Response Letter

For

Manuscript ID: **egusphere-2025-1983**

"Hybrid Lake Model (HyLake) v1.0: unifying deep learning and physical principles for simulating lake-atmosphere interactions"

Editor:

This manuscript is overall of good quality. However, reviewers have essential concerns on the steps of methods and robustness. For instance, the developed method relies on training data from a single site without cross-validation and provides insufficient explanation of model development and validation. In addition, the writing is unclear and inconsistent, with repetitive and confusing terminology, an inadequate abstract, and numerous sentences that are misleading or difficult to understand. The reviewers have provided detailed feedback and recommendations. Please carefully address all their concerns before resubmitting the manuscript.

Response: We thank the Editor and four anonymous reviewers for their comments and suggestions. We apologize for resubmitting the response due to our misunderstanding of the Geoscientific Model Development submission process. Changes have been made in the revised manuscript based on the comments provided by the four reviewers. The manuscript has been improved to meet the standards of Geoscientific Model Development. The reviewer's comments are all accepted and **Relisted in black**, followed by our **Replies in blue** and **Revisions in red (highlighted revisions in bold)**. Before point-by-point responses, we would like to address the editorial thresholds as follows:

(1) Additional evidence on model validation and intercomparison

To show the model transferability, we intercompare the computational efficiency among all models, cross-validate the performance of LSTM-based surrogates that were trained with different lake sites, and applied HyLake v1.0 to Lake Chaohu, another inland lake with distinct characteristics.

(a) We applied the established HyLake v1.0 to Lake Chaohu, which is one of several large shallow lakes in the middle and lower reaches of the Yangtze River in southeastern China, but with distinct characteristics than Lake Taihu. Simulated results showed that HyLake v1.0 outperformed FLake in lake surface temperature, demonstrating its potential in model transferability. Relevant revisions were listed in Materials and Methodology (Section 2.3.1, Lines 286-289, Lines 308-314), Discussion (Section 4.1, Lines 560-567, Lines 609-611; Section 4.2, Lines 647-648), Table 3, and Figure S7-9.

(b) We compared the computational efficiency of all models and found that HyLake v1.0 has relatively higher costs, although it has better performance than FLake. Supporting results can be found in Table 3.

(c) The cross-validation of LSTM-based surrogates was executed. Specifically, we used observations from 5 lake sites in Lake Taihu to search for optimized BO-BSTM-based surrogates individually and embedded them into physical principles. Results indicated that HyLake v1.0 and its adapted MLW-trained surrogate are more reliable than others. The associated revisions were listed in Discussion (Section 4.2, Lines 660-666).

(2) Revising the abstract and selected misleading statements

We have revised specific terms, such as "evaluation" vs. "validation", "experiments" vs. "models". Here, we used "validation" to assess the model accuracy (e.g., RMSE, MAE, R) while "evaluation" was used to assess models' abilities (e.g., transferability). In the current manuscript, "models" included FLake, PBBM, Baseline, TaihuScene, and HyLake v1.0. "Experiments" represents the models used in different regions or forcing datasets, including MLW, Taihu-obs, Taihu-ERA5 and Chaohu experiment. The associated revisions were listed in the Abstract and Tables 2-3. In addition, selected misleading statements have been revised according to the comments throughout the manuscript.

Reviewer #1:

The manuscript presents HyLake v1.0, a hard-coupled hybrid lake model in which an LSTM surrogate replaces the implicit-Euler surface-temperature solver embedded within an in-house one-dimensional physical backbone. The surrogate is trained at the MLW site on Lake Taihu and then applied to five other sites that differ in both biological characteristics and meteorological forcing. Although the hybrid framework outperforms several process-based and deep-learning-based benchmarks, its validation strategy and treatment of uncertainty require further refinement. Overall, the paper is clearly written and could be suitable for publication after moderate revision.

Response: We thank Reviewer #1 for his/her positive feedback and constructive comments. The comments are all accepted and **Relisted in black**, followed by our **Replies in blue** and **Revisions in red (highlighted revisions in bold)**. According to the comments, we modified the manuscript, particularly on the capacity of HyLake v1.0 from the aspects of model transferability, computational efficiency, and future improvements. Major changes are summarized as follows:

No.	Major Revisions	Important Messages
1	Applied HyLake v1.0 for Lake Chaohu and validated based on MYD11A1 imagery.	HyLake v1.0 outperformed FLake using ERA5 forcing dataset in Lake Chaohu **(Discussion)**.
2	Intercompared computational efficiency between FLake, Baseline, TaihuScene, and HyLake v1.0.	Computational efficiency depends on surrogate architecture. HyLake v1.0 costed fewer than Baseline and TaihuScene but much than FLake in all experiments **(Discussion)**.
3	Cross-validated different BO-BLSTM-based surrogates that were individually trained with five lake site observations.	MLW observations are the most suitable site for training BO-BLSTM-based surrogate, which achieved the best performance in cross-validation **(Discussion)**.

Major comments

1. To address the model's generality, the authors should apply HyLake to at least one morphologically distinct lake or extend the simulation period to include additional years.

Response: Good point. We agree that the capacity of HyLake v1.0's transferability is important. Therefore, we conducted one additional experiment via FLake and HyLake v1.0, and then validated the lake surface temperature (LST) on a morphologically distinct lake - Lake Chaohu, a large, shallow lake in southeastern China (Fig. R1). The average depth and area of Lake Chaohu are 2.7 m and 768 km², which is one of several large shallow lakes in the middle and lower reaches of the Yangtze River, and is densely populated and strongly influenced by human activities (Wei et al., 2022; Zhang et al., 2022). The observation was derived by MYD11A1 MODIS/aqua daily products with 1 km spatial resolution. The ERA5 dataset was used as the forcing dataset to drive the Chaohu experiment. ERA5 forcing dataset used 4 black grids to cover Lake Chaohu, while MODIS observations covered Lake Chaohu by red points in Figure S7.

Our assessments indicated that HyLake v1.0 performed better than the FLake model, showing the promise of applying it to other lakes (Table 3, Figure S8-9). Changes can be found in Materials and Methodology (Section 2.3.1, Lines 286-289, Lines 308-314), Discussion (Section 4.1, Lines 560-567, Lines 609-611; Section 4.2, Lines 647-648), Table 3, and Figure S7-9.

References:

Wei, Z., Yu, Y., and Yi, Y.: Spatial distribution of nutrient loads and thresholds in large shallow lakes: the case of Chaohu Lake, China, *J. Hydrol.*, *613*, 128466, https://doi.org/10.1016/j.jhydrol.2022.128466, 2022.

Zhang, J., Gao, J., Zhu, Q., Qian, R., Zhang, Q., and Huang, J.: Coupling mountain and lowland watershed models to characterize nutrient loading: An eight-year investigation in Lake Chaohu Basin, *J. Hydrol.*, *612*, 128258, https://doi.org/10.1016/j.jhydrol.2022.128258, 2022.

Revision:

"To address the generalization and transferability of HyLake v1.0 in studied (MLW) and ungauged lake sites (DPK, BFG, XLS and PTS) (Table 1), this study further conducted three numerical experiments, **including MLW experiment, Taihu-obs experiment, Taihu-ERA5 experiment, and Chaohu experiment,** using distinct **models** and forcing datasets (Table **2 and 3**), including FLake, Baseline, and TaihuScene for intercomparison." (Section 2.3.1, Lines 286-289)

"Furthermore, this study implemented the HyLake v1.0 into Lake Chaohu, the 5th-largest shallow freshwater lake in China, which has experienced heavy eutrophication and harmful algal blooms (Yang et al., 2020), to assess its transferability to other lakes. A LST dataset in Lake Chaohu was obtained from MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V061 imageries (MYD11A1, https://www.earthdata.nasa.gov/data/catalog/lpcloud-mod11a1-061), which were used to validate the performance of LST derived from HyLake v1.0. The computational efficiency for each 1-time prediction was recorded using a 16G 10-Core Apple M4 processor based on the established HyLake v1.0 model in this study. The training of the above-mentioned surrogates was run using a 24G NVIDIA GeForce RTX 4090 GPU." (Section 2.3.1, Lines 308-314)

References added:

Yang, C., Yang, P., Geng, J., Yin, H., and Chen, K.: Sediment internal nutrient loading in the most polluted area of a shallow eutrophic lake (Lake Chaohu, China) and its contribution to lake eutrophication, *Environ. Pollut.*, 262, 114292, https://doi.org/10.1016/j.envpol.2020.114292, 2020.

"To address issues related to model performance, generalization, and transferability in ungauged locations, three additional numerical experiments, including FLake, Baseline, and TaihuScene, were proposed **for intercomparison and a framework for applying HyLake v1.0 to another lake, such as Lake Chaohu, with a deeper lake depth of 3.06 m and lake area of 760 km^2 (Figure S6, Jiao et al., 2018), to validate the potential capacity of model application.** These experiments were compared using **observed meteorological datasets and ERA5 datasets, then** validated for both spatiotemporal patterns of LST at Lake Taihu **and Lake Chaohu (Tables 2-3). Similarly, ERA5 dataset-derived HyLake v1.0 outperformed FLake in estimating LST (R = 0.97, RMSE = 2.07 °C, MAE = 1.57 °C) in Lake Chaohu, compared to MYD11A1 datasets (Table 3 and Figures S7-9).**" (Section 4.1, Lines 560-567)

References added:

Jiao, Y., Yang, C., He, W., Liu, W. X., and Xu, F. L.: The spatial distribution of phosphorus and their correlations in surface sediments and pore water in Lake Chaohu, China, *Environ. Sci. Pollut. Res.*, 25, 25906-25915, https://doi.org/10.1007/s11356-018-2606-x, 2018.

"HyLake v1.0, developed based on *in situ* observations from Lake Taihu, has been proven to be reliable and rigorously validated in Lake Chaohu (Table 3), demonstrating a faster and more accurate framework for enhancing the understanding of hybrid hydrological modeling." (Section 4.1, Lines 609-611)

"HyLake v1.0 has been applied to Lake Chaohu and achieved superior performance in comparison to the MYD11A1 LST observations, showing a promising way for more applications." (Section 4.2, Lines 647-648)

"**Table 3: Intercomparison of model performance across different experiments conducted in diverse regions with different forcing datasets. Observations from all lake sites (MLW, DPK, BFG, XLS, and PTS) on Lake Taihu, were used to drive models in the Taihu-obs experiment. Bold values in the table present the best-performing model with each group of experiments. Computational efficiency is reported as the runtime for a single simulation.**

Exp	Model	Forcing	R			RMSE			MAE			Efficiency
			LST	LE	HE	LST	LE	HE	LST	LE	HE	(s)
MLW	PBBM	MLW	0.98	0.85	0.89	1.78	38.34	9.37	1.38	23.54	6.22	189.49
	FLake	MLW	0.98	0.82	0.84	1.76	42.73	7.24	1.35	24.76	5.01	**16.40**
	Baseline	MLW	0.96	0.74	0.75	2.71	51.77	14.63	2.11	33.52	9.30	151.46
	HyLake v1.0	MLW	**0.99**	**0.94**	**0.93**	**1.08**	**24.65**	**7.15**	**0.85**	**15.18**	**4.73**	270.21
Taihu-obs	FLake	All sites	0.97	0.61	0.74	2.24	15.46	69.11	1.69	41.95	10.44	**89.00**
	TaihuScene	All sites	**0.99**	**0.82**	0.89	1.52	14.93	43.49	1.23	29.53	10.63	6928.44
	HyLake v1.0	All sites	**0.99**	0.81	**0.90**	**1.36**	**11.19**	**39.20**	**1.03**	**24.79**	**7.88**	2693.23
Taihu-ERA5	FLake	ERA5	0.98	0.63	0.69	1.82	12.31	67.24	1.46	50.94	9.68	**19.60**
	TaihuScene	ERA5	**0.99**	0.68	0.73	1.60	13.00	64.83	1.29	47.78	10.11	652.25
	HyLake v1.0	ERA5	**0.99**	**0.71**	**0.78**	**1.12**	**11.05**	**49.48**	**0.90**	**35.02**	**7.97**	236.78
Chaohu	FLake	ERA5	**0.97**	\	\	2.28	\	\	1.76	\	\	**70.40**
	HyLake v1.0	ERA5	**0.97**	\	\	**2.07**	\	\	**1.57**	\	\	972.83

” (Table 3)

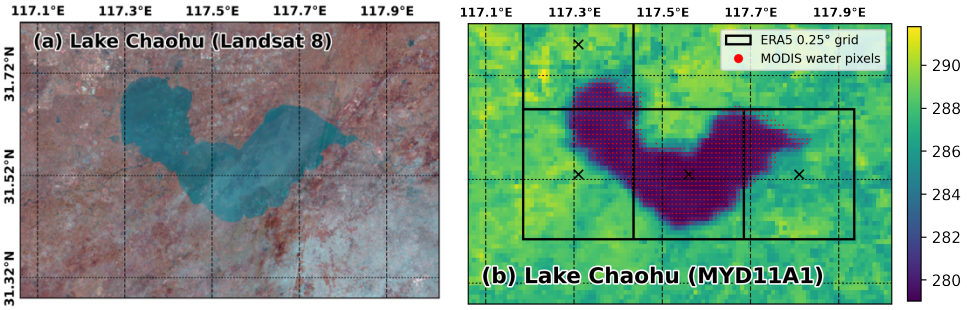
“ 

Figure S7: The locations of Lake Chaohu overlaid on a true-color image from (a) Landsat 8 and daily land surface temperature from (b) MYD11A1 product.” (Figure S7 in Supplementary materials)

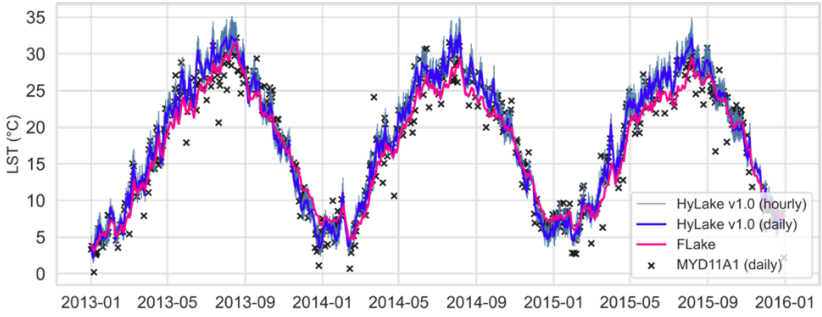
“ 

Figure S8: Time series of daily grid-average LST on Lake Chaohu derived from MYD11A1, FLake simulation, and HyLake v1.0 from 2013 to 2015. HyLake v1.0 provides daily and hourly simulations.” (Figure S8 in Supplementary materials)

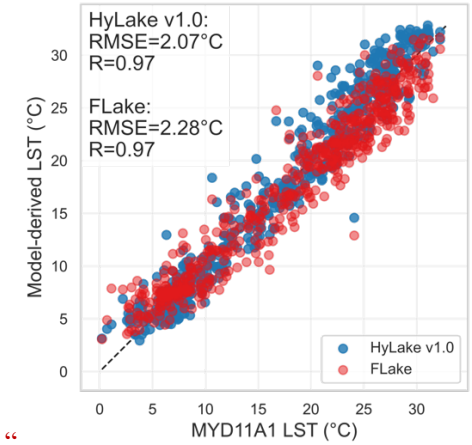
“ 

Figure S9: The intercomparison of daily LST between model simulations (FLake and HyLake v1.0) and MYD11A1 observations on Lake Chaohu from 2013 to 2015.” (Figure S9 in Supplementary materials)

2. The study employs Bayesian optimization to optimize network depth, width, optimizer, and learning rate, but ignore the critical information. Please provide the search ranges of hypermeters, objective function, stopping criterion, and computational cost.

Response: We apologize for the missing hypermeters, which are being searched for in the space of Bayesian Optimization and computational cost. In the revised manuscript, we have added information about surrogate training and compared the computational efficiency of each model in all experiments. We also employed an EarlyStopping strategy to optimize the best set of hyperparameters using Bayesian Optimization. The comparison results of computational efficiency for each model indicated that the computational costs depended on the surrogate's architecture, suggesting that the BO-BLSTM-based surrogate in TaihuScene, which has a larger network, required more computational resources than HyLake v1.0 and the Baseline. The associated revisions can be found in Materials and Methodology (Section 2.2.3, Lines 276-279) and Table 3.

Revision:

"The hyperparameter space included the number of hidden layers (ranging from 1 to 8), neurons per layer (ranged from 16 to 1,024), optimizer (Adam, or RMSprop), batch size (ranging from 8 to 256), and learning rate (ranging from 1E-6 to 1E-2). The hyperparameters in BO-BLSTM-based surrogates were optimized using BO with a maximum of 100 iterations, 1000 epochs for each iteration, and 50 patience in an EarlyStopping strategy." (Section 2.2.3, Lines 276-279)

"Table 3: Intercomparison of model performance across different experiments conducted in diverse regions with different forcing datasets. Observations from all lake sites (MLW, DPK, BFG, XLS, and PTS) on Lake Taihu, were used to drive models in the Taihu-obs experiment. Bold values in the table present the best-performing model with each group of experiments. Computational efficiency is reported as the runtime for a single simulation.

Exp	Model	Forcing	R			RMSE			MAE			Efficiency
			LST	LE	HE	LST	LE	HE	LST	LE	HE	(s)
MLW	PBBM	MLW	0.98	0.85	0.89	1.78	38.34	9.37	1.38	23.54	6.22	189.49
	FLake	MLW	0.98	0.82	0.84	1.76	42.73	7.24	1.35	24.76	5.01	**16.40**
	Baseline	MLW	0.96	0.74	0.75	2.71	51.77	14.63	2.11	33.52	9.30	151.46
	HyLake v1.0	MLW	**0.99**	**0.94**	**0.93**	**1.08**	**24.65**	**7.15**	**0.85**	**15.18**	**4.73**	270.21
Taihu-obs	FLake	All sites	0.97	0.61	0.74	2.24	15.46	69.11	1.69	41.95	10.44	**89.00**
	TaihuScene	All sites	**0.99**	**0.82**	0.89	1.52	14.93	43.49	1.23	29.53	10.63	6928.44
	HyLake v1.0	All sites	**0.99**	0.81	**0.90**	**1.36**	**11.19**	**39.20**	**1.03**	**24.79**	**7.88**	2693.23
Taihu-ERA5	FLake	ERA5	0.98	0.63	0.69	1.82	12.31	67.24	1.46	50.94	9.68	**19.60**
	TaihuScene	ERA5	**0.99**	0.68	0.73	1.60	13.00	64.83	1.29	47.78	10.11	652.25
	HyLake v1.0	ERA5	**0.99**	**0.71**	**0.78**	**1.12**	**11.05**	**49.48**	**0.90**	**35.02**	**7.97**	236.78
Chaohu	FLake	ERA5	**0.97**	\	\	2.28	\	\	1.76	\	\	**70.40**
	HyLake v1.0	ERA5	**0.97**	\	\	**2.07**	\	\	**1.57**	\	\	972.83

" (Table 3)

3. HyLake performed well at MLW, PTS, and XLS, whereas TaihuScene outperforms it at BFG and DPK. Discuss possible causes and advise when multi-site versus single-site training is preferable.

Response: Good point. We were surprised to find that the BO-BLSTM-based surrogate, which used a larger training dataset encompassing all lake sites in Lake Taihu, performed worse than one trained solely with MLW observations. The references found two hypothesis that might explain this issue: (1) The scaling laws for deep learning models determined that model performance will not continue to increase indefinitely with stacking of neurons and layers (Hestness et al., 2017); (2) Training data with high representation samples helps improve model performance, while using all over datasets would neglect heterogeneous functional samples and hinder the model's performance gains (Wang et al., 2025). We agreed that the surrogate in HyLake v1.0 meets the hypotheses. However, a comprehensive assessment of how many samples should be adapted at different scales (e.g., individual-lake, regional, or global scale) needs to be discussed in the future. The current manuscript provides additional explanations regarding these two hypotheses, although it does not reach a definitive conclusion. The associated revisions are listed in Discussion (Section 4.1, Lines 654-670).

References:

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., et al.: Deep-learning scaling is predictable, empirically, *arXiv* [preprint], arXiv:1712.00409, https://doi.org/10.48550/arXiv.1712.00409, 2017.

Wang, S., Yu, L., Gao, C., Zheng, C., Liu, S., Lu, R., et al.: Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning, arXiv [preprint], arXiv:2506.01939, 2025.

Revision:

"Future development of HyLake v1.0 **needs to** collect more observations, including heat fluxes and water temperature, and searching for more variables in datasets to train LSTM-based surrogates and acquire more general models at a larger scale. However, it is important to note that **the** performance of HyLake may not always improve with an increase in the training data size. **The training datasets with higher representation of physical principles help**

improve the model performance. Similar phenomenon has already been observed in many deep-learning-based large models, demonstrating that directly training models using all datasets would neglect heterogeneous functional samples and thereby hinder performance gains (Wang et al., 2025). Therefore, this study assumed that an individual-site-trained LSTM-based surrogate would have better capacity in representing lake-atmosphere interactions, which was collectively matched to the above-mentioned hypotheses. Due to insufficient observations at other lake sites (DPK, PTS, and XLS), to some degree, the surrogates trained on their datasets performed closely in estimating ΔLST except for XLW (Table S1). For the relatively complete observed datasets in BFG (although its biological characteristics cannot represent the whole Lake Taihu), the surrogate performed poorer than the proposed BO-BLSTM-based surrogate in terms of diurnal patterns of LST of HyLake v1.0 (Figure S10). As HyLake is extended to larger scales or more lakes, the computational architecture will need to accommodate large training datasets, which may limit performance for specific lakes. Specifically, the scaling laws for deep-learning models indicate that model performance does not continue to increase indefinitely with the stacking of neurons and layers (Hestness et al., 2017). Adapting more powerful deep-learning-based surrogates will further improve HyLake v1.0 performance, leading to a better representation of lake-atmosphere interactions in ungauged lakes." (Section 4.1, Lines 654-670)

References added:

Wang, S., Yu, L., Gao, C., Zheng, C., Liu, S., Lu, R., et al.: Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning, arXiv [preprint], arXiv:2506.01939, 2025.

4. The discussion regarding computational efficiency of HyLake is inadequate. Provide a detailed table comparing training time and wall-clock simulation speed for HyLake, FLake, and any other relevant models.

Response: We have provided the computational efficiency of models in all numerical experiments, including PBBM, Baseline, FLake, TaihuScene, and HyLake v1.0. As for the training time of LSTM-based surrogates, we do not compare each other because it is evident that the BO-BLSTM-based surrogate in TaihuScene, which used larger datasets to train, costs more than that in HyLake v1.0. Therefore, we discussed the computational efficiency of process-based models and hybrid lake models in response to this comment. The results indicated that HyLake v1.0 required higher resources compared to the traditional process-based models, including PBBM and FLake, in some cases, but cost less than TaihuScene. We found that their computational efficiency depends on the architecture of their surrogates. It is undeniable that the surrogate with deeper and broader networks requires more resources to train and predict. Therefore, we need to develop more deep-learning-based approaches to simulate results accurately and rapidly in the future. The associated revision can be found in Table 3 and Discussion (Section 4.1, Lines 615-624; Section 4.2, Lines 671-680).

Revision:

"**However, we found that HyLake v1.0 required slightly higher computational costs compared to process-based models, which depend on the hyperparameters of LSTM-based surrogates, despite achieving greater performance (Table 3). In an individual case of MLW prediction, HyLake v1.0 took about 9 times longer to run compared to FLake, with a cost of 151.46 seconds. To compare different experiments of hybrid lake models, Baseline, coupled to an LSTM-based surrogate with 1 layer and 256 neurons per layer, indicated the lowest cost. While TaihuScene, constructed by an LSTM-based surrogate with 7 layers and 836 neurons per layer, showed the most expensive in predictions. Given the sophisticated architecture of LSTM-based surrogates, which inevitably leads to higher costs in training and prediction, developing novel algorithms for approximating LSTMs is urgently needed. Furthermore, the recent research progress demonstrated that** LSTM-based surrogates are more suited for short-term predictions compared to **the prevalent Transformer-based family**, which **is** suited for long-term predictions and commonly used in global weather forecasting systems (K. F. Bi et al., 2023; L. Chen et al., 2023)." (Section 4.1, Lines 615-624)

"BO-BLSTM-based surrogate exhibits superior performance in estimating LST for HyLake v1.0. This study adapted BO and EarlyStopping strategies to ensure BLSTM provides accurate and reliable estimates in prediction but increases the computational demands for training due to its ability to converge from its more complex Bayesian

architecture (Peng et al., 2025; Ferianc et al., 2021). In addition, the mere 1 Bayesian fully connected layer that was adapted in this surrogate only captures limited data uncertainty, which may lose several aspects of probabilistic prediction (Klotz et al., 2022). Given the importance of uncertainty quantification for BLSTM, it is worth noting that HyLake v1.0 has the potential to assess the variance of predictions and probabilities of lake extreme events occurrence by developing its surrogate in future (Kar et al., 2024; Gawlikowski et al., 2023). Major limitations, including high computational demands and insufficient model performance, should be addressed by developing a novel deep-learning-based surrogate based on a more efficient architecture and larger datasets." (Section 4.2, Lines 671-680)

References added:

Ferianc, M., Que, Z., Fan, H., Luk, W., and Rodrigues, M.: Optimizing Bayesian recurrent neural networks on an FPGA-based accelerator, in: 2021 International Conference on Field-Programmable Technology (ICFPT), IEEE, December, 1-10, 2021.

Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., and Nearing, G.: Uncertainty estimation with deep learning for rainfall–runoff modeling, *Hydrol. Earth Syst. Sci.*, 26, 1673–1693, https://doi.org/10.5194/hess-26-1673-2022, 2022.

Peng, Z., Mo, S., Sun, A. Y., Wu, J., Zeng, X., Lu, M., and Shi, X.: An explainable Bayesian TimesNet for probabilistic groundwater level prediction, *Water Resour. Res.*, 61, e2025WR040191, https://doi.org/10.1029/2025WR040191, 2025.

"Table 3: Intercomparison of model performance across different experiments conducted in diverse regions with different forcing datasets. Observations from all lake sites (MLW, DPK, BFG, XLS, and PTS) on Lake Taihu, were used to drive models in the Taihu-obs experiment. Bold values in the table present the best-performing model with each group of experiments. Computational efficiency is reported as the runtime for a single simulation.

Exp	Model	Forcing	R			RMSE			MAE			Efficiency
			LST	LE	HE	LST	LE	HE	LST	LE	HE	(s)
MLW	PBBM	MLW	0.98	0.85	0.89	1.78	38.34	9.37	1.38	23.54	6.22	189.49
	FLake	MLW	0.98	0.82	0.84	1.76	42.73	7.24	1.35	24.76	5.01	**16.40**
	Baseline	MLW	0.96	0.74	0.75	2.71	51.77	14.63	2.11	33.52	9.30	151.46
	HyLake v1.0	MLW	**0.99**	**0.94**	**0.93**	**1.08**	**24.65**	**7.15**	**0.85**	**15.18**	**4.73**	270.21
Taihu-obs	FLake	All sites	0.97	0.61	0.74	2.24	15.46	69.11	1.69	41.95	10.44	**89.00**
	TaihuScene	All sites	**0.99**	**0.82**	0.89	1.52	14.93	43.49	1.23	29.53	10.63	6928.44
	HyLake v1.0	All sites	**0.99**	0.81	**0.90**	**1.36**	**11.19**	**39.20**	**1.03**	**24.79**	**7.88**	2693.23
Taihu-ERA5	FLake	ERA5	0.98	0.63	0.69	1.82	12.31	67.24	1.46	50.94	9.68	**19.60**
	TaihuScene	ERA5	**0.99**	0.68	0.73	1.60	13.00	64.83	1.29	47.78	10.11	652.25
	HyLake v1.0	ERA5	**0.99**	**0.71**	**0.78**	**1.12**	**11.05**	**49.48**	**0.90**	**35.02**	**7.97**	236.78
Chaohu	FLake	ERA5	**0.97**	\	\	2.28	\	\	1.76	\	\	**70.40**
	HyLake v1.0	ERA5	**0.97**	\	\	**2.07**	\	\	**1.57**	\	\	972.83

" (Table 3)

Minor comments

Line 164: Define the acronym LWT on first use.

Response: Corrected.

Revision: "T_s (℃) accounts for LST solved by 1-D vertical **lake water temperature (LWT)** transport equation; …" (Section 2.2.1, Lines 176-177)

Figure 3: Text are crowded in d-f. Consider summarizing the model accuracy in a table.

Response: We have added Table 3 to summarize the model's performance. The citations to Table 3 have already been added to the manuscript.

Revision:

"Table 3: Intercomparison of model performance across different experiments conducted in diverse regions

with different forcing datasets. Observations from all lake sites (MLW, DPK, BFG, XLS, and PTS) on Lake Taihu, were used to drive models in the Taihu-obs experiment. Bold values in the table present the best-performing model with each group of experiments. Computational efficiency is reported as the runtime for a single simulation.

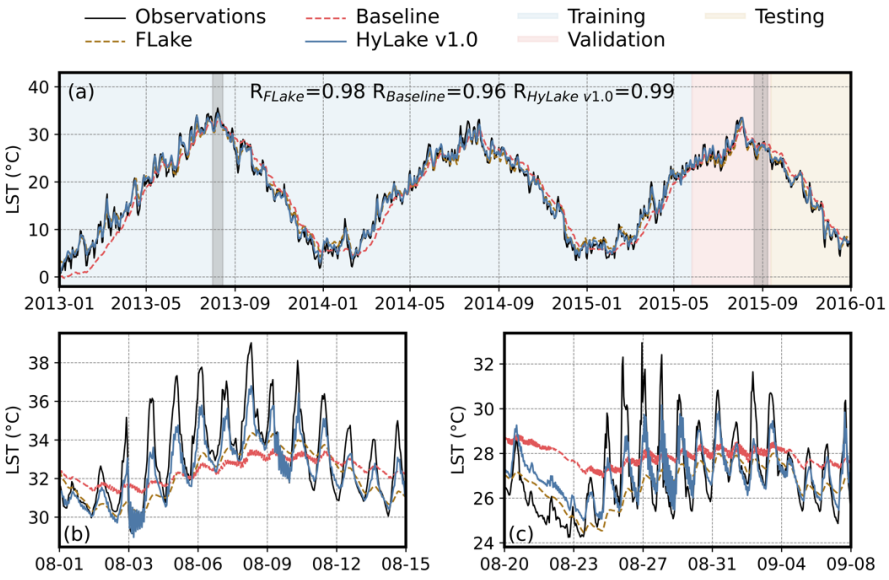
Exp	Model	Forcing	R			RMSE			MAE			Efficiency
			LST	LE	HE	LST	LE	HE	LST	LE	HE	(s)
MLW	PBBM	MLW	0.98	0.85	0.89	1.78	38.34	9.37	1.38	23.54	6.22	189.49
	FLake	MLW	0.98	0.82	0.84	1.76	42.73	7.24	1.35	24.76	5.01	**16.40**
	Baseline	MLW	0.96	0.74	0.75	2.71	51.77	14.63	2.11	33.52	9.30	151.46
	HyLake v1.0	MLW	**0.99**	**0.94**	**0.93**	**1.08**	**24.65**	**7.15**	**0.85**	**15.18**	**4.73**	270.21
Taihu-obs	FLake	All sites	0.97	0.61	0.74	2.24	15.46	69.11	1.69	41.95	10.44	**89.00**
	TaihuScene	All sites	**0.99**	**0.82**	0.89	1.52	14.93	43.49	1.23	29.53	10.63	6928.44
	HyLake v1.0	All sites	**0.99**	0.81	**0.90**	**1.36**	**11.19**	**39.20**	**1.03**	**24.79**	**7.88**	2693.23
Taihu-ERA5	FLake	ERA5	0.98	0.63	0.69	1.82	12.31	67.24	1.46	50.94	9.68	**19.60**
	TaihuScene	ERA5	**0.99**	0.68	0.73	1.60	13.00	64.83	1.29	47.78	10.11	652.25
	HyLake v1.0	ERA5	**0.99**	**0.71**	**0.78**	**1.12**	**11.05**	**49.48**	**0.90**	**35.02**	**7.97**	236.78
Chaohu	FLake	ERA5	**0.97**	\	\	2.28	\	\	1.76	\	\	**70.40**
	HyLake v1.0	ERA5	**0.97**	\	\	**2.07**	\	\	**1.57**	\	\	972.83

” (Table 3)

Figure 4-6: Add the relevant validation statistics directly to the LST, LE, and HE plots for clarity.

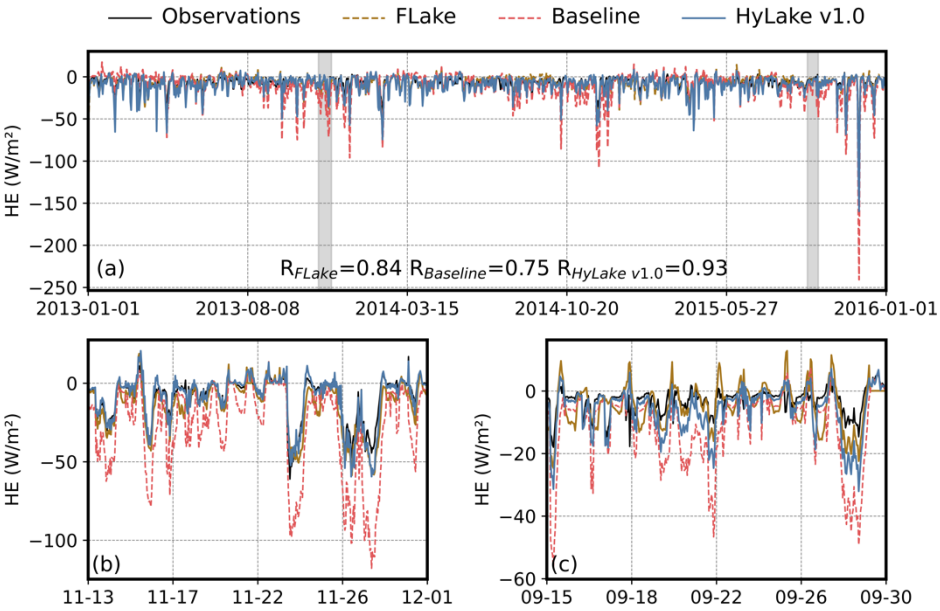
Response: We have added the Pearson coefficient in Figure 4-6 (now Figure 6-8 in the revised manuscript) due to the aesthetic appeal of the figures. The detailed information about model validation can be found in Table 3.

Revision:



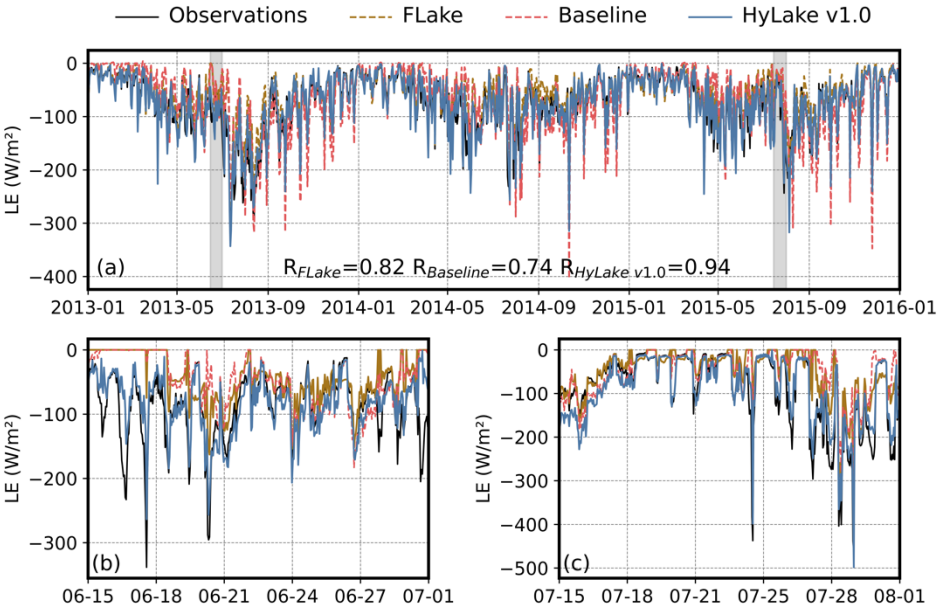
“

Figure 6: Comparison of observations and predictions by FLake, Baseline, and HyLake v1.0 in temporal trends of LST. Comparison of (a) the full time series and (b-c) partial time series of models derived LST and observations from 2013 to 2015. All results in (a) were presented at a daily-average scale by resampling. Blue, red, and yellow regions represent the period for the train, validation, and test datasets, respectively. Black solid, brown dashed, red dashed, and blue solid lines represent LST from observations, FLake, Baseline, and HyLake v1.0, respectively.” (Figure 6)



“

Figure 7: Comparisons of observations and predictions by FLake, Baseline, and HyLake v1.0 in temporal trends for LE. Comparison of (a) full and (b-c) partial time series of model derived LE and observations from 2013 to 2015.” (Figure 7)



“

Figure 8: Comparisons of observations and predictions by FLake, Baseline, and HyLake v1.0 in temporal trends for HE. Comparison of (a) full and (b-c) partial time series of model derived HE and observations from 2013 to 2015.” (Figure 8)

Figure 9: a-c and d-i represent distinct data types and should not share a single figure label.

Response: Corrected. Figure 9 was now spitted into Figure 10 and 11.

Revision:

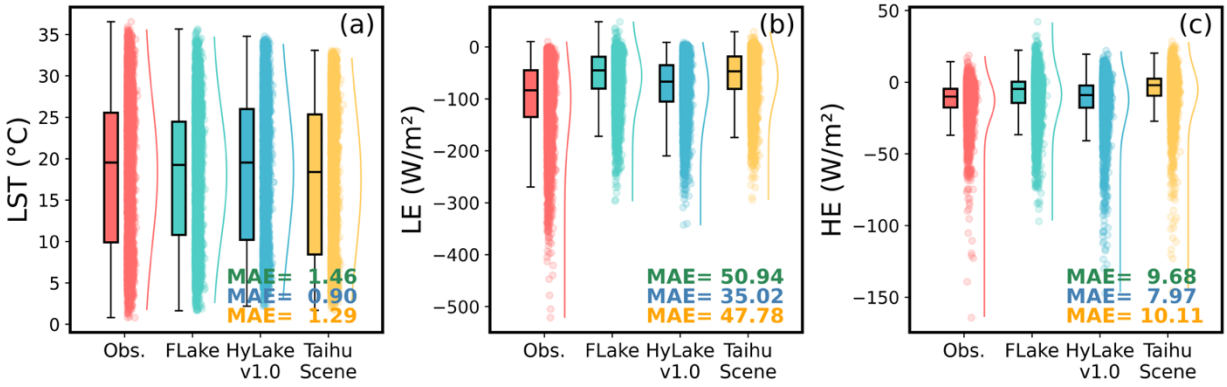


Figure 10: The statistical characteristics and spatial average of LST, LE and HE for observations, FLake, HyLake v1.0 and TaihuScene in all sites using ERA5 forcing datasets. Green, blue and yellow texts in figures represent the MAEs of LST, LE and HE for FLake, HyLake v1.0 and TaihuScene, respectively.” (Figure 10)

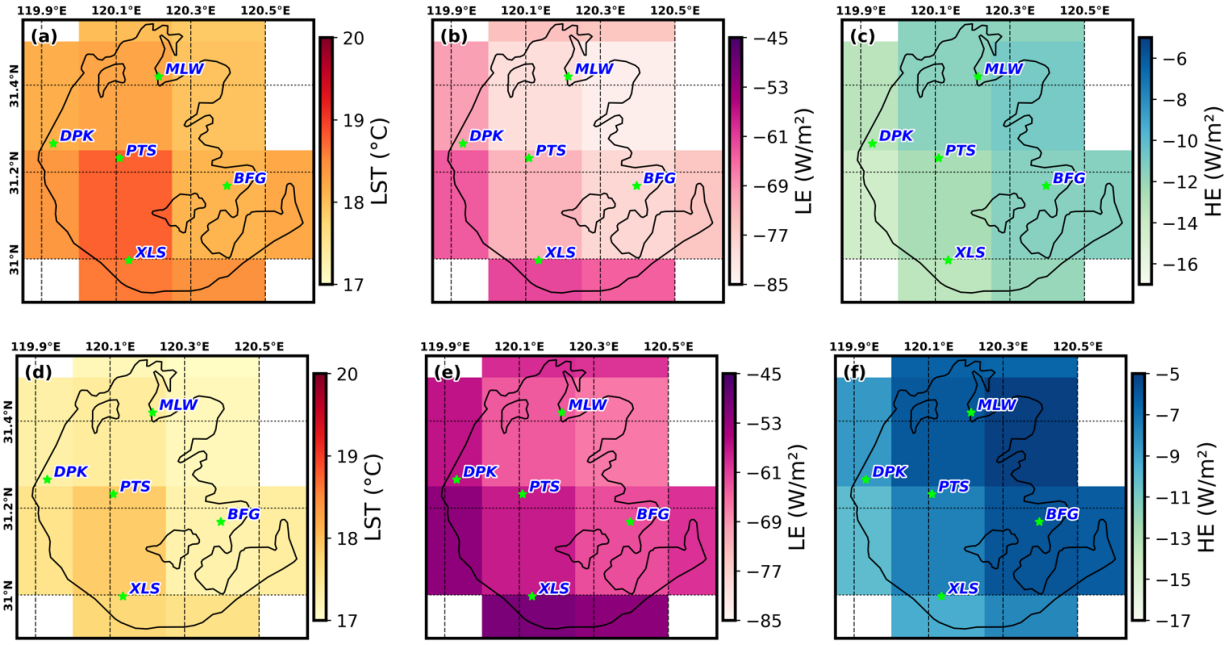


Figure 11: The statistical characteristics and spatial average of predicted LST, LE and HE for HyLake v1.0 and TaihuScene drove by ERA5 forcing datasets. (a-c) represent LST, LE and HE for HyLake v1.0, respectively; (d-f) represent LST, LE and HE for TaihuScene, respectively. The green stars noted in all figures are lake sites in Lake Taihu.” (Figure 11)

Line 847: correct the citation format for code repositories with the rest of the reference list.

Response: Corrected as “He, Y.: Code and datasets of paper "Hybrid Lake Model (HyLake) v1.0: unifying deep learning and physical principles for simulating lake-atmosphere interactions", Zenodo [code and data set], https://doi.org/10.5281/zenodo.15289113, 2025.”

Reviewer #2:

The manuscript entitled "Hybrid Lake Model (HyLake) v1.0: unifying deep learning and physical principles for simulating lake-atmosphere interactions" written by Yuan He and Xiaofan Yang (egusphere-2025-1983) presented the HyLake v1.0 hybrid model, which performed better than other models. The manuscript is generally well-written, which will be within the scope of GMD. Please clarify the following points before the possible publication.

Response: We thank Reviewer #2 for the positive and constructive comments. All the comments have been accepted and **Relisted in black**, followed by our **Replies in blue** and **Revisions in red (highlighted revisions in bold)**. According to the comments, we particularly discussed the reasons of using individual-site observations to train LSTM-based surrogates and explained the details of model development. Major changes are summarized in the following table:

No.	Major Revisions	Key Messages
1	Discussed the selection of MLW observations to train LSTM surrogate in HyLake v1.0.	As one of the long-term monitored lake sites in Lake Taihu, MLW has high-quality of observations and highly represents the eutrophic status of Lake Taihu. We cross-validated the performance using different observations from lake sites to train LSTM surrogate and confirmed that observations of the MLW are reliable **(Materials and methodology; Discussion)**.
2	Described how to fill the data gaps using ERA5 dataset	We used meteorological variables in ERA5 dataset to fill the missing data in the lake site that existed missing in their time series. These observations were used to force lake models to predict lake surface temperature and heat fluxes **(Materials and methodology)**.

Major comments:

Line 117-119 (and Table 1): This might be the trial and error in the authors and was not presented explicitly within the manuscript, but why was only the MLW site used for training and other sites used for validation? I missed the information, but why was the cross-validation not attempted in the process? I am wondering about the robustness of the developed model based on the training data from one site.

Response: We apologize for the confusion regarding our use of MLW observations to train an LSTM-based surrogate. Specific reasons are listed as follows:

(1) Reason for choosing MLW observations to train LSTM-based surrogate. There are five sites in Lake Taihu where hydrometeorological variables are observed via the lake Eddy Flux Network, including air temperature, rainfall rate, net longwave and shortwave radiation, wind speed, surface pressure, relative humidity, lake surface temperature, and latent and sensible heat fluxes. Air temperature, rainfall rate, net longwave and shortwave radiation, wind speed, surface pressure, and relative humidity were adapted to force lake models, while lake surface temperature, latent and sensible heat fluxes were used to validate the model performance. Among these observations, we found there are data gaps to different degrees in these sites. For example, about 475-time steps (~1.36%) of observed surface pressure were found to be lacked in DPK site during 2012 and 2015; 7,959 time steps (~22.71%) of all observed variables were missing in XLS site; 12,539 time steps (~35.78%) of all observed variables were missing in PTS site during 2013-2015. We think these lake sites were not suitable for training the LSTM-based surrogates. Given the MLW and BFG sites, citations have evidenced that Lake Taihu and its MLW site are quintessential examples of severe eutrophication in China (Yan et al., 2024; Wang et al., 2019), which differs from BFG's biological characteristics. The association descriptions can be found in Material and methodology (Section 2.1, Lines 114-139).

(2) Cross-validation between each lake site in training LSTM-based surrogates: The MLW observations are the most reliable among the 5 lake sites after our rigorous validation. We also used observations from 5 lake sites to individually search for optimized BO-BSTM-based surrogates, respectively. The validation results are given in Figure 4, S10, R1 and Table S1. The results indicated that, theoretically, the surrogates trained with MLW, BFG, DPK, and PTS

performed well in validation, while those trained with XLS performed poorer than the other surrogates (Figure 4, R1 and Table 1). Considering the absence of DPK and PTS observations, we only choose the surrogate trained with BFG to couple to the PBBM backbone model (namely HyLake-BFG in this comment) and then compare its performance to HyLake v1.0 in MLW and BFG site. Results indicated that HyLake v1.0 outperformed HyLake-BFG in both MLW and BFG site (Figure S10), indicating that using MLW observations to train surrogate helps hybrid lake models learn physical knowledge and improve their accuracy. Therefore, we decided to developed HyLake v1.0 based on MLW observations to according to the comprehensive validation. The associated revisions were listed in Discussion (Section 4.2, Lines 660-666).

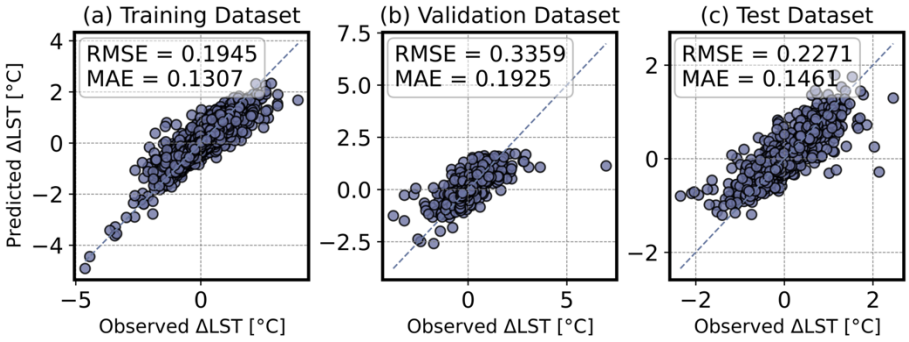
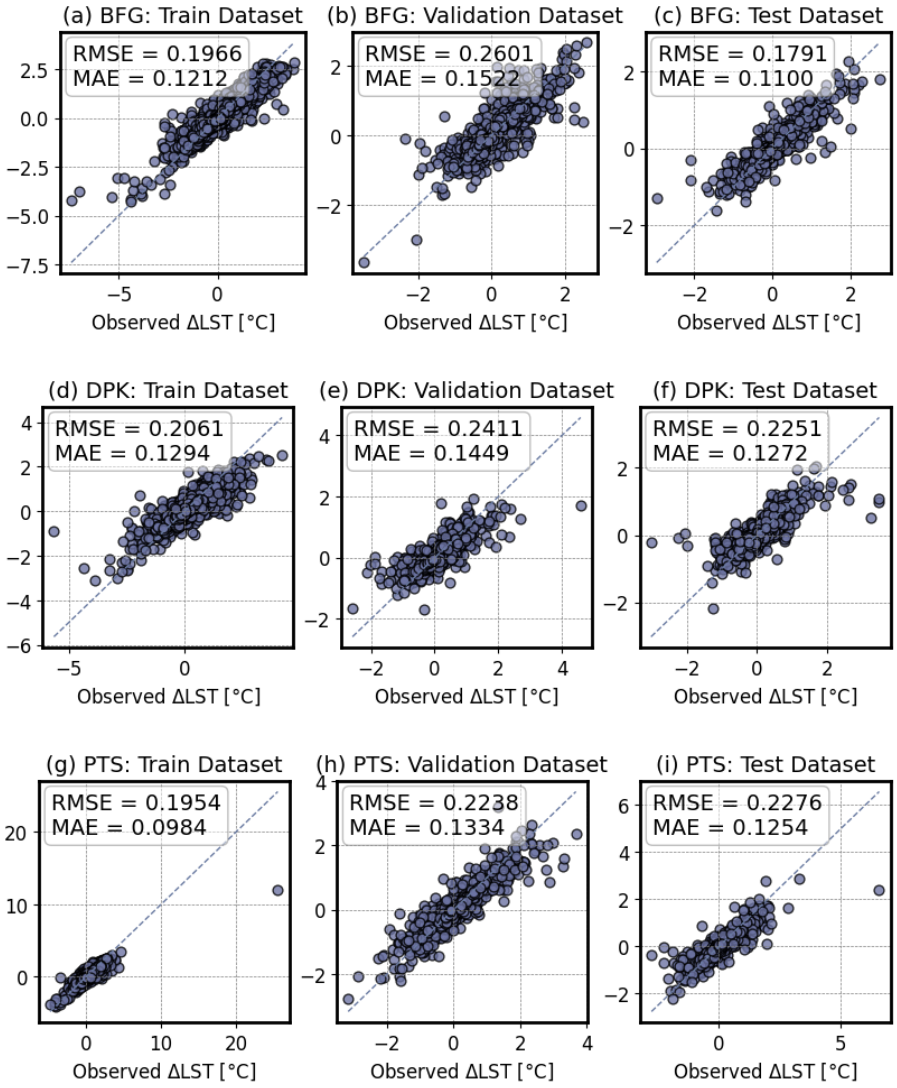


Figure 4: Validation of BO-BLSTM-based surrogate trained with MLW observations in HyLake v1.0 for (a) training, (b) validation and (c) test datasets.



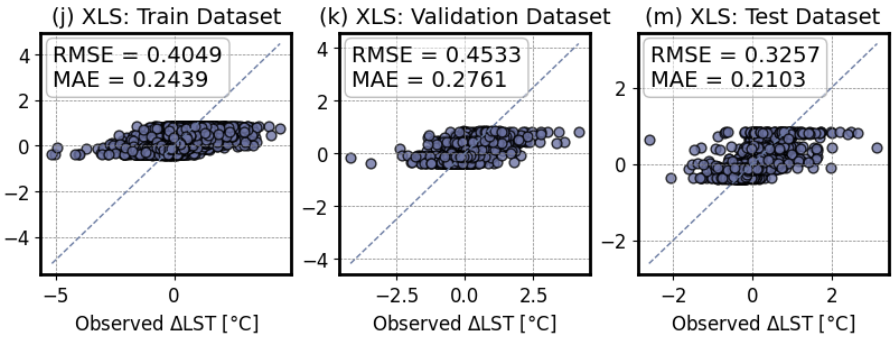


Figure R1: Validation of BO-BLSTM-based surrogates trained with (a-c) BFG, (d-f) DPK, (g-i) PTS, and (j-m) XLS observations in train, validation and test datasets, respectively.

Table S1: Model specifications of BO-BLSTM-based surrogates that trained with BFG, DPK, PTS, and XLS observations and performance in training sets, validation sets, and test sets of MLW. The RMSE for each surrogate was calculated from the difference between their training datasets.

NO.	Training dataset	Model specifications				RMSE (℃)		
		Number of layers	Neurons per layer	Batch size	Learning rate	Train	Validation	Test
1	MLW	4	467	64	9.6E-4	0.19	0.34	0.23
2	BFG	5	30	94	2.5E-3	0.20	0.26	0.18
3	DPK	5	94	124	3.0E-3	0.21	0.24	0.23
4	PTS	6	143	124	7.5E-4	0.20	0.22	0.23
5	XLS	5	170	29	1.0E-2	0.40	0.45	0.33
6	Whole	7	836	145	2.5E-2	0.24	0.33	0.23

References:

Wang, J., Fu, Z., Qiao, H., and Liu, F.: Assessment of eutrophication and water quality in the estuarine area of Lake Wuli, Lake Taihu, China. *Sci. Total Environ.*, *650*, 1392-1402, https://doi.org/10.1016/j.scitotenv.2018.09.137, 2019.

Yan, X., Xia, Y. Q., Ti, C. P., Shan, J., Wu, Y. H., and Yan, X. Y.: Thirty years of experience in water-pollution control in Taihu Lake: a review, *Sci. Total Environ.*, 914, 169821, https://doi.org/10.1016/j.scitotenv.2023.169821, 2024.

Revision:

"**The datasets included two parts: (1) hydrometeorological variables observed from the Taihu Lake Eddy Flux Network to force and validate the models, and (2) meteorological variables from ERA5 datasets to fill the gaps of observations and force the models. Within the network,** each site is equipped with an eddy covariance system that continuously monitors **LE and HE** using sonic anemometers and thermometers (Model CSAT3A; Campbell Scientific, Logan, UT, USA) positioned 3.5 to 9.4 m above the lake surface. Hydrometeorological variables, including air humidity and temperature (Model HMP45D/HMP155A; Vaisala, Helsinki, Finland), wind speed (Model 03002; R.M. Young Co., Traverse City, MI, USA), and net radiation components (Model CNR4; Kipp & Zonen, Delft, the Netherlands), are also measured. **These meteorological variables were used to force lake models while LE, HE and LST from observations were used to validate the results of each numerical experiment, on top of which, the** inferred radiative LST were collected at 30-minute intervals that are publicly accessible via Harvard DataVerse (Lee, 2004; Zhang et al., 2020; https://doi.org/10.7910/DVN/HEWCWM). The dataset spans from 2012 to 2015 **and contains several data gaps across these lake sites. Specifically, 475 time steps (~1.36%) of observed surface pressure were found missing at the DPK site during 2012 and 2015; 7,959 time steps (~22.71%) of all observed variables were missing at the XLS site; 12,539 time steps (~35.78%) of all observed variables were missing at the PTS site. Observations at the MLW and BFG sites were complete during the entire study periods. For the model evaluation of Taihu-obs experiment, the data gaps of observed variables in these lake sites were directly filled by ERA5**

datasets at the corresponding time steps, which were used to predict lake-atmosphere interactions. In this study, **observed meteorological** variables from the MLW site**, an eutrophic lake site that presents the trophic status of Lake Taihu (Table 1, Wang et al., 2019),** are used to train the Long Short-Term Memory (LSTM)-based surrogates (Sect. 2.2); while data from the remaining sites serve to evaluate the generalization of HyLake v1.0 and train the LSTM-based surrogates. To **further** address the generalization **and transferability** of HyLake v1.0 across different forcing datasets, this study utilized **8 meteorological variables that obtained from** hourly ERA5 datasets from 2012 to 2015, with a spatial resolution of 0.25° at a single level to **force** HyLake v1.0. These datasets, available from the Climate Data Store (Hersbach et al., 2020; http://cds.climate.copernicus.eu), include variables such as air temperature, dew point temperature, surface pressure, wind speed, and surface net longwave and shortwave radiation**, which has similar probability distribution to observations across Lake Taihu (Figure S1).**The ERA5 datasets are **also individually** used to **force FLake and TaihuScene for comparison and predict lake-atmosphere interactions** in Lake Taihu, providing insights into the model's generalization, transferability and performance **using** different climatic forcing datasets." (Section 2.1, Lines 114-139)

References added:

Wang, J., Fu, Z., Qiao, H., and Liu, F.: Assessment of eutrophication and water quality in the estuarine area of Lake Wuli, Lake Taihu, China. *Sci. Total Environ.*, *650*, 1392-1402, https://doi.org/10.1016/j.scitotenv.2018.09.137, 2019.

"Therefore, this study assumed that an individual-site-trained LSTM-based surrogate would have better capacity in representing lake-atmosphere interactions, which was collectively matched to the above-mentioned hypotheses. Due to insufficient observations at other lake sites (DPK, PTS, and XLS), to some degree, the surrogates trained on their datasets performed closely in estimating ΔLST except for XLW (Table S1). For the relatively complete observed datasets in BFG (although its biological characteristics cannot represent the whole Lake Taihu), the surrogate performed poorer than the proposed BO-BLSTM-based surrogate in terms of diurnal patterns of LST of HyLake v1.0 (Figure S10)." (Section 4.2, Lines 660-666)

"**Table S1: Model specifications of BO-BLSTM-based surrogates that trained with BFG, DPK, PTS, and XLS observations and performance in training sets, validation sets, and test sets of MLW. The RMSE for each surrogate was calculated from the difference between their training datasets.**

NO.	Training dataset	Model specifications				RMSE (℃)		
		Number of layers	Neurons per layer	Batch size	Learning rate	Train	Validation	Test
1	MLW	4	467	64	9.6E-4	0.19	0.34	0.23
2	BFG	5	30	94	2.5E-3	0.20	0.26	0.18
3	DPK	5	94	124	3.0E-3	0.21	0.24	0.23
4	PTS	6	143	124	7.5E-4	0.20	0.22	0.23
5	XLS	5	170	29	1.0E-2	0.40	0.45	0.33
6	Whole	7	836	145	2.5E-2	0.24	0.33	0.23

" (Table S1 in Supplementary Materials)

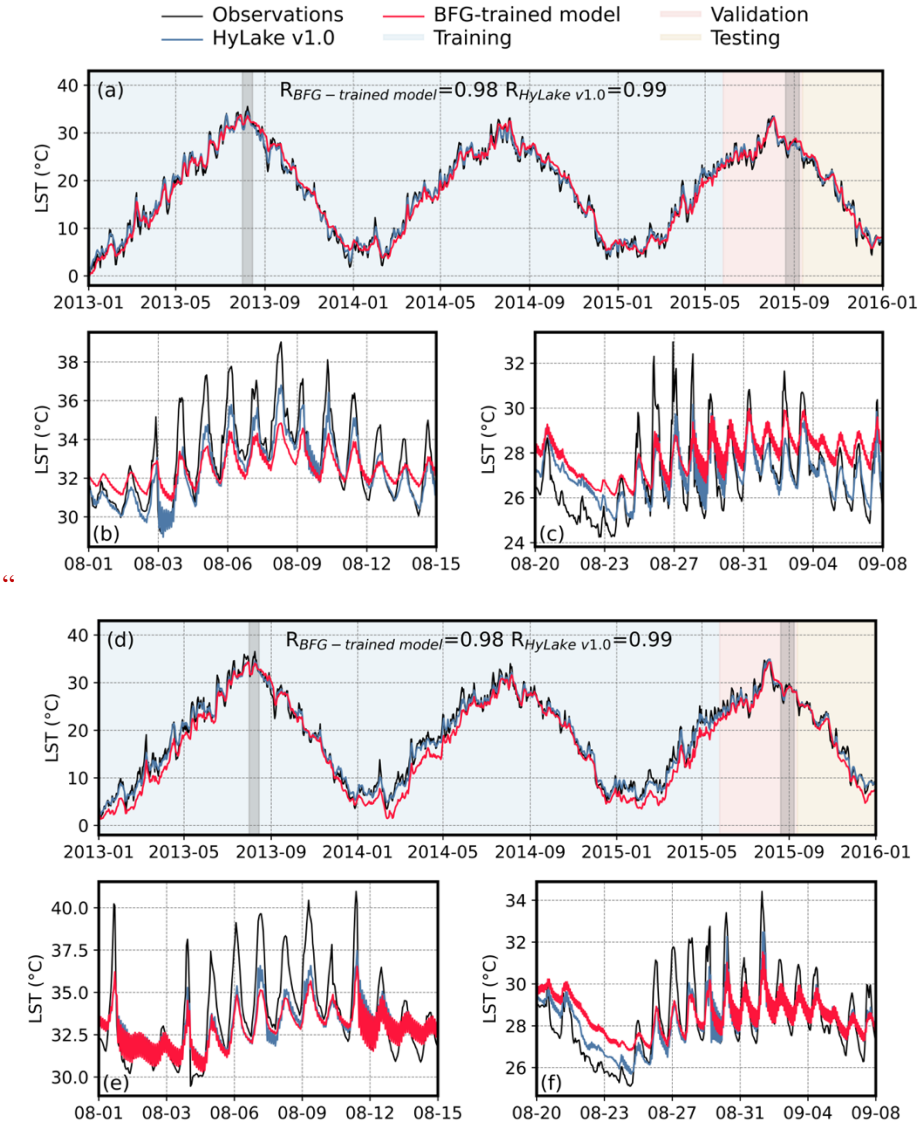


Figure S10: Comparison between HyLake v1.0 used MLW-train surrogate and BFG-trained surrogate in temporal trends of LST. (a-c) and (d-f) present the time series comparison at MLW and BFG site, respectively. Comparison of (a, and d) the full time series and (b-c, and e-f) partial time series of models derived LST and observations from 2013 to 2015. Blue, red, and yellow regions represent the period for the training, validation, and test datasets, respectively." (Figure S10 in Supplementary Materials)

Specific comments:

Line 40-42: The reference for each process-based model will be better.

Response: Corrected.

Revision: "Process-based lake **thermodynamics** models, such as the Freshwater Lake model (FLake) **(Mironov et al., 2010)**, the General Lake Model (GLM) **(Hipsey et al., 2019)**, and **the lake thermodynamics model in Weather Research & Forecasting Model (WRF-Lake) (Gu et al., 2015)**, are built on relationships between climate variables and LST, often employing simplified assumptions based on empirical physical principles (Mironov et al., 2010; Piccolroaz et al., 2024; L. J. Xu et al., 2016)." (Section 1, Lines 42-46)

References added:

Gu, H., Jin, J., Wu, Y., Ek, M. B., and Subin, Z. M.: Calibration and validation of lake surface temperature simulations with the coupled WRF-lake model. *Clim. Change*, *129*(3), 471-483, https://doi.org/10.1007/s10584-013-0978-y, 2015.

Hipsey, M. R., Bruce, L. C., Boon, C., Busch, B., Carey, C. C., Hamilton, D. P., Hanson, P. C., Read, J. S., de Sousa, E., Weber, M., and Winslow, L. A.: A General Lake Model (GLM 3.0) for linking with high-frequency sensor data from the

Global Lake Ecological Observatory Network (GLEON), Geosci. Model Dev., 12, 473–523, https://doi.org/10.5194/gmd-12-473-2019, 2019.
Mironov, D., Heise, E., Kourzeneva, E., Ritter, B., Schneider, N., and Terzhevik, A.: Implementation of the lake-parameterization scheme FLake into the numerical-weather-prediction model COSMO, *Boreal Environ. Res.*, 15, 218–230, 2010.

Line 116-117 and 120-125: How to fill the gap by the ERA5 reanalysis dataset was ambiguous. For example, what was the deficit rate from 2012 to 2015? Please rewrite this explanation.

Response: Sorry for missing this information. We first provided a conceptual figure to describe the way of using the ERA5 dataset to fill the observations (Figure R2). We used the most straightforward method, which involved checking and replacing missing data in observations with ERA5 datasets for each variable, as the two datasets share a similar probability distribution in their meteorological variables (Figure S1). Then, we calculated the deficit rate (missing length/length of time series) for observations at each lake site from 2012 to 2015. Specifically, 475 time steps (~1.36%) of observed surface pressure were found to be lacking in the DPK site during 2012 and 2015; 7959 time steps (~22.71%) of all observed variables were missing in the XLS site; 12539 time steps (~35.78%) of all observed variables were missing in the PTS site; Observations at the MLW and BFG sites were complete during the study periods. More details about using ERA5 datasets in this study were provided in Materials and methodology (Section 2.1, Lines 114-139) and Figure S1.

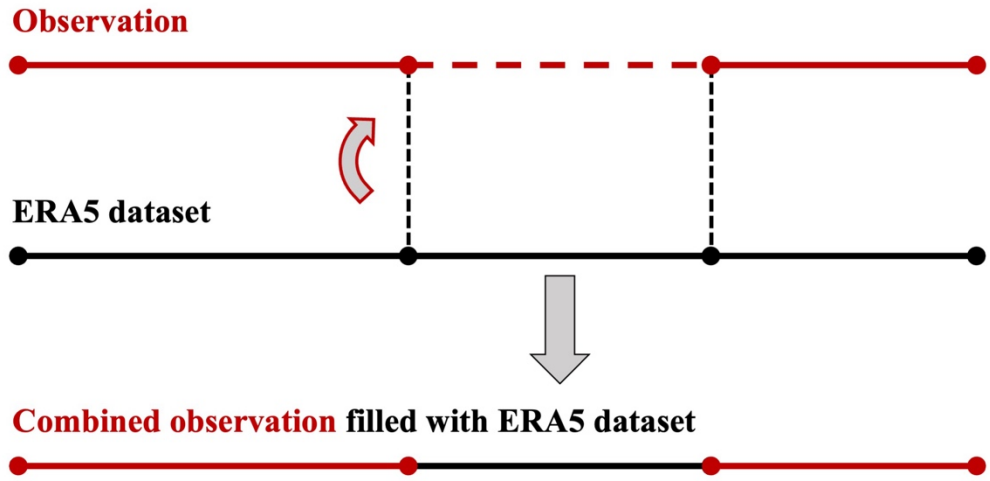


Figure R2: Conceptual diagram for gap filling of the observations by using ERA5 dataset.

Revision: "The datasets included two parts: (1) hydrometeorological variables observed from the Taihu Lake Eddy Flux Network to force and validate the models, and (2) meteorological variables from ERA5 datasets to fill the gaps of observations and force the models. Within the network, each site is equipped with an eddy covariance system that continuously monitors **LE and HE** using sonic anemometers and thermometers (Model CSAT3A; Campbell Scientific, Logan, UT, USA) positioned 3.5 to 9.4 m above the lake surface. Hydrometeorological variables, including air humidity and temperature (Model HMP45D/HMP155A; Vaisala, Helsinki, Finland), wind speed (Model 03002; R.M. Young Co., Traverse City, MI, USA), and net radiation components (Model CNR4; Kipp & Zonen, Delft, the Netherlands), are also measured. **These meteorological variables were used to force lake models while LE, HE and LST from observations were used to validate the results of each numerical experiment, on top of which, the** inferred radiative LST were collected at 30-minute intervals that are publicly accessible via Harvard DataVerse (Lee, 2004; Zhang et al., 2020; https://doi.org/10.7910/DVN/HEWCWM). The dataset spans from 2012 to 2015 **and contains several data gaps across these lake sites. Specifically, 475 time steps (~1.36%) of observed surface pressure were found missing at the DPK site during 2012 and 2015; 7,959 time steps (~22.71%) of all observed variables were missing at the XLS site; 12,539 time steps (~35.78%) of all observed variables were missing at the PTS site. Observations at the MLW and BFG sites were complete during the entire study periods. For the model evaluation**

of Taihu-obs experiment, the data gaps of observed variables in these lake sites were directly filled by ERA5 datasets at the corresponding time steps, which were used to predict lake-atmosphere interactions. In this study, **observed meteorological** variables from the MLW site**, an eutrophic lake site that presents the trophic status of Lake Taihu (Table 1, Wang et al., 2019),** are used to train the Long Short-Term Memory (LSTM)-based surrogates (Sect. 2.2); while data from the remaining sites serve to evaluate the generalization of HyLake v1.0 and train the LSTM-based surrogates. To **further** address the generalization **and transferability** of HyLake v1.0 across different forcing datasets, this study utilized **8 meteorological variables that obtained from** hourly ERA5 datasets from 2012 to 2015, with a spatial resolution of 0.25° at a single level to **force** HyLake v1.0. These datasets, available from the Climate Data Store (Hersbach et al., 2020; http://cds.climate.copernicus.eu), include variables such as air temperature, dew point temperature, surface pressure, wind speed, and surface net longwave and shortwave radiation**, which has similar probability distribution to observations across Lake Taihu (Figure S1).**The ERA5 datasets are **also individually** used to **force FLake and TaihuScene for comparison and predict lake-atmosphere interactions** in Lake Taihu, providing insights into the model's generalization, transferability and performance **using** different climatic forcing datasets." (Section 2.1, Lines 114-139)

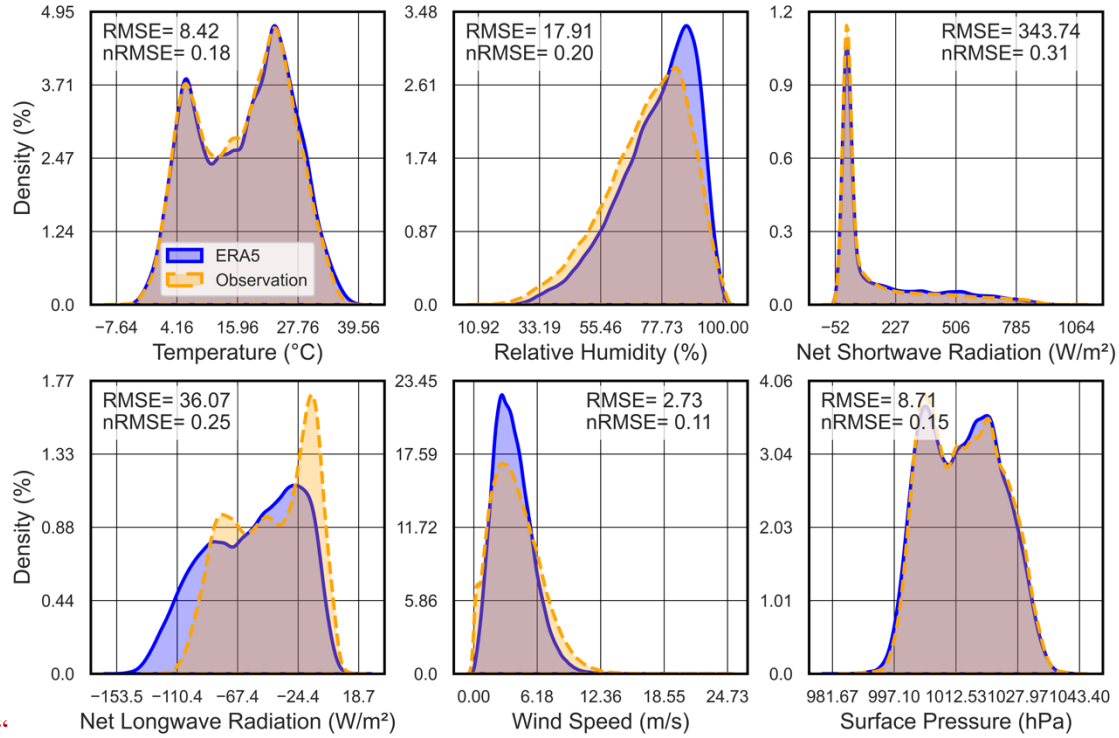


Figure S1: The probability density distribution of meteorological variables from observation and ERA5 reanalysis datasets in MLW, BFG, DPK, PTS, and XLS site during 2012 to 2015. A normalized RMSE (nRMSE) was assigned to assess the error between observation and ERA5 reanalysis datasets." (Figure S1 in Supplementary Materials)

Line 120-125: In addition to the above comment, when the ERA5 reanalysis gaps the data at the MLW site, is this the self-validation? Please clarify.

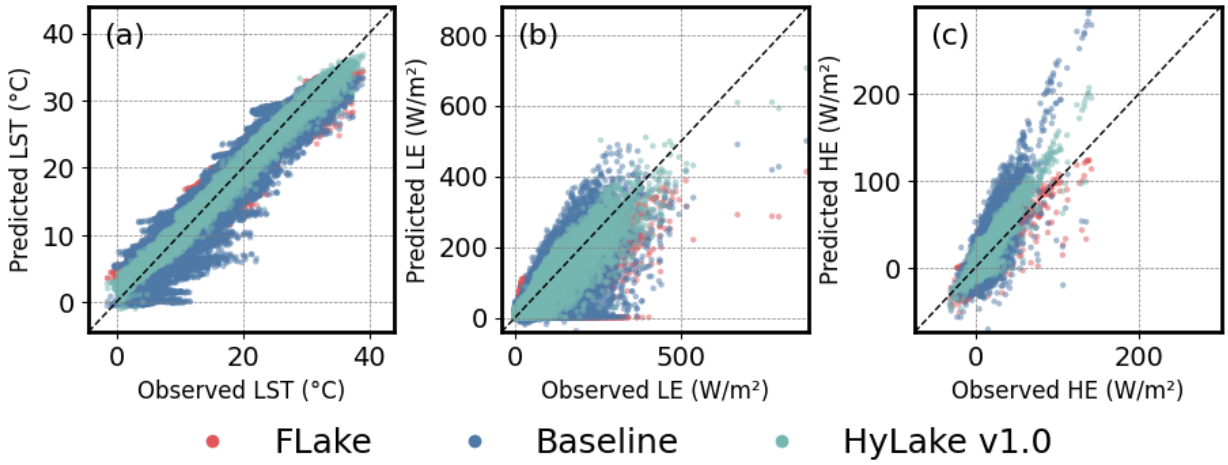
Response: The MLW site has complete observations for 2013 and 2015, which DOESN'T require any gap filling with ERA5 datasets. We only checked and filled the meteorological variables from observations, including air temperature, relative humidity, surface pressure, wind speed, and surface net longwave and shortwave radiation, with ERA5 datasets to force HyLake v1.0 and other lake models in DPK, PTS, and XLS sites during the studied period if needed.

Revision: "In the evaluation of all observations-forced experiments, the data gaps of observed variables in these lake sites were directly filled by ERA5 datasets at the corresponding time steps to predict lake-atmosphere interactions." (Section 2.1, Lines 127-129)

Line 345: The legend "HyLake-baseline" will be confusing. I would like to recommend expressing "Baseline".

Response: Corrected.

Revision:



"

Figure 5: Comparison of predicted (a) LST, (b) LE and (c) HE by using FLake (red points), Baseline (blue points) and HyLake v1.0 (green points) in MLW experiments." (Figure 5)

Technical comments:

Line 29: "surface water temperature" will not match the abbreviation of "LST". Is this "lake surface temperature"? Please confirm.

Response: Thanks. Corrected to "lake surface temperature (LST)".

Line 110: No need to repeat these abbreviations.

Response: Corrected.

Revision: "**Within the network, e**ach site is equipped with an eddy covariance system that continuously monitors **LE and HE** using sonic anemometers and thermometers (Model CSAT3A; Campbell Scientific, Logan, UT, USA) positioned 3.5 to 9.4 m above the lake surface." (Lines 116-118)

Reviewer #3:

This manuscript presents HyLake v1.0, a hybrid lake–atmosphere model that embeds a Bayesian-optimized bidirectional LSTM surrogate within a process-based 1-D vertical transport framework to simulate lake surface temperature and surface fluxes. The work addresses a key challenge in environmental modeling: integrating data-driven surrogates with physical principles. The extensive validation on Lake Taihu (2012–2015) against FLake demonstrates clear performance gains, and the hybrid approach represents a meaningful methodological advance for lake modeling. While the methodology is sound and the Lake Taihu validation is comprehensive, the authors should more clearly discuss the requirements and limitations for applying this approach to other lake systems. The current multi-site validation within Lake Taihu provides good evidence of transferability, but broader applicability claims should be more cautiously framed.

Response: We sincerely thank Reviewer #3 for the constructive comments. In revision, we particularly discussed the requirements and limitations for HyLake v1.0 and presented an example using another morphologically distinct lake to show its transferability. All comments are accepted and **Relisted in black**, followed by our **Replies in blue** and **Revisions in red (highlighted revisions in bold)**. Before point-by-point response, we summarized major revisions followed by Reviewer #3's comments as:

No.	Major Revisions	Important Messages
1	Presented a test case for applying HyLake v1.0 to another morphologically distinct lake.	The revised manuscript employed HyLake v1.0 to simulate lake-atmosphere interactions in another morphologically distinct lake, Lake Chaohu, and discussed the potential challenges in model application (**Materials and methodology; discussion**).
2	Discussed the limitations of deep-learning-based surrogates.	We discussed the cons and pros of computational requirements, BO-BLSTM-based surrogate, and the choice of lake surface temperature module (**Discussion**).
3	Provided future directions to improve HyLake v1.0.	We mainly discussed the uncertainty of Bayesian algorithms. Future improvements should focus on development of surrogates by using novel techniques. The employment of a Bayesian fully connected layer in surrogates could also provide probabilistic predictions by quantifying uncertainties in the future (**Discussion**).

Specific Comments

The multi-site validation within Lake Taihu is convincing but add discussion of what adaptations would be needed for different lake types (e.g., deeper lakes, different climate zones, varying trophic states). Consider outlining a framework for applying the methodology to new lake systems.

Response: Good point! We agree that applying HyLake v1.0 to other lakes is essential. Therefore, we utilized it to another lake in the middle and lower reaches of the Yangtze River Plain—Lake Chaohu and discussed potential limitations for model application.

(1) Applying HyLake v1.0 to another lake: Lake Chaohu is the 5th-largest shallow freshwater lake in China, with a deeper lake depth of 3.06 m and smaller lake area of 760 km^2 than Lake Taihu (Jiao et al., 2018), which has experienced heavy eutrophication and harmful algal blooms (Yang et al., 2020). Given the difficulty that Lake Chaohu does not have sufficient observations, unlike Taihu, we outlined a framework that utilized ERA5 datasets to force HyLake v1.0 and the MOD11A1 land surface temperature dataset for validating lake surface temperature changes. The results indicated that HyLake v1.0 performed well in Lake Chaohu, with an R^2 of 0.97, RMSE of 2.07 °C, and MAE of 1.57 °C, outperforming FLake compared to the MOD11A1 datasets (Figure S7-9). The successful attempt of HyLake v1.0 in Lake Chaohu demonstrated that HyLake v1.0 is promising to apply in ungauged lakes. The associated revisions can be found in Materials and methodology (Section 2.3.1, Lines 286-289, Lines 308-314) and Discussion (Section 4.1, Lines

560-567, Lines 609-611; Figure S7-S9).

(2) Discussing potential challenges for model application: Although HyLake v1.0 succeeded in estimating lake-atmosphere interactions in Lake Chaohu, it still has several limitations. Considering the diverse lake types worldwide, it remains challenging to validate the performance of HyLake v1.0 in every case due to the limited observations and simplified physical principles. The quantitative restriction on observations hampers our ability to improve the model's performance in regional cases by retraining or fine-tuning the LSTM-based surrogates for each lake type. Additionally, the inaccurate relationships between lake surface conditions (e.g. friction velocity, surface roughness length) and climate change pose a challenge to HyLake v1.0. Specifically, we found that there are biases in the surface roughness length (z_0) and friction velocity (u^*) between observations and predictions (Figure S6). These potential differences were hard to quantify due to data scarcity in the current process-based models, which impeded us to improve the understanding of lake-atmosphere interactions. Therefore, the physical principles between lake surface conditions and climate change should be focused in the future using novel process-based or data-driven techniques. The associated revisions are listed in Discussion (Section 4.2, Lines 647-654).

To sum up, HyLake v1.0 provided a novel method for improving the understanding of lake-atmosphere interactions on most lakes. However, the current limitations of data and physical principles restrict the generalization ability for all unknown lake types. We aim to expand the modules and functions of HyLake v1.0 and validate it in additional lakes in the future, to accurately predict lake-atmosphere interactions for a broader range of lake types.

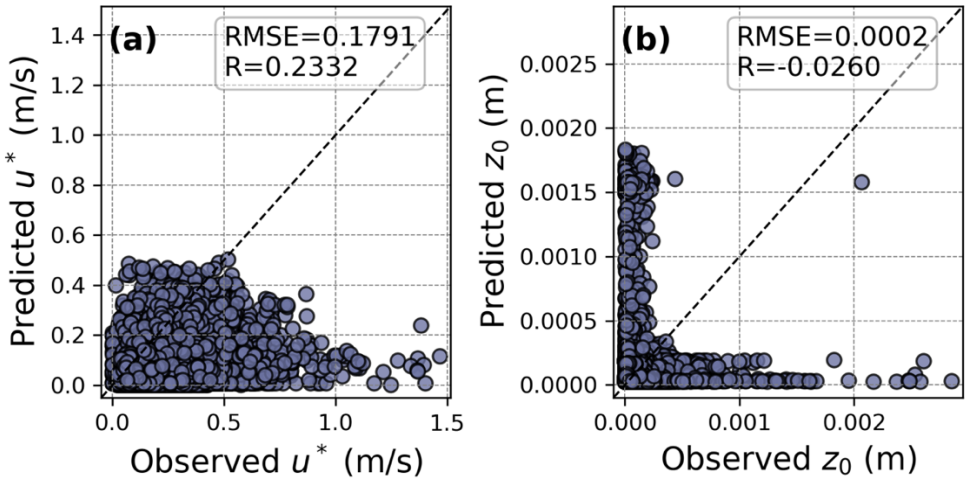


Figure S6: The comparison of friction velocity (u*) and surface roughness length (z_{0m}, m) in MLW lake site between simulation derived from PBBM and HyLake v1.0 and observations.

References:

Jiao, Y., Yang, C., He, W., Liu, W. X., and Xu, F. L.: The spatial distribution of phosphorus and their correlations in surface sediments and pore water in Lake Chaohu, China, *Environ. Sci. Pollut. Res.*, 25, 25906-25915, https://doi.org/10.1007/s11356-018-2606-x, 2018.

Yang, C., Yang, P., Geng, J., Yin, H., and Chen, K.: Sediment internal nutrient loading in the most polluted area of a shallow eutrophic lake (Lake Chaohu, China) and its contribution to lake eutrophication, *Environ. Pollut.*, 262, 114292, https://doi.org/10.1016/j.envpol.2020.114292, 2020.

Revision:

"To address the generalization and transferability of HyLake v1.0 in studied (MLW) and ungauged lake sites (DPK, BFG, XLS, and PTS) (Table 1), this study further **conducted** three numerical experiments**, including MLW experiment, Taihu-obs experiment, Taihu-ERA5 experiment, and Chaohu experiment,** using distinct **models** and forcing datasets (Table **2 and 3**), including FLake, **Baseline, and TaihuScene to intercompare.**" (Section 2.3.1, Lines 286-289)

"Furthermore, this study implemented the HyLake v1.0 into Lake Chaohu, the 5th-largest shallow freshwater lake in China, which has experienced heavy eutrophication and harmful algal blooms (Yang et al., 2020), to assess its transferability to other lakes. A LST dataset in Lake Chaohu was obtained from MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V061 imageries (MOD11A1,

https://www.earthdata.nasa.gov/data/catalog/lpcloud-mod11a1-061), which were used to validate the performance of LST derived from HyLake v1.0. The computational efficiency for each 1-time prediction was recorded using a 16G 10-Core Apple M4 processor based on the established HyLake v1.0 model in this study. The training of the above-mentioned surrogates was run using a 24G NVIDIA GeForce RTX 4090 GPU." (Section 2.3.1, Lines 308-314)

"**Table 3: Intercomparison of model performance across different experiments conducted in diverse regions with different forcing datasets. Observations from all lake sites (MLW, DPK, BFG, XLS, and PTS) on Lake Taihu, were used to drive models in the Taihu-obs experiment. Bold values in the table present the best-performing model with each group of experiments. Computational efficiency is reported as the runtime for a single simulation.**

Exp	Model	Forcing	R			RMSE			MAE			Efficiency
			LST	LE	HE	LST	LE	HE	LST	LE	HE	(s)
MLW	PBBM	MLW	0.98	0.85	0.89	1.78	38.34	9.37	1.38	23.54	6.22	189.49
	FLake	MLW	0.98	0.82	0.84	1.76	42.73	7.24	1.35	24.76	5.01	**16.40**
	Baseline	MLW	0.96	0.74	0.75	2.71	51.77	14.63	2.11	33.52	9.30	151.46
	HyLake v1.0	MLW	**0.99**	**0.94**	**0.93**	**1.08**	**24.65**	**7.15**	**0.85**	**15.18**	**4.73**	270.21
Taihu-obs	FLake	All sites	0.97	0.61	0.74	2.24	15.46	69.11	1.69	41.95	10.44	**89.00**
	TaihuScene	All sites	**0.99**	**0.82**	0.89	1.52	14.93	43.49	1.23	29.53	10.63	6928.44
	HyLake v1.0	All sites	**0.99**	0.81	**0.90**	**1.36**	**11.19**	**39.20**	**1.03**	**24.79**	**7.88**	2693.23
Taihu-ERA5	FLake	ERA5	0.98	0.63	0.69	1.82	12.31	67.24	1.46	50.94	9.68	**19.60**
	TaihuScene	ERA5	**0.99**	0.68	0.73	1.60	13.00	64.83	1.29	47.78	10.11	652.25
	HyLake v1.0	ERA5	**0.99**	**0.71**	**0.78**	**1.12**	**11.05**	**49.48**	**0.90**	**35.02**	**7.97**	236.78
Chaohu	FLake	ERA5	**0.97**	\	\	2.28	\	\	1.76	\	\	**70.40**
	HyLake v1.0	ERA5	**0.97**	\	\	**2.07**	\	\	**1.57**	\	\	972.83

" (Table 3)

"To address issues related to model performance, generalization, and transferability in ungauged locations, three additional numerical experiments, including FLake, Baseline, and TaihuScene, were proposed for **intercomparison and a framework for applying HyLake v1.0 to another lake, such as Lake Chaohu, with a deeper depth of 3.06 m and area of 760 km^2 (Figure S7, Jiao et al., 2018), to validate the potential capacity of transferability further.** These experiments were compared using **observed meteorological datasets, and ERA5 datasets** and **then** validated for both spatial and temporal patterns at Lake Taihu **and Lake Chaohu (Tables 2-3). Similarly, ERA5 dataset-derived HyLake v1.0 outperformed FLake in estimating LST (R = 0.97, RMSE = 2.07 °C, MAE = 1.57 °C) in Lake Chaohu, compared to MOD11A1 datasets (Table 3 and Figures S7-9).**" (Section 4.1, Lines 560-567)

"HyLake v1.0, developed based on *in situ* observations from Lake Taihu, has been proven to be reliable and rigorously validated in Lake Chaohu (Table 3), demonstrating a faster and more accurate framework for enhancing the understanding of hybrid hydrological modeling." (Section 4.1, Lines 609-611)

"**HyLake v1.0 has been applied to Lake Chaohu and achieved superior performance in comparison to the MYD11A1 LST observations, showing a promising way for more applications.** Future improvements to HyLake v1.0 should focus on investigating the scaling laws of datasets, development of surrogate architectures, and extension of coupled modules. Currently, HyLake v1.0 has been validated primarily in Lake Taihu, utilizing high-quality training data provided by the Lake Taihu Eddy Flux Network (Zhang et al., 2020). **However, in some exceptional cases, the lake may be influenced by regional inflows/outflows, or it may be covered by snow/ice for a long period, and the processes at the lake-air interface may differ from those in our experiments (Woolway et al., 2020). As a result, our model may not be quantifiable for these situations. Its surrogate will be required for more high-quality local datasets to retrain or finetune.**" (Section 4.2, Lines 647-654)

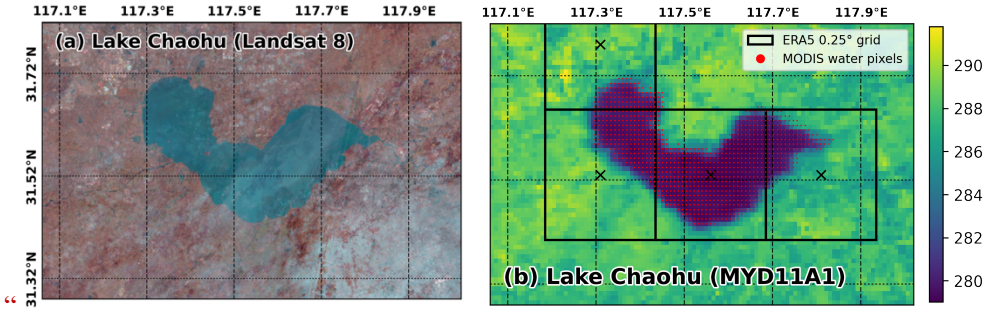


Figure S7: The locations of Lake Chaohu overlaid on a true-color image from (a) Landsat 8 and daily land surface temperature from (b) MYD11A1 product." (Figure S7 in Supplementary materials)

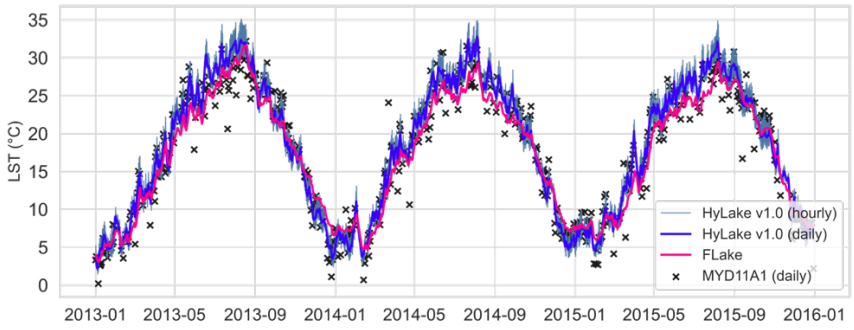


Figure S8: Time series of daily grid-average LST on Lake Chaohu derived from MYD11A1, FLake simulation, and HyLake v1.0 from 2013 to 2015. HyLake v1.0 provides daily and hourly simulations.

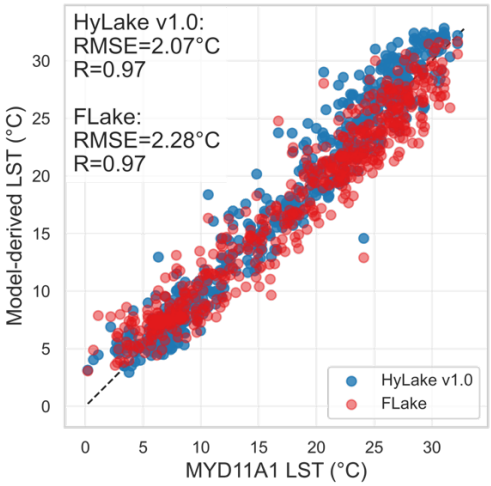


Figure S9: The intercomparison of daily LST between model simulations (FLake and HyLake v1.0) and MYD11A1 observations on Lake Chaohu from 2013 to 2015." (Figures S7-S9 in Supplementary Material)

Better justify the choice of BO-BLSTM over simpler alternatives. provide clearer explanation of why Bayesian optimization and bidirectional LSTM architecture were chosen over deterministic alternatives.

Response: Thanks for the suggestions. Using the LSTM-based surrogate with the best group of hyperparameters based on Bayesian Optimization (BO), integrated the abilities of LSTM for time series forecasting and the high computational efficiency of Bayesian Optimization (BO) to represent the physical principles of lake surface temperature changes significantly.

(1) The selection of BO-BLSTM over simpler alternatives: LSTM is one of Recurrent Neural Networks (RNNs) that learn from past data by using several gates in their network architecture to remember the past data (Siami-Namini et al., 2019). It becomes feasible for long-term time series forecasting due to the ability to learn many-step dependencies and handle variable-length input sequences in fields such as hydrology (Liu et al., 2024). It outperformed traditional, process-based, and machine learning models in many cases, including predictions of soil moisture, streamflow, water temperature, and groundwater levels (Mao et al., 2021; Feng et al., 2020; Papacharalampous et al., 2018). Meanwhile,

previous studies have shown that LSTM-based models outperform other traditional deep-learning models in auto-regressive predictions, supporting this study in predicting lake surface temperature changes robustly and reliably (Siami-Namini et al., 2019). Bayesian LSTM (BLSTM), an improved version of LSTM, adapts probability-distributed weight parameters, which reduce model overfitting and provide robust predictions in hydrology (Li et al., 2021; Lu et al., 2019). In comparison to these models in BO, we ultimately selected BLSTM-based surrogates to address the challenges in this autoregressive prediction task. Nevertheless, we agree that the surrogate should be improved due to its lower computational efficiency, which will be discussed in the future. Here we explained the advantages of LSTM and BLSTM in Materials and Methodology (Section 2.2.2, Lines 209-214) and discussed the limitations and potential improvements in Discussion (Section 4.1, Lines 614-624; Section 4.2, Lines 671-680).

(2) Using Bayesian Optimization and Bayesian LSTM over deterministic alternatives: In this study, we selected BO to search for the best group of hyperparameters in Bayesian LSTM (not Bidirectional LSTM) models. BO is a hyperparameter tuning algorithm based on the Bayesian theorem, which can significantly improve the performance and efficiency of deep learning models by building the relationships between model performance and their hyperparameters (Victoria et al., 2021; Wu et al., 2019). Previous studies have established that deep learning models often tune their hyperparameters using manual search or automatic search methods (Wu et al., 2019). Manual search methods depend on expert knowledge and are hard to reproduce and find the optimized hyperparameters. Traditional automatic search methods, such as grid search, train models with each combination of hyperparameters, which is exhaustive searching (Wu et al., 2019; Bergstra et al., 2012). BO adapted a random search technique to fit the data and update the posterior distribution of functions based on Gaussian processes and the Bayesian theorem (Victoria et al., 2021; Wu et al., 2019). Wu et al. (2019) compared the accuracy and costs between BO and grid search methods, finding that both methods performed almost equally well in the same case, while BO runs 12 times faster than grid search.

To summarize, given the large variability and complex relationships of the observations in this study, we would like to employ a more computationally efficient method to help users identify the most robust surrogate within a large hyperparameter space as soon as possible. Considering that the selection of optimization methods is not a focus of this study, the current manuscript provides detailed information about the hyperparameter space for each surrogate to help readers understand. The associated revisions are listed in Materials and Methodology (Section 2.2.3, Lines 276-279).

References:

Bergstra, J. and Bengio, Y.: Random search for hyper-parameter optimization, *J. Mach. Learn. Res.*, 13, 281–305, https://dl.acm.org/doi/10.5555/2188385.2188395, 2012.

Liu, J., Bian, Y., Lawson, K., and Shen, C.: Probing the limit of hydrologic predictability with the Transformer network, *J. Hydrol.*, 637, 131389, https://doi.org/10.1016/j.jhydrol.2024.131389, 2024.

Feng, D., Fang, K., and Shen, C.: Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales, *Water Resour. Res.*, 56, e2019WR026793, https://doi.org/10.1029/2019WR026793, 2020.

Mao, G., Wang, M., Liu, J., Wang, Z., Wang, K., Meng, Y., et al.: Comprehensive comparison of artificial neural networks and long short-term memory networks for rainfall–runoff simulation, *Phys. Chem. Earth, A/B/C*, 123, 103026, https://doi.org/10.1016/j.pce.2021.103026, 2021.

Papacharalampous, G., Tyralis, H., and Koutsoyiannis, D.: One-step ahead forecasting of geophysical processes within a purely statistical framework, *Geosci. Lett.*, 5, 12, https://doi.org/10.1186/s40562-018-0111-1, 2018.

Siami-Namini, S., Tavakoli, N., and Namin, A. S.: The performance of LSTM and BiLSTM in forecasting time series, in: 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 9–12 December 2019, 3285–3292, https://doi.org/10.1109/BigData47090.2019.9006190, 2019.

Wu, J., Chen, X. Y., Zhang, H., Xiong, L. D., Lei, H., and Deng, S. H.: Hyperparameter optimization for machine learning models based on Bayesian optimization, *J. Electron. Sci. Technol.*, 17, 26–40, https://doi.org/10.11989/JEST.1674-862X.80904120, 2019.

Victoria, A. H. and Maragatham, G.: Automatic tuning of hyperparameters using Bayesian optimization, *Evol. Syst.*, 12, 217–223, https://doi.org/10.1007/s12530-020-09345-2, 2021.

Revision:

"It has been demonstrated that LSTM could capture historical time-step dependencies and handle variable-length input sequences using gradient optimization combined with backpropagation in hydrological applications (J. Liu et al., 2024). Bayesian LSTM (as an improved LSTM) adapts probability distributed weight parameters, which reduces the model overfitting, thereby providing robust predictions in hydrology (D. Li et al., 2021; Lu et al., 2019). The development of LSTM-based surrogates offers the possibility of accurate predictions in addressing the critical processes in lake-atmosphere modeling systems." (Section 2.2.2, Lines 209-214)

References:

Li, D., Marshall, L., Liang, Z., Sharma, A., and Zhou, Y.: Bayesian LSTM with stochastic variational inference for estimating model uncertainty in process-based hydrological models. *Water Resour. Res.*, *57*(9), e2021WR029772, https://doi.org/10.1029/2021WR029772, 2021.

Liu, J., Bian, Y., Lawson, K., and Shen, C.: Probing the limit of hydrologic predictability with the transformer network, *J. Hydrol.*, 637, 131389, https://doi.org/10.1016/j.jhydrol.2024.131389, 2024.

Lu, D., Liu, S., and Ricciuto, D.: An efficient bayesian method for advancing the application of deep learning in earth science, in: *Proceedings of the 2019 International Conference on Data Mining Workshops (ICDMW)*, IEEE, November, 270-278, https://doi.org/ 10.1109/ICDMW.2019.00048, 2019.

"The hyperparameter space included the number of hidden layers (ranging from 1 to 8), neurons per layer (ranged from 16 to 1,024), optimizer (Adam, or RMSprop), batch size (ranging from 8 to 256), and learning rate (ranging from 1E-6 to 1E-2). The hyperparameters in BO-BLSTM-based surrogates were optimized using BO with a maximum of 100 iterations, 1000 epochs for each iteration, and 50 patience in a EarlyStopping strategy." (Section 2.2.3, Lines 276-279)

"**However, we found that HyLake v1.0 required slightly higher computational costs compared to process-based models, which depend on the hyperparameters of LSTM-based surrogates, despite achieving greater performance (Table 3). In an individual case of MLW prediction, HyLake v1.0 took about 9 times longer to run compared to FLake, with a cost of 151.46 seconds. To compare different experiments of hybrid lake models, Baseline, coupled to an LSTM-based surrogate with 1 layer and 256 neurons per layer, indicated the lowest cost. While TaihuScene, constructed by an LSTM-based surrogate with 7 layers and 836 neurons per layer, showed the most expensive in predictions. Given the sophisticated architecture of LSTM-based surrogates, which inevitably leads to higher costs in training and prediction, developing novel algorithms for approximating LSTMs is urgently needed. Furthermore, the recent research progress demonstrated that** LSTM-based surrogates are more suited for short-term predictions compared to **the prevalent Transformer-based family**, which **is** suited for long-term predictions and commonly used in global weather forecasting systems (K. F. Bi et al., 2023; L. Chen et al., 2023)." (Section 4.1, Lines 614-624)

"BO-BLSTM-based surrogate exhibits superior performance in estimating LST for HyLake v1.0. This study adapted BO and EarlyStopping strategies to ensure BLSTM provides accurate and reliable estimates in prediction but increases the computational demands for training due to its ability to converge from its more complex Bayesian architecture (Peng et al., 2025; Ferianc et al., 2021). In addition, the mere 1 Bayesian fully connected layer that was adapted in this surrogate only captures limited data uncertainty, which may lose several important aspects of probabilistic prediction (Klotz et al., 2022). Given the importance of uncertainty quantification for BLSTM, it is worth noting that HyLake v1.0 has the potential to assess the variance of predictions and probabilities of lake extreme events occurrence by developing its surrogate in future (Kar et al., 2024; Gawlikowski et al., 2023). Major limitations, including high computational demands and insufficient model performance, should be addressed by developing a novel deep-learning-based surrogate based on a more efficient architecture and larger datasets." (Section 4.2, Lines 671-680)

References:

Ferianc, M., Que, Z., Fan, H., Luk, W., and Rodrigues, M.: Optimizing Bayesian recurrent neural networks on an FPGA-based accelerator, in: 2021 International Conference on Field-Programmable Technology (ICFPT), IEEE, December, 1-10, 2021.

Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., and Nearing, G.:

Uncertainty estimation with deep learning for rainfall–runoff modeling, *Hydrol. Earth Syst. Sci.*, 26, 1673–1693, https://doi.org/10.5194/hess-26-1673-2022, 2022.
Peng, Z., Mo, S., Sun, A. Y., Wu, J., Zeng, X., Lu, M., and Shi, X.: An explainable Bayesian TimesNet for probabilistic groundwater level prediction, *Water Resour. Res.*, 61, e2025WR040191, https://doi.org/10.1029/2025WR040191, 2025.

Discuss how the surrogate maintains physical consistency and whether energy balance is preserved through the hybrid coupling. Consider briefly addressing this in the discussion section.

Response: Good point. Indeed, we considered which processes in the lake model can be replaced by a deep-learning-based surrogate. Energy balance and lake water temperature approximations are two individual modules in process-based models, which are difficult to replace with deep-learning-based models simultaneously, while also ensuring numerical stability. There are two reasons to address this issue and listed in Discussion (Section 4.1, Lines 586-611):

(1) Inadequate observations to build relationships between surface conditions and heat fluxes. The energy balance equations are integrated modeling systems based on the bulk aerodynamic method from the Monin–Obukhov similarity theory, which covers the calculation of surface conditions (e.g., surface roughness length, friction velocity), as well as water and heat fluxes (e.g., latent heat, sensible heat, evaporation, precipitation-induced heat). Specifically, the latent and sensible heat fluxes are functions of transfer coefficients, which are iteratively updated using the Monin–Obukhov length, surface roughness length, and friction velocity, based on bulk flux algorithms (Verburg and Antenucci, 2010; Woolway et al., 2015). They performed well in estimating heat fluxes from the evidence in previous studies, which has been widely applied in process-based models (Woolway et al., 2020; Thiery et al., 2014). However, surface conditions in current research were always obtained from calculation instead of direct observations. Fewer studies have focused on monitoring surface conditions due to limited equipment, which hinders our ability to construct generalized lake models that reflect their potential relationships. Moreover, there is a large difference between observed surface conditions and predictions by Monin–Obukhov similarity theory, although the Lake Taihu Eddy Flux Network has monitored these conditions for a long time (Figure S6). In the future, high-quality observations and physical principles at the land-air interface should focus on addressing the significant discrepancies between observations and simulations.

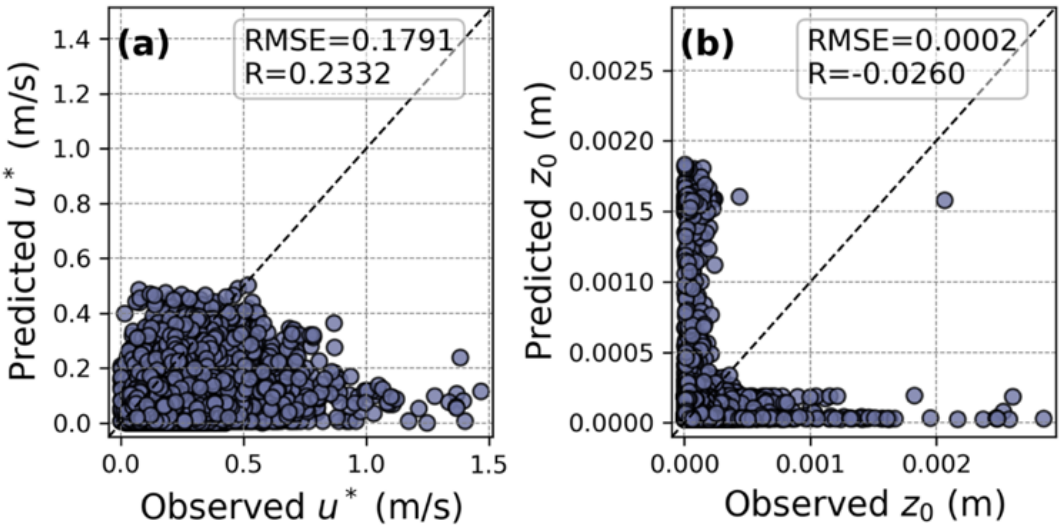


Figure S6: The comparison of friction velocity (u*) and surface roughness length (z_{0m}, m) in MLW lake site between simulation derived from PBBM and HyLake v1.0 and observations.

(2) Lake surface temperature governing equations existed uncertainly. The lake water temperature module is suitable for replacement by a deep-learning-based surrogate due to the rich and easily accessible observations and simplified schemes. Until now, accurately predicting lake water temperature using a generalized framework has remained a challenge due to the significant regional differences among lakes. Several researchers have attempted to approximate lake water temperature changes using complex integrated neural networks, such as physics-informed neural networks (PINNs), physics-guided neural networks (PGNNs), and modular networks (He et al., 2025; Ladwig et al., 2024; Read et al., 2019). These models may exhibit superior performance in specific tasks but require high computational power for

pretraining or fine-tuning, and are challenging to predict untrained variables, such as latent heat, sensible heat fluxes, and evaporation. Choose this module to replace in this study, which hopes to propose a generalized integrated framework that combines physical principles and deep learning, and then improve the understanding of lake-atmosphere interactions in finer resolutions.

References:

He, Y., and Yang, X.: A physics-informed deep learning framework for estimating thermal stratification in a large deep reservoir, *Water Resour. Res.*, 61, e2025WR040592, https://doi.org/10.1029/2025WR040592, 2025.

Ladwig, R., Daw, A., Albright, E. A., Buelo, C., Karpatne, A., Meyer, M. F., et al.: Modular compositional learning improves 1-D hydrodynamic lake-model performance by merging process-based modelling with deep learning, *J. Adv. Model. Earth Syst.*, 16, e2023MS003953, https://doi.org/10.1029/2023MS003953, 2024.

Read, J. S., Jia, X. W., Willard, J. D., Appling, A. P., Zwart, J. A., Oliver, S. K., et al.: Process-guided deep-learning predictions of lake-water temperature, *Water Resour. Res.*, 55, 9173–9190, https://doi.org/10.1029/2019WR024922, 2019.

Thiery, W. I. M., Stepanenko, V. M., Fang, X., Jöhnk, K. D., Li, Z., Martynov, A., et al.: LakeMIP Kivu: evaluating the representation of a large, deep tropical lake by a set of one-dimensional lake models, *Tellus A: Dyn. Meteorol. Oceanogr.*, 66(1), 21390, https://doi.org/10.3402/tellusa.v66.21390, 2014.

Verburg, P., and Antenucci, J. P.: Persistent unstable atmospheric boundary layer enhances sensible- and latent-heat loss in a tropical great lake: Lake Tanganyika, *J. Geophys. Res.-Atmos.*, 115, D11109, https://doi.org/10.1029/2009JD012839, 2010.

Woolway, R. I., Kraemer, B. M., Lenters, J. D., Merchant, C. J., O'Reilly, C. M., and Sharma, S.: Global lake responses to climate change, *Nat. Rev. Earth Environ.*, 1, 388–403, https://doi.org/10.1038/s43017-020-0067-5, 2020.

Woolway, R. I., Jones, I. D., Hamilton, D. P., Maberly, S. C., Muraoka, K., Read, J. S., et al.: Automated calculation of surface-energy fluxes with high-frequency lake-buoy data, *Environ. Model. Softw.*, 70, 191–198, https://doi.org/10.1016/j.envsoft.2015.04.013, 2015.

Revision:

"Moreover, simplified parameterizations in traditional process-based lake models are commonly adopted (Golub et al., 2022; Mooij et al., 2010), which influence the coupling strategies in HyLake **v1.0. The two critical components, including energy balance equations and 1-D vertical lake water temperature transport equations, compose the physical principles of lake-atmosphere interaction modeling systems, which also possess simplification to some degrees.** For example, the calculation of friction velocity (u^*) and surface roughness length (z_{0m}) in surface flux solutions has improved over time from constant empirical models to iterative routines (Hostetler et al., 1993; Woolway et al., 2015), but substantial discrepancies still exist between simulation results and observations (Figure S**6**), which in turn influence the physical principles between land surface conditions and LST. **The current approaches for solving energy balance equations uses bulk aerodynamic method based on the Monin–Obukhov similarity theory (Monin and Obukhov, 1954), and is the vital module in process-based lake models (e.g., FLake (Mironov et al., 2010), GLM (Hipsey et al., 2019), WRF-Lake (Gu et al., 2015)). However, it remains challenges to construct explainable approaches to quantify the relationships between surface conditions and fluxes and LST due to inadequate observations.** These potential differences in physical processes lead to uncertainties in training deep-learning-based surrogates, contributing to the **insufficient/limited** knowledge during model training and thereby introducing large uncertainties in hybrid models. Furthermore, the long-term trends and diurnal variations in lake water temperature profiles remain challenging to accurate approximate using the finite difference method (e.g., Crank-Nicholson solution, implicit Euler scheme) (Piccolroaz et al., 2024; Sarovic et al., 2022; Subin et al., 2012). **On top of the extensive observations of water temperature,** several hybrid models that integrate deep-learning-based and process-based models have been constructed in previous studies, achieving improved performance in model comparisons (**He et al., 2025**; Ladwig et al., 2024; Read et al., 2019). These models and their training strategies generally perform better on training and test datasets **due to their complex coupling strategies and higher computational requirements**, while their generalization and transferability need further validation. Lake Taihu, **as one of typical shallow**, eutrophic, and

large **Chinese lakes** with almost complete mixing throughout the year and subject to complex chemical and biological influences in its aquatic ecosystem, requires a suitable model as part of the temperature-solving module in the water column to predict lake water temperature and estimate other potential ecological implications under thermodynamic changes. **HyLake v1.0, developed based on *in situ* observations from Lake Taihu, has been proven to be reliable and rigorously validated in Lake Chaohu (Table 3), demonstrating a faster and more accurate framework for enhancing the understanding of hybrid hydrological modeling.**" (Section 4.1, Lines 586-611)

References added:

He, Y., and Yang, X.: A physics-informed deep learning framework for estimating thermal stratification in a large deep reservoir, *Water Resour. Res.*, 61, e2025WR040592, https://doi.org/10.1029/2025WR040592, 2025.

Monin, A. S., and Obukhov, A. M.: Basic laws of turbulent mixing in the surface layer of the atmosphere, *Contrib. Geophys. Inst. Acad. Sci. USSR*, *151*(163), e187, 2954, 1954.

Mironov, D., Heise, E., Kourzeneva, E., Ritter, B., Schneider, N., and Terzhevik, A.: Implementation of the lake-parameterization scheme FLake into the numerical-weather-prediction model COSMO, *Boreal Environ. Res.*, 15, 218–230, 2010.

Hipsey, M. R., Bruce, L. C., Boon, C., Busch, B., Carey, C. C., Hamilton, D. P., Hanson, P. C., Read, J. S., de Sousa, E., Weber, M., and Winslow, L. A.: A General Lake Model (GLM 3.0) for linking with high-frequency sensor data from the Global Lake Ecological Observatory Network (GLEON), *Geosci. Model Dev.*, 12, 473–523, https://doi.org/10.5194/gmd-12-473-2019, 2019.

Gu, H., Jin, J., Wu, Y., Ek, M. B., and Subin, Z. M.: Calibration and validation of lake surface temperature simulations with the coupled WRF-lake model. *Clim. Change*, *129*(3), 471-483, https://doi.org/10.1007/s10584-013-0978-y, 2015.

While full uncertainty quantification may be beyond the current scope, briefly discuss the uncertainty implications of the Bayesian surrogate and how this could be leveraged in future applications.

Response: The proposed Bayesian LSTM (BLSTM) in this study, an improved version of LSTM that replaces the last fully connected layer with a Bayesian fully connected layer, provides robust predictions by utilizing probability-distributed weight parameters in networks (D. Li et al., 2021; Lu et al., 2019). However, it inevitably causes uncertainties from challenging data sources and network architecture (Gawlikowski et al., 2023) and increases the computational requirements due to the complex architecture (Peng et al., 2025; Ferianc et al., 2021). The uncertainties caused by BLSTM's probability-distributed parameters, which have been widely used for assessing the variance of predictions and the probability of extreme events occurring when using out-of-bag samples, thereby improving the accuracy of decision-making for users (Kar et al., 2024; Gawlikowski et al., 2023). We are expected to improve the surrogate in HyLake v1.0 and quantify its uncertainties to further enhance our understanding of the occurrence and frequency of lake extreme events in the future. The associated revisions can be found in Materials and Methodology (Section 2.2.2, Lines 209-214), and Discussion (Section 4.2, Lines 671-680).

References:

Ferianc, M., Que, Z., Fan, H., Luk, W., and Rodrigues, M.: Optimizing Bayesian recurrent neural networks on an FPGA-based accelerator, in: 2021 International Conference on Field-Programmable Technology (ICFPT), IEEE, December, 1-10, 2021.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., et al.: A survey of uncertainty in deep neural networks. *Artif. Intell. Rev.*, *56*, 1513-1589, https://doi.org/10.1007/s10462-023-10562-9, 2023.

Kar, S., McKenna, J. R., Sunkara, V., Coniglione, R., Stanic, S., and Bernard, L.: XWaveNet: enabling uncertainty quantification in short-term ocean wave height forecasts and extreme event prediction. *Appl. Ocean Res.*, *148*, 103994, https://doi.org/10.1016/j.apor.2024.103994, 2024.

Li, D., Marshall, L., Liang, Z., Sharma, A., and Zhou, Y.: Bayesian LSTM with stochastic variational inference for estimating model uncertainty in process-based hydrological models. *Water Resour. Res.*, *57*(9), e2021WR029772, https://doi.org/10.1029/2021WR029772, 2021.

Lu, D., Liu, S., and Ricciuto, D.: An efficient bayesian method for advancing the application of deep learning in earth

science, in: *Proceedings of the 2019 International Conference on Data Mining Workshops (ICDMW)*, IEEE, November, 270-278, https://doi.org/ 10.1109/ICDMW.2019.00048, 2019.

Peng, Z., Mo, S., Sun, A. Y., Wu, J., Zeng, X., Lu, M., and Shi, X.: An explainable Bayesian TimesNet for probabilistic groundwater level prediction, *Water Resour. Res.*, 61, e2025WR040191, https://doi.org/10.1029/2025WR040191, 2025.

Revision:

"It has been demonstrated that LSTM could capture historical time-step dependencies and handle variable-length input sequences using gradient optimization combined with backpropagation in hydrological applications (J. Liu et al., 2024). Bayesian LSTM (as an improved LSTM) adapts probability distributed weight parameters, which reduce the model overfitting, thereby providing robust predictions in hydrology (D. Li et al., 2021; Lu et al., 2019). The development of LSTM-based surrogates offers the possibility of accurate predictions in addressing the critical processes in lake-atmosphere modeling systems." (Section 2.2.2, Lines 209-214)

"BO-BLSTM-based surrogate exhibits superior performance in estimating LST for HyLake v1.0. This study adapted BO and EarlyStopping strategies to ensure BLSTM provides accurate and reliable estimates in prediction but increases the computational demands for training due to its ability to converge from its more complex Bayesian architecture (Peng et al., 2025; Ferianc et al., 2021). In addition, the mere 1 Bayesian fully connected layer that was adapted in this surrogate only captures limited data uncertainty, which may lose several important aspects of probabilistic prediction (Klotz et al., 2022). Given the importance of uncertainty quantification for BLSTM, it is worth noting that HyLake v1.0 has the potential to assess the variance of predictions and probabilities of lake extreme events occurrence by developing its surrogate in future (Kar et al., 2024; Gawlikowski et al., 2023). Major limitations, including high computational demands and insufficient model performance, should be addressed by developing a novel deep-learning-based surrogate based on a more efficient architecture and larger datasets." (Section 4.2, Lines 671-680)

References added:

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., et al.: A survey of uncertainty in deep neural networks. *Artif. Intell. Rev.*, *56*, 1513-1589, https://doi.org/10.1007/s10462-023-10562-9, 2023.

Kar, S., McKenna, J. R., Sunkara, V., Coniglione, R., Stanic, S., and Bernard, L.: XWaveNet: enabling uncertainty quantification in short-term ocean wave height forecasts and extreme event prediction. *Appl. Ocean Res.*, *148*, 103994, https://doi.org/10.1016/j.apor.2024.103994, 2024.

Minor Comments

Terminology: Define LE (latent heat) and HE (sensible heat) at first mention.

Response: We have defined LE and HE in the Introduction.

Revision: "Lake-atmosphere interactions represent a tightly coupled system (B. B. Wang et al., 2019), where process-based models traditionally approximate the interdependence between LST, **latent heat (LE) and sensible heat (HE) fluxes**." (Section 1, Lines 77-79)

References: Standardize citation formats (e.g., "Hersbach et al. (2020)" vs. "Hersbach et al., 2020").

Response: We have checked the manuscript. We adopted "Hersbach et al. (2020)" when discussing their contributions and used "Hersbach et al., 2020" to cite their conclusions.

Section Organization: Consider moving deep implementation details (e.g., GUI remarks) into a Supplement or Code & Data Availability section.

Response: We now provided example bash scripts to run HyLake v1.0 and other models (e.g., Baseline, TaihuScene) in Taihu and Chaohu experiments. The example script for run these models was given by Figure R1. The information of data and code availability was given in Lines 728-732.

```
# --- Example 1: Run PBBM in MLW experiment ------------------------
python HyLake.py \
  --data_source MLW \
  --model "PB"
# --- Example 2: Run Baseline in MLW experiment --------------------
python HyLake.py \
  --data_source MLW \
  --model "Baseline"
# --- Example 3: Run HyLake v1.0 in MLW experiment -----------------
python HyLake.py \
  --data_source MLW \
  --model "LSTM"
# --- Example 4: Run TaihuScene in MLW experiment ------------------
python HyLake.py \
  --data_source MLW \
  --model "Taihu_LSTM"
# --- Example 5: Run HyLake v1.0 in Taihu-ERA5 experiment-----------
python HyLake.py \
  --data_source ERA5 \
  --model "LSTM"
# --- Example 5: Run TaihuScene in Taihu-ERA5 experiment------------
python HyLake.py \
  --data_source ERA5 \
  --model "Taihu_LSTM"
# --- Example 6: Run HyLake v1.0 in 1 grid of Chaohu experiment ----
python HyLake.py \
  --Lake_Name Chaohu \
  --Lake_Lat 31.53 \
  --Lake_Lon 117.31 \
  --Lake_altitude -4 \
  --Lake_depth 4 \
  --SimLength 35040 \
  --initial_temp 5.0 \
  --model "LSTM" \
  --data_source custom \
  --csv_path "./data/chaohu/ERA5_forcings/forcing_csv/Chaohu_forcing_lat31.53_lon117.31.csv" \
  --exp "Chaohu_Lake_lat31.53_lon117.31" \
  --col_indices 1,2,3,4,5,6,7
```

Figure R1: The example scripts for run HyLake v1.0 for MLW, Taihu-Obs and Taihu-ERA5 experiments.

Revision: "***Code and data availability.*** The datasets, codes **and scripts** of HyLake v1.0 and other models **(e.g., Baseline, TaihuScene)** used in this study are available at https://doi.org/10.5281/zenodo.15289113 (He et al., 2025). **FLake model was run via LakeEmsemblR tool (https://aemon-j.github.io/LakeEnsemblR/).** The ERA5 reanalysis datasets can be downloaded from the Climate Data Store (https://cds.climate.copernicus.eu/). Observations of lake surface water temperature, latent and sensible heat fluxes at Lake Taihu are available at Harvard Dataverse (https://doi.org/10.7910/DVN/HEWCWM; Zhang et al., 2020)." (Code and data availability, Lines 728-732)

Caption Detail: Enhance figure captions to specify whether plotted values are observed or simulated and note dataset origins (real vs. semi-synthetic).

Response: We have improved the captions to clearly describe the datasets in Figures 5 to 11.

Reviewer #4:

He and Yang present a new model, the Hybrid Lake Model v1.0. With this model He et al. want to approximate LST changes which are a crucial indicator of climate change in the Earth system. Their model combines process-based with deep learning methods. Their results show that HyLake outperforms other models. The study is interesting and may be published, but the manuscript needs major revisions before it can be considered for publication. It is essential that the content of the study and presented results become more clear and the study understandable for a broader readership. Without that it is quite tough to assess the quality of the here presented results.

Response: We thank Reviewer #4 for the careful review and constructive comments. We have reorganized this manuscript and provided a point-by-point response. All comments are accepted and **Relisted in black**, followed by our **Replies in blue** and **Revisions in red (highlighted revisions in bold)**. The following table summarizes the major changes addressing the reviewer's comments.

No.	Major Revisions	Important messages
1	Clarified the usage of words, including model vs. experiments, evaluation vs. validation, and etc.	(1) This study inter-compared 5 lake thermodynamics models, including PBBM, FLake, Baseline, TaihuScene and HyLake v1.0, via 4 suites of numerical experiments against observations (MLW, Taihu-obs, Taihu-ERA5 and Chaohu) to assess the models' performance (Materials and Methodology). (2) We considered about the usage of words in the full text. Specifically, this study used "validation" to assess the model accuracy (e.g., RMSE, MAE, R), while "evaluation" was used to assess models' abilities (e.g., transferability). The "model" included FLake, PBBM, Baseline, TaihuScene, and HyLake v1.0, "experiments" means the models using in different regions or forcing datasets, including MLW, Taihu-obs, Taihu-ERA5 and Chaohu experiment.
2	Explained the datasets used in this study	There are 2 datasets used in this study, including hydrometeorological variables from 5 lake sites in Lake Taihu eddy flux network, and meteorological variables from ERA5 datasets. Specifically, meteorological variables in these two datasets were used to force models in different experiments, hydrometeorological variables, such as lake surface temperature, latent and sensible heat fluxes, were used to validate the models for each experiment (Materials and Methodology).
3	Improved language and presentation throughout the manuscript	We have one-by-one improved and rephrased the sentences in the manuscript according to comments.

General comments:

There are many weird sentences in the text which do not make sense or are misleading. I will provide several examples in the specific comments.

Response: Sorry for the confusion. We have carefully revised the statements based on the comments, provided more details to explain methods and results, and improved the language and presentation throughout the manuscript.

The abstract should be significantly improved and clearly state what has been done in this study and what are the major results.

Response: Thanks for the careful review. We have revised the Abstract to clearly describe what has been done and the key findings of this study.

Revision: "Abstract: Lake-atmosphere interactions, which significantly modulate the impacts of climate change

on land-air water and heat exchange, play a critical role in Earth system dynamics. However, modeling key indicators of these interactions, lake surface temperature (LST) and latent heat (LE) and sensible heat (HE) fluxes, remains challenging. This stems from oversimplified physics in traditional process-based models and the limited interpretability of purely data-driven "black-box" structure. **Hybrid models unifying physical principles with sparse observations offer a promising solution for simultaneously predicting lake-atmosphere interactions.**
This study **presents the** Hybrid Lake Model v1.0 (HyLake v1.0), which integrates a Bayesian Optimized Bidirectional Long Short-Term Memory-based (BO-BLSTM-based) surrogate trained **on data** from Meiliangwan (MLW) site in Lake Taihu to approximate LST **dynamics. LE and HE are subsequently derived using** surface energy balance equations. **We intercompare HyLake v1.0 against the Freshwater Lake (FLake) model** and hybrid lake models **using** different surrogates **(Baseline and TaihuScene) across multiple Lake Taihu sites. Forcing datasets include eddy flux covariance observations and ECMWF Reanalysis v5 (ERA5) datasets.**
Results demonstrate HyLake v1.0's capability to predict lake-atmosphere interactions with satisfactory performance. At MLW, HyLake v1.0 **outperformed all models, achieving** R and RMSE of 0.99 and 1.08 °C **for** LST, R and RMSE of 0.94 and 24.65 W/m^2 **for** LE and R and RMSE of 0.93 and 7.15 W/m^2 **for** HE**, respectively.** To assess model generalization and transferability in ungauged lake sites, HyLake v1.0 exhibited superior performance **across all lake sites compared to FLake,** with MAEs of 0.85 °C **(LST)**, 21.56 W/m^2 **(LE)** and 6.63 W/m^2 **(HE)**. **When forced by** ERA5 datasets, HyLake v1.0 **outperformed benchmarks** for 14 of 15 variables (including LST, LE, and HE across 5 lake sites), **yielding** MAEs of 0.90 °C **(LST)**, 35.02 W/m^2 **(LE)** and 7.97 W/m^2 **(HE). It indicates strong capacity for application with unlearned forcing data. HyLake v1.0 exhibits excellent skill in estimating interactions for untrained lake sites, supporting its potential for extending applications to other ungauged lakes. This advancement promotes hybrid modeling techniques in Earth system science, enhancing understanding of land-atmosphere interaction dynamics.**" (Abstract, Lines 9-30)

The entire manuscript needs are clear writing and thus needs to be rewritten. There are many repetitions on one hand, but on the other hand a mixed terminology is used as e.g. evaluation and validation; model, model results and model experiments; surrogates so that it does not become clear to the reader what has been used and what exactly has been done and which models/data sets are compared.

Response: We have double-checked and improved the terminology of the manuscript, making sure that the terms are consistent and precise. The usage of terminology was listed as follows:

(1) Evaluation vs. Validation: After careful consideration of the usage of "Evaluation" and "Validation", we believe that "evaluation" is used to assess the model's ability, such as its generalization and transferability; while "validation" is used to validate the model's accuracy, which is represented by R, RMSE, and MAE. In the current manuscript, we have revised the usage of these two terms to help readers understand.

(2) Model vs. Experiment: We are sorry for the incorrect usage of "model" and "experiment". Now, we reorganized this manuscript and corrected the usage of these two terms. This study inter-compared 5 lake thermodynamics models, including PBBM, FLake, Baseline, TaihuScene and HyLake v1.0, via 4 suites of numerical experiments against observations (MLW, Taihu-obs, Taihu-ERA5 and Chaohu) to assess the models' performance. Specific information is relisted in Table 2 and 3.

(3) Surrogate for models: Surrogates are deep-learning-based models used to replace the Euler Scheme in traditional process-based models. They are individual modules for different hybrid lake models. For example, Baseline was coupled to an LSTM-based surrogate that was trained on the outputs of PBBM; HyLake v1.0 was coupled to a BO-BLSTM-based surrogate that was trained on the MLW observations.

Revision: "Table 2. Specification of each model for intercomparison.

Model	Forcing datasets	Surrogate	Training datasets	Description
PBBM	\	\	\	Backbone for HyLake v1.0
FLake	ERA5; observations	\	\	A process-based freshwater lake

				model for intercomparison
Baseline	MLW	LSTM	PBBM outputs	A baseline experiment using PBBM outputs for model intercomparison
TaihuScene	ERA5; observations	BO-BLSTM	All observations	A numerical experiment using large train dataset to train surrogate
HyLake v1.0	ERA5; observations	BO-BLSTM	MLW observations	Proposed hybrid lake model in this study

" (Table 2)

"**Table 3: Intercomparison of model performance across different experiments conducted in diverse regions with different forcing datasets. Observations from all lake sites (MLW, DPK, BFG, XLS, and PTS) on Lake Taihu, were used to drive models in the Taihu-obs experiment. Bold values in the table present the best-performing model with each group of experiments. Computational efficiency is reported as the runtime for a single simulation.**

Exp	Model	Forcing	R			RMSE			MAE			Efficiency
			LST	LE	HE	LST	LE	HE	LST	LE	HE	(s)
MLW	PBBM	MLW	0.98	0.85	0.89	1.78	38.34	9.37	1.38	23.54	6.22	189.49
	FLake	MLW	0.98	0.82	0.84	1.76	42.73	7.24	1.35	24.76	5.01	**16.40**
	Baseline	MLW	0.96	0.74	0.75	2.71	51.77	14.63	2.11	33.52	9.30	151.46
	HyLake v1.0	MLW	**0.99**	**0.94**	**0.93**	**1.08**	**24.65**	**7.15**	**0.85**	**15.18**	**4.73**	270.21
Taihu-obs	FLake	All sites	0.97	0.61	0.74	2.24	15.46	69.11	1.69	41.95	10.44	**89.00**
	TaihuScene	All sites	**0.99**	**0.82**	0.89	1.52	14.93	43.49	1.23	29.53	10.63	6928.44
	HyLake v1.0	All sites	**0.99**	0.81	**0.90**	**1.36**	**11.19**	**39.20**	**1.03**	**24.79**	**7.88**	2693.23
Taihu-ERA5	FLake	ERA5	0.98	0.63	0.69	1.82	12.31	67.24	1.46	50.94	9.68	**19.60**
	TaihuScene	ERA5	**0.99**	0.68	0.73	1.60	13.00	64.83	1.29	47.78	10.11	652.25
	HyLake v1.0	ERA5	**0.99**	**0.71**	**0.78**	**1.12**	**11.05**	**49.48**	**0.90**	**35.02**	**7.97**	236.78
Chaohu	FLake	ERA5	**0.97**	\	\	2.28	\	\	1.76	\	\	**70.40**
	HyLake v1.0	ERA5	**0.97**	\	\	**2.07**	\	\	**1.57**	\	\	972.83

" (Table 3)

Specific comments:

P1, L13: What exactly do you mean with "has yet to fully benefit from the integration of process-based and deep learning based models."? Do you mean these processes need to be still integrated in these models? Please rephrase the sentence to be more clear.

Response: This sentence describes the advantages of developing hybrid models by integrating process-based and deep-learning-based models. We have rephrased this sentence to be clearer.

Revision: "Hybrid models that unifying physical principles with sparse observations offer a promising solution for simultaneously predicting lake-atmosphere interactions." (Abstract, Lines 13-14)

P1, L16: What is "FLake"? Is this a ML model or a process-based model?

Response: FLake is a traditional process-based lake model (Mironov et al., 2010). It has been widely coupled to land surface models and applied in many lakes. The associated information has been added to the Abstract.

References:

Mironov, D., Heise, E., Kourzeneva, E., Ritter, B., Schneider, N., and Terzhevik, A.: Implementation of the lake-parameterization scheme FLake into the numerical-weather-prediction model COSMO, *Boreal Environ. Res.*, 15, 218–230, 2010.

Revision: "**We intercompare HyLake v1.0 against the Freshwater Lake (FLake) model** and hybrid lake models **using** different surrogates **(Baseline and TaihuScene) across multiple Lake Taihu sites. Forcing datasets include eddy flux covariance observations and ECMWF Reanalysis v5 (ERA5) datasets.**" (Abstract, Lines 17-20)

P1, L17-19: Compared to what does HyLake outperform other models? What has been used as reference?

Response: HyLake v1.0 outperformed other models in lake surface temperature, latent heat and sensible heat fluxes compared to the observations. This sentence has been rephrased.

Revision: "**Results demonstrate HyLake v1.0's capability to predict lake-atmosphere interactions with satisfactory performance. At MLW,** HyLake v1.0 **outperformed the best among all models, achieving** R and RMSE of 0.99 and 1.08 °C **for** LST, R and RMSE of 0.94 and 24.65 W/m^2 **for** LE and R and RMSE of 0.93 and 7.15 W/m^2 **for** HE**, respectively.**" (Abstract, Lines 21-23)

P1, L21: What do you mean with "Under ERA5 reanalysis datasets"? This does not make any sense and needs to be rephrased.

Response: In the Taihu-ERA5 experiment, we used meteorological variables obtained from ERA5 datasets to force FLake, TaihuScene, and HyLake v1.0. The results indicated that HyLake v1.0 performed the best, demonstrating that HyLake v1.0 has a strong capability to apply to the unlearned forcing datasets. We have rephrased this sentence in Abstract (Lines 25-28). The detailed information can be found in Materials and Methodology (Section 2.1, Lines 114-139), which will be given in the following response.

Revision: "**When forced by ERA5 datasets, HyLake v1.0 outperformed benchmarks** for 14 of 15 variables **(including LST, LE, and HE across 5 lake sites), yielding** MAEs of 0.90 °C **(LST)**, 35.02 W/m^2 **(LE)** and 7.97 W/m^2 **(HE).** It indicates strong **capacity for application with unlearned forcing datasets.**" (Abstract, Lines 25-28)

P1, L21-23: What is meant with "generalization and transferability"? Concerning what is HyLake indicating a strong generalization and transferability?

Response: Generalization and transferability are the most important features and functions of deep learning. Specifically, the generalization ability of deep-learning-based models presents test-time performance. Successful deep artificial neural networks can exhibit a remarkably small gap between training and test performance (Zhang et al., 2021). Transferability of deep-learning-based models means the ability of models to be applied to cross-domain tasks (Long et al., 2016). Here, we assess the generalization and transferability of HyLake v1.0 using four groups of experiments. In the Abstract, we rephrased these sentences to help readers understand what we have done in this study. The specific information about model evaluation of generalization and transferability will be explained in the following comments.

References:

Long, M., Cao, Y., Wang, J., and Jordan, M.: Learning transferable features with deep adaptation networks. In *International conference on machine learning*. June, PMLR, 97-105, 2015.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O.: Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, *64*(3), 107-115, 2021.

Revision: "**When forced by ERA5 datasets, HyLake v1.0 outperformed benchmarks** for 14 of 15 variables **(including LST, LE, and HE across 5 lake sites), yielding** MAEs of 0.90 °C **(LST)**, 35.02 W/m^2 **(LE)** and 7.97 W/m^2 **(HE).** It indicates strong **capacity for application with unlearned forcing datasets.**" (Abstract, Lines 25-28)

P1, L26-27: The last sentence is in my opinion a repetition of what has been said before and is thus obsolete.

Response: This sentence has been rephrased to highlight the contribution and potential of HyLake v1.0 proposed in this study.

Revision: "This advancement promotes hybrid modeling techniques in Earth system science, enhancing understanding of land-atmosphere interaction dynamics" (Abstract, Lines 29-30)

P3, L76-77: The abbreviations HE and LE should be introduced here once again.

Response: Corrected.

Revision: "Lake-atmosphere interactions represent a tightly coupled system (B. B. Wang et al., 2019), where process-

based models traditionally approximate the interdependence between LST, **latent heat (LE) and sensible heat (HE) fluxes.**" (Introduction, Lines 77-79)

P3, L82-82: "where differ significantly in its biological characteristics" is not clear and the sentence should be rephrased.
Response: It has been rephrased.
Revision: "Traditional lake models seem challenging to be generalized in ungauged lake or even regions in a large lake. Lake Taihu, the third largest freshwater lake in China**, which indicates a significant regional difference in its biological characteristics (Table 1),** has experienced **severe** deterioration in water quality, thereby significantly threatening drinking water security (Zhang et al., 2020; Yan et al., 2024)." (Introduction, Lines 83-86)

P3, L88-89: What is the difference here between "validate" and "evaluate". To which data sets has HyLake been evaluated or validated?
Response: "Validate" was used to assess the model accuracy (e.g., R, RMSE, MAE) in this study, while "evaluate" was used to assess the models' abilities (e.g., transferability). The models' generalization and transferability are both assessed by statistical metrics that are calculated from observations and predictions. ERA5 datasets were used as forcing datasets to fill the data gaps of observations and individually force the models. Here we improved these sentences to describe the research objectives in this study clearly.
Revision: "**To improve novel hybrid modeling techniques and enhance the understanding of lake-atmosphere interactions, t**he objectives of this study are to (1) develop a novel hybrid lake model HyLake v1.0 by embedding LSTM-based surrogate into process-based model; (2) validate the performance of HyLake v1.0 in LST, LE, and HE **based on observations from Taihu Lake Eddy Flux Network**; and (3) evaluate the transferability of HyLake v1.0 in **ungauged** lake sites with different biological characteristics **using ECMWF Reanalysis v5 (ERA5) forcing** datasets." (Introduction, Lines 89-95)
P3, L90: What do you mean with "under ERA5 reanalysis datasets"? This does not make any sense. Please rephrase the sentence.
Response: ERA5 datasets provided meteorological variables, including air temperature, dew point temperature, wind speed, net radiation fluxes, surface pressure, and precipitation, which were used to force the lake models in the Taihu-ERA5 and Chaohu experiments and to fill data gaps at some lake sites in the Lake Taihu eddy flux network. This sentence has been rephrased to describe the functions of ERA5 datasets clearly.
Revision: "**To improve novel hybrid modeling techniques and enhance the understanding of lake-atmosphere interactions, t**he objectives of this study are to (1) develop a novel hybrid lake model HyLake v1.0 by embedding LSTM-based surrogate into process-based model; (2) validate the performance of HyLake v1.0 in LST, LE, and HE **based on observations from Taihu Lake Eddy Flux Network**; and (3) evaluate the transferability of HyLake v1.0 in **ungauged** lake sites with different biological characteristics **using ECMWF Reanalysis v5 (ERA5) forcing** datasets." (Introduction, Lines 89-95)

P3, L90-91: Why? Is this a result from your validation/evaluation?
Response: We have reorganized this sentence to show the promising of the development of HyLake v1.0.
Revision: "The **results will provide reliable evidence for improving lake-atmosphere interactions modeling by unifying physical principles and deep learning in ungauged regions.**" (Introduction, Lines 95-96)

P3, L95: rapid increase of what? The water temperature? Please be more clear. Additional questions I have are if this increase is based on observations and if these are climate change induced increases or increases due to other reasons.
Response: Sorry for the missing information. Zhang et al. (2018) indicated that lake water temperature would increase at a rate of ~0.37 °C per decade based on the observations, which has been corrected in Materials and Methodology (Section 2.1, Lines 98-100). There are no conclusions about the attribution of lake warming; however, its trends are consistent with the increase in air temperature, with a rate of 0.36 °C per decade. Therefore, we preferred that climate

change is the primary factor influencing lake thermodynamics. We hope to further elucidate the potential causes of lake warming in the future by utilizing more advanced tools.

References:

Zhang, Y. L., Qin, B. Q., Zhu, G. W., Shi, K., and Zhou, Y. Q.: Profound changes in the physical environment of Lake Taihu from 25 years of long-term observations: implications for algal-bloom outbreaks and aquatic-macrophyte loss, *Water Resour. Res.*, 54, 4319–4331, https://doi.org/10.1029/2017WR022401, 2018.

Revision: "Lake Taihu (30.12–32.22°N, 119.03–121.91°E), located in the Yangtze Delta, is the third-largest freshwater lake in China, covering an area of 2,400 km² with an average depth of 1.9 m, with a rapid increasing rate of ~0.37 °C/decade **in LST** (Yan et al., 2024; Zhang et al., 2020; Zhang et al., 2018)." (Section 2.1, Lines 98-100)

P3, L96-97: Sentence grammatically not correct, please improve.

Response: Corrected.

Revision: "As a typical urban lake, **Lake Taihu is** situated in one of the most densely populated regions of China. **It** has experienced significant eutrophication, **characterized** by **recurrent** algae blooms that threaten local drinking water security (Yan et al., 2024)." (Section 2.1, Lines 100-102)

P4, L110: The introduction of the abbreviations LE and HE should be done already in L76 (see my comment above).

Response: Corrected.

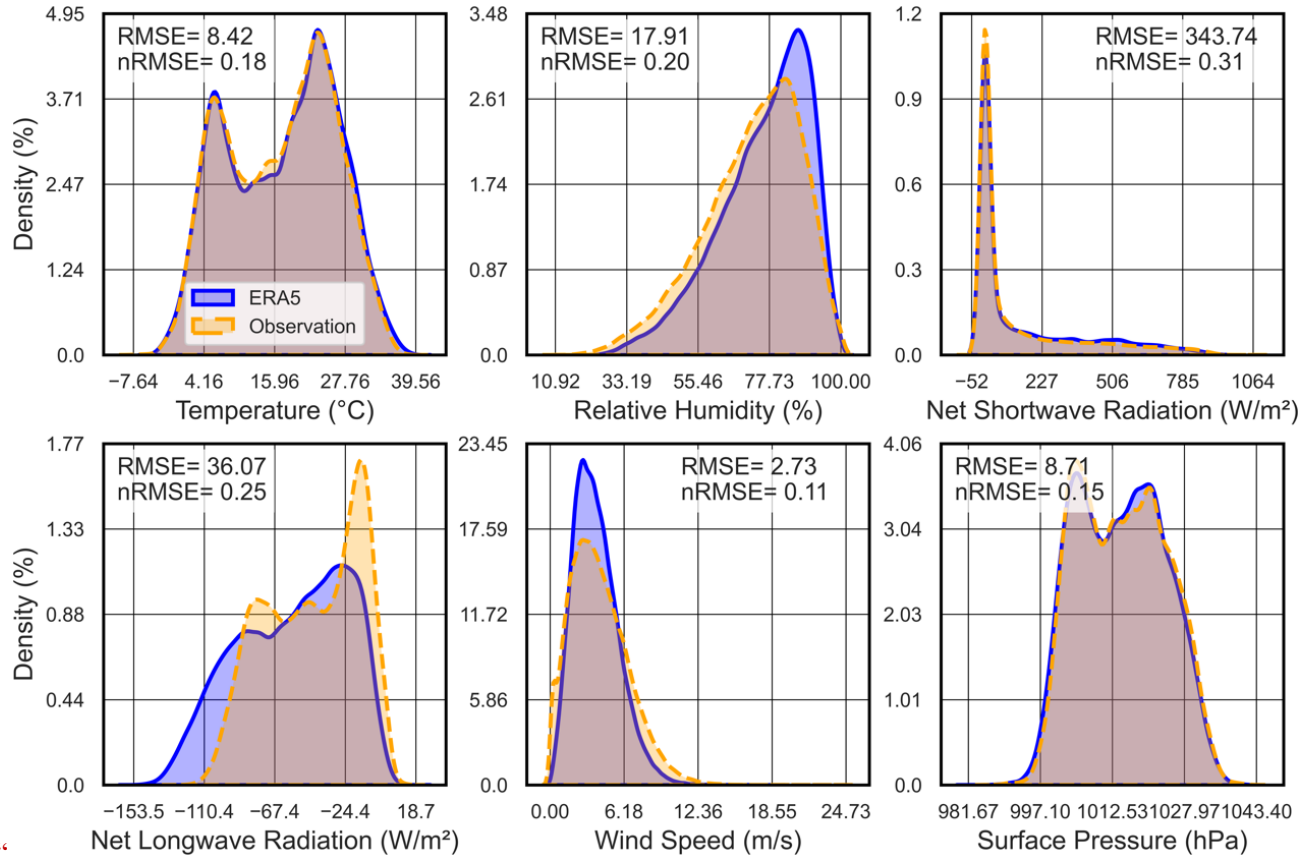
Revision: "**Within the network, e**ach site is equipped with an eddy covariance system that continuously monitors **LE and HE** using sonic anemometers and thermometers (Model CSAT3A; Campbell Scientific, Logan, UT, USA) positioned 3.5 to 9.4 m above the lake surface." (Section 2.1, Lines 116-118)

P4, L117: "using ERA5 reanalysis data sets". Which parameters are used from ERA5? HE and LE? How accurate is the data?

Response: We used meteorological variables from ERA5 datasets, including air temperature, wind speed, net radiation fluxes, surface pressure, dew point temperature, and precipitation. These data were used to (1) force lake models in Taihu-ERA5 and Chaohu experiments, and (2) fill data gaps in the observations in the Lake Taihu eddy flux network. The LE and HE are obtained from the Lake Taihu eddy flux network, which is observed to validate the model accuracy in lake sites. The meteorological variables from ERA5 are in great agreement with the observations, as shown in Figure S1. The associated revisions can be found in Materials and Methodology (Section 2.1, Lines 114-137).

Revision: "**The datasets included two parts: (1) hydrometeorological variables observed from the Taihu Lake Eddy Flux Network to force and validate the models, and (2) meteorological variables from ERA5 datasets to fill the gaps of observations and force the models. Within the network,** each site is equipped with an eddy covariance system that continuously monitors **LE and HE** using sonic anemometers and thermometers (Model CSAT3A; Campbell Scientific, Logan, UT, USA) positioned 3.5 to 9.4 m above the lake surface. Hydrometeorological variables, including air humidity and temperature (Model HMP45D/HMP155A; Vaisala, Helsinki, Finland), wind speed (Model 03002; R.M. Young Co., Traverse City, MI, USA), and net radiation components (Model CNR4; Kipp & Zonen, Delft, the Netherlands), are also measured. **These meteorological variables were used to force lake models while LE, HE and LST from observations were used to validate the results of each numerical experiment, on top of which, the** inferred radiative LST were collected at 30-minute intervals that are publicly accessible via Harvard DataVerse (Lee, 2004; Zhang et al., 2020; https://doi.org/10.7910/DVN/HEWCWM). The dataset spans from 2012 to 2015 **and contains several data gaps across these lake sites. Specifically, 475 time steps (~1.36%) of observed surface pressure were found missing at the DPK site during 2012 and 2015; 7,959 time steps (~22.71%) of all observed variables were missing at the XLS site; 12,539 time steps (~35.78%) of all observed variables were missing at the PTS site. Observations at the MLW and BFG sites were complete during the entire study periods. For the model evaluation of Taihu-obs experiment, the data gaps of observed variables in these lake sites were directly filled by ERA5 datasets at the corresponding time steps, which were used to predict lake-atmosphere interactions.** In this study,

observed meteorological variables from the MLW site**, an eutrophic lake site that presents the trophic status of Lake Taihu (Table 1, Wang et al., 2019),** are used to train the Long Short-Term Memory (LSTM)-based surrogates (Sect. 2.2); while data from the remaining sites serve to evaluate the generalization of HyLake v1.0 and train the LSTM-based surrogates. To **further** address the generalization **and transferability** of HyLake v1.0 across different forcing datasets, this study utilized **8 meteorological variables that obtained from** hourly ERA5 datasets from 2012 to 2015, with a spatial resolution of 0.25° at a single level to **force** HyLake v1.0. These datasets, available from the Climate Data Store (Hersbach et al., 2020; http://cds.climate.copernicus.eu), include variables such as air temperature, dew point temperature, surface pressure, wind speed, and surface net longwave and shortwave radiation**, which has similar probability distribution to observations across Lake Taihu (Figure S1).**" (Section 2.1, Lines 114-137)



"

Figure S1: The probability density distribution of meteorological variables from observation and ERA5 reanalysis datasets in MLW, BFG, DPK, PTS, and XLS site during 2012 to 2015. A normalized RMSE (nRMSE) was assigned to assess the error between observation and ERA5 reanalysis datasets." (Figure S1 in Supplementary Materials)

P4, L124-125: If ERA5 is used to fill up data gaps , it should not be used for evaluation.

Response: ERA5 datasets were not only used for gap filling but also forcing models. This sentence has been rephrased.

Revision: "The ERA5 datasets are **also individually** used to **force FLake and TaihuScene for comparison and predict lake-atmosphere interactions** in Lake Taihu, providing insights into the model's generalization, transferability and performance **using** different climatic forcing datasets." (Section 2.1, Lines 137-139)

P5, L134: What is meant with "variants"? Do you mean variables? What exactly are you doing here? Are you using different set-ups, thus performing sensitivity simulations?

Response: We have deleted this sentence. The variants of HyLake v1.0 refer to the hybrid lake models coupled to different surrogates. Specifically, Baseline and TaihuScene are variants of HyLake v1.0. Baseline used an LSTM-based surrogate that was trained on the outputs of PBBM; TaihuScene used a BO-BLSTM-based surrogate that was trained on observations from 5 lake sites in the Lake Taihu eddy flux network. Both of these surrogates are different from the BO-BLSTM-based surrogate in HyLake v1.0, which was trained on MLW observations.

P6, L143-148: Hasn't the same you have written here been written in slightly different wording already in the previous paragraph? Please avoid repetitions.

Response: We have reorganized this paragraph to avoid the repetitions. The associated revisions can be found in Materials and Methodology (Section 2.2.1, Lines 155-160).

Revision: "**A process-based backbone lake model (PBBM) is separately constructed to serve as the backbone of HyLake v1.0, which** referred to the process-based lake models based on the governing equations and parameterization schemes of previously validated lake physical processes (Sarovic et al., 2022). The conceptual model of PBBM is depicted in Figure 2 and Table 2. Specifically, the lake-atmosphere modeling system in PBBM primarily involves energy balance equations for solving **LE and HE** at the lake-atmosphere interface and the 1-D vertical lake water temperature transport equations within the water column **for solving LST** (Piccolroaz et al., 2024)." (Section 2.2.1, Lines 155-160)

P8, L197-198: This sentence does not make sense. "LST" is a parameter while the "Euler scheme" is a method.

Response: This constructed several LSTM-based surrogates to solve ΔLST (changes in LST) for each time step, instead of using the Euler scheme in traditional process-based models. The LST for each time step (t) can be calculated from the LST at the previous time step (t-1) plus ΔLST derived from surrogates. We have been reorganized this sentence in Materials and Methodology (Section 2.2.2, Lines 214-218).

Revision: "HyLake v1.0 and **other hybrid lake models,** including Baseline and TaihuScene, employed LSTM-based surrogates rather than the implicit Euler scheme **in process-based models to solve LST** for each time step (Figure 3a). Specifically, several sequence-to-one LSTM-based surrogates are adapted to be trained to approximate **ΔLST (the difference of LST between two time steps)** based on dynamic inputs, including time series of historical 24-step variables of LST, friction velocity (u^*, m/s), surface roughness length (z_{0m}, m), and $G(0)$." (Section 2.2.2, Lines 214-218)

P8, L199: What is meant with "increments in LST"? Do you mean components that affect LST?

Response: Corrected. It means ΔLST, which is the difference between LST in current (t) and previous time step (t-1).

Revision: "Specifically, several sequence-to-one LSTM-based surrogates are adapted to be trained to approximate **ΔLST (the difference of LST between two time steps)** based on dynamic inputs, including time series of historical 24-step variables of LST, friction velocity (u^*, m/s), surface roughness length (z_{0m}, m), and $G(0)$." (Section 2.2.2, Lines 215-218)

P8, L200: What is G(0)?

Response: It is the net heat flux, which was defined in Section 2.2.1 and Eq. (1) (Lines 161-170):

The changes in LST are primarily driven by the net heat fluxes entering the lake surface. Therefore, the net heat flux is imposed as a Neumann boundary condition at the upper boundary of the water column. Following Piccolroaz et al. (2024), the net heat flux $G(0)$ (W/m²) into the lake surface can be expressed by the energy balance equation:

$$G(0) = (1 - r_s)H_s + (1 - r_a)H_a + H_c + H_e + H_p \quad (1)$$

where H_s (W/m^2) and H_a (W/m^2) represent net downward shortwave and longwave radiation (also referred to the net solar and thermal radiation in ERA5), respectively; r_s and r_a account for the shortwave and longwave albedos of water; the HE and LE are denoted by H_c (W/m^2) and H_e (W/m^2); H_p represent the heat flux (W/m^2) brought from precipitation, often calculated via an empirical equation to quantify (Sarovic et al., 2022). All heat fluxes are considered positive in downward direction. The net shortwave and longwave radiation are derived from observation in Lake Taihu eddy flux network and ERA5 reanalysis datasets.

P8, L204-205: The sentence is not clear and needs to be rephrased. What do you mean with different models? To my understanding you are not using different models, these are rather different model runs.

Response: *NN(·)* donates different LSTM-based surrogates within HyLake v1.0, Baseline and TaihuScene. This study constructed the above-mentioned 3 hybrid lake models, which have different LSTM-based surrogates. Specifically, Baseline is coupled to an LSTM-based surrogate trained on the outputs of PBBM. TaihuScene is another hybrid lake

model that is coupled to a BO-BLSTM-based surrogate trained on observations from all sites (MLW, BFG, DPK, PTS, and XLS) in Lake Taihu, which differs from HyLake v1.0. This sentence has been corrected.

Revision: "where *NN(·)* donates **different** LSTM-based surrogates within **HyLake v1.0, Baseline and TaihuScene**, which will activate to approximate the increment of lake surface temperature for each time step. " (Section 2.2.2, Lines 222-223)

P10, L225: For non LSTM users it should be explained what the "forget gate" is.

Response: The LSTM unit comprises three gates: the forget gate, the input gate, and the output gate, which control whether information should be retained or updated (Figure R1). Specifically, the forget gate decides what information we're going to throw away from the cell state; the input gate, as the second gate, decides what new information we're going to store in the cell state; the output gate decides what we're going to output. These 3 gates control the data in and out. Considering we did not improve the LSTM architecture, we don't think that we should explain more about the fundamental concept of LSTM.

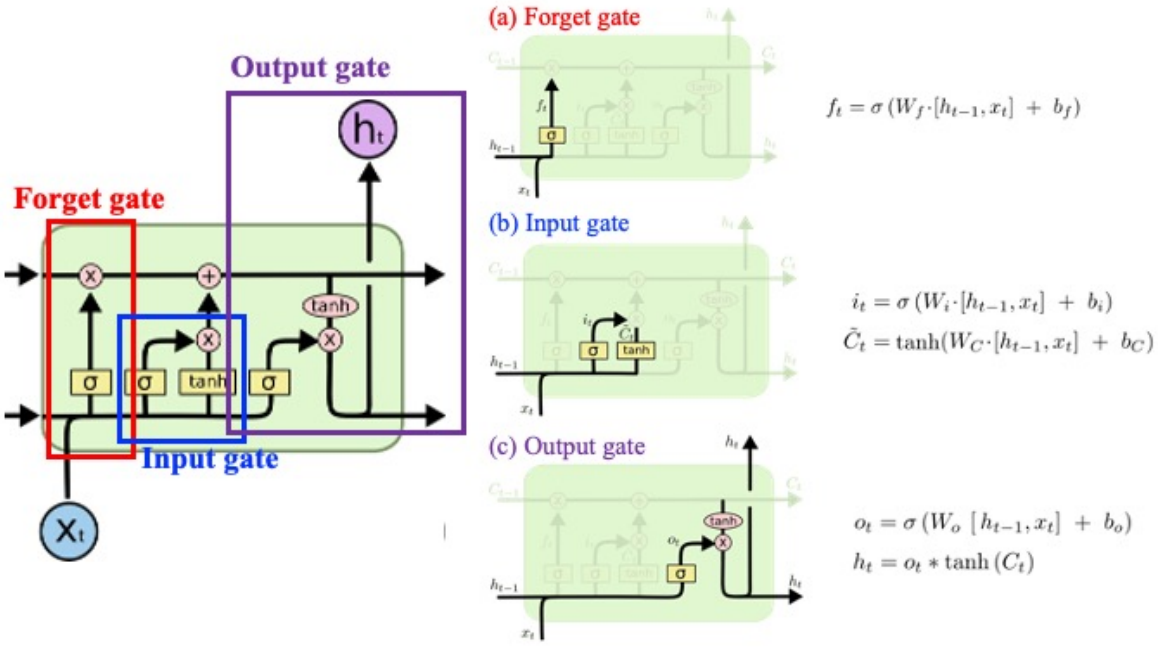


Figure R1: The architecture LSTM unit and 3 gates inside the unit.

P11, L250: What is an "Adam optimizer"?

Response: It is one of the common Adaptive optimization algorithms that are used to help LSTM-based surrogates minimize the loss. These algorithms aim to automatically adapt the learning rate to different parameters based on the statistics of the gradient. In this study, it was found that using the Adam Optimizer in LSTM-based surrogates of the Baseline is the best through manual adjustment of the optimizers.

References:

Zhang, Z.: Improved adam optimizer for deep neural networks. In: 2018 IEEE/ACM 26th international symposium on quality of service (IWQoS), IEEE, June, 1-2, 2018.

Adam, K. D. B. J.: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 1412(6), 2014.

P11, L259-261: 10% and 10% correct? I think it would be easier for the reader if the information listed here would be put into a table.

Response: It is correct. This study divided the studied period (2013-2015) into three parts, including the training period (2013-01-01 00:00:00 to 2015-05-26 04:00:00), the validation period (2015-05-26 04:00:00 to 2015-09-12 14:00:00), and the test period (2015-09-12 14:00:00 to 2015-12-30 23:00:00), according to 80%, 10% and 10%. The associated

information was updated in Materials and Methodology (Section 2.2.3, Lines 280-283).

Revision: "Training, validation, and test datasets for each lake site were divided by 80%, 10% and 10% **of the length of time series (2013-2015), respectively. They are** divided into 2013-01-01 00:00:00 to 2015-05-26 04:00:00, 2015-05-26 04:00:00 to 2015-09-12 14:00:00, and 2015-09-12 14:00:00 to 2015-12-30 23:00:00." (Section 2.2.3, Lines 280-283)

P11, L266-267: Rephrase/Improve sentence "The briefly introduction……"

Response: We have deleted the unclear description and reorganized this paragraph (Section 2.3.1, Lines 286-291).

Revision: "To address the generalization and transferability of HyLake v1.0 in studied (MLW) and ungauged lake sites (DPK, BFG, XLS, and PTS) (Table 1), this study further **conducted** three numerical experiments**, including MLW experiment, Taihu-obs experiment, Taihu-ERA5 experiment, and Chaohu experiment,** using distinct **models** and forcing datasets (Table **2 and 3**), including FLake, **Baseline, and TaihuScene to intercompare. Baseline and TaihuScene serve as extended models of HyLake v1.0 that are composed of the same physical principles and distinct LSTM-based surrogates using different training strategies were used to intercompare with HyLake v1.0. The descriptions of these models are described as follows:**" (Section 2.3.1, Lines 286-291)

P11, L275: Rather "used" than "proposed". Compared to what is the improvement of HyLake compared?

Response: Baseline used a surrogate than trained on the outputs of PBBM, which is different from HyLake v1.0. This sentence has been corrected.

Revision: "• Baseline is **a** hybrid lake model that **is** coupled **to** an LSTM-based surrogate trained on outputs of PBBM, which is **used** to intercompare the performance with HyLake v1.0." (Section 2.3.1, Lines 298-299)

P12, L278: For me it is not clear what the difference to HyLake v1.0 is. Please clarify and improve the text.

Response: TaihuScene used a BO-BLSTM-based surrogate that was trained on observations from 5 sites in the Lake Taihu eddy flux network, which is different from HyLake v1.0. The proposal of TaihuScene aims to intercompare the performance in Taihu-obs and Taihu-ERA5 experiments because the magnitude of the training datasets is larger. We think it is worth revealing that the difference between the same hybrid models used with surrogates trained on different observations. This sentence clearly describes the difference.

Revision: "• TaihuScene is another hybrid lake model that **is** coupled **to** a BO-BLSTM-based surrogate trained on observations from **all sites (MLW, BFG, DPK, PTS, and XLS) in** Lake Taihu, which is different from the HyLake v1.0. The purpose of TaihuScene is to **compare** the performance by using **a** larger **training dataset to train a surrogate model with that of using a** small dataset from HyLake v1.0." (Section 2.3.1, Lines 300-303)

P12, L282: What exactly has been intercompared? For me it is still not clear for what ERA5 has been used.

Response: LST, LE and HE that were calculated from FLake, Baseline, HyLake v1.0 and TaihuScene in all experiments were intercompared. ERA5 was used to force these models to evaluate the generalization and transferability in Lake Taihu and Chaohu. The associated revisions were given in Materials and Methodology (Section 2.3.1, Lines 304-307).

Revision: "The PBBM performed like FLake in MLW site, indicating a high reliability and accuracy (Figure S**2**). Except for PBBM, **the LST, LE and HE calculated from models** in all experiments were initially intercompared in each lake site from Lake Taihu. FLake and TaihuScene was additionally **intercompared** using the forcing datasets from ERA5 datasets **in Taihu-ERA5 experiment.**" (Section 2.3.1, Lines 304-307)

P12, L283: "almost validated" does not make any sense. Either you have validated your experiments or not. Additionally to this unclear phrasing, this sentence is a repetition of what has been said at the begin of the paragraph.

Response: This sentence has been deleted.

P12, L284: The specification of what can be found in Table 2?

Response: Corrected.

Revision: "The specification **of the datasets used, surrogate, and the descriptions for each model** can be found in Table 2." (Section 2.3.1, Line 307)

"**Table 2. Specification of each model for intercomparison.**

Model	Forcing datasets	Surrogate	Training datasets	Description
PBBM	\	\	\	Backbone for HyLake v1.0
FLake	ERA5; observations	\	\	A process-based freshwater lake model for intercomparison
Baseline	MLW	LSTM	PBBM outputs	A baseline experiment using PBBM outputs for model intercomparison
TaihuScene	ERA5; observations	BO-BLSTM	All observations	A numerical experiment using large train dataset to train surrogate
HyLake v1.0	ERA5; observations	BO-BLSTM	MLW observations	Proposed hybrid lake model in this study

" (Table 2)

P12,Table 2: It is still not clear if these are different "models" or "model experiments" since throughout the manuscript different wording is used and what exactly has been done has not properly been explained.

Response: It has been corrected. Models included PBBL, FLake, Baseline, TaihuScene, and HyLake v1.0. Experiments covered different models using different forcing datasets. For example, in the MLW experiment, models used forcing meteorological variables from MLW observations; in the Taihu-obs experiment, models used forcing meteorological variables from observations for each lake site; in the Taihu-ERA5 and Chaohu experiments, models used ERA5 forcing datasets. This information was summarized in Table 2 and 3.

Revision: "**Table 2. Specification of each model for intercomparison.**

Model	Forcing datasets	Surrogate	Training datasets	Description
PBBM	\	\	\	Backbone for HyLake v1.0
FLake	ERA5; observations	\	\	A process-based freshwater lake model for intercomparison
Baseline	MLW	LSTM	PBBM outputs	A baseline experiment using PBBM outputs for model intercomparison
TaihuScene	ERA5; observations	BO-BLSTM	All observations	A numerical experiment using large train dataset to train surrogate
HyLake v1.0	ERA5; observations	BO-BLSTM	MLW observations	Proposed hybrid lake model in this study

" (Table 2)

"**Table 3: Intercomparison of model performance across different experiments conducted in diverse regions with different forcing datasets. Observations from all lake sites (MLW, DPK, BFG, XLS, and PTS) on Lake Taihu, were used to drive models in the Taihu-obs experiment. Bold values in the table present the best-performing model with each group of experiments. Computational efficiency is reported as the runtime for a single simulation.**

Exp	Model	Forcing	R			RMSE			MAE			Efficiency
			LST	LE	HE	LST	LE	HE	LST	LE	HE	(s)
MLW	PBBM	MLW	0.98	0.85	0.89	1.78	38.34	9.37	1.38	23.54	6.22	189.49
	FLake	MLW	0.98	0.82	0.84	1.76	42.73	7.24	1.35	24.76	5.01	**16.40**
	Baseline	MLW	0.96	0.74	0.75	2.71	51.77	14.63	2.11	33.52	9.30	151.46
	HyLake v1.0	MLW	**0.99**	**0.94**	**0.93**	**1.08**	**24.65**	**7.15**	**0.85**	**15.18**	**4.73**	270.21

Taihu-obs	FLake	All sites	0.97	0.61	0.74	2.24	15.46	69.11	1.69	41.95	10.44	**89.00**
	TaihuScene	All sites	**0.99**	**0.82**	0.89	1.52	14.93	43.49	1.23	29.53	10.63	6928.44
	HyLake v1.0	All sites	**0.99**	0.81	**0.90**	**1.36**	**11.19**	**39.20**	**1.03**	**24.79**	**7.88**	2693.23
Taihu-ERA5	FLake	ERA5	0.98	0.63	0.69	1.82	12.31	67.24	1.46	50.94	9.68	**19.60**
	TaihuScene	ERA5	**0.99**	0.68	0.73	1.60	13.00	64.83	1.29	47.78	10.11	652.25
	HyLake v1.0	ERA5	**0.99**	**0.71**	**0.78**	**1.12**	**11.05**	**49.48**	**0.90**	**35.02**	**7.97**	236.78
Chaohu	FLake	ERA5	**0.97**	\	\	2.28	\	\	1.76	\	\	**70.40**
	HyLake v1.0	ERA5	**0.97**	\	\	**2.07**	\	\	**1.57**	\	\	972.83

" (Table 3)

P12, L286-287: "validation" or "evaluation"? Only one of the terms should be used. Since you assess the quality of models (or model experiments) "evaluate" would be the correct term.

Response: We agreed that there should use "evaluation". It has been corrected.

Revision: "**2.3.2 Metrics for model evaluation and intercomparison**" (Section 2.3.2, Line 316)

P12, L290; It should rather read "assess if the models over or underestimate the observations"

Response: Corrected.

Revision: "Specifically, the R is proposed to measure the linear correlation of the observed and modeled values, RMSE and MAE **assess if the models over or underestimate the observations** with the same data units (Piccolroaz et al., 2024)." (Section 2.3.2, Lines 319-321)

P12, L292-294: If you calculate the difference between model and observations it should also read in the equations for RMSE and ME x_i-y_i. Please check and correct.

Response: Corrected.

Revision: "The calculation of R, RMSE, and MAE can be expressed by:

$$R = \frac{\sum(x_i-\bar{x}_i)(y_i-\bar{y}_i)}{\sqrt{\sum(x_i-\bar{x}_i)^2\sum(y_i-\bar{y}_i)^2}} \quad (17)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y}_i)^2} \quad (18)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i-\bar{y}_i| \quad (19)$$

where $\boldsymbol{x_i}$ and $\boldsymbol{\bar{x}_i}$ are the observations and its average; while $\boldsymbol{y_i}$ and $\boldsymbol{\bar{y}_i}$ are the results of model and its average; n represents the length of time series.

" (Section 2.3.2, Lines 321-326)

P13, L305-307: Improve sentence and make clear if these are different models or model experiments.

Response: In Section 3.1 of Results (Lines 335-336), we first compared the validation results of surrogates in Baseline, TaihuScene and HyLake v1.0 because their surrogates are modules of models, which should be individually validated. We have rephrased these sentences and reorganized the number of figures to make it clear.

Revision: "**This study separately validated Baseline, TaihuScene, HyLake v1.0 and their adapted LSTM-based surrogates using MLW observations to address the performance of integrated models** (Figure 4**, 5** and Figure S**3**)." (Section 3.1, Lines 335-336)

P13, L313: Here it is still not clear to what the comparisons are done. Are you comparing these to observations or to other models/model experiments?

Response: BO-BLSTM-based Surrogate in HyLake v1.0 was compared its performance to the surrogates of Baseline and TaihuScene, which are the other hybrid models, based on MLW observations in MLW experiments. These statements

have been improved (Section 3.1, Lines 336-337, Lines 341-344). Furthermore, the subtitle of Section 3.1 (Line 328) has been improved.

Revision: "3.1 **Validation** of HyLake v1.0 **in MLW experiment**" (Section 3.1, Line 328)

"Firstly, the results from HyLake v1.0 and its BO-BLSTM-based surrogate was individually validated based on MLW observations." (Section 3.1, Lines 336-337)

"Specifically, the **ΔLST** results for the BO-BLSTM-based surrogate showed RMSE values of 0.1945 °C and MAE of 0.1306 °C in the **training dataset,** RMSE of 0.3359 °C and MAE of 0.192**5** °C in the validation **dataset**, and RMSE of 0.2271 °C and MAE of 0.1461 °C in the test **dataset**, respectively." (Section 3.1, Lines 341-344)

P13, L324: What is meant with "feasible" and "reasonable and robust way"? This terminology does not make any sense here and the text should be rewritten.

Response: This sentence has been deleted.

P13, L328-329: Which are the surrogates? I still cannot follow. Here it sounds like that Flake and HyLake have been integrated to Baseline and HyLake 1.0? However, aren't Baseline and HyLake 1.0 model experiments?

Response: We have rephrased this sentence. Surrogates are individual modules of lake models. For example, an LSTM-based surrogate was coupled to Baseline, a BO-BLSTM-based surrogate was coupled to HyLake v1.0, while another BO-BLSTM-based surrogate was coupled to TaihuScene. These surrogates were only used to predict ΔLST. The backbone of hybrid models was used to process the outputs of surrogates from ΔLST to LST, LE, and HE. Therefore, Baseline, TaihuScene, and HyLake can predict LST, LE, and HE, while their surrogates can only predict ΔLST, which has been validated in the above paragraphs. FLake provided simulations of LST, LE, and HE, which are used to intercompare with outputs of hybrid models. Baseline and HyLake 1.0 are models in MLW experiment in this section.

Revision: "**After validating the accuracy of all** LSTM-based surrogates in Baseline, TaihuScene and HyLake v1.0, this study **conducted MLW experiments to predict** the LST, LE and HE **by using Baseline and HyLake v1.0 that integrated these surrogates**, then compared w**ith the outputs of traditional process-based FLake model using MLW observations** (Figure **5** and **Table 3**)." (Section 3.1, Lines 357-359)

"**Table 3: Intercomparison of model performance across different experiments conducted in diverse regions with different forcing datasets. Observations from all lake sites (MLW, DPK, BFG, XLS, and PTS) on Lake Taihu, were used to drive models in the Taihu-obs experiment. Bold values in the table present the best-performing model with each group of experiments. Computational efficiency is reported as the runtime for a single simulation.**

Exp	Model	Forcing	R			RMSE			MAE			Efficiency
			LST	LE	HE	LST	LE	HE	LST	LE	HE	(s)
MLW	PBBM	MLW	0.98	0.85	0.89	1.78	38.34	9.37	1.38	23.54	6.22	189.49
	FLake	MLW	0.98	0.82	0.84	1.76	42.73	7.24	1.35	24.76	5.01	**16.40**
	Baseline	MLW	0.96	0.74	0.75	2.71	51.77	14.63	2.11	33.52	9.30	151.46
	HyLake v1.0	MLW	**0.99**	**0.94**	**0.93**	**1.08**	**24.65**	**7.15**	**0.85**	**15.18**	**4.73**	270.21
Taihu-obs	FLake	All sites	0.97	0.61	0.74	2.24	15.46	69.11	1.69	41.95	10.44	**89.00**
	TaihuScene	All sites	**0.99**	**0.82**	0.89	1.52	14.93	43.49	1.23	29.53	10.63	6928.44
	HyLake v1.0	All sites	**0.99**	0.81	**0.90**	**1.36**	**11.19**	**39.20**	**1.03**	**24.79**	**7.88**	2693.23
Taihu-ERA5	FLake	ERA5	0.98	0.63	0.69	1.82	12.31	67.24	1.46	50.94	9.68	**19.60**
	TaihuScene	ERA5	**0.99**	0.68	0.73	1.60	13.00	64.83	1.29	47.78	10.11	652.25
	HyLake v1.0	ERA5	**0.99**	**0.71**	**0.78**	**1.12**	**11.05**	**49.48**	**0.90**	**35.02**	**7.97**	236.78
Chaohu	FLake	ERA5	**0.97**	\	\	2.28	\	\	1.76	\	\	**70.40**
	HyLake v1.0	ERA5	**0.97**	\	\	**2.07**	\	\	**1.57**	\	\	972.83

" (Table 3)

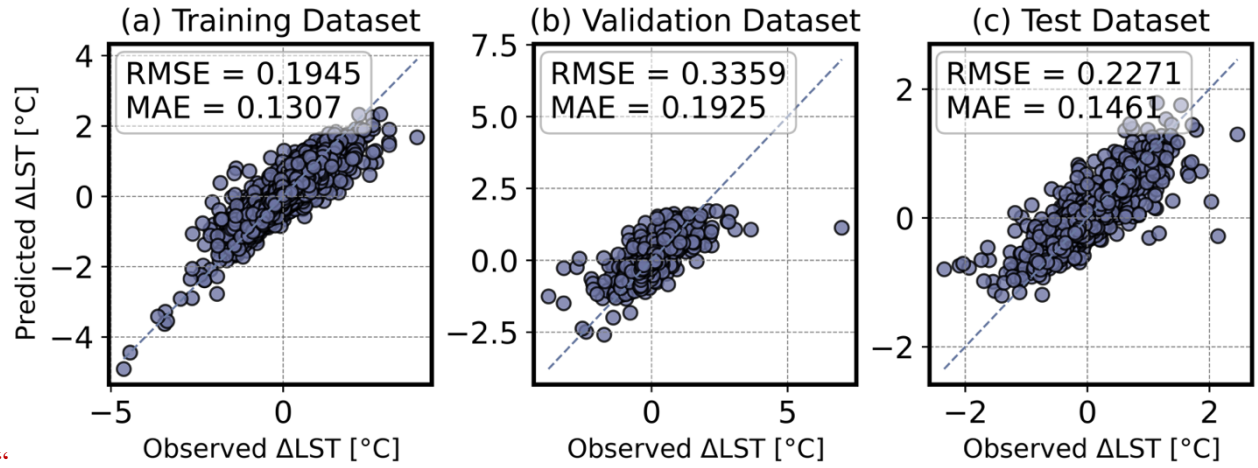
P14, L330-341: It still did not become clear to me to which data set the model has been compared.

Response: Sorry for the missing. In this section, we compared the outputs (LST, LE and HE) of all models (FLake, Baseline, and HyLake v1.0) to MLW observations from the Lake Taihu eddy flux network.

P14, Figure 4: I would suggest to make two figures instead of one figure.

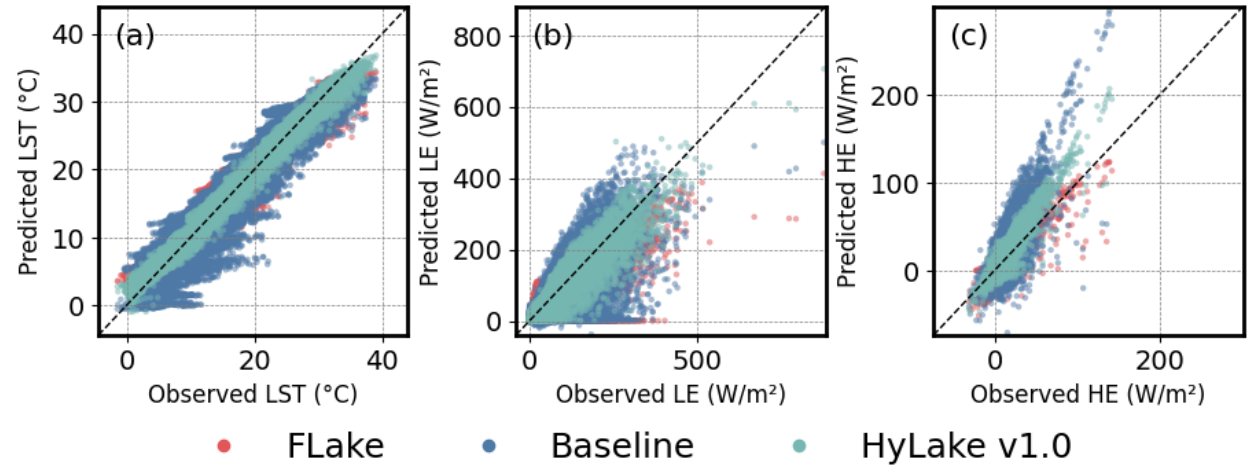
Response: Corrected. The captions of these figures were reorganized.

Revision:



“

Figure 4: The validation of BO-BLSTM-based surrogate in HyLake v1.0 for (a) training, (b) validation and (c) test datasets.” (Figure 4)



“

Figure 5: Comparison of predicted (a) LST, (b) LE and (c) HE by using FLake (red points), Baseline (blue points), HyLake v1.0 (green points) and observations in MLW experiments.” (Figure 5)

P15, 349: Which experiments? Model experiments? Are Flake, Baselin and HyLakev1.0 model experiments or models?

Response: Corrected. Here, we used models instead of experiments. Models included FLake, Baseline, and HyLake v1.0 were included in MLW experiments that were forced by meteorological variables derived from the MLW site.

Revision: “This study conducted a comprehensive intercomparison of daily and hourly trends in LST, LE and HE from **MLW experiment** in the MLW site during the period from 2013 to 2015, including FLake, Baseline, and HyLake v1.0 (Figures **6-8**).” (Section 3.2, Lines 382-383)

P15, L351: Are you referring here to one period or several periods? Which period exactly is considered?

Response: These periods were mentioned in Materials and Methodology (Section 2.2.3, Lines 280-283) that divided by the training, validation and test datasets. Here we highlighted it again in Results (Section 3.2, Lines 384-386).

Revision: “Training, validation, and test datasets for each lake site were divided by 80%, 10% and 10% **of the length**

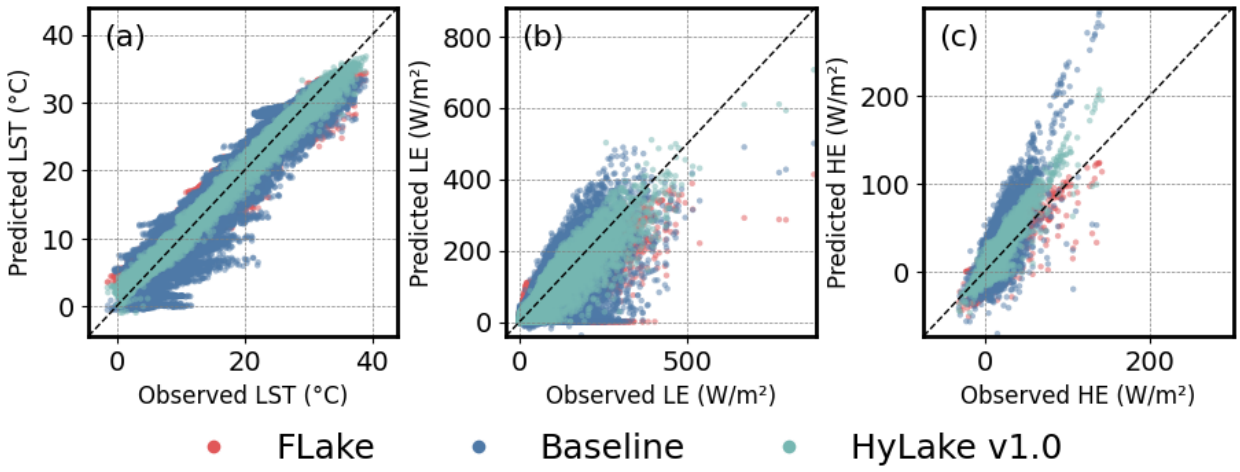
of time series (2013-2015), respectively. They are divided into 2013-01-01 00:00:00 to 2015-05-26 04:00:00, 2015-05-26 04:00:00 to 2015-09-12 14:00:00, and 2015-09-12 14:00:00 to 2015-12-30 23:00:00." (Section 2.2.3, Lines 280-283)

"As shown in Figure **6**, the temporal changes in LST for the period of surrogates training **(2013-01-01 00:00:00 to 2015-05-26 04:00:00)**, validation **(2015-05-26 04:00:00 to 2015-09-12 14:00:00)**, and test datasets **(2015-09-12 14:00:00 to 2015-12-30 23:00:00)** were compared." (Section 3.2, Lines 384-386)

P15, L356: What is meant with daily scale? On a daily basis or the daily cycle? What exactly is hwon in Figure 4 needs to be better explained.

Response: Daily scale means resampling to an average of 24 hours of outputs every day. Here, it has been corrected to "daily-average scale" to help readers understand. Figure 4 has been reorganized to Figure 5; the results in the legends before have been moved to Table 3.

Revision: "Specifically, FLake provided a good match to observations at a **daily-average scale**, which, however, showed poorer performance in capturing diurnal variations of LST (R = 0.98, RMSE = 1.76 °C, Figure 5a)." (Section 3.2, Lines 390-391)



"

Figure 5: Comparison of predicted (a) LST, (b) LE and (c) HE by using FLake (red points), Baseline (blue points), HyLake v1.0 (green points) and observations in MLW experiments." (Figure 5)

"Table 3: Intercomparison of model performance across different experiments conducted in diverse regions with different forcing datasets. Observations from all lake sites (MLW, DPK, BFG, XLS, and PTS) on Lake Taihu, were used to drive models in the Taihu-obs experiment. Bold values in the table present the best-performing model with each group of experiments. Computational efficiency is reported as the runtime for a single simulation.

Exp	Model	Forcing	R			RMSE			MAE			Efficiency
			LST	LE	HE	LST	LE	HE	LST	LE	HE	(s)
MLW	PBBM	MLW	0.98	0.85	0.89	1.78	38.34	9.37	1.38	23.54	6.22	189.49
	FLake	MLW	0.98	0.82	0.84	1.76	42.73	7.24	1.35	24.76	5.01	**16.40**
	Baseline	MLW	0.96	0.74	0.75	2.71	51.77	14.63	2.11	33.52	9.30	151.46
	HyLake v1.0	MLW	**0.99**	**0.94**	**0.93**	**1.08**	**24.65**	**7.15**	**0.85**	**15.18**	**4.73**	270.21
Taihu-obs	FLake	All sites	0.97	0.61	0.74	2.24	15.46	69.11	1.69	41.95	10.44	**89.00**
	TaihuScene	All sites	**0.99**	**0.82**	0.89	1.52	14.93	43.49	1.23	29.53	10.63	6928.44
	HyLake v1.0	All sites	**0.99**	0.81	**0.90**	**1.36**	**11.19**	**39.20**	**1.03**	**24.79**	**7.88**	2693.23
Taihu-ERA5	FLake	ERA5	0.98	0.63	0.69	1.82	12.31	67.24	1.46	50.94	9.68	**19.60**
	TaihuScene	ERA5	**0.99**	0.68	0.73	1.60	13.00	64.83	1.29	47.78	10.11	652.25
	HyLake v1.0	ERA5	**0.99**	**0.71**	**0.78**	**1.12**	**11.05**	**49.48**	**0.90**	**35.02**	**7.97**	236.78
Chaohu	FLake	ERA5	**0.97**	\	\	2.28	\	\	1.76	\	\	**70.40**

	HyLake v1.0	ERA5	**0.97**	\	\	**2.07**	\	\	**1.57**	\	\	972.83

" (Table 3)

P15, L374: What is meant with "peak and valley values"? Please rephrase.

Response: Sorry for the confusion. The "peak and valley values" mentioned in the manuscript present higher and lower values among the near-time steps. This sentence has been corrected.

Revision: "**The LE predicted by HyLake v1.0 reproduces both the peak and trough magnitudes more closely to the MLW observations than FLake and Baseline models** (Figure **7b-c)**, indicating its overall superior capacity for describing the diurnal variations. Still, some biases persisted in the validation and test periods." (Section 3.2, Lines 407-411)

P15, L376-377: Also this sentence needs to be revised.

Response: Rephrased.

Revision: "For example, HyLake v1.0 **overestimated to the observations during** 2015-08-20 and 2015-08-23 (Figure **7c**)." (Section 3.2, Lines 410-411)

P18, L403: developed? Isnt' that a model experiment?

Response: Sorry for the mistakes. TaihuScene is a model rather than an experiment.

Revision: "To address these challenges with HyLake v1.0, this study specially developed a TaihuScene (Table 2), **another hybrid lake model** which enlarges the size of training datasets by incorporating data from **5** lake sites in Lake Taihu to train **its** BO-BLSTM-based surrogates and evaluate the potential difference **from HyLake v1.0**." (Section 3.3, Lines 440-443)

P18, L410ff: It would be much more concise if these ME and RMSE values would be listed in a table.

Response: A table was given to show the results of model intercomparison for all experiments (Table 3). The bolded numbers in the table represent the best-performing results among all the models in their respective experiments.

Revision: "**Table 3: Intercomparison of model performance across different experiments conducted in diverse regions with different forcing datasets. Observations from all lake sites (MLW, DPK, BFG, XLS, and PTS) on Lake Taihu, were used to drive models in the Taihu-obs experiment. Bold values in the table present the best-performing model with each group of experiments. Computational efficiency is reported as the runtime for a single simulation.**

Exp	**Model**	**Forcing**	**R**			**RMSE**			**MAE**			**Efficiency**
			LST	LE	HE	LST	LE	HE	LST	LE	HE	(s)
MLW	PBBM	MLW	0.98	0.85	0.89	1.78	38.34	9.37	1.38	23.54	6.22	189.49
	FLake	MLW	0.98	0.82	0.84	1.76	42.73	7.24	1.35	24.76	5.01	**16.40**
	Baseline	MLW	0.96	0.74	0.75	2.71	51.77	14.63	2.11	33.52	9.30	151.46
	HyLake v1.0	MLW	**0.99**	**0.94**	**0.93**	**1.08**	**24.65**	**7.15**	**0.85**	**15.18**	**4.73**	270.21
Taihu-obs	FLake	All sites	0.97	0.61	0.74	2.24	15.46	69.11	1.69	41.95	10.44	**89.00**
	TaihuScene	All sites	**0.99**	**0.82**	0.89	1.52	14.93	43.49	1.23	29.53	10.63	6928.44
	HyLake v1.0	All sites	**0.99**	0.81	**0.90**	**1.36**	**11.19**	**39.20**	**1.03**	**24.79**	**7.88**	2693.23
Taihu-ERA5	FLake	ERA5	0.98	0.63	0.69	1.82	12.31	67.24	1.46	50.94	9.68	**19.60**
	TaihuScene	ERA5	**0.99**	0.68	0.73	1.60	13.00	64.83	1.29	47.78	10.11	652.25
	HyLake v1.0	ERA5	**0.99**	**0.71**	**0.78**	**1.12**	**11.05**	**49.48**	**0.90**	**35.02**	**7.97**	236.78
Chaohu	FLake	ERA5	**0.97**	\	\	2.28	\	\	1.76	\	\	**70.40**
	HyLake v1.0	ERA5	**0.97**	\	\	**2.07**	\	\	**1.57**	\	\	972.83

" (Table 3)

P19, 442: The transferability and generalization is too often mentioned without explaining what actually is meant with that.

Response: Generalization and transferability are both important abilities for deep-learning-based models when applied to general or specific tasks. Successful deep artificial neural networks can exhibit a remarkably small gap between training and test performance (Zhang et al., 2021). Transferability of deep-learning-based models means the ability of models to be applied to cross-domain tasks (Long et al., 2016). Here, we found that HyLake v1.0 performed well in other lake sites where the data was not included in the training datasets of its surrogate, which demonstrates strong transferability for HyLake v1.0 in ungauged regions. Other reviewers strongly affirmed our study and emphasized the importance of generalization and transferability, and suggested that we extend the application to other lakes. According to their comments, we additionally conducted a Chaohu experiment and implemented ERA5 datasets forced HyLake v1.0 into Lake Chaohu, an eutrophic lake in the middle and lower reaches of the Yangtze River basin, then used MODIS imagery to validate the predicted LST. Results demonstrated that HyLake v1.0 in Lake Chaohu, which is an unknown region for the model, outperforms the FLake model with ~10% accuracy improvements. We reorganized this sentence to highlight the results in this paragraph.

References:

Long, M., Cao, Y., Wang, J., and Jordan, M.: Learning transferable features with deep adaptation networks. In *International conference on machine learning*. June, PMLR, 97-105, 2015.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O.: Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, *64*(3), 107-115, 2021.

Revision: "It is clear that HyLake v1.0 demonstrated outstanding capacity **to apply for** ungauged regions, surpassing traditional lake-atmosphere interaction models such as FLake in prediction accuracy for each variable**, which demonstrated a strong transferability for future applications.**" (Section 3.3, Lines 480-482)

P19, Figure 8 caption: Which overall datasets? What is meant with each variable? HE, LE and LST. If yes, then please clearly write this.

Response: We deleted the wrong expression and corrected captions of Figure 8 (now is Figure 9).

Revision:

"

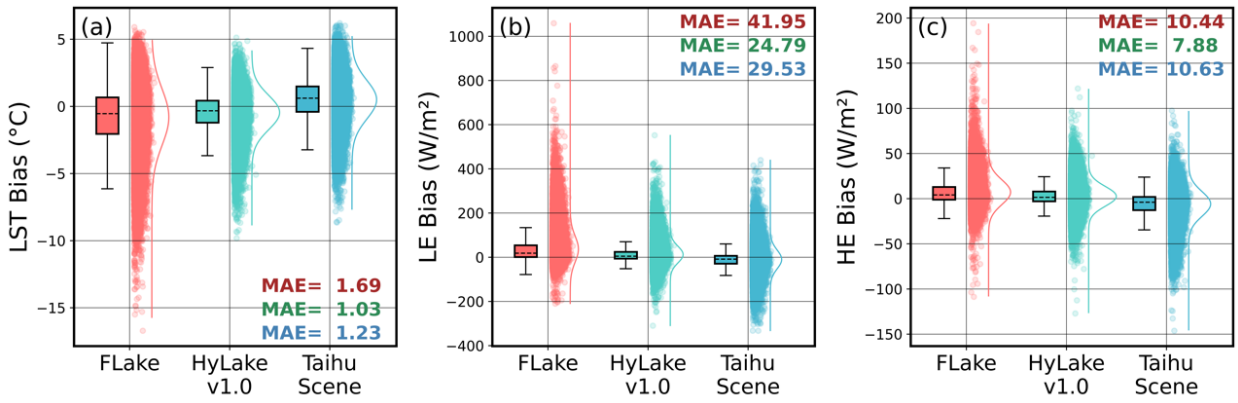


Figure 9: Comparisons of (a) LST, (b) LE, and (c) HE between observations, FLake, HyLake v1.0 and TaihuScene in five sites (MLW, BFG, DPK, PTS, and XLS) of Lake Taihu based on the Taihu-obs experiment. Dashed lines in boxplot represent median biases between observations and predictions simulated by FLake, HyLake v1.0, and TaihuScene, respectively. The scatterplots and probability distribution curves illustrate the data distribution of LST, LE and HE. The Numbers at the top or bottom right of subfigures with same color to boxes indicate the MAE of outputs for FLake, HyLake v1.0, and TaihuScene, respectively." (Figure 9)

P19, Figure 8 caption: There seem to be some repetitions. Please check and improve the figure caption.

Response: It has been corrected (see our response and revision to the previous comment).

P19, L455: This is mentioned to often. Please avoid to many repetitions.

Response: The repetitions has deleted. These sentences have been reorganized to clearly describe what we have done in Section 3.4 of Results (Lines 492-494).

Revision: "**3.4 Performance comparison of models in Lake Taihu based on ERA5 datasets**

This study additionally conducted Taihu-ERA5 experiment to demonstrate transferability of HyLake v1.0**, which proves its** superior capability **to apply for the ungauged locations based on different forcing datasets.**" (Section 3.4, Lines 492-494)

P19, L459: using ERA5 for what? As forcing data set? If yes, what has then been used in the other results presented?

Response: ERA5 datasets are used as forcing datasets to force models in the Taihu-ERA5 experiment. ERA5 datasets have always been merely forced datasets.

Revision: "**The meteorological variables from ERA5 dataset,** which are widely used as forcing datasets for process-based models (Albergel et al., 2018; Hersbach et al., 2020)**, were selected to force FLake, TaihuScene and HyLake v1.0 and then compared their performance on LST, LE and HE observations from the Lake Taihu Eddy Flux Network**." (Section 3.4, Lines 494-497)

P20, L479: Repetition -> sentence obsolete.

Response: It is an important conclusion in Section 3.4, which should not be obsoleted. We have rephrased this sentence.

Revision: "**HyLake v1.0 in Taihu-ERA5 experiments exhibited superior performance for each lake site, showing a strong transferability using ERA5 datasets** (Figure S5)." (Section 3.4, Lines 519-520)

P20, L486: Same as for L479.

Response: Rephrased.

Revision: "HyLake v1.0 still performed considerable **well in ungauged sites** by learning physical principles from **MLW observations** (Figure S4d-o)." (Section 3.4, Lines 525-526)

P21, L502: Same here.

Response: Corrected.

Revision: "Overall, both HyLake v1.0 and TaihuScene showed reliable **performance** across lake sites in Lake Taihu. Specifically, HyLake v1.0 performed the best in 14 of 15 variables **(included LST, LE and HE for 5 lake sites)** in Lake Taihu among these 3 **models**; …" (Section 3.4, Lines 540-541)

P22, L528: Which are the different forcing data sets? Only ERA5 is mentioned.

Response: The Lake Taihu eddy flux network also provided forcing datasets. In our experiments, the MLW experiment utilized MLW observations from the network to force FLake, Baseline, and HyLake v1.0, whereas the Taihu-ERA5 and Chaohu experiments used ERA5 datasets to force FLake, TaihuScene, and HyLake v1.0. This sentence has been rephrased to highlight our contributions.

Revision: "These experiments were compared using **observed meteorological datasets and ERA5 datasets, then** validated for both spatial and temporal patterns at Lake Taihu and **Lake Chaohu (Tables 2-3).**" (Section 4.1, Lines 564-566)

P24, 578: Not clear, before it was always stated that HyLake v1.0 outperformed the other models. Now, here the opposite is stated.

Response: Sorry for the confusion. This sentence is to express that the explainability of HyLake v1.0 is weaker than process-based models due to their deep-learning-based surrogates. This sentence has been corrected.

Revision: "While **the interpretability of** HyLake v1.0 is better than that of purely data-driven models due to its hard-coupling structure**, which retains the energy balance equations and utilizes a BO-BLSTM-based surrogate to solve**

LST, it still lags behind process-based models." (Section 4.1, Lines 635-637)

P25, L624: Are these forcing data sets observations or model simulations?

Response: Corrected.

Revision: "Additionally, this study used **different** forcing datasets, **including observations from** 5 lake sites **in** Lake Taihu and **the** ERA5 datasets, to **evaluate** the and transferability of HyLake v1.0 in ungauged regions **and unlearned datasets.**" (Conclusion, Lines 704-706)

P26, L640: Which 15 variables have been used? These have nowhere been mentioned.

Response: There are 3 variables for 5 lakes sites, a total of 15 variables.

Revision: "Regarding the capability of spatial **transferability** using ERA5 forcing datasets, results indicated HyLake v1.0 performed the most closely matched the observations in Lake Taihu compared to FLake and TaihuScene in 14 of 15 variables **(LST, LE and HE in 5 lake sites)**." (Conclusion, Lines 719-721)

Technical corrections:

P1, L13: proposed -> proposes

Response: To highlight what we have done in this study, we have reorganized this sentences.

Revision: "This study **presents** the Hybrid Lake Model v1.0 (HyLake v1.0), which integrates a Bayesian Optimized Bidirectional Long Short-Term Memory-based (BO-BLSTM-based) surrogate trained **on data** from Meiliangwan (MLW) site in Lake Taihu to approximate LST **dynamics. LE and HE are subsequently derived using** surface energy balance equations." (Abstract, Lines 15-17)

P4, L104: with -> has

Response: Corrected to "has".

P5, Figure 1 caption: valid -> validate

Response: Corrected.

Revision:

"

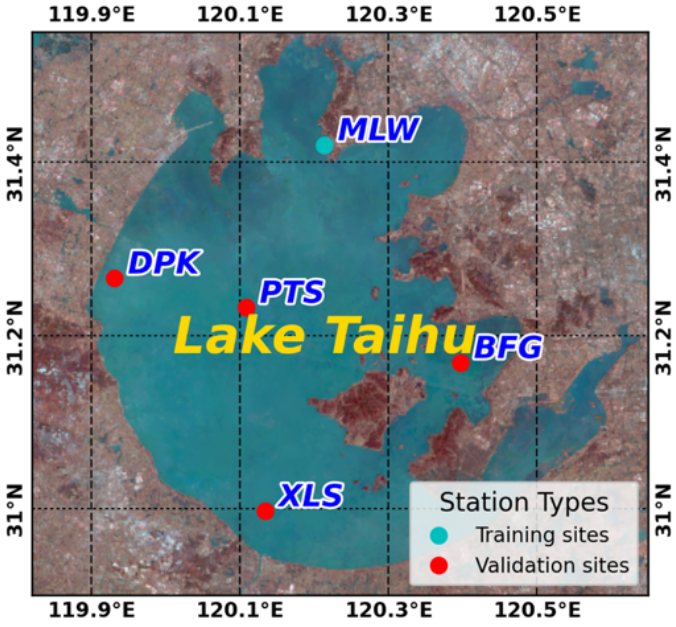


Figure 1: The locations of Lake Taihu and the five eddy covariance lake sites (MLW, DPK, BFG, XLS, and PTS) are shown in cyan and red bubbles, overlaid on a true-color image from Landsat 8. MLW as a training site was used to train BO-BLSTM-based surrogate, while the other validation sites were adapted as ungauged sites to

validate the HyLake v1.0 performance." (Figure 1)

P5, L133: couple a LSTM-based -> coupled to a LSTM-based
Response: Added an "a" in this sentence.

P5, L133: which is shown -> as schematically shown
Response: Corrected.
Revision: "HyLake v1.0 **is** constructed in this study based on the backbone of physical principles from process-based lake models and then couple **to** a LSTM-based surrogate for LST approximation to further solve the untrained variables (e.g., LE, HE), **as schematically shown** in Figure 2." (Section 2.2, Lines 146-148)

P8, L184 and L208: Equation -> Eq.
Response: Checked and corrected the full text.

P10, L233: same here as for L184 and L208.
Response: Corrected.

P11, L262: designment -> design
Response: Title was corrected to "2.3 Numerical experiments design and evaluation metric".

P12, L279: using larger train datasets against -> using a larger training dataset compared to a
Response: Corrected.
Revision: "• The purpose of TaihuScene is to **compare** the performance of using **a** larger **training dataset to train a surrogate model with that of using a** small dataset from HyLake v1.0." (Section 2.2, Lines 301-303)

P12, L287: a -> the
Response: Corrected.

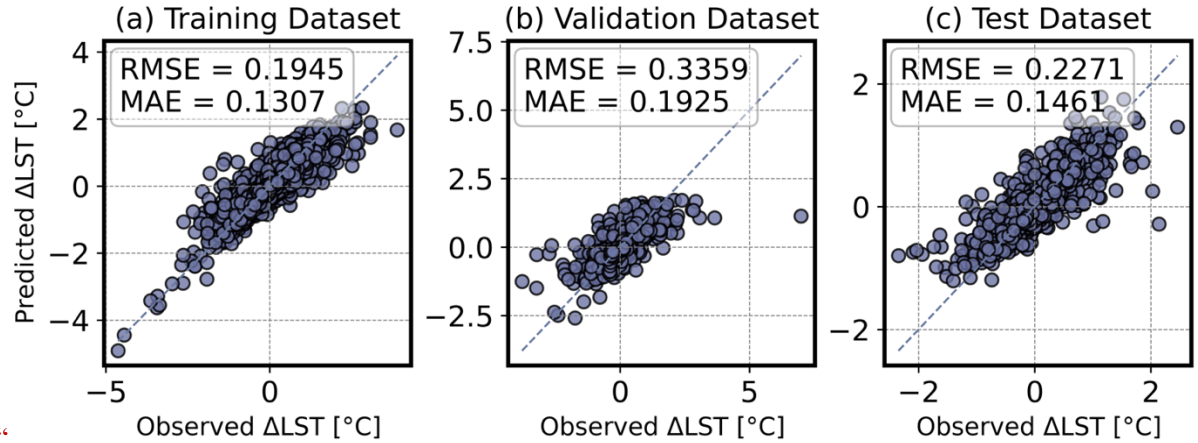
P13, L305: Delete "After that" and start sentence with "To evaluate".
Response: This sentence has been rephrased.

P13, L308: train -> training
Response: Checked and corrected all.

P13, L313: train set -> training data set
Response: Corrected to training dataset.

P13, L313: validation set -> validation data set
Response: Corrected to validation dataset.

P13, L323: relatively -> somewhat (?). Check wording and improve.
Response: This sentence has been corrected.
Revision: "These results were **somewhat** lower than HyLake v1.0 due to the larger dataset size in training for ΔLST." (Section 3.1, Lines 353-354)
P14, Figure 4 caption: train -> training
Response: The caption of Figure 4 has been corrected.
Revision:



“

Figure 4: The validation of BO-BLSTM-based surrogate in HyLake v1.0 for (a) training, (b) validation and (c) test datasets.” (Figure 4)

P18, L420: it is worthy -> it is worthy to note (?)

Response: Corrected.

Revision: “But it is worthy **to note** that TaihuScene still far outperformed FLake, …” (Section 3.3, Line 458)

P18, L404: train -> training

Response: Checked and corrected all.

P18, L429: appeared -> apparent

Response: Corrected.

P22, L527: proposed to intercompare -> proposed for intercomparison

Response: Corrected.