus, although the ML approach leads to a slight improvement in the estimation of extreme values, their acute underestimation persists, confirming that this remains one of the main challenges to be addressed in reservoir emission modeling.

The bias analysis for CH_4 models mirrors that of the R^2. Comparing results achieved by SVR and G-res models, the approaches seem similar, even if G-res show a lower tendency toward underestimation. In particular, the median bias in the entire sets amounts to -2.14 gCO_{2eq} m^{-2} yr^{-1} and -1.6 gCO_{2eq} m^{-2} yr^{-1} for SVR algorithm and G-res, respectively. Overall, the results of CH_4 models reveal that SVR better captures the average tendency and provides greater stability; G-res model, on the other hand, seems more flexible and able to capture outliers.

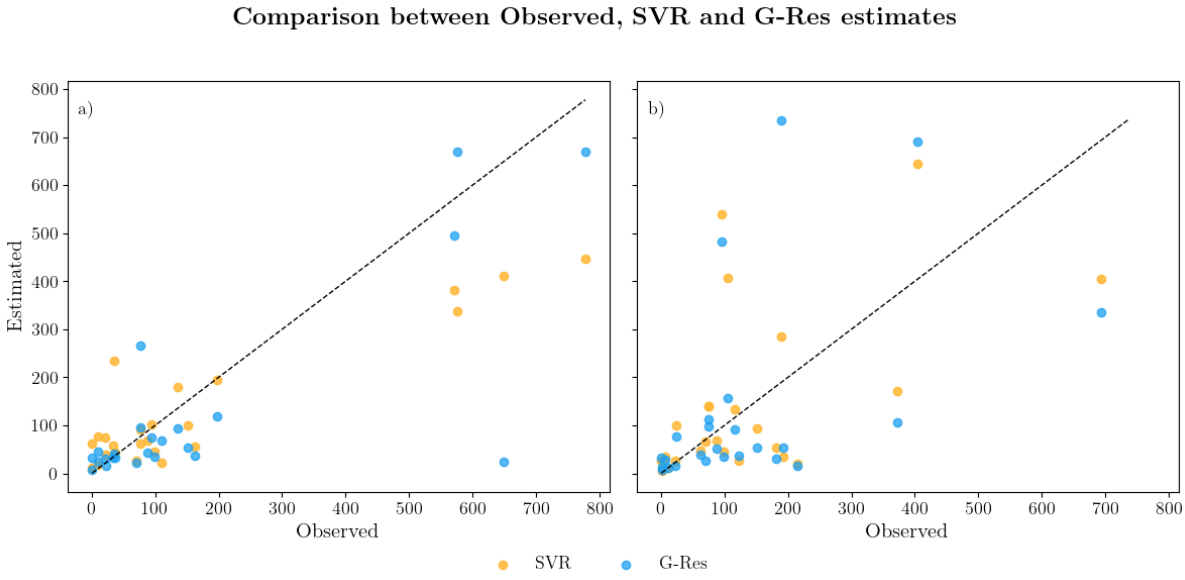

Figure 5.8: Comparison of SVR (G-res set) and G-res CH_4 estimates against observations. Evaluation sets in panel *a* and *b* are respectively the ones corresponding to R^2 best and worst value for SVR configuration. Emissions are expressed in gCO_{2eq} m^{-2} yr^{-1}.

Considering the comparable performance metrics, the predictive capacity of SVR and G-res models is further examined by comparing estimates and observations. Figure 5.8 displays the test sets in which the SVR configuration shows the best (panel *a*) and worst

(panel *b*) results. The comparison in the best case corroborates this observation. Both models well predict the emissions, with SVR better estimating the average value while G-res the extremes. In panel b), neither model is able to capture the emission pattern, with G-res model making even higher errors, suggesting that this particular subset cannot be accurately reproduced by the available predictor variables.

5.1.3. Final SVR models using G-res variables

Once the performance and stability of the SVR algorithm were assessed and compared to G-res, final ML models for CO_2 and CH_4 were trained for future applications. Given the stability of the SVR algorithm across data splits, the final models are expected to reproduce a similar level of accuracy, consistent with the results obtained in the iterative analysis, regardless of the specific data partition.

This is confirmed by the evaluation metrics reported in Table 5.1. As expected, the R^2 values are consistent with those observed in the SVR configuration analysis. The models show robustness on the training and full datasets, while the degree of overfitting is moderate, considering the limited and skewed distribution of the data. Specifically, the CO_2 model yields R^2 values of 0.41 on the training set and 0.21 on the test set, whereas the CH_4 model achieves 0.59 and 0.36, respectively.

Set	**CO_2 model**			**CH_4 model**		
	R^2	**MSE**	**Bias**	**R^2**	**MSE**	**Bias**
Training set	0.413	681668	-127	0.593	1142	-22
Test set	0.209	414661	183	0.360	2367	-83
Full set	0.393	630363	-63	0.543	1397	-35

Table 5.1: Metrics of SVR final models using G-res covariates. MSE and Bias are expressed in $gCO_{2eq}\ m^{-2}\ yr^{-1}$.

The only unexpected outcome is the positive bias value on the test set for CO_2 model, amounting to 183 $gCO_{2eq}\ m^{-2}\ yr^{-1}$. An investigation of its composition reveals that it mainly includes emission observations lying in the low-to-middle range of the distribution. Therefore, model behavior reflects the results obtained by the best SVR configuration in the subset below the 80^{th} percentile (panel *a*, Figure 5.9).

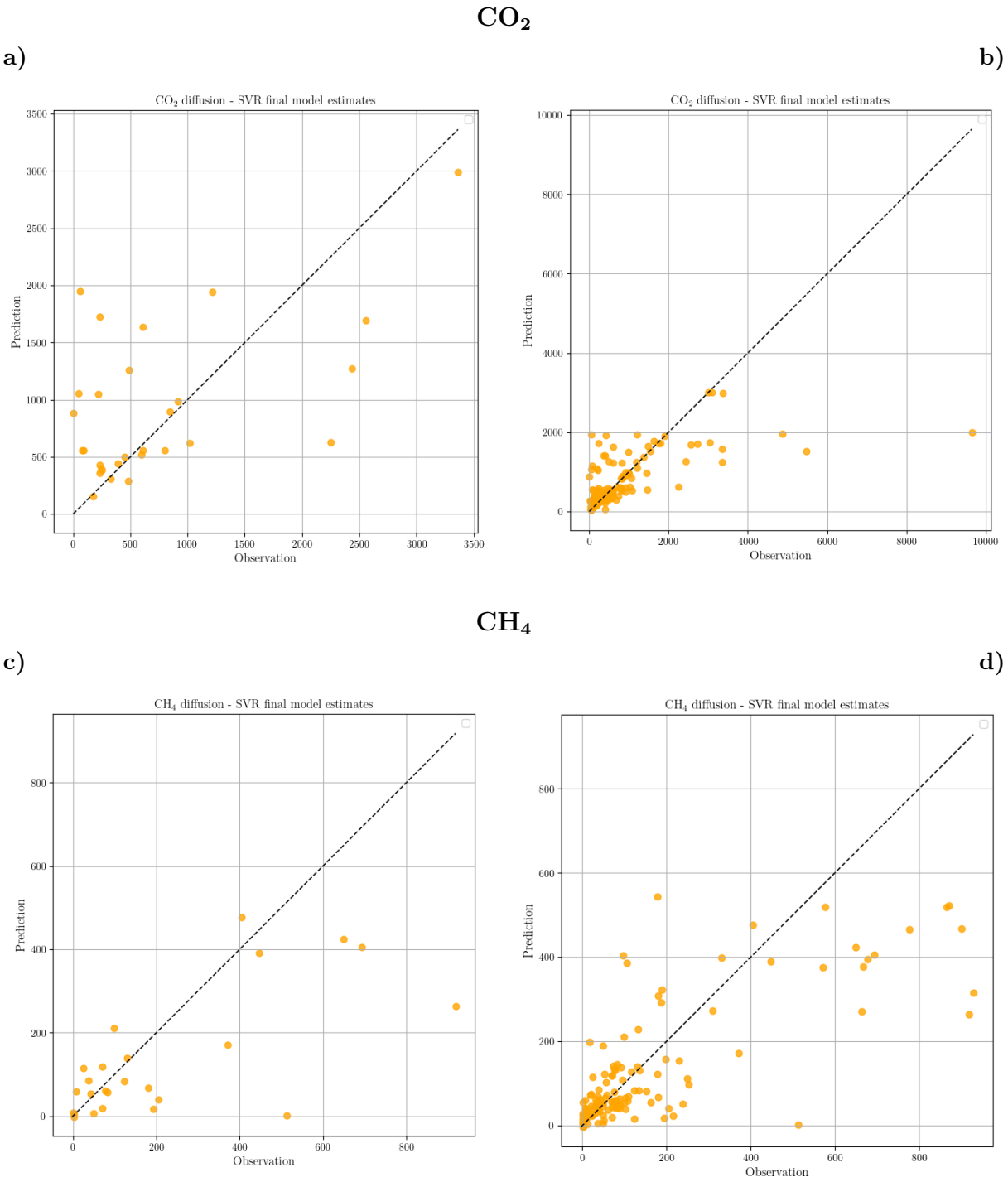

Figure 5.9: Comparison of ML final model estimates and observations, when ML uses G-res covariates (panel *a* and *c* test set, panel *b* and *d* entire set). The *y*-axis represents the estimated values, while the *x*-axis represents the observations. Emissions are expressed in gCO_{2eq} m^{-2} yr^{-1}.

Considering the limited size of the test set, the model predictive capacity is also evaluated on the entire set (panel *b* and *d*). In both CO_2 and CH_4 cases, the models reproduce the general behavior already observed for the SVR algorithm. Models predict reasonably well most of the emissions in the low-to-middle distribution, showing relatively few cases that deviate from the 1:1 line. However, systematic underestimation emerges for high-emission observations, in line with the already discussed limitations of SVR in representing extreme values (Section 5.1.1). This issue is particularly pronounced for CO_2 diffusive fluxes above 4000 gCO_{2eq} m^{-2} yr^{-1}, where the underestimation becomes acute.

Finally, Figure 5.10 shows the predicted emission intensity for the reservoirs included in the G-res dataset, providing a spatial representation of CO_2 and CH_4 emissions. In line with the literature and with the identified emission drivers (Section 3.1.1), the highest values are predominantly predicted in tropical regions (Brazil, Colombia, Central Africa, South-East Asia), whereas lower intensities are estimated elsewhere, highlighting the models' ability to reproduce the actual spatial distribution of emissions.

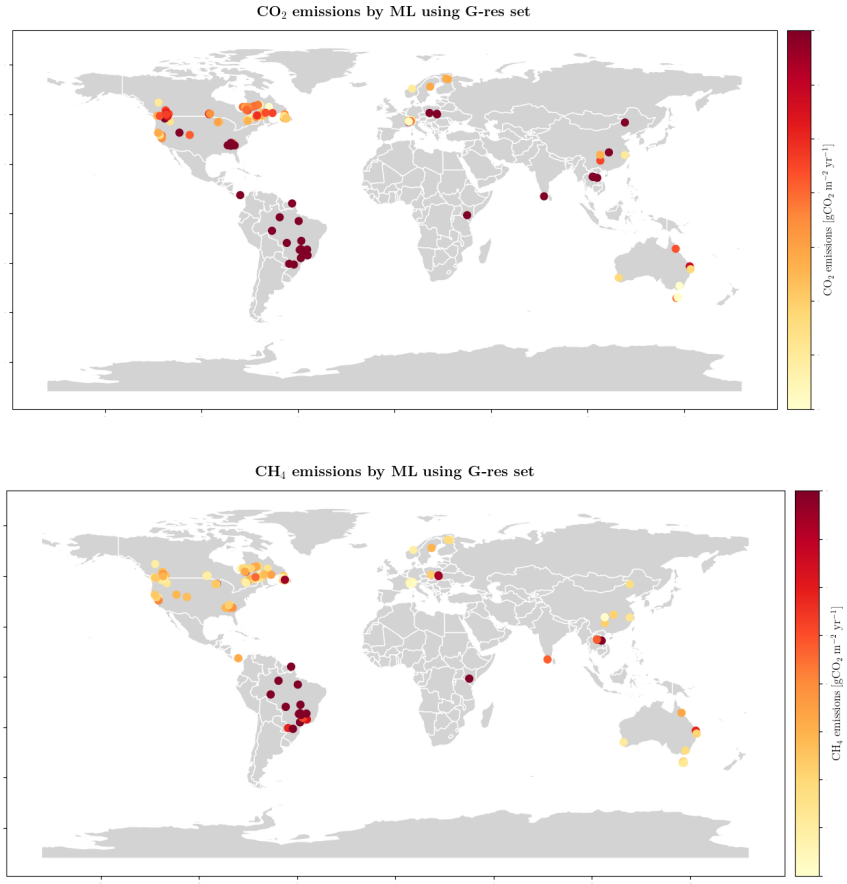

Figure 5.10: Spatial distribution of emission intensity of reservoirs included in G-res dataset, estimated by ML models using G-res variables (gCO_{2eq} m^{-2} yr^{-1})

5.2. SVR algorithm applied using new variables

The performance analysis of the SVR algorithm on the G-res dataset addresses the first key question of this study, demonstrating how Machine Learning can support the modeling of GHG emissions from reservoirs. Subsequently, the SVR is applied to the newly developed dataset, with the objective of assessing potential of readily available data.
The following sections present the results, structured analogously to those obtained for the G-res dataset.

5.2.1. SVR algorithm performance

Our analysis begins by showing the new set of identified covariates and analyzing the performance of the SVR algorithm in modeling CO_2 and CH_4 emissions.

CO_2 application

The set of covariates selected from the new dataset, associated to their permutation importance, is illustrated in Figure 5.11.

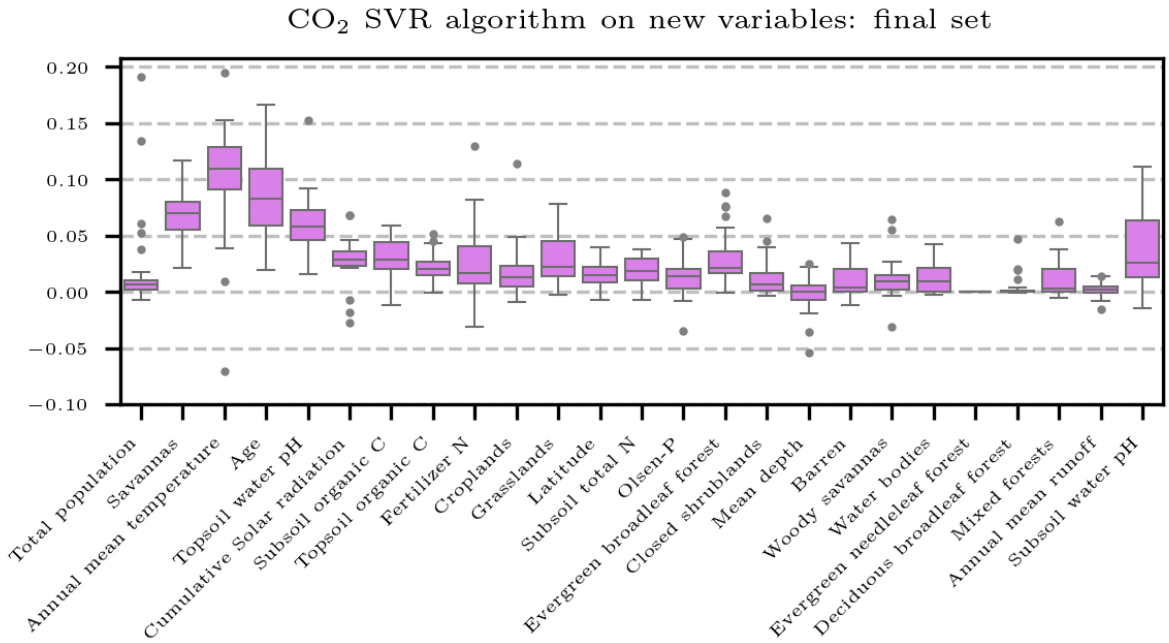

Figure 5.11: Permutation importance of the final set of covariates selected from new variables for reproducing CO_2 diffusion.

The number of predictors included in the model configuration increases substantially when considering the new variables.
Age and annual mean temperature confirm their central role in the CO_2 diffusive process, remaining the most influential factors in this configuration as well. The same consistency is observed for organic inputs, included in the model both through the content of soil organic carbon and nitrogen and, indirectly, through land cover classes. Figure 5.11 shows that among the land cover types, Savannas and Forests have the highest relevance, as they are typically associated with high primary productivity. The total population also emerges as a relevant predictor, as it is representative of the anthropogenic input of new organic matter. However, its contribution is less stable, assuming also negative importance in some cases.

The corresponding R^2 values' range in training, test and entire are respectively 0.31-0.54, 0.18-0.57, and 0.33-0.41 and are illustrated in Figure 5.12.

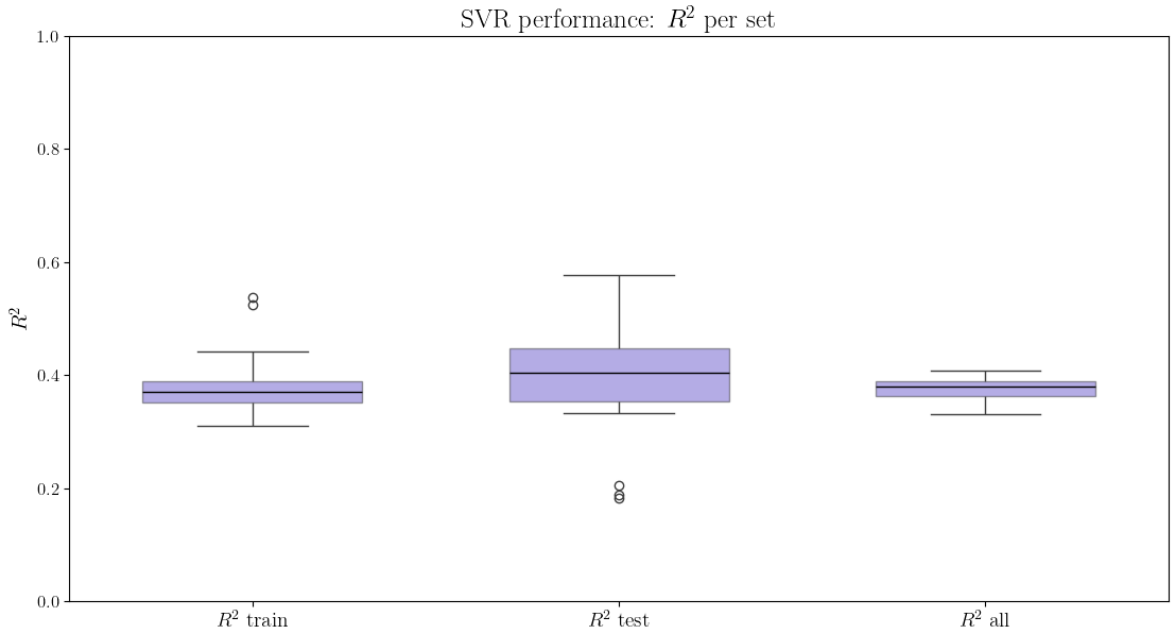

Figure 5.12: R^2 values computed on the training, test and the entire set for the 25 CO_2 SVR models on new covariates to reproduce CO_2 fluxes.

Similar to the SVR trained with G-res variables, the R^2 values in the test sets show a higher variability than in the training sets. However,the median R^2 values are essentially equivalent, 0.41 in the test and 0.38 in training sets, indicating that the algorithm con-

figuration is robust and the risk of overfitting is limited. Three isolated cases deviate from the trend. As in the worst cases of the SVR applied to G-res variables, the presence of an isolated extremely high emission value partially biases the metric, jeopardizing the models' ability to accurately reproduce the emission process (Figure 5.13).

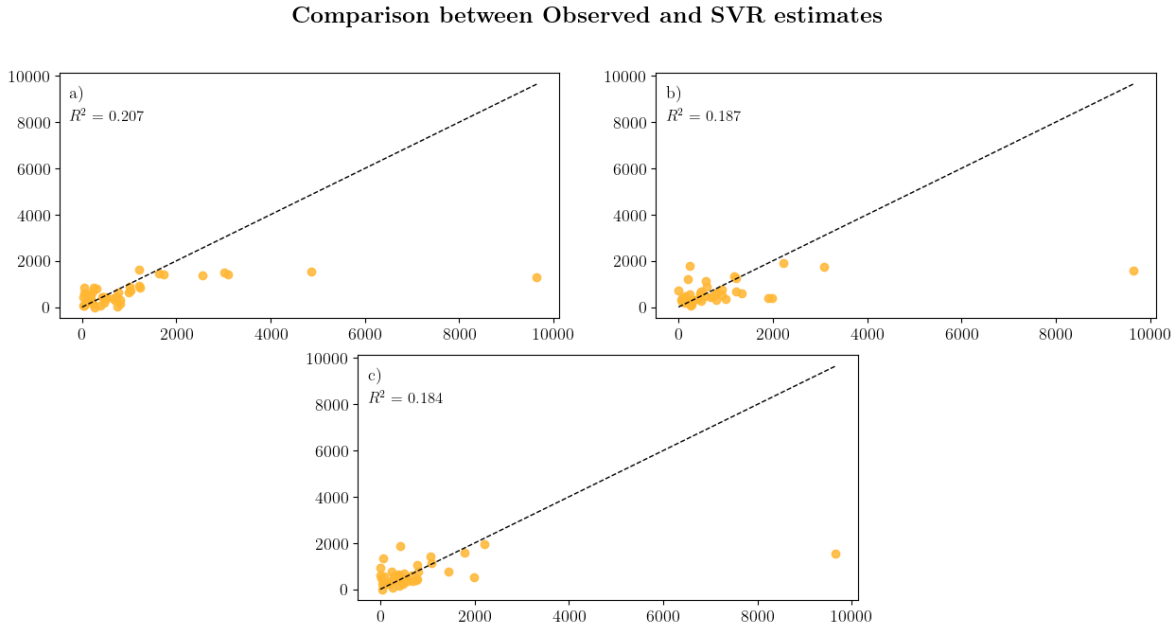

Figure 5.13: Comparison between CO_2 (gCO_{2eq} m^{-2} yr^{-1}) emissions predicted by the SVR (new dataset covariates) and the observed values in the test set where the SVR performs worst. The y-axis represents the estimated values, while the x-axis represents the observations.

Finally, when comparing the results of the SVR algorithm using the new and G-res covariates, adopting the new dataset seems to improve the SVR configuration's performance in terms of both stability and R^2 values. The SVR applied in the new dataset achieves higher R^2 values and shows narrower R^2 ranges, reflecting a weaker dependence of the algorithm performance on the specific train–test split.

CH_4 application

Consistently with G-res and with the SVR on G-res variables (Section 5.1.1), the analysis on the new predictors importance also highlights temperature as the main driver of GHG emissions (Figure 5.14). Although ranging in a wide interval across iterations, its relevance emerges above the others. In contrast with the SVR on G-res set case, in the new set the biochemical variables provide a more relevant contribution. Particularly relevant are

the factors representing soil primary productivity, namely nitrogen and phosphorus based fertilizers inputs and cropland coverage. This in accordance with literature (Ion and Ene, 2021). In fact, they are proxy indicators for new organic input to reservoir waters which are one of the main driver of CH_4 production.

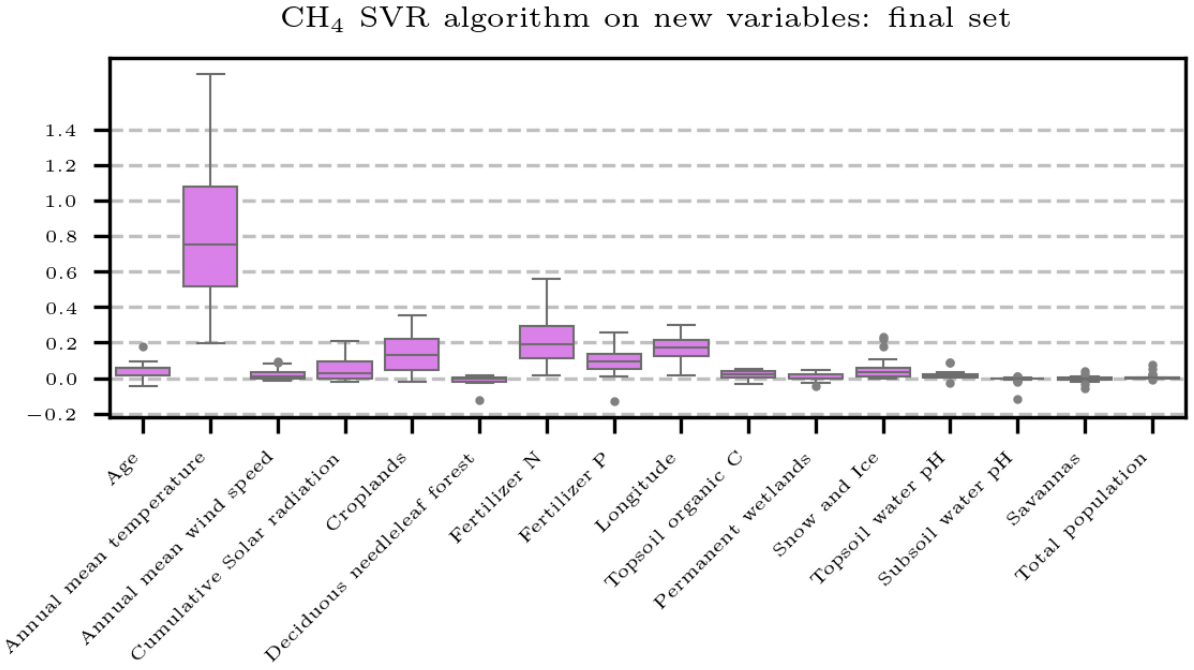

Figure 5.14: Permutation importance of the final set of covariates selected from new variables for reproducing CH_4 diffusion.

Figure 5.15 displays the R^2 obtained by the SVR identified configuration. It confirms that the new set of variables enhances the stability of the SVR algorithm (Section 5.2.1), whose implementation appears fairly robust across the 25 different sets. The three boxplots, corresponding to intervals of 0.37-0.55 in training sets, 0.27-0.61 in test sets, and 0.40-0.54 in entire sets, are relatively narrow compared to the ones obtained training the SVR on G-res set (Figure 5.5).
By contrast, the median R^2 values in the three sets (0.47 in training, 0.42 in test and 0.47 in full) are comparable to those achieved by the SVR on G-res set, suggesting that the use of new predictors does not lead to any improvement in accuracy.

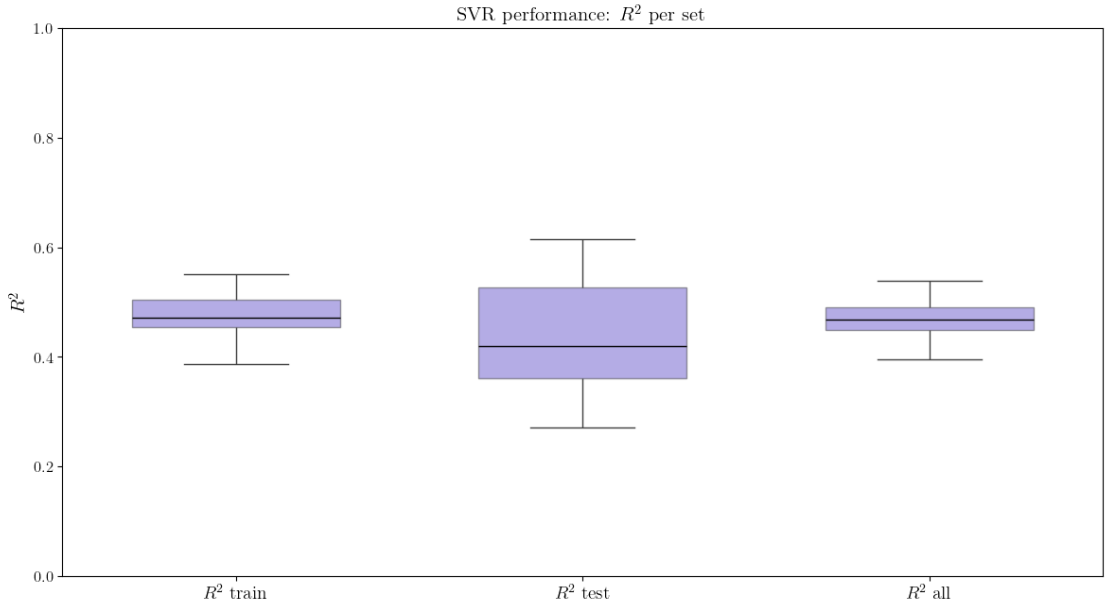

Figure 5.15: R^2 values computed on the training, test and the entire set for the 25 SVR models on new covariates to reproduce CH_4 fluxes.

5.2.2. SVR algorithm and G-res: performance comparison

Figure 5.16 compares the R^2 values obtained by SVR and G-res original models in corresponding subsets. The analysis of the R^2 box plots confirms that the ML approach reduces the variability of performance across data distribution.

This effect is particularly evident for CH_4, where the R^2 values range from 0.60 to 0.31, compared with the wider range of 0.75 to 0.18 observed for G-res (panel *b*, Figure 5.16). The overall behavior of the SVR trained on the new variables is comparable to that of the SVR based on G-res variables. Consequently, the SVR algorithm demonstrates greater stability and an improved ability to reproduce the overall emission pattern, whereas G-res exhibits a relatively stronger tendency to capture extreme values. In terms of bias, the average value in the test sets is -34 gCO_{2eq} m^{-2} yr^{-1} for SVR and -26 gCO_{2eq} m^{-2} yr^{-1} for G-res.

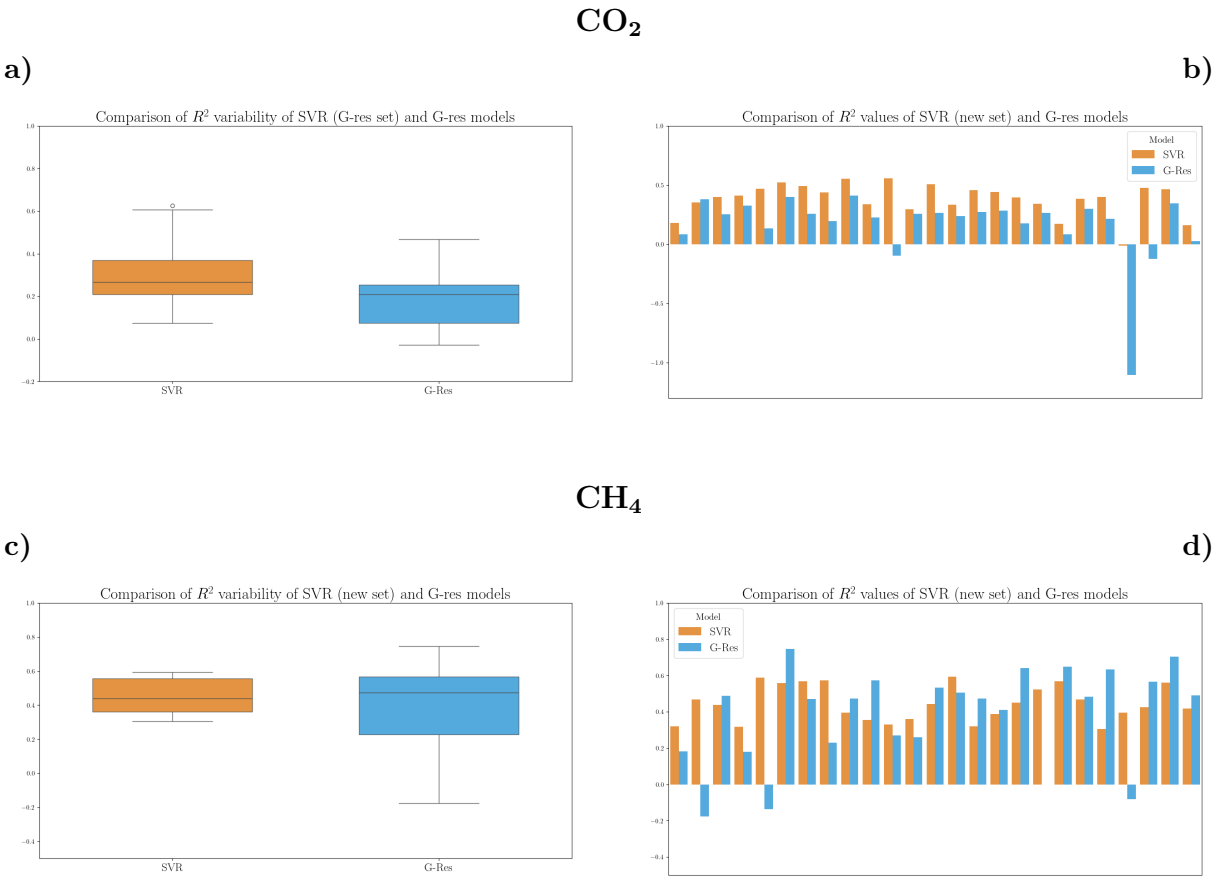

Figure 5.16: Comparison of R^2 values achieved by SVR and G-res approaches in corresponding subset. Panel *a* and *c* reveals the metric variability while panel *b* and *d* display subset-by-subset comparison.

Similarly to the SVR trained on the G-res dataset, the ML approach improves the CO_2 flux model of G-res but still severely underestimates the highest emissions.
Bars comparison (panel *b*, Figure 5.16) reveals that in only one case out of 25 G-res model yields an higher R^2 than the SVR, showing, in addition, severely poor performances in several cases.
Figure 5.17 demonstrates the underestimation of the extremely high values—more pronounced for G-res—and the slight tendency of SVR to overestimate low-to-middle emissions. For the full set, the average bias is equal to -112 gCO_{2eq} m^{-2} yr^{-1} for the SVR and -213 gCO_{2eq} m^{-2} yr^{-1} for the G-res model, for the upper 20% of emissions, it is -856 gCO_{2eq} m^{-2} yr^{-1} and -1177 gCO_{2eq} m^{-2} yr^{-1}, while for the lower 80% it is 117 gCO_{2eq} m^{-2} yr^{-1} and 86 gCO_{2eq} m^{-2} yr^{-1}, respectively.

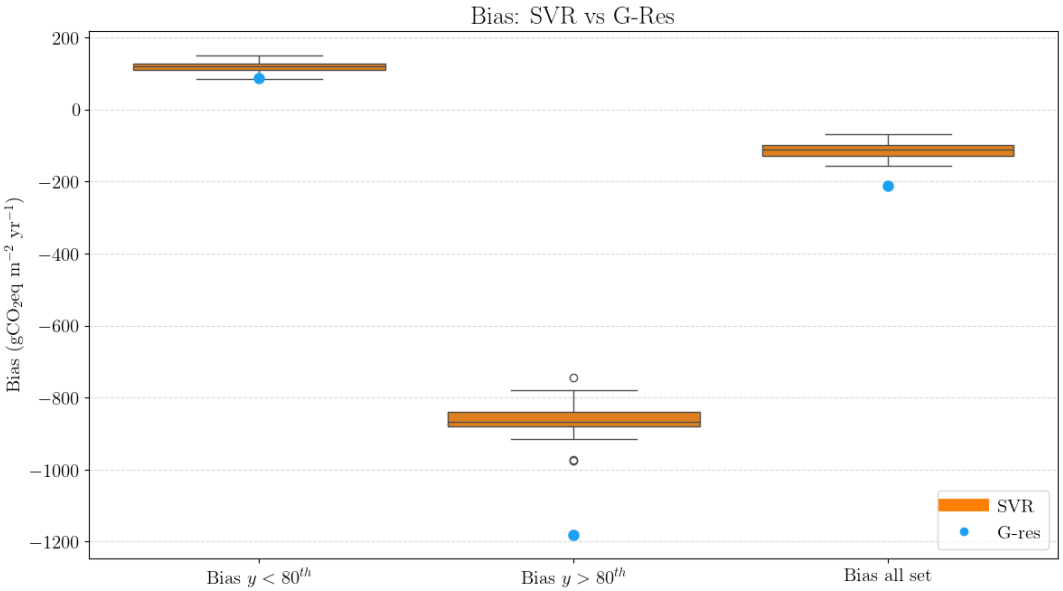

Figure 5.17: Comparison of bias values between SVR (new dataset covariates) and G-res models for three data groups: emissions below the 80^{th} percentile (*Bias* $y < 80^{th}$), emissions above the 80^{th} percentile (*Bias* $y > 80^{th}$), and the entire dataset (*Bias all set*).

5.2.3. Final SVR models using new variables

Once analysed SVR algorithm performance using the new variables and compared to G-res, final models are built. The corresponding performance criteria are reported in Table 5.2.

Set	**CO_2 model**			**CH_4 model**		
	R^2	**MSE**	**Bias**	**R^2**	**MSE**	**Bias**
Training set	0.359	434385	-76	0.608	1132	-37
Test set	0.371	945643	-266	0.345	1483	-51
Full set	0.362	659101	-118	0.565	1204	-40

Table 5.2: Metrics of SVR final models using new covariates. MSE and Bias are expressed in gCO_2-eq m^{-2} yr^{-1}.

R^2 for training, test, and full set of CO_2 model are nearly equivalent (0.36, 0.37 and 0.36, respectively), indicating model robustness and flexibility, with no evidence of overfitting. On the other hand, a moderate overfitting is observed for the CH_4 model, with R^2 values

of 0.61 in training and 0.35 in the test set. However, data availability for CH_4 diffusion process is lower, therefore explaining this different observed behavior.
The ML models reflect the behavior of SVR algorithm observed both with new and G-res variables. They are able to reasonably well reproduce the low-to-middle emission range but still underestimate the highest emission values. Bias values are consistently negative—although to varying degrees across sets—and the comparison between estimates and observations confirms this tendency.

The spatial distribution of GHG fluxes (Figure 5.18) reproduced by the ML models trained on the new variables is also similar to that previously obtained with the G-res variables (Figure 5.10), confirming the models' ability to capture the actual emission distribution.

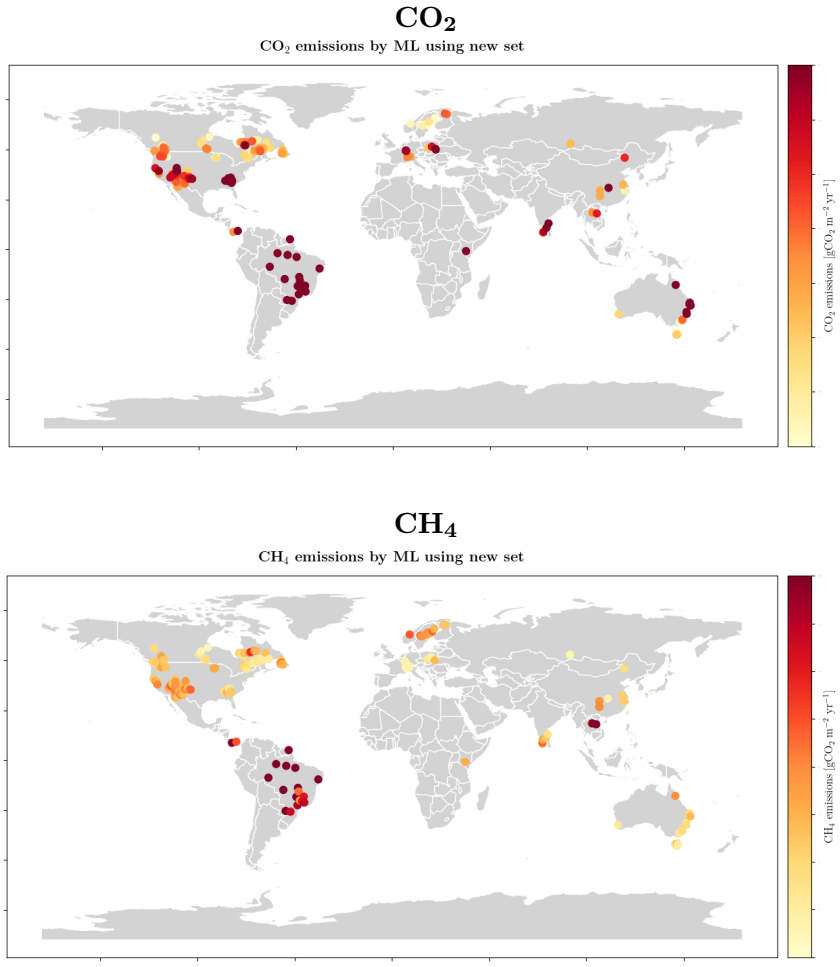

Figure 5.18: Spatial distribution of emission intensity of reservoirs included in G-res dataset, estimated by ML models using new variables (gCO_{2eq} m^{-2} yr^{-1})

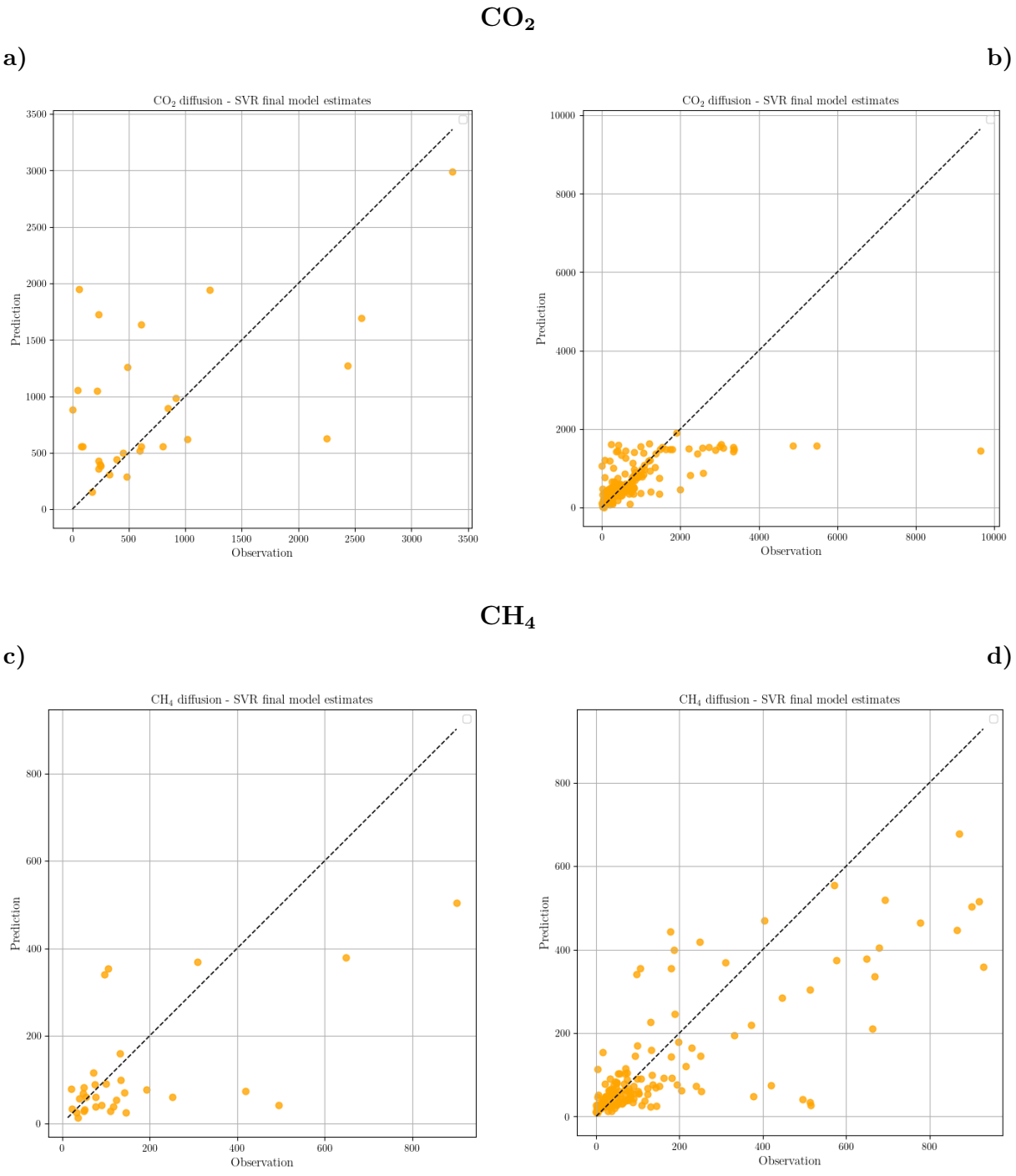

Figure 5.19: Comparison of ML final model estimates and observations, when ML uses G-res covariates (panel *a* and *c* test set, panel *b* abd *d* entire set). The *y*-axis represents the estimated values, while the *x*-axis represents the observations. Emissions are expressed in gCO_{2eq} m^{-2} yr^{-1}.

5.3. Assessment of ML and G-res models predictive capacity

Finally, the performances of final ML models, using new and G-res variables, and the original G-Res models are compared. This step aims to comprehensively answer both the research questions guiding our study, allowing for an evaluation of the behavior of ML models relative to G-res and in relation to the two datasets.
One of the main outcomes of the present analysis is the comparable behavior exhibited by the ML models trained on the two sets of covariates, for both CO_2 and CH_4. This consistency, already noted in the individual assessment of the SVR algorithm (Sections 5.2.1, 5.2.2), becomes even more evident when the final models are directly compared. These findings confirm that the new set of covariates ensures a predictive capacity equivalent to that of the G-res variables, making it a solid alternative to the empirically derived predictors adopted in G-res.

5.3.1. ML and G-res models: CO_2 emission underestimation

A major limitation of the CO_2 models, common across all approaches, is their restricted ability to capture the highest emission values. Since these values may strongly affect the estimation of the carbon footprint of large dams, the model comparison focuses on their tendency to underestimate CO_2 diffusive fluxes.

Initially, models' bias are compared, considering also data partition in below and above the 80^{th} percentile threshold.

Model	Bias full set	Bias high set	Bias low set
G-res original model	-224	-1458	84
SVR on G-res variables	-61	-945	160
SVR on new variables	-139	-1175	120

Table 5.3: Bias values computed on the entire set, the high-emission subset, and the low-emission subset. The sets includes only emission values estimate simultaneously by the three CO_2 models. Emission are expressed in gCO_{2eq} m^{-2} yr^{-1}.

The SVR using variables from the G-res dataset shows the best performance in terms of bias. Although underestimation is pronounced in all models, it is less acute in both the full and high-emission subsets for this configuration. Conversely, its positive bias is the highest among all models, yet remains limited.

Figure 5.20 compares the emission predictions and observations and allows a better interpretation of the metric.

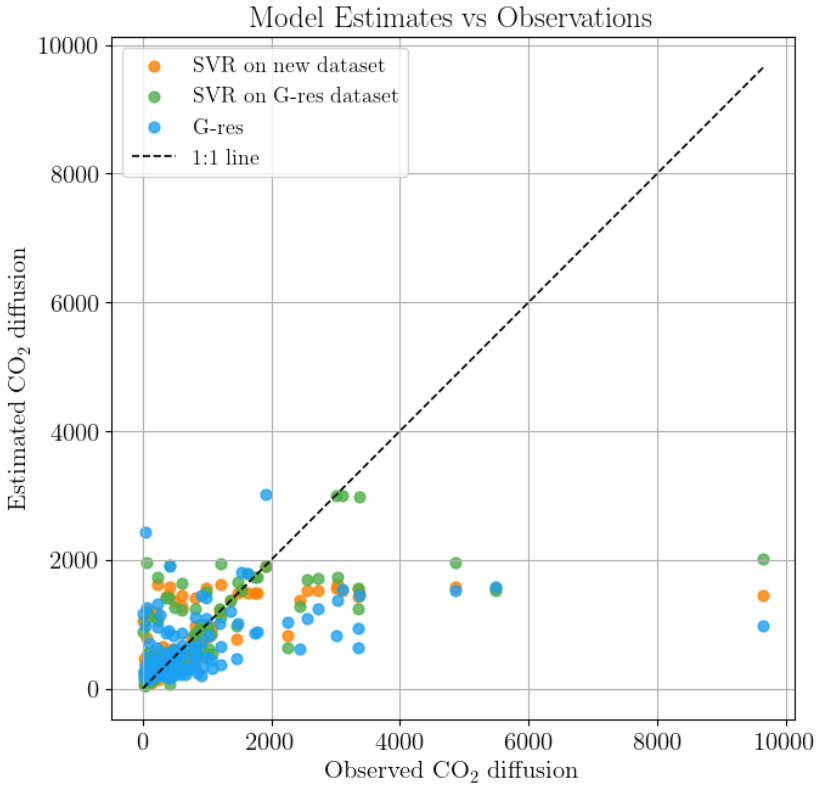

Figure 5.20: CO_2 emissions estimated by all three models — SVR trained on the new dataset, SVR trained on the G-res dataset covariates, and the G-res model — compared to observed values (gCO_{2eq} m^{-2} yr^{-1}). The y-axis represents the estimated values, while the x-axis represents the observations.

The scatterplot highlights the ability of all models to predict reasonably well the lowest emission values (under 2000 gCO_{2eq} m^{-2} yr^{-1}). Among them, the SVR trained on the new variables is the one that most closely follows the 1:1 line, whereas the G-res model shows larger deviations both toward overestimation and underestimation. The SVR on the G-res variables is able to better capture the middle-range emissions (2000 gCO_{2eq} m^{-2} yr^{-1} - 4000 gCO_{2eq} m^{-2} yr^{-1}). This pattern is consistent with the lower negative bias observed for both the high and full sets. Even if for most middle-to high emission values ML models provide slightly better estimates than G-res, their underestimation remains severe and confirms the challenge of accurately reproducing extreme emission values in modeling.

Globally, the differences in prediction shown by the three models are reflected in cumulative values that amount to 24.94 $MtCO_{2eq}$ yr^{-1} for the model using new variables,

29.24 $MtCO_{2eq}$ yr^{-1} for the model using G-res variables and 25.54 $MtCO_{2eq}$ yr^{-1} for the original G-res model. These results are obtained by including only the estimates simultaneously produced by all models and by averaging the estimates corresponding to the same reservoirs at different ages. The corresponding observed value amounts to 27.77 $MtCO_{2eq}$ yr^{-1}.

On this basis, each of the three models deviates from the observed values by approximately 2 $MtCO_{2eq}$ yr^{-1}. To explore the potential implications of such bias at the global scale, this discrepancy was first normalized per reservoir and then extrapolated to the total number of large dams reported in the GRanD database. The propagated error amounts to approximately 132 $MtCO_{2eq}$ yr^{-1}, which corresponds to only 0.83% of the CO_2 emissions from the power sector in 2015 (EDGAR). Although this calculation is based on a strong simplification—namely the assumption of a uniform per-reservoir error—it provides a useful order-of-magnitude indication of the uncertainty that model inaccuracies could introduce in global assessments.

Although the uniform-bias assumption highlights the need for further refinement of predictive accuracy, from a broader perspective, this result suggests that the models are sufficiently robust to support large-scale evaluations of reservoirs emission.

5.3.2. ML and G-res models: reproduction of CH_4 emissions

Previous comparisons between the SVR algorithm and G-res revealed two general trends: SVR tends to capture the overall emission pattern more accurately, whereas G-res is more effective in reproducing extreme values (Sections 5.1.1, 5.2.1). Nonetheless, the overall performance of the two approaches remains largely comparable. For this reason, the comparison considered both dimensions of predictive capacity, namely the reproduction of emission patterns and the extent of systematic over- or underestimation. Consistently with the previous analyses, neither model clearly outperforms the other.

Set	R^2	Bias
SVR on G-res set	0.35	-75
SVR on new set	0.39	-73
G-res	0.40	-59

Table 5.4: Metrics of ML using G-res and new dataset final model. Bias is expressed in $gCO_{2\text{-}eq}$ m^{-2} yr^{-1}.

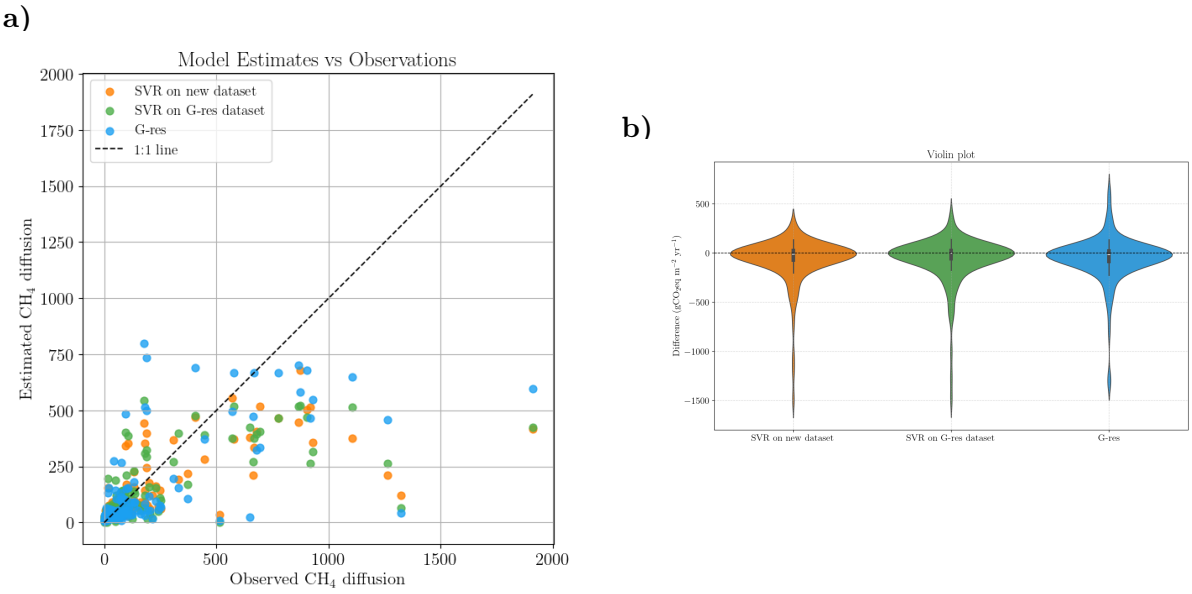

Figure 5.21: Panel a) compares models estimates against observations. Panel b) represents the errors distribution of the three model. ML model on new set is in orange, ML model on G-res set is in green, G-res original model is in blue and emissions and differences are expressed in gCO_{2eq} m^{-2} yr^{-1}.

The ML models behave similarly regardless of the set of predictors used. Both fit the observations in the lower range fairly well, while, as emission values increase, their predictive capacity is exceeded by the G-res model. In particular, G-res performs significantly better in the medium range (500 - 1000 gCO_{2eq} m^{-2} yr^{-1}), whereas for the most extreme values a strong underestimation is observed for all models. This is reflected in bias values, which amounts -59 gCO_{2eq} m^{-2} yr^{-1} for G-res, -75 gCO_{2eq} m^{-2} yr^{-1} for ML model on G-res set and -73 gCO_{2eq} m^{-2} yr^{-1} for ML on new set.
Panel b) of Figure 5.21 shows the error distribution, with violin width indicating the error frequency of occurrence. Consistent with the overall ability of the models to reproduce the emission process reasonably well, most of the errors lie close to zero. While the general underestimation of the models was already evident, the violin plot further emphasizes the pronounced upper tail of the G-res distribution, reflecting occasional but relevant overestimation. Looking at the direct comparison between estimates and observations (panel *a*), these cases are in the lower tile of emissions distribution.
The analysis of the cumulative values shows that the underestimation amounts to 0.2 $MtCO_{2eq}$ m^{-2} yr^{-1} for the G-res model, 1.0 $MtCO_{2eq}$ m^{-2} yr^{-1} for the ML model trained on G-res variables, and 0.7 $MtCO_{2eq}$ m^{-2} yr^{-1} for the ML model trained on the new variables. The comparatively better performance of the G-res model can be attributed both

to its greater ability to capture extreme values, but also to the tendency toward overestimation evidenced by the violin plot. The larger error is made by the ML model applied on G-res variables. However, applying the error propagation procedure outlined in Section 5.3.1 shows that this error accounts for less than 0.035% of global CH_4 emissions from the power sector (EDGAR data about CH_4 global emissions), thus supporting the applicability of the models to global assessments.

5.4. Assessment of models applicability

Once the performance of the ML and G-res models had been assessed, the comparison was extended to their global applicability, a crucial aspect for large-scale applications. The analysis first focused on the models' ability to reproduce the spatial distribution of emissions, and then evaluated their inclusiveness in terms of the number of reservoirs covered.

5.4.1. CO_2 models applicability

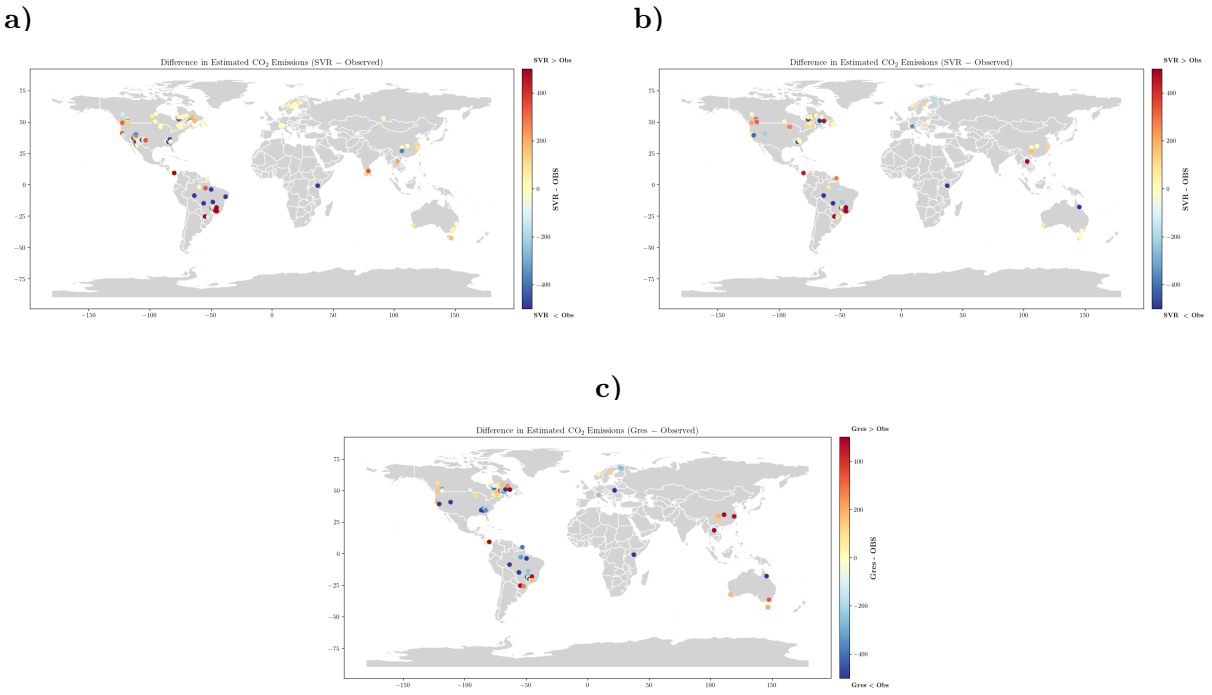

Figure 5.22: Difference in emission values ($gCO_{2eq}\ m^{-2}\ yr^{-1}$). Panel *a* shows the ML model using the new variables, panel *b* the ML model using the G-res variables, and panel *c* the original G-res model.

Figure 5.22 displays models errors across global reservoirs. Dots color represents the magnitude of the difference between predicted and observed values, where dark blue indicates severe underestimation.

As expected, most of the blue dots are in the tropical areas (South America, Central Africa) where the highest reservoir emission rates are reported. Despite models are able to recognize that these areas correspond to the highest GHG fluxes (Figures 5.10 and 5.18), models are far from reproducing such high values. By contrast, no spatial pattern for emissions overestimation can be identified. The red dots are, in fact, scattered across regions. Excluding the tropical area, ML models maps show that our models capture the spatial distribution of the reservoir emissions. The 40% of the dots have an error of 50 gCO_{2eq} m^{-2} yr^{-1} (yellow colors). Consequently, ML models reliably cover most regions worldwide. Moreover, the comparison with the G-res map, characterized by bright dots, highlights the enhancement brought by Machine Learning techniques in estimating global reservoir carbon footprint.
Between ML two models, using G-res and new variables, the former seems to provide more accurate predictions, especially in the tropical area. This is in agreement with the performance analysis previously discussed (Section 5.3.1).

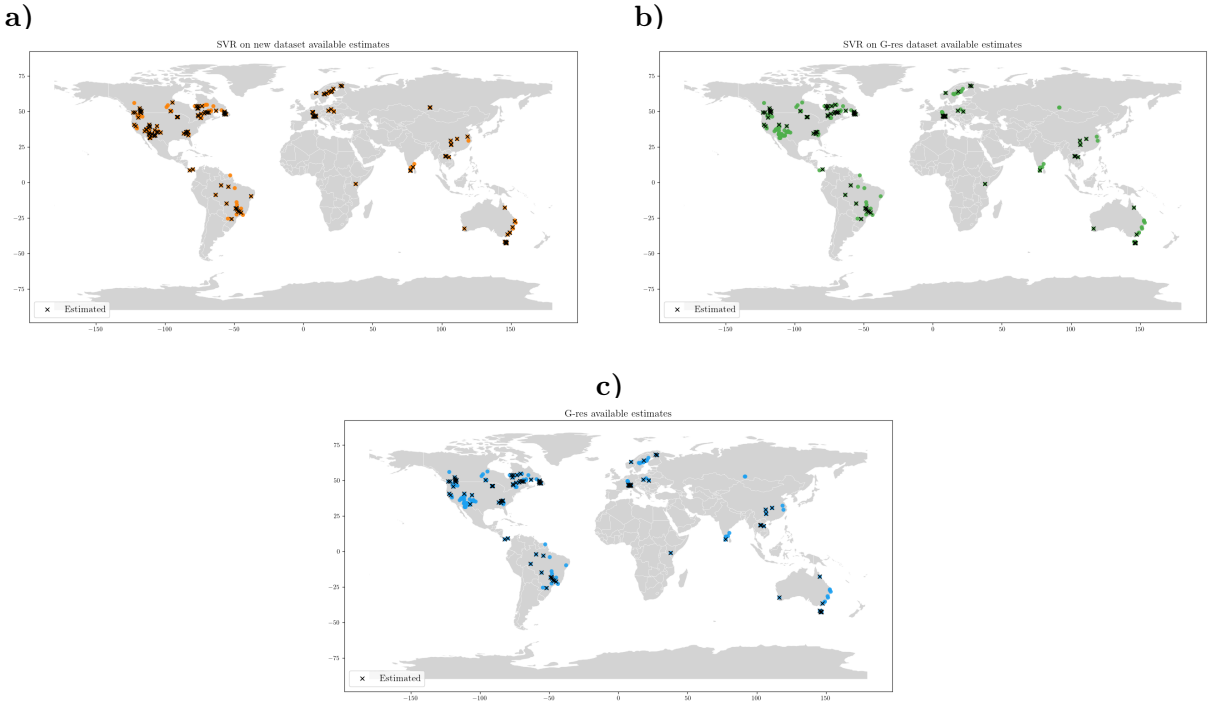

Figure 5.23: Comparison between the number of total dams in the dataset and the available estimates.

Examining their capacity for reservoir inclusion, the use of variables from the G-res dataset limits models global applicability.
Our ML model using G-res variables predicts emission values from 119 out of 288 samples (panel *b*, Figure 5.22) while G-res original model 207 (panel *c*, Figure 5.22). The model built with the new dataset allows, by contrast, for 280 estimates (panel *a*, Figure 5.22), corresponding to an increase of approximately 25% in coverage.
With regard to model robustness, the discussion in Section 5.3.1 identifies the SVR trained on G-res variables as the slightly better-performing model. However, considering the notable higher applicability shown by the use of the new set of predictors, and the relatively small difference between the two ML models in most of the emission distribution and cumulative values, SVR built from the newly developed dataset model can be considered the most effective approach. It outperforms the G-res model and enables wider applicability for assessing emissions from reservoirs worldwide.

5.4.2. CH_4 models applicability

The analysis of CH_4 models reflects the results of CO_2 models applicability evaluation. Panel *a*, panel *c* and panel *d* in Figure 5.24 displays that larger errors occur in tropical areas, characterized by the presence of the darker dots, both blue and red. The best predicted climate zone is the temperate (North America), for which more observations are available. Especially the ML using the new set appears capable to represent this area.

By contrast, models appear significantly different in terms of reservoirs inclusiveness. Panel *b*, *d* and *f* compares the number of reservoirs for which emission predictions are available and the total reservoirs. Clearly, the ML using new variables enables the inclusion of a higher number of reservoirs. Quantitatively, it estimates 279 emission values, G-res model 239, while the ML model trained on G-res set only 193.

Taken together, these results indicate that, as already observed for CO_2, the ML model based on the new set of readily available variables represents the most effective approach for CH_4 emission assessment, combining comparable predictive performance with broader applicability in terms of reservoir coverage.

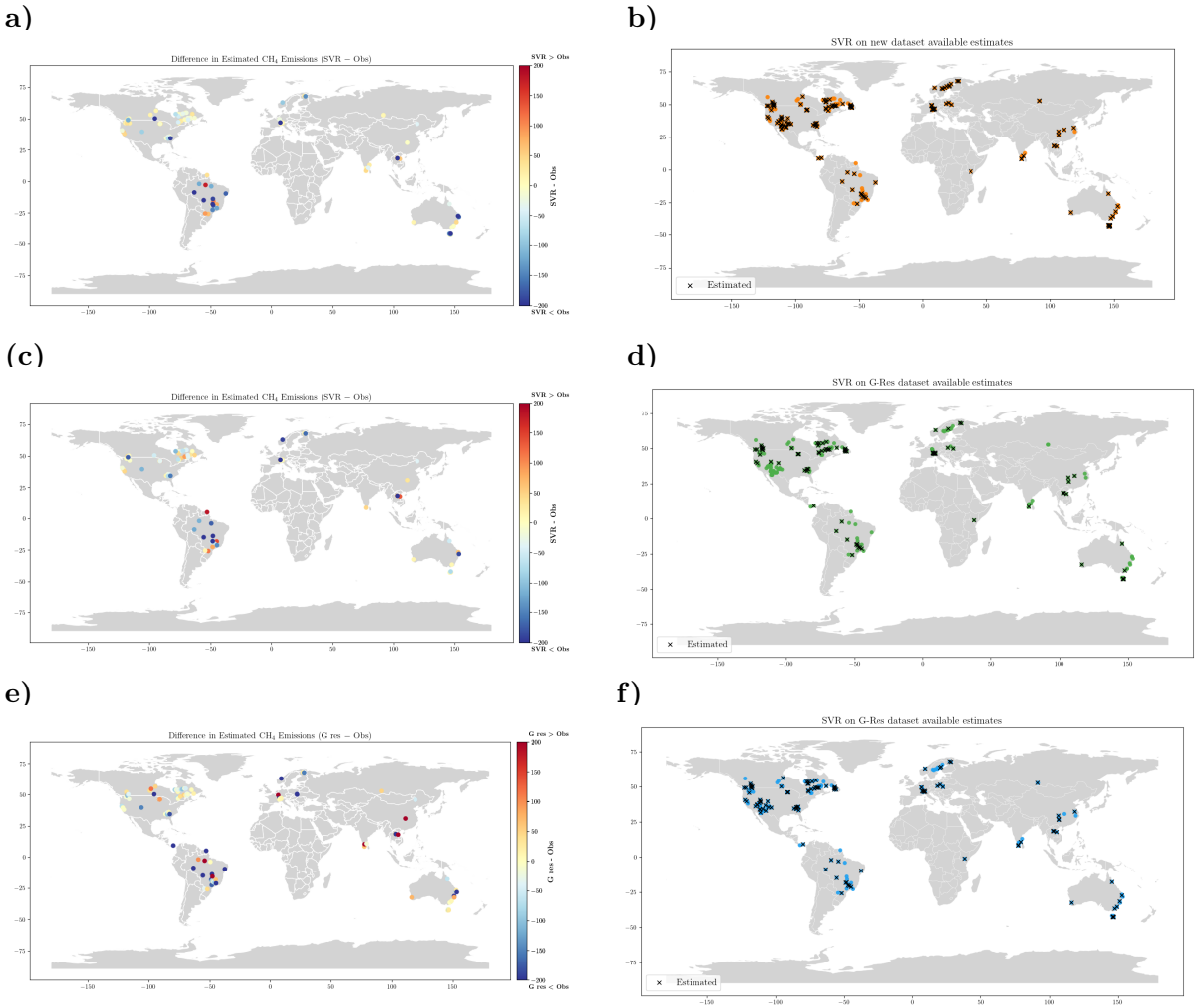

Figure 5.24: Difference in terms of emission (gCO_{2eq} m^{-2} yr^{-1}) to compare the models (panel *a*, *c* and *e*). Panel *b*, *d* and *f* compares the number of total dams in the dataset and the available estimates. First row shows results for ML model on new variables, second row for ML on G-res variables and third for G-res original model

5.5. SVR model application on the European case study

The relevance of our work is revealed in the real application of our models.
Both ML and G-res models are applied to estimate the carbon intensity of European large dams to examine how models perform on a large scale assessment, from data collection to the final estimate of the GHG footprint.
The analysis begins with an evaluation of the magnitude of biogenic emissions from artificial reservoirs and subsequently concentrates on hydropower reservoirs, in order to assess

the contribution of these fluxes to the carbon impact of hydropower in Europe.
Among the ML approaches, we implemented only the models based on the new variable set. Results (Sections 5.3 and 5.4) show the comparability of ML models when trained on the two sets of variables. By contrast, relying on the new dataset considerably enhances the applicability of the models (CO_2 and CH_4), making it the most reasonable choice for large-scale applications.

The first step of the European large-dams study regards data collection. As a starting point, we rely on the GRanD database (Global Dam Watch), which represents the standard reference for global studies on reservoirs and dams. It provides their coordinates, year of impoundment, main uses, and features (area and depth). Climatic variables, basin descriptors and soil biogeochemical properties are extracted applying the procedures detailed in Section 4.6, after the definition of the catchment areas.
Missing predictors needed by the G-res models were estimated using the G-res empirical submodels. The application of the original G-res models additionally required a number of assumptions, the most relevant of which concerned soil carbon content. G-res uses as a predictor the concentration of soil organic carbon prior to impoundment (CO_2 model, Eq 3.5). Since extracting this information for each reservoir is not feasible at a large scale, the variable was assumed constant over time. Without this assumption, the number of estimates produced by G-res would decrease drastically, clearly highlighting the limitations of G-res models in large-scale applications.

5.5.1. Estimating the emission from European large dams

After defining the European dataset, the models were applied. In Europe, the ML model estimates biogenic emissions from artificial reservoirs being 18.78 $MtCO_{2eq}$ yr^{-1} for CO_2 diffusion and 4.99 $MtCO_{2eq}$ yr^{-1} for CH_4 diffusion. The corresponding G-res estimates are 13.63 $MtCO_{2eq}$ yr^{-1} and 1.68 $MtCO_{2eq}$ yr^{-1}, respectively.

Figure 5.25 examines this divergence from the point of view of model behavior, illustrating the reservoir-level differences between the ML model and G-res across Europe.
Analysing models tendency in CO_2 application (panel *a*, Figure 5.25), ML model predicts more intense fluxes compared to G-res in northern Europe, while an opposite trend emerges in southern regions. To assess which model better reproduces the real reservoirs emission distribution, we compared the estimated values with the available observations from reservoirs situated in the corresponding regions of the training dataset. Unfortunately, the available flux measurements do not include samples in southern European areas (Figure 4.2) so a direct comparison between estimates and observations is not pos-

sible. However, the relatively warm temperatures, together with the widespread presence of croplands and grasslands characterizing the areas, suggest high emission rates. On the other hand, in northern Europe, the ML model demonstrates a good predictive skill (panel *a*, Figure 5.22), while G-res tend to underestimate (panel *e*, Figure 5.22).

Panel b) of Figure 5.25 examines the model differences in CH_4 fluxes. Darker dots are clustered in the United Kingdom, where our model predictions are significantly higher than G-res. A direct comparison with CH_4 observations is not possible as our dataset does not cover that area. The boreal and temperate climate of the area, in addition to the generally modest organic loads due to the location of the reservoirs, suggests low CH_4 fluxes (McClure et al., 2020); in contrast, the limited size of UK reservoirs favors CH_4 release (Yang et al., 2023). Consequently, it is not possible to determine with confidence which model is more representative of the emission dynamics in this region.
In southern Europe, especially South Spain and the Balkan Peninsula, G-res estimates are higher than the ML model. Accordingly with the analysis of the emission drivers made for CO_2 diffusion in these areas, higher fluxes seem more consistent. In contrast, the Scandinavian peninsula is characterized by red dots. The higher estimates of the ML model seems to be more reliable, considering the underestimation of the emission rate of the region (panels *a* and *c*, Figure 5.24).

a) CO_2 fluxes (ML-G-res) b) CH_4 fluxes (ML-G-res)

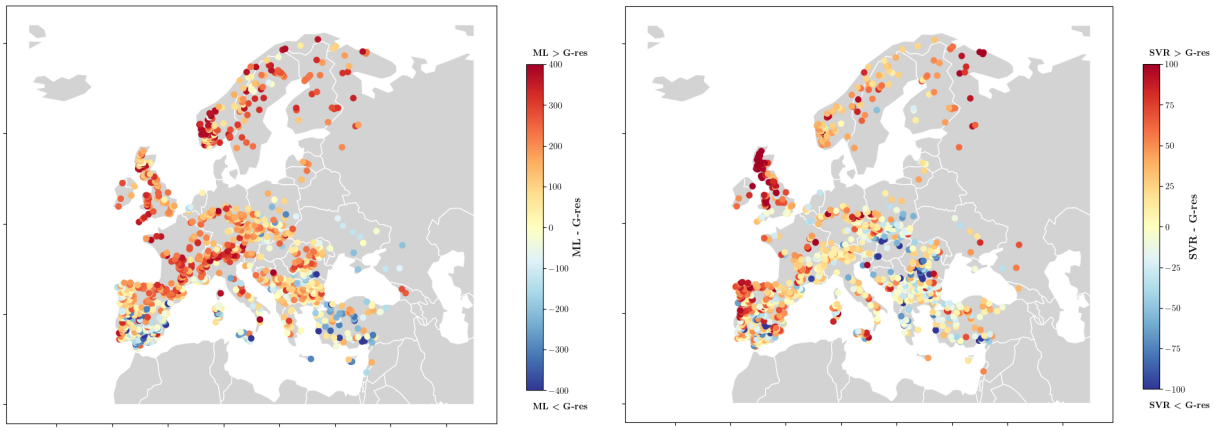

Figure 5.25: Differences in terms of emission intensity estimated by the models.

The analysis of model applicability highlights their differing inclusiveness, which further explains the discrepancies observed in the cumulative values. In line with the findings discussed in Section 5.4, the ML model demonstrates substantially higher applicability than G-res. Specifically, it is able to estimate both CO_2 and CH_4 diffusive emissions for 1281 reservoirs. In contrast, G-res covers 1221 reservoirs in the CO_2 analysis and only 1112 in the CH_4 assessment. Additionally, as mentioned, predictions for CO_2 are possible

only assuming soil carbon content constant before and after the impoundment.
The cumulative values clearly highlight the impact of the more limited spatial coverage of G-res on the total assessment of the biogenic carbon footprint of reservoirs, reducing it by 27.8% for CO_2 and by 66.4% for CH_4.

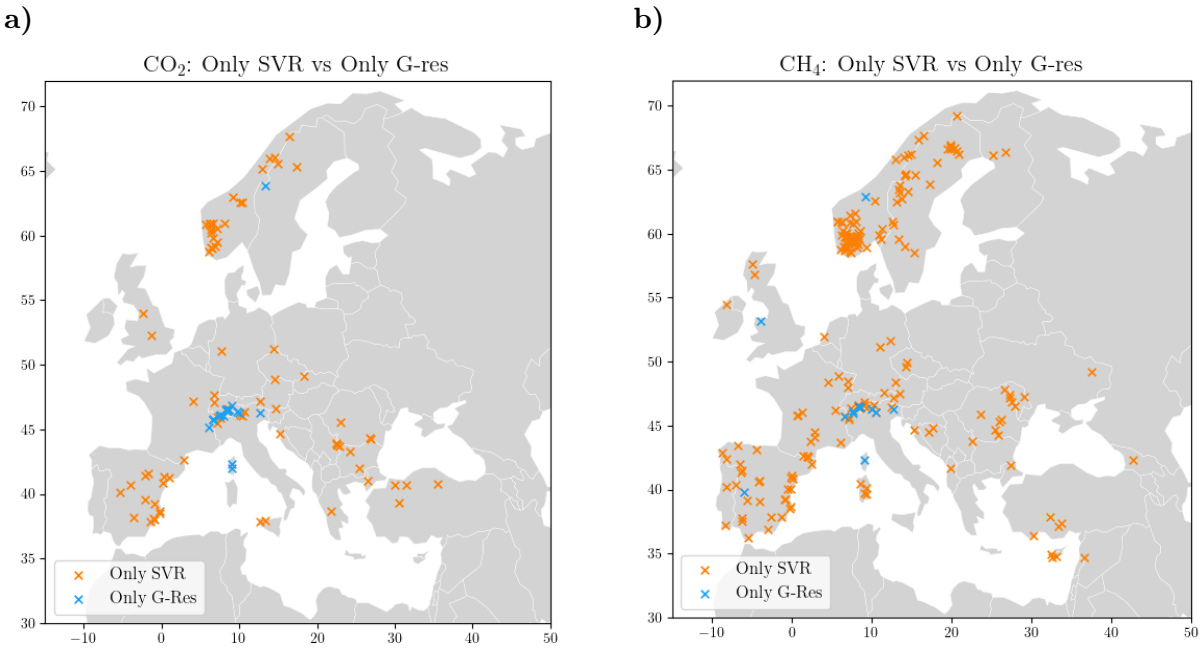

Figure 5.26: Panel *a* and panel *b* represents which reservoirs are estimated by only ML model (orange cross) and G-res model (light blue cross) in CO_2 and CH_4 application, respectively.

The wider coverage of our model enables a more comprehensive analysis of reservoir behavior under site-specific conditions. Panel a) of Figure 5.26 shows that the CO_2 analysis performed by G-res model excludes a considerable number of reservoirs in Spain and in Balkan Peninsula, where the highest fluxes are expected, as previously mentioned. A similar pattern is evident in the CH_4 analysis (panel *b*, Figure 5.26), which is additionally affected by the low coverage in Norway.
The examination of the spatial distribution of emissions provides a key tool for energy planning, as it could support decision-making on both new infrastructure and mitigation strategies. Identifying where the highest fluxes are expected helps to avoid siting reservoirs in highly sensitive areas and guides the selection of renewable sources with lower biogenic carbon impact.

For consistency, we report also the cumulative values computed on the subset of reservoirs estimated simultaneously by the two models. CO_2 total emissions amount to

18.64 $MtCO_{2eq}$ yr^{-1} by the ML model and 13.26 $MtCO_{2eq}$ yr^{-1} by G-res, whereas CH_4 to 3.6 $MtCO_{2eq}$ yr^{-1} and 1.7 $MtCO_{2eq}$ yr^{-1}, respectively. The results highlights that, even considering the same number of reservoirs, G-res tends to predict a lower GHG reservoirs intensity.

Totally, the biogenic impact from European artificial reservoirs estimated by the ML model amounts to 23.77 $MtCO_{2eq}$ yr^{-1}, partitioned into 26% from diffusive CH_4 and 74% from CO_2. This confirms CO_2 diffusion as the dominant emission pathway, in line with Section 4.3.2.
With reference to EDGAR data about GHG emissions, biogenic emissions from European reservoirs account for 4.46% of total European GHG emissions (excluding LULUCF). It should also be noted that this analysis excludes CH_4 bubbling and degassing emissions, whose relevance has been recently emphasized (McClure et al., 2020). These results clearly highlight the contribution of biogenic emissions and the urgent need to improve their quantification and understanding.
Due to the inconsistent number of reservoirs included by the G-res model in the CO_2 and CH_4 assessments, conducting the same analysis is not feasible.

5.5.2. Computing European hydropower footprint

Once assessed the relevance of the European biogenic emission intensity, we focus on the impact that they have on the hydropower carbon footprint. Currently, hydropower reservoirs, either as the main or secondary use, account for nearly 50% of European large dams (768 out 1322). Our model is able to include in its assessment 738 of them, with their flux intensity shown shown in Figure 5.27.

We estimate a continental biogenic release of 20.49 $MtCO_{2eq}$ yr^{-1}, with CO_2 diffusion accounting for the 83%.
This amount relevantly contributes to the European hydropower GHG intensity. NREL (National Renewable Energy Laboratory) reports that hydropower generation causes an average emission of 22 gCO_{2eq} kWh^{-1} yr^{-1} during its life cycle (construction, operation, and disposal phases). While variations may arise due to management practices, materials or construction techniques, the global estimate can be regarded as reasonably representative and sufficiently robust to be used as a reference also in the European context. Multiplying this value by the annual production of European hydropower, equal to 1174077 GWh (IRENA), we can estimate a CO_2 release of 25.32 $MtCO_{2eq}$ yr^{-1}. Adding the biogenic fluxes from reservoirs, the total carbon footprint nearly doubles, reaching an amount of 45.81 $MtCO_{2eq}$ yr^{-1}.

When biogenic emissions are taken into account, the carbon intensity of hydropower plants increases substantially—by 46.82% compared to the LCA value—reaching 46 gCO_{2eq} kWh^{-1}. This adjustment significantly reduces the distinct advantage hydropower previously held over other renewables, placing it instead within their range: wind (7–56 gCO_{2eq} kWh^{-1}) and geothermal (6–79 gCO_{2eq} kWh^{-1}). While hydropower still performs markedly better than non-renewable sources, which emit 290–1689 gCO_{2eq} kWh^{-1} (IPCC), it should be considered at least more directly comparable to other renewable technologies.

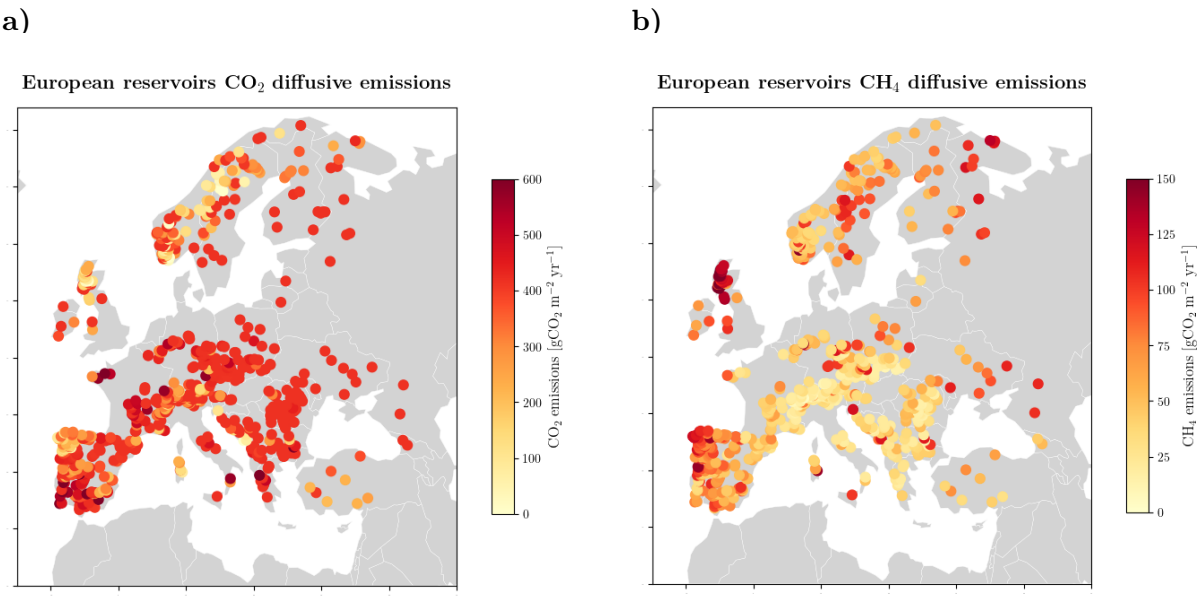

Figure 5.27: CO_2 (panel *a*) and CH_4 (panel *b*) diffusive emissions from European hydropower reservoirs (gCO_{2eq} m^{-2} yr^{-1}).

6 | Conclusions and future research

This study addresses the assessment of biogenic emissions from reservoirs, with a focus on models reliability and accessibility.
GHG fluxes from reservoirs result from complex interactions among environmental, climatic, and biological variables. To date, their representation has relied on linear models, whose inherent simplicity prevents them from capturing the underlying relationships between emissions and their drivers, while data scarcity excludes several important variables. The G-res tool is currently the most advanced approach. It addresses the problem of data scarcity by using empirical models to estimate predictors, but still relies on variables that are available only at local rather than global scale. Its dependence on empirical and non-globally accessible variables reduces both the applicability of the model and the reliability of global estimates of the reservoir GHG footprint.

Our objectives are to improve G-res models accuracy and applicability.
To achieve this, we implement the Support Vector Regression, a Machine Learning algorithm. As result of its greater flexibility, the ML approach is expected to better capture the relationships between predictors and targets, thereby improving model quality. Additionally, it enables the use of a larger set of variables and proxy indicators, making it possible to replace the empirical or unavailable factors used by G-res with globally accessible ones.
This approach results in four final models, two for CO_2 and two for CH_4 diffusion. A first set of CO_2 and CH_4 models use predictors included within G-res dataset, whereas a second set relies on new globally accessible variables. The effectiveness of a ML approach with respect to the linear one is assessed comparing CO_2 and CH_4 models performance with the original G-res ones.

The results of the CO_2 application reveals that both ML models (G-res and new variables) better predict the emissions from reservoirs compared to G-res. Their R^2 are 0.56

and 0.52 using G-res and new variables respectively, whereas G-res R^2 amounts to 0.42. In addition, they also show a lower tendency towards underestimation corresponding to bias values of -61, -139 and -224 gCO_{2eq} m^{-2} yr^{-1}. Both ML models capture the CO_2 emission patterns reasonably well, demonstrating their ability to account for complex and non-linear relationships among target and predictors. Model underestimation remains relevant, highlighting the larger errors made by all models on the extreme values. However, the ML algorithm slightly enhances this aspect too, especially the case of ML model applied on G-res variables.

In the assessment of CH_4, the enhancements observed are limited. In terms of performance, in fact, the improvements expected from the implementation of ML techniques appear negligible. The R^2 and bias values are 0.35 and –75 gCO_{2eq} m^{-2} yr^{-1} for the ML model trained on G-res variables, 0.39 and –73 gCO_{2eq} m^{-2} yr^{-1} for the ML model using new variables, and 0.40 and –59 gCO_{2eq} m^{-2} yr^{-1} for G-res. The three approaches yield comparable results, reproducing the overall trend of CH_4 emissions but showing a tendency towards underestimation. This underestimation is even more pronounced in the ML models compared to G-res. Interestingly, a direct comparison between model estimates and observations shows that G-res exhibits greater flexibility, in contrast to what would normally be expected from empirical linear models versus Machine Learning approaches.

The results obtained by ML models applied to new and G-res dataset are similar both in CO_2 and CH_4 application.

This outcome suggests that the new variables explain GHG emissions to the same degree as those included in G-res. Moreover, as the new dataset excludes variables which are not accessible at large scale, it indicates that the proxy variables adopted are able to represent them effectively.

The major difference between the use of the two sets lies in the applicability of the corresponding models.

The results clearly highlight the improvement in inclusiveness brought by the new dataset. In CO_2 application, the use of the G-res dataset restricts the number of estimates to 119 for the ML model and 207 for the original G-res model, whereas the ML model trained on the new variables successfully produces 280 estimates. The corresponding analysis for CH_4 results in 279 estimates for ML model using the new dataset, 193 for ML model using G-res 239 using G-res original model. This demonstrates that the use of the new dataset increases the number of estimates by approximately 25% for CO_2 and 14% for CH_4, compared to the G-res–based approaches.

Considering the performance and the substantially higher applicability, the ML model built on new variables proves to be suitable for large-scale, even global, applications.

In conclusion, the use of Machine Learning represents a clear advancement in modeling biogenic emissions from reservoirs. While the improvement in model estimates is not always decisive, though still relevant, the main achievement lies in the higher applicability of the models. ML makes it possible to bypass the empirical submodels employed by G-res to estimate covariates, which inherently introduce an additional degree of uncertainty. Most importantly, the new approach relies on raw and easily accessible data, making the models suitable for large-scale and even global applications.
The broad coverage of the new dataset allows for a more comprehensive inclusion of reservoirs and ensures a more reliable quantification of their total carbon footprint. Moreover, it enables for a extensive examination of the spatial distribution of emissions, which is particularly relevant for hydropower reservoirs. Identifying where high emissions are most likely to occur provides practical support for planning future projects, guiding decision-making towards locations with the lowest potential impact.

The results obtained by our models clearly show their relevance in the analysis of large European dams. The ML model based on the new set of variables demonstrates a substantially higher inclusiveness than G-res in this large-scale application, being able to account for 97% of artificial reservoirs. On the other hand, the applicability of G-res reveals its limitations, as it estimates emissions for only 1112 out of 1322 reservoirs. This reduced coverage leads to an underestimation of the carbon footprint of European artificial reservoirs, which is predicted to be 27% and 66% lower than the ML model for CO_2 and CH_4 diffusion, respectively.

The biogenic carbon intensity from large dams demonstrates its significance. Particularly, we estimate that European artificial reservoirs cumulatively emit 23.77 $MtCO_{2eq}\ yr^{-1}$, of which 18.78 $MtCO_{2eq}\ yr^{-1}$ as diffusive CO_2 and 5.99 $MtCO_{2eq}\ yr^{-1}$ as diffusive CH_4. These results confirm the predominant role of CO_2 in the biogenic GHG intensity of reservoirs, while also revealing that CH_4 contribution cannot be ignored.

The analysis focused on hydropower dams predicts that they cumulatively emit 20.49 $MtCO_{2eq}\ yr^{-1}$. These emissions significantly affect the carbon intensity of hydropower plants, doubling the emission attributed to dam construction, operations and disposal. When added to the life-cycle carbon intensity of dams reported by NREL, the GHG emissions from reservoirs raise the average carbon intensity of European hydropower plant to 46 $gCO_{2eq}\ kWh^{-1}$. The inclusion of biogenic emissions in the carbon footprint of hydropower reduces part of its perceived advantage over other energy sources, thereby increasing its comparability with them. Whereas previously hydropower enjoyed a clear advantage, this gap is now less pronounced.

Currently, hydropower-related emissions, including both biogenic and life cycle contributions, represent the 0.86% of total European GHG emissions (EDGAR), thereby supporting its pivotal role in the global energy transition. In Europe, over 600 hydropower projects are currently planned or under construction (VGBE), particularly in the Balkan Peninsula. Spatial analysis of biogenic fluxes indicates that this area is linked to the highest GHG emissions, raising concerns that the carbon intensity of future hydropower developments could be significantly high compared with other renewable technologies. In this context, the application of our model offers a practical support for energy planning, helping to identify areas where hydropower is expected to be most emission-intensive and to consider alternative solutions.

6.1. Limitations and Further Developments

Applying Machine Learning to biogenic emission modeling has proven to yield significant improvements, particularly with regard to its broader applicability. Nevertheless, additional efforts are required to enhance both the accuracy and robustness of the estimates. Particularly, as discussed in Sections 5.1.2 and 5.2.2, the main limitations are related to model instability and a tendency to underestimate emissions.
The challenges encountered in this work guide future research toward the steps needed to improve model quality and advance assessments of the biogenic carbon footprint.

Firstly, expanding the number of flux observations is a clear priority. Limited data availability constrains ML model performance, as these approaches heavily rely on representative datasets. Additional information would strengthen the ability of ML algorithms to capture the relationships between predictors and fluxes, thereby improving the robustness of emission estimates. This effect is already evident in our study. Although the CO_2 dataset includes only 44 more samples than the CH_4 set, it reproduces noticeably better the emission process and improves sensitively the linear approach. Hence, strengthening observational efforts is the first and urgent step toward advancing biogenic GHG modeling.
Data scarcity is also responsible for the exclusion of bubbling and degassing emissions from current assessments of the GHG carbon footprint of reservoirs. Although recent studies have highlighted their relevance (Deemer et al., 2016), the available observations remain too limited to allow for a robust application of ML approaches. Future sampling campaigns will be crucial to expand ML modeling to these emission pathways and to yield more realistic assessment of the biogenic impact of hydropower reservoirs.

Secondly, the most critical aspect of our model is the underestimation of high-emission fluxes. While some improvements have been achieved in the CO_2 diffusion model, ML approaches still tend to severely underestimate biogenic emissions in tropical regions. Improving their quantification is essential, as Africa and South America are expected to experience the highest growth in hydropower capacity (Zarl and Tydecks, 2014) with numerous new plants under construction. Consequently, it is urgent to provide reliable estimates of the real impact of new plants in terms of biogenic emissions.
Once again, the main cause is the scarcity of flux observations in tropical regions, underlining the need for new measurement campaigns.

In line with our findings, additional investigation of the contribution of biogenic emissions to the carbon footprint of hydropower is suggested.
Our assessment of European hydropower reservoirs shows that these fluxes can double the carbon intensity of hydropower, even larger increases are expected in regions characterized by higher emission rates. Consequently, we suggest performing further analyses of the spatial variability of carbon intensity, with a particular focus on tropical areas, an aspect that could not be explored within the timeframe of this thesis.
It should also be noted that the revised hydropower carbon intensity estimates presented here exclude CH_4 bubbling and degassing emissions, which are likely to further increase the overall values. As a consequence, we emphasize the need for future efforts to model these pathways in order to provide a realistic assessment of the climate impact of hydropower.

Finally, for a proper assessment of artificial reservoirs GHG impact, only anthropogenic emissions—those caused by impoundment and reservoir management—should be considered (Section 2.1). A partition between pre- and post-impoundment emissions has already been proposed by G-res, which computes net emissions by subtracting the soil emissions that would have occurred in the absence of the reservoir, estimated through a zonal statistics approach.
Moreover, recent studies have shown that part of the emissions observed at the reservoir site would have occurred even in the absence of the dam, as downstream emissions tend to be reduced. In other words, these fluxes do not represent additional emissions but rather a spatial shift, occurring upstream in the reservoir instead of further downstream. Consequently, this portion also needs to be discounted (Yan et al., 2022).
This represents a limitation of our ML approach, which could be addressed in future work by expanding the use of ML to estimate and account for non-anthropogenic emissions.

For future research, and in light of the pivotal role that hydropower is expected to play in the global energy transition, a key challenge will be to estimate the emissions associated

with future development scenarios. The use of environmental variables such as temperature, wind speed, land use, and population offers the possibility to extend the analysis beyond present conditions and to project the future carbon footprint of hydropower under different scenarios.

Ultimately, such foresight offers critical guidance for energy planning and development, helping to ensure that the transition progresses along a truly sustainable trajectory.

Bibliography

T. Abbasi and S. A. Abbasi. A model to forecast methane emissions from tropical and subtropical reservoirs on the basis of artificial neural networks. *Water*, 2020.

J. Almeida, Nobrega and Figueiredo. High primary production contrasts with intense carbon emission in a eutrophic tropical reservoir. *Frontiers*, 2016.

M. C. V. T. Almeida and P. S. Coelho. An integrated approach based on the correction of imbalanced small datasets and the application of machine learning algorithms to predict total phosphorus concentration in rivers. *Ecological Informatics*, 2023.

L. A. W. Bambace, F. M. Ramos, I. B. T. Lima, and R. R. Rosa. Mitigation and recovery of methane emissions from tropical hydroelectric dams. *Energy*, 2007.

N. Barros, J. J. Cole, L. J. Tranvik, Y. T. Prairie, D. Bastviken, V. L. M. Huszar, P. del Giorgio, and F. Roland. Carbon emission from hydroelectric reservoirs linked to reservoir age and latitude. *Nature Climate Change*, 2011.

S. Bragadeesh. Linearity and non-linearity in machine learning, 2020.

J. Brownlee. Nested cross-validation for machine learning with python, 2021.

Copernicus Climate Change Service (C3S). Era5 monthly averaged data on single levels from 1979 to present, 2017.

S. R. De Faria, Jaramillo and Barros. Estimating greenhouse gas emissions from future amazonian hydroelectric reservoirs. *Environmental research*, 2015.

B. R. Deemer, J. A. Harrison, S. Li, J. J. Beaulieu, T. DelSontro, N. Barros, J. F. Bezerra-Neto, S. M. Powers, M. A. dos Santos, and J. A. Vonk. Greenhouse gas emissions from reservoir water surfaces: a new global synthesis. *BioScience*, 2016.

T. DelSontro. *Quantifying Methane Emissions from Reservoirs: From Basin-scale to Discrete Analyses with a Focus on Ebullition Dynamics*. Ph.d. dissertation, Université du Québec à Montréal, 2011.

T. DelSontro and P. A. del Giorgio. Co_2 emissions from reservoirs: A global perspective on emissions and their controls. *Limnology and Oceanography Letters*, 2018.

T. DelSontro, J. J. Beaulieu, and J. A. Downing. Greenhouse gas emissions from lakes and impoundments: Upscaling in the face of global change. *Limnology and Oceanography Letters*, 2018.

T. Diem, S. Koch, S. Schwarzenbach, B. Wehrli, and C. J. Schubert. Greenhouse gas emissions (co, ch, and no) from several perialpine and alpine hydropower reservoirs by diffusion and loss in turbines. *Aquatic Sciences*, 2012.

EDGAR. Edgar - emissions database for global atmospheric research.

Ember. Global electricity review 2025. Technical report, Ember, 2025.

European Commission, Joint Research Centre (JRC). Copernicus global human settlement layer – download portal.

M. Forsberg and Dunne. The potential impact of new andean dams on amazon fluvial ecosystems. *Plos ONE*, 2017.

G-res Team. G-res tool technical document, version 3.3, 2022.

G-res Tool. G-res Tool: The Hydropower Greenhouse Gas Emissions Tool.

GeeksforGeeks. Support vector regression (svr) using linear and non-linear kernels in scikit-learn.

Global Dam Watch. Grand - global reservoir and dam database, version 1.3.

C. Gudasz, D. Bastviken, K. Steger, K. Premke, S. Sobek, and L. J. Tranvik. Temperature-controlled organic carbon mineralization in lake sediments. *Nature*, 2010.

E. G. Hertwich. Ghg emissions from the global hydropower reserve: A critical review and future perspectives. *Environmental Science & Technology*, 2010.

E. G. Hertwich. Addressing biogenic greenhouse gas emissions from hydropower in lca. *Environmental Science & Technology*, 2013.

A. Hou, L. Duan, J. Li, Q. Zhang, C. He, W. Qu, X. Chen, and J. Zhang. Methane emissions from the surface of the three gorges reservoir in china. *Environmental Science & Technology*, 2013.

HydroSHEDS Team. HydroSHEDS: Hydrological data and maps based on SHuttle Elevation Derivatives at multiple Scales.

IHA. International hydropower association.

Ion and Ene. Evaluation of greenhouse gas emissions from reservoirs: A review. *Sustanability*, 2021.

IPCC. Ipcc - intergovernmental panel on climate change.

IRENA. IRENA-Renewable Capacity Statistics 2025.

ISRIC – World Soil Information. Soilgrids: Global gridded soil information, 2024.

H. I. Jager, R. M. Pilla, C. H. Hansen, P. G. Matson, B. Iftikhar, and N. A. Griffiths. Understanding how reservoir operations influence methane emissions: A conceptual model. *Environmental Science Technology*, 2023.

T. Janus. reemission: Ghg emissions from reservoirs, 2023.

W. Jin. Nested cross validation explained, August 25 2018.

M. S. Johnson, E. Matthews, D. Bastviken, B. Deemer, J. Du, and V. Genovese. Spatiotemporal methane emission from global reservoirs. *Journal of Geophysical Research: Biogeosciences*, 2021.

M. Kang, L. Liu, and H.-P. Grossart. Spatio-temporal variations of methane fluxes in sediments of a deep stratified temperate lake. *Science of The Total Environment*, 2023.

F. Karambelkar and Ames. Hydropower reservoir greenhouse gas emissions: State of the science and roadmap for further research to improve emission accounting and mitigation. *Sustanability*, 2025.

C. A. Kelly, J. W. M. Rudd, V. L. St. Louis, and A. Heyes. Changes in fluxes of greenhouse gases and methyl mercury with flooding of an experimental boreal forest reservoir. *Environmental Science & Technology*, 1997.

A. Levasseur, S. Mercier-Blais, Y. Prairie, A. Tremblay, and C. Turpin. Improving the accuracy of electricity carbon footprint: Estimation of hydroelectric reservoir greenhouse gas emissions. *Renewable and Sustainable Energy Reviews*, 2021.

C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 2013.

C. Li, Y. Wang, Y. Yi, X. Wang, C. A. G. Santos, and Q. Liu. A review of reservoir carbon cycling: Key processes, influencing factors and research methods. *Renewable and Sustainable Energy Reviews*, 2024.

Y. Li and R. He. Carbon intensity of global existing and future hydropower reservoirs. *Renewable and Sustainable Energy Reviews*, 2022.

C. Lu and H. Tian. Global nitrogen and phosphorus fertilizer use for agriculture production in the past half century: Shifted hot spots and nutrient imbalance. *Earth System Science Data*, 2017.

A. Maeck, T. Delsontro, D. F. McGinnis, H. Fischer, S. Flury, M. Schmidt, P. Fietzek, and A. Lorke. Sediment trapping by dams creates methane emission hot spots. *Environmental Science & Technology*, 2013.

R. P. McClure, M. E. Lofton, S. Chen, K. M. Krueger, J. C. Little, and C. C. Carey. The magnitude and drivers of methane ebullition and diffusion vary on a longitudinal gradient in a small freshwater reservoir. *Ecosystems*, 2020.

R. McDowell. Global available soil phosphorus database, 2023.

S. Mercier-Blais. Reservoir emissions from 1900–2060: How does the timeline look?, 2024.

R. Mustafa, M. T. Ahmad, A. Kumar, S. Kumar, N. K. Sah, and A. Kumar. Prediction of central deflection and slenderness limit for lateral stability of simply supported concrete beam using machine learning techniques. *Asian Journal of Civil Engineering*, 2024.

NASA Earthdata. Modis terra+aqua land cover type yearly l3 global 500m sin grid version 6.1 (mcd12q1), 2024.

National Renewable Energy Laboratory. About nrel.

J. R. Paranaíba, N. Barros, R. Mendonça, A. Linkhorst, A. Isidorova, F. Roland, R. M. Almeida, and S. Sobek. Spatially resolved measurements of co_2 and ch_4 concentration and gas-exchange velocity highly influence carbon-emission estimates of reservoirs. *Environmental Science & Technology*, 2018.

W. W. Phyoe and F. Wang. A review of carbon sink or source effect on artificial reservoirs. *International Journal of Sustainable Development and Planning*, 2019.

D. A. Pisner and D. M. Schnyer. Support vector machine. 2020.

Y. T. Prairie, J. Alm, A. Harby, S. Mercier-Blais, R. Nahas, V. Chanudet, J. A. Harrison, and C. Soued. A new modelling framework to assess the biogenic emissions of greenhouse gases from reservoirs. *Environmental Modelling & Software*, 2017.

Y. T. Prairie, J. Alm, J. Beaulieu, N. Barros, T. Battin, J. Cole, P. del Giorgio, T. DelSontro, F. Guérin, A. Harby, J. Harrison, S. Mercier-Blais, D. Serça, S. Sobek, and

D. Vachon. Greenhouse gas emissions from freshwater reservoirs: What does the atmosphere see? *Ecosystems*, 2018.

P. A. Raymond, J. Hartmann, R. Lauerwald, S. Sobek, C. McDonald, M. Hoover, D. Butman, R. G. Striegl, E. Mayorga, C. Humborg, P. Kortelainen, H. H. Dürr, M. Meybeck, P. Ciais, and P. Guth. Global carbon dioxide emissions from inland waters. *Science*, 2013.

F. Roland, L. O. Vidal, F. S. Pacheco, N. O. Barros, A. Assireu, J. P. H. B. Ometto, A. C. P. Cimbleris, and J. J. Cole. Variability of carbon dioxide flux from tropical (cerrado) hydroelectric reservoirs. *Aquatic Sciences*, 2010.

D. Rosalina, G. Abril, J. Deborde, F. Wit, R. Delmas, M. Demarty, W. Rode, A. Vongkhamsao, V. Chanudet, S. Descloux, and F. Guérin. Variability of greenhouse gas emissions from tropical reservoirs over daily to monthly time scales. 2016.

S. Savale. Machine learning: An overview. *International Journal of Research - Granthaalayah*, August 2021.

L. Scherer and S. Pfister. Hydropower's biogenic carbon footprint. *PLoS ONE*, 2016.

C. L. Schneider, C. Vatovec, J. P. D. Abbatt, W. S. Compton, J. Junker, et al. Carbon dioxide (co_2) fluxes from terrestrial and aquatic systems in headwater catchments: A combined field and modelling approach. *Journal of Geophysical Research: Biogeosciences*, 2020.

A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 2004.

V. L. St. Louis, C. A. Kelly, E. Duchemin, J. W. M. Rudd, and D. M. Rosenberg. Reservoir surfaces as sources of greenhouse gases to the atmosphere: A global estimate. *Environmental Science Technology*, 2000.

J. A. K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Artificial Intelligence Review*, 1999.

T. Talaei Khoei and N. Kaabouch. Machine learning: Models, challenges, and research directions. *Future Internet*, 2023.

VGBE. VGBE Energy Association – technical association for generation and storage of electricity and heat.

Z. Wang, F. K. S. Chan, M. Feng, and M. F. Johnson. Greenhouse gas emissions from

hydropower reservoirs: emission processes and management approaches. *Environmental Research Letters*, 2024.

J. Xu et al. A comprehensive framework for estimating greenhouse gas emissions from reservoirs. *Renewable and Sustainable Energy Reviews*, 2021.

D. Yan, X. Li, G. Zhao, J. Zhang, H. Wang, Z. Li, and S. Liu. Reservoirs change pco_2 and water quality of downstream rivers: Evidence from three reservoirs in china. *Water Research*, 2022.

X. Yan, V. Thieu, and J. Garnier. Long-term evolution of greenhouse gas emissions from global reservoirs. *Global Biogeochemical Cycles*.

H. Yang, G. He, H. Chen, and L. Tong. Spatial and temporal variations of methane emissions from reservoirs in china. *Biogeosciences*, 2014.

P. Yang, Y. Zhang, H. Yang, Q. Guo, D. Y. Lai, G. Zhao, L. Li, and C. Tong. Ebullition was a major pathway of methane emissions from the aquaculture ponds in southeast china. *Water Research*, 2023.

Y. Yang, S. Zhang, W. Wang, and H. Liu. Understanding the carbon footprint of hydropower reservoirs: A global assessment based on machine learning approach. *Science of The Total Environment*, 2024.

Z. Yang, Lu and Wang. Progress in the studies on the greenhouse gas emissions from reservoirs. *Acta Ecologica Sinica*, 2014.

B. Zarl, Lumsdon and Tydecks. A global boom in hydropower dam construction. *Acquatic science*, 2014.

List of Figures

List of Tables

A | Acknowledgements

Questa tesi segna la conclusione del percorso magistrale intrapreso presso il Politecnico di Milano, che è stato per me fonte di crescita sia accademica che personale.
Colgo quindi questa occasione per ringraziare tutte le persone che mi hanno accompagnato lungo questo percorso.

La mia gratitudine va innanzitutto al Prof. Andrea Castelletti, per avermi dato l'opportunità di lavorare a questo progetto. Desidero ringraziare anche il Prof. Rafael Schmitt, che mi ha ospitato presso l'Università UC Santa Barbara, e che è stato un prezioso supporto in tutte le fasi del lavoro, e il Dr. Bruno Invernizzi, il cui aiuto è stato fondamentale per superare ogni difficoltà.

Un grazie speciale va alla mia famiglia senza la quale questa esperienza non sarebbe stata possibile, e ai miei nonni, che sono per me il punto di riferimento da sempre.

Infine, non posso che ringraziare i miei amici, chi c'è da sempre e chi ho incontrato durante questo percorso, per aver portato leggerezza e supporto quando servivano.

A.1. English Version

This thesis marks the conclusion of the Master's program undertaken at Politecnico di Milano, which has been for me a source of both academic and personal growth. I would therefore like to take this opportunity to thank all the people who have accompanied me along this journey.

My gratitude goes first and foremost to Prof. Andrea Castelletti, for giving me the opportunity to work on this project. I would also like to thank Prof. Rafael Schmitt, who welcomed me at UC Santa Barbara and provided invaluable support throughout all phases of the work, and Bruno Invernizzi, whose help was essential in overcoming every challenge.

A special thanks goes to my family, without whom this experience would not have been possible, and to my grandparents, who have always been my point of reference.

Finally, I warmly thank my friends—those who have always been there and those I met along the way—for bringing lightness and support whenever it was most needed.