Revista Brasileira de Recursos Hídricos
Brazilian Journal of Water Resources
Versão On-line ISSN 2318-0331
RBRH, Porto Alegre, v. 26, e30, 2021
Scientific/Technical Article

https://doi.org/10.1590/2318-0331.262120210072

State-transition matrices as an analysis and forecasting tool applied to water quality in reservoirs

Matrizes de transição de estados como ferramenta de análise e previsão aplicada a qualidade da água de reservatórios

João Marcos Carvalho[1] &Tobias Bleninger[1]

[1]Universidade Federal do Paraná, Curitiba, PR, Brasil
E-mails: joao.huf.carvalho@gmail.com (JMC), tobias.bleninger@gmail.com (TB)

Received: May 18, 2021 – Revised: July 15, 2021 – Accepted: August 20, 2021

ABSTRACT

Water reservoirs have the function to control the temporal variability of the water availability, thus bringing greater security over these resources. The water quality of these systems must be adequate for their multiple uses, and one of the main tools to understand it, is mathematical modelling. Given the importance of the water quality, the goal of this paper is to develop an analysis that takes into account the randomness of the variables that affect the thermal and/or biochemical regimes of a reservoir. For this, it is proposed a combination of deterministic and statistical analysis, where the probabilities of occurrence of a given event are considered. Difficult factors, such as the lack of data on the water quality and other variables, were considered, which increases the replicability of the method. The research method is divided into three groups: Modelling, Scenarios and Compilation of these scenarios. Through modelling, a base layout is created, enabling the use of scenarios, which are statistically analysed, and compiled into a state-transition matrix. With this, a more robust tool to understand the dynamics of water quality in a system is obtained, since it is not heavily dependent on field measurements and is easily adaptable and replicable.

Keywords: State-transition matrix; Water quality; Reservoirs.

RESUMO

Reservatórios de água têm como função controlar a variabilidade temporal da disponibilidade hídrica, trazendo um aumento de segurança sobre esse recurso. A qualidade da água desses sistemas deve ser adequada para os seus múltiplos usos, sendo uma das principais ferramentas para o seu entendimento, a modelagem matemática. Dada a importância da qualidade da água, o objetivo deste artigo é desenvolver uma análise que considera a aleatoriedade das variáveis que afetam os regimes térmicos e/ou biogeoquímicos de reservatórios. Para isso, foi feita a combinação de análises determinísticas e estatísticas, de modo que as probabilidades de ocorrência de um determinado evento sejam consideradas. Fatores dificultantes, como a falta de dados de qualidade da água e outras variáveis, foram considerados, o que aumenta a capacidade de replicabilidade da análise. O método de pesquisa é dividido em três grupos: Modelagem, Cenarização e Compilação. Através da modelagem, um modelo é criado, possibilitando o uso de cenários, que são estatisticamente gerados, e compilados em matrizes de transição de estados. Com isso, obtêm-se uma ferramenta mais robusta para entender a dinâmica da qualidade da água em um sistema, dado que o método não é altamente dependente de medições de campo e é facilmente adaptável e replicável.

Palavras-chave: Matriz de transição de estados; Qualidade da água; Reservatórios.

INTRODUCTION

Reservoirs in general are fundamental systems for the development of society, since they increase the availability of a given resource, facilitating its management in the face of natural variability. In the case of water reservoirs, the ability to guarantee a certain supply, in other words, the chance that they will fail during their operation, is directly linked to the percentage of the average long-term flow that they manage to regulate, the variance of the flow and the volume of the reservoir.

Two of the various methods for determining the capacity of regularization and size of a reservoir are those of Moran (1954) and Gomide (1975). Both have similar bases and simplifications, and can be summarized as follows: The chance that a reservoir will fail ("dry up") depends on three factors - the regulated outflow, the probability distribution of the inflows and the volume of the reservoir. As the only random variable considered in the method is the inflow, the probability of the reservoir's response to a certain inflow, is equal to the probability of the inflow. Through these hypotheses, it is possible to use *n* different inflows to *n* volume scenarios, in order to allow a view of the general behavior of the system. When organizing these *n* scenarios and their probabilities in a matrix, a matrix that can indicate the probability that the reservoir is with a volume i and transits to a volume j is obtained, and this new matrix is the state-transition matrix of volumes for a given reservoir (Figure 1).

As a means of solving these difficulties, the application of models is used as a representation of the system; these representations are based on both physical equations, mathematical approaches and statistical adjustments. Currently with the advancement of data science, many of the new forecasting methods are based on machine learning techniques, (Tiyasha et al. 2020), both exclusively (Chen et al. 2018; Barzegar et al. 2018; Arefinia et al. 2020; Elkiran et al. 2019) and (Najah Ahmed et al. 2019), sometimes also coupled to models with physical principles (Jia et al. 2019). However, the major problem with this type of approach is the data available, since a large amount of well sampled data is required for the proper calibration and validation of the method (Erichson et al. 2020) and (Callaham et al. 2019).

With respect to water quality parameters, data is not only scarce, but also not necessarily well distributed and capable of adequately representing the main processes involved, (Chapin, 2015; Kozak, 2016) and (Ferreira et al. 2019). As an example, it is common for conventional water quality sampling methods to have fewer samples during extreme events, as there is a large security risk involved for the monitoring team to collect samples during these situations. Other techniques, such as the use of satellite images and remote sensing, have helped to reduce this database limitation, but there are still problems when dealing with some locations across the globe, (Damania et al. 2019).

With this type of sample limitations in mind, this work aims to establish a method that allows the understanding of the uncertainties of a system and also a reasonable ability to predict its behavior, without the need for a large amount of data for the objectives of the study. The use of state-transition matrix is not restricted to the method above, but it is widely implemented on a whole lot of situations: Thompson et al. (2002) applied this kind of approach to create a probabilistic characterization of tidal mixing for coastal regions, while Zhang et al. (2011) uses it to analyze changes in wetlands area. In the case of monitoring water quality, Kim et al. (2020) developed a transition matrix to assess and monitor chlorophyll-a on the Lower Nakdong River. The main difference identified was that those studies were based on a large number of observed data, whereas our novel approach is using a combination of observed data, complemented with intensive modeling data. As this opens up a large number of new opportunities, we believe that our approach brings a more generalized way of creating these robust analyses.

Therefore, an adaptation of the methods of Moran (1954) and Gomide (1975) was made, using complex mathematical models for the creation of state-transition matrices. Through this adapted methodology, a case study was carried out to assess the water quality of the Jurumirim hydroelectric plant reservoir in south Brazil, where it was possible to create a transition matrix that takes into account the uncertainties and dynamics within this system, through the use of scenarios run by the General Lake Model (GLM). Thus, establishing a method capable of helping in the management and decision-making about the conditions in which a reservoir may be in the future, taking into account the uncertainties of the natural processes that affect it.

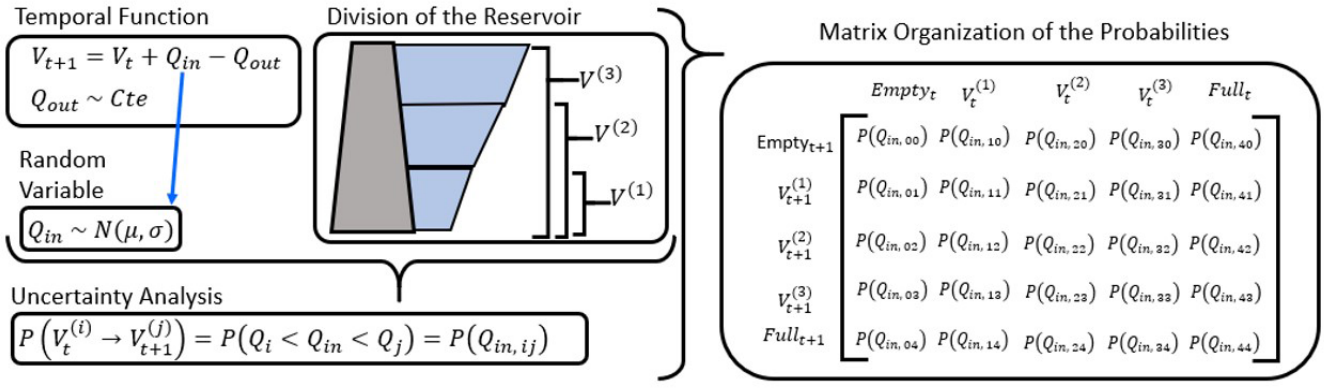

Figure 1. Flowchart of Moran (1954) and Gomide (1975) methods using state-transition matrices.

METHODS

State-transition matrix

The transition matrix aims to describe a Markov chain in a visual and grouped way, (Gomide, 1975), where normally the columns of the matrix represent the possible initial states, and the lines the possible final states of a system. Matrix analytical methods are popular as modeling tools, as they offer the ability to build and analyze, in a unified way, a wide class of stochastic models (Latouche & Ramaswami, 1999).

With this type of configuration, the sum of each column of the matrix must have a value equal to 1, since all the possibilities of the system must be included in the matrix, therefore there is no possibility of transition beyond the states used in it. Figure 2 exemplifies and illustrates how a transition matrix is made using the following hypothetical situation: Considering that the chance of occurrence of, for example, an algae bloom in a reservoir in any given year is 20%, and that this occurrence can be considered as statistically independent; the behavior of the reservoir in the face of this phenomenon can be described using a transition matrix. As an algae bloom is a high-risk situation from a management point of view, a single occurrence will be seen as a total management failure. For this to be incorporated into the matrix, it is considered that in the event of a bloom, the probability of the system staying itself with the bloom (management failure) is 100%.

In general, the assembly of a state-transition matrix requires the following items:

I. Definition of the States of the System: It is the most easily fulfilled, as it is enough to determine a criterion according to the need to understand the variations of the system (In the example of Figure 2 the states are: Normal and Bloom);

II. Calculation of the Transition Probabilities of each State: Several methods can be used in this step, however the greater the number of variables that govern the transition of states, the more complex its determination becomes. (In the example of Figure 2 these probabilities are pre-defined);

III. Markovian Memory System: The future state of the system should depend only on the previous state. (In the example of Figure 2 this criterion is met, because of the assumption of events being statistically independent).

Moran (1954) uses the water mass balance equation, Equation 1, together with the probability distribution of the random variables, inflow, to assemble the transition matrix, considering **1** time step for the transition. Since the process is considered Markovian, the multiplication of this matrix **n** times, provides the transition probabilities for **n** time steps in the future. With these characteristics, it is possible to predict the possibility of failure of a reservoir for **n** future years.

$$V_{t+1} = V_t + Q_t - R_t \tag{1}$$

Where,

V_t = Volume in the time step t

R_t = Regularized flow/Outflow (constant)

Q_t = Inflow: Following a Normal Distribution ~ N(μ, σ^2)

This type of approach is only possible due to the availability of an equation, or system of equations, that is reasonably easy to solve, making an application of this technique limited to more simplified models.

Another situation where a transition matrix can be created, is when there is enough data directly related to the main variable, as it was the case in the studies of recent publications, such as (Kim et al., 2020) and (Zhang et al., 2011). With this data a direct stochastic model can be applied to the variable. Nevertheless, this is a very restrictive case, since the lack of environmental data is common around the world. Given the difficulty of this types of direct approach to water quality in reservoirs, physical-chemical modeling of scenarios was used to remedy the lack of robust and simple equations, capable of being analyzed without using numerical methods. As each scenario is able to inform its initial and final state, since the model itself has a time scale, it is possible to arrange the results of **n** scenarios within a state-transition

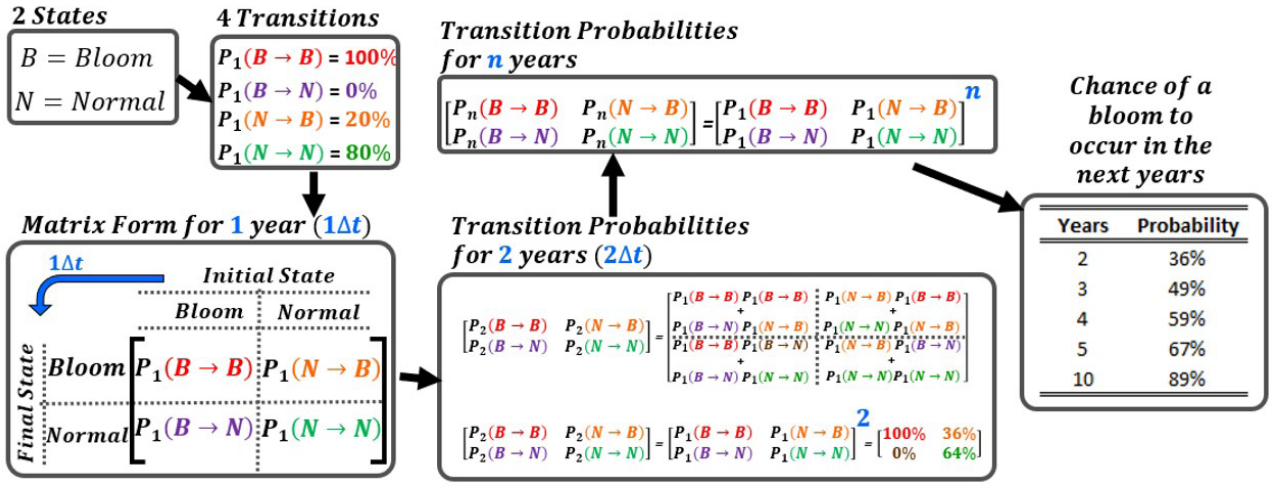

Figure 2. Example of the creation and properties of a state-transition matrix.

matrix; for this case study the General Lake Model (Hipsey et al. 2019) was used.

In this work, the states of the system are discrete intervals of the variables evaluated in the water quality modeling, while the probabilities are given by the statistical analysis of the model inputs. The input variables were considered statistically independent, so the multiplication of the probabilities of each input is equal to the transition probability of the model itself, fulfilling the last criterion.

For the determination of the probabilities, a normal distribution was adjusted for the annual averages of each input variable of the model. The distributions were then discretized in intervals, so that the probability of occurrence of a certain value is considered as the probability of occurrence of the interval in which the value is found. Figure 3 shows this process of discretization and attribution of the probability.

Reservoir modelling - General Lake Model (GLM)

The General Lake Model (GLM) is an open source and one-dimensional vertical model for the solution of the mass, energy and heat balance in natural and artificial lakes (Hipsey et al. 2019). The model was developed by the Aquatic Ecodynamics research group (AED) of the University of Eastern Australia through the initiative and collaboration of the Global Lake Ecological Observatory Network (GLEON) and the Aquatic Ecosystem MOdelling Network (AEMON).

The model has a fast execution time and can be integrated with several water quality models such as the LakeAnalyser (LA) (Read et al. 2011), the Framework for Aquatic Biogeochemical Models (FABM) and the Aquatic Ecodynamics (AED2) (Bruggeman & Bolding, 2014), thus complementing GLM. The model is ideal for systems with little horizontal variation, since it is 1D vertically; and presents great performance for modelling long periods of time or a large number of scenarios, since it requires little computational power, Hipsey et al. (2019).

The amount of input information that the model can absorb is high and can be seen in the Figure 4, however given its broad configuration options, not all of this data is essential, depending only on the degree of complexity required by each project. The main time series that govern the model are presented in Table 1. The source of this time series was the Hydro Power Plant (HPP) Jurumirim operator (China Three Gorges Brasil, 2020), National Water Resources Information System database (Agência Nacional de Águas, 2020a) and from the Environmental Company of the State of São Paulo (Companhia Ambiental do Estado de São Paulo, 2020).

The Level x Area x Volume curve of the reservoir was obtained from the Sistema de Acompanhamento de Reservatórios (Agência Nacional de Águas, 2020b), and is shown in Figure 5.

The model implemented did not use cloud and solar radiation data - as they were calculated by the GLM itself using relative humidity, air temperature and reservoir latitude. This choice was made because the intention is to create a model as simple as possible. The last information needed is the precipitation; this data is found indirectly inside the flow data, due to a water balance that will be shown later.

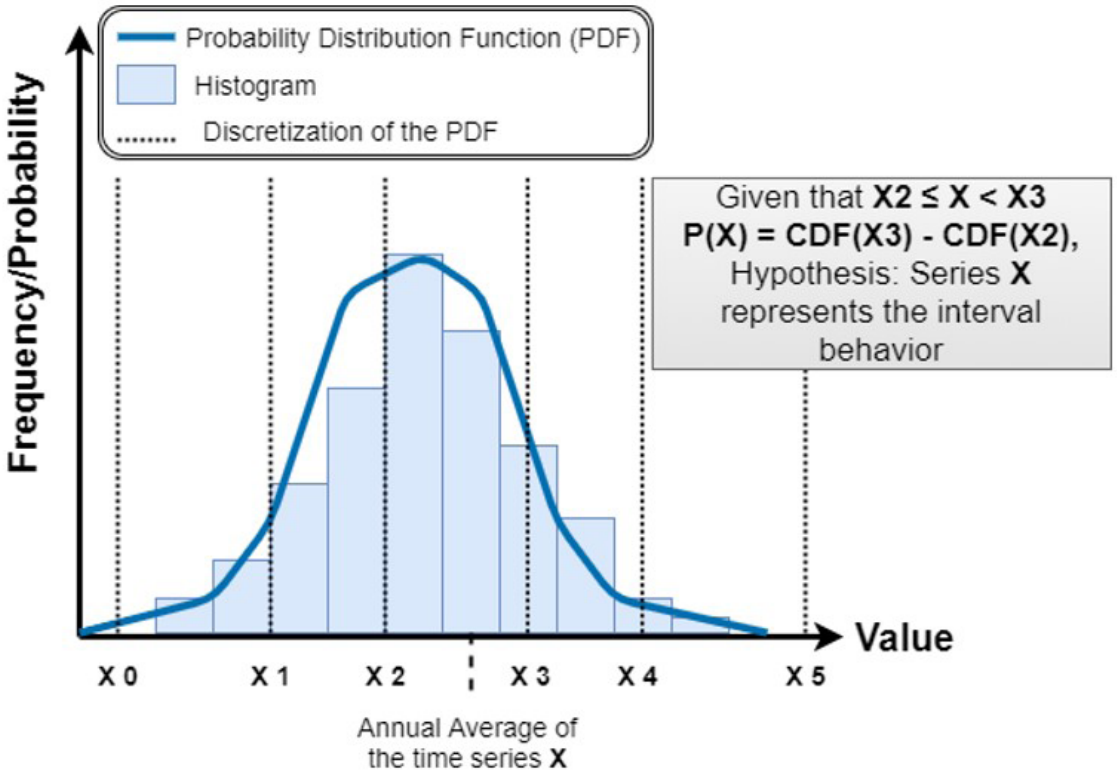

Figure 3. Example of discretization and probability assignment for a given value.

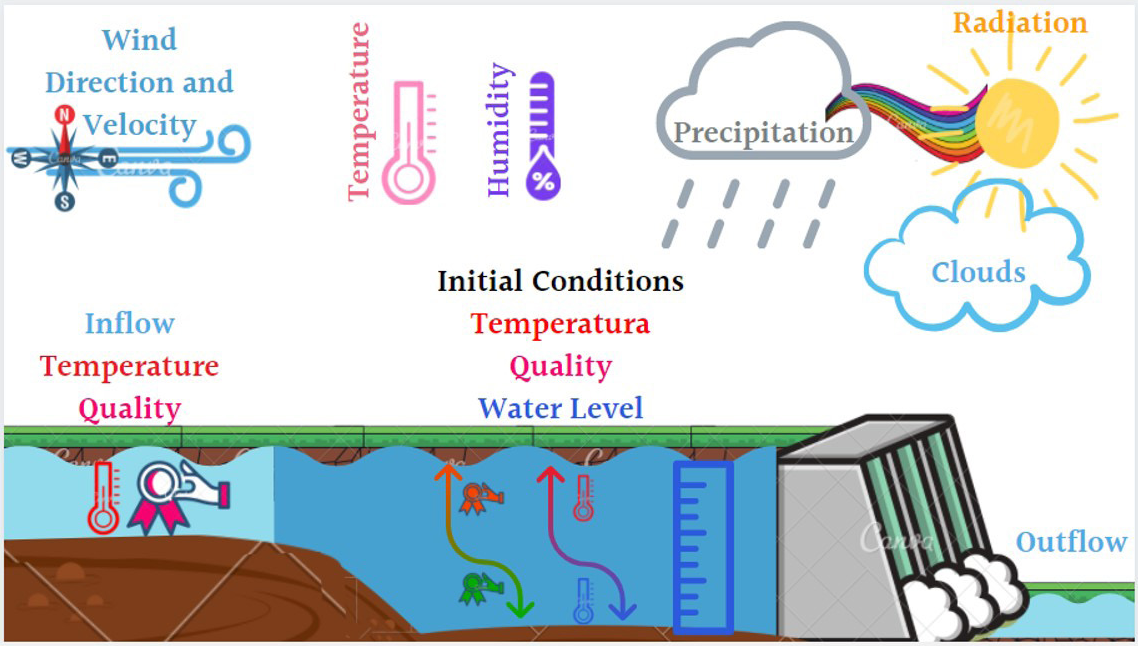

Figure 4. Layout of the necessary inputs for the model.

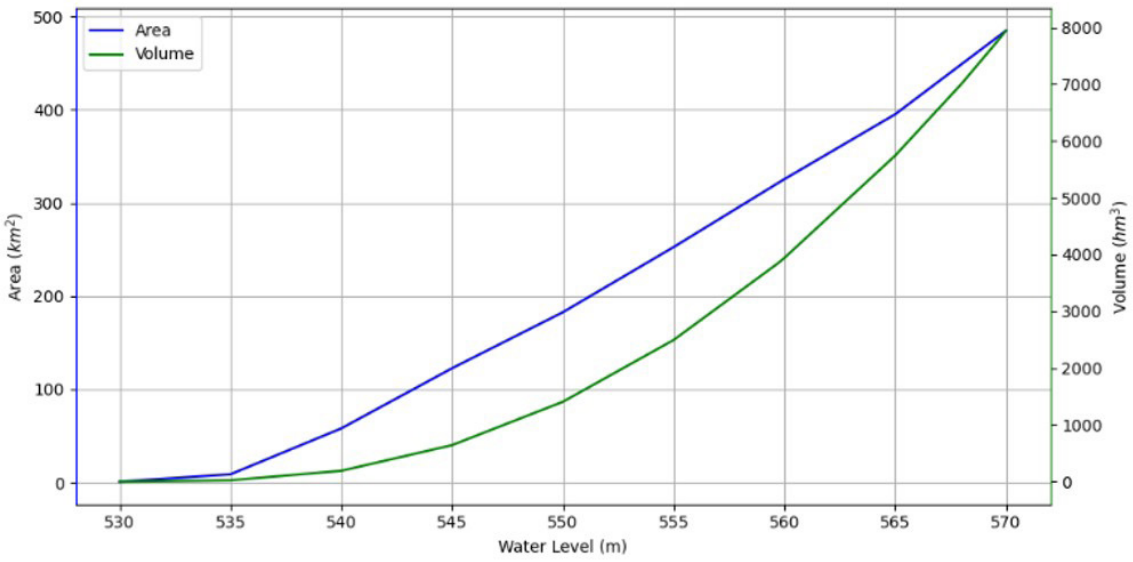

Figure 5. Level x Area x Volume curve of the Hydro Power Plant Jurumirim reservoir.

Table 1. Source of each temporal input used.

Source	Variable	Reference
CTG Brasil	Total Inflow Total Outflow Water Level	CTG-Brasil (China Three Gorges Brasil, 2020)
CFSR Model (ds0930.0 and ds0940.0)	Air Temperature Relative Humidity Wind Velocity Inflow Water Temperature	Saha et al.(2010) and Saha et al. (2011)
Water Quality Gauges 64080000 (HidroWeb), 64081000 (HidroWeb), 64082000 (HidroWeb) and PARP02100 (CETESB)	Dissolved Oxygen Total Phosphorus Nitrate Ammoniacal Nitrogen	CETESB (Companhia Ambiental do Estado de São Paulo, 2020) and HidroWeb (Agência Nacional de Águas, 2020a)

Case study – hydro power plant Jurumirim

As a case study for the water quality modelling, the reservoir of the HPP Jurumirim was selected, being the first large reservoir of the Paranapanema river cascade. It was chosen because it was part of a R&D project between the Federal University of Paraná and the Brazilian National Water Agency, with focus on hydrodynamic modelling and classification of water quality in the Paranapanema river basin.

The reservoir has a volume of 7007.1 hm^3, maximum operating depth of 38 m and a flooded area of 499 km^2. The reservoir has an inter annual regularization capacity and different uses, such as hydroelectric power generation, aquaculture and regularization of flow for the Paranapanema river. Power generation is its most important function, with an installed capacity of 100.9 MW. The reservoir is inside the Paranapanema river basin, and is located in the state of São Paulo, Brazil.

The water body has two large tributaries, the Paranapanema and Taquari river, responsible for more than 80% of the total inflow, the first one larger and more expressive than the second one. The water level and long-term average inflow of the reservoir are 35.6 m and 255 m3/s, respectively. The hydrographic basin has approximately 17.900 km^2, and the length of the Paranapanema river until the reservoir is 321.2 km.

For the model calibration the following data was used: water quality and morphology data from ANA (Agência Nacional de Águas, 2020a); hydrological time series from the reservoir operator (CTG-Brasil); meteorological data from the CFSR reanalysis meteorological model (Saha et al. 2010) and (Saha et al. 2011). The Figure 6 presents the reservoir, along with the geographic points of the time series that were used.

The final setup of the model consisted of a single point of entry and exit of water. The single entry and exit points were considered as an equivalent representation of all the entries and exits of the reservoir. This assumption was made on one hand, because of the lack of information about water quality and quantity on the different punctual and diffuse entrances along the reservoir. On the other hand, commonly reservoirs are analyzed as a whole, not considering spatial distributions. This is justified by the fact that most reservoirs act as sinks of substances, showing considerable variations only over the vertical, usually represented as thermal and chemical stratifications. Those vertical variations are affected by in and outflowing total mass (water and substances), as well as by processes (i.e. heat transfer and nutrient fluxes) at the air-water and water-sediment interfaces. It is thus not important to know at which location mass or thermal energy enters the system, but to know their total temporal variations.

The setup was configured for a better understanding of the region closest to the dam, since it is the place with the greatest depths and potential for vertical stratification. However other depths can easily be analyzed by considering the vertical distribution of temperature and concentrations.

Calibration of the model

The basecase modelling was done for the period between the years 1990 and 2018, using a warm-up time of 1 year for

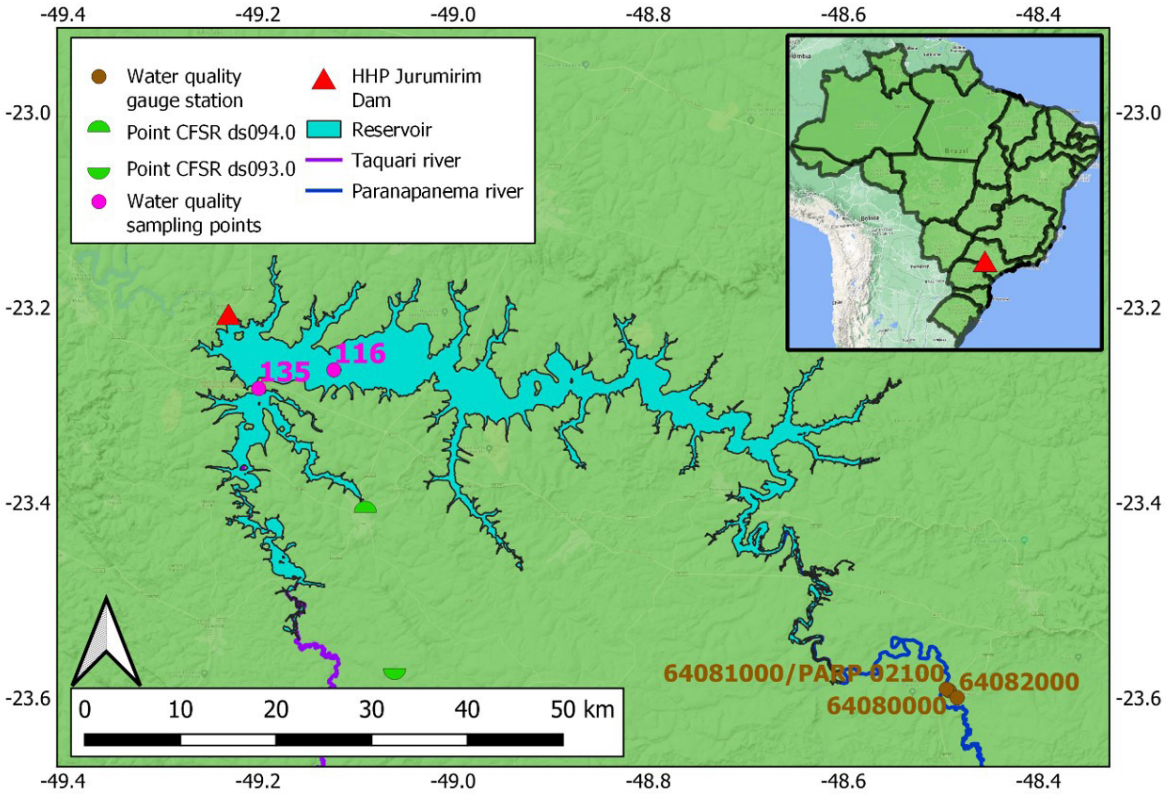

Figure 6. Positioning of the reservoir, water quality monitoring stations and monitoring points inside the reservoir used for calibration.

convenience of configuring the model; this value being almost 4 times greater than the minimum warm-up time previously estimated. The model was calibrated using temperature and water quality profiles measured in 04/2011 and 09/2011, which are the only data available about the water quality of the reservoir.

Figure 7 shows the results obtained in the calibration, with the dotted lines representing the measured data, the continuous ones the calculated data and the color indicates if the profiles have the same date. In order to add the space-time interpretation of the calibration, a comparative grouping of the measured and modelled data by intervals of depth was added, where the lines with a confidence interval of 5% and 95% represent the data from the model and the dots the measured data.

Since the observed data presents a lot of variations on a very short periods of time and space, which were not consistent with literature and physical-chemical relations, the calibration focused on trying to follow average patterns of the observed data. These extreme variations can be seen for all the water quality parameters. As an example: On 09/2011 the oxygen of the bottom layer of the reservoir (0 – 10 m], presented observed data of 5.5 mg/l and 8.5 mg/l.

Given this observed data, and the type of representation by GLM, average behavior of the layers, temperature was the only parameter that was well represented by the model. GLM was able to represent the average patterns of dissolved oxygen, and in some degree also nitrate. The model was unable to represent the measured total phosphorus profiles. Given the lack of observed data on 09/2011 for these last two parameters, it was not possible to evaluate inter-annual trends of nitrate and phosphorus.

The calculated profiles showed mean absolute errors of 1.54 °C, 0.79 mg/l, 0.14 mg/l and 0.04mg/l, respectively for water temperature, dissolved oxygen, nitrate and total phosphorus concentrations. As the model follows physical-chemical principles, it is capable to reproduce well their relative variations, due to the varied forcings. The results thus were considered of being enough for the specific case, even not matching absolute values with high accuracy. Much more improved calibration efforts furthermore would not change the results of the applied methodology to

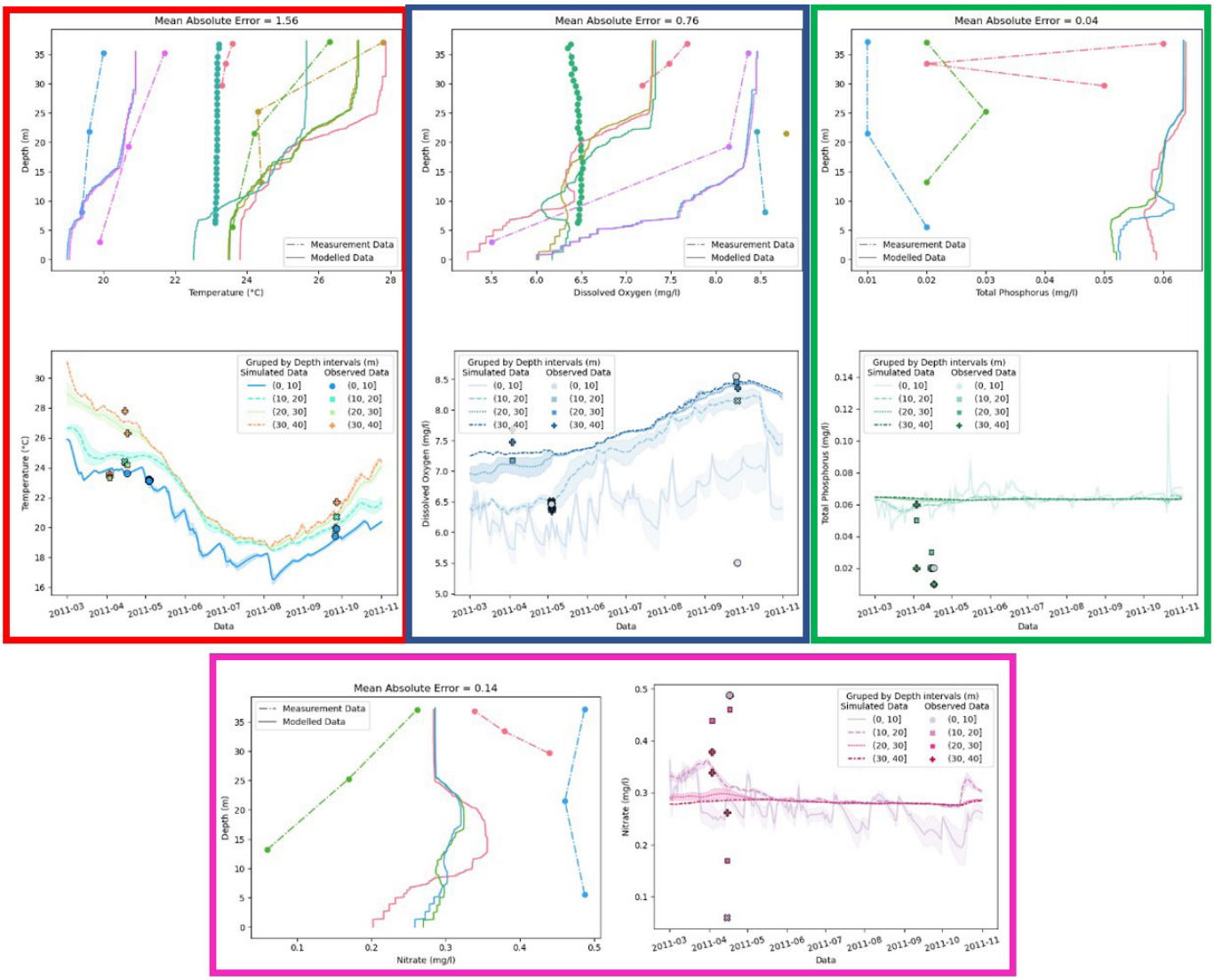

Figure 7. Calibration results considering the measurement points ID 135 and ID 116.

obtain a transition matrix, but only would create a shift in those. Remembering that the objective of this paper is not to perfectly calibrate a water quality model to represent absolute values with high accuracy, but to present an approach to assess uncertainties from the combination of deterministic modelling and statistical/stochastic analysis.

Probability distribution and statistical independence

The variables considered as random were flow, air temperature, relative humidity and wind speed. Due to the reduced amount of data, the water quality parameters were not considered random, being used as time series on a daily scale, generated from a Markovian model and the mean and variance of the observed data.

In total, four normal probability distribution functions (PDF) were adjusted, one for each random variable. To determine the normality of the series, the Kolmogorov-Smirnov (KS), Anderson-Darling (AD) and Shapiro-Wilk (SW) tests were used, with a significance level of 5%. Table 2 presents the p-values obtained along with the acceptance or non-acceptance of the normality of each variable, while Figure 8 contains a compilation of the histograms, with the adjusted probability distributions.

The test was also applied to a significance level of 5%, in order to assess the statistical independence between the four variables, Table 3. This was done because if the statistical independence between these four random variables is accepted, the probability of occurrence of a scenario generated by the combination of these variables, can be taken as being the product

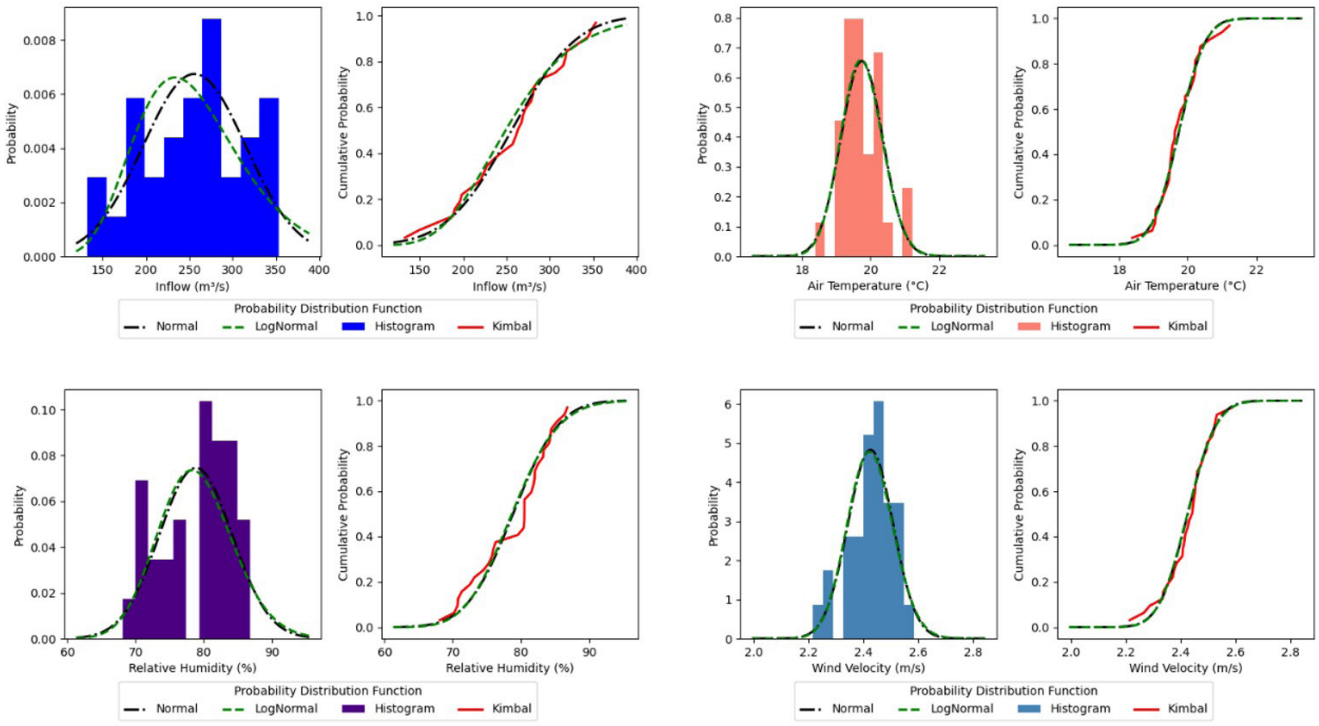

Figure 8. Histograms together with Normal, Lognormal and Empirical (Kimbal) PDF's of the random variables.

Table 2. Acceptance of the normality of each variable and p-values of the tests.

Variable		Kolmogorov-Smirnov	Shapiro–Wilk	D'Agostino
Annual Averages	Inflow	~96% Yes	~67% Yes	~60% Yes
	Air Temperature	~91% Yes	~76% Yes	~50% Yes
	Relative Humidity	~16% Yes	~4.6% No	~12% Yes
	Wind Velocity	~77% Yes	~56% Yes	~24% Yes

Table 3. Relationship of dependencies between variables and p-values of the test.

Variable	Inflow	Air Temperature	Relative Humidity	Wind Velocity
Inflow	0% Depend.	16% Independ.	49% Independ.	28% Independ.
Air Temperature	16% Independ.	0% Depend.	17% Independ.	44% Independ.
Relative Humidity	49% Independ.	17% Independ.	0% Depend.	34% Independ.
Wind Velocity	28% Independ.	44% Independ.	34% Independ.	0% Depend.

of the individual probability of the variables that created the scenario. Such approach avoids the need to adjust a joint probability distribution function for the variables, simplifying the process.

Almost all variables were accepted as normally distributed by the 3 applied tests, where only the relative humidity did not pass the Shapiro-Wilk test. As at least two of the three tests applied were accepted for normality, all four random variables were assumed to be normally distributed. The four variables were also considered independent, according to the χ^2 test for a significance level of 5%.

From these previous statistical results, individual normal distributions were adjusted for each of the variables, and these distributions served as the basis for the formulation of the modeling scenarios and the inference of the transition probabilities used in the creation of the state-transition matrix.

Mass generation of models/scenarios

As it is not possible to evaluate an explicit equation for this kind of reservoir modelling problem, scenarios were used as a method to overcome this problem. The scenarios were created from the observed series, being selected series that represent a certain interval of the adjusted probability distribution of each random variable. The outflow and initial water level were treated in such a way that they were always grouped with their respective observed inflow, the same was done to the inflow water temperature and the air temperature. This type of grouping was adopted to avoid inconsistencies during the creation of the scenarios, since the variables within these two groups are directly dependent.

The time series used were selected from defined intervals within each adjusted normal distribution. The intervals were created as follows: By defining a number **n** of intervals, **n-1** points equidistant on the x-axis of the probability distribution function are estimated, and these points are the limits of each interval of the PDF. In this work an **n** equal to 10 was used, and since the normal distribution is contained between $-\infty$ and ∞, the first and last created points were forced to be the points where the cumulative probability distribution function (CDF) is equal to 1% and 99%.

The time series were used on a daily scale, with approximately 1.1 years in length, to guarantee the transition from January one year to January of the next year, plus 5 months of warm-up to the model. The warm-up time of 5 months was defined and tested as sufficient to prevent interference from the initial conditions within the results of each scenario.

In addition, as a simplifying hypothesis, each interval was associated with a series of observed data, where this series is considered representative of all the possible behavior from series in that interval. The selected series were those that had their annual average closest to the center of the interval.

In this process, not necessarily all intervals may have observed data, and if an interval does not have any measurement, the closest of all series to the center of the interval is chosen and adjusted by a factor, in order to create a series that is inside the interval. The correction factor used was the division of the value of the center of the interval by the average of the chosen series, so the average of the new series is mandatorily equal to the center value of the interval.

Figure 9 presents the intervals and the position of the average of the selected series, where the black dots are observed data and the red dots the adjusted data by a factor.

With this approach 10,000 scenarios were generated by the combination of the 4 random variables and their 10 representative series, and modelled by using the basecase setup created during the calibration procedure of GLM.

Compilation

For the final arrangement of the data and assembly of the transition matrix, the results of the modeled scenarios were compiled together with the statistical analysis of the series that generated each scenario.

In order to mitigate possible extremes and make the matrix more usable, from the point of view of reservoir management, the initial and final state of each water quality variable was considered as the average of the first and second January of the scenario. In this way, a single outlier is not able to distort the positioning of a scenario within the matrix, in addition to providing a more interesting temporal resolution for the management of the system.

Each model has an initial and final value of the water quality variables, thus defining its position within the grid of the transition matrix. Each scenario also has an associated probability of occurrence, which is obtained through the product of the probabilities of the series from the random variables that generate the scenario. Remembering that each of the series represents a range of probabilities within the normal distribution of flow probabilities, air temperature, wind speed and relative air humidity.

The criterion regarding the time interval of the state transition is malleable, as an example, instead of using a 1 year interval between two consecutive Januarys, like in this work, transition probabilities with an interval of 2 days or 10 seconds or 50 decades could also be evaluated. The definition of this period is given by the type of analysis that is intended to be done.

With the information of initial state, final state and probabilities it is possible to create the state-transition matrix, since the states represent the coordinates within the matrix, while the probability fills the cells of the matrix. If two or more scenarios have the same coordinates within the matrix (initial and final state), the probability of occurrence of that coordinate is taken as the sum of the probabilities of those scenarios.

As the reservoir of the HPP Jurumirim presents thermal and chemical stratification, regions of analysis for the matrices were defined: depth averaged vertical profile; Upper Third of water column; Middle Third of water column; Lower Third of water column.

RESULTS AND DISCUSSIONS

General matrix

From the information of initial, final state and probability of each scenario, a transition matrix with discrete intervals was created.

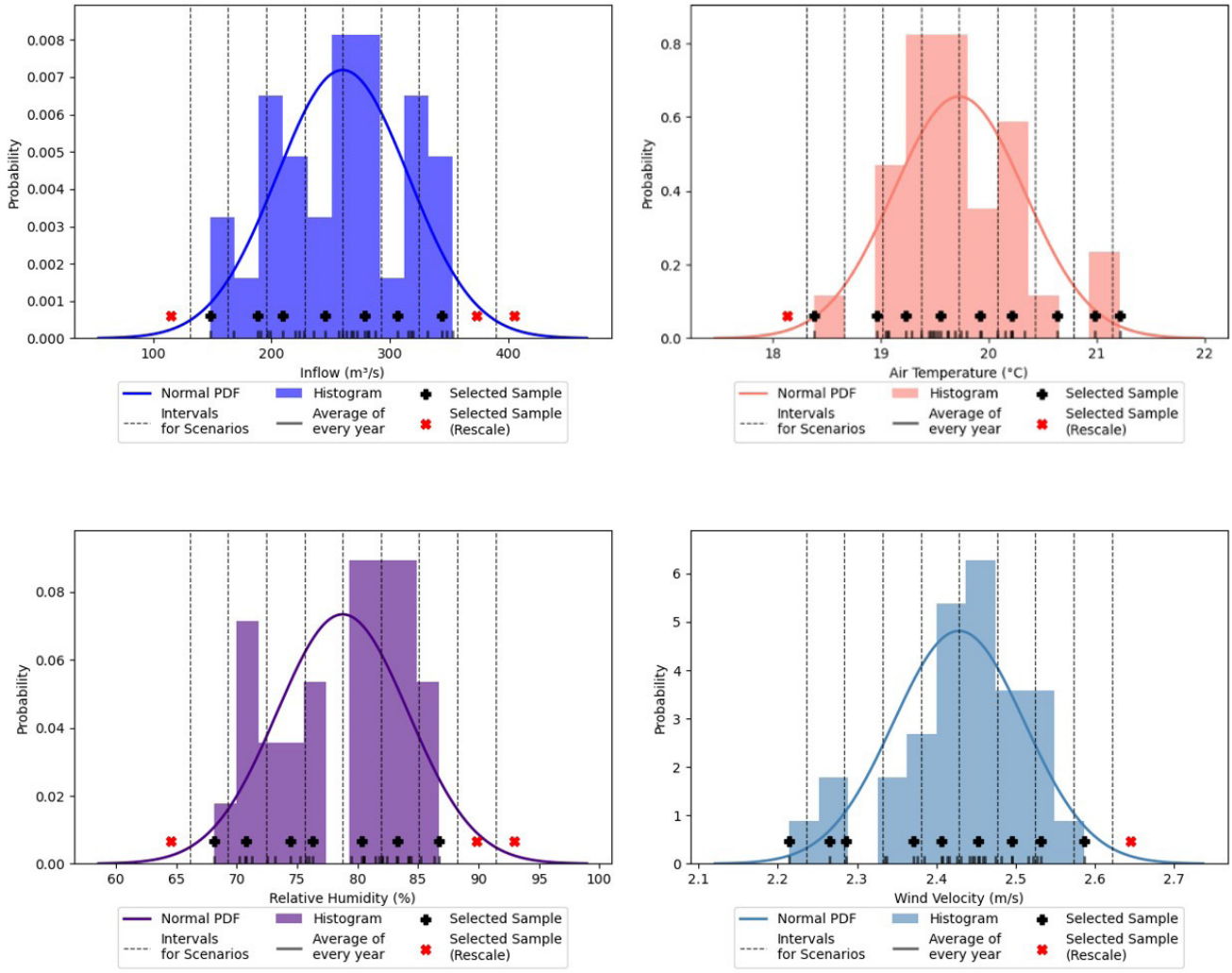

Figure 9. Adjusted Normal distributions with 10 intervals for the creation of scenarios, together with the series that the behavior of the interval.

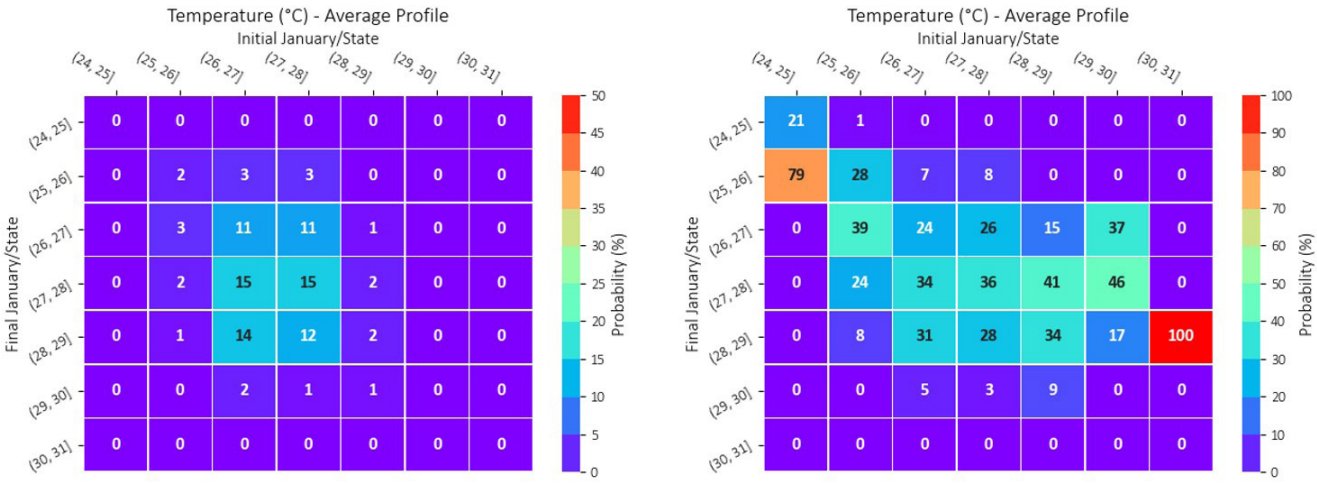

Figure 10. Probability matrix (left) and state-transition matrix (right) – Depth average temperature profile of the water column.

The columns of the matrix present the possible initial states and the lines the possible final states of the reservoir. Figure 10 presents this process for the average water temperature in the reservoir. As an example, the interpretation can be done as follows: The probability that the reservoir will have an average temperature within the range (27°C, 28°C] in January of the year X_i and between (28°C, 29°C] in

January of the year X_{i+1} is 12%, that is, the chance that this type of temperature transition will occur on two consecutive January is 12%.

The initially generated matrix is not exactly a state-transition matrix, it is necessary to carry out a marginalization process for each of the columns to obtain the state-transition matrix. The marginalization process consisted of normalizing each probability by the sum of the column in which the probability is found. Again, Figure 10 brings the results of this marginalization. In this case the interpretation must be made as follows: Since the system's initial temperature is within the range (27°C, 28°C], the probability that there will be a transition to the range (28°C, 29°C] is 28%.

As the presentation of all the generated state-transition matrices would become repetitive, due to the large number of parameters and the division of the reservoir into regions, some cases were selected as the most interesting for analysis (the complete results can be found in Carvalho & Bleninger, 2021).

In Figure 11 the transition matrices for dissolved oxygen and the Upper Third, Middle Third, Lower Third regions of the water column can be observed. These regions were selected, because oxygen presented a large vertical variation. Also, this parameter presents a stratification similar to thermal stratification, thus representing the general behavior of both parameters.

As expected, the decay of the dissolved oxygen along the water column is noted in the matrices. In the surface and middle layers, the stability of the system is greater than that of the bottom, since it has a small chance of transition of more than one state in the matrix. The upper third (surface) has a greater chance to have concentrations between 7.5 and 8 mg/l, while the middle third (center) between 7 and 7.5 mg/l.

The results for nitrate, ammoniacal nitrogen and total phosphorus are shown in Figure 12, with the matrices of the depth averaged profile and the lower third region. These regions were

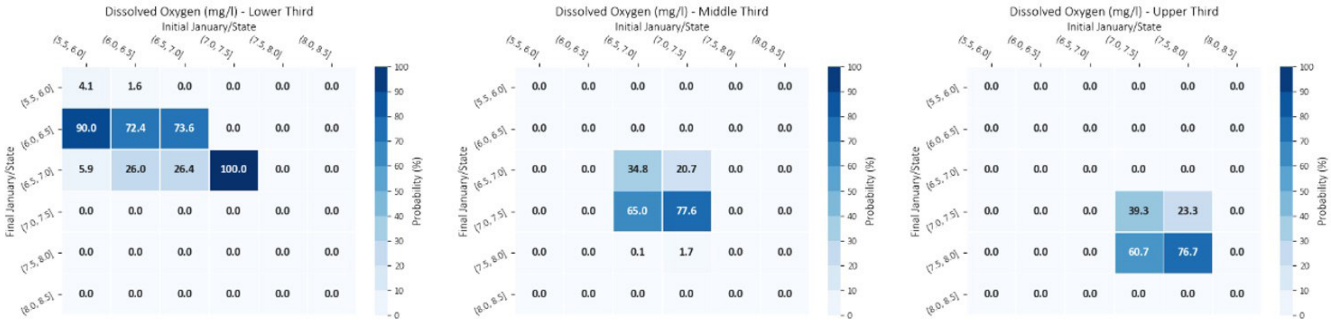

Figure 11. Transition matrices for the upper third, middle third and lower third of the water column – Dissolved Oxygen.

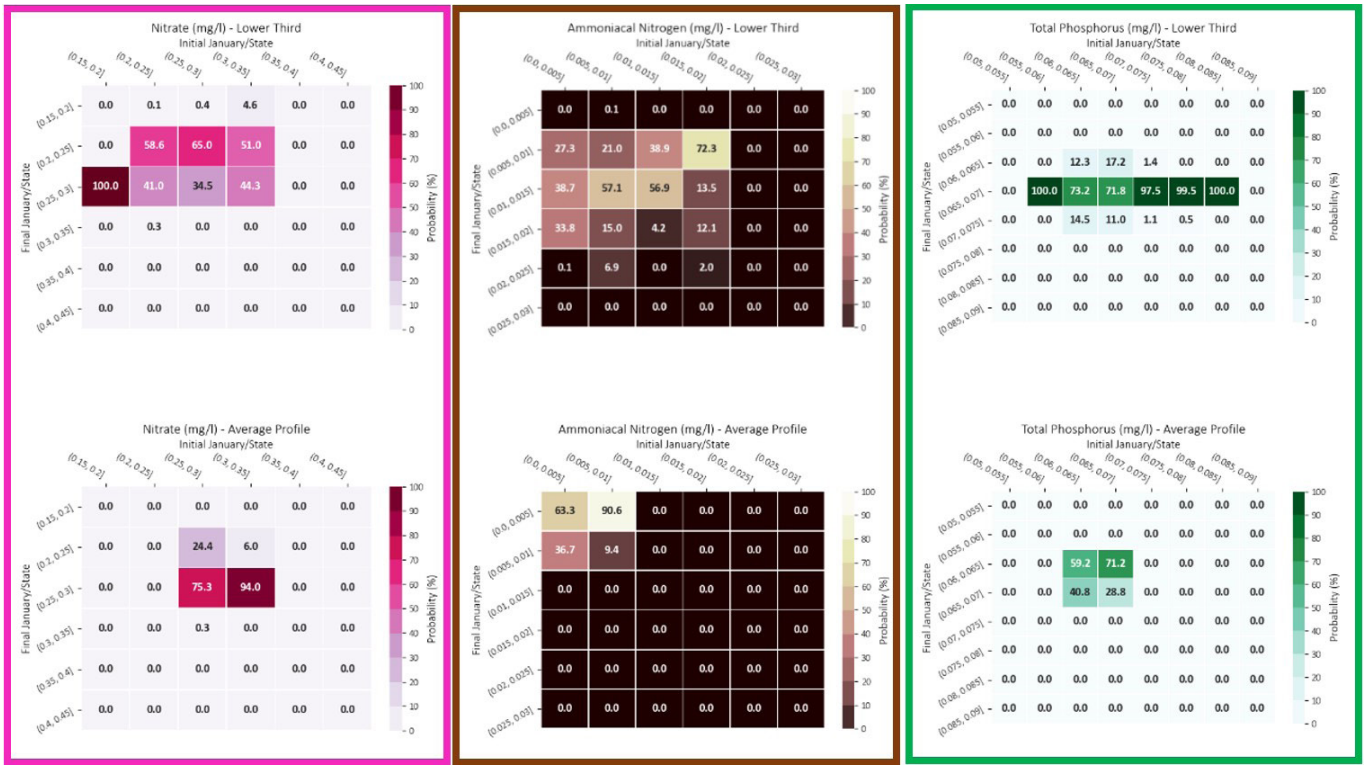

Figure 12. Transition matrices for the Depth averaged profile (first line of plots) and for the bottom third (second line of plots) region of the water column.

chosen because these parameters have relatively stable average profiles, with greater variations only at the bottom of the reservoir, due to oxygen consumption of the sediment.

All 3 parameters have a much more pronounced variability in the bottom third region of the reservoir, and compared to the average profile, the number of states doubles from 2 to 4.

Censored matrix

The matrices presented here provide the global information of the system, however, yet another advantage of this approach is that a priori assumptions can provide more insight into the analysis. As an example, only the scenarios where the average annual flow was greater than 250 m^3/s and the relative humidity superior to 80% were analyzed, simulating a case that the reservoir will pass through a year with higher precipitation than the average. In this way, only the scenarios that meet these two conditions were used, totaling 2000 scenarios.

Figure 13 presents the comparison between the complete and the censored state-transition matrices for dissolved oxygen and the depth averaged profile.

Through this process it can be noted that a priori knowledge of the possible future behavior of the system, results can be narrowed as presented by the matrix. Using Figure 13 as an example, it is observed that for the censored matrix, it is no longer necessary to take into account events where the dissolved oxygen in the system may become less than 6.5 mg/l. This type of analysis is useful for decision-making processes, as it allows a more precise analysis of the system's uncertainties.

Simplified validation

One of the disadvantages of the method presented so far is the difficulty in validating the results obtained, since one of its purposes is to remedy the lack of information on a variable that is capable of being mathematically modeled.

As there is no measured data for consecutive January about the water quality of the Jurumirim HPP reservoir, a different approach was made to validate part of the hypothesis adopted. The nonexistent observed data was replaced by the data modelled continuously between 01/1990 and 01/2019, with the same configurations already used in the GLM.

This way 30 years (with 30 information for Januarys) were simulated, creating 29 pairs of transitions that represent the average behavior of the reservoir. With this approach it is possible to at least identify if the simplications used, like the warm-up time and the definition of representative time series by intervals in the PDF's, can be accepted.

The evaluation of the validation was qualitative only, using the plot of the probability matrix together with the 29 pairs of transitions of each water quality parameter. Thus, if the 29 transitions observed in the continuous model are similar to those in the probability matrix, the model can be considered as sufficiently representative. The Figure 14 presents these plots.

Only for the parameter of total phosphorus a slight displacement was observed in the 29 pairs of transitions, approximately 0.05 mg/l of displacement. This may have occurred due to insufficient warm-up time of the model, since total phosphorus was the parameter that took the longest time to converge during the warming-up tests of the method. However, the general results are still considered satisfactory.

Multiplication of the matrix

Another capability of using the transition matrix is the following: Given that the system has no autocorrelation on the chosen time scale, it is possible to raise the transition matrix by a number **n**, so the time interval of the transition evaluated by the matrix is equal to **n**. As an example, all matrices shown until now are defined by the time interval of 1 year, and in this case **n** is equal to 1.

For the transition matrix to have this property, it is necessary to verify the non-autocorrelation between consecutive states, because in this way the state can be considered time independent.

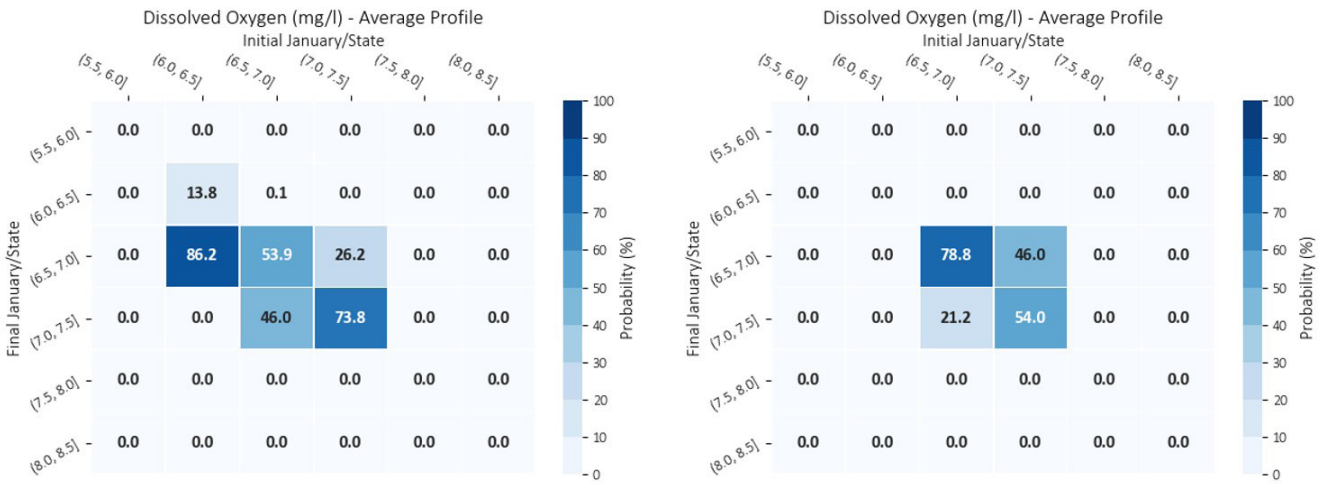

Figure 13. Transition matrix (left) and censored transition matrix (right) – Average of all the profiles of dissolved oxygen.

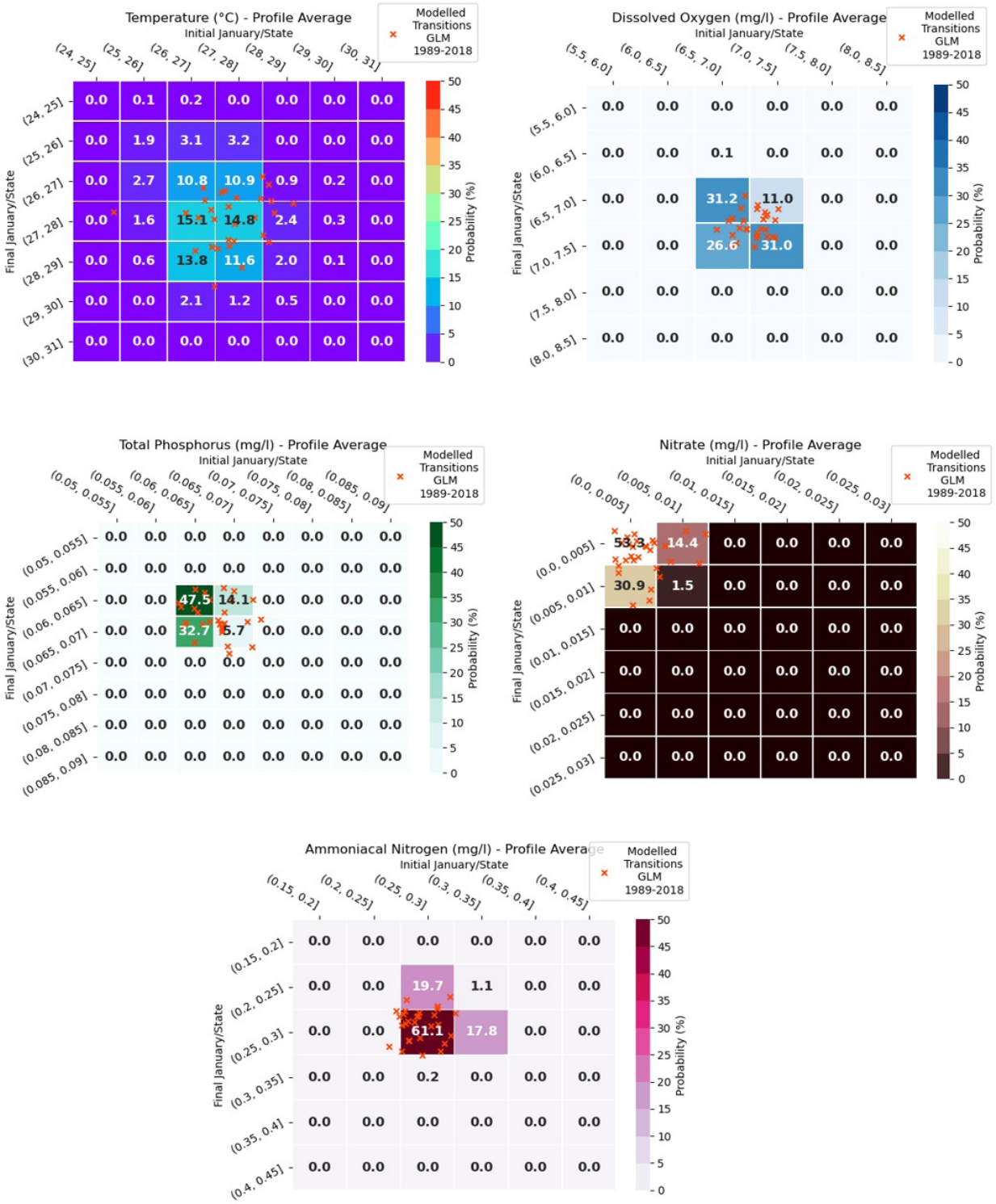

Figure 14. Transition matrices together with the 29 transitions modelled.

To fulfill this criterion, the autocorrelation function (ACF) was plotted for the 5 water quality parameters, as illustrated in Figure 15. Through the 5 graphs it is possible to identify that the hypothesis of non-autocorrelation can be accepted for all parameters; enabling the use of the multiplicative process of the state-transition matrix.

It is worth remembering that for non-chaotic systems like the present model, after a certain value of n, occurs a stabilization of the matrix probabilities. Figure 16 presents this process for the depth averaged profile of water temperature. The convergence time of this multiplicative process was approximately the same for all parameters, between 7 to 9 multiplications.

This procedure can also be extended to censored matrices, however with limitations, given that a decrease in the number of scenarios can generate transition matrices with discontinuities,

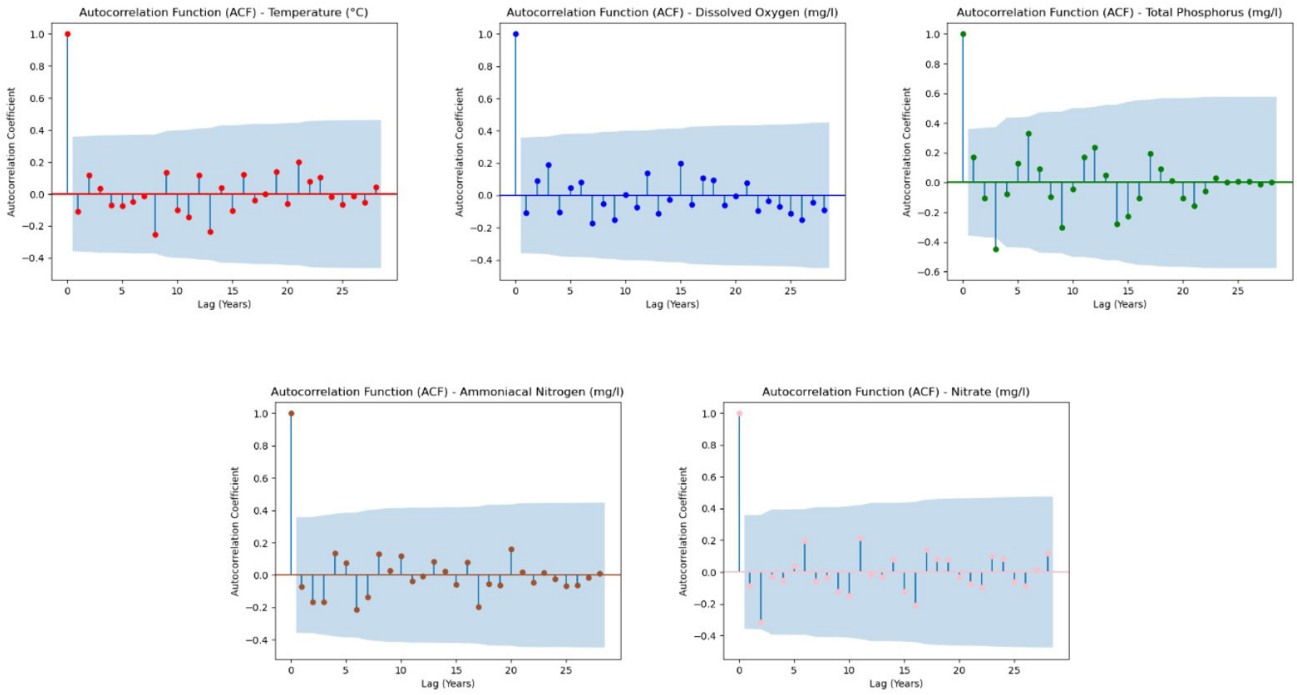

Figure 15. Autocorrelation functions for all the water quality parameters modelled.

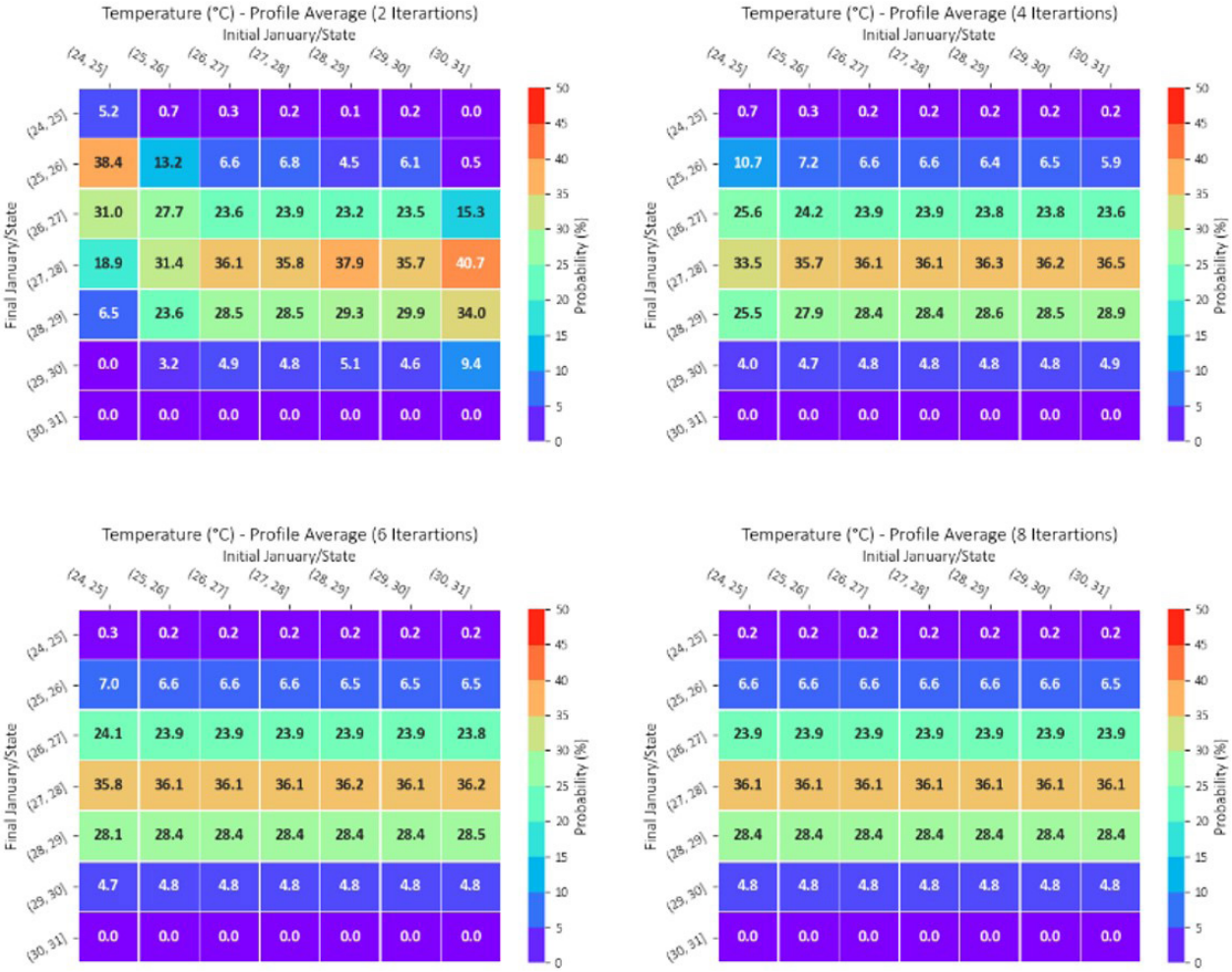

Figure 16. Transition matrices powered by n's equal 2, 4, 6 and 8 – Water temperature for the profile average.

causing problems of convergence. Figure 17 illustrates an example of this limitation, where a discontinuity appears in the transition matrix, in column (25 ° C, 26 ° C] and line (27 ° C, 28 ° C]; because of this, when powering the matrix by n, all the matrix converges to zero, thus showing unreliable results.

Application example

As an example of the application of the transition matrix, a hypothetical scenario was created. For this, the nonexistent observed data was replaced by the data modelled continuously between 01/1990 and 01/2019, and already presented in the validation chapter.

The idea of the example is to imagine being a reservoir manager at the beginning of the year 1990, but with the transition matrix in hand as a management tool. The objective is to use the matrix as a support tool for understanding how the water quality of the reservoir will be in the next year. The parameters of dissolved oxygen and temperature were used in this example.

Figure 18 has an example of the proposed merger between time series and state-transition matrix, where for each new sample, the matrix is consulted and the probabilities of the next states are inserted within the time series. As time passes, the matrix is consulted in each new January (initial state), and the column of the matrix referring to the interval that this new January is in, is inserted within the time series, right in front of the data. Thus, facilitating the continuous analysis of the system's chances to transition to each future state.

Four graphs were assembled following this model, two of them considering the general transition matrix of temperature and oxygen, already presented in Figure 10 and Figure 13, respectively. The results of these two graphs are presented in Figure 19. It is possible to observe that the matrix is able to provide reasonable information on the future behavior of the reservoir, and in most of the time steps the state of highest probability of occurrence was also the one that occurred.

The other two graphs were made using the same principle, but with a censored matrix for each time step. The censorship was done as follows:

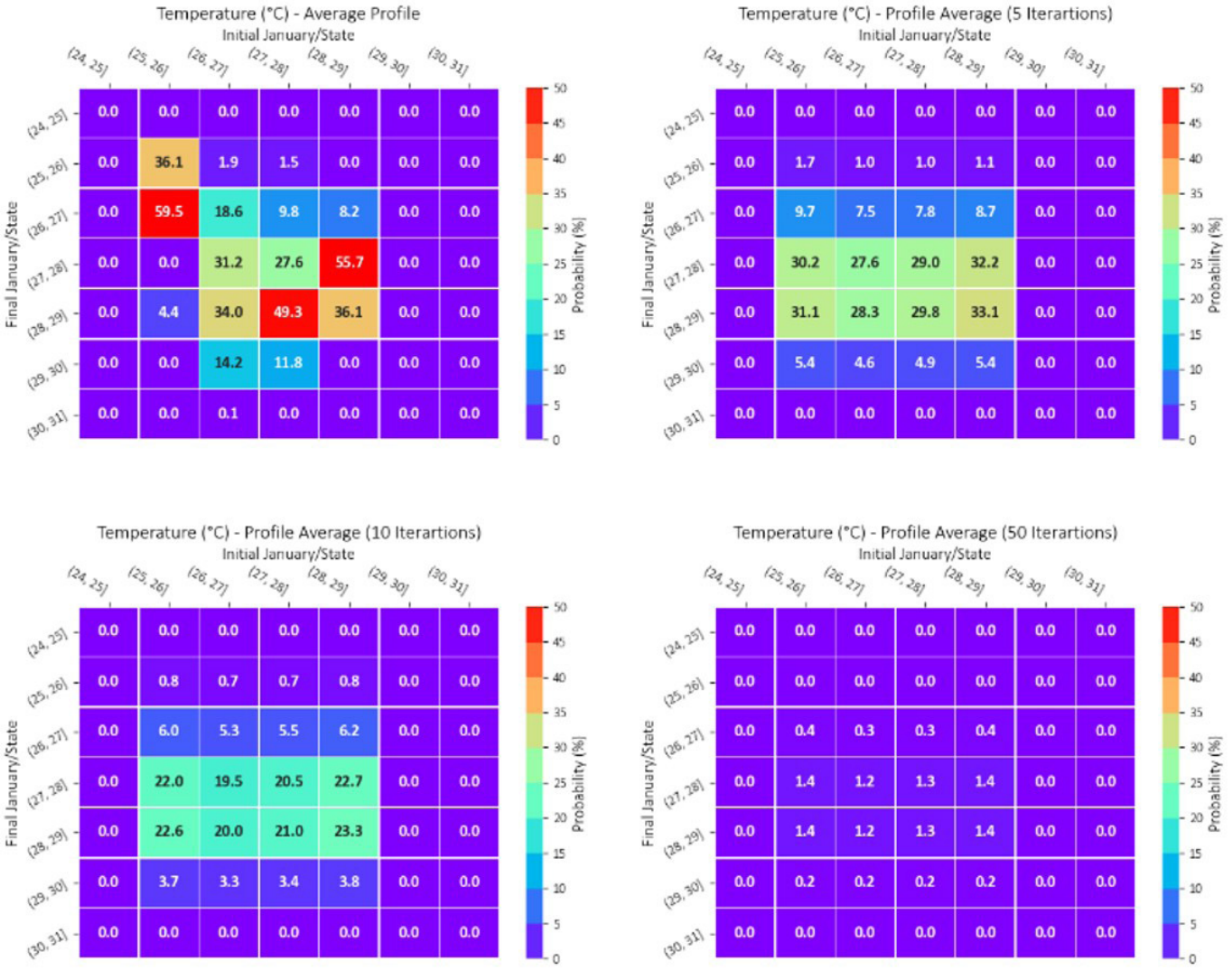

Figure 17. Censored transition matrices powered n equals 1, 5, 10 and 50 - water temperature for the profile average.

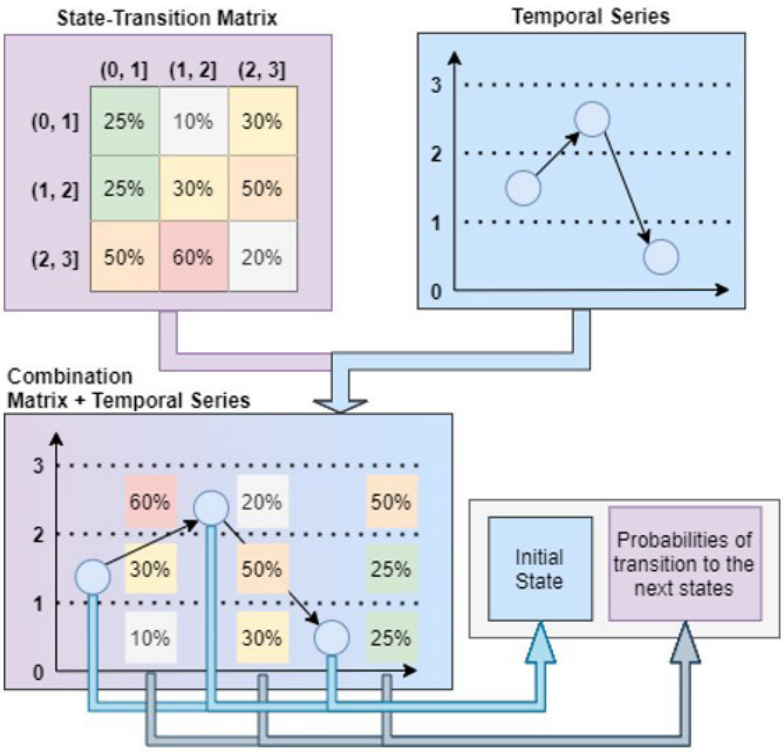

Figure 18. Example of combined interpretation of the transition matrix and observed data.

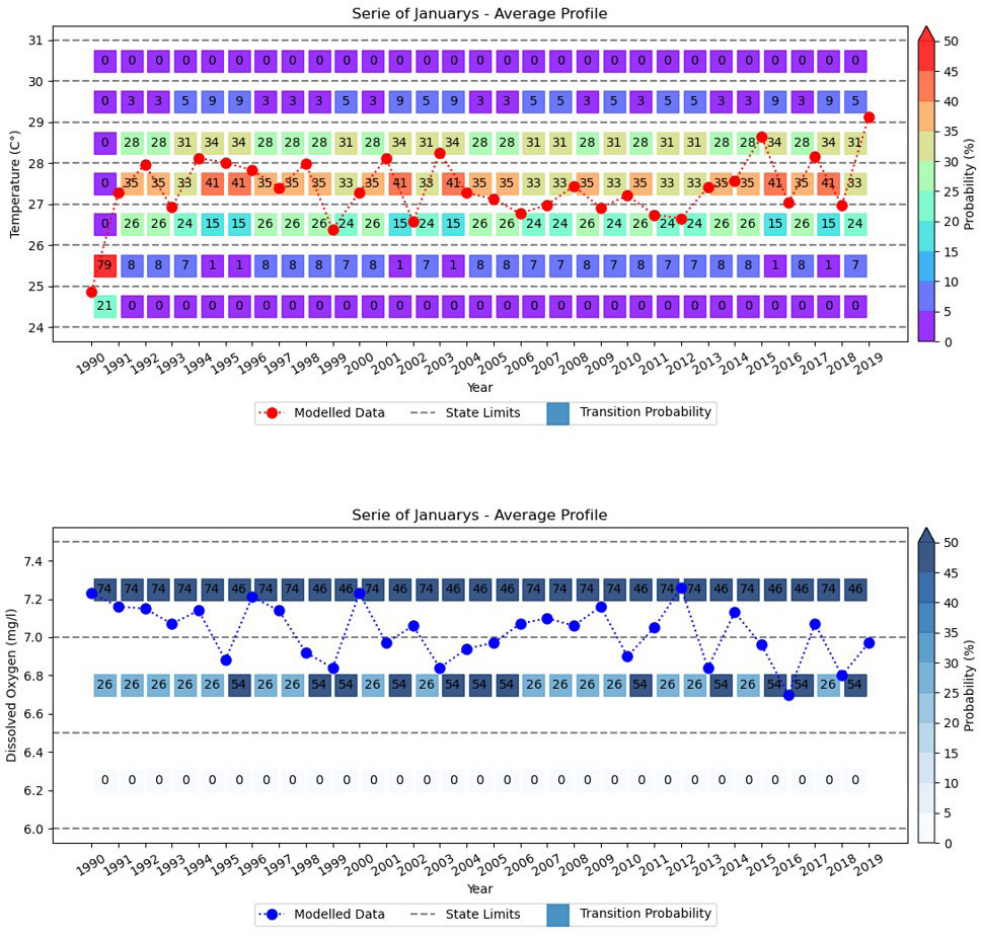

Figure 19. Temporal series merged with the transition matrix information – depth averaged profile of water temperature and dissolved oxygen.

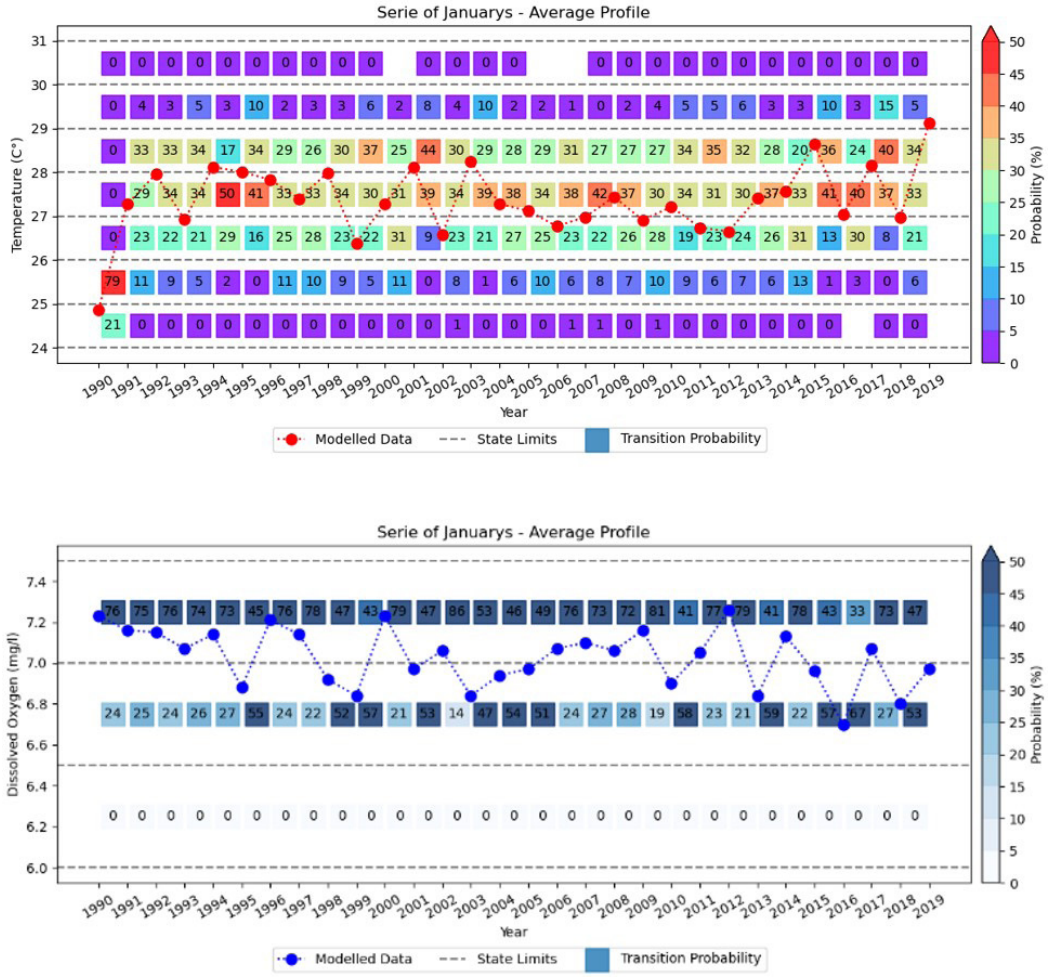

Figure 20. Temporal series merged with censored transition matrix information – depth averaged profile of water temperature and dissolved oxygen.

I. Separation of the averages of the input time series from the previous year;
II. Identification of the next interval below and above those averages in the PDF's;
III. Selection of all scenarios within this input range;
IV. Generation of the matrix with the selected scenarios;
V. Repetition of the process until the penultimate sample.

The idea behind this is to improve the understanding of the system, using as a hypothesis that the next events will not be more or less similar from those that occurred in the previous year. This way, information from 1989 is used to better understand the transition process between 1990 and 1991, and going on until 2019. The complete results of this approach are also shown in Figure 19 and Figure 20.

The examples applying the censored matrix obtained similar results to those with the uncensored matrix, showing only punctual improvements. These results were the opposite of what was expected from the censored matrix, since more specific data tend to improve the results obtained. This may have occurred due to an insufficient number of scenarios selected in the censorship process, which may result in an unrepresentative matrix. Since this single example is not the main theme of the article, but rather a complement to how the transition matrix can be applied and shaped, more details about it were not analyzed yet. However, there remains the opportunity for future research that may come to explore the best type of censorship to be applied to obtain more accurate predictions.

CONCLUSIONS

The method presented here is a viable alternative for complex mathematical representations of a given phenomenon, and increases the possibilities of using state-transition matrices in different systems. However, it has some limitations, such as the need to create and execute a large number of scenarios, which must be well defined to avoid loss of information with respect to the modeled phenomena.

The probabilities presented in the matrices may appear to be inaccurate, however this occurs mostly because this method tends to present the uncertainties inherent to natural processes more explicitly.

It is worth remembering that the method uses events that have already occurred to estimate the behavior of a system;

therefore, its use as a forecasting and management tool is only valid for stationary systems. Otherwise, the methodology must be adapted to correct and consider non-stationary processes.

Despite its limitations, this type of approach is advantageous and provides an interesting tool for reservoir operation and forecasting, as the matrix has a simple and intuitive visual interpretation, which carries a large amount of information about a system and its risks.

REFERENCES

Agência Nacional de Águas – ANA. (2020a). *HidroWeb. Sistema de informações Hidrológicas*. Brasília: ANA. Retrieved in 2020, June 15, from https://www.snirh.gov.br/hidroweb/apresentacao

Agência Nacional de Águas – ANA. (2020b). *SAR – Sistema de Acompanhamento de Reservatórios*. Brasília: ANA. Retrieved in 2020, June 15, from: https://www.ana.gov.br/sar

Arefinia, A., Bozorg-Haddad, O., Oliazadeh, A., & Loáiciga, H. A. (2020). Reservoir water quality simulation with data mining models. *Environmental Monitoring and Assessment*, *192*(7), 1-13.

Barzegar, R., Asghari Moghaddam, A., Adamowski, J., & Ozga-Zielinski, B. (2018). Multi-step water quality forecasting using a boosting ensemble multi-wavelet extreme learning machine model. *Stochastic Environmental Research and Risk Assessment*, *32*(3), 799-813.

Bruggeman, J., & Bolding, K. (2014). A general framework for aquatic biogeochemical models. *Environmental Modelling & Software*, *61*, 249-265.

Callaham, J. L., Maeda, K., & Brunton, S. L. (2019). Robust flow reconstruction from limited measurements via sparse representation. *Physical Review Fluids*, *4*(10), 103907.

Carvalho, J. M., & Bleninger, T. (2021). *Matrizes de transição como ferramenta de análise e previsão da qualidade da água em reservatórios - estudo de caso: Reservatório da UHE Jurumirim* (Dissertação de mestrado). Universidade Federal do Paraná, Curitiba.

Chapin, T. P. (2015). High-frequency, long-duration water sampling in acid mine drainage studies: a short review of current methods and recent advances in automated water samplers. *Applied Geochemistry*, *59*, 118-124.

Chen, S., Fang, G., Huang, X., & Zhang, Y. (2018). Water quality prediction model of a water diversion project based on the improved artificial bee colony–backpropagation neural network. *Water (Basel)*, *10*(6), 806.

China Three Gorges Brasil. (2020). Retrieved in 2020, August 15, from https://www.ctgbr.com.br/

Companhia Ambiental do Estado de São Paulo – CETESB. (2020). *Programa de monitoramento de água interiores*. Retrieved in 2020, June 15, from https://cetesb.sp.gov.br/aguas-interiores/programa-de-monitoramento

Damania, R., Desbureaux, S., Rodella, A. S., & Russ, J. (2019). *Quality unknown: the invisible water crisis*. Washington: World Bank Publications.

Elkiran, G., Nourani, V., & Abba, S. I. (2019). Multi-step ahead modelling of river water quality parameters using ensemble artificial intelligence-based approach. *Journal of Hydrology (Amsterdam)*, *577*, 123962.

Erichson, N. B., Mathelin, L., Yao, Z., Brunton, S. L., Mahoney, M. W., & Kutz, J. N. (2020). Shallow neural networks for fluid flow reconstruction with limited sensors. *Proceedings of the Royal Society of London. Series A*, *476*(2238), 20200097.

Ferreira, D. M., Fernandes, C. V. S., Kaviski, E., & Fontane, D. (2019). Water quality modelling under unsteady state analysis: strategies for planning and management. *Journal of Environmental Management*, *239*, 150-158.

Gomide, F. L. S. (1975). *Range and deficit analysis using Markov chains* (Doctoral dissertation). Fort Collins: Colorado State University.

Hipsey, M. R., Bruce, L. C., Boon, C., Busch, B., Carey, C. C., Hamilton, D. P., Hanson, P. C., Read, J. S., de Sousa, E., Weber, M., & Winslow, L. A. (2019). A General Lake Model (GLM 3.0) for linking with high-frequency sensor data from the Global Lake Ecological Observatory Network (GLEON). *Geoscientific Model Development*, *12*(1), 473-523.

Jia, X., Willard, J., Karpatne, A., Read, J., Zwart, J., Steinbach, M., & Kumar, V. (2019, May). Physics guided RNNs for modeling dynamical systems: a case study in simulating lake temperature profiles. In *Proceedings of the 2019 SIAM International Conference on Data Mining* (pp. 558-566). Filadélfia: Society for Industrial and Applied Mathematics.

Kim, K. B., Jung, M. K., Tsang, Y. F., & Kwon, H. H. (2020). Stochastic modeling of chlorophyll-a for probabilistic assessment and monitoring of algae blooms in the Lower Nakdong River, South Korea. *Journal of Hazardous Materials*, *400*, 123066.

Kozak, C. (2016). *Water quality assessment and its effects on diffuse pollution considering a new water quality and quantity approach* (PhD thesis). Universidade Federal do Paraná, Curitiba. Retrieved in 2020, June 15, from https://acervodigital.ufpr.br/handle/1884/43629

Latouche, G., & Ramaswami, V. (1999). *Introduction to matrix analytic methods in stochastic modeling*. Filadélfia: Society for Industrial and Applied Mathematics. http://dx.doi.org/10.1137/1.9780898719734.

Moran, P. A. P. (1954). A probability theory of dams and storage systems. *Australian Journal of Applied Science*, *5*, 116-124.

Najah Ahmed, A., Binti Othman, F., Abdulmohsin Afan, H., Khaleel Ibrahim, R., Ming Fai, C., Shabbir Hossain, M., Ehteram, M., & Elshafie, A. (2019). Machine learning methods for better water quality prediction. *Journal of Hydrology (Amsterdam)*, *578*, 124084.

Read, J. S., Hamilton, D. P., Jones, I. D., Muraoka, K., Winslow, L. A., Kroiss, R., Wu, C. H., & Gaiser, E. (2011). Derivation of lake mixing and stratification indices from high-resolution lake buoy data. *Environmental Modelling & Software*, *26*(11), 1325-1336.

Saha, S., Moorthi, S., Pan, H., Wu, X., Wang, J., Nadiga, S., & Goldberg, M. (2010). *NCEP Climate Forecast System Reanalysis (CFSR) Selected Hourly Time-Series Products, January 1979 to December 2010*. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. https://doi.org/10.5065/D6513W89.

Saha, S., Moorthi, S., Pan, H., Wu, X., Wang, J., Nadiga, S., & Goldberg, M. (2011). *NCEP Climate Forecast System Version 2 (CFSv2) 6-hourly Products*. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. https://doi.org/10.5065/D61C1TXF.

Thompson, K. R., Dowd, M., Shen, Y., & Greenberg, D. A. (2002). Probabilistic characterization of tidal mixing in a coastal embayment: a Markov Chain approach. *Continental Shelf Research*, *22*(11-13), 1603-1614.

Tiyasha, Tung, T. M., & Yaseen, Z. M. (2020). A survey on river water quality modelling using artificial intelligence models: 2000-2020. *Journal of Hydrology (Amsterdam)*, *585*, 124670.

Zhang, R., Tang, C., Ma, S., Yuan, H., Gao, L., & Fan, W. (2011). Using Markov chains to analyze changes in wetland trends in arid Yinchuan Plain, China. *Mathematical and Computer Modelling*, *54*(3-4), 924-930.

Authors contributions

João Marcos Carvalho: Responsible by the development and application of the methodologies implemented.

Tobias Bleninger: Reviewer, consultant and supervisor of work in general.

Editor in-Chief: Adilson Pinheiro

Associated Editor: Iran Eduardo Lima Neto