From hydrometeorology to water quality: can a deep learning model learn the dynamics of dissolved oxygen at the continental scale?

Wei Zhi[1], Dapeng Feng[1], Wen-Ping Tsai[1], Gary Sterle[2], Adrian Harpold[2], Chaopeng Shen[1], Li Li[1,*]

[1], Department of Civil and Environmental Engineering, The Pennsylvania State University, State College, PA 16802, USA

[2], Department of Natural Resources & Environmental Science, The University of Nevada, Reno, NV, 89557, USA

* Correspondence to lili@engr.psu.edu

Classification: *Physical Sciences, Environmental Sciences*

Keywords: *dissolved oxygen, water quality, deep learning, LSTM, big data, continental-scale model*

Author Contributions

G. S. and A. H. compiled the CAMELS-chem database. C. S. and L. L. conceived the idea. W. Z. performed the research and analysis with help from D. F. and W. T. on model training. W. Z. wrote the first draft of the manuscript, which was heavily revised by LL. C. S. and A. H. helped edit the manuscript.

Abstract: Dissolved oxygen (DO) sustains aquatic life and is an essential water quality measure. Our capabilities of forecasting DO levels, however, remain elusive. Unlike the increasingly intensive earth surface and hydroclimatic data, water quality data often have large temporal gaps and sparse areal coverage. Here we ask the question: can a Long Short-Term Memory (LSTM) deep learning model learn the spatio-temporal dynamics of stream DO from intensive hydroclimatic and sparse DO observations at the continental scale? That is, can the model harvest the power of big hydroclimatic data and use them for water quality forecasting? Here we used data from CAMELS-chem, a new dataset that includes sparse DO concentrations from 236 minimally-disturbed watersheds. The trained model can generally learn the theory of DO solubility under specific temperature, pressure, and salinity conditions. It captures the bulk variability and seasonality of DO and exhibits the potential of forecasting water quality in ungauged basins without training data. It however often misses concentration peaks and troughs where DO level depends on complex biogeochemical processes. The model surprisingly does not perform better where data are more intensive. It performs better in basins with low streamflow variations, low DO variability, high runoff-ratio (> 0.45), and precipitation peaks in winter. This work suggests that more frequent data collection in anticipated DO peak and trough conditions are essential to help overcome the issue of sparse data, an outstanding challenge in the water quality community.

Significance Statement. Sufficient dissolved oxygen (DO) level is essential to maintain a healthy aquatic ecosystem. Yet DO is challenging to forecast because of complex hydroclimatic and biogeochemical conditions that vary spatially and temporally. Here we harvested the power of intensive hydrometeorological data and developed a continental-scale deep-learning model that captured the bulk variability and DO seasonality across poorly gauged and ungauged basins of diverse climate, geology, and vegetation conditions in the United States. The model surprisingly does not perform better in places with most data. Instead, it often performs better in basins with low streamflow and DO variability and high runoff-ratio. The model suggests that more intensive data collection at DO peaks and troughs is needed to capture the full dynamics.

1. Introduction

Dissolved oxygen (DO) sustains aquatic life and is a critical water quality measure (1, 2). Low DO levels arising from eutrophication, or hypoxia, have caused dead zones worldwide (3, 4). DO concentrations also affect nutrient availability and metal mobilization. Low DO conditions favor denitrification and facilitate the release of reactive phosphorus and toxic metals (e.g., As, Cr) from polluted sediments (5-7). The capability of forecasting DO is essential for monitoring aquatic ecosystem health and water management (2, 8, 9). DO levels are controlled by the relative magnitude of source and sinks. DO is supplied by the O_2 dissolution and aquatic photosynthesis (6). DO solubility follows Henry's law and decreases as temperature increases; it also drops as partial pressure of O_2 decreases at high elevation (10) and as salinity increases (11, 12). DO is consumed via aquatic respiration (13), often peaking at summer times when biological activities culminate. Generally speaking, shallow streams with actively running water can mix sufficiently and often have abundant light for photosynthesis, leading to high DO levels compared to deeper and stagnant waters. Streams with rich organic materials are often more depleted in DO due to carbon decomposition (14).

The past decades have witnessed extensive studies of DO in human-affected estuaries and coastal waters with elevated nutrient loading (e.g., agriculture, fertilizer) and organic input (e.g., sewage treatment plants) (3, 6, 15). Water quality forecasting has traditionally used process-based modelling. Irby et al. (2016) (16) evaluated eight process-based, well-calibrated DO models in the Chesapeake Bay and their results .and found that capturing observed spatial variability is much more challenging than capturing temporal variability. Stefan and Fang (1994) (17) developed a regional process-based DO model for seven lakes and obtained the standard errors ranges from 1.2 to 2.3 mg/L (mean = 1.9 mg/L) after calibration at individual sites.

With greater data availability and computational power, deep learning approaches have recently shown promises in applications in a wide range of fields. Deep learning models do not rely on a prior model assumption, are computationally efficient and less scale-dependent than traditional process-based models that solve differential equations (18, 19). The emerging Long Short-Term Memory (LSTM) is a type of recurrent neural network structure that learns directly from sequential data (20, 21). The LSTM can learn to carry useful information from the past input to the next time step for time-series

prediction. Recent work has shown that the LSTM can capture soil moisture dynamics (22-24) and predict streamflow at the regional (25) and continental scale (26). A few studies have used LSTM for DO prediction in individual sites from small aquaculture ponds (27-30) to a lake (31), with RMSE in the range of 0.5 to 1 mg/L. These studies used intensively measured, high frequency (e.g., 15 minutes) water quality data that not only include DO but also other measures such as pH, organic carbon, and nutrients. At a larger scale in the Chesapeake Bay, Ross and Stock (2019) (32) developed a machine learning model for DO and showed RMSE values of 0.5 – 2 mg/L for multiple stations. To the best of our knowledge, no attempt has been made to use LSTM to understand and forecast the DO spatial-temporal dynamics at the continental scale.

The recent decades have seen significant increase in hydroclimatic data, discharge data, and earth surface characteristics data (33). Discharge data are often collected at minutes to daily resolution, capturing detailed nuances of streamwater dynamics. Water quality data, including DO and nutrients, are typically collected at much coarser temporal resolution (e.g., weekly to quarterly sampling frequency). Water quality data often have large temporal gap (e.g., multi-year to decade gap) and relatively small spatial coverage, presenting major challenges for applying machine-driven models and for forecasting water quality. Here we ask the question: Can a uniform machine learning model learn the spatial-temporal dynamics of DO from intensive hydroclimatic data, watershed characteristics, and sparse DO data at the continental scale? Can a model learn enough to forecast DO levels in ungauged basins without any DO measurements? To answer these questions, we used DO records from the CAMELS-chem database for 236 sites across the U.S. and trained a single LSTM model. The CAMELS-chem database is a new nationwide water chemistry database for minimally disturbed streams that build upon the well-known CAMELS hydrology dataset (Catchment Attributes and Meteorology for Large-sample Studies) (34, 35). The dataset includes sparse DO data, time series of climate forcing data (e.g., precipitation, temperature, solar radiation, pressure) and watershed attributes (e.g., topography, land cover, vegetation, geology, and soil) that can be important determinants of DO levels. The DO levels in these minimally-disturbed watersheds also offer contrasts and reference points for extensively-studied coastal and urban areas where human influences are vast and deep. We also examined the conditions under which the LSTM performs well to provide insights on DO process dynamics and to identify strategies for future data collection.

2. Results

2.1 Spatial patterns of DO and model performance

Mean DO concentrations (Figure 1b) generally are higher at higher latitude, with the highest (11 – 12 mg/L) in the Northeast and Northwest. The lowest mean DO occurs in Florida and is surprisingly low (i.e., 4 – 5 mg/L) for these minimally-disturbed watersheds, close to the hypoxia limit of 3 mg/L. DO variations (Figure 1c) are lower in the West compared to the East. For the core evaluation group of 84 sites, the model achieved satisfactory performance (NSE >= 0.4) for 74% sites (Table S1). Specifically, the good performance (NSE >= 0.7, 38% of the 84 evaluated sites) and fair performance (0.4 <= NSE < 0.7, 36%) groups exhibit a mean (median) NSE of 0.77 (0.75) and 0.53 (0.55), respectively. For the whole evaluation group, the model achieves a mean (median) NSE, RMSE, and Pcorr of 0.51 (0.57), 1.2 (1.1) mg/L, and 0.78 (0.82) respectively. Note that even the low NSE performance group achieved a mean (median) Pcorr of 0.61 (0.60), indicating a robust capture of DO seasonality (Table S1).

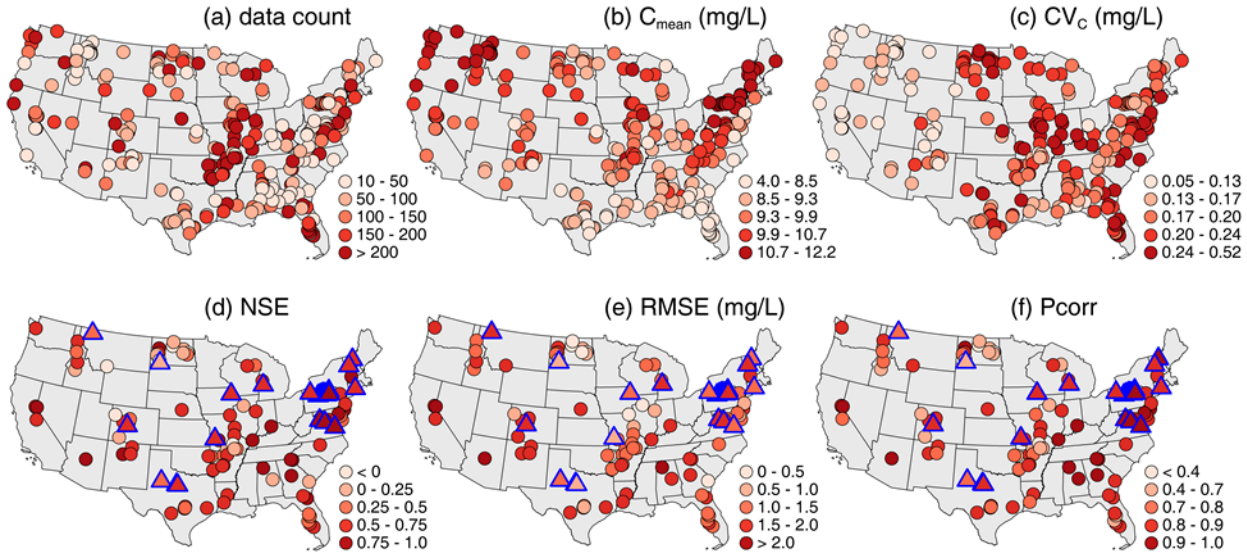

Figure 1. DO records (top panel, 236 sites) from the CAMELS-chem database and model performance (bottom panel, core 84 sites + ungauged 24 sites): (a) data count; (b) mean DO concentration (mg/L); (c) coefficient of variation of DO concentration (mg/L); (d) Nash-Sutcliffe Efficiency (NSE); (e) Root Mean Square Error (RMSE); (f) Pearson's correlation coefficient (Pcorr). Triangles with blue border indicate ungauged basins lacking training data (Figure 2, the last two rows). Darker red color in the performance panel (d, e, f) suggests better performance.

Figure 1d-f shows NSE, RMSE, and Pcorr maps for the core evaluation group (84 sites) and the out-of-training group (i.e., 24 ungauged sites). The Northeast and Eastern regions exhibit the best NSE performance. California and New Mexico also show consistent (>= 3 sites) good NSE performance of 0.71 and 0.70, respectively. For the covered 33 states, only three states show unsatisfying state-averaged performance (NSE < 0.4), i.e., Montana = 0.18, North Dakota = 0.18, Colorado = 0.29. North Dakota has the largest model error of 1.82 mg/L and a low correlation of 0.59; the state-averaged NSE of Montana and Colorado are lowered by one or two sites with lower NSE and larger biases (0.9 – 1.6 mg/L). The mean and median of modeled DO errors quantified by RMSE (Figure 1e) across the 108 sites are 1.2 and 1.1 mg/L, respectively. The model also captures DO seasonality at the continental scale (Figure 1f), with mean and median Pearson' correlation coefficient of 0.80 and 0.83, respectively.

A detailed look further confirms that the model captures the seasonal DO dynamics across diverse conditions (Figure 2), despite large data gaps that may have challenged the training process. In general, the model covers the bulk variability of DO concentrations of 5 - 15 mg/L. In some cases, it misses the DO peaks and troughs. While a few occasional mismatches at extremes may have a limited impact on NSE values in basins with abundant data (Figure 2d, i), they could lead to a large penalty on NSE in basins with less testing data points (e.g., low performance group, Figure 2l, o). Such larger biases at DO extremes could be attributed to several reasons. High and low extremes are often rare and thus are underrepresented in the training process. Second, DO values are instantaneously measured, reflecting temperature conditions at the time of measurements, which may differ substantially from conditions represented by the daily average of temperature, among other meteorological conditions. On the other hand, low DO conditions are often driven by biogeochemical processes, in particular in-stream decomposition of organic carbon. These processes vary spatially and temporally and hinge on local conditions (36, 37) and are challenging to capture (16). For example, Bailey and Ahmadi (2014) (36) revealed that biogeochemical processes (e.g., algal respiration, chemical oxidation) vary substantially with seasonal variations of water temperature and solar radiation, and specific locations within a stream network. These figures also indicate that model performance does not necessarily depend on the number of data points and / or the length of the record, contradicting conclusions from many existing studies. Some

sites with sparse data perform well (Figure 2a); some sites with dense and long records perform poorly (Figure 2k).

For the ungauged basins, the trained model captures the seasonalities with high mean and median Pearson's correlation coefficients of 0.85 and 0.89 (Figure 1f, blue triangles), respectively. The model also captures the bulk variations in DO ranges for most of the time (Figure 2, last two rows) with some occasionally missed peak and troughs. Overall, the model captures the DO trend and achieves an above-the-average performance with mean and median NSE of 0.60 and 0.78, respectively. Out of the 24 ungauged basins, only 3 basins (two of them shown in Figure 2n, o) exhibit low performance (NSE < 0.4).

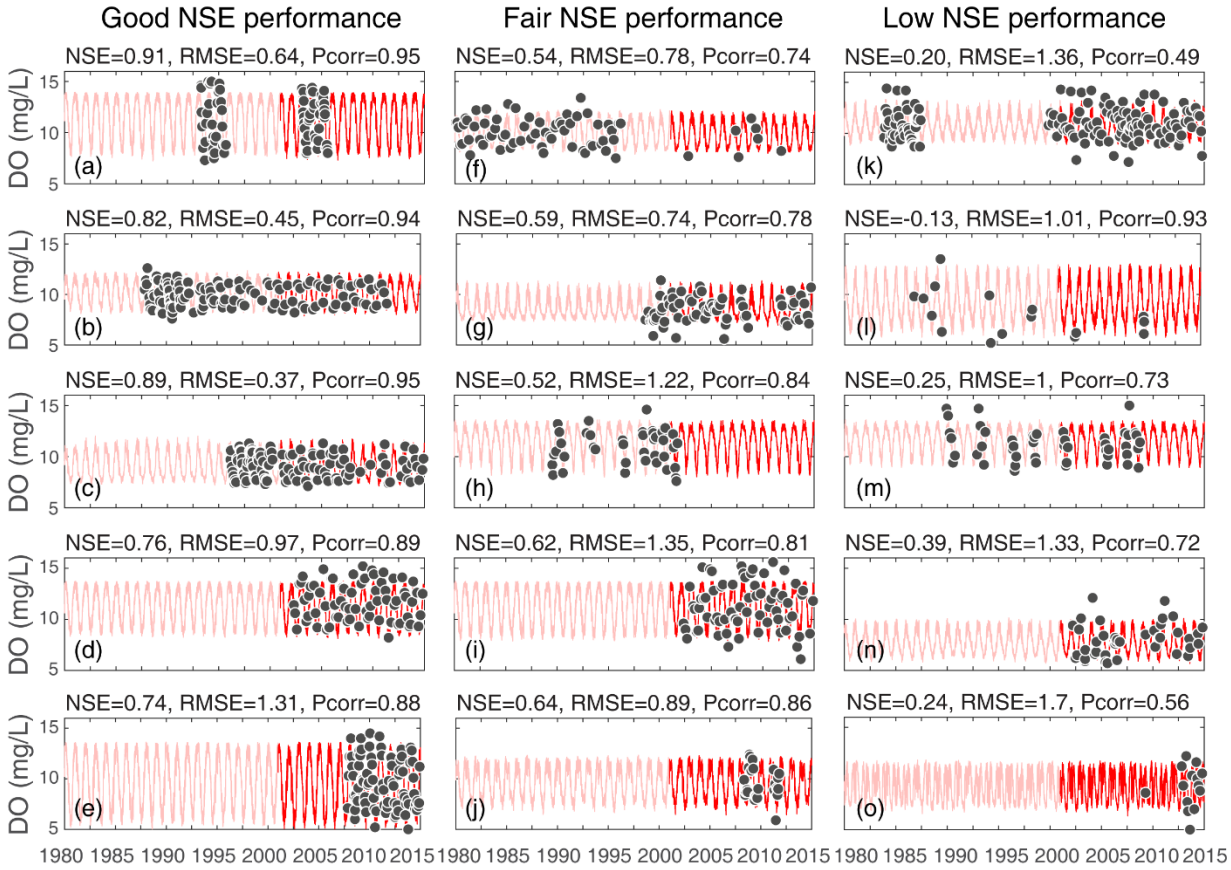

Figure 2. Temporal DO dynamics for randomly selected basins in good NSE (>= 0.7), fair NSE (0.4 – 0.7), and low NSE (< 0.4) performance group. Black dots are measurements; lines are model predictions. The top three rows are core evaluation sites with data in both training (light red line, 1980-2000) and testing period (red line, 2001-2014). The last two rows are ungauged basins with data only in testing period (also in Figure 1).

2.2 Model performance for reproducing C-Q relationships.

Concentration-discharge (C-Q) relationships are often used to understand the response of solute concentration to changing streamflow (38-40) and offer clues about catchment structure and biogeochemical processes (41-43). Shallow pink circles in Figure S2 are DO outputs from the model with measured stream discharge. The figure shows that the good and fair performance occur when the model captures the full range of DO levels (left and middle columns). The sites with low performance are those that missed the peaks and troughs. The figure also indicates that DO measurements occur mostly in low to intermediate-high discharge regimes, covering 70-80% on logQ but often miss the concentrations at high Q regimes that could largely determine C-Q patterns.

2.3 Reproducing Concentration-Temperature relationships

Temperature controls DO solubility and biological activities in water and is often the dominant driver for DO variations. Both data and model output show DO decrease as temperature increases at all sites (Figure 3). The theoretical prediction line of DO solubility was based on the Benson and Krause Equations under the specific altitude (barometric pressure) and salinity conditions at individual sites (44, 45). The model outputs show that for most sites, the model can learn the DO solubility theory and produce at least part of the concentration versus temperature curves (Figure 3), especially under low T conditions. At higher T conditions, DO levels often drop to levels much lower than the solubility prediction (Figure 3, middle row), which is expected. Biological activities such as decomposition of organic carbon are much higher under high T conditions. At some sites, the DO data follow the prediction lines at all T range (Figure 3, top row), indicating minimal biological activities that consume DO. At some other sites, DO can be much higher than the solubility line (Figure 3, bottom row), potentially indicating significant aquatic photosynthesis in streams. The model generally performs better where the DO has a relatively narrow range (i.e., less scattering in black dots, Figure 3a-c). In some sites (Figure 3g), for example, at the same 20 deg C, DO concentrations can vary from 5 to 13 mg/L, indicating significant supply or consumption. The model often does not do well in these sites (Figure 3g-i).

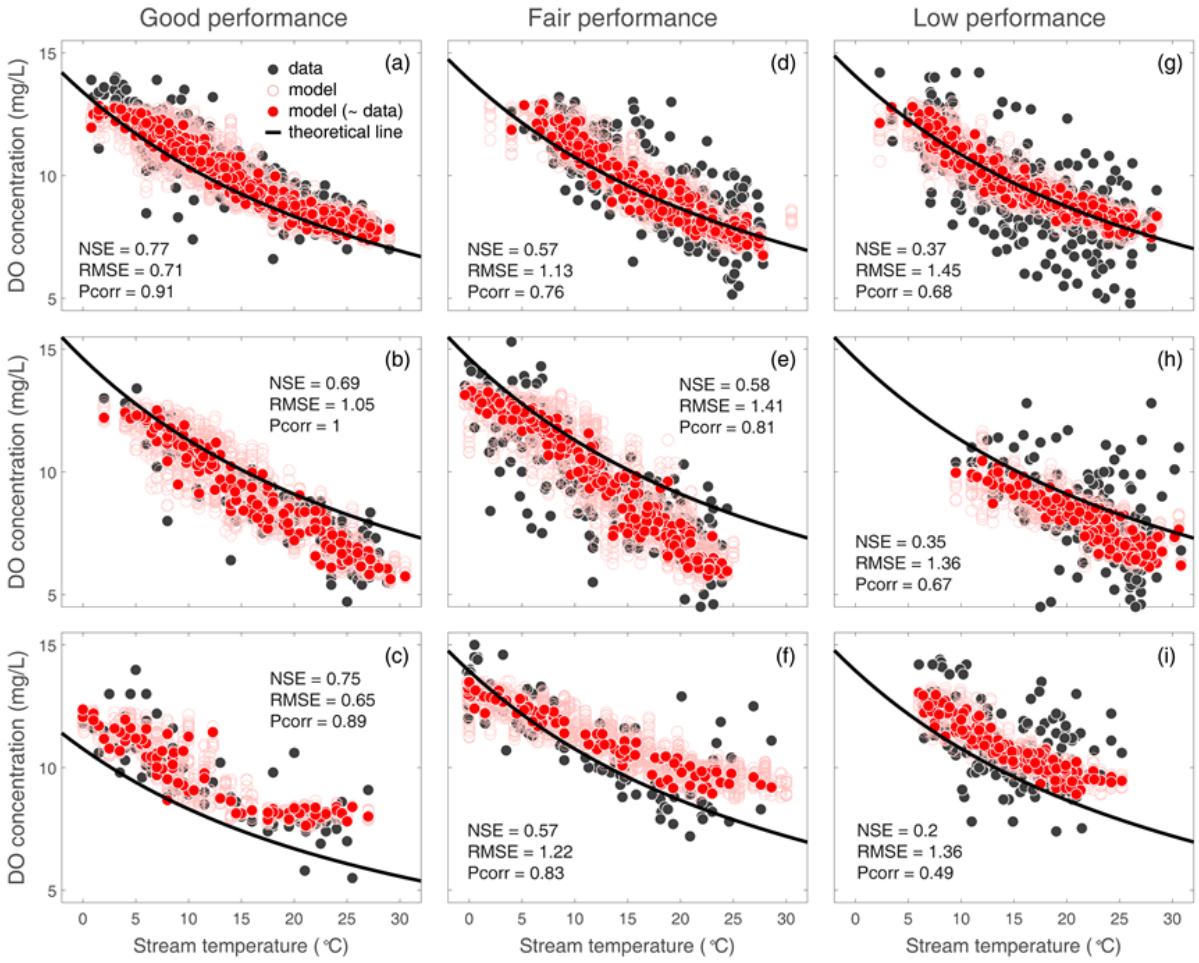

Figure 3. Concentration-temperature relationships for a few selected sites from the good, fair, and low NSE performance group. The black solid line is the theoretical prediction of DO solubility by the Benson and Krause Equations (Benson & Krause Jr, 1980; Benson & Krause Jr, 1984) based on local water temperature, altitude (barometric pressure) and salinity conditions. The specific conductance (SC) that typically ranges from 10 to 1000 um/cm for reference stream (Griffith, 2014) was used for salinity estimation by the equation: $Salinity = 5.572 \times 10^{-4}(SC) + 2.02 \times 10^{-9}(SC)^2$ (American Public Health Association, 2005). The empty light red circles represent the 7-day modeled DO concentrations within the week when the stream water temperature was measured. The filled red circle represents the modeled daily DO concentration on the same date with the measured stream water temperature. The top row includes sites with data and modeled DO falling along theoretical prediction lines, indicating minimum DO consumption process in the stream. The middle row includes sites with lower DO that theoretical DO solubility, especially at higher temperature, indicating significant DO consumption in rivers. The bottom row includes cases with DO higher than theoretical DO solubility, indicating potentially significant aquatic photosynthesis in the stream.

2.4 Controls on model performance

Correlation between predictors and RMSE (as a loss function during the training process) were analyzed to understand important controls on the model performance. Results show RMSE values are positively correlated with the variation of DO concentrations (Figure 4a), i.e., standard deviation (R^2 = 0.52) and coefficient of variation (R^2 = 0.46). This indicates that basins with smaller DO variations have lower RMSE and perform better. The model performance does not correlate to the number of data points (R^2 = 0.03, not shown). RMSE values are negatively correlated with minimum DO values (R^2 = 0.40, Figure 4b), indicating the model does not do well at sites and under conditions of high DO consumption. Model training performance also depends on hydroclimatic characteristics (Figure 4c). The model performs in regions with high runoff-ratios (> 0.45) and negative p_seasonality where precipitation peaks in winter. Lower performance sites co-occurred with positive p_seasonality when precipitation peaks in summer. The low-performance basins in North Dakota (Figure 1d) could be due to their low runoff ratios as they are challenging to model even just for streamflow (26). The training performance was also weakly to moderately (R^2 = 0.19 – 0.24) correlated with other hydrological processes including stream flow and baseflow (Figure S3a, b). This indicates that basins with large variations in streamflow tended to perform not as well, as more seasonal dominance and less short-term transient processes are harder to model. The ungauged basins (i.e., blue triangles) follow the same correlation trends. The capturable dependence of error on these attributes mean that the errors may be predictable with deep-learning-based uncertainty methods (46, 47).

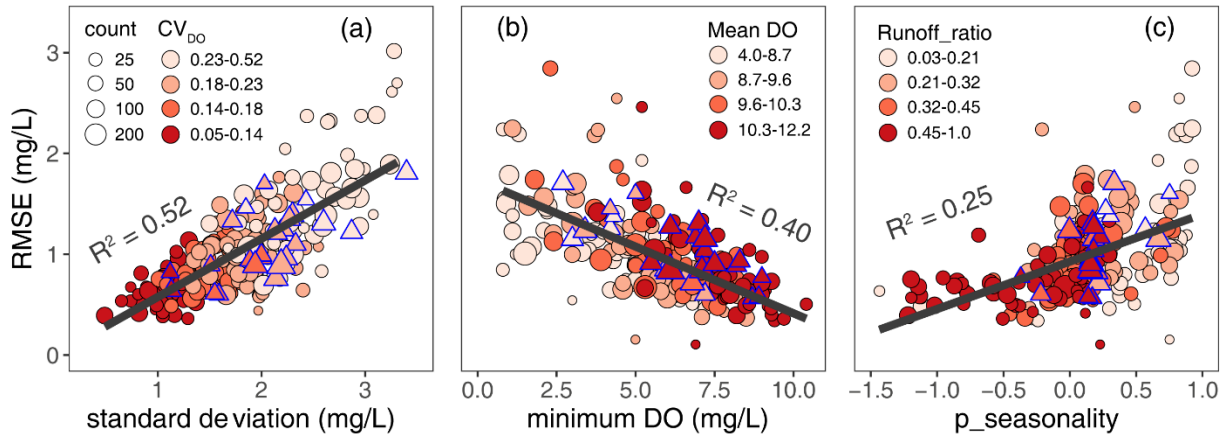

Figure 4. Deep learning model training insights: relationships between model RMSE performance and watershed predictors for all 236 sites. Blue triangles are 24 ungauged basins in Figure 1 (also in Figure 2). The "CV_{DO}" in (a) is the coefficient of variation of DO concentration. The "runoff_ratio" (c) is the ratio of mean daily discharge to mean daily precipitation. The "p_seasonality" is the precipitation seasonality: positive (negative) values indicate that precipitation peaks in summer (winter), values close to 0 indicate similar precipitation throughout the year.

3. Discussion

3.1 Can the model learn the spatio-temporal dynamics of DO from intensive hydrometeorology data and sparse water quality data?

The continental-scale model developed here has good and fair performance for 74% of the basins. Although the low NSE performance group (26% basins) has relatively larger model errors (i.e., 1.7 mg/L RMSE), they still generally agree with observed DO seasonality (i.e., mean Pcorr = 0.61). The largest errors occur at DO peaks and troughs, similar to conclusions from LSTM DO forecasting models developed for individual sites. For example, Irby et al. (2016) (16) found it challenging to reproduce spatial DO variability than seasonality, because the models have difficulty capturing the dynamics of biological drivers. Even with intensive DO and other water quality data, DO modeling errors generally increase with spatial scales, from 0.5 – 0.7 mg/L in individual ponds (27, 31) to 1 – 2 mg/L at the regional scale (e.g., Chesapeake Bay, Twin Cities region) (16, 17, 32). A meta-analysis of well-calibrated water quality models (e.g., N, P) at smaller scales (i.e., pilot to watershed scale) shows the monthly NSE ranges from 0.2 to 0.6 (mean = 0.44) (48). This indicates that the continental scale model here achieved comparable performance with other DO or water quality models for individual sites and regional scale. Note that existing LSTM models for individual sites often have much more data, typically with high frequency DO data at hour to daily frequency other water quality data such as pH, water temperature, specific conductance, carbon (27, 31, 49). This suggests the LSTM model can learn the DO dependence on the hydro-meteorological conditions modulated by site conditions, and forecast DO with confidence only using sparse DO data. This is significant because meteorological and hydrological data are much more available and at much higher frequency than water quality data.

It is somewhat surprising that the training performance was not correlated with sample count (R^2 = 0.03). Instead, the model performance correlates more with DO

variability (Figure 4) and performs better in places with low DO variability without extremely low DO troughs, relatively stable flow conditions (high baseflow index, low streamflow flashness, and high runoff ratio). This is because more DO variations and low DO troughs indicate complex biogeochemical processes that form and decompose organic matter (50).

3.2 Potentials for DO prediction in ungauged and poorly gauged basins

Prediction in ungauged basin for streamflow has been highlighted as a grand challenge for decades (51). Water quality modelling in ungauged basins has generally lagged far behind and received much less attention (52). In fact, 35% of the 236 sites in this work have less than 60 counts for the entire study period of 35 years and the mean (median) count density of the 236 sites is 4 (3) count/year. Such data sparsity in fact makes many of these sites poorly gauged basins for water quality even though they may be well gauged for streamflow. The satisfactory model performance in most of the ungauged basins suggests that the continental-scale model can learn spatio-temporal dynamics across diverse climatic and watershed conditions, and is promising for forecasting DO dynamics in diverse ungauged basins. The model could be used to fill temporal data gaps in many poorly gauged basins as well. It can also be used to improve temporal resolution (e.g., from monthly to daily). The more complete datasets can be used for trend analysis to understand the response of water chemistry and earth surface system to changing environmental and human-induced perturbations such as acid rain, urbanization, and changing climate (53). They can also be used as constraints or predictor (similar to pH) for modelling other chemicals such as nutrients (e.g., denitrification, phosphorus remobilization) or aquatic activity (e.g., fish kill). On the other hand, data gaps beyond a decade probably need to be filled with caution as the influential factors (e.g., climate change, land cover change) for DO dynamics may be changing. GIS analysis of stream network information and near stream vegetation/shading information could probably help with model prediction.

3.3 Implications for water quality measurements

The challenge of sparse data is omnipresent for water quality measures that are costly to collect, analyze, and monitoring (54). Unlike streamflow data that are often measured at minutes to daily time scales, water quality and chemistry observations are typically at monthly to quarterly frequency but at a moment in time. Large data gaps from multiple years to decades are also common. Studies using hydrological models (55) have indicated that, when streamflow data are carefully chosen, 10% of streamflow measurements are sufficient for parameter calibration, meaning that approximately 90% of the data are redundant for parameter identification. The number of data points is often not critical in discriminating between parameter sets. What matters is often the information content and the efficiency with which information is extracted (56, 57). Model uncertainty and data analysis studies are common in the hydrology literature but much rarer in biogeochemistry and water quality literature.

This work shows that current monthly and bimonthly sampling, together with high-frequency hydrometeorological data, is generally sufficient for daily prediction for these reference sites with minimum human disturbance. Although larger data do not necessarily lead to improved performance, results also suggest that DO peak and trough values are critical measures. Increased sampling frequency in seasonal highs and lows is not only useful for model training but also offer opportunities for mechanistic understanding of processes (58). The map (Figure 2) shows DO data is unevenly distributed between the East and the West, where the East has 2.4 times more basins with better NSE performance (East mean: 0.56) than the West (West mean: 0.41). This suggests the DO model could benefit from more data from the West that is relatively underrepresented in the current data.

Results further show that basin performance varies from site to site and largely depend on local DO variability that is regulated by climate, hydrological, and biogeochemical processes (Figure 4). In regions with low runoff ratio and summer precipitation peaks where the model generally has lower performance, more measurements covering the DO variability at full flow regime could be useful. In humid regions (e.g., Pennsylvania, Vermont) where the model already captures the seasonality and bulk variability, sampling campaign could be relaxed in frequency and directed more toward summer time where low DO levels are often expected. Basins with flashy

hydrology tend to have larger model errors (Figure S3). This indicates watersheds with steep slope, highly conductive soil, or thin regolith and shallow water storage (thus flashy streamflow) (59) should be sampled more frequently to alleviate data limitation and cover the full range of more discharge and DO variability. In basins with chemodynamic C-Q patterns (e.g., flushing patterns in Figure S2b, c), more measurements could be planned for hydrological events such as snow-melting floods in spring or large storms in summer, and in the post-melt dry periods where DO can drop low (e.g., typical in western high-mountain regions such as Colorado). This is particularly the case in predicting hypoxia under human influences. Moatar et al. (2020) (60) suggests that optimal sampling frequency largely depends on the flashiness of streamflow and the response of solute concentration variation to streamflow variations. For solutes that are sensitive to streamflow variation such as total suspended solids (TSS), sub-daily frequency is necessary. For solutes that are insensitive to streamflow variations such as total dissolved solutes (TDS), monthly sampling may be sufficient.

Methodology

CAMELS-Chem database

The CAMELS-Chem dataset is built as a supplement to the Catchment Attribute for Large Sample Studies (CAMELS) (34, 35) that includes hydrologic, climatic, and catchment attributes (61). CAMEL-Chem compiles U.S. Geological Survey (USGS) water chemistry and instantaneous discharge from 1980 through 2014 in 493 headwater catchments. It includes many solutes other than DO. Here we used a total of 236 sites with at least 10 records from 1980 to 2014. The drainage areas varied from 5.4 to 25,791 km^2, with mean and median areas of 1,111 and 489 km^2, respectively.

DO data density varies significantly from site to site, with denser and longer records in the East. Data points (Figure 1a) vary from 10 to 842 with the mean and median of 139 and 96, respectively. Although a handful of sites have > 400 records, 16% and 35% sites have less than 30 and 60 data points from 1980 to 2014, respectively. Record length (Figure S1) varies from 0.9 to 98 years, with mean and median of 27 and 27 years, respectively. Data gaps from multi-year to decade are common. Record density varies from 0.25 to 19 count/year, with mean and median of 4 and 3 count/year, respectively, close to an overall frequency of quarterly sampling.

Deep learning algorithm: Long Short-Term Memory network

The Long Short-Term Memory (LSTM) network (20, 21) solves the problem of vanishing gradients in traditional RNNs for time-series tasks (e.g., speech recognition, time-series forecasting) (19). The LSTM layer consists of a set of recurrently connected blocks (i.e., memory cells) to store and pass sequential information. Each LSTM memory cell has three information gates (i.e., input gate, forget gate, and output gate in $Eqn\ 3-5$) and two states (i.e., cell state and hidden state in $Eqn\ 6-7$) to control what to flow in, what to forget, and what to memorize across time steps, allowing the network to learn long-term dependencies (e.g., water storage). The forward pass of the LSTM model is described by the $Eqn\ 1-8$. The LSTM network was implemented in the open source machine learning framework PyTorch (62) while other LSTM details are contained in Feng et al. (2020) (26).

$$Input\ transformation: x^t = ReLU(W_I I^t + b_I) \quad (Eqn\ 1)$$

$$Input\ node: g^t = tanh\big(D(W_{gx}x^t) + D(W_{gh}h^{t-1}) + b_g\big) \quad (Eqn\ 2)$$

$$Input\ gate: i^t = \sigma(D(W_{ix}x^t) + D(W_{ih}h^{t-1}) + b_i) \quad (Eqn\ 3)$$

$$Forget\ gate: f^t = \sigma\big(D(W_{fx}x^t) + D(W_{fh}h^{t-1}) + b_f\big) \quad (Eqn\ 4)$$

$$Output\ gate: o^t = \sigma(D(W_{ox}x^t) + D(W_{oh}h^{t-1}) + b_o) \quad (Eqn\ 5)$$

$$Cell\ state: s^t = g^t \odot i^t + s^{t-1} \odot f^t \quad (Eqn\ 6)$$

$$Hidden\ state: h^t = tanh\ (s^t) \odot o^t \quad (Eqn\ 7)$$

$$Output: y^t = W_{hy}h^t + b_y \quad (Eqn\ 8)$$

where the superscript t represents the time step for time-dependent variables, I^t represents the raw input to the model, x^t is the input vector to the LSTM cell, D is the dropout operator to reduce overfitting, $tanh$ and σ in $Eqn\ 2-5, and\ 7$ are tangent and sigmoidal function, respectively, W and b with different subscripts in $Eqn\ 1-5$ represent the weights and bias parameters, respectively, $\odot$ in $Eqn\ 6-7$ is the element-wise multiplication operator, i^t, f^t, and o^t are the input, forget, and output gates, respectively, g^t is the output of the input node, s^t and h^t are the memory cell state and hidden state, respectively, y^t represents the predicted output at the time step of t.

Model training and evaluation

The DO records were split between 1980-01-01 to 2000-12-01 for training (i.e., 21 years) and 2001-01-01 to 2014-12-31 for testing (i.e., 14 years). This split corresponds to 67% and 33% data for training and testing, respectively. The model was trained using daily time-series of six meteorological features (i.e., precipitation, solar radiation, maximum and minimum air temperature, vapor pressure, day length), with 49 watershed attributes directly imported from CAMELS (e.g., topography, land cover, geology, and soil) (attribute details in Addor et al., 2017 (34)), and one set of air temperature attributes (i.e., T_{avg}, T_{max}, T_{min}) calculated from the Daymet forcing dataset (https://daymet.ornl.gov). With large data gaps and discontinuity in data record in many sites, a core group of 84 sites with at least 6 DO records in both training and testing period was selected for model

performance evaluation. Another out-of-training group of 24 sites with no data points in training periods (1980 - 2000) but has data in testing period (2001 - 2014) were used for evaluating model performance in these as ungauged basins without training data. A total of 89 sites with data only in the training period were not reported for testing performance. The remaining 39 sites with < 6 records in training and testing were not included for performance evaluation due to lower statistical power and potential bias in evaluation metrics. For example, the Nash-Sutcliffe Efficiency (NSE) is sensitive to extreme values due to the squared difference term (Eqn 9) (48). In basins with only a few data points in the testing period, model biases at DO peaks could result in large, overestimated drop in the NSE. This work intends to provide a glimpse of general model performance as well as some across-site understanding on large water-quality dataset, so we did not carry out a systematic optimization on model performance. We tried both daily and monthly (i.e., aggregating daily DO data into monthly average) resolution and they did not differ much in overall performance. The monthly resolution setup took around 20 minutes of computational time for training (i.e., 300 epochs) with an NVIDIA 1080 Ti graphical processing unit (GPU) while the daily setup took around 60 minutes. The choice of 67%-33% data splitting allows more data points in the testing period to be fairly evaluated for more basins (e.g., fewer data points could result in biased NSE values). The model could be further improved by a systematic hyperparameter search and by the inclusion of more basins and data in the training process.

Three statistical metrics were used to measure the model performance: Nash-Sutcliffe Efficiency (NSE), Root Mean Square Error (RMSE), and Pearson's correlation coefficient (Pcorr). NSE ranges from $-\infty$ to 1, with 1 being the perfect match between model prediction and observation. NSE values between 0 and 1 are generally considered as acceptable, whereas NSE < 0 indicates unacceptable performance where model prediction is worse than mean observations (63). Here NSE values >= 0.4 were considered as satisfactory with NSE >= 0.7 and 0.4 <= NSE < 0.7 being good and fair performance, respectively (48) (Table S1). RMSE ranges from 0 to ∞ with lower values indicate better model performance and 0 being the perfect match. Here we found that RMSE < 1.0 mg/L, 1.0 <= RMSE < 1.5 mg/L, RMSE >= 1.5 mg/L are generally associated with good performance (NSE >= 0.7), fair performance (0.4 <= NSE < 0.7), and low performance (NSE < 0.4), respectively. Pcorr ranges from -1 to 1 for perfect negative and

positive correlation, respectively. It is useful for assessing the model capture of seasonality, with values close to 1 indicating better capture of seasonality.

$$NSE = 1 - \frac{\sum_{i=1}^{n} |y_{mod} - y_{obs}|^2}{\sum_{i=1}^{n} (y_{mod} - \overline{y_{obs}})^2} \quad (Eqn\ 9)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_{mod} - y_{obs})^2}{n}} \quad (Eqn\ 10)$$

$$Pcorr = \frac{\sum_{i=1}^{n} (y_{mod} - \overline{y_{mod}})(y_{obs} - \overline{y_{obs}})}{\sqrt{\sum_{i=1}^{n} (y_{mod} - \overline{y_{mod}})^2 \sum_{i=1}^{n} (y_{obs} - \overline{y_{obs}})^2}} \quad (Eqn\ 11)$$

Where y_{mod} and y_{obs} model prediction and observation, respectively, $\overline{y_{mod}}$ and $\overline{y_{obs}}$ are the model predication mean and observation mean, n is the total number of paired model prediction and observation in the testing period.

Data and Code Availability

The CAMELS-Chem DO dataset is deposited at the GitHub repository at https://github.com/ZhiWei2020/CAMELS-Chem-DO-dataset. The hydrometeorological time-series data and watershed attributes are available at the CAMELS data website (https://ral.ucar.edu/solutions/products/camels). The deep-learning LSTM code is also available from the GitHub at https://github.com/mhpi/hydroDL.

Acknowledgements

This study was supported by a seed grant from the Penn State Institute of Computation and Data Science and the U.S. Department of Energy (DOE) Subsurface Biogeochemical Research (SBR) program (DE-SC0016221).

Reference

1. D. Miller, S. Poucher, L. Coiro, Determination of lethal dissolved oxygen levels for selected marine and estuarine fishes, crustaceans, and a bivalve. *Marine Biology* **140**, 287-296 (2002).
2. E. Sánchez *et al.*, Use of the water quality index and dissolved oxygen deficit as simple indicators of watersheds pollution. *Ecological Indicators* **7**, 315-328 (2007).
3. R. J. Diaz, R. Rosenberg, Spreading dead zones and consequences for marine ecosystems. *Science* **321**, 926-929 (2008).
4. N. N. Rabalais, R. E. Turner, W. J. Wiseman Jr, Gulf of Mexico Hypoxia, A.K.A. "The Dead Zone". *Annual Review of Ecology and Systematics* **33**, 235-263 (2002).
5. J. L. Banks, D. J. Ross, M. J. Keough, B. D. Eyre, C. K. Macleod, Measuring hypoxia induced metal release from highly contaminated estuarine sediments during a 40 day laboratory incubation experiment. *Science of the Total Environment* **420**, 229-237 (2012).
6. M. Pena, S. Katsev, T. Oguz, D. Gilbert, Modeling dissolved oxygen dynamics and hypoxia. *Biogeosciences* **7** (2010).
7. S. Wang, X. Jin, Q. Bu, L. Jiao, F. Wu, Effects of dissolved oxygen supply level on phosphorus release from lake sediments. *Colloids and Surfaces A: Physicochemical and Engineering Aspects* **316**, 245-252 (2008).
8. W. Ni, M. Li, A. C. Ross, R. G. Najjar, Large projected decline in dissolved oxygen in a eutrophic estuary due to climate change. *Journal of Geophysical Research: Oceans* **124**, 8271-8289 (2019).
9. A. J. Tesoriero, S. Terziotti, D. B. Abrams, Predicting Redox Conditions in Groundwater at a Regional Scale. *Environmental Science & Technology* **49**, 9657-9664 (2015).
10. J. E. Girard, Principles Of Environmental Chemistry. (2013).
11. H. E. Garcia, L. I. Gordon, Oxygen solubility in seawater: Better fitting equations. *Limnology and oceanography* **37**, 1307-1312 (1992).
12. R. F. Weiss (1970) The solubility of nitrogen, oxygen and argon in water and seawater. in *Deep sea research and oceanographic abstracts* (Elsevier), pp 721-735.
13. B. A. Cox, A review of dissolved oxygen modelling techniques for lowland rivers. *Science of The Total Environment* **314-316**, 303-334 (2003).
14. R. E. Turner, N. N. Rabalais, D. Justic, Gulf of Mexico Hypoxia: Alternate States and a Legacy. *Environmental Science & Technology* **42**, 2323-2327 (2008).
15. W. Kemp, J. M. Testa, D. J. Conley, D. Gilbert, J. D. Hagy, Temporal responses of coastal hypoxia to nutrient loading and physical controls. *Biogeosciences* **6**, 2985-3008 (2009).
16. I. D. Irby *et al.*, Challenges associated with modeling low-oxygen waters in Chesapeake Bay: a multiple model comparison. *Biogeosciences* **13**, 2011 (2016).
17. H. G. Stefan, X. Fang, Dissolved oxygen model for regional lake analysis. *Ecological modelling* **71**, 37-68 (1994).
18. C. Shen, A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists. *Water Resources Research* **54**, 8558-8593 (2018).

19. C. P. Shen *et al.*, HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community. *Hydrology and Earth System Sciences* **22**, 5639-5656 (2018).
20. K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, J. Schmidhuber, LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems* **28**, 2222-2232 (2016).
21. S. Hochreiter, J. Schmidhuber, Long short-term memory. *Neural computation* **9**, 1735-1780 (1997).
22. K. Fang, C. Shen, Near-Real-Time Forecast of Satellite-Based Soil Moisture Using Long Short-Term Memory with an Adaptive Data Integration Kernel. *Journal of Hydrometeorology* **21**, 399-413 (2020).
23. K. Fang, M. Pan, C. P. Shen, The Value of SMAP for Long-Term Soil Moisture Estimation With the Help of Deep Learning. *Ieee T Geosci Remote* **57**, 2221-2233 (2019).
24. K. Fang, C. Shen, D. Kifer, X. Yang, Prolongation of SMAP to Spatiotemporally Seamless Coverage of Continental U.S. Using a Deep Learning Neural Network. *Geophysical Research Letters* **44**, 11,030-011,039 (2017).
25. Z. Xiang, J. Yan, I. Demir, A rainfall‐runoff model with LSTM‐based sequence‐to‐sequence learning. *Water resources research* **56**, e2019WR025326 (2020).
26. D. Feng, K. Fang, C. Shen, Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. *Water Resources Research* 10.1029/2019WR026793 (2020).
27. J. J. Dabrowski, A. Rahman, A. George (2018) Prediction of dissolved oxygen from pH and water temperature in aquaculture prawn ponds. in *Proceedings of the Australasian joint conference on artificial intelligence-workshops*, pp 2-6.
28. Z. Hu *et al.*, A water quality prediction method based on the deep LSTM network considering correlation in smart mariculture. *Sensors* **19**, 1420 (2019).
29. L. Michieletto, B. Ouyang, P. S. Wills (2020) Investigation of water quality using transfer learning, phased LSTM and correntropy loss. in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, p 113950P.
30. W. Li *et al.*, Prediction of dissolved oxygen in a fishery pond based on gated recurrent unit (GRU). *Information Processing in Agriculture* https://doi.org/10.1016/j.inpa.2020.02.002 (2020).
31. R. Barzegar, M. T. Aalami, J. Adamowski, Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model. *Stochastic Environmental Research and Risk Assessment* **34**, 415-433 (2020).
32. A. C. Ross, C. A. Stock, An assessment of the predictability of column minimum dissolved oxygen concentrations in Chesapeake Bay using a machine learning model. *Estuarine, Coastal and Shelf Science* **221**, 53-65 (2019).
33. M. F. McCabe *et al.*, The future of Earth observation in hydrology. *Hydrology and Earth System Sciences* **21**, 3879 (2017).
34. N. Addor, A. J. Newman, N. Mizukami, M. P. Clark, The CAMELS data set: catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences (HESS)* **21**, 5293-5313 (2017).
35. A. Newman *et al.*, Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and

assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences* **19**, 209 (2015).
36. R. T. Bailey, M. Ahmadi, Spatial and temporal variability of in-stream water quality parameter influence on dissolved oxygen and nitrate within a regional stream network. *Ecological Modelling* **277**, 87-96 (2014).
37. O. Langman, P. Hanson, S. Carpenter, Y. Hu, Control of dissolved oxygen in northern temperate lakes over scales ranging from minutes to days. *Aquatic Biology* **9**, 193-202 (2010).
38. S. E. Godsey, J. W. Kirchner, D. W. Clow, Concentration-discharge relationships reflect chemostatic characteristics of US catchments. *Hydrological Processes* **23**, 1844-1864 (2009).
39. W. Zhi *et al.*, Distinct Source Water Chemistry Shapes Contrasting Concentration-Discharge Patterns. *Water Resources Research* **55**, 4233-4251 (2019).
40. W. Zhi, L. Li, The shallow and deep hypothesis: subsurface chemical contrasts shape nitrate export patterns from different land uses. *Environmental Science & Technology* 10.1021/acs.est.0c01340 (2020).
41. S. E. Godsey, J. Hartmann, J. W. Kirchner, Catchment chemostasis revisited: Water quality responds differently to variations in weather and climate. *Hydrological Processes* **33**, 3056-3069 (2019).
42. B. W. Abbott *et al.*, Unexpected spatial stability of water chemistry in headwater stream networks. *Ecology Letters* **21**, 296-308 (2018).
43. F. Moatar, B. W. Abbott, C. Minaudo, F. Curie, G. Pinay, Elemental properties, hydrology, and biology interact to shape concentration‐discharge curves for carbon, nutrients, sediment, and major ions. *Water Resources Research* **53**, 1270-1287 (2017).
44. B. B. Benson, D. Krause Jr, The concentration and isotopic fractionation of gases dissolved in freshwater in equilibrium with the atmosphere. 1. Oxygen. *Limnology and Oceanography* **25**, 662-671 (1980).
45. B. B. Benson, D. Krause Jr, The concentration and isotopic fractionation of oxygen dissolved in freshwater and seawater in equilibrium with the atmosphere. *Limnology and oceanography* **29**, 620-632 (1984).
46. K. Fang, D. Kifer, K. Lawson, C. Shen, *Evaluating the potential and challenges of an uncertainty quantification method for long short-term memory models for soil moisture predictions* (2020), doi:10.1002/essoar.10503330.1, pp. 80.
47. A. Kendall, Y. Gal (2017) What uncertainties do we need in bayesian deep learning for computer vision? in *Advances in neural information processing systems*, pp 5574-5584.
48. D. N. Moriasi, M. W. Gitau, N. Pai, P. Daggupati, Hydrologic and water quality models: Performance measures and evaluation criteria. *T Asabe* **58**, 1763-1785 (2015).
49. P. Liu, J. Wang, A. K. Sangaiah, Y. Xie, X. Yin, Analysis and Prediction of Water Quality Using LSTM Deep Neural Networks in IoT Environment. *Sustainability* **11**, 2058 (2019).
50. E. V. Yakushev *et al.*, Analysis of the water column oxic/anoxic interface in the Black and Baltic seas with a numerical model. *Marine Chemistry* **107**, 388-410 (2007).

51. M. Sivapalan *et al.*, IAHS Decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences. *Hydrological sciences journal* **48**, 857-880 (2003).
52. J. Strömqvist, B. Arheimer, J. Dahné, C. Donnelly, G. Lindström, Water and nutrient predictions in ungauged basins: set-up and evaluation of a model at the national scale. *Hydrological Sciences Journal* **57**, 229-247 (2012).
53. S. S. Kaushal, A. J. Gold, S. Bernal, J. L. Tank, Diverse water quality responses to extreme climate events: an introduction. *Biogeochemistry* **141**, 273-279 (2018).
54. C. R. Levine *et al.*, Evaluating the efficiency of environmental monitoring programs. *Ecological Indicators* **39**, 94-101 (2014).
55. J. A. Vrugt, W. Bouten, H. V. Gupta, S. Sorooshian, Toward improved identifiability of hydrologic model parameters: The information content of experimental data. *Water Resources Research* **38**, 48-41-48-13 (2002).
56. V. K. Gupta, S. Sorooshian, The relationship between data and the precision of parameter estimates of hydrologic models. *Journal of Hydrology* **81**, 57-77 (1985).
57. S. Sorooshian, V. K. Gupta, Automatic calibration of conceptual rainfall‐runoff models: The question of parameter observability and uniqueness. *Water Resources Research* **19**, 260-268 (1983).
58. D. A. Burns *et al.*, Monitoring the riverine pulse: Applying high-frequency nitrate data to advance integrative understanding of biogeochemical and hydrological processes. *WIREs Water* **6**, e1348 (2019).
59. J. Ackerer *et al.*, Determining how Critical Zone structure constrains hydrogeochemical behavior of watersheds: learning from an elevation gradient in California's Sierra Nevada. *Frontiers in Water* **2**, 23 (2020).
60. F. Moatar *et al.*, Stream Solutes and Particulates Export Regimes: A New Framework to Optimize Their Monitoring. *Frontiers in Ecology and Evolution* **7** (2020).
61. G. Sterle *et al.*, Augmenting CAMELS (Catchment Attributes and Meteorology for Large-sample Studies) with Atmospheric and Stream Water Chemistry Data. *In preparation* (2020).
62. A. Paszke *et al.* (2019) Pytorch: An imperative style, high-performance deep learning library. in *Advances in neural information processing systems*, pp 8026-8037.
63. D. N. Moriasi *et al.*, Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *T Asabe* **50**, 885-900 (2007).

Supplementary Information for

From hydrometeorology to water quality: can a deep learning model learn the dynamics of dissolved oxygen at the continental scale?

Wei Zhi[1], Dapeng Feng[1], Wen-Ping Tsai[1], Gary Sterle[2], Adrian Harpold[2], Chaopeng Shen[1], Li Li[1,*]

[1], Department of Civil and Environmental Engineering, The Pennsylvania State University, State College, PA 16802, USA

[2], Department of Natural Resources & Environmental Science, The University of Nevada, Reno, NV, 89557, USA

* Correspondence to lili@engr.psu.edu

This PDF file includes:

Figures S1 to S3

Tables S1

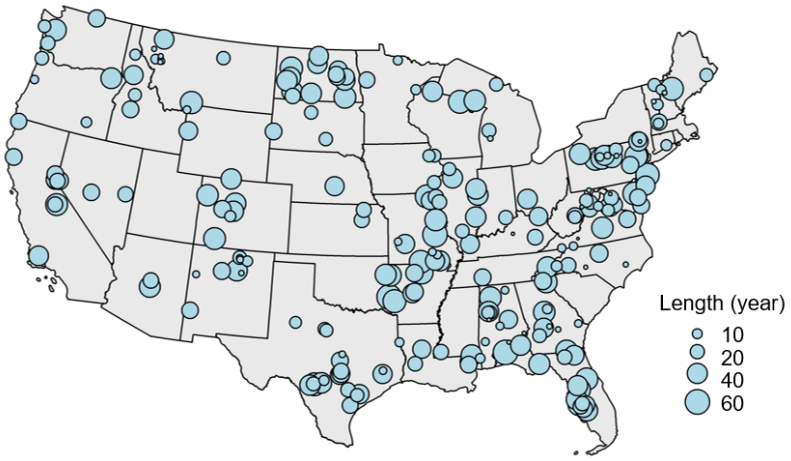

Figure S1. DO record length (year) of 236 sites (>= 10 records) from the CAMELS-chem database.

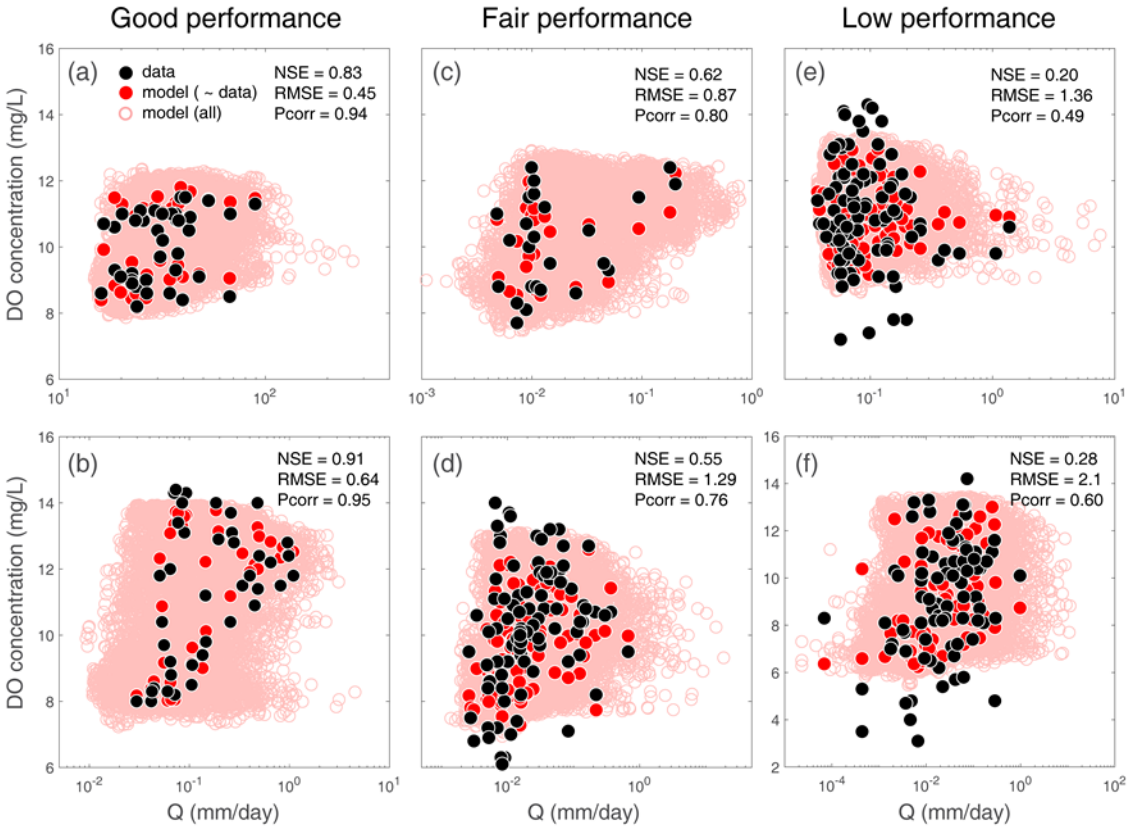

Figure S2. Model performance in reproducing concentration-discharge (C-Q) relationships for a few selected sites from the good, fair, and low performance group. Note only DO concentrations (not Q) were modeled. Empty pink circles are all modeled daily DO concentration during the entire study period (1980 to 2014) while solid black and red circles are measurements and corresponding modeled values, respectively.

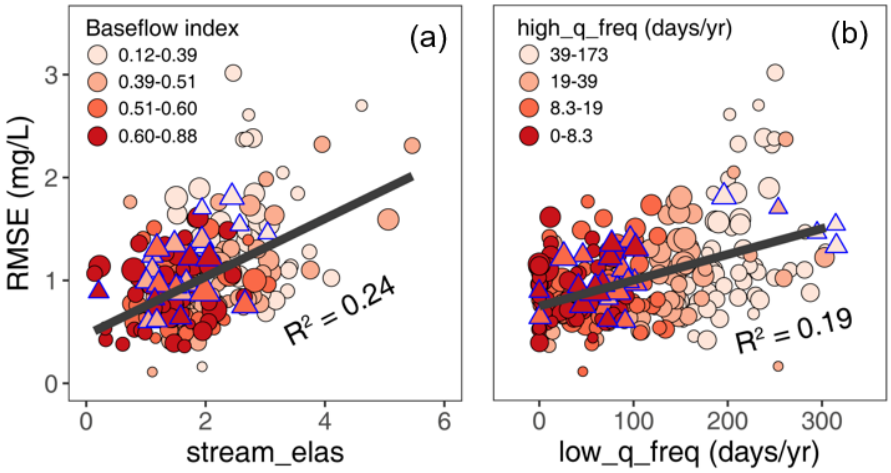

Figure S3. Deep learning model training insights: relationships between model RMSE performance and watershed predictors for all 236 sites. Blue triangles are 24 ungauged basins in Figure 1 (also in Figure 2). The “stream_elas” (a), or streamflow elasticity, is the sensitivity of streamflow to changes in precipitation at the annual time scale; “baseflow_index” is the ratio of mean daily baseflow to mean daily discharge with larger values indicate higher fraction of baseflow to the stream. The “low_q_freq” and “high_q_freq” in (b) mean frequency of low-flow days (< 0.2 times the mean daily flow) and high-flow days (> 9 times the median daily flow), respectively.

Table S1. Model testing performance for the core evaluation group (84 sites)

Metric	Good performance (NSE >= 0.7)			Fair performance (NSE ~ 0.4-0.7)			Low performance (NSE < 0.4)		
	site (%)	range	mean (median)	site (%)	range	mean (median)	site (%)	range	mean (median)
NSE	32 (38%)	0.70 - 0.93	0.77 (0.75)	30 (36%)	0.4 – 0.65	0.53 (0.55)	22 (26%)	-0.5 – 0.29	0.08 (0.17)
RMSE		0.37 – 1.7	0.87 (0.78)		0.64 – 2.0	1.2 (1.2)		0.79 – 2.9	1.7 (1.7)
Pcorr		0.82 – 0.99	0.90 (0.89)		0.67 – 0.93	0.78 (0.80)		0.18 – 0.90	0.61 (0.60)