A CONTEXT-AWARE FRAMEWORK FOR BIOLOGICAL HEALTH INDEX HARMONIZATION AND MACHINE LEARNING-AIDED HEALTH FORECASTING OF WATERBODIES *

Kazi Fouzia Akter Anita
Kingston University London
United Kingdom
anitafouzia104@gmail.com

Sakib Shahriar Shafin
Federation University Australia
Churchill, VIC
ss.shafin@federation.edu.au

Priyambada Joshi
Federation University Australia
Ararat, VIC
priyam.joshi@federation.edu.au

ABSTRACT

National biological monitoring repositories, such as the U.S. EPA's STORET/WQX, accumulate ecological health scores of water bodies from many agencies using heterogeneous biological indices and scoring conventions. Due to the different standards these agencies adhere to, these scores are not directly comparable. This difference in health score scales and benchmarks undermines large-scale ecological analysis and the application of machine learning models to forecast waterbody health. Existing ML studies predominantly model indices from a single standard framework or utilize physio/chemical parameters for analysis. Studies on composite water quality indices are rare, who also assume homogeneous response variables and regular data sampling. Yet a cursory look into STORET shows regular data sampling to assess the ecological health of waterbodies is incredibly rare. To develop a harmonization framework for multi-agency biological and habitat indices into a bounded site-level ecological health score, we propose Contextual Temporal Deficit Aggregation (CTDA) that retains assessment context and temporal history. CTDA orients scores, applies rank-based normalization within each index context, performs adaptive temporal smoothing based on the sampling interval, and accumulates persistent deficits before aggregating multiple biological contexts. On a curated national dataset of 17880 index records from 2500 monitoring locations and 38 organizations, CTDA achieved the strongest concordance with independent agency condition classes (Spearman $\rho = 0.844$), and more than doubled the separation between transient and persistent low-condition events. We then designed an ELM–LSTM hybrid model that combines contextual feature mapping with sequence modelling over CTDA trajectories to forecast health scores. Across rolling forecasting windows, ELM–LSTM outperformed all competing models, achieving the lowest MAE (0.109), RMSE (0.141) and R2 (0.759). By harmonizing diverse biological parameters and coupling them with context-aware sequence forecasting, this study provides a pathway for national-level ecological health trajectory analysis, which boosts targeted management and early warning systems.

Keywords Biological Indices · Waterbody · Index Harmonization · Machine Learning

1 Introduction

Freshwater ecosystems drive society's drinking water security, agricultural production, cultural values, and maintain biodiversity. The continuous degradation is now recognized as one of the most severe and least publicly appreciated environmental problems of the present century [1]. Biological communities integrate the effects of multiple physical and chemical stressors, and as such, biological monitoring is widely regarded as a more robust indicator of long-term ecological condition than snapshot chemical sampling alone [2, 3]. Multi-metric biological indices based on fish, macroinvertebrates, algae, or macrophytes are now central to regulatory bio-assessment frameworks in many regions, including the United States, Europe, and parts of Asia [4]. Large national data infrastructures now make it possible to

* *Citation*: **KFA Anita, SS Shafin, P Joshi**

revisit ecological conditions at much broader spatial and temporal scales than was previously practical. In the United States, the Water Quality Portal and the broader STORET/WQX ecosystem were created precisely to bring together water quality, biological, habitat, metrics, and index data submitted by many agencies and organizations, creating a powerful foundation for national-scale biological monitoring and trend analysis [5].

Yet this opportunity inherently contains a methodological problem that limits the potential of the data from being fully realizable. Biological assessment data in large repositories are not generated under a single index system [2]. Different agencies report different biological indices, define them within different institutional contexts, and apply scoring schemes that are often not directly comparable even when they refer to similar ecological constructs [3]. As a result, the same numerical field may contain scores from fish-based integrity indices, macro-invertebrate indices, habitat evaluation tools, and other assessment frameworks whose scales and meanings differ substantially. This creates a very specific gap for large-scale ecological analysis. Though repositories such as WQX provide access to extensive biological monitoring records, there is still no widely adopted workflow for converting multi-agency biological index scores into a common ecological health scale that remains faithful to local assessment context while being suitable for examination [6].

Recent work in predictive ecology and bio-assessment has demonstrated that machine learning (ML) can recover complex relationships between environmental drivers and biological indices [7]. In doing so, the interpretability of biological responses to individual environmental components can be better understood. Furthermore, ML can also unearth non-linear ecological responses and identify patterns that may be missed by conventional statistical analysis [8]. For national monitoring repositories such as STORET/WQX where the size and heterogeneity of the data often make manual inspection or site-specific modelling impractical, an appropriate ML model can be of great benefit.

However, the current use of ML in freshwater assessment has largely developed in a different data setting. Existing studies predict biological quality or ecosystem response within a single frame of reference, which may be an individual biological index framework or a dataset [9]. In such settings, the target variable is already coherent, and the modelling task is to explain or predict variation in that single index. But for large-scale repositories such as WQX, the reported ecological health scores are derived from multiple index systems, each defined within a specific organizational and methodological context. As a result, the numerical score recorded for one observation is not directly equivalent in meaning to the score recorded for another [10]. An ML model trained on these scores with any processing will create semantic inconsistency when generating insights from such a dataset. Before prediction or trajectory analysis can be carried out in a defensible way, the scores must first be transformed into a common ecological health representation that preserves local assessment context while making the response variable comparable across records [6].

From a thorough review of existing literature on this topic, we identify two gaps. Firstly, there is no widely adopted framework for converting multi-agency biological index scores into a unified ecological health scale suitable for contextual ML analysis. Existing national-scale integration methods mainly address data formatting and structural cleaning, and not resolving the ecological incompatibility of health scores assessed by different organizations with different standards [11]. As such, any large-scale trend analyses across available records require careful harmonization. However, once such a normalization is achieved, an ML-aided analysis becomes plausible, which can improve the health prediction of biological bodies including lakes: i) ecological health score prediction from organizational and temporal context, and ii) health trajectory prediction of ecological sites under irregular long-term sampling, the latter being a setting in which conventional forecasting methods are known to degrade as gaps and uneven intervals accumulate [12].

In this paper, we develop a Contextual Temporal Deficit Aggregation (CTDA) method to normalize heterogeneous biological index scores and derive a comparable measure of waterbody ecological health. CTDA applies deficit-sensitive aggregation within each index context and smooths successive observations to reduce the effect of isolated measurements and retain persistent changes in condition. The normalized scores are arranged into time-ordered site histories that represent the current health state of each waterbody and its change across sequential observations. An ELM-LSTM hybrid model is then used for prediction and forecasting. The ELM captures nonlinear relationships among the organizational, contextual, and temporal features while LSTM models dependencies across each site's observation sequence. The model consequently forecasts future health scores to identify site health trajectories from the harmonized historical scores. Therefore, this paper:

- Develops CTDA, a contextual aggregation algorithm that aligns and normalizes biological index scores within their assessment context, applies time-gap-aware smoothing to irregular site histories, and combines geometric health aggregation with a persistent-deficit penalty.

- Designs an ELM-LSTM forecasting model to combine nonlinear feature learning with sequence modelling to predict future waterbody ecological health from harmonized monitoring scores.

- Curates a temporally ordered STORET/WQX biological monitoring dataset by harmonizing multi-agency records and deriving contextual, temporal, and site-history features. The resulting dataset supports score-normalization tasks and broader machine learning analyses of waterbody health.

2 Related Works

The use of large ecological datasets has expanded the range of methods used to assess environmental health. Previous studies have investigated macroinvertebrate-based river assessment, composite indicators for ecological health, and ML-based approaches for predicting quality and health scores of waterbodies. This section reviews the relevant literature and identifies the research that provides the basis for our study.

2.1 Ecological bioassessment and biological indices

Biological monitoring programmes routinely use macroinvertebrate, fish and habitat indices to translate taxonomic observations into operational classifications of ecological condition. However, these indices are usually calibrated within specific regulatory frameworks and river types, and their scores cannot be assumed to represent comparable ecological health across agencies or regions. The following group of work examines how existing biotic indices perform in different contexts and how reliably they capture stress gradients and recovery.

Mureithi et al. [4] survey benthic macroinvertebrate bio-assessment initiatives across East African rivers and demonstrate that regional programs remain constrained by sparse baseline data and fragmented institutional responsibilities. They propose a stepwise pathway from national pilot studies through cross-country harmonization to Afrotropical policy frameworks, which highlights the dependence of bio-assessment on locally calibrated criteria and stakeholder input. However, the study lacks empirical testing of index behaviour or statistical evaluation, therefore highlighting the fact that many regions still rely on narrative frameworks rather than data-driven calibration. On the other hand, studies focusing on a single waterbody site offer more detailed assessments of how alternative benthic indices respond to pollution and habitat stress. Long-term macroinvertebrate and metal data from the Odiel Estuary were studied in [10] who shows that multimetric indices such as M-AMBI and BENFES track spatial and temporal gradients in metal contamination and taxonomic diversity more effectively. Their use of ordination and canonical correspondence analysis demonstrates that multi-indexing frameworks can detect improvement in impacted bodies, but also reveals that different indices deliver divergent classifications over the same monitoring record. Similarly, Majaneva et al. [13] compare the organic pollution oriented indices with the pressure-independent IBIBI scoring system in a hydropower-regulated Norwegian river and find that where pollution index reports “Good–High” condition, IBIBI classifies sites as “Bad–Moderate”. This discrepancy illustrates how indices that are tuned to specific stressors can misrepresent ecological status when applied beyond their original design.

Additional case studies such as [14, 15] reinforce the need for context-aware index evaluation. [14] show that, in a Turkish stream, pollution and pressure-based scoring systems correspond better to physicochemical gradients than others, despite all being drawn from established European frameworks. Nahli et al. [15]apply Iberian and Belgian indices to wastewater-impacted Moroccan waterbodies and infer resilience from increased macroinvertebrate data and improved index scores compared to historical descriptions. However, this is done using a single sampling period and imported index thresholds. Taken together, these studies demonstrate that biotic indices can be informative about spatial gradients and recovery, yet they remain tightly bound to their regional calibration. None of them attempt to reconcile scores produced by different agencies for different biological groups.

2.2 Composite indicators and indicator aggregation

Composite indicators aggregate multiple measurements into a single index intended to represent stream or river health. But their construction is complicated by substantive choices about indicator orientation and scaling. Those choices determine if high values in one aspect can offset poor values in another and how uncertainty propagates into the final score.

Zhang et al. [16] develop a composite river-health model for the Huai River Basin by integrating hydromorphological attributes, chemical parameters and biotic diversity, and weighting them using chaotic particle-swarm optimization. Their fuzzy approach shows how data-driven weights can reduce reliance on purely expert judgment while retaining a multi-domain representation of condition. The study in [6] finds that for a macroinvertebrate biotic index in the Yangtze River, continuous scores only become operationally meaningful when converted into manually defined quality classes. They derive five biological water-quality classes by using the 75th percentile of values at least-disturbed waterbodies to set the "good" boundary and partitioning the remaining range. Both studies emphasise that composite and biotic indices require explicit, context-dependent benchmarks and cannot be interpreted without reference to local data distributions.

Methodological analysis of composite indices further clarify the implications of weighting and aggregation choices. [17] outline stages for constructing health indices for streams that include theoretical framing, indicator standardization, weighting, and then aggregation. They find that additive aggregation is highly dominant, allowing strong performance in one indicator to mask persistent deficits in another. However, geometric aggregation is less compensatory and shows more focus on poorly performing components. Using simulated monitoring datasets, the authors find that indices built from the full set of available indicators tend to have lower bias than those based on arbitrary subsets. [18] extend this argument by demonstrating that nominal weights rarely correspond to appropriate influence on the composite. Correlated indicators can dominate the index even under equal weighting, while distinctive indicators may become numerically marginal. The nonlinear sensitivity index and optimization strategies implemented here show that indicator correlations must be considered when designing weight structures. Otherwise, a supposedly neutral composite can inadvertently emphasize redundant information.

In parallel, an instructive cross-index comparison is performed in [10] by examining how six benthic biotic indices measure long-term ecological quality in a polluted estuary. They find that M-AMBI and BENFES are markedly more sensitive to recovery and diversity than AMBI or BENTIX. The insight underlines that aggregation schemes and metric selection change ecological inferences and that no single index is universally reliable across contexts. For large repositories, this implies that biological scores must be understood within their own assessment frameworks before any attempt is made to combine them. From these studies, it is clear that composite health indices require accurate standardization, benchmarking, and transparent aggregation. However, existing approaches generally assume a shared indicator framework or a single parameter index. They do not address the challenge of combining already established scores produced by multiple agencies.

2.3 Machine learning for water quality and biological indices

ML methods have been widely applied to water quality prediction and classification using physical/chemical variables and composite indices such as WQI as targets. These studies demonstrate the utility of sequence models, ensemble classifiers and Deep Learning (DL) models for dense monitoring records. However, they generally operate on homogeneous index definitions and assume regular sampling schedules. Wang et al. [19] combine entropy weighting and Pearson correlation for feature selection, followed by LSTM modelling to achieve accurate short-term predictions of dissolved oxygen (DO) and nutrient concentrations at a single waterbody site. The model achieves high R2 values across many experiments, confirming that recurrent models can capture local temporal dependence in regularly sampled data of this kind. Tree-based models such as decision trees, random forest (RF) and deep cascade forest were evaluated in [20] to reproduce national regulatory water quality classes from weekly measurements of pH, dissolved oxygen and many others. Across 124 stations, the combined model obtains near-perfect F1 scores for the most degraded classes even with fewer samples. Similarly, [21] report high accuracies when classifying Water Quality Index (WQI) classes using multinomial logistic regression and tree-based models. In each case, however, the response classes are derived directly from the same variables supplied to the models, so the tasks primarily replicate deterministic index calculations rather than forecasting harmonized ecological condition independently.

Studies also aim to reduce reliance on broad parameter panels or reconstruct incomplete records. Such as, [22] design an oxygen-focused support vector regression (SVR) "Super Model" to predict WQIs from small sets of oxygen-related variables and apply SHAP to interpret feature contributions across multiple Iranian rivers. Often times broad datasets on ecological indicators will have incomplete records, which is another arena of interest. Rodríguez et al. [23] compare a wide range of statistical and ML algorithms for imputing highly incomplete water quality records in a Uruguayan river. The usage of inverse distance weighting seemed to achieve accurate imputation performance across variables with 50–70% missing data.

In the aforementioned studies, ML models are deployed as standalone mechanisms that operate on singular sites or moderately numbered multi-site datasets using regularly spaced measurements. The response variables are either chemical concentrations, regulatory classes or composite indices constructed directly from the predictors. None of them combine complementary algorithms to separately address contextual nonlinear feature representation and sequence forecasting for health scores. This gap motivates the hybrid ELM–LSTM design used in our work.

3 Data Source, Study Population, and Curation

Biological and habitat index records of waterbodies were obtained through the U.S. Environmental Protection Agency's Water Quality Exchange and Water Quality Portal. The following sections define the study population and explain the curation process utilized to preserve assessment context for the CTDA methodology.

3.1 Data infrastructure and source extraction

This study assembled a national-scale biological and habitat index dataset from the United States Environmental Protection Agency's Storage and Retrieval Data Warehouse (STORET) via the Water Quality Exchange (WQX) infrastructure and the public Water Quality Portal (WQP). WQX provides a common exchange schema through which federal and non-governmental monitoring organizations record ambient quality information of waterbodies. The components include a diverse set, encompassing chemical components and taxonomic observations that are merged into biological index scores. Each record in the portal represents a biological or habitat assessment score reported by the respective organization.

Data collection was restricted to biological and habitat index records for monitoring locations in the United States. All records returned by this query were retained in the raw extract, with subsequent screening for numerical and temporal analyses described in Sections 4.2 and 4.3. The recorded assessment dates ranged from 19 August 1974 to 6 November 2025. 17880 index records were extracted with 18 parameters, which, over the years, have been monitored by 38 distinct agencies. The list of authorities include state environmental protection agencies (e.g. Texas Commission on Environmental Quality, Ohio EPA), volunteer monitoring programmes (e.g. NJDEP AmeriCorps), and state natural resources departments. The extract covered 2500 distinct monitoring locations with 64 biological or habitat index types reported within 26 assessment contexts. Assessment dates ranged from 19 August 1974 to 6 November 2025, spanning more than five decades of opportunistic biological monitoring across aquatic scenarios.

Each record in the collected dataset represents a biological or habitat assessment reported for a particular monitoring location on a particular date. Alongside the numerical score, the record retains information about the reporting organization, the type of index applied, the assessment method, and any associated qualifiers or comments. An assessment context here is the specific reporting framework under which a score was produced. It is defined by the reporting organization together with the index type and its associated context, because similarly named indices may be implemented differently across agencies. Let i denote a monitoring location, o a reporting organization, c an assessment context, m an index type, and t the assessment date. A record is then represented as:

$$O_{i,c,m,t} = (x_{i,c,m,t}, a_i, o_i, r_{c,m}, q_{i,c,m,t}) \tag{1}$$

where $x_{i,c,m,t}$ is the reported numerical index score, a_i identifies the monitoring location, o_i identifies the reporting organization, $r_{c,m}$ contains the available index-method metadata, and $q_{i,c,m,t}$ contains score qualifiers and narrative comments. This structure therefore separates each reported score from the assessment framework that produced it. Retaining this information will support appropriate score harmonization using CTDA and identification of records that may not be directly comparable.

3.2 Study-population definition and geographic attribution

The collected record set from the STORET portal identifies the reporting organization and monitoring location associated with each assessment but does not provide additional details about the waterbody. Geographic and hydrological attributes were therefore obtained by extracting site metadata from the portal and linking each location to its biological index records. The linkage used the location as the common key, with the portal site identifier checked against the identifiers retained in the biological metric dataset. The site metadata provided: i) the geographic coordinates, ii) reported waterbody type, iii) hydrological unit information, and other administrative attributes for each location. These data were used to describe the distribution of monitoring sites across geographic and hydrological settings. Finally, the waterbody type was obtained from the explicit site classification reported by the portal.

The study population includes every monitoring location represented in the collected dataset for which a biological or habitat index score was reported. It encompasses rivers and streams, lakes, Great Lakes sites, canals, estuaries, springs, reservoirs, and wetlands. Waterbody type was retained as descriptive metadata. Records that could not be linked to waterbody type metadata were retained and flagged as "missing type attribution", without inferring the missing classification. The normalized score obtained from the CTDA methodology therefore represents a relative ecological condition of a site given the metrics reported and agency-inferred score, rather than being an absolute measure of ecological health.

3.3 Index diversity and semantic non-equivalence

The collected dataset accumulates scores generated by many biological and habitat assessment initiatives that consider very different sets of parameters. Separate agencies evaluate different components of aquatic ecosystems and are designed for different regulatory and maintenance processes. For example, some records include fish or macroinvertebrate indices, periphyton assessments, while others focus on vegetation quantities, chemical composition, and others.

Table 1: Fields retained from the raw WQX biological-index extract and their role in the analysis.

Attribute group	Retained source fields	Role in the analysis
Data provenance	`OrganizationIdentifier`, `ProviderName`	Identifies the organisation that submitted each record and its data provider. These fields were retained to characterise variation among reporting programmes and to support organisation-level validation.
Monitoring-location identity	`MonitoringLocationIdentifier`	Provides the unique site identifier used to assemble site histories. It was also used to ensure that records from the same site were not divided between training and test data.
Index identity and assessment context	`IndexIdentifier`, `IndexTypeIdentifier`, `IndexTypeIdentifierContext`, `IndexTypeName`	Identifies the biological index, its reported type, and the assessment context in which it was calculated. Together, these fields define the context-specific score distributions used in CTDA.
Score and accompanying interpretation	`IndexScoreNumeric`, `IndexTypeScaleText`, `IndexQualifierCode`, `IndexCommentText`	Contains the reported numerical score, stated scale where available, qualification codes, and descriptive comments. The fields were used to identify non-numeric placeholders, incompatible scoring directions or ranges, and records subject to method-specific limitations.
Assessment time	`IndexCalculatedDate`	Supplies the date used to order observations within each site history and to calculate the time interval between successive assessments for gap-aware state updating.
Method documentation	`ResourceTitleName`, `ResourceCreatorName`, `ResourceSubjectText`, `ResourcePublisherName`, `ResourceDate`, `ResourceIdentifier`	Links records to available documentation for the underlying assessment method. These fields supported manual checking of frequently reported indices and contexts for which index identity or interpretation was unclear.

A numerical index score therefore has meaning only within the assessment framework that produced it. An index may be based on a particular taxonomic group or calibrated against local reference sites, and interpreted using thresholds defined by the responsible agency. Higher values may indicate better condition for one index, whereas the opposite may apply to another. Scores may also differ in their theoretical range or be reported without a stated scale. Consequently, two records with the same numerical value cannot be assumed to represent the same ecological condition. This issue is critical in a national dataset. Direct or almost blind aggregation of all raw scores would mix ecological variation with differences in index design. Any apparent national pattern could then reflect reporting practices as much as differences in ecological condition. To retain this information, each score was interpreted within the context of when the record was measured. An assessment context variable c keeps track of this during aggregation, defined as the combination of the reporting organisation, the index type, and the associated index context:

$$c = (\text{organization}, \text{index context}, \text{index type})\,. \tag{2}$$

The reporting organization was included as an index label alone does not necessarily identify a common assessment procedure for others. For example, a score of 40 from one agency's fish index was not assumed to have the same meaning as a score of 40 from another agency's habitat or macroinvertebrate assessment. Several assessments were also reported for the same location and date. Such records provide complementary evidence rather than duplicate measurements. Each of the fish index, the benthic macroinvertebrate index, and the habitat assessment describes a different aspect of the same site. These records were retained separately during preparation and were only combined through our proposed CTDA method, which aims to preserve the identity and interpretation of each index while providing a defensible basis for deriving comparable indicators of relative ecological condition.

3.4 Data screening and construction of final dataset

To create the dataset for CTDA assessment and subsequent forecasting analysis, the collected records underwent multiple stages of review. A preliminary review kept the imported records file and the waterbody site metadata file from the WQX portal intact, and then we considered the integrity of site and organisation identifiers, the availability of the dates and index parameters, and linkage to the waterbody metadata. Each source record was assigned a row reference before any processing was undertaken. Derived tables were created separately, and the source reference was retained throughout. Row counts were recorded before and after every operation. Thus, where a record could not contribute to a particular temporal or numerical procedure, it remained in the original dataset with the additional tags.

Table 2: Raw data quality of the collected WQX records

Audit item	Count	Proportion (%)
Raw metric–index records	17,880	100.00
Records with a numerical score	17,536	98.08
Records without numerical score	344	1.92
Records with valid calculated date	15,655	87.56
Records without calculated date	2,225	12.44
Records with a valid numerical score and calculated date	15,505	86.72
Records with declared score scale	1,831	10.24
Records without declared score scale	16,049	89.76
Records containing comments	8,133	45.49
Negative numerical scores	193	1.08
Zero numerical scores	478	2.67
Exact duplicate source records	0	0.00

The records were then processed by rows to be placed within a temporal sequence for applicability into the aggregation and forecasting tasks. Assessment dates were converted to a consistent date format and checked for parsing errors. Records without a usable date were retained but marked as unavailable for chronological analyses. Similarly, index scores were checked for numerical representation. Empty entries were recorded as unavailable for numerical processing, while non-numeric values were label coded accordingly.

The records were then sorted by monitoring location, index assessment context c, and date. We examined repeated records using the organization, location, index context, index type, date, and reported score. However, from our row-by-row analysis of the 17800 records, we did not find any duplicated records of a waterbody measurement taken at the exact same date that measures the exact same parameters. For same day records of a certain waterbody with different parameters, values were represented by their median score so that repeated entries did not receive disproportionate influence during temporal smoothing. As individual sites are monitored by one agency at a time, creating the median did not require any cross-agency averaging. The site metadata were then linked to the biological metric records using the monitoring/location identifier. The linkage was checked by comparing the number of distinct site identifiers before and after the merge and by reviewing unmatched identifiers. Geographic and waterbody descriptors were added where available.

After these processes, the final dataset was accumulated for score harmonization and temporal analysis. No biological context was inferred where it had not been measured, and no missing score was imputed. The resulting analytical dataset therefore retains the traceability of the original records while providing a consistent basis for the subsequent CTDA calculations and forecasting analyses.

4 Methodology of this research

This section presents the framework developed to transform heterogeneous WQX index records into temporally ordered, comparable estimates of waterbody ecological health. The framework first preserves the assessment context associated with each reported score, then applies Contextual Temporal Deficit Aggregation (CTDA) to orient, normalize and aggregate indices across irregular sampling intervals. The resulting health sequences of sites are subsequently used to train an ELM-LSTM model for forecasting future ecological condition and identifying changes in site health over time.

4.1 Ecological bioassessment and biological indices

Biological monitoring agencies report a diverse set of ecological indices such as fish integrity scores, benthic macroinvertebrate indices, habitat quality scores, and lake vegetation indices under their own scoring schemes and regulatory contexts. Let $x_{i,c,t}$ denote the raw biological index score recorded at monitoring location i, under assessment context c, at time t. The assessment context encompasses both the formal index type and its reporting framework, c, so that all observations belonging to the same c share a common scoring standard and scale.

For a given waterbody i, the STORET/WQX biological metric table thus contains a set of irregular, multi-index time series $\{x_{i,c,t}\}_{c,t}$ rather than a single coherent system. Direct application of ML to these raw scores would convolute their meaning and undermine any attempt at national-scale health analysis. The first objective of this study is thus to construct a harmonized, context-aware ecological health score for each site available in the STORET/WQX dataset. These scores are subsequently used to forecast future ecological health and classify each site's temporal trajectory, which will allow overseeing agencies to take actions to improve their conditions.

To address objective 1, we introduce the CTDA algorithm. For forecasting and classification tasks, we design an ELM–LSTM model that combines fast nonlinear feature learning of ELM with sequence modelling over CTDA-derived scores using LSTM. Recent studies [24, 25, 26] demonstrate the efficacy of combining multiple DL models to address separate needs of a water quality forecasting problem instead of a single model's general usage, which inspires our approach.

4.2 Contextual Temporal Deficit Aggregation (CTDA)

The CTDA algorithm is constructed in three stages: i) context-specific score orientation and normalization, ii) data-adaptive temporal smoothing with contextual deficit memory, and iii) aggregation of available biological contexts into a site-level health score.

4.2.1 Context-specific normalization

Before aggregation, all index scores within each context c are oriented such that larger values consistently indicate better ecological condition. For any index whose original direction is reversed, the score is transformed in reverse before subsequent processing. Let $r_{i,c,t}$ be the ascending rank of $x_{i,c,t}$ among the n_c observations belonging to context c. The normalized health fraction is defined using the adjusted empirical distribution

$$h_{i,c,t} = \frac{r_{i,c,t}}{n_c + 1}. \tag{3}$$

This transformation maps every observation to the open interval (0,1) while preserving the original score ordering. The fraction $h_{i,c,t}$ expresses the score's relative position among observations in that context. It does not, by itself, indicate whether the waterbody meets an ecological health standard. Invariance to monotonic rescaling ensures that CTDA is unaffected by differences in score ranges or units across agencies. For example, converting an index from a 0–100 scale to a 0–20 scale does not alter its normalized health position.

4.2.2 Data-driven temporal smoothing

Biological observations in STORET/WQX are irregularly distributed in time. A site may contain observations separated by several weeks during one monitoring period and by several years during another. A fixed smoothing coefficient would therefore impose an external temporal assumption that may not be appropriate for all index systems.

For each context c, let $\widetilde{\Delta}_c$ denote the median of all positive inter-observation intervals recorded in that context. This quantity is estimated directly from the data and represents the typical monitoring interval associated with the given assessment framework. The gap-adaptive update weight is then defined as

$$\alpha_{i,c,t} = \frac{\Delta t_{i,c,t}}{\Delta t_{i,c,t} + \widetilde{\Delta}_c}, \tag{4}$$

where $\Delta t_{i,c,t}$ is the elapsed time between the current and preceding observations at a site i, under context c. When the elapsed interval is typical for that context, the current and previous observations receive equal influence (0.5). More closely spaced observations retain more of the preceding state, whereas observations separated by longer gaps place greater emphasis on the newly observed condition.

The temporally smoothed health state is calculated through geometric recursion:

$$s_{i,c,t} = h_{i,c,t}^{\alpha_{i,c,t}} s_{i,c,t-1}^{1-\alpha_{i,c,t}}, \tag{5}$$

with $s_{i,c,1} = h_{i,c,1}$. The geometric recursion keeps the normalized health state within its defined range while giving persistent low-condition observations greater influence than isolated high-condition observations. As a result, occasional improvements do not completely mask an extended period of poor ecological condition.

4.2.3 Contextual deficit memory

The degradation of a waterbody is represented in CTDA with a deficit score. From a standard reference, the deficit indicates the extent of a health score falling below the context-specific reference. We construct this score to include as much scenario details as possible, to deviate from being just a statistical outlier.

If a responsible agency provides a documented biological-condition boundary g_c, such as a threshold for acceptable, good, or impaired condition, it is converted to the common relative health scale using the empirical distribution function

$F_c(\cdot)$. However, from our research we find that such thresholds are not consistently published or directly recoverable from STORET/WQX metadata. For contexts without a verifiable boundary, a contextual median is used as a neutral reference point. Accordingly, the contextual reference fraction q_c is defined as

$$q_c = \begin{cases} F_c(g_c), & \text{if a boundary } g_c \text{ is available,} \\ 0.5, & \text{otherwise.} \end{cases} \tag{6}$$

The instantaneous deficit is then calculated as

$$d_{i,c,t} = \max\left(0, q_c - s_{i,c,t}\right). \tag{7}$$

Thus, $d_{i,c,t} = 0$ when the smoothed health state meets or exceeds the reference, while positive values show the magnitude of condition below that reference. Where documented criteria exist, the reference retains the agency's ecological interpretation; otherwise, CTDA describes relative condition within the observed distribution of that assessment context.

To retain evidence of persistent poor condition, the deficit burden is updated using the same gap-adaptive weight derived for temporal smoothing:

$$B_{i,c,t} = \alpha_{i,c,t} d_{i,c,t} + \left(1 - \alpha_{i,c,t}\right) B_{i,c,t-1}, \tag{8}$$

where $B_{i,c,1} = d_{i,c,1}$. Repeated observations below the reference increase the accumulated deficit. Later observations at or above the reference reduce the effect of earlier deficits, with longer monitoring gaps giving greater weight to the new observation.

4.2.4 Final CTDA score

The final CTDA score D combines the temporally smoothed health state with the normalized persistent deficit:

$$D_{i,c,t} = s_{i,c,t}\left(1 - \frac{B_{i,c,t}}{q_c}\right). \tag{9}$$

As $0 \leq B_{i,c,t} \leq q_c$, the final score stays confined within $[0, 1]$. A site with no accumulated deficit retains its smoothed health state, whereas repeated observations below the contextual reference progressively reduce the CTDA score.

However, a given site and assessment date, several biological contexts such as fish, benthic, habitat, or vegetation indices may be available. These context-level CTDA values should be combined into one site-level health score $D_i(t)$:

$$\mathrm{D}_{i,t}^{\mathrm{site}} = \left(\prod_{c \in \mathcal{C}_{i,t}} \mathrm{D}_{i,c,t}\right)^{1/|\mathcal{C}_{i,t}|}, \tag{10}$$

where $\mathcal{C}_{i,t}$ is the set of assessment contexts recorded for site i at time t. The geometric mean retains the bounded scale of the component scores and prevents high condition in one available biological context from fully offsetting poor condition in another.

Each parameter within the CTDA methodology is derived from the observed monitoring record or obtained from published regulatory guidance. Due to the availability of historical monitoring data, a data-derived methodology avoids imposing any fixed assumptions about the waterbody which may undermine the accuracy of the score. Instead, each site's data are considered to build their score to create an accurate representation.

4.3 Combining ELM and LSTM for Forecasting and Classification of Waterbody Health

After the context-specific CTDA scores have been aggregated into singular scores for each site $D_{i,t}$, an ELM-LSTM model is used for subsequent forecasting. Concurrent literature shows that hybrid architectures consistently outperform single model methods in water quality prediction [24, 27]. Usage of different models to address specific obstacles and issues in the given scenarios yields better results than employing a general model as an all-purpose solution [28].

4.3.1 Extreme Learning Machine for contextual feature mapping

ELM is a single-hidden-layer feedforward network in which the input weights and hidden-layer biases are randomly assigned and fixed, while the output weights are estimated through a closed-form linear solution [29]. This structure allows ELM to learn nonlinear relations from diverse data representations with substantially lower training cost than iteratively trained neural networks. The model avoids repeated gradient optimization, which reduces the chances of overfitting. ELM has consequently been widely applied to environmental monitoring and water quality prediction problems [30].

For each site i at time t, the ELM receives an input vector $\mathbf{z}_{i,t}$ comprising organizational, contextual, temporal, and historical health information. Categorical variables are encoded before model fitting, while numerical variables are scaled using statistics estimated from the training data only. The ELM hidden layer maps $\mathbf{z}_{i,t}$ to a nonlinear representation $\tilde{\mathbf{z}}_{i,t}$. This captures interactions between assessment context, monitoring time, and recent ecological condition that may not be expressed through the original features alone [a]. The resulting representations are ordered by time within each site and provided to the LSTM component.

4.3.2 LSTM for site-history modelling

Long Short-Term Memory (LSTM) networks are recurrent neural networks designed to retain and update information across sequential observations through gated memory cells [31, 32]. Their ability to model temporal dependence has increased their adoption in water-quality prediction, where environmental processes are nonlinear, and measurements can be non-stationary [33].

For each site i, the temporally ordered sequence

$$\{\tilde{\mathbf{z}}_{i,1}, \tilde{\mathbf{z}}_{i,2}, \ldots, \tilde{\mathbf{z}}_{i,T_i}\} \tag{11}$$

is supplied to the LSTM together with the corresponding site health history $\{D_{i,t}\}_{t=1}^{T_i}$. The monitoring interval is retained as an input feature so that the model can account for irregular observation times rather than treating all consecutive records as equally separated. For forecasting, the LSTM estimates the health score at the third sequential monitoring state. The predicted score is evaluated against the observed value $D_{i,t+3}$ using regression measures such as mean absolute error, root mean squared error, and the coefficient of determination.

4.3.3 Forecasting dataset construction and evaluation protocol

Following the CTDA score aggregation process, each site was represented by a sequence of harmonized health scores, $\{D_{i,t}\}_{t=1}^{T_i}$. The site records were individualized into mini-subsets for the forecasting experiment. Forecasting was formulated as a continuous regression problem using rolling windows of at least four consecutive assessments. For a window ending at time t, the model predicted the CTDA score at the third subsequent monitoring event, $D_{i,t+3}$. Each input assessment was represented by four features: i) the CTDA score, ii) contextual-rank score, iii) adaptive weight $\alpha_{i,t}$, and iv) the log-transformed interval since the preceding assessment. However, windows were excluded when the target assessment occurred more than five years after the previous input observation, as such a gap will not produce informed forecasts. This restriction also avoided treating long gaps in monitoring as short-term forecast horizons.

Samples were divided chronologically by target date into training, validation, and test sets (60%, 20%, and 20%, respectively). Predictor standardization was fitted using the training data and then applied to the validation and test data. The ELM-LSTM model received the four-observation sequence and estimated $\hat{D}_{i,t+3}$. Performance was compared with persistence, standalone LSTM, Random Forest (RF), Histogram-based Gradient Boosting (HGB), and Deep Neural Network (DNN) models. The persistence model used the most recent score as the forecast, $\hat{D}_{i,t+3} = D_{i,t}$. Ridge regression and Random Forest were fitted to flattened windows. The ELM used 256 hidden units, while the LSTM and ELM-LSTM models used two recurrent layers with 64 and 32 units. Forecast accuracy was evaluated on the test set using MAE, RMSE, and R^2, where lower MAE and RMSE and higher R^2 indicate better performance. The baselines were selected to compare persistence, static nonlinear regression, and recurrent sequence modelling, rather than to exhaust all available algorithms. Standard Naïve Bayes was not included because the target is continuous and Support Vector Classifiers can struggle with sparse site histories with irregular intervals, as they are more suited to high-volume data [34].

5 Performance Evaluation

This section evaluates the proposed framework in two stages. First, CTDA is assessed as a score harmonization method by examining its agreement with agency-reported ecological condition classes collected in the dataset, followed by its ability to distinguish isolated low-condition observations of waterbodies from persistent deficits. Second, the resulting site-level health scores are used in a chronologically separated forecasting experiment to assess whether the ELM-LSTM model improves prediction of subsequent ecological condition over conventional methods.

5.1 Score Aggregation Evaluation

This subsection evaluates the efficacy of our proposed CTDA algorithm on creating a useful ecological health representation after biological records from different waterbodies situated in the United States have been harmonized. Two experiments were conducted. First, the derived scores were compared against pre-existing condition classes available for 1,785 datapoints, provided by the responsible monitoring agency. Second, their ability to distinguish transient degradation events from persistent deficits was examined across 430 transient and 3,806 persistent events. We compare the CTDA's performance with contextual rank (CR) normalization, an arithmetic temporal aggregation, and a geometric temporal aggregation method.

5.1.1 Concordance With Agency Condition Classes

The datapoints with agency classifications (described ecological condition as *Very Poor*, *Poor*, *Fair*, *Good*, or *Excellent*) were collected into a separate dataset for the first experiment. Due to the scarcity of references for comparing performance, these labels were used exclusively as an external validation criterion. They are treated as target labels and were not used during score computation. Since the *Very Poor* class contained only seven records, it was combined with *Poor* for class-level interpretation. Spearman's rank correlation was computed to measure agreement between each continuous score and the reference condition, while the widely used Kruskal-Wallis statistic assessed separation among class-specific score distributions [35]. The hypothesis is that, after harmonizing scores across different scales and assessment contexts, CTDA may accurately match the condition classifications assigned by the contributing agencies. If it is able to achieve that, then its derived scores are expected to be reliable for unlabeled records that cannot be directly validated.

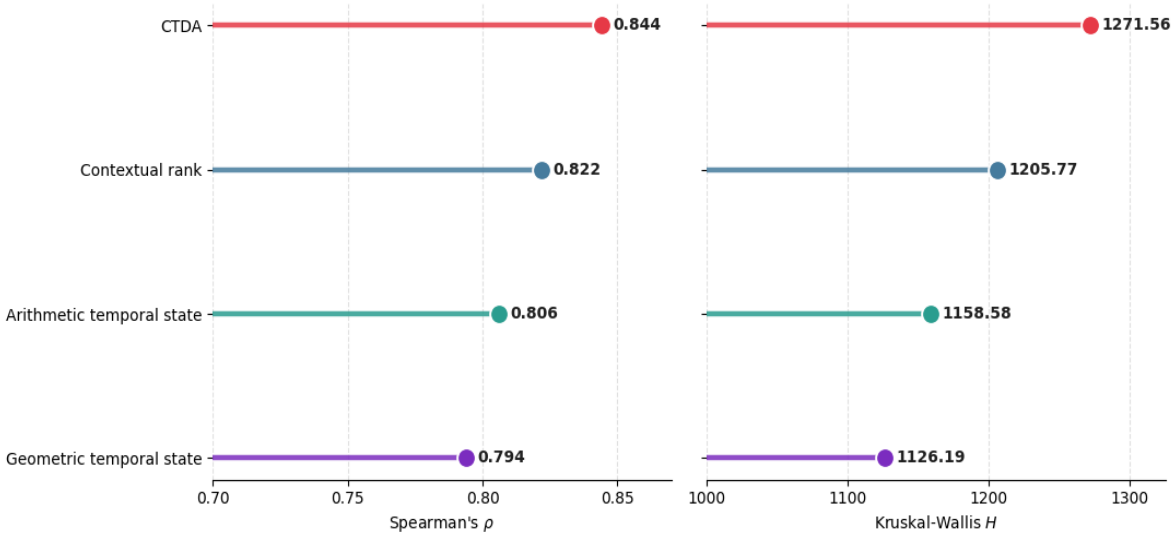

Figure 1: Agency condition agreement of evaluated aggregation methods.

As shown in Fig. 1, CTDA achieved the highest concordance with the agency classes ($\rho = 0.844$), exceeding contextual rank and the two temporal aggregation methods by 0.022, 0.038, and 0.05 respectively. The result indicates that CTDA preserves the non-persistent ecological information contained in agency-specific biological assessments while adding temporal information absent from rank-wise normalization. Especially the CTDA medians increased from 0.044 for the pooled Poor/Very Poor class to 0.201 for Fair, 0.590 for Good, and 0.878 for Excellent. Thus, incorporating recent data samples into CTDA scoring avoided the excessive attenuation produced by the preliminary deficit-only formulation,

while maintaining a clear ordinal relationship with external ecological evidence. Contextual rank stands second best from deriving score solely from the biological score reported at the same assessment event without including temporal trends. As such, CTDA's performance evidences the need to integrate both context and historical data to create a standard score.

5.1.2 Persistent Deficit Discrimination

The second set of experiment examined whether the aggregation methods could separate a short-lived decline in ecological condition from a persistent decline of a site's ecological index. For each site and biological index context, a low score was treated as *transient* when the preceding and subsequent data indicated acceptable condition, i.e., improvement. In contrast, a low score was treated as *persistent* when it followed an earlier low assessment at the same site. This comparison is important because a single unfavorable datapoint may reflect brief decline due to one-off events, whereas repeated low results provide stronger evidence of sustained ecological degradation.

Table 3: Discrimination between transient and persistent low-condition events.

Method	Transient median	Persistent median	Difference
Contextual rank	0.361	0.246	-0.115
Arithmetic temporal state	0.489	0.251	-0.238
Geometric temporal state	0.467	0.230	-0.238
CTDA	0.419	0.179	-0.240

Table 3 shows that all temporally informed methods assigned lower scores to persistent than to transient adverse events. CTDA produced a median score of 0.419 for transient events, which fell to 0.179 when ecological indices kept falling (a reduction of 0.240). This was more than twice the reduction obtained from contextual rank alone (0.115). As CR does not utilize any temporal trend, the results prove that a score based only on the current assessment cannot adequately identify isolated deterioration from an ongoing problem.

The arithmetic and geometric models produced reductions of 0.238, closely matching CTDA. CTDA nevertheless remains the strongest contender for the present application because it achieved the strongest agreement with agency condition classes (Section 5.1) while retaining this clear response to persistent poor condition.

5.2 Ecological Health Score Forecasting

Following the CTDA aggregation experiment and its evaluation with competing methods, these scores were utilized to predict future states of the ecological sites. Forecasting is framed as a continuous regression task where a model can analyze and preserve the temporal trends of a site's ecology. We used the rolling window dataset described in Section 4.3.3, where each sample contained four consecutive CTDA assessments. Our proposed ELM-LSTM model was compared with: i) the persistence model, which is a naïve baseline assuming the most recent score remains unchanged, ii) two ensemble models, Random Forest (RF) and Histogram Gradient Boosting (HGB), iii) a standalone LSTM, and iv) a general Deep Neural Network (DNN) model. All models used the same input features: CTDA score, contextual-rank score, gap-adaptive weight, and log-transformed inter-assessment interval.

Performance was assessed using MAE, RMSE, R^2, and mean signed bias (MSB), calculated as $\text{Bias} = \frac{1}{n}\sum_{j=1}^{n}(\hat{y}_j - y_j)$ [36]. Lower MAE and RMSE, together with higher R^2, indicate improved forecast accuracy. As for MSB, the target is to achieve neutrality. Positive bias indicates systematic over-prediction, whereas negative bias indicates under-prediction.

Table 4 presents the out-of-sample forecasting performance of all models. The persistence benchmark achieved a robust MAE = 0.116 and $R^2 = 0.698$ which indicates there are some levels of continuity between successive CTDA health scores. This is expected as CTDA incorporates a temporally smoothed ecological state with normalized context. However, the proposed ELM-LSTM achieved the best performance across all primary metrics, with MAE = 0.109, RMSE = 0.141, $R^2 = 0.759$, and near-zero bias (0.003). Relative to persistence, our model shows a 6.0% reduction in MAE and a 9.6% reduction in RMSE. Thus, the improvement reflects more than simply predicting a higher or lower average score.

The difference between the ELM-LSTM and the standalone LSTM is noticeable. While the LSTM improved RMSE from 0.156 to 0.144 and increased R^2 to 0.746, its MAE of 0.115 remained close to that of the persistence benchmark. The LSTM's temporal recurrence is observed to reduce larger errors. However, the ELM transformation boosts predictions of typical site-level changes, thereby allowing the hybrid model to distinguish observations with similar current CTDA scores but different contextual ranks or inter-assessment intervals. These distinctions are highly needed because the same current score can represent stable condition, recovery, or decline depending on its preceding sequence.

Table 4: Predictive performance of the evaluated forecasting models.

Model	MAE ↓	RMSE ↓	R^2 ↑	Bias
Persistence	0.116	0.156	0.698	0.006
LSTM	0.115	0.144	0.746	0.011
ELM–LSTM	**0.109**	**0.141**	**0.759**	**0.003**
RF	0.116	0.151	0.716	−0.012
HGB	0.121	0.156	0.698	−0.025
DNN	0.249	0.285	−0.006	−0.023

The ensemble models did not produce the same improvement. RF produced an RMSE of 0.151 and R^2 of 0.716. It has almost double the bias of the benchmark, which suggests the model captured some non-linear associations between the features. HGB performed less favorably, with more than double that of RF (-0.025). The model particularly has worsening performance where recovery or upward transitions occur. Static boosting models are usually limited in representing ordered temporal dependence when the same observed score can imply different future states depending on the preceding trajectory. For these exact scenarios, our ELM-LSTM architecture was chosen. The DNN produced the weakest result, with an MAE of 0.249, RMSE of 0.285, and a slightly negative R^2 of -0.006. The negative R^2 indicates that, on the held-out period, the DNN did not improve upon predicting a constant mean score. Even though DNN receives the same instantaneous CTDA-derived predictors as the other models, it has no recurrent mechanism through which to preserve the sequence of previous ecological states. Consequently, it cannot distinguish a currently low score following gradual decline.

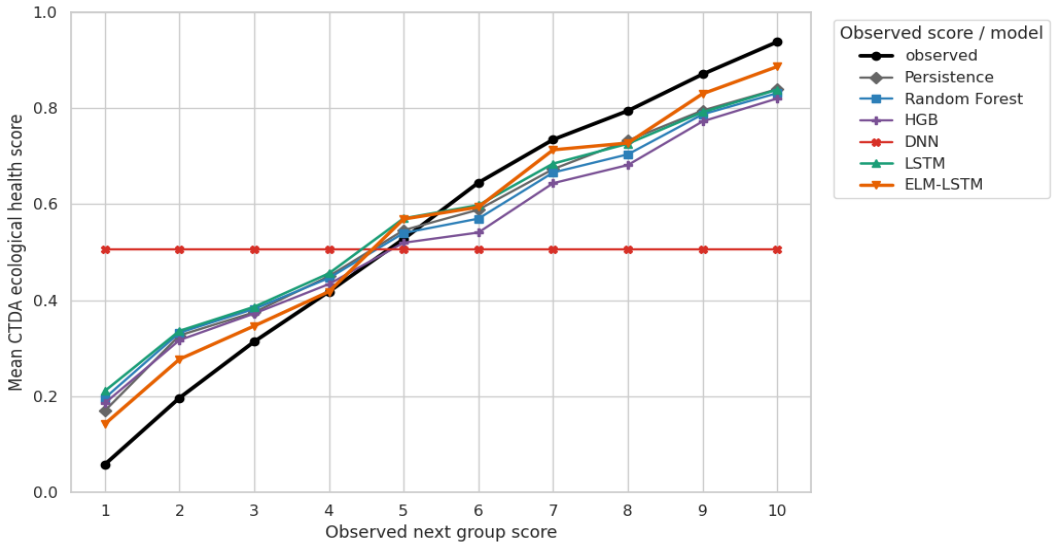

Figure 2: Calibrated forecasting of ecological health score of different models

Fig. 2 examines how well each model forecasted the variation across the observed ecological health score range in the held-out test set. Test observations were ranked according to their observed future CTDA scores and divided into ten equally sized groups, from lowest to highest observed condition. For each group, we calculated the mean observed CTDA score and the mean forecast produced by each model. Agreement between these two values indicates that a model separates sites with relatively low, intermediate, and high subsequent ecological health scores, rather than predicting average scores across the board that give no meaningful insight. Observed mean scores increased from approximately 0.05 in the first group to 0.94 in the tenth. Among the evaluated models, ELM-LSTM most closely reproduced this gradient. Its mean forecasting increased from 0.14 in the lowest group to 0.88 in the highest group. Agreement was the closest with the actual data during the fourth group, where both values were approximately 0.42. During the seventh group also the predicted mean of ELM-LSTM was 0.71, closely trailing the observed mean of 0.73.

Most evaluating models tended to predict scores towards the middle of the health range instead of the observed. They predicted better condition than was observed for low-scoring sites and poorer condition than was observed for sites. This compression of forecasts was least pronounced for ELM-LSTM. In the highest observed score group (sites with

best ecological health comparatively), the model forecasted 0.88, when the other models predicted around 0.82 and 0.84. Though the model did not forecast the score perfectly, it was better able to distinguish healthy waterbodies than the others. The performance of persistence, RF, HGB and LSTM were largely similar for most parts while DNN mostly failed to separate between any low and high condition scores.

Overall, the ELM-LSTM most effectively forecasted subsequent ecological-health scores by learning nonlinear temporal patterns across recent CTDA scores and the context.

6 Conclusion

Large-scale ecological monitoring programmes now actively combine diverse biological indices, habitat scores and complementary chemical measurements to assess waterbody health. Yet their reported health scores remain locked within different agency-specific frameworks that impede advanced modelling and analysis on a national scale. In this work we addressed the central problem of semantic non-equivalence across biological indices by developing CTDA, a data-driven aggregation method that harmonizes waterbody health scores from different agency scales while considering: i) assessment context, ii) irregular sampling patterns and iii) persistent ecological deficits. The CTDA evaluation against independent agency condition classes showed that the harmonized scores not only retain the original information encoded by standard-specific thresholds, but also strengthen their alignment with external evidence obtained from waterbody site metadata. Spearman concordance of 0.844 exceeded contextual rank and average temporal aggregation methods, which shows that combining context, gap-adaptive smoothing and deficit memory yields a health representation that is both interpretable and empirically robust. CTDA distinguishes transient deterioration from persistent degradation better effectiveness compare to rank-derived scores, which proves that explicit temporal deficit accumulation is necessary to avoid confusing brief disturbances with chronic ecological stress. Building on the harmonized trajectories, the proposed ELM–LSTM forecasting model delivered consistent gains over competing models. Compared to prevalent ML/DL models, the hybrid model reduced MAE by 6% and RMSE by nearly 10% against observed scores. Here, the ELM learns nonlinear patterns in the data, while LSTM analyzes previous temporal patterns to find trends beyond simple continuity. The calibrated group analysis performed afterwards further showed that ELM–LSTM best reproduced the gradient between low and high future CTDA scores, reducing the compression toward average predictions observed in RF, HGB and standalone LSTM models. These forecasts could help agencies identify sites for closer monitoring, given the operational values are validated.

Our methodology complements recent ML studies that focus on imputing missing hydrochemical data, predicting individual water quality parameters or modelling health indices using tree-based, neural and sequence models. Those approaches typically assume a single, homogeneous index and regular measurement schedules, whereas CTDA explicitly tackles heterogeneous biological indices, irregular monitoring gaps and multi-context aggregation before forecasting. Together, these developments inspire the development of integrated systems in which harmonized ecological scores, physiological or chemical parameters jointly inform environmental decisions of the waterbodies.

CTDA provides a transparent mechanism for converting biological indices calculated with different scales and standards into a common relative benchmark. This allows national repositories to be used for appropriate trend analysis and ML-assisted forecasting without discarding agency-specific interpretation. The ELM–LSTM model illustrates that lightweight sequence models can outperform both static ensemble models and deep networks on harmonized health trajectories. Their deployment in operational settings is thus extremely valuable where training cost and interpretability matter. Future work would incorporate uncertainty estimates for contextual ranks and deficit burdens within CTDA, enabling probabilistic assessments of ecological condition and forecast confidence. The aggregation scheme could also be extended to fish and macroinvertebrate data to test its transferability beyond STORET/WQX.

References

[1] James S Albert, Georgia Destouni, Scott M Duke-Sylvester, Anne E Magurran, Thierry Oberdorff, Roberto E Reis, Kirk O Winemiller, and William J Ripple. Scientists' warning to humanity on the freshwater biodiversity crisis. *Ambio*, 50(1):85–94, 2021.

[2] Tor E Eriksen, John E Brittain, Geir Søli, Dean Jacobsen, Peter Goethals, and Nikolai Friberg. A global perspective on the application of riverine macroinvertebrates as biological indicators in africa, south-central america, mexico and southern asia. *Ecological Indicators*, 126:107609, 2021.

[3] Morwenna McKenzie, Judy England, Ian Foster, and Martin Wilkes. Evaluating the performance of taxonomic and trait-based biomonitoring approaches for fine sediment in the uk. *Ecological Indicators*, 134:108502, 2022.

[4] Priscilla Wagari Mureithi, Amon Aine, Rose Basooma, Judith Namumbya, Florence Nansumbi, Mourine Jessie Yegon, Harald Meimberg, and Wolfram Graf. Advancements in macroinvertebrate-based river bioassessment

research in the afrotropical region: review and steps towards a regional framework. *Environmental monitoring and assessment*, 197(8):935, 2025.

[5] Md Rajaul Karim, MM Mahbubul Syeed, Ashifur Rahman, Khondkar Ayaz Rabbani, Kaniz Fatema, Razib Hayat Khan, Md Shakhawat Hossain, and Mohammad Faisal Uddin. A comprehensive dataset of surface water quality spanning 1940-2023 for empirical and ml adopted research. *Scientific Data*, 12(1):391, 2025.

[6] Jiwei Zhang, Pinghong Jiang, Kai Chen, Siwen He, Beixin Wang, and Xiaowei Jin. Development of biological water quality categories for streams using a biotic index of macroinvertebrates in the yangtze river delta, china. *Ecological Indicators*, 117:106650, 2020.

[7] Dimitris Politikos, Kostas Stefanidis, George Varlas, Anastasios Papadopoulos, and Elias Dimitriou. Predicting the ecological quality of rivers: A machine learning approach and a what-if scenarios tool. *Environmental Modeling & Assessment*, 29(6):1059–1077, 2024.

[8] Mohammad Najafzadeh, Elahe Sadat Ahmadi-Rad, and Daniel Gebler. Ecological states of watercourses regarding water quality parameters and hydromorphological parameters: deriving empirical equations by machine learning models. *Stochastic Environmental Research and Risk Assessment*, 38(2):665–688, 2024.

[9] Daniel Gebler, Pedro Segurado, Maria Teresa Ferreira, and Francisca C Aguiar. Predicting freshwater biological quality using macrophytes: A comparison of empirical modelling approaches. *Environmental Science and Pollution Research*, 31(56):65092–65108, 2024.

[10] J Emilio Sánchez-Moyano, Mario López-Cepeda, and Isabel García-Asencio. Evaluating the effectiveness of biotic indices for long-term ecological quality assessment in a heavily polluted estuary. *Environmental Monitoring and Assessment*, 197(10):1100, 2025.

[11] Justin Bousquin and Cristina A Mullin. harmonize-wq: Standardize, clean and wrangle water quality portal data into more analytic-ready formats. *Journal of open source software*, 9(102):7305, 2024.

[12] Bethany Johnson and Stephan B Munch. An empirical dynamic modeling framework for missing or irregular samples. *Ecological Modelling*, 468:109948, 2022.

[13] Markus Majaneva, Line Elisabeth Sundt-Hansen, Hege Brandsegg, Rolf Sivertsgård, Terje Bongard, and Frode Fossøy. Comparing methods and indices for biodiversity and status assessment in a hydropower-regulated river. *Hydrobiologia*, 851(9):2139–2157, 2024.

[14] Alperen Ertas, Tugba Boz, and Inci Tüney Kızılkaya. Comparative analysis of benthic macroinvertebrate-based biotic and diversity indices used to evaluate the water quality of kozluoluk stream (west anatolia of turkey). *Community Ecology*, 22(3):381–390, 2021.

[15] Abdelmottalib Nahli, Said Oubraim, and Mohamed Chlaida. Application of the biotic indices for water quality and resilience assessment of a disturbed stream (casablanca, morocco). *Biologia*, 77(10):2887–2904, 2022.

[16] Xiang Zhang, Yu Meng, Jun Xia, Bi Wu, and Dunxian She. A combined model for river health evaluation based upon the physical, chemical, and biological elements. *Ecological Indicators*, 84:416–424, 2018.

[17] Melissa J Dobbie and David Dail. Robustness and sensitivity of weighting and aggregation in constructing composite indices. *Ecological Indicators*, 29:270–277, 2013.

[18] David Lindén, Marco Cinelli, Matteo Spada, William Becker, Patrick Gasser, and Peter Burgherr. A framework based on statistical analysis and stakeholders' preferences to inform weighting in composite indicators. *Environmental Modelling & Software*, 145:105208, 2021.

[19] Xianhe Wang, Ying Li, Qian Qiao, Adriano Tavares, and Yanchun Liang. Water quality prediction based on machine learning and comprehensive weighting methods. *Entropy*, 25(8):1186, 2023.

[20] Kangyang Chen, Hexia Chen, Chuanlong Zhou, Yichao Huang, Xiangyang Qi, Ruqin Shen, Fengrui Liu, Min Zuo, Xinyi Zou, Jinfeng Wang, et al. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water research*, 171:115454, 2020.

[21] Md Mehedi Hassan, Md Mahedi Hassan, Laboni Akter, Md Mushfiqur Rahman, Sadika Zaman, Khan Md Hasib, Nusrat Jahan, Raisun Nasa Smrity, Jerin Farhana, M Raihan, et al. Efficient prediction of water quality index (wqi) using machine learning algorithms. *Human-Centric Intelligent Systems*, 1(3):86–97, 2021.

[22] Amin Arzhangi and Sadegh Partani. Water quality index prediction via a robust machine learning model using oxygen-related indices for river water quality monitoring. *Scientific Reports*, 16(1):6102, 2026.

[23] Rafael Rodríguez, Marcos Pastorini, Lorena Etcheverry, Christian Chreties, Mónica Fossati, Alberto Castro, and Angela Gorgoglione. Water-quality data imputation with a high percentage of missing values: A machine learning approach. *Sustainability*, 13(11):6318, 2021.

[24] KP Rasheed Abdul Haq and VP Harigovindan. Water quality prediction for smart aquaculture using hybrid deep learning models. *Ieee Access*, 10(1):60078–60098, 2022.

[25] Wenxian Luo, Leijun Huang, Jiabin Shu, Hailin Feng, Wenjie Guo, Kai Xia, Kai Fang, and Wei Wang. Predicting water quality in municipal water management systems using a hybrid deep learning model. *Engineering Applications of Artificial Intelligence*, 133:108420, 2024.

[26] Sakib Shahriar Shafin, Gour Karmakar, Iven Mareels, Venki Balasubramanian, and Ramachandra Rao Kolluri. Whose data are reliable: Sensor declared data reliability. In *2023 19th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, pages 249–254. IEEE, 2023.

[27] Joseph Malisaba, Eisa Babiker Omer, Erheyovwe Bubu, and Perez Mukasa. A review of advanced machine learning and deep learning technologies for water quality management. *Discover Applied Sciences*, 2026.

[28] Heelak Choi, Sang-Ik Suh, Su-Hee Kim, Eun Jin Han, and Seo Jin Ki. Assessing the performance of deep learning algorithms for short-term surface water quality prediction. *Sustainability*, 13(19):10690, 2021.

[29] Xiaoxuan Lu, Han Zou, Hongming Zhou, Lihua Xie, and Guang-Bin Huang. Robust extreme learning machine with its application to indoor positioning. *IEEE transactions on cybernetics*, 46(1):194–205, 2015.

[30] Rahim Barzegar, Asghar Asghari Moghaddam, Jan Adamowski, and Bogdan Ozga-Zielinski. Multi-step water quality forecasting using a boosting ensemble multi-wavelet extreme learning machine model. *Stochastic environmental research and risk assessment*, 32(3):799–813, 2018.

[31] Sakib Shahriar Shafin, Gour Karmakar, and Iven Mareels. Obfuscated memory malware detection in resource-constrained iot devices for smart city applications. *Sensors*, 23(11):5348, 2023.

[32] Shahriar Kaisar, Md Mamunur Rashid, Abdullahi Chowdhury, Sakib Shahriar Shafin, Joarder Kamruzzaman, and Abebe Diro. Enhancing telemarketing success using ensemble-based online machine learning. *Big Data Mining and Analytics*, 7(2):294–314, 2024.

[33] Yuheng Jia, Sam Kwong, and Ran Wang. Applying exponential family distribution to generalized extreme learning machine. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 50(5):1794–1804, 2018.

[34] Md Sarwar M Haque, Ghazanfar Latif, Md Rafiul Hasan, Md Arifuzzaman, Shakib S Shafin, and Quazi A Rahman. Scalable parallel svm on cloud clusters for large datasets classification. In *2nd Smart Cities Symposium (SCS 2019)*, page 13. IET, 2019.

[35] Carolina Cristiane Pinto, Giovanna Moura Calazans, and Sílvia Corrêa Oliveira. Assessment of spatial variations in the surface water quality of the velhas river basin, brazil, using multivariate statistical analysis and nonparametric statistics. *Environmental Monitoring and Assessment*, 191(3):164, 2019.

[36] Vinay Kumar Gautam, Chaitanya B Pande, Kanak N Moharir, Abhay M Varade, Nitin Liladhar Rane, Johnbosco C Egbueri, and Fahad Alshehri. Prediction of sodium hazard of irrigation purpose using artificial neural network modelling. *Sustainability*, 15(9):7593, 2023.