Supplementary information for Global-scale reconstruction of daily river dissolved oxygen using transformer

Kai Tan[1,2], Jianrong Ma[1,2*],Wei Hu[1,2]

[1] *Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 400714, China*
[2] *Chongqing College, University of Chinese Academy of Sciences, Chongqing 400714, China*

This PDF file includes:
Table S1, S2
Text S1
Figure S1, S2, S3, S4

Table S1. Data sources detailed description

Dataset	**URL**	**Reference**
Global River Water Quality Archive (GRQA)	`https://zenodo.org/records/7056647`	[1]
Global river chemistry database (GLORICH)	`https://doi.pangaea.de/10.1594/PANGAEA.902360`	[2]
GEMStat	`https://gemstat.org/`	[3]
OLIGOTREND	`https://portal.edirepository.org/nis/mapbrowse?packageid=edi.1778.3`	[4]

Continued on next page

*Corresponding Author.
E-mail: majianrong@cigit.ac.cn

Table S1 – continued from previous page

Dataset	URL	Reference
An extensive spatiotemporal water quality dataset covering four decades (1980-2022) in China	`https://figshare.com/articles/dataset/A_water_quality_dataset_in_China/22584742`	[5]
ERA5-Land Daily Aggregated	`https://developers.google.com/earth-engine/datasets/catalog/ECMWF_ERA5_LAND_DAILY_AGGR`	[6]
MCD12Q1.061: MODIS Land Cover Type Yearly Global 500m	`https://lpdaac.usgs.gov/products/mcd12q1v061/`	[7]
HydroATLAS	`https://www.hydrosheds.org/hydroatlas`	[8]
Gridded Population of the World Version 4.11	`https://developers.google.com/earth-engine/datasets/catalog/CIESIN_GPWv411_GPW_UNWPP-Adjusted_Population_Density`	[9]

Table S2. Hyperparameter search space and optimization results for the models

model	window Size	batch Size	d_model	nhead	num_layers	dropout rate
Transformer	120	128	128	8	3	0.1
BiLSTM	60	64	64	–	4	0.2
LSTM	90	128	256	–	6	0.3
GRU	60	128	128	–	3	0.2

Search space: **window size**: (60, 90, 120); **batch size**: (64, 128, 256); **d_model**: dimension of the model (64, 128, 256); **nhead**: number of attention heads in transformer (4, 8, 16); **num_layers**: number of hidden layers (3 - 8); **dropout rate**: (0.1, 0.2, 0.3).

Text S1. The pseudocode for the transformer model architecture.

Algorithm Transformer model

```
function TRANSFORMER(input)
    x ← Linear(input, output_dim = 128)                          ▷ embedding layer
    x ← LayerNorm(x)
    x ← Dropout(x, p = 0.1)

    x ← PositionalEncoding(x)

    for i = 1 to 3 do                                            ▷ num_layers = 3
        attn_output ← MultiHeadAttention(x, x, x)
        attn_output ← Dropout(attn_output, p = 0.1)
        x ← LayerNorm(x + attn_output)

        ff_output ← Linear(x, output_dim = 512)                  ▷ feed forward
network
        ff_output ← Dropout(ff_output, p = 0.1)
        ff_output ← Linear(ff_output, output_dim = 128)
        ff_output ← Dropout(ff_output, p = 0.1)
        x ← LayerNorm(x + ff_output)
    end for

    x ← Linear(x, output_dim = 64)                               ▷ output layer
    x ← Dropout(x, p = 0.1)
    x ← Linear(x, output_dim = 1)
    output ← ReLU(x)
    return output
end function
```

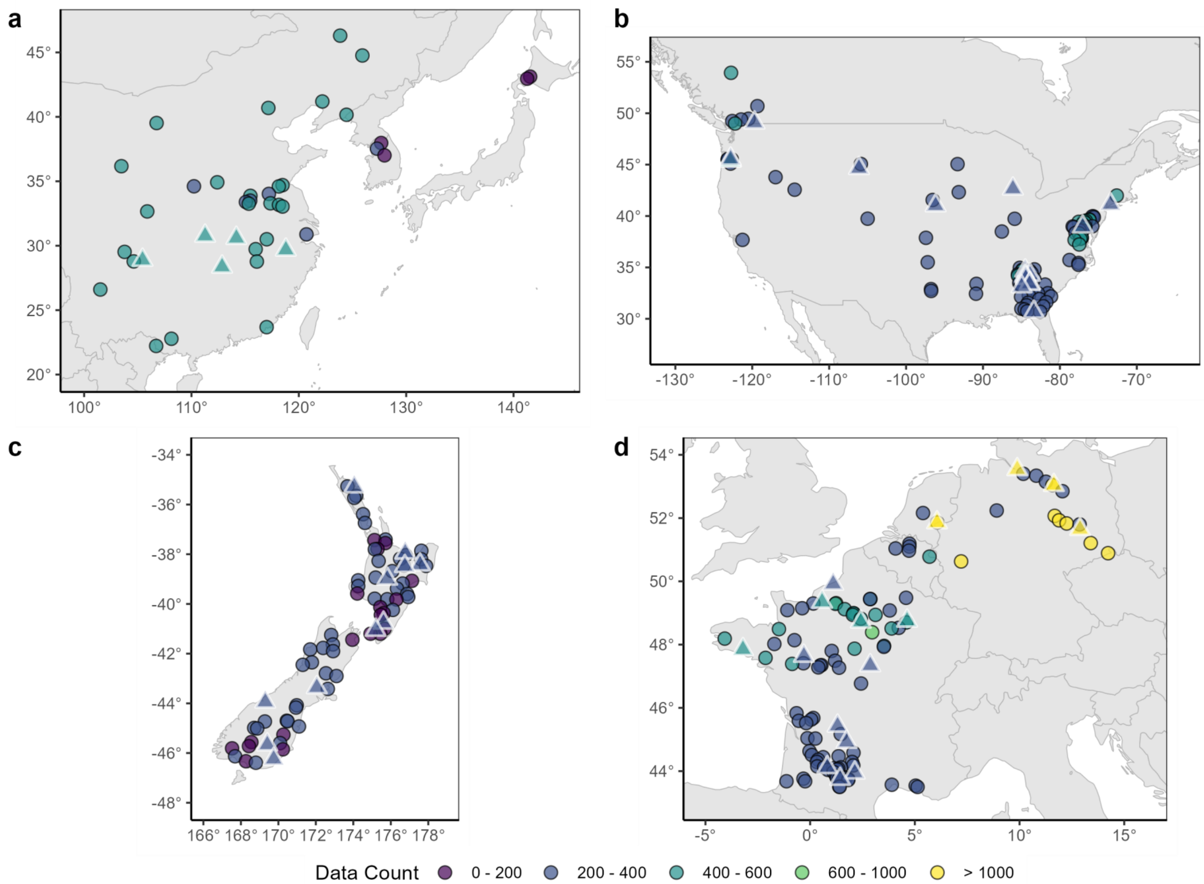

Fig. S1. Enlarged regional maps showing the distribution of monitoring stations in East Asia (EA), North America (NA), New Zealand (NZ), and Western Europe (WE). Circles denote monitoring stations used for training and testing, while triangles indicate hold-out validation stations.

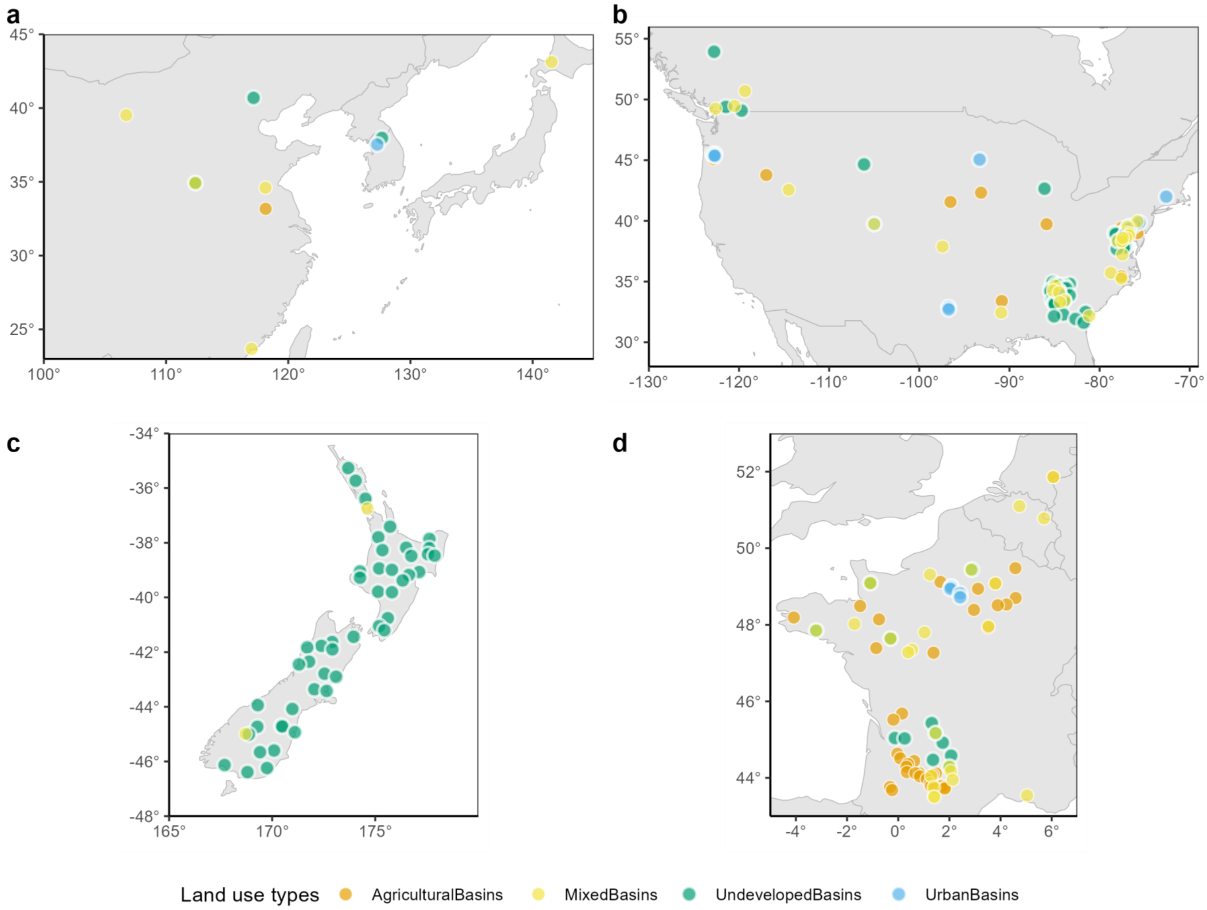

Fig. S2. Spatial distribution of land use types at monitoring stations in in East Asia (EA), North America (NA), New Zealand (NZ), and Western Europe (WE) with Nash-Sutcliffe efficiency (NSE) ≥ 0.5.

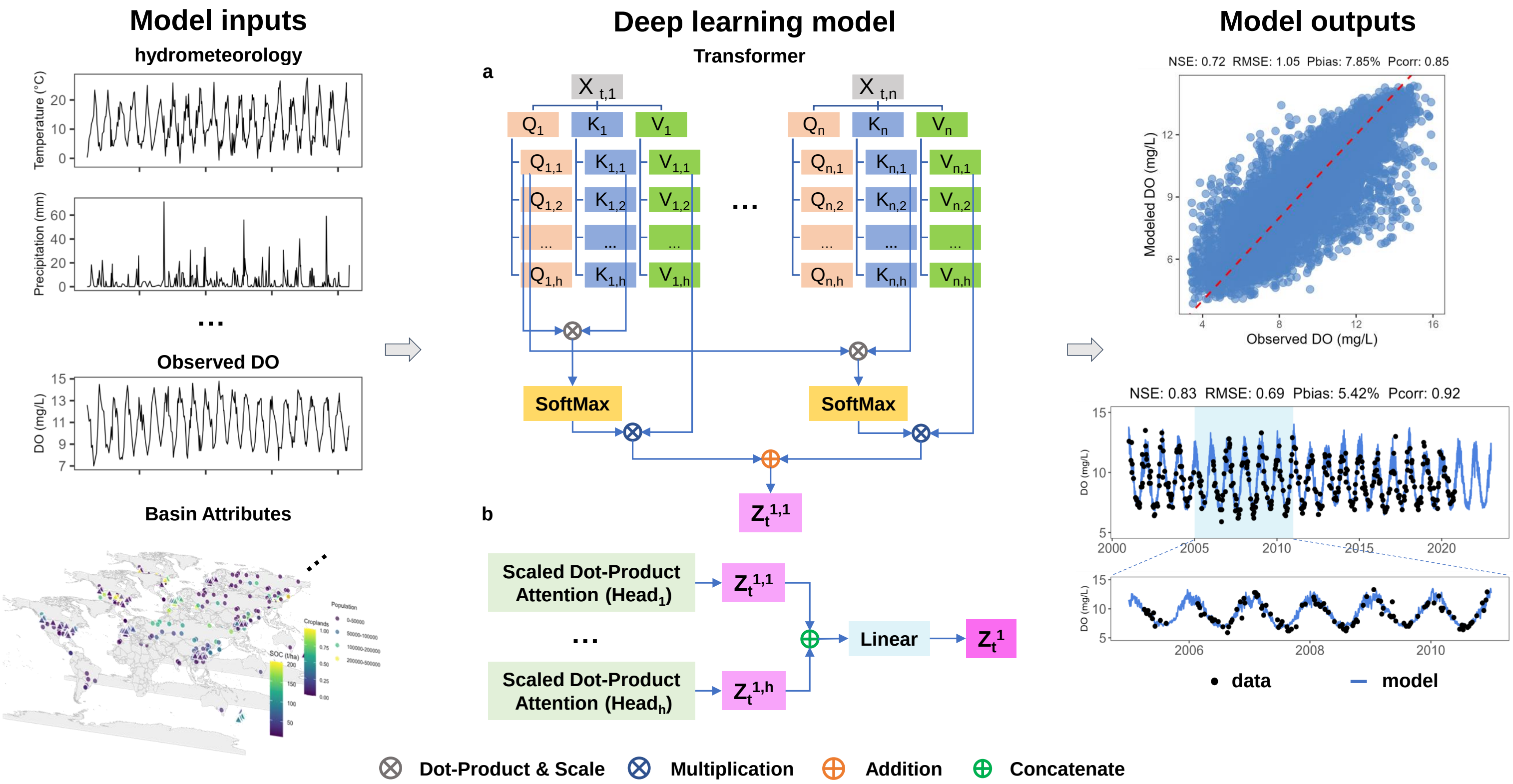

Fig. S3. Workflow of the transformer model for DO prediction. The model integrates hydrometeorological drivers and basin attributes as inputs, processes them through a multi-head self-attention mechanism, and outputs the predicted DO concentrations. At each time step t, input embeddings $X_{t,1}, \ldots, X_{t,n}$ are linearly projected into query (Q), key (K) and value (V) vectors for each of the h attention heads. Scaled dot-product attention outputs from all heads $Z_t^{1,1}, \ldots, Z_t^{1,h}$ are concatenated and passed through a linear layer to produce the final representation Z_t^1. Multi-head self-attention enables the model to capture temporal dependencies and interactions among drivers.

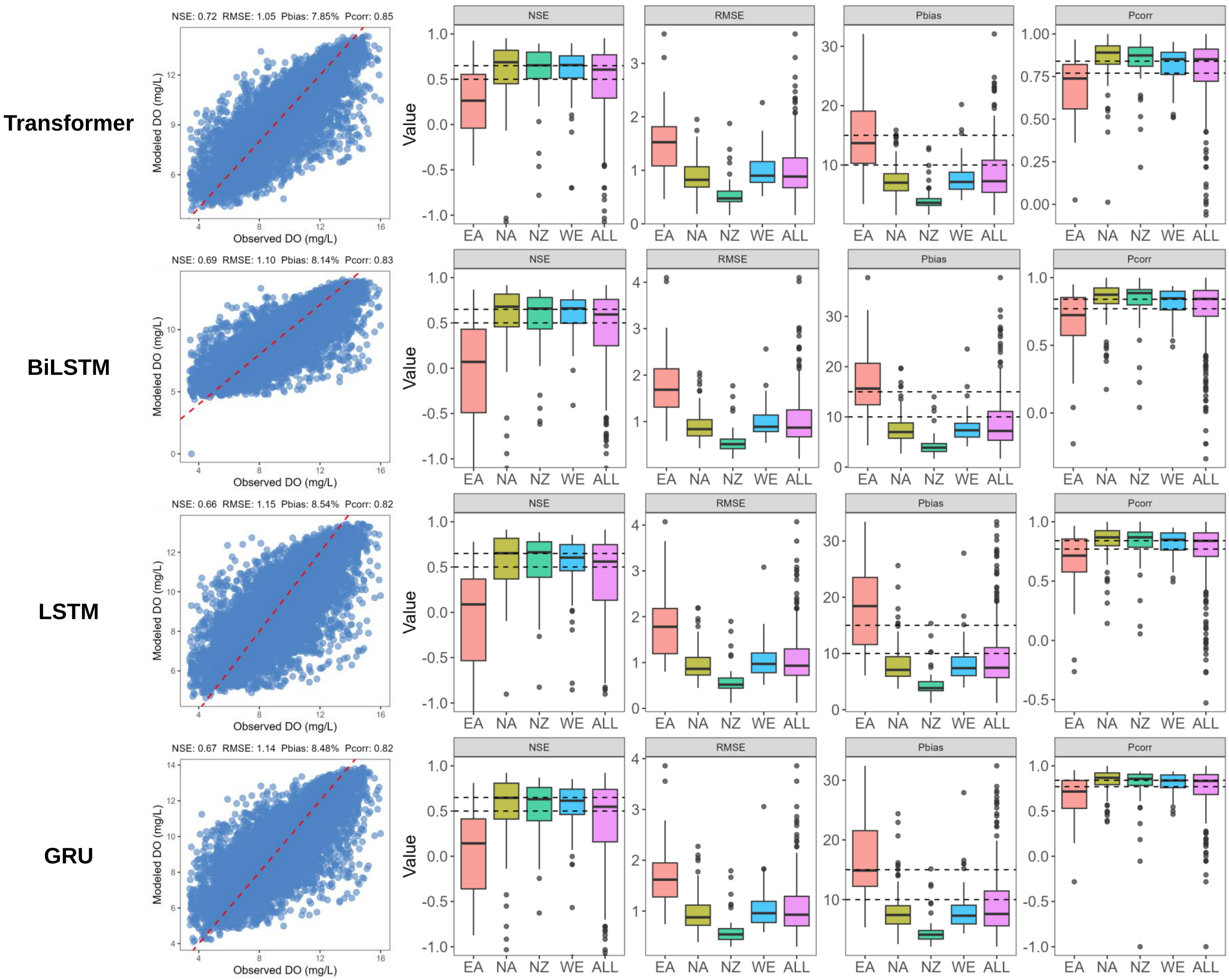

Fig. S4. Benchmark Results. The scatter plot illustrates the model's performance across all test data, without distinguishing between individual monitoring stations. The red dashed line represents the 1:1 identity line, indicating perfect agreement between predicted and observed values. The box plots summarize the performance metrics—NSE, RMSE, Pbias , and Pcorr—during the testing phase across East Asia (EA), North America (NA), New Zealand (NZ), Western Europe (WE), and all stations at the global scale (ALL). Each box plot displays the median (central line) and the interquartile range (IQR), spanning from the first quartile (Q1) to the third quartile (Q3). The whiskers extend to the smallest and largest values within 1.5 times the IQR from Q1 and Q3, respectively. Dashed lines in the box plots indicate thresholds for good and high performance. For clarity, outliers beyond the display range are omitted.

References

[1] Holger Virro, Giuseppe Amatulli, Alexander Kmoch, Longzhu Shen, and Evelyn Uuemaa. Grqa: global river water quality archive. *Earth System Science Data Discussions*, 2021:1–30, 2021.

[2] Jens Hartmann, Ronny Lauerwald, and Nils Moosdorf. Glorich-global river chemistry database. 2019.

[3] Moritz Heinle, Dmytro Lisniak, and Philipp Saile. UNEP GEMS/Water Global Freshwater Quality Archive [data set], 2024.

[4] Camille Minaudo, Andras Abonyi, Carles Alcaraz, Jacob Diamond, Nicholas JK Howden, Michael Rode, Estela Romero, Vincent Thieu, Fred Worrall, Qian Zhang, et al. Oligotrend, towards a global database of multi-decadal chlorophyll-a and water quality timeseries for rivers, lakes and estuaries. *Earth System Science Data Discussions*, 2025:1–28, 2025.

[5] Jingyu Lin, Peng Wang, Jinzhu Wang, Youping Zhou, Xudong Zhou, Pan Yang, Hao Zhang, Yanpeng Cai, and Zhifeng Yang. An extensive spatiotemporal water quality dataset covering four decades (1980–2022) in china. *Earth System Science Data*, 16(2):1137–1149, 2024.

[6] Joaquín Muñoz-Sabater, Emanuel Dutra, Anna Agustí-Panareda, Clément Albergel, Gabriele Arduini, Gianpaolo Balsamo, Souhail Boussetta, Margarita Choulga, Shaun Harrigan, Hans Hersbach, et al. Era5-land: A state-of-the-art global reanalysis dataset for land applications. *Earth system science data*, 13(9):4349–4383, 2021.

[7] M Friedl and Damien Sulla-Menashe. Modis/terra+ aqua land cover type yearly l3 global 500m sin grid v061. *NASA EOSDIS Land Processes DAAC*, 10, 2022.

[8] Simon Linke, Bernhard Lehner, Camille Ouellet Dallaire, Joseph Ariwi, Günther Grill, Mira Anand, Penny Beames, Vicente Burchard-Levine, Sally Maxwell, Hana Moidu, et al. Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution. *Scientific data*, 6(1):283, 2019.

[9] CIESIN. Gridded population of the world, version 4 (gpwv4): Population density adjusted to match 2015 revision un wpp country totals, revision 11, 2018.