LakeBeD-US: *a benchmark dataset for lake water quality time series and vertical profiles*

Bennett J. **McAfee** et al.

Correspondence: **Bennett J. McAfee (bennettjmcafee@gmail.com)**

Abstract.

Water quality in lakes is an emergent property of complex biotic and abiotic processes that differ across spatial and temporal scales.

[If the authors mean that the processes vary over spatial and temporal scales then either the number or processes vary or the nature of each process may vary. Then "differ", as the plural form, seems an appropriate choice. However, if the authors mean that the water quality in lakes varies over spatial and temporal scales "differs" as a singular form seems more appropriate. Can the authors clarify their meaning please?**]**

Water quality is also a determinant of ecosystem services that lakes provide,[no need for a comma after 'provide'] and thus is of great interest to ecologists. ~~Increasingly, m~~**M**achine learning and other computer science techniques are [insert 'increasingly'] being used to predict water quality dynamics as well as to gain a greater understanding of water quality patterns and controls. To benefit both the sciences of ecology and computer science, we have created a benchmark dataset of lake water quality time series and vertical profiles. LakeBeD-US contains over 500 million unique observations of lake water quality collected by multiple long-term monitoring organizations across 17 water quality variables in 21 lakes in the United States. There are two published versions of LakeBeD-US: an "Ecology Edition" published in the Environmental Data Initiative repository, and a "Computer Science Edition" published in the Hugging Face repository. Each edition is formatted in a manner conducive to inquiries and analyses specific to each domain. For ecologists, LakeBeD-US provides an opportunity to study the spatial and temporal dynamics of several lakes with varying water quality, ecosystem, and landscape characteristics. For computer scientists, LakeBeD-US acts as a benchmark dataset that enables the advancement of machine learning for water quality prediction. [in the previous 2 sentences both refer to the same name 'LakeBeD-US' but each is recognised in the preceding text as either the '"Ecology Edition" or the "Computer Science Edition". To minimise confusion, especially to an "AI" algorithm, there needs to be clarity here and in the whole paper. Would "**LakeBed-US-E**" and **"LakeBeD-US-CS**" suffice as unambiguous acronyms of each program? **Can the authors please clarify and alter the rest of the MS accordingly?**]

1 Introduction

Water quality is a critical determinant of the ecosystem services provided by lakes (Keeler et al., 2012; Angradi et al., 2018). Water quality varies across spatial and temporal scales (Hanson et al., 2006; Langman et al., 2010; Soranno et al., 2017) due to a variety of interacting physical and biological processes. For example, hypolimnetic anoxia (low oxygen) in lakes decreases habitat for cold-water fish species (Arend et al., 2011; Jane et al., 2024). Anoxia can be fueled by the product of another water quality problem, the formation of toxic phytoplankton blooms (Jane et al., 2021). Both of these water quality phenomena emerge at the ecosystem scale as a consequence of multiple physical-biological interactions, driven by external nutrient loads and weather conditions (Paerl and Huisman, 2009; Snortheim et al., 2017; Ladwig et al., 2021; Jane et al., 2021). While there is mechanistic understanding of how these water quality phenomena evolve for well-studied lake systems, predicting their occurrence under scenarios of change or in large numbers of systems with sparse data remains challenging (Guo et al., 2021; Miller et al., 2023). To meet this challenge, we need scalable water quality models that are supported by observational data of sufficient spatiotemporal resolution to recreate key water quality dynamics (Ejigu, 2024; Varadharajan et al., 2022). Knowledge-guided machine learning (KGML) has emerged as a powerful technique for incorporating both ecological knowledge and observational data within a model (Karpatne et al., 2017, 2024). By fusing machine learning with physical and ecological principles, KGML has proven effective for assessing lake surface area change (Wander et al., 2024), modeling lake temperature (Read et al., 2019; Daw et al., 2014; Ladwig et al., 2024; Chen et al., 2024b), phytoplankton (chlorophyll) forecasting (Lin et al., 2023; Chen et al., 2024a), and predicting lake phosphorus concentrations

(Hanson et al., 2020). ~~As has been shown~~ **Thus**, a variety of modeling techniques within and beyond KGML are required to advance water quality understanding and prediction (Wai et al., 2022; Lofton et al., 2023). Creative approaches will likely spring from interdisciplinary collaborations of both lake ecologists and computer scientists (Carey et al., 2019) and will need diverse, high volume, high quality observational data that are easily accessible to researchers from multiple disciplines. Predicting the evolution of water quality through time and space requires treating lakes as dynamical systems that operate across many scales. Studies that have addressed the temporal dynamics of water quality at broad spatial scales are few in number **(but see**, Wilkinson et al., 2022; Zhao et al., 2023; Meyer et al., 2024), due in large part to the nature of data collection and research project design focusing on one scale at a time.[some clumsy sentence construction here, I assume this document developed from an amalgam of text from both ecologists and computer scientists, each with unique styles of expression. I respectfully suggest a scan of the draft by an independent person with good US English grammar and composition skills to enable a smooth flow of explanation through the whole document] Datasets that capture spatial gradients (Soranno et al., 2017; Pollard et al., 2018), temporal gradients (Magnuson et al., 2006; Goodman et al., 2015), or both have been curated manually to produce harmonized derived products (Read et al., 2017). Few examples of lake water quality data exist that harmonize both manually sampled and autonomously sampled high-frequency data across key gradients in space and across decadal timescales. A benchmark dataset for lake water quality that has well-resolved temporal data spanning multiple variables would be invaluable to both limnologists and computer scientists for simultaneously advancing both water quality modeling and KGML. Benchmark datasets are curated and cleaned datasets used in computation-heavy fields to test new operational methods and compare their performances (Peters et al., 2018). High-quality benchmark datasets are a significant effort to create (Sarkar et al., 2020) but are of fundamental importance to the field of computer science (Li et al., 2024). These datasets are becoming more prevalent in the field of ecology (e.g., Weinstein et al., 2021; Schur et al., 2023). Ecological benchmark datasets are vital as environmental data, including water quality data, exhibit properties such as prevalent missing values and non-normal distributions (Helsel, 1987; Lim and Surbeck, 2011) that are not typically represented in machine learning benchmark datasets. Benchmark datasets exist within the field of hydrology (e.g., Addor et al., 2017; Demir et al., 2022) and some recent limnology datasets advertise machine learning as a potential application (e.g., Spaulding et al., 2024), but benchmark datasets are rare in the field of limnology. This scarcity has caused some limnological studies to use non-limnological benchmark datasets to test their machine learning methods (e.g., Kadkhodazadeh and Farzin, 2021). This paper introduces LakeBeD-US, a dataset of lake water quality time series and vertical profiles intended as a benchmark for comparative methodological analysis for water quality modeling. LakeBeD-US harmonizes water quality data from long-term water quality monitoring programs, including the North Temperate Lakes Long-TermEcological Research program(NTLLTER), National Ecological Observatory Network (NEON), Niwot Ridge Long-Term Ecological Research program (NWTLTER), and the Carey Lab at Virginia Tech as part of the Virginia Reservoirs Long-Term Research in Environmental Biology.