

SUPPLEMENTARY MATERIALS

Long-Term (2007–2024) Thermal and Water Quality Dynamics in Lake Tisza (Kisköre Reservoir), Hungary: A Shallow Freshwater Ecosystem Under Climate Pressure

David Matamoros ¹, György Szabó ¹, Eduárd Csépes ², Borbála Benkhard ¹, Emőke Kiss ¹, Mária Vasvári ¹, Péter Csorba ¹ and Tamás Mester ^{1,*}

- ¹ Department of Landscape Protection and Environmental Geography, Institute of Earth Science, Faculty of Science and Technology, University of Debrecen, 4032 Debrecen, Hungary; nicolas.matamoros@science.unideb.hu (D.M.); szabo.gyorgy@science.unideb.hu (G.S.); benkhard.borbala@science.unideb.hu (B.B.); kiss.emoke@science.unideb.hu (E.K.); vasvari.maria@science.unideb.hu (M.V.); csorba.peter@science.unideb.hu (P.C.)
- ² Regional Laboratory of the Middle-Tisza District Water Directorate, Tiszaiget 9688/3, 5000 Szolnok, Hungary; csepese@kotivizig.hu
- * Correspondence: mester.tamas@science.unideb.hu

This appendix provides detailed descriptions of inferential and multivariate statistical methods (linear regression, Spearman correlation, PCA, clustering, IDW) and data handling procedures referenced in Section 2.4 of the main manuscript. Descriptive visualizations (heatmaps, boxplots, violin plots) are also included. R code for all visualizations is available from the corresponding author upon request.

SUPPLEMENTARY S1: DETAILED DESCRIPTION OF STATISTICAL METHODS

S1.1	Temperature dynamics across the hydrological regimes (Heatmap visualization)	2
S1.2	Temporal evolution of the average annual temperature	2
S1.3	Critical Thermal-Year Proportions Across Time Periods	2
S1.4	Annual trend of mean water temperature	2
S1.5	Boxplot water temperature distribution.....	3
S1.6	Spearman's rank correlation for temperature vs. dissolved oxygen	3
S1.7	Principal Component Analysis (PCA)	4
S1.8	Hierarchical clustering (Ward's method).....	5
S1.9	Water Quality Index (WQI)	5
S1.10	Inverse Distance Weighting (IDW) interpolation	6
S1.11	Handling of missing data.....	7
S1.12	Software and reproducibility	7

S1.1 Temperature dynamics across the hydrological regimes (Heatmap visualization)

The heat map was created using averages for each period from 2007-2012, 2013-2018, 2019-2024, distributed homogeneously by month, site, and water regime. R packages such as ggplot, dplyr, stats and tidyr were used to filter the months of interest, group by site(s), month(m), period(p), and water body, and calculate the average for subsequent color-coded representation indicating higher or lower temperatures allowing seasonal comparisons by the following formula.

$$\bar{T} = \frac{1}{n} \sum_{k=1}^n T \quad (S1)$$

Where the average temperature from the same monitoring site, month and period its determined by the sum of all temperatures(n) divided by the number of measurements(T).

S1.2 Temporal evolution of the average annual temperature

For each monitoring site, the annual average temperature was calculated for all observations, and then the annual values were averaged by hydrological regime (lake-transition-river), observing multi-line time series.

S1.3 Critical Thermal-Year Proportions Across Time Periods

To determine the percentage of each type of water body that falls within classification type (1-7) (Table 2), the temporal evolution of critical thermal conditions was evaluated using the proportion of annual observations in the periods (2007–2012, 2013–2018, 2019–2024) by the expression:

$$\%Critical_{w,p} = 100 \cdot \frac{\sum I_{i,t}}{N_{w,p}} \quad (S2)$$

Counting the years that were critical by group, divided by the total number of observations expressed as a percentage, referring to $I_{i,t}$ if the observation falls under critical either 1 or 7 from the table 2 and $N_{w,p}$ the total number of observations by period and water body.

S1.4 Annual trend of mean water temperature

(Referenced in main text Section 2.4)

To quantify the rate of warming (in °C per decade) across the 18-years study period (2007-2024), ordinary least squares linear regression was applied to annual mean temperatures at each monitoring site (Table 4).

Formula:

For each site i , the annual mean temperature $T_i(t)$ in year t was modeled as:

$$T_i(t) = \beta_{0,i} + \beta_{1,i}t + \varepsilon_{i,t} \quad (S3)$$

where:

- $\beta_{0,i}$ is the intercept
- $\beta_{1,i}$ is the slope (rate of change in °C per year)
- $\varepsilon_{i,t}$ is the residual error

The slope $\beta_{1,i}$ was estimated using the least squares formula:

$$\beta_{1,i} = \frac{\sum_t (t - \bar{t}) [T_i(t) - \bar{T}_i]}{\sum_t (t - \bar{t})^2} \quad (S4)$$

where:

- \bar{t} is the mean year of the time series

- \bar{T}_i is the mean temperature at site i across all years

Conversion to °C/decade

$$Trend_i = 10x\beta_{1,i} \quad (S5)$$

Interpretation criteria

- $\beta_{1,i} > 0 \rightarrow$ Warming
- $\beta_{1,i} < 0 \rightarrow$ Cooling
- $\beta_{1,i} \approx 0 \rightarrow$ No clear trend

Assumptions checked

1. Independence of residuals (Durbin-Watson test)
2. Normality of residuals (Shapiro-Wilk test, $p > 0.05$ after transformation when necessary)
3. Homoscedasticity (Breusch-Pagan test, $p > 0.05$)

Example

For lake station TV/1, the model yielded $\beta_{1,i} = 0.0897$ °C/year, corresponding to 0.897 °C/decade ($p < 0.001$).

S1.5 Boxplot water temperature distribution

The distribution of water temperature among the types of water regimes for the period 2007-2024 was analyzed by grouping temperatures by type in conjunction with the violin method, which represents the density of observations, in addition to calculating the median, interquartile range, and extreme values, whose logic is $IQR = Q3 - Q1$ with its lower limit = $Q1 - 1.5 \times IQR$ and upper limit = $Q3 + 1.5 \times IQR$.

S1.6 Spearman's rank correlation for temperature vs. dissolved oxygen

(Referenced in main text Section 2.4)

Preliminary analysis showed that neither temperature nor dissolved oxygen followed a normal distribution (Shapiro-Wilk, $p < 0.05$ for both variables). Spearman's rank correlation is non-parametric and robust to non-normality and outliers, making it appropriate for these data.

Formula

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (S6)$$

where:

- d_i is the difference between the rank of temperature and the rank of dissolved oxygen for observation i
- n is the number of observations

Interpretation

- ρ close to $+1$: Strong positive monotonic relationship

- ρ close to -1: Strong negative monotonic relationship
- ρ close to 0: No monotonic relationship

Confidence intervals

95% bootstrap confidence intervals were calculated using 1,000 resamples with replacement.

Application by water body type

- River stations (n = 486)
- Transition stations (n = 1,134)
- Lake stations (n = 3,078)

Results are reported in Figure 11 of the main text.

S1.7 Principal Component Analysis (PCA)

(Referenced in main text Section 2.4)

To identify latent spatial gradients in water quality and reduce the dimensionality of the six physicochemical/microbiological variables: conductivity, pH, BOD₅, total nitrogen, total phosphorus, total coliforms and temperature, dissolved oxygen as a supplementary variable.

Standardization (z-scores)

Variables were measured in different units ($\mu\text{S}/\text{cm}$, mg/L , $\text{UFC}/100\text{mL}$, etc.). To give each variable equal weight, values were standardized:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \quad (\text{S7})$$

where:

- x_{ij} is the original value of variable j at site i
- \bar{x}_j is the mean of variable j across all sites
- s_j is the standard deviation of variable j

After standardization, each variable has mean = 0 and variance = 1.

PCA procedure

1. Correlation matrix was computed from standardized variables
2. Eigenvalues and eigenvectors were extracted
3. Principal components (PCs) were ordered by decreasing eigenvalue
4. Components with eigenvalues > 1.0 were retained (Kaiser criterion)
5. Loadings (correlations between original variables and PCs) were examined

Variance explained

PC1 and PC2 together explained 85% of total variance across all three periods (G1: 2007-2012, G2: 2013-2018, G3: 2019-2024).

Interpretation of loadings

-Loading > 0.6: Strong contribution to the component

-Loading between 0.3-0.6: Moderate contribution

-Loading < 0.3: Weak contribution

Results are visualized in Figure 13 of the main text.

S1.8 Hierarchical clustering (Ward's method)

(Referenced in main text Section 2.4)

To group monitoring sites with similar water quality characteristics, complementing the PCA with a classification approach.

Dissimilarity measure (Euclidean distance)

$$d_{i,j} = \sqrt{\sum_{k=1}^n (Z_{ik} - Z_{jk})^2} \quad (S8)$$

where z_{ik} is the standardized value of variable k at site i (using Equation S5 above).

Clustering algorithm (Ward's minimum variance method), at each step, the algorithm merges the pair of clusters that minimizes the increase in total within-cluster variance. The increase in variance for merging clusters A and B is:

$$\Delta(A, B) = \left[\frac{n_A n_B}{n_A + n_B} \right] x ||\mu_A - \mu_B||^2 \quad (S9)$$

where:

- n_A , n_B are cluster sizes

- μ_A , μ_B are cluster centroids (mean vectors)

Cut height selection

The dendrogram was cut at a height chosen by:

1. Visual inspection of the dendrogram (elbow/knee method)
2. Plot of within-cluster sum of squares vs. number of clusters

Number of clusters retained

Three clusters (river, transition, lake) for all periods, though internal heterogeneity was observed within the lake cluster.

Results are visualized in Figure 14 of the main text.

S1.9 Water Quality Index (WQI)

(Referenced in main text Section 2.4)

To synthesize six parameters into a single score comparable across sites and periods, following the Canadian Border Water Quality Index (CB-WQI) approach.

Parameter scoring (discretization)

Each parameter value $X_{i,p}$ was assigned a score based on thresholds from Table 1 of the main text:

$$C_{i,p} = \begin{cases} \text{Optimal}[100]; X_{i,p} \leq L_{p,opt} \\ \text{Suboptimal}[60]; L_{p,opt} < X_{i,p} \leq L_{p,sub} \\ \text{Critical}[20]; X_{i,p} > L_{p,sub} \end{cases} \quad (S10)$$

where $L_{p,opt}$ and $L_{p,sub}$ are thresholds defined in Table 1 for each water body type (river, transition, lake).

The reason behind 100-60-20 (Optimal-suboptimal-critical) scoring follows established simplified WQI approaches [1,2]. The three categories correspond directly to regulatory classifications ("good", "moderate", "poor" status under the EU Water Framework Directive) to reduce uncertainty arising from incomplete data, (ii) to avoid overinterpretation of small differences between measurements, (iii) to maintain consistency with regulatory thresholds of the EU Water Framework Directive, (iv) to facilitate aggregation of multiple parameters into a single index, and (v) to improve clarity in spatial and temporal interpretation of results¹, ensuring the robustness and interpretability of the index. While this loses fine-grained information, it provides management-relevant outputs and is robust to measurement noise. Limitations are discussed in Section 4 of the main text.

Aggregation (arithmetic mean)

$$CB - WQI_i = \frac{1}{n_i} \sum_{p=i}^{n_i} S \cdot I_{i,p} \quad (S11)$$

where n_i is the number of parameters available for site i (Minimum 5 of 6 required for calculation).

S1.10 Inverse Distance Weighting (IDW) interpolation

(Referenced in main text Section 2.4.2)

To generate continuous temperature and dissolved oxygen surfaces from point measurements at the 29 monitoring stations for spatial visualization (Figures 4, 5, and 10 of the main text).

Formula

The weight assigned to each monitoring point i is:

$$w_i = \frac{1}{d_i^p} \quad (S12)$$

where d_i is the distance between the monitoring site and the interpolated point (power parameter $p = 2$).

The interpolated temperature $T(x)$ at location x is:

$$T(x) = \frac{\sum_{i=1} w_i T_i}{\sum_{i=1} w_i} \quad (S13)$$

where T_i is the measured temperature at site i .

Implementation

IDW was performed using the `gstat` package in R and the IDW module in QGIS 3.34. The output grid resolution was set to 50 m × 50 m.

¹ Aggregation using the arithmetic mean (Equation 9) prevents overcompensation between parameters, unlike multiplicative indices or those based on Euclidean distances [3].

S1.11 Handling of missing data

(Referenced in main text Section 2.4)

Extent of missingness

The original dataset contained 4.7% missing values (523 out of 11,136 possible observations). Missing values were not randomly distributed: they were more frequent in winter months (December-February: 7.2% missing vs. 3.1% in summer) and in two lake stations (TA/4 and TK/3) during 2007-2010 (11.4% missing).

No imputation or interpolation was applied to avoid introducing artificial autocorrelation or bias.

S1.12 Software and reproducibility

All statistical analyses were performed using:

- RStudio 2024.12.0+467

Key packages with versions

- ggplot2_3.5.0 (visualization)
- dplyr_1.1.4 (data manipulation)
- tidyr_1.3.1 (data tidying)
- stats_4.3.2 (base statistical functions)
- factoextra_1.0.7 (PCA visualization)
- gstat_2.1-1 (IDW interpolation)

Reproducibility

The complete R script and anonymized data are available from the corresponding author upon reasonable request.

References cited in this Appendix

(Note: These references are also included in the main reference list of the article)

1. Chidiac, S.; El Najjar, P.; Ouaini, N.; El Rayess, Y.; El Azzi, D. A comprehensive review of water quality indices (WQIs): history, models, attempts and perspectives. *Rev. Environ. Sci. Bio/Technol.* 2023, 22, 349-395.
2. Sutadian, A.D.; Muttill, N.; Yilmaz, A.G.; Perera, B.J.C. Development of river water quality indices—a review. *Environ. Monit. Assess.* 2016, 188, 1-29.
3. Helsel, D.R.; Hirsch, R.M.; Ryberg, K.R.; Archfield, S.A.; Gilroy, E.J. *Statistical Methods in Water Resources*; U.S. Geological Survey Techniques and Methods, Book 4, Chapter A3; 2020.