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"

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.

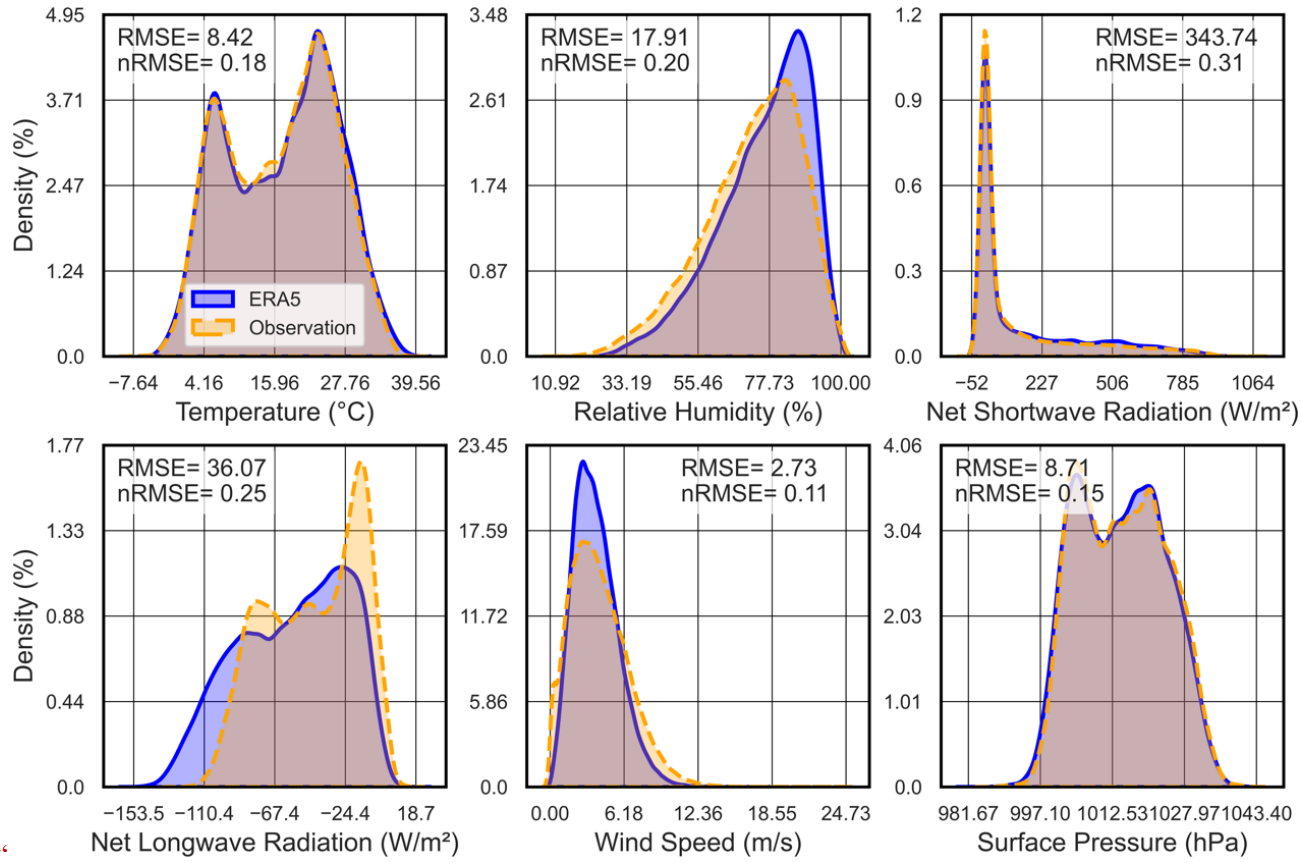
Revision: "**Within the network, e**ach site is equipped with an eddy covariance system that continuously monitors **LE and HE** using sonic anemometers and thermometers (Model CSAT3A; Campbell Scientific, Logan, UT, USA) positioned 3.5 to 9.4 m above the lake surface." (Section 2.1, Lines 116-118)

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.

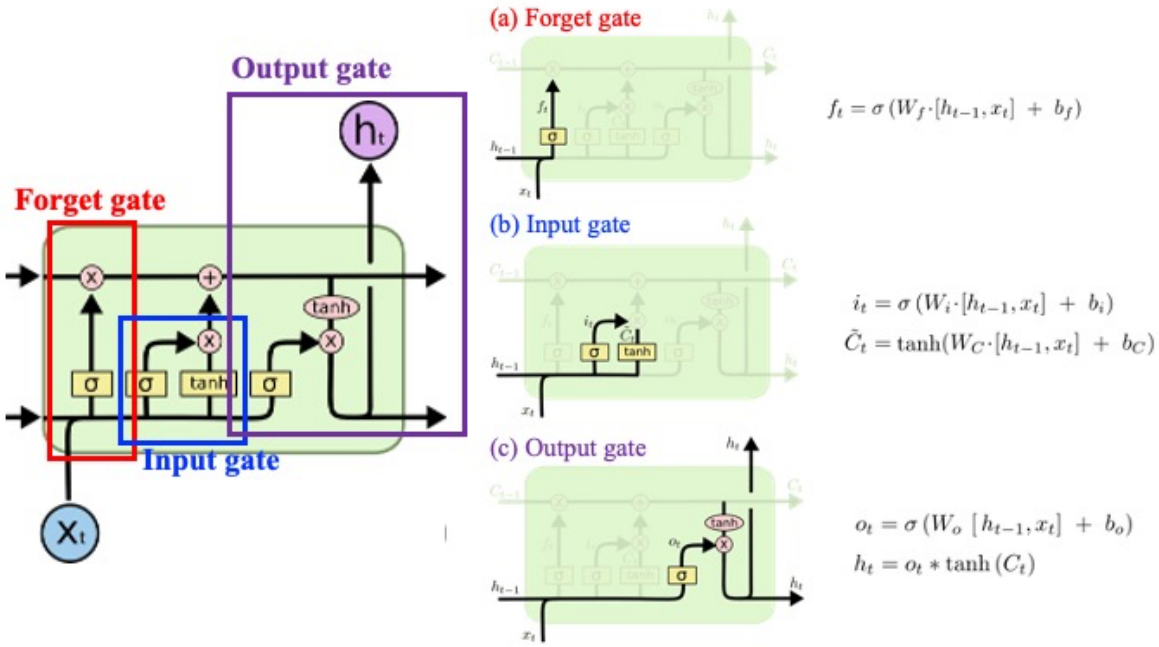


Figure R1: The architecture LSTM unit and 3 gates inside the unit.

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 $\bar{\boldsymbol{x}}_{\boldsymbol{i}}$ are the observations and its average; while $\boldsymbol{y_i}$ and $\bar{\boldsymbol{y}}_{\boldsymbol{i}}$ 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:

“

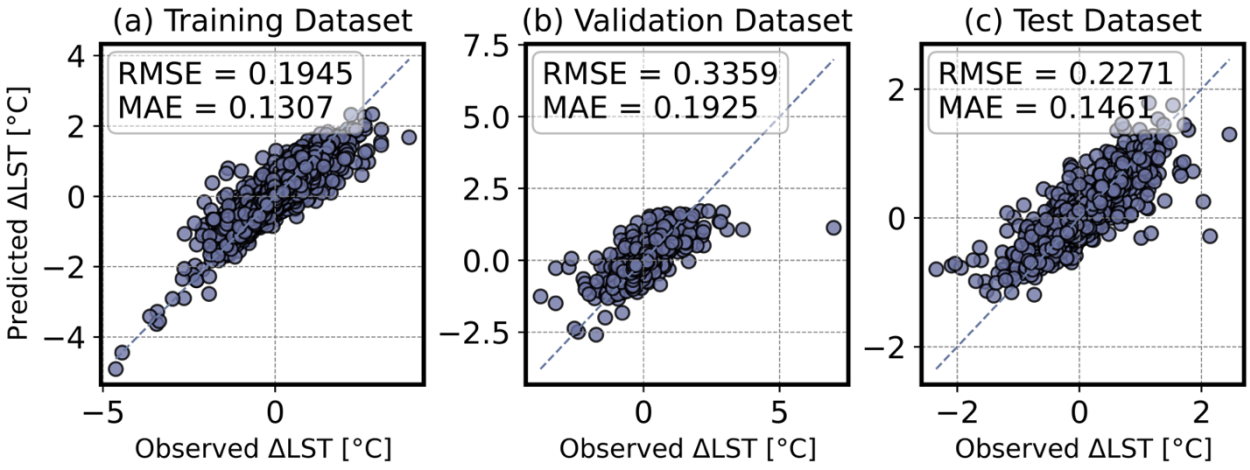


Figure 1: The validation of BO-BLSTM-based surrogate in HyLake v1.0 for (a) training, (b) validation and (c) test datasets.” (Figure 4)

“

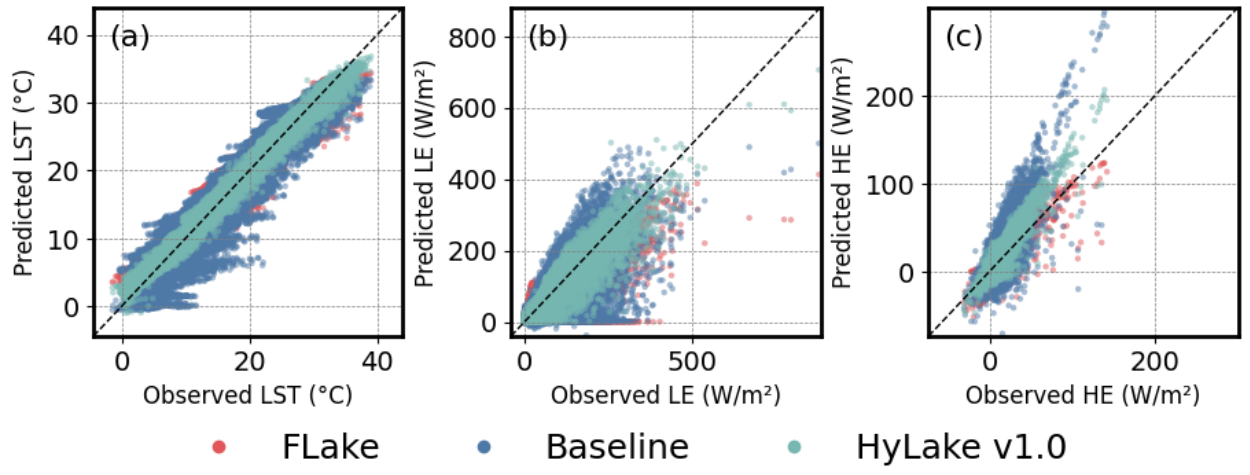


Figure 2: Comparison of predicted (a) LST, (b) LE and (c) HE by using FLake (red points), Baseline (blue points), HyLake v1.0 (green points) and observations in MLW experiments.” (Figure 5)

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).

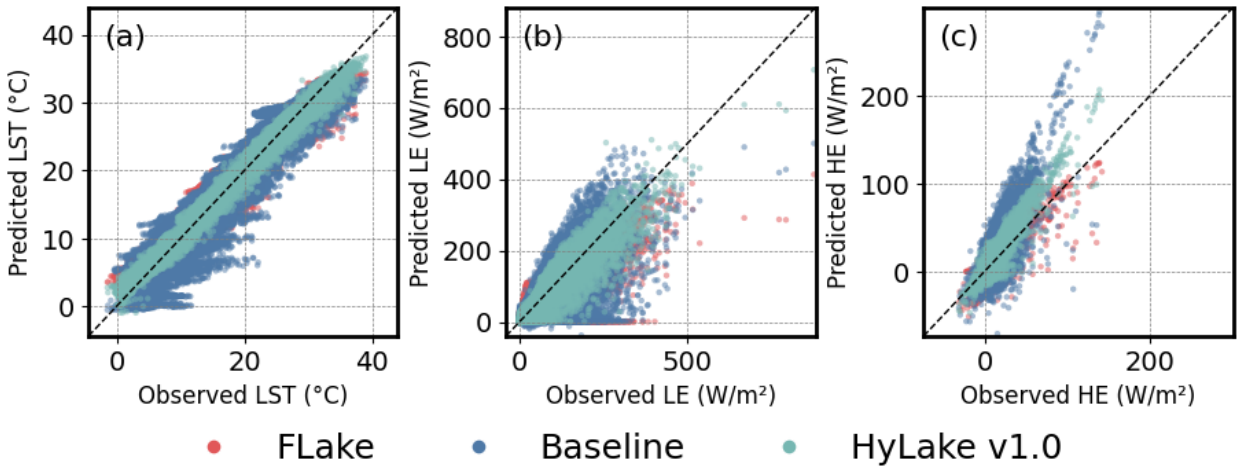
Revision: "Training, validation, and test datasets for each lake site were divided by 80%, 10% and 10% **of the length of time series (2013-2015), respectively. They are** divided into 2013-01-01 00:00:00 to 2015-05-26 04:00:00, 2015-05-26 04:00:00 to 2015-09-12 14:00:00, and 2015-09-12 14:00:00 to 2015-12-30 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 3: 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

Taihu-obs	**HyLake v1.0**	MLW	**0.99**	**0.94**	**0.93**	**1.08**	**24.65**	**7.15**	**0.85**	**15.18**	**4.73**	270.21
	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:

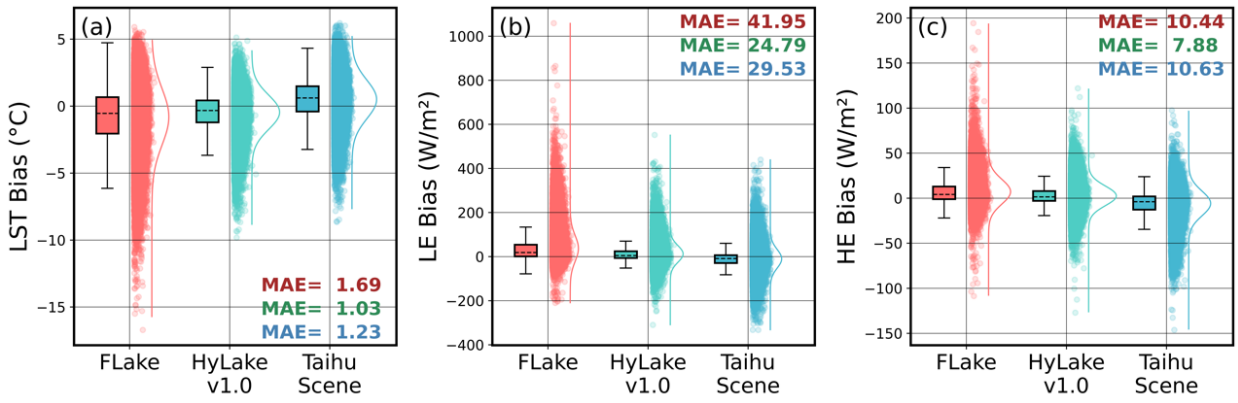


Figure 4: Comparisons of (a) LST, (b) LE, and (c) HE between observations, FLake, HyLake v1.0 and TaihuScene in five sites (MLW, BFG, DPK, PTS, and XLS) of Lake Taihu based on the Taihu-obs experiment. Dashed lines in boxplot represent median biases between observations and predictions simulated by FLake, 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)

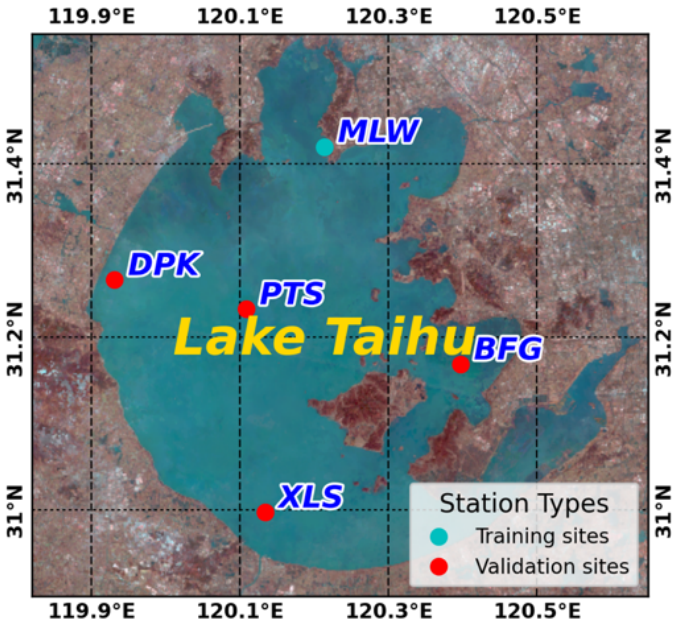
P4, L104: with -> has

Response: Corrected to "has".

P5, Figure 1 caption: valid -> validate

Response: Corrected.

Revision:



“

Figure 5: 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:

“

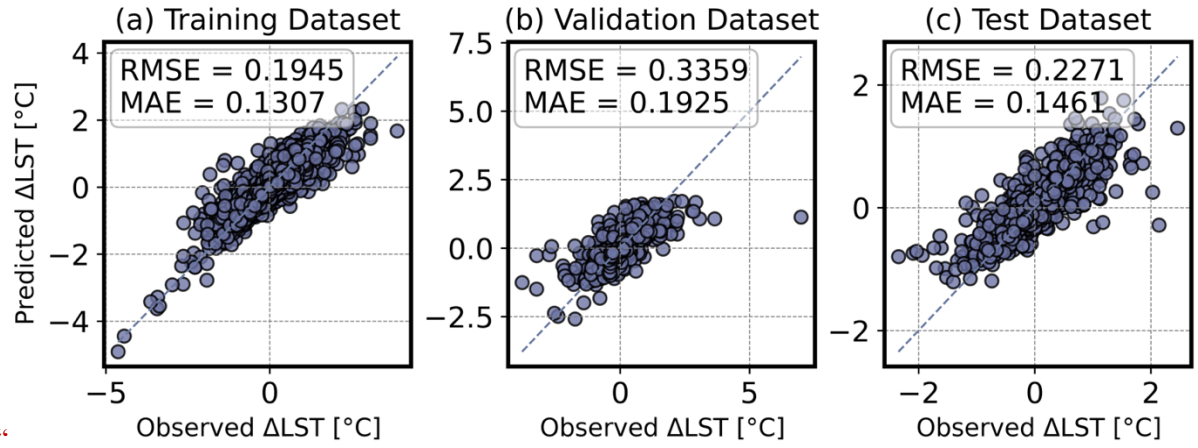


Figure 4: The validation of BO-BLSTM-based surrogate in HyLake v1.0 for (a) training, (b) validation and (c) test datasets.” (Figure 4)

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.