hydrology

MDPI

Article

A Data-Driven Approach to Close the Water Balance of a Cascaded Multiple Reservoir–Lake System

Máté Chappon [1], Katalin Bene [1] and Richard Ray [2,*]

[1] National Laboratory for Water Science and Water Security, Széchenyi István University, Department of Transport Infrastructure and Water Resources Engineering, Egyetem tér 1, H-9026 Győr, Hungary; chappon.mate@sze.hu (M.C.); benekati@sze.hu (K.B.)
[2] Széchenyi István University, Department of Structural and Geotechnical Engineering, Egyetem tér 1, H-9026 Győr, Hungary
* Correspondence: ray@sze.hu

Abstract

This study presents a data-driven methodology to develop and close the continuous water balance of a cascaded hydrological system consisting of two upstream regulating reservoirs and a downstream lake. The analysis is based on monthly hydrometeorological and hydrographic data for the period 1998–2024. A baseline water balance model using raw measured data was first implemented, followed by two bias correction approaches based on multiple linear regression coefficient estimation. The first approach applies uniform coefficients across all months, while the second accounts for seasonal variability by estimating coefficients only for the December–May period, when systematic errors are most pronounced. Model performance was evaluated using calibration (1998–2017) and validation (2018–2024) periods, with Nash–Sutcliffe efficiency (NSE) and residual diagnostics used as evaluation metrics. The baseline model showed substantial cumulative deviations due to systematic errors (NSE: −0.28 and −5.13 for the two reservoirs and −18.70 for the lake), confirming the need for bias correction. Both correction approaches significantly improved model performance, with NSE values up to 0.89 for reservoirs and 0.83–0.86 for the lake during calibration. In validation, the seasonally adjusted model performed more stably in simulating water levels for Lake Velence (NSE = 0.72) than the uniform coefficient model (NSE = 0.60), particularly under extreme hydrological conditions. Residual analysis further indicated improved independence and homoscedasticity when seasonal structure was considered. The results demonstrate that water balance closure in the system is affected by distributed errors across multiple components rather than a single dominant source. The proposed methodology provides a practical, scalable framework for reconstructing consistent water balance time series in data-limited, regulated systems and supports the development of water management scenarios.

Keywords: water balance; multiple linear regression; bias correction; Lake Velence; water management

Check for updates

Academic Editor: Minxue He

Received: 31 May 2026
Revised: 9 July 2026
Accepted: 11 July 2026
Published: 16 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Lakes and reservoirs constitute essential freshwater resources, supporting municipal, industrial, agricultural, and ecological functions [1]. They represent a major component of the global hydrological system, storing a significant portion of the Earth's accessible surface freshwater and providing critical ecosystem services [2]. Due to their multifunctional

role and sensitivity to both climatic variability and human intervention, their sustainable management is increasingly important [3].

A key requirement for effective water resources management is the ability to understand past water level dynamics and reliably predict future water availability [4,5]. The water balance equation provides a fundamental framework for this purpose by linking changes in storage to inflows, outflows, and an aggregated error term [6]. In theory, the long-term application of the water balance requires that the error term has a zero mean, implying that only random measurement errors are present and that they cancel out over time [7]. In practice, however, water balance components are affected by both random and systematic errors, resulting in persistent imbalances that accumulate over time and ultimately compromise the reliability of water balance models [8].

Addressing these imbalances is therefore a central challenge in hydrology [9]. Researchers have proposed various approaches, including integrating multiple data sources [10], applying physically based or data assimilation methods [11], and redistributing residual errors among water balance components [12]. While these methods can improve consistency, they are often computationally demanding, data-intensive, or difficult to apply in operational water management contexts [13,14].

In parallel, data-driven approaches, such as multiple linear regression (MLR), have been widely used in hydrology due to their simplicity, transparency, and ability to identify relationships between variables [15–17]. However, their application has primarily focused on modeling hydrological responses, such as rainfall–runoff relationships, rather than directly addressing the closure of water balance equations. Additionally, most studies focus on individual lakes or reservoirs, whereas few examine cascaded systems with multiple connected reservoirs and lakes. In such systems, errors may propagate both temporally and spatially, further complicating the achievement of a consistent system-wide water balance [18].

Lake Velence and its upstream regulating reservoirs represent a typical example of such a cascaded system [19]. Reservoir operations, regulated inflows, and anthropogenic pressures exert a strong influence on the lake, while existing water balance calculations exhibit significant systematic errors and rely partly on retrospective adjustments [20]. These limitations restrict the direct use of observed and calculated time series for long-term modeling and scenario analysis.

Therefore, there is a need for a practical, operationally applicable methodology that enables consistent reconstruction and closure of the water balance in interconnected and regulated systems while remaining applicable in data-limited operational contexts. This study addresses this need by proposing a data-driven bias correction framework based on multiple linear regression, applied directly to the water balance components of a cascaded reservoir–lake system. In contrast to previous applications of regression methods, which typically focus on hydrological response modeling, the proposed approach aims to close the system's continuous monthly water balance. Two different bias correction model formulations are evaluated, including one that explicitly accounts for seasonal patterns in water balance errors, thereby addressing the temporal structure of systematic biases.

The objective of this study is to develop a data-driven methodology for constructing and closing the continuous water balance of a cascaded reservoir–lake system using long-term monthly data. The study first constructs a baseline water balance model using raw measured data, followed by two bias correction approaches based on multiple linear regression coefficient estimation. The performance of these models is evaluated through calibration and validation, with particular attention to the role of seasonal error patterns. The resulting closed water balance provides a consistent basis for subsequent water management scenario analysis.

2. Literature Review

2.1. Lakes, Reservoirs, and Water Resources Management

Lakes and reservoirs are fundamental components of the hydrological system and play a critical role in water resources management. Lakes represent a major storage of freshwater, as they store the largest liquid water mass on the terrestrial surface [2,21]. At the same time, reservoirs actively regulate hydrological processes by redistributing river flow in space and time [5,22]. Given their importance, water managers must understand the availability and sustainability of these systems, for which water balance modeling provides a fundamental analytical framework [6].

Fluctuations in lake water levels are particularly informative indicators of both climatic variability and anthropogenic impacts [23]. Identifying the drivers behind these fluctuations is therefore crucial for effective lake management [24]. Consequently, water balance modeling has become a widely used and essential tool for analyzing lake dynamics and supporting water management decisions [18]. Its relevance is further demonstrated across a range of applications, including agricultural water management [25], flood and drought risk mitigation through reservoir regulation [5], climate change impact assessment and scenario analysis [26,27], and the analysis of lake level dynamics and their driving mechanisms [18,24].

2.2. Water Balance Calculation, Uncertainty, and Error Sources

The water balance provides the conceptual foundation for hydrological analysis; however, its practical application involves significant uncertainties [7]. Although the concept is straightforward, water budgets are often difficult to determine accurately due to limited data availability and measurement precision [6]. A persistent issue in the literature reveals that errors in individual water balance components are often not analyzed in sufficient detail, despite their substantial influence on results [7]. Studies of large lake systems, such as Lake Victoria [23], highlight that the primary objective of water balance modeling is to explain major water level fluctuations. However, due to limited observational data, these estimates often disclose considerable uncertainty. Similar challenges arise in reservoir systems, where water balance estimation may become ill-posed when key fluxes are not directly measured [28].

Further complications arise from sparse monitoring networks, particularly in large or remote basins, which limit the reliability of precipitation and other input data [8]. In addition, modeling errors can propagate over time due to the inherent inertia of large water bodies, thereby amplifying inaccuracies in long-term simulations [24]. These factors demonstrate that uncertainties in water balance modeling are both structural and data-driven and must be explicitly addressed to ensure reliable interpretation.

2.3. Methods for Closing the Water Balance

Given their inherent uncertainties, various methods have been developed to achieve closure in water balance calculations and improve model reliability. One common approach introduces additional constraints, thereby reducing errors in individual water balance components and improving overall system consistency [9]. More advanced techniques focus on analyzing and redistributing residual errors, such as precipitation decorrelation and residual redistribution methods, which aim to identify and correct systematic biases in the water balance [12]. Calibration-based approaches are also widely applied, in which model parameters are adjusted to achieve the best agreement between observed and simulated water level changes [16].

In addition, probabilistic frameworks can explicitly represent uncertainties and biases in input data, thereby enabling a more consistent reconciliation of multiple data sources [29].

When data availability is limited, structural approaches that constrain hydrological models using available observations can help address ill-posed water balance problems [28].

Data assimilation techniques aim to achieve full water balance closure by integrating ground-based measurements with Earth observation data. They ensure consistency with governing equations and correct biases in individual components, for example, through Bayesian hierarchical modeling [11]. Overall, these approaches demonstrate that achieving water balance closure typically requires a combination of statistical, physical, and data-driven methods.

2.4. Application of Multiple Linear Regression in Closing the Water Balance

A simpler and widely used data-driven approach for bias correction and supporting water balance closure is multiple linear regression (MLR). Regression-based methods enable the identification and correction of systematic biases in water balance components by relating them to observable variables. The main advantage of MLR lies in its simplicity and transparency, making it well suited for identifying relationships between model residuals and explanatory variables [16]. Such approaches can enhance water balance modeling by providing additional insight into system behavior and enabling data-driven correction of errors [16]. Regression analysis has examined key water balance components to support water resource management decisions across various lake systems [16,30,31].

The application of MLR is particularly relevant in cases where input data require adjustment, as demonstrated in studies that bias-correct climate model outputs before their use in water balance simulations [23]. However, the use of regression methods requires careful model specification and validation, including assessing parameter significance and verifying model robustness [32,33].

2.5. Research Gap and Novelty

To address limited observations and measurement errors, previous studies have proposed a range of approaches, including data assimilation techniques, probabilistic frameworks, and residual redistribution methods, to improve consistency among water balance components [9,11,12,29]. While these methods have proven effective in reducing uncertainty, they are often computationally complex, data-intensive, or difficult to implement in operational water management contexts.

At the same time, regression-based methods, particularly MLR, have been widely applied in hydrology due to their simplicity, transparency, and ability to capture relationships between hydrological variables [16,30]. However, their application has largely focused on modeling hydrological responses (e.g., rainfall–runoff relationships) rather than directly addressing water balance closure problems. Only limited examples exist where regression approaches support water balance consistency or correct individual components [31], and even in these cases, their role remains secondary rather than central.

Another limitation in the literature is the lack of studies on cascaded water systems with multiple connected reservoirs and lakes. Most existing work focuses on single lakes or reservoirs, while interconnected systems introduce additional complexity due to error propagation and dependency between upstream and downstream components. As demonstrated in large lake systems, inaccuracies can propagate over time due to system inertia [24]. However, this issue is even more critical in cascaded systems, where errors transfer spatially as well.

Studies that explicitly consider regulated systems highlight the importance of operational rules and human interventions. For instance, in the Lake Victoria system, dam management scenarios have been shown to exert a stronger influence on water levels than climate forcing alone [26].

Therefore, this study addresses these gaps by proposing a data-driven bias correction framework based on MLR, applied directly to the water balance components of a cascaded reservoir–lake system. The study evaluates two different MLR-based model formulations: the first models the system continuously throughout the year. In contrast, the second explicitly accounts for seasonal patterns observed in water balance errors.

3. Site Description and Data

3.1. Study Area

Lake Velence is a shallow lake located in Central Hungary. The total catchment area (including the lake) is 602.3 km^2, covering the south-eastern slope of the Vértes Mountains, the northern part of Mezőföld, and the Velence Mountains. Upper Carboniferous granite forms most of the Velence Mountains, while Upper Eocene andesite volcanics also occur in the north-northeastern part of the mountain range. Mesozoic carbonates are present in the north-northwestern part of the watershed, whereas Pannonian and Quaternary formations dominate in the central and southern parts [34].

Hydrologically, the catchment consists of three main sub-basins. The largest (383 km^2) is the Császár-víz watershed, whose upper ~75 km^2 karstic area is partially inactive. The second main tributary is the Vereb–Pázmándi watercourse (105 km^2), while the remaining direct catchment of the lake covers 114.3 km^2. Among the inflows of the lake, only the Császár-víz is a permanent stream.

The hydrological structure of the system is further characterized by two upstream reservoirs along the Császár-víz, forming a cascaded reservoir–lake system (Figure 1).

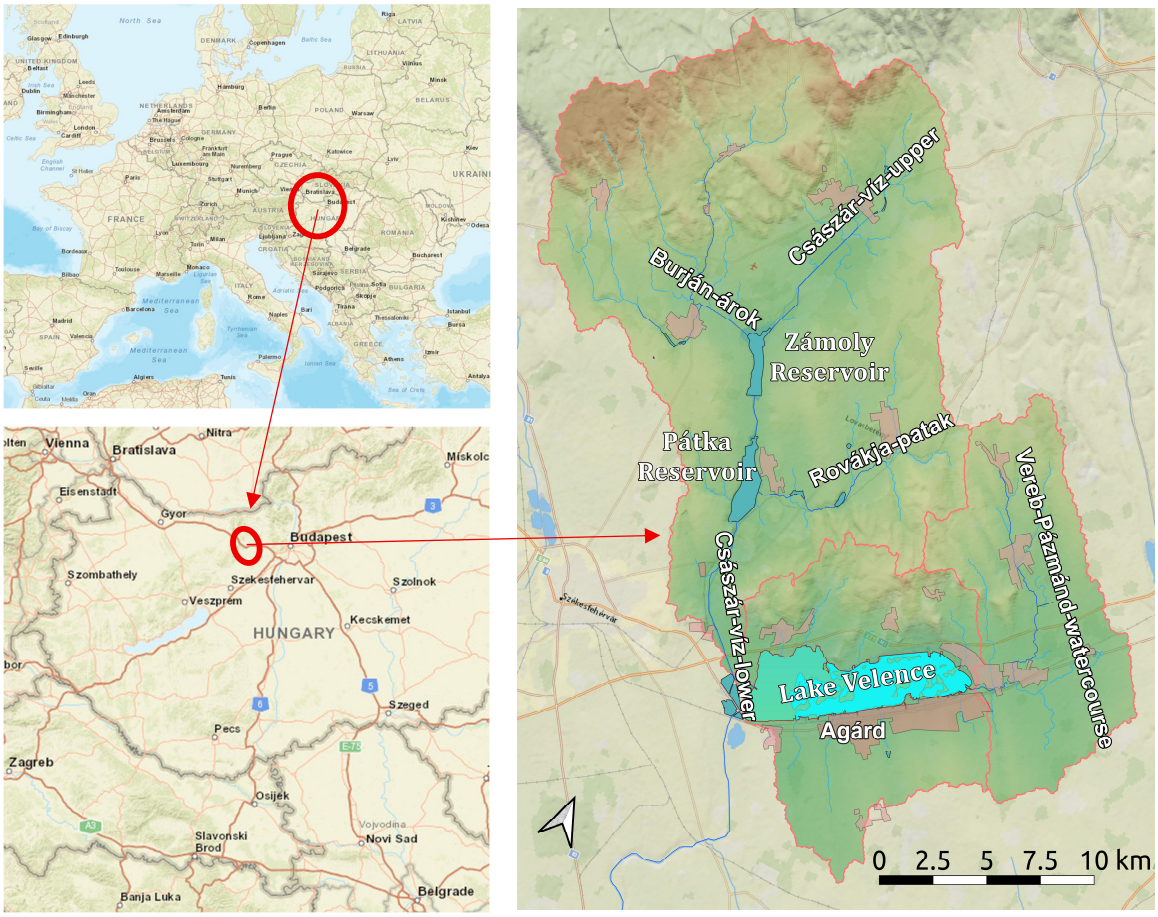


Figure 1. Lake Velence catchment area with the two large reservoirs on the Császár-víz: Zámoly Reservoir (upstream) and Pátka Reservoir (downstream).

3.2. Water Management and System Characteristics

The lake has a surface area of 24.2 km^2 at a water level of 160 cm on the Agárd gauge. It is approximately 15 km long, 2–3 km wide, and has an average depth of 1.9 m.

The lake exhibits strong spatial and functional heterogeneity. Reed beds largely cover the western part and serve as a protected ecological area, whereas the eastern part consists of open water and provides an intensively used recreational area. As a result, two upstream reservoirs and a controlled outflow tightly regulate water levels in Lake Velence. (See the water level time series in Figure 2.)

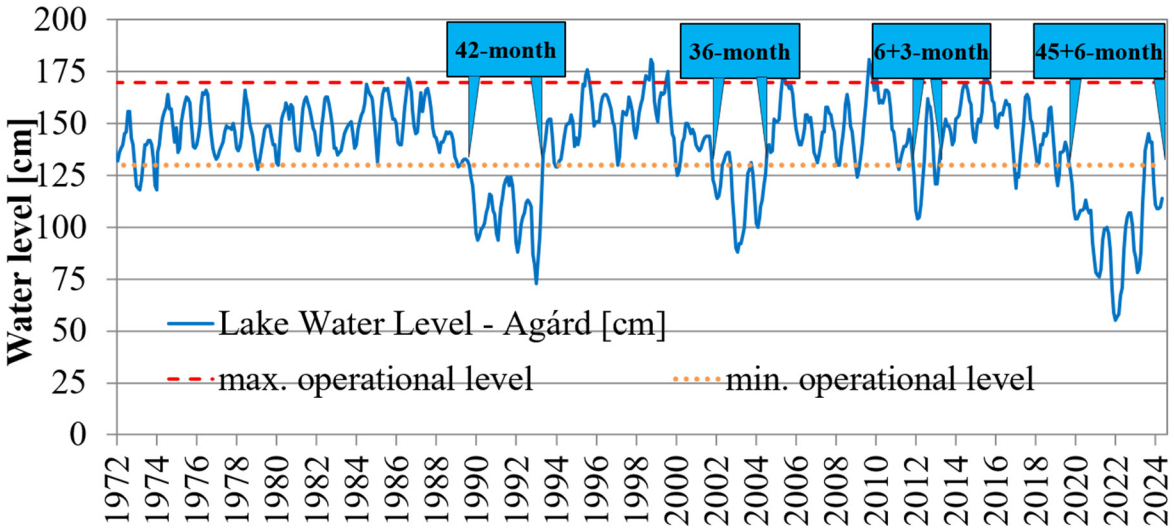


Figure 2. Lake Velence water level time series from 1972 to 2024. Blue boxes highlight periods with water levels below the minimum operational level.

The two upstream reservoirs (Zámoly and Pátka), constructed in the 1970s, were originally designed to support water replenishment during dry periods by storing water during wetter seasons (November–April). Over time, their role has expanded beyond water regulation to include fisheries and recreational uses, which have contributed to increased nutrient loading. Combined with external inputs from agriculture and wastewater, this has led to eutrophication problems, occasionally limiting the usability of stored water for lake replenishment. In certain years, the reservoirs' interception of streamflow has reduced direct runoff to the lake, effectively disconnecting a significant portion of the catchment.

Recent hydrological extremes have further highlighted the system's vulnerability. Following a prolonged drought from 2020 to 2022, lake water levels reached a historical low of 53 cm in September 2022. This event triggered increased management pressure and a range of interventions, including partial wetland drainage and modifications to reservoir operation.

The Zámoly reservoir was emptied in 2021, while water from the Pátka reservoir was diverted towards the lake during the wet winter of 2023–2024. As of 2026, both reservoirs operate primarily as empty storage areas with limited and temporary active retention.

3.3. Data Sources

The Central Transdanubian Water Directorate (KDTVIZIG) provided the hydrometeorological and hydrographic data shown in Table 1 and Figure 3. The dataset covers the period 1998–2024 and includes water levels, bathymetric relationships, streamflow, precipitation, and evaporation. The dataset integrates measurements from multiple components of the system, including the two reservoirs, the lake, and their tributaries and outflows.

Bathymetric curves complement water level observations by describing the relationships among water level, surface area, and storage volume for each water body.

Table 1. Hydrometeorological data in the catchment of Lake Velence (source: KDTVIZIG).

Location	Symbols	Parameter	Frequency	Timeframe
Water level and bathymetric curves				
Zámolyi reservoir	H_{ZR} (Obs) $A_{ZR} = f(H_{ZR})$ $V_{ZR} = f(H_{ZR})$	H = Water level (cm) Observed	Monthly	1998–2024
Pátkai reservoir	H_{PR} (Obs) $A_{PR} = f(H_{PR})$ $V_{PR} = f(H_{PR})$	A = Water surface (m^2) as a function of H		
Lake Velence	H_{LV} (Obs) $A_{LV} = f(H_{LV})$ $V_{LV} = f(H_{LV})$	V = Water volume (m^3) as a function of H		
Streamflow				
Császár-víz (upper)—Csákvár	Q_{Cs}	Q = Discharge (m^3/s)	Daily, Monthly	1998–2024
Burján-árok—Zámoly	$Q_{Bá}$			
Rovákja-patak—Pátka	Q_{Rp}			
Császár-víz—Kisfalud	$Q_{in\text{-}LV\text{-}1}$	$Q_{in\text{-}LV} = Q_{in\text{-}LV\text{-}1} + Q_{in\text{-}LV\text{-}2}$		
Vereb-Pázmánd watercourse—Kápolnásnyék	$Q_{in\text{-}LV\text{-}2}$			
Zámolyi Reservoir outflow	$Q_{out\text{-}ZR}$		—	
Pátkai Reservoir outflow	$Q_{out\text{-}PR}$		Monthly	
DIT (nursery pond) water abstraction	Q_{Use}			
Lake Velence outflow structure—Dinnyés	$Q_{out\text{-}LV}$			
Precipitation				
Zámoly	P_Z	P = Rainfall (mm)	Monthly	1998–2024
Pátka	P_P			
Dinnyés	P_D	Average rainfall: $P_{LV} = \frac{P_D+P_A+P_V}{3}$ (mm)	Daily, Monthly	
Agárd	P_A			
Velence	P_V			
Agárd	E_{LV}	Actual lake evaporation (mm)—calculated	Monthly	1998–2024

Streamflow data include measured inflows, reservoir releases, and outflows from the lake system. Multiple stations within the catchment record precipitation, and lake-wide precipitation is calculated as a spatial average. Evaporation data are derived from measurements at the Agárd station. All variables were processed to ensure consistency in subsequent modeling. Streamflow values (m^3/s) were converted to monthly volumes, while precipitation and evaporation (mm) were transformed into volumetric units using the corresponding water surface areas.

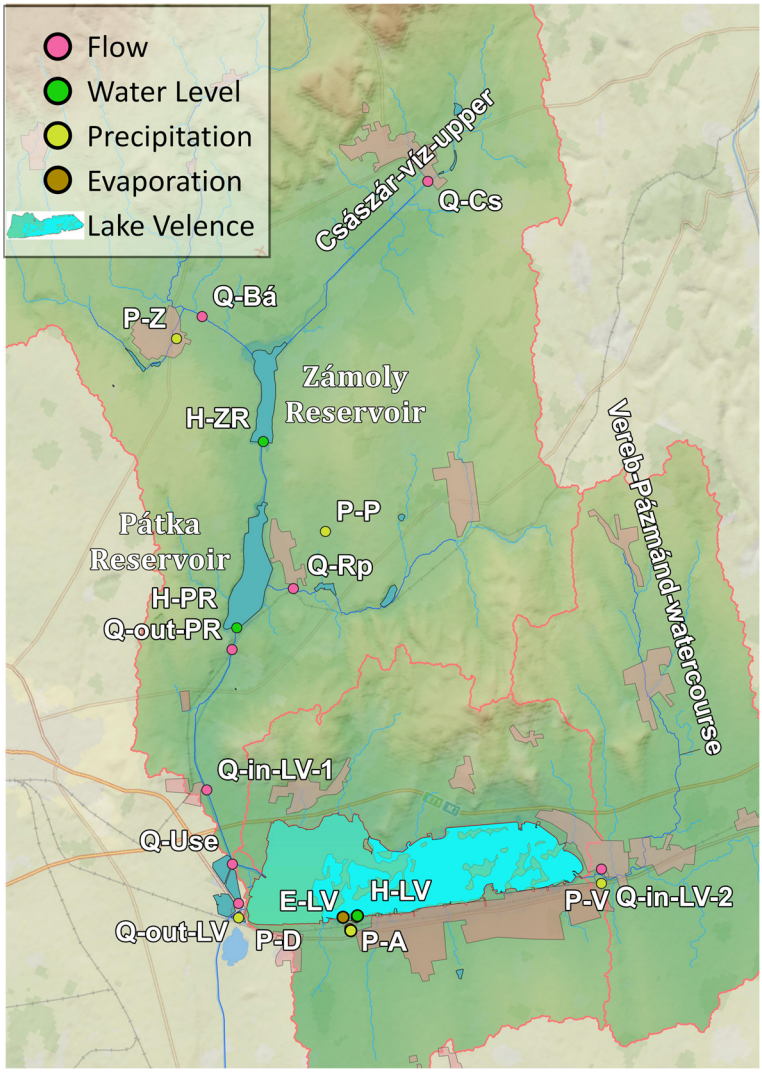


Figure 3. Overview map of the Lake Velence catchment area. Hydrometeorological stations corresponding to entries in Table 1 are shown.

4. Water Balance Calculations

4.1. Current Water Balance Calculation Method

The Water Directorate has calculated the monthly water balance of Lake Velence since 1986 using the following equation:

$$P_{LV} + Q_{in\text{-}LV} + Q_{out\text{-}PR} + Z_{month} = E_{LV} + Q_{out\text{-}LV} + Q_{Use} \pm \Delta H_{LV}$$

where

P_{LV} is the monthly precipitation falling on Lake Velence (lake-mm);

$Q_{in\text{-}LV}$ is the total monthly inflow from tributaries with multipliers to account for unmeasured inflow;

$Q_{out\text{-}PR}$ represents releases from the upstream reservoir system;

Z_{month} is the residual error term calculated by subtracting the observed water level from the calculated water level;

E_{LV} is lake evaporation;

$Q_{out\text{-}LV}$ is the outflow from the lake;

Q_{Use} is water abstraction;

ΔH_{LV} is the observed change in lake water level.

Figure 4 presents the contribution of water budget inflow and outflow elements to the water balance of Lake Velence between 1998 and 2024 according to the current calculation method.

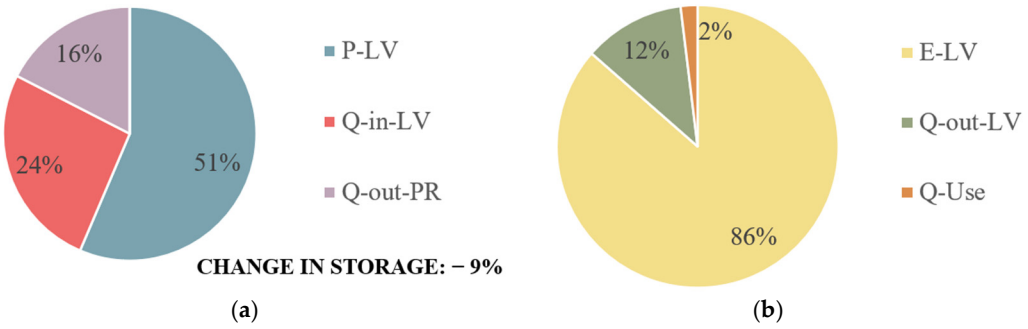


Figure 4. Contribution of Lake Velence water balance inflow and outflow components during the 1998–2024 period—resulting in a 9% reduction in storage change. All percentage values are relative to the total outflow volume (100%) during these 27 years, as calculated under the current water balance methodology. (**a**) Average contribution of inflow elements relative to total outflow; (**b**) average contribution of outflow elements relative to total outflow.

Note that reservoir releases are treated separately from natural inflows, and upstream reservoir dynamics are not explicitly integrated into the balance. Since the water balances for the two upstream reservoirs are not systematically calculated, the current method does not fully capture the mass balance of the coupled reservoir–lake system.

Monthly residual errors (Z) of the current water balance calculation result from subtracting observed water levels from the calculated water levels. In practice, residual errors are redistributed retrospectively at the end of each year among selected components (precipitation, evaporation, and inflow) based on expert judgment to achieve annual closure. While this ensures formal closure, it does not resolve underlying systematic errors.

Previous analyses have shown that, before error redistribution, the average annual water balance error was approximately −70 lake-mm/year for the 1986–2022 period [35]. A negative error indicates that the calculated lake water level is lower than the observed water level. The magnitude of this error is comparable to the average annual variation in the lake water level. Furthermore, it corresponds to approximately 13% of the mean annual precipitation. These results indicate the presence of significant systematic errors or potentially missing components in the water balance equation.

Further analysis of error characteristics revealed relationships between water balance components and residual errors [35,36]. A weak negative correlation was identified between precipitation and the error term, indicating that the water balance tends to underestimate lake water levels during wetter periods and overestimate them during drier conditions. Figure 5 illustrates the relationship between precipitation and water balance error.

This finding initially suggested systematic underestimation of precipitation measurements. However, a more detailed analysis of the temporal structure of the errors provided additional insight. Average monthly errors were close to zero during the June–November period, while consistently negative values occurred between December and May, with a peak of −18.8 lake-mm/month in April [36]. Figure 6 shows the monthly components of water balance and errors.

This pronounced seasonal pattern contradicts the assumption of purely random errors and indicates the presence of systematic, time-dependent biases or missing processes affecting the winter–spring period. The temporal structure of the errors suggests that precipitation bias alone cannot fully explain the observed discrepancies.

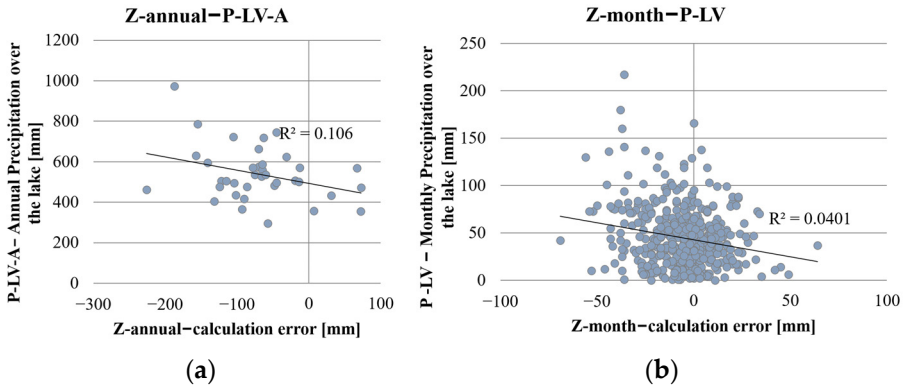


Figure 5. (**a**) Annual and (**b**) monthly error terms plotted against annual and monthly precipitation over Lake Velence.

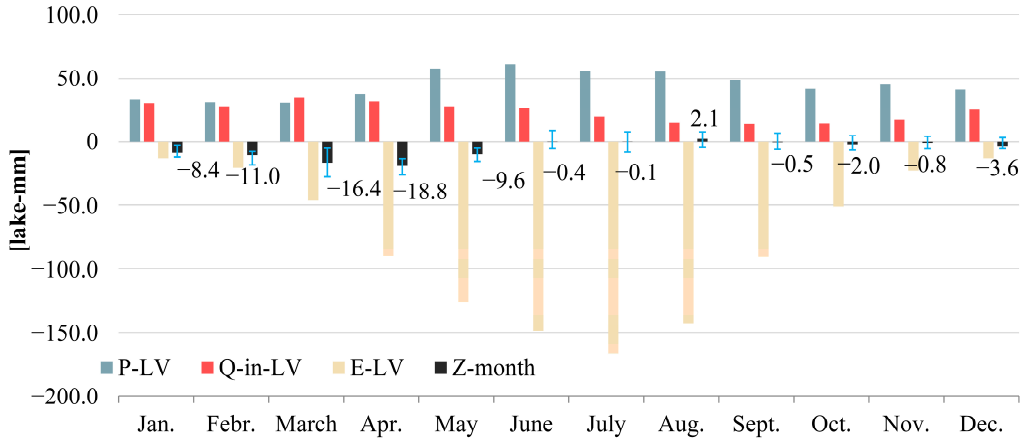


Figure 6. Average monthly precipitation P-LV, surface inflow Q-in-LV, evaporation E-LV, and residual error Z-month in the water balance calculation according to the current methodology.

As an alternative explanation, recent studies have investigated the role of groundwater exchange. Baják et al. [37] applied a three-dimensional transient groundwater flow model to estimate subsurface inflows and outflows, suggesting a net groundwater inflow of approximately 47 lake-mm/year. Incorporating this component would substantially reduce the observed annual deficit, although the temporal distribution has not yet been compared to residual errors. The current water balance calculation method exhibits significant uncertainty, reliance on retrospective adjustments, and a lack of consistent representation of the full reservoir–lake system. Therefore, the raw time series of measured and calculated components are not directly suitable for continuous modeling or for evaluating water management scenarios.

These limitations motivated the development of a revised methodology capable of closing the joint water balance of the cascaded system.

4.2. Methodological Framework

The proposed methodology aims to calculate and close the joint, continuous water balance of a cascaded system comprising two upstream reservoirs and a downstream lake at a monthly time step over multiple decades.

The workflow consists of five sequential stages:

1. Baseline modeling (Model-A): Construction of a continuous water balance using raw measured data without bias correction.
2. Bias correction (Model-B): Application of a linear correction framework with uniform coefficients to reduce systematic errors.

3. Seasonal correction (Model-C): Extension of Model-B by introducing seasonally varying coefficients based on observed error patterns.
4. Residual diagnostics: Evaluation of model assumptions and identification of remaining biases.
5. Validation and performance assessment: Comparison of model performance over an independent validation period.

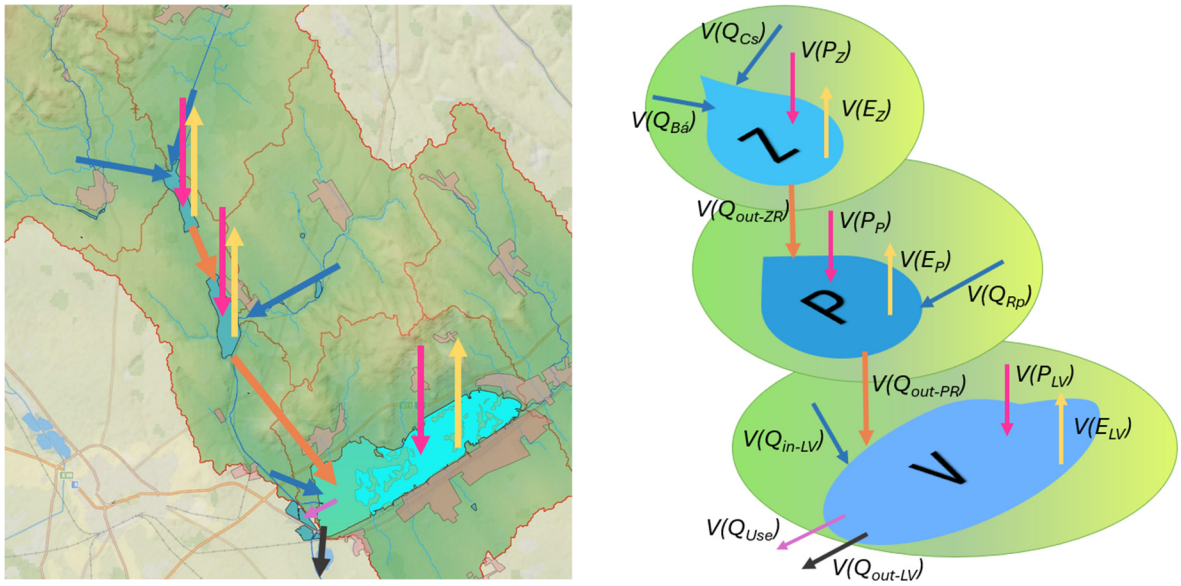
All models operate on a monthly time step, and storage is updated sequentially along the cascade.

4.3. Model-A: Baseline Water Balance Model

4.3.1. System Equations

Model-A represents the joint water balance of the cascaded reservoir–lake system using measured input data. The system represents a continuous mass balance model with a monthly time step, in which storage in each water body is updated sequentially along the cascade (from the upstream reservoir to the downstream lake) based on inflows and outflows. For each month *i*, the change in storage (ΔV) is calculated from all inflow and outflow components, and total storage is updated cumulatively from the initial observed state. Water levels at the beginning of each time step are then derived from storage using bathymetric relationships.

Figure 7 illustrates the conceptual structure of the cascade system.



(**a**) (**b**)

Figure 7. (**a**) Overview map and (**b**) conceptual model of the joint continuous water balance of the Zámoly Reservoir–Pátka Reservoir–Lake Velence cascade in the case of Model-A. Blue arrows represent inflow (Q), orange, outflow (Q_{out}), red arrows are precipitation (P), and yellow are evaporation (E).

The calculation process uses two different indices. Index *k* represents the month of interest, while index *i* represents a general month in the summation. The remaining terms in the following equations are presented in Table 1 and Figure 7b.

The calculated water level of Zámoly Reservoir at the beginning of month *k* uses its bathymetric curves:

$$H_{ZR}(C)_k = f(V_{ZR}(C)_k) \quad (1)$$

The calculated volume of Zámoly Reservoir at the beginning of month k is:

$$V_{\mathbf{Z}R}(C)_k = \max\begin{Bmatrix} V_{\mathbf{Z}R}(Obs)_0 + \sum_{i=0}^{k-1} \Delta V_{\mathbf{Z}R}(C)_i \\ 0 \end{Bmatrix} \quad (2)$$

where $V_{\mathbf{Z}R}(Obs)_0$ is the initial observed water level of the Zámoly reservoir on 1/1/1998.

The calculated change in volume for Zámoly Reservoir in month i is (see conceptual model, Figure 7b, and Table 1):

$$\Delta V_{\mathbf{Z}R}(C)_i^{Model-A} = V(P_{\mathbf{Z},i}) + V(Q_{Bá,i}) + V(Q_{Cs,i}) - V\left(E_{\mathbf{Z},i}\right) - \{V(Q_{out-\mathbf{Z}R,i})\} \quad (3)$$

where V denotes volume calculated from measured discharge, precipitation, or evaporation data. The final term in {} connects the cascade between Zámoly and Pátka Reservoirs.

The calculated water level of Pátka Reservoir using its bathymetric curves, at the beginning of month k, is:

$$H_{\mathbf{P}R}(C)_k = f(V_{\mathbf{P}R}(C)_k) \quad (4)$$

The calculated volume of Pátka Reservoir at the beginning of month k is:

$$V_{\mathbf{P}R}(C)_k = \max\begin{Bmatrix} V_{\mathbf{P}R}(Obs)_0 + \sum_{i=0}^{k-1} \Delta V_{\mathbf{P}R}(C)_i \\ 0 \end{Bmatrix} \quad (5)$$

where $V_{\mathbf{P}R}(Obs)_0$ is the initial observed water level of Pátka Reservoir on 1/1/1998.

The calculated change in volume for Pátka Reservoir in the i-th month is:

$$\Delta V_{\mathbf{P}R}(C)_i^{Model-A} = V(P_{\mathbf{P},i}) + V\left(Q_{Rp,i}\right) + \{V(Q_{out-\mathbf{Z}R,i})\} - V(E_{\mathbf{P},i}) - [\![V(Q_{out-\mathbf{Z}R,i})]\!] \quad (6)$$

where V denotes volume calculated from measured discharge, precipitation, or evaporation data. The {} term from Equation (3) is now an input, and the ⟦ ⟧ term cascades to Lake Velence.

The calculated water level of Lake Velence using its bathymetric curves, at the beginning of the k-th month, is:

$$H_{\mathbf{LV}}(C)_k = f(V_{\mathbf{LV}}(C)_k) \quad (7)$$

The calculated volume of Lake Velence at the beginning of month k is:

$$V_{\mathbf{LV}}(C)_k = V_{\mathbf{LV}}(Obs)_0 + \sum_{i=0}^{k-1} \Delta V_{\mathbf{LV}}(C)_i \quad (8)$$

where $V_{\mathbf{LV}}(Obs)_0$ is the initial observed water level of Lake Velence on 1/1/1998.

The calculated change in volume for Lake Velence in month i is:

$$\begin{aligned} \Delta V_{\mathbf{LV}}(C)_i^{Model-A} = {} & V(P_{\mathbf{LV},i}) + V(Q_{in\ \mathbf{LV},i}) + [\![V(Q_{out-\mathbf{P}R,i})]\!] - V(E_{\mathbf{LV},i}) \\ & - V(Q_{out-\mathbf{LV},i}) - V(Q_{use,i}) \end{aligned} \quad (9)$$

where the ⟦ ⟧ term from Equation (6) connects Pátka Reservoir to Lake Velence.

4.3.2. System Coupling (Joint Water Balance)

The water balance model is formulated as a joint system, in which the three water bodies are hydraulically connected. Outflows from the upstream Zámoly Reservoir, $V(Q_{\text{out-ZR, i}})$, are treated as inflows to the downstream Pátka Reservoir (Equations (3) and (6)). Similarly, releases from Pátka Reservoir, $\boldsymbol{V(Q_{out-PR,\,i})}$, are considered as inflows to Lake Velence (Equations (6) and (9)). The two creeks connecting the three water bodies have

negligible volumes compared to the reservoirs and the lake. Their travel times are short when viewed on a monthly basis. This configuration ensures that mass is conserved across the entire cascade and that interactions between upstream and downstream components are explicitly represented.

4.3.3. Time Step and Unit Conversion (Monthly Resolution)

The model operates on a monthly time step, with all fluxes aggregated over each month i. Streamflow variables measured in m^3/s (e.g., $Q_{out-ZR,i}$) are converted to monthly volumes $V(Q_{out-ZR,i})$ by multiplying by the number of seconds in the respective month.

Precipitation volumes (e.g., $V(P_{Z,i})$) are calculated by multiplying monthly precipitation depth (mm) by the water surface area of the respective water body at month i. Similarly, evaporation volumes are calculated using the evaporation depth calculated from measured data at the Agárd station and the surface area of each reservoir or lake.

All variables are consistently converted to volumetric units (m^3) to ensure compatibility within the mass balance framework. In larger systems, time lags between connected water bodies may need to be considered. However, no time lag was introduced in this study, as the water travel time within the system is on the order of days, which is negligible compared to the monthly modeling time step.

4.3.4. Continuous Storage Calculation and Residuals

The model is formulated as a continuous storage system, in which water volumes are updated cumulatively over time. For each water body, total storage at the beginning of month k is calculated by summing all monthly storage changes from the initial observed volume up to month $k-1$ (Equations (2), (5) and (8)). Water levels are then derived from storage using bathymetric relationships (Equations (1), (4) and (7)). Model residuals are calculated as:

$$H_{\mathbf{ZR}}(Obs)_k = H_{\mathbf{ZR}}(C)_k - \varepsilon_{\mathbf{ZR},\,k} \rightarrow \varepsilon_{\mathbf{ZR},\,k} = H_{\mathbf{ZR}}(C)_k - H_{\mathbf{ZR}}(Obs)_k \quad (10)$$

$$H_{\mathbf{PR}}(Obs)_k = H_{\mathbf{PR}}(C)_k - \varepsilon_{\mathbf{PR},\,k} \rightarrow \varepsilon_{\mathbf{PR},\,k} = H_{\mathbf{PR}}(C)_k - H_{\mathbf{PR}}(Obs)_k \quad (11)$$

$$H_{\mathbf{LV}}(Obs)_k = H_{\mathbf{LV}}(C)_k - \varepsilon_{\mathbf{LV},\,k} \rightarrow \varepsilon_{\mathbf{LV},\,k} = H_{\mathbf{LV}}(C)_k - H_{\mathbf{LV}}(Obs)_k \quad (12)$$

Since water levels are derived from cumulative storage, these residuals represent the cumulative effect of errors in all water balance components up to the *k-th* time step.

4.3.5. Representation of Reservoir Operation

Reservoir operation is explicitly incorporated to reproduce observed emptying conditions. When, according to observations, the reservoir contains water at the end of a time step, the measured outflow is applied. When the reservoir is empty, outflow is constrained to prevent simulated storage from exceeding the physically available water. The outflow is therefore limited by the maximum of the following amounts: (1) measured outflow, (2) available inflow plus precipitation minus evaporation, and (3) initial storage at the beginning of the time step. This formulation enforces consistency between observed and simulated reservoir states while preserving mass balance.

4.4. Model-B: Linear Bias Correction Model with Uniform Coefficients

The results from the baseline water balance model (Model-A) indicate significant systematic bias when using raw measured data, preventing reliable long-term simulations. (See Section 5.) Therefore, new models that carry out bias correction are introduced. In a first attempt, Model-B applies multiple linear regression (MLR) bias correction, in which

coefficients are used to adjust water balance components. In Model-B, these coefficients are assumed to be constant over time and are applied uniformly across all months.

Equations (13)–(15) define the corrected storage change (ΔV_i) equations for Zámoly Reservoir, Pátka Reservoir, and Lake Velence, respectively. These equations replace the corresponding mass balance equations of Model-A (Equations (3), (6) and (9)) while preserving the original structure of inflow and outflow components. Figure 8 presents conceptual figures of the MLR models.

$$\Delta V_{\boldsymbol{ZR}}(C)_i^{Model-B,C} = \beta_{ZR,0} + \beta_{ZR,1}\, V(P_{\boldsymbol{Z},i}) + \beta_{ZR,2}\, V(Q_{Cs,i}) + \beta_{ZR,3}\, V(Q_{Bá,i}) - \beta_{ZR,4}\, V(E_{\boldsymbol{Z},i}) - \beta_{ZR,5}\, \{V(Q_{out-\boldsymbol{ZR},i})\} \quad (13)$$

$$\Delta V_{\boldsymbol{PR}}(C)_i^{Model-B,C} = \beta_{PR,0} + \beta_{PR,1}\, V(P_{\boldsymbol{P},i}) + \beta_{PR,2}\, V(Q_{Rp,i}) + \beta_{PR,3}\, \{V(Q_{out-\boldsymbol{ZR},i})\} - \beta_{PR,4}\, V(E_{\boldsymbol{P},i}) - \beta_{PR,5}\, [\![V(Q_{out-\boldsymbol{PR},i})]\!] \quad (14)$$

$$\Delta V_{\boldsymbol{LV}}(C)_i^{Model-B} = \beta^B_{LV,0} + \beta^B_{LV,1}\, V(P_{\boldsymbol{LV},i}) + \beta^B_{LV,2} V(Q_{in,i}) + \beta^B_{LV,3}\, [\![V(Q_{out-\boldsymbol{PR},i})]\!] - \beta^B_{LV,4} V(E_{\boldsymbol{LV},i}) - \beta^{B,C}_{LV,5}\, V(Q_{out-\boldsymbol{LV},i}) - \beta^{B,C}_{LV,6} V(Q_{use,i}) \quad (15)$$

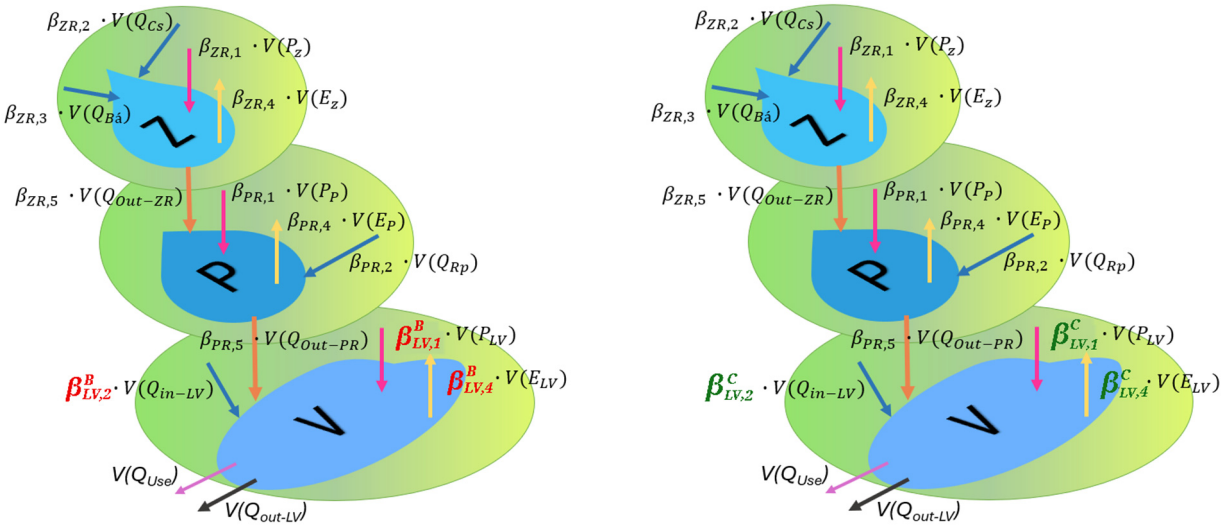


Figure 8. Conceptual figures of multiple linear regression models: (**a**) Model-B (red coefficients) and (**b**) Model-C (green coefficients). Blue arrows are inflow, orange arrows are outflow, red arrows are precipitation, and yellow arrows are evaporation.

In the equations, the subscripts of the β regression coefficients denote the associated water body (ZR: Zámoly Reservoir, PR: Pátka Reservoir, LV: Lake Velence), and balance components: 1 represents precipitation, 2 and 3 represent inflow, 4 represents evaporation, and 5 represents outflow from the respective water body.

4.4.1. Parameter Selection and Model Structure

Preliminary statistical analysis identified precipitation, inflow, and evaporation as significant predictors based on Student's *t*-tests. Furthermore, pairwise correlation analysis and Variance Inflation Factors (VIFs) were evaluated to verify that multicollinearity among the candidate predictors remained sufficiently low for reliable coefficient estimation [38]. As a result, the corresponding coefficients ($\beta_{ZR,1}$, $\beta_{ZR,2}$, $\beta_{ZR,3}$; $\beta_{ZR,4}$; $\beta_{PR,1}$, $\beta_{PR,2}$, $\beta_{PR,4}$; $\beta_{LV,1}$, $\beta_{LV,2}$, $\beta_{LV,4}$) were included in the calibration. In addition, coefficients for reservoir outflow ($\beta_{ZR,5} = \beta_{PR,3}$; and $\beta_{PR,5} = \beta^{B,C}{}_{LV,3}$) were calibrated to approximate overall mass balance across the cascade. MLR coefficients of the remaining components ($\beta^{B,C}{}_{LV,5}$, $\beta^{B,C}{}_{LV,6}$) were

fixed at 1, while the intercepts (constant terms), $\beta_{ZR,0}$, $\beta_{PR,0}$, and $\beta_{LV,0}$, were considered 0. Retaining only statistically significant predictors and verifying low multicollinearity improves parameter identifiability and reduces the likelihood of unstable coefficient estimates [38].

4.4.2. Calibration Procedure

Coefficients were estimated sequentially for each water body, starting from Zámoly Reservoir (upstream), followed by Pátka Reservoir (intermediate), and ending with Lake Velence (downstream). This sequential calibration ensured that corrected outflows from upstream components correctly propagated as inflows to downstream elements. Calibration sought to minimize the root-mean-square error (RMSE) between simulated and observed water levels. RMSE was selected because it quantifies prediction error in the physical units of the state variable (cm) and gives greater weight to larger deviations [39], which are particularly important in long-term water balance simulations where cumulative volume errors propagate through time.

Although the regression coefficients are included linearly in the water balance equations, the calibration problem itself is nonlinear. This is because simulated storage is updated recursively, so each month's prediction depends on the corrected storage changes from all previous months. Consequently, parameter estimation was performed using the Generalized Reduced Gradient (GRG) nonlinear optimization algorithm [40], implemented in Microsoft Excel [41]. To reduce the risk of convergence to local minima, a multi-start procedure was applied, with a population size of 100 and a fixed random seed value of 10 for each water body. Coefficient bounds (see Table 2) were defined to reflect physical plausibility and constrain the solution space, accounting for expected uncertainties across different water balance components.

Table 2. Boundaries of MLR coefficients used when calibrating Model-B and Model-C.

MLR Coefficient	Lower Boundary	Upper Boundary
$\beta_{ZR,1}$, $\beta_{ZR,2}$, $\beta_{ZR,3}$, $\beta_{ZR,5}$	0.60	2.00
$\beta_{ZR,4}$, $\beta_{PR,4}$, $\beta_{PR,5}$	0.70	1.50
$\beta_{PR,1}$, $\beta_{PR,2}$, $\beta_{LV,1}$, $\beta_{LV,2}$, $\beta_{LV,4}$	0.75	1.30

The imposed limits prevented the use of physically unrealistic correction factors while allowing sufficient flexibility to compensate for systematic bias. As shown in Table 2, coefficients β_{ZR},1, β_{ZR},2, β_{ZR},3, and β_{ZR},5 (related to Zámoly Reservoir) were allowed wider bounds to account for the greater uncertainty in this reservoir's water balance. This uncertainty is likely caused by unmeasured surface inflows entering Zámoly Reservoir between the Csákvár discharge gauging station and the reservoir itself (see Figure 2), resulting in an apparent imbalance between the measured inflows and the recorded reservoir releases.

4.5. Model-C: Seasonal Bias Correction Model

Previous studies, as summarized in Section 4.1, have shown that long-term average monthly water balance errors of Lake Velence exhibit a clear seasonal pattern, with significant negative errors during the December–May period and values close to zero during June–November [36]. Based on this observation, the underlying hypothesis of Model-C is that systematic errors are temporally structured and primarily affect the winter–spring period. Therefore, in Model-C, bias correction is applied selectively only during these months, while unmodified measured water balance components are used during June–November. The formulation for Lake Velence is given in Equation (16), replacing Equation (15) of Model-B.

$$\Delta V_{\boldsymbol{LV}}(C)_i^{Model-C} = \text{if} \begin{cases} (Dec.-May)\ \beta_{LV,0}^{C} + \beta_{LV,1}^{C}\, V(P_{\boldsymbol{LV},i}) + \beta_{LV,2}^{C}\, V(Q_{in,i}) + \beta_{LV,3}^{C} V(Q_{out-\boldsymbol{PR},i}) \\ \qquad -\beta_{LV,4}^{C} V(E_{\boldsymbol{LV},i}) - \beta_{LV,5}^{B,C}\, V(Q_{out-\boldsymbol{LV},i}) - \beta_{LV,6}^{B,C}\, V(Q_{Use,i}) \\ (June-Nov.)\ V(P_{\boldsymbol{LV},i}) + V(Q_{in,i}) + V(Q_{out-\boldsymbol{PR},i}) - V(E_{\boldsymbol{LV},i}) \\ \qquad -V(Q_{out-\boldsymbol{LV},i}) - V(Q_{use,i}) \end{cases} \quad (16)$$

In this formulation, MLR coefficients for Lake Velence ($\beta_{\boldsymbol{LV}}$,x) are applied only during the period with identified systematic bias, while the original mass balance formulation of Model-A is retained during months with negligible average error. The subscripts of the regression coefficients are associated with the same water balance components as in Model-B.

A review of the data comparing release volumes from the reservoirs to errors concerning their volume estimates found no discernible pattern, seasonal or otherwise. Figure 9 shows outflow volumes and Model-A residuals for the Zámoly and Pátka reservoirs from 1998 to 2017. During this time, zero outflow occurred in 189 of the 240 months for Zámoly and in 168 for Pátka. The figure illustrates the absence of seasonal cycles, particularly given that the majority of outflow volumes were zero. The residuals from Model-A show a random distribution with the magnitude of outflow volume as well. Therefore, the analysis did not consider seasonal adjustment factors for either reservoir. This decision was further reinforced by the calibration and verification runs, which showed no cyclic behavior in the residuals from these two reservoirs. The seasonal correction approach was applied only to Lake Velence. For the Zámoly and Pátka reservoirs, monthly storage changes were calculated using Model-B (Equations (13) and (14)), and the resulting corrected outflows were used as inputs for Lake Velence. This hybrid approach allowed the model to incorporate observed seasonal error patterns while maintaining consistency in the upstream components of the cascade.

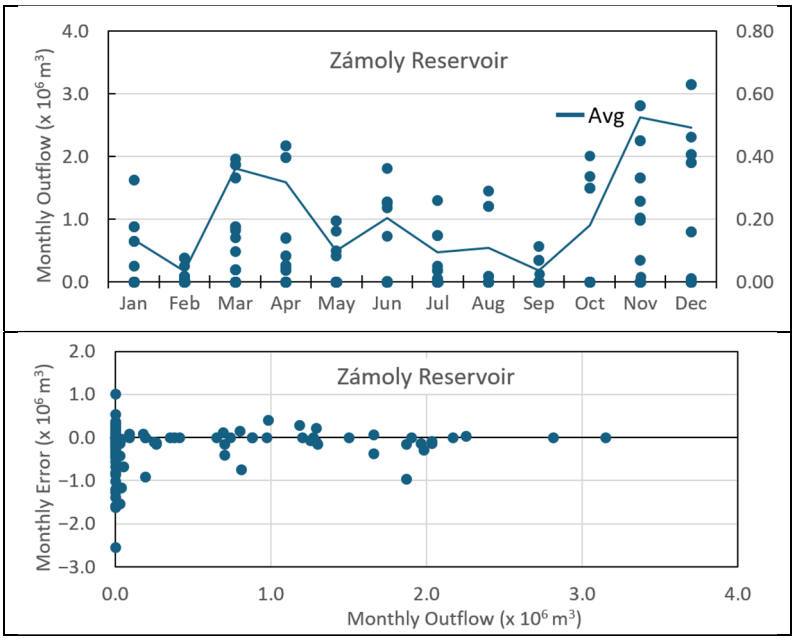


Figure 9. *Cont.*

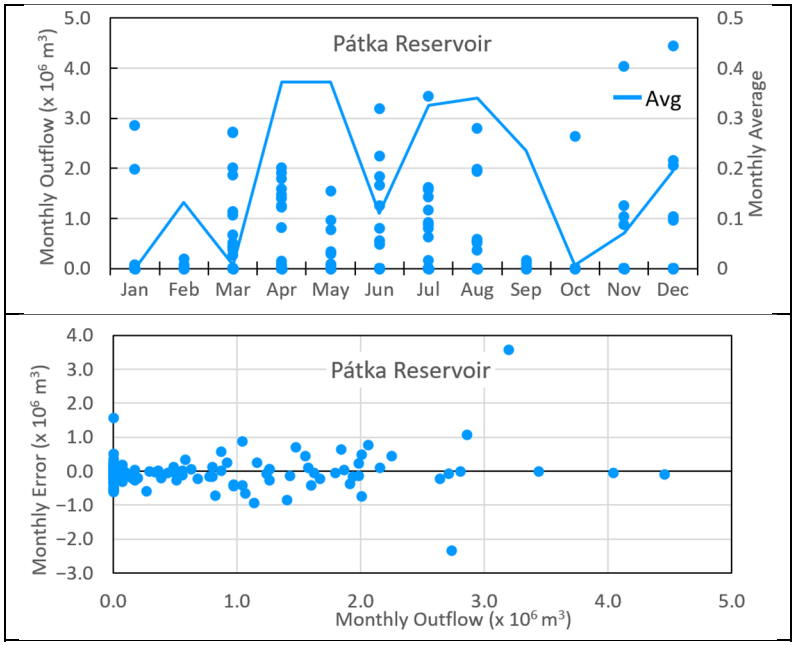


Figure 9. Data from Zámoly and Pátka Reservoirs showing monthly outflow volumes and residuals vs. monthly outflow volumes from 1998 to 2017. Note that there are a significant number of months with zero outflow volume.

4.6. Residual Diagnostics and Model Assumptions

Model performance was further evaluated through residual diagnostics. The criteria were assessed by (1) multicollinearity of input variables (Variance Inflation Factor), (2) autocorrelation of residuals (e.g., autocorrelation function), (3) normality of residuals (distribution analysis), and (4) homoscedasticity (variance consistency).

These diagnostics were used to evaluate the validity of regression assumptions and to identify remaining model deficiencies. Schmidt and Finan [33] point out that, for large sample sizes, violating the normality assumption does not markedly affect the results, and the focus should be on the independence and low multicollinearity of the inputs, as well as the homoscedasticity of the residual errors.

4.7. Validation and Performance Assessment

The models were calibrated for the period 1998–2017. The stability of the calibrated coefficients was assessed by applying the calibrated parameter sets without modification during the independent validation period (2018–2024), thereby evaluating their transferability beyond the calibration dataset. For both the calibration and validation periods, model performance was evaluated using Nash–Sutcliffe Efficiency (NSE) [42].

5. Results

5.1. Performance of the Baseline Model (Model-A)

The water levels of Zámoly Reservoir, Pátka Reservoir, and Lake Velence were first simulated using the baseline water balance formulation (Model-A), based solely on measured and calculated but uncorrected input data, as shown in Figure 10a–c with a black line. The results show that Model-A produces rapidly diverging and physically unrealistic water levels within a few years. The accumulation of systematic errors in the water balance components propagates through the continuous storage calculation. The performance of Model-A for Zámoly Reservoir is NSE-$A_{ZR} = -0.28$; for Pátka Reservoir, NSE-$A_{PR} = -5.13$; and for Lake Velence, NSE-$A_{LV} = -18.69$. As a result, Model-A is unsuitable for long-term simulation of the cascaded system.

The divergence is observed consistently across all three water bodies, indicating that errors are systematic and propagate through the interconnected cascade. Since storage is computed continuously, even small biases in individual components accumulate over time, leading to large deviations in simulated water levels. The accumulation confirms the violation of the zero-mean residuals assumption, so the current measurement-based water balance cannot be used for long-term simulation without correction. Importantly, the divergence is observed consistently across all three water bodies, indicating that the issue is not localized but affects the entire cascaded system. This observation supports the hypothesis that errors propagate both temporally and spatially within interconnected systems.

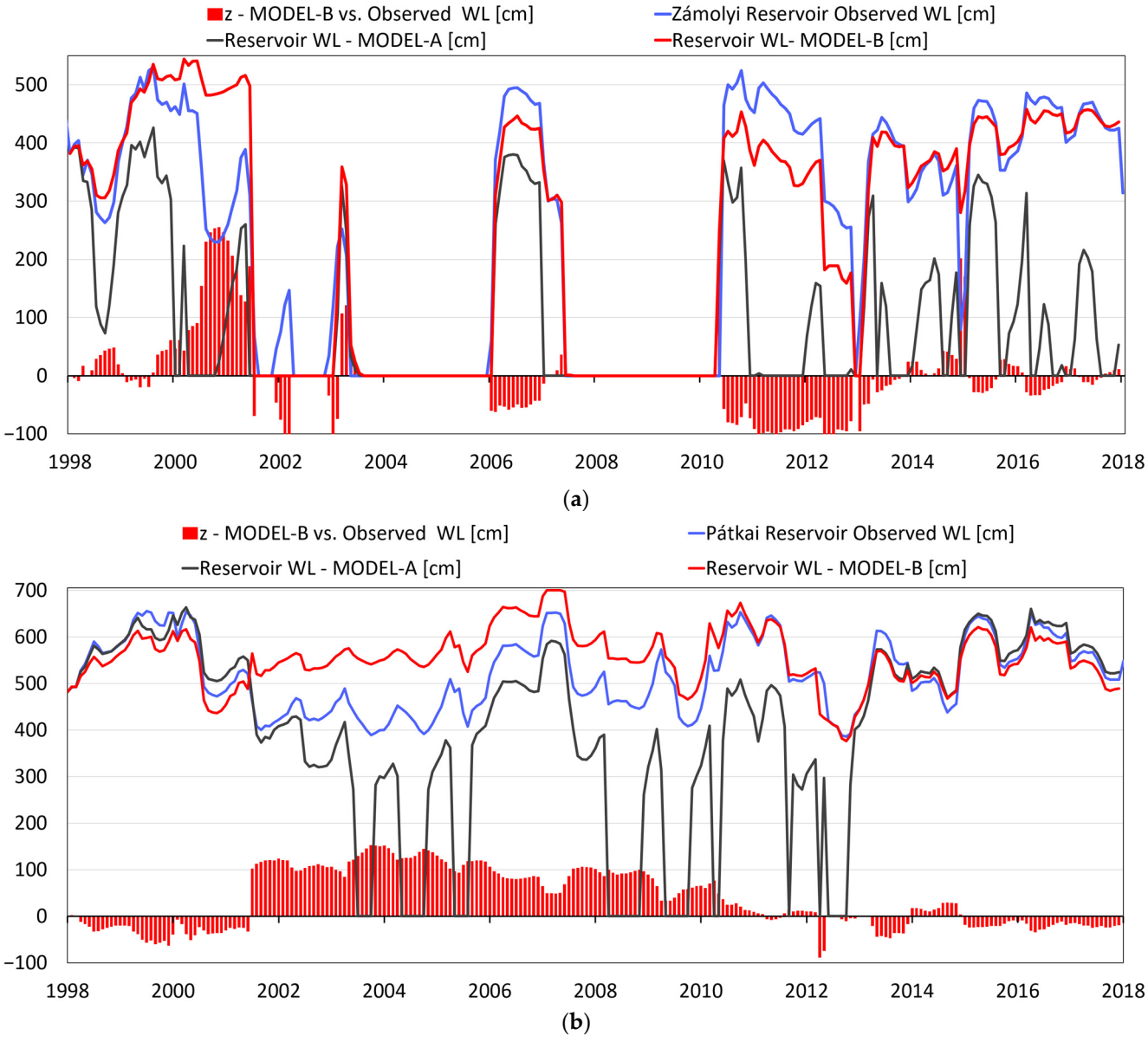


Figure 10. *Cont.*

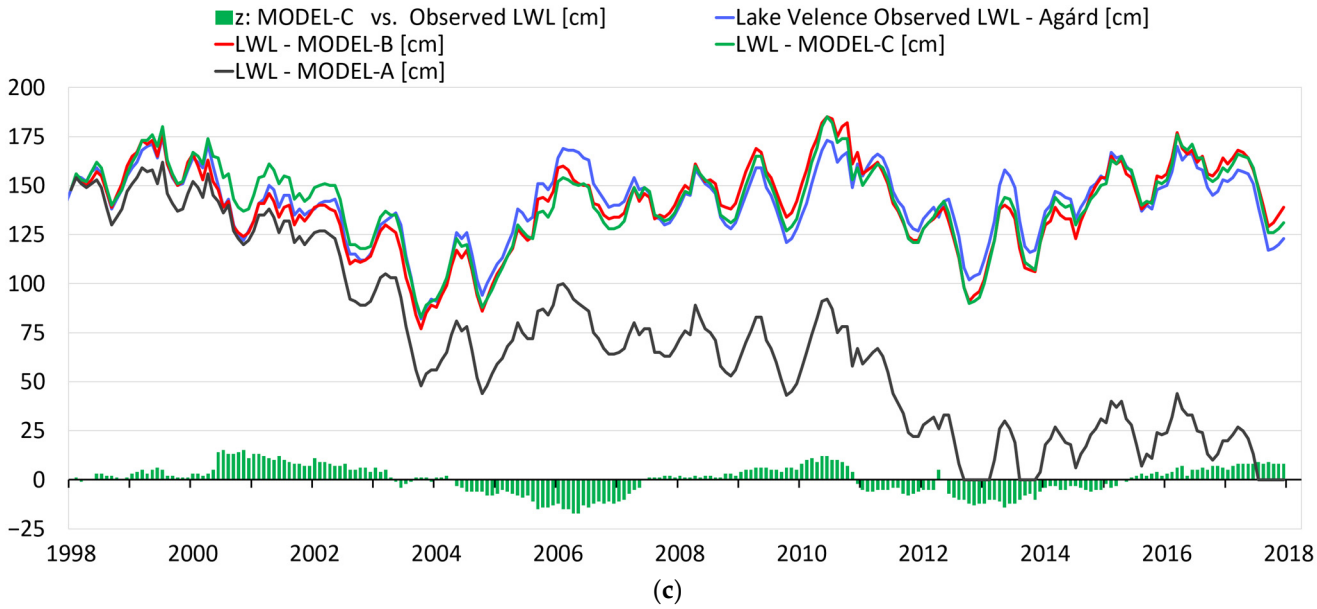


Figure 10. (**a**)—Water level comparison for Zámoly reservoir. Observed water levels (blue) and water levels calculated with Model-A (black) and Model-B (red). Model-B errors (calculated water level—observed water level, red columns). (**b**)—Water level comparison for Pátkai Reservoir. Observed water levels (blue) and water levels calculated with Model-A (black) and Model-B (red). Model-B errors (calculated water level—observed water level, red columns). (**c**)—Water level comparison for Lake Velence. Observed water levels (blue), water levels calculated with Model-A (black), Model-B (red), and Model-C (green), and errors for Model-C (calculated water level—observed water level, green columns).

5.2. Input Variable Analysis and Model Setup

Before applying the two MLR models, the statistical properties of the input variables were evaluated to ensure the validity of the modeling framework. Pairwise correlations and multicollinearity were assessed for each water body, as shown in Figures 11 and 12.

System		P()	Q()	Q()	E()	Q()	
Zámoly	P(ZR)	1.00	0.32	0.26	0.64	0.17	
	Q(Cs)	0.32	1.00	0.72	0.06	0.42	
	Q(Ba)	0.26	0.72	1.00	−0.04	0.33	
	E(ZR)	0.64	0.06	−0.04	1.00	0.05	
	Q(out-ZR)	0.17	0.42	0.33	0.05	1.00	
Pátka	P(PR)	1.00	0.39	0.08	0.24	0.09	
	Q(Rp)	0.39	1.00	0.33	−0.22	0.25	
	Q(out-ZR)	0.08	0.33	1.00	−0.06	0.59	
	E(PR)	0.24	−0.22	-0.06	1.00	0.14	
	Q(out-PR)	0.09	0.25	0.59	0.14	1.00	
Velence	P(LV)	1.00	0.23	0.02	0.18	0.17	0.12
	Q(in-LV)	0.23	1.00	0.33	−0.11	0.69	0.09
	Q(out-PR)	0.02	0.33	1.00	0.11	0.47	0.20
	E(LV)	0.18	−0.11	0.11	1.00	−0.11	0.09
	Q(out-LV)	0.17	0.69	0.47	−0.11	1.00	0.01
	Q(use)	0.12	0.09	0.20	0.09	0.01	1.00

−1.00 −0.80 −0.60 −0.40 −0.20 −0.00 1.00 0.80 0.60 0.40 0.20 0.00

Figure 11. Pairwise Pearson correlation heat maps for Zámoly, Pátka, and Velence.

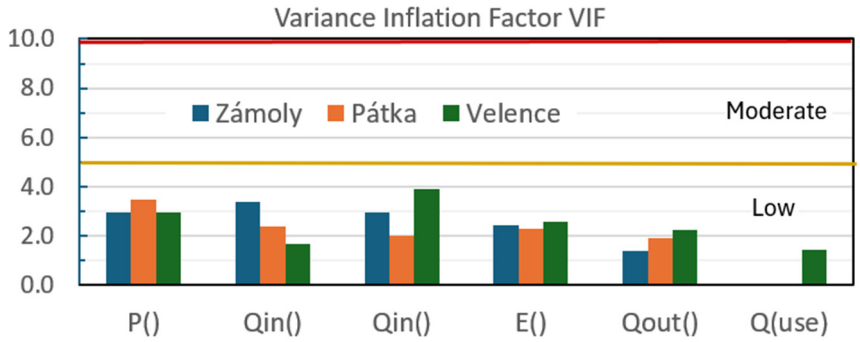


Figure 12. Variance Inflation Factors for Zámoly, Pátka, and Velence.

Moderate correlations were observed between inflow and outflow variables, reflecting the physical connectivity of the cascade system. However, multicollinearity remained within acceptable limits, with all Variance Inflation Factors (VIFs) below 5. For Lake Velence, most variables were weakly correlated, except inflow (Q_{in}) and outflow (Q_{out}). Since Q_{out} is directly measured, while Q_{in} partly includes estimated components, Q_{out} was considered more reliable. In addition, based on statistical testing (Student-t), Q_{in} showed higher explanatory power; therefore, Q_{out} was excluded from the regression formulation.

5.3. Performance of the Bias-Corrected Model (Model-B)

Model-B (red curves in Figure 10a–c) results in a substantial improvement over Model-A. The simulated water levels show improved agreement with observations: Zámoly Reservoir: NSE = B_{ZR} = 0.89, Pátka Reservoir: NSE-B_{PR} = 0.17, and Lake Velence: NSE-B_{LV} = 0.83.

The relatively low NSE for the Pátka Reservoir resulted from known inconsistencies in the measured reservoir release data from Zámoly and Pátka Reservoirs in June 2001. The large absolute error caused by this one month persisted over several years, up until 2010. Meanwhile, the model approximated monthly changes in water storage well. Overall, Model-B significantly reduces systematic bias and stabilizes long-term simulations. Table 3 presents the calculated coefficients.

Table 3. Final MLR coefficients used when calibrating Model-B and Model-C.

MLR Coefficient	Zámoly Reservoir	Pátka Reservoir	Lake Velence	
			Model-B	Model-C
Precipitation	$\beta_{ZR,1}$ = 1.040	$\beta_{PR,1}$ = 1.300	$\beta^{B}_{LV,1}$ = 1.270	$\beta^{C}_{LV,1}$ = 1.260
Inflow-1	$\beta_{ZR,2}$ = 1.940	$\beta_{PR,2}$ = 0.870	$\beta^{B}_{LV,2}$ = 0.815	$\beta^{C}_{LV,2}$ = 0.815
Inflow-2	$\beta_{ZR,3}$ = 1.200	$\beta_{PR,3}$ = 0.700 [1]	$\beta^{B}_{LV,3}$ = 0.820 [2]	$\beta^{C}_{LV,3}$ = 0.820 [2]
Evaporation	$\beta_{ZR,4}$ = 0.914	$\beta_{PR,4}$ = 1.130	$\beta^{B}_{LV,4}$ = 0.988	$\beta^{C}_{LV,4}$ = 0.970
Outflow	$\beta_{ZR,5}$ = 0.700 [1]	$\beta_{PR,5}$ = 0.820 [2]	$\beta^{B}_{LV,5}$ = 1.000	$\beta^{C}_{LV,5}$ = 1.000

[1] Outflow from Zámoly = Inflow-2 to Pátka. [2] Outflow from Pátka = Inflow-2 to Velence.

As discussed earlier, Table 3 shows that the upstream outflow coefficients equal the corresponding downstream inflow coefficients, ensuring consistency within the cascaded system. The forced agreement (two coefficients), setting the Velence to ground truth (one coefficient = 1.0), and forcing zero intercepts (three offset coefficients = 0.0) reduce the number of adjustable MLR coefficients from 18 to 12.

The improvement introduced by Model-B demonstrates that MLR bias correction of water balance components can effectively reduce systematic errors.

5.4. Performance of the Seasonal Model (Model-C)

For Lake Velence, Model-C (see green curve in Figure 10c) provides a moderate improvement over Model-B during the calibration period, with NSE increasing by 0.03 points from NSE-B_{LV} = 0.83 to NSE-C_{LV} = 0.86.

Although the increase in NSE is modest, the improvement introduced by Model-C is structurally important. By applying correction only during the December–May period, the model directly targets the time window where systematic errors are known to occur. This selective correction reduces seasonal bias without overfitting periods where the original data already perform well. As a result, Model-C provides a more balanced representation of the system, as shown later in the model performance assessment and final verification.

5.5. Model Performance Assessment

5.5.1. Model Accuracy and Structural Assessment

The agreement between the simulated and observed water levels was evaluated using scatter plots, including a 45° reference line and fitted regression lines (Figure 13). The 45° line represents perfect agreement, while deviations from this line indicate bias or scaling errors.

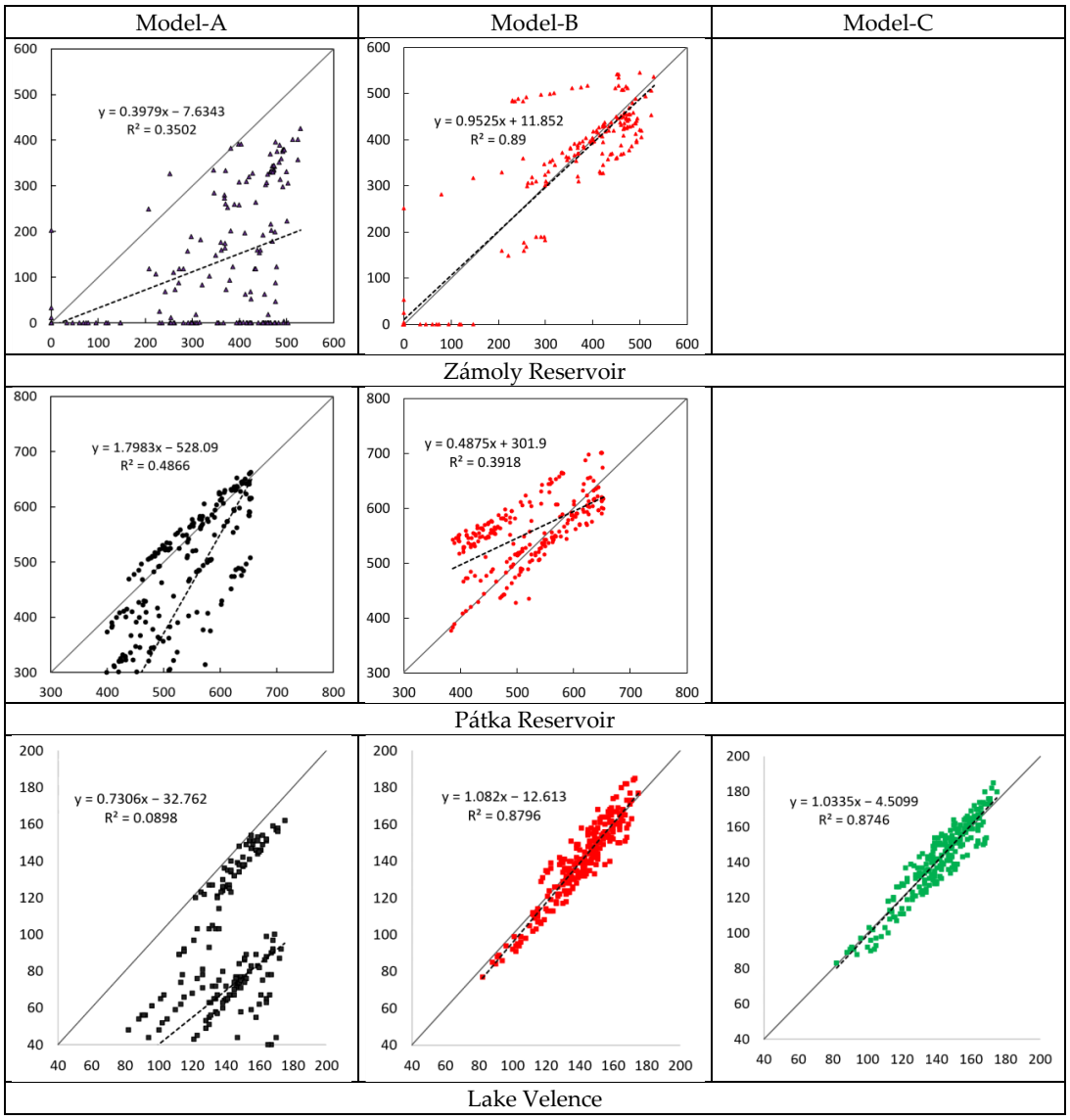


Figure 13. Comparison of measured and predicted water levels with Models A, B, and C in Zámoly, Pátka Reservoirs, and Lake Velence.

For Model-A, the regression results show substantial deviation from the 45° line, with slopes far from unity and low coefficients of determination. The result indicates both poor representation of variability and significant systematic bias. For Model-B, the regression slopes are close to 1, and intercepts are relatively small, particularly for Zámoly Reservoir (slope ≈ 0.95, R^2 ≈ 0.89), indicating that both the magnitude and variability in water levels are well reproduced. For Pátka Reservoir, the regression slope remains notably different due to the inconsistency in reservoir outflow for a certain month, as mentioned earlier. For Lake Velence, both Model-B and Model-C show slopes close to unity and high R^2 values, indicating good agreement with the observed dynamics. However, Model-C further improves the intercept, indicating improved bias correction. Model-C results for the reservoirs are identical to Model-B; therefore, their regression relationships remain identical.

5.5.2. Residual Analysis

Residual diagnostics reveal clear differences in model performance across water bodies and model formulations. Figure 14 shows histograms of the residuals for Model-B and Model-C. The histograms for Zámoly and Pátka Reservoirs show distributions that are somewhat non-normal. The Zámoly data contain a significant number of zero values (317/444) for the time period studied (see Figure 9). The histograms for Lake Velence show a more normal distribution with Model-C.

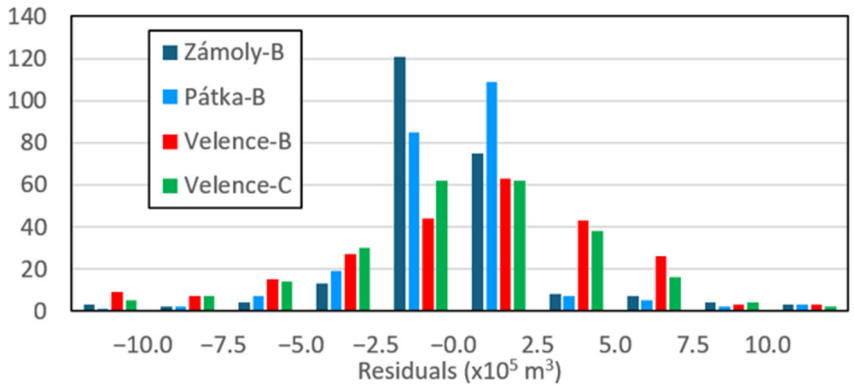


Figure 14. Histogram plots of residual values for Zámoly, Pátka (Model-B), and Lake Velence (Model-B and Model-C).

Figure 15 shows the autocorrelation functions (ACFs) and distribution of residuals for Zámoly and Pátka Reservoirs. Zámoly Reservoir exhibits some periodicity in its autocorrelation results; however, the ACF magnitude is modest. The Pátka Reservoir results show no pattern and low magnitudes. The residual distributions follow the same pattern, with Zámoly showing heteroscedasticity: negative fitted values yield positive residuals, and positive values yield negative residuals. Pátka presents a more random distribution; however, residuals may tend to increase with increasing fitted values, but the distribution is generally homoscedastic.

In contrast, the residuals for Lake Velence under Model-B are homoscedastic and closer to normality, but they exhibit autocorrelation, indicating temporal dependence. Model-C leads to a substantial improvement in residual behavior for Lake Velence. Residuals become more normally distributed, autocorrelation reduces significantly, and variance remains consistent over time, indicating homoscedasticity. Residual analyses for Lake Velence are shown in Figure 16.

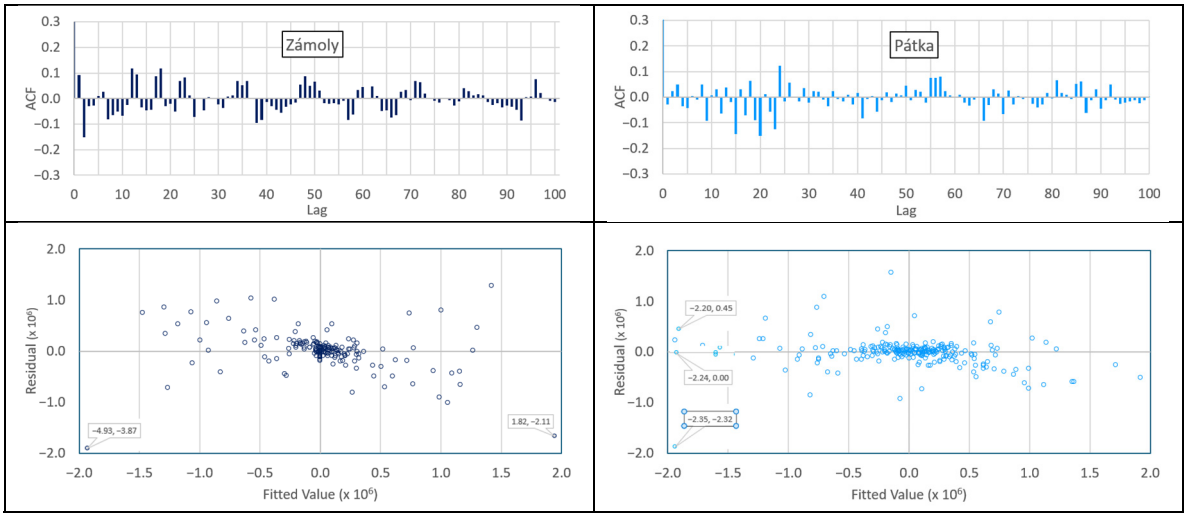


Figure 15. Residual diagnostics for Zámoly Reservoir (**left**) and Pátka Reservoir (**right**).

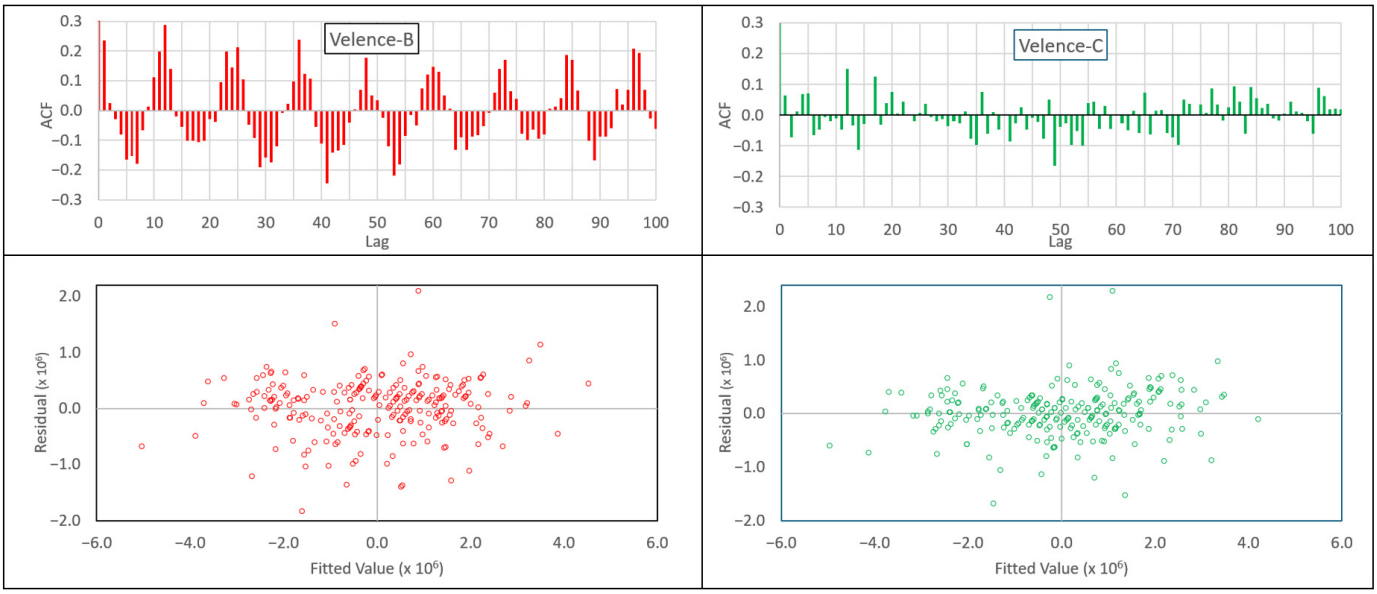


Figure 16. Residual diagnostics for Lake Velence: Model-B (**left**) and Model-C (**right**).

5.6. Validation Results (2018–2024)

Model performance was evaluated on an independent validation period (2018–2024). For the reservoirs, Model-B produced outstanding results, exceeding the calibration results. For Zámoly Reservoir, $NSE^{B}_{ZR\text{-}valid} = 0.91$, and for Pátka Reservoir, $NSE^{B}_{PR\text{-}valid} = 0.92$.

The observed and simulated water level time series are shown in Figure 17 for Zámoly Reservoir and in Figure 18 for Pátka Reservoir.

For Lake Velence, Model-C outperformed Model-B (Figure 19): for Model-B, $NSE^{B}_{LV\text{-}valid} = 0.60$, and for Model-C, $NSE^{C}_{LV\text{-}valid} = 0.72$.

The validation results highlight a key difference between Model-B and Model-C. While Model-B achieves a strong fit during the initial part of the validation period, its performance deteriorates under changing hydrological conditions, indicating limited robustness. In contrast, Model-C provides more stable performance across the entire validation period, with reduced long-term drift. The better performance suggests that accounting for seasonal bias improves model performance under non-stationary conditions. The improved validation

results of Model-C are consistent with the residual analysis and support the hypothesis that temporally structured errors are a dominant source of uncertainty in the system.

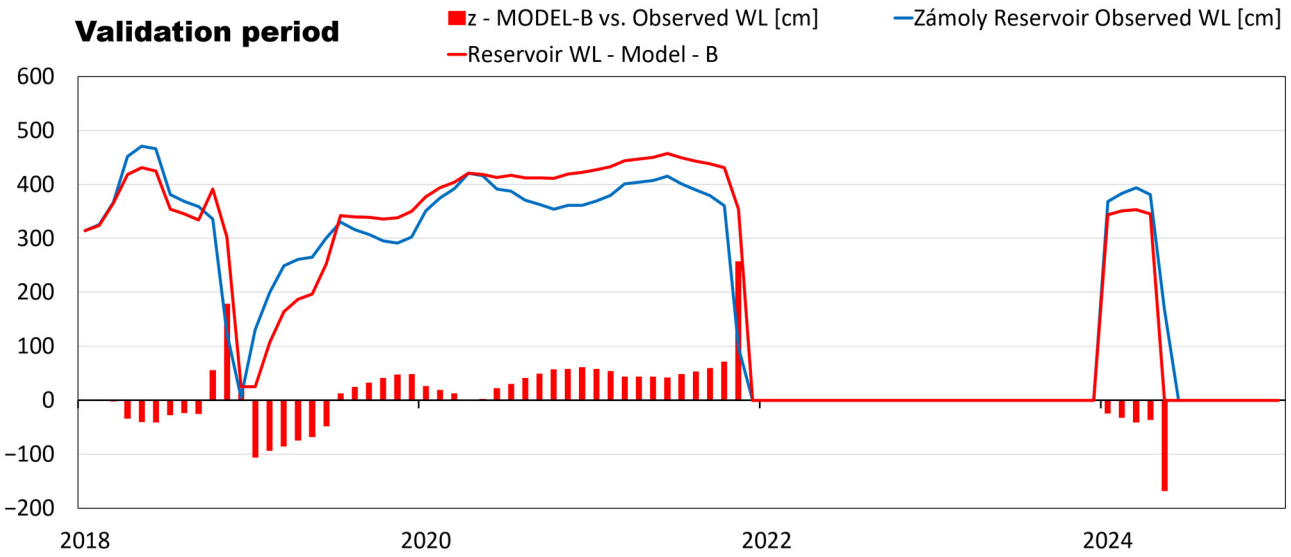


Figure 17. Validation results for Zámoly Reservoir. Observed water levels (blue) and water levels calculated with Model-B (red). Model-B errors = calculated water level—observed water level (red columns).

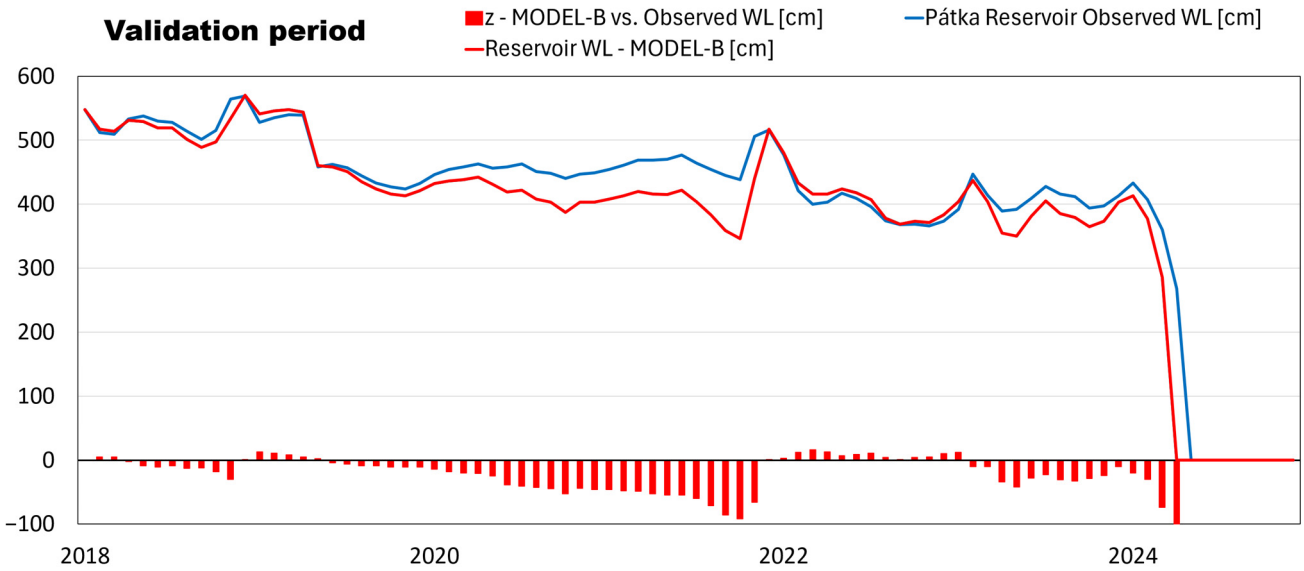


Figure 18. Validation results for Pátka Reservoir. Observed water levels (blue) and water levels calculated with Model-B (red). Model-B errors = calculated water level—observed water level (red columns).

The results show a clear progression in model performance. Model-A diverges due to cumulative bias, Model-B stabilizes the system through global correction, and Model-C further improves consistency by addressing the seasonal error structure.

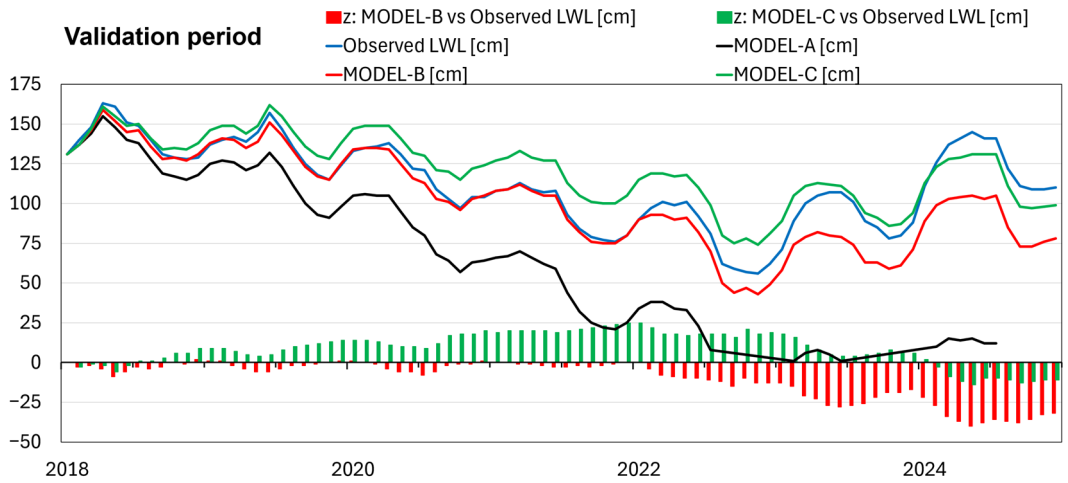


Figure 19. Validation results for Lake Velence. Observed water levels (blue) and water levels calculated with Model-A (black), Model-B (red), Model-B errors (red columns), Model-C (green), and Model-C errors (green columns).

6. Discussion

The results demonstrate that the water balance of the cascaded reservoir–lake system is affected by systematic and distributed errors rather than purely random variability. This conclusion is supported by the failure of Model-A, which results in physically unrealistic water levels within a relatively short period. Such behavior indicates that the assumption of zero-mean residuals is violated and that systematic biases are present in the input data. A key observation is that these errors cannot be attributed to a single component but instead arise from inconsistencies across multiple water balance terms, including inflow, precipitation, evaporation, and reservoir releases. The multiple terms are particularly evident in Zámoly Reservoir, where measured outflows substantially exceed the sum of inflows and precipitation minus evaporation. This imbalance suggests unmeasured inflows, uncertainties in discharge measurements, or structural limitations in the monitoring network. The relatively large correction applied to inflow-related coefficients in Model-B further supports the interpretation that inflow underestimation plays a major role in the observed discrepancies.

An important outcome of this study is the demonstration that errors propagate both temporally and spatially within cascaded hydrological systems. In such systems, inaccuracies introduced in upstream components are transmitted downstream through reservoir releases, thereby affecting all subsequent elements of the cascade. This mechanism explains the observed differences in model performance between the three water bodies. The upstream Zámoly Reservoir shows the best agreement after correction. At the same time, performance deteriorates at the intermediate Pátka Reservoir and then stabilizes again at Lake Velence due to the buffering effect of its larger storage capacity. These results highlight that water balance errors are not isolated but are dynamically transferred through the system. This finding extends previous studies that focused primarily on individual lakes or reservoirs and underscores the need for system-level approaches to address water balance closure in interconnected systems.

The application of multiple linear regression as a bias correction tool significantly improves model performance and stabilizes long-term simulations. However, the regression coefficients should not be interpreted solely as empirical fitting parameters. Instead, they provide insight into the underlying inconsistencies within the water balance components. Coefficients deviating from unity indicate the presence of systematic measurement errors, missing processes, or structural deficiencies of the initial water balance model. For instance,

elevated coefficients associated with inflow suggest that measured inflows underestimate actual water contributions, while deviations in precipitation coefficients may reflect spatial representativeness issues or methodological limitations.

In contrast, evaporation coefficients remained consistently close to unity across all calibrated models, indicating that the evaporation estimates were generally reliable and required only minor adjustment. This consistency suggests that evaporation contributes less to the observed water balance errors than precipitation or inflow estimates. In this sense, the regression-based correction serves not only as a modeling tool but also as a diagnostic framework for identifying uncertainties in the input data.

One of the most significant findings of the study is the identification of a pronounced seasonal pattern in water balance errors, with systematic deviations occurring primarily during the December–May period. The improved performance of Model-C confirms that these errors are temporally structured and cannot be adequately addressed using time-invariant correction factors. The seasonal behavior suggests the influence of processes not fully captured by the current water balance formulation, including seasonal biases in precipitation measurements, potential groundwater interactions, and/or operational effects of reservoir management.

The role of precipitation uncertainty deserves particular attention. Spatial variability of rainfall across the catchment may introduce systematic bias when a limited number of stations are used to represent basin-wide conditions. Kebede et al. [43] reported similar findings that spatial rainfall gradients can lead to overestimation of lake levels when not properly accounted for. For Lake Velence, the weak negative correlation between precipitation and residual error suggests that precipitation inputs may be underestimated during wetter periods. However, precipitation bias alone cannot explain the magnitude and seasonal structure of the observed water balance deficit.

A more plausible explanation involves an unobserved groundwater component. As highlighted by Dorigo et al. [9], a water balance cannot be accurately closed if one or more components are missing. The groundwater modeling results of Baják et al. [37] provide strong evidence for this, indicating a net groundwater inflow of approximately 47 lake mm/year. This magnitude is comparable to the observed long-term annual error and would substantially reduce the imbalance if included in the water balance equation. The omission of groundwater exchange, therefore, represents a major limitation of the current modeling framework. The absence of an explicit groundwater term has important implications for interpreting the MLR-based correction coefficients as well. In the present model, regression coefficients implicitly compensate for missing or biased components, including groundwater fluxes. As a result, the calibrated coefficients represent aggregated corrections rather than purely physical scaling factors. Explicitly incorporating groundwater would likely reduce the magnitude of these corrections and improve the model's physical interpretability while allowing a clearer distinction between measurement error and actual hydrological processes.

Residual diagnostics provide further insight into model adequacy. Model-B residuals exhibit heteroscedasticity, deviations from normality, and temporal autocorrelation, particularly for Lake Velence. These characteristics indicate that important processes remain unresolved and that model assumptions are only partially satisfied. The introduction of seasonal correction in Model-C significantly improves residual behavior by reducing autocorrelation, stabilizing variance, and bringing residuals closer to normality. These results confirm that seasonality is a key driver of model error structure.

The findings confirm that water balance closure in regulated systems cannot be achieved solely through the direct use of measured data. Instead, systematic errors embedded within multiple components must be addressed simultaneously. The proposed

methodology provides a practical framework for achieving this by combining physical consistency with data-driven correction.

Despite these improvements, several limitations remain. The model does not explicitly account for groundwater exchange, and the regression-based correction does not provide a comprehensive physical explanation for the observed discrepancies. In addition, model performance depends on the quality of input data and the accuracy of reservoir operation records. Future research should focus on integrating groundwater fluxes, exploring probabilistic or Bayesian approaches for uncertainty representation [11], and combining regression-based correction with conceptual hydrological models. Although the present study demonstrates the applicability of the proposed methodology for reconstructing a consistent water balance, an explicit assessment of parameter uncertainty and sensitivity would be a valuable prerequisite before applying the calibrated models to evaluate alternative reservoir operation scenarios, as it would allow the robustness of management decisions to be quantified.

Overall, the proposed approach provides a scalable and operationally applicable method for closing the water balance in cascaded systems. By correcting systematic biases and ensuring consistency across connected water bodies, the methodology enables more reliable long-term simulations and supports the development of water management scenarios.

7. Conclusions

This study presented a data-driven methodology for constructing and closing the joint water balance of a cascaded reservoir–lake system using long-term monthly observations. A baseline water balance model based on uncorrected measured data (Model-A) was first implemented, followed by two bias correction approaches using multiple linear regression (Model-B and Model-C).

The results demonstrate that applying measured water balance components directly leads to significant cumulative errors, rendering the baseline model unsuitable for long-term simulation. These discrepancies arise from the combined effect of systematic errors across multiple components rather than from a single dominant source. The application of MLR correction with uniform coefficients across all months (Model-B) substantially improves model performance and stabilizes the simulated water level time series across all water bodies. However, residual analysis reveals that model errors exhibit temporal structure, particularly at the lake. Incorporating this seasonal behavior (Model-C) further improves model consistency by reducing autocorrelation and improving residual characteristics, resulting in more reliable long-term simulations.

The findings highlight that water balance closure in cascaded systems requires a system-level approach that accounts for both spatial connectivity and temporal variability of errors. The proposed framework provides a transparent and computationally efficient method for correcting systematic biases while preserving the information contained in measured time series.

Despite these improvements, limitations remain, particularly regarding the representation of groundwater exchange and other unobserved processes. Future work should focus on integrating additional physical components and on exploring probabilistic approaches to represent uncertainty explicitly.

Overall, the developed methodology enables the reconstruction of consistent water balance time series and provides a robust foundation for subsequent water management scenario analysis in regulated and interconnected lake and reservoir systems.

Author Contributions: Conceptualization, M.C.; methodology, M.C.; formal analysis, M.C. and K.B.; investigation, M.C.; resources, K.B. and R.R.; writing—original draft preparation, M.C.; writing—review and editing, K.B. and R.R.; visualization, M.C. and R.R.; supervision, K.B.; project administra-

tion, K.B.; funding acquisition, K.B. All authors have read and agreed to the published version of the manuscript.

Funding: The research presented in the article was carried out within the framework of the Széchenyi Plan Plus program with the support of the RRF 2.3.1 21 2022 00008 project.

Data Availability Statement: Data from this study is available upon request to the authors.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study, in the collection, analysis, or interpretation of data, in the writing of this manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

MLR	multiple linear regression
KDTVIZIG	Central Transdanubian Water Directorate
ACF	autocorrelation function

References

1. Gleick, P.H. *Water-In-Crisis*; Oxford University Press: New York, NY, USA, 1993.
2. Crétaux, J.-F.; Abarca-del-Río, R.; Bergé-Nguyen, M.; Arsen, A.; Drolon, V.; Clos, G.; Maisongrande, P. Lake Volume Monitoring from Space. *Surv. Geophys.* **2016**, *37*, 269–305. [CrossRef]
3. Williamson, C.E.; Saros, J.E.; Vincent, W.F.; Smol, J.P. Lakes and reservoirs as sentinels, integrators, and regulators of climate change. *Limnol. Oceanogr.* **2009**, *54*, 2273–2282. [CrossRef]
4. Rodell, M.; Beaudoing, H.K.; L'Ecuyer, T.S.; Olson, W.S.; Famiglietti, J.S.; Houser, P.R.; Adler, R.; Bosilovich, M.G.; Clayson, C.A.; Chambers, D. The observed state of the water cycle in the early twenty-first century. *J. Clim.* **2015**, *28*, 8289–8318. [CrossRef]
5. Sun, J.; Chen, W.; Hu, B.; Xu, Y.J.; Zhang, G.; Wu, Y.; Hu, B.; Song, Z. Roles of reservoirs in regulating basin flood and droughts risks under climate change: Historical assessment and future projection. *J. Hydrol. Reg. Stud.* **2023**, *48*, 101453. [CrossRef]
6. USGS. *Water Budgets: Foundations for Effective Water-Resources and Environmental Management*; USGS: Reston, VA, USA, 2007.
7. Winter, T.C. Uncertainties in estimating the water balance of lakes. *JAWRA J. Am. Water Resour. Assoc.* **1981**, *17*, 82–115. [CrossRef]
8. Safeeq, M.; Bart, R.R.; Pelak, N.F.; Singh, C.K.; Dralle, D.N.; Hartsough, P.; Wagenbrenner, J.W. How realistic are water-balance closure assumptions? A demonstration from the southern sierra critical zone observatory and kings river experimental watersheds. *Hydrol. Process.* **2021**, *35*, e14199. [CrossRef]
9. Dorigo, W.; Dietrich, S.; Aires, F.; Brocca, L.; Carter, S.; Cretaux, J.-F.; Dunkerley, D.; Enomoto, H.; Forsberg, R.; Güntner, A.; et al. Closing the water cycle from observations across scales where do we stand? *Bull. Am. Meteorol. Soc.* **2021**, *102*, E1897–E1935. [CrossRef]
10. Wang, K.; Hu, T.; Zhang, P.; Huang, W.; Mao, J.; Xu, Y.; Shi, Y. Improving Lake Level Prediction by Embedding Support Vector Regression in a Data Assimilation Framework. *Water* **2022**, *14*, 3718. [CrossRef]
11. Schoups, G.; Nasseri, M. GRACEfully Closing the Water Balance: A Data-Driven Probabilistic Approach Applied to River Basins in Iran. *Water Resour. Res.* **2021**, *57*, e2020WR029071. [CrossRef]
12. Trask, J.C.; Fogg, G.E.; Puente, C.E. Resolving hydrologic water balances through a novel error analysis approach, with application to the Tahoe basin. *J. Hydrol.* **2017**, *546*, 326–340. [CrossRef]
13. Rusli, S.; Weerts, A.; Taufiq, A.; Bense, V. Estimating water balance components and their uncertainty bounds in highly groundwater-dependent and data-scarce area: An example for the Upper Citarum basin. *J. Hydrol. Reg. Stud.* **2021**, *37*, 100911. [CrossRef]
14. Durnford, D.; Fortin, V.; Smith, G.C.; Archambault, B.; Deacu, D.; Dupont, F.; Dyck, S.; Martinez, Y.; Klyszejko, E.; MacKay, M.; et al. Toward an operational water cycle prediction system for the great lakes and St. Lawrence river. *Bull. Am. Meteorol. Soc.* **2018**, *99*, 521–546. [CrossRef]
15. Jang, D.; Choi, G.; Park, H. Adaptation of multiple regression analysis to identify effective factors of water losses in water distribution systems. *Smart Water* **2019**, *4*, 1. [CrossRef]
16. Somogyvári, M.; Scherer, D.; Bart, F.; Fehrenbach, U.; Okujeni, A.; Krueger, T. A hybrid data-driven approach to analyze the drivers of lake level dynamics. *Hydrol. Earth Syst. Sci.* **2024**, *28*, 4331–4348. [CrossRef]
17. Hinegk, L.; Adami, L.; Piccolroaz, S.; Amadori, M.; Moretti, M.; Tubino, M.; Toffolon, M. Multidecadal analysis of Lake Garda water balance. *J. Limnol.* **2023**, *82*, 149–169. [CrossRef]

18. Moknatian, M.; Piasecki, M. Development of predictive models for water budget simulations of closed-basin lakes: Case studies of lakes azuei and enriquillo on the island of hispaniola. *Hydrology* **2021**, *8*, 148. [CrossRef]
19. Koris, K.; Nagy, B. Regulation model of the Lake Velence Reservoir System. *Period. Polytech. Civ. Eng.* **1977**, *21*, 147–155.
20. KDTVIZIG. *Velencei-Tó Vízmérleg, 2002 (Lake Velence Water Balance, 2002)*; Central Transdanubian Water Directorate, Annual Report; Közép-dunántúli Vízügyi Igazgatóság: Székesfehérvár, Hungary, 2002; 19p.
21. Wang, L.; Wang, J.; Wang, L.; Zhu, L.; Li, X. Lake Evaporation and Its Effects on Basin Evapotranspiration and Lake Water Storage on the Inner Tibetan Plateau. *Water Resour. Res.* **2023**, *59*, e2022WR034030. [CrossRef]
22. López-Moreno, J.I.; Vicente-Serrano, S.M.; Beguería, S.; García-Ruiz, J.M.; Portela, M.M.; Almeida, A.B. Dam effects on droughts magnitude and duration in a transboundary basin: The Lower River Tagus, Pain and Portugal. *Water Resour. Res.* **2009**, *45*, e2008WR007198. [CrossRef]
23. Vanderkelen, I.; van Lipzig, N.P.M.; Thiery, W. Modelling the water balance of Lake Victoria (East Africa)—Part 1: Observational analysis. *Hydrol. Earth Syst. Sci.* **2018**, *22*, 5509–5525. [CrossRef]
24. Lima-Quispe, N.; Ruelland, D.; Rabatel, A.; Lavado-Casimiro, W.; Condom, T. Modeling Lake Titicaca's water balance: The dominant roles of precipitation and evaporation. *Hydrol. Earth Syst. Sci.* **2025**, *29*, 655–682. [CrossRef]
25. Tran, H.Q.; Fehér, Z.Z. Water balance calculation capability of hydrological models. *Acta Agrar. Kaposváriensis* **2022**, *26*, 37–53. [CrossRef]
26. Vanderkelen, I.; van Lipzig, N.P.M.; Thiery, W. Modelling the water balance of Lake Victoria (East Africa)—Part 2: Future projections. *Hydrol. Earth Syst. Sci.* **2018**, *22*, 5527–5549. [CrossRef]
27. Nigatu, Z.M.; Rientjes, T.; Haile, A.T. Hydrological Impact Assessment of Climate Change on Lake Tana's Water Balance, Ethiopia. *Am. J. Clim. Change* **2016**, *5*, 27–37. [CrossRef]
28. Song, J.; Her, Y.; Kang, M. Estimating Reservoir Inflow and Outflow From Water Level Observations Using Expert Knowledge: Dealing with an Ill-Posed Water Balance Equation in Reservoir Management. *Water Resour. Res.* **2022**, *58*, e2020WR028183. [CrossRef]
29. Gronewold, A.D.; Smith, J.P.; Read, L.K.; Crooks, J.L. Reconciling the water balance of large lake systems. *Adv. Water Resour.* **2020**, *137*, 103505. [CrossRef]
30. Khatibi, R.; Ghorbani, M.; Naghipour, L.; Jothiprakash, V.; Fathima, T.; Fazelifard, M. Inter-comparison of time series models of lake levels predicted by several modeling strategies. *J. Hydrol.* **2014**, *511*, 530–545. [CrossRef]
31. Ciampittiello, M.; Dresti, C.; Saidi, H. Water resource management through understanding of the water balance components: A case study of a sub-alpine shallow lake. *Water* **2021**, *13*, 3124. [CrossRef]
32. Xu, C.-Y.; Singh, V.P. A Review on Monthly Water Balance Models for Water Resources Investigations. *Water Resour. Manag.* **1998**, *12*, 20–50. [CrossRef]
33. Schmidt, A.F.; Finan, C. Linear regression and the normality assumption. *J. Clin. Epidemiol.* **2018**, *98*, 146–151. [CrossRef] [PubMed]
34. Gyalog, L.; Horváth, I. (Eds.) *Geology of the Velence Hills and the Balatonfó: Explanatory Book of the Geological Map of the Velence Hills (1: 25 000) and the Geological Map of Pre-Sarmatian Surface of the Balatonfó-Velence Area (1: 100 000)*; Geological Institute of Hungary: Budapest, Hungary, 2004.
35. Chappon, M.; Bene, K. A Velencei-tavi vízmérlegszámítás záróhibáinak elemzése és a számítási módszertan felülvizsgálata. *Magy. Hidrol. Társ. Vándorgyűl.* **2023**, *XL*, 1–20.
36. Chappon, M.; Bene, K. Improving methods to calculate the monthly water budget for Lake Velence, Hungary. In *EGU General Assembly Conference Abstracts*; Copernicus Publications: Göttingen, Germany, 2023; p. EGU-8334. [CrossRef]
37. Baják, P.; Csepregi, A.; Szabó, P.; Chappon, M.; Tóth, Á.; Hegedűs-Csondor, K.; Erőss, A. Quantifying the overlooked groundwater component in the water budget of a shallow soda lake in Hungary amidst climate change concerns. *J. Hydrol. Reg. Stud.* **2024**, *56*, 101961. [CrossRef]
38. Montgomery, D.C.; Peck, E.A.; Vining, G.G. *Introduction to Linear Regression Analysis*, 5th ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2012.
39. Chai, T.; Draxler, R.R. Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* **2014**, *7*, 1247–1250. [CrossRef]
40. Lasdon, L.S.; Fox, R.L.; Ratner, M.W. Nonlinear optimization using the generalized reduced gradient method. *Rev. Fr. D'automatique Inform. Rech. Opérationnelle* **1974**, *8*, 73–103. Available online: https://www.numdam.org/article/RO_1974__8_3_73_0.pdf (accessed on 10 April 2026). [CrossRef]
41. Microsoft Corporation. *Microsoft Excel, 2026*, version 16; Microsoft Corporation: Redmond, DC, USA, 2026. Available online: https://office.microsoft.com/excel (accessed on 8 July 2026).

42. Nash, J.E.; Sutcliffe, J.V. *River Flow Forecasting Through Conceptual Models Part I—A Discussion of Principles*; North-Holland Publishing Co.: Amsterdam, The Netherlands, 1970.
43. Kebede, S.; Travi, Y.; Alemayehu, T.; Marc, V. Water balance of Lake Tana and its sensitivity to fluctuations in rainfall, Blue Nile basin, Ethiopia. *J. Hydrol.* **2006**, *316*, 233–247. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.