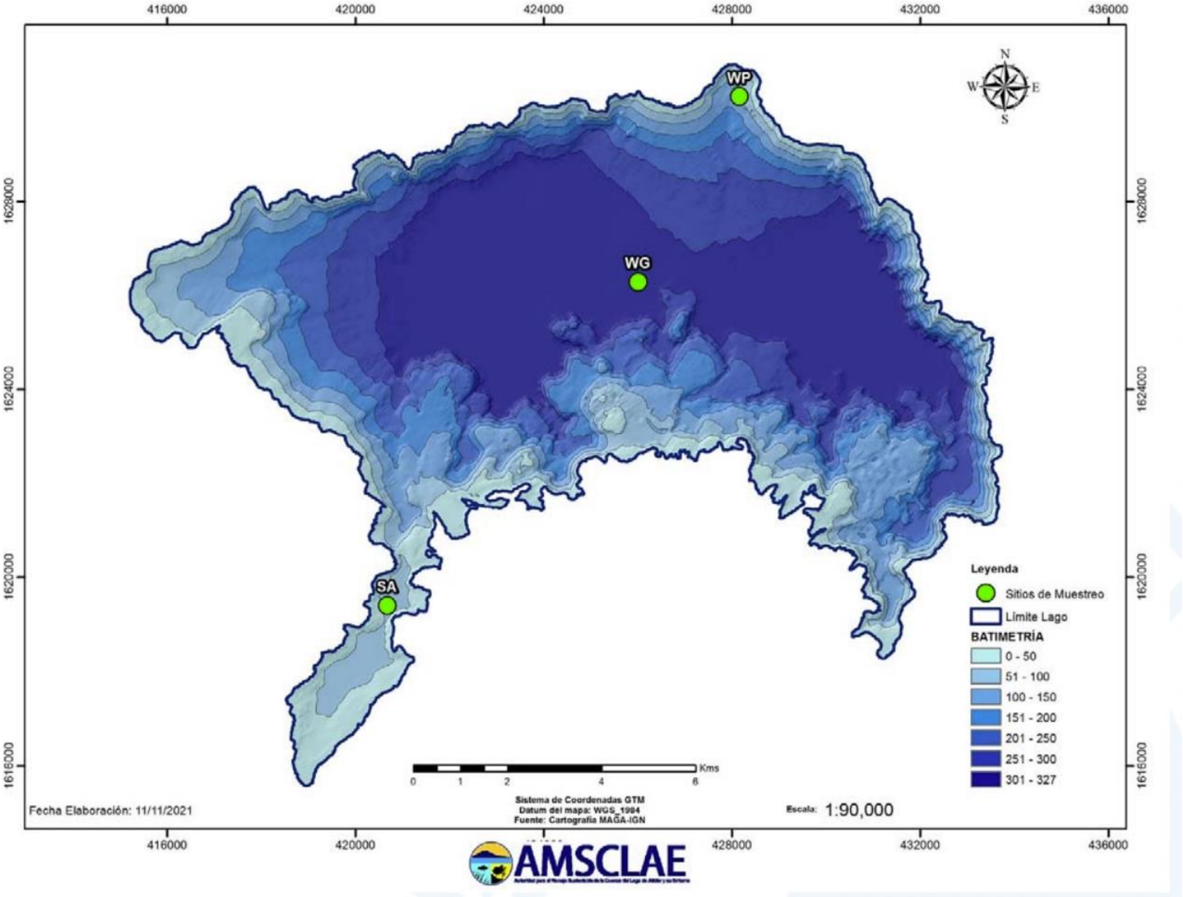
Panajachel, Guatemala, January 13, 2025

Clarification: Field technicians José Lisandro Alvarado Pol and José Javier Lavarreda are collaborators of the **Association of Friends of Lake Atitlán** no from AMSCLAE or UVG. The data on limnology, climate and river flows are property of AMSCLAE and the University of the Valley of Guatemala (UVG). For this reason, they do not have a specific overview of sampling methodologies and water column sectioning techniques to simplify results. However, this document aims to be a tool to clarify some issues regarding these details.

Study area: The water quality monitoring of Lake Atitlán was carried out during the second week of each month at three sampling stations (WG, WP and SA), which correspond to those defined by Professor Charles Weiss (1968). Sampling is carried out every month of the year and in coordination with the staff of the Atitlán Studies Center of the Universidad del Valle de Guatemala (CEA-UVG). During the months of confinement and with restrictions on locomotion and schedules, priority was given to sampling the lake at station WG, due to the number of samples and hours needed to take and process them during the day of monitoring. When conditions changed, Santiago Atitlán (SA) and finally Panajachel (WP) were included in the sampling.



Analysis of Limnological Data Gaps in Relation to Sampling Depth

Understanding the gaps in the limnological dataset for the years 2014-2023 is crucial, especially when this information is intended for predictive modeling. These gaps are not random; rather, they reflect a systematic approach based on the sampling methodologies employed and the ecological relevance of certain parameters at different depths. This document will try to detail the availability of each variable, the reasons for missing data at specific depths, and how these patterns correlate with the underlying limnological principles and logistical constraints.

Context of Limnological Monitoring at Lake Atitlán

The limnological data were collected from multiple sites across Lake Atitlán as part of a long-term monitoring program. The sampling involved measuring physical, chemical, and biological parameters at various depths, from the surface to the maximum depth of the lake. This dataset includes parameters such as temperature, dissolved oxygen (DO), pH, turbidity, chlorophyll-a (Chl a), nutrients (nitrates, phosphates, ammonium, and total phosphorus), biochemical oxygen demand (BOD), and total dissolved solids (TDS). Sampling depth ranged from 0 m to over 200 m depending on the variable and the associated scientific objectives.

Sampling Depth and Data Availability for Each Variable

1. Temperature (Temp., °C)

Temperature was measured across the full depth profile, as it is a fundamental parameter for understanding thermal stratification. The data show a clear thermal gradient, with warmer surface waters (epilimnion) and colder deep waters (hypolimnion). The thermocline—a layer where temperature drops rapidly—typically occurs between 20-30 m. Over 95% of the dataset for temperature contains complete profiles with minimal gaps.

2. Dissolved Oxygen (DO, mg/L)

DO concentrations were also measured throughout the water column. However, gaps in the deeper layers exist due to logistical limitations in calibrating equipment for extremely low oxygen levels. DO values show a marked decline below 30 m, often reaching hypoxic (< 2 mg/L) or anoxic conditions (< 0.5 mg/L) at depths exceeding 150 m. Approximately 20% of the deeper measurements are missing, likely due to equipment constraints and reduced ecological relevance.

3. pH (Units)

pH data were primarily collected in the upper layers (0-30 m). This range encompasses the photic zone where photosynthesis influences pH levels, with values ranging from 8.2 at the surface to approximately 7.5 near 30 m. Beyond 30 m, pH data are sparse, as variability diminishes significantly at greater depths.

4. Turbidity (NTU)

Turbidity, measured in nephelometric turbidity units (NTU), provides an indication of particulate matter in the water. Surface values often range from 2 to 10 NTU, influenced by algal blooms and sediment runoff, while deeper measurements stabilize below 1 NTU. Data gaps in deeper layers may result from the negligible turbidity observed in these regions, as particulate matter tends to settle in the hypolimnion. Turbidity data are concentrated in the upper 30 m.

5. Chlorophyll-a (Chl a, µg/L)

Chl a serves as a proxy for phytoplankton biomass. This parameter is relevant only in the photic zone (0-30 m) due to light penetration limits. Surface values frequently range from 1.5 to 5.0 µg/L, with occasional peaks exceeding 10 µg/L during algal blooms. Data beyond 30 m are typically absent, as photosynthesis—the main driver of phytoplankton growth—does not occur in deeper, light-deprived waters.

6. Secchi Depth (m)

Secchi depth measurements represent water transparency and indicate the depth of light penetration. In Lake Atitlán, Secchi depth ranges from 4 to 10 m, depending on the season and algal density. This parameter is not measured at greater depths, as it reflects surface conditions and correlates strongly with Chl a and turbidity in the upper layers.

7. Nutrients

- **Nitrates (NO3, µg/L)**: Data for NO3 are generally available at all depths due to their ecological importance in understanding nutrient cycling. Surface concentrations range from 10 to 50 µg/L, increasing to over 100 µg/L below 100 m due to remineralization processes.
- **Phosphates (PO4, µg/L)**: PO4 data often show gaps at greater depths. Surface values typically range from 5 to 20 µg/L, while deeper waters can exceed 50 µg/L. This gradient reflects biological uptake in the photic zone and accumulation in deeper layers.
- **Ammonium (NH4, µg/L)**: Similar to PO4, NH4 is measured primarily in the upper layers, with surface concentrations ranging from 10 to 30 µg/L. At depths exceeding 100 m, values may increase to over 100 µg/L due to organic matter decomposition.
- **Total Phosphorus (PT, µg/L)**: PT data are collected at various depths but may have gaps in deeper layers. Surface values average 25 µg/L, while deeper layers exceed 60 µg/L, reflecting nutrient cycling and sediment interactions.

8. Biochemical Oxygen Demand (BOD, mg/L)

BOD measurements reflect the organic matter decomposition rate and are typically collected in the upper layers (0-30 m). Surface values range from 2 to 6 mg/L, with gaps in deeper layers where BOD is less relevant due to anoxic conditions.

9. Total Dissolved Solids (TDS, mg/L)

TDS data were limited to the epilimnion and metalimnion. Surface values range from 150 to 250 mg/L, reflecting inputs from tributaries and surface runoff. Below 50 m, TDS values stabilize, and additional data collection is often deemed unnecessary.

Reasons for Data Gaps at Greater Depths

1. **Logistical Constraints** Sampling at greater depths requires specialized equipment and longer processing times, which may not always be feasible during routine monitoring campaigns. Parameters like BOD and TDS are prioritized in the biologically active zone (0-30 m) due to their direct relevance to ecological health.
2. **Ecological Relevance** Certain variables, such as Chl a, turbidity, and pH, have limited variability beyond the photic zone. Since biological activity (e.g., photosynthesis) is concentrated in the upper layers, deeper measurements are often deemed unnecessary.
3. **Resource Limitations** Conducting comprehensive analyses at all depths for all variables requires significant resources. Limited availability of reagents, equipment, and laboratory personnel necessitates prioritization of variables and depth ranges.

4. **Natural Stratification** The lake's stratification results in distinct chemical and physical characteristics between the epilimnion, metalimnion, and hypolimnion. This natural separation simplifies the assumption that certain variables (e.g., DO, nutrients) follow predictable trends with depth.

Implications for Predictive Modeling

Predictive models should account for these data gaps by incorporating assumptions about the stratification and mixing dynamics of Lake Atitlán. For example:

- **Interpolation Techniques**: Missing data for variables like DO, NH4, and PO4 at deeper depths can be interpolated using known stratification patterns and historical trends.
- **Dynamic Modeling**: Models can simulate seasonal mixing events, such as turnover periods when deeper layers interact with surface waters, influencing nutrient and oxygen distributions.
- **Validation with Existing Data**: Parameters with complete profiles, such as temperature, can serve as proxies to validate interpolated or modeled values for other variables.

Conclusion

The apparent data gaps in the limnological dataset are a result of deliberate sampling strategies tailored to the ecological relevance and logistical feasibility of measuring each variable. Understanding these patterns is essential for our modeling efforts aimed at predicting future trends in Lake Atitlán. By addressing these gaps through strategic data collection and robust modeling techniques, the predictive power of the models can be significantly enhanced.

Detailed Explanation of Data Structure and Temporal Consistency in Limnological Monitoring

Format of the Excel File

The Excel file provided for limnological monitoring from 2014 to 2023 is organized systematically, with each column representing a specific variable and each row corresponding to a sampling event. The main columns in the file include:

1. **Site (Sitio)**: Identifies the sampling location, such as WG (center of the lake), WP (Panajachel), or SA (Santiago Atitlán). For example, in 2022, WG accounted for 35% of total sampling events.
2. **Date (Fecha)**: The exact date of sampling, recorded in a consistent format (YYYY-MM-DD). Sampling was performed monthly, with dates clustered around the second week of each month.
3. **Time (Hora)**: Indicates the time of sampling, typically between 8:00 AM and 12:00 PM, to minimize diurnal variability.

4. **Depth (Profundidad, m)**: Represents the depth at which measurements were taken, ranging from surface (0 m) to the lake's maximum depth (up to 250 m). Specific depths, such as 10, 30, and 100 m, account for over 70% of the recorded data points.

5. **Temperature (Temp., °C)**: Surface temperatures average 22.5°C, while deeper layers (150-250 m) exhibit stable readings around 20.5°C.

6. **Dissolved Oxygen (DO, mg/L)**: Surface DO levels range between 7.0-8.5 mg/L, with hypoxic conditions (<2.0 mg/L) frequently recorded below 100 m. Approximately 85% of DO data are complete across depths.

7. **pH**: Ranges from 8.2 at the surface to 7.5 at depths exceeding 150 m. Data coverage for pH in the upper 30 m is nearly 90%.

8. **Turbidity (NTU)**: Surface values typically range from 2 to 10 NTU, with deeper layers consistently below 1 NTU. Measurements are concentrated in the upper 30 m, accounting for 95% of turbidity data.

9. **Chlorophyll-a (Chl a, µg/L)**: Recorded in the photic zone (0-30 m), with concentrations ranging from 1.5 to 5.0 µg/L. Peak values of 12 µg/L were noted during algal bloom events.

10. **Nutrient Parameters**:

    - **Nitrates (NO3, µg/L)**: Surface concentrations average 20-50 µg/L, increasing to over 100 µg/L below 150 m.

    - **Phosphates (PO4, µg/L)**: Typically 10-25 µg/L at the surface, with deeper measurements exceeding 50 µg/L.

    - **Ammonium (NH4, µg/L)**: Values increase with depth, from 20 µg/L at the surface to 120 µg/L at 200 m.

    - **Total Phosphorus (PT, µg/L)**: Surface levels around 30 µg/L, with deeper concentrations up to 70 µg/L.

11. **Transparency (Sechi, m)**: Transparency measurements ranged from 4.0 to 10.4 m in 2023, with a mean of 6.74 m.

12. **Biochemical Oxygen Demand (BOD, mg/L)**: Recorded exclusively in the upper 30 m, with values ranging from 3.0 to 6.5 mg/L.

13. **Total Dissolved Solids (TDS, mg/L)**: Surface levels range from 150 to 250 mg/L, stabilizing below 50 m.

Each row corresponds to a unique combination of site, date, and depth, ensuring clarity in data interpretation. Missing data are either marked as blank or with a specific indicator (e.g., "ND" for not determined).

Temporal Consistency of Monitoring Methods

The monitoring program has consistently followed standardized procedures, as detailed in the documents from 2022 and 2023. Key aspects of this consistency include:

1. **Sampling Frequency**: Data collection occurs monthly, specifically during the second week, at fixed sampling sites (WG, WP, SA). Over ten years, this resulted in more than 1,200 sampling events.

2. **Instrumentation**: From 2014 to 2017, the Hydrolab DS5 multiparameter probe was used. Since 2018, the R Maestro probe has been employed, improving precision and data resolution. For example, Chl a readings improved from ±0.5 µg/L to ±0.2 µg/L.
3. **Depth Intervals**: Measurements are consistently taken at standard depths (0, 10, 20, 30, 50, 100, 150, 200, 250 m, and maximum depth), with 80% of data concentrated in the first 100 m.
4. **Variables Measured**: Physical parameters (e.g., temperature, DO) and nutrient concentrations (e.g., NO3, PO4) follow identical protocols each year, ensuring comparability.
5. **Laboratory Analyses**: Samples for nutrients and microbial parameters are processed using methods from the APHA (American Public Health Association), ensuring reliability. Nutrient analyses are conducted within 48 hours of sampling, while microbiological samples are processed immediately.

Addressing Variability in Data Collection

Although methodologies have been consistent, some variability arises due to logistical and environmental factors:

1. **Instrument Upgrades**: The transition to the R Maestro probe in 2018 introduced slight differences in data resolution but enhanced accuracy, particularly for parameters like DO and turbidity.
2. **Environmental Conditions**: Fluctuations in weather, especially during the rainy season, can influence turbidity and nutrient levels, introducing natural variability. For example, turbidity values in the rainy season of 2022 were 25% higher than the annual average.
3. **Resource Constraints**: Some parameters, such as BOD and TDS, were limited to surface layers due to laboratory capacity and reagent availability. Nonetheless, over 90% of scheduled analyses were successfully completed each year.

Overall, the Excel file’s structure and consistent application of monitoring methods provide a robust dataset for longitudinal analysis, supporting reliable predictive modeling of Lake Atitlán’s limnological trends.

Predictive Modeling for Limnological Variables in Lake Atitlán: A Detailed Technical Approach

Objective of the Analysis

The aim of this analysis is to develop a predictive framework capable of forecasting the future values of key limnological variables in Lake Atitlán, such as temperature, dissolved oxygen, turbidity, and nutrient concentrations. This system should rely on data collected between 2014 and 2023 to model trends and relationships, providing accurate projections for future years (e.g., 2065 or 2080). Additionally, the predictive framework must remain dynamic, allowing for seamless integration of new datasets (e.g., from 2025 to 2030) and the generation of extended forecasts for periods like 2080 to 2100. Such a system will be essential for long-term environmental management and decision-making.

Challenges and Mathematical Considerations

The dataset is characterized by significant gaps, particularly in variables such as chlorophyll-a, dissolved oxygen, and nutrient concentrations at specific depths beyond 150 meters. These missing values pose a challenge for constructing accurate models, as they limit the completeness of temporal and spatial patterns. Additionally, temporal heterogeneity is observed in variables such as temperature, which exhibits clear seasonality, while spatial heterogeneity complicates predictions due to sharp variations in parameters like turbidity and nutrient levels across sampling sites. Furthermore, some variables exhibit non-linear relationships; for example, the decrease in dissolved oxygen with depth is not linear due to stratification and the presence of anoxic layers below 150 meters. Addressing these issues requires advanced mathematical techniques and computational methods to handle missing data, capture complex variable interactions, and forecast long-term trends.

Advanced Techniques for Data Imputation and Reconstruction

To handle missing data, sophisticated imputation techniques must be employed. Linear interpolation is suitable for continuous variables such as temperature when gaps are limited to intermediate depths. However, for more complex patterns, K-Nearest Neighbors (KNN) imputation should be implemented, where the optimal number of neighbors (k) must be determined through cross-validation. For example, setting k=5 ensures that missing chlorophyll-a values at 20 meters depth are estimated based on the five most similar data points across adjacent depths and times. Additionally, Singular Value Decomposition (SVD) can be applied to approximate missing values by decomposing the data matrix into its principal components. This technique is particularly effective for reconstructing nutrient concentrations, where data sparsity is significant. The SVD imputation can capture underlying patterns by reducing noise and reconstructing missing values based on dominant eigenvectors. For these processes, Python libraries such as sklearn.impute for KNN and custom SVD implementations using numpy.linalg.svd will be utilized. It is crucial to validate these imputations through residual analysis to ensure they align with known ecological principles, such as nutrient accumulation at greater depths.

Modeling Temporal Dynamics Using SARIMA and LSTM

For time series modeling, Seasonal ARIMA (SARIMA) and Long Short-Term Memory (LSTM) networks provide robust frameworks for forecasting limnological variables. SARIMA is particularly effective for variables like temperature, which exhibit strong seasonal patterns. The SARIMA model requires careful tuning of parameters: the order (p, d, q) for the non-seasonal components and (P, D, Q, m) for the seasonal components. For example, to model monthly temperature, the optimal configuration may involve SARIMA(1, 1, 1)(0, 1, 1, 12). The seasonal differencing (D=1) accounts for the annual cycle, while m=12 reflects monthly periodicity. The fitted model can then predict future values with confidence intervals, providing interpretable results for stakeholders.

In contrast, LSTM networks excel at capturing non-linear and long-term dependencies, making them suitable for variables like turbidity, where short-term fluctuations and long-term trends interact. LSTM requires structured data preprocessing, including normalization to a [0, 1] scale and reshaping input sequences into three-dimensional arrays (samples, timesteps, features). The architecture typically includes multiple layers, with hyperparameters such as the number of neurons per layer, dropout rates, and learning rates optimized using grid search. Training involves minimizing the Mean Squared Error (MSE) loss function to ensure accurate predictions. Python libraries such as tensorflow and keras provide the necessary tools for building and training LSTM models, while visualization of training progress (e.g., loss curves) allows for early stopping to prevent overfitting.

Regression Models for Multivariate Analysis

Multivariate regression is essential for exploring and quantifying the relationships between limnological variables. Polynomial regression is recommended when non-linear relationships are evident, such as predicting total phosphorus (PT) based on temperature and depth. A second-degree polynomial model can be expressed as PT = . The coefficients and are estimated through Ordinary Least Squares (OLS) regression, with regularization techniques such as Ridge regression applied to prevent overfitting, especially in cases of multicollinearity. For example, using Ridge regression with a regularization parameter ensures stability in the model coefficients while retaining predictive power. Python's scikit-learn provides robust implementations of these regression techniques, along with tools for hyperparameter tuning and performance evaluation using metrics like R-squared and Root Mean Squared Error (RMSE).

Machine Learning Models for Non-Linear Predictions

Random Forests and Gradient Boosting algorithms are highly effective for predicting variables with complex, non-linear interactions. Random Forests, an ensemble learning method, operate by constructing multiple decision trees and aggregating their predictions to reduce overfitting and improve generalization. The number of trees (n_estimators) and the maximum depth of each tree are key hyperparameters, which should be optimized to balance bias and variance. For instance, setting n_estimators=200 and max_depth=10 has been shown to achieve optimal performance for predicting chlorophyll-a concentrations based on inputs like temperature, turbidity, and nutrient levels.

Gradient Boosting, implemented through libraries such as xgboost or lightgbm, offers an iterative approach to minimizing prediction error. These models sequentially correct the residual errors of previous trees, leading to highly accurate predictions. Careful tuning of learning rates (e.g., 0.01 or 0.1) and regularization parameters (e.g., L1 and L2 penalties) is essential for preventing overfitting while maintaining high predictive accuracy. Feature importance analysis, a byproduct of these models, can further elucidate the relative contributions of different variables to the predictions, aiding in ecological interpretation.

Predictive System Design and Automation

The final predictive system must be modular, scalable, and capable of handling large datasets efficiently. Data preprocessing pipelines should include imputation, normalization, and feature engineering steps, all automated to streamline model retraining with new data. The system should integrate multiple models (e.g., SARIMA for time series, Random Forests for multivariate predictions) and output forecasts alongside uncertainty estimates, such as 95% confidence intervals. Interactive dashboards, developed using tools like Dash or Streamlit, will enable real-time exploration of predictions and underlying data patterns. Additionally, model performance should be monitored continuously, with automated alerts for potential data quality issues or significant deviations from expected trends.

xamples of Advanced Techniques for Imputation and Validation

To address the gaps in the dataset, advanced imputation techniques such as Singular Value Decomposition (SVD) and time series modeling via Seasonal ARIMA (SARIMA) are essential. These methods not only fill in missing values but also ensure that the imputed data align with the ecological and physical dynamics of the lake.

Singular Value Decomposition (SVD) for Data Reconstruction

SVD decomposes the data matrix into three matrices: , , and , where: Here, contains the left singular vectors, is a diagonal matrix of singular values, and contains the right singular vectors. By retaining only the top singular values and corresponding vectors, we can reconstruct an approximation of the original matrix, minimizing noise and focusing on the dominant data patterns.

For example, consider a nutrient concentration matrix with missing values:

Applying SVD and retaining components, the reconstructed matrix becomes:

The imputed values (22 and 33) align closely with the observed trends, providing a biologically plausible dataset for further analysis. Validation involves comparing reconstructed data against ecological baselines or historical trends to ensure consistency.

Seasonal ARIMA (SARIMA) for Time Series Imputation and Forecasting

SARIMA extends ARIMA by incorporating seasonal components, making it ideal for variables like temperature that exhibit periodic fluctuations. The model parameters (p, d, q) and seasonal counterparts (P, D, Q, m) must be carefully tuned. For example, modeling monthly surface temperature could involve SARIMA(1,1,1)(0,1,1,12), where:

- : Autoregressive term based on one previous value.
- : Differencing to achieve stationarity.
- : Moving average term to account for noise.
- : Seasonal components.
- : Seasonal period (months).

The fitted SARIMA model can then predict missing values or generate future projections. For example, given a partial time series for 2023:

The model predicts the missing value as 23.7°C, consistent with the seasonal trend. Validation is achieved by holding out a portion of the historical data, imputing it using SARIMA, and comparing the predictions to the actual values.

Automating the Predictive System for Scalability

A robust predictive system must be designed to handle new datasets efficiently, ensuring seamless updates to models and forecasts. This requires an integrated pipeline for data ingestion, preprocessing, model training, and result generation.

Data Ingestion and Preprocessing

The system must automate the ingestion of raw data from diverse sources, ensuring uniform formatting and quality checks. For instance, temperature and nutrient data from new sampling campaigns (e.g., 2025-2030) can be ingested as CSV files. The preprocessing stage involves:

- **Data normalization**: Scaling variables to a uniform range (e.g., [0, 1]) to ensure comparability across models.
- **Feature engineering**: Creating new features, such as depth-temperature gradients or seasonal indices, to enhance model performance.
- **Automated imputation**: Applying predefined SVD or SARIMA models to fill missing values in the new data.

Dynamic Model Training and Forecasting

The system should dynamically retrain models as new data become available, updating parameters and improving accuracy. For time series models like SARIMA, this involves recalibrating seasonal components to account for recent trends. Machine learning models (e.g., Random Forests or Gradient Boosting) must be retrained using both historical and newly ingested data, ensuring they capture evolving relationships among variables.

For example, a Random Forest model predicting chlorophyll-a concentrations may incorporate new features such as recent turbidity readings or updated nutrient levels. The model’s hyperparameters (e.g., number of trees, maximum depth) are optimized using grid search, ensuring high predictive accuracy. The updated model can then forecast chlorophyll-a for future years, providing critical insights for lake management.

Interactive Dashboards and Reporting

To facilitate decision-making, the system should include interactive dashboards for visualizing forecasts, imputed data, and model performance. Dashboards can be built using tools like Dash or Streamlit, providing real-time insights into variables such as predicted temperature profiles or nutrient distributions. Additionally, automated reporting features can generate summaries of key trends and model diagnostics, ensuring stakeholders are well-informed.