中国科学院南京地理与湖泊研究所
NANJING INSTITUTE OF GEOGRAPHY & LIMNOLOGY, CHINESE ACADEMY OF SCIENCES

To

Hydrology and Earth System Sciences

Editorial office

Nanjing, 05.24.2026

Cover letter for revised manuscript egusphere-2025-6442

Dear Prof. Nadia Ursino,

We have thoroughly revised our manuscript ("Four decades of full-depth profiles reveal layer-resolved drivers of reservoir thermal regimes and episodic hypolimnetic warming") according to the comments from the three reviewers and we hereby resubmit the completely revised version for the further evaluation.

The revised manuscript has been substantially strengthened while retaining the original focus of the study. We now place XGBoost in a broader methodological context by benchmarking it against SVR, Random Forest, and MARS under a common modelling protocol, with clearer documentation of the predictor–target dataset, train–test consistency, hyperparameter settings, and statistical comparison of model errors. We also expanded the CE-QUAL-W2 description to make the physical basis of the process model more transparent, including its calibration and evaluation strategy, bathymetric representation, boundary forcing, and assumptions of the laterally averaged setup. We further refined the attribution framework. Causal wording has been moderated throughout the manuscript, and the SHAP-based interpretation is now supported by complementary explainability analyses and targeted CE-QUAL-W2 counterfactual simulations. These additions specifically test the two central attribution signals: the subsurface importance of antecedent shortwave radiation and the contribution of intensified late-season bottom withdrawals to episodic hypolimnetic warming. The revision also includes a redesigned methodological workflow figure and a clearer separation between long-term epilimnetic warming and event-scale hypolimnetic warming. We now feel very comfortable with the reworked version, which has improved a lot from our point of view.

Detailed response letters to the reviewer's comments, where all changes in the manuscript are explained and further information is provided, are attached below. We also include two versions of the new manuscript, one with all changes marked in colored text and another one with all changes accepted. We hope that our revision fulfills the requirements to make our work a contribution to *Hydrology and Earth System Sciences*.

With kind regards,
Xiangzhen Kong
On behalf of all coauthors

地址：南京市江宁区创展路 299 号　　邮编：211135
网址：http：//www.niglas.ac.cn
电话：025-86882010　86882020　传真：025-57714759
电子信箱：niglas@niglas.ac.cn

Address: 299#Chuangzhan Road, Nanjing 211135, China
http://www.niglas.ac.cn
Tel: +86-25-86882010　86882020　Fax: +86-25-57714759
E-mail: niglas@niglas.ac.cn

Reply on RC1

The present study applies two types of modelling approaches namely the CE-QUAL-W2 numerical model and a machine learning model based on XGBoost to predict the thermal structure of a reservoir. The machine learning model is trained using data generated from the CE-QUAL-W2 simulations. The modelling framework is clearly structured, with the ML model used to predict temperature-related variables at several depths, including epilimnetic (5 m), metalimnetic (15 m), hypolimnetic (30 m), and bottom (50 m) temperatures, as well as key thermal structure indicators such as Schmidt stability, bottom-to-surface temperature difference, and mixed layer depth. Based on certain assumptions, the authors restricted the analysis to the May-October period, which corresponds to the stratified season of the reservoir. A set of meteorological variables was used as predictive features, and 30-day moving average predictors were introduced to represent the cumulative influence of atmospheric and hydraulic forcing on thermal conditions across different depths.

Although the topic of the manuscript is relevant and demonstrates a satisfactory degree of originality, the overall organization of the paper particularly the Results section does not yet meet the scientific standards expected for publication in a high-quality journal. Substantial improvements are required before the manuscript can be considered for publication. In its current form, the study requires major revision. The following issues should be addressed carefully by the authors:

We sincerely thank the reviewer for the careful and constructive evaluation of our manuscript. We appreciate the reviewer's positive assessment of the relevance and originality of the study, as well as the clear identification of aspects that need to be strengthened. The reviewer's comments mainly concern methodological benchmarking, transparency of the dataset and model setup, and the depth of the interpretability analysis. These are valuable points, and we agree that the current version can be substantially improved in these respects. Please find below our detailed responses and corresponding revisions.

1. Lack of baseline comparison. It is neither appropriate nor convincing to present a machine learning prediction framework without a solid baseline for comparison. The present study relies essentially on a single ML model, which makes the analysis incomplete and weak from a methodological standpoint. Additional models (e.g., classical regression models or alternative ML approaches) should be included to provide a meaningful benchmark.

We agree that the original approach provides an incomplete benchmark for the XGBoost framework. Our original intention was not to conduct an exhaustive algorithm-comparison study, but rather to build an interpretable process-ML attribution framework on top of the CE-QUAL-W2 outputs. We fully agree that the performance of XGBoost should be better contextualized against alternative baseline models.

In the revised manuscript, we have implemented this suggestion by extending the modelling framework beyond XGBoost and adding three additional data-driven benchmark models (see Section 2.4): Multivariate Adaptive Regression Splines (MARS), Support Vector Regression (SVR), and Random Forest (RF). These models were applied to the same target variables as XGBoost, including water temperatures at 5, 15, 30, and 50 m depths, Schmidt stability, bottom-to-surface temperature difference, and mixed-layer depth. To ensure a rigorous and directly comparable benchmark, all four models were trained and evaluated using the same input predictors, the same train-test partitioning strategy, and the same performance metrics. We have also added the descriptions of the benchmark models and their hyperparameter settings to the Methods section. In response to this comment and the related comments below, we further provide comprehensive result tables summarizing the training and testing performance separately for all target variables and all models (see the new Table 1). These additions allow the predictive skill of XGBoost to be evaluated against both regression-based and nonlinear machine-learning alternatives in a transparent and reproducible manner. Overall, XGBoost consistently achieved the highest predictive accuracy for both layer-specific water temperatures and stratification metrics, thereby providing a stronger methodological basis for using the XGBoost outputs in the subsequent SHAP-based driver-attribution analysis.

2. Absence of comprehensive result tables. The manuscript does not contain adequate tables summarizing the performance of the models. For a study focused on machine learning applications, detailed quantitative comparisons presented in well-structured tables are essential.

We agree that the original manuscript relied too heavily on figures and did not provide sufficiently compact quantitative summaries of model performance. To address this comment, we have prepared two structured benchmark tables for inclusion in the revised manuscript. Table 1 reports coefficients of determination (R^2), and Table 2 reports root mean square error (RMSE), both presented separately for the training and testing datasets across all target variables. The comparison now includes all four data driven models evaluated in the revised study, namely XGBoost, SVR, Random Forest, and MARS, and covers water temperature at 5, 15, 30, and 50 m, bottom to surface temperature difference, mixed layer depth, and Schmidt stability. This revision provides the detailed quantitative comparison requested by the reviewer and makes the relative predictive performance across models, targets, and datasets much more transparent. Moreover, we have added a new supplementary table (Table S4) to document the performance of CE-QUAL-W2 in reproducing the observed hydrodynamic and thermal structure of Rappbode Reservoir. This table summarizes the agreement between simulated and observed water level and depth-specific water temperatures using R^2 and RMSE as performance metrics. We believe these additions substantially improve the clarity and rigor of the Results section.

3. Insufficient description of the dataset. A much clearer presentation of the dataset used for model development is required. This should include:

3.1. Detailed description of input features and output variables

- Descriptive statistical analysis
- Explanation of the data splitting strategy (training and testing datasets)
- Distributional comparisons between training and testing sets using Kolmogorov-Smirnov tests (empirical CDF comparison)
- Jensen–Shannon distance comparisons between training and testing sets
- Energy distance comparisons to further evaluate distributional consistency.

We fully agree with the reviewer that the description of the machine learning dataset should be more systematic and transparent. Although the current manuscript defines the main predictors and targets, it does not yet present them in a sufficiently structured way. Following the suggestion, we now expand the dataset description in both Methods and Supplement.

First, as mentioned above, a new table (Table S2) has been added to the Supplementary Materials that summarizes the descriptive statistics of the hydro-meteorological predictors and thermal target variables used in machine-learning model development.

Second, data splitting strategy was indicated in the Methods section, where the full dataset was randomly divided into 80% training and 20% testing subsets, and hyperparameter tuning was performed within the training data using 5-fold rolling-origin cross-validation combined with Bayesian optimization. In the revised manuscript, we now clarify this procedure more explicitly in section 2.4 and also report training and testing results separately in the benchmark tables to improve transparency.

Third, to further document the consistency of the implemented train/test partitions, we have added a new supplementary table (Table S3) reporting the Kolmogorov-Smirnov statistic and p value, Jensen-Shannon distance, and energy distance for the actual training and testing subsets associated with each of the seven target-specific models. Across all seven targets, these diagnostics indicate no appreciable distributional shift between the training and testing subsets. Taken together, these additions substantially improve the clarity and completeness of the dataset description.

3.2. Hyperparameter specification. The hyperparameters used in the ML models should be clearly reported and justified to ensure reproducibility.

We fully agree with the reviewer that the hyperparameter settings should be reported more clearly to ensure reproducibility. In the revised manuscript, we have addressed this point by adding explicit hyperparameter information for all four data-driven models. For XGBoost, the optimized values of the key tuning parameters are now reported directly in the Methods section. For the three newly added benchmark models, namely SVR, Random Forest, and MARS, we now provide concise model descriptions together with their main hyperparameter settings. Details are now provided in the revised Section 2.4. These additions make the benchmark framework more transparent and reproducible.

3.3. Clear separation of training and testing results. Model performance should be presented distinctly for both training and testing datasets, using appropriate evaluation metrics and visualizations.

We agree that the separation of training and testing performance should be made more explicit. To address this point, the revised manuscript now reports model skill separately for the two datasets throughout the benchmark analysis. Specifically, Table 1 presents training and testing R^2 values, and Table 2 presents the corresponding RMSE values, for XGBoost and the three benchmark models (SVR, RF, and MARS) across all seven target variables, including four depth specific temperatures and three stratification metrics. This separation is also reflected in the revised Figures 4 and 5, which provide the corresponding visual representation of model performance for thermal and stratification dynamics. We believe these additions substantially improve the interpretability of the results.

3.4. Need for a methodological flowchart. A clear flowchart summarizing the overall framework of the study from data preparation to model development, validation, and interpretation should be provided.

We agree that the methodological workflow can be made more explicit. The current Fig. 1 illustrates the coupling among long-term observations, the process-based CE-QUAL-W2 model, and the XGBoost/SHAP framework. However, we agree that its role as a stepwise workflow is not sufficiently clear. In the revised version, we have redesigned Fig. 1 to emphasize the sequential workflow of the study. Specifically, the figure has been reorganized to show the progression from long-term hydro-meteorological and temperature monitoring, CE-QUAL-W2 process-based simulation and gap filling, XGBoost model training and testing and SHAP-based interpretation of dominant driving factors controlling reservoir thermal structure. This revised layout makes the connection between data sources, model development, validation, and interpretation more explicit and easier for readers to follow. Please see the revised Figure 1 and caption below.

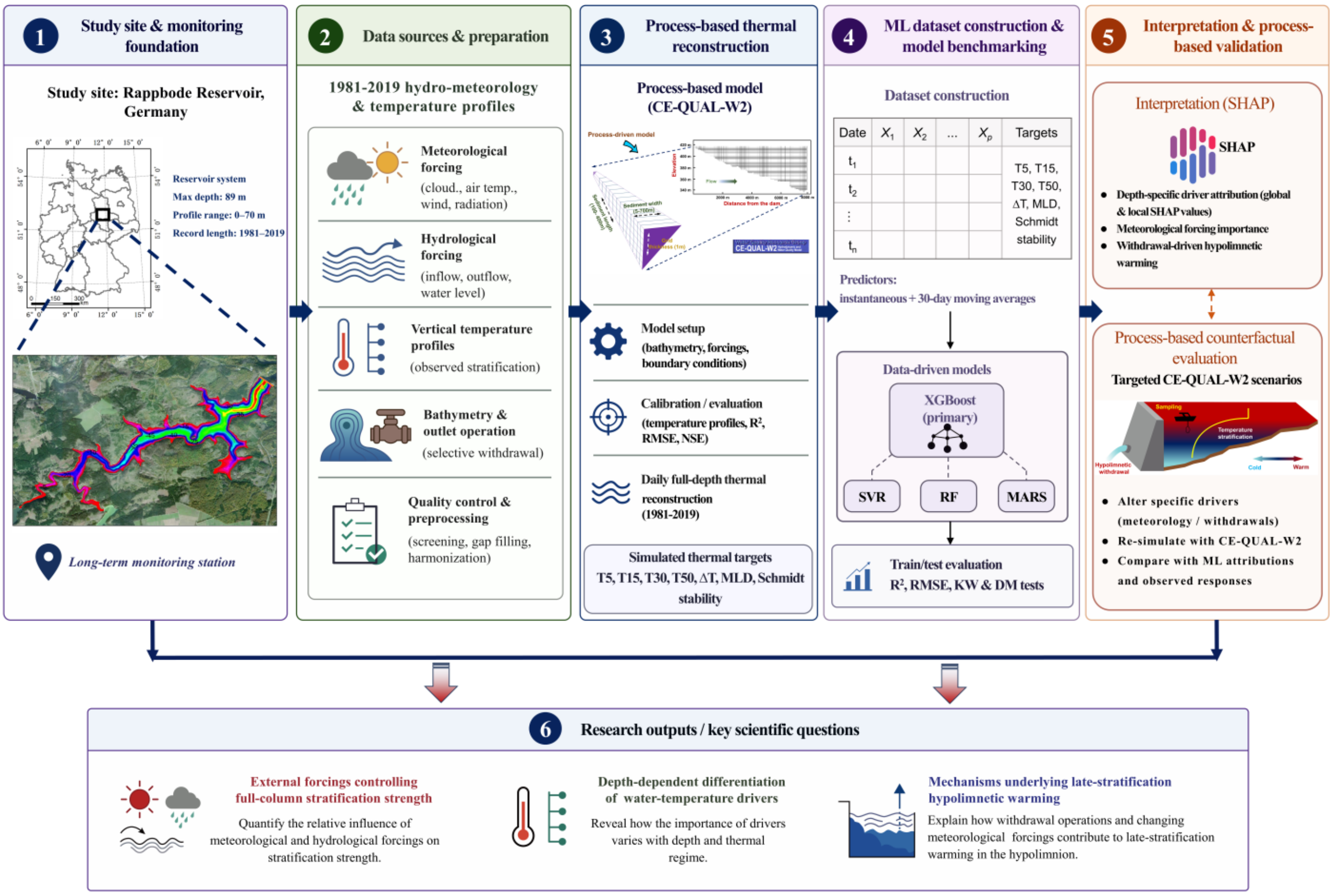

Fig. 1. Study site and methodological workflow of the integrated CE-QUAL-W2 and machine-learning framework. The workflow links long-term monitoring and hydro-meteorological forcing with CE-QUAL-W2 thermal reconstruction, data-driven model benchmarking, SHAP-based attribution, and targeted process-based validation. The final outputs address stratification-strength controls, depth-dependent thermal drivers, and mechanisms of episodic hypolimnetic warming.

3.5. Incomplete interpretability analysis. Since the authors devoted a significant portion of the manuscript to SHAP-based analysis (SHAP), both global and local interpretability analyses should be conducted. Currently, the local interpretability component is missing.

We fully agree with the reviewer that the local interpretability component should be made more explicit. The current manuscript provided extensive global SHAP analyses across the seven thermal indicators (Figs. 6–7) and stage-specific

SHAP analyses for bottom-water temperature (Fig. 9a–d), with Fig. 9d showing sample-specific temporal variation in the three dominant late-stratification drivers. To address the reviewer’s concern, we further add a supplementary figure (attached below, the new Fig. S6) showing explicit local SHAP decompositions for two representative late-stratification bottom-water temperature cases. Specifically, the anomalous warming case corresponds to 2007-10-02, which exhibited the highest bottom-water temperature among the four anomalous years, whereas the reference case corresponds to 1988-09-11, whose bottom-water temperature was closest to the median of the non-anomalous late-stratification cases. These local explanations show that the anomalous event was driven overwhelmingly by strong positive contributions from antecedent and contemporaneous outflow, whereas the reference case lacked such a concerted positive withdrawal signal and remained close to the normal late-stratification temperature range. These details have been added to the Results section (Section 3.3). We believe this targeted addition directly addresses the reviewer's concern by complementing the existing global SHAP analysis with explicit single-event attribution.

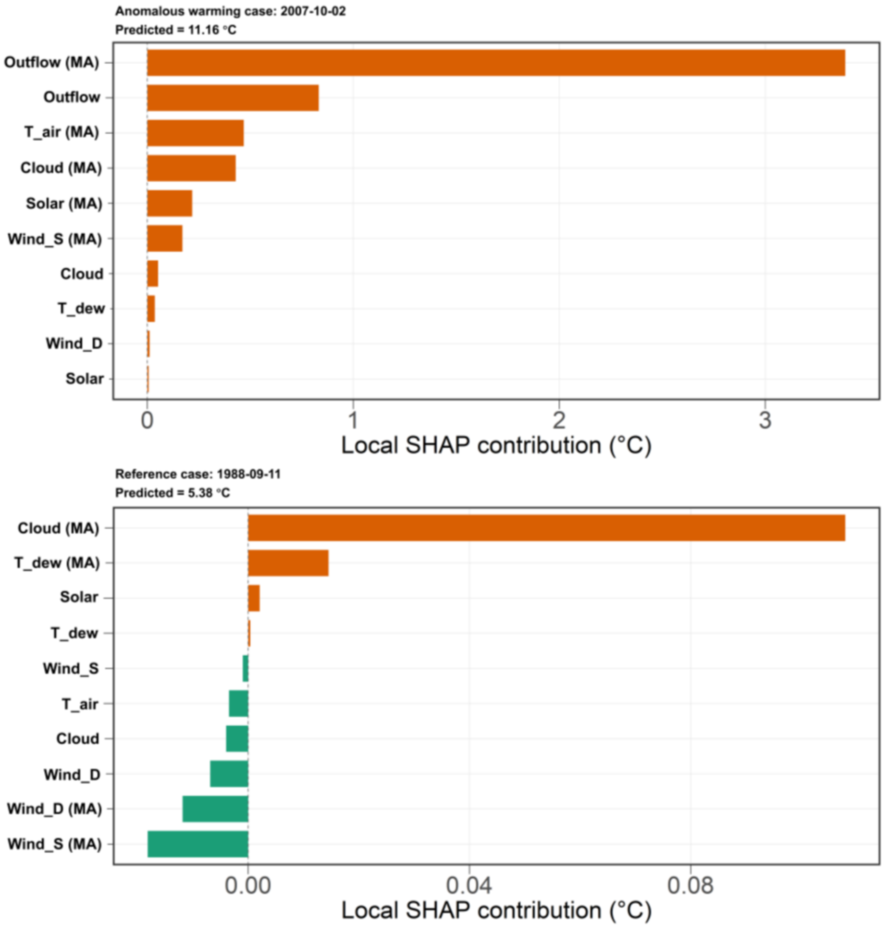

Fig. S6. Local SHAP decompositions for two representative late-stratification cases of bottom-water temperature at 50 m. The anomalous warming case (2007-10-02) was selected as the highest bottom-water-temperature day among the four anomalous years (1981, 2001, 2002, and 2007), whereas the reference case (1988-09-11) was selected as the non-anomalous late-stratification day whose bottom-water temperature was closest to the median of all non-anomalous cases. Bars show the contribution of each predictor to the model prediction relative to the base value; positive and negative SHAP values indicate upward and downward effects on predicted bottom-water temperature, respectively.

4. Additional explainability analyses. To improve the understanding of feature importance and interactions, the study should incorporate complementary explainability techniques, including Permutation Feature Importance (PFI) for global

ranking, Partial Dependence Plots (PDP) for marginal effects, and Individual Conditional Expectation (ICE) plots to capture response heterogeneity. Including these analyses would significantly strengthen the scientific contribution of the work.

We highly appreciate this constructive suggestion and agree that complementary explainability analyses can strengthen the interpretation of the model results. In response, we retained SHAP as the primary interpretability framework, because it is central to the attribution logic of this study, but we have now added a targeted set of supplementary diagnostics that directly address the reviewer's comment. Specifically, we added permutation feature importance (PFI) analyses for the depth-specific temperature models at 5, 15, and 30 m, as well as for the late-stratification bottom-water temperature model at 50 m. In addition, for the late-stratification 50 m model, which is the most mechanistically important target in this manuscript, we added partial dependence plots (PDPs) and individual conditional expectation (ICE) plots for the three dominant drivers, namely antecedent outflow, antecedent air temperature, and antecedent shortwave radiation.

These additional analyses provide further support for the conclusions derived from the SHAP analyses (see Section 3.2). Across the depth-specific temperature models, the PFI results confirm the same vertical reordering of dominant controls inferred from SHAP (see the newly added Fig. S4): the 5 m model is governed primarily by antecedent atmospheric forcing, whereas shortwave radiation becomes dominant below the surface and the importance of antecedent outflow increases with depth, reaching a prominent role by 30 m. For the late-stratification 50 m bottom-water temperature model, PFI again identifies antecedent outflow as the dominant predictor, in agreement with the SHAP results, while antecedent air temperature and shortwave radiation remain secondary controls (see the newly added Fig. S4). The PDP and ICE curves further clarify the nature of these effects by showing a strong positive and nonlinear response of bottom-water temperature to increasing antecedent outflow, in contrast to the weaker and more heterogeneous effects of antecedent air temperature and shortwave radiation (see the newly added Fig. S5, also attached below).

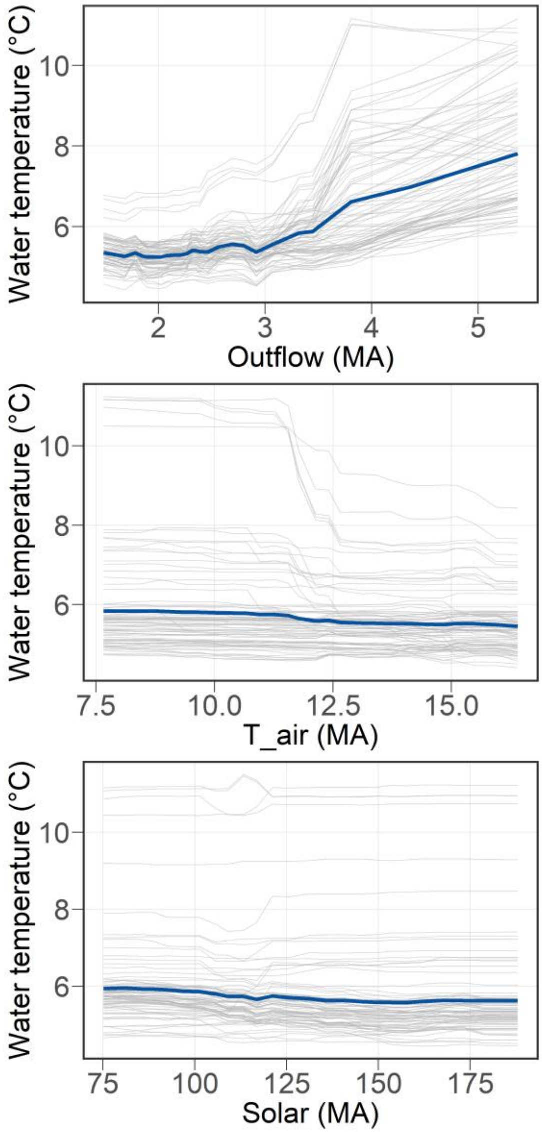

Fig. S5. Partial dependence plots (blue lines) and individual conditional expectation curves (gray lines) for the three dominant predictors of late-stratification bottom-water temperature at 50 m, namely antecedent outflow (top), antecedent air temperature (middle), and antecedent shortwave radiation (bottom). The panels show the modeled marginal response of bottom-water temperature to each predictor and the corresponding sample-specific heterogeneity.

5. Statistical comparison of models. Once a robust baseline comparison is established among several models, the authors should perform appropriate statistical tests such as the Kruskal-Walli's test and the Diebold-Mariano test to assess whether differences in predictive performance between models are statistically significant.

We appreciate this important suggestion and agree that, once additional benchmark models are introduced, the comparison should not rely solely on point estimates of predictive performance. Following the reviewer's recommendation, we have added rigorous statistical comparisons based on the test-set prediction errors of the four data-driven models. Specifically, for each of the seven target variables, we applied the Kruskal-Wallis test to assess whether overall differences in error distributions were present among XGBoost, SVR, RF, and MARS, and we further performed pairwise Diebold-Mariano tests to compare XGBoost against each benchmark model.

These additional analyses substantially strengthen the benchmark comparison. The Kruskal-Wallis tests indicate significant overall differences among the four models for all seven target variables. The Diebold-Mariano tests further show that the predictive advantage of XGBoost over SVR and MARS is statistically significant across all seven targets. Relative to RF, the advantage of XGBoost is statistically significant for water temperature at 15 and 50 m, mixed-layer depth, and Schmidt stability, but not for water temperature at 5 and 30 m or for the bottom-to-surface temperature difference. This pattern is fully consistent with the revised benchmark tables: XGBoost delivers the best point performance across all targets, as indicated by the highest test R^2 values and lowest test RMSE values, while its margin over RF is comparatively small for a subset of variables.

To make these results transparent, we have added a new supplementary table (Table S5) reporting the Kruskal-Wallis and Diebold-Mariano *p*-values for all seven target variables. We believe such addition directly addresses the reviewer's concern by demonstrating that the benchmark comparison is supported not only by descriptive performance metrics, but also by formal statistical evidence.

Reply on RC2

Mi and colleagues used a long data record of thermal profiles from Rappbode reservoir to fit a two-dimensional hydrodynamic model, and then used a machine learning to link patterns in temperature and stratification dynamics to potential causes. The study is well-written, clear, and the methodology is reasonably novel and could be applied outside the presented study case. The study suits the audience well and could eventually be published, but I have two major issues that need to be addressed first. I list them below, including some additional minor comments further down.

We sincerely thank the reviewer for the constructive assessment of our manuscript. We greatly appreciate the reviewer's positive evaluation of the manuscript's novelty and broader applicability, as well as the identification of two key issues that require strengthening before the study can be considered further. We fully agree that these points are central to the credibility of the manuscript, particularly with respect to the physical grounding of the ML-based attribution. In the revised manuscript, we have addressed these concerns through targeted additional analyses, clearer methodological documentation, and more cautious wording where needed. Please find below our detailed point-by-point responses and the corresponding revisions.

My first main comment relates to the aim of coupling the machine learning (ML) model XGBoost to the process-based (PB) model CE-QUAL-W2. The ML model is trained on the PB model and is employed specifically to investigate external drivers and causal mechanisms (e.g. L. 32-33). Using data-driven approaches to investigate causality can absolutely be done, but in the present paper, there is no evidence presented on the accuracy of this, although the authors do try to link the outputs from the ML model to processes in the Discussion. Moreover, this could be tested even better than in usual cases, because the PB model that the ML model tries to emulate, has the processes included by design. The authors rightfully argue that the ML-approach primarily gains upon PB scenario runs in terms of time (L. 465-466), but because the novelty of the approach, I argue that the conclusions from the ML model should be validated using PB scenarios first. Some of the causalities highlighted by the ML model (e.g. increasing importance of shortwave radiation with depth; importance of withdrawal volume for end-of-stratification deep-water temperature increase) can be explored with the PB model as well. This should be done in order to build trust in the accuracy of the presented method. If the conclusions between the two methods deviate, the underlying reason should be discussed. The current method has a strong risk of mixing correlation and causality (for instance in telling apart the influence of air temperature and shortwave radiation), and this risk should be mitigated by validation using the PB model. For example, the increasing importance of shortwave radiation with depth is surprising, and this validation could shed light on the reliability. This validation exercise could be presented in the supplementary material and referred to in the main text.

We fully agree that the original manuscript does not yet provide sufficiently explicit physical support for the machine learning-derived attributions. This is especially critical, as the novelty of the framework rests on connecting interpretable machine learning with process-based models, rather than employing ML as a purely predictive black box. In addition, we agree with the reviewer that the manuscript does not yet distinguish sufficiently between statistical attribution and causal inference. In this study, XGBoost/SHAP was intended as a fast, depth-resolved attribution layer built on a physically consistent CE-QUAL-W2 reconstruction, rather than as a stand-alone tool for causal demonstration. The reviewer is therefore correct that the strongest ML-derived inferences should be checked more directly against the process-based framework.

To address this concern, we have strengthened the manuscript in three ways. First, we moderated causal wording throughout the manuscript so that the manuscript distinguishes more clearly among statistical attribution, mechanistic interpretation, and direct process-based validation. Second, to test the SHAP-derived result that antecedent shortwave radiation becomes the highest-ranked meteorological predictor below the surface layer, we added a targeted CE-QUAL-W2 counterfactual analysis in the Supplement. This analysis was designed to evaluate relative forcing importance within each depth-specific temperature model, rather than to imply that shortwave radiation directly heats deeper layers more strongly than the epilimnion.

Specifically, starting from the calibrated baseline simulation, we constructed four anomaly-removal scenarios. In each scenario, only one SHAP-identified key meteorological factor influencing the thermal structure, namely shortwave radiation, air temperature, dew-point temperature, or cloud cover, was modified. For every year during 1st April-31st October, the daily value of the target forcing was replaced by its 1981-2019 long-term median for the same day of year, while all other meteorological and hydrological forcings were kept unchanged. The replacement window began on 1 April because the ML predictors include 30-day antecedent moving-average terms, and the model response was evaluated over May-October, consistent with the SHAP analysis. To evaluate whether the SHAP-derived meteorological rankings were consistent with the process-based model responses, we designed a coupling-weighted counterfactual response index (CWCRI) for each depth (z) and forcing (k):

$$CWCRI_k(z)=nRMSE_k(z)\times|\rho(T_{BASE,z},X_{k,30d})| \quad (1)$$

where $nRMSE_k(z)$ represents the standardized W2 temperature response to removing the anomaly of forcing k, and $|\rho(T_{BASE,z},X_{k,30d})|$ represents the historical coupling between baseline W2 temperature and the corresponding 30-day moving-average forcing used in the SHAP model. This indicator evaluates the relevance of each meteorological forcing by integrating two components: the temperature response simulated after removing its anomaly and its historical coupling with temperature variability at each depth. Therefore, a higher CWCRI indicates stronger process-consistent influence of the given meteorological forcing on simulated temperature at a given depth, reflecting both a larger standardized CE-QUAL-W2 response to anomaly removal and tighter coupling with baseline-simulated temperature variations.

The resulting W2-response ranking closely reproduced the key SHAP-derived pattern: air temperature ranked highest at 5 m, while shortwave radiation ranked highest at subsurface layers. At 15 and 30 m, which directly address the reviewer's concern about the increasing relative importance of shortwave radiation below the surface layer, the W2-response ranking was Solar > T_dew > T_air > Cloud, identical to the SHAP ranking in Fig. 6. This agreement indicates that the elevated SHAP importance of antecedent shortwave radiation in the subsurface temperature models is also reflected in the independent CE-QUAL-W2 counterfactual response. The full calculation and rank comparison are provided in Table S6 (also attached below).

Table S6. Coupling-weighted CE-QUAL-W2 counterfactual response index (CWCRI) for meteorological forcings across representative depths. MA indicates moving-average features.

Target variable	T_air (MA)	T_dew (MA)	Solar (MA)	Cloud (MA)
Water temp (5 m depth)	0.209	0.119	0.173	0.027
Water temp (15 m depth)	0.040	0.065	0.153	0.002
Water temp (30 m depth)	0.022	0.023	0.159	0.016
Water temp (50 m depth)	0.007	0.040	0.144	0.011

Furthermore, we have added two W2 withdrawal counterfactual scenarios to test whether the SHAP-derived dominance of outflow discharge during late stratification is consistent with the process-based model response. The analysis focused on the four years with anomalously elevated late-stratification hypolimnetic temperature (1981, 2001, 2002, and 2007). The withdrawal perturbation was applied from 1st August to 31st October, while the thermal response was evaluated from 1st September to 31st October. This design accounts for the 30-day antecedent discharge term used in the SHAP analysis while matching the timing of the observed hypolimnetic warming events. The reference withdrawal statistics were calculated for each calendar day using all years except the four anomalous years. In the P50-normalized scenario, bottom withdrawal from 1st August to 31st October at the four years was replaced by the corresponding non-anomalous calendar-day median. In the P90-normalized scenario, bottom withdrawal was replaced by the corresponding non-anomalous calendar-day 90th percentile. In both scenarios, the difference in bottom withdrawal relative to the baseline was redistributed to the upper outlet, so that total daily outflow remained identical to the baseline simulation. All meteorological and inflow boundary conditions were kept unchanged. The results showed that reducing anomalously high bottom withdrawals strongly suppressed late-stratification warming at 50 m. Across the four anomalous years, mean September-October 50 m temperature decreased from 8.05 °C in the baseline simulation to 6.30 °C in the P90-normalized scenario and 6.00 °C in the P50-normalized scenario. The corresponding maximum 50 m temperature decreased from 11.37 °C to 8.43 °C (P90) and 6.99 °C (P50). These CE-QUAL-W2 counterfactuals provide process-based support that intensified bottom withdrawals substantially amplified bottom-water warming, consistent with the SHAP-derived outflow dominance during late stratification.

All scenario details, diagnostic metrics, and results have now been added to the supplementary material and are referenced in the revised manuscript (S4.1 and S4.2). We believe that these targeted W2 validation exercises greatly strengthen the robustness and physical credibility of the SHAP-based attribution, because the two key SHAP signals highlighted by the reviewer are now supported by independent process-based counterfactual scenarios. We sincerely thank the reviewer for this important suggestion, which helped us clarify the distinction between statistical attribution and process-based mechanistic interpretation and strongly improved the evidential basis of the manuscript.

My second main comment relates to the lack of information on the calibration and validation, primarily of CE-QUAL-W2. What periods were used for training and validation, what parameters were adjusted, what were the calibration targets, what method was used? The terms “site-validated” (L. 124) and “site-calibrated CE-QUAL-W2 model” (L. 457) suggest that this was done before, but it is not clearly indicated where this information can be found. If a model from another publication is used, this should be clearly stated in the data availability statement.

We fully agree with the reviewer that the original manuscript did not describe the CE-QUAL-W2 calibration and evaluation procedure with sufficient details in the main text. The original manuscript included a brief description of the calibration variables and the full parameter set in Section S1 of the supplementary material, and we agree that these details should be stated explicitly in the Methods section of the main text. In the revised manuscript, we have expanded Section 2.3 in response to the reviewer's suggestion to clarify the relevant methodological details. Specifically, the W2 model was not "trained" in the machine-learning sense. It was manually calibrated through a trial-and-error procedure against the observed vertical thermal structure of Rappbode Reservoir over the 1981–2019 period. The calibration focused on site-specific optical and meteorological exposure parameters, namely Dynamic shading coefficient (SHADE), Wind sheltering coefficient (WSC), and Extinction coefficient for pure water (EXH2O). The final values were SHADE=1, WSC=1, and EXH2O=0.55 m^{-1}, while all remaining parameters were retained at standard CE-QUAL-W2 settings because they are physically based and not typically subject to site calibration (Cole and Wells 2006). The main calibration targets were the temperatures at different depths, and water level was used as an additional system-scale check of the water-balance representation. No separate split-sample validation period was defined. Instead, the full 1981-2019 observational record was used for continuous calibration/evaluation across a wide range of hydroclimatic conditions. This choice follows the calibration guidance in the official CE-QUAL-W2 User Manual (Wells 2026), which emphasizes evaluating model behavior across broad observed conditions rather than relying on an arbitrary calibration/verification split. We now state this explicitly and avoid implying an independent validation period where none was used. We also revised "site-validated" and "site-calibrated" to more precise wording, namely "calibrated and evaluated," and clarified that the present application builds on earlier Rappbode CE-QUAL-W2 configurations (Mi et al. 2023; Mi et al. 2020) but was re-run and evaluated for the 1981–2019 dataset used in this study. The calibrated parameter set remains provided in Table S1.

Minor comments:

Title: "event-scale hypolimnetic warming" is not clear. Suggest to change to "episodic hypolimnetic warming" as you did in the abstract (or another term that you find more suitable).

Agree. The title has been revised to "Four decades of full-depth profiles reveal layer-resolved drivers of reservoir thermal regimes and episodic hypolimnetic warming" following the reviewer's suggestion.

L. 37-38: change to "by the 30-day antecedent moving average of shortwave radiation and air temperature" (or similar).

This sentence in the Abstract has been modified according to the reviewer's suggestion. Thanks!

L. 40: change to "episodic hypolimnetic warming by up to 10 °C…"

We have revised the text following this suggestion.

L. 42: use of "compensatory-flow pathway" not clear in this context.

We agree that "compensatory-flow pathway" was not sufficiently clear in the Abstract. We replaced it with a more explicit physical description: intensified deep withdrawals weakened the density gradient and promoted downward transport of warm upper-layer water into the hypolimnion.

L. 56: I could not find information on the effect of vertical temperature structure on temperature-sensitive organisms in Carr et al., although some of the references in there might have looked at this. Please assess if this citation is appropriate or cite the original source if possible.

Good point! In the revised manuscript, Carr et al. (2019) has been replaced with Kraemer et al. (2021) to support this statement.

L. 72-73: The GLTC initiative is mentioned but not used further in the manuscript. Suggest to remove this (but keep the Sharma et al. reference and rest of the sentence).

We agree with the reviewer's point. In the revised manuscript, we have removed the mention of the GLTC initiative while retaining the relevant citation and associated sentence.

L. 97-100: I think that "temperate" is a too broad term here to say that hypolimnetic temperatures are around 4-6 degrees. Boehrer & Schultze also do not support the generality of this rule for the temperate climate zone. Suggest to change to lakes that cool to the temperature of maximum density.

We agree that the original wording overgeneralized this pattern to temperate lakes and reservoirs. We therefore revised the sentence to refer specifically to deep, seasonally stratified systems that cool toward the temperature of maximum freshwater density during winter circulation. We also retained Boehrer and Schultze (2008) only for the underlying stratification physics and added Lewis Jr et al. (2019) and Schwefel et al. (2025) to better support the cold and relatively stable hypolimnetic thermal regime relevant to this study.

L. 108-110: Bouffard et al. (2013) does not seem to provide support for the general statement that biological and chemical reaction rates roughly double per 10 °C. Please find a more appropriate reference or change the statement to be more representative of the findings in that paper.

We agree that Bouffard et al. (2013) does not directly support the broad statement that biological and chemical reaction rates roughly double per 10 °C.

We therefore removed this general Q10-style formulation and revised the sentence to focus more specifically on hypolimnetic warming, temperature-sensitive mineralization, and sediment–water solute exchange. We also replaced the citation with more targeted lake/reservoir studies, including Dadi et al. (2023) and Nkwalale et al. (2023), which better support the link between warmer bottom waters, oxygen depletion, and enhanced biogeochemical fluxes.

L. 112 (and elsewhere, such as L. 375): General editing comment that there should be spaces around the dash

We agree and have checked the manuscript for dash spacing. Where the dash was used to introduce an explanatory clause, we either added spaces around it or replaced it with a colon/semicolon for clearer scientific prose.

L. 147-150: What data did you use for the depth-varying outflow discharge? Was this based on data from the reservoir operator?

We thank the reviewer for requesting clarification. The outflow discharge and the corresponding withdrawal elevation information were provided by the Rappbode Reservoir authority (Talsperrenbetrieb Sachsen-Anhalt). We have revised Section 2.3 to make clear that the outflow boundary was prescribed directly from operator records and was not inferred from the simulated temperature field or adjusted during calibration.

L. 185: “an monitoring buoy” -> “a monitoring buoy”

We have fixed this point in the revised version, thanks!

L. 262-266 (and links to comment on L. 147-150): In this part you show an impressive R2 value for water level (0.99), but without knowing more about the data sources (see previous comment on L. 147-150) and optional calibration/validation (e.g. added or scaled inflows/outflows to better fit the water level) (see 2nd main comment), it is difficult to judge the extent to which this builds trust in the model.

We fully agree with the reviewer that the water-level result required clearer contextualization, particularly with respect to the hydrological data sources and the treatment of residual water-balance terms. We have expanded Section 2.3 to clarify that inflow discharge, inflow temperature, outflow discharge, and the prescribed outlet elevation were provided by the Rappbode Reservoir authority and used directly as model boundary conditions. To close the reservoir water balance, residual inflows required to match the observed water-level evolution were represented as distributed tributary inflows, following the CE-QUAL-W2 water-balance framework. These residual inflows were used only for water-balance closure and were distributed along the main reservoir branch to minimize local thermal impacts. The recorded outflow discharge and prescribed outlet elevation were not scaled, inferred from the simulated thermal structure, or adjusted during temperature calibration.

L. 276-279: Add something like "…at increasing depth" at the end of the sentence.

We agree and revised the sentence to make the depth dependence explicit. The revised text now states that the intra-annual temperature amplitude diminished sharply with depth, exceeding 15 °C at 5 m but declining to <10 °C at 30 m and <8 °C at 50 m.

L. 292: It is very important here that XGBoost was compared to W2, NOT to observations. The term "error" might be a confusing concept here, so I would underline this by explicitly saying "the RMSE between W2 and XGBoost".

Good point! We have clarified explicitly in the revised manuscript that the reported RMSE here refers to the difference between CE-QUAL-W2 and XGBoost, not between XGBoost and observations.

L. 301: I wouldn't describe the mixed depth as a "tight clustering along the 1:1 line". It is better to quantify this by showing the R2 value (both in the figure and the text).cxe

We agree with the reviewer and have revised this part accordingly. In the revised manuscript, we removed the qualitative description of mixed-layer depth and now quantify the agreement explicitly. The Results section reports the R^2 and RMSE for all stratification metrics, as summarized in Tab. 1. In addition, Fig. 5 now displays the R^2 values directly in the corresponding comparison panels, so that the agreement between XGBoost and CE-QUAL-W2 outputs is quantitatively shown in both the text and the figure.

L. 311: I think “decreases”, “reduces” or “diminishes” are clearer terms than “attenuates” in this context.

We agree and have replaced “attenuates” with “decreases” in this place.

L. 367: “combinng” -> “combining”

Thank you for pointing this out. We have corrected this typo in the revised manuscript.

L. 371: “bridges an inter-decadal gap” could suggest a gap-filling exercise. I suggest to change the wording.

We agree that “bridges an inter-decadal gap” could be misread as referring to a gap-filling exercise. We revised the wording to avoid this ambiguity and now state that the 39-year full-profile archive offers a robust empirical basis for multi-decadal hydrodynamic modelling.

L. 376-377: the part “and the pan-European synthesis of temperate lakes” in not integrated with the rest of the sentence; please rewrite.

We agree that the original wording did not integrate the comparison with previous lake-warming syntheses clearly. We rewrote the sentence to separate the comparison with the regional air-temperature trend from the comparison with large-scale lake-temperature syntheses.

L. 383: “enhances” -> “enhance”

This has been corrected in the revised manuscript. Thanks!

L. 400: According to the equation and the provided units, shouldn't the unit of thermal inertia be time per Kelvin, instead of only time? Currently the units do not check out (though I guess a ΔT of 1 K is assumed). I did a quick scan of Imberger & Patterson but could not find this formula; please reply with the equation number in Imberger & Patterson or alternatively how this equation and the units were derived from the reference.

We highly appreciate the reviewer for this important observation. We agree that the original formulation was dimensionally incomplete: $\rho c_p h/H_{net}$ has units of time per Kelvin, not time. We have revised the equation to make explicit that the estimate refers to the characteristic timescale for a 1 K mixed-layer temperature response, i.e., $\tau_{1K} = \rho c_p h \Delta T/|H_{net}|$, with $\Delta T = 1$ K. We also removed the implication that this exact equation comes from Imberger and Patterson (1989). The revised text now describes the estimate as derived from the heat budget of the mixed layer and following the same physical basis as the lake temperature formulation of Piccolroaz et al. (2013). This revision clarifies both the dimensional interpretation and the reference basis.

L. 424-427: Can you please clarify how increased vulnerability to mixing events relate to the importance of deep-layer withdrawals?

We thank the reviewer for noting this ambiguity. Our intention was not to treat mixing events as an independent mechanism separate from deep-layer withdrawals. We have revised Section 4.3 to clarify that late-season weakening of stratification provides the background condition under which bottom withdrawals become thermally more effective. Specifically, reduced Schmidt stability lowers the buoyancy resistance across the metalimnion, so bottom-withdrawal forcing can generate stronger withdrawal-induced entrainment and compensatory downward transport of warmer upper-layer water. Thus, the increased vulnerability to mixing refers to the late-season susceptibility of the density structure to withdrawal-induced vertical exchange, rather than to an additional driver.

L. 427: “event” -> “events”

We agree. During the revision of this paragraph to clarify the relationship between late-season stratification weakening and withdrawal-induced vertical exchange, the original phrase containing “mixing event” was removed. The revised text no longer contains this singular/plural inconsistency.

L. 428-435: I missed the argument that a significantly lowered hypolimnetic volume if the mixed layer deepens. Water withdrawn from a smaller hypolimnion will more quickly lower the thermocline. This could be quantified using hypsographic information.

That is a good suggestion! We fully agree with the reviewer that mixed-layer deepening also reduces the remaining deep-water volume and that this volumetric effect should be made explicit. We therefore quantified the water volume below the mixed-layer base using the W2 volume-elevation relationship, daily water level, and daily mixed-layer depth. The revised Section 4.3 now reports that mean mixed-layer depth increased from 6 m in May-June to 18.0 m in September-October, while the mean water volume below the mixed-layer base decreased from 8.78×10^7 to 4.67×10^7 m^3, a 46.8% reduction. This addition clarifies that, during late stratification, a given bottom-withdrawal discharge acts on a much smaller deep-water volume, increasing the withdrawal-to-volume ratio and strengthening the physical explanation for the enhanced sensitivity of bottom-water temperature to deep withdrawals.

Figure 8: the x-axes are not aligned.

We have aligned the x-axes in the revised manuscript. Thanks!

Reply on RC3

Greetings. I have revised the manuscript entitled 'Four decades of full-depth profiles reveal layer-resolved drivers of reservoir thermal regimes and event-scale hypolimnetic warming. The paper deals with Germany's largest drinking-water reservoir. By comparing two modelling approaches: the CE-QUAL-W2 numerical model and a machine learning XGBoost-based model, the goal is to predict the thermal structure of a reservoir, assessing the performance of each method. The paper is suitable for publication in the HESS journal, whereas it presents some lacks to be fixed before going further down the publication way. These can be listed as follows:

We sincerely thank the reviewer for the careful assessment of our manuscript and for considering the study potentially suitable for publication in HESS after revision. We appreciate the reviewer's comments on methodological framing, physical system description, and the presentation of long-term variability. These points are helpful, and we agree that the manuscript can be strengthened by better positioning the XGBoost–SHAP approach within the broader methodological literature, and by making the CE-QUAL-W2 system setup more explicit in the main text. Please find below our detailed point-by-point responses and the corresponding revisions.

1. The paper only presents one type of Machine Learning approach and one kind of model. There is a need for at least a small paragraph that frames, in a general way, machine learning approaches. Moreover, there is no presence of reference to Genetic Algorithms and Metaheuristics ones, that have been proved to outperform other ML methods (see Rajwar et al., 2023; Schiavo and Pedretti, 2026). This does not mean that these approaches should be implemented, nor that the paper should compare these results with the offered ones, but at least citing their existence and weight in the literature framework, and clearly explaining why the XGBoost method has been preferred to Genetic algorithms.

We fully agree that the original manuscript did not sufficiently situate

XGBoost within the broader family of machine-learning and optimization-based approaches. To address this point, and in response to Reviewer 1's request, the revised manuscript now provides a broader methodological framing and, importantly, includes new benchmark comparisons with three additional data-driven models: SVR, Random Forest, and MARS. These benchmark models were trained and evaluated using the same predictor set, train-test split, and performance metrics as XGBoost, thereby placing the XGBoost results in a clearer methodological context.

In the revised manuscript, we have expanded Section 2.4 by adding a concise methodological framing paragraph that acknowledges regression-based methods, kernel methods, tree ensembles, and metaheuristic optimization approaches. In the revised Section 2.4, we also cite the reviewer-suggested references by Rajwar et al. (2023) and Schiavo and Pedretti (2026) to acknowledge the importance of genetic algorithms and related metaheuristics in search, optimization, parameter calibration, and hybrid Earth-science modelling. In the present study, XGBoost was selected because the central objective was interpretable, depth- and phase-resolved attribution of reservoir thermal drivers from tabular hydro-meteorological predictors. XGBoost provides strong predictive performance for nonlinear tabular data, computational efficiency, and direct compatibility with SHAP, which is essential to the attribution framework used here. We now state this rationale more explicitly in the revised manuscript.

2. Some recent literature works assessed how, for some earth sciences applications, ‘usual’ geostatistical methods (e.g. kriging or co-kriging-based ones) still perform better than Neural Network algorithms or XGBoost methods (Brckovic et al., 2025), or at least they are good enough at predicting the target variable that they should be involved in ML algorithms (Grey et al., 2025) to achieve reliable predictions. I suggest discussing this point further.

We highly appreciate the reviewer for raising this important and insightful point! We agree that geostatistical methods remain widely used in Earth-science prediction, particularly where spatial covariance, interpolation among monitoring sites, or prediction at unsampled locations is central. In response, we have revised Section 4.4 to explicitly acknowledge the value of kriging, co-kriging, and hybrid frameworks integrating geostatistics and machine learning, and we now cite the reviewer-suggested studies by Brcković et al. (2025) and Grey et al. (2025).

We also clarified the methodological scope of the present study. Our analysis focuses on depth-resolved temporal attribution at a fixed long-term reservoir monitoring station, using a physically consistent CE-QUAL-W2 reconstruction and a prescribed set of hydro-meteorological predictors, rather than on spatial interpolation across monitoring locations. Therefore, geostatistical methods were not implemented in the present framework. We now discuss these methods as valuable future extension (see Section 4.4), especially for applications involving multi-station observations, spatially distributed reservoir monitoring, or reservoir-network-scale thermal prediction.

3. I am not sure that a strong enough system conceptualization is given. Thus, I am a bit reluctant about the goodness of the results without grounding them to physically-based description, meshing, and imposition of boundary conditions. Maybe the Authors can clarify this point, underlining the (i) assumptions they have made, (ii) giving more info about the geological structure, and (iii) providing hydrogeological or geological sections to support their claims. Indeed, a max depth of 80 m (Figure 2) requires strong geological proof of concepts.

We thank the reviewer for this important comment! We fully agree that the original manuscript did not make the physical conceptualization of the CE-QUAL-W2 application sufficiently explicit. We have revised Sections 2.1-2.3 to clarify the geological setting, the bathymetric basis of the model domain, the depth range of the monitoring profiles, and the main physical assumptions of the W2 application.

Specifically, we have added a short description of the physical setting, noting that the reservoir is impounded within a narrow valley in the Harz Mountains and that the regional catchment is underlain mainly by consolidated Paleozoic bedrock, including greywacke, clay schist, and diabase. We also clarify that this geological setting provides the valley framework for the reservoir bathymetry, whereas the hydrodynamic domain in CE-QUAL-W2 is defined by reservoir bathymetry and water-level boundary conditions rather than by subsurface geological sections.

In Section 2.3, we now explicitly state the assumptions made in applying CE-QUAL-W2: lateral variations within each segment are assumed to be secondary to longitudinal and vertical gradients, allowing the reservoir to be represented by a laterally averaged longitudinal-vertical model; groundwater exchange and subsurface hydrogeological flow are not explicitly represented. We also clarified that the maximum depth refers to the bathymetric water-column depth of the reservoir, while the long-term routine temperature profiles used for model evaluation cover the monitored 0-70 m range. These additions provide the physical grounding requested by the reviewer without implying that geological cross-sections are required inputs for the open-water CE-QUAL-W2 hydrodynamic model.

4. Temperature fluctuations seem extremely noisy and subject to historical trends. Why have these not been performed and offered in the results?

We fully agree with the reviewer that the long-term trend behaviour should be presented more explicitly, especially because the deeper temperature records show strong interannual variability. In the revised Results section, we now state more clearly that monotonic trends were evaluated using annual temperature series at representative depths (see Section 3.1 and 4.1). The revised text reports significant warming at 5 m in both observations and simulations, whereas temperatures at 15-50 m show pronounced interannual variability but no statistically significant monotonic trend. We also revised the Discussion to distinguish this multidecadal trend behaviour from the event-scale analysis of episodic hypolimnetic warming. This revision makes clear that the manuscript treats long-term epilimnetic warming and episodic hypolimnetic warming as related but distinct thermal features of the reservoir.

Reference

Cole, T. M. & S. A. Wells, 2006. CE-QUAL-W2: A two-dimensional, laterally averaged, hydrodynamic and water quality model, version 3.5.

Mi, C., T. Shatwell, X. Kong & K. Rinke, 2023. Cascading climate effects in deep reservoirs: Full assessment of physical and biogeochemical dynamics under ensemble climate projections and ways towards adaptation. Ambio 54, 385–401.

Mi, C., T. Shatwell, J. Ma, V. C. Wentzky, B. Boehrer, Y. Xu & K. Rinke, 2020. The formation of a metalimnetic oxygen minimum exemplifies how ecosystem dynamics shape biogeochemical processes: A modelling study. Water Research, 115701.

Wells, S. A., 2026. CE-QUAL-W2: A Two-Dimensional, Laterally Averaged, Hydrodynamic and Water Qual-ity Model, Version 4.5. 5 User Manual: Part 5–Model Utilities.