UNIVERSIDADE FEDERAL DE MINAS GERAIS
Escola de Engenharia
Programa de Pós-Graduação em Saneamento, Meio Ambiente e Recursos Hídricos

Marina Salim Dantas

MACHINE LEARNING ALGORITHMS FOR ASSESSMENT AND PREDICTION OF THE PERFORMANCE OF WASTEWATER TREATMENT PLANTS

Belo Horizonte
2024

Marina Salim Dantas

MACHINE LEARNING ALGORITHMS FOR ASSESSMENT AND PREDICTION OF THE PERFORMANCE OF WASTEWATER TREATMENT PLANTS

Tese apresentada ao Programa de Pós-graduação em Saneamento, Meio Ambiente e Recursos Hídricos da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Doutora em Saneamento, Meio Ambiente e Recursos Hídricos.

Área de concentração: Meio Ambiente

Linha de pesquisa: Caracterização, prevenção e controle da poluição

Orientadora: Profa. Dra. Sílvia Maria Alves Corrêa Oliveira

Coorientador: Prof. Dr. Cristiano Christofaro Matosinhos (UFVJM)

Belo Horizonte
2024

Dantas, Marina Salim.
D102m Machine learning algorithms for assessment and prediction of the performance of wastewater treatment plants [recurso eletrônico] / Marina Salim Dantas. – 2024.
1 recurso online (203 f. : il., color.) : pdf.

Orientadora: Sílvia Maria Alves Corrêa Oliveira.
Coorientador: Cristiano Christofaro Matosinhos.

Tese (doutorado) - Universidade Federal de Minas Gerais, Escola de Engenharia.

Apêndices: f. 196-203.

Bibliografia: f. 183-195.

1. Engenharia sanitária - Teses. 2. Meio ambiente - Teses. 3. Inteligência artificial - Teses. 4. Ciência de dados - Teses. 5. Aprendizado do computador - Teses. 6. Estação de Tratamento de Esgoto (ETE) - Teses. I. Oliveira, Sílvia Maria Alves Corrêa. II. Matosinhos, Cristiano Christófaro. III. Universidade Federal de Minas Gerais. Escola de Engenharia. IV. Título.

CDU: 628(043)

Ficha catalográfica elaborada pelo bibliotecário Reginaldo César Vital dos Santos - CRB-6/2165
Biblioteca Prof. Mário Werneck - Escola de Engenharia da UFMG

UNIVERSIDADE FEDERAL DE MINAS GERAIS

Escola de Engenharia
Curso de Pós-Graduação em Saneamento, Meio Ambiente e Recursos Hídricos

"Machine Learning Algorithms For Assessment And Prediction Of The Performance Of Wastewater Treatment Plants"

MARINA SALIM DANTAS

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores:

Prof. CRISTIANO CHRISTOFARO MATOSINHOS - COORIENTADOR

Prof. MARCELO CARDOSO

Profa MARIA CRISTINA DE ALMEIDA SILVA

Profa KARLA PATRICIA SANTOS OLIVEIRA RODRÍGUEZ ESQUERR

Profa ARIUSKA KARLA BARBOSA AMORIM

Aprovada pelo Colegiado do PG SMARH Versão Final aprovada por

Profa. Priscilla Macedo Moura Oliveira
Coordenadora

Profª. Sílvia Maria Alves Corrêa
Orientadora

Belo Horizonte, 20 de agosto de 2024.

Documento assinado eletronicamente por **Karla Patricia Santos Oliveira Rodriguez Esquerre**, **Usuária Externa**, em 20/08/2024, às 15:18, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Ariuska Karla Barbosa Amorim**, **Usuária Externa**, em 20/08/2024, às 16:14, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Cristiano Christofaro Matosinhos**, **Usuário Externo**, em 20/08/2024, às 16:15, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Maria Cristina de Almeida Silva**, **Usuária Externa**, em 20/08/2024, às 19:34, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Marcelo Cardoso**, **Professor do Magistério Superior**, em 22/08/2024, às 12:06, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Eduardo Coutinho de Paula**, **Coordenador(a) de curso de pós-graduação**, em 09/10/2024, às 10:16, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3459665** e o código CRC **200F5A93**.

Referência: Processo nº 23072.245238/2024-08 SEI nº 3459665

Aos meus pais e doutores da vida, Marta e Fernando

AGRADECIMENTOS

Depois de quatro anos de muito trabalho, olho para trás e não posso deixar de reconhecer o apoio de tantas pessoas e instituições que tive para chegar até aqui.

Agradeço à UFMG, minha segunda casa, universidade que me recebeu desde a graduação, no mestrado e no doutorado. Ao Programa de Pós-Graduação em Saneamento, Meio Ambiente e Recursos Hídricos, pela minha formação. Aos professores, pelos ensinamentos valiosos.

Agradeço à professora Sílvia, minha orientadora. Nesses anos todos trabalhando juntas, meu carinho e admiração só cresceram a cada dia. Obrigada pela confiança depositada em mim e meu trabalho e pelo auxílio inestimável na tese. Ao meu coorientador, professor Cristiano, pela dedicação, pelos ensinamentos e pelo incentivo de sempre. Obrigada por cada reunião, discussão sobre o trabalho e cada leitura atenta da tese, com importantes contribuições.

Aos colegas do Grupo de Estudos para Tratamento Estatístico de Dados Ambientais (GETEDA) pela parceria. Agradeço em especial às professoras Carol e Lenora pelo apoio na pesquisa.

Aos membros da banca, professores Maria Cristina, Karla, Ariuska e Marcelo, pela disponibilidade em avaliar o trabalho e pelas contribuições valiosas para a melhoria do documento.

À Companhia de Saneamento Ambiental do Distrito Federal (Caesb) e ao Metropolitan Water Reclamation District of Greater Chicago (MWRD) pela disponibilização dos dados, sem os quais esta pesquisa não seria possível.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo amparo concedido para o desenvolvimento da pesquisa.

À Baylor University e ao Department of Statistical Science pela acolhida durante meu doutorado sanduíche. À minha orientadora do período nos Estados Unidos, professora Mandy, por me receber com carinho não somente em seu grupo de pesquisas, mas

também pela acolhida em Waco. Obrigada por me possibilitar viver essa experiência de tantos aprendizados.

À Fulbright, por viabilizar a realização do doutorado sanduíche e me permitir fazer parte dessa rede de pesquisadores e profissionais de excelência.

Aos meus pais, por todo o suporte, incentivo aos estudos e à minha formação. À minha mãe, Marta, por ser fonte constante de amor, cuidado e bondade. Obrigada por cada conversa, conselho amigo, e por estar sempre ao meu lado. Ao meu pai, Fernando, por ser exemplo de dedicação à formação acadêmica, sempre com boas ideias e discussões. Obrigada pelo apoio constante em cada etapa da minha vida. Ao meu irmão, François, por, desde sempre, ser meu exemplo e referência de ser humano, de amizade e companheirismo. Obrigada em especial por ter compartilhado a jornada do doutorado comigo, dividindo angústias, desafios, e alegrias do processo. Ao Vinícius, por me incentivar em cada aventura da vida, por me encorajar a seguir os meus sonhos e sempre acreditar em mim. Obrigada por me impulsionar e ao mesmo tempo ser meu porto seguro.

A Deus, por me proteger e guiar e iluminar o meu caminho.

“It’s impossible”, said pride.

“It’s risky”, said experience.

“It’s pointless”, said reason.

“Give it a try”, whispered the heart.

RESUMO

O lançamento de efluentes domésticos brutos ou tratados de maneira insuficiente representa a maior pressão sobre os recursos hídricos no Brasil. Estações de tratamento de esgotos (ETEs) são essenciais para a garantia da saúde pública e ambiental. No entanto, problemas de projeto e operação podem levar à geração de efluentes em desconformidades com padrões ambientais. Desta forma, o monitoramento dos processos de tratamento é de fundamental importância, permitindo a análise dos dados gerados nos programas de monitoramento e ampliando o conhecimento sobre os sistemas em operação, com reflexos positivos no desempenho dos processos. Técnicas que utilizam o aprendizado de máquina, definido como um conjunto de ferramentas que utilizam computadores para transformar dados em conhecimento, são promissoras para o monitoramento de processos de tratamento de esgotos. Dentre os algoritmos existentes, as redes neurais artificiais (RNAs) têm sido amplamente empregadas no setor, uma vez que são robustas e eficientes para lidar com dados não lineares e complexos, como os dados de monitoramento de estações de tratamento de esgotos. A configuração dos modelos e seus hiperparâmetros é uma tarefa essencial para um desempenho satisfatório dos modelos preditivos. Apesar disso, não existe um direcionamento sobre as melhores práticas de configuração dos modelos de RNAs, e esta tarefa é deixada a cargo da experiência do pesquisador. Neste estudo, uma revisão sistemática da literatura sobre o uso de RNAs na previsão do desempenho de ETEs em escala real foi conduzida. Publicações que modelaram diferentes sistemas de tratamento de efluentes domésticos em diversos países foram levantadas e selecionadas e as principais características dos modelos utilizados foram identificadas. Em seguida, o estudo selecionou uma ETE em operação no Brasil e uma nos Estados Unidos como estudos de caso para aplicação da ferramenta. A ETE Brasília Sul foi escolhida por ser a maior ETE do Distrito Federal em termos de vazão de projeto e apresentar tecnologia de tratamento avançada para a remoção de nitrogênio e fósforo. A ETE John E. Egan, em operação em Illinois/EUA, foi selecionada por possuir porte similar à ETE Brasília Sul, mas banco de dados com diferentes características, o que poderia levar a melhor desempenho do modelo de RNA. Análises aprofundadas foram conduzidas nos dados de monitoramento para compreender o desempenho das ETEs e identificar importantes definições para a etapa de aplicação dos modelos. A última etapa da pesquisa foi a aplicação das RNAs para a previsão das concentrações efluentes, principalmente de nutrientes, nas ETEs Brasília Sul e John E. Egan. De maneira geral, os modelos tiveram adequado desempenho preditivo. No entanto, para os modelos da ETE Brasília Sul, houve maior sobreajuste aos dados de treinamento, o que levou a um declínio no desempenho durante o teste dos modelos. Os resultados podem ser explicados devido às diferentes tecnologias de tratamento e distintas características dos bancos de dados das estações, como número de observações e frequências de monitoramento. Espera-se que este trabalho contribua para a melhor gestão dos sistemas de esgotamento sanitário no Brasil, principalmente no que diz respeito à avaliação e previsão do desempenho de processos.

Palavras-chave: inteligência artificial; redes neurais artificiais; ciência de dados; aprendizado de máquina; tratamento de esgotos.

ABSTRACT

The discharge of untreated or inadequately treated domestic effluents represents the main source of pressure on water resources in Brazil. Wastewater treatment plants (WWTPs) are essential for ensuring public and environmental health. However, in Brazil, design and operation issues can lead to the generation of effluents that do not meet environmental standards. Therefore, monitoring treatment processes is critical, as analysis of data generated from monitoring programs allows for a deeper understanding of the treatment systems in operation and improvements in process performance. Techniques that use machine learning, defined as a set of tools that use computers to transform data into knowledge, are promising for monitoring wastewater treatment processes. Among the existing algorithms, artificial neural networks (ANNs) have been widely employed in the sector as they are robust and efficient in dealing with nonlinear and complex data, such as wastewater treatment monitoring data. Adequate model configuration and hyperparameter settings are essential as they significantly affect predictive model performance. However, there is no guidance on best practices for configuring ANN models, leaving this task to the researcher's experience. This study conducted a systematic literature review on the use of ANNs in predicting the performance of full-scale WWTPs. Publications that modeled different sewage treatment systems in various countries were selected, and the main characteristics of the developed ANN models were identified. Next, the study selected WWTPs in operation in Brazil and the United States as case studies for applying the technique. Brasília Sul WWTP was chosen because it is the largest facility in the Federal District in terms of design flow and employs advanced treatment technology for removing nitrogen and phosphorus. John E. Egan WWTP, operating in Illinois/USA, was selected as it has a similar size to the Brasília Sul facility but a dataset with different characteristics, which could lead to a better performance of the ANN model. Extensive analyses were conducted on the monitoring data to understand the WWTPs' performance and identify important definitions for the next step of model application. The final stage was the application of ANNs to predict effluent concentrations, especially nutrients, in Brasília Sul and John E. Egan facilities. In general, models had adequate predictive performance. However, Brasília Sul WWTP models had some overfitting to the training data, which led to a worse performance during model testing. The results can be explained by the different treatment technologies and the distinct characteristics of the datasets of the facilities, such as the number of observations and monitoring frequencies. This work is expected to contribute to better sanitation systems management in Brazil, mainly concerning the evaluation and prediction of process performance.

Keywords: artificial intelligence; artificial neural networks; data science; machine learning; sewage treatment.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

A^2O – Anaerobic/anoxic/oxic process

Adasa – Agência Reguladora de Águas, Energia e Saneamento Básico do Distrito Federal (Regulatory Agency for Water, Energy, and Basic Sanitation of the Federal District of Brazil)

AHS – American Housing Survey

ANA – Agência Nacional de Águas e Saneamento Básico (National Water and Sanitation Agency)

AI – Artificial intelligence

ANFIS – Adaptive neuro-fuzzy inference system

ANFIS-GA – Adaptive neuro-fuzzy inference system coupled with genetic algorithm

ANN – Artificial neural network

BOD – Biochemical oxygen demand

Caesb – Companhia de Saneamento Ambiental do Distrito Federal (Environmental Sanitation Company of the Federal District)

CBOD – Carbonaceous biochemical oxygen demand

COD – Chemical oxygen demand

CODf – Filtered chemical oxygen demand

CONAMA – Conselho Nacional do Meio Ambiente (National Environment Council)

CV – Coefficient of variation

CWNS – Clean Watersheds Needs Survey

DCB – Deep cascade-forward backpropagation networks

DFFNN – Deep feedforward neural network

DSAE-NN-GA – Deep learning which combines stacked autoencoders with neural network and genetic algorithm

EBPR – Enhanced biological phosphorus removal

EC – Electrical conductivity

E. coli – *Escherichia coli*

EDA – Exploratory data analysis

ELM – Extreme learning machine

ELM-GA – Extreme learning machine coupled with genetic algorithm

FFNN – Feedforward neural network

GA – Genetic algorithm

GRNN – Generalized regression neural networks

HELM – Hierarchical extreme learning machine

IQR – Interquartile range

JMP – Joint Monitoring Programme for Water Supply, Sanitation and Hygiene

LSTM – Long short-term memory

LSTM-AM – Long short-term memory based on attention mechanism

MAE – Mean absolute error

MWRD – Metropolitan Water Reclamation District of Greater Chicago

MLP – Multilayer perceptron network

MLP-GA – Multilayer perceptron network coupled with genetic algorithm

MLR – Multiple linear regression

MLSS – Mixed-liquor suspended solids

MSE – Mean square error

NARX – Nonlinear autoregressive with exogenous neural network

NH_4-N – Ammonia nitrogen

NO_2-N – Nitrite nitrogen

NO_3-N – Nitrate nitrogen

NPDES – National Pollutant Discharge Elimination System

O&G – Total oil and grease

PAO – Phosphorus accumulating organisms

PHB – Polyhydroxybutyrate

PO_4 – Phosphate/orthophosphate

pH – Potential hydrogen

Q – Flow rate

R – Correlation coefficient

R^2 – Coefficient of determination

RBF – Radial basis function neural network

RBF-GA – Radial basis function neural network coupled with genetic algorithm

RHONN – Recurrent high-order neural network

RMSE – Root mean square error

RVFL – Random Vector Functional Link Networks

SNIS – Sistema Nacional de Informações Sobre Saneamento (National System of Information on Sanitation)

SO_4 – Sulfate

SO-RBF – Self-organizing radial basis function neural network

SS – Sedimentable solids

SWNN – Small-world neural network

T – Temperature

TA – Total alcalinity

TFS – Total fixed solids

TKN – Total Kjeldahl nitrogen

TN – Total nitrogen

TP – Total phosphorus

TS – Total solids

TSS – Total suspended solids

TTC – Thermotolerant coliforms

TVS – Total volatile solids

UASB – Upflow anaerobic sludge blanket

UNICEF – United Nations Children's Fund

USEPA – United States Environmental Protection Agency

WHO – World Health Organization

WRP – Water reclamation plant

WWTP – Wastewater treatment plant

SUMÁRIO

CHAPTER 1: INTRODUCTION

1.1 BACKGROUND AND JUSTIFICATION

Population growth in recent decades, especially in urban areas, has led to growing pressures on the environment. In particular, the increased amount of sewage generated has put additional pressure on water resources. The demand for improved public health and environmental quality has motivated researchers and practitioners to focus their attention on wastewater treatment plants (WWTPs) (PHAM et al., 2020). Although WWTPs play a critical role in maintaining the water quality of receiving water bodies, they face significant challenges due to escalating demand and stricter environmental regulations (HANSEN; STOKHOLM-BJERREGAARD; DURDEVIC, 2022; XU et al., 2023).

Wastewater treatment methods involve a combination of physical, chemical, and biological processes to remove contaminants from effluents. The performance of a WWTP is linked to a series of complex processes, such as its size, the treatment technology adopted, and the operational environment of the processes (WANG et al., 2024b). Each wastewater treatment technology has its own advantages and constrains in terms of feasibility, efficiency, practicability, reliability, environmental impact, sludge production, operation level, and pre-treatment requirements (CRINI; LICHTFOUSE, 2019). Therefore, there is no single optimum solution for all scenarios, and the objectives of the treatment must be clearly defined when selecting the technology to be adopted in each situation. The requirements for the effluent quality to be achieved depend on the specific local legislation, which may define quality standards for effluent discharge and the receiving water body (VON SPERLING, 2014). Other factors that influence the conception and design of treatment technologies include local aspects, such as the form of occupation of the cities, population size and density, socioeconomic conditions, area availability, land topography, and soil and subsoil characteristics (ANA, 2020).

In a WWTP, several successive steps are carried out to remove pollutants from the liquid phase. Although not all WWTPs contain all stages in their process, they are described below. The first step is the preliminary treatment, which aims to remove coarse solids through physical mechanisms (CRINI; LICHTFOUSE, 2019; VON SPERLING, 2014). The primary treatment seeks to remove settleable solids and some of the organic matter, and physical mechanisms are also predominant in this step (VON

SPERLING, 2014). The secondary treatment aims at removing organic matter and possibly nutrients (nitrogen and phosphorus) through biological processes (VON SPERLING, 2014). The tertiary treatment seeks to remove specific pollutants or to complement the removal of pollutants that were not sufficiently removed in the secondary treatment (VON SPERLING, 2014), such as nutrients. Mainly physical and chemical mechanisms are involved in the tertiary treatment (CRINI; LICHTFOUSE, 2019).

In Brazil, 52.2% of the domestic effluent generated was treated in 2022, according to the National System of Information on Sanitation (SNIS, 2023). This was consistent with the evaluation made by the National Water and Sanitation Agency, which reported that 46.5% of the Brazilian urban population was served by wastewater collection and treatment services in 2019 (ANA, 2020).

The National Water and Sanitation Agency identified 3,668 WWTPs in Brazil in 2019 (ANA, 2020). Most of these systems use anaerobic reactors (representing 37% of the systems), primarily upflow anaerobic sludge blanket (UASB) reactors, that may or may not be followed by post-treatment. Stabilization ponds (35%) constitute the second most widely used technology in Brazil. Other less popular processes that are utilized include simplified processes (such as septic tank systems) (12%) and activated sludge (10%). WWTPs that adopt a combination of biological and chemical processes represent 2% of the systems. Filtration systems or other configurations represent 2% of the WWTPs. The remaining 2% are inactivated WWTPs or undefined treatment technologies (ANA, 2020). There is a great diversity of secondary level treatment technologies in the country; however, less than 5% of the Brazilian WWTPs identified in 2013 have been designed to remove nutrients (ANA, 2017).

Comparison with indices from developed countries, such as the United States, demonstrates Brazil's low level of sewage services coverage. According to the U.S. Census Bureau’s American Housing Survey (AHS), in 2021, from the 128,503,000 households in the United States, 84.5% were served by a public sewer, 15.2% by an individual decentralized system, such as septic tank or cesspool, 0.1% did not have access to wastewater treatment systems, and 0.2% had other system or not reported information (USCB, 2021).

According to the Joint Monitoring Programme for Water Supply, Sanitation and Hygiene (JMP) of the World Health Organization (WHO) and the United Nations Children's Fund (UNICEF), in 2022, 97.04% of the population in the USA had safely managed sanitation, 2.59% had basic service level, and 0.37% unimproved level (WHO; UNICEF, 2022a). “Safely managed” means the “*use of improved facilities that are not shared with other households and where excreta are safely disposed of in situ or transported and treated offsite*”. Sewage that undergoes at least secondary treatment or primary treatment with a long ocean outfall is considered “safely managed”. “Basic” service level means “*use of improved facilities that are not shared with other households*”. “Unimproved” level means the “*use of pit latrines without a slab or platform, hanging latrines or bucked latrines*” (WHO; UNICEF, 2022b).

The country with the largest number of WWTPs in the world is the United States (WANG et al., 2024b). The Clean Watersheds Needs Survey (CWNS) from the United States Environmental Protection Agency (USEPA, 2024) found 17,544 publicly owned WWTPs in 2022 that served 270.4 million people, or 82% of the population. Of these, 6,167 facilities had advanced wastewater treatment level (greater than secondary), serving 139.3 million people (42% of the population in 2022). The report also found that 2.543 of the facilities (serving 33.0 million people, or 10% of the population) did not discharge the final effluent into surface water but instead reused for beneficial purposes such as spray disinfection, and groundwater recharge (USEPA, 2024).

As can be seen from the data on the population served by wastewater treatment and the number and level of treatment plants in Brazil and the United States, there is a high discrepancy between the countries. The assessment of two countries in distinct conditions is potentially valuable for comparing the two scenarios and formulating recommendations for priority investments in the sanitation sector in Brazil. Besides that, the environmental scenarios faced by developing countries are of interest to a wide range of scientists. The low proportion of the population served by sanitation services, insufficient amount of wastewater treatment, and degradation level of surface water quality are problems that affect Brazil and several other developing countries.

Many Brazilian wastewater treatment systems face design, operation, and maintenance challenges, which can lead to the discharge of contaminated effluents

into receiving water bodies. Disposal of polluted effluents into these water bodies causes severe environmental and public health risks (MJALLI; AL-ASHEH; ALFADALA, 2007; SINGH et al., 2010; YU et al., 2024). The disposal of untreated or insufficiently treated domestic wastewater is the primary source of pressure on Brazilian water bodies (SNIS, 2021).

Domestic effluents, even after treatment, impose physical, chemical, and biological alterations to the receiving water bodies and may have significant impacts (DRURY; ROSI-MARSHALL; KELLY, 2013). Therefore, it is essential to understand and improve the performance of WWTPs. Because of the nature of the information, access to WWTP monitoring data is strict and controlled by companies operating Brazilian WWTPs. Datasets are not publicly available in Brazil and other developing countries, making them difficult to obtain for analysis; therefore, there is a lack of knowledge and experience in this area. Scientific information must instead be disseminated from studies that have performed statistical and in-depth evaluations of WWTP monitoring data.

Law No. 14,026/2020 has established a new regulatory framework for basic sanitation in Brazil that aims to attain universal access to basic sanitation by 2033. Another goal of the framework is to ensure an increase in the operating efficiency and quality of services provided. Therefore, various investments in both public and private sectors will take place in the coming years. Considering existing WWTPs, the monitoring and operation of these systems require improvement to allow for a greater generation of data. Large amounts of data from WWTPs must be properly transformed into relevant information to enhance operations (COROMINAS et al., 2018), allowing for more robust tools for data analysis by sanitation service providers.

Fluctuations in effluent quality can happen due to changes in influent quality and quantity, weather conditions, and operational changes (ROOHI; NAZIF; RAMAZI, 2024). Proper analysis of data obtained from monitoring WWTPs can enhance the reliability and performance of the treatment processes (ELMAADAWY et al., 2021). As WWTPs are highly complicated and dynamic systems, adequate operation, control, and simulation are crucial for ensuring public and environmental health (HEJABI et al., 2021; NOURANI; ELKIRAN; ABBA, 2018).

Developing a modeling tool to predict the performance of WWTPs based on previous observations of key quality parameters is necessary for effective control of the process (HEJABI et al., 2021; MATHUR et al., 2024). Reliable and convenient modeling tools play an essential role in simulating and monitoring the performance of WWTPs and describing the overall phenomena occurring in the entire system (NOURANI; ELKIRAN; ABBA, 2018).

However, WWTPs consist of complex physical, biological, and chemical processes, which usually exhibit nonlinear behavior (LEE et al., 2011; NOURANI; ASGHARI; SHARGHI, 2021), making it challenging to develop accurate models for WWTPs (PHAM et al., 2020). The complex behavior of WWTPs cannot be easily predicted or explained by classical linear models (NOURANI; ELKIRAN; ABBA, 2018).

In this context, artificial intelligence (AI) is a simulation of the human brain by machines (MALVIYA; JASPAL, 2021). AI techniques offer superior performance in simulating nonlinear and complex engineering problems (SAFEER et al., 2022). In recent decades, AI approaches have been used as effective tools to investigate environmental engineering issues (HEJABI et al., 2021), such as the wastewater treatment processes. Despite rapid advancements, AI applied to wastewater treatment is still in its primary stage (LI et al., 2024). Data-driven models can simulate a wide range of situations based on data and algorithms (WANG et al., 2024a).

Machine learning is a process of extracting knowledge from data (KHATRI; KHATRI; SHARMA, 2020) and is a central subfield of AI (MATHUR et al., 2024) that provides a set of tools that use computers to transform data into actionable knowledge (LANTZ, 2013). Among the existing machine learning algorithms, artificial neural networks (ANNs) have been adopted to solve many practical engineering problems in WWTPs (QIAO; HU; LI, 2016). Bahramian et al. (2023) conducted a comprehensive literature review on the state-of-the-art in the application of data-driven models in WWTPs. They searched publications from 2000 to 2021 and selected 281 studies for qualitative assessment. The ANNs were identified as the most popular model among the studies and were commonly used as a prediction model focusing on the removal of pollutants (BAHRAMIAN et al., 2023).

The concept of ANNs is based on the human brain (MATHUR et al., 2024). This model applies as a mathematical structure to demonstrate the nonlinear relationships between the inputs and outputs. ANNs comprise an input layer, one or more hidden layers, and an output layer (RAHMATI; TISHEHZAN; MOAZED, 2021). When multiple hidden layers in a network are used, the ANN technique is called deep learning (COROMINAS et al., 2018). Analogous to the human brain, ANNs are composed of neurons and their connections. The neurons are processing units present on each layer and are connected through weighted connections (NEZHAD et al., 2016; RAHMATI; TISHEHZAN; MOAZED, 2021).

The determination of the ANN structure involves configuring hyperparameters prior to training the model, such as the number of hidden layers and neurons in each hidden layer. Determining an appropriate network structure is an important task involved in ANN modeling since the network structure significantly affects the predicted result and neural network performance (GAYA et al., 2014; LEE et al., 2011). Usually, the structure is selected through a trial-and-error process (GAYA et al., 2014), where several network structures are developed, and their performances are compared (ALSULAILI; REFAIE, 2021). However, due to the requirement for design expertise in determining the optimum model structure (PHAM et al., 2020), gaining knowledge from previous studies in the literature on how ANN models were configured for similar prediction problems may help in the search for an optimized model configuration. Although there are literature review studies on the use of AI and machine learning techniques in the wastewater treatment sector (BAHRAMIAN et al., 2023; COROMINAS et al., 2018; HADJIMICHAEL; COMAS; COROMINAS, 2016; MALVIYA; JASPAL, 2021; ZHAO et al., 2020), no literature review specifically focused on the ANNs was found. Conducting a literature review with a more focused approach on ANNs, the most widely employed model in the wastewater treatment sector, allows for addressing this gap. This review is essential to identify the key characteristics of model configurations developed in previous studies and, thus, assist in decision-making for studies with similar prediction objectives.

Another gap identified in the literature was the lack of application of ANN models to WWTP treating domestic wastewater in Brazil. Despite the development of studies that use neural networks to model WWTPs data, most have occurred in relation to

developed countries. A thorough literature search was conducted using the systematic review method, and no studies have been found that assessed a WWTP in Latin American countries. Advanced data science techniques may help improve the performance of existing WWTPs in Brazil and promote advances in sanitation in that country.

1.2 HYPOTHESES

The following hypotheses are proposed and investigated in this study: (i) There are typical ANN structures and hyperparameters configurations used in studies that apply ANN models to predict WWTP performance; (ii) A quantitative relationship exists between influent qualitative and quantitative characteristics, operational factors, and variables in treated effluent, which can be utilized to develop an ANN model for predicting the performance of a WWTP; (iii) The main contributors to WWTP performance can be identified through a sensitivity analysis of the developed models; and (iv) ANN developed for predicting the performance of a WWTP with larger datasets, with more observations to train and test the models, exhibit improved performance.

1.3 OBJECTIVES

1.3.1 General objective

Assess and model the performance of full-scale wastewater treatment plants, especially regarding effluent nutrient concentrations.

1.3.2 Specific objectives

The specific objectives are:

- Conduct a systematic literature review on the use of artificial neural networks for the prediction of the performance of full-scale WWTPs;
- Explore and analyze the datasets under study using statistical analyses and data visualization techniques, with the goal of gaining a deeper understanding of the treatment processes;
- Develop neural network models for the prediction of effluent quality of WWTPs, considering raw sewage and operational variables as input variables;

- Determine the main operational factors and conditions responsible for the performance of WWTPs; and
- Compare the performance of the neural network models developed to predict the effluent quality of two distinct WWTPs.

1.4 DOCUMENT STRUCTURE

This document is divided into six chapters. The first, “Chapter 1: Introduction,” presents the background, justification, hypotheses, and objectives of the study. The second chapter, “Chapter 2: Systematic literature review,” relates to specific objective 1. Chapter 2 aims to present a systematic review of the use of artificial neural networks in predicting the effluent quality and removal efficiencies of full-scale wastewater treatment plants. This review was conducted to gain a deep understanding of the use of these advanced models in previous studies in the field to improve the efficiency of determining the optimum setting and the performance of future models. The systematic review has been published in the journal Water Science & Technology (Appendix A).

The third chapter, titled “Chapter 3: Case studies, data description, and exploratory analyses,” introduces the selected WWTPs located in Brazil and the United States as the case studies for the application of the data analysis techniques. The chapter describes the datasets provided by the sanitation companies and the preliminary analyses conducted to understand the datasets behavior. Chapter 3 corresponds to specific objective 2.

In the fourth chapter, “Chapter 4: Use of artificial neural network models for the prediction of the performance of wastewater treatment plants: the case studies of Brasília Sul and John E. Egan WWTPs,” the modeling development process is presented, utilizing the preprocessed data from Chapter 3. Chapter 4 corresponds to specific objectives 3, 4, and 5. Chapters 3 and 4 will result in other manuscripts that, at the time of this thesis’s publication, are being prepared for submission to scientific journals.

The fifth chapter, “Chapter 5: Final considerations”, provides the conclusions of this study and suggestions for future work. Finally, the sixth chapter presents the references used in this work.

CHAPTER 2: SYSTEMATIC LITERATURE REVIEW

2.1 INTRODUCTION

Recent concerns regarding environmental issues have induced specialists to focus their attention on the efficient operation and control of wastewater treatment plants (WWTPs) (MJALLI; AL-ASHEH; ALFADALA, 2007; PHAM et al., 2020). WWTPs are highly complex and dynamic systems that require consistent high performance despite hourly, daily, and seasonal fluctuations (COROMINAS et al., 2018).

The treatment of wastewater is affected by several chemical, physical, and microbiological factors. The complexity of wastewater treatment technology results in uncertainty and variation in the treatment system, leading to fluctuations in effluent quality and environmental risks to the receiving water (ZHANG et al., 2023; ZHAO et al., 2020). Hence, proper operation and control are essential for safeguarding public health and protecting the environment (NOURANI; ELKIRAN; ABBA, 2018).

Safe operation and control of WWTPs can be achieved through the development of a robust and appropriate mathematical model for predicting plant performance based on past observations of key quality parameters (HAMED; KHALAFALLAH; HASSANIEN, 2004; NASR et al., 2012; SINGH et al., 2010). Modeling is widely used to assess the performance of WWTPs (HAMED; KHALAFALLAH; HASSANIEN, 2004; MJALLI; AL-ASHEH; ALFADALA, 2007; SINGH et al., 2010); however, the complexity and dynamics of treatment systems make it difficult to perform predictions and simulations using traditional linear methods (NOURANI; ELKIRAN; ABBA, 2018).

Artificial intelligence (AI) has become a powerful tool for minimizing the complexities in wastewater treatment (MALVIYA; JASPAL, 2021; ZHANG et al., 2023; ZHAO et al., 2020). Zhao et al. (2020) conducted a bibliometric analysis of the trends in artificial intelligence technology as applied to wastewater treatment. Those authors found that the number of published articles utilizing AI in wastewater treatment research was 19 times greater in 2019 than that in 1995. Most AI techniques have been modeled using experimental data to simulate, predict, confirm, and optimize contaminant removal in wastewater treatment processes (ZHAO et al., 2020).

Machine learning is a central subfield of artificial intelligence. Machine learning algorithms are increasingly used and play a fundamental role in the operation of

WWTPs (DE CANETE et al., 2021). Machine learning approaches have become powerful tools for dealing with the complexities of uncertain and dynamic problems. Therefore, these techniques are becoming common for modeling complex environmental problems, such as that of wastewater treatment and optimization of wastewater (GUO et al., 2015; YE et al., 2020; ZHAO et al., 2020). These approaches maximize the knowledge obtained from data and operational experience and help strengthen the management and control of WWTPs, thereby improving the performance of these facilities (ZHAO et al., 2020).

Machine learning methods can be supervised or unsupervised. Supervised methods are used to build predictive models that characterize the link between explanatory and response variables. These models predict the response variable of interest (output) using the explanatory variables (inputs) of the dataset (COROMINAS et al., 2018; LANTZ, 2013; NEWHART et al., 2019; NEWHART; HERING; CATH, 2022). Supervised machine learning includes models such as naïve Bayes, regression trees, artificial neural networks (ANNs), and support vector machines (LANTZ, 2013). Unsupervised methods are used to build descriptive models. They are applied when the goal is to identify patterns in the data without any advanced knowledge of the possible relationships involved (NEWHART et al., 2019).

Previous literature reviews have identified ANNs as the most employed in the wastewater treatment sector. Hadjimichael, Comas and Corominas (2016) conducted a literature review on the application of artificial intelligence methods (mainly machine learning) to the urban water sector. Those authors found 1,394 papers on wastewater published between 1935 and 2016, and ANNs were found to be the most common method used in various sectors of water-related research, including that of wastewater treatment. ANNs have emerged as an attractive option for predicting and classifying water systems as well as for modeling and optimizing performance (HADJIMICHAEL; COMAS; COROMINAS, 2016).

Corominas et al. (2018) performed a literature review of computer-based techniques for data analysis to improve the operation of WWTPs. Those authors described various methods that enable the transformation of data into pertinent information. According to Corominas et al. (2018), the European Union is the leading region in this field with the

largest number of studies (61%), followed by Asia-Oceania (34%), and North America (12%). A minority of studies (less than 4%) have been conducted by South American or African research groups. Among the 340 selected papers (published up to 2015), ANN was the most commonly used technique, particularly for predicting process performance, soft sensing, and control (COROMINAS et al., 2018).

Zhao et al. (2020) conducted a bibliometric analysis covering 1995 to 2019 of trends in applying artificial intelligence technology to wastewater treatment. According to those authors, research has mainly focused on AI technology in relation to pollutant removal. The majority of studies utilized ANN models to simulate and predict the performance of biological WWTP, and there has been an increase in the number of publications using this technique in recent years (ZHAO et al., 2020).

Soft measurement estimates variables that are difficult to measure by correlating them with available variables that are more readily measured (OSMAN; LI, 2020). Ching, So and Morck (2021) conducted a literature review covering 102 studies on the development of soft sensors for wastewater treatment. Those authors showed that neural networks were the most common modeling approach. These methods have remained the dominant methodologies for soft sensor development since the early 2000s, and it appears that ANNs will continue to predominate in the coming years (CHING; SO; MORCK, 2021).

Bahramian et al. (2023) conducted a comprehensive literature review on the state-of-the-art in the application of data-driven models in WWTPs. They searched publications from 2000 to 2021 and selected 281 studies for qualitative assessment. The ANNs were identified as the most popular model among the studies and were commonly used as a prediction model focusing on the removal of pollutants (BAHRAMIAN et al., 2023).

Zhang et al. (2023) provided a summary of the status and trends in AI research as applied to wastewater treatment, based on published papers and patents from 2000 to 2022. According to the authors, ANN is the most common and widely used model for AI in wastewater treatment (ZHANG et al., 2023).

The parameters of wastewater treatment monitoring data tend to share nonlinear and complex chemical relationships (CHING; SO; MORCK, 2021). The nonlinear nature of

an ANN can accurately predict pollutant removal in WWTPs (YE et al., 2020). The wide usage of ANNs in water-related research relates to their ability to learn (through training process) complex nonlinear and multi-input/output relationships between process parameters using historical data (MADIĆ; RADOVANOVIĆ, 2011). ANNs can also be applied when there is insufficient knowledge of the process to construct a mechanistic model of the wastewater treatment system, which relies on fundamental material and energy balances and empirical correlations that are often inaccurate (MJALLI; AL-ASHEH; ALFADALA, 2007). Many simplifications and assumptions are required to ensure mechanistic models are tractable and computable, and accordingly they have many limitations (WANG et al., 2021).

ANN models consist of predefined mathematical functions that effectively capture the nonlinear relationships between variables in complex systems (CIVELEKOGLU et al., 2009). ANNs require historical data during training, after which they should have the ability to extrapolate correlations to new data (PALANI; LIONG; TKALICH, 2008). The ANN learns from the training data and captures the relationships between data points, which can be used for simulation, prediction, and optimization (ZHAO et al., 2020).

The concept of an ANN was based on the biological human brain and its learning processes. ANNs are numerical structures comprising nodes (neurons) and connections (weights) (MJALLI; AL-ASHEH; ALFADALA, 2007; NEZHAD et al., 2016). The artificial neural network architecture is the overall structure and manner in which information flows from one layer to another (CHEN et al., 2020). The architecture consists mainly of the number of neurons and the manner in which they are interconnected (MJALLI; AL-ASHEH; ALFADALA, 2007). An ANN includes a variety of hyperparameters that must be tuned during model development, including the number of hidden layers, number of neurons in each hidden layer, and activation functions that are applied (CHING; SO; MORCK, 2021).

The main task in designing a robust neural network is to determine the appropriate model architecture to minimize overall model error (MADIĆ; RADOVANOVIĆ, 2011; NEZHAD et al., 2016). Selecting a network structure (e.g., a feedforward neural network with one hidden layer and five neurons in the hidden layer that are connected by a sigmoid activation function, or a deep neural network with multiple hidden layers

and multiple parameters) is a crucial step in the design of ANNs. The structure must be optimized for reducing computer processing, achieving adequate performance, and avoiding overfitting (MJALLI; AL-ASHEH; ALFADALA, 2007).

There is a limited theoretical and practical background to assist in the systematic selection of ANN hyperparameters through model development and training processes (MADIĆ; RADOVANOVIĆ, 2011). Therefore, most studies choose the appropriate ANN model structure using a trial-and-error approach (CHEN et al., 2020; MADIĆ; RADOVANOVIĆ, 2011; MJALLI; AL-ASHEH; ALFADALA, 2007; PALANI; LIONG; TKALICH, 2008), whereby several networks are trained and compared (MADIĆ; RADOVANOVIĆ, 2011; MJALLI; AL-ASHEH; ALFADALA, 2007), which is challenging and time-consuming (LEE et al., 2011; ZAGHLOUL; ACHARI, 2022). Choosing the ANN architecture and selecting the training algorithm (which is used to minimize the error between the observed and predicted output) and related parameters is primarily related to the experience of the designer (MADIĆ; RADOVANOVIĆ, 2011).

Previous literature reviews investigated the use of artificial intelligence in the field of wastewater treatment, and they have identified the neural network as the data-driven technique and machine learning model most applied in the wastewater sector (BAHRAMIAN et al., 2023; COROMINAS et al., 2018; HADJIMICHAEL; COMAS; COROMINAS, 2016; MALVIYA; JASPAL, 2021; ZHAO et al., 2020). However, these studies did not focus exclusively on neural networks and no specific literature review has been found on the use of ANN in the wastewater treatment sector. Previous studies were broad in scope, providing a general overview of the use of AI models, such as which ones are most employed, the types of applications they are used for, the countries where they are applied, and trends in publications, including journals and areas of publication. Nevertheless, since these studies did not delve deeply into neural networks, a specific study on them is necessary. This is because neural networks are complex and involve a wide range of hyperparameters that must be carefully configured. Therefore, the current investigation may improve the configuration of models based on studies in this field. Understanding the hyperparameter tuning process from datasets of WWTPs might improve the efficiency of determining the optimum setting and the performance of future models.

2.2 METHODS

2.2.1 Review objective and research question

With the increased use of neural network methods for predictions, it is important to study their role in predicting WWTP performance. The various ANN structures and hyperparameters used in the wastewater treatment sector have not been adequately studied. Therefore, a systematic review was conducted to develop an understanding of WWTP performance predictions using an artificial neural network.

A systematic review is a literature review based on clearly formulated questions. It identifies relevant studies and summarizes evidence using an explicit methodology (KHAN et al., 2003). A systematic review differs from a traditional general review, as it adopts a replicable, scientific, and transparent process (QAZI et al., 2015). The current study followed the guidelines and protocols for systematic reviews (KHAN et al., 2003; PULLIN; STEWART, 2006).

The first step in a systematic review is to formulate a specific question (KHAN et al., 2003; PULLIN; STEWART, 2006). The following research question was the basis of this review: “What are the main architectures and hyperparameters of ANN models used to predict the performance of different types of full-scale WWTPs?”

2.2.2 Search strategy

The next step is to identify relevant studies (KHAN et al., 2003) by formulating a formal search strategy. The systematic review design reported here was initiated in August 2021. After several refinements and improvements, the publication search began in February 2022.

The ScienceDirect, Scopus, and Web of Science databases were searched, and the results restricted to peer-reviewed articles published in journals from 2011 through 2021 in English.

Pilot searches were performed to refine the keywords, and the following final search strategy was used, based on document titles, abstracts, and author-specified keywords:

(“wastewater treatment plant” OR “sewage treatment plant” OR WWTP) AND (“neural network” OR ANN)

2.2.3 Selection criteria

The study selection criteria flow directly from the review questions and should be previously specified. The reasons for inclusion and exclusion were recorded (KHAN et al., 2003). The eligibility criteria were designed to focus exclusively on the use of ANNs for predicting the performance of WWTPs in terms of effluent quality or removal efficiencies. The goal was to gather a comprehensive set of studies specifically focused on the application of ANNs in this context. The selection process was structured as follows:

Inclusion Criteria:

a. Studies using ANNs: Only studies that employed ANNs as the modeling tool for predicting the effluent quality or removal efficiencies of WWTPs were considered for inclusion. Other machine learning algorithms and modeling techniques were excluded to maintain a specific focus on ANNs.

b. Full-scale WWTPs: Only studies involving full-scale WWTPs were included in the review. Pilot- and bench-scale plants were excluded to ensure relevance to real operational conditions.

c. Domestic effluent treatment: The review was limited to studies that focused on WWTPs specifically designed to treat domestic effluent. Industrial plants were excluded.

Exclusion Criteria:

a. Studies using ANNs for other purposes: Studies utilizing ANNs for purposes other than predicting WWTP performance in terms of effluent quality or removal efficiencies (e.g., energy consumption control, process optimization) were excluded, as they deviated from the primary research focus.

b. Non-journal publications: Publications such as book chapters, conference papers, and lecture notes were excluded from the review as journal publications were the focus.

Language and data criteria:

a. Language: Articles published in languages other than English were excluded.

b. Data availability: During the full text screening, an additional criterion was applied to assess whether the selected papers contained the necessary data and information to effectively answer the main research question (ESPINOSA et al., 2020). Studies lacking relevant data were excluded.

After selecting documents based on the search strategy, duplicates were removed using Mendeley software. Then, non-journal publications were excluded. Subsequently, articles were screened for exclusion criteria based on their titles. The abstracts were then evaluated for the inclusion and exclusion criteria, and the remaining articles were subsequently screened based on their full text for the eligibility criteria.

2.2.4 Data extraction and analysis

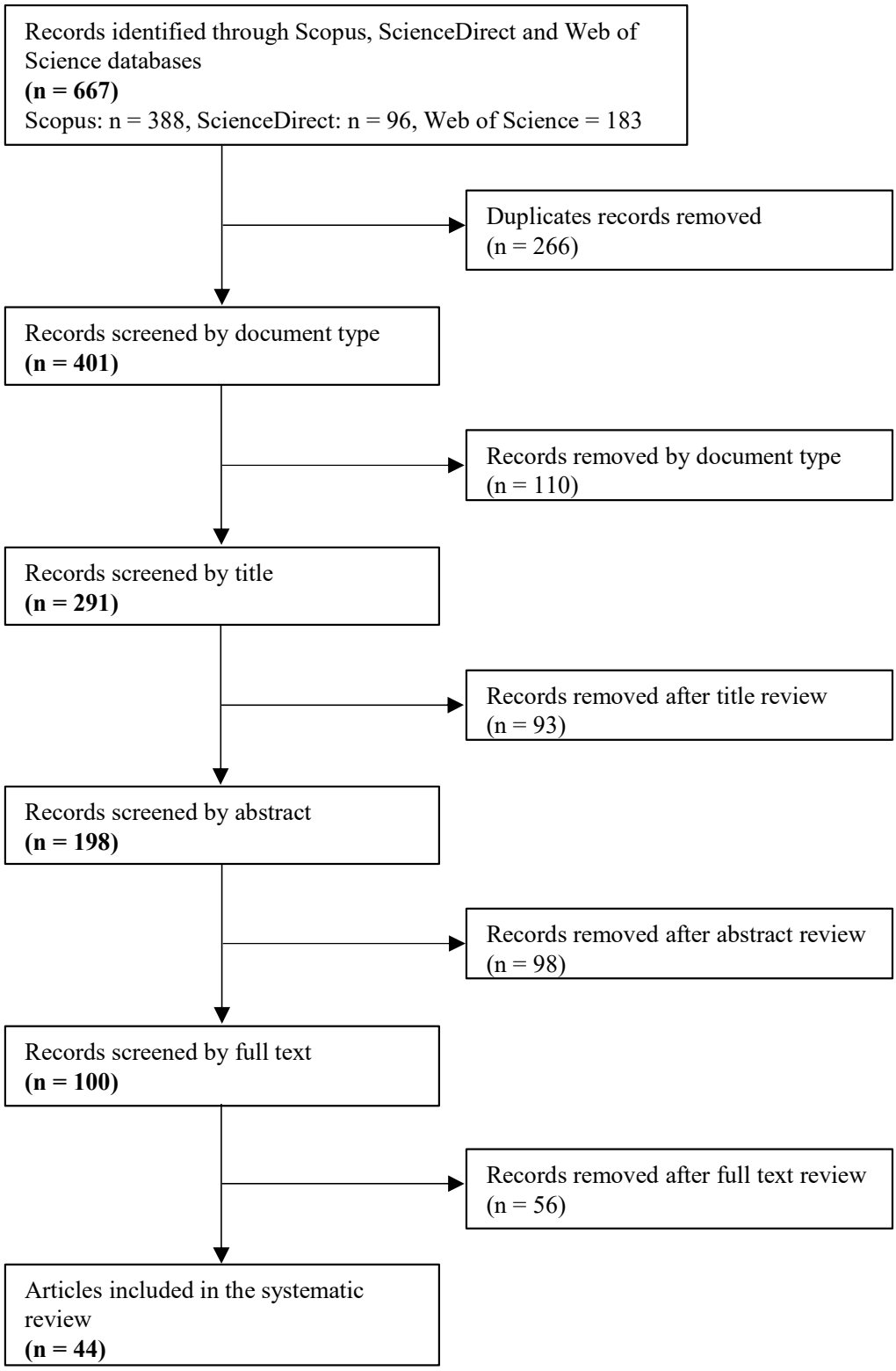
The next step was to extract data from the final selected papers by identifying relevant information related to the research question (QAZI et al., 2015). A detailed investigation was conducted, and data from papers were extracted and presented in a table with the following fields: (i) reference (author(s), year, journal, and paper title); (ii) country of the study; (iii) wastewater treatment technology and inflow rate/design flow of the facility; (iv) monitoring frequency and period; (v) number of samples; (vi) data division into training/validation/testing datasets (%); (vii) input and output variables; (viii) data preprocessing methods; (ix) neural network architectures and hyperparameters (ANN methods, training algorithms, number of hidden layers, number of neurons in each hidden layer, and activation functions); and (x) metrics of model performance.

2.3 RESULTS AND DISCUSSION

2.3.1 Search results

A total of 667 articles were identified by searching the three databases. Duplicates (266 records) and publications from books, book chapters, conference articles, and lecture notes (110 documents) were removed. In the next step, 291 records were screened based on their titles, and 93 were excluded. The main reasons for exclusion at this stage were that the studies focused on the gas and solid phases of WWTPs; other operating conditions of the systems, such as energy consumption, treatment cost, odor, or membrane fouling; or industrial effluents. Following this, 198 papers were screened based on their abstracts, and 98 were excluded. The main exclusions occurred in relation to studies not conducted on full-scale WWTPs (pilot plants, bench-scale, or benchmark simulation models); models of influent conditions (quality and quantity) or other operating conditions (such as aeration control); studies on industrial effluents; or articles that did not use ANNs. Subsequently, 100 papers were screened based on the full text, and 56 were excluded, mainly for not including the appropriate data to answer the research question or not assessing full-scale WWTPs (pilot plants, laboratory scale, or benchmark simulation models). The remaining 44 studies were included in this review (Figure 1).

Figure 1 – Flow diagram of the systematic review on the use of ANNs to predict the performance of WWTPs



There was no observed increase in the number of publications selected among the years. However, 13 papers (30%) were published in 2021 (Figure 2), which may be related to the COVID-19 pandemic. Due to the lockdown, researchers in many fields

were able to commit more time to writing and submitting papers to peer-reviewed journals. In addition, researchers were hindered from conducting laboratory research, and studies were more focused on statistics and mathematical modeling using secondary data.

Figure 2 – Number of publications over the years (n = 44)

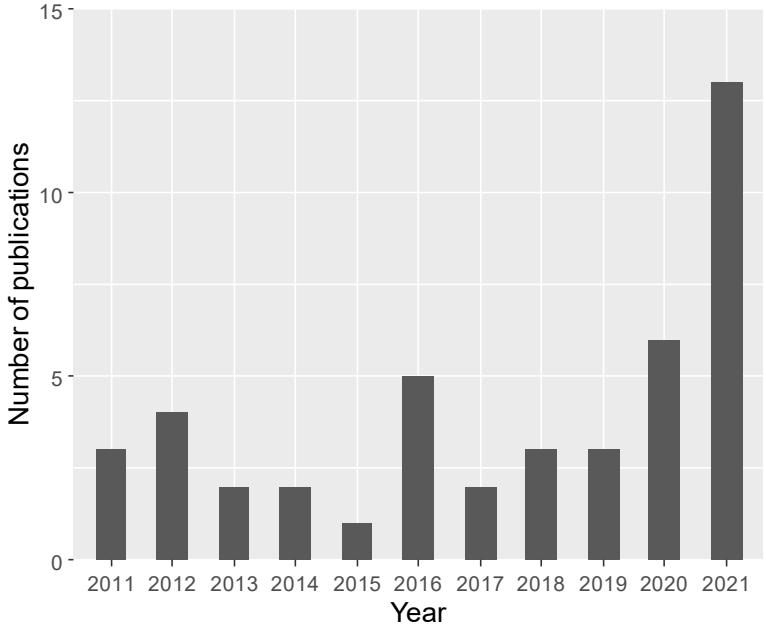
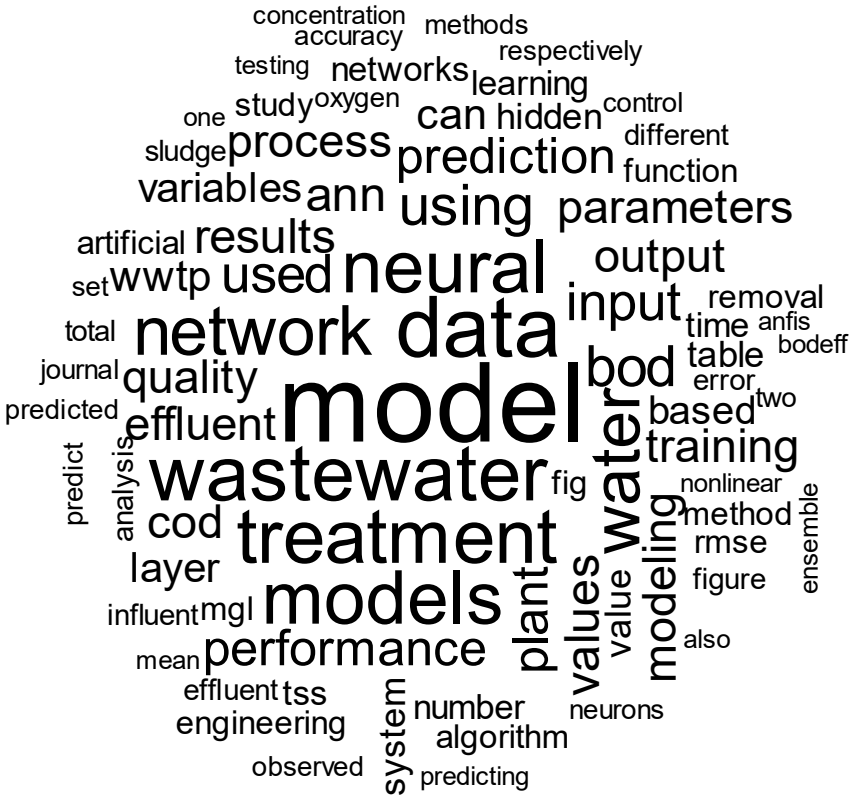


Figure 3 shows the word cloud generated from the 44 selected papers using the package “wordcloud” (FELLOWS, 2018) of the R programming language. The size of a word is proportional to its frequency in the texts. Some terms (artificial, neural, network, ANN, wastewater, treatment, plant, and WWTP) were expected to be in the word cloud, as they were used in the search strategy of the systematic review. The word “model” was highlighted, which was not used in the search strategy but was the most frequently used term in the papers. Other words related to the modeling process also appeared such as algorithm, modeling, data, predict, predicted, predicting, prediction, training, testing, learning, method, function, hidden, layer, nonlinear, ensemble, accuracy, RMSE, error, neurons, output, input, and ANFIS. Another category included terms related to the wastewater treatment process such as engineering, effluent, influent, quality, system, water, removal, concentration, control, process, oxygen, sludge, BOD, COD, TSS, and BODeff.

Figure 3 – Word cloud generated from the 44 selected papers of the systematic review



A complete and detailed table was elaborated with detailed information extracted from the 44 papers. The spreadsheet is shown in the Supplemental Material of the published paper (Appendix A). The following sections discuss the main results extracted from the table.

2.3.2 WWTPs characteristics

One study (GE et al., 2020) assessed two WWTPs, while the remaining 43 studies evaluated only one treatment facility. Three papers (7%) did not provide information on the treatment technology adopted in the WWTP under investigation. The conventional activated sludge process was the most common and was found in 18 papers (41%), followed by anaerobic/anoxic/oxic processes in 5 articles (11%). The remaining 41% of the studies included WWTPs that employed different activated sludge configurations (anoxic/oxic processes, activated sludge with coagulation/flocculation, extended aeration activated sludge, aerated lagoon followed by activated sludge, step-feed activated sludge processes, sequential batch reactors, intermittent cycle extended aeration-sequential batch reactors), two-stage trickling filters, constructed wetlands, and membrane bioreactors. The activated sludge

processes had higher prevalence in the selected studies because this is the most employed wastewater treatment technology globally (SIN; AL, 2021).

Sixteen (36%) of the selected papers did not mention the size of the WWTP under study. The remaining 28 papers (64%) reported the inflow rate, design flow of the WWTPs, or both. The sizes of the WWTPs were variable, ranging from 52.1 to 11,574 L/s. However, most studies assessed large WWTPs. Fourteen WWTPs had inflow rates or design flows above 1,000 L/s. The inclusion of large facilities in the studies may be because large systems have better monitoring schemes with more data to train the ANN models. Larger WWTPs also have improved operational control, which encourages the development of models for predicting system performance.

Figure 4 shows the locations of the WWTPs studied. The country of the authors was considered in four papers that did not mention the WWTP site. This was an acceptable criterion, as the WWTPs were located in the same countries as the authors in papers that presented that information. The 44 selected publications for the systematic review originated in 15 countries, with the largest contribution from China (20% of the papers). The publications were concentrated in northern countries. Further research should be conducted in countries from other regions with other socioeconomic and climatic characteristics that lead to different wastewater treatment operational conditions. These distinct conditions may aid in providing important information on the use of ANNs for predicting WWTP performance.

Figure 4 – Distribution of the 44 publications included in the systematic review according to the country where the studies were conducted

2.3.3 Datasets characteristics

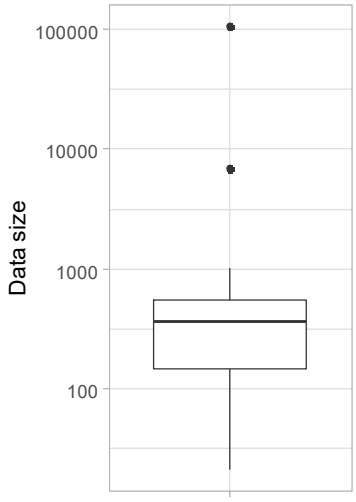
The WWTP data are collected at various time intervals, from continuous online sensor measurements to quarterly laboratory results (NEWHART et al., 2019). In the WWTPs under study, 11 (25%) of the publications did not include the monitoring frequency, while three presented more than one frequency, from daily to monthly.

Four papers (9%) had samples collected at a frequent temporal resolution, such as every 10 min, every hour, or three times a day. The most common data collection period was daily (20 papers, 45%). Other studies collected samples every two or three days, or three days a week (three papers, 7%); weekly (four papers, 9%); monthly (five papers, 11%); and biweekly, once or twice a month, or every two weeks (two papers, 5%).

Six (14%) studies did not provide the period in which the data was collected. The remaining 38 studies had distinct time frames, from three months (GE et al., 2020) to more than 15 years (HEJABI et al., 2021). Most studies (50%) assessed one to two years of a dataset.

There are no strict standards for the amount of experimental data required to train a prediction model for reliable results (YE et al., 2020). Data size information was not detailed in 15 papers (34%). The remaining 29 reported the number of samples (also called data points, instances, and records), which varied from 21 (HAZALI; WAHAB; IBRAHIM, 2017) to 105,763 (WANG et al., 2021) illustrating that ANN models are capable of dealing with different sized datasets (CHEN et al., 2020). However, most studies presented relatively small samples, and the median considering all papers that reported this information was 361.5 (Figure 5).

Figure 5 – Data size considering all 29 papers that included the number of samples



2.3.4 Data preprocessing methods

An important preprocessing method is to normalize the data, and most reviewed papers used this step. Twelve studies did not mention whether normalization was performed on the data, while six conducted this step but did not identify the specific method used. This information should be clearly defined because different methods affect the final result of the model differently (CHEN et al., 2020). Among the papers that provided details about data normalization, the most used method (16 papers) was min-max normalization in the range [0, 1]. Distinct range values were also used, but less frequently, namely, [-1, 1] (two papers), [-0.9, 0.9] (one paper), [0.1, 0.9] (one paper), and [0.05, 1] (two papers). The min-max technique normalizes the data using Equation (1).

$$y = \left(\frac{x - x_{min}}{x_{max} - x_{min}}\right)(new\_max_x - new\_min_x) + new\_min_x \quad (1)$$

where *y* is the normalized data, *x* is the measured data, x_{min} and x_{max} are the minimum and maximum values of the measured data, respectively, and [new_min_x, new_max_x] is the range to which the data are normalized (GE et al., 2020).

Another normalization method is Z-score normalization, in which the variables are standardized to have a zero mean and unit variance (JAMI; MUJELI; KABBASHI, 2011). This approach was used in four papers, and Equation (2) shows the Z-score transformation. The results were in accordance with those of Chen et al. (2020), who mentioned that range scaling and standardization are two common categories in data normalization.

$$y = \frac{x - \bar{x}}{\sigma} \quad (2)$$

where *y* is the normalized data; *x* is the measured data; and $\bar{x}$ and σ are the mean and the standard deviation of the variable, respectively (JAMI; MUJELI; KABBASHI, 2011).

Other methods of preprocessing used were the removal of outliers, abnormal data, noise, or errors in the data (ALSULAILI; REFAIE, 2021; HAN; WANG; QIAO, 2014; JAMI et al., 2012; JAMI; MUJELI; KABBASHI, 2011; KUSIAK; WEI, 2013; QIAO; HU; LI, 2016; YAQUB et al., 2020; ZHAO; CHAI; YUAN, 2012); the estimation, interpolation, or imputation of missing points (ALDAGHI; JAVANMARD, 2021; HAN; WANG; QIAO, 2014; LIU et al., 2021; ZHAO; CHAI; YUAN, 2012); and the use of multivariate statistical analyses, such as clustering methods and principal component analyses (ABBA; ELKIRAN; NOURANI, 2021; HAN et al., 2018; QIAO; HU; LI, 2016; SHARGHI et al., 2019; YASMIN et al., 2017; ZHAO et al., 2016), mainly for the selection of input variables of the models.

2.3.5 Modeling development

2.3.5.1 Data dividing

Data division is an important step in modeling (CHEN et al., 2020). Most studies (30 papers, 68%) divided the dataset into training and testing subsets. The training dataset is used to develop the model, that is, to accomplish network learning and fit the network weights. The testing dataset is used to evaluate how well the model generalizes to unseen data, that is, how accurately the network predicts targets for inputs that are not in the training set (LANTZ, 2013; MJALLI; AL-ASHEH; ALFADALA, 2007; ZHAO et al., 2020). Of these 30 papers, 26 mentioned the proportion of data division. The most common allocation (used in eight studies) was 75% for training and 25% for testing.

A different approach was adopted in 12 (27%) articles that divided the dataset into training, validation, and testing subsets, and the validation dataset was used to optimize the model (ZHAO et al., 2020) by adjusting the hyperparameters (CHEN et al., 2020). Most of these (seven studies) divided the dataset into 70% for training, 15% for validation, and 15% for testing.

Different approaches have accomplished data division. Nine papers divided data in chronological order, in which the first data points were used for training, and the remainder for validation and testing. Another nine papers randomly divided the dataset.

For larger samples, it was expected that a greater percentage would be destined to train the model. However, there was no significant correlation between the number of samples and the percentage used for training ($p = 0.27$ and Pearson correlation coefficient = 0.22). This confirms that there are no uniform rules for dividing the dataset, and most researchers divided the data either by domain knowledge or arbitrarily (CHEN et al., 2020).

2.3.5.2 Input and output parameters

Forty papers (91%) used effluent quality indicators as the target parameters, and four (9%) had removal efficiencies as the targets. The majority (28 papers, 64%) of the studies had more than one output parameter in single-output models (20 papers), multi-output models (seven papers), or both (one paper).

Table 1 shows that biochemical oxygen demand (BOD) and chemical oxygen demand (COD) effluent concentrations were the outputs in most papers. Other target parameters commonly used in the models were effluent concentrations of solids (total suspended solids, TSS) and effluent concentrations of nutrients (ammonia nitrogen, NH_4-N, total nitrogen, TN, and total phosphorus, TP). The three most used output variables appeared in the word cloud generated from the 44 selected papers of the systematic review (Figure 3). Among these three, the largest term in the word cloud was BOD, followed by COD and TSS, which is accordingly with Table 1. According to Alsulaili and Refaie (2021), most studies have utilized BOD, COD, and TSS to predict the performance of WWTPs using ANN-based models.

Key variables in wastewater treatment must be evaluated to control pollution (OSMAN; LI, 2020), and their use as targets in the models confirms that they are important for assessing the performance of a WWTP. BOD and COD reflect organic water pollution and are considered the most important parameters for effluent quality control (NOURANI; ASGHARI; SHARGHI, 2021). BOD is difficult to measure online, and laboratory measurements are time-consuming, as they are calculated by a 5-day off-line delay (OSMAN; LI, 2020; RAHMATI; TISHEHZAN; MOAZED, 2021), which

reinforces the importance of the development of predictive models for this parameter. TSS is another important variable, as excess TSS depletes dissolved oxygen in effluent water (VERMA; WEI; KUSIAK, 2013). There has been a continuous increase in the number of studies concerning nutrient removal (CHING; SO; MORCK, 2021) due to the control of effluents to prevent eutrophication of water bodies. According to Ching, So and Morck (2021), the various parameters involved in the nitrogen removal process are consistent areas of interest in soft sensor development. In comparison, there are fewer sensor studies on phosphorus removal processes. The significance of phosphorus as a wastewater parameter depends on the local abundance or shortage of this nutrient (CHING; SO; MORCK, 2021).

Table 1 – Number of publications for each output variable of the ANN models

Target variable	Number of publications
Effluent BOD	25
Effluent COD	21
Effluent TSS	19
Effluent NH_4-N	10
Effluent TN	7
Effluent TP	7
Effluent pH	4
Effluent quality index	2
Removal efficiency of NH_4-N	2
Effluent CBOD	1
Effluent biodegradable dissolved organic nitrogen	1
Effluent total coliform	1
Effluent fecal streptococci	1
Effluent TKN	1
Effluent PO_4	1
Effluent NO_2	1
Effluent NO_3	1
Effluent T	1
Effluent EC	1
Removal efficiency of fecal coliform	1
Removal efficiency of total coliform	1
Removal efficiency of arsenic	1
Removal efficiency of TN	1
Removal efficiency of TP	1
Removal efficiency of TSS	1
Removal efficiency of COD	1
Removal efficiency of BOD	1
Removal efficiency of sulfide	1

BOD: biochemical oxygen demand; COD: chemical oxygen demand; TSS: total suspended solids; NH_4-N: ammonia nitrogen; TN: total nitrogen; TP: total phosphorus; pH: potencial hydrogen; CBOD: carbonaceous biochemical oxygen demand; TKN: total Kjeldahl nitrogen; PO_4: phosphate/orthophosphate; NO_2: nitrate; NO_3: nitrite; T: temperature; EC: electrical conductivity

The quality of the treated effluent depends on the influent quality and process parameters of the WWTP (KHATRI; KHATRI; SHARMA, 2020). The explanatory variables (input) of the models were highly changeable in the studies, as many affect

WWTP performance. Most papers (52%) had influent wastewater quality and quantity indicators as input variables. This means that the majority of studies used influent characteristics to predict effluent wastewater quality, demonstrating the value of using ANNs to represent the complex and nonlinear relationship between raw influent and treated effluent water quality measurements (SALEH, 2021). For example, Bekkari and Zeddouri (2019) used the influent variables pH, temperature (T), TSS, total Kjeldahl nitrogen (TKN), BOD, and COD as inputs. The purpose of that study was to predict the performance of an activated sludge WWTP in Algeria in terms of effluent COD. In evaluating WWTP soft sensors, Ching, So and Morck (2021) also found that influent quality parameters were used in most cases as input variables for modeling effluent quality.

Other approaches included using treated effluent quality indicators as input variables to predict a different effluent indicator as the output, wastewater quality indicators sampled at different locations in the treatment train, and combinations of influent quality indicators and operational variables (such as returned sludge flow rate, sludge volume index, food/microorganism ratio, sludge retention time, and energy and chemical products consumption). For example, to predict the effluent concentrations of TP, BOD, COD, TSS, and NH_4-N in a WWTP (Harbin, China), Zhao et al. (2016) developed an ANN model using raw wastewater quality data (influent concentrations of TP, BOD, COD, TSS, NH_4-N, and influent pH) and energy consumption parameters (electricity consumption, coagulant, and flocculants) as the input variables.

Table 2 shows the most common input variables, all of which were included in more than 20% of the papers, highlighting their importance as predictors of WWTP performance in the ANN models. The majority of studies included indicators of organic matter, BOD and COD, as both input (influent concentrations, Table 2) and output (effluent concentrations, Table 1) variables. According to Ching, So and Morck (2021), COD is one of the strongest estimators for BOD; hence, most studies use COD concentrations as inputs for BOD models.

Other important input parameters in the models were influent TSS concentration, pH, nutrients concentration (NH_4-N, TN, and TP), and flow (Q). The choice of these variables may be related to their ease of measurement (such as pH and Q) or the

ability to develop models to predict some indicators in the treated effluent using the same indicator measured in the influent as one of the explanatory variables.

Table 2 – Number of publications with the most used input variables in the ANN models of the selected papers of the systematic review

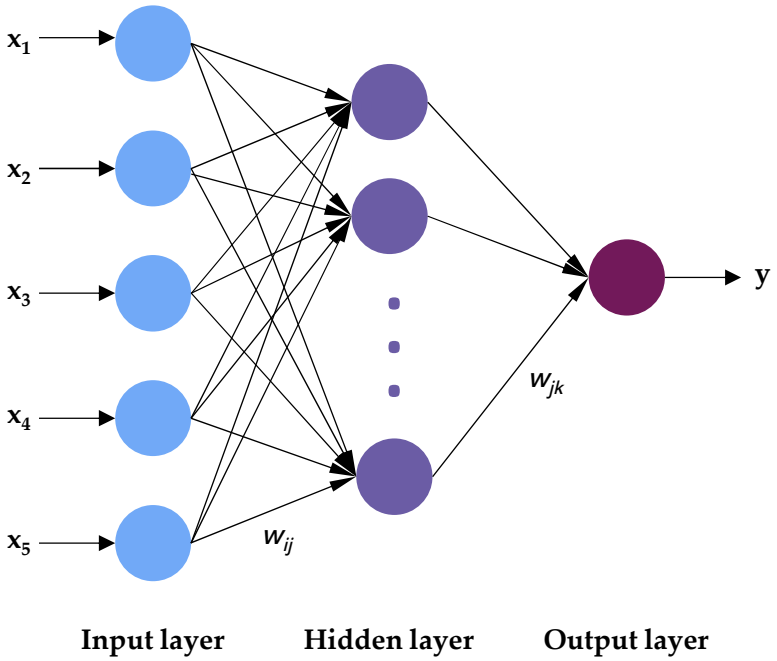
Input variable	Number of publications
Influent COD	31
Influent TSS	28
Influent BOD	25
Influent pH	20
Influent NH_4-N	17
Influent TN	11
Influent Q	10
Influent TP	9

2.3.5.3 ANN methods

There are several different classifications of ANNs (YE et al., 2020), and the most used model structure is the traditional feedforward neural network (FFNN), which was adopted in 21 papers (48%). This structure consists of one input layer, one or more hidden layers, and one output layer (Figure 6). The term feedforward describes the method in which the output of the neural network is calculated layer by layer from its input throughout the network (COROMINAS et al., 2018; MJALLI; AL-ASHEH; ALFADALA, 2007; PALANI; LIONG; TKALICH, 2008). Information is transmitted from one layer to another through serial operations (CIVELEKOGLU et al., 2009; PALANI; LIONG; TKALICH, 2008). According to Chen et al. (2020), most researchers use the FFNN for water quality prediction in WWTP systems, which may be because this method provides a good analysis of these systems.

The other commonly used neural network types are described next. A multilayer perceptron network (MLP) is a type of FFNN (BAGHERI et al., 2015), and was used in seven (16%) studies. According to Newhart, Hering and Cath (2022), a neural network that uses sigmoid functions in the hidden layer and a linear function in the output layer is more commonly referred to as an MLP. Nevertheless, most studies use MLP as equivalent to the traditional FFNN, and it can be considered a broader term that can be applied to various types of activation functions in the hidden and output layers.

Figure 6 – Typical neural network structure with one hidden layer



Index: x_i is the input variable; w_{ij} is the weight between input i and hidden neuron j; w_{jk} is the weight of the connection of neuron j in the hidden layer to neuron k in the output layer, and y is the output variable.

A radial basis function neural network (RBF) is another type of FFNN (BAGHERI et al., 2015) that uses radial basis activation functions in the hidden layer (CHEN et al., 2020). Although Newhart, Hering and Cath (2022) mentioned that RBF is increasingly used, it was adopted in only three (7%) papers in this systematic review.

An extreme learning machine (ELM) was used in four studies (9%). An ELM consists of a single hidden layer FFNN (ABBA; ELKIRAN; NOURANI, 2021) where the values of the weights between the input and hidden layers are randomly selected and the weights between the hidden and output layers are analytically characterized (PHAM et al., 2020). As an ELM only needs to learn the output weight, it can reduce computation problems because the weights of the input and hidden layers do not require adjustment (CHEN et al., 2020).

Deep learning refers to the use of multiple hidden layers in a network (COROMINAS et al., 2018) and is suitable for modern applications with highly complex processes (OSMAN; LI, 2020). Deep learning methods were used in three studies (7%). One of these (OSMAN; LI, 2020) was published in 2020, and the other two (EL-RAWY et al., 2021; WANG et al., 2021) in 2021. This result indicates that deep learning is a recent

technique. Corominas et al. (2018) did not find any advances in the identification of deep learning methods for wastewater treatment applications in papers published up to 2015.

Recurrent neural networks were used in three papers (7%), two of them utilizing long short-term memory (LSTM) methods. Recurrent neural networks are distinguished by their internal memory features, which allow observations to be considered in an ordered sequence (NEWHART; HERING; CATH, 2022). Recurrent neural networks allow signals to travel in both directions using loops to learn highly complex patterns (LANTZ, 2013). LSTM is capable of learning sequences of events over a period of time and can capture long-term dependencies in the data. Therefore, LSTM is frequently used to deal with time-series tasks, including those of wastewater data (LIU et al., 2021).

An adaptive neuro-fuzzy inference system (ANFIS) is a hybrid learning method that combines neural and fuzzy methods. It integrates the learning capacities of the ANN with fuzzy logic reasoning abilities to map the input-output relationships (YE et al., 2020). ANFIS uses a hybrid of backpropagation and least-squares algorithms to train the parameters and automatically generate “If/Then” rules (ZHAO et al., 2020). ANFIS was used in seven papers (16%).

2.3.5.4 Network structure

As shown in Figure 6, each layer of a neural network structure contains a certain number of neurons, also known as nodes. The numbers of input and output nodes are the number of features in the input data and the number of output variables to be modeled, respectively. The number of hidden layers and neurons in this(ese) layer(s) are configured by the user before training the model, and depend on the difficulty of the problem (SALEH, 2021). An insufficient number of hidden layer neurons may reduce prediction accuracy, causing underfitting problems. However, an excessive amount of neurons may lead to overfitting, whereby the error on the training set is driven to a small value and the test data are presented to the network with a large error. This implies that the generalization ability of the neural network was affected (CHEN et al., 2020; GAYA et al., 2014; YE et al., 2020).

In most studies (27 papers, 61%), the authors tuned the network structure using a trial-and-error approach, whereby ranges of values for the number of hidden layers and hidden neurons were tested to search for the optimum architecture. In some cases, other configurations were also tested by trial-and-error, such as the proportions of samples allocated to the training, validation, and testing subsets, and the training algorithms and activation functions to be used. In this trial-and-error approach, several ANNs are developed and compared to select the best result. For example, Sharghi et al. (2019) developed FFNN models to predict effluent BOD concentrations in an activated sludge WWTP. Those authors adopted one hidden layer, and the optimal hidden layer was determined by varying the number of nodes from 1 to 10. The authors observed the best results in a model with five neurons in the hidden layer.

Five of the 27 papers that adopted the trial-and-error approach established the range of hidden neurons and/or hidden layers to be tested using equations from the literature. To some extent, the use of equations may contribute to determining the model structure as they guide researchers based on previous studies (CHEN et al., 2020).

Another approach to determine the best network structure was adopted in five (11%) papers that used hybrid learning and combined various neural network methods (MLP, ANFIS, ELM, RBF, or deep learning) with a genetic algorithm (GA). A GA is an efficient search algorithm that can be applied to identify the combination of hyperparameters that will result in the best model performance (CHING; SO; MORCK, 2021). These hybrid models use a GA to iteratively optimize the parameters in the neural network to increase the problem-solving ability (ZHAO et al., 2020).

Table 3 shows the final and complete network structures of the papers that presented this information. The structure column indicates the number of neurons in the input layer, each hidden layer, and the output layer. For example, Jami, Mujeli and Kabbashi (2011) developed a model using the influent BOD concentration, NH_4-N concentration, pH, and Q as explanatory variables (four input neurons), with 15 neurons in the single hidden layer of the FFNN, to predict the effluent concentrations of NH_4-N (one output neuron) in a sequential batch reactor WWTP in Malaysia.

Table 3 – Neural network structure from 31 papers that presented this information

Reference	Output parameter(s)	Structure
Jami, Mujeli and Kabbashi (2011)	Effluent NH_4-N	4-15-1 [a]
Lee et al. (2011)	Effluent BOD	8-19-14-1 [b]
	Effluent COD	8-27-1 [b]
	Effluent SS	8-3-6-1 [b]
	Effluent TN	8-17-23-1 [b]
Qiao, Yang and Yuan (2011)	Effluent COD, BOD, SS, & NH_4-N (multi-output model)	8-4-8-4 [c]
Zhang and Hu (2012)	Effluent BOD	5-2-3-8-1 [d]
Chen and Lo (2012)	Effluent Q, BOD, COD, & SS (multi-output model)	4-16-4 [e]
Jami et al. (2012)	Effluent BOD, SS, COD (single-output models)	1-20-1 [a] or 3-30-1 [a]
Kusiak and Wei (2013)	Effluent CBOD	5-3-1 [e]
	Effluent TSS	5-10-1 [e]
Liu, Huang and Yoo (2013)	Effluent COD	9-54-6-6-1 [f]
Han, Wang and Qiao (2014)	Effluent BOD	5-150-1 [g] and 5-180-1 [g]
Gaya et al. (2014)	Effluent COD, SS, NH_4-N (single-output models)	5-10-1 [a]
Bagheri et al. (2015)	Effluent COD, TN, TSS (single-output models)	5-10-1 [b]; 5-5-1 [h]
Simsek (2016)	Effluent biodegradable dissolved organic nitrogen	4-10-1 [e]
Zhao et al. (2016)	Effluent TP, BOD, COD, SS, & NH_4-N (multi-output model)	9-19-5 [a], 9-19-5 [a], 9-16-5 [a], 9-14-5 [a], and 9-15-5 [a]
Nezhad et al. (2016)	Effluent quality index	8-7-1 [a]
Hazali, Wahab and Ibrahim (2017)	Effluent TN, TP, NH_4-N (single-output models)	6-6-1 [i]
Nourani, Elkiran and Abba (2018)	Effluent BOD, COD, TN (single-output models)	5-3-1 [a]
Elfanssi et al. (2018)	Effluent TSS, BOD, COD, total coliform, & fecal streptococci (multi-output model)	5-7-8-7-5 [a]
Sharghi et al. (2019)	Effluent BOD	3-5-1 [a]
Khatri, Khatri and Sharma (2019)	Effluent TSS	7-4-1 [a]
	Effluent pH, COD, TKN (single-output models)	7-5-1 [a]
	Effluent BOD, NH_4-N, TP (single-output models)	7-6-1 [a]
Bekkari and Zeddouri (2019)	Effluent COD	6-50-1 [a]
Khatri, Khatri and Sharma (2020)	Removal efficiency of fecal coliform	10-6-1 [a]
	Removal efficiency of total coliform	10-8-1 [a]
Ge et al. (2020)	Removal efficiency of arsenic	4-3-1 [a]
Al-Obaidi (2020)	Effluent quality index	5-3-1 [a]
Osman and Li (2020)	Effluent BOD	19-13-13-13-1 [j]
El-Rawy et al. (2021)	Removal efficiency of TSS, COD, BOD, NH_4-N, sulfide (single-output models)	5-8-1 [a]; 5-10-10-10-10-1 [k], 5-10-10-10-10-1 [l]
Wang et al. (2021)	Effluent TSS	32-128-256-128-1 [k]
	Effluent PO_4	32-256-128-128-1 [k]
Nourani, Asghari and Sharghi (2021)	Effluent BOD, COD (single-output models)	5-3-1 [a]
Alsulaili and Refaie (2021)	Effluent BOD	3-17-17-17-1 [a]
	Effluent COD	3-13-13-13-1 [a]
	Effluent TSS	3-11-11-11-11-1 [a]
Aldaghi and Javanmard (2021)	Effluent Q, BOD, COD, TSS, pH, T, TP, NO_3, TN, NO_2, NH_4-N, & EC (multiple-output model)	12-25-12 [e]

Reference	Output parameter(s)	Structure
Saleh (2021)	Effluent COD	9-6-6-1 [a]
	Effluent BOD	9-6-6-6-1 [a]
	Effluent TSS	9-6-6-6-1 [a]
	Effluent COD, BOD & TSS (multiple-output model)	7-6-6-3 [a]
Abba, Elkiran and Nourani (2021)	Effluent BOD	6-6-1 [e]
	Effluent COD, TN, TP (single-output models)	9-10-1 [e]

Obs.: Neural network methods [a] FFNN; [b] MLP-GA; [c] RHONN; [d] SWNN; [e] MLP; [f] ANFIS-GA; [g] HELM; [h] RBF-GA; [i] SO-RBF; [j] DSAE-NN-GA; [k] DFFNN; [l] DCB.

Although some recent studies have used deep learning, most developed shallow neural networks with a single hidden layer (Table 3). Other review papers have identified that most ANN models use a single hidden layer (COROMINAS et al., 2018; YE et al., 2020) as this is usually sufficient to investigate many problems (SALEH, 2021). There was a wide range in the number of neurons in the hidden layer(s) of the studies, from two to 256.

Considering the studies that developed single-output models for both BOD and COD effluent concentrations (the two most common target variables in the studies, Table 1), the same network structure for the two variables was adopted in three papers (JAMI et al., 2012; NOURANI; ASGHARI; SHARGHI, 2021; NOURANI; ELKIRAN; ABBA, 2018). In the other four papers, more complex structures were used to model BOD effluent concentrations, with greater numbers of hidden layers (LEE et al., 2011; SALEH, 2021) or hidden neurons (ALSULAILI; REFAIE, 2021; KHATRI; KHATRI; SHARMA, 2019). Only one study (ABBA; ELKIRAN; NOURANI, 2021) had a larger number of hidden neurons for the COD model. This result shows that modeling BOD concentrations may be more complex than modeling COD concentrations, with more intricate network structures required to map the relationship between the input and output phases.

2.3.5.5 Activation functions

In a neural network, each artificial neuron in the hidden and output layers calculates the weighted sum of its inputs and produces an output value using predefined activation functions, also known as transfer functions (ELFANSSI et al., 2018; MJALLI; AL-ASHEH; ALFADALA, 2007). Therefore, the activation function is applied to a certain layer to obtain the output of that layer, which is then used as the input for the next layer (SHARMA; SHARMA; ANIDHYA, 2020).

Activation functions introduce nonlinearity into the neural network. The choice of the activation function is important because it affects the prediction performance of the neural network (SHARMA; SHARMA; ANIDHYA, 2020).

From the papers in the systematic review that included this information, the most common activation functions in the hidden layer were the logistic sigmoid (nine studies) and hyperbolic tangent (12 studies) functions. This result is in accordance with Corominas et al. (2018), who mentioned hyperbolic tangent and sigmoid functions among the typically applied activation functions in ANN models for nonlinear classification and regression problems in wastewater treatment research. Newhart, Hering and Cath (2022) stated that the most widely used ANN activation function in environmental engineering is the logistic sigmoid function.

The logistic sigmoid function is given by Equation (3), where *x* is the input of the activation function. The curve resembles an S-shape, and the returned values range from 0 to 1 (FENG; LU, 2019).

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

The hyperbolic tangent function is given by Equation (4). The output values range from -1 to 1. The function is symmetric around the origin, which makes its outputs more likely to be closer to zero than those of the sigmoid function, leading to faster convergence. For this reason, it is often used in hidden layers of ANNs (FENG; LU, 2019), which may be the reason that it was the most used in the studies of this systematic review.

$$f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} = \frac{2}{1 + e^{-2x}} - 1 \tag{4}$$

The output layer is the layer in the neural network model that directly returns a prediction. The most used activation function in the output layer of the neural network models of this systematic review (15 studies) was the linear activation function, also called “identity function” or “no activation.” The linear activation function is given by Equation (5) and is represented by a straight line. It does not change the weighted sum of the previous layer, but only returns the value directly. The outputs can range from -∞ to +∞ (FENG; LU, 2019).

$$f(x) = x \quad (5)$$

2.3.5.6 Training algorithms

The training of a neural network is performed by adjusting the neurons weights to minimize the error between the observed data and network output (MJALLI; AL-ASHEH; ALFADALA, 2007; NASR et al., 2012). The most common learning algorithm used for this purpose is backpropagation, which involves working backward layer by layer from the output to adjust the weights accordingly and reduce the average error across all layers (MJALLI; AL-ASHEH; ALFADALA, 2007; NEWHART et al., 2019; NEZHAD et al., 2016). The backpropagation algorithm was used in 27 papers in this systematic review (61%).

Backpropagation is the most widely used ANN training algorithm (CHEN et al., 2020; NEWHART; HERING; CATH, 2022; YE et al., 2020; ZHAO et al., 2016), and is commonly applied in the field of environmental pollution control (YE et al., 2020). The majority of applications of neural networks in engineering or wastewater treatment problems use the FFNN architecture with a backpropagation training algorithm because of its accuracy and capability (AL-GHAZAWI; ALAWNEH, 2021).

The standard backpropagation algorithm uses the gradient descent optimization method to perform calculations (CHEN; LO, 2012; ZHAO et al., 2016). This method involves the network weight value moving along a negative gradient of the performance function. Hence, the weight and bias values are continually renewed to minimize the performance function (CHEN; LO, 2012).

2.3.5.7 Software tools

Sixteen studies (36%) did not specify which software tools were used for model development. Among the papers that provided this information, the most frequently used was MATLAB, which was used in 21 publications (48%). Other tools included R (two studies, 5%), SPSS (two studies, 5%), Python (one study, 2%), MATLAB integrated with C++ (one study, 2%), and MATLAB integrated with C# (one study, 2%). Among these, MATLAB and SPSS are paid, and R, Phyton, C++, and C# are free tools.

MATLAB was also found by Corominas et al. (2018), Ye et al. (2020), and Bahramian et al. (2023) to be the most popular software platform in the literature for modeling WWTPs with AI techniques. According to the authors, the wide usage of this software platform is due to its packages and toolboxes, which are user-friendly and convenient for users with minimal knowledge of data science (BAHRAMIAN et al., 2023; YE et al., 2020).

2.3.6 Model performance

Since machine learning models produce a biased solution to the learning problem, it is essential to assess how effectively the algorithm has learned from its experience (LANTZ, 2013). To evaluate the accuracy of prediction methods in a real-world application, various metrics can be employed during the training and testing phases of the models. While there is no standard in the literature as to which specific metrics are used, it is common practice to rely on a measure of both training and testing error to assess model fit and performance (NEWHART; HERING; CATH, 2022).

The model performance indicates the results of a comparison of the experimental data with the predicted data (ZHAO et al., 2020). The performance of the models in the studies was calculated using various statistical metrics, including error (mainly mean square error [MSE] and root mean square error [RMSE]) and goodness-of-fit (mainly correlation coefficient [R] and coefficient of determination [R^2]). The MSE and RMSE indicators identify the errors between the experimental values and model output, with smaller results signifying higher accuracy. The metrics R and R^2 indicate the degree of correlation between the observed and predicted values, with higher R or R^2 values indicating better prediction performance (YE et al., 2020).

Some papers presented the metrics separately for the training and testing subsets, each target variable being modeled, and each type of model used. For this reason, there are many results for model performance.

The following discusses the results of the performance of the models. As the metrics of errors, RMSE and MSE, depend on the unit of the variable or if they are presented as normalized data, the R and R^2 results are presented in Table 4. This data highlights

the large variability in the results, with R ranging from -0.018 to 0.998 and R^2 from 0.260 to 0.998.

It is unfeasible to determine the reasons for the differences between the studies because the context of each application is different, with distinct methods, target parameters, and datasets (CHING; SO; MORCK, 2021). Even when a single study developed different types of neural network methods for the same target variable, various situations were observed. For example, Yasmin et al. (2017) observed a better prediction accuracy of the ANFIS model compared with the FFNN method when modeling the pH effluent. In contrast, Nourani, Asghari and Sharghi (2021) achieved similar results with ANFIS and FFNN when modeling the same output parameters (effluent BOD and COD concentrations). This highlights that the advantage of one method over another may be due to the context of the application, the differences in the dataset used, and the configuration settings in the model of each study.

Table 4 – Model performance in terms of goodness-of-fit indicators

Reference	Output parameter(s)	ANN methods	Model performance
Jami, Mujeli and Kabbashi (2011)	Effluent NH_4-N	FFNN	R=0.7980
Zhao, Chai and Yuan (2012)	Effluent BOD	Selective ensemble ELM-GA	R^2=0.7576
	Effluent COD		R^2=0.7729
	Effluent SS		R^2=0.5957
	Effluent NH_4-N		R^2=0.8273
Chen and Lo (2012)	Effluent Q	MLP	R=0.9781
	Effluent BOD		R=0.6963
	Effluent COD		R=-0.0178
	Effluent SS		R=0.1031
Jami et al. (2012)	Effluent BOD	FFNN	R=0.346948
	Effluent COD		R=0.052622
	Effluent SS		R=0.158717
Liu, Huang and Yoo (2013)	Effluent COD	ANFIS-GA	R^2=0.800
	Effluent TN		R^2=0.577
	Effluent TP		R^2=0.284
Gaya et al. (2014)	Effluent COD	FFNN	R=0.647
	Effluent SS		R=0.512
	Effluent NH_4-N		R=0.425
	Effluent COD	ANFIS	R=0.847
	Effluent SS		R=0.995
	Effluent NH_4-N		R=0.948
Bagheri et al. (2015)	Effluent COD	MLP-GA	R^2=0.98044
	Effluent TN		R^2=0.98479
	Effluent TSS		R^2=0.95484
	Effluent COD	RBF-GA	R^2=0.97232
	Effluent TN		R^2=0.98325
	Effluent TSS		R^2=0.95217

Simsek (2016)	Effluent biodegradable dissolved organic nitrogen	ANFIS MLP RBF GRNN	R^2=0.94 R^2=0.78 R^2=0.66 R^2=0.97
Heddam, Lamda and Filali (2016)	Effluent BOD	GRNN	R=0.922
Nezhad et al. (2016)	Effluent quality index	FFNN	R=0.96
Hazali, Wahab and Ibrahim (2017)	Effluent TP Effluent TN Effluent NH_4-N	SO-RBF	R^2=0.8442 R^2=0.7282 R^2=0.2833
Yasmin et al. (2017)	Effluent pH	FFNN ANFIS	R=0.39698 R=0.70868
Nourani, Elkiran and Abba (2018)	Effluent BOD Effluent COD Effluent TN	FFNN	R^2=0.6600 R^2=0.9363 R^2=0.9022
	Effluent BOD Effluent COD Effluent TN	ANFIS	R^2=0.7640 R^2=0.9260 R^2=0.9410
Sharghi et al. (2019)	Effluent BOD	FFNN	R^2=0.67
Khatri, Khatri and Sharma (2019)	Effluent pH Effluent BOD Effluent COD Effluent TSS Effluent TKN Effluent NH_4-N Effluent TP	FFNN	R=0.816 R=0.649 R=0.656 R=0.457 R=0.670 R=0.493 R=0.748
Bekkari and Zeddouri (2019)	Effluent COD	FFNN	R=0.8781
Khatri, Khatri and Sharma (2020)	Removal efficiency of fecal coliform Removal efficiency of total coliform	FFNN	R=0.986 R=0.977
Ge et al. (2020)	Removal efficiency of arsenic	FFNN	R^2=0.851
Al-Obaidi (2020)	Effluent quality index	FFNN	R^2=0.998
Osman and Li (2020)	Effluent BOD	DSAE-NN-GA	R^2=0.987
El-Rawy et al. (2021)	Removal efficiency of BOD Removal efficiency of COD Removal efficiency of TSS Removal efficiency of NH_4-N Removal efficiency of sulfide	FFNN	R=0.55564 R=0.90859 R=0.52105 R=0.95459 R=0.9866
	Removal efficiency of BOD Removal efficiency of COD Removal efficiency of TSS Removal efficiency of NH_4-N Removal efficiency of sulfide	DFFNN	R=0.76327 R=0.66487 R=0.70718 R=0.99427 R=0.92402
	Removal efficiency of BOD Removal efficiency of COD Removal efficiency of TSS Removal efficiency of NH_4-N Removal efficiency of sulfide	DCB	R=0.77167 R=0.94572 R=0.80847 R=0.97696 R=0.98585
Al-Ghazawi and Alawneh (2021)	Effluent BOD Effluent COD	FFNN	R^2=0.48 R^2=0.45

	Effluent SS Effluent NH_4-N		R^2=0.44 R^2=0.26
Wang et al. (2021)	Effluent TSS Effluent PO_4	DFFNN	R^2=0.920 R^2=0.872
Nourani, Asghari and Sharghi (2021)	Effluent BOD Effluent COD	FFNN	R^2=0.7182 R^2=0.7178
	Effluent BOD Effluent COD	ANFIS	R^2=0.7203 R^2=0.7148
Elmaadawy et al. (2021)	Effluent BOD Effluent TSS	RVFL	R^2=0.924 R^2=0.917
Alsulaili and Refaie (2021)	Effluent BOD Effluent COD Effluent TSS	FFNN	R^2=0.752 R^2=0.6115 R^2=0.6308
Abba et al. (2021)	Effluent TSS Effluent pH	NARX	R^2=0.9846 R^2=0.6293
Hejabi et al. (2021)	Effluent BOD Effluent COD Effluent TSS	FFNN	R^2=0.760 R^2=0.715 R^2=0.632
Liu et al. (2021)	Effluent COD	LSTM-AM	R^2=0.869
Aldaghi and Javanmard (2021)	Effluent Q, BOD, COD, TSS, pH, T, TP, NO_3-N, TN, NO_2-N, NH_4-N, & EC	MLP	R=0.5804
Saleh (2021)	Effluent BOD Effluent COD Effluent TSS	FFNN	R=0.99782 R=0.77301 R=0.8317
Abba, Elkiran and Nourani (2021)	Effluent BOD Effluent COD Effluent TN Effluent TP	ELM	R^2=0.6341 R^2=0.9742 R^2=0.9656 R^2=0.8807
	Effluent BOD Effluent COD Effluent TN Effluent TP	MLP	R^2=0.5776 R^2=0.9555 R^2=0.86662 R^2=0.72544
Rahmati, Tishehzan and Moazed (2021)	Effluent BOD	FFNN ANFIS	R=0.897 R=0.930

2.3.7 Limitations of the review and future perspectives

The ever-evolving nature of machine learning techniques leads to numerous possibilities for applications in the wastewater treatment sector. This systematic review focused specifically on the use of ANNs for predicting the performance of WWTPs in terms of effluent quality and removal efficiencies. This more focused approach was necessary due to the rigorous methods employed in a systematic review, allowing for thorough selection, screening, and analysis of publications, facilitating a deeper understanding of the main architectures, hyperparameter configurations of the models, and assessment of the studies. It is important to note that the implementation of the

models in real-world WWTPs was not the primary focus of this work. However, it is worth mentioning that one of the main challenges in implementing these models remains the availability of high-quality data (COROMINAS et al., 2018; FAISAL et al., 2023; RAY et al., 2023).

Other systematic reviews should be conducted for other specific applications of neural networks and even other machine learning algorithms in the wastewater treatment sector. For instance, neural networks and different machine learning approaches have been utilized for the optimization of WWTPs, including operational cost and energy consumption optimization, automation, control of operational conditions, real-time monitoring, forecasting of membrane fouling or operational failure (RAY et al., 2023), fault detection, and multi-objective control strategies that aim to maintain effluent quality while reducing energy consumption (FAISAL et al., 2023). Each of these applications could serve as the focus of new systematic reviews.

Still considering the constantly evolving nature of machine learning and its applications, according to Zhang et al. (2023), future AI research applied to wastewater treatment will continue to focus on the removal of phosphorus, organic pollutants, and emerging contaminants. Promising directions for research include exploring microbial community dynamics, achieving multi-objective optimization, improving the performance of WWTPs to remove various pollutants, and predicting water quality under specific conditions (ZHANG et al., 2023).

2.4 CONCLUSIONS

The results of the systematic review of the use of ANN models for the prediction of the performance of full-scale WWTPs, considering 44 relevant papers that were extracted and assessed accordingly, indicated the main trends and applications in the field. Most studies modeled a large activated sludge facility because they have better monitoring and control schemes. The datasets usually included a monitoring period of one to two years, with daily samplings, resulting in relatively small datasets (median = 361.5). Prior to training the models, the most common preprocessing method was the min-max normalization in the range [0, 1], and data division was achieved mainly with either 75% for training and 25% for testing the model, or 70% for training, 15% for validation, and 15% for testing.

The publications used influent indicator qualities as the input variables for neural network models to predict WWTP effluent quality, mainly those of organic matter concentrations. Although other methods were utilized, such as MLP, RBF, hybrid learning, and in recent years, deep learning, the FFNN architecture with a backpropagation training algorithm was the most common. In general, shallow networks with single hidden layers were used, and good performance was achieved.

Not all models must be tuned in the same manner, as they vary according to the dataset characteristics and study objectives. However, the findings of this research may act as a starting point and provide highly beneficial information to industry and research practitioners in the search for an optimum design modeling process in future studies with similar prediction problems.

CHAPTER 3: CASE STUDIES, DATA DESCRIPTION, AND EXPLORATORY ANALYSES

3.1 PRESENTATION / INTRODUCTION

This chapter describes the Brasília Sul and the John E. Egan WWTPs, which were selected as the case studies for the application of the artificial neural network models for performance prediction. The monitoring datasets were presented, along with the preliminary analyses and exploration of the data.

The WWTP of Brazil selected for the case study has a national relevance as it is the largest facility in Brazil's capital municipality in terms of design flow and employs a tertiary treatment level (GDF, 2017c), which is advanced compared to typical practices in Brazil (ANA, 2020). Despite the importance of this WWTP, few studies have been conducted to evaluate its performance, with a comprehensive analysis of its monitoring data.

A WWTP in operation in the United States, the John E. Egan, was selected to integrate the survey data to allow a comparative assessment of the results observed in the two countries. In developing countries, researchers in wastewater treatment face the problem of data scarcity (VON SPERLING; VERBYLA; OLIVEIRA, 2020). The dataset from a treatment system in a developed country may present a higher monitoring frequency and more variables being available. Besides that, in developed countries, sanitation systems were established decades ago (KHALID et al., 2018), which could result in longer time series of monitoring data. All these characteristics could lead to a larger number of samples in the dataset, which is important for a better understanding of the plant's performance and for model development.

The examination of the data in detail is an important step prior to developing a machine learning model, as a better understanding of the data allows better definitions in the machine learning process. The exploration of the data demands a great effort and a great deal of human intervention (LANTZ, 2013) and also requires familiarity with the source of the data and the process itself (NEWHART et al., 2019). The data exploration involves identifying the structure and characteristics of the dataset, calculating summary statistics of numeric variables (such as measures of central tendency, spread, and amplitude of data), and using data visualization techniques (LANTZ, 2013). Some examples of data visualization are time series plots to visualize

observations recorded over time, and box-plot graphs to visualize data distribution (LANTZ, 2013; NEWHART et al., 2019).

The term “Exploratory Data Analysis” (EDA) refers to the process of understanding the data to find patterns, test hypotheses, visualize the data, and use summary statistics (DEMING; DEKKATI; DESAMSETTI, 2018). This investigation and exploration of the data should precede any predictive modeling since before constructing models, it is necessary to become familiar with the data and prepare it for modeling analysis (DEMING; DEKKATI; DESAMSETTI, 2018). For example, in this chapter, variables that have more consistent monitoring, with fewer gaps, were identified to be included in the next chapter’s modeling process.

Many existing wastewater treatment systems are poorly investigated, and modeling the system can provide a concise method to capture the existing system's process. This information can then be used to help decision-makers manage the system or evaluate it with the aim of improving it (PHAM et al., 2020).

This chapter presents the datasets, along with statistical analyses and data visualization methods to explore the monitoring datasets and gain a deeper understanding of the Brasília Sul and the John E. Egan plants' processes and performances. The preliminary analyses of this chapter allowed a better understanding of the behavior of the WWTPs under investigation and their monitoring datasets. This is a crucial step in the study, as having a thorough understanding of the dataset subsidized decisions and criteria for the next stage, of the models' development, such as selecting variables for use as input and output for the models and data dividing for model training and testing.

3.2 METHODS

3.2.1 Case studies

The study began with a comprehensive and detailed description of the WWTPs case studies in Brazil and the United States. This included information about their location and the treatment process employed, which helped to underscore their significance in each national context and justify their selection.

3.2.1.1 Case study 1: the Brasília Sul wastewater treatment plant

The Federal District comprises an area of around 5,780 km² and is the smallest federative unit of Brazil. It has a resident population of over three million, with an urbanization rate of 96% (CAESB, 2021). The Federal District is an autonomous federative unit that accumulates the functions of both a state and a municipality. Its territory is divided into Administrative Regions (CAESB, 2021). In its area is located the federal capital of Brazil, Brasília, which is also the seat of the government of the Federal District (GDF, 2017a).

Despite its location in the Central-West region of Brazil, at the headwaters of the tributaries of three of the largest Brazilian rivers - the Maranhão River (a tributary of the Tocantins River), the Preto River (a tributary of the São Francisco River) and the São Bartolomeu, Descoberto and São Marcos rivers (tributaries of the Paraná River) - the water bodies in the Federal District have low water flow (CAESB, 2021), which decreases significantly during drought periods (GDF, 2017a).

In the Federal District, 89.3% of the population is served by wastewater collection and treatment, making it the federal unit with the highest coverage rate in Brazil (ANA, 2020). The Federal District has 15 wastewater treatment plants with various technologies and capacities, which are operated by the Caesb (Environmental Sanitation Company of the Federal District, in Portuguese Companhia de Saneamento Ambiental do Distrito Federal) (ADASA, 2020). These WWTPs are located in four watersheds of the Federal District (CAESB, 2021). Of the total volume of wastewater collected, 86% undergoes tertiary treatment to remove nutrients (CAESB, 2021), reflecting the higher level of demand for the operation of WWTPs compared to the rest of the country. This more advanced coverage in sanitation and higher wastewater treatment level is due to the water-limited characteristics of the region’s water bodies and the presence of lakes, which are lentic environments that are sensitive water bodies for effluent disposal (GDF, 2017b). Strict operational control is necessary to meet the requirements to preserve these receiving water bodies (CAESB, 2021).

The Brasília Sul WWTP is the largest plant in the Federal District regarding design flow, with a capacity of 1,500 L/s and a population equivalent of 460,000 inhabitants (CAESB, 2021). This WWTP was selected for this study because it is located in Brazil's

capital and discharges effluent into a sensitive receiving water body, Lake Paranoá (Figure 7). To prevent eutrophication, Brasília Sul WWTP adopts a tertiary treatment aimed at nitrogen and phosphorus removal, which is considered advanced compared with the majority of Brazilian treatment systems. As a result, this WWTP has more stringent operational procedures and standards and a higher demand for its system operation and monitoring. Thus, the dataset resulting from this monitoring contains a large amount of information, allowing for more robust tools for data analysis.

Lake Paranoá was formed by the damming of the Paranoá River in 1959. The Lake had various functions such as enhancing the landscape of Brasília, providing leisure activities, improving the microclimate, and generating electricity. Over time, it started to have other functions such as serving as a means of transportation, supporting commercial and subsistence fishing, and receiving and diluting sanitary effluents. As a result, Lake Paranoá is considered an artificial lake with multiple uses (PINHO; SANTOS, 2016).

The Brasília Sul WWTP was inaugurated in 1962 to serve a population of 150,000 inhabitants employing conventional activated sludge technology (GDF, 2017c). However, the accelerating eutrophication process in the 1970s and 1980s prompted the implementation of a program for the recovery and maintenance of Lake Paranoá's water quality, which included the expansion and modernization of WWTPs that discharge effluents into the Lake. The current Brasília Sul WWTP has a capacity of 1,500 L/s and employs tertiary treatment for the removal of phosphorus and nitrogen, having started operations in 1993 (GDF, 2017c; PINHO; SANTOS, 2016).

Figure 7 – Location of the WWTP under study (Brasília Sul)

The original function of Lake Paranoá was not to supply water for the population (CAESB, 2021). However, with the improvement in the water quality of Lake Paranoá, the Caesb company started to use the watershed for abstraction for water supply in 2009. Given the water crisis that affected the Federal District from 2015 to 2018, the company was granted permission to extract water from Lake Paranoá. In October 2017, Caesb implemented water treatment using ultrafiltration membranes in the system that collects raw water directly from the lake, with a design flow rate of 700 L/s (BERTOLOSSI; NEDER; BRANDÃO, 2021).

Given the use of Lake Paranoá as a source of drinking water, there is an urgent need for studies to evaluate the quality of treated wastewater that is discharged into the lake, including the effluent from Brasília Sul WWTP (BERTOLOSSI; NEDER; BRANDÃO, 2021).

Figure 8 shows the current wastewater treatment process at the Brasília Sul WWTP, emphasizing the liquid phase. In the preliminary treatment, coarse solids and sand are removed. After compacting, the coarse solids are destined for disposal at the Brasília landfill, along with the sand. All preliminary treatment units are covered and equipped with a gas collection network, which is directed to the gas treatment system, with the objective of odor control.

The next stage is the primary treatment, which is performed using three primary settlers, each with a diameter of 32 m, a liquid depth of 3.5 m, and a volume of 3,110 m³ (REBELO, 2019). The solids removed at this step constitute the primary sludge, which is directed to the solid treatment, while the liquid phase is directed to the next step.

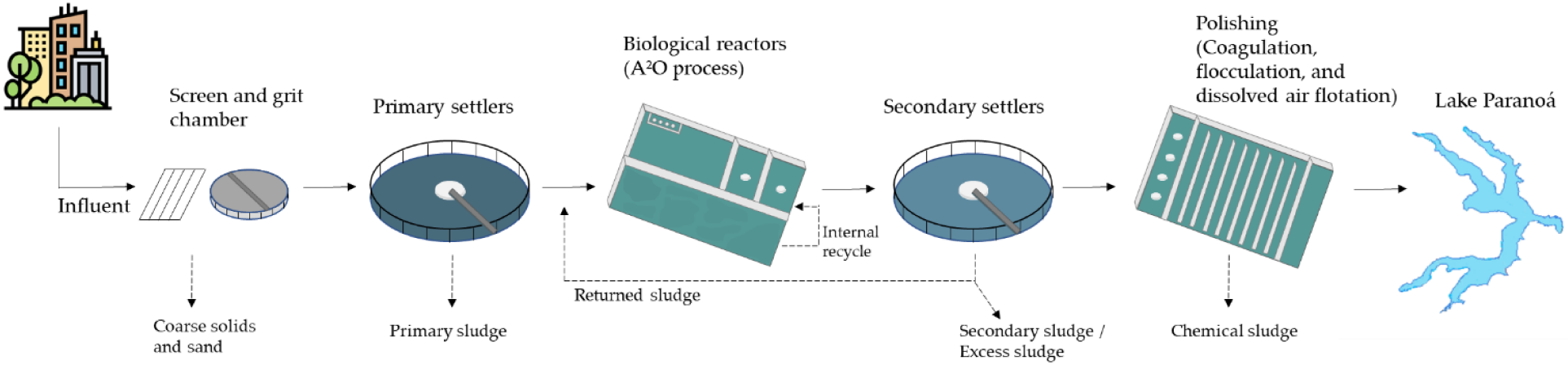
The biological treatment in this WWTP is the A²O (anaerobic/anoxic/oxic) activated sludge process, also called Phoredox. There are four biological reactors with a capacity of 8,245 m³ (REBELO, 2019). This process not only removes organic matter and solids, but also combines the removal of nitrogen and phosphorus. The biological nutrient removal occurs through the action of microorganisms under different redox conditions: aerobic, anaerobic, and anoxic. Nitrification, denitrification, and biological phosphorus removal are key metabolic processes involved. To effectively remove

nitrogen, aerobic-anoxic condition is ideal, whereas for phosphorus removal, an alternating anaerobic-aerobic condition is necessary (ROUT et al., 2021).

The Brasília Sul WWTP has 12 secondary settlers (GDF, 2017c), where the settling of the solids, or biomass, takes place. A portion of the solids is recirculated to the anaerobic zone of the biological reactors as returned sludge to maintain a high biomass concentration in the reactors. The biomass concentration in the reactors is controlled by withdrawing the remaining part of the solids, or excess sludge, from the system. These solids, referred to as biological sludge, are directed to the sludge treatment stage (VON SPERLING, 2007).

Since biological phosphorus removal is achieved through the incorporation of excess amounts of phosphorus into the bacterial biomass, the loss of suspended solids in the effluent from the secondary treatment can lead to an increase in phosphorus concentrations (VON SPERLING, 2007). Thus, as WWTP Brasília Sul discharges the effluent into a lake, low phosphorus levels are needed, and polishing stages are adopted for further removal of suspended solids. This is done through coagulation, flocculation (using aluminum sulfate as coagulant and anionic polyelectrolyte as flocculant), and dissolved air flotation. The solids produced (chemical sludge) are then removed by surface scrapers and directed to solid phase treatment. The final effluent is then discharged into Lake Paranoá.

Figure 8 – Schematic diagram of the Brasília Sul WWTP process (liquid phase)



Source: author's own elaboration

3.2.1.2 Case study 2: the John E. Egan wastewater treatment plant

Illinois is a state in the Midwestern region of the United States. It has a land area of 143,793 km² and its population was 12,812,508 in 2020 (USCB, 2023a). The capital of Illinois is Springfield. However, the largest city is Chicago, with a population of 2,746,388 in 2020, being the third largest city in the USA (USCB, 2023a).

Chicago is in Cook County, which has a land area of 2,447 km² and had a population of 5,275,541 people in its 135 municipalities in 2020 (USCB, 2023a). The Cook County is the second most populous county in the United States and is part of the Greater Chicago, which is the third largest metropolitan area of the nation (USCB, 2023b).

Chicago is located in between two major watersheds, namely the Great Lakes and the Mississippi River watersheds. The Des Plaine River is in the Mississippi River basin, which ends up in Gulf of Mexico. The Chicago River is in the Great Lakes watershed as it is a tributary of Lake Michigan. The historical water problems of the region and their solutions are associated with Chicago's unique location, topography, and hydrography (MACAITIS, 1985).

The Metropolitan Water Reclamation District of Greater Chicago (MWRD) is a government agency responsible for stormwater management and wastewater collection and treatment in Cook County. The MWRD originated in 1889, originally named as the Sanitary District of Chicago, in response to the rapid population growth and consequent health concerns in the region. The MWRD had originally the goal of reversing the flow of the Chicago and the Calumet rivers through channels in order to protect the water quality of Lake Michigan (the main source of water supply in the metropolitan area) from sewage that was discharged into surface water and mitigate the flooding in the region (MWRD, 2024a; MWRD, 2024b; PLUTH et al., 2021). The flow was reversed, and water started to flow into the Des Plaines River, which is in the Mississippi River watershed (MWRD, 2024b). More than 98 km of canals and waterways were constructed and are known as the Chicago Area Waterway System (MWRD, 2024a).

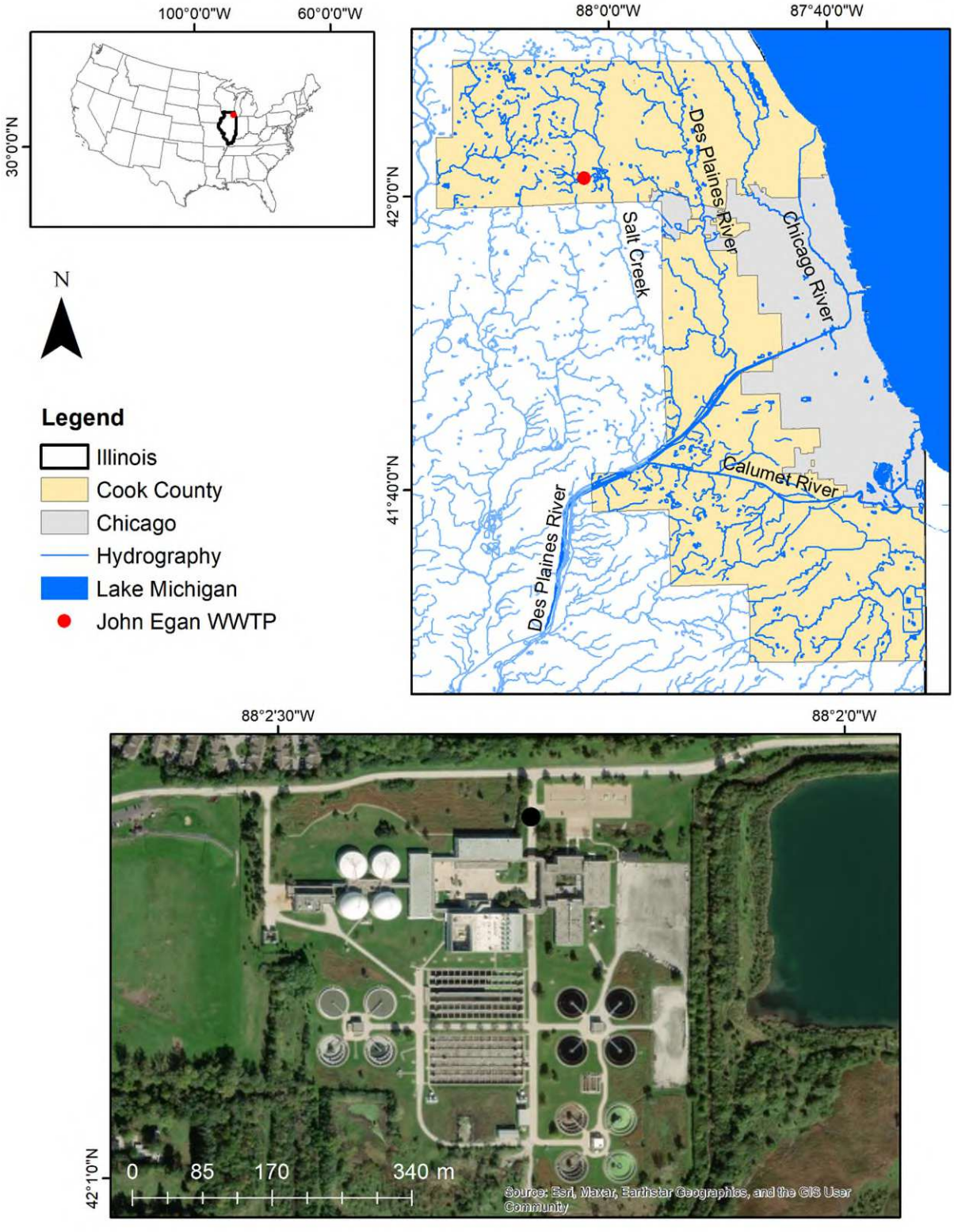
However, until the beginning of the 20th century, the effluents were only collected and removed away from residences but were not treated. In the early 20th century, the

MWRD started to build pilot plants to develop technologies for the treatment of wastewater while constructing the intercepting sewers. The first full-scale WWTP (the Calumet WWTP) was implemented in 1922 and another six plants were implemented after that. One of the WWTPs operated by the MWRD, namely the Stickney WWTP, is one of the largest plants in the world, with a capacity to treat more than 63,000 L/s (MWRD, 2024a). The treatment plants are called Water Reclamation Plants (WRPs) by the agency. However, to remain consistent with the terminology of this study, they will be named WWTPs.

As there are seven WWTPs in Cook County that are operated by the MWRD, including some of the largest in the world, it was selected a facility for this study with the closest size to the Brasília Sul WWTP in Brazil, so they are better comparable. John E. Egan WWTP has capacity to treat 2,190 L/s, with an average flow of 1,314 L/s. It serves 160,735 residents of the municipalities of Arlington Heights, Elk Grove Village, Hoffman Estates, Inverness, Palatine, Rolling Meadows, Roselle, and Schaumburg. John Egan WWTP has been in operation since 1975 (MWRD, 2019). It treats mainly domestic sewage collected through separate sewer systems (ZHANG et al., 2006; ZHANG et al., 2008).

John E. Egan WWTP is a single stage nitrification activated sludge with tertiary filtration and disinfection process. Wastewater undergoes seasonal disinfection through chlorination and dichlorination, and tertiary filtration is carried out as a polishing step for further removal of solids using tertiary filters made of dual media of sand and anthracite (MWRD, 2015). The main treatment units for the liquid phase treatment include screens, aerated grit tanks, primary settling tanks, activated sludge aeration tanks, secondary clarifiers, and tertiary filters (ZHANG et al., 2008). Figure 9 shows the location of the facility.

Figure 9 – Location of the WWTP under study (John E. Egan)



The average hydraulic detention time of the facility is 7.8 hours (MWRD, 2015). The plant typically operates its aeration tanks with mixed liquor suspended solids ranging from 2000 to 2500 mg/L, solids retention time of 8 to 15 days, and relatively constant return flow at approximately 70% of the influent rate (ZHANG et al., 2006). After

treatment, the final effluent is discharged into the Salt Creek (MWRD, 2019), which is a tributary of the Des Plaines River. During dry weather, John Egan WWTP effluent constitutes the major inflow to the creek for approximately four kilometers (ZHANG et al., 2008).

3.2.2 Monitoring datasets

3.2.2.1 Brasília Sul WWTP

Data from the Brasília Sul WWTP were obtained from the Environmental Sanitation Company of the Federal District (Caesb, in Portuguese Companhia de Saneamento Ambiental do Distrito Federal). In the initial analyses all variables associated with the liquid phase that are monitored in the process were included. The monitoring frequency was analyzed since it differed by variable. The monitoring period was from January 2020 to March 2022. The monitoring dataset included quality parameters of samples collected from the influent and final effluent, as presented in Table 5.

Table 5 – Monitored parameters in each site at Brasília Sul WWTP

Site	Variables
Influent	BOD, COD, *E. coli*, O&G, pH, TA, TN, TP, TFS, TS, TVS, TSS
Effluent	BOD, COD, CODf, *E. coli*, O&G, pH, NH_4-N, TN, TP, PO_4, SO_4, SS, TSS

BOD: biochemical oxygen demand; COD: chemical oxygen demand; CODf: filtered COD; *E. coli*: *Escherichia coli*; NH_4-N: ammonia nitrogen; O&G: total oil and grease; pH: potential hydrogen; PO_4: phosphate; SO_4: sulfate; SS: sedimentable solids; TA: total alkalinity; TFS: total fixed solids; TN: total nitrogen; TP: total phosphorus; TS: total solids; TSS: total suspended solids; TVS: total volatile solids

A second spreadsheet obtained from the Brasília Sul WWTP monitoring dataset contained data on operational variables. All variables associated with the liquid phase are presented in Table 6. As this study focuses on the liquid phase of the treatment, the variables associated with the solid phase and energy consumption were not assessed.

Table 6 – Operational variables presented in the dataset

Characteristic	Variables	Frequency
Flow	Influent flow (m^3), Effluent flow (m^3), Bypass (m^3)	Daily
Chemical products consumption	Aluminum sulfate (kg), Anionic polyelectrolyte (kg)	

The dataset contained information on daily influent and effluent flow. Only in four records, differences were found between the values of influent and effluent flow. For

this reason, only influent flow was assessed, and data points were converted from m³/d to L/s unit. The chemical products aluminum sulfate and anionic polyelectrolyte are used in the tertiary treatment of the liquid phase. Data on the chemical products consumption is also registered daily; therefore, their consumption is recorded in kg/d.

Bypass is the intentional diversion of untreated sewage to the receiving water body, mainly during events of extraordinary inflow rates, which occurs primarily during heavy rains (LEE; BECK; BÜRGMANN, 2022; REBELO, 2019). However, the dataset showed that bypass was recorded as zero in all observations. According to Rebelo (2019), Brasília Sul WWTP has four biological reactors and operates three of them. Since 2018, the fourth reactor, which is deactivated, has been used to route excess flow. With this operational procedure, the bypass was eliminated (REBELO, 2019). For this reason, the bypass variable was excluded from the analysis.

3.2.2.2 John E. Egan WWTP

Data from the John E. Egan WWTP were obtained from the MWRD website. Data made available were displayed in spreadsheets that contained monitoring data for each year and each sampling point. The data were then compiled to create a dataset for the complete study period.

The monitoring period selected was from January 2001 to June 2023, and the sampling frequency was mostly daily, with samples collected from the influent and final effluent. In the routine monitoring program of the facility, 24-hour composite samples are collected from midnight to midnight (ZHANG et al., 2008). This large time series of 22.5 years and the high sampling frequency resulted in 8,216 observations.

There were 62 variables collected in the influent and 70 in the final effluent. However, some of these variables had mostly missing data, such as metals. As the objective of including a WWTP from the USA in this study was to have more data to assess comparable to the WWTP from Brazil, it was selected only the variables monitored through the entire study period, and with less than 50% of missing data for all analyses. Although thermotolerant coliform in the effluent had 58% of missing data, it was also included due to its importance, especially because of the disinfection process at John Egan WWTP. Table 7 shows the selected parameters.

Table 7 – Selected monitored parameters in each site at John E. Egan WWTP

Site	Variables
Influent	pH, BOD, CBOD, TS, TSS, TKN, NH_4-N, TP
Effluent	Flow, pH, BOD, CBOD, TSS, TKN, NH_4-N, TP, SP, TTC

pH: potential hydrogen; BOD: biochemical oxygen demand; CBOD: carbonaceous BOD; TS: total solids; TSS: total suspended solids; TKN: total Kjeldahl nitrogen; NH_4-N: ammonia nitrogen; TP: total phosphorus; TTC: thermotolerant coliforms; SP: soluble phosphorus.

3.2.3 Exploratory data analysis

The analyses described next were mostly common for both facilities, but they were conducted and presented separately by each WWTP. The descriptive statistics were calculated for each variable, with values of minimum, 1st quantile, median, mean, 3rd quantile, maximum, number of samples, number of missing data, percentage of missing data, and coefficient of variation (CV, standard deviation divided by the mean). Since some variables in the John E. Egan monitoring dataset had censored values, the percentage of censored data was also presented in the descriptive statistics of this facility.

Considering the samples collected in the treated effluent, compliance with locally applicable environmental standards was assessed for the Brasília Sul WWTP, in which the violation percentages to the maximum permissible limits were calculated. Compliance with locally applicable environmental discharge standards was not assessed for John E. Egan WWTP. The IL0036340 is the National Pollutant Discharge Elimination System (NPDES) permit for this facility, which the Illinois Environmental Protection Agency regulates. However, the discharge standards are not as straightforward to analyze as the ones in Brazil. For example, in the permit, for a single variable, there are standards on both monthly averages and maximum observation limits. There are also standards on both load (kg/day) and concentrations (mg/L) for a single variable. Some standards are variable according to the period of the year. Besides that, as a long-time series was selected for this facility, it is possible that different standards were ruled for different years.

For variables monitored at both the influent and effluent sites, the median and mean removal efficiencies were calculated considering the entire study period. The CV and interquartile range (IQR) measures of dispersion were also calculated for the removal

efficiencies. For the John E. Egan WWTP, the median and mean removal efficiencies were also presented for each year since many years were assessed.

For a deeper understanding of the performance of Brasília Sul WWTP, parameters with less than 30% of missing data, thus resulting in larger samples for both the influent and effluent variables, were selected for further investigation. These variables are sampled every two or three days in Brasília Sul WWTP, allowing temporal variation analysis. For John E. Egan, it was not necessary to select variables for further investigation since they were previously selected (variables with less than 50% of missing data) and are sampled from two times a week to daily. Box-plot and time series graphs were created to visualize the data distribution and temporal variability. In the time series graphs, a smoothing line was added to help visualize the overall temporal trend and patterns. The smoothing line uses the method “loess” (locally weighted polynomial) and represents a curve that best fits the relationship between the variable and time, along with a grey area around the curve that represents the 95% confidence interval.

In addition to the results of the time series graphs, statistical tests were applied to assess the influence of different periods or seasons in the year on the influent, effluent samples, and operational variables. For the Brasília Sul WWTP, the Mann-Whitney non-parametric test was applied at a 5% significant level to determine if there was a significant difference between the dry (April to September) and rainy (October to March) periods. For the John E. Egan WWTP, since seasons of the year are more clearly defined in the region, the Kruskal-Wallis non-parametric test was applied at a 5% significance level to compare the four seasons. If a significant difference was observed, the Dunn non-parametric test was applied at a 5% significance level to determine where the difference was found.

Since the treatment technologies of the two WWTPs under study differ, their performance was not directly compared, and results were presented separately by each facility. The only comparable results were the CVs in the influent and effluent variables and the CV and IQR of the removal efficiencies. The CV indicates the degree of variability of the data (OLIVEIRA; VON SPERLING, 2008) and, in a wastewater treatment plant, its stability (VON SPERLING; VERBYLA; OLIVEIRA, 2020).

The R programming language was used in all stages of data structuring, graph creation, and statistical analyses. It is a free and open-source software for data science (R CORE TEAM, 2020).

3.3 RESULTS AND DISCUSSION

3.3.1 Case study 1: the Brasília Sul wastewater treatment plant

3.3.1.1 Data of influent and effluent samples

The sampling frequency of the monitoring of influent and effluent parameters at Brasília Sul WWTP differed by variable. Table 8 shows the sampling frequency of each variable and Figure 10 the heatmap of the number of samples for each variable in each month and year.

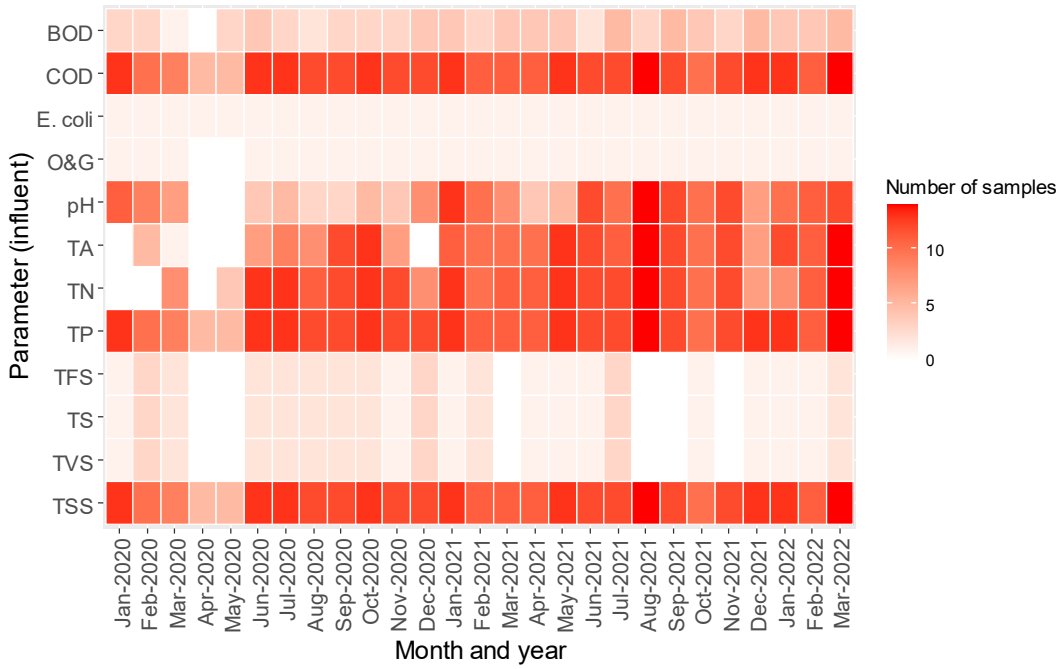
The highest frequency of data collection was every two or three days, and the smaller frequency was monthly. Besides this difference, there were gaps of more than one month with no data collection for some variables in both influent and effluent data. These dataset characteristics are challenging to deal with due to the varying levels of information available. Variables collected less frequently and/or with gaps may lack sufficient data points to capture trends or relationships accurately.

Table 9 and Table 10 present the descriptive statistics for all monitored variables at the influent and effluent sites, respectively. The dataset was structured to input data collected every two or three days. For this reason, a variable collected less frequently had a higher percentage of missing data in the descriptive statistics. This does not necessarily mean missing data since their sampling frequency differs (VON SPERLING; VERBYLA; OLIVEIRA, 2020). For example, *E. coli* is measured once a month (Table 8) and had 92% of missing data (Table 9 and Table 10).

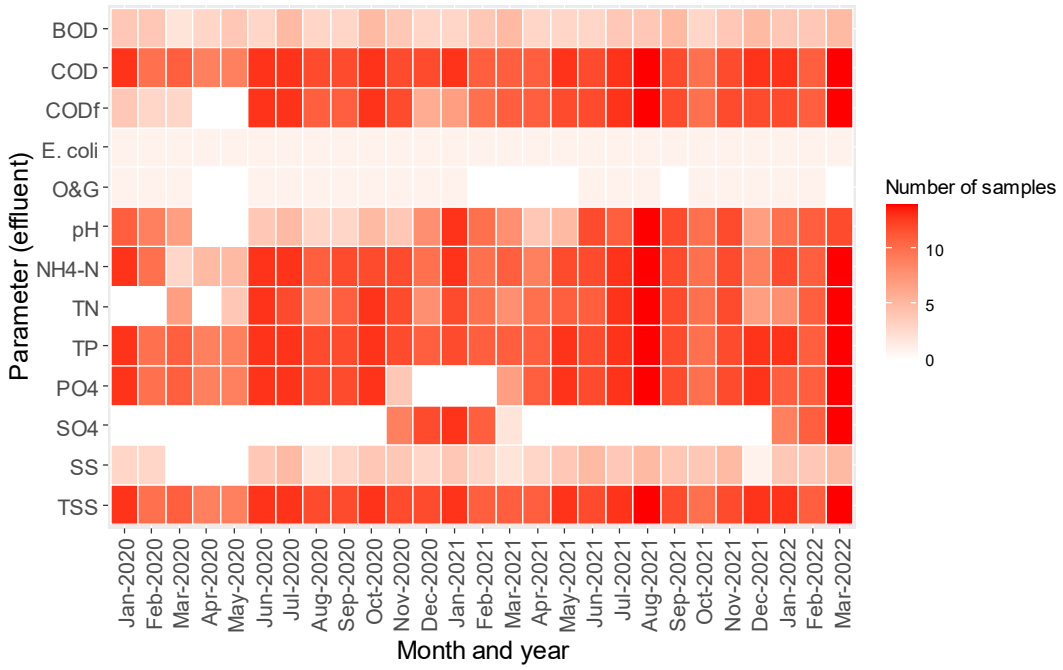
Table 8 – Frequency of monitoring of influent and effluent variables of the Brasília Sul WWTP

Site	Variable	Frequency
Influent	BOD	Weekly or twice a month
	COD	Every two or three days
	E. coli	Monthly
	O&G	Monthly
	pH	Every two or three days or weekly
	TA	Every two or three days (gap for three months)
	TN	Every two or three days (gap for two months)
	TP	Every two or three days
	TFS	Variable, mostly twice a month (with gaps)
	TS	Variable, mostly twice a month (with gaps)
	TVS	Variable, mostly twice a month (with gaps)
	TSS	Every two or three days
Effluent	BOD	Weekly or twice a month
	COD	Every two or three days
	CODf	Every two or three days or weekly
	E. coli	Monthly
	O&G	Monthly (gap for almost five months)
	pH	Every two or three days or weekly (gap for almost three months)
	NH_4-N	Every two or three days
	TN	Every two or three days (gap for two months)
	TP	Every two or three days
	PO_4	Every two or three days or weekly (gap for four months)
	SO_4	Every two or three days (gap for more than ten months)
	SS	Weekly (gap for three months)
	TSS	Every two or three days

Figure 10 – Heatmap of the number of samples by variable and year in the influent (a) and final effluent (b) of the Brasília Sul WWTP



(a)



(b)

Table 9 – Descriptive statistics of the influent variables of the Brasília Sul WWTP

Variable	Abbreviation	Unit	Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum	n	na	% missing	CV
Biochemical oxygen demand	BOD	mg/L	100	282	340	362	400	1219	91	240	73	0.41
Chemical oxygen demand	COD	mg/L	102.8	437.0	506.8	534.5	595.0	1874.4	311	20	6	0.36
Escherichia coli	*E. coli*	MPN/100 mL	3.43E+06	7.07E+06	9.21E+06	1.06E+07	1.36E+07	1.99E+07	27	304	92	0.43
Total oil and grease	O&G	mg/L	4.3	16.3	22.4	30.6	33.1	126.0	25	306	92	0.86
Potential hydrogen	pH	-	6.95	7.34	7.42	7.42	7.52	7.82	209	122	37	0.02
Total alcalinity	TA	mg/L	111.2	169.1	188.2	190.0	210.3	290.5	231	100	30	0.16
Total nitrogen	TN	mg/L	13.0	57.0	63.0	63.2	70.0	150.0	264	67	20	0.20
Total phosphorus	TP	mg/L	2.80	5.33	6.30	6.44	7.25	14.43	311	20	6	0.25
Total fixed solids	TFS	mg/L	120	198	254	262	288	656	35	296	89	0.37
Total solids	TS	mg/L	336	476	548	615	668	2372	35	296	89	0.54
Total volatile solids	TVS	mg/L	110	257	292	353	364	1716	35	296	89	0.73
Total suspended solids	TSS	mg/L	34.0	167.8	204.0	218.3	242.0	1150.0	311	20	6	0.48

Table 10 – Descriptive statistics of the effluent variables of the Brasília Sul WWTP

Variable	Abbreviation	Unit	Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum	n	na	% missing	CV
Biochemical oxygen demand	BOD	mg/L	3	5	8	9	13	35	102	240	70	0.64
Chemical oxygen demand	COD	mg/L	1.8	19.5	23.2	23.9	27.5	106.6	322	20	6	0.38
Filtered chemical oxygen demand	CODf	mg/L	1.2	15.4	19.7	19.4	22.9	54.0	262	80	23	0.34
Escherichia coli	*E. coli*	MPN/100 mL	1.60E+02	1.08E+04	1.73E+04	5.08E+04	4.23E+04	3.65E+05	27	315	92	1.61
Total oil and grease	O&G	mg/L	0.4	1.0	2.5	4.7	6.3	21.9	19	323	94	1.21
Potential hydrogen	pH	-	6.50	7.02	7.14	7.14	7.27	7.64	210	132	39	0.03
Ammonia nitrogen	NH_4-N	mg/L	0.12	0.90	2.41	3.03	4.14	15.90	293	49	14	0.90
Total nitrogen	TN	mg/L	2.4	5.8	7.5	8.1	9.5	41.0	252	90	26	0.49
Total phosphorus	TP	mg/L	0.05	0.11	0.17	0.23	0.24	5.92	320	22	6	1.62
Phosphate	PO_4	mg/L	0.05	0.05	0.05	0.10	0.10	3.00	272	70	20	2.27
Sulfate	SO_4	mg/L	16.94	45.94	54.12	54.09	62.01	94.57	81	261	76	0.26
Sedimentable solids	SS	mL/L	0.05	0.05	0.05	0.23	0.20	1.00	88	254	74	1.54
Total suspended solids	TSS	mg/L	1.1	2.7	3.6	4.5	5.0	72.0	322	20	6	1.05

Raw sewage characteristics may vary according to climate, population habits, per capita water consumption, and the socioeconomic status of the population. von Sperling (2014) reported the typical range of raw domestic sewage in developing countries, and most data of Brasília Sul WWTP (Table 9) had values within the usual range. In a study of other WWTPs in the same region, the data showed results close to those observed in the Brasília Sul WWTP (BARROS, 2013).

Effluent COD, TSS, NH_4-N, and TP concentrations (Table 10) were lower in Brasília Sul WWTP compared to the results of Brasília Norte WWTP in the study of Barros (2013). Both plants employ the same treatment technology. However, it is worth noting that the study of Barros (2013) covered the period from 2000 to 2011, and updated results would be necessary for a more accurate comparison.

The concentrations of *E. coli* in the effluent (Table 10) were similar to those reported by Dantas, Barroso and Oliveira (2021) for activated sludge facilities in Minas Gerais, Brazil. One of the goals defined in the Basic Sanitation Plan of the Federal District was to implement ultraviolet disinfection in Brasília Sul WWTP by 2037 (GDF, 2017b). This step is necessary to meet the bathing water standards in Lake Paranoá (ANTERO, 2020) and to ensure the safety of the water supply, as the lake is also a source of water.

According to Rout et al. (2021), in A²O process, typical effluent concentrations of TN range from 7.3 to 9.0 mg/L and TP from 0.025 to 0.98 mg/L. At the Brasília Sul WWTP, 30% of the samples had TN concentrations higher than 9.0 mg/L. Therefore, more attention should be paid to nitrogen removal at this facility. With regards to TP, 2% of the samples showed concentrations above 0.98 mg/L, which can be attributed to the use of tertiary treatment in Brasília Sul WWTP, leading to higher levels of phosphorus removal.

The effluent TP and TSS concentrations of Brasília Sul WWTP (Table 10) were comparable to those reported by Kwon et al. (2015) at the Jinju WWTP in South Korea. This plant has a capacity of 2,200 L/s and employs the A²O process followed by flocculation and dissolved air flotation. The authors found that the dissolved air flotation step played a crucial role in removing TP but was ineffective in removing TN (KWON et al., 2015).

In Brazil, the National Environment Council (CONAMA) Resolution n. 430/2011 establishes conditions for the disposal of effluents into water bodies at the national level. Considering the parameters assessed in the study, only BOD, pH, sedimentable solids (SS), and total oil and grease (O&G) have legislated standards for sanitary sewage discharge (Table 11). The requirements of CONAMA 430/2011 for municipal wastewater discharge are met for all the samples. The complete compliance is because this legislation is flexible compared to international standards, and Brasília Sul WWTP presents an advanced treatment level compared to other Brazilian treatment systems.

Table 11 – Standards for municipal wastewater discharge established by the CONAMA Resolution n. 430/2011

Variable	Standard
BOD	Maximum of 120 mg/L or removal efficiency of at least 60%
pH	Between 5.0 and 9.0
SS	Maximum of 1 mL/L (for discharges into lakes, the sedimentable materials must be virtually absent)
O&G	Maximum of 100 mg/L

Another Brazilian legal requirement is the granting of rights to use water resources, including for the discharge of effluents. These standards are defined based on the characteristics of the receiving water body and the effluent, the water body's uses, and the presence of abstractions and other discharges. The Resolution n. 11/2016 of the Regulatory Agency for Water, Energy, and Basic Sanitation of the Federal District of Brazil (Adasa, in Portuguese Agência Reguladora de Águas, Energia e Saneamento Básico do Distrito Federal) outlines the criteria for the concession of Brasília Sul WWTP (ADASA, 2020). Table 12 presents the limits for each variable and the violation percentage during the study period.

Table 12 – Standards for the discharge of effluents from Brasília Sul WWTP established by the Adasa Resolution n. 11/2016, and the violation percentages

Variable	Limit (mg/L)	Violation to maximum permissible limits (%)
BOD	27.8	1
TP	0.3	15
TN	8.7	37

The granting concession criteria for the Brasília Sul facility are more stringent than those specified in CONAMA Resolution n. 430/2011 regarding BOD concentrations and include standards for TP and TN, which the CONAMA Resolution does not require (Table 11 and Table 12). This is because the effluent from Brasília Sul WWTP is

discharged into a lentic water body, and the risk of eutrophication is considered when granting rights to use water resources (ADASA, 2021).

During the study period from January 2020 to March 2022, only 1% of the samples exceeded the limit of 27.8 mg/L for BOD in the effluent. A previous study conducted by Antero (2020) from January 2017 to September 2020, found no violations of the BOD discharge limit of Brasília Sul WWTP. This indicates that the A²O activated sludge technology with final polishing used in the facility effectively removes organic matter.

Regarding TP, 15% of the samples demonstrated higher concentrations than the discharge limit. The biological phosphorus removal in the A²O activated sludge process is achieved due to the presence of anaerobic and aerobic zones in the treatment line.

Phosphorus accumulating organisms (PAOs) are bacteria capable of accumulating excess amounts of phosphorus as polyphosphates. In the anaerobic zone, a portion of the biodegradable organic matter (soluble BOD) is converted, through fermentation processes, into simple organic molecules of low molecular weight, such as volatile fatty acids. These volatile fatty acids become available in the liquid medium and are preferred by the PAOs, which rapidly assimilate and store them inside their cells (VON SPERLING, 2007). PAOs assimilate these fermentation products better than the other organisms commonly found in the activated sludge process, leading to a selection of the PAO population in the anaerobic zone. The energy required for substrate transport and for the formation and storage of organic metabolic products, such as polyhydroxybutyrate (PHB), is supplied by the release of phosphate that was previously accumulated by the organisms in the aerobic stage. In the aerobic zone, PHB is oxidized into carbon dioxide and water, and the soluble phosphate is removed from the solution and stored in the PAO cells. This leads to an increase in the PAO population (VON SPERLING, 2007). This process is known as “luxury uptake of phosphorus” or “enhanced biological phosphorus removal” (EBPR) (ROUT et al., 2021).

As PAOs incorporate large amounts of phosphorus from the liquid medium into their cells, the removal of phosphorus from the system is accomplished by the removal of excess biological sludge (VON SPERLING, 2007). The alternating of anaerobic and aerobic conditions required by PAOs is achieved by recirculating a fraction of sludge

from the secondary settlers to the anaerobic zone of the biological reactors (HAN et al., 2021; VON SPERLING, 2007).

Several factors influence the biological removal of phosphorus, including environmental conditions (dissolved oxygen, temperature, pH, nitrate levels in the anaerobic zone), design parameters (sludge age, detention time and configuration of the anaerobic zone, detention time in the aerobic zone, and excess sludge treatment methods), and the characteristics of the influent sewage, and suspended solids in the effluent (VON SPERLING, 2007).

In addition to biological nutrient removal, the Brasília Sul WWTP also employs coagulation, flocculation, and dissolved air flotation for phosphorus removal. In coagulation and flocculation steps, the particles are destabilized to form flocs, allowing for efficient collisions and aggregation between bubbles and flocs to occur in the flotation unit (PENETRA, 2003).

The dissolved air flotation process is a separation technique that uses air bubbles as the transport medium. In the process, a portion of the treated liquid is recirculated, pressurized, and saturated with air in a saturation chamber. The recirculated liquid is introduced into the flotation chamber through a series of injectors and mixed with the flocculated incoming effluent. At the injectors, the pressure is reduced to atmospheric pressure, causing the air to be released in the form of micro-bubbles. The air bubbles attach to the flocs, causing them to rise to the surface of the flotation chamber and form the chemical sludge (PENETRA, 2003), which is removed through surface scrapers at Brasília Sul WWTP. The efficiency of dissolved air flotation is influenced by various factors such as coagulation and flocculation conditions (e.g. the type and dosage of coagulant used, mixing conditions, and pH), the amount of air supplied, the diameter of the air bubbles diameter (usually between 10 and 120 μm), and the removal of chemical sludge (PENETRA, 2003).

With regards to TN, 37% of the samples showed concentrations that exceeded the discharge limit (Table 12). In the A²O activated sludge process, nitrogen removal is accomplished through a two-step process that involves nitrification followed by denitrification. Nitrification involves the oxidation of ammonia to nitrite and then to nitrate. The nitrification takes place at the aerobic (use of oxygen as electron

acceptors) zone of the reactor (VON SPERLING, 2007). Denitrification involves the reduction of nitrates to nitrite and then to gaseous nitrogen, which escapes to the atmosphere. This step takes place at the anoxic zone of the reactor. In anoxic conditions (absence of oxygen but presence of nitrates), the nitrates are used by facultative heterotrophic microorganisms as electron acceptors (ROUT et al., 2021; VON SPERLING, 2007). These microorganisms require a source of organic carbon for denitrification, which is supplied by the wastewater. The nitrates are directed to the anoxic zone through internal recirculation from the aerobic zone to the anoxic zone (VON SPERLING, 2007).

In warm-climate conditions like Brazil, nitrification occurs almost systematically in activated sludge plants. However, various factors such as sludge age, temperature, pH, dissolved oxygen, and the presence of toxic or inhibiting substances can impact nitrification (VON SPERLING, 2007). While denitrifying bacteria are relatively less sensitive to environmental conditions, the rate of denitrification can still be influenced by dissolved oxygen, temperature, pH, and toxic or inhibiting substances (VON SPERLING, 2007). Further investigation into the impact of these factors on nitrogen removal at Brasília Sul WWTP is necessary for a better understanding of the process.

As shown in Table 5, not all parameters are sampled at both the influent and effluent sampling sites. Figure 11 shows influent and effluent values for the parameters that are monitored at both locations, and Table 13 presents the median, mean, CV, and IQR of the removal efficiencies for these parameters (except pH) considering the entire study period.

Figure 11 – Influent and effluent values of Brasília Sul WWTP

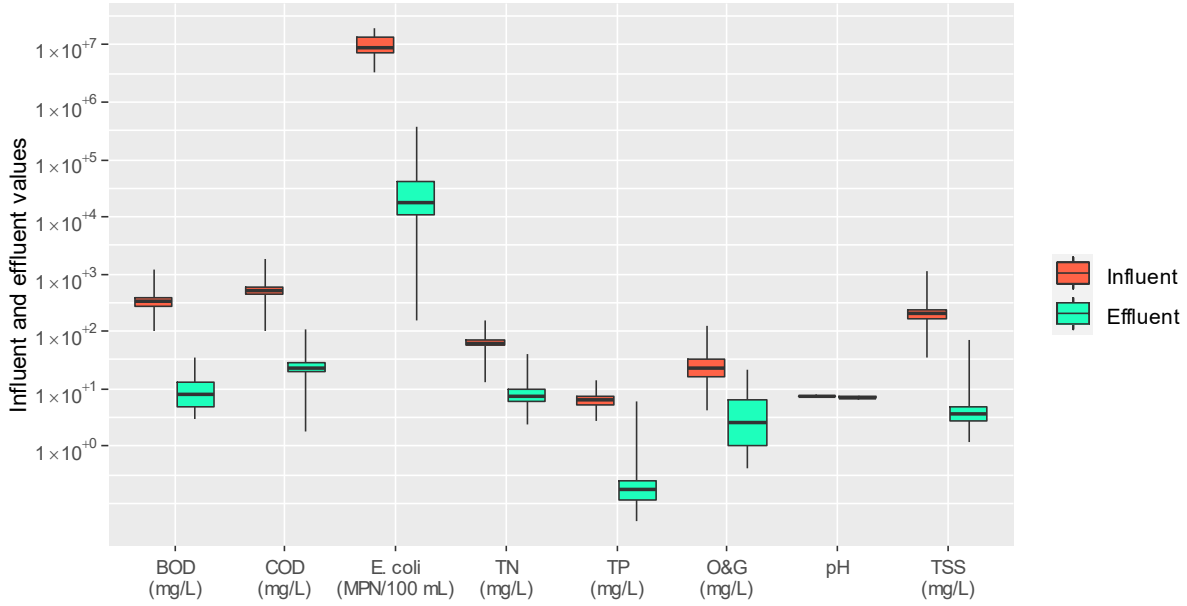


Table 13 – Median, mean, CV and IQR of the removal efficiencies of Brasília Sul WWTP, considering the entire study period

Variable	Median removal efficiencies (%)	Mean removal efficiencies (%)	CV	IQR
BOD	97.9	97.3	0.02	1.80
COD	95.4	95.0	0.04	1.83
E. coli	99.814 (log removal of 2.73)	99.447 (log removal of 2.26)	0.01	0.53
TN	87.7	86.8	0.09	6.55
TP	97.3	96.1	0.08	2.13
O&G	84.5	78.4	0.30	30.28
TSS	98.3	97.0	0.12	1.49

The removal efficiencies are considerably high (Table 13) when compared to other WWTPs in Brazil of various sizes and technologies (DANTAS, 2020; SILVA, 2020). This confirms the higher treatment level and operational control of Brasília Sul WWTP.

The median removal efficiencies for BOD, TSS, TP, and *E. coli* were comparable to those found by Barros (2013) for the Brasília Norte WWTP, which also operates in the Federal District and employs the same treatment technology. The mean removal efficiencies for COD, TP, and TN were higher in Brasília Sul WWTP than those reported by Zhang et al. (2011) in four A²O activated sludge facilities in China. According to the authors, the low nutrient removal efficiencies were due to low nutrient levels in the influent of these Chinese WWTPs (ZHANG et al., 2011).

There were only two instances of negative removal efficiencies in the study period for Brasília Sul WWTP, which were -111.8% for TSS and -18.4% for TP. These negative values were recorded on the same date (March 12, 2020), when other parameters also showed their lowest removal efficiencies in the period (49.0% for COD and 6.8% for TN).

Antero (2020) assessed data from August 2017 to October 2020 of Brasília Sul WWTP and found global mean removal efficiencies in the period (Table 14) close to the efficiencies found in this work. The TN removal efficiencies were higher in this work. This could mean that there could have been recent operational improvements since this study period (January 2020 to March 2022) is more recent.

Antero (2020) had access to samples collected at more points in the system, including influent, effluent from the primary treatment, effluent from the secondary treatment, and final effluent. Thus, the mean removal efficiencies of each treatment step from August 2017 to October 2020 are shown in Table 14. The importance of the final polishing step for removing the remaining COD, TSS, and TP is highlighted, and TN removal occurs mainly during the biological treatment.

Table 14 – Mean removal efficiencies of COD, TSS, TN, and TP in the period August 2017 to October 2020 of Brasília Sul WWTP

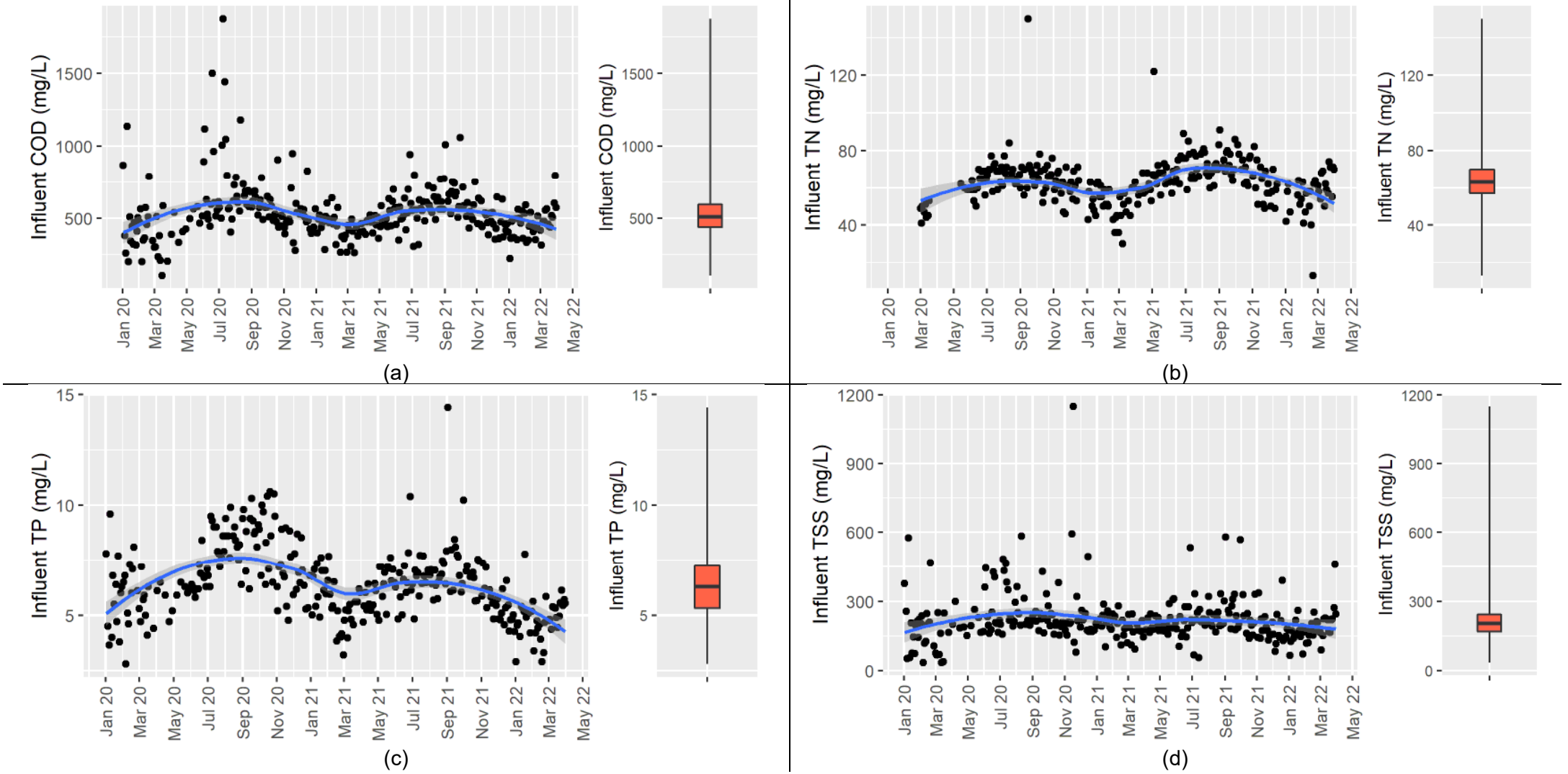
	Removal efficiencies (%)			
Treatment step	**COD**	**TSS**	**TN**	**TP**
Primary	36	61	20	24
Secondary	70	28	66	53
Final polishing	74	89	11	89
Global	95	97	76	96

Source: Adapted from Antero (2020)

Although various parameters are monitored at Brasília Sul WWTP, the number of samples is highly variable depending on the parameter, as indicated in Table 9 and Table 10. This is due to gaps in the data and differences in sampling frequencies for each parameter. The next analyses were conducted for variables with less than 30% of missing data for both influent and effluent concentrations, allowing temporal variation analysis.

Figure 12 shows the time series and box-plot graphs of influent concentrations of COD, TN, TP, and TSS. For all four variables, there is variability in the influent concentrations throughout the period, with higher values during the dry season (April to September) and lower values during the rainy season (October to March), particularly for COD, TN, and TP.

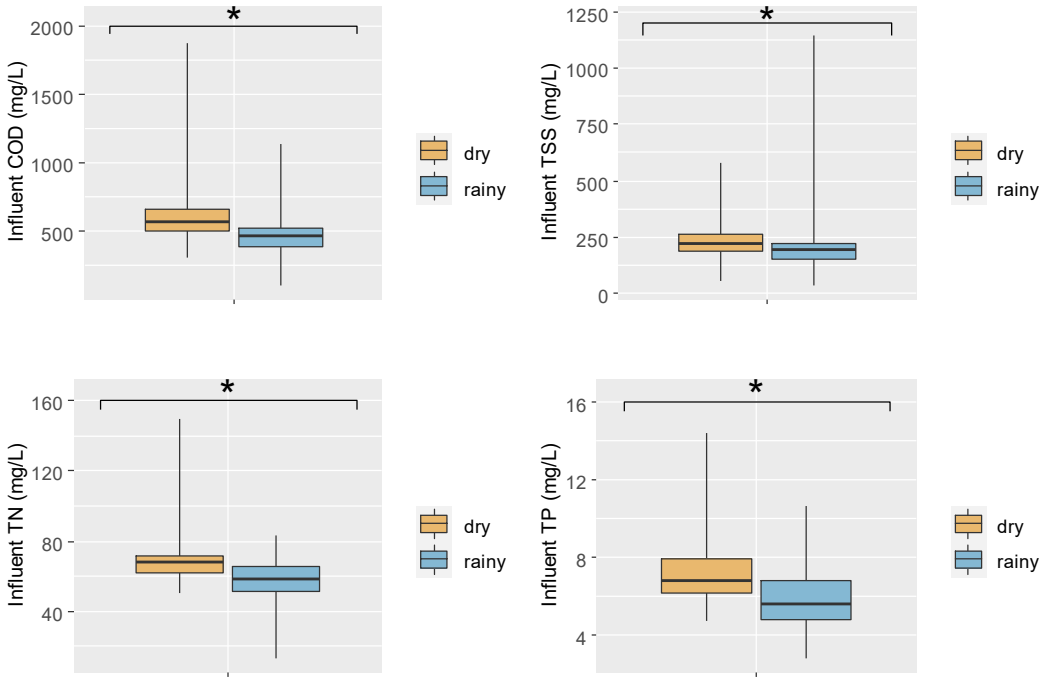
Figure 12 – Time series and box-plot graphs of influent COD (a), TN (b), TP (c), and TSS (d) of Brasília Sul WWTP. The blue line is the smoothing line of the trend in the data, and the grey area is the 95% confidence region for the fit



The results of the Mann-Whitney test showed a significant ($p < 0.05$) difference between the dry and rainy periods for all four influent parameters, with higher values observed during the dry period (Figure 13). This is consistent with the findings of Barros (2013), who also found significantly higher concentrations of raw wastewater during the dry period compared with the rainy period in three other WWTPs located in the Federal District.

The sewage collection systems in Brazil are designed as absolute separator systems, meaning that sewage and rainwater should be collected, transported, and treated separately. In practice, this does not always occur; there are connections between sewage and rainwater systems, mainly through clandestine connections or accidental interceptions (OLIVEIRA; SOARES; HOLANDA, 2020). This situation also occurs in Brasília municipality (ADASA, 2016). Therefore, the rainfall contributions may have led to lower concentrations during the rainy period.

Figure 13 – Influent concentrations during the dry and rainy periods in Brasília Sul WWTP and the result of the Mann-Whitney test



* Significant ($p<0.05$ in Mann-Whitney test)

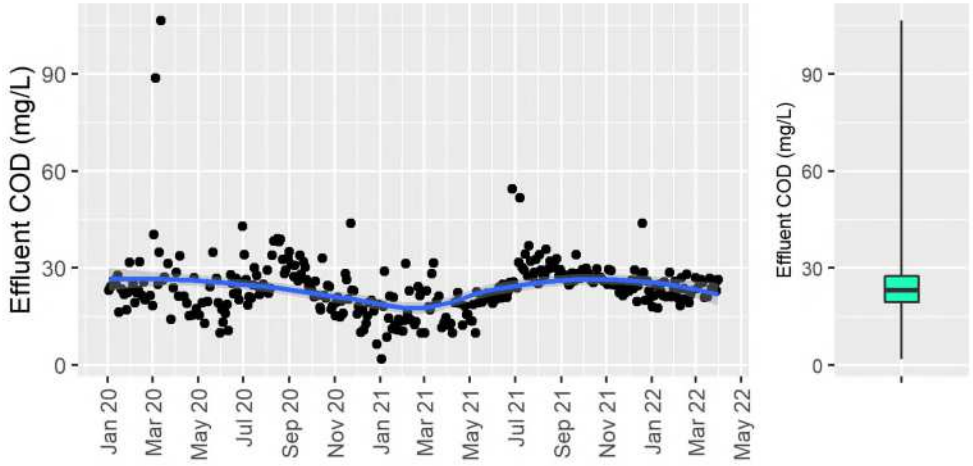
Figure 14 and Figure 15 show the time series and box-plot graphs of effluent concentrations of COD, CODf, NH_4-N, and TN, and effluent concentrations of TP, PO_4, and TSS, respectively.

The highest effluent concentrations of COD, TN, TP, PO_4, and TSS were recorded on March 12, 2020, when negative removal efficiencies for TP and TSS were observed. These outliers are evident in the time series graphs (Figure 14 and Figure 15) and confirm that there was some operational issue on this date. CODf and NH_4-N were not monitored on March 12, 2020, so their maximum values were recorded on other dates.

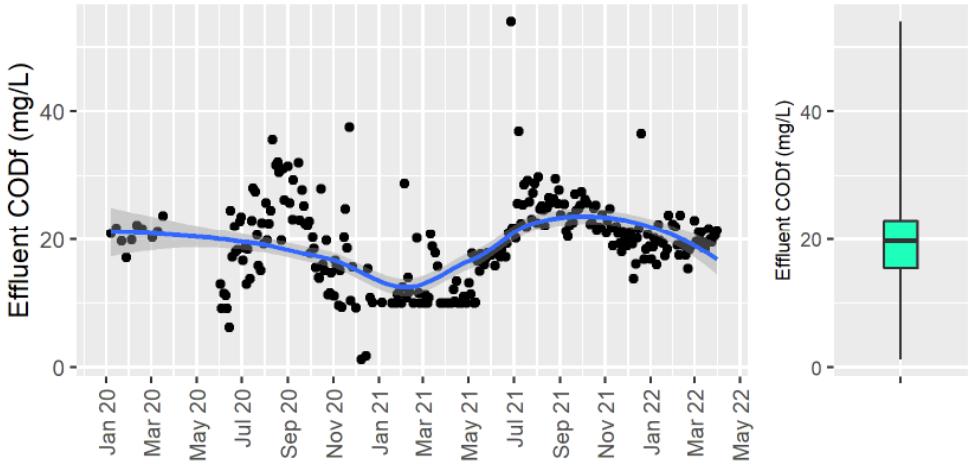
While effluent concentrations exhibited temporal variability (Figure 14 and Figure 15), potentially due to factors such as seasonality and operational conditions at the WWTP, it is not ideal for the effluent quality to be variable. A WWTP is expected to perform in a stable and reliable manner despite the challenge of dynamic environmental and operational conditions (VON SPERLING; VERBYLA; OLIVEIRA, 2020). Regarding effluent TN, which had the highest violation percentage according to discharge standards in this work, Portela (2018) found great variability in nitrogen removal at Brasília Sul WWTP from 2016 to 2018, probably related to the instability of the system. There was also a noticeable variability in TN effluent concentrations in this work; however, there was a general trend of reduction over the years (Figure 14 (d)).

In Figure 16, the results of the Mann-Whitney test comparing effluent parameters between the rainy and dry periods are presented. The test revealed significant difference ($p < 0.05$) in COD, CODf, PO_4, and TP, with higher values observed during the dry period. As mentioned, the influence of seasonality on treated effluent quality was not expected. However, it is worth mentioning that the frequency at which data is collected (every two or three days) increases the likelihood of the statistical test to detect a significant difference.

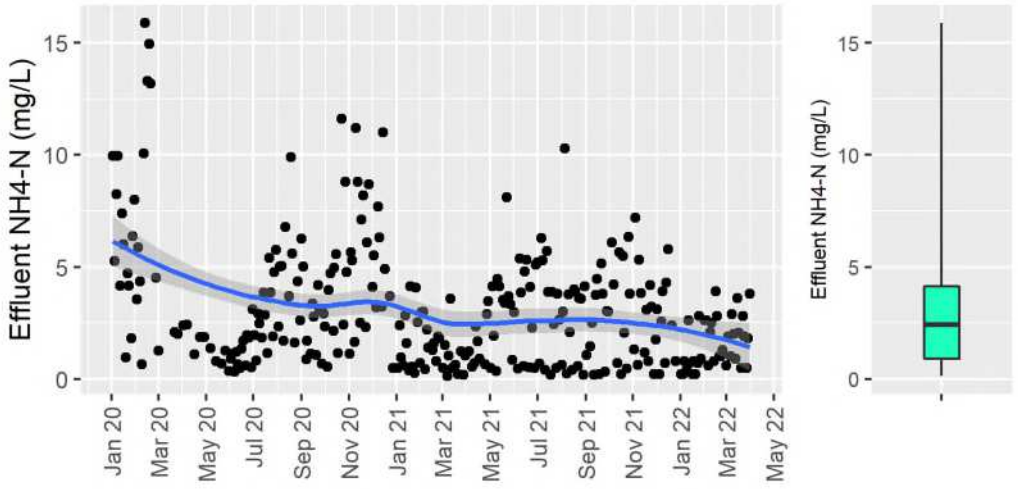
Figure 14 – Time series and box-plot graphs of effluent COD (a), CODf (b), NH_4-N (c), and TN (d) of Brasília Sul WWTP. The blue line is the smoothing line of the trend in the data, and the grey area is the 95% confidence region for the fit



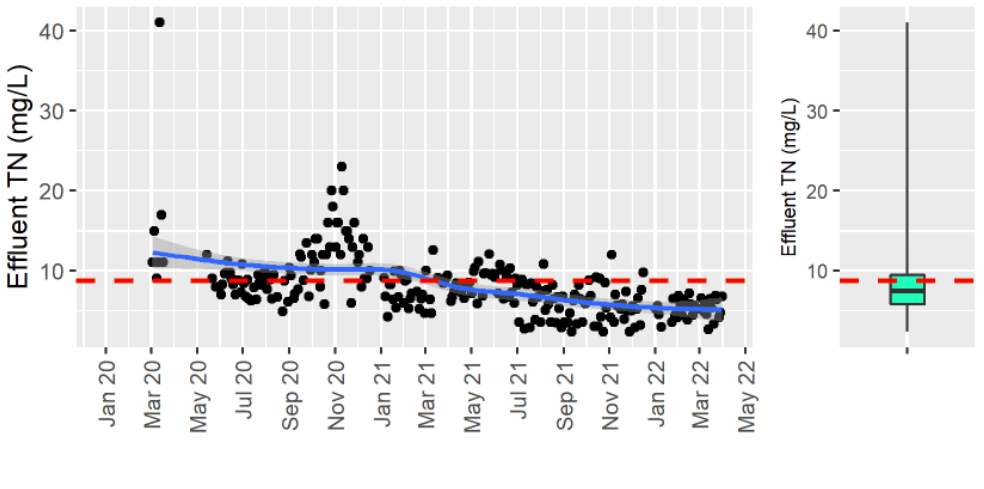
(a)



(b)



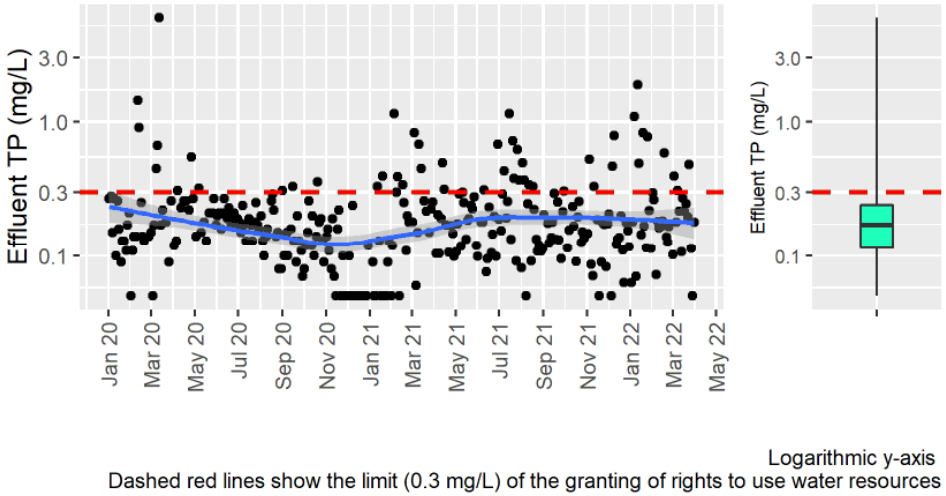
(c)



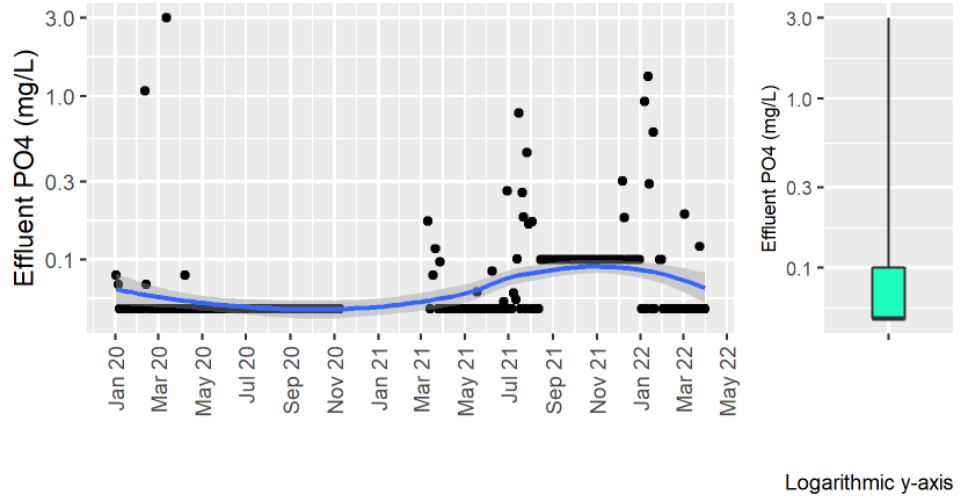
Dashed red lines show the limit (8.7 mg/L) of the granting of rights to use water resources

(d)

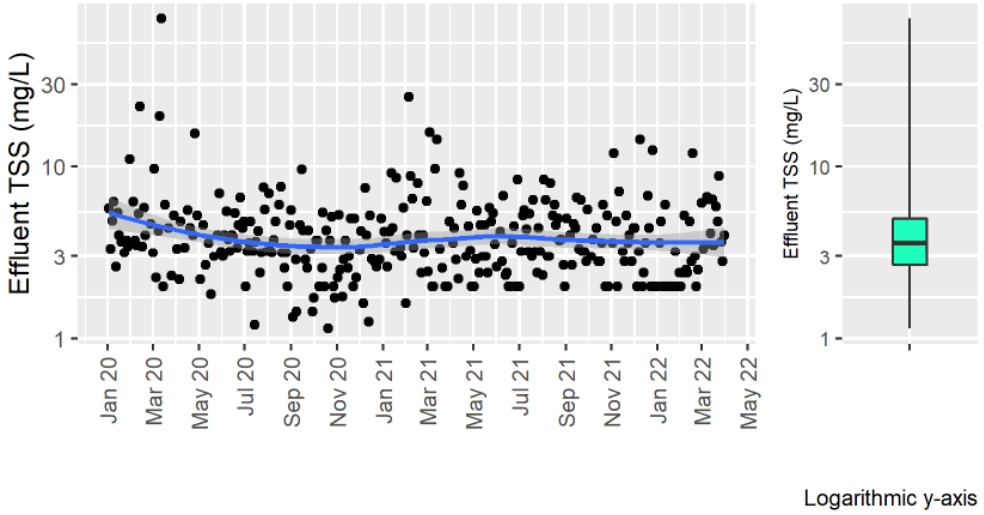
Figure 15 – Time series and box-plot graphs of effluent TP (a), PO_4 (b), and TSS (c) of Brasília Sul WWTP. The blue line is the smoothing line of the trend in the data, and the grey area is the 95% confidence region for the fit



(a)

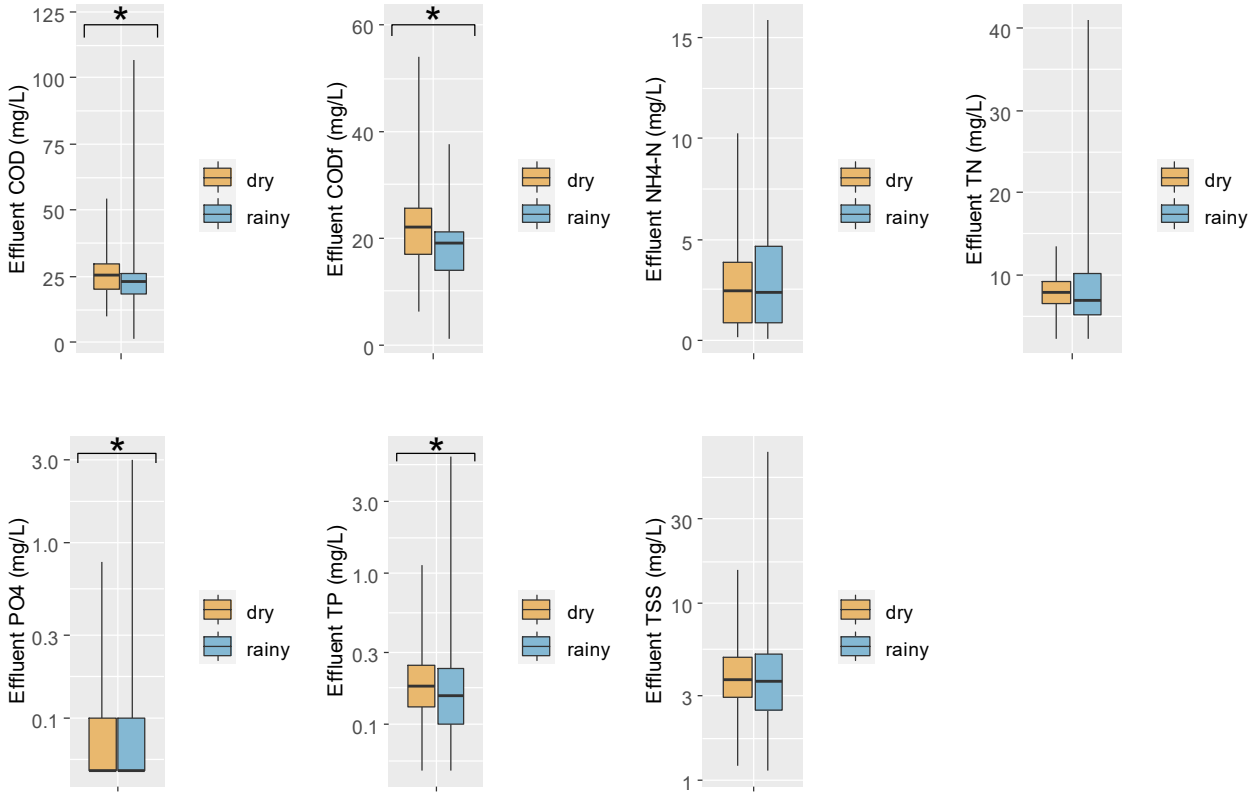


(b)



(c)

Figure 16 – Effluent concentrations during the dry and rainy periods in Brasília Sul WWTP and the result of the Mann-Whitney test



* Significant (p<0.05) in Mann-Whitney test
Logarithm y-axis in PO4, TP, and TSS graphs

3.3.1.2 Data of operational variables

Table 15 presents the descriptive statistics for the operational variables of the Brasília Sul WWTP. Upon initial inspection of the dataset, an extreme outlier of 4,180 kg/d was found in the consumption of anionic polyelectrolyte. This observation was recorded on February 29, 2020, the last day of the month, and all daily records for that month were reported as zero. It was concluded that 4,180 kg/d represented the monthly sum of the daily consumption of this chemical product in February 2020, which is consistent with the average monthly consumption of 3,887 kg/d over the study period. For this reason, all observations of this variable for February 2020, including the outlier, were excluded (replaced by missing data) as it was not possible to estimate the daily consumption of the anionic polyelectrolyte in this month. The descriptive statistics (Table 15) were calculated after this initial screening and preprocessing of the dataset.

Table 15 – Descriptive statistics of the operational variables of the Brasília Sul WWTP

Variable	Unit	Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum	n
Influent flow	L/s	804	1,131	1,248	1,275	1,378	2,342	821
Aluminum sulfate	kg/d	0	7,800	10,400	11,149	14,300	33,150	821
Anionic polyelectrolyte	kg/d	0	80	120	127	160	480	792

Figure 17 shows the time series and box-plot graphs of the influent flow, aluminum sulfate consumption, and anionic polyelectrolyte consumption of Brasília Sul WWTP. The granting of rights to use water resources of Brasília Sul defines the maximum limit of 1,500 L/s for the flow (ADASA, 2020), which is equivalent to the treatment capacity of the plant. The average flow (1,275 L/s, Table 15) from January 2020 to March 2022 was below Brasília Sul WWTP's design flow. The limit is presented as dashed red lines in Figure 17 (a). During the study period, 12% of the data points showed inflow rates above the limit of the granting criteria. However, as mentioned before, there was no bypass when the influent flow exceeded the WWTP's design flow.

The influence of seasonality in the influent flow of the Brasília Sul WWTP is evident, with higher values during the rainy season and lower values during the dry season (Figure 17 (a)). This influence explains the significant difference in influent wastewater quality between the dry and rainy periods (Figure 13). Since there is no flow equalization system at the Brasília Sul WWTP, there is significant fluctuation in the flow rates, negatively affecting the system.

In the study of the Basic Sanitation Plan of the Federal District (GDF, 2017c), the hourly influent flow in 2015 was assessed. The study found that the average influent flow (1,330 L/s in 2015) was below Brasília Sul WWTP's design flow (1,500 L/s) but concluded that the capacity of the plant was systematically exceeded by analyzing the hourly flows (GDF, 2017c).

Araujo and de Jesus (2018) investigated the influence of rainy days on the hourly influent flow of Brasília Sul WWTP and found that there is an expressive increase in the influent flow during rainy events. The authors also studied the relationship between the monthly treated volume of wastewater and precipitation in the region from 2012 to 2016. They confirmed that the treated volumes were higher in the rainy periods compared to the dry months (ARAUJO; DE JESUS, 2018).

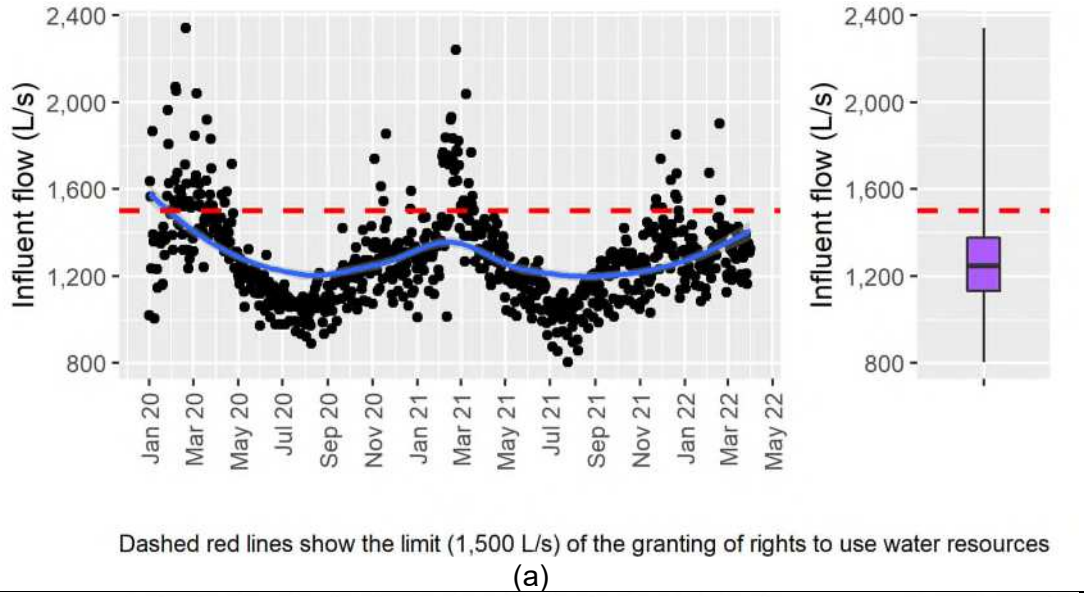
High variability was observed in the consumption of aluminum sulfate (Figure 17 (b)) and anionic polyelectrolyte (Figure 17 (c)), which may be linked to operational conditions such as dosage control.

According to Rebelo (2019), aluminum sulfate is applied in Brasília Sul WWTP prior to tertiary treatment. The dosage is done manually by an operator who must monitor the orthophosphate and pH levels in the final effluent and decide to increase or decrease the dosage according to the result of the measured parameters.

The consumption of aluminum sulfate in Brasília Sul WWTP is high (Figure 17 (b)), with an average daily consumption of 11,149 kg/d (Table 15). According to Rebelo (2019), the Caesb company investigated the possible reasons for the high consumption in this plant. The investigation was based on data from 2018, and the average daily sulfate consumption was 9,753 L/day that year. The high consumption was mainly due to problems in the secondary and tertiary treatment steps, such as out-of-operation blowers and maintenance on final polishing chambers; difficulties inherent to manual dosing and decision-making about when to interfere in the dosing, submitting to subjective criteria of operators' perception; and increased influent load and flow. The investigation concluded with the importance of dosing automation as one of the main factors that could reduce aluminum sulfate consumption and cost (REBELO, 2019).

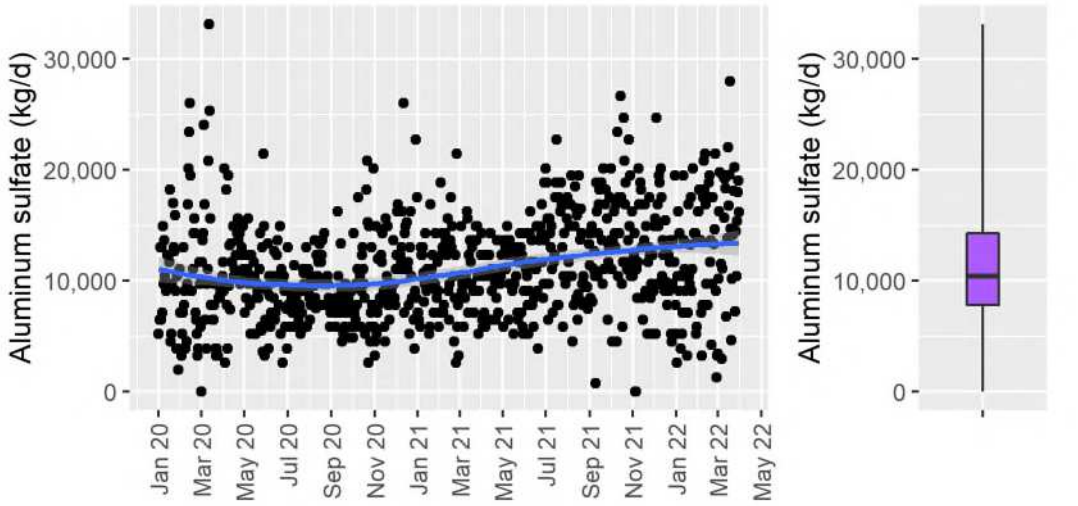
According to Rebelo (2019), the anionic polyelectrolyte is used as a flocculant in the final polishing steps of the Brasília Sul WWTP, agglutinating particles from the secondary treatment. The dosage of this polymer is done automatically using the concentration of the solution as the input parameter in the dosing equipment. The dosage of this polymer has been almost uniform since the automatic dosing equipment was installed, only increasing in cases of very high influent loads (REBELO, 2019). Rebelo (2019) found that among the products studied (aluminum sulfate, cationic polyelectrolyte, anionic polyelectrolyte, and lime), the consumption of anionic polyelectrolyte was the most stable, as it has an automatic dosage system that only changes in exceptional cases of high influent loads (REBELO, 2019).

Figure 17 – Time series and box-plot graphs of the influent flow (a), aluminum sulfate consumption (b), and anionic polyelectrolyte consumption (c) of Brasília Sul WWTP. The blue line is the smoothing line of the trend in the data

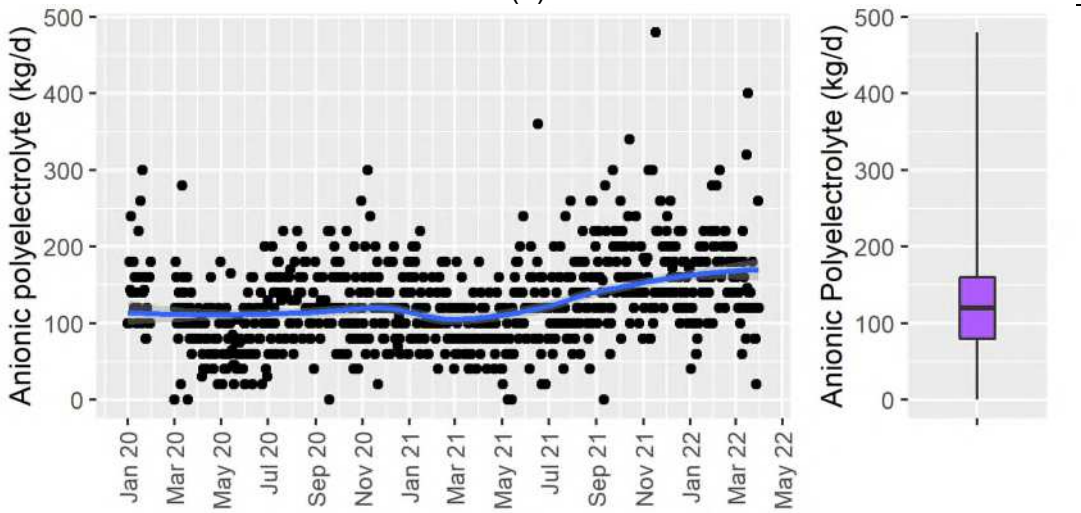


Dashed red lines show the limit (1,500 L/s) of the granting of rights to use water resources

(a)



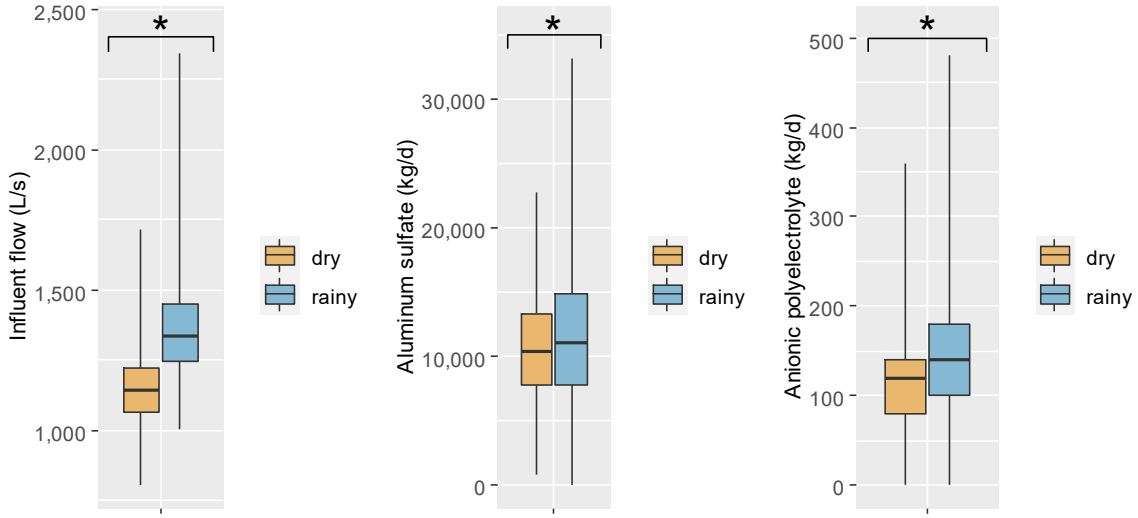
(b)



(c)

The results of the Mann-Whitney test showed a significant difference ($p < 0.05$) in the influent flow and the consumption of the chemical products, aluminum sulfate and anionic polyelectrolyte, between the rainy and dry periods (Figure 18). Araujo and de Jesus (2018) conducted an analysis of the period from 2012 to 2016 and found, through graphical analysis, a difference in the influent flow of the Brasília Sul WWTP between the dry and rainy periods. The authors also observed a rise in the consumption of aluminum sulfate and anionic polyelectrolyte as the influent flow increased (ARAUJO; DE JESUS, 2018).

Figure 18 – Operational variables during the dry and rainy periods in Brasília Sul WWTP and the result of the Mann-Whitney test



* Significant (p<0.05 in Mann-Whitney test)

3.3.2 Case study 2: the John E. Egan wastewater treatment plant

Table 16 and Table 17 show the descriptive statistics of the influent and effluent variables of the John E. Egan WWTP, respectively.

Table 16 – Descriptive statistics of the influent variables of the John Egan WWTP

Variable	Abbreviation	Unit	Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum	n	na	% missing	% censored	CV
Potential hydrogen	pH	-	6.3	7.2	7.3	7.3	7.4	7.8	8212	4	0	0	0.02
Biochemical oxygen demand	BOD	mg/L	24	139	169	178	205	3039	7222	994	12	0	0.51
Carbonaceous BOD	CBOD	mg/L	10	91	110	113	130	1341	6897	1319	16	0	0.40
Total solids	TS	mg/L	358	780	852	885	948	6771	8165	51	1	0	0.23
Total suspended solids	TSS	mg/L	11	136	168	195	217	6050	8192	24	0	0	0.84
Total Kjeldahl nitrogen	TKN	mg/L	5.6	24.2	28.9	28.8	32.7	189.6	7811	405	5	0	0.29
Ammonia nitrogen	NH_4-N	mg/L	2.2	13.3	16.2	16.0	18.8	42.1	7995	221	3	0	0.26
Total phosphorus	TP	mg/L	1.2	4.9	5.9	6.1	6.9	55.7	7811	405	5	0	0.38

Table 17 – Descriptive statistics of the effluent variables of the John Egan WWTP

Variable	Abbreviation	Unit	Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum	n	na	% missing	% censored	CV
Flow	-	L/s	618	903	1012	1092	1174	2957	8216	0	0	0	0.27
Potential hydrogen	pH	-	6.4	7	7.1	7.1	7.2	7.9	4917	3299	40	0	0.02
Biochemical oxygen demand	BOD	mg/L	0	2	2	3	3	26	5025	3191	39	60	0.59
Carbonaceous BOD	CBOD	mg/L	0	2	2	2	2	14	5028	3188	39	83	0.28
Total suspended solids	TSS	mg/L	2	2	2	3	3	42	4867	3349	41	34	0.56
Total Kjeldahl nitrogen	TKN	mg/L	0.1	1.0	1.3	1.5	1.8	6.3	4821	3395	41	4	0.45
Ammonia nitrogen	NH_4-N	mg/L	0.0	0.1	0.1	0.2	0.2	5.4	4750	3466	42	32	1.62
Total phosphorus	TP	mg/L	0.0	2.0	3.0	2.8	3.6	10.1	4822	3394	41	0	0.45
Soluble phosphorus	SP	mg/L	0.0	1.9	2.8	2.6	3.4	10.0	4653	3563	43	0	0.46
Thermotolerant coliforms	TTC	CNTS/100 ml	9	10	10	640	20	3.00E+05	3437	4779	58	52	9.11

Considering influent variables commonly monitored at both the Brasília Sul and John Egan WWTPs (pH, BOD, TSS, TP), as well as similar variables (TN versus TKN and NH_4-N and *E. coli* versus thermotolerant coliforms), their CV values were higher for all variables at the John Egan facility, except for pH, which was the same for both facilities (Table 9 and Table 16). On the other hand, for effluent variables commonly monitored at both WWTPs (BOD, TSS, NH_4-N, and TP), the CVs were higher at the Brasília Sul WWTP, except for NH_4-N, which was higher at the Egan facility (Table 10 and Table 17). This indicates that, although experiencing greater variability in the influent conditions, the performance of John Egan WWTP is generally more stable.

Although samples were collected mostly daily, the monitoring frequency was variable according to the variable and year, which resulted in different numbers of samples, as shown in Figure 19.

Figure 19 – Heatmap of the number of samples by variable and year in the influent (a) and final effluent (b) of the John Egan WWTP

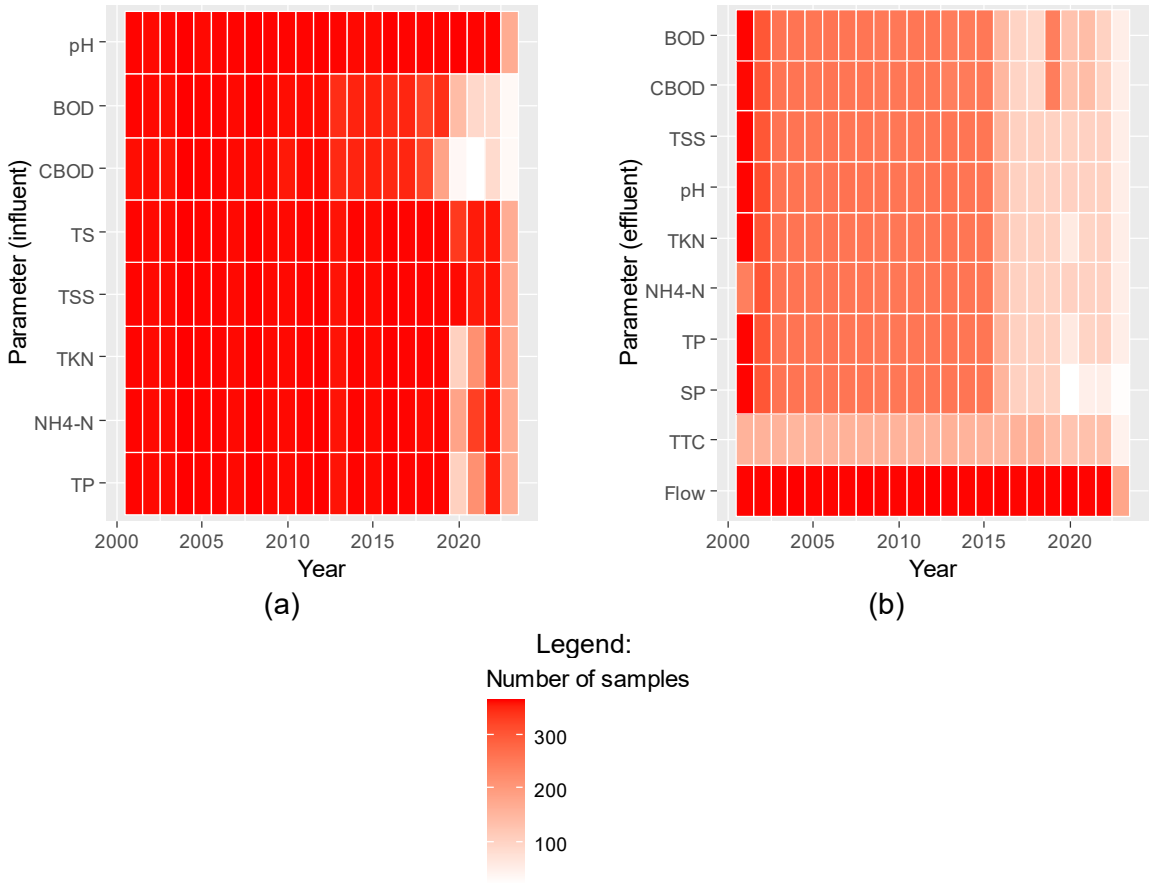


Table 18 shows the monitoring frequency by variable measured at the influent and final effluent by period throughout the time series. For the influent, samples were collected daily until March 2020, when some parameters started to be measured twice a week to weekly. This could be attributed to the COVID-19 pandemic, likely due to challenges in sample collection and analysis in the laboratory, as well as potential flexibility from the environmental agency during this period.

For the effluent, samples were collected daily until May 2016, when samples started to be measured mainly twice a week or weekly. This could mean that there was a change in the sampling protocol. Regarding thermotolerant coliforms, samples are collected five times a week from May to October, which are the months in which John Egan WWTP undergoes the disinfection process, and weekly from November to April, when disinfection is not conducted.

Table 18 – Monitoring frequency by variable and period at John Egan WWTP

		Jan 2001 - Apr 2001	May 2001 - Apr 2016	May 2016 - Feb 2020	Mar-20	Apr 2020 - Jan 2021	Feb 2021 - May 2021	Jul-21	Aug 2021 - Jun 2023
Influent	pH								
	BOD								
	CBOD								
	TS								
	TSS								
	TKN								
	NH_4-N								
	TP								
Effluent	Flow								
	pH								
	BOD								
	CBOD								
	TSS								
	TKN								
	NH_4-N								
	TP								
	SP								

* Obs.: Effluent TTC had weekly monitoring from November to April (with gaps from 2019 to 2023) and five times a week from May to October throughout the time series

Legend:

- Daily/five times a week
- Two times a week
- Weekly
- Interrupted

Figure 20 shows the influent and effluent values of the variables that are sampled at both locations. Table 19 presents the median and mean removal efficiencies for these parameters (except pH) considering each year and the entire study period.

The removal efficiencies for BOD, CBOD, TKN, NH_4-N, and TSS were high for the entire study period. However, the removal efficiencies of TP were low as there is no specific strategy for phosphorus removal at John Egan WWTP.

Nitrate was not included in this study since its monitoring was interrupted in November 2015. However, it is important to mention that all nitrate removal efficiencies were negative. This highlights that this facility is designed for the removal of ammonia (hence high removal efficiencies - Table 19) through nitrification, but there is no additional denitrification step for removing nitrate and, with that, total nitrogen.

Figure 20 – Influent and effluent values of John Egan WWTP

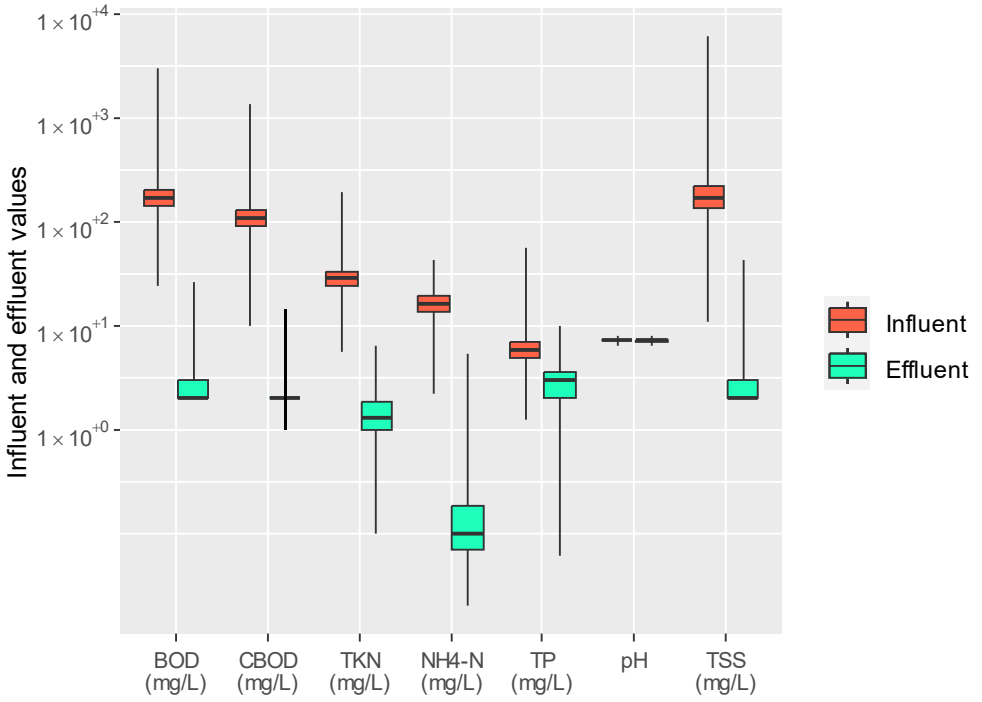


Table 19 – Median and mean removal efficiencies of John Egan WWTP, considering each year and the entire study period

	Median removal efficiencies (%)						Mean removal efficiencies (%)					
Year	**BOD**	**CBOD**	**TKN**	**NH_4-N**	**TP**	**TSS**	**BOD**	**CBOD**	**TKN**	**NH_4-N**	**TP**	**TSS**
2001	98.5	98.0	95.0	99.6	50.3	98.7	98.0	97.6	94.3	99.4	49.4	98.5
2002	98.5	98.1	95.0	99.5	50.1	98.7	97.9	97.9	94.4	97.5	50.9	98.5
2003	98.9	98.4	95.7	99.4	48.1	98.9	98.8	98.3	95.3	98.5	49.8	98.8
2004	98.9	98.4	96.1	99.4	55.4	98.9	98.6	98.3	95.0	97.6	55.8	98.8
2005	99.0	98.4	96.4	99.5	51.5	99.0	98.9	98.3	96.2	99.2	51.7	98.9
2006	99.0	98.3	95.4	99.6	45.0	98.9	98.9	98.3	95.3	99.5	45.5	98.8
2007	99.0	98.2	96.5	99.6	95.0	99.0	98.9	98.1	96.2	99.3	89.5	98.9
2008	98.7	98.0	96.2	99.7	94.0	98.8	98.7	97.9	96.0	99.5	92.6	98.7
2009	98.9	98.1	95.5	99.7	47.2	98.8	98.8	98.0	95.2	99.5	49.3	98.7
2010	98.8	98.3	95.4	99.3	49.2	98.7	98.6	98.2	95.3	99.0	51.1	98.6
2011	98.6	98.0	94.2	99.3	52.6	98.6	98.2	97.8	93.8	98.8	53.8	98.3
2012	98.6	98.3	94.6	99.5	50.0	98.7	98.2	98.2	94.5	99.2	51.3	98.4
2013	98.3	98.1	94.3	99.4	47.9	98.4	98.0	97.9	94.2	98.9	48.9	98.1
2014	98.5	98.1	96.2	99.4	51.9	98.6	98.3	98.0	95.2	98.7	53.5	98.4
2015	98.7	98.1	96.3	99.4	48.9	98.7	98.5	98.0	95.4	98.7	51.3	98.6
2016	98.6	98.1	96.2	99.4	48.7	98.6	98.4	98.0	95.5	98.6	49.4	98.5
2017	98.4	97.9	95.7	99.4	44.0	98.4	98.2	97.8	95.1	99.1	46.2	97.8
2018	97.6	97.5	94.1	97.6	45.4	97.7	97.3	97.2	94.1	97.2	46.7	96.9
2019	97.8	97.7	96.3	97.6	47.3	98.3	97.3	97.4	95.0	96.0	48.7	98.1
2020	97.1	96.9	96.4	98.1	49.0	98.7	96.9	96.6	95.7	97.7	49.7	98.5
2021	97.1	98.2	88.2	98.3	41.0	98.2	96.7	98.1	88.8	97.5	42.2	98.0
2022	96.0	97.3	87.5	97.7	38.9	98.0	95.3	96.6	87.5	96.9	38.6	97.9
2023	96.5	97.6	91.7	97.4	48.6	98.2	95.7	97.0	91.0	95.1	49.2	98.1
Global	**98.7**	**98.2**	**95.5**	**99.4**	**50.8**	**98.7**	**98.2**	**97.9**	**94.8**	**98.6**	**54.6**	**98.5**

Table 20 shows the CV and IQR for the removal efficiencies of Egan WWTP for the entire study period. For variables commonly monitored at both Brasília Sul and John Egan WWTPs (BOD, TP, and TSS), as well as similar variables (TN versus TKN and NH4-N), lower values of CV and IQR were recorded for the removal efficiencies of John Egan. The only exception was the TP values of CV and IQR, which were lower at Brasília Sul (Table 13 and Table 20), since this facility has a treatment step to remove phosphorus, and John Egan does not.

Table 20 – Coefficient of variation (CV) and interquartile range (IQR) of the removal efficiencies of John Egan WWTP, considering the entire study period

	BOD	**CBOD**	**TKN**	**NH_4-N**	**TP**	**TSS**
CV	0.01	0.01	0.03	0.03	0.32	0.01
IQR	0.94	0.70	2.55	0.73	16.00	0.72

Figure 21 and Figure 22 show the time series and box-plot graphs of the influent concentrations of BOD, CBOD, TS, and TSS, and influent values of TKN, NH_4-N, pH, and TP, respectively. The influent concentrations are within the typical composition of untreated wastewater in the United States (METCALF & EDDY, 2014).

Influent BOD and CBOD show a similar pattern (Figure 21 (a) and (b)) and slightly smaller concentrations in recent years. There was a gap in the monitoring of influent CBOD from March 20, 2020, to July 31, 2021.

It is clear that there is a pattern every year for most influent variables, especially TS (Figure 21 (c)), TKN, NH_4-N, and TP (Figure 22 (a), (b), and (d)), which could be due to the influence of the seasonality.

Figure 21 – Time series and box-plot graphs of influent BOD (a), CBOD (b), TS (c), and TSS (d) of John Egan WWTP. The blue line is the smoothing line of the trend in the data

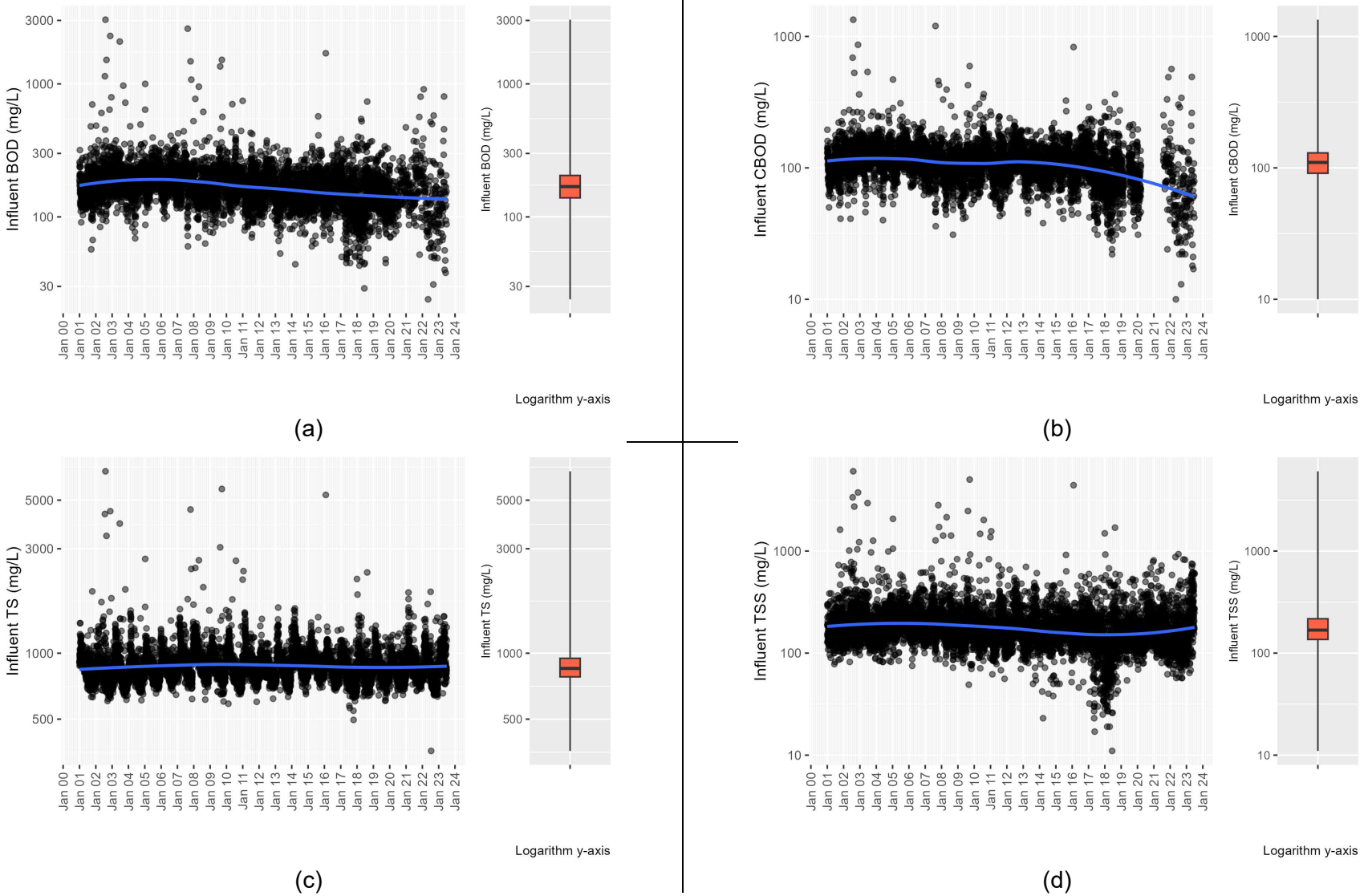
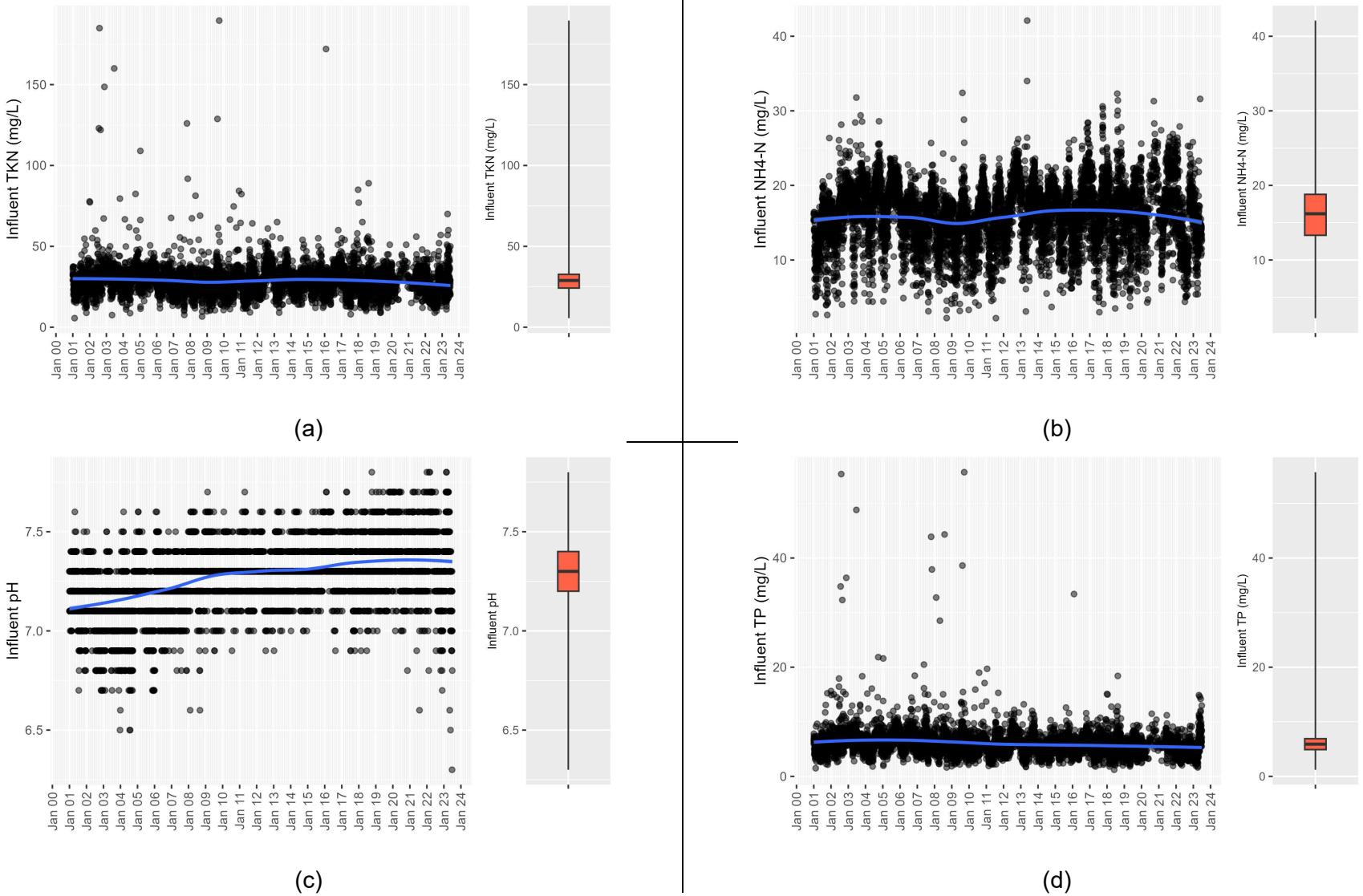


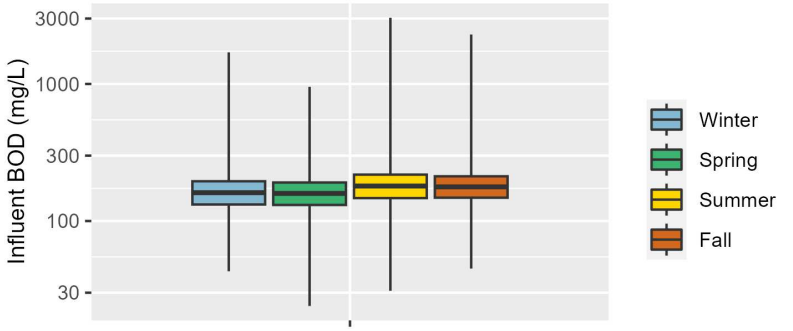
Figure 22 – Time series and box-plot graphs of influent TKN (a), NH_4-N (b), pH (c), and TP (d) of John Egan WWTP. The blue line is the smoothing line of the trend in the data



To assess possible differences in the influent values according to the period of the year, it was considered all four seasons since they are more distinctly from each other than in Brazil. In the Chicago region, climate is typically continental with cold winters with snow, warm summers, and both cold and warm days during spring and fall. Months with more precipitation, above 80 mm (based on data from 1991 to 2020), are from April to October (ILLINOIS STATE CLIMATOLOGIST, 2024).

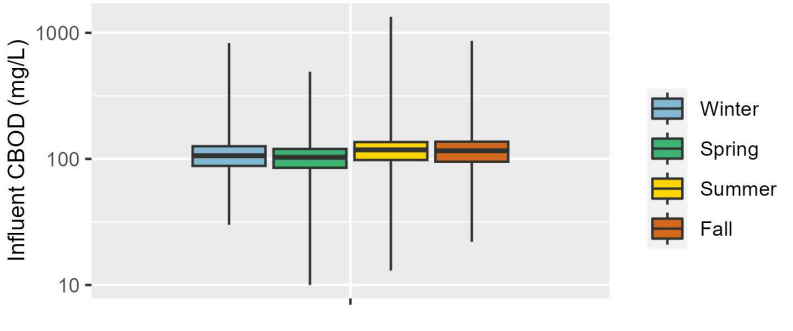
Figure 23 and Figure 24 show the box-plot graphs of influent values in different seasons of the year and the results of the statistical tests. Since there was a significant difference for all parameters in Kruskal-Wallis test ($p < 0.05$), Table 21 summarizes the results of the Dunn statistical test for all variables and comparison of all seasons.

Figure 23 – Influent concentrations in the seasons in John Egan WWTP and results of Kruskal-Wallis and Dunn statistical tests



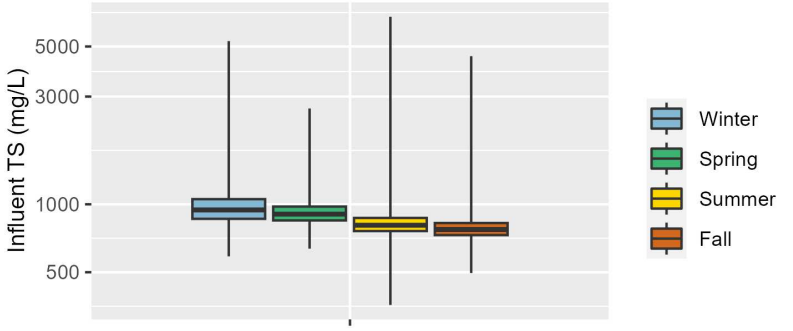
Logarithm y-axis
* p < 0.05 (Kruskal-Wallis test)
Winter ≠ Summer, Winter ≠ Fall, Spring ≠ Fall, Spring ≠ Summer (p < 0.05 in Dunn test)

(a)



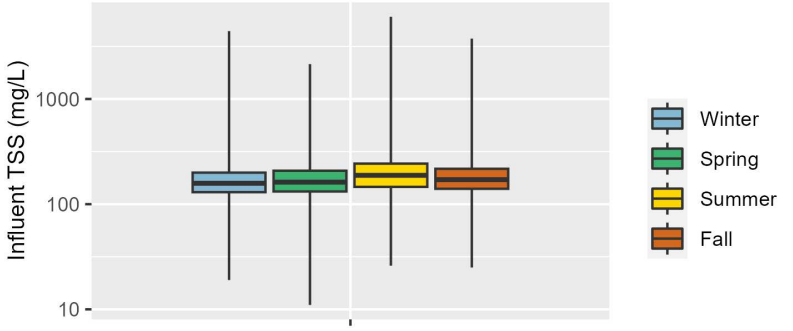
Logarithm y-axis
* p < 0.05 (Kruskal-Wallis test)
Winter ≠ Spring, Winter ≠ Summer, Winter ≠ Fall,
Spring ≠ Summer, Spring ≠ Fall (p < 0.05 in Dunn test)

(b)



Logarithm y-axis
* p < 0.05 (Kruskal-Wallis test)
All seasons are significantly different from each other (p < 0.05 in Dunn test)

(c)



Logarithm y-axis
* p < 0.05 (Kruskal-Wallis test)
All seasons are significantly different from each other (p < 0.05 in Dunn test)

(d)

Figure 24 – Influent values in the seasons in John Egan WWTP and results of Kruskal-Wallis and Dunn statistical tests

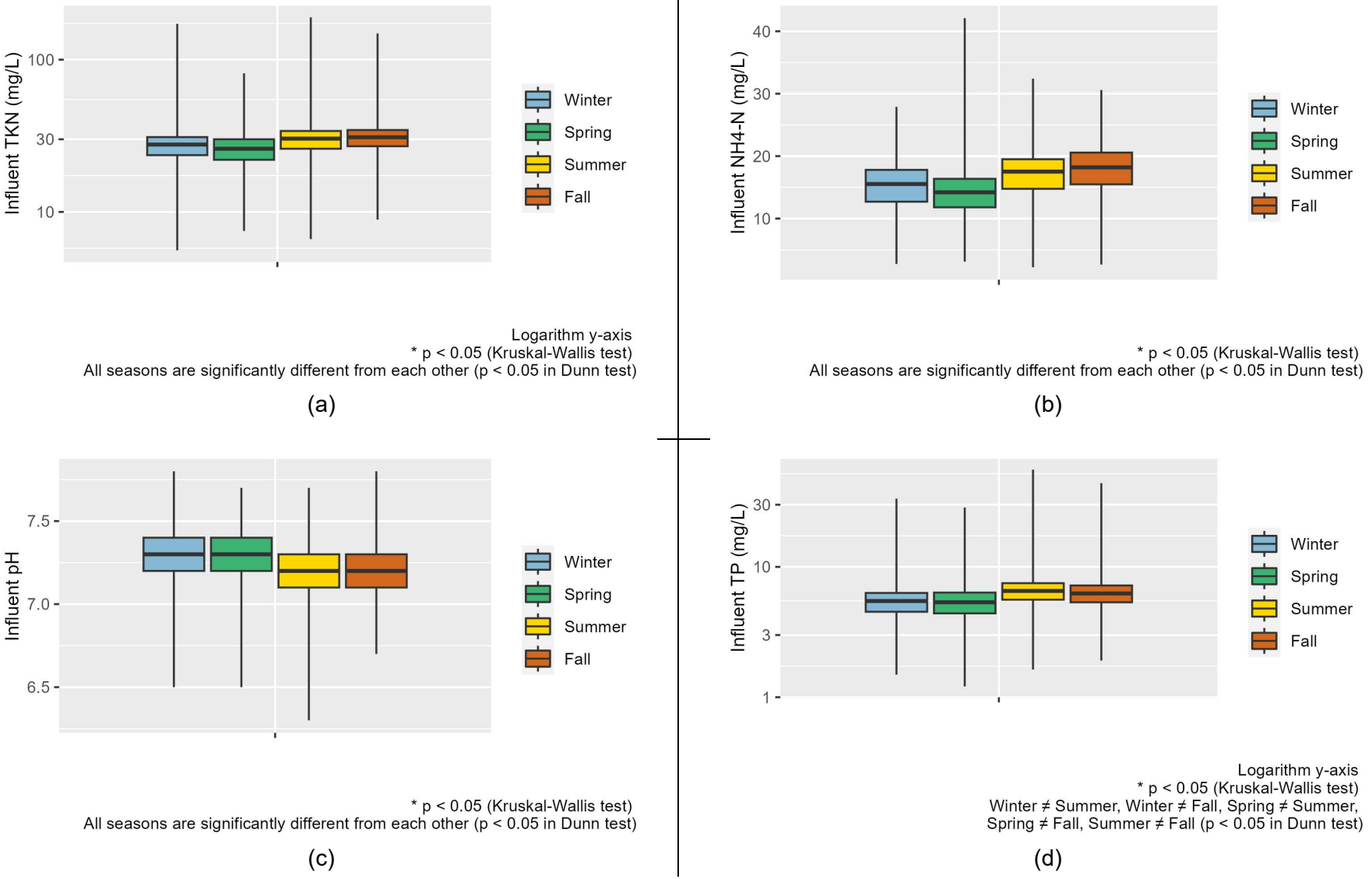


Table 21 – Results of the Dunn test for comparison of all seasons for all influent variables after the result of statistical significance (p < 0.05) of the Kruskal-Wallis test

Comparison	BOD	CBOD	TS	TSS	TKN	NH_4-N	pH	TP
Winter x spring	=	≠	≠	≠	≠	≠	≠	=
Winter x summer	≠	≠	≠	≠	≠	≠	≠	≠
Winter x fall	≠	≠	≠	≠	≠	≠	≠	≠
Spring x summer	≠	≠	≠	≠	≠	≠	≠	≠
Spring x fall	≠	≠	≠	≠	≠	≠	≠	≠
Summer x fall	=	=	≠	≠	≠	≠	≠	≠

Dunn test
≠ significantly different ($p < 0.05$)
= not significantly different ($p > 0.05$)

For most variables, there was a significant difference ($p < 0.05$ in Kruskal-Wallis and Dunn test) among the seasons. The only exceptions (when no statistical difference was found) were influent BOD and TP between winter and spring, and BOD and CBOD between summer and fall.

The volume of wastewater is affected by the sewerage system, whether separate or combined (SALA-GARRIDO; MOLINOS-SELANTE; HERNÁNDEZ-SANCHO, 2012). Although some systems in Cook County are combined sewer systems, John Egan is a separate system, which means that stormwater is drained through storm sewers and sanitary sewage is intercepted to sewers and is directed to John Egan WWTP (MWRD, 2024c). For this reason, the influent values should not be much influenced by the rainy or snowy periods as in combined systems.

Wang et al. (2017) found significant differences between warm and cold seasons in the influent wastewater characteristics of a WWTP in Norway. Most variables had lower concentrations during the cold season. As the studied area had a combined sewerage system, the difference was attributed to the dilution that happened in the period due to snow melting (WANG et al., 2017).

The lower influent concentrations of BOD, CBOD, TKN, NH_4-N, and TP observed during spring and winter at John Egan WWTP may be attributed to the higher flow rates in these seasons, leading to dilution of the sewage. The result of the flow will still be presented in the following results, as it is measured in the effluent.

Figure 25, Figure 26 and Figure 27 show the time series and box-plot graphs of the effluent values of BOD, CBOD, TSS, and pH, the effluent concentrations of TKN, NH_4-N, TP, SP, and the effluent values of TTC and flow, respectively.

Very low concentrations were observed for effluent BOD and CBOD, in which 60% and 83% (Table 17) were censored values of < 2 mg/L, respectively. In this study, when data were censored, the detection limit was utilized as the recorded value, resulting in consistently low concentrations (Figure 25 (a) and (b)).

Although the percentage of censored data was not very high for effluent TSS (34%, Table 17) and NH_4-N (32%, Table 17), the censored data can also be seen in Figure 25 (c) and Figure 26 (b), respectively. For TSS, the detection limit was 2 mg/L and for NH_4-N was variable, namely 0.02, 0.03, 0.04, 0.1, 0.3, and 0.5 mg/L.

Nevertheless, the constant values in the graphs are not only due to censored data. For example, effluent TKN had only 4% of censored data (Table 17) and from 2012 until the end of the study period, the observed values were limited to 1, 2, 3, and 4 mg/L, resembling a categorical variable, and some censored values of 3 and 4 mg/L (Figure 26 (a)). In an email exchange with the plant manager and the laboratory manager of the Egan facility, they mentioned that there was no change in the sampling protocol or analytical methods used to measure TKN. According to the professionals, prior to 2012, TKN values were reported to one decimal place. Starting January 1, 2012, the reporting method was adjusted to whole numbers, which explains the constant values of effluent TKN.

Effluent TP and SP had a very similar pattern, with cyclical trends in each year, which could be attributed to seasonality (Figure 26 (c) and (d)). Most of the effluent phosphorus is in the soluble form, with a median of 96% of the range of SP with regards to TP. Dueñas et al. (2003) studied two WWTPs in Spain with no specific strategy for phosphorus removal and found that the TP in the final effluent was mainly SP since particulate phosphorus was almost completely removed in the primary and secondary clarifiers (DUEÑAS et al., 2003).

Figure 25 – Time series and box-plot graphs of effluent BOD (a), CBOD (b), TSS (c), and pH (d) of John Egan WWTP. The blue line is the smoothing line of the trend in the data

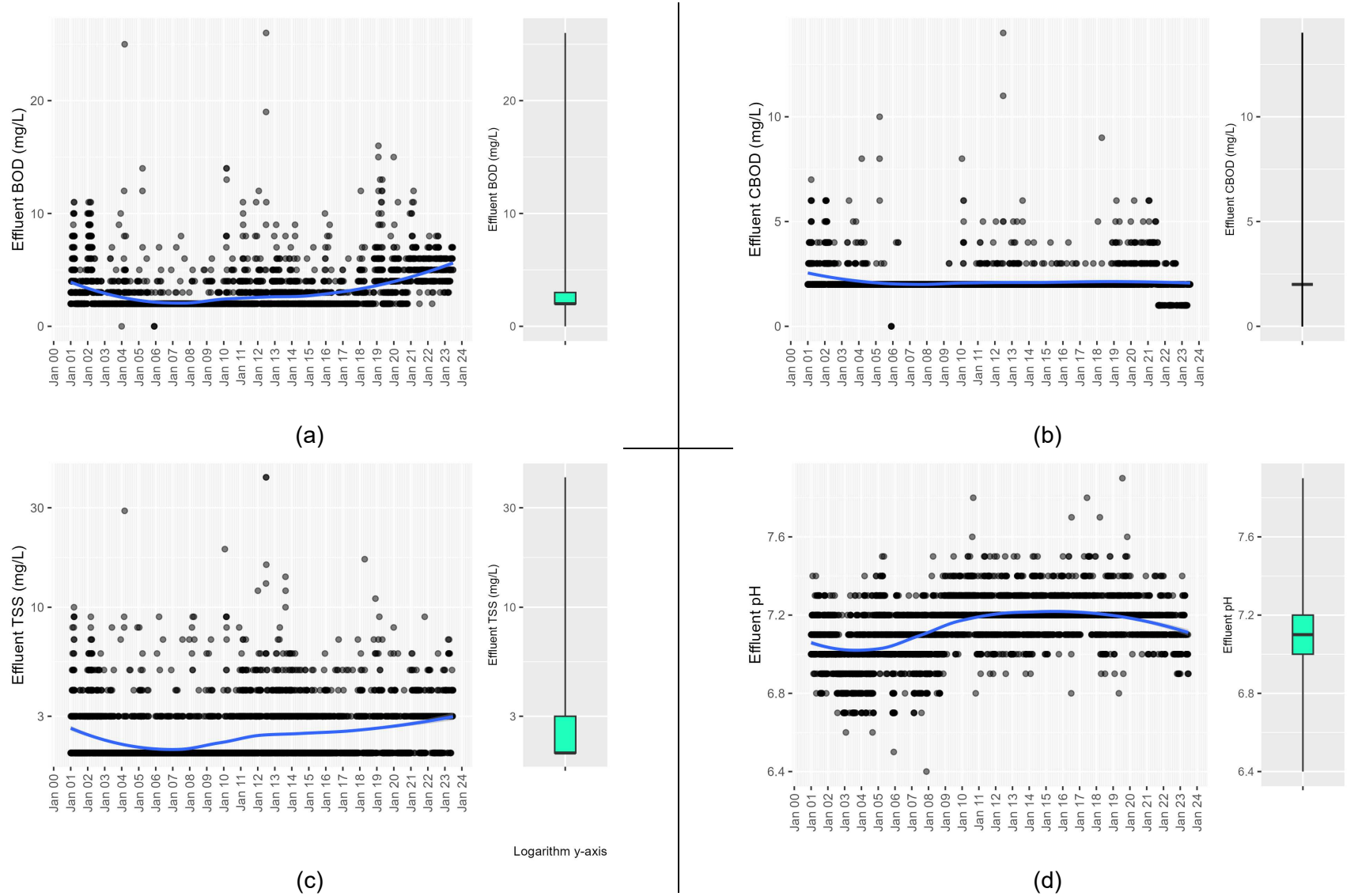


Figure 26 – Time series and box-plot graphs of effluent TKN (a), NH_4-N (b), TP (c), and SP (d) of John Egan WWTP. The blue line is the smoothing line of the trend in the data

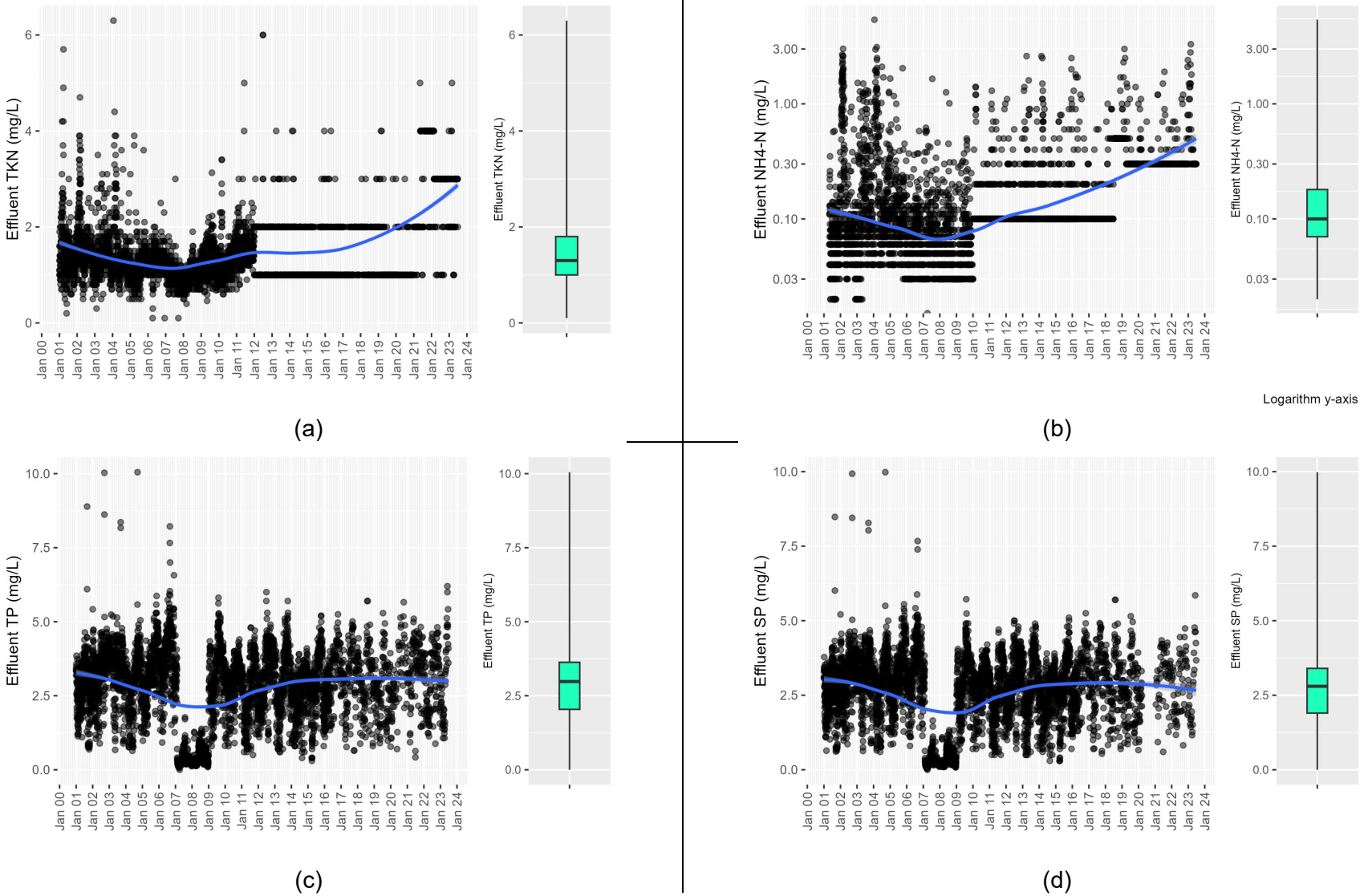
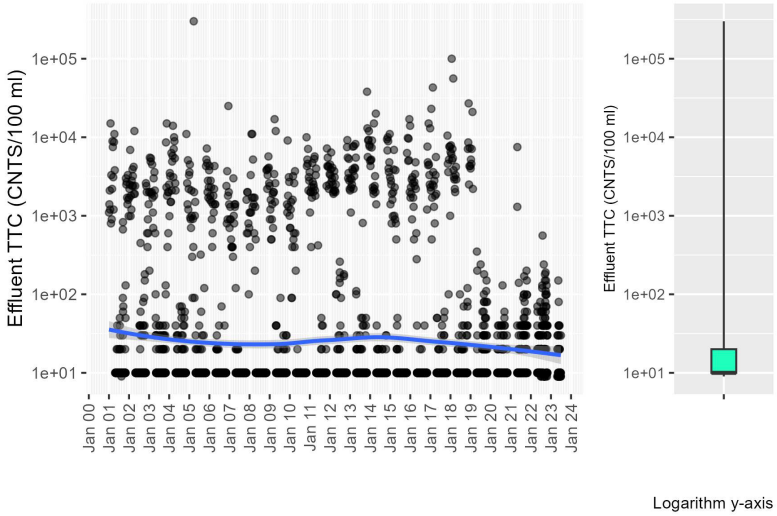
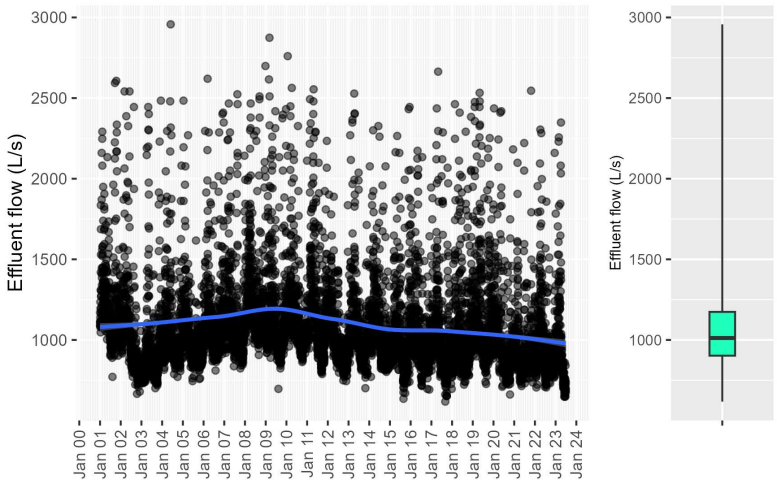


Figure 27 – Time series and box-plot graphs of effluent TTC (a) and flow (b) of John Egan WWTP. The blue line is the smoothing line of the trend in the data



(a)



(b)

From July 2007 to January 2009 the effluent concentrations of TP and SP were much lower than in the remaining period (Figure 26 (c) and (d)), which indicates that there was probably a change in the operation of the facility during this period. From February 2007 to February 2008, a full-scale demonstration study was conducted at the John Egan facility to investigate phosphorus removal. It was a limited-duration project agreed upon between MWRD and the Illinois Environmental Protection Agency to enhance the understanding of the effects of TP reduction to 0.5 mg/L at the Egan WWTP on water and sediment chemistry and aquatic communities in Salt Creek (ZHANG et al., 2008). Ferric chloride ($FeCl_3$) was added to the mixed liquor at the end of the aeration tanks, which caused the mixing of $FeCl_3$ with wastewater. The chemical

reaction resulted in the particulate phosphorus to precipitate in the secondary clarifiers. In the twelve-month period, the average TP concentration in the final effluent was 0.34 mg/L (ZHANG et al., 2008).

The recently reissued NPDES for the John E. Egan facility requires a discharge standard of 1 mg/L for the monthly average of TP with associated load limits. Several studies are being conducted by the MWRD to understand the feasibility of various technologies for TP removal (MWRD, 2022).

The seasonal disinfection results in a high variability in thermotolerant coliforms concentrations, which can achieve very low concentrations, below the detection limit of 10 CNTS/100 mL in the months that the facility undergoes disinfection, and very high concentrations, achieving a maximum of 3.0×10^5 CNTS/100 mL in the remaining months (Figure 27 (a)). There is also a high variability in the effluent flow of the facility (Figure 27 (b)), which could be attributed to the influence of seasonality. The design flow of the John Egan WWTP is 2,190 L/s, and only 1.5% of the samples were above this value.

Understanding the influence of seasonality on the effluent quality of WWTPs is important to guide water quality management decisions (COMBER; GARDNER; ELLOR, 2019). Figure 28, Figure 29 and Figure 30 show the box-plot graphs of effluent values in different seasons of the year and the results of the statistical tests. Table 22 summarizes the results of the Dunn statistical test for all variables and comparison of all seasons.

Figure 28 – Effluent values in the seasons in John Egan WWTP and results of Kruskal-Wallis and Dunn statistical tests

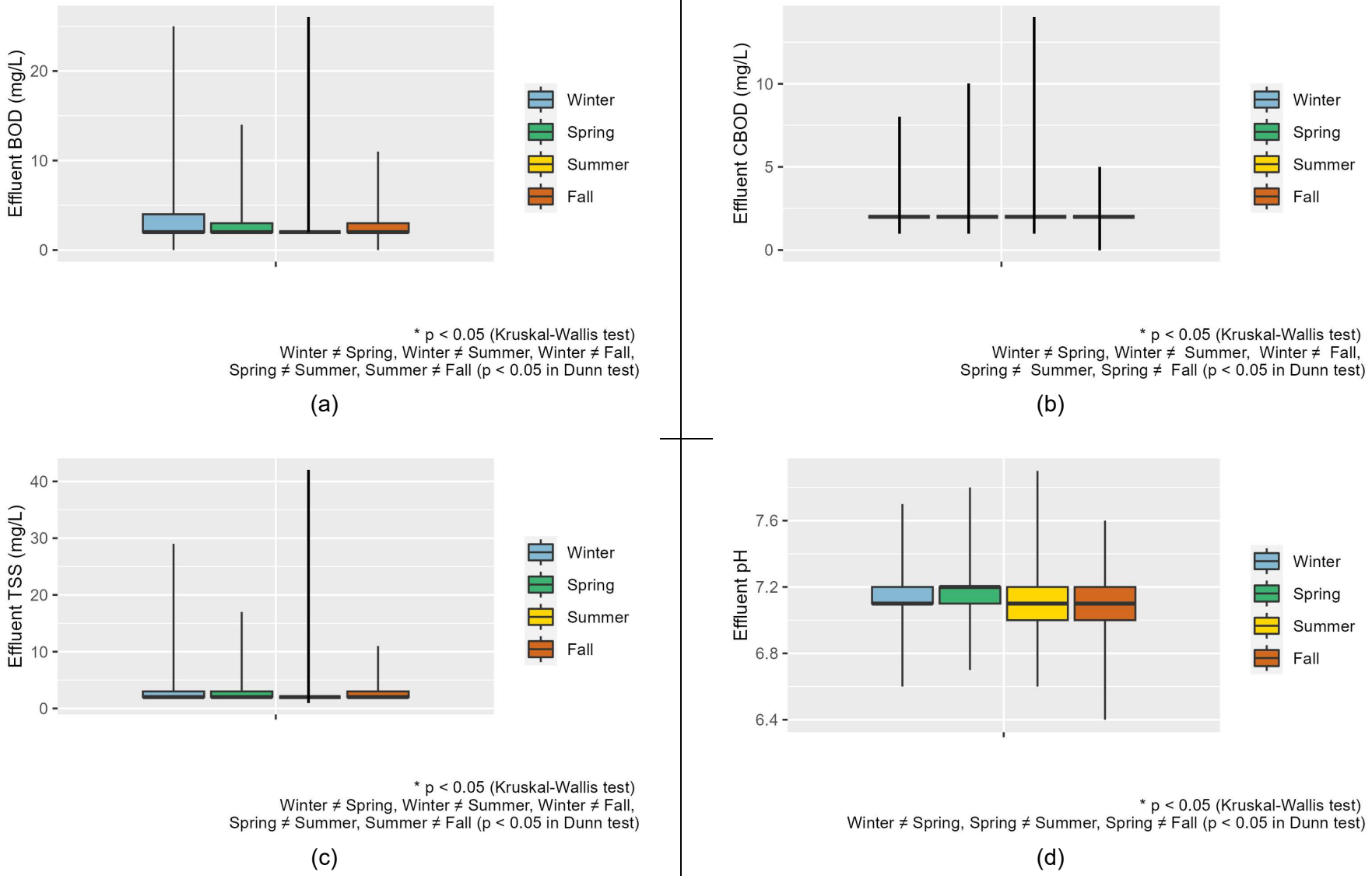


Figure 29 – Effluent values in the seasons in John Egan WWTP and results of Kruskal-Wallis and Dunn statistical tests

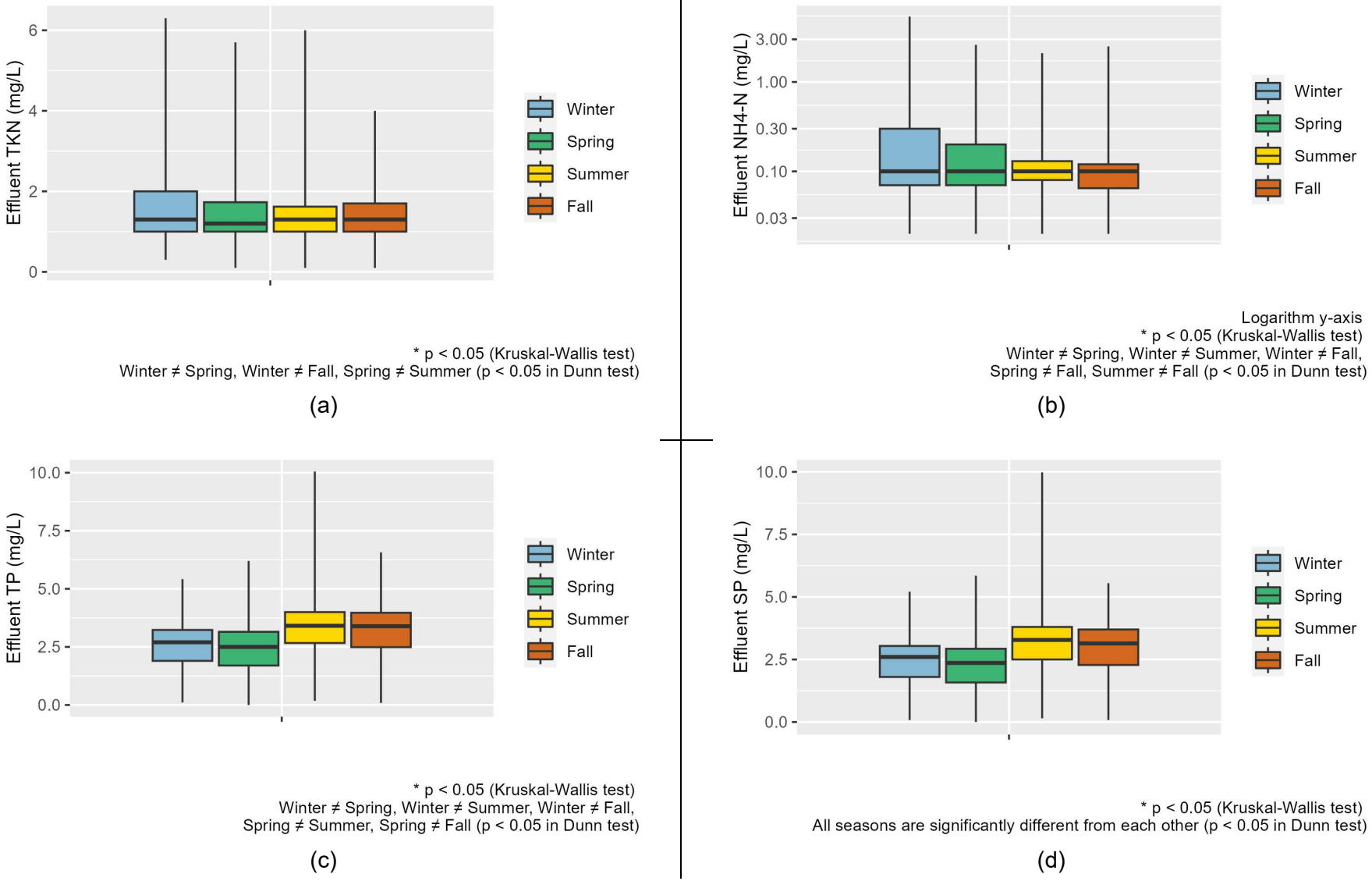
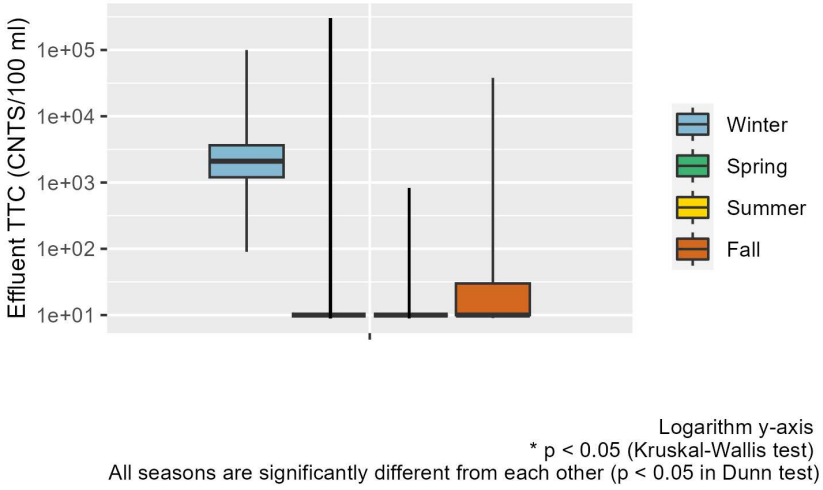
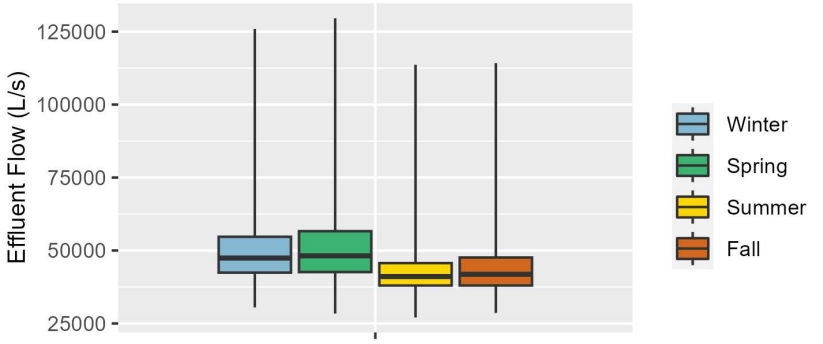


Figure 30 – Effluent values in the seasons in John Egan WWTP and results of Kruskal-Wallis and Dunn statistical tests



Logarithm y-axis
* p < 0.05 (Kruskal-Wallis test)
All seasons are significantly different from each other (p < 0.05 in Dunn test)

(a)



* p < 0.05 (Kruskal-Wallis test)
Winter ≠ Summer, Winter ≠ Fall, Spring ≠ Summer,
Spring ≠ Fall, Summer ≠ Fall (p < 0.05 in Dunn test)

(b)

Table 22 – Results of the Dunn test for comparison of all seasons for all effluent variables after the result of statistical significance (p < 0.05) of the Kruskal-Wallis test

Comparison	BOD	CBOD	TSS	pH	TKN	NH_4-N	TP	SP	TTC	Flow
Winter x spring	≠	≠	≠	≠	≠	≠	≠	≠	≠	=
Winter x summer	≠	≠	≠	=	=	≠	≠	≠	≠	≠
Winter x fall	≠	≠	≠	=	≠	≠	≠	≠	≠	≠
Spring x summer	≠	≠	≠	≠	≠	=	≠	≠	≠	≠
Spring x fall	=	≠	=	≠	=	≠	≠	≠	≠	≠
Summer x fall	≠	=	≠	=	=	≠	=	≠	≠	≠

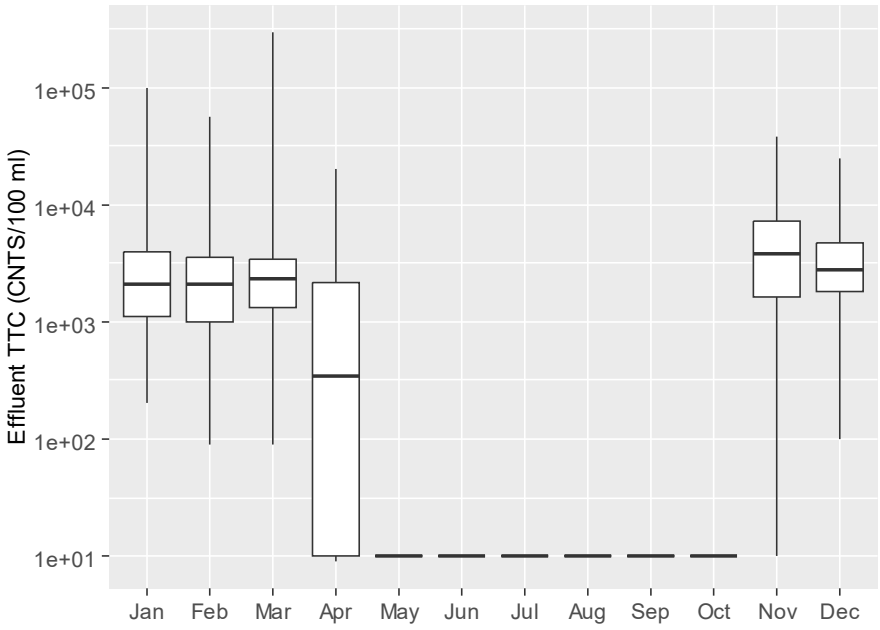
Dunn test
≠ significantly different (p < 0.05)
= not significantly different (p > 0.05)

Higher concentrations of effluent BOD, TKN, and NH_4-N during winter could be due to the dependence of the temperature for the biological process (COMBER; GARDNER; ELLOR, 2019). However, for effluent TP and SP, the highest concentrations were observed during summer and fall. As there is no specific phosphorus removal step at the John E. Egan WWTP, the influence of the temperature for the biological removal could not be verified as for other compounds. The influent concentrations of TP were also higher in the summer and fall (Figure 24 (d)), which can explain the higher effluent concentrations in these seasons.

Regarding effluent flow, as the John Egan is a separate sewer system, it is not expected that the flow is substantially influenced by rain, snow melt, or other hydrological factors. However, considering the region's climate, higher flows were expected during warm seasons, in which the population potentially showers more often. However, the highest flows were recorded during spring and winter (Figure 30 (b)), which could be associated with other population dynamics and habits. The existence of holidays, such as spring and winter breaks, can also potentially lead to a higher contribution of the inflow rate during these seasons.

The disinfection at John E. Egan is performed in warmer seasons, when the population uses the water bodies for recreation activities, such as swimming or boating, with direct contact with water. For this reason, there was a significant difference in the thermotolerant coliforms concentrations among all seasons (Table 22). Figure 31 shows the TTC concentrations in the months of the year. In months in which disinfection is performed, the concentrations were censored values, below 10 CNTS/100 mL. This result highlights the importance of disinfection as it can achieve very low concentrations.

Figure 31 – Effluent concentrations of thermotolerant coliforms in the months of the year at the John Egan WWTP



3.4 CONCLUSIONS

The description of the Brasília Sul WWTP process and intrinsic characteristics highlighted its national importance as it is the largest plant in the capital city of Brazil and employs A²O activated sludge process and tertiary treatment to remove nutrients, which is an advanced technology compared to the national scenario.

Data exploration techniques have proven essential in understanding the data characteristics, WWTP process, and performance. The study found that the influent quality of the Brasília Sul facility was significantly influenced by seasonality, resulting in high variability in the data. Although not expected, high variability was also observed in effluent quality, likely due to seasonality, influent characteristics, and operational conditions. Even though high removal efficiencies were recorded, the WWTP still did not fully comply with discharge standards for nitrogen and phosphorus concentrations.

The operational variables of Brasília Sul WWTP data, namely influent flow, aluminum sulfate consumption, and anionic polyelectrolyte consumption, also showed high variability over the study period, with seasonality significantly influencing these variables.

The John E. Egan WWTP, which is in operation in the United States, served as the second case study in the research. Situated in a region of national relevance due to its historical sanitation development, the facility shares a comparable size with the Brasília Sul WWTP but employs distinct treatment technology. The study evaluated over 20 years of monitoring data, resulting in a large dataset.

The influent concentrations at the John Egan WWTP exhibited high variability, which was attributed to seasonal influences. When comparing the CV of the influent variables, higher values were found at John Egan than at Brasília Sul WWTP, which highlights the variability in John Egan's influent conditions.

High removal efficiencies were observed for all constituents at John Egan except phosphorus. Effluent concentrations were generally low, with a notable proportion falling below detection limits for several variables. For most variables, lower CV and IQR values of the effluent concentrations and removal efficiencies were found for John Egan compared to Brasília Sul's results. This highlighted that, although influent conditions are highly variable, John Egan has high stability. However, compliance with local discharge standards was not assessed for this facility.

The variability and complexity of the data and the wastewater treatment process highlighted the importance of preliminary statistical analyses and data visualization techniques, particularly since there is a lack of comprehensive studies on the Brasília Sul and John E. Egan WWTPs in the literature.

CHAPTER 4: USE OF ARTIFICIAL NEURAL NETWORK MODELS FOR THE PREDICTION OF THE PERFORMANCE OF WASTEWATER TREATMENT PLANTS: THE CASE STUDIES OF BRASÍLIA SUL AND JOHN E. EGAN WWTPs

4.1 PRESENTATION

This chapter will cover the use of artificial neural network models for the prediction of the performance of two WWTPs in terms of effluent quality. This chapter was developed using the knowledge and expertise that were built up in Chapters 2 and 3.

Chapter 2 presented a systematic literature review in which the key characteristics of the models developed in the literature with similar prediction objectives were identified. Understanding the typical ANN models used, their structures and configurations, preprocessing techniques, and metrics for assessing model performance was crucial for conducting a better modeling approach for the WWTPs in Brazil and the United States.

Chapter 3 introduced the monitoring dataset of the Brasília Sul and John E. Egan WWTPs, which were selected as the case studies. To effectively use the predictive models, a comprehensive understanding of the monitored parameters, the datasets characteristics, and the plants' performances was essential. For instance, parameters with higher monitoring frequency or more consistent monitoring were identified, resulting in a larger number of samples, which is required for machine learning training.

This chapter presents the use of ANN to predict the effluent concentrations of TN, NH_4-N, TP, COD, and TSS at the Brasília Sul facility and effluent concentrations of TP at the John E. Egan WWTP. The modeling approach is described, which includes preprocessing techniques, modeling development, and assessment of the model performance. Additionally, a sensitivity analysis was conducted to understand which variables are most responsible for the WWTPs' performance.

4.2 INTRODUCTION

Wastewater treatment plants play a crucial role in safeguarding environmental and public health. However, improper operation of these systems may cause serious risks since discharging an insufficiently treated effluent into a receiving water body can cause or spread diseases to human beings and impact the aquatic ecosystem (SHARGHI et al., 2019).

It is essential to implement effective monitoring and control of WWTPs (SHARGHI et al., 2019). Key variables in wastewater treatment need to be evaluated to control pollution (OSMAN; LI, 2020). In this context, nitrogen and phosphorus require special attention, as their removal in a wastewater treatment process is crucial for preventing the eutrophication of water bodies. There has been a growing number of studies on nutrient removal in recent years (CHING; SO; MORCK, 2021).

Although it is crucial to maintain high performance to achieve adequate removal efficiencies of pollutants, including nutrients, WWTPs are highly complex systems. They are influenced by numerous physicochemical and microbiological aspects (SAFEER et al., 2022), resulting in a high variability in influent quality, flow rate, pollutant load, and hydraulic conditions. This complexity makes WWTP operation challenging (SAFEER et al., 2022; WANG et al., 2021).

Better control of a WWTP may be achieved by proposing a modeling tool to predict the performance of the WWTP based on previous observations of key quality parameters (HEJABI et al., 2021; ZHANG; HU, 2012). These modeling approaches can even improve the knowledge about how operational factors affect effluent quality, which is extremely valuable in engineering scenarios for improving WWTP performance (WANG et al., 2021). Quantitative models have been developed to assess the impacts of physical, chemical, and biological conditions on the removal of carbon, nitrogen, and phosphorus from sewage (WANG et al., 2024b). However, modeling a WWTP is a challenging task due to the nonlinear characteristics resulting from the complexities of the biological processes involved (LIU; HUANG; YOO, 2013; ZHANG; HU, 2012).

In recent decades, artificial intelligence approaches have been used as effective tools for investigating environmental engineering issues (HEJABI et al., 2021). These techniques are robust and can handle data with nonlinear characteristics such as those from WWTPs (SHARGHI et al., 2019). AI methods can predict the performance of WWTPs based on past observations of quality parameters (SHARGHI et al., 2019).

Safeer et al. (2022) found that there has been increasing research interest in AI techniques for wastewater treatment in recent years. According to the authors, ANN has been a rapidly growing field among AI techniques in the past decade (SAFEER et al., 2022). ANNs are mathematical structures that capture the nonlinear relationships

between inputs and outputs (RAHMATI; TISHEHZAN; MOAZED, 2021). These structures consist of an input layer, one or more hidden layers, and an output layer that processes the weighted average and bias in the hidden layers to generate a specific output corresponding to the response variable being predicted. There are some neurons in each of these layers that are interconnected in the form of a network (SAFEER et al., 2022).

Previous studies have employed ANN techniques to predict the effluent nutrient concentrations in full-scale WWTPs, including total nitrogen (ABBA; ELKIRAN; NOURANI, 2021; BAGHERI et al., 2015; HAZALI; WAHAB; IBRAHIM, 2017; LEE et al., 2011; LIU; HUANG; YOO, 2013; NOURANI; ELKIRAN; ABBA, 2018), ammonia nitrogen (AL-GHAZAWI; ALAWNEH, 2021; GAYA et al., 2014; HAZALI; WAHAB; IBRAHIM, 2017; JAMI; MUJELI; KABBASHI, 2011; KHATRI; KHATRI; SHARMA, 2019; ZHAO; CHAI; YUAN, 2012), and total phosphorus (ABBA; ELKIRAN; NOURANI, 2021; HAZALI; WAHAB; IBRAHIM, 2017; KHATRI; KHATRI; SHARMA, 2019; LIU; HUANG; YOO, 2013). Besides nutrients, organic matter and solids are contaminants targeted for removal in WWTPs as they can impact water resources and public health (GHOLIZADEH et al., 2024). Previous studies have also developed ANN models for the prediction of effluent chemical oxygen demand (AL-GHAZAWI; ALAWNEH, 2021; ALSULAILI; REFAIE, 2021; BAGHERI et al., 2015; EL-RAWY et al., 2021; HEJABI et al., 2021; KHATRI; KHATRI; SHARMA, 2019; NOURANI; ASGHARI; SHARGHI, 2021) and total suspended solids (ELMAADAWY et al., 2021; GAYA et al, 2014; GHOLIZADEH et al., 2024; KUSIAK; WEI, 2013; SALEH, 2021; WANG et al., 2021).

However, no studies have been found that used ANN predictive models for the performance of domestic WWTPs represented by the effluent quality or removal efficiencies in Brazil or other Latin American countries. The development of such a technique for a Brazilian WWTP may expand the knowledge about the potential of ANN in predicting effluent quality under the country's climatic and socioeconomic conditions. This study presents the application of ANN to an important Brazilian WWTP that discharges its effluent into a sensitive water body, highlighting the importance of modeling nutrient removal at this facility. As a form of comparison with a different scenario, an ANN application was also presented for a WWTP in the USA, with a

different process and dataset characteristics, such as a longer monitoring period and higher monitoring frequency.

4.3 METHODS

4.3.1 Wastewater treatment plants and monitoring datasets description

The Brasília Sul WWTP, which is located in Brasília, Brazil's capital, is the largest wastewater treatment facility in the municipality in terms of design flow. It has a design flow of 1,500 L/s and a population equivalent of 460,000 inhabitants. The treated effluent from the plant is discharged into Lake Paranoá, a water body of multiple uses, including recreation and water supply. This lentic water body is also sensitive to the eutrophication process, which can cause severe environmental damage.

To mitigate the potential impacts of nutrient pollution, the Brasília Sul WWTP employs a technology aimed at nitrogen and phosphorus removal. The treatment process begins with a preliminary stage to remove coarse solids and sand. Next, the effluent undergoes primary treatment, which involves three primary settlers to remove a portion of the solids and organic matter. The effluent is then directed to three biological reactors that use the A^2O (anaerobic/anoxic/oxic) activated sludge process to biologically remove nutrients, organic matter, and solids. This secondary treatment step includes 12 secondary settlers, which allow the biomass to sediment, with a portion being recirculated to the anaerobic zone of the biological reactors. The final step of the treatment process is the tertiary treatment, which involves coagulation with aluminum sulfate as the coagulant and anionic polyelectrolyte as the flocculant, followed by flocculation and dissolved air flotation. This polishing step is necessary for further removal of phosphorus, which is a critical component for preventing eutrophication.

A collaboration was established with Caesb, the sanitation company responsible for operating the Brasília Sul WWTP. The company provided a monitoring dataset that included influent and effluent samples, as well as operational variables, collected from January 2020 to March 2022. The monitoring frequency varied depending on the variable being measured, with those that were monitored more frequently having a larger amount of data available. When selecting the variables for the study, those with

a higher frequency of monitoring were prioritized. In addition to this criterion, the effluent variables (output of the models) were selected considering the importance of nutrient removal control in this facility.

A second WWTP was included in the study to compare different scenarios. The facility is in the United States and has a similar size to Brasília Sul WWTP but employs different treatment processes and its monitoring dataset has different characteristics. John E. Egan WWTP is located in Schaumburg, Illinois. It has a capacity to treat 2,190 L/s and serves 160,735 residents of Cook County. The treated effluent is discharged into Salt Creek.

John E. Egan WWTP is a single stage nitrification activated sludge with a tertiary filtration and disinfection process. Wastewater undergoes seasonal disinfection through chlorination and dichlorination, and tertiary filtration is carried out as a polishing step for further removal of solids using tertiary filters made of dual media of sand and anthracite (MWRD, 2015). The main treatment units for the liquid phase treatment include screens, aerated grit tanks, primary settling tanks, activated sludge aeration tanks, secondary clarifiers, and tertiary filters (ZHANG et al., 2008).

Monitoring data were obtained from the Metropolitan Water Reclamation District of Greater Chicago (MWRD), the agency responsible for operating the facility. The data were collected from January 2001 to June 2023. The monitoring frequency varied according to the variables and timeframe within the series. In selecting variables for the study, those with a higher frequency, lower percentage of missing and censored data, and fewer gaps were selected. In addition, certain variables had consistently constant values throughout the time series, and for this reason, they were not included in the modeling study. In addition to this criterion, the effluent variable (output of the models) was selected because of the implementation of a new discharge limit for the facility (1 mg/L for the monthly average of TP with associated load limits), highlighting the importance of understanding the removal of this pollutant in the process.

Table 23 displays variables that were ultimately selected for the study for both WWTPs.

Table 23 – Selected variables for the study

WWTP	Type	Variables	Unit	Frequency
Brasília Sul	Influent	COD, TN, TP, TSS	mg/L	Every two or three days
	Effluent	TN, TP, NH_4-N, COD, TSS	mg/L	Every two or three days
	Operational	Influent flow Aluminum sulfate, anionic polyelectrolyte	L/s kg/d	Daily
John E. Egan	Influent	BOD, TS, TSS, TKN, NH_4-N, TP	mg/L	Daily to two times a week
	Effluent	TP	mg/L	Daily to two times a week
	Operational	Effluent flow	L/s	Daily to two times a week

COD: chemical oxygen demand; TN: total nitrogen; TP: total phosphorus; TSS: total suspended solids; NH_4-N: ammonia nitrogen; BOD: biochemical oxygen demand; TS: total solids; TKN: total Kjeldahl nitrogen

4.3.2 Development of artificial neural network models

The R programming language was employed throughout the entire process of data preprocessing, model development, and evaluation. The “h2o” package (LEDELL et al., 2022) was used for the model development.

4.3.2.1 Input and output variables

From January 2020 to March 2022, 15% of the samples collected from the Brasília Sul WWTP had TP effluent concentrations exceeding the discharge limit of 0.3 mg/L, while 37% exceeded the TN effluent concentration limit of 8.7 mg/L. Given the critical importance of nutrient removal in this facility, nitrogen and phosphorus compounds were selected as the response variables for the single-output ANN models. Therefore, three separate models were developed to predict the effluent concentrations of TN, NH_4-N, and TP at the Brasília Sul WWTP. Additionally, models were developed to predict the effluent concentrations of COD and TSS. The biological phosphorus removal is achieved through the incorporation of excess amounts of phosphorus into the bacterial biomass. Therefore, the loss of suspended solids in the effluent from the secondary treatment can lead to an increase in phosphorus concentrations (VON SPERLING, 2007). For this reason, the removal of organic matter and solids is also an indication of the removal of phosphorus in this case study.

The quality of the treated effluent depends on the concentration of the influent and process parameters of the WWTP (KHATRI; KHATRI; SHARMA, 2020). The input variables for all the models of Brasília Sul WWTP comprised the influent concentrations of COD, TN, TP, and TSS and the influent flow rate. These variables were selected

because of the higher availability of data points in the monitoring dataset. Understanding the relationship between raw sewage characteristics, namely influent quality and influent flow, and the effluent quality is crucial for optimizing the performance of a WWTP.

The consumption of the chemical products aluminum sulfate and anionic polyelectrolyte used in the tertiary treatment of Brasília Sul WWTP were additional input variables for the prediction models of effluent TP, COD, and TSS. The models for nitrogen compounds did not include variables related to chemical product consumption, as the polishing step is ineffective in removing TN and NH_4-N (BARROS, 2013; KWON et al., 2015), which was confirmed by Portela (2018) while analyzing the Brasília Sul facility process.

At John E. Egan WWTP, there is currently no current specific phosphorus removal step. However, the new standards of the facility require a discharge limit of 1 mg/L for the monthly average TP, with associated load limits. For this reason, it is important to understand the current factors influencing the removal of phosphorus at the facility for further improvement. Besides that, in the dataset of the John Egan facility, phosphorus was the effluent parameter with the most consistent results since the other variables had a high percentage of censored data or constant values. For this reason, a single-output model was developed to predict the effluent concentrations of TP at the John E. Egan facility.

The input variables of the John E. Egan WWTP model comprised the influent concentrations of BOD, TS, TSS, TKN, NH_4-N, TP, and effluent flow. These variables were better and more consistently monitored, with fewer missing data points and gaps throughout the time series. The predictive models for effluent phosphorus from the two WWTPs did not have the exact same input variables, since the selection of explanatory variables depended on the availability of data in each dataset. However, the number of explanatory variables for the two WWTP models was equal (seven variables), enabling a better comparison.

Table 24 presents the input variables for each output of the models for both WWTPs.

Table 24 – Input variables for each output of the models

WWTP	Output	Input
Brasília Sul	TN_{efl} $NH_4\text{-}N_{efl}$	COD_{infl}, TN_{infl}, TP_{infl}, TSS_{infl}, influent flow
	TP_{efl} COD_{efl} TSS_{efl}	COD_{infl}, TN_{infl}, TP_{infl}, TSS_{infl}, influent flow, aluminum sulfate, anionic polyelectrolyte
John E. Egan	TP_{efl}	BOD_{infl}, TS_{infl}, TSS_{infl}, TKN_{infl}, $NH_4\text{-}N_{infl}$, TP_{infl}, effluent flow

COD: chemical oxygen demand; TN: total nitrogen; TP: total phosphorus; TSS: total suspended solids; NH_4-N: ammonia nitrogen; BOD: biochemical oxygen demand; TS: total solids; TKN: total Kjeldahl nitrogen; infl: influent; efl: effluent

As the John Egan WWTP had a longer time series of data and higher monitoring frequency, the number of observations for training and testing the model was much larger than that for the Brasília Sul models. For a more appropriate comparison, a second model was developed to predict the effluent TP concentration at John Egan WWTP by using an equal number of data points as in the model for WWTP Brasília Sul and selecting the most recent observations from the dataset.

4.3.2.2 Data preprocessing

Data preparation before feeding it to a model is an essential step in machine learning techniques (YAQUB et al., 2020). Model development involves several steps, including data collection and preprocessing, model design, model training, testing, and model execution (PHAM et al., 2020). Missing records cannot be used for training or testing a neural network model, so they were excluded from the dataset.

Specifically for the John E. Egan WWTP data, from July 2007 to January 2009, the effluent concentrations of TP were much lower than those during the remaining time series, as the facility's operation changed during this period (results shown in Chapter 3). For this reason, the data from July 2007 to January 2009 were removed from the dataset before developing the ANN model.

Data division is a necessary step in modeling (CHEN et al., 2020). The training dataset was used to perform network learning and adjust the network weights, while the testing dataset was utilized to evaluate the model's ability to generalize to new data (LANTZ, 2013; MJALLI; AL-ASHEH; ALFADALA, 2007; WANG et al., 2021; ZHAO et al., 2020). The data were randomly divided into 75% for training and 25% for testing, which is the

same partitioning ratio used in previous studies on ANN models for predicting the quality of WWTP effluent nutrient concentrations (ABBA; ELKIRAN; NOURANI, 2021; AL-GHAZAWI; ALAWNEH, 2021; NOURANI; ELKIRAN; ABBA, 2018).

Normalization should be conducted after dividing the data, as noted by Chen et al. (2020), so training and testing subsets were normalized separately. Range scaling and standardization are two common categories of data normalization (CHEN et al., 2020). These two methods were tested in this study and are described below. After prediction, the predicted values were denormalized (WANG et al., 2024a), so the error metrics were presented in the original unit (mg/L) for all models' outputs.

The most common practice in the field is the range scaling method by the min-max normalization in the range [0, 1], as described by Equation (6). This preprocessing technique has been used in previous studies on ANN models for predicting the quality of WWTP effluent, including nutrient concentrations (AL-GHAZAWI; ALAWNEH, 2021; BAGHERI et al., 2015; GAYA et al., 2014; NOURANI; ELKIRAN; ABBA, 2018; QIAO; YANG; YUAN, 2011; YAQUB et al., 2020).

$$y = \left(\frac{x - x_{min}}{x_{max} - x_{min}}\right) \quad (6)$$

Where y is the normalized data, x is the measured data, and x_{min} and x_{max} are the minimum and maximum values of the measured data, respectively (GE et al., 2020).

The Z-score normalization accomplished the standardization method, in which the variables are standardized to have a mean of zero and standard deviation of one (JAMI; MUJELI; KABBASHI, 2011), as shown in Equation (7). This approach was used in previous studies by Jami, Mujeli, and Kabbashi (2011) and Zhao, Chai, and Yuan (2012) to predict NH_4-N concentrations in effluents from full-scale WWTPs. Xu et al. (2024) found the Z-score normalization to be the best among the various normalization methods for predicting effluent TP concentrations.

$$y = \frac{x - \bar{x}}{\sigma} \quad (7)$$

Where y is the normalized data; x is the measured data; and $\bar{x}$ and σ are the mean and the standard deviation of the variable, respectively (JAMI; MUJELI; KABBASHI, 2011).

4.3.2.3 ANN methods and structures

There are several classifications of ANNs (YE et al., 2020). In this study, feedforward neural networks (FFNNs) with backpropagation training algorithms were developed for each output variable. This type of model structure is widely used in engineering and wastewater treatment problems due to its accuracy and capabilities (AL-GHAZAWI; ALAWNEH, 2021).

The FFNNs consist of one input layer, one or more hidden layers, and one output layer, with each layer containing a specific number of interconnected neurons that determine the neural system's architecture (ELFANSSI et al., 2018). The number of neurons in the input and output layers corresponds to the number of input features and output variables, respectively. The user should specify the number of hidden layers and neurons before training the ANN. A trial-and-error approach was used in this study to tune the network structure, testing all possible combinations of one or two hidden layers and one to ten neurons in each hidden layer. These hyperparameters' ranges were selected based on commonly used structures in similar prediction problems in the literature.

The number of epochs is another hyperparameter that can be adjusted to optimize the model’s performance. An epoch refers to a complete pass through the entire training dataset during the training process of a machine learning model. During one epoch, the model processes every sample in the training set and adjusts the network parameters to improve its performance on the task it is being trained for (AHMED et al., 2019; ALKMIM et al., 2019). For each output variable, models with 10, 100, 200, 500, and 1,000 epochs were tested.

The hyperbolic tangent function, with output values ranging from -1 to 1, was used as the activation function in the hidden layers. This function is commonly employed in hidden layers of ANNs (FENG; LU, 2019). The identity or linear activation function was used as the activation function in the output layer.

To optimize the trial-and-error procedure, a grid search was conducted, considering all combinations of the numbers of hidden layers and neurons and the number of epochs, resulting in a total of 550 models for each normalization method and output variable. The best hyperparameter combination, resulting in the model with the lowest error in

the training set, was defined based on the grid search results. The best model was then tested on the testing set to evaluate its generalization capacity. To prevent overfitting, simpler model structures with good performance in the grid search results were also tested to determine if they yielded better performance on the testing set.

In addition to testing simpler structures with the grid search, various and exhaustive attempts were made to test other network configurations and try to minimize model overfitting. Since there were numerous attempts, these methods will not be described in greater detail, as they did not achieve better results or minimize the models' overfitting, but they are described next. Other activation functions were tested in the hidden layers, namely the logistic sigmoid and the rectified linear unit (ReLU) functions. Outliers were removed from the data before starting the training process. The division of the data in a chronological order, rather than a random one, was tested. Cross-validation was also tested, in which the dataset was split into five subsets (folds), and the model was trained on four of these folds while being tested on the remaining one. This process was repeated five times, and the results were averaged. By using cross-validation, the risk of overfitting can be reduced. However, as mentioned, these techniques did not improve model performance or reduce overfitting of overfitted models.

4.3.2.4 Model performance

The following metrics were calculated to measure the prediction performance of the models: the root mean square error (RMSE, Equation (8)), mean absolute error (MAE, Equation (9)), and coefficient of determination (R^2, Equation (10)). The performance criteria formulas are as follows (NEWHART; HERING; CATH, 2022):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{8}$$

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \tag{9}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{10}$$

where y_i is the observed value, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the average of the observed values, and n is the number of observations.

4.3.2.5 Sensitivity analysis

After establishing the ANN models and assessing their performance, a sensitivity analysis was conducted to understand the relationship between the input and output phases of the ANN model (HAMEED et al., 2016). Sensitivity analysis approaches are used to rank the general relevance of the input variables for the output variable of the neural network model (PAPADOKONSTANTAKIS; LYGEROS; JACOBSSON, 2006). Hence, the influence of each water quality parameter and operational variable on the effluent quality was analyzed using this procedure to determine the variables with greater contributions to WWTP performance.

The sensitivity analysis was conducted using the Gedeon method (GEDEON, 1997), which was performed by the function “h2o.varimp” of the “h2o” package in R. The function displays each variable’s importance after it has been scaled between 0 and 1.

In the Gedeon method, the importance of an input neuron is determined by observing the effect of small perturbations or changes in the input neuron on the output of the neural network. If these perturbations lead to substantial changes in the output, it indicates that the input variable corresponding to that neuron is important and makes a notable contribution to the model's predictions (GEDEON, 1997).

4.3.3 Development of multiple linear regression models

As a form of comparison with the ANNs, multiple linear regression (MLR) models were developed for both the WWTPs. Developing simpler models such as MLR is an important practice, as it serves as a baseline to evaluate the performance of more complex models, such as ANNs. If the ANNs demonstrate better performance, their use is justified due to their higher accuracy and superior handling of nonlinear relationships between variables.

The same input and output variables were used in the MLR for each facility (Table 24). After removing the records with missing values, the data were also randomly divided

into 75% for training and 25% for testing. The data were not normalized, as this step is not necessary for MLR models. The function “lm” of base R was used to train the MLR models.

MLR is a statistical technique that aims to investigate and model the linear relationship between more than one explanatory variable and one response variable (MATHUR et al., 2024). The MLR model with *n* predictors is given by Equation (11).

$$y = a + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n + error \quad (11)$$

where y is the response variable; a is the intercept; $x_1, x_2, \ldots, x_n$ are the independent variables; $b_1, b_2, \ldots, b_n$ are the slope coefficients; and error is the difference between the observed and predicted y (VON SPERLING; VERBYLA; OLIVEIRA, 2020).

The MLR's performance was assessed using the same metrics as the ANN models (RMSE, MAE, and R^2) for both training and testing datasets. As the focus of the study is the ANN, the MLR results were presented in an appendix and discussed and compared with the ANNs in the text.

4.4 RESULTS AND DISCUSSION

4.4.1 Brasília Sul wastewater treatment plant

4.4.1.1 Total nitrogen model

After selecting all the input variables and output (TN_{efl}), the removal of observations with missing data from the selected dataset resulted in 251 out of the initial 331 samples. The data were then divided into 191 observations for training the model and 60 observations for testing its performance.

After conducting multiple tests to identify the best model architecture, the chosen model employed Z-score normalization as the preprocessing technique. The neural network was trained with 1,000 epochs and consisted of two hidden layers with 9 neurons each. The selected model's structure is presented in Figure 32. The figures illustrating the model structures in this chapter, including Figure 32, were created using NN-SVG (Neural-Network Scalable Vector Graphics), a tool developed by LeNail (2019).

Figure 32 – Structure of the neural network model for the prediction of effluent TN concentrations at the Brasília Sul WWTP

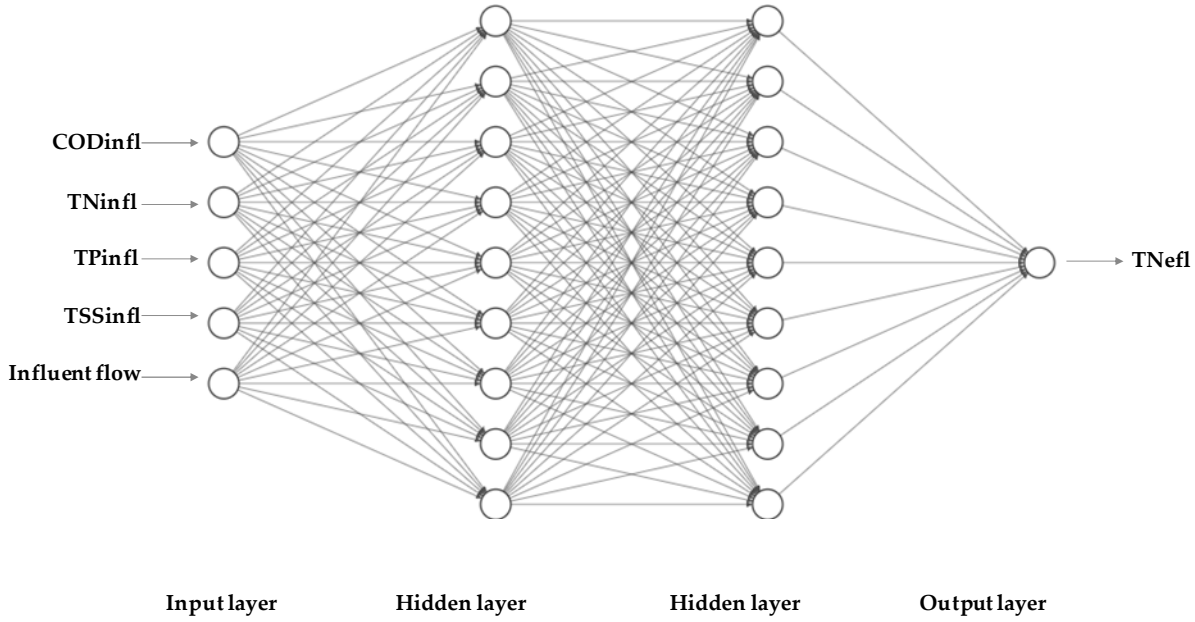


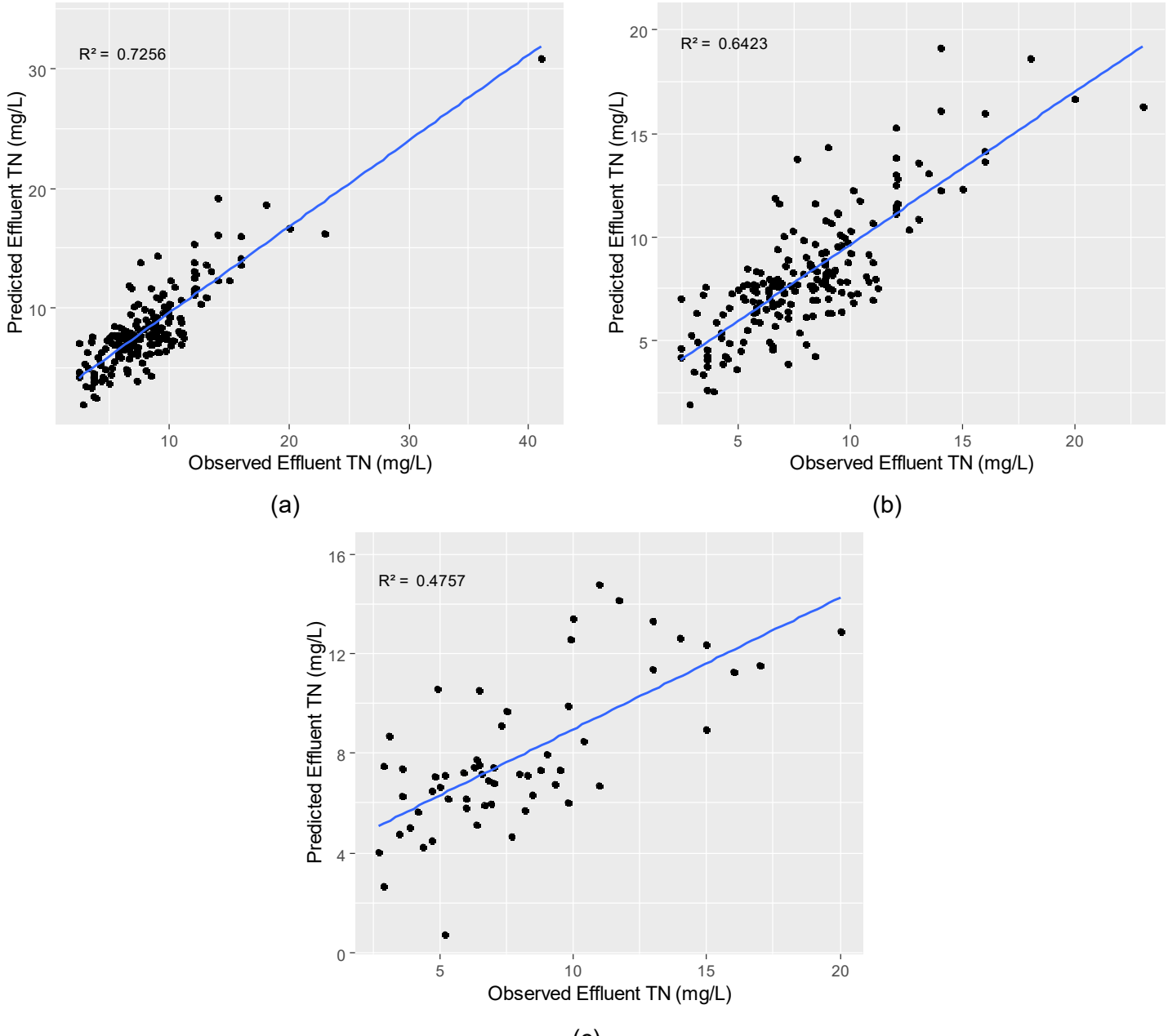
Table 25 presents the results of RMSE and MAE for both training and testing datasets. Figure 33 depicts the regression plot between the predicted and observed data of effluent TN also for the training and testing datasets. As the outlier of the observed TN of 41.0 mg/L was affecting the R^2 result in the training data (Figure 33 (a)), making the data fit appear better than it actually is, this outlier was removed. Then, a new plot was created and R^2 was recalculated (Figure 33 (b)) to show the more realistic performance of the model.

Table B.1 and Figure B.1 (Appendix B) show the MLR model's results for the effluent TN. The ANN model outperformed the MLR model in both the training and testing datasets.

Table 25 – Performance metrics of the model for training and testing datasets for the prediction of effluent TN

	Training (mg/L)	Testing (mg/L)
RMSE	2.099	2.738
MAE	1.565	2.157

Figure 33 – Regression plot between predicted and observed data of effluent TN for (a) training, (b) training excluding the outlier, and (c) testing datasets



Considering the error metrics of the ANN model, the MAE values were lower than the RMSE values for both training and testing subsets. This result was anticipated because

the MAE depends on absolute errors, as opposed to the squared errors of the RMSE, so it is less influenced by large differences between the actual and modeled values (NEWHART; HERING; CATH, 2022).

The low values of RMSE and MAE and the high value of R^2 for the training data indicate that the model effectively captured the complexity of the training data and can explain the variability in the dependent variable (effluent TN) well. However, the model performed better on the training subset compared to the testing subset, which suggests the possibility of overfitting. Overfitting occurs when the model has more parameters than necessary to capture the overall pattern (NEWHART; HERING; CATH, 2022) and ANNs are highly prone to overfitting (BAHRAMIAN et al., 2023). Models with fewer neurons in each hidden layer or fewer epochs were tested, and the selected one (1,000 epochs and 9 neurons in each hidden layer) had the best performance even when testing the models.

Real-world processes are subjected to a lot of fluctuations in the influent stream and operating conditions, mainly due to natural limitations and varying process dynamics. This high variability in the data may not be captured by the prediction models, which can lead to inaccurate results for the models trained using historical data from full-scale WWTPs (BAHRAMIAN et al., 2023; SAFEER et al., 2022; WANG et al., 2024a).

In their study, Liu, Huang and Yoo (2013) employed a hybrid learning method that combined a genetic algorithm with an adaptive neuro-fuzzy inference system (ANFIS-GA) to estimate effluent nutrient concentrations in a WWTP located in the Republic of Korea. The data used for training and testing the model comprised 357 samples that were measured between March 2007 and February 2008. The variables input of the models included influent flow, TSS, BOD, COD, TN, and TP, as well as the effluent COD, TN, and TP from the previous day. The results obtained for predicting effluent TN were $R^2 = 0.796$ for the training dataset and $R^2 = 0.577$ for the testing dataset (LIU; HUANG; YOO, 2013), which are similar to the results obtained in this study.

Nourani, Elkiran and Abba (2018) utilized a FFNN model to predict TN_{efl} in a WWTP in Cyprus, using daily data obtained from two years of monitoring. The input variables of the model comprised the influent pH, electrical conductivity, BOD, COD, and TN. The authors obtained an R^2 value of 0.9343 for the training dataset; and an R^2 value of

0.9022 for the testing dataset (NOURANI; ELKIRAN; ABBA, 2018). These results demonstrate higher accuracy than the model developed for the Brasília Sul WWTP.

Figure 34 shows the result of the relative variable importance of the explanatory variables for the prediction of the effluent TN. The influent quality indicators were more important than the influent quantity (flow) for this model. It would be expected that influent TN concentrations would be the most important predictor for effluent TN concentrations (LIU et al., 2023). However, this was not observed in this model.

Figure 34 – Relative variable importance for the TN effluent model

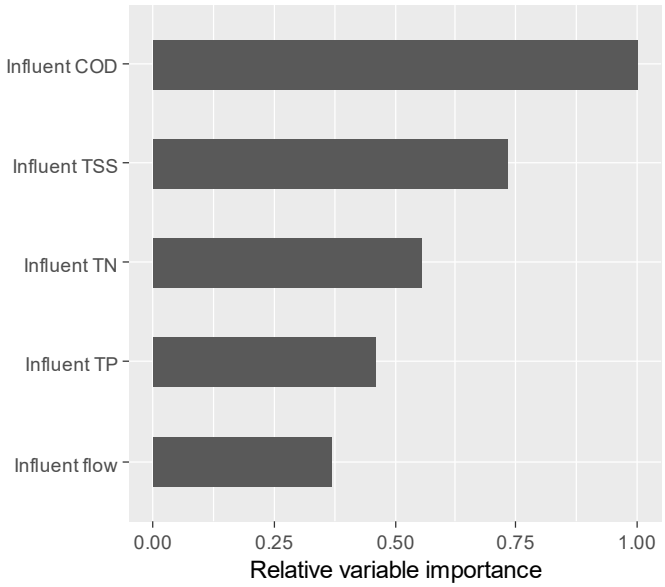
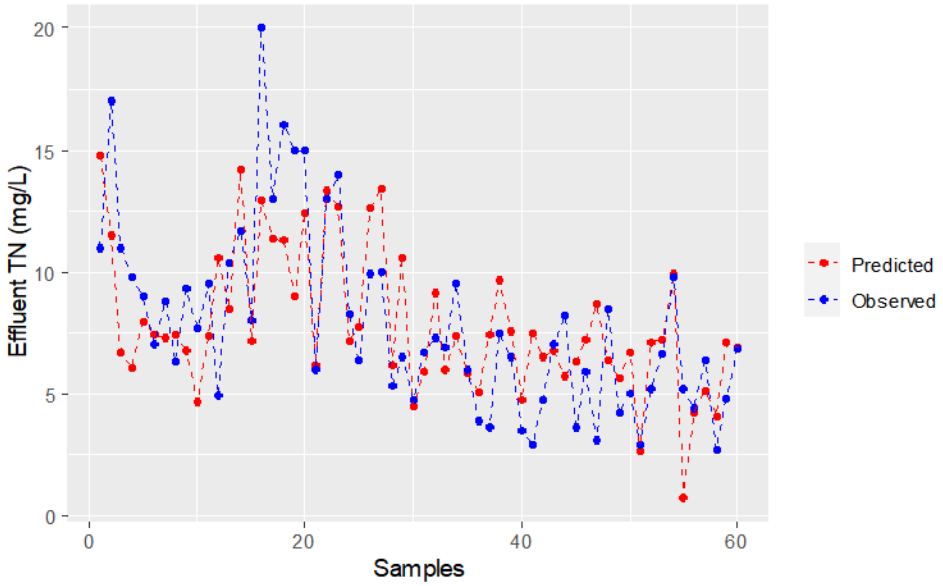


Figure 35 depicts a comparison between the predicted and observed effluent TN concentrations in the testing subset of the Brasília Sul WWTP model. Although there were some differences between the observed and predicted values, as indicated by the error metrics in Table 25, it is evident that the model was able to capture certain variability in the data. Specifically, the model was able to capture trends when the concentrations increased or decreased. The models developed using data preprocessed by the min-max normalization technique had a lower ability to capture higher concentrations, including those that exceeded the discharge limit of 8.7 mg/L. This was another criterion for selecting the Z-score normalization method.

Figure 35 – Comparative plot between the predicted and observed effluent TN concentrations for the testing dataset

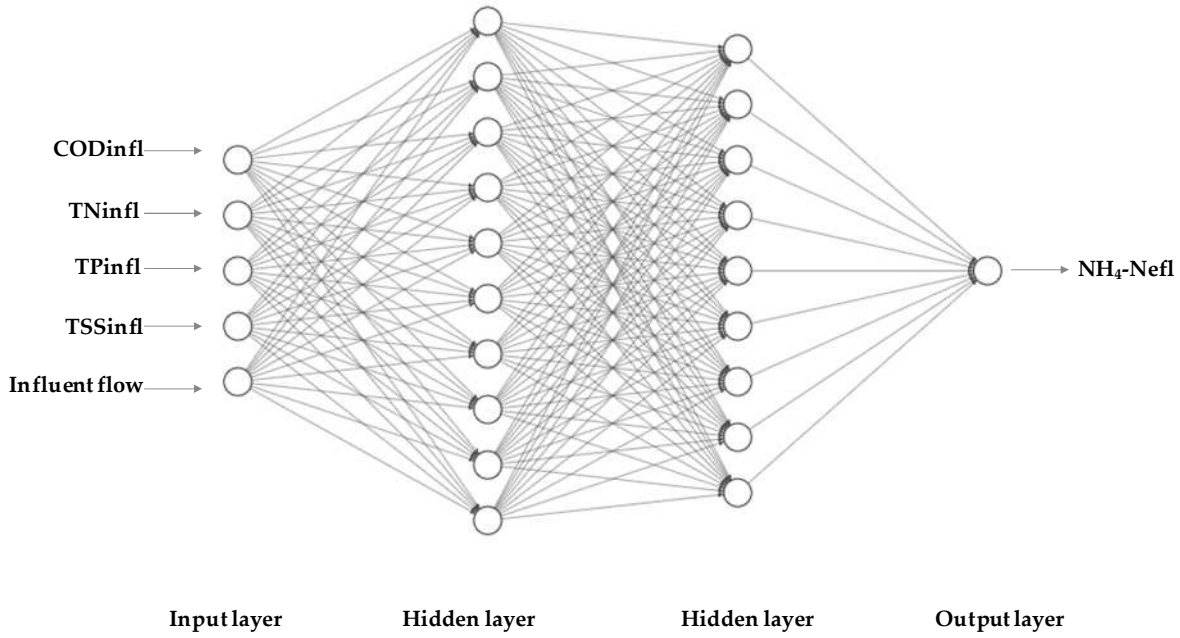


4.4.1.2 Ammonia nitrogen model

After selecting all input variables and the output variable (NH_4-N_{efl}), the removal of missing records from the selected dataset reduced the sample size from 331 to 250. The data were then divided into two sets, one for training and the other for testing the model. The training set comprised 178 observations, while the testing set comprised 72 observations.

After conducting numerous tests to determine the most effective model architecture, the selected model utilized Z-score normalization as the preprocessing technique. This model consisted of a neural network with two hidden layers, containing 10 and 9 neurons, respectively, and was trained over 1,000 epochs. The structure of the selected model is illustrated in Figure 36.

Figure 36 – Structure of the neural network model for the prediction of effluent NH_4-N concentrations at the Brasília Sul WWTP



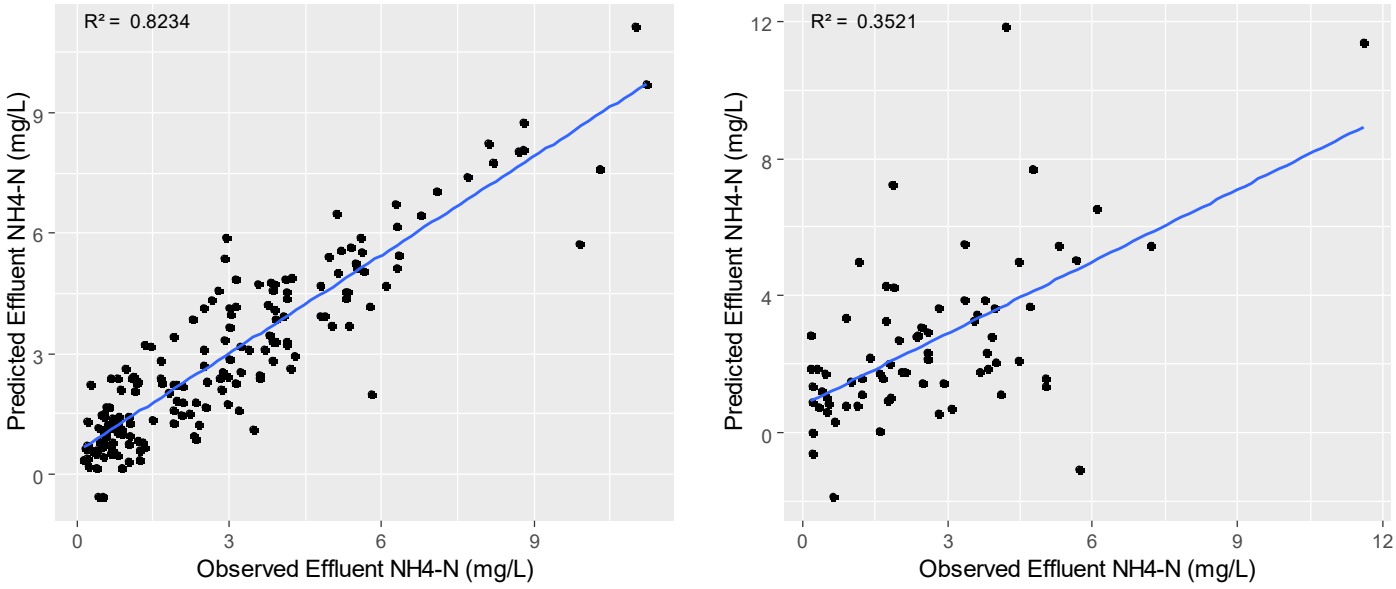
The results of the RMSE and MAE for the training and testing datasets are presented in Table 26. As observed in the TN model, the MAE results were lower than RMSE for both training and testing of the NH_4-N model. This result shows that there are some large individual errors that increase the overall error of the model.

Table 26 – Performance metrics of the model for training and testing datasets for the prediction of effluent NH_4-N

	Training (mg/L)	Testing (mg/L)
RMSE	1.007	1.996
MAE	0.764	1.350

Figure 37 depicts the regression plot between the predicted and observed effluent NH_4-N data for both the training and testing datasets.

Figure 37 – Regression plot between predicted and observed data of effluent NH_4-N for (a) training and (b) testing datasets



(a) (b)

The NH_4-N model exhibited similar behavior to the TN model, as a superior performance was obtained for the training set. Nevertheless, simpler model architectures produced comparable errors and goodness-of-fit metrics in evaluating the testing set. Consequently, achieving an optimized configuration that would enhance the algorithm's ability to generalize better was not feasible.

Table B.2 and Figure B.2 (Appendix B) show the performance of the MLR model for the prediction of effluent NH_4-N. The ANN performed better than MLR during training, with lower error metrics and higher R^2, showing that a more complex algorithm, such as ANN, is necessary to model the complex and nonlinear relationships that exist in a full-scale WWTP. However, the error metrics were slightly lower for the MLR while testing the model, which highlights the overfitting in the training data that ANN was subjected to.

Jami, Mujeli and Kabbashi (2011) developed a FFNN model to predict the effluent NH_4-N concentrations in a WWTP in Malaysia. The model was based on weekly monitoring data collected over a four-year period, with influent BOD, NH_4-N, pH, and inflow rate as inputs. The network consisted of a single hidden layer with 15 neurons, and the authors reported R^2 values of 0.6861 and 0.6368 for training and testing, respectively (JAMI; MUJELI; KABBASHI, 2011). In comparison, the NH_4-N model trained using data

from Brasília Sul WWTP demonstrated a higher R^2 value for the training dataset relative to the study conducted by Jami, Mujeli and Kabbashi (2011). However, the testing dataset for the Brasília Sul WWTP model exhibited poorer performance.

Zhao, Chai and Yuan (2012) developed an extreme learning machine (ELM) neural network model to predict the effluent quality of a WWTP in China. The authors employed 17 variables as input to the model, including influent quality indicators and operational variables such as oxidation-reduction potential in the anoxic and aerobic tanks, aeration flow in the bioreactor, and dissolved oxygen in the anoxic tank. The authors reported the prediction results of the effluent NH_4-N concentration by the normalized RMSE, which was 1.9736, and the R^2, which was 0.8273 during testing (ZHAO; CHAI; YUAN, 2012).

Gaya et al. (2014) developed a model for an activated sludge WWTP in Malaysia. The authors employed influent BOD, COD, TSS, NH_4-N, and O&G as inputs to the FFNN and ANFIS models. The ANFIS model accurately predicted effluent NH_4-N, with R^2 values of 0.9980 for training and 0.8987 for testing. The FFNN achieved high accuracy during training (R^2 = 0.9960) but demonstrated lower generalization ability (R^2 = 0.1806 for testing) (GAYA et al., 2014), which was even lower than the value obtained for Brasília Sul WWTP in this study.

Hazali, Wahab and Ibrahim (2017) developed a model for a WWTP using a self-organizing radial basis function neural network. The input variables included influent concentrations of COD, total organic carbon, TN, TP, NH_4-N, and mixed-liquor suspended solids (MLSS). During training, the model achieved a high R^2 value for predicting effluent NH_4-N concentrations (R^2 = 0.9195). However, achieving good generalization was difficult, as the R^2 value during testing was low (R^2 = 0.2833) (HAZALI; WAHAB; IBRAHIM, 2017), similar to the behavior observed in modeling the NH_4-N effluent concentrations of Brasília Sul WWTP.

Khatri, Khatri and Sharma (2019) developed FFNN models to predict effluent quality parameters of a WWTP in India based on influent characteristics, including pH, BOD, COD, TSS, TKN, NH_4-N, and TP. The authors obtained a dataset of 180 samples from the monitoring period between January 2017 and December 2017. The optimum structure for predicting the effluent NH_4-N concentration was a single hidden layer with

six neurons, resulting in R^2 values of 0.2540 for training and 0.2430 for the complete dataset. Although the authors concluded that the ANN modeling showed a strong correlation between the measured values of the influents and the predicted values of effluents (KHATRI; KHATRI; SHARMA, 2019), the low R^2 values indicated that there was no good fitting of the predicted values to the observed values.

Al-Ghazawi and Alawneh (2021) developed a FFNN to model an extended aeration activated sludge WWTP in Jordan. The dataset for the study comprised 487 records obtained from daily monitoring between June 2001 and October 2002. The dataset was divided into training (75%) and testing (25%) sets. The model used influent characteristics, namely flow rate, temperature, pH, BOD, COD, TSS, and NH_4-N to predict effluent NH_4-N concentrations. Although the model had a high accuracy of R^2 = 0.965 during training, its generalization to unseen data was not very accurate, with an R^2 of 0.260 during testing (AL-GHAZAWI; ALAWNEH, 2021).

Wang et al. (2024a) fitted an RBF model for the prediction of effluent NH_4-N in a WWTP in China. Input variables were influent COD, TN, TP, pH, and NH_4-N collected over a period of two years, resulting in 600 observations. The authors found an R^2 of 0.58 when testing the model and an RMSE of 0.1407 mg/L, concluding that the RBF model predicted effluent ammonia nitrogen well (WANG et al., 2024a). Nevertheless, an R^2 value of 0.58 suggests that the model's predictive accuracy may not be as robust as the authors claimed.

Figure 38 shows the relative importance of the input variables for the prediction of the effluent NH_4-N concentrations at Brasília Sul WWTP. Al-Ghazawi and Alawneh (2021) found that the ANN model to predict effluent NH_4-N concentrations was highly sensitive to influent temperature, pH, BOD, and COD; moderately sensitive to influent NH_4-N; and slightly sensitive to influent flow and TSS. Although the explanatory variables in the present study are not the same, for the Brasília Sul WWTP model, the influent TN (which is the most similar to influent NH_4-N) was the most important variable, and influent COD the least important, which differs from the results of Al-Ghazawi and Alawneh (2021). However, in both studies, the influent flow was among the least important variables.

Figure 38 – Relative variable importance for the NH_4-N effluent model

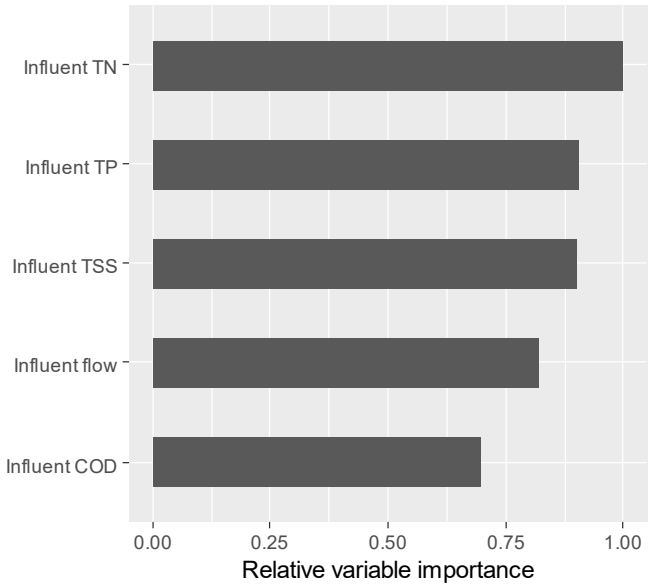
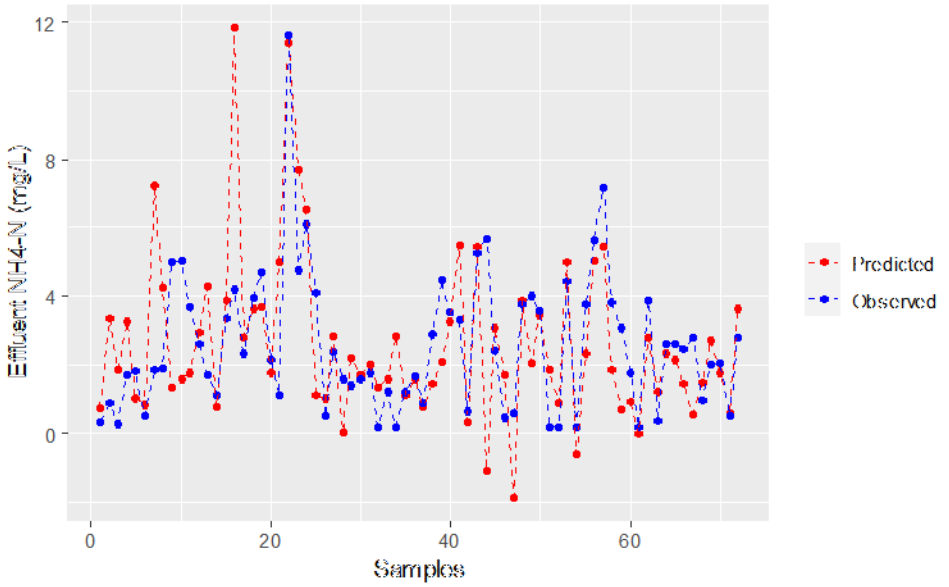


Figure 39 presents a comparison of the predicted and observed effluent NH_4-N concentrations for the testing dataset of the Brasília Sul WWTP. As observed in the TN model, the NH_4-N model exhibited some error between the predicted and observed values (Table 26). However, the model demonstrated a certain ability to capture the variability in the data (Figure 39). The chosen model, trained and tested with data preprocessed by the Z-score normalization method, could better capture the high NH_4-N concentrations, sometimes even overestimating them. It is crucial to model high values accurately as they represent the environmental risk in effluent discharge. Therefore, the Z-score normalization method was selected instead of the min-max normalization technique.

Figure 39 – Comparative plot between the predicted and observed effluent NH_4-N concentrations for the testing dataset



4.4.1.3 Total phosphorus model

After selecting the input variables and the output (TP_{efl}), the removal of records with missing values in the selected dataset resulted in 263 out of the 331 initial samples. Subsequently, the data was divided into 196 observations for training and 67 observations for testing the model.

After conducting multiple tests to determine the optimal model architecture, the selected model consisted of a neural network with 1,000 epochs and two hidden layers, each with 9 neurons. The model also utilized Z-score normalization as the preprocessing technique. Figure 40 displays the structure of the chosen model.

Figure 40 – Structure of the neural network model for the prediction of effluent TP concentrations at the Brasília Sul WWTP

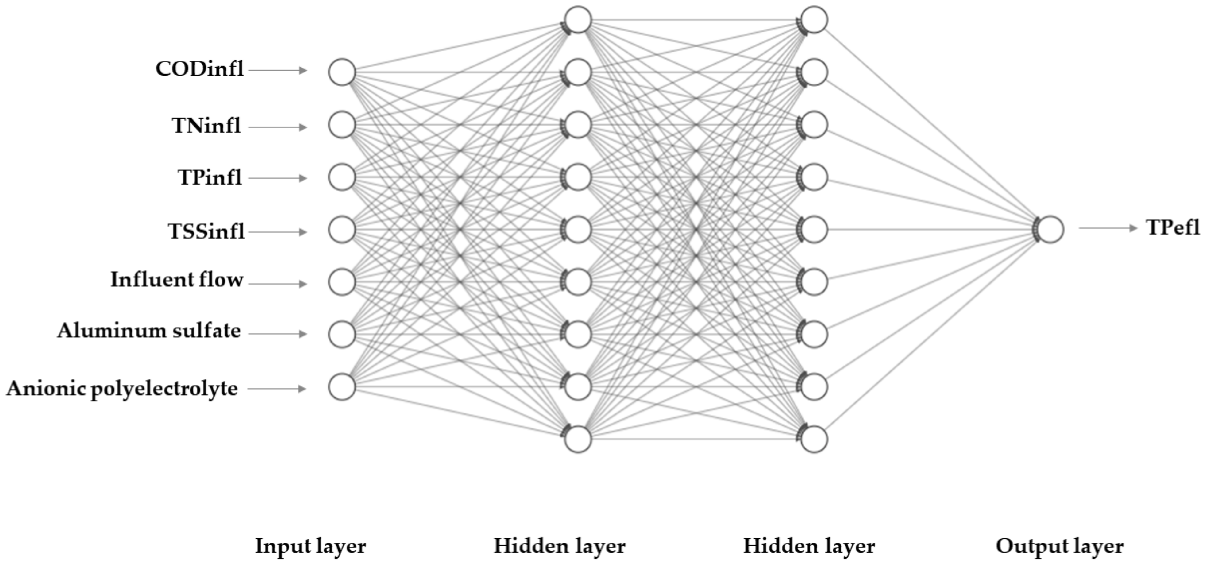


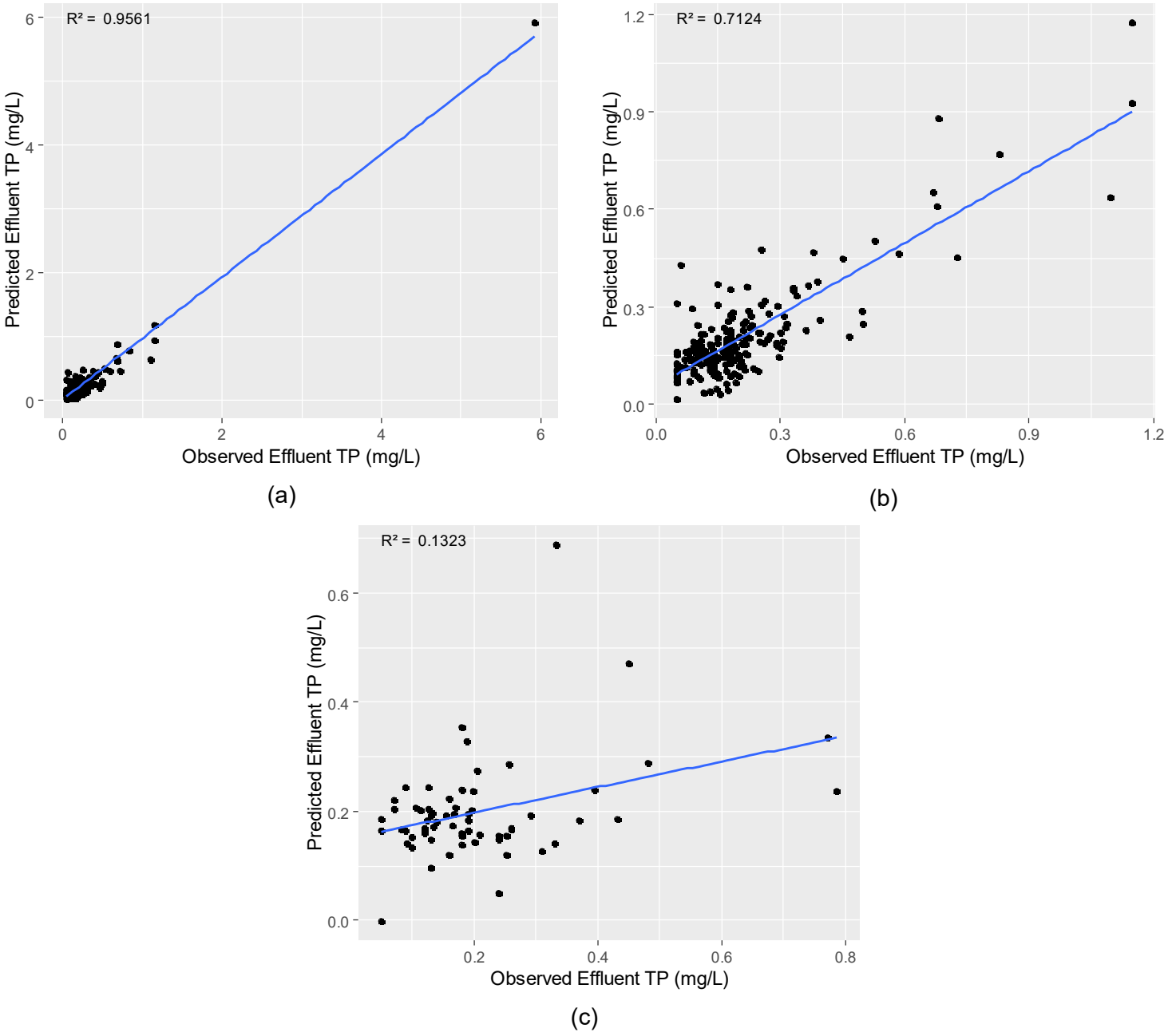
Table 27 presents the RMSE and MAE values for training and testing datasets. Figure 41 displays the regression plot depicting the predicted and observed data of effluent TP for both the training and testing datasets. The outlier of 5.92 mg/L of the observed TP in the training data was making the data fit appear better than it is (Figure 41 (a)). It was removed, a new plot was created, and the R^2 was recalculated (Figure 41 (b)).

Table B.3 and Figure B.3 (Appendix B) display the performance of the MLR model for predicting the effluent TP concentrations. For all metrics (RMSE, MAE, and R^2) the results were better for the ANN model when compared to the MLR results.

Table 27 – Performance metrics of the model for training and testing datasets for the prediction of effluent TP

	Training (mg/L)	Testing (mg/L)
RMSE	0.093	0.137
MAE	0.066	0.098

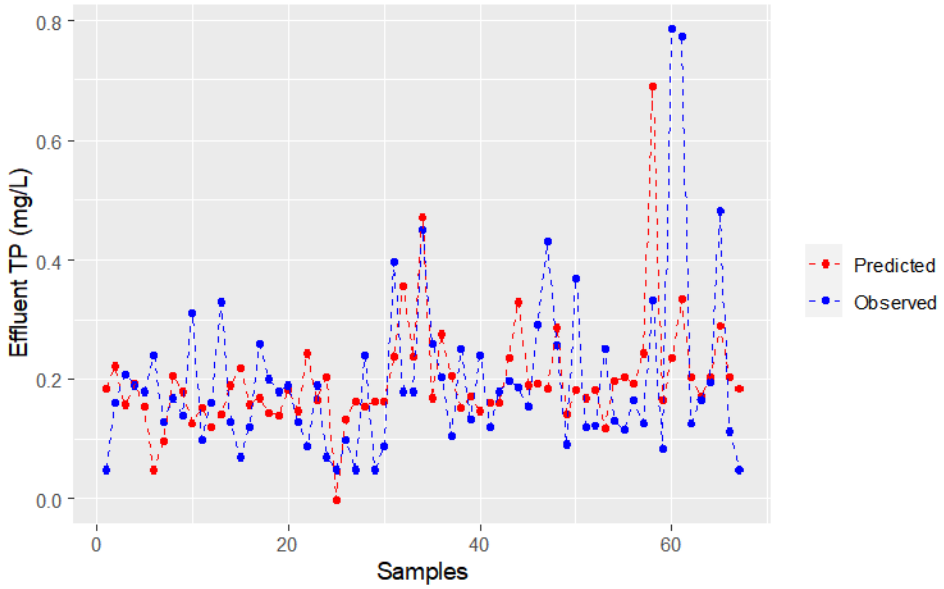
Figure 41 – Regression plot between predicted and observed data of effluent TP for (a) training, (b) training excluding the outlier, and (c) testing datasets



Based on the results, it is evident that the ANN model could train accurately. When it came to testing the model on new observations, although low RMSE and MAE values were obtained, showing that the model had a low error when predicting values, a low value or R^2 was obtained. The R^2 metric indicates how well the model captures the variability in the data. So, while the errors were small, the model did not capture the underlying relationships or patterns in the data well. Figure 42 shows the comparison

of the predicted and observed effluent TP concentrations for the testing dataset of the Brasília Sul WWTP.

Figure 42 – Comparative plot between the predicted and observed effluent TP concentrations for the testing dataset



It is often difficult to find the optimal structure for ANNs. Network learning is performed over the training data, which can differ greatly from testing data. In this scenario, the model's generalization ability is reduced when applied to unseen data (BAHRAMIAN et al., 2023). For the development of the TP_{efl} model of the Brasília Sul WWTP, the removal of missing records resulted in the reduction of 20% of the samples of the dataset. This led to a low number of records for testing the model, which may not represent accurately the variability of the training dataset. Roohi, Nazif, and Ramazi (2024) found that, as the percentage of missing values increased in the datasets used to fit machine learning models for predicting effluent quality in WWTPs, the models' performance was affected, especially when the missing data exceeded 10% (ROOHI; NAZIF; RAMAZI, 2024). The impact of missing data is even higher for small datasets.

Liu, Huang and Yoo (2013) developed an ANFIS-GA model to predict effluent TP in a full-scale WWTP. The input variables were the same as those used for the TN model, including influent flow, TSS, BOD, COD, TN, and TP, as well as the effluent COD, TN, and TP of the previous day. The authors achieved a training R^2 of 0.595 and a testing R^2 of 0.284, indicating inadequate performance for both training and testing. According

to the authors, the reason for the low accuracy of the model was that the effluent TP showed relatively regular variations that were not influenced by seasonality, whereas the independent variables had seasonal variations, resulting in significant prediction errors (LIU; HUANG; YOO, 2013). This phenomenon is not the reason for the low accuracy of the model for TP_{efl} at the Brasília Sul WWTP, as TP effluent concentrations showed significant differences between the rainy and dry periods (results shown in Chapter 3). In the study of Liu, Huang and Yoo (2013), the authors recommended incorporating additional measurement variables, such as effluent TSS and BOD and chemical dosage for TP control, to enhance their model's accuracy (LIU; HUANG; YOO, 2013).

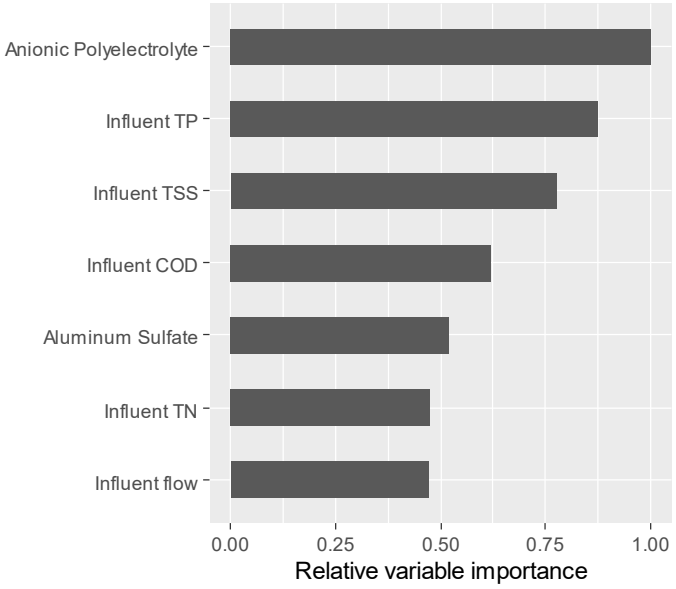
Khatri, Khatri and Sharma (2019) developed FFNN models to predict the effluent quality parameters of a WWTP in India based on influent characteristics. The optimal structure for predicting the effluent TP concentration was a single hidden layer with six neurons, using the same input variables as the NH_4-N model. The TP modeling resulted in a R^2 of 0.5944 for training and a R^2 of 0.5595 for the complete dataset (KHATRI; KHATRI; SHARMA, 2019).

Yaqub et al. (2020) developed a long short-term memory (LSTM)-based neural network to predict the removal efficiency of NH_4-N, TN, and TP in an A^2O MBR system in the Republic of Korea. The input variables considered were influent characteristics and operating parameters, such as dissolved oxygen, oxidation-reduction potential, and MLSS. Each parameter was recorded hourly throughout the one-year study, resulting in 6,876 samples used to train and test the model. The average MSE for the removal efficiencies were 2.85% for NH_4-N, 17.15% for TN, and 8.88% for TP (YAQUB et al., 2020), indicating that the model performed well in predicting NH_4-N and TP removal but had relatively higher errors for TN removal efficiencies.

Xu et al. (2024) employed machine learning models to predict effluent TP in a small-scale activated sludge WWTP in Illinois, USA, using nine years of data. Fifteen input variables were used, including water quality parameters, temperature, and influent flow rate. ANN model had $R^2 = 0.69$ in the training data and $R^2 = 0.55$ in the testing data, which implied overfitting of the model, according to the authors (XU et al., 2024).

Figure 43 shows the relative importance of the explanatory variables for the prediction of the effluent TP. Influent TP was among the most important variables for the model, as expected. Influent quality indicators and the consumption of chemical products were more important than the influent flow. However, since the model was overfitted to the training data, the result of the variable importance is valid for the data used to train the model, and it may not be adequate to extrapolate to other data.

Figure 43 – Relative variable importance for the TP effluent model



4.4.1.4 Chemical oxygen demand model

After selecting the input variables and the output (COD_{efl}), the removal of records with missing data in the selected dataset resulted in 264 out of the 331 initial samples. Subsequently, the data was divided into 201 observations for training and 63 observations for testing the model.

After conducting multiple tests to determine the optimal model architecture, the selected model consisted of a neural network with 500 epochs and two hidden layers, with 8 and 4 neurons, respectively. The model also utilized Z-score normalization as the preprocessing technique. Figure 44 displays the structure of the chosen model.

Figure 44 – Structure of the neural network model for the prediction of effluent COD concentrations at the Brasília Sul WWTP

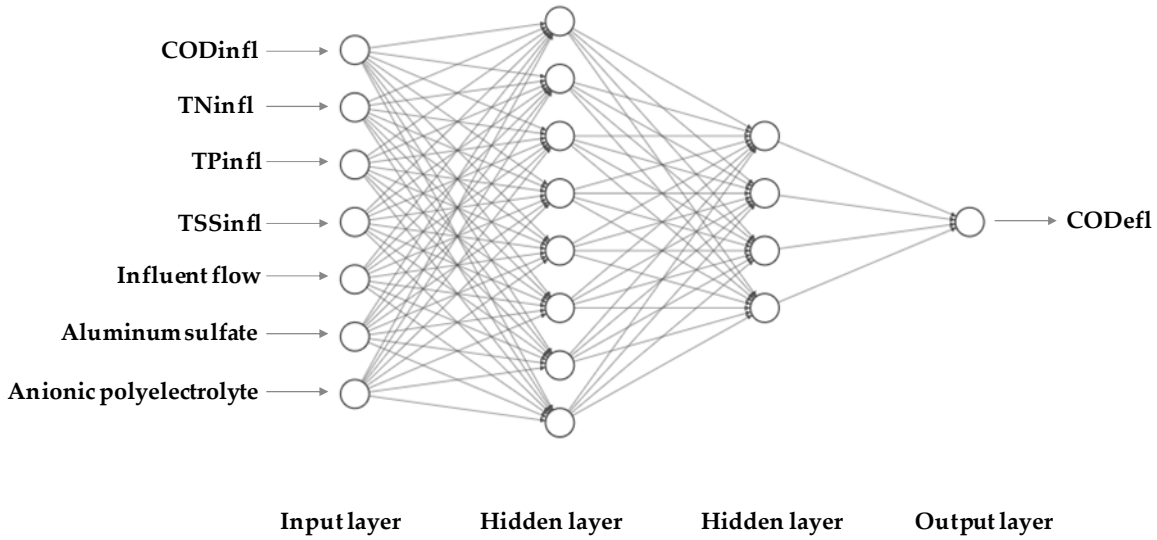


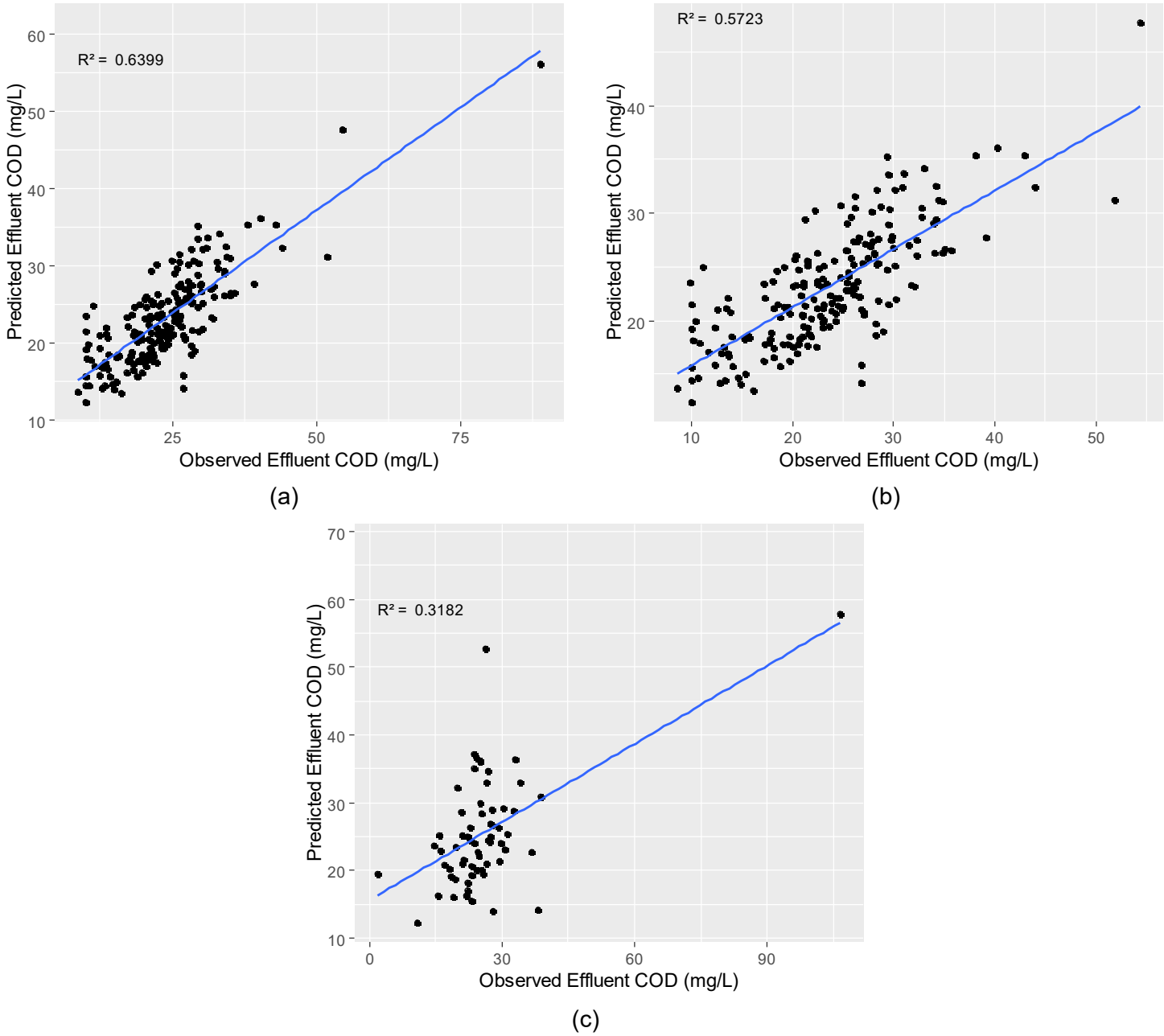
Table 28 presents the results of RMSE and MAE for both training and testing datasets. Figure 45 illustrates the regression plot between the predicted and observed data of effluent COD for both the training (with – Figure 45(a) – and without – Figure 45 (b) – the outlier of 106.6 mg/L of observed effluent COD) and testing (Figure 45 (c)) datasets.

Table B.4 and Figure B.4 (Appendix B) show the result of the MLR model for effluent COD. Except for the MAE value for testing, all other metrics showed better performance for the ANN model (lower RMSE and MAE and higher R^2) for both training and testing datasets. Even though it performed better than the MLR, the ANN model did not achieve satisfactory results for predicting effluent COD.

Table 28 – Performance metrics of the model for training and testing datasets for the prediction of effluent COD

	Training (mg/L)	Testing (mg/L)
RMSE	5.415	10.007
MAE	3.960	6.549

Figure 45 – Regression plot between predicted and observed data of effluent COD for (a) training, (b) training excluding the outlier, and (c) testing datasets



Alsulaili and Refaie (2021) studied a WWTP in Kuwait over seven years of data and 1,032 observations. Using influent temperature, pH, electrical conductivity, TSS, COD, and BOD as input variables, a FFNN model to predict COD_{efl} was constructed. The optimal network structure consisted of three hidden layers with 13 neurons in each, resulting in an R^2 of 0.6115. However, the authors did not specify whether this R^2 value pertained to the training dataset or the test dataset, which is concerning as it makes it difficult to accurately assess the model's generalization performance.

Saleh (2021) used a FFNN with a backpropagation training algorithm to model the effluent COD concentrations in a WWTP in Iraq. Daily data collected over two years from influent BOD, NH_4-N, TN, PO_4, NO_3-N, NO_2-N, pH, TSS, and temperature were used to train the model. The authors found an $R^2 = 0.62517$ in the training data, similar to the Brasília Sul WWTP model. However, they obtained a better generalization of their model as the R^2 was 0.59754 in the testing data.

Abba, Elkiran and Nourani (2021) fitted MLP and ELM models for the prediction of effluent COD concentrations in a WWTP in Cyprus. Influent pH, electrical conductivity, BOD, COD, TN, TP, NH_4-N, and TSS were used as explanatory variables, and data was collected daily, resulting in 392 observations. For the MLP model, the results in the training were $R^2 = 0.9617$ and RMSE = 0.0689, and for testing $R^2 = 0.9555$ and RMSE = 0.0648. For the ELM model, the training data resulted in $R^2 = 0.9757$ and RMSE = 0.0208, and the testing in $R^2 = 0.9742$ and RMSE = 0.0515, highlighting the high level of accuracy for both models.

Figure 46 shows the comparison of the predicted and observed effluent COD concentrations for the testing dataset of the Brasília Sul WWTP. The model did not fit well to an outlier with a value much higher than the other observed values. However, even with a low degree of model fit, there is some ability to predict data trends (Figure 46).

Figure 46 – Comparative plot between the predicted and observed effluent COD concentrations for the testing dataset

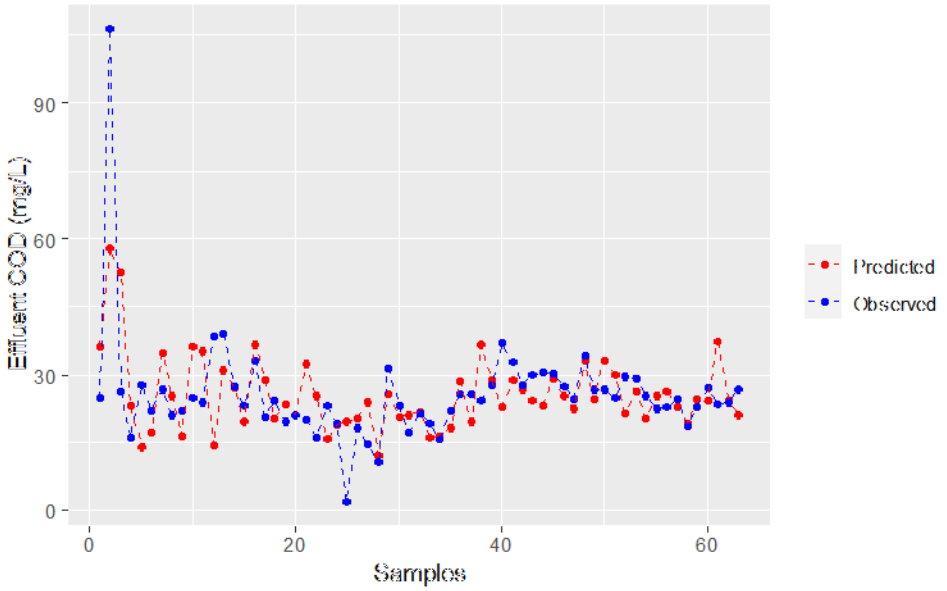
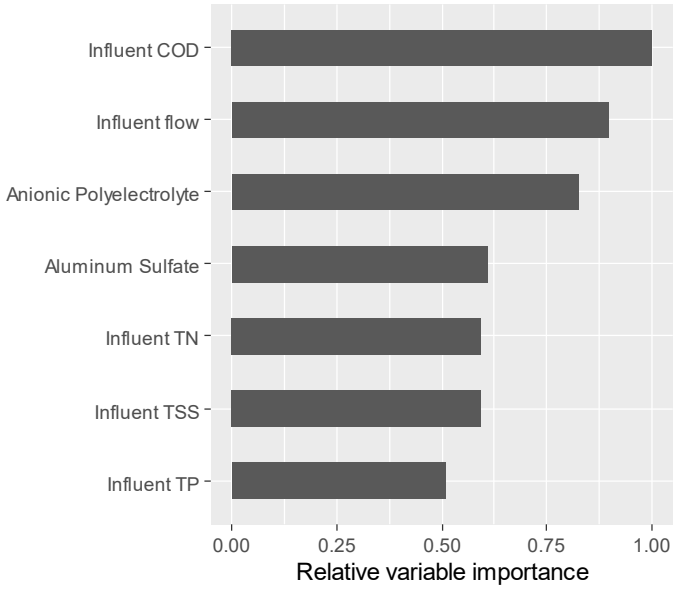


Figure 47 demonstrates the relative importance of the explanatory variables of the model developed to predict the effluent COD concentrations at Brasília Sul WWTP. The influent COD was the most important variable to predict the effluent COD concentrations. Roohi, Nazif, and Ramazi (2024) also found that an influent characteristic is the most important predictor in the model for predicting the same variable in the effluent, which is expected.

Figure 47 – Relative variable importance for the COD effluent model



4.4.1.5 Total suspended solids model

After selecting the input variables and the output (TSS_{efl}), the removal of observations with missing data resulted in 264 out of the 331 initial samples. Subsequently, the data was divided into 202 observations for training and 62 observations for testing the model.

After conducting multiple tests to determine the optimal model architecture, the selected model consisted of a neural network with 1,000 epochs and two hidden layers, with 10 and 6 neurons, respectively. The model also utilized Z-score normalization as the preprocessing technique. Figure 48 displays the structure of the chosen model.

Figure 48 – Structure of the neural network model for the prediction of effluent TSS concentrations at the Brasília Sul WWTP

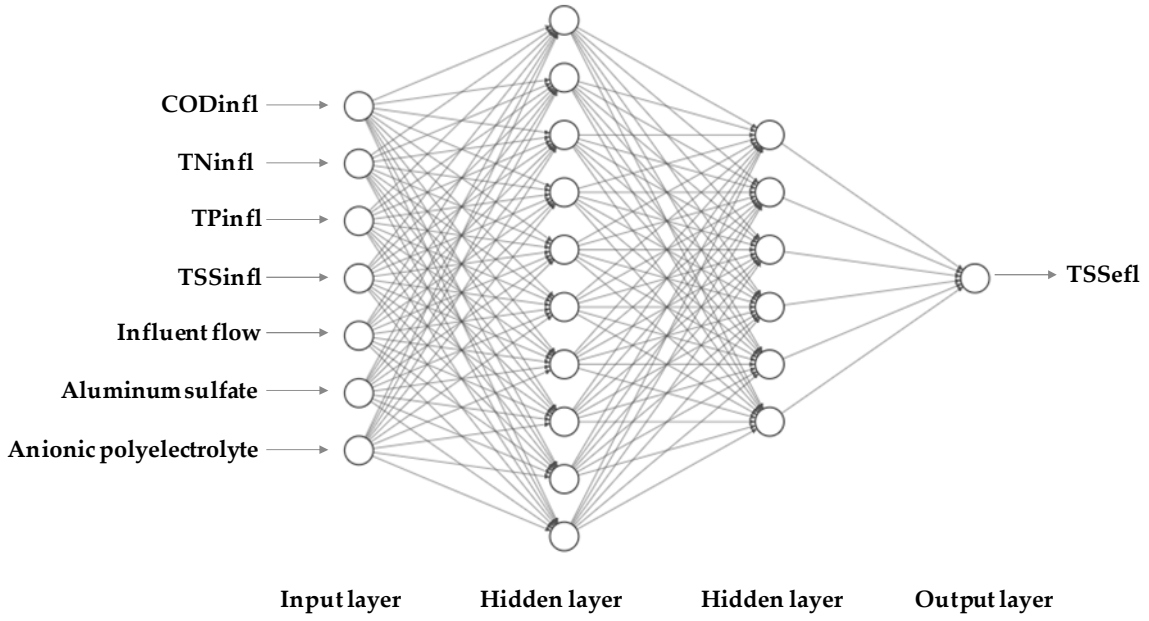
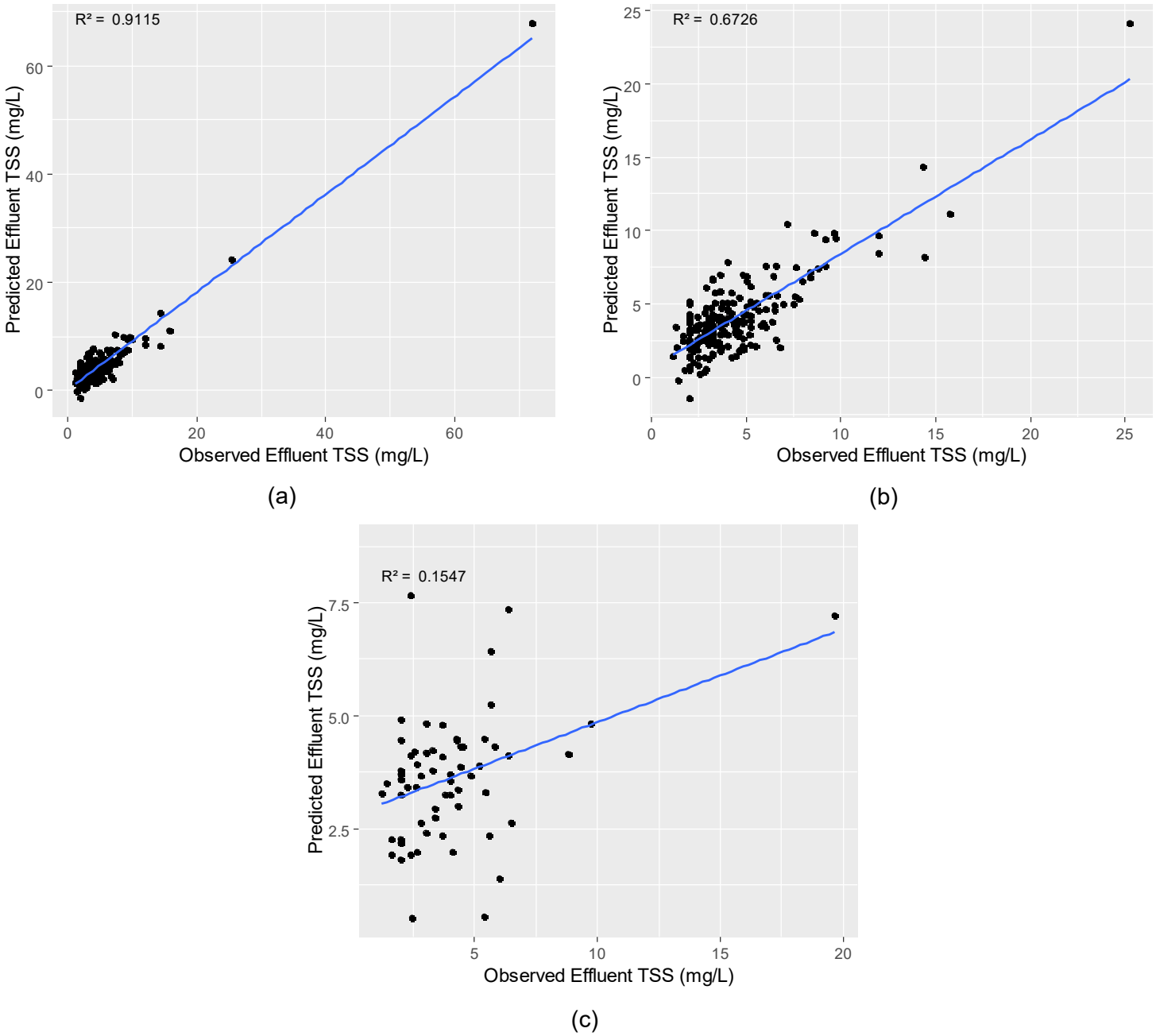


Table 29 presents the results of RMSE and MAE for both training and testing datasets. Figure 49 illustrates the regression plot between the predicted and observed data of effluent TSS for both the training (with – Figure 49 (a) – and without – Figure 49 (b) – the outlier of 72.0 mg/L of observed effluent TSS) and testing (Figure 49 (c)) datasets. Table B.5 and Figure B.5 (Appendix B) show the results of the MLR model for the effluent TSS concentrations. The ANN performed better for all metrics, highlighting the nonlinear behavior of the data.

Table 29 – Performance metrics of the model for training and testing datasets for the prediction of effluent TSS

	Training (mg/L)	Testing (mg/L)
RMSE	1.664	2.486
MAE	1.278	1.615

Figure 49 – Regression plot between predicted and observed data of effluent TSS for (a) training, (b) training excluding the outlier, and (c) testing datasets



Al-Ghazawi and Alawneh (2021) developed a FFNN to predict effluent TSS using daily data from an extended aeration activated sludge WWTP in Jordan. The input parameters were influent flow, temperature, pH, BOD, COD, TSS, and NH_4-N. The dataset contained 487 daily records. The results indicated a certain overfitting of the ANN, as during training the model performed well with $R^2 = 0.79$, but during testing there was a worsening in the performance, with $R^2 = 0.44$.

Alsulaili and Refaie (2021) developed a FFNN to predict TSS_{efl} in a WWTP using influent temperature, pH, electrical conductivity, TSS, COD, and BOD as input variables. The optimal network structure consisted of four hidden layers with 11 hidden neurons in each and resulted in an R^2 of 0.6308.

Gholizadeh et al. (2024) applied an MLP ANN to predict TSS concentrations in a WWTP effluent in Iran. Data was collected daily from 2016 to 2020, resulting in 654 samples in the dataset. The study adopted feature selection methods to select the most efficient scenario, which resulted in influent concentrations of BOD and TN and effluent concentrations of BOD and COD as input variables. This scenario resulted in an $R^2 = 0.78$ in training and $R^2 = 0.80$ in testing, showing that the ANN efficiently predicted effluent TSS.

Figure 50 shows the comparison of the predicted and observed effluent TSS concentrations for the testing dataset of the Brasília Sul WWTP. As with the COD model, the TSS model did not fit well to an outlier with a value much higher than the other observed values. As for the other observed values, the model had a certain fit to the trends in the data.

Figure 50 – Comparative plot between the predicted and observed effluent TSS concentrations for the testing dataset

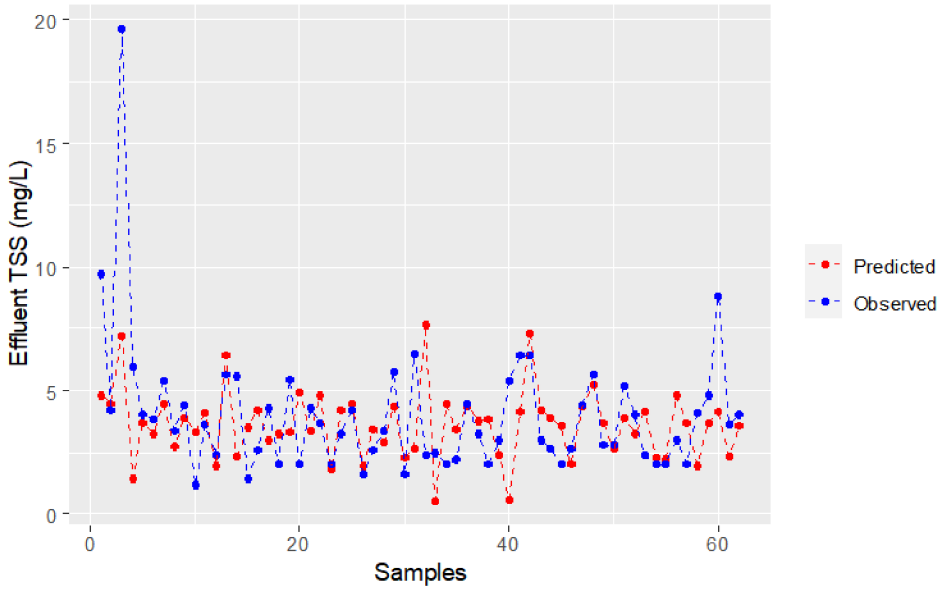
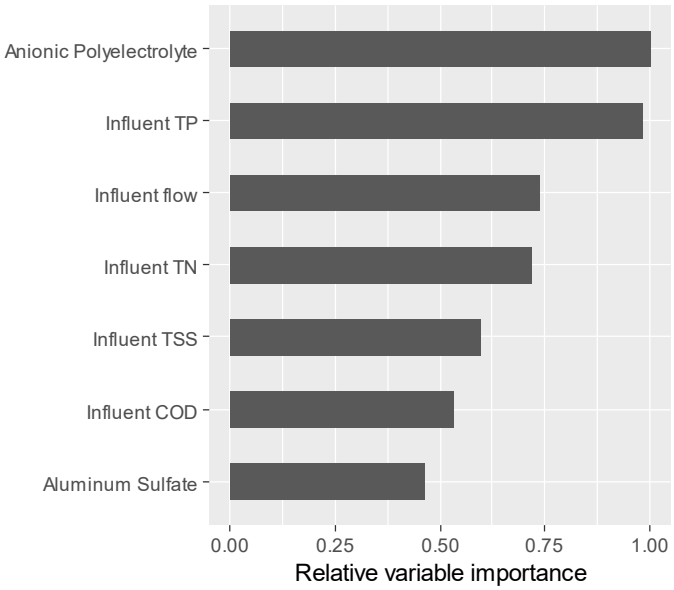


Figure 51 demonstrates the relative variable importance of the explanatory variables of the model developed to predict the effluent concentrations of TSS at Brasília Sul WWTP. In the study of Alsulaili and Refaie (2021), the effluent TSS was best predicted by the temperature, influent BOD, and influent TSS. Contrary to what was expected, the influent TSS was not among the most important variables for the Brasília Sul TSS_{efl} model. Another unexpected result is that the anionic polyelectrolyte was the most important variable, while the aluminum sulfate was the least. However, it is important to note that these results are valid for the training data that was used in this model. Because overfitting of the ANN to the training data occurred, it is not ideal to extrapolate the findings to conclude that the most important variables in Figure 51 are consistently the most influential in the removal of TSS at the facility.

Figure 51 – Relative variable importance for the TSS effluent model



4.4.2 John E. Egan wastewater treatment plant

4.4.2.1 Total phosphorus model

After selecting the input variables and the output (TP_{efl}) and removing data from the period with much lower effluent TP concentrations due to operational changes at the John E. Egan facility (from July 2007 to January 2009), the dataset contained 7,521 observations. Removing observations with missing values resulted in 4,181 observations. Subsequently, the data were divided into 3,092 observations for training and 1,089 observations for testing the model.

After conducting multiple tests to determine the optimal model architecture, the selected model consisted of a neural network with 1,000 epochs and two hidden layers, with 10 and 9 neurons, respectively. The model also utilized Z-score normalization as the preprocessing technique. Figure 52 displays the structure of the chosen model.

Figure 52 – Structure of the neural network model for the prediction of effluent TP concentrations at the John Egan WWTP

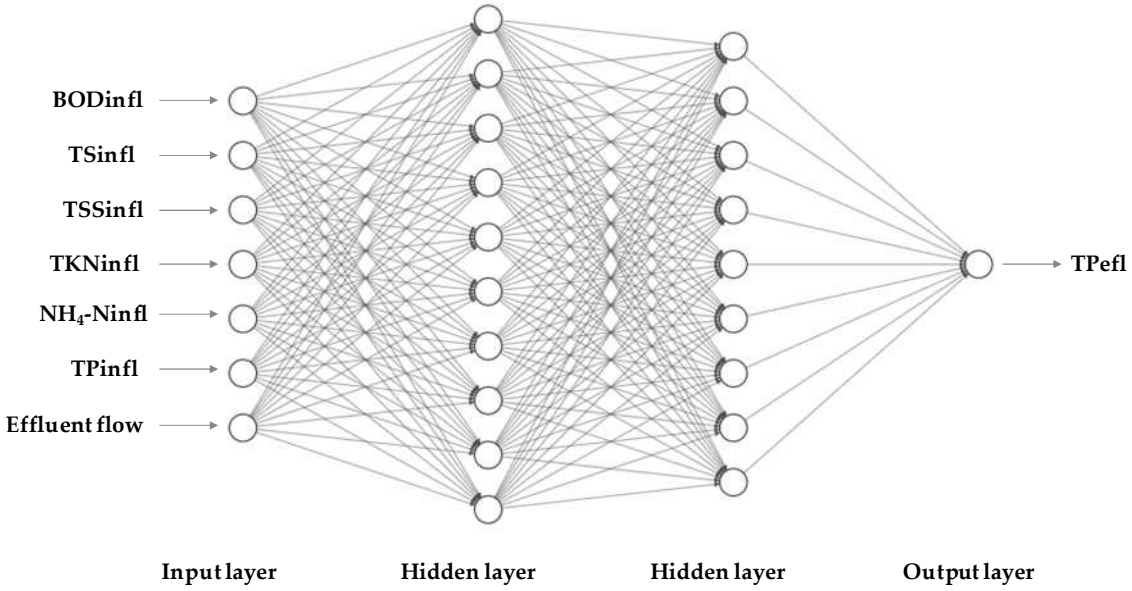
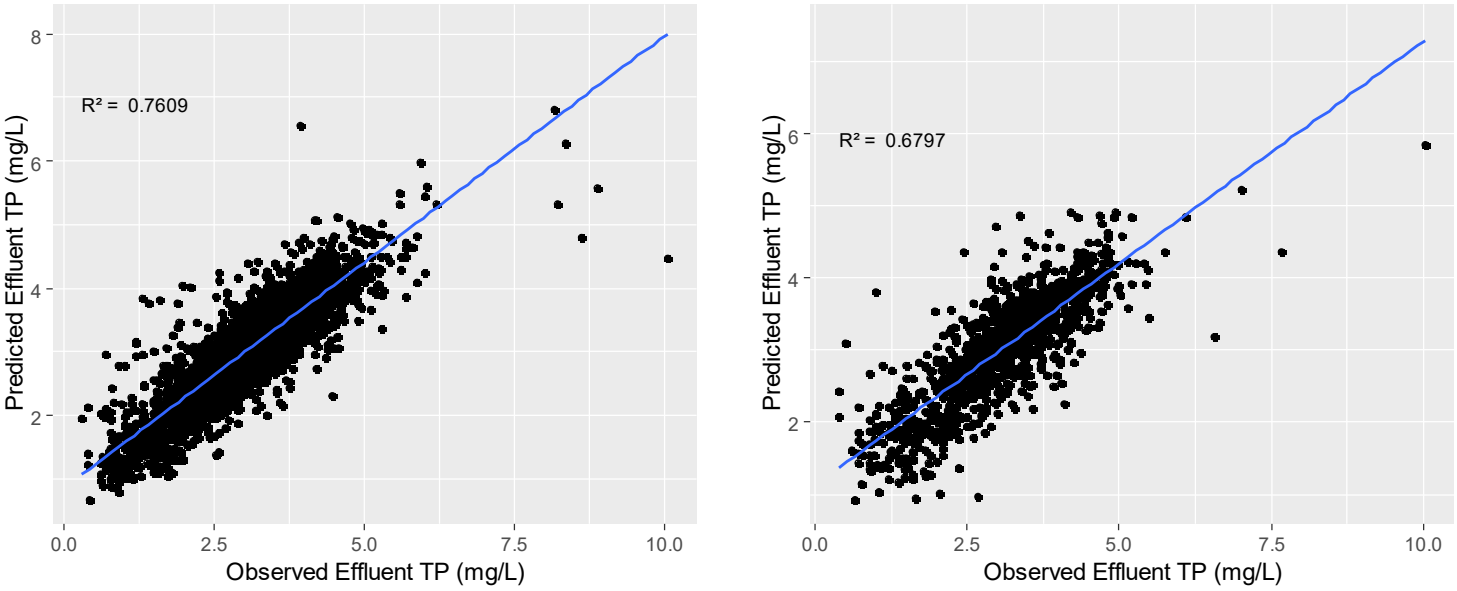


Table 30 presents the results of RMSE and MAE for both training and testing datasets. Figure 53 shows the regression plot between the predicted and observed effluent TP data for both the training and testing datasets. Table C.1 and Figure C.1 (Appendix C) show the results of the performance of the MLR model for predicting effluent TP concentrations at John Egan WWTP. The ANN model had a better prediction accuracy as the error metrics were lower and R^2 was higher for both training and testing datasets.

Table 30 – Performance metrics of the model for training and testing datasets for the prediction of effluent TP at John Egan WWTP

	Training (mg/L)	Testing (mg/L)
RMSE	0.494	0.585
MAE	0.361	0.428

Figure 53 – Regression plot between predicted and observed data of effluent TP at John Egan WWTP for (a) training and (b) testing datasets



(a) (b)

The ANN model developed to predict effluent TP concentrations at John Egan WWTP performed better, with higher R^2 (0.6797) for testing the model when compared to the Brasília Sul WWTP and other studies that developed predictive models for effluent TP (KHATRI; KHATRI; SHARMA, 2019; LIU; HUANG; YOO, 2013; XU et al., 2024). This could be due to the larger amount of data since over 4,000 observations were used to train and test the model of John E. Egan WWTP. For Brasília Sul WWTP, 263 observations were used, while the studies of Liu, Huang and Yoo (2013) had 357 samples and Khatri, Khatri and Sharma (2019), 180 samples.

Figure 54 shows a comparison between the predicted and observed effluent TP concentrations in the testing subset of the John E. Egan WWTP ANN model. This result demonstrates the high degree of fit between the predicted data and observed data during model testing, indicating a high level of generalization. The trained model showed a strong capability to generalize the results to unseen data previously unknown to the algorithm, demonstrating excellent predictive performance. Extremely high TP concentrations were not well captured (Figure 54) and the same was observed in the TP_{efl} model of the study of Xu et al. (2024).

Figure 54 – Comparative plot between the predicted and observed effluent TP concentrations at John Egan WWTP for the testing dataset

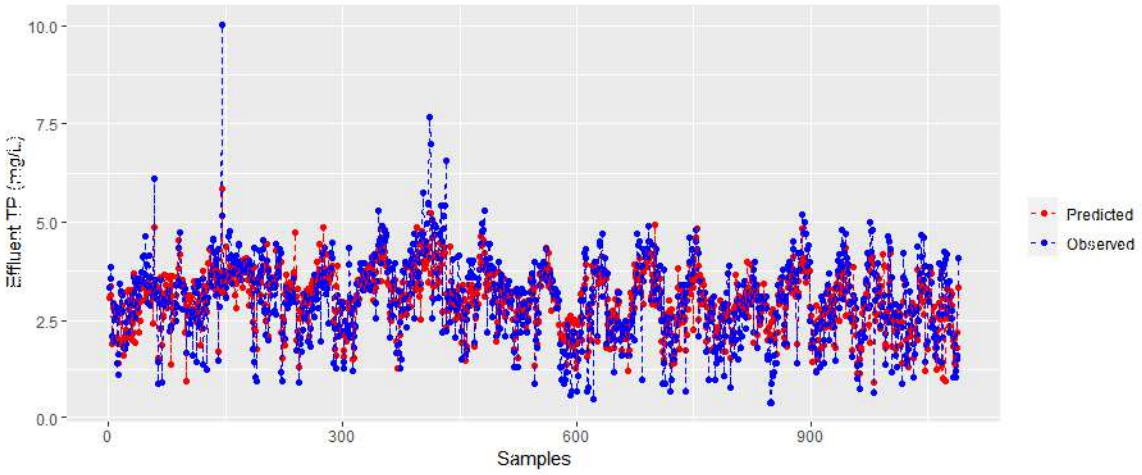
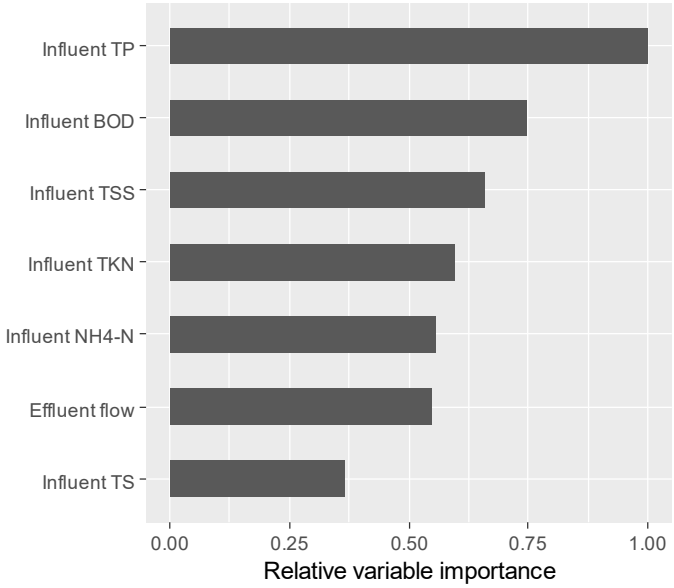


Figure 55 presents the relative variable importance for the model developed to predict effluent concentrations of TP at John E. Egan. The most important variable was the phosphorus sampled at the influent wastewater. The influent quality indicators were mostly more important than the effluent flow.

Figure 55 – Relative variable importance for the TP effluent model at John Egan WWTP



4.4.2.2 Total phosphorus model: number of observations equal to that of Brasília Sul WWTP model

To test the hypothesis that better predictions of the effluent TP made at the John Egan WWTP compared to the Brasília Sul facility were because of the higher data availability, a new model was developed for the Egan facility. The same variables from

the previous model were used to train the TP_{efl} model for the John Egan WWTP, but only the 263 most recent observations (monitored from October 11, 2022, to June 30, 2023) were used. This was the same number of observations used in the Brasília Sul WWTP model (Topic 4.4.1.3). Subsequently, the data were divided into 202 observations for training and 61 for testing the model.

After conducting multiple tests to determine the optimal model architecture, the selected model consisted of a neural network with 500 epochs and two hidden layers, with 9 and 8 neurons, respectively. The model also utilized Z-score normalization as the preprocessing technique. Figure 56 displays the structure of the chosen model.

Figure 56 – Structure of the neural network model for the prediction of effluent TP concentrations at the John Egan WWTP (n = 263)

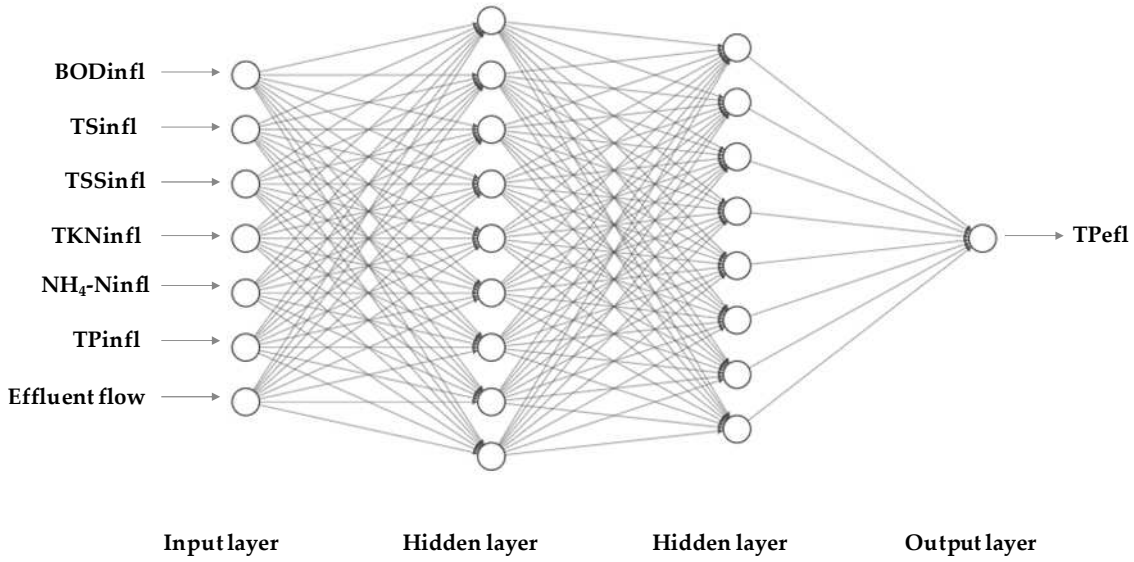
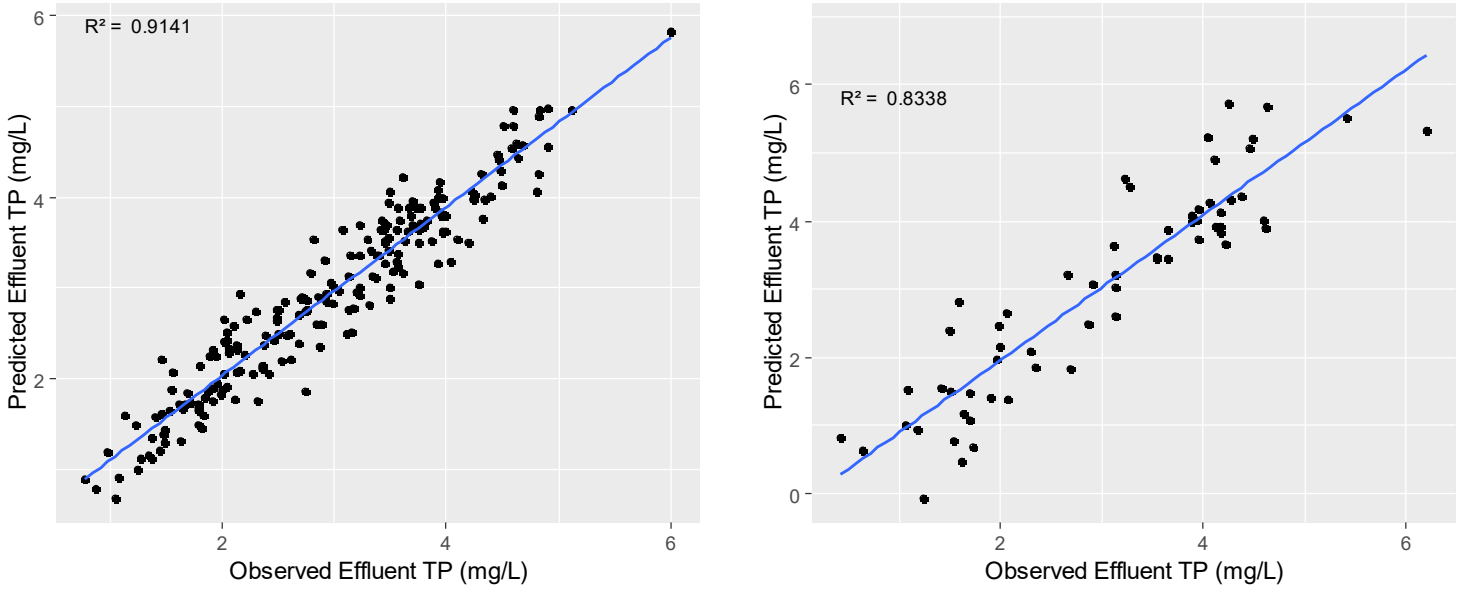


Table 31 displays the results of RMSE and MAE for both training and testing datasets. Figure 57 presents the regression plot between the predicted and observed effluent TP concentrations for both the training and testing datasets. Table C.2 and Figure C.2 show the results of the MLR model for the same data. The ANN performed better during training and the MLR performed slightly better for testing the model. Both the models efficiently predicted TP_{efl} concentrations.

Table 31 – Performance metrics of the model for training and testing datasets for the prediction of effluent TP at John Egan WWTP (n = 263)

	Training (mg/L)	Testing (mg/L)
RMSE	0.308	0.623
MAE	0.242	0.478

Figure 57 – Regression plot between predicted and observed data of effluent TP at John Egan WWTP for (a) training and (b) testing datasets (n = 263)

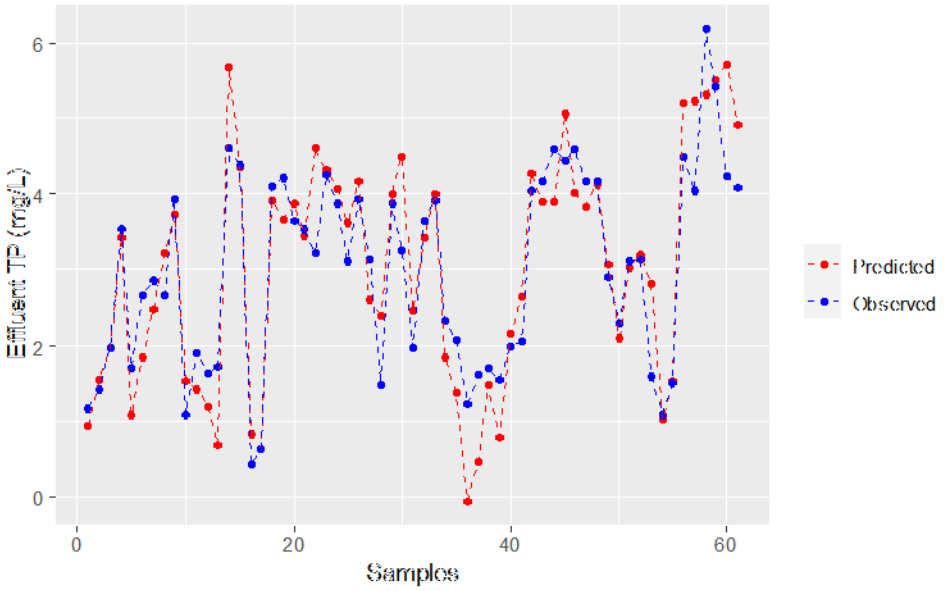


(a) (b)

The performance of the ANN model with fewer observations (n = 263) for both training and testing was similar to that developed for the entire time series (n = 4,181) for all the metrics. The R^2 values were even higher for the model with fewer observations during both training and testing, showing that data quality is more important than sample size for modeling. This could be due to better sensor calibration at the John E. Egan, resulting in more representative data to be trained and tested in the models.

Figure 58 shows a comparison between the predicted and observed effluent TP concentrations in the testing subset of the John E. Egan WWTP model. This result demonstrates that the model was capable of effectively capturing the trends and variability in the data, highlighting the good performance and accuracy of the model and high generalization capability.

Figure 58 – Comparative plot between the predicted and observed effluent TP concentrations at John Egan WWTP for the testing dataset (n = 263)



Due to these findings, it is not feasible to affirm that the best fit for the TP model of John E. Egan shown in Topic 4.4.2.1 compared to the Brasília Sul model (Topic 4.4.1.3) was exclusively owing to the larger amount of data used to train and test the ANN. Even with fewer observations, the model developed for John Egan WWTP was highly accurate and explained the variability in the testing data well.

These results highlight the complexity of the ANN models developed to predict the effluent quality of WWTPs. The performance of the models varies according to the explanatory variables used to train the models, the degree of the relationship between the explanatory and response variables in each facility, the amount of data used to develop the models, the data quality, the monitoring frequency of the data, the variability in the data, and the treatment process employed in each facility and operating conditions. Since the MLR also performed well in the scenario of n = 263 for the John Egan TP concentrations (Table C.2 and Figure C.2), this could indicate that there are fewer complexities and nonlinearities in this data.

As various processes in a WWTP are highly intricate, modeling solutions can serve differently for each problem (GHOLIZADEH et al., 2024). Depending on the complexity of each treatment process and each dataset, as well as the degree of the relationship between the variables involved, in certain cases, it may be necessary to have a larger amount of data to train and test the algorithms and with that, improve the capability of

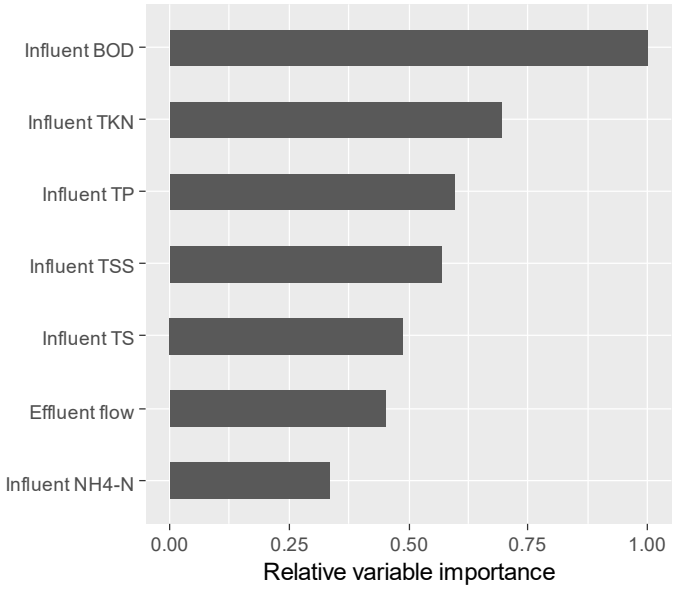
the modeling of the response variable (PHAM et al., 2020; SIMSEK, 2016) and/or it may necessary to add additional explanatory variables. However, collecting additional data on Brasília Sul WWTP to expand the dataset and attempt to improve the model performance was unfeasible.

In the review conducted by Bahramian et al. (2023) on the application of data-driven models in WWTPs, in most of the selected publications, the researchers reported that they were left with no choice but to work with the available data. This can impact the type, structure, efficiency, and accuracy of data-driven models, including ANNs (BAHRAMIAN et al., 2023). These findings highlight the challenges of working with secondary monitoring data from WWTPs.

Roohi, Nazif, and Ramazi (2024) fitted several machine learning algorithms to predict the effluent quality in three WWTPs. The authors tested different numbers of samples, explanatory variables, and algorithms to understand their impact on models' performances. For all the response variables (such as COD_{efl}, TSS_{efl}, NH_4-N_{efl}, TN_{efl}, and TP_{efl}), the authors found that with an increase in the number of observations used to train the models, there was a decrease in the error in the testing data. This happened because a larger training dataset contains more information regarding data trends and variable correlations. According to the authors, there was a sudden drop in the performance of the models when the training data were smaller than 250 samples (ROOHI; NAZIF; RAMAZI, 2024). This highlights that for each WWTP and scenario, the minimum number of samples necessary to train the model could vary, since for the John E. Egan WWTP TP_{efl} model, training the model with 202 observations had a high level of prediction accuracy.

Figure 59 shows the relative importance of the input variables for the TP_{efl} model at John Egan with fewer observations. Although the results were distinct from the model developed with all observations (Figure 55), the four most important variables were the same (influent BOD, TKN, TP, and TSS).

Figure 59 – Relative variable importance for the TP effluent model at John Egan WWTP (n = 263)



4.5 CONCLUSIONS

In this study, neural network models were used to model the effluent quality of two WWTPs, the Brasília Sul facility in Brazil and the John E. Egan facility in the USA. The developed ANN models revealed the relationship between raw wastewater quality, operational variables, and treated effluent quality.

For Brasília Sul WWTP, the models were fitted to predict the effluent concentrations of TN, NH_4-N, TP, COD, and TSS. It was possible at a certain level to represent the complex and nonlinear relationship between the input variables (wastewater influent quality and operational variables) and the treated effluent water quality measurements. The training outcomes of all the five models displayed relatively low errors and high R^2 values. The effective training performance highlights the potential of the ANN technique in modeling WWTPs, mainly because it outperformed the MLR technique for most models. However, a lower accuracy was exhibited during the testing of the models, which indicated overfitting of the ANN. Less complex structures of the neural networks were tested but did not achieve better predictions. This result could be attributed to the high percentage of missing data in the dataset, which reduced the number of samples for model training and testing, limited number of input variables, and irregular monitoring frequency considering all the variables.

The dataset for the John E. Egan WWTP comprised more than 20 years of monitoring, collected mostly on a daily basis. In this dataset, phosphorus was the effluent parameter with the most consistent results, as other variables had a high percentage of censored data or constant values. For this reason, the ANN model was fitted only for the prediction of effluent TP at the John E. Egan facility. The ANN algorithm used to model the effluent TP concentrations performed well, and the model had high accuracy in explaining the variability in the data.

As a much larger dataset (n = 4,181) was used compared to the Brasília Sul WWTP model for the same output variable (n = 263), a second model was fitted to the TP_{efl} of John Egan facility. For this second model, the 263 most recent observations were selected, and the ANN was trained and tested using this dataset. Although it was anticipated that the model would not perform effectively with a smaller sample size, the ANN had a high accuracy on both training and testing datasets. This result emphasizes the complexity of ANN models in predicting the effluent concentrations of full-scale WWTPs. Their performance can vary according to the data size, data quality, the degree of relationship between the input and output variables, the treatment process employed, the stability of the process, and the variability in the data. For this reason, there is no single solution that would fit all scenarios, and new studies can be required for facilities with different influent characteristics and operational schemes and conditions.

CHAPTER 5: FINAL CONSIDERATIONS

5.1 CONCLUSIONS

This work investigates the use of machine learning algorithms, specifically artificial neural networks, in wastewater treatment. Chapter 2 addressed specific objective 1, in which, through a systematic review of the literature, the use of the ANN technique for assessing and predicting the performance of full-scale WWTPs was investigated. In this investigation, it was possible to understand how these models are configured in the field, the main types of ANNs being used, and the performance of the models developed by previous studies. Most studies have used FFNN with a backpropagation training algorithm to predict effluent quality using influent quality indicators as input variables. To train and test the models, most studies have used daily data collected over a period of one to two years, resulting in relatively small datasets (median of 361.5 observations). The most common allocation was 75% of the dataset used for training and 25% for testing the model. Although the use of deep learning was identified, most studies have used one to two hidden layers and less than ten neurons in each hidden layer. The hyperbolic tangent function was the most common function in the hidden layers, and the linear activation function was used in the output layer to train the models. The findings from the systematic literature review confirmed hypothesis (i) and guided the next steps of applications of ANN models to WWTPs under study, enabling a better modeling approach.

In the literature review, no studies were found that had applied the ANN technique to a Brazilian WWTP. Therefore, the Brasília Sul WWTP was selected as the first case study in the country for predicting plant performance using neural networks. Before developing the modeling approach, in Chapter 3, a deep analysis of the monitoring dataset of the facility was conducted to gain a comprehensive understanding of the plant's performance and define key quality parameters. This chapter addressed specific objective 2. The evaluation revealed that, despite the adoption of advanced wastewater treatment technology, the Brasília Sul WWTP still had a considerable violation percentage of nitrogen and phosphorus discharge standards. Since this facility discharges its effluent into a lake, removing nutrients is crucial. In order to investigate a WWTP in a distinct scenario for comparison, the John E. Egan facility, operating in the United States, was included in the study. This facility has a similar size to Brasília Sul but has different treatment process and monitoring data characteristics.

A longer time series of data was made available, and a deep analysis of this dataset was also conducted. A high variability in influent conditions was found, but overall, with stable operation and performance. Removing phosphorus was also highlighted as an important goal since recent discharge standards have started to rule for this facility, and there is no specific step in the treatment process to remove phosphorus.

Using our knowledge of how similar models have been developed in the literature (Chapter 2) and the understanding of the WWTPs datasets and performance (Chapter 3), in Chapter 4, ANN models were developed to predict effluent quality at the Brasília Sul and John E. Egan WWTPs. In Chapter 4, specific objectives 3, 4, and 5 were met. Hypothesis (ii) was partially confirmed since, in Brasília Sul WWTP models, the potential of the ANNs for predicting effluent quality was observed at some level. This was especially evident due to the high accuracy in the predictions during the training process, highlighting the ability of this method to model complex datasets and processes. However, the predictive performance of the models worsened during testing, indicating overfitting. Other structures were tested but did not improve the generalization ability of the algorithms. Hypothesis (iii) was also partially accepted since, for overfitted models, the results of the sensitivity analysis are valid for the training data but not necessarily for the treatment process as a whole.

The model developed to predict effluent TP concentrations at John Egan WWTP showed an adequate performance in both training and testing subsets, highlighting its better generalization ability. Since a large dataset (n = 4,181) was used in this model, a second ANN model for this facility was fitted using the same number of observations as the TP_{efl} model of the Brasília Sul facility (n = 263). The second model also had high accuracy in predicting the effluent TP values and captured the variability in the data well. This result showed that the size of the dataset is not the only responsible for the model performance since it can be impacted by the input variables being used, the monitoring frequency, data quality, treatment process, and variability in the data, among other factors. For this reason, hypothesis (iv) was refuted since the number of observations is not the only reason for a better predictive performance.

5.2 SUGGESTIONS FOR FUTURE WORK

The following studies are suggested for future work:

- Conduct a systematic review on the use of other machine learning algorithms for WWTP performance prediction, such as regression trees, random forest, support vector machine, and hybrid models;
- Develop other machine learning algorithms for WWTP performance prediction and compare the results with the ANN models;
- Design ANN models for WWTP performance prediction with additional predictors, including more operational variables and samples collected at a greater number of points in the system, such as between treatment units (as opposed to only at influent);
- Assess the impact of different monitoring frequencies on ANN model performance;
- Develop ANN models for the performance prediction of other WWTPs in operation in Brazil, in distinct locations and with different treatment technologies and sizes.

5.3 EXPERIENCE ABROAD – VISITING PH.D. RESEARCH AT BAYLOR UNIVERSITY

From August 2023 to May 2024, part of the research was conducted at Baylor University (Waco, Texas, United States) with funding from the Fulbright Program. Due to the complexity of the development of artificial intelligence models, establishing an international collaboration for advanced data analysis was of great importance.

During the visiting Ph.D. research period, it was possible to develop several activities under the supervision of Professor Amanda S. Hering, a professor at the Department of Statistical Science at Baylor University. I sat in on two courses at Baylor University, Methods in Statistics I and Methods in Statistics II, and one course at Colorado School of Mines, namely Water and Wastewater Treatment Units Processes. I also participated in the Paper Clubs led by graduate students of the Department of Statistical Science at Baylor University.

I attended two conferences while in the USA: the WEFTEC (Water Environment Federation's Technical Exhibition and Conference) in Chicago and the Water Quality Technology Conference hosted by the American Water Works Association (AWWA) in Dallas. At these conferences, I learned more about water and wastewater treatment

practices and innovations in the US and other countries and established connections in the field.

I also attended the “Bears Building Bridges” Research Symposium at Baylor University, hosted by the Biology, Environmental Science, and Anthropology departments. At the event, I presented a poster with some of the results from the research project I’ve conducted at Baylor.

Besides the poster presentation, during my exchange program, I gave four oral presentations: the first for the research group I participated in at Baylor University, the second for an undergraduate biology class at Baylor, the third for the Department of Civil and Environmental Engineering at Colorado School of Mines, and the fourth at Boulder Water Resource Recovery Facility. In some of these presentations, I talked about water and wastewater treatment and surface water quality in Brazil and Minas Gerais. I shared information about the panorama of these complex topics that require advancements in Brazil. I also discussed my Ph.D. research in Brazil and the research project I developed in the USA.

I had the opportunity to visit Colorado alongside Professor Amanda Hering. Our visit included a tour of the pilot water and wastewater treatment plants at the Colorado School of Mines coordinated by Professor Tzahi Cath. Additionally, I delivered a presentation at the university and visited a full-scale WWTP, the Boulder Water Resource Recovery Facility.

The main activities of the program were continuing my Ph.D. dissertation, in which I obtained data from the John E. Egan WWTP to be incorporated into the survey, and I learned more about machine learning and neural networks to attempt to improve the models developed to predict the performance of the WWTPs. I also participated in the research project titled “*Assessing the Long-Term Performance of Data-Driven Forecasting Models of Ammonia in a Full-Scale Wastewater Treatment Plant*”. In this project, we analyzed data collected every five minutes for nearly three years, totaling over 200,000 observations. A hybrid model combining a diurnal trend, linear regression, and artificial neural network was developed to forecast ammonia in the aeration basin of the Boulder facility. The manuscript resulting from this project is in development.

From all the experiences I had in the course of the international program, I improved my R and English skills, especially in the academic environment. I met new researchers, and I improved my knowledge of statistical concepts. I also contributed through my experience of statistically analyzing different environmental, social, and sanitary contexts of a developing country. Besides that, after spending some time conducting research in the US, I have come to appreciate the quality of scientific research in Brazil even more.

Conducting part of my dissertation in the US as a visiting Ph.D. student was important for developing my research and advancing my academic career. I expanded my knowledge and vision from an international perspective. Moreover, this period also improved my personal development and cultural exchange.

REFERENCES

ABBA, S. I.; ABDULKADIR, R. A.; GAYA, M. S.; SAMMEN, S. S.; GHALI, U.; NAWAILA, M. B.; OĞUZ, G.; MALIK, A.; AL-ANSARI, N. Effluents quality prediction by using nonlinear dynamic block-oriented models: a system identification approach. **Desalination and Water Treatment**, v. 218, p. 52–62, 2021.

ABBA, S. I.; ELKIRAN, G.; NOURANI, V. Improving novel extreme learning machine using PCA algorithms for multi-parametric modeling of the municipal wastewater treatment plant. **Desalination and Water Treatment**, v. 215, p. 414–426, 2021.

AGÊNCIA NACIONAL DE ÁGUAS – ANA. **Atlas Esgotos: Atualização da base de dados de estações de tratamento de esgotos no Brasil**. Brasília: ANA, 2020, 44p.

AGÊNCIA NACIONAL DE ÁGUAS – ANA. **Atlas Esgotos: Despoluição de Bacias Hidrográficas**. Brasília: ANA, 2017, 88p.

AGÊNCIA REGULADORA DE ÁGUAS, ENERGIA E SANEAMENTO BÁSICO DO DISTRITO FEDERAL – ADASA. **Fiscalização ETE Brasília Sul referente ao aparecimento de algas no Lago Paranoá**. Brasília: ADASA, 2016, 14p.

AGÊNCIA REGULADORA DE ÁGUAS, ENERGIA E SANEAMENTO BÁSICO DO DISTRITO FEDERAL – ADASA. **Manual Técnico e Administrativo de Outorga de Direito de Uso de Recursos Hídricos no Distrito Federal**. Brasília: ADASA, 2021.

AGÊNCIA REGULADORA DE ÁGUAS, ENERGIA E SANEAMENTO BÁSICO DO DISTRITO FEDERAL – ADASA. **Relatório de monitoramento do atendimento ao padrão de lançamento outorgado pelas ETEs no Distrito Federal**. Brasília: ADASA, 2020.

AHMED, A. N.; OTHMAN, F. B.; AFAN, H. A.; IBRAHIM, R. K.; FAI, C. M.; HOSSAIN, M. S.; EHTERAM, M.; ELSHAFIE, A. Machine learning methods for better water quality prediction. **Journal of Hydrology**, v. 578, p. 1–18, 2019.

AL-GHAZAWI, Z.; ALAWNEH, R. Use of artificial neural network for predicting effluent quality parameters and enabling wastewater reuse for climate change resilience – A case from Jordan. **Journal of Water Process Engineering**, v. 44, p. 1–10, 2021.

AL-OBAIDI, B. H. K. Predicting municipal sewage effluent quality index using mathematical models in the Al-Rustamiya sewage treatment plant. **Journal of Engineering Science and Technology**, v. 15, n. 6, p. 3571–3587, 2020.

ALDAGHI, T.; JAVANMARD, S. The evaluation of wastewater treatment plant performance: a data mining approach. **Journal of Engineering, Design and Technology**, 2021.

ALKMIM, A. R.; ALMEIDA, G. M.; CARVALHO, D. M.; AMARAL, M. C. S.; OLIVEIRA, S. M. A. C. Improving Knowledge about Permeability in Membrane Bioreactors through Sensitivity Analysis using Artificial Neural Networks Improving knowledge about permeability in membrane bioreactors through sensitivity analysis using artificial neural networks. **Environmental Technology**, v. 41, n. 19, p. 1–15, 2019.

ALSULAILI, A.; REFAIE, A. Artificial neural network modeling approach for the prediction of five-day biological oxygen demand and wastewater treatment plant performance. **Water Supply**, v. 21, n. 5, p. 1861–1877, 2021.

ANTERO, G. V. C. **Avaliação do desempenho operacional e potencial de recuperação de fósforo da estação de tratamento de esgoto Brasília Sul**. 2020. Monografia de projeto final (Graduação em Engenharia Ambiental) – Departamento de Engenharia Civil e Ambiental – Universidade de Brasília.

ARAUJO, J. E. M.; DE JESUS, M. C. **Estudo do impacto das águas pluviais no sistema de tratamento de esgotos da ETE Sul Brasília**. 2018. Monografia de projeto final (Graduação em Engenharia Civil) – Departamento de Engenharia Civil e Ambiental – Universidade de Brasília.

BAGHERI, M.; MIRBAGHERI, S. A.; EHTESHAMI, M.; BAGHERI, Z.; KAMARKHANI, A. M. Analysis of variables affecting mixed liquor volatile suspended solids and prediction of effluent quality parameters in a real wastewater treatment plant. **Desalination and Water Treatment**, v. 57, n. 45, p. 1–14, 2015.

BAHRAMIAN, M.; DERELI, R. K.; ZHAO, W.; GIBERTI, M.; CASEY, E. Data to intelligence: The role of data-driven models in wastewater treatment. **Expert Systems with Applications**, v. 217, p. 1–20, 2023.

BARROS, I. P. A. F. **Proposta de um sistema de indicadores de desempenho para avaliação de estações de tratamento de esgotos do Distrito Federal**. 2013. Dissertação (Mestrado) – Programa de Pós-Graduação em Saneamento, Meio Ambiente e Recursos Hídricos – Universidade Federal de Minas Gerais.

BEKKARI, N.; ZEDDOURI, A. Using artificial neural network for predicting and controlling the effluent chemical oxygen demand in wastewater treatment plant. **Management of Environmental Quality: An International Journal**, v. 30, n. 3, p. 593–608, 2019.

BERTOLOSSI, V. M.; NEDER, T. F.; BRANDÃO, C. C. S. Avaliação de ultrafiltração como alternativa à flotação por ar dissolvido no pós-tratamento do efluente de lodos ativados – estudo em escala piloto na estação de tratamento de esgoto Brasília Norte. **Engenharia Sanitaria e Ambiental**, v. 26, n. 6, p. 1003–1014, 2021.

CHEN, H.-M.; LO, S. L. Neural network-based multi-back-propagation prediction model of a domestic wastewater treatment plant for an under-construction sewer system. **Journal of the Chinese Institute of Engineers**, v. 35, n. 7, p. 815–826, 2012.

CHEN, Y.; SONG, L.; LIU, Y.; YANG, L.; LI, D. A Review of the Artificial Neural Network Models for Water Quality Prediction. **Applied Sciences**, v. 10, n. 17, p. 1–49, 2020.

CHING, P. M. L.; SO, R. H. Y.; MORCK, T. Advances in soft sensors for wastewater treatment plants: A systematic review. **Journal of Water Process Engineering**, v. 44, p. 1–11, 2021.

CIVELEKOGLU, G.; YIGIT, N. O.; DIAMADOPOULOS, E.; KITIS, M. Modelling of COD

removal in a biological wastewater treatment plant using adaptive neuro-fuzzy inference system and artificial neural network. **Water Science and Technology**, v. 60, n. 6, p. 1475–1488, 2009.

COMBER, S. D. W.; GARDNER, M. J.; ELLOR, B. Seasonal variation of contaminant concentrations in wastewater treatment works effluents and river waters. **Environmental Technology**, v. 41, n. 21, p. 2716–2730, 2019.

COMPANHIA DE SANEAMENTO AMBIENTAL DO DISTRITO FEDERAL – CAESB. **Plano de Exploração – Diagnóstico e Caracterização**. Brasília: CAESB, 2021, 122p.

COROMINAS, L.; GARRIDO-BASERBA, M.; VILLEZ, K.; OLSSON, G.; CORTÉS, U.; POCH, M. Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques. **Environmental Modelling & Software**, v. 106, p. 89–103, 2018.

CRINI, G.; LICHTFOUSE, E. Advantages and disadvantages of techniques used for wastewater treatment. **Environmental Chemistry Letters**, v. 17, n. 1, p. 145–155, 2019.

DANTAS, M. S. **Análise integrada do tratamento de esgotos domésticos e da qualidade das águas superficiais nas bacias hidrográficas do rio das Velhas e dos rios Jequitaí e Pacuí - MG**. 2020. Dissertação (Mestrado) – Programa de Pós-Graduação em Saneamento, Meio Ambiente e Recursos Hídricos – Universidade Federal de Minas Gerais.

DANTAS, M. S.; BARROSO, G. R.; OLIVEIRA, S. C. Performance of sewage treatment plants and impact of effluent discharge on receiving water quality within an urbanized area. **Environmental Monitoring and Assessment**, v. 193, n. 5, p. 1–21, 2021.

DE CANETE, J. F.; SAZ-OROZCO, P.; GÓMEZ-DE-GABRIEL, J.; BARATTI, R.; RUANO, A.; RIVAS-BLANCO, I. Control and soft sensing strategies for a wastewater treatment plant using a neuro-genetic approach. **Computers and Chemical Engineering**, v. 144, p. 1–14, 2021.

DEMING, C.; DEKKATI, S.; DESAMSETTI, H. Exploratory Data Analysis and Visualization for Business Analytics. **Asian Journal of Applied Science and Engineering**, v. 7, n. 1, p. 93–100, 2018.

DRURY, B.; ROSI-MARSHALL, E.; KELLY, J. J. Wastewater treatment effluent reduces the abundance and diversity of benthic bacterial communities in urban and suburban rivers. **Applied and Environmental Microbiology**, v. 79, n. 6, p. 1897–1905, 2013.

DUEÑAS, J. F.; ALONSO, J. R.; REY, À. F.; FERRER, A. S. Characterisation of phosphorous forms in wastewater treatment plants. Journal of Hazardous Materials, v. 97, n. 1–3, p. 193–205, 2003.

EL-RAWY, M.; ABD-ELLAH, M. K.; FATHI, H.; AHMED, A. K. A. Forecasting effluent and performance of wastewater treatment plant using different machine learning techniques. **Journal of Water Process Engineering**, v. 44, p. 1-15, 2021.

ELFANSSI, S.; OUAZZANI, N.; LATRACH, L.; HEJJAJ, A.; MANDI, L. Phytoremediation of domestic wastewater using a hybrid constructed wetland in mountainous rural area. **International Journal of Phytoremediation**, v. 20, n. 1, p. 75–87, 2018.

ELMAADAWY, K.; ELAZIZ, M. A.; ELSHEIKH, A. H.; MOAWAD, A.; LIU, B.; LU, S. Utilization of random vector functional link integrated with manta ray foraging optimization for effluent prediction of wastewater treatment plant. **Journal of Environmental Management**, v. 298, p. 1–9, 2021.

ESPINOSA, M. F.; SANCHO, A. N.; MENDONZA, L. M.; MOTA, C. R.; VERBYLA, M. E. Systematic review and meta-analysis of time-temperature pathogen inactivation. **International Journal of Hygiene and Environmental Health**, v. 230, p. 1–9, 2020.

FAISAL, M.; MUTTAQI, K. M.; SUTANTO, D.; AL-SHETWI, A. Q.; KER, P. J.; HANNAN, M. A. Control technologies of wastewater treatment plants: The state-of-the-art, current challenges, and future directions. **Renewable and Sustainable Energy Reviews**, v. 181, p. 1–28, 2023.

FELLOWS, I. (2018). **wordcloud: Word Clouds. R package version 2.6**. https://CRAN.R-project.org/package=wordcloud.

FENG, J.; LU, S. Performance Analysis of Various Activation Functions in Artificial Neural Networks. **Journal of Physics: Conference Series**, v. 1237, n. 2, p. 1–6, 2019.

GAYA, M. S.; WAHAB, N. A.; SAM, Y. M.; SAMSUDIN, S. I. ANFIS modelling of carbon and nitrogen removal in domestic wastewater treatment plant. **Jurnal Teknologi**, v. 67, n. 5, p. 29–34, 2014.

GE, J.; GUHA, B.; LIPPINCOTT, L.; CACH, S.; WEI, J.; SU, T.; MENG, X. Challenges of arsenic removal from municipal wastewater by coagulation with ferric chloride and alum. **Science of The Total Environment**, v. 725, p. 1–9, 2020.

GEDEON, T. D. Data mining of inputs: analysing magnitude and functional measures. **International Journal of Neural Systems**, v. 8, n. 2, p. 209–218, 1997.

GHOLIZADEH, M.; SAEEDI, R.; BAGHERI, A.; PAEEZI, M. Machine learning-based prediction of effluent total suspended solids in a wastewater treatment plant using different feature selection approaches: A comparative study. **Environmental Research**, v. 246, p. 1–8, 2024.

GOVERNO DO DISTRITO FEDERAL – GDF. **Plano Distrital de Saneamento Básico do Distrito Federal: Produto 7**. Brasília: GDF, 2017a.

GOVERNO DO DISTRITO FEDERAL – GDF. **Plano Distrital de Saneamento Básico do Distrito Federal: Relatório síntese**. Brasília: GDF, 2017b.

GOVERNO DO DISTRITO FEDERAL – GDF. **Plano Distrital de Saneamento Básico do Distrito Federal: Tomo IV - Produto 2**. Brasília: GDF, 2017c.

GUO, H.; JEONG, K.; LIM, J.; JO, J.; KIM, Y. M.; PARK, J.; KIM, J. H.; CHO, K. H. Prediction of effluent concentration in wastewater treatment plant using machine learning models. **Journal of Environmental Sciences**, v. 32, p. 90–101, 2015.

HADJIMICHAEL, A.; COMAS, J.; COROMINAS, L. Do machine learning methods used in data mining enhance the potential of decision support systems? A review for the urban water sector. **AI Communications**, v. 29, n. 6, p. 747–756, 2016.

HAMED, M. M.; KHALAFALLAH, M. G.; HASSANIEN, E. A. Prediction of wastewater treatment plant performance using artificial neural networks. **Environmental Modelling and Software**, v. 19, n. 10, p. 919–928, 2004.

HAMEED, M.; SHARQI, S. S.; YASEEN, Z. M.; AFAN, H. A.; HUSSAIN, A.; ELSHAFIE, A. Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia. **Neural Computing and Applications**, v. 28, n. 1, p. 893–905, 2016.

HAN, H.-G.; WANG, L.-D.; QIAO, J.-F. Hierarchical extreme learning machine for feedforward neural network. **Neurocomputing**, v. 128, p. 128–135, 2014.

HAN, H.; ZHU, S.; QIAO, J.; GUO, M. Data-driven intelligent monitoring system for key variables in wastewater treatment process. **Chinese Journal of Chemical Engineering**, v. 26, n. 10, p. 2093–2101, 2018.

HAN, H. G.; ZHANG, J.; DU, S.; SUN, H.; QIAO, J. Robust optimal control for anaerobic-anoxic-oxic reactors. **Science China Technological Sciences**, v. 64, n. 7, p. 1485–1499, 2021.

HANSEN, L. D.; STOKHOLM-BJERREGAARD, M.; DURDEVIC, P. Modeling phosphorous dynamics in a wastewater treatment process using Bayesian optimized LSTM. **Computers and Chemical Engineering**, v. 160, p. 1–13, 2022.

HAZALI, N.; WAHAB, N. A.; IBRAHIM, S. Modelling and evaluation of sequential batch reactor using artificial neural network. **International Journal of Electrical and Computer Engineering**, v. 7, n. 3, p. 1620–1627, 2017.

HEDDAM, S.; LAMDA, H.; FILALI, S. Predicting Effluent Biochemical Oxygen Demand in a Wastewater Treatment Plant Using Generalized Regression Neural Network Based Approach: A Comparative Study. **Environmental Processes**, v. 3, n. 1, p. 153–165, 2016.

HEJABI, N.; SAGHEBIAN, S. M.; AALAMI, M. T.; NOURANI, V. Evaluation of the effluent quality parameters of wastewater treatment plant based on uncertainty analysis and post-processing approaches (case study). **Water Science and Technology**, v. 83, n. 7, p. 1633–1648, 2021.

ILLINOIS STATE CLIMATOLOGIST. **Climate of Chicago**. 2024. Available at: <https://stateclimatologist.web.illinois.edu/climate-of-illinois/climate-of-chicago/>. Accessed 14 March 2024.

JAMI, M. S.; HUSAIN, I. A. F.; KABASHI, A. N.; ABDULLAH, N. Multiple inputs artificial neural network model for the prediction of wastewater treatment plant performance. **Australian Journal of Basic and Applied Sciences**, v. 6, n. 1, p. 62–69, 2012.

JAMI, M. S.; MUJELI, M.; KABBASHI, N. A. Simulation of ammoniacal nitrogen effluent using feedforward multilayer neural networks. **African Journal of Biotechnology**, v. 10, n. 81, p. 18755–18762, 2011.

KHALID, S.; SHAHID, M.; NATASHA; BIBI, I.; SARWAR, T.; SHAH, A. H.; NIAZI, N. K. A review of environmental contamination and health risk assessment of wastewater use for crop irrigation with a focus on low and high-income countries. In International **Journal of Environmental Research and Public Health**, v. 15, n. 5, p. 1–36, 2018.

KHAN, K. S.; KUNZ, R.; KLEIJNEN, J.; ANTES, G. Five steps to conducting a systematic review. **Journal of the Royal Society of Medicine**, v. 96, n. 3, p. 118–121, 2003.

KHATRI, N.; KHATRI, K. K.; SHARMA, A. Artificial neural network modelling of faecal coliform removal in an intermittent cycle extended aeration system-sequential batch reactor based wastewater treatment plant. **Journal of Water Process Engineering**, v. 37, p. 1–8, 2020.

KHATRI, N.; KHATRI, K. K.; SHARMA, A. Prediction of effluent quality in ICEAS-sequential batch reactor using feedforward artificial neural network. **Water Science and Technology**, v. 80, n. 2, p. 213–222, 2019.

KUSIAK, A.; WEI, X. Optimization of the activated sludge process. **Journal of Energy Engineering**, v. 139, n. 1, p. 12–17, 2013.

KWON, S. B.; KIM, D. I.; GUAN, Y. T.; DOCKKO, S. Reduction of phosphorous at WWTP combined with DAF and A2/O. **Desalination and Water Treatment**, v. 54, p. 1090–1097, 2015.

LANTZ, B. **Machine Learning with R**. Birmingham: Packt Publishing, 2013. 375 p.

LEDELL, E.; GILL, N.; AIELLO, S.; FU, A.; CANDEL, A.; CLICK, C.; KRALJEVIC, T.; NYKODYM, T.; ABOYOUN, P.; KURKA M.; MALOHLAVA M. (2022). **h2o: R Interface for the 'H2O' Scalable Machine Learning Platform. R package version 3.38.0.1**. https://CRAN.R-project.org/package=h2o.

LEE, J.-W.; SUH, C.; HONG, Y.-S. T.; SHIN, H.-S. Sequential modelling of a full-scale wastewater treatment plant using an artificial neural network. **Bioprocess and Biosystems Engineering**, v. 34, n. 8, p. 963–973, 2011.

LEE, J.; BECK, K.; BÜRGMANN, H. Wastewater bypass is a major temporary point-source of antibiotic resistance genes and multi-resistance risk factors in a Swiss river.

Water Research, v. 208, p. 1–12, 2022.

LENAIL, A. NN-SVG: Publication-Ready Neural Network Architecture Schematics. **Journal of Open Source Software**, v. 4, n. 33, p. 747, 2019.

LI, X.; SU, J.; WANG, H.; BOCZKAJ, G.; MAHLKNECHT, J.; SINGH, S. V.; WANG, C. Bibliometric analysis of artificial intelligence in wastewater treatment: Current status, research progress, and future prospects. **Journal of Environmental Chemical Engineering**, v. 12, n. 4, p. 1–13, 2024.

LIU, H.; HUANG, M.; YOO, C. A fuzzy neural network-based soft sensor for modeling nutrient removal mechanism in a full-scale wastewater treatment system. **Desalination and Water Treatment**, v. 51, p. 6184–6193, 2013.

LIU, W.; LIU, T.; LIU, Z.; LUO, H.; PEI, H. A novel deep learning ensemble model based on two-stage feature selection and intelligent optimization for water quality prediction. **Environmental Research**, v. 224, p. 1–13, 2023.

LIU, X.; SHI, Q.; LIU, Z.; YUAN, J. Using LSTM Neural Network Based on Improved PSO and Attention Mechanism for Predicting the Effluent COD in a Wastewater Treatment Plant. **IEEE Access**, v. 9, p. 146082–146096, 2021.

MACAITIS, W. Chicago's Water Problems and Solutions. **GeoJournal**, v. 11, n. 3, p. 229–237, 1985.

MADIĆ, M. J.; RADOVANOVIĆ, M. R. Optimal Selection of ANN Training and Architectural Parameters Using Taguchi Method: A Case Study. **FME Transactions**, v. 39, n. 2, p. 79–86, 2011.

MALVIYA, A.; JASPAL, D. Artificial intelligence as an upcoming technology in wastewater treatment: a comprehensive review. **Environmental Technology Reviews**, v. 10, n. 1, p. 177–187, 2021.

MATHUR, R.; SHARMA, M. K.; LOGANATHAN, K.; ABBAS, M.; HUSSAIN, S.; KATARIA, G.; ALQAHTANI, M. S.; RAO, K. S. Modeling of two-stage anaerobic onsite wastewater sanitation system to predict effluent soluble chemical oxygen demand through machine learning. **Scientific Reports**, v. 14, p. 1–14, 2024.

METCALF & EDDY. **Wastewater engineering: treatment and resource recovery**. 5. ed. New York: Metcalf & Eddy, Inc., 2014, 2018p.

METROPOLITAN WATER RECLAMATION DISTRICT OF GREATER CHICAGO – MWRD. **A History of Protecting Our Water Environment**. 2024a. Available at: <https://mwrd.org/what-we-do/history-protecting-our-water-environment>. Accessed 13 March 2024.

METROPOLITAN WATER RECLAMATION DISTRICT OF GREATER CHICAGO – MWRD. **A Utility of the Future**. MWRD, 2024b, 11p.

METROPOLITAN WATER RECLAMATION DISTRICT OF GREATER CHICAGO –

MWRD. **Phosphorus removal feasibility study for the Kirie Water Reclamation Plant**. MWRD, 2022, 178p.

METROPOLITAN WATER RECLAMATION DISTRICT OF GREATER CHICAGO – MWRD. **Service area**. 2024c. Available at: <https://mwrd.org/what-we-do/service-area>. Accessed 14 March 2024.

METROPOLITAN WATER RECLAMATION DISTRICT OF GREATER CHICAGO – MWRD. **John E. Egan Water Reclamation Plant – Fact Sheet**. MWRD, 2019, 2p.

METROPOLITAN WATER RECLAMATION DISTRICT OF GREATER CHICAGO – MWRD. **Press Release: MWRD's Egan Water Reclamation Plant celebrates 40 years of service and innovation in enhancing water quality and pioneering technology**. MWRD, 2015, 3p.

MJALLI, F. S.; AL-ASHEH, S.; ALFADALA, H. E. Use of artificial neural network black-box modeling for the prediction of wastewater treatment plants performance. **Journal of Environmental Management**, v. 83, n. 3, p. 329–338, 2007.

NASR, M. S.; MOUSTAFA, M. A. E.; SEIF, H. A. E.; KOBROSY, G. E. Application of Artificial Neural Network (ANN) for the prediction of EL-AGAMY wastewater treatment plant performance-Egypt. **Alexandria Engineering Journal**, v. 51, n. 1, p. 37–43, 2012.

NEWHART, K. B.; HERING, A. S.; CATH, T. Y. Data science tools to enable decarbonized water and wastewater treatment systems. In: **Pathways to Water Sector Decarbonization, Carbon Capture and Utilization**. IWA Publishing, 2022. p. 275–301.

NEWHART, K. B.; HOLLOWAY, R. W.; HERING, A. S.; CATH, T. Y. Data-driven performance analyses of wastewater treatment plants: A review. **Water Research**, v. 157, p. 498–513, 2019.

NEZHAD, M. F.; MEHRDADI, N.; TORABIAN, A.; BEHBOUDIAN, S. Artificial neural network modeling of the effluent quality index for municipal wastewater treatment plants using quality variables: south of Tehran wastewater treatment plant. **Journal of Water Supply: Research and Technology**, v. 65, n. 1, p. 18–27, 2016.

NOURANI, V.; ASGHARI, P.; SHARGHI, E. Artificial intelligence based ensemble modeling of wastewater treatment plant using jittered data. **Journal of Cleaner Production**, v. 291, p. 1–15, 2021.

NOURANI, V.; ELKIRAN, G.; ABBA, S. I. Wastewater treatment plant performance analysis using artificial intelligence – an ensemble approach. **Water Science and Technology**, v. 78, n. 10, p. 2064–2076, 2018.

OLIVEIRA, D. B. C. DE; SOARES, W. DE A.; HOLANDA, M. A. C. R. DE. Effects of rainwater intrusion on an activated sludge sewer treatment system. **Revista Ambiente e Água**, v. 15, n. 3, p. 1–12, 2020.

OLIVEIRA, S. C.; VON SPERLING, M. Reliability analysis of wastewater treatment plants. **Water Research**, v. 42, n. 4-5, p. 1182–1194, 2008.

OSMAN, Y. B. M.; LI, W. Soft Sensor Modeling of Key Effluent Parameters in Wastewater Treatment Process Based on SAE-NN. **Journal of Control Science and Engineering**, v. 2020, p. 1–9, 2020.

PALANI, S.; LIONG, S.-Y.; TKALICH, P. An ANN application for water quality forecasting. **Marine Pollution Bulletin**, v. 56, n. 9, p. 1586–1597, 2008.

PAPADOKONSTANTAKIS, S.; LYGEROS, A.; JACOBSSON, S. P. Comparison of recent methods for inference of variable influence in neural networks. **Neural Networks**, v. 19, n. 4, p. 500–513, 2006.

PENETRA, R. G. **Flotação Aplicada Ao Pós-Tratamento Do Efluente De Reator Anaeróbio De Leito Expandido Tratando Esgoto Sanitário**. 2003. Tese (Doutorado) – Escola de Engenharia de São Carlos – Universidade de São Paulo.

PHAM, Q. B.; GAYA, M. S.; ABBA, S. I.; ABDULKADIR, R. A.; ESMAILI, P.; LINH, N. T. T.; SHARMA, C.; MALIK, A.; KHOI, D. N.; DUNG, T. D.; LINH, D. Q. Modeling of Bunus regional sewage treatment plant using machine learning approaches. **Desalination and Water Treatment**, v. 203, p. 80–90, 2020.

PINHO, A. L. V. DE; SANTOS, D. S. **Avaliação qualitativa da pluma de contaminação do Braço do Riacho Fundo, Lago Paranoá - DF**. 2016. Monografia de projeto final (Graduação em Engenharia Ambiental) – Departamento de Engenharia Civil e Ambiental – Universidade de Brasília.

PLUTH, T. B.; BROSE, D. A.; GALLAGHER, D. W.; WASIK, J. Long-Term Trends Show Improvements in Water Quality in the Chicago Metropolitan Region With Investment in Wastewater Infrastructure, Deep Tunnels, and Reservoirs. **Water Resources Research**, v. 57, n. 6, p. 1–20, 2021.

PORTELA, M. F. P. **Avaliação das eficiências de remoção de poluentes na estação de tratamento de esgotos Brasília Sul com enfoque no fósforo**. 2018. Monografia de projeto final (Graduação em Engenharia Ambiental) – Departamento de Engenharia Civil e Ambiental – Universidade de Brasília.

PULLIN, A. S.; STEWART, G. B. Guidelines for Systematic Review in Conservation and Environmental Management. **Conservation Biology**, v. 20, n. 6, p. 1647–1656, 2006.

QAZI, A.; FAYAZ, H.; WADI, A.; RAJ, R. G.; RAHIM, N. A.; KHAN, W. A. The artificial neural network for solar radiation prediction and designing solar systems: a systematic literature review. **Journal of Cleaner Production**, v. 104, p. 1–12, 2015.

QIAO, J.; HU, Z.; LI, W. Soft measurement modeling based on chaos theory for biochemical oxygen demand (BOD). **Water**, v. 8, n. 12, p. 1–21, 2016.

QIAO, J.; YANG, W.; YUAN, M. Recurrent high order neural network modeling for

wastewater treatment process. **Journal of Computers**, v. 6, n. 8, p. 1570–1577, 2011.

R CORE TEAM (2020). **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

RAHMATI, M. G.; TISHEHZAN, P.; MOAZED, H. Determining the best and simple intelligent models for evaluating BOD5 of Ahvaz wastewater treatment plant. **Desalination and Water Treatment**, v. 209, p. 242–253, 2021.

RAY, S. S.; VERMA, R. K.; SINGH, A.; GANESAPILLAI, M.; KWON, Y.-N. A holistic review on how artificial intelligence has redefined water treatment and seawater desalination processes. **Desalination**, v. 546, p. 1–14, 2023.

REBELO, L. S. **Estudo sobre a necessidade de implantação do sistema de equalização na estação de tratamento de esgotos Brasília Sul**. 2019. Monografia de projeto final (Graduação em Engenharia Civil) – Departamento de Engenharia Civil e Ambiental – Universidade de Brasília.

ROOHI, A. M.; NAZIF, S.; RAMAZI, P. Tackling data challenges in forecasting effluent characteristics of wastewater treatment plants. **Journal of Environmental Management**, v. 354, p. 1–14, 2024.

ROUT, P. R.; SHAHID, M. K.; DASH, R. R.; BHUNIA, P.; LIU, D.; VARJANI, S.; ZHANG, T. C.; SURAMPALLI, R. Y. Nutrient removal from domestic wastewater: A comprehensive review on conventional and advanced technologies. **Journal of Environmental Management**, v. 296, p. 2–16, 2021.

SAFEER, S.; PANDEY, R. P.; REHMAN, B.; SAFDAR, T.; AHMAD, I.; HASAN, S. W.; ULLAH, A. A review of artificial intelligence in water purification and wastewater treatment: Recent advancements. **Journal of Water Process Engineering**, v. 49, p. 1–18, 2022.

SALA-GARRIDO, R.; MOLINOS-SENANTE, M.; HERNÁNDEZ-SANCHO, F. How does seasonality affect water reuse possibilities? An efficiency and cost analysis. Resources, **Conservation and Recycling**, v. 58, p. 125–131, 2012.

SALEH, H. A. AL. Wastewater Pollutants Modeling Using Artificial Neural Networks. **Journal of Ecological Engineering**, v. 22, n. 7, p. 35–45, 2021.

SHARGHI, E.; NOURANI, V.; ASHRAFI, A. A.; GÖKÇEKUŞ, H. Monitoring effluent quality of wastewater treatment plant by clustering based artificial neural network method. **Desalination and Water Treatment**, v. 164, p. 86–97, 2019.

SHARMA, S.; SHARMA, S.; ANIDHYA, A. Activation Functions in Neural Networks. **International Journal of Engineering Applied Sciences and Technology**, v. 4, n. 12, p. 310–316, 2020.

SILVA, L. F. M. **Desempenho de estações de tratamento de esgoto e impactos de seus efluentes em corpos de água receptores em Minas Gerais**. 2020.

Dissertação (Mestrado) – Programa de Pós-Graduação em Saneamento, Meio Ambiente e Recursos Hídricos – Universidade Federal de Minas Gerais.

SIMSEK, H. Mathematical modeling of wastewater-derived biodegradable dissolved organic nitrogen. **Environmental Technology**, v. 37, n. 22, p. 2879–2889, 2016.

SIN, G.; AL, R. Activated sludge models at the crossroad of artificial intelligence – A perspective on advancing process modeling. **Clean Water**, v. 4, n. 16, p. 1–7, 2021.

SINGH, K. P.; BASANT, N.; MALIK, A.; JAIN, G. Modeling the performance of "up-flow anaerobic sludge blanket" reactor based wastewater treatment plant using linear and nonlinear approaches – A case study. **Analytica Chimica Acta**, v. 658, n. 1, p. 1–11, 2010.

SISTEMA NACIONAL DE INFORMAÇÕES SOBRE SANEAMENTO – SNIS. **Diagnóstico Temático: Serviços de Água e Esgoto - Visão Geral**. Brasília: SNIS, 2021.

SISTEMA NACIONAL DE INFORMAÇÕES SOBRE SANEAMENTO – SNIS. **Diagnóstico Temático: Serviços de Água e Esgoto - Visão Geral: Ano de referência 2022**. Brasília: SNIS, 2023.

UNITED STATES CENSUS BUREAU – USCB. **American Housing Survey (AHS)**. 2021. Available at: <https://www.census.gov/programs-surveys/ahs/data/interactive/ahstablecreator.html?s_areas=00000&s_year=2021&s_tablename=TABLE4&s_bygroup1=1&s_bygroup2=1&s_filtergroup1=1&s_filtergroup2=1>. Accessed 8 March 2024.

UNITED STATES CENSUS BUREAU – USCB. **Metropolitan and Micropolitan Statistical Areas Population Totals: 2020-2022**. 2023b. Available at: < https://www.census.gov/data/tables/time-series/demo/popest/2020s-total-metro-and-micro-statistical-areas.html>. Accessed 13 March 2024.

UNITED STATES CENSUS BUREAU – USCB. **QuickFacts: Cook County, Illinois; Chicago city, Illinois; Illinois**. 2023a. Available at: < https://www.census.gov/quickfacts/fact/table/cookcountyillinois,chicagocityillinois,IL/PST045222>. Accessed 12 March 2024.

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY – USEPA. **Clean Watersheds Needs Survey 2022: Report to Congress**. USEPA, 2024.

VERMA, A.; WEI, X.; KUSIAK, A. Predicting the total suspended solids in wastewater: A data-mining approach. **Engineering Applications of Artificial Intelligence**, v. 26, n. 4, p. 1366–1372, 2013.

VON SPERLING, M. **Activated sludge and aerobic biofilm reactors**. London: IWA Publishing, 2007.

VON SPERLING, M. **Introdução à qualidade das águas e ao tratamento de esgotos**. 1. ed. Belo Horizonte: Editora UFMG, 2014.

VON SPERLING, M.; VERBYLA, M. E.; OLIVEIRA, S. M. A. C. **Assessment of Treatment Plant Performance and Water Quality Data**. London: IWA Publishing, 2020.

WANG, D.; THUNÉLL, S.; LINDBERG, U.; JIANG, L.; TRYGG, J.; TYSKLIND, M.; SOUIHI, N. A machine learning framework to improve effluent quality control in wastewater treatment plants. **Science of The Total Environment**, v. 784, p.1–11, 2021.

WANG, J.; GUO, Y.; PENG, S.; WANG, Y.; ZHANG, W.; ZHOU, X.; JIANG, L.; LAI, B. Prediction of effluent ammonia nitrogen in wastewater treatment plant based on self-organizing hybrid neural network. **Journal of Water Process Engineering**, V. 59, p. 1–10, 2024a.

WANG, Z.; DAI, H.; CHEN, B.; CHENG, S.; SUN, Y.; ZHAO, J.; GUO, Z.; CAI, X.; WANG, X.; LI, B.; GENG, H. Effluent quality prediction of the sewage treatment based on a hybrid neural network model: Comparison and application. **Journal of Environmental Management**, v. 351, p. 1–11, 2024b.

WANG, X.; KVAAL, K.; RATNAWEERA, H. Characterization of influent wastewater with periodic variation and snow melting effect in cold climate area. **Computers and Chemical Engineering**, v. 106, p. 202–211, 2017.

WORLD HEALTH ORGANIZATION – WHO; United Nations Children's Fund – UNICEF. 2022a. **Joint Monitoring Programme for Water Supply, Sanitation and Hygiene – Data**. Available at: <https://washdata.org/data/household#!/>. Accessed 11 March 2024.

WORLD HEALTH ORGANIZATION – WHO; United Nations Children's Fund – UNICEF. 2022a. **Joint Monitoring Programme for Water Supply, Sanitation and Hygiene – Guidance note to facilitate country consultation on JMP estimates**. WHO; UNICEF, 2022b, 6p.

XU, B.; POOI, C. K.; TAN, K. M.; HUANG, S.; SHI, X.; NG, H. Y. A novel long short-term memory artificial neural network (LSTM)-based soft-sensor to monitor and forecast wastewater treatment performance. **Journal of Water Process Engineering**, v. 54, p. 1–9, 2023.

XU, Y.; WANG, Z.; NAIRAT, S.; ZHOU, J.; HE, Z. Artificial Intelligence-Assisted Prediction of Effluent Phosphorus in a Full-Scale Wastewater Treatment Plant with Missing Phosphorus Input and Removal Data. **ACS ES&T Water**, v. 4, n. 3, p. 880–889, 2024.

YAQUB, M.; ASIF, H.; KIM, S.; LEE, W. Modeling of a full-scale sewage treatment plant to predict the nutrient removal efficiency using a long short-term memory (LSTM) neural network. **Journal of Water Process Engineering**, v. 37, p. 1–11, 2020.

YASMIN, N. S. A.; GAYA, M. S.; WAHAB, N. A.; SAM, Y. M. Estimation of pH and MLSS using neural network. **Telkomnika (Telecommunication Computing**

Electronics and Control), v. 15, n. 2, p. 912–918, 2017.

YE, Z.; YANG, J.; ZHONG, N.; TU, X.; JIA, J.; WANG, J. Tackling environmental challenges in pollution controls using artificial intelligence: A review. **Science of the Total Environment**, v. 699, p. 1–28, 2020.

YU, Y.; CHEN, Y.; HUANG, S.; WANG, R.; WU, Y.; ZHOU, H.; LI, X.; TAN, Z. Enhancing the effluent prediction accuracy with insufficient data based on transfer learning and LSTM algorithm in WWTPs. **Journal of Water Process Engineering**, v. 62, p. 1–11, 2024.

ZAGHLOUL, M. S.; ACHARI, G. Application of machine learning techniques to model a full-scale wastewater treatment plant with biological nutrient removal. **Journal of Environmental Chemical Engineering**, v. 10, n. 3, p. 1–18, 2022.

ZHANG, H.; JAIN, J. S.; BRAND, M.; PERKOVICH, B.; LAI, K.; CARMODY, S.; URGUN-DEMIRTAS, M.; PAGILLA, K. Full Scale Test on Chemical P Removal during a Step Feed BNR Study at John E. Egan Water Reclamation Plant. *In*: **Proceedings of the Water Environment Federation**, p. 5176–5184, 2006.

ZHANG, H.; JAIN, J. S.; O'CONNOR, C.; GRANATO, T.; BRAND, M.; LAI, K.; FORD, J.; CARMODY, S. Simple Retrofitting for Phosphorus Removal and Its Impact on Plant Performance at the John E. Egan Water Reclamation Plant. *In*: **Proceedings of the Water Environment Federation**, p. 7401–7410, 2008.

ZHANG, R.; HU, X. Effluent quality prediction of wastewater treatment system based on small-world ANN. **Journal of Computers**, v. 7, n. 9, p. 2136–2143, 2012.

ZHANG, S.; JIN, Y.; CHEN, W.; WANG, J.; WANG, Y.; REN, H. Artificial intelligence in wastewater treatment: A data-driven analysis of status and trends. **Chemosphere**, v. 336, p. 1–8, 2023.

ZHANG, Z.; LI, H.; ZHU, J.; WEIPING, L.; XIN, X. Improvement strategy on enhanced biological phosphorus removal for municipal wastewater treatment plants: Full-scale operating parameters, sludge activities, and microbial features. **Bioresource Technology**, v. 102, n. 7, p. 4646–4653, 2011.

ZHAO, L.-J.; CHAI, T.-Y.; YUAN, D.-C. Selective ensemble extreme learning machine modeling of effluent quality in wastewater treatment plants. **International Journal of Automation and Computing**, v. 9, n. 6, p. 627–633, 2012.

ZHAO, L.; DAI, T.; QIAO, Z.; SUN, P.; HAO, J.; YANG, Y. Application of artificial intelligence to wastewater treatment: A bibliometric analysis and systematic review of technology, economy, management, and wastewater reuse. **Process Safety and Environmental Protection**, v. 133, n. 92, p. 169–182, 2020.

ZHAO, Y.; GUO, L.; LIANG, J.; ZHANG. M. Seasonal artificial neural network model for water quality prediction via a clustering analysis method in a wastewater treatment plant of China. **Desalination and Water Treatment**, v. 57, n. 8, p. 3452–3465, 2016.

APPENDIX A – Paper published in Water Science & Technology

Check for updates

Water Science & Technology

IWA PUBLISHING

Water Science & Technology Vol 88 No 6, 1447 doi: 10.2166/wst.2023.276

Artificial neural networks for performance prediction of full-scale wastewater treatment plants: a systematic review

Marina Salim Dantas [a,*], Cristiano Christofaro [b] and Sílvia Corrêa Oliveira [a]

[a] Department of Sanitary and Environmental Engineering, Federal University of Minas Gerais, Av. Presidente Antônio Carlos, 6627, Belo Horizonte, MG CEP 31270-901, Brazil
[b] Department of Forestry Engineering, Federal University of Jequitinhonha and Mucuri Valleys, Road MG 367, 5000, Diamantina, MG CEP 39100-000, Brazil
*Corresponding author. E-mail: marina-dantas@hotmail.com

MSD, 0000-0002-7084-7798; CC, 0000-0002-9957-202X; SCO, 0000-0003-4286-3667

ABSTRACT

Wastewater treatment plants (WWTPs) are complex systems that must maintain high levels of performance to achieve adequate effluent quality to protect the environment and public health. Artificial intelligence and machine learning methods have gained attention in recent years for modeling complex problems, such as wastewater treatment. Although artificial neural networks (ANNs) have been identified as the most common of these methods, no study has investigated the development and configuration of these models. We conducted a systematic literature review on the use of ANNs to predict the effluent quality and removal efficiencies of full-scale WWTPs. Three databases were searched, and 44 records of the 667 identified were selected based on the eligibility criteria. The data extracted from the papers showed that the majority of studies used the feedforward neural network model with a backpropagation training algorithm to predict the effluent quality of plants, particularly in terms of organic matter indicators. The findings of this research may help in the search for an optimum design modeling process for future studies of similar prediction problems.

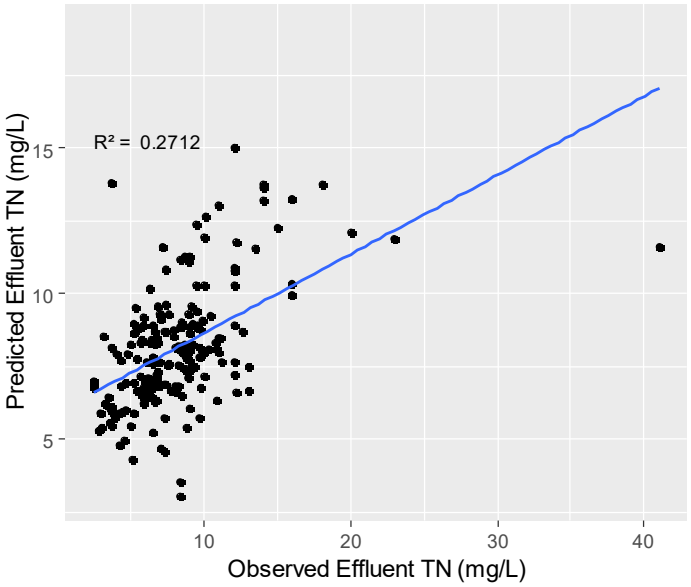
Key words: ANN, artificial intelligence, data science, literature review, machine learning, WWTP

APPENDIX B – Multiple linear regression models results of the Brasília Sul WWTP

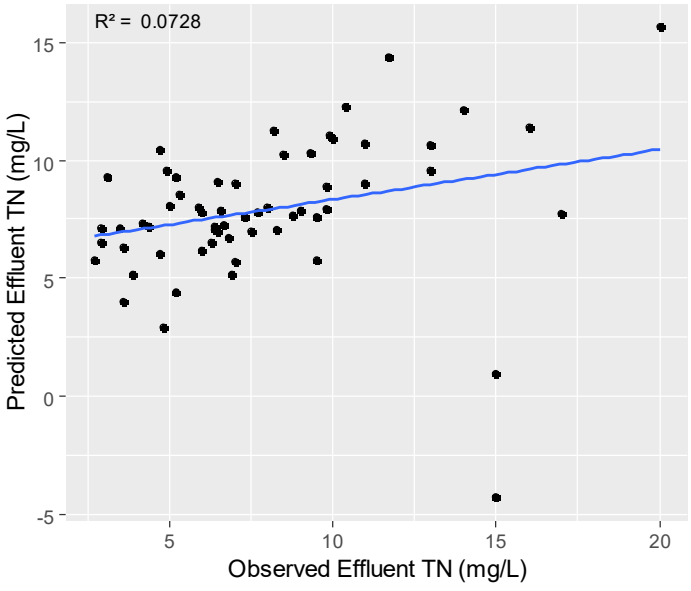
Table B.1 – Performance metrics of the MLR model for training and testing datasets for the prediction of effluent TN

	Training (mg/L)	Testing (mg/L)
RMSE	3.414	4.112
MAE	2.185	2.638

Figure B.1 – Regression plot of MLR between predicted and observed data of effluent TN for (a) training and (b) testing datasets



(a)

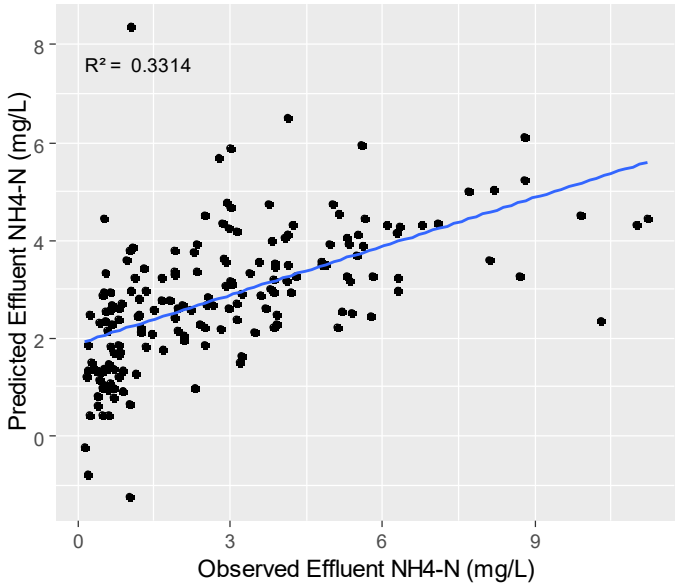


(b)

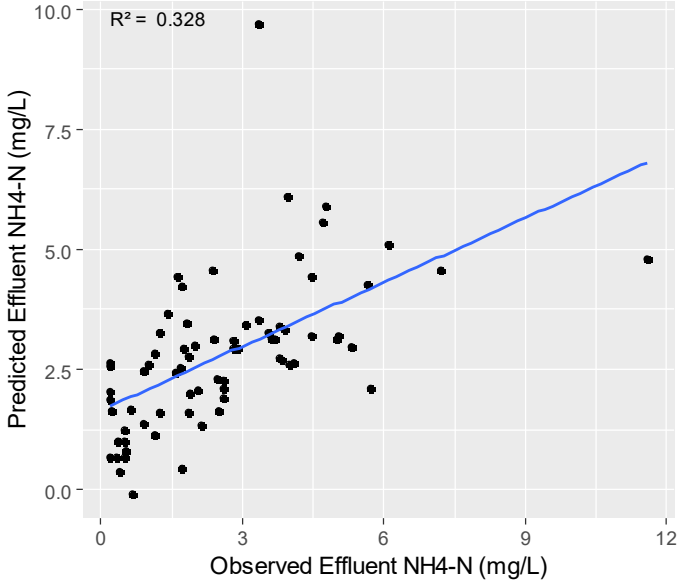
Table B.2 – Performance metrics of the MLR model for training and testing datasets for the prediction of effluent NH_4-N

	Training (mg/L)	Testing (mg/L)
RMSE	1.956	1.706
MAE	1.429	1.204

Figure B.2 – Regression plot of MLR between predicted and observed data of effluent NH_4-N for (a) training and (b) testing datasets



(a)

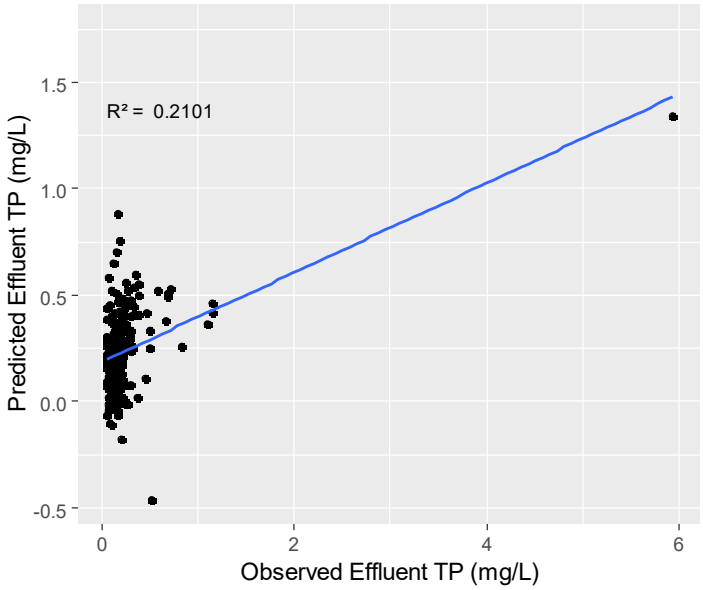


(b)

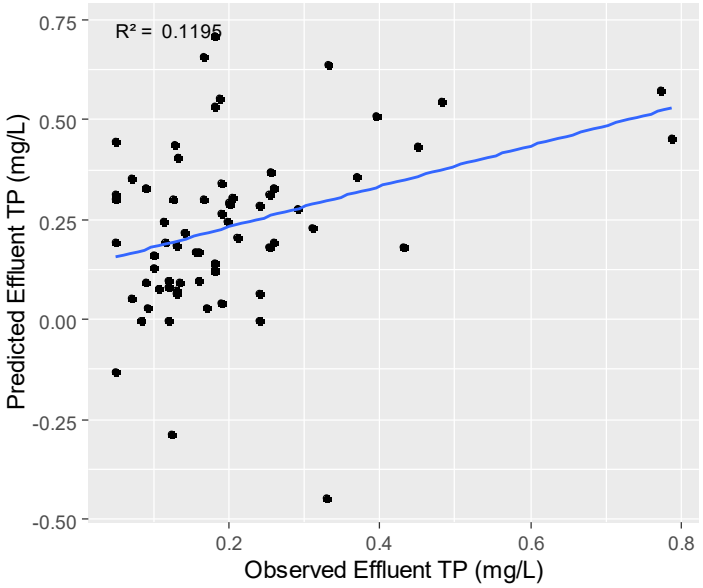
Table B.3 – Performance metrics of the MLR model for training and testing datasets for the prediction of effluent TP

	Training (mg/L)	Testing (mg/L)
RMSE	0.393	0.208
MAE	0.180	0.147

Figure B.3 – Regression plot of MLR between predicted and observed data of effluent TP for (a) training and (b) testing datasets



(a)

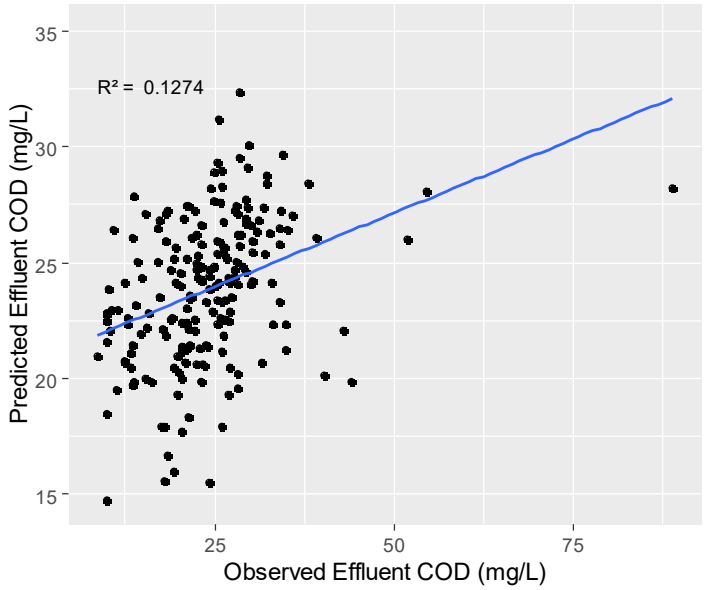


(b)

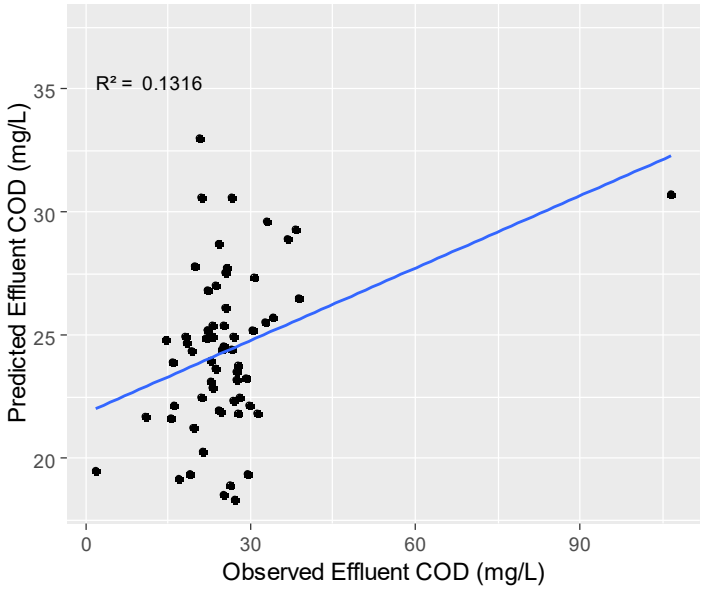
Table B.4 – Performance metrics of the MLR model for training and testing datasets for the prediction of effluent COD

	Training (mg/L)	Testing (mg/L)
RMSE	8.191	11.303
MAE	5.458	5.958

Figure B.4 – Regression plot of MLR between predicted and observed data of effluent COD for (a) training and (b) testing datasets



(a)

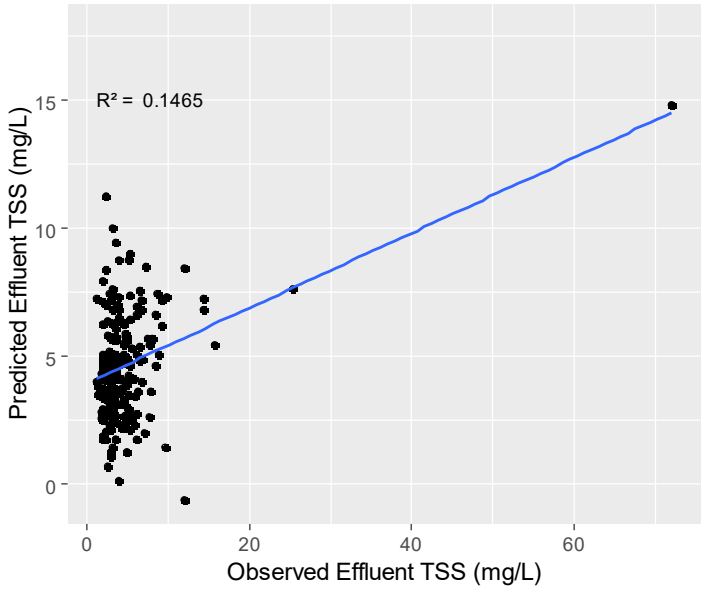


(b)

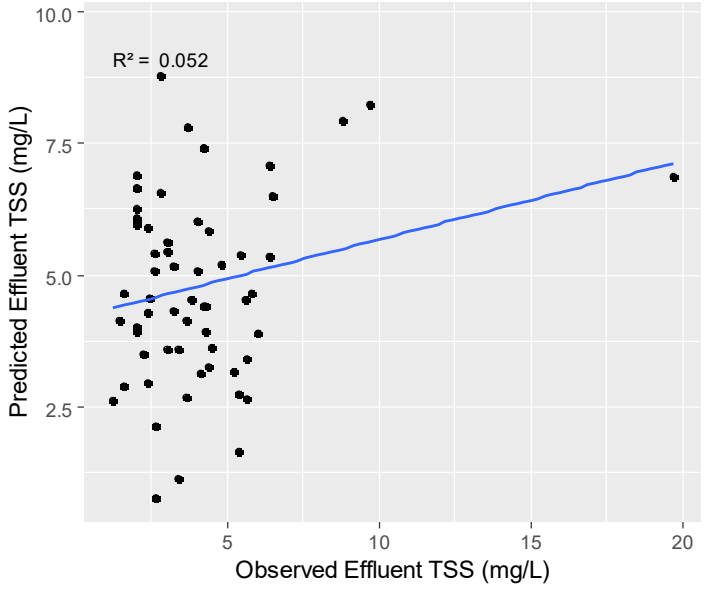
Table B.5 – Performance metrics of the MLR model for training and testing datasets for the prediction of effluent TSS

	Training (mg/L)	Testing (mg/L)
RMSE	5.072	2.919
MAE	2.494	2.181

Figure B.5 – Regression plot of MLR between predicted and observed data of effluent TSS for (a) training and (b) testing datasets



(a)



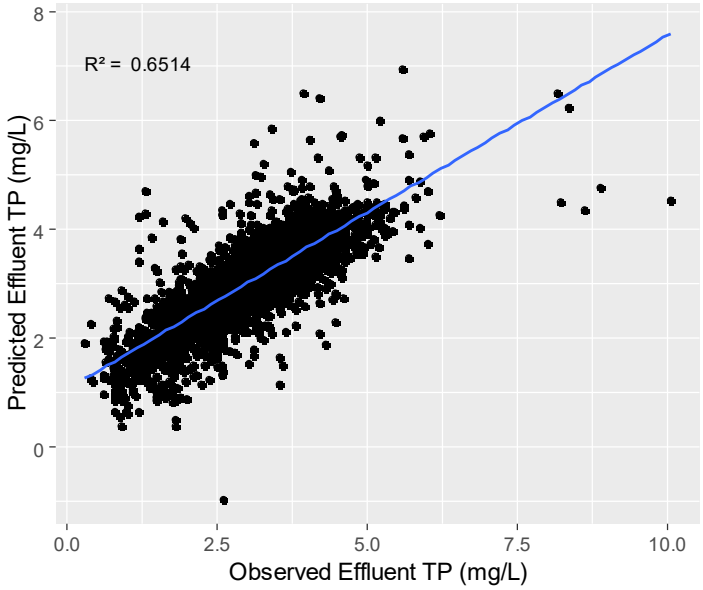
(b)

APPENDIX C – Multiple linear regression models results of the John E. Egan WWTP

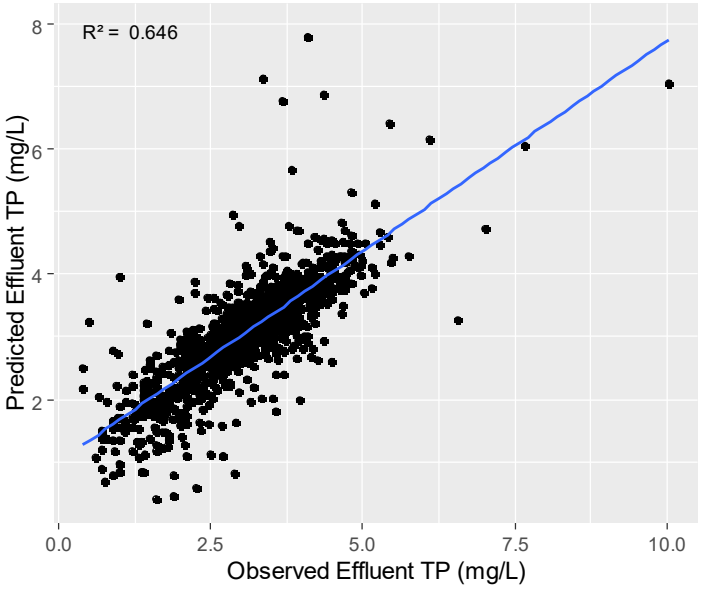
Table C.1 – Performance metrics of the MLR model for training and testing datasets for the prediction of effluent TP

	Training (mg/L)	Testing (mg/L)
RMSE	0.591	0.609
MAE	0.426	0.435

Figure C.1 – Regression plot of MLR between predicted and observed data of effluent TP for (a) training and (b) testing datasets



(a)

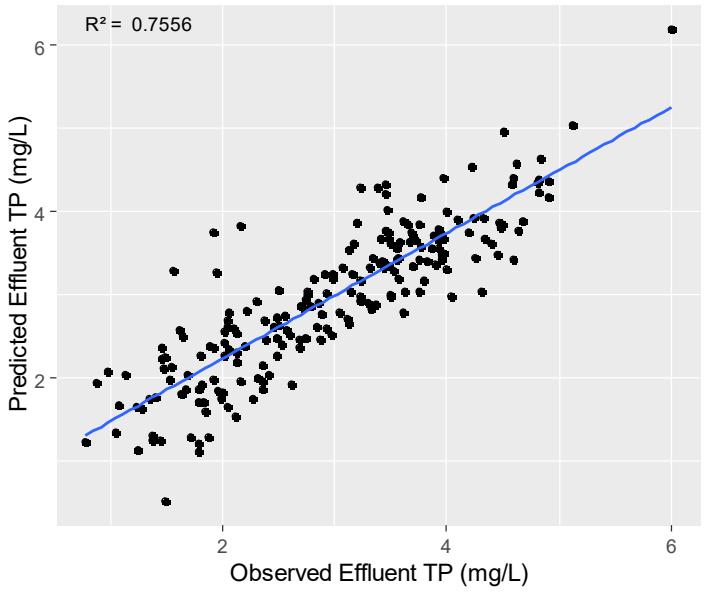


(b)

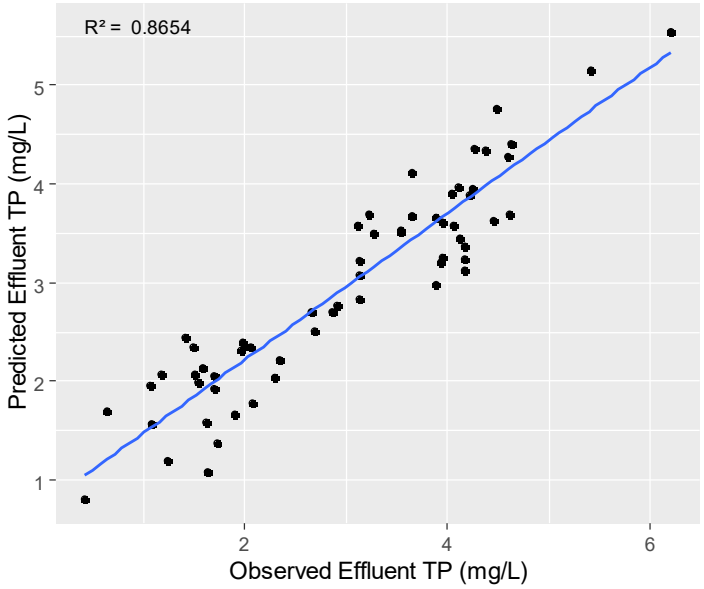
Table C.2 – Performance metrics of the MLR model for training and testing datasets for the prediction of effluent TP (n = 263)

	Training (mg/L)	Testing (mg/L)
RMSE	0.517	0.508
MAE	0.406	0.410

Figure C.2 – Regression plot of MLR between predicted and observed data of effluent TP for (a) training and (b) testing datasets (n = 263)



(a)



(b)