Supporting Information for

Prediction of algal blooms via data-driven machine learning models in a dimictic mesotrophic lake

Shuqi Lin[1], Don Pierson[1], Jorrit P. Mesman [1,2]

1. Erken Laboratory and Limnology Department, Uppsala University, Uppsala, Sweden
2. Département F.-A. Forel des sciences de l'environnement et de l'eau, Université de Genève, Genève, Switzerland

* Correspondence to Shuqi.lin@ebc.uu.se

This PDF file includes:

Text S1-3

Table S1-5

Figures S1, S2, S3, S4, S5, S6, S7, S8, S9

Text S1 Monitoring methods used at Lake Erken

A meteorological station on an island offshore from Uppsala University's Erken Laboratory provides measurements of wind speed, solar radiation, and air temperature. An automated water temperature monitoring system records water temperature profiles at a depth of 15 m with sensors placed at 0.5 m intervals. Water discharge is measured entering the lake from the largest input at Kristineholm, and the outflow at Stensta (Fig. 1). These data have been further quality controlled and combined with data from other nearby meteorological stations to provide a long-term dataset that is suitable as input for model simulations (Moras et al., 2019). Since 1991, a consistent (1-2 week) monitoring program has collected integrated water samples from the epilimnion and hypolimnion during stratified conditions or from the entire water column during isothermal conditions. Stream samples are collected from the main inflow at Kristineholm and the outflow of the lake. All samples are analyzed by the Erken Laboratory for all major nutrient concentrations (e.g., NO_X, NH_4, PO_4, Total P, Si, etc.), dissolved oxygen (O_2), and *Chl* concentration. Water and nutrient loads input to the GOTM/SELMAPROTBAS model were calculated from the discharge and nutrient concentrations measured at Kristineholm (Fig. 1) which accounted for 50.7 % of the lake watershed. Inputs from the remaining watershed were estimated from the measured Kristineholm inputs that were scaled by area to account for the remaining 49.3 % of the watershed area. Further details of meteorological and hydrological data processing can be found in Moras et al. (2019) and Mesman et al. (2022).

Text S2 Hyperparameters setting in ML models

The hyperparameters in GBR are optimized via *RandomizedSearchCV* function within Python Scikit-Learn library. The loss function of model is chosen as 'huber', which is a combination of the squared error and absolute error of regression. Since the target variable in our research *Chl* concentration has peak values during algal blooms which could be regarded as outliers, the 'huber' loss function is more robust and gives greater weight to peak values than the mean squared error function.

Essentially, the LSTM model defines a transition relationship for a hidden representation through a LSTM cell which combines the input features at each time step with the inherited information from previous time steps. There are 3 hidden LSTM layers with 100 neurons in each layer, and each of them is followed by a dropout layer with 0.01-0.03 dropout rate for regularizing the network. The numbers of batch and epoch are set as 10 and 100, respectively. Thus, the training samples are divided into 10 batches, and the internal model parameters will update after working through one batch. And the deep learning algorithm will work through the entire training dataset 100 (epochs) times. The 'MinMaxScaler' was used to pre-process the data for generalization purposes, and 'Mean Absolute Error' was used as loss function.

Text S3 Calculations of hydrodynamic features

The mixing layer depth (z_e) was computed using the GOTM simulated vertical eddy diffusivity (K_z) profiles, and was defined as the first depth, from the lake surface, where K_z fell below the predefined threshold value (Wilson et al., 2020), and can be describe as

$$z_e = z_i + (K_Z^{threshold} - K_{zi})(\frac{z_{i+1}-z_i}{K_{zi+1}-K_{zi}}),$$

where z_i and K_{zi} are the depth from the lake surface, and the eddy diffusivity, respectively, in the ith layer within the model. The threshold value $K_z^{threshold}$ was set to 5×10^{-5} m^{2}s^{-1}, based on the value described in Wüest and Lorke (2009) and Lin et al. (2021).

Unlike the dynamically varying mixing layer depth derived from the modelled K_z profiles, the calculation of the seasonal thermocline depth was estimated using Lake Analyzer (Read et al., 2011) based on the modelled temperature profile. A movement of thermocline can allow nutrient released from the sediment to enter the upper water column, leading to nutrient enrichment. It also can lead to resuspension of cells or dormant forms of cyanobacteria into the water column, encouraging bloom development (Reichwaldt and Ghadouani, 2012).

The Wedderburn number W_n, introduced by Thompson and Imberger (1980), is used to estimate the chance of upwelling occurring in the lake. It is written as

$$W_n = \frac{g' z_e^2}{u_*^2 L_s},$$

where $g' = g\frac{\Delta\rho}{\rho_h}$ is the reduced gravity due to the change in water density $\Delta\rho$ between the hypolimnion (ρ_h) and epilimnion (ρ_e). L_s is the lake fetch length (2700 m for Lake Erken) and u_* is the wind stress induced water friction velocity, defined as

$$u_* = \sqrt{\frac{\tau_w}{\rho_e}},$$

where τ_w is the wind shear (N m^{-2}) on the water surface, computed by $\tau_w = C_D \rho_{air} U^2$. U is wind speed (m s^{-1}) measured at 10 m above the water surface. C_D is drag coefficient, given as 10^{-3} for $U < 5$ m s^{-1}, and 1.5×10^{-3} for $U >= 5$ m s^{-1}.

Table S1. List of training features and target variables in each workflow. Blue indicates training features, red indicates target variables, purple indicates the variables are the target variables in step 1 used to produce daily a training feature for use in step 2. The order of nutrient model sequence is from the top to bottom based on its position in the table (NOx to Si).

variables	Sample interval	workflow 1	workflow 2		workflow 3	
			Step 1	Step 2	Step 1	Step 2
Inflow	Daily					
Meteorological data (Air temperature, wind speed, shortwave radiation, precipitation, humidity, cloud cover)	Daily					
ΔT	Daily					
Ice duration	Daily					
Days from ice-off date	Daily					
z_e	Daily					
W_n	Daily					
thermD	Daily					
NOx	1-2 weeks					
O_2	1-2 weeks					
PO_4	1-2 weeks					
Total P	1-2 weeks					
NH_4	1-2 weeks					
Si	1-2 weeks					
Chl	1-2 weeks					

Table S2. Confusion matrix and metrics based on it.

	Modeled onset	Modeled no onset
Observed onset	True Positive (TP): Model predicted the bloom onset when there was an onset	True Negative (TN): Model predicted no bloom onset when there was no onset.
Observed on onset	False Positive (FP): Model predicted the bloom onset when there was no onset	False Negative (FN): Model did not predict bloom onset when in fact there was an onset
True positive rate (TPR) = TP / (TP+FN); What proportion of all events were correctly detected		
False positive rate (FPR) = FP / (TN+FP); What proportion of no events were incorrectly defined as bloom onset		
Kappa = $(P_o-P_e)/(1-P_e)$; The modified accuracy that considers the possibility of the agreement occurring by chance. P_o = (TP + TN)/(TP+TN+FP+FN); Actual accuracy P_e = ((TP+FP)/(TP+TN+FP+FN) * (FN+TN)/(TP+TN+FP+FN)) + ((TP+FN)/(TP+TN+FP+FN) * (FP+TN)/(TP+TN+FP+FN)); Chance agreement		

Table S3 Comparisons of ML models' performance based on *RMSE*, *MAE*, and *R2* in training dataset (via 5-fold cross validation) and testing dataset.

Scenario	GBR			LSTM		
	MAE	**RMSE**	**R2**	**MAE**	**RMSE**	**R2**
1 (training)	2.86	4.30	0.18	2.66	4.38	0.31
1 (testing)	3.55	5.77	0.13	3.58	5.64	0.20
2 (training)	2.78	4.07	0.33	2.71	4.73	0.31
2 (testing)	4.22	6.27	0.05	3.87	6.00	0.13
3 (training)	2.79	4.10	0.32	2.64	4.51	0.40
3 (testing)	3.99	5.94	0.14	3.71	5.81	0.18

Table S4 Coefficient of variation of evaluating metrics in shuffling training years to test 2019-2020.

Model	***MAE* (%)**	***RMSE* (%)**	**TPR (%)**	**FPR (%)**	**Kappa (%)**
GBR	4.49	4.00	23.98	31.77	4.53
LSTM	5.80	5.21	16.36	21.41	6.30

Table S5 Coefficient of variation of *MAE*s, *RMSE*s, and TPRs in shuffling year data sparsity test.

Model	Sample interval	MAE (%)	RMSE (%)	TPR (%)
GBR	Original	13.82	12.88	31.62
	7 days	18.60	17.08	34.63
	14 days	15.17	15.12	43.94
	21 days	15.73	15.22	59.51
	28 days	18.30	20.65	77.09
	35 days	13.63	14.11	118.61
LSTM	Original	20.52	16.98	62.12
	7 days	15.71	13.05	91.63
	14 days	15.97	14.32	113.53
	21 days	19.83	13.08	107.39
	28 days	19.15	15.81	110.40
	35 days	14.44	16.12	106.99

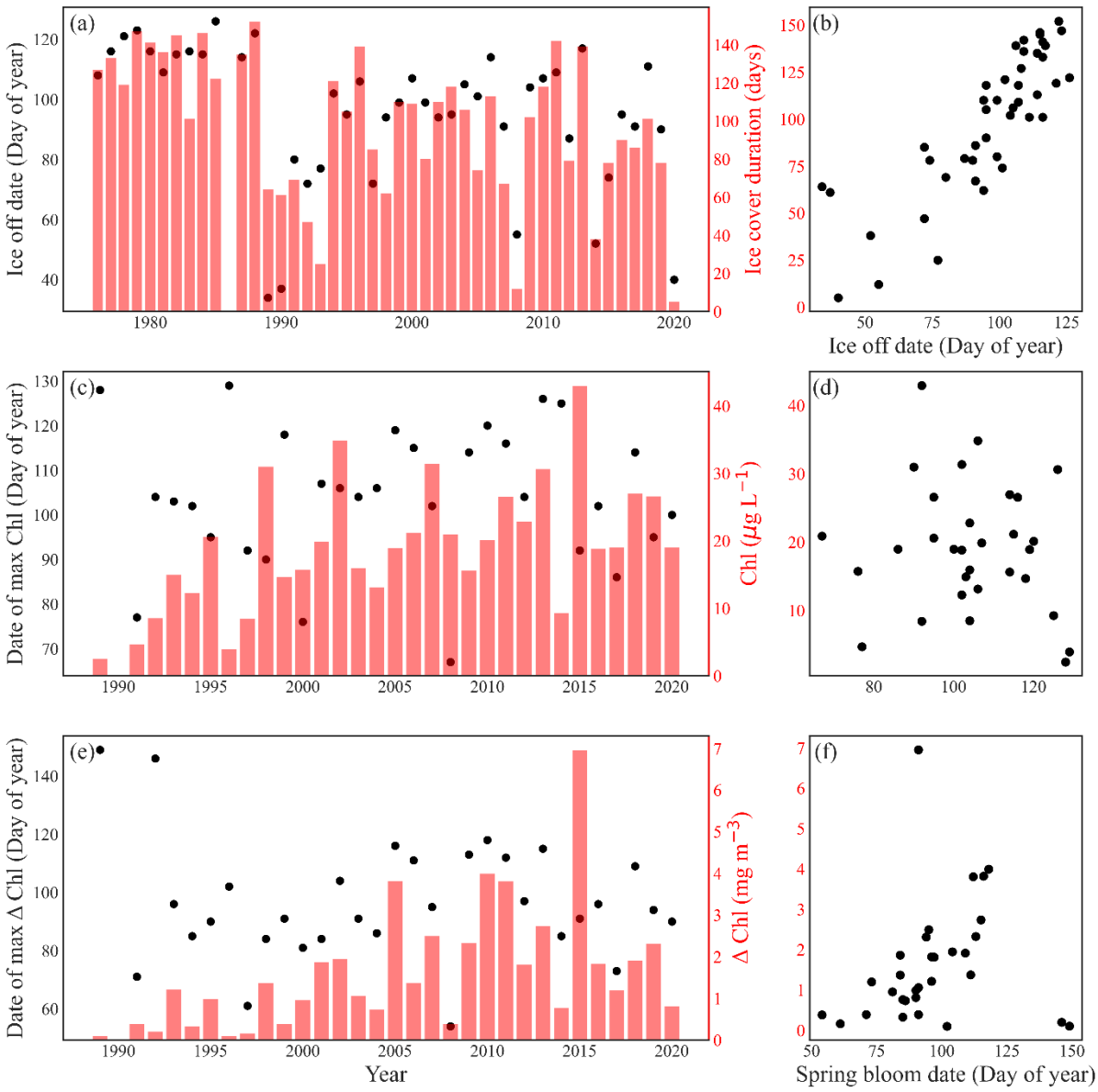

Figure S1. (a, b) Ice break-up dates and ice cover durations since 1975 (Part of data from Weyhenmeyer et al. 1999). The timing of spring bloom in Lake Erken defined by (c, d) maximum Chl peak, and (e, f) steepest daily change of Chl.

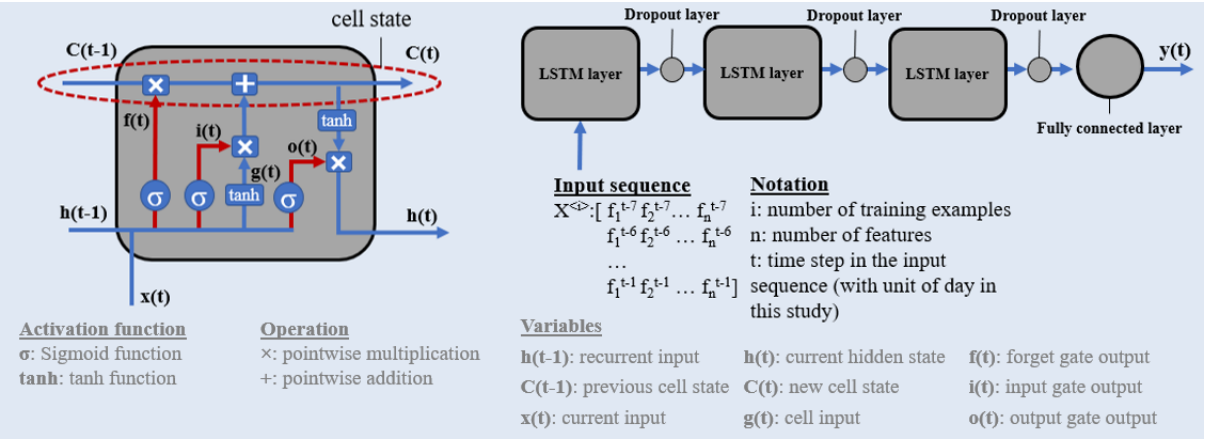

Figure S2. Left: Detail of a LSTM cell. Right: The LSTM model architecture (based on Hochreiter and Schmidhuber, 1997).

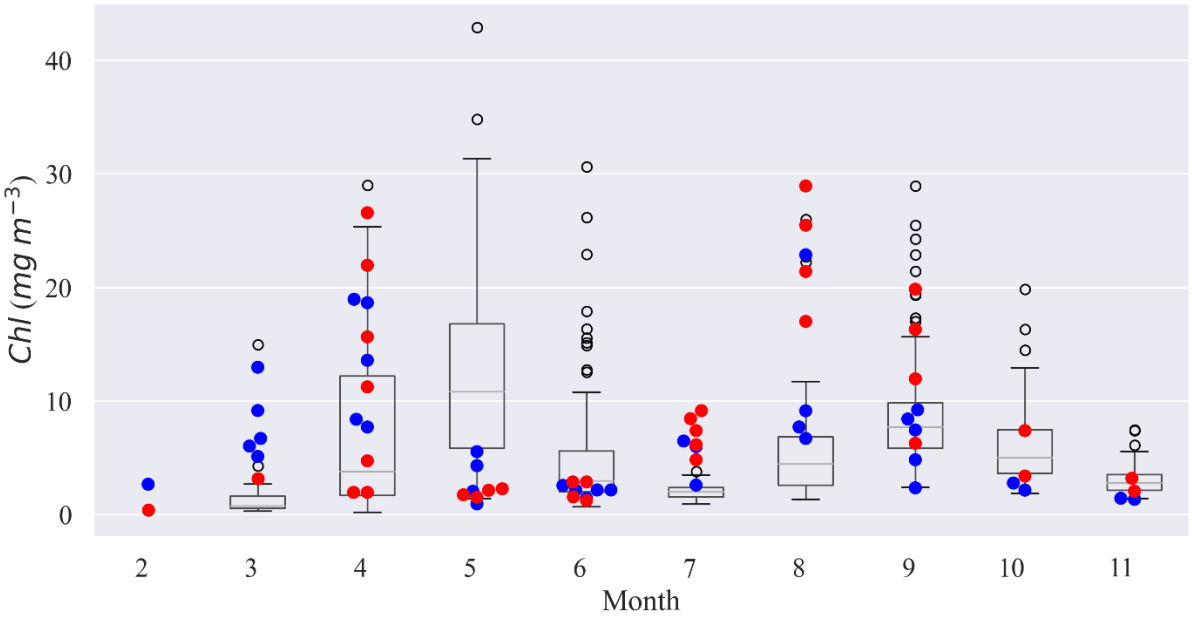

Figure S3. Comparison of *Chl* concentrations in every month over 2004-2020, the red and blue dots represent the data from 2019 and 2020, respectively.

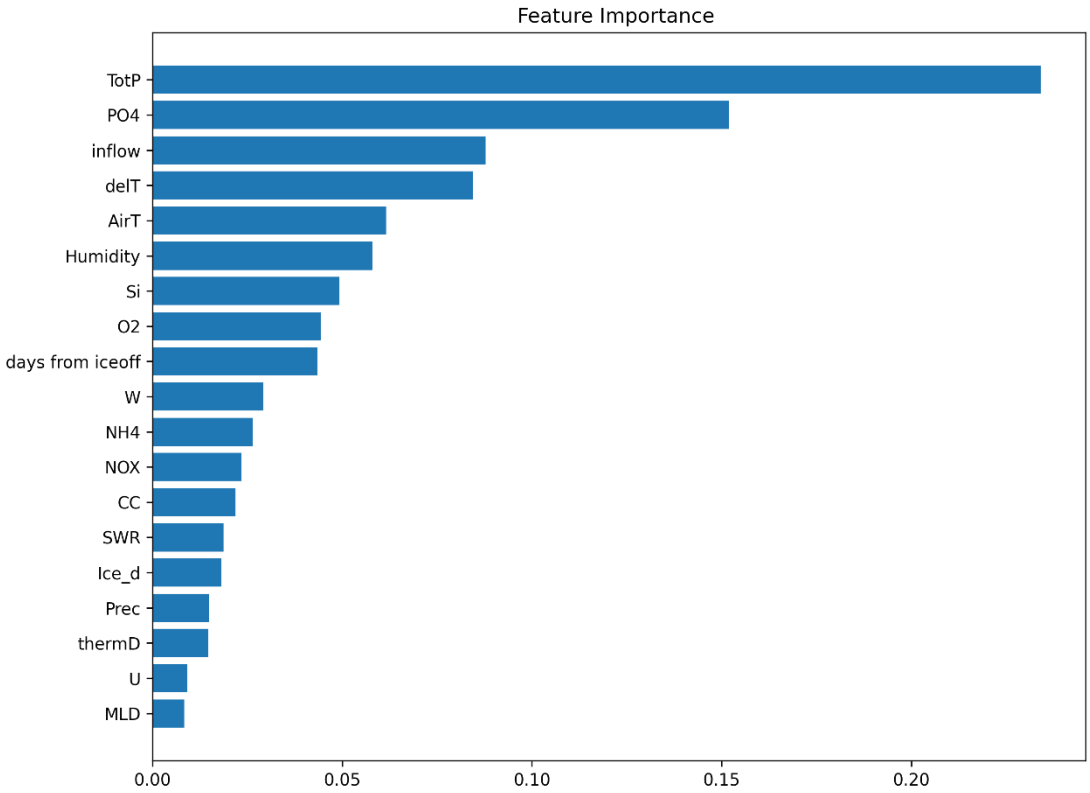

Figure S4. Feature importance based on '*feature_importances_*' function from GBR model in scenario 1.

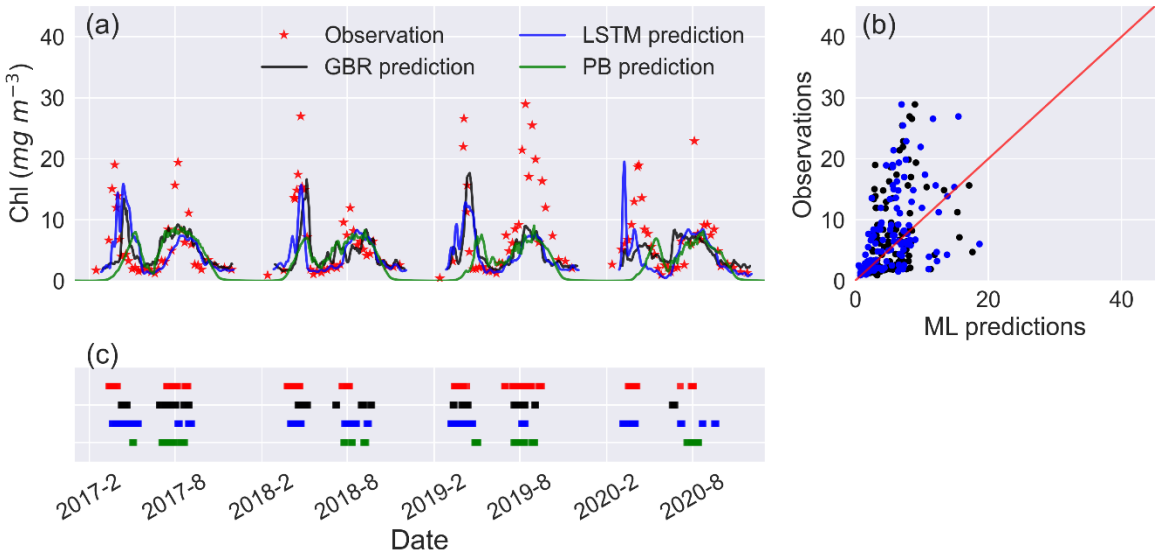

Figure S5. (a) Timeseries of observed and predicted *Chl* from GBR and LSTM models in workflow 2, (b) scatter plots of observations vs GBR and LSTM models. Penal (c) shows the observed and predicted algal bloom onsets in 2017-2020 using the same color coding as the previous panels. Results from the PB model simulation in Mesman et al. (2022) are also shown in (a) and (c).

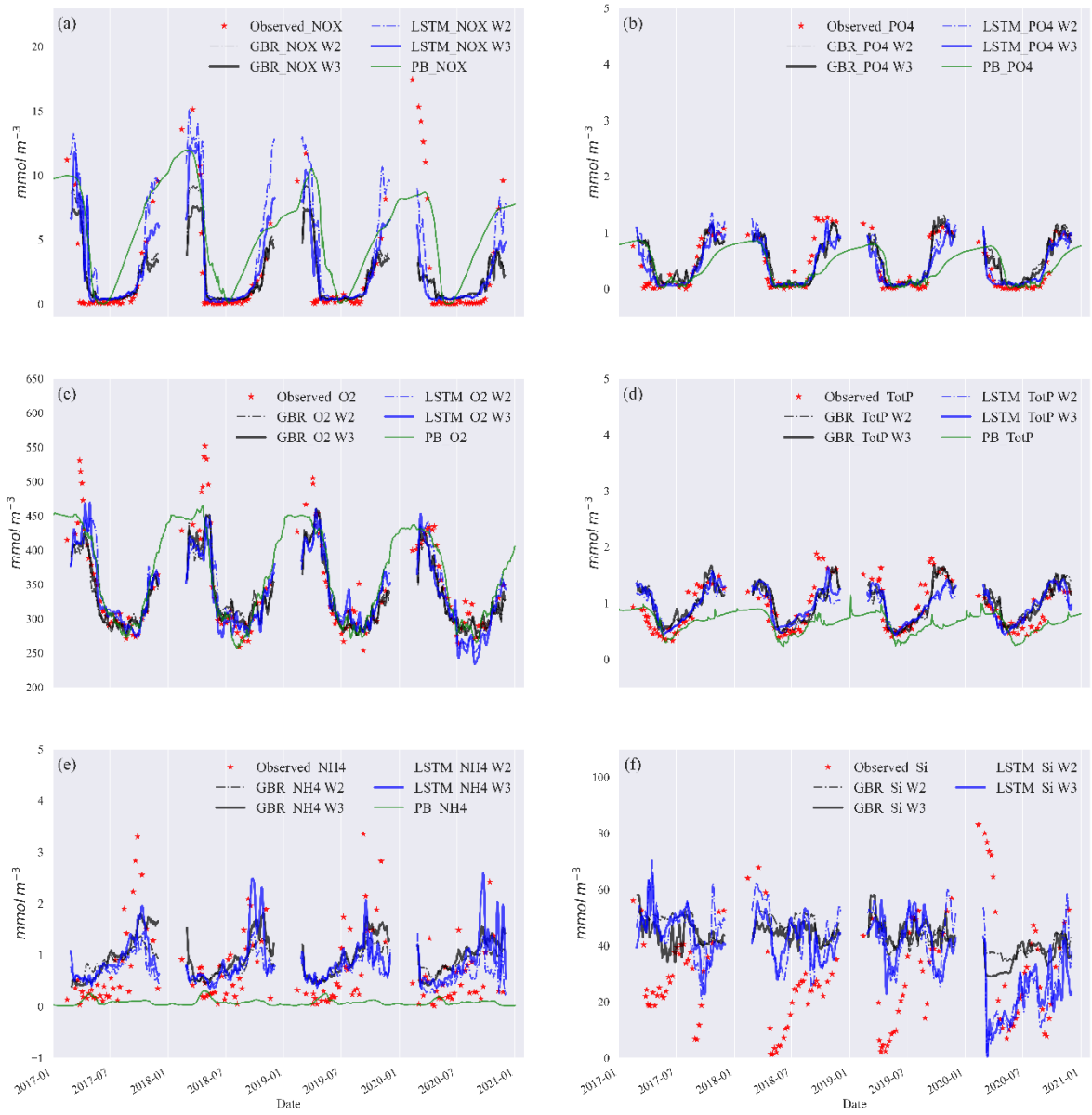

Figure S6. Timeseries of six observed and predicted nutrients (a) NOX, (b) PO4, (c) O2, (d) Total P, (e) NH4, (f) Si, at surface (-3 m) from GBR, LSTM in workflow 2 (W2) and 3 (W3), and PB models. The Si simulations in the PB model had not been optimized, so these are not shown in the figure.

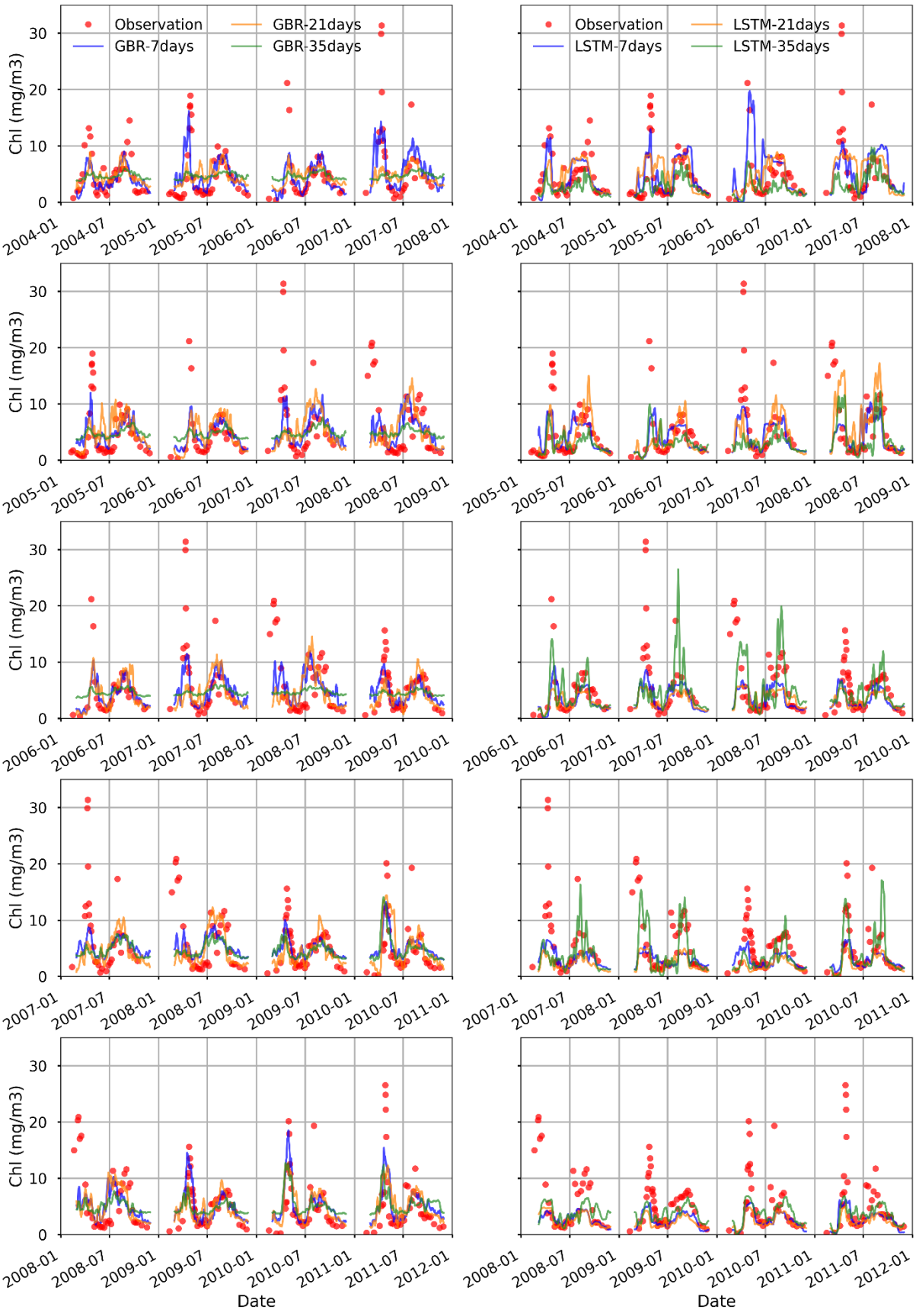
Observation
GBR-7days
GBR-21days
GBR-35days
LSTM-7days
LSTM-21days
LSTM-35days
Chl (mg/m3)
Date

Figure S7. Timeseries of observed and predicted *Chl* from GBR (panels on the left) and LSTM (panels on the right) models based on 7-day, 21-day, and 35-day sample intervals, via leave-four-year-out shuffling year test. Each row is a different 4-year period.

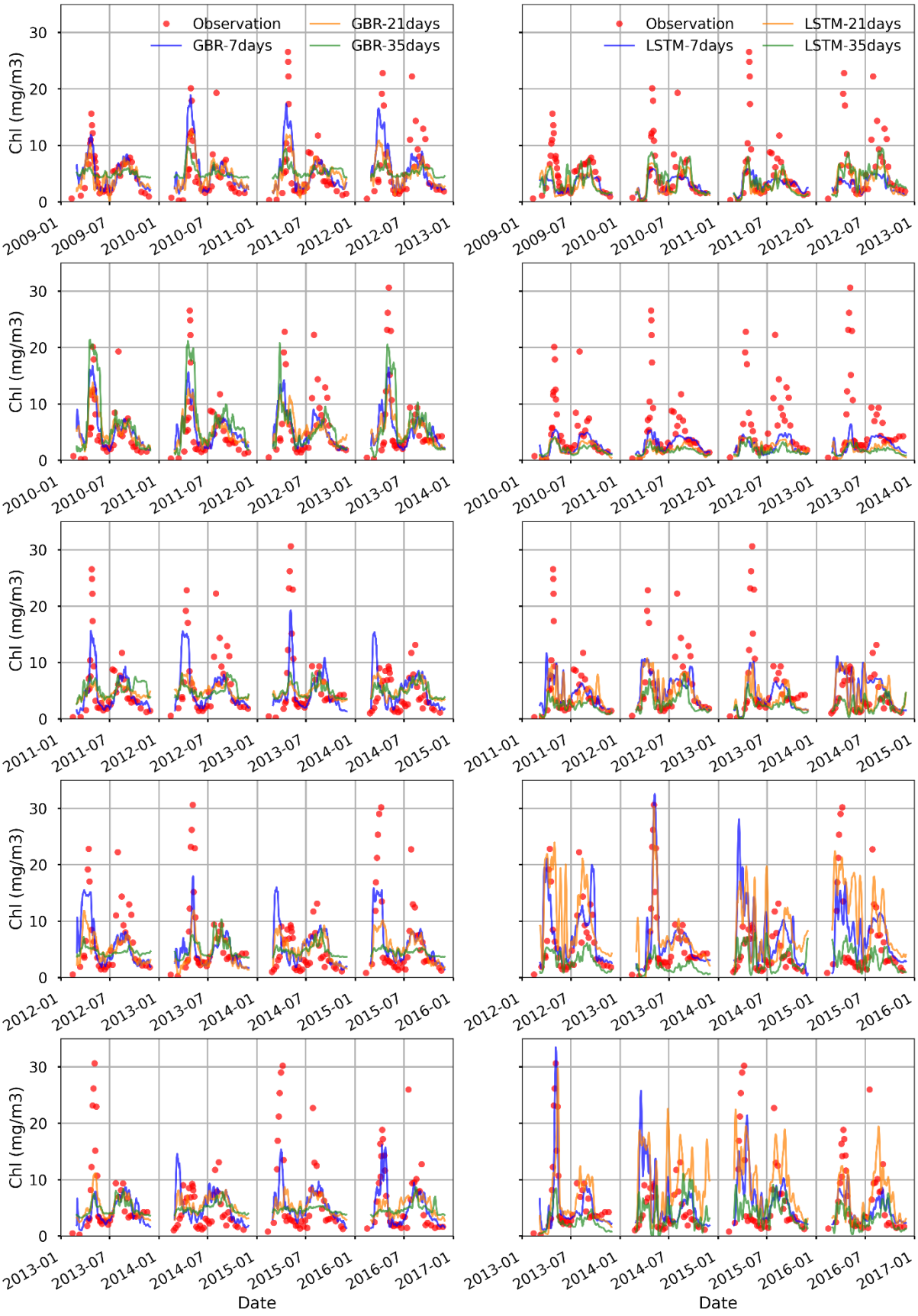
Observation
GBR-7days
GBR-21days
GBR-35days
LSTM-7days
LSTM-21days
LSTM-35days
Chl (mg/m3)
Date

Figure S8 Timeseries of observed and predicted *Chl* from GBR (panels on the left) and LSTM (panels on the right) models based on 7-day, 21-day, and 35-day sample intervals, via leave-four-year-out shuffling year test (Same as Figure S6, but with different x-axis).

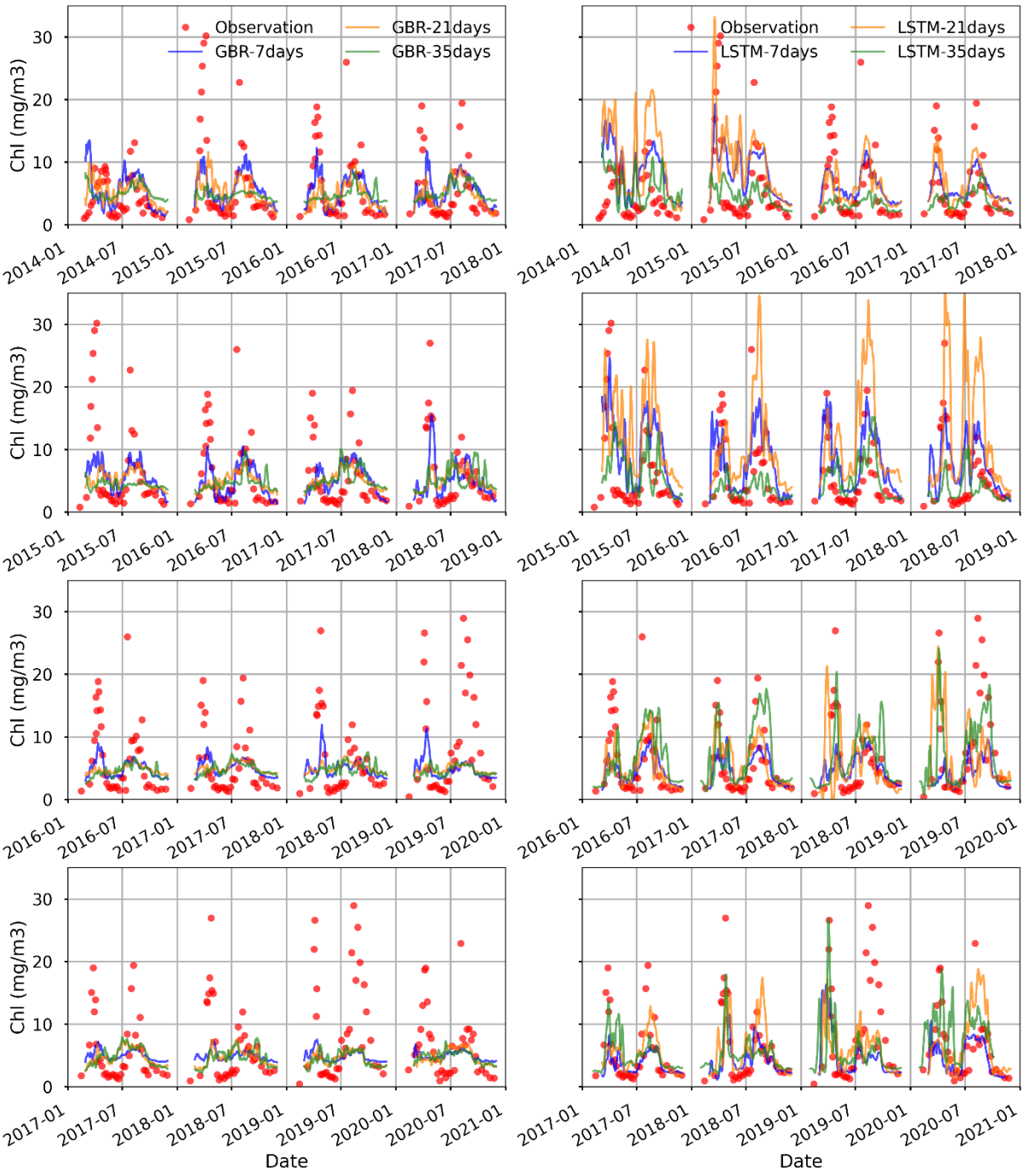

Figure S9. Timeseries of observed and predicted *Chl* from GBR (panels on the left) and LSTM (panels on the right) models based on 7-day, 21-day, and 35-day sample intervals, via leave-four-year-out shuffling year test (Same as Figure S6, but with different x-axis).

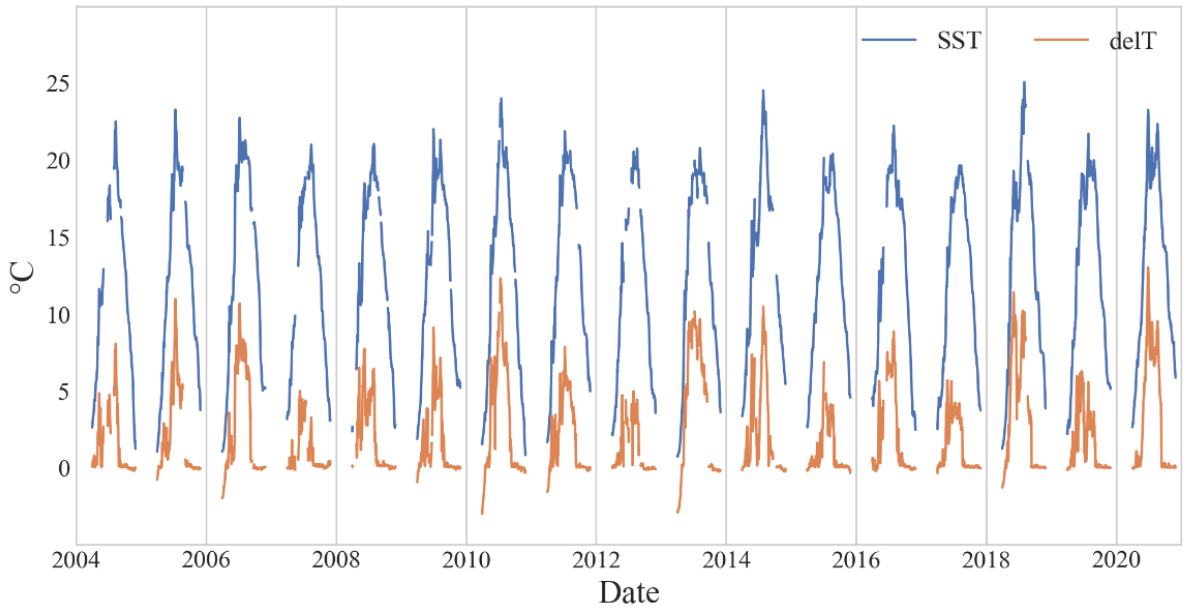

Figure S10. Timeseries of observed surface water temperature and difference between surface water (averaged over the upper 3 m) and bottom water (15 m).