Harmful algal bloom forecast via machine learning (GBR) and deep learning (LSTM) models

New site: Lake Ekoln

Shuqi Lin, PhD

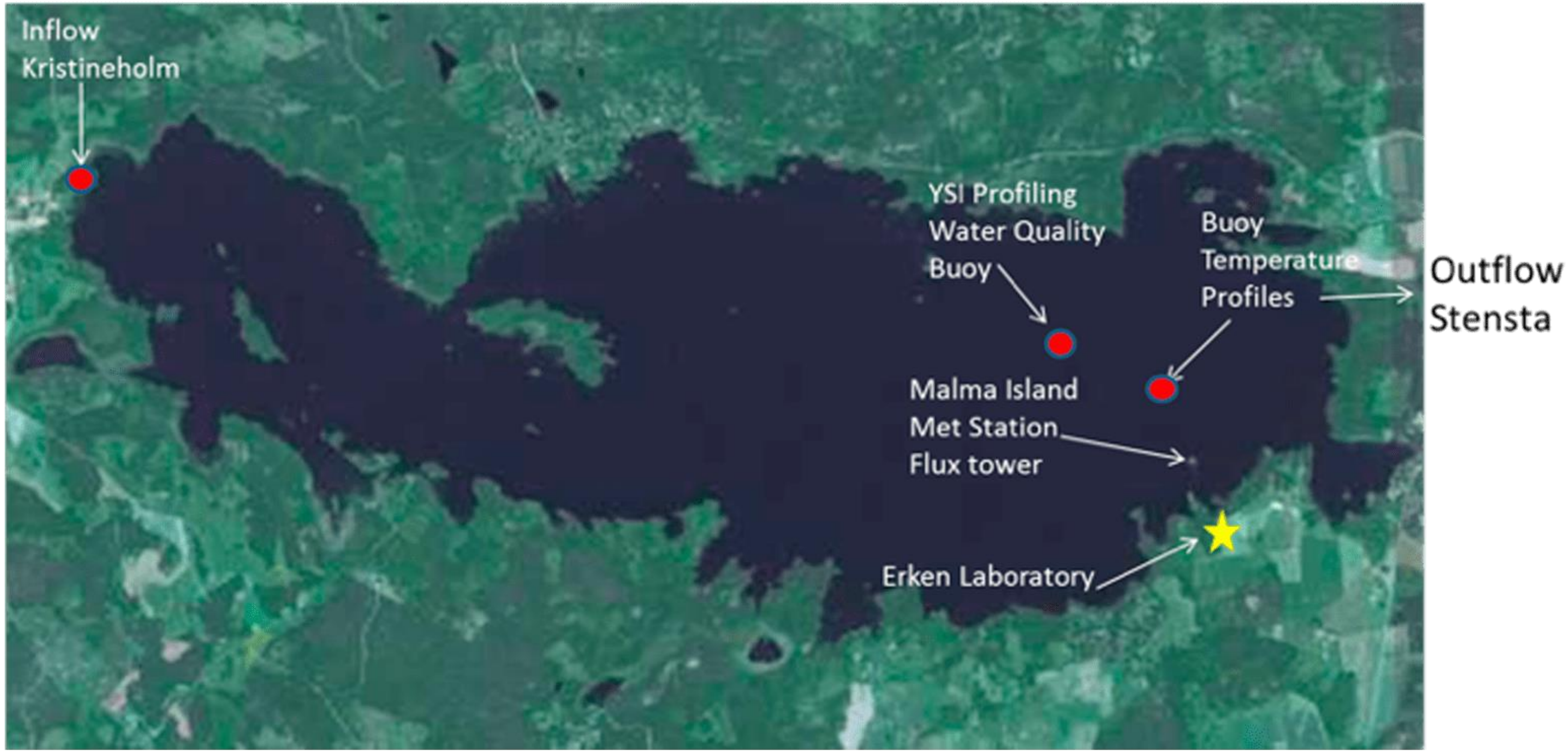
Postdoctoral Fellow
Erken Laboratory and Limnology Department
Uppsala University
Uppsala, Sweden

Data-driven models

- **Gradient Boost Regressor (GBR) ←Tree model**
- **Long Short-Term Memory network (LSTM) ← RNN model**
- **Three scenario:**
  - ➢ Direct data-driven models based on observations of physical factors and less frequently measured nutrients
  - ➢ Two-step data-driven models based on observed physical factors and **pre-generated daily nutrients**
  - ➢ Two-step data-driven models based on observed physical factors, **pre-generated daily nutrients**, and **hydrodynamic features** from the process-based (PB) model

Lake Erken

- Surface area of 24 km^2
- Average depth of 9 m
- Maximum depth of 21 m



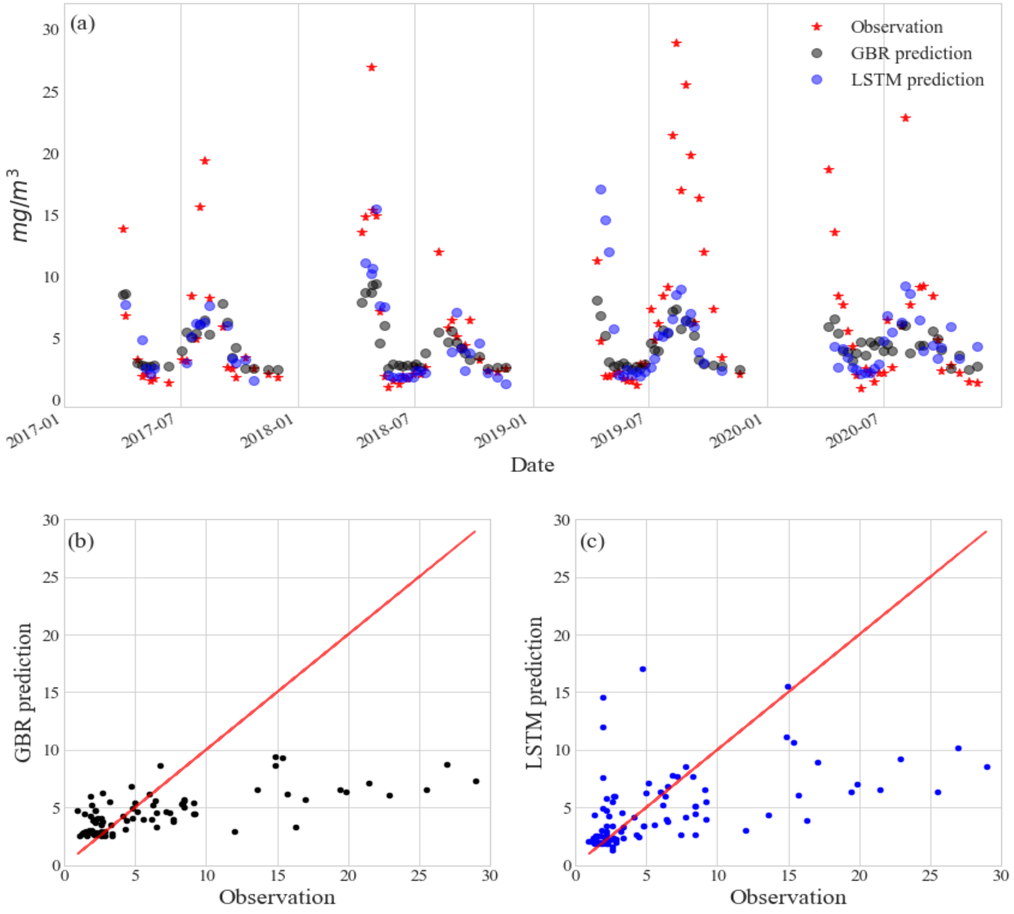
- Meteorological data
- Water temperature profiles
- Water discharge
- Water samples (1-2 weeks)

- Training data: 2004-2016
- Testing data: 2017-2020

S1: Direct data-driven models based on observations of physical factors and less frequently measured nutrients

- Features: Inflow, AirT, Prec, U, Humidity, CC, swr, Ice_d, days from iceoff, delT,

NOX, NH4, PO4, TotP, Si, O2 ← Weekly data



GBR evaluation:

RMSE: 5.55 mg/m^3

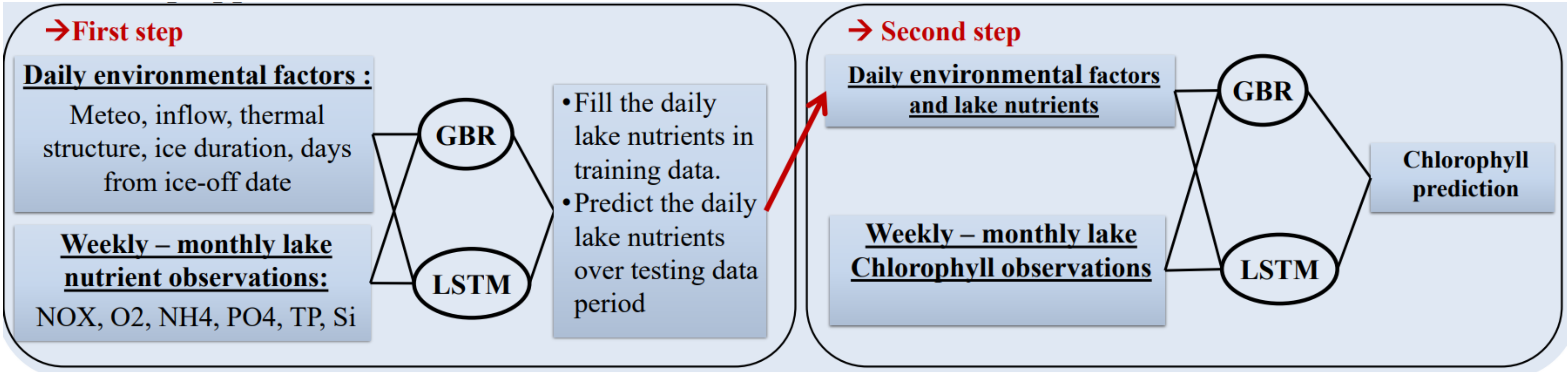
R2 0.21

GBR evaluation:

RMSE: 5.82 mg/m^3

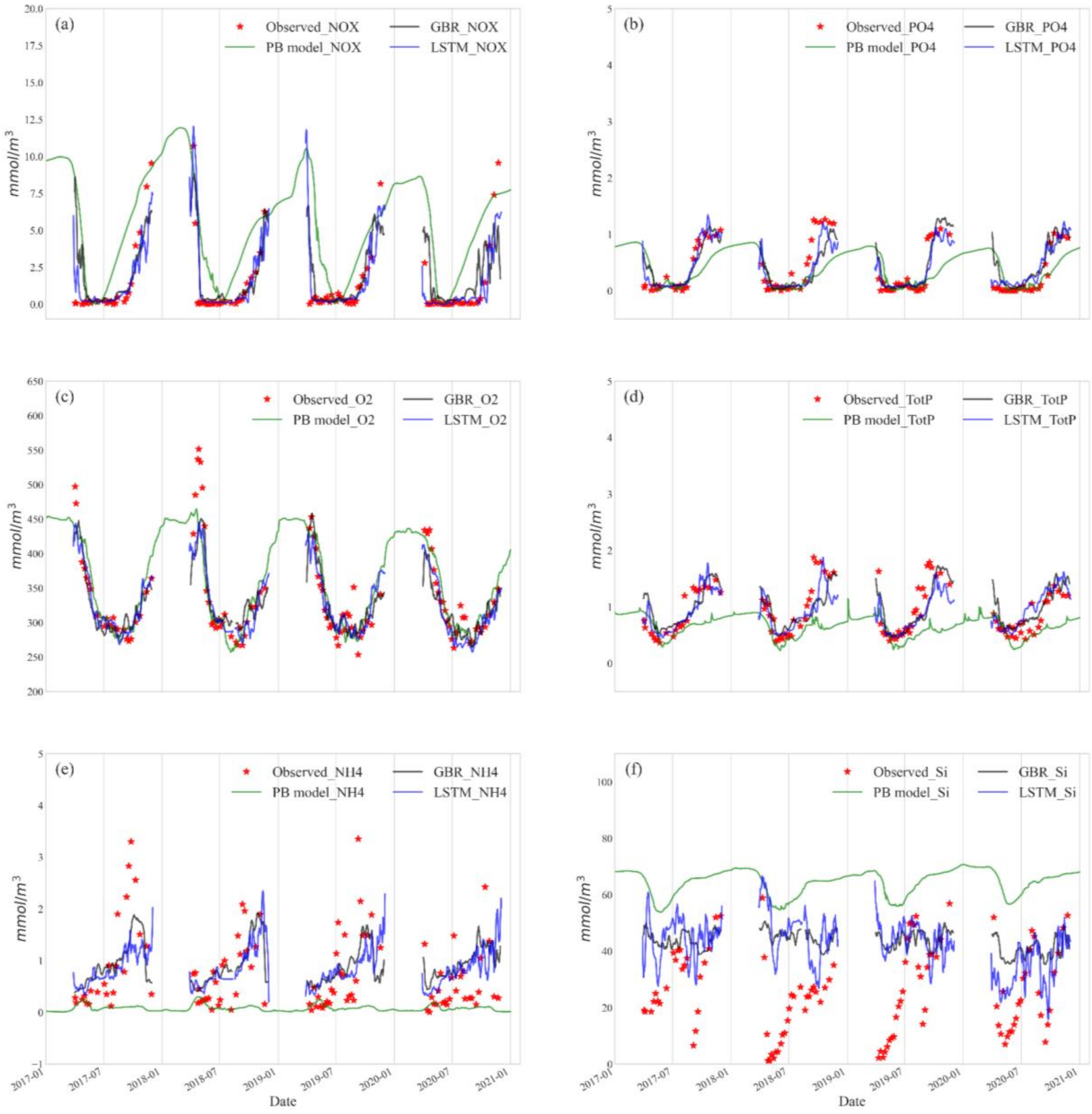
R2 0.17

S2: Two-step data-driven models based on pre-generated daily nutrients and observed physical factors

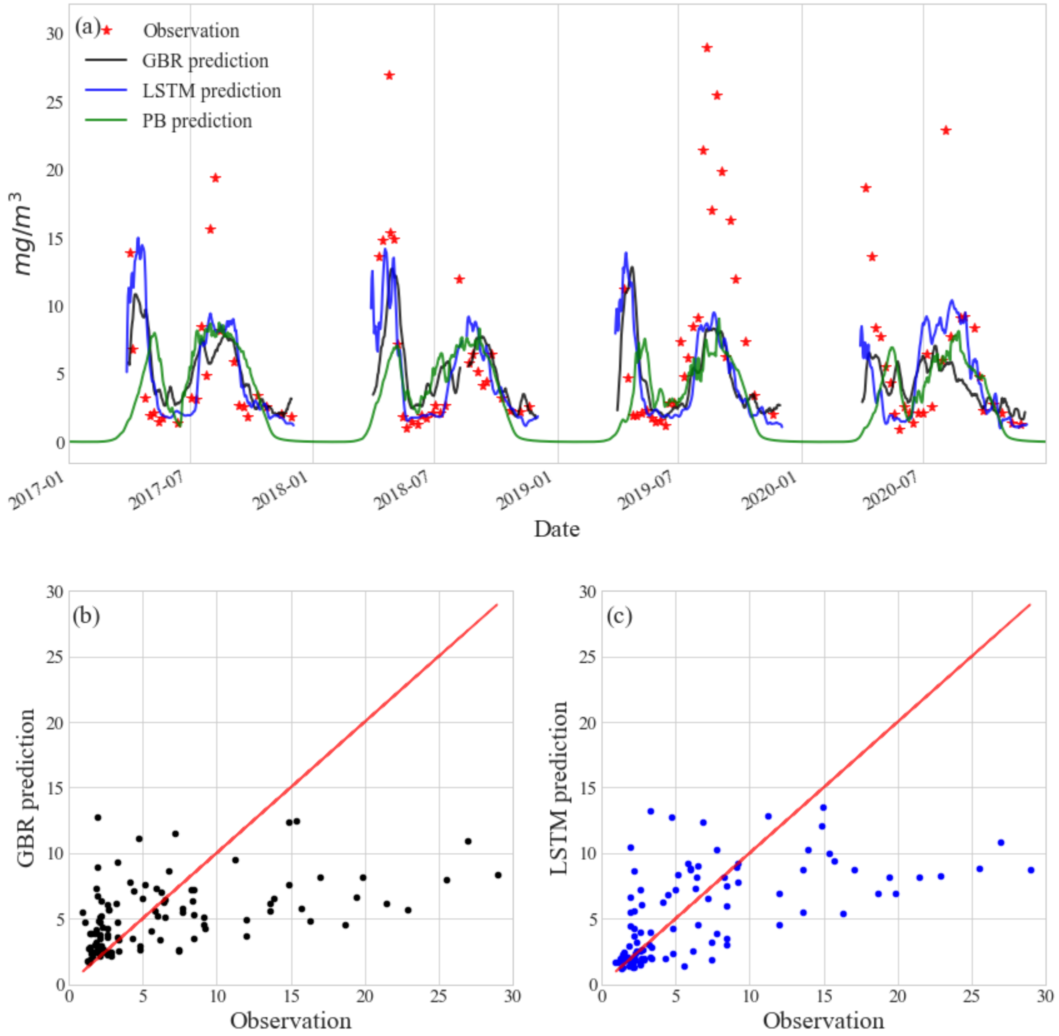


- 10 daily training features: Inflow, AirT, Prec, delT, U, Humidity, CC, swr, Ice_d, days from iceoff
- **Time_steps = 7 for LSTM model**

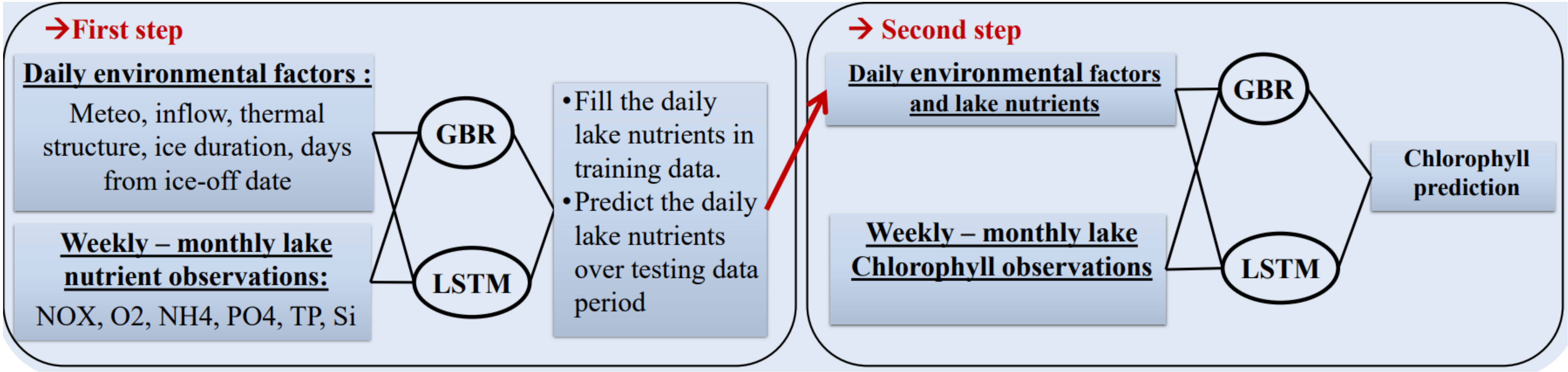
S2: Two-step data-driven models based on pre-generated daily nutrients and observed physical factors



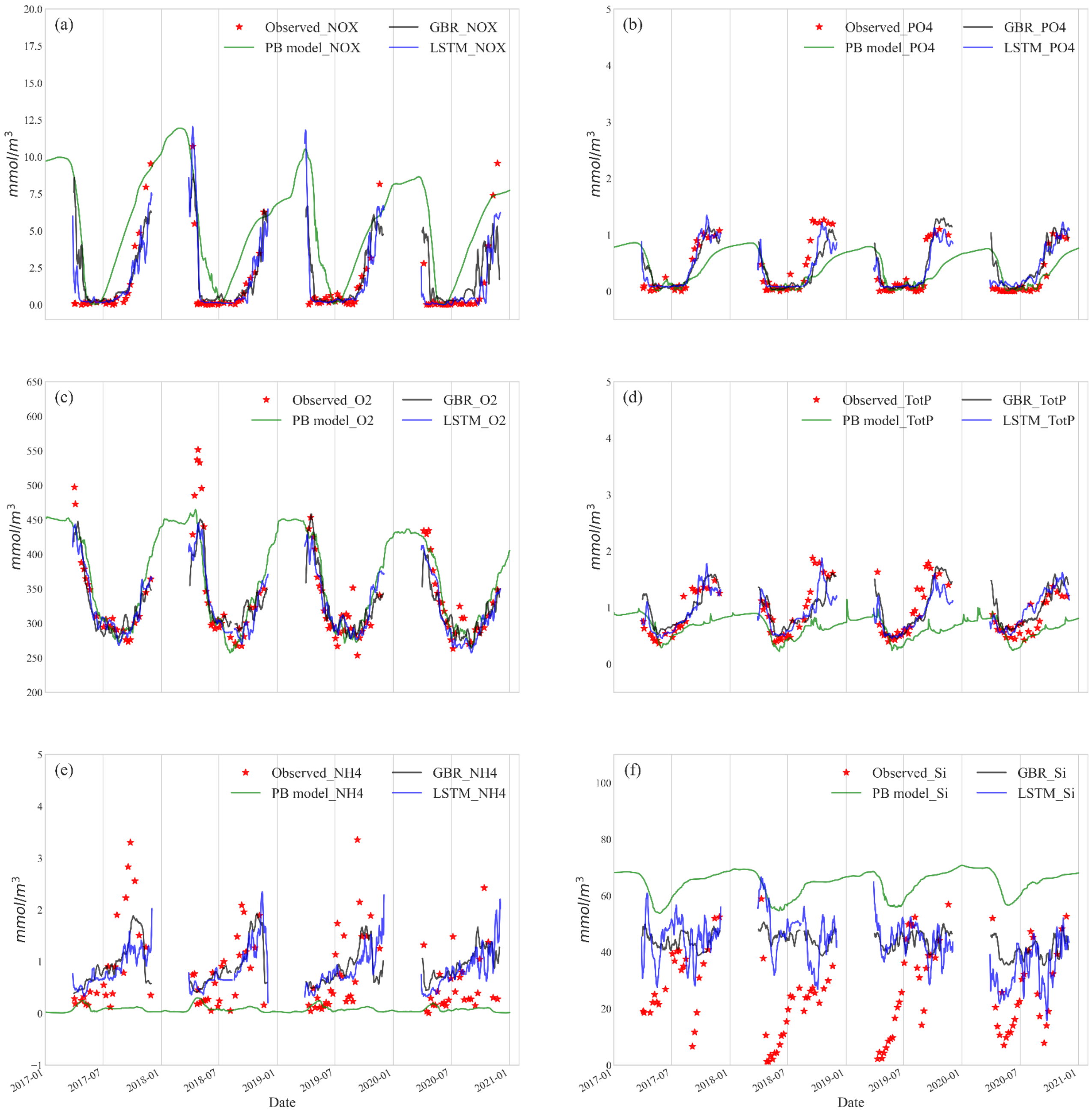
Chl

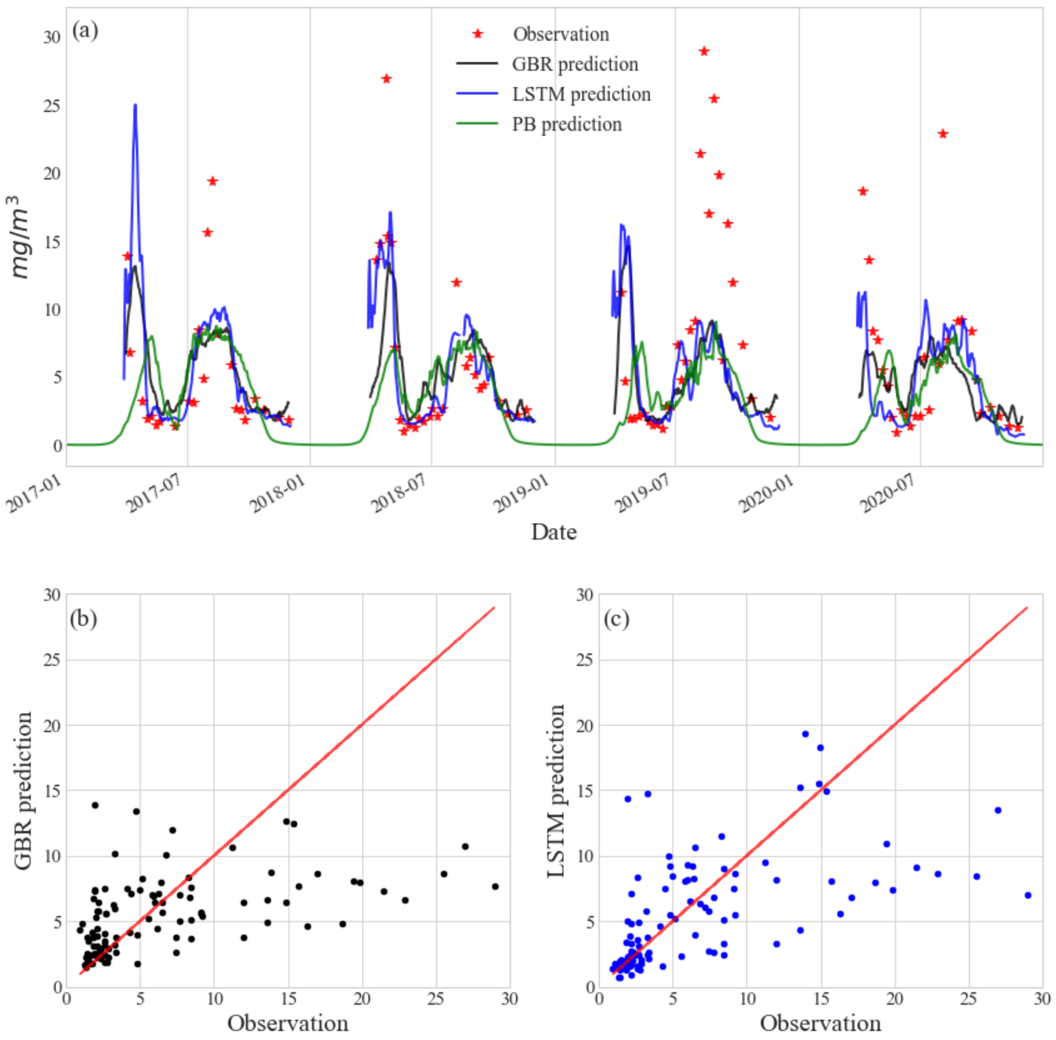


S3: Two-step data-driven models based on pre-generated daily nutrients, observed physical factors and hydrodynamic features from the process-based model



- 13 training features
- Additional daily environmental factors: **Mixing layer depth, Wedderburn number, thermocline depth**

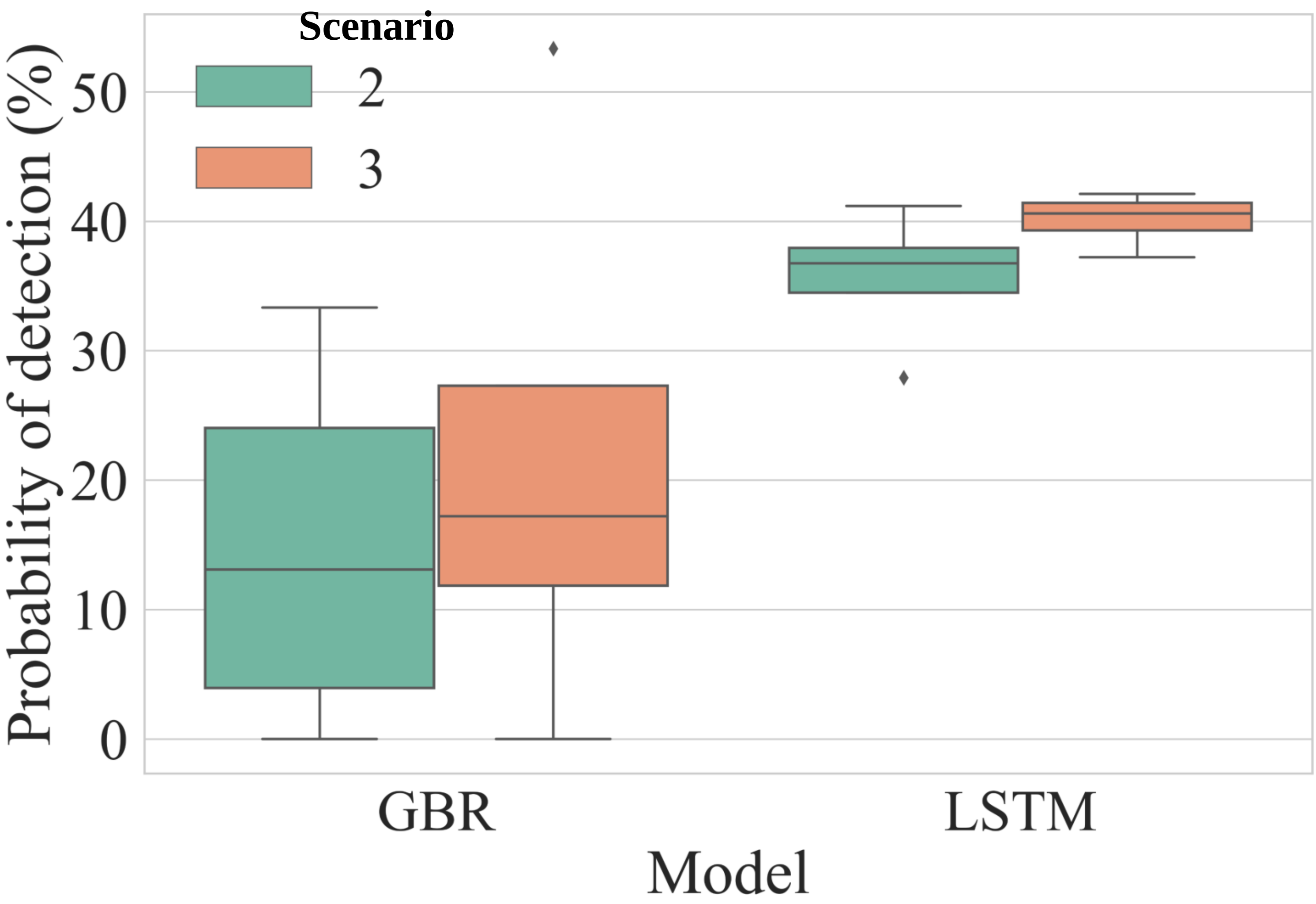




	Testing RMSE	Testing R2
GBR	5.2	0.31
LSTM	5.22	0.31

Probability of detection: P_d = Hits/(Hits + Misses)

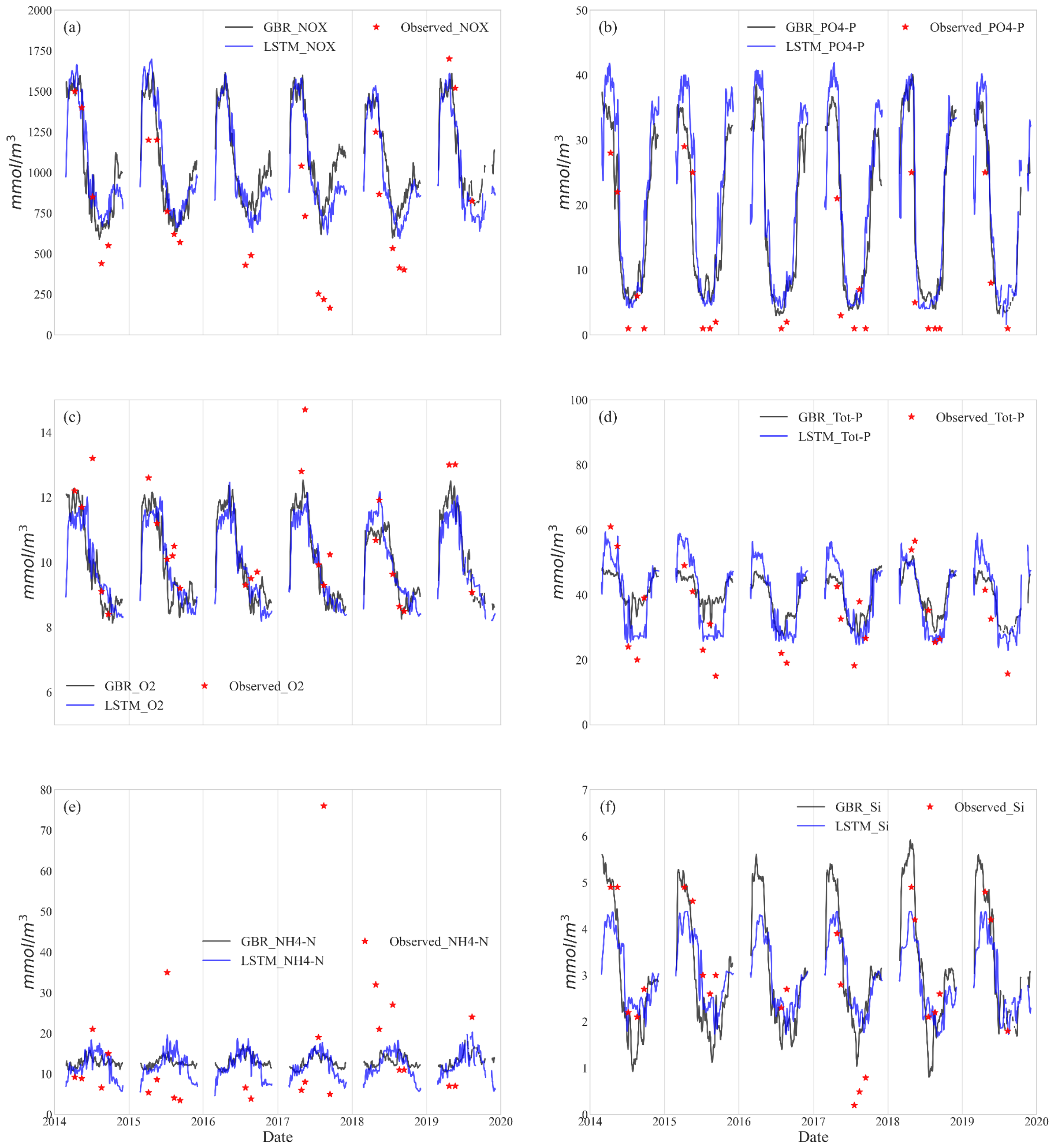
Hit ← Both predicted ΔChl and observed ΔChl >0.3
Miss ← Observed ΔChl >0.3 but predicted ΔChl <=0.3

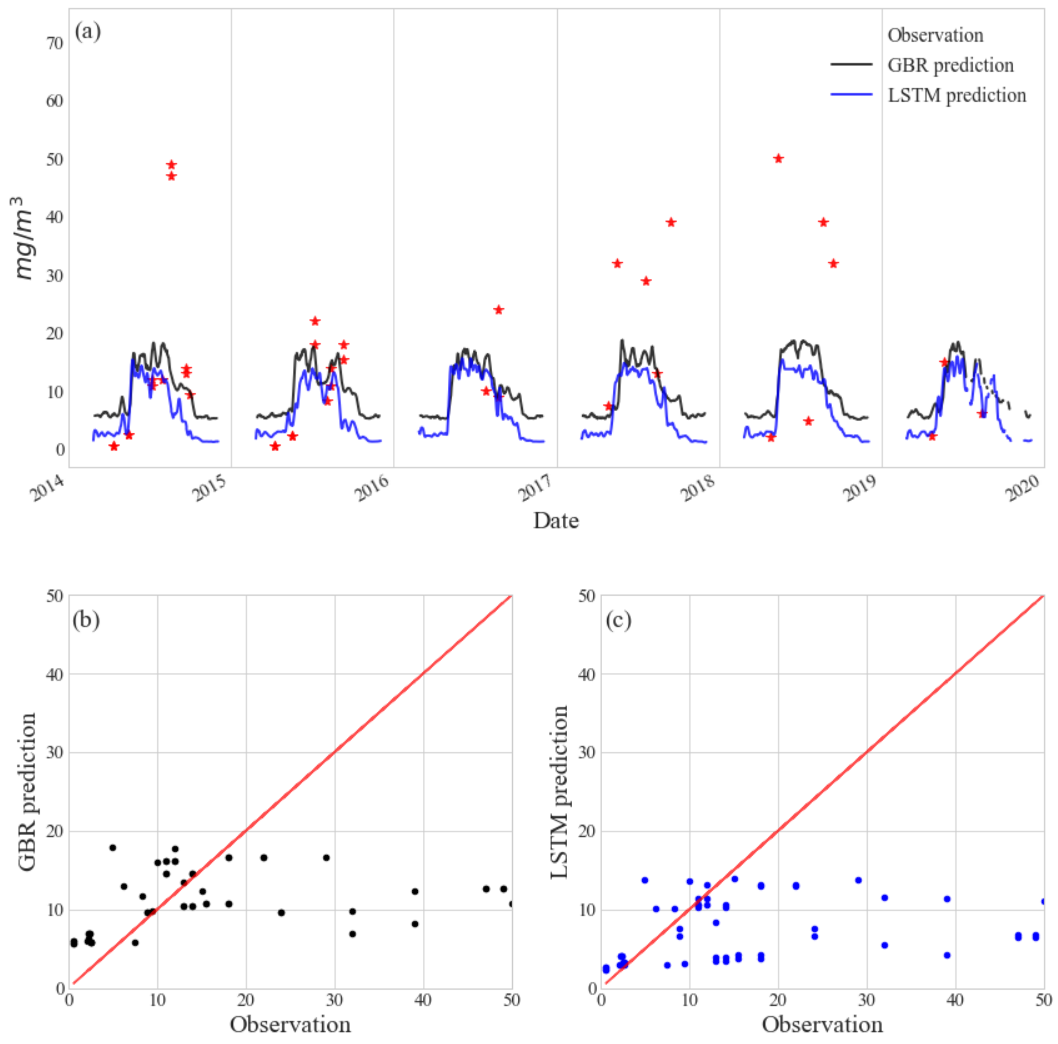


Lake Ekoln

- Less monitored (monthly lake nutrients sampling)
- Data source (1985-2016)
  - ✓ SMHI (https://www.smhi.se/; Meteorological data)
  - ✓ SLU Environmental data MVM (https://miljodata.slu.se/MVM; Station name: Ekoln Vreta Udd; water temperature and lake nutrient data)
  - ✓ GOTM model → Ice duration, days from iceoff date, MLD, W, thermD.
- Training data: 1985-2013
- Testing data: 2014-2020

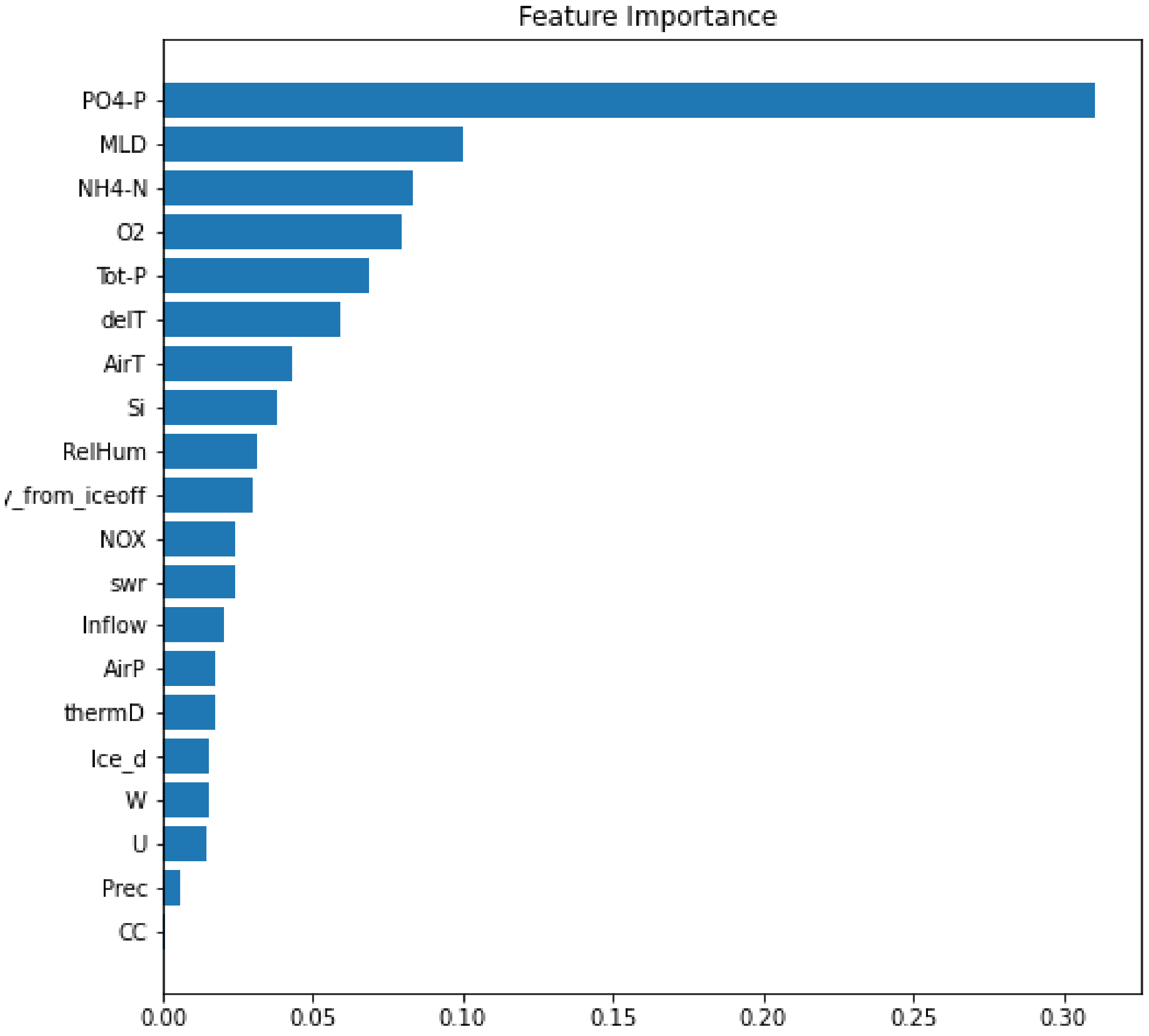
ΔChl >0.3

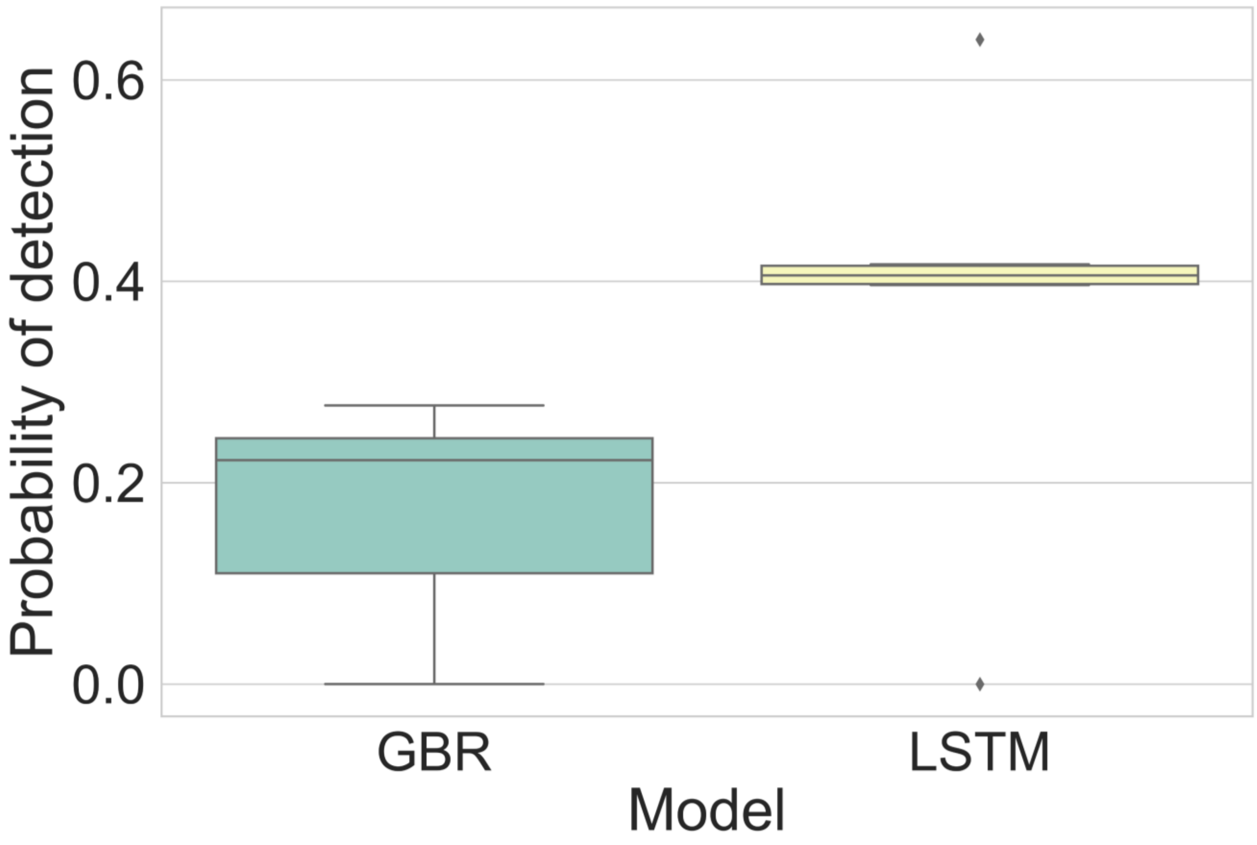




	Testing RMSE	Testing R2
GBR	14.2	-0.02
LSTM	15.48	-0.27

S3: Two-step data-driven models based on pre-generated daily nutrients, observed physical factors and hydrodynamic features from the process-based model

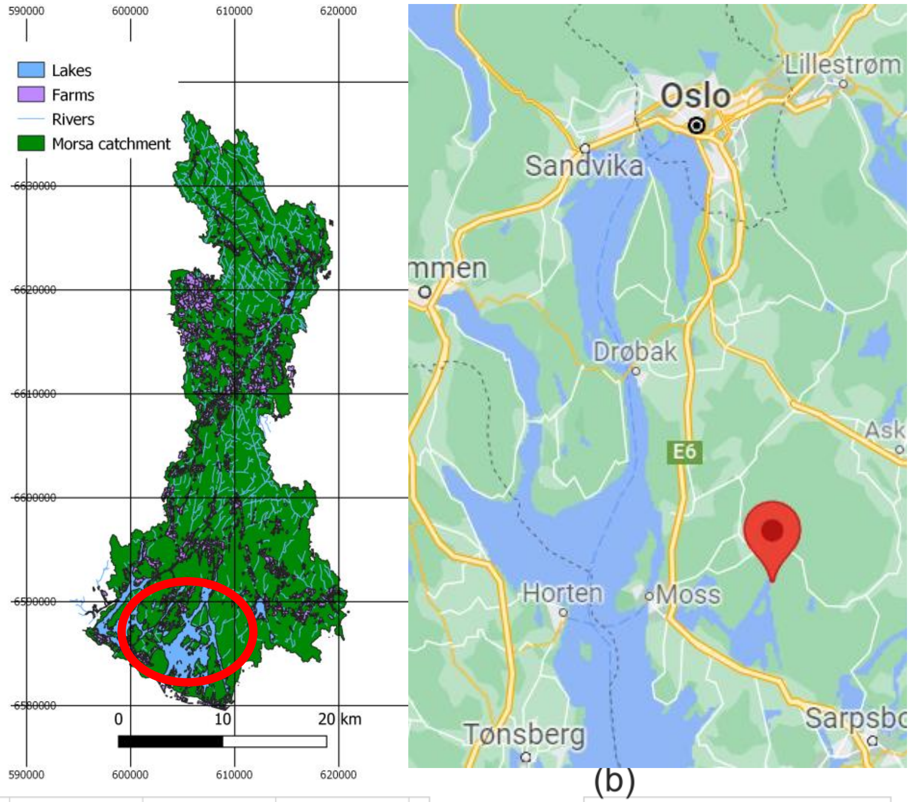


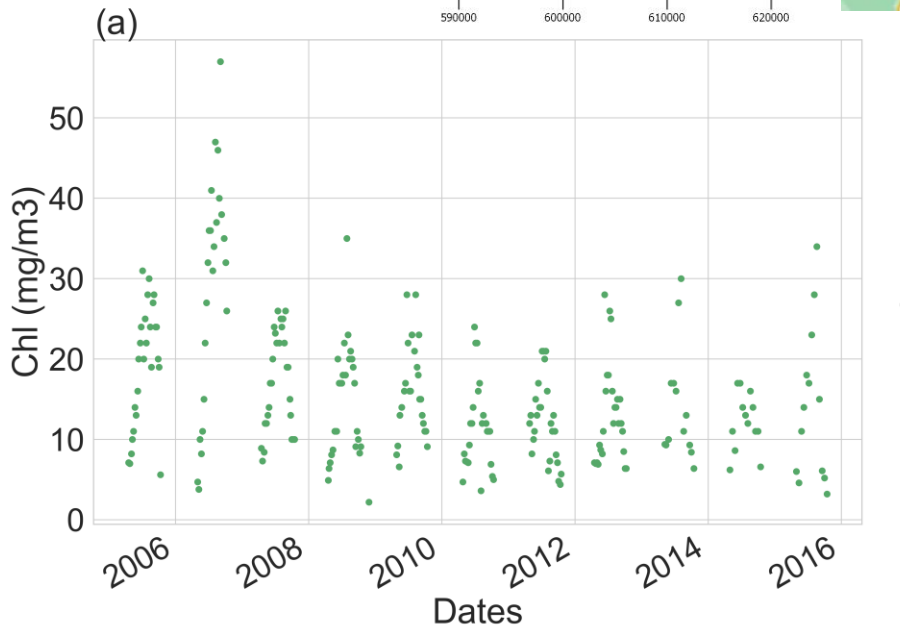


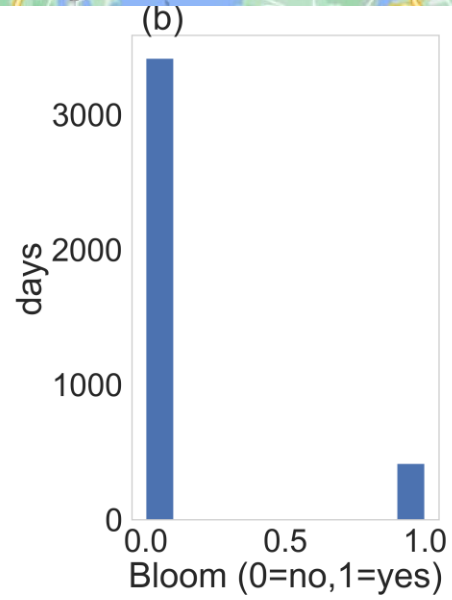
Lake Vansjø

- Surface area of 12 km^2
- Maximum depth of 21 m
- Daily meteorological, inflow, temperature profiles with 0.5 m interval, annual ice record, and weekly lake nutrients data.
- No hydrodynamic model → Use observed temperature to generate MLD, W, thermD.

- Training data: 2005-2011 (7 yrs)
- Testing data: 2012-2015 (4 yrs)

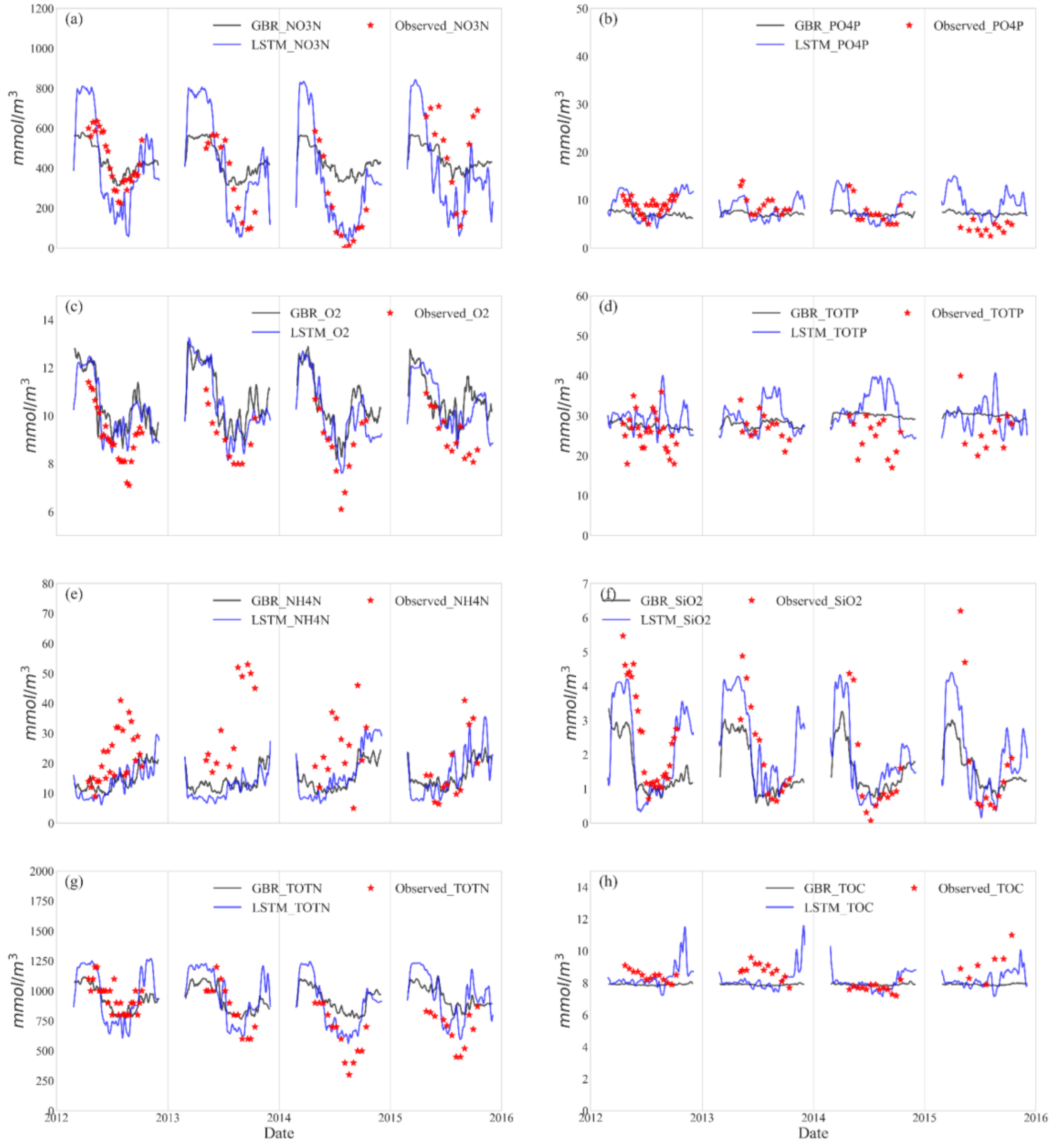


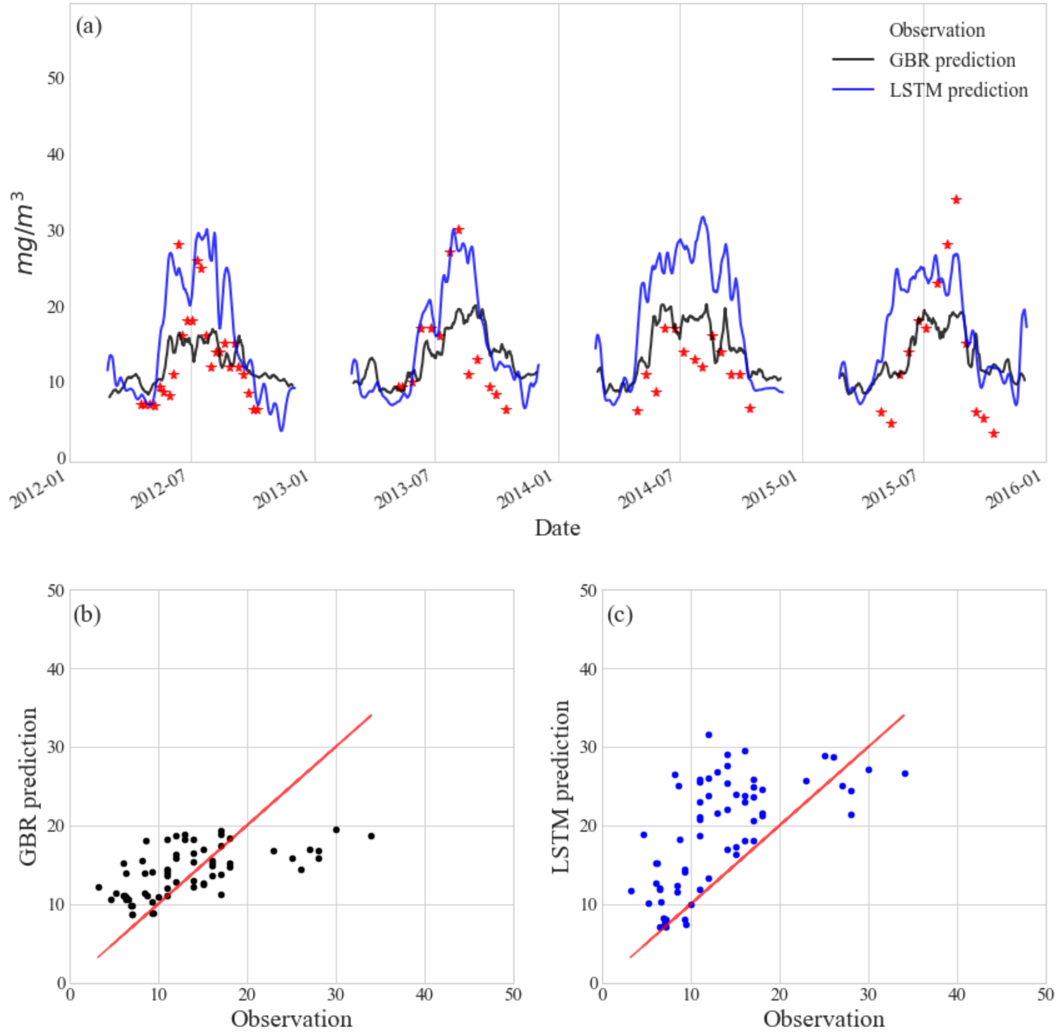




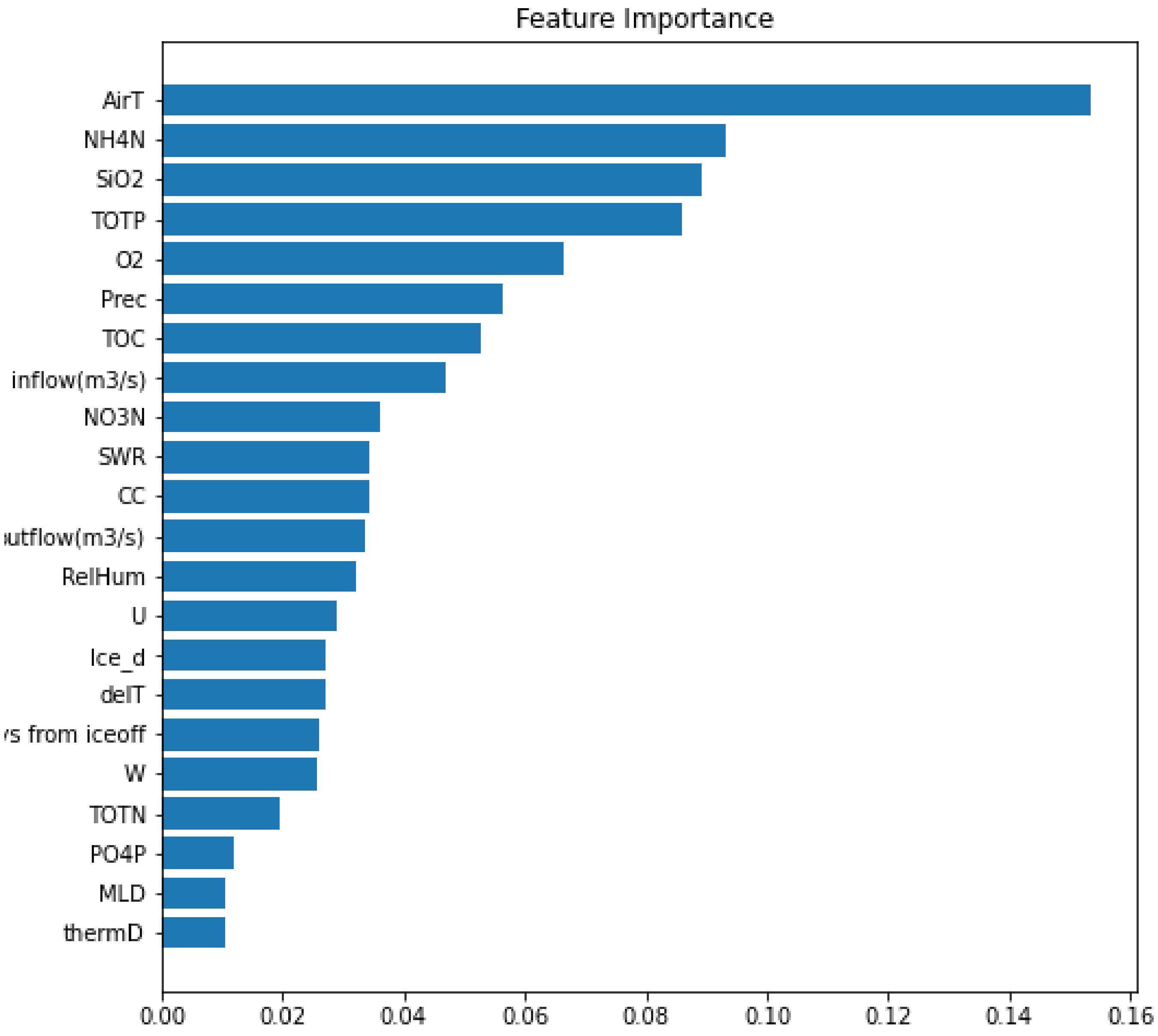
ΔChl >0.3

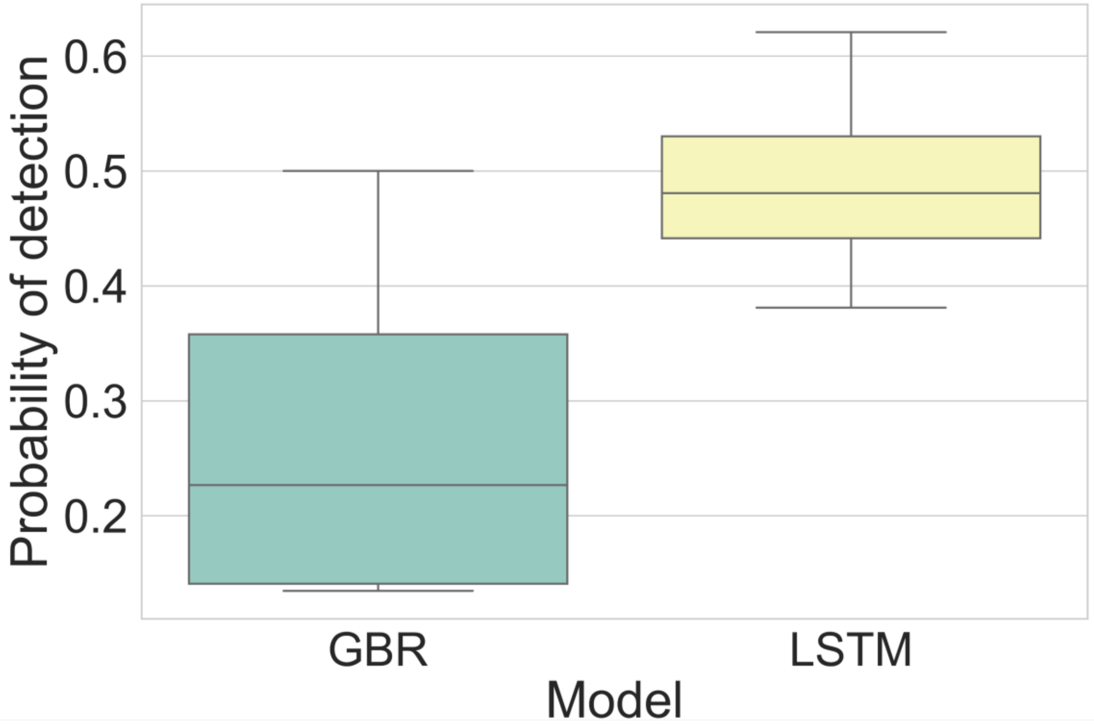
8 lake nutrients: NO3, PO4, O2, Total P, NH4, SiO2, Total N, Total organic carbon





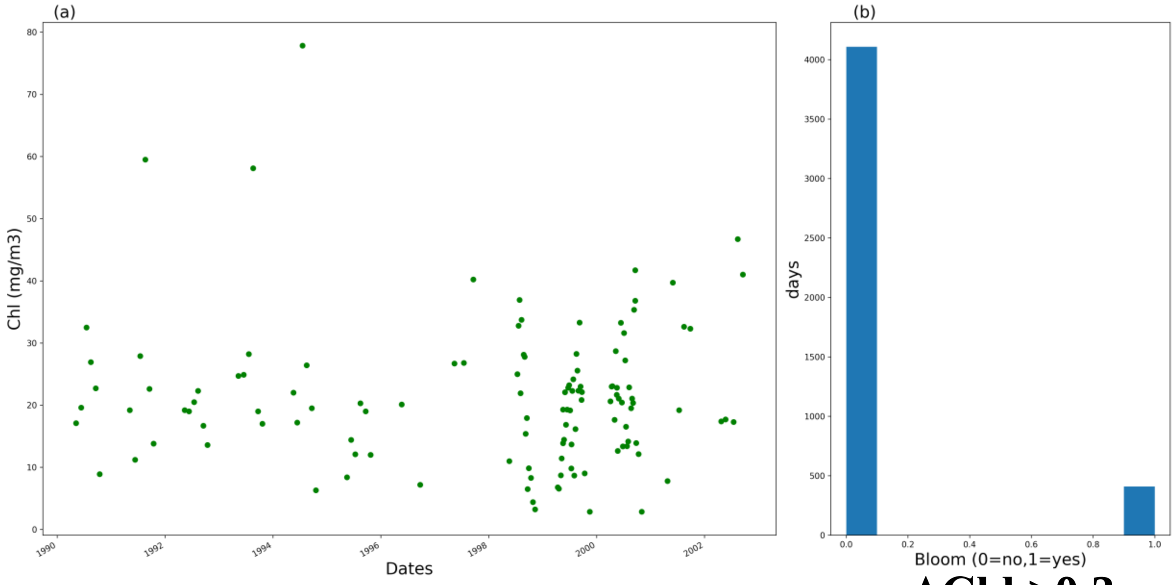
	Testing RMSE	Testing R2
GBR	5.82	0.32
LSTM	8.56	-0.66


Feature Importance
AirT
NH4N
SiO2
TOTP
O2
Prec
TOC
inflow(m3/s)
NO3N
SWR
CC
utflow(m3/s)
RelHum
U
Ice_d
delT
s from iceoff
W
TOTN
PO4P
MLD
thermD
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16


Probability of detection
0.6
0.5
0.4
0.3
0.2
GBR
LSTM
Model

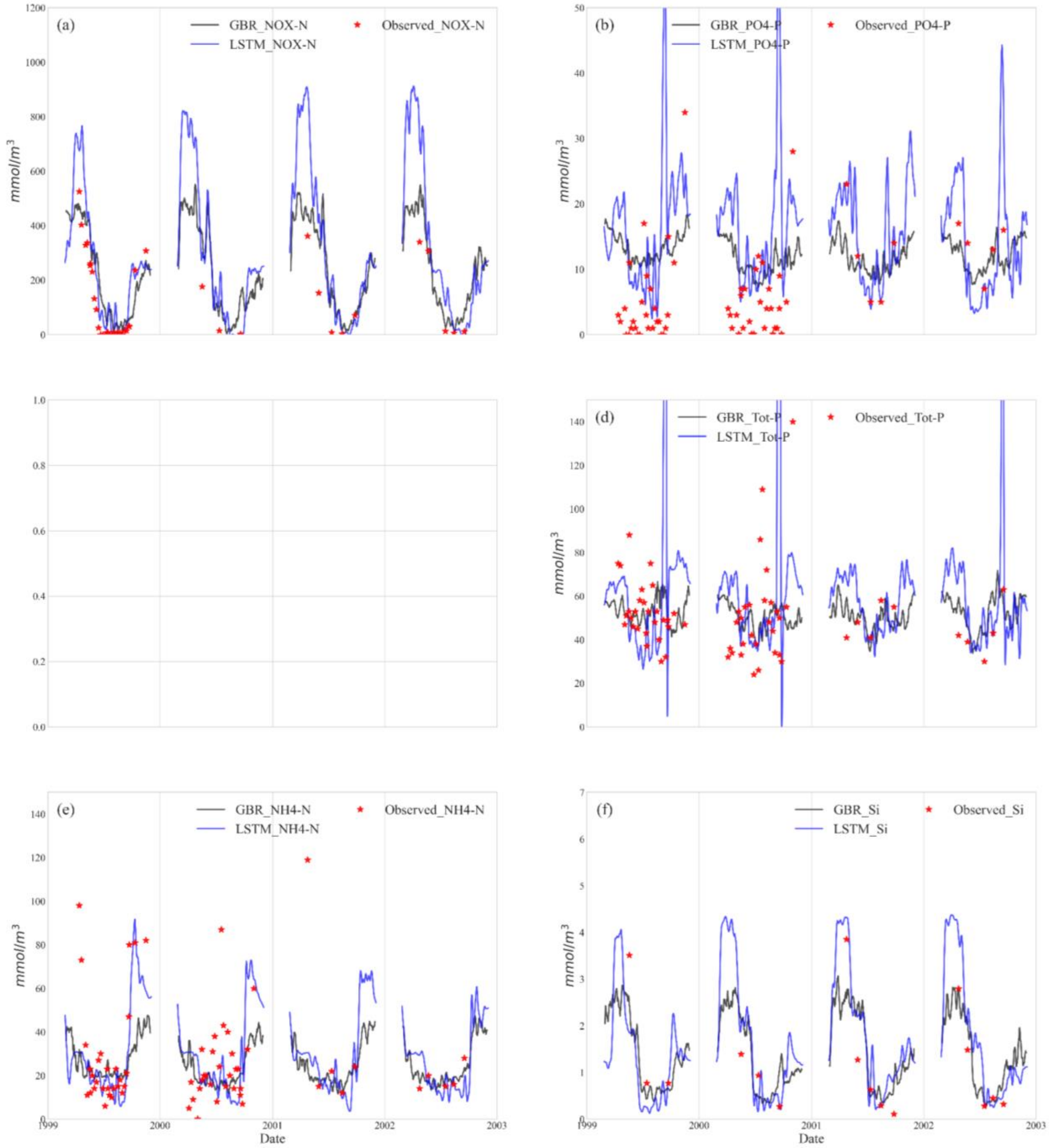
Lake Galten

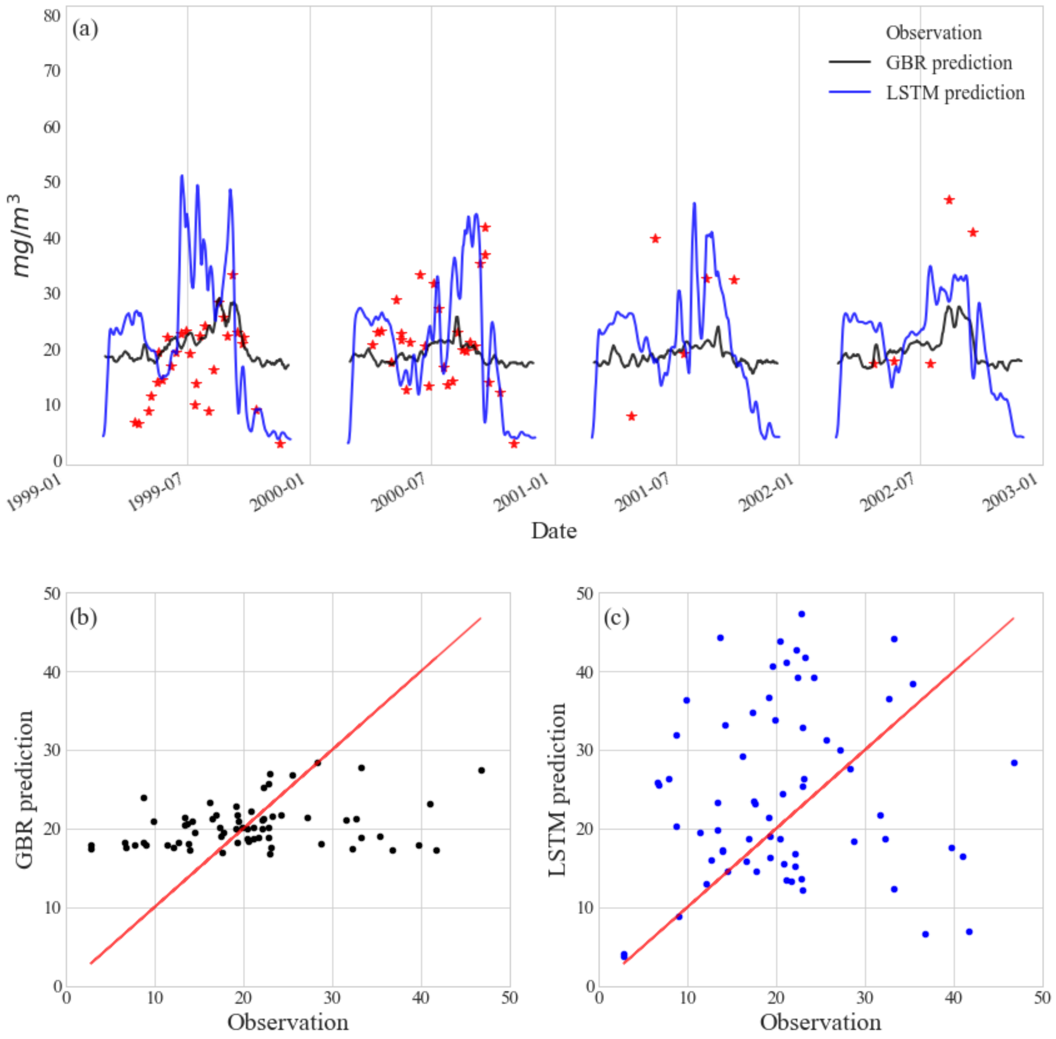
- Monthly lake nutrients sampling except 1999 and 2000 with weekly data.
- Data source (1990-2002)
  - ➢ Grided meoteological model → Meteorological data
  - ➢ SLU Environmental data MVM (Station name: Galten) → water temperature and lake nutrient data
  - ➢ GOTM model → Ice duration, days from iceoff date, MLD, W, thermD.
- Training data: 1990-1998 (9 yrs)
- Testing data: 1999-2002 (4 yrs)



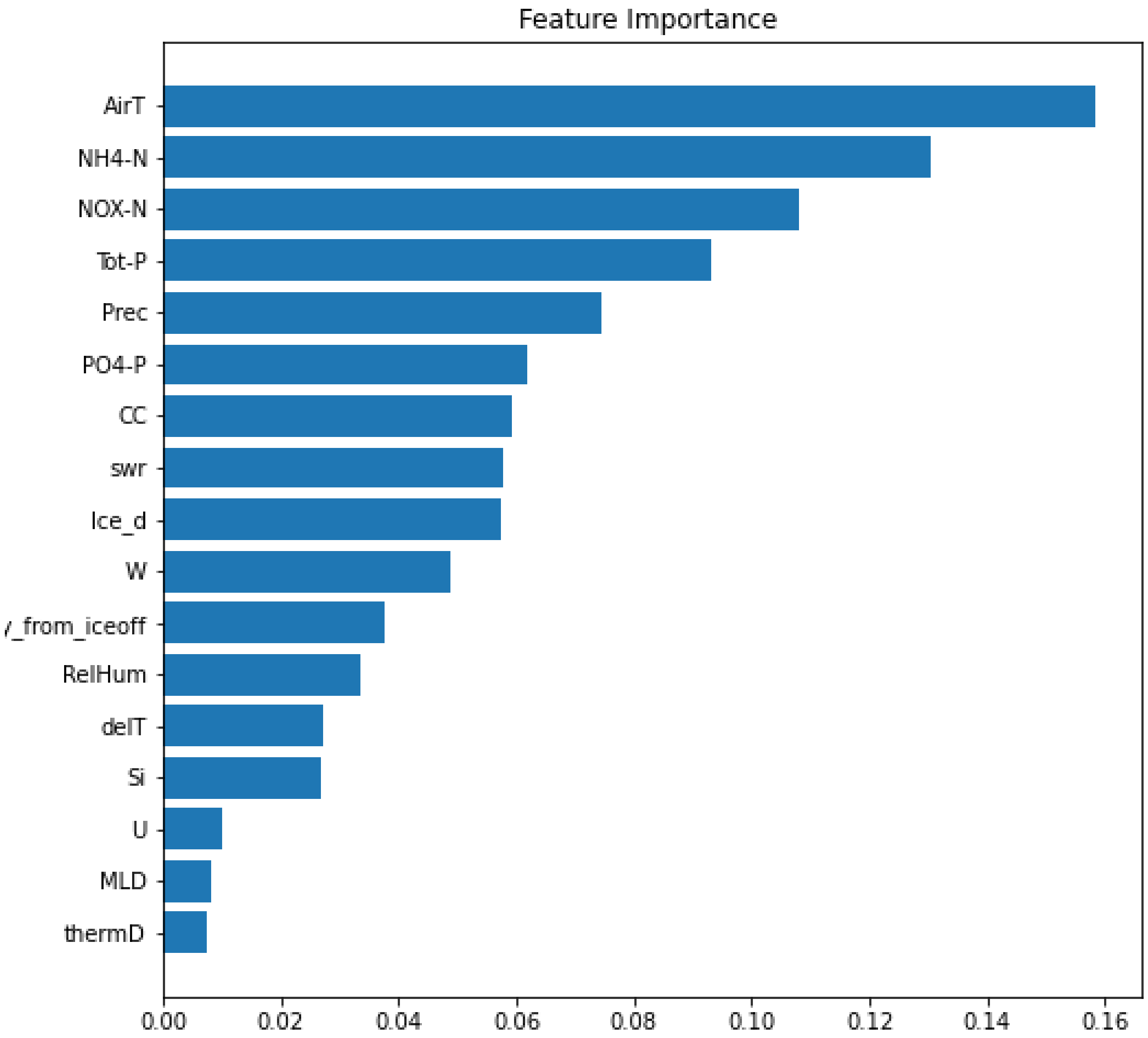
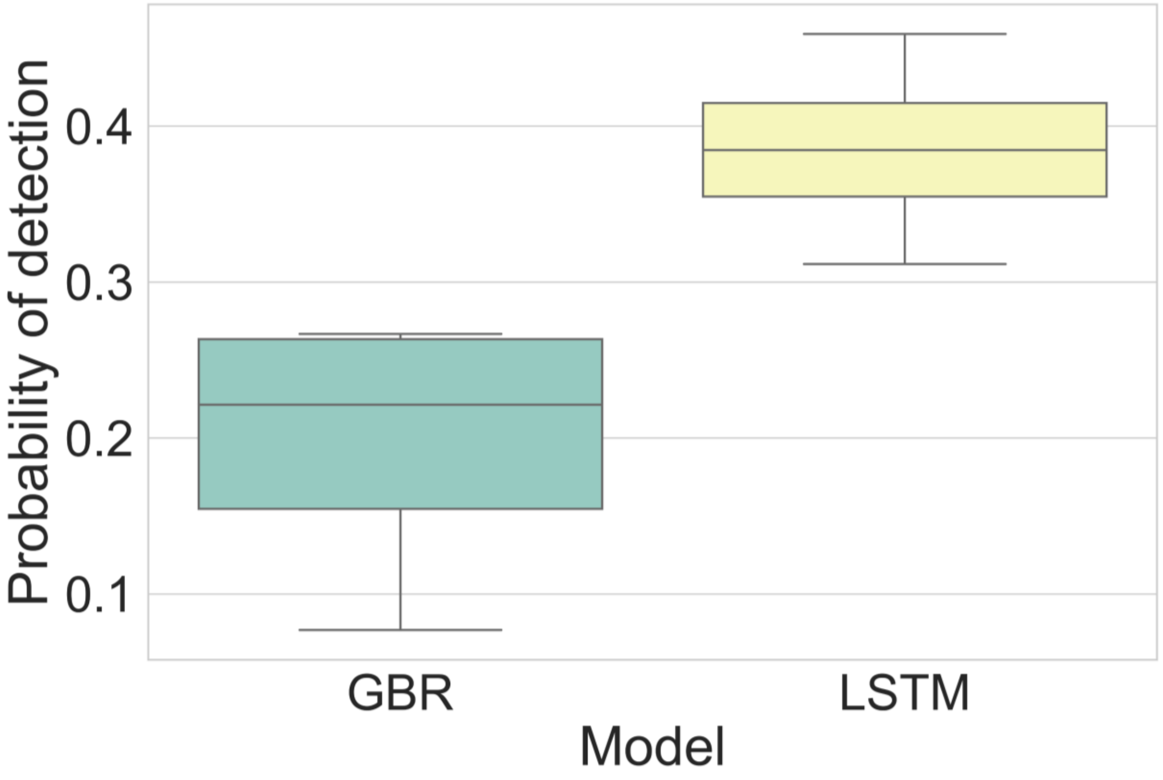
ΔChl >0.3

5 lake nutrients: NOX, PO4, Total P, NH4, Si



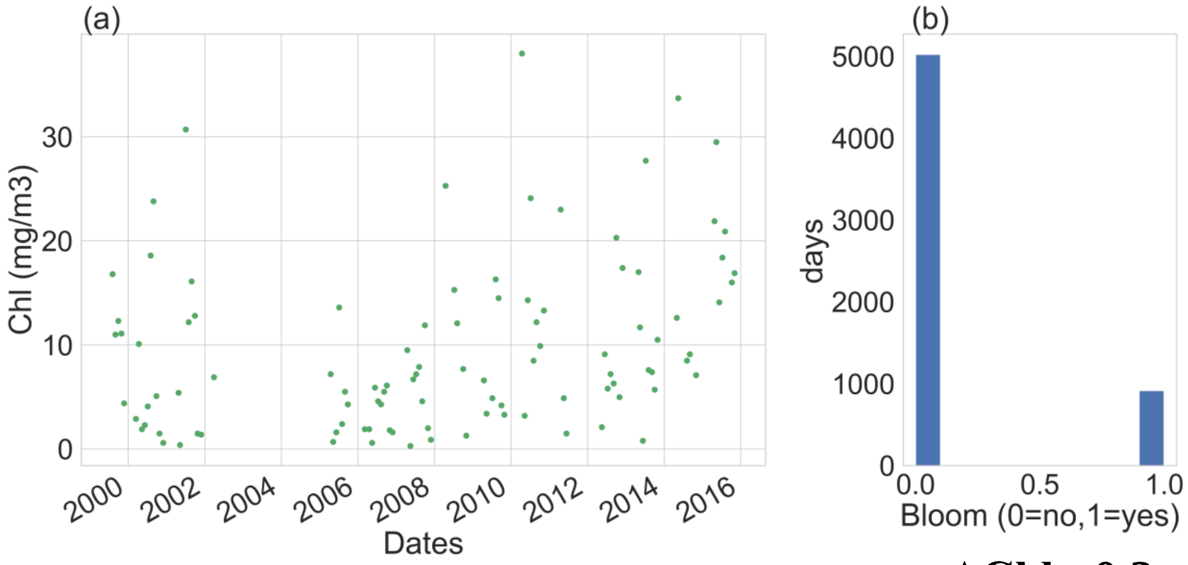


	Testing RMSE	Testing R2
GBR	8.9	0.09
LSTM	15.35	-1.72


Feature Importance
AirT
NH4-N
NOX-N
Tot-P
Prec
PO4-P
CC
swr
Ice_d
W
/_from_iceoff
RelHum
delT
Si
U
MLD
thermD
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16

Probability of detection
0.4
0.3
0.2
0.1
GBR
LSTM
Model

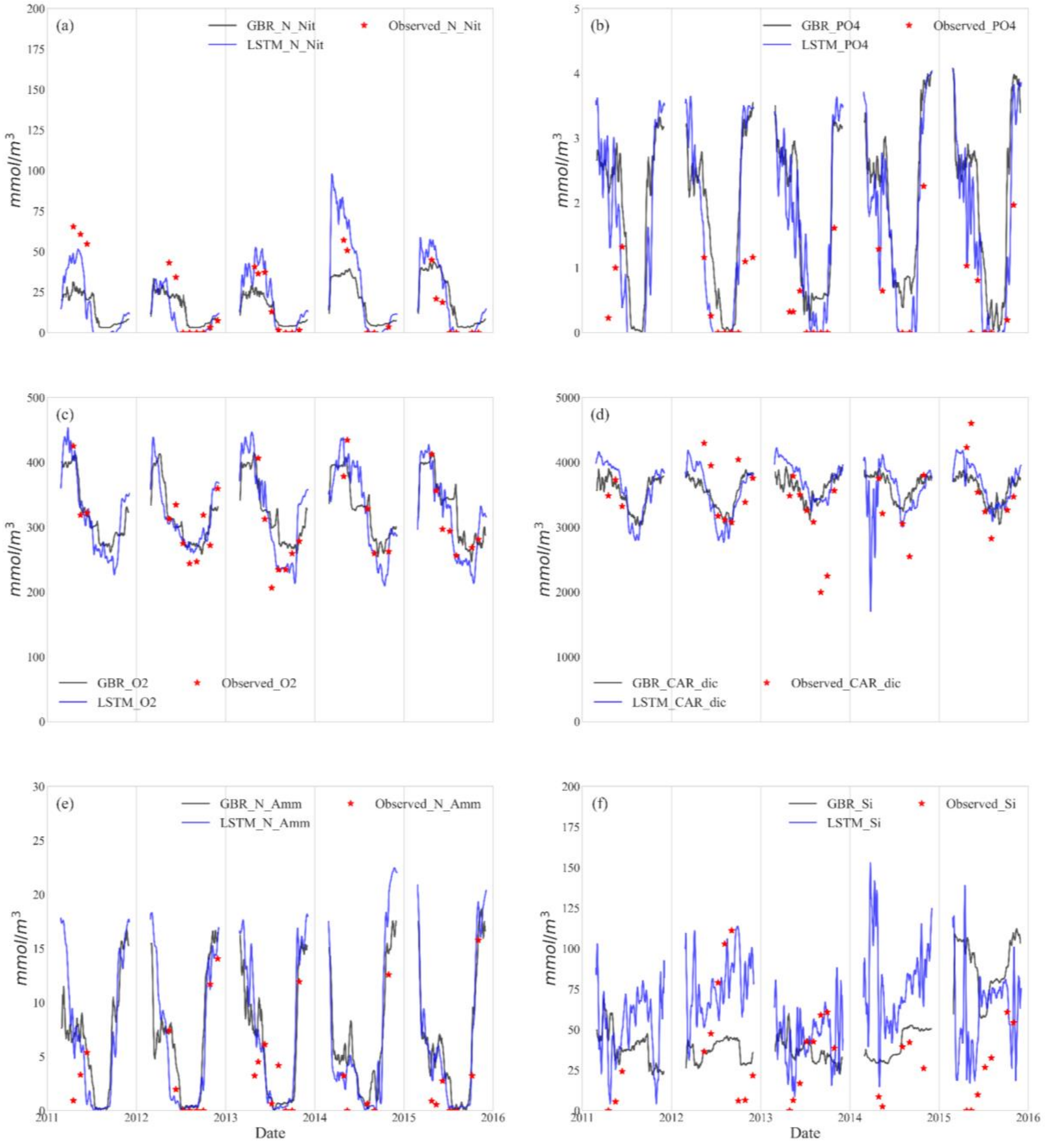
Lake Mendota

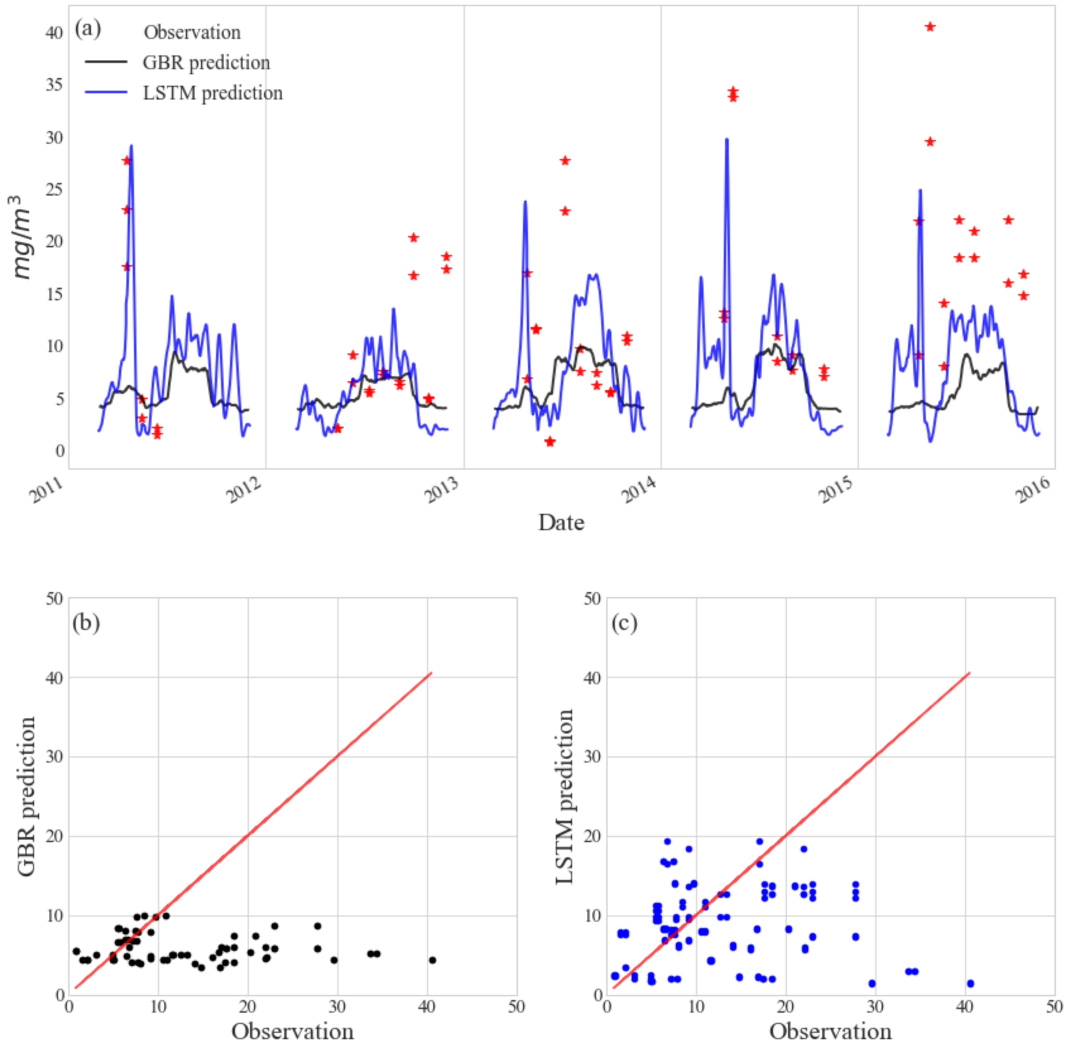
- Daily meteorological, inflow, annual ice record, and weekly lake nutrients data.
- GLM → MLD, W.
- Training data: 1999-2002, 2005-2010 (10 yrs)
- Testing data: 2011-2015 (5 yrs)



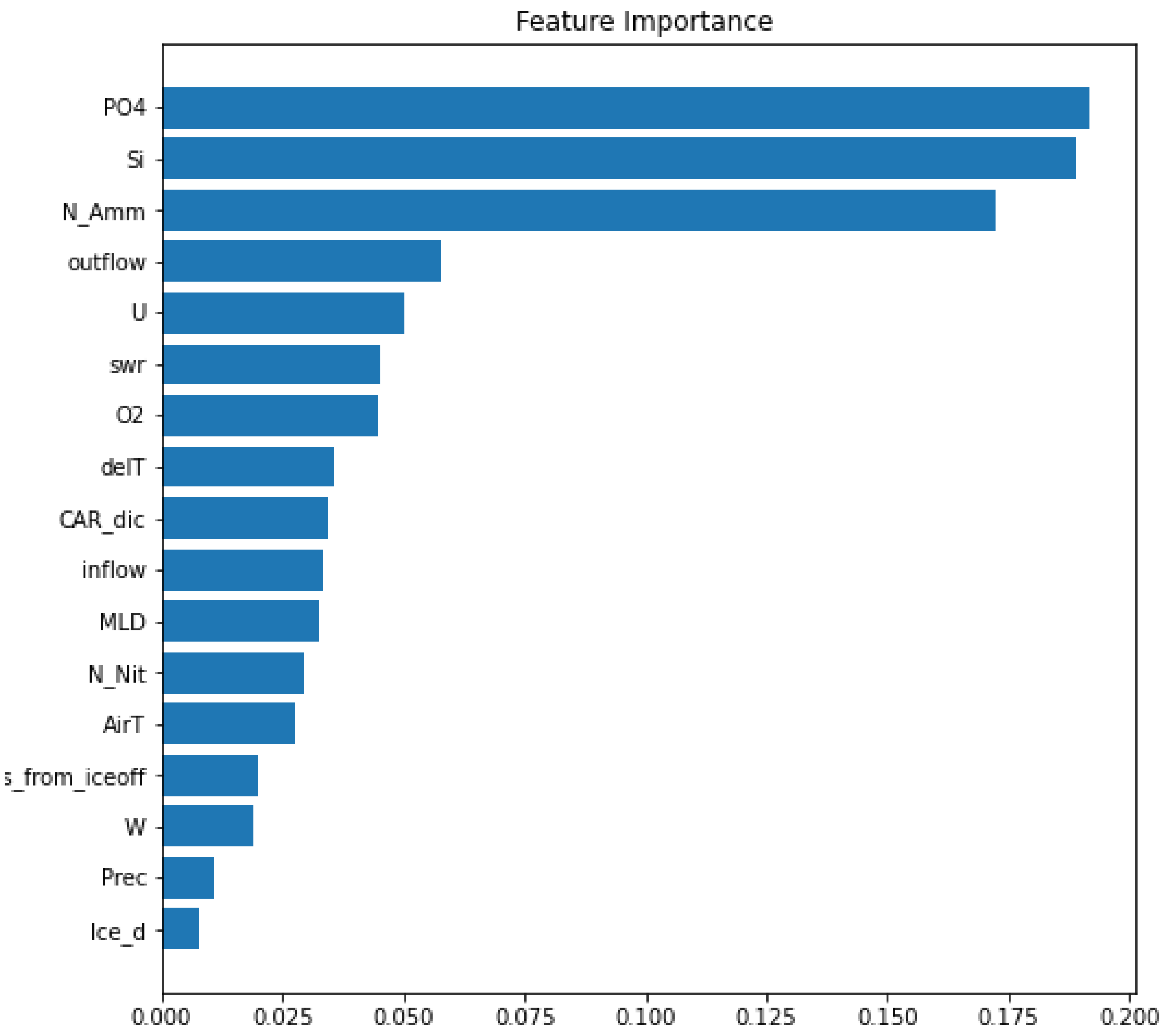
Δ**Chl >0.3**

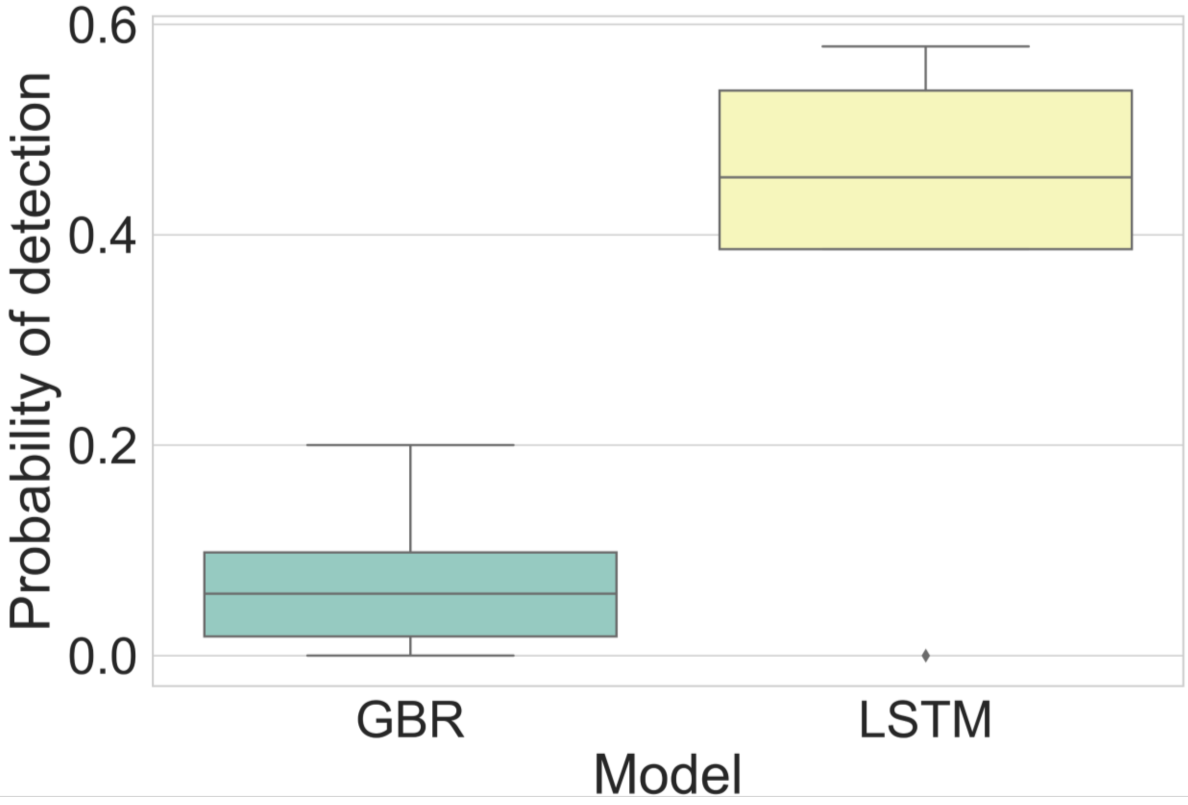
6 lake nutrients: NO3, PO4, O2, Total P, NH4, Si, CAR_dic





	Testing RMSE	Testing R2
GBR	11.46	-0.68
LSTM	11.37	-0.65


Feature Importance
PO4
Si
N_Amm
outflow
U
swr
O2
delT
CAR_dic
inflow
MLD
N_Nit
AirT
s_from_iceoff
W
Prec
Ice_d
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200


Probability of detection
0.6
0.4
0.2
0.0
GBR
LSTM
Model

Preliminary conclusion and discussion

- The two-step approach (pre-generate nutrients)
  - ➢ Partially overcome the limitation of sparse nutrient observations.
  - ➢ More applicable in real-time algal bloom forecast.
  - ➢ Benefit the water quality prediction.
- By adding the features reproduced by process-based model, the performance of machine learning models improved.
- Based on the evaluating metrics (RMSE, R^2, P_d),
  - ➢ LSTM outperforms GBR and process-based model.
  - ➢ LSTM shows less uncertainty.
- The runtime ratio of LSTM and GBR is 18:1.
- The model may be improved by adding external factors to indicate the characteristics of lakes and the abnormal events that lead to sudden increase or decrease of nutrient loads to the lakes.