LIMNOLOGY
and
OCEANOGRAPHY: METHODS

ASLO

Limnol. Oceanogr.: Methods 17, 2019, 97–111

doi: 10.1002/lom3.10302

Improving estimates and forecasts of lake carbon dynamics using data assimilation

Jacob A. Zwart,[1]* **Oleksandra Hararuk**,[2,3] **Yves T. Prairie**,[4] **Stuart E. Jones**,[5] **Christopher T. Solomon**[3]

[1]Integrated Information Dissemination Division, United States Geological Survey, Middleton, Wisconsin
[2]Department of Natural Resource Sciences, McGill University, Montreal, Quebec, Canada
[3]Cary Institute of Ecosystem Studies, Millbrook, New York
[4]Département des Sciences Biologiques, Université du Québec à Montréal, Montreal, Québec, Canada
[5]Department of Biological Sciences, University of Notre Dame, Notre Dame, Indiana

Abstract

Lakes are biogeochemical hotspots on the landscape, contributing significantly to the global carbon cycle despite their small areal coverage. Observations and models of lake carbon pools and fluxes are rarely explicitly combined through data assimilation despite successful use of this technique in other fields. Data assimilation adds value to both observations and models by constraining models with observations of the system and by leveraging knowledge of the system formalized by the model to objectively fill observation gaps. In this article, we highlight the utility of data assimilation in lake carbon cycling research by using the ensemble Kalman filter to combine simple lake carbon models with observations of lake carbon pools and fluxes. We demonstrate that data assimilation helps reduce uncertainty in estimates of lake carbon pools and fluxes and more accurately estimate the true carbon pool size compared to estimates derived from observations alone. Data assimilation techniques should be embraced as valuable tools for lake biogeochemists interested in learning about ecosystem dynamics and forecasting ecosystem states and processes.

Lakes are areas of intense carbon (C) processing. Current estimates of C exported annually from terrestrial ecosystems to inland waters are on par with annual global land net ecosystem production (Randerson et al. 2002; Drake et al. 2017). Nearly 50% of this C is transferred to the atmosphere and about 20% is buried, forming a sediment pool that is now larger than the remainder of the terrestrial biosphere (e.g., land plants and soils; Tranvik et al. 2009; Cole 2013). Clearly, lakes play an integral role in global C cycling, and it is important to understand the drivers of magnitudes and variability of C pools and fluxes.

Observations of lake C pools and fluxes have fundamentally advanced our understanding of lake C cycling. For example, Cole et al. (1994) demonstrated that a vast majority of lakes are supersaturated with CO_2, contributing significantly to regional C cycles as net sources of C to the atmosphere. Long-term data have revealed that dissolved organic carbon (DOC) concentration has been increasing in numerous lakes, initiating a wave of research on the impacts of elevated DOC on lake ecosystem functioning (Monteith et al. 2007). Additionally, through advances in sensor technology, observations of lake metabolic processes have been shown to be meaningfully heterogeneous both within and across lakes (Coloso et al. 2008; Van de Bogert et al. 2012; Solomon et al. 2013; Obrador et al. 2014; Giling et al. 2017).

Models are useful for exploring the implications suggested by observations despite that they simplify complexities of reality and focus on key processes regulating system dynamics. Akin to the observational studies mentioned above, several models have also advanced our understanding of lake C cycling. For example, a scaling study demonstrated that lakes are an important component of the global C cycle (Cole et al. 2007), justifying the inclusion of lakes in the Intergovernmental Panel on Climate Change's Fifth Assessment Report on global C budget (IPCC 2013). A dynamical modeling study demonstrated that allochthonous sources of C can support a large portion of secondary production in lakes through utilization of low-molecular–weight compounds (Berggren et al. 2010). Additionally, a study using first principles of physical limnology showed that gas exchange between lakes and the

*Correspondence: jayzlimno@gmail.com

Additional Supporting Information may be found in the online version of this article.

Author contribution Statement: CTS and YTP designed the study; JAZ collected and analyzed the data with input on methodologies from CTS and SEJ; JAZ developed the model with significant input from OH, CTS, and SEJ; JAZ wrote the first draft of the manuscript; and all authors contributed to the final version.

atmosphere was dominated by convective mixing in small lakes and wind shear mixing in larger lakes (Read et al. 2012).

Despite the significant advances in lake C cycling research highlighted above, both observations and models have weaknesses that can hinder their utility. Observations are snapshots of dynamic pools and processes, which often require gap filling in order to be appropriately scaled across space and time. For example, infrequent observations of lake CO_2 fluxes or concentrations may miss important periods of CO_2 emissions, such as during or after extreme precipitation events (Ojala et al. 2011; Vachon and del Giorgio 2014). Models are simplified representations of reality and can produce errors due to misrepresentation or omission of important processes and uncertainties in parameterization owing to data limitations. For instance, the choice of gas flux model formulation can substantially affect the estimates of lake metabolic balances and CO_2 emissions when models are not constrained with observations (Dugan et al. 2016).

Data assimilation is a framework used to overcome the limitations of models and observations while also capitalizing on their strengths. Data assimilation adds value to models by informing model errors and parameters with observations of the system and adds value to observations by leveraging knowledge of the system contained in the model to fill information gaps. Sequential data assimilation techniques, like ensemble Kalman filter (EnKF), iteratively incorporate information from the observations into our understanding of the system, updating model states and parameter estimates, which have been calibrated with older observations, as new observations are collected and assimilated. Unlike traditional model calibration, data assimilation requires explicit accounting of model and observation uncertainty so that uncertainty can be propagated into the forecasting step. Iteratively informing models with observations using data assimilation techniques has improved understanding of a system and facilitated forecasting its states in fields such as meteorology and terrestrial ecology (Dee et al. 2011; Niu et al. 2014). However, data assimilation techniques are rarely used in aquatic biogeochemical research, which, given the amount of observations and hypotheses about controls of C dynamics, suggests a missed opportunity to combine complementary information from observations and models of these important ecosystems.

In this article, we demonstrate benefits of data assimilation for aquatic C cycling research and show how data assimilation can reveal important ecosystem processes that are not evident from observations or models alone. We use the EnKF, a sequential data assimilation technique often used in meteorology and hydrology, to combine simple lake C process models with real and synthetic observations of lake C pools and fluxes. Using real observations, we show that the EnKF can estimate relevant ecosystem parameters that are hard to measure or were not measured at high temporal frequency (e.g., turnover rate of DOC) and that it significantly reduces uncertainty in lake C pools and fluxes compared to observation uncertainty. Using synthetic observations, we also show that the EnKF technique estimates the true state of the system more accurately than estimates derived solely from observations, even when the model structure is a simplification of the true process. By highlighting these benefits, we encourage lake biogeochemists to take advantage of data assimilation techniques like the EnKF; as such, we explain in detail this data assimilation technique (EnKF) and provide open-source R code of the model in order to reduce analytic barriers.

Materials and procedures

Dataset description

We used both synthetic and real observations of lake C pools and fluxes to demonstrate the value of data assimilation for aquatic biogeochemical studies. The observations of lake C pools and fluxes assimilated into our lake C process models were from East Long Lake during the open-water period in 2014 and included C loading, export, gross primary production (GPP), and in-lake C pool estimates. East Long Lake is located at the University of Notre Dame Environmental Research Center (46°13′N 89°32′W), and it has an area of 3.2 ha and mean depth of 4 m. East Long Lake is a dimictic, mesotrophic (total phosphorus: 15.9 μg L^{-1}; chlorophyll *a*: 7.9 μg L^{-1}) lake with an average water residence time of 296 d (Zwart et al. 2016, 2017). Light is attenuated fairly quickly in the water column (light extinction coefficient: 2.86 m^{-1}) due to the high concentration of DOC. The lake is ice-covered annually and thermally stratifies shortly after ice out, which typically occurs in April or early May. A full description of the data collection methodologies was provided by Zwart et al. (2016, 2017). The description of the synthetic lake C pools follows our description of the lake C model below.

Lake carbon dynamics model

We modeled epilimnetic DOC and CO_2 pools as a function of inputs (e.g., inlet stream, precipitation, and atmospheric flux), within-lake processing (e.g., respiration and primary production), entrainment, and outputs (e.g., outlet stream and atmospheric flux) at a daily time scale. For model simplicity, we considered a single inorganic C pool consisting of CO_2, because, in soft waters such as East Long Lake, the net mass exchange between CO_2 and the other inorganic forms (bicarbonate + carbonate ions) is small and because including carbonate equilibria dynamics in the model structure had very little impact on estimates of the DOC and CO_2 pools. For a detailed description of a model version that includes carbonate equilibria dynamics as well as model results, see the Supporting Information and the *dic_co2_model* branch of our Github repository (https://github.com/jzwart/lake_C_EnKF/tree/dic_co2_model). Although our daily time step model does not explicitly account for diel variation in CO_2, our assimilated observations were collected in the morning (between 10:00 and 11:40 h) and likely represent a reasonable average concentration for the day since this time period is a balance

between peak net CO_2 consumption (midday) and peak net CO_2 production (nighttime; see Supporting Information Fig. S6). For DOC, we considered either a single homogenous pool or separate fast and slow decomposing pools. We used observations of the state variables (DOC and CO_2) and forcing variables (e.g., hydrologic inputs and outputs and temperature) to fit the model by assimilating synthetic or real observations to inform parameters describing the respiration rate of DOC or the partitioning of inflowing DOC between fast and slow decomposing pools.

Limnologists have dealt with the nature of DOC in many ways, from treating DOC as a black box (one DOC pool) to model molecular formula for each compound, as well as a continuum of complexity between these two extremes. Embracing DOC chemical and reactivity diversity has led to an understanding of how terrestrial C is assimilated into lake secondary production (Berggren et al. 2010), how hydrology influences lake DOC reactivity and CO_2 production (Vachon et al. 2016), and interactions between bacterial community composition and DOC degradation (Logue et al. 2016). We balance the heterogeneity of DOC compounds and reactivity with model computational efficiency by using a simple two-compartment representation of variation in DOC reactivity (two DOC pools) in our lake C process model and compare the model output with our observations of CO_2 and DOC when assimilating real observations. When assimilating synthetic observations, we represent DOC either as a single homogenous pool or separate fast and slow decomposing pools depending on the complexity of the model (See *Synthetic Observation Simulation and Assimilation* section). Representing DOC as a single pool or two pools potentially produces different dynamics of total DOC decay (turnover rate multiplied by DOC pool) as the one-pool model turnover rate *constant* is independent of DOC pool size, while the two-pool model turnover rate is a function of the relative size of the recalcitrant and labile DOC pools. In the one DOC pool model, we estimated the turnover rate of DOC (standardized to 20 °C, d_{20}) using the EnKF. In the two DOC pool model, we fixed the turnover rate constants of the fast and slow decomposing pools due to equifinality when estimating both pool's respiration rates, and instead estimated the partitioning of inflowing DOC between fast and slow decomposing DOC pools (*fracFast*). We calculated the emergent turnover rate of the total DOC pool in the two DOC pool model as the pool-weighted average, which allowed for direct comparison to the turnover rate of DOC estimated in the one DOC pool model.

We ran the lake C process models at a daily time step to estimate either two (total DOC [*DOC*] and CO_2; unit = mol C) or four state variables (slow decomposing DOC [DOC^s], fast decomposing DOC [DOC^f], total DOC [*DOC*], and CO_2; unit = mol C). The lake C dynamics were expressed as:

$$\mathbf{x_{t+1}} = \mathbf{B_t x_t} + \mathbf{C_t u_t} \quad (1)$$

where $\mathbf{x_t}$ was either a 2 × 1 or 4 × 1 vector of lake C pools at time *t*:

$$\mathbf{x_t} = \begin{bmatrix} CO_{2,t} \\ DOC_t \end{bmatrix}; \text{for Models } 2-5 \quad (2)$$

$$\mathbf{x_t} = \begin{bmatrix} CO_{2,\mathrm{t}} \\ DOC_t^{\mathrm{s}} \\ DOC_t^{\mathrm{f}} \\ DOC_t \end{bmatrix}; \text{for Models } 1, 6-9 \quad (3)$$

$\mathbf{B_t}$ was either a 2 × 2 or 4 × 4 matrix (unit = fraction d^{-1}) describing C pool-dependent processes including hydrologic C export downstream, vertical entrainment, C decay, and atmospheric losses. DOC and CO_2 stream and groundwater export from the lake was estimated as the quotient of measurements of stream and groundwater discharge out (Q_{out}; unit = $m^3\ d^{-1}$) and epilimnetic volume (*V*; unit = m^3). CO_2 efflux to the atmosphere was estimated as the quotient of gas piston velocity for CO_2 (*k*; unit = m d^{-1}; modeled from Vachon and Prairie [2013]) and epilimnetic depth (*zMix*; unit = m). Vertical loss of DOC and CO_2 to the hypolimnion if *zMix* decreases (*Loss* = 1 if *zMix* decreases, otherwise 0) estimated as the quotient of vertical entrainment water volume (*Vert*; unit = $m^3\ d^{-1}$) and *V* (Vachon et al. 2017). The DOC transformation into CO_2 was estimated as the respiration of either the total DOC pool into CO_2 (d_{20}; unit = d^{-1}; estimated using the EnKF described below) or both the slow ($d_{\mathrm{slow},20}$ = 0.004 d^{-1}) and fast ($d_{\mathrm{fast},20}$ = 0.3 d^{-1}) DOC pools into CO_2. We fix $d_{\mathrm{slow},20}$ and $d_{\mathrm{fast},20}$ and estimate the partitioning between inflowing DOC into fast and slow decomposing DOC pools (*fracFast*; unit = fraction; Eq. 9) due to equifinality when estimating both $d_{\mathrm{slow},20}$ and $d_{\mathrm{fast},20}$. *d*, d_{slow}, and d_{fast} were standardized to 20°C using the mean epilimnion temperature at time *t* (*EpiT*; unit = °C) and Eq. 4 to take into account temperature influence on the mineralization rate of DOC (Holtgrieve et al. 2010; Solomon et al. 2013):

$$d_t = d_{20,t} \times 1.047^{\left(EpiT_t - 20\right)} \quad (4)$$

$$\mathbf{B_t} = \begin{bmatrix} 1 - \left(\frac{Q_{\mathrm{out},t}}{V_t}\right) - \left(\frac{k_t}{zMix_t}\right) - \left(\frac{Loss_t \times Vert_t}{V_t}\right) & d_t \\ 0 & 1 - d_t - \left(\frac{Q_{\mathrm{out},t}}{V_t}\right) - \left(\frac{Loss_t \times Vert_t}{V_t}\right) \end{bmatrix}; \text{for Models } 2-5 \quad (5)$$

$$\mathbf{B_t} = \begin{bmatrix} 1-\left(\frac{Q_{\text{out},t}}{V_t}\right)-\left(\frac{k_t}{zMix_t}\right)-\left(\frac{Loss_t \times Vert_t}{V_t}\right) & d_{\text{slow},t} & d_{\text{fast},t} & 0 \\ 0 & 1-d_{\text{slow},t}-\left(\frac{Q_{\text{out},t}}{V_t}\right)-\left(\frac{Loss_t \times Vert_t}{V_t}\right) & 0 & 0 \\ 0 & 0 & 1-d_{\text{fast},t}-\left(\frac{Q_{\text{out},t}}{V_t}\right)-\left(\frac{Loss_t \times Vert_t}{V_t}\right) & 0 \\ 0 & 1-d_{\text{slow},t}-\left(\frac{Q_{\text{out},t}}{V_t}\right)-\left(\frac{Loss_t \times Vert_t}{V_t}\right) & 1-d_{\text{fast},t}-\left(\frac{Q_{\text{out},t}}{V_t}\right)-\left(\frac{Loss_t \times Vert_t}{V_t}\right) & 0 \end{bmatrix}; \quad (6)$$

for Models 1, 6–9

$\mathbf{C_t}$ was either a 2 × 6 or 4 × 7 matrix and $\mathbf{u_t}$ was either a 6 × 1 or 7 × 1 vector describing pool-independent processes including atmospheric influx of C, vertical entrainment, and C loads. The elements in the resulting matrix product of $\mathbf{C_t u_t}$ had unit of mol C d^{-1}. Atmospheric CO_2 ([$atmCO_2$]; unit = mol C m^{-3}) influx to the lake was estimated as the quotient of gas piston velocity for CO_2 (k; unit = m d^{-1}) and epilimnetic depth (*zMix*; unit = m), where [$atmCO_2$] is 400 ppm converted to mol C m^{-3}. CO_2 and DOC loading ($loadCO_2$ and *loadDOC*, respectively; unit = mol C d^{-1}) to the lake was estimated from stream, groundwater, precipitation, and overland flow measurements multiplied by the fraction that enters the epilimnion (*fracEpi;* estimated using the EnKF described below). The fraction of *loadDOC* that enters the fast-decomposing DOC pool (*fracFast*; unit = fraction) was estimated using the EnKF described below. Vertical entrainment of DOC and CO_2 from the hypolimnion if *zMix* increases (*Gain* = 1 if *zMix* increases, otherwise 0) was estimated as the product of vertical entrainment water volume (*Vert*; unit = m^3 d^{-1}) and hypolimnic CO_2 ($hypoCO_2$; unit = mol C m^{-3}) or DOC (*hypoDOC*; unit = mol C m^{-3}) concentration (fraction of hypolimnion DOC pool that is fast decomposing set to 0.01, $fracFast_0$; Guillemette and del Giorgio 2011; Guillemette et al. 2013; Mostovaya et al. 2016). The loss of CO_2 to primary production accounting for the fraction of GPP that is respired quickly was set to 0.85 (R_{GPP}; unit = fraction; Quay et al. 1986; Cole et al. 2002; Hanson et al. 2004). Finally, DOC exudate from phytoplankton production included both slow ($Exude^s$ = 0.03) and fast ($Exude^f$ = 0.07) decomposing DOC fractions from phytoplankton exudate or the sum of the two, termed *Exude* (Baines and Pace 1991; Hanson et al. 2004).

$$\mathbf{C_t} = \begin{bmatrix} \frac{k_t}{zMix_t} & 1 & -GPP_t & 0 & 0 & hypoCO_{2,t} \\ 0 & 0 & 0 & GPP_t & 1 & hypoDOC_t \end{bmatrix}; \text{ for Models } 2-5 \quad (7)$$

$$\mathbf{u_t} = \begin{bmatrix} [atmCO_2] \times V_t \\ loadCO_{2,t} \times fracEpi \\ 1-R_{\text{GPP},t} \\ Exude_t \\ loadDOC_t \times fracEpi \\ Gain_t \times Vert_t \end{bmatrix}; \text{ for Models } 2-5 \quad (8)$$

$$\mathbf{C_t} = \begin{bmatrix} \frac{k_t}{zMix_t} & 1 & -GPP_t & 0 & 0 & 0 & hypoCO_{2,t} \\ 0 & 0 & 0 & GPP_t & 0 & 1-fracFast_t & hypoDOC_t \times \left(1-fracFast_{0,t}\right) \\ 0 & 0 & 0 & 0 & GPP_t & fracFast_t & hypoDOC_t \times fracFast_{0,t} \\ 0 & 0 & 0 & GPP_t & GPP_t & 1 & hypoDOC_t \end{bmatrix}; \text{ for Models } 1, 6-9 \quad (9)$$

$$\mathbf{u_t} = \begin{bmatrix} [atmCO_2] \times V_t \\ loadCO_{2,t} \times fracEpi \\ 1-R_{\text{GPP},t} \\ Exude_t^s \\ Exude_t^f \\ loadDOC_t \times fracEpi \\ Gain_t \times Vert_t \end{bmatrix}; \text{ for Models } 1, 6-9 \quad (10)$$

Ensemble Kalman filter

We used the EnKF (Evensen 1994) to estimate the states of the lake system $\mathbf{x_t}$ and the parameters (either $d_{20,t}$ and $fracEpi_t$ or $fracFast_t$ and $fracEpi_t$) from Eq. 1. EnKF uses an iterative two-step process to estimate the state of a system: a forecast

step, during which the state of a system (DOC and CO_2) is predicted by a model (Eq. 1), and an update step, during which the state variables and parameters are adjusted using observations (DOC and CO_2) and the Kalman gain matrix. The EnKF uses a Monte Carlo sampling technique to produce an ensemble of model states and parameters, which allows us to represent the error statistics in the model estimates as a sample covariance matrix. In our study, we set ensemble size to 100 (N_e = 100), as such a number of ensemble members has been shown to be sufficient in previous studies (e.g., Gao et al. 2011; Huang et al. 2013), and through a sensitivity analysis, we observed little effect of ensemble size on model performance (N_e = 50–1000). We implemented the EnKF algorithm as described by Gao et al. (2011).

We generated the initial 100 ensemble members for the parameters by drawing from a normal distribution with a mean and standard deviation based on literature values and previous research conducted on East Long Lake (Zwart et al. 2016, 2017) for the fraction of inflowing DOC loaded into the epilimnion and the turnover rate of the total DOC pool standardized to 20°C or the fraction of loaded DOC that is fast decomposing (*fracEpi*, mean = 0.1, SD = 0.04; d_{20}, mean = 0.007, SD = 0.0038; *fracFast*, mean = 0.30, SD = 0.099). We also estimate a covariance inflation factor (σ; mean = 1, SD = 1.2) using the EnKF, which prevents the model from becoming overconfident in its predictions and ignoring assimilated observations, also known as "filter divergence" (Li et al. 2009; Dietze 2017). Filter divergence is especially important to consider in sequential data assimilation as process error becomes reduced as more data are assimilated through time. Ideally, the model should be able to predict abrupt changes in the modeled ecosystem; however, if certain ecosystem dynamics are unknown or are not explicitly included in the model structure, it may be best to inflate process variance to anticipate unknown ecosystem changes not captured by the model (Anderson 2007). Other options include using Bayesian Model Averaging which allows for different model structures to be run in a forecast and weighted based on their uncertainty (Hipsey et al. 2015; Dietze 2017).

The initial values for model state variables in each ensemble member were drawn from a normal distribution of CO_2 and DOC with the mean set to the earliest observations of CO_2 and DOC and standard deviations described below (Eq. 18). If estimating four C states, the initial DOC^f pool was set to 1% of the initial *DOC* pool for each ensemble while initial DOC^s was the remainder (Guillemette and del Giorgio 2011; Guillemette et al. 2013; Mostovaya et al. 2016). We concatenated parameter estimates (**p**) and state estimates (**x**) into one vector (**y**), and we used this vector for forecasting and updating in the EnKF.

$$\mathbf{y_{i,t}} = \begin{bmatrix} \mathbf{p_{i,t}} \\ \mathbf{x_{i,t}} \end{bmatrix} \tag{11}$$

where i was the ith ensemble member of the model. $\mathbf{p_{i,t}}$ was a 3 × 1 vector containing the parameter estimates (either $d_{20,t}$, $fracEpi_t$, and σ_t or $fracLabile_t$, $fracEpi_t$, and σ_t) and $\mathbf{x_{i,t}}$ was either a 2 × 1 or 4 × 1 vector containing the state estimates (either CO_2 and *DOC* or CO_2, DOC^s, DOC^f, and *DOC*). The **y** vector was propagated through time using our simple C model (Eq. 1) and parameter estimates from the previous time step, where the forecasted **y** vector was denoted as $\mathbf{y^f}$.

$$\mathbf{y^f_{i,t}} = \begin{bmatrix} \mathbf{p_{i,t-1}} \\ \mathbf{x_{i,t}} \end{bmatrix} \tag{12}$$

When one or more observations were available at a time step t, the **y** vector was updated using a Kalman gain. The updated **y** vector ($\mathbf{y^u}$) was expressed as:

$$\mathbf{y^u_{i,t}} = \mathbf{y^f_{i,t}} + \mathbf{K_t}\left(\mathbf{x_t^{obs}} - \mathbf{H_t}\mathbf{y^f_{i,t}}\right) (i = 1,2,\ldots N_e) \tag{13}$$

where $\mathbf{x_t^{obs}}$ was a 2 × 1 vector of observation data (CO_2 and DOC); $\mathbf{H_t}$ was either a 2 × 5 or 2 × 7 measurement operator matrix with 1s when observations of C pools were available and 0s otherwise.

$$\mathbf{H_t} = \begin{bmatrix} 0 & 0 & 0 & CO_{2,\mathrm{obs},t} & 0 \\ 0 & 0 & 0 & 0 & DOC_{\mathrm{obs},t} \end{bmatrix}; \text{ for Models } 2-5 \tag{14}$$

$$\mathbf{H_t} = \begin{bmatrix} 0 & 0 & 0 & CO_{2,\mathrm{obs},t} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & DOC_{\mathrm{obs},t} \end{bmatrix}; \text{ for Models } 1, 6-9 \tag{15}$$

$\mathbf{K_t}$ was the Kalman gain weighting matrix modified by the estimated covariance inflation, which was expressed as:

$$\mathbf{K_t} = \frac{1}{N_e - 1}\sigma_t \times \mathbf{\Delta Y_t \Delta Y_t}^\top \mathbf{H_t}^\top \left(\frac{1}{N_e - 1}\sigma_t \times \mathbf{H_t \Delta Y_t \Delta Y_t}^\top \mathbf{H_t}^\top + \mathbf{R_t}\right)^{-1} \tag{16}$$

where $\mathbf{R_t}$ was a 2 × 2 observation error covariance matrix and expressed as:

$$\mathbf{R_t} = \begin{bmatrix} Variance\, CO_{2,\mathrm{obs},t} & 0 \\ 0 & Variance\, DOC_{\mathrm{obs},t} \end{bmatrix} \tag{17}$$

We model C pools to propagate different sources of observation error (variance) including error in C concentration (CO_2 or DOC) and epilimnetic volume. These sources of error were propagated using the formula:

$$Variance\, C_t = \left(C_t \times \sqrt{cv[C]_t^{\,2} + cvLA_t^{\,2} + cvzMix_t^{\,2}}\right)^2 \tag{18}$$

where $VarianceC_t$ was C pool variance at time t; C_t was the C pool observation; and $cv[C]_t$, $cvLA_t$, and $cvzMix_t$ were the coefficient of variation (CV) for C concentration, lake area, and epilimnetic depth, respectively. Standard deviations for observations of DOC (0.106 mol C m^{-3}) and CO_2 (0.00503 mol C m^{-3}) concentrations were estimated using 8 DOC and 12 CO_2

replicates from a sampling time point on 30 July 2014 from the adjacent lake basin (West Long Lake, description in Zwart et al. [2016]). Standard deviation in lake area was conservatively set to 4000 m^2, and standard deviation in the epilimnetic depth was conservatively set to 0.25 m based on the accuracy of the USGS National Hydrography Dataset and Onset HOBO temperature pendants (Onset Computer Corporation), respectively.

$\mathbf{\Delta Y_t}$ in Eq. 16 was either a 5 x N_e (Models 2–5) or 7 x N_e (Models 1 and 6–9) matrix of all ensemble deviations from the mean of estimated states and parameters at time t ($\overline{y}_t$), expressed as:

$$\Delta \mathbf{Y_t} = \left[\Delta \mathbf{y_{1,t}} \ldots \Delta \mathbf{y_{i,t}} \ldots \Delta \mathbf{y_{N_e,t}}\right] \quad (19)$$

where the ith column of $\mathbf{\Delta Y_t}$ was:

$$\mathbf{\Delta y_{i,t}} = y_{i,t} - \overline{y}_t \quad (20)$$

All modeling and subsequent analyses were conducted using the R statistical package (R Core Team 2016). The development code is available on GitHub (https://github.com/jzwart/lake_C_EnKF), and the dataset and code used to generate results in this manuscript were from v1.0 of this repository (http://doi.org/10.5281/zenodo.1322130).

Evaluating EnKF performance with real observations

When assimilating real observations, we withheld every other observation from assimilation to be used to evaluate model performance. In order to evaluate the model performance, we compared model-estimated states to the nonassimilated observed data using root mean squared error (RMSE) and coefficient of determination (r^2). We estimate the uncertainty in lake C states by calculating the CV of the ensemble state estimates and observations across the model run.

Synthetic observation simulation, assimilation, and evaluation

While we can evaluate the performance of simple lake C process models by comparing their output with observations, we cannot be sure that these models fully capture the "true" underlying ecosystem processes. Given that the true underlying process will never be fully represented by any model under consideration, should we accept data assimilation results as improved estimates of the truth or disregard them as artifacts of the model structure and instead trust our observations? To explore this question, we created "true" ecosystem states of DOC and CO_2 (hereafter termed *true states*) using a known process model of the ecosystem (hereafter termed *true process*). We then simulated the collection of observations on these true states (hereafter termed *synthetic observations*), and assimilated the synthetic observations of DOC and CO_2 into eight different process models ranging from the true process to highly simplified representations of the true process (Fig. 1). This type of analysis is akin to observing system simulation experiments (Masutani et al. 2010), which aim to identify how many and what types of observations are needed to reach a desired model performance given limited research funds. As our data collection had already occurred, the primary goal of our synthetic data assimilation experiment was to show how data assimilation can combine information from observations and simplified models (which are imperfect representations of the true states and true process, respectively) to produce estimates of C pools that are closer to the true C pools than direct observation of these pools.

To create the true states, we considered the true process to be the two DOC pool model structure described above (Eqs. 1,

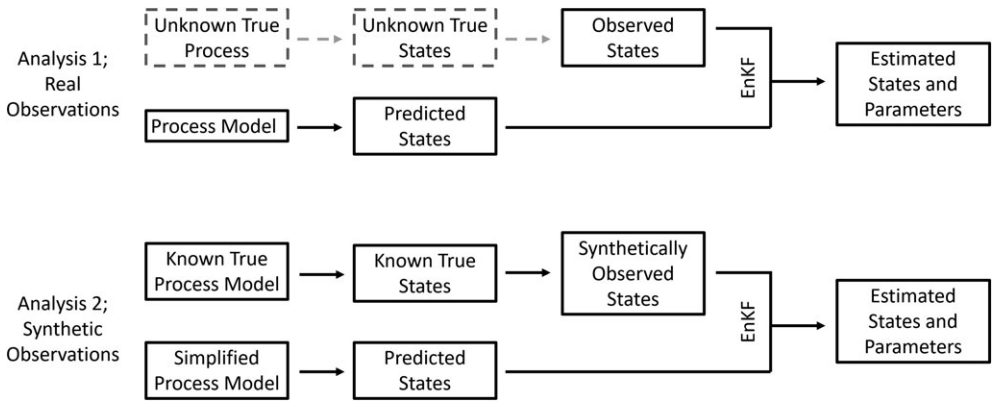


Fig. 1. Schematic of how the EnKF was used in two different analyses using real observations and synthetic observations. In our first analysis, we used real observations to demonstrate that the EnKF can estimate relevant ecosystem parameters and significantly reduce uncertainty in lake carbon pools and fluxes compared to observation uncertainty. In our second analysis, we generated true states using a true process model (Model 9; Table 1), and then sampled these true states to create synthetic observations which were assimilated into simplified versions of the true process. For this analysis, we showed that the EnKF technique improved estimates of the true state of the system compared to estimates derived solely from observations, even when the model structure was a simplification of the true process.

Table 1. Model structure of the nine different lake carbon models used in this study. Models either estimated two (total DOC [*DOC*] and CO_2) or four state variables (slow decomposing DOC [DOC^s], fast decomposing DOC [DOC^f], total DOC [*DOC*], and CO_2) while estimating either the turnover rate of DOC (d_{20}), fraction of loaded C into the epilimnion (*fracEpi*), and inflation coefficient (σ) or the partitioning of loaded DOC into fast and slow decomposing pools (*fracFast*), *fracEpi*, and σ, respectively. We compared eight different model structures when we assimilated synthetic observations, where Model 9 was the true process which was used to generate the true states. Model estimates of d_{20} and the fixed parameters d_{slow} and d_{fast} were modeled with or without temperature dependence (Eq. 4) and with or without Michaelis–Menten kinetics (Eq. 21).

Model	State variables	Parameters estimated	Data	Temperature dependence	Michaelis–Menten kinetics
1	CO_2, *DOC*, DOC^l, DOC^r	*fracLabile*, *fracEpi*, σ	Real	True	False
2	CO_2, *DOC*	d_{20}, *fracEpi*, σ	Synthetic	False	False
3	CO_2, *DOC*	d_{20}, *fracEpi*, σ	Synthetic	True	False
4	CO_2, *DOC*	d_{20}, *fracEpi*, σ	Synthetic	False	True
5	CO_2, *DOC*	d_{20}, *fracEpi*, σ	Synthetic	True	True
6	CO_2, *DOC*, DOC^l, DOC^r	*fracLabile*, *fracEpi*, σ	Synthetic	False	False
7	CO_2, *DOC*, DOC^l, DOC^r	*fracLabile*, *fracEpi*, σ	Synthetic	True	False
8	CO_2, *DOC*, DOC^l, DOC^r	*fracLabile*, *fracEpi*, σ	Synthetic	False	True
9	CO_2, *DOC*, DOC^l, DOC^r	*fracLabile*, *fracEpi*, σ	Synthetic	True	True

3, 6, 9, and 10), with an additional complexity of the turnover rate of DOC (d, d_{slow}, and d_{fast}) following Michaelis–Menten kinetics (d_{MM}) as has been used to describe DOC decay previously (e.g., Søndergaard and Middelboe 1995).

$$d_{MM,t} = \frac{d_t \times \frac{12 \times DOC_t}{V_t}}{\left(K + \frac{12 \times DOC_t}{V_t}\right)} \quad (21)$$

where K is the half-saturation constant, set to 4 g m^{-3} for the total DOC pool. We chose to add d_{MM} to the true process model in order to increase the number of simplified models of DOC decay that we could consider, which included either one or two DOC pools, temperature-dependent or temperature-independent DOC decay, and Michaelis–Menten decay kinetics or not (Table 1, Models 2–9). We created the true states of in-lake C pools using the true process model (Model 9 in Table 1) with a single parameterization and forcing data from the dataset described in Zwart et al. (2016, 2017) (e.g., loading of DOC and CO_2, hydrologic outflow, and primary production). We created synthetic observations of the true states by sampling from a normal distribution with the true states as the mean and standard deviations set to our observation error (Eq. 18), thereby mimicking DOC and CO_2 sampling from our simulated lake.

Sampling protocol (e.g., sampling interval and sample replicates) may influence the model performance during data assimilation as process models gain more information as more observations are assimilated. Therefore, we additionally compared each model's performance across varying sampling protocols, which included varying sampling interval (1–35 d sampling intervals) and sample replicates (1–6 replicates for each sample time point) of the true states. We assimilated synthetic observations using each model across all combinations of these sampling protocols and calculated the mean model performance by running each model 100 times for each sampling protocol to reduce the influence of random draws of initial C pools and random sampling of the true states on model performance.

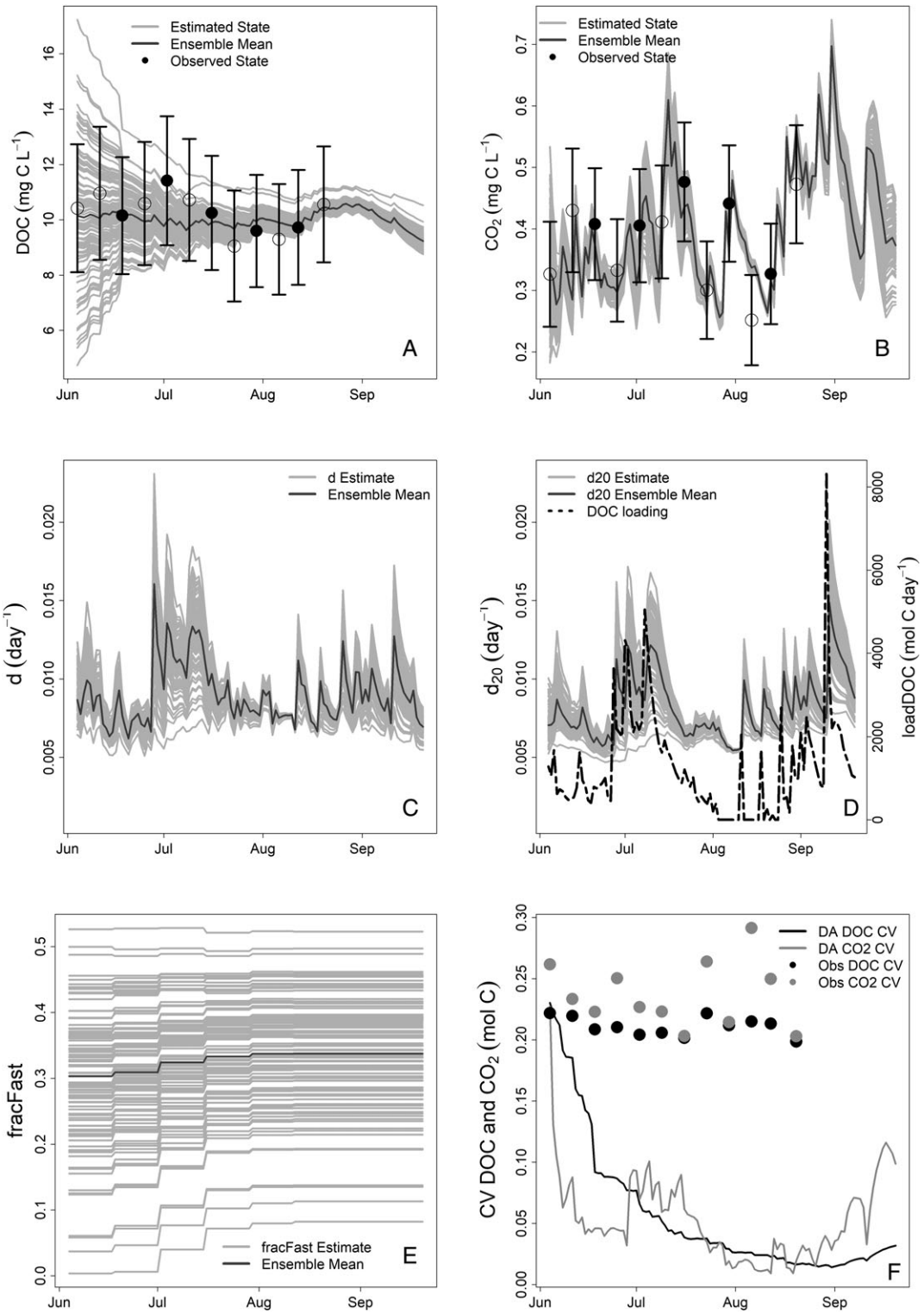
We asked whether the data assimilation estimates or synthetic observations better capture the true states by comparing states estimated by data assimilation as well as the synthetic observation states to the true states using RMSE (termed DA RMSE and Obs RMSE, respectively). We regressed the difference in Obs RMSE and DA RMSE against process model structure, sampling interval, and sample replicates to ask whether sampling protocol and/or model structure significantly influenced data assimilation performance compared to observation performance in estimating the true ecosystem state.

Results and discussion

Real observation assimilation

Using real observations, we captured East Long Lake DOC and CO_2 dynamics well by assimilating data with a two DOC pool model using the EnKF (Fig. 2A,B; CO_2 r^2 = 0.41, CO_2 RMSE = 0.074 g C m^{-3}, DOC r^2 = 0.24, DOC RMSE = 0.65 g C m^{-3}). Our simple representation of the DOC reactivity continuum produced dynamic estimates of the turnover rate of DOC, as the turnover rate of DOC standardized to 20°C (d_{20}) ranged from 0.0054 to 0.0153 d^{-1} and responded to hydrologic loads of DOC. Lakes which have short hydrologic residence times (HRTs), and which therefore are fairly continuously resupplied with fresh allochthonous inputs, should have higher emergent mineralization rates of the dissolved resource, as a greater proportion of the emergent mineralization rate will be derived from mineralization of fast decomposing resources (Jones et al. 2018). In these low HRT lakes, this can be thought of as a sustained reactivity

distribution, such that the reactivity of the DOC amalgamation within the lake will be similar to its inflowing water source, whereas the reactivity of DOC in a long HRT lake will be much lower than its inflowing water source. Indeed, recent research points to control of hydrology on the emergent property of the turnover rate of DOC. For example, Vachon et al. (2016) have developed an explicit formulation describing the effect of HRT on the apparent decay rate of DOC as a



function of the initial reactivity of the DOC inputs. Similarly, the turnover rate of DOC has been shown to be negatively related to HRT in a broad cross section of systems (Catalán et al. 2016; Evans et al. 2017). Finally, the turnover rate of DOC was enhanced during periods of short HRT following extreme precipitation events (Zwart et al. 2017).

The uncertainty in our estimated C pools generally decreased as more observations were assimilated, while observed uncertainty in C pools remained roughly the same throughout the study period (Fig. 2F). Reducing uncertainty in lake C pools and flux estimates helps researchers better understand how lakes fit into regional C budgets and interpret lake C responses to interannual and intra-annual variations in driving variables (e.g., weather). Repeated observations of ecosystems increase our confidence in the measured states and our understanding of ecosystem processes. However, monetary and time constraints can limit the number of observations of the ecosystem across a sampling time period and/or during a sampling event. Data assimilation makes the best use of limited observations by reducing our uncertainty in estimated states and parameters by combining information from the collected data and models. Additionally, power analyses may be performed to determine how many and what types of observations are needed to reduce forecast uncertainty below a desired threshold (Dietze 2017). Mathematically, combining two pieces of information with Gaussian error distributions such as in the EnKF leads to a combined error that is less than or equal to the minimum error of the two pieces of information (Lahoz et al. 2010). We were confident that the reduced uncertainty in estimated C pools was meaningful and informative since there was no relationship between time and the RMSE of modeled DOC or CO_2 for our model runs ($p > 0.5$; linear regression), indicating that there was not a systematic divergence of models and observed data (filter divergence; Dietze 2017). Furthermore, our estimated covariance inflation factor was near one and varied little throughout the study period (mean = 0.97; CV = 0.01).

In addition to reducing the uncertainty in estimates of C pool sizes, the data assimilation analysis provides informed estimates, with quantified uncertainty, of C pool sizes at time points when observations were not made. Without a process model to perform this gap filling, we would often resort to linear interpolation to estimate C pool sizes a time points when observations were not available; however, alternative options to linear interpolation are available for gap filling such as using generalized additive models. If observation frequency is low and/or C pools are variable through time, then linearly interpolated C pools will likely be dissimilar to the more accurate process model-estimated C pools. For example, even with weekly sampling rate, linearly interpolated CO_2 pool could deviate by as much as 51% from the model estimated CO_2, whereas linearly interpolated DOC pool had a maximum difference of 15% from the model estimated DOC pool. This can have important implications for estimating how much CO_2 is emitted to the atmosphere from a lake at a seasonal time scale, especially if lakes are not sampled during times of dramatic increases or decreases in CO_2 such as following extreme precipitation events (Ojala et al. 2011; Vachon and del Giorgio 2014), shortly after ice-off in temperate and boreal lakes (Striegl et al. 2001; Baehr and DeGrandpre 2004; Vachon and del Giorgio 2014), or during phytoplankton blooms in eutrophic lakes (Balmer and Downing 2011).

Synthetic observation assimilation

For an example of synthetic observations assimilated into a simplified process model and compared to the true state of the ecosystem, see Fig. 3. In this example, a simplified process model informed with synthetic observations produced lake C pool estimates that were closer to the true state than the synthetic observations themselves were (Fig. 3A,B). The simplified process model compared to synthetic observations had an RMSE of 0.71 g C m^{-3} for DOC and 0.030 g C m^{-3} for CO_2. This was similar to the RMSE calculated between our model estimated states and observations when assimilating real observations (RMSE DOC: 0.65 g C m^{-3}; RMSE CO_2: 0.074 g C m^{-3}). Furthermore, the uncertainty in the C pools was reduced after assimilating synthetic observations, while synthetic observations remained fairly uncertain throughout the time period (Fig. 3D).

Across all simplified model structures and sampling protocols of the true states, the data assimilation estimates of DOC were closer to the true state than the synthetic observations (Fig. 4A,C,E). This was not the case for data assimilation estimates of CO_2, as only 49% of the data assimilation runs were closer to the true CO_2 states than the synthetic observations were to the true ecosystem state (Fig. 4B,D,F). However, the

Fig. 2. Data-assimilation–estimated and observed states of **(A)** DOC and **(B)** CO_2 for the two DOC pool process model (Model 1; Table 1). Filled in black circles are the assimilated observations and the open circles are the withheld observations used for evaluating model performance. Black error bars (panels A and B) for the observed states of DOC and CO_2 are 1 standard deviation. The temperature corrected turnover rate of DOC **(C)** and the turnover rate of DOC standardized to 20°C **(D)** varied throughout the time period and responded to loaded DOC, indicated by the black dashed line in panel D. The EnKF estimates the partitioning of loaded DOC into the labile and recalcitrant DOC pools, where 0.3 represents that 30% of the loaded DOC is fast decomposing **(E)**. The gray lines (panels A–E) represent data-assimilation–estimated states or parameters for each ensemble ($n = 100$), and the dark gray line represents the mean of all ensemble estimates. At each time point, we calculated the CV of the observed states, as well as the CV of the data-assimilation–estimated states across ensemble model runs as an indication of the uncertainty in our C pool size estimates. The uncertainty in data assimilation estimates of DOC and CO_2 pools decreased as more data are assimilated (DA DOC CV and DA CO_2 CV in panel F), while the uncertainty in observed C pools remained fairly constant throughout the time period (Obs DOC CV and Obs CO_2 CV in panel F). Note that the CV in panel F includes uncertainty in epilimnetic volume as well as C concentration.

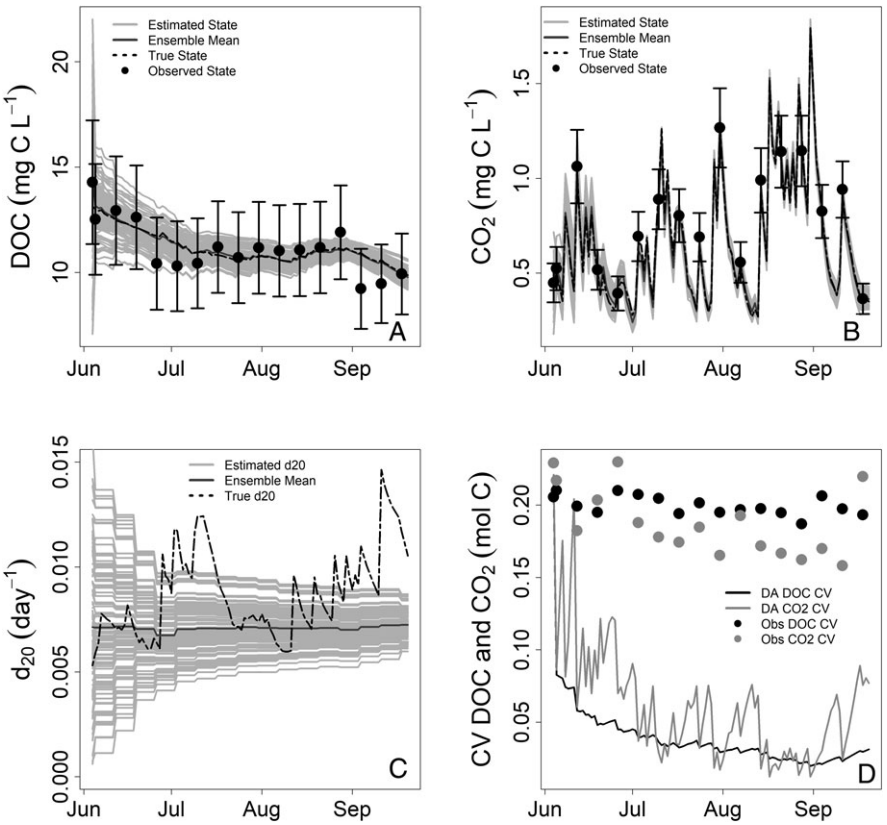


Fig. 3. Data assimilation example of a simplified process model (Model 2; Table 1) run with synthetic observations that were sampled every 7 d with two replicates at each sampling time point. In this example, the data-assimilation–estimated DOC and CO_2 states were closer to the true DOC and CO_2 states than the synthetic observations themselves (DA RMSE for DOC: 0.15 mg L^{-1}; Obs RMSE for DOC: 0.80 mg L^{-1}; DA RMSE for CO_2: 0.029 mg L^{-1}; Obs RMSE for CO_2: 0.035 mg L^{-1}). The data-assimilation–estimated turnover rate of DOC (d_{20}, dark gray line, panel C) varied throughout the model run ranging from 0.0067 to 0.0072 d^{-1}. The true d_{20} (dashed black line, panel C) responded to hydrologic loads of DOC (not shown) and varied throughout the time period. At each time point, we calculated the CV of the synthetic observation states, as well as the CV of the data-assimilation–estimated states across ensemble model runs as an indication of the uncertainty in C pool size estimates. The uncertainty of the estimated states was reduced through time via data assimilation, while the uncertainty of the observed states remained roughly constant (panel D).

synthetic observations of CO_2 were closer to the true state only when more than one sample replicate was taken, meaning that data assimilation was most valuable when few replicate samples were collected, even if the process model used for data assimilation was a simplified version of the true process. As multiple inaccurate model structures resulted in constrained predictions and acceptable results, it raises the question as to how we can use data assimilation to gain knowledge of processes governing the system? Indeed, sometimes it is difficult to distinguish which processes are important for accurately predicting system dynamics due to equifinality in estimated parameters or insensitive model structures. For example, Hararuk et al. (2018) found it difficult to distinguish between biodegradation and photodegradation rates as these rates were additive and drivers for each process were positively correlated. In such cases, assimilating additional sources of information to constrain processes or structure can help; for example, in situ incubation experiments could be conducted to help constrain rates of photodegradation and biodegradation.

Model structure had a significant effect on data assimilation performance compared to synthetic observation performance for both DOC and CO_2 ($p < 0.001$; analysis of variance). Also, sample replicates and sampling interval both had significant, negative effects on the difference between synthetic observation RMSE and data assimilation RMSE ($p < 0.001$ for both; linear regression; Fig. 4). The slight bimodal shape in some of the violin plots in Fig. 4 represents an observation performance gain where the observations were getting closer to the true C state by averaging over more replicate samples. The largest performance gain was observed when transitioning from one to two replicate samples (Fig. 4C,D). This indicates that data assimilation was most valuable when sampling interval was short and sample replicates were low. This has implications for how we best use high-frequency sampling (e.g., automated sensors), which typically includes just one sample replicate since automated sensors are usually expensive. By assimilating synthetic observations, we demonstrate that unless there are multiple sample replicates, high-frequency sampling will be further from the true states than

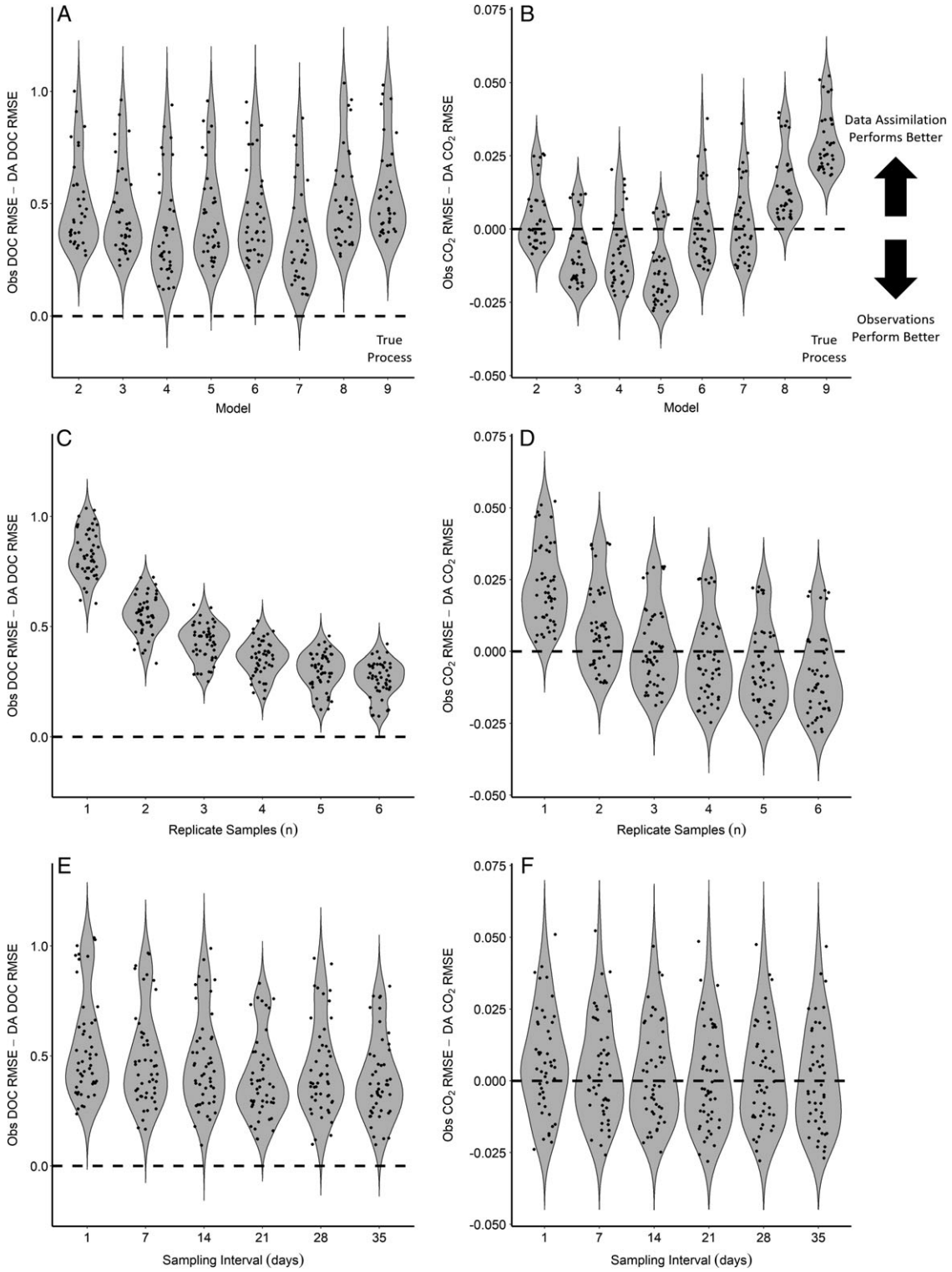


Fig. 4. The difference between RMSE for synthetic observations compared to true states (Obs RMSE) and RMSE for data-assimilation–estimated states compared to true states (DA RMSE). The horizontal dashed line indicates when there was no difference between data-assimilation–estimated states and observed states in terms of accurately capturing the true states, and positive values indicate that data-assimilation–estimated states were more accurate while negative values indicate that observed states were more accurate. Violin plots show the distribution of each model's performance across all sampling protocols for both DOC **(A)** and CO_2 **(B)**, with models ordered in increasing complexity ending with Model 9 as the true process. Data assimilation was most valuable with fewer replicate samples as more replicates decreased the models' performance over observations for both DOC **(C)** and CO_2 **(D)** across all model structures. Shorter sampling intervals also increased the models' performance over observations for both DOC **(E)** and CO_2 **(F)** across all model structures.

the data-informed model estimates. Therefore, although automated sensors are valuable tools for lake scientists, assimilating the data they collect into a process model increases their utility as the information from these sensors coupled with knowledge of ecological and biogeochemical interactions captured by a process model provides estimates closer to the true state compared to observation-derived estimates. Additionally, autocorrelation in high-frequency observations such as those from automated sensors reduces the effective sample size of assimilated observations as these observations are not independent.

We show that by combining information from both the synthetic observations and process model through data assimilation, we can often estimate the true state of the system more accurately than our observations of the true state, even if our representation of the true process is simplified. However, the accuracy of the data assimilation estimates depends on the states that are estimated. In our case, it was much easier to estimate the true DOC state than the true CO_2 state; in fact, the data-assimilation–estimated DOC states were always closer to the true DOC states than the observations themselves were, regardless of sampling frequency or replication. This may be because the size of the C flux (DOC decay into CO_2) was much larger relative to the CO_2 pool than to the DOC pool, and any inaccuracies in estimated parameters of C flux affect the CO_2 pool much more than the DOC pool. Accurate ecological forecasts are needed to guide our decision making on important economic, societal, and environmental issues (Clark et al. 2001; Dietze et al. 2018). However, there is often debate on how much detail to include in ecological models (e.g., Grimm and Railsback 2005; Vallino 2010; Travis et al. 2014) and even whether ecological forecasting using process models is a useful endeavor (Schindler and Hilborn 2015). We show that these synthetic data assimilation experiments can help identify which modeled ecosystem states are most sensitive to variations in model structure, and help guide which type of data to collect, at what frequency, and how many replicates in order to accurately predict lake C states.

Comments and recommendations

Although we include many essential processes for describing lake C dynamics when assimilating real observations, we did not explicitly test different model structures (other than including carbonate equilibria dynamics, see Supporting Information), which is another utility of data assimilation. For an example of a robust model structure comparison (102 different models) using data assimilation with the same lake dataset used in this analysis, see Hararuk et al. (2018). While our model structure performed well when assimilating real observations from East Long Lake, it is unclear if the same model structures used here and in Hararuk et al. (2018) will perform equally well across different lakes. For example, including carbonate equilibria processes did not affect our estimates of the DOC and CO_2 pools (Supporting Information); however, applying a model without explicit consideration of carbonate equilibria processes to hardwater lakes and lakes with different pH dynamics than East Long Lake could be quite problematic. Identifying essential processes to include in lake C models across a diverse set of lakes is an important next step for advancing broad-scale modeling of lake C cycling. Likewise, forecasting lake C states and fluxes and confronting these forecasts with new observations through data assimilation can accelerate lake C research by identifying processes we do and do not know well, and what data we need to collect to help us learn more about these important ecosystems (Dietze et al. 2018).

Given the important and changing role of lakes in the global C cycle, it is imperative that we understand the processes that regulate lake C pools and fluxes, estimate those pools and fluxes as accurately as possible, and quantify the uncertainty in those estimates. The last decade has seen dramatic increases in networked lake science (Hanson et al. 2016), high-frequency sensor data availability (Porter et al. 2012), continental-scale lake sampling efforts (USEPA 2009), and harmonized water quality databases (Soranno et al. 2015; Read et al. 2017) to help understand lake responses to environmental change. However, lake biogeochemists have yet to embrace data assimilation techniques when using these data, which presents a missed opportunity to combine information from data and models to help distinguish between competing model structures, reduce uncertainty in model estimates and forecasts, and identify sampling efforts needed. By assimilating data into simple lake C process models, we demonstrate that we can reduce uncertainty in our estimates of lake C pools and fluxes and more accurately estimate the true C states compared to observations, even with simplified representations of the true process. We echo the call from Hipsey et al. (2015) to utilize data assimilation techniques within lake biogeochemistry to more accurately forecast ecosystem states and processes, learn about ecosystem dynamics, and maximize use of scientific knowledge and observations of these important ecosystems. In order to reduce analytical barriers to using such tools, we provide open-source R code of the model and EnKF we used in this analysis in v1.0 (http://doi.org/10.5281/zenodo.1322130) of our GitHub repository (https://github.com/jzwart/lake_C_EnKF).

Data availability statement

All information on GitHub are available at https://github.com/jzwart/lake_C_EnKF

References

Anderson, J. L. 2007. An adaptive covariance inflation error correction algorithm for ensemble filters. Tellus Ser. A **59**: 210–224. doi:10.1111/j.1600-0870.2006.00216.x

Baehr, M. M., and M. D. DeGrandpre. 2004. In situ pCO_2 and O_2 measurements in a lake during turnover and stratification: Observations and modeling. Limnol. Oceanogr. **49**: 330–340. doi:10.4319/lo.2004.49.2.0330

Baines, S. B., and M. L. Pace. 1991. The production of dissolved organic matter by phytoplankton and its importance to bacteria: Patterns across marine and freshwater systems. Limnol. Oceanogr. **36**: 1078–1090. doi:10.4319/lo.1991.36.6.1078

Balmer, M. B., and J. A. Downing. 2011. Carbon dioxide concentrations in eutrophic lakes: Undersaturation implies atmospheric uptake. Inland Waters **1**: 125–132. doi:10.5268/IW-1.2.366

Berggren, M., L. Ström, H. Laudon, J. Karlsson, A. Jonsson, R. Giesler, A.-K. Bergström, and M. Jansson. 2010. Lake secondary production fueled by rapid transfer of low molecular weight organic carbon from terrestrial sources to aquatic consumers. Ecol. Lett. **13**: 870–880. doi:10.1111/j.1461-0248.2010.01483.x

Catalán, N., R. Marcé, D. N. Kothawala, and L. J. Tranvik. 2016. Organic carbon decomposition rates controlled by water retention time across inland waters. Nat. Geosci. **9**: 501–506. doi:10.1038/NGEO2720

Clark, J. S., and others. 2001. Ecological forecasts: An emerging imperative. Science **293**: 657–660. doi:10.1126/science.293.5530.657

Cole, J. J. 2013, Freshwater ecosystems and the carbon cycle. International Ecology Institute.

Cole, J. J., N. F. Caraco, G. W. Kling, and T. K. Kratz. 1994. Carbon dioxide supersaturation in the surface waters of lakes. Science **265**: 1568–1570. doi:10.1126/science.265.5178.1568

Cole, J. J., S. R. Carpenter, J. F. Kitchell, and M. L. Pace. 2002. Pathways of organic carbon utilization in small lakes: Results from a whole-lake ^{13}C addition and coupled model. Limnol. Oceanogr. **47**: 1664–1675. doi:10.4319/lo.2002.47.6.1664

Cole, J. J., and others. 2007. Plumbing the global carbon cycle: Integrating inland waters into the terrestrial carbon budget. Ecosystems **10**: 172–185. doi:10.1007/s10021-006-9013-8

Coloso, J. J., J. J. Cole, P. C. Hanson, and M. L. Pace. 2008. Depth-integrated, continuous estimates of metabolism in a clear-water lake. Can. J. Fish. Aquat. Sci. **65**: 712–722. doi:10.1139/F08-006

Dee, D. P., and others. 2011. The ERA-interim reanalysis: Configuration and performance of the data assimilation system. Q. J. R. Meteorol. Soc. **137**: 553–597. doi:10.1002/qj.828

Dietze, M. C. 2017, Ecological forecasting. Princeton Univ. Press.

Dietze, M. C., et al. 2018. Iterative near-term ecological forecasting: Needs, opportunities, and challenges. Proc. Natl. Acad. Sci. USA. **115**: 1424–1432. doi:10.1073/pnas.1710231115

Drake, T. W., P. A. Raymond, and R. G. Spencer. 2017. Terrestrial carbon inputs to inland waters: A current synthesis of estimates and uncertainty. Limnol. Oceanogr. Lett. **3**: 132–142. doi:10.1002/lol2.10055

Dugan, H. A., and others. 2016. Consequences of gas flux model choice on the interpretation of metabolic balance across 15 lakes. Inland Waters **6**: 581–592. doi:10.5268/IW-6.4.836

Evans, C. D., M. N. Futter, F. Moldan, S. Valinia, Z. Frogbrook, and D. N. Kothawala. 2017. Variability in organic carbon reactivity across lake residence time and trophic gradients. Nat. Geosci. **10**: 832–837. doi:10.1038/ngeo3051

Evensen, G. 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J. Geophys. Res. **99**: 10143–10162. doi:10.1029/94JC00572

Gao, C., H. Wang, E. Weng, S. Lakshmivarahan, Y. Zhang, and Y. Luo. 2011. Assimilation of multiple data sets with the ensemble Kalman filter to improve forecasts of forest carbon dynamics. Ecol. Appl. **21**: 1461–1473. doi:10.1890/09-1234.1

Giling, D. P., et al. 2017. Delving deeper: Metabolic processes in the metalimnion of stratified lakes. Limnol. Oceanogr. **62**: 1288–1306. doi:10.1002/lno.10504

Grimm, V., and S. F. Railsback. 2005, Indiviual-based modeling and ecology. Princeton Univ. Press. doi:10.1021/jp0450540

Guillemette, F., and P. A. del Giorgio. 2011. Reconstructing the various facets of dissolved organic carbon bioavailability in freshwater ecosystems. Limnol. Oceanogr. **56**: 734–748. doi:10.4319/lo.2011.56.2.0734

Guillemette, F., S. L. McCallister, and P. A. del Giorgio. 2013. Differentiating the degradation dynamics of algal and terrestrial carbon within complex natural dissolved organic carbon in temperate lakes. J. Geophys. Res.: Biogeosci. **118**: 963–973. doi:10.1002/jgrg.20077

Hanson, P. C., A. I. Pollard, D. L. Bade, K. Predick, S. R. Carpenter, and J. A. Foley. 2004. A model of carbon evasion and sedimentation in temperate lakes. Glob. Change Biol. **10**: 1285–1298. doi:10.1111/j.1365-2486.2004.00805.x

Hanson, P. C., K. C. Weathers, and T. K. Kratz. 2016. Networked lake science: How the global Lake ecological observatory network (GLEON) works to understand, predict, and communicate lake eocsystem response to global change. Inland Waters **6**: 543–554. doi:10.5268/IW-6.4.904

Hararuk, O., J. A. Zwart, S. E. Jones, Y. Prairie, and C. T. Solomon. 2018. Model-data fusion to test hypothesized drivers of lake carbon cycling reveals importance of physical controls. J. Geophys. Res.: Biogeosci. **123**: 1130–1142. doi:10.1002/2017JG004084

Hipsey, M. R., and others. 2015. Predicting the resilience and recovery of aquatic systems: A framework for model evolution within environmental observatories. Water Resour. Res. **51**: 7023–7043. doi:10.1002/2015WR017175

Holtgrieve, G. W., D. E. Schindler, T. A. Branch, and Z. T. A'mar. 2010. Simultaneous quantification of aquatic ecosystem metabolism and reaeration using a Bayesian statistical model of oxygen dynamics. Limnol. Oceanogr. **55**: 1047–1063. doi:10.4319/lo.2010.55.3.1047

Huang, J., J. Gao, J. Liu, and Y. Zhang. 2013. State and parameter update of a hydrodynamic-phytoplankton model using ensemble Kalman filter. Ecol. Model. **263**: 81–91. doi: 10.1016/j.ecolmodel.2013.04.022

IPCC. 2013. Carbon and other biogeochemical cycles, p. 465–570. *In* , IPCC, Climate Change 2013: The physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge Univ. Press. doi: 10.1080/09084282.2012.670171

Jones, S. E., J. A. Zwart, P. T. Kelly, and C. T. Solomon. 2018. Hydrologic context constrains lake heterotrophy and terrestrial carbon fate. Limnol. Oceanogr. Lett. **3**: 256–264. doi: 10.1002/lol2.10054

Lahoz, W., B. Khattotov, and R. Menard [eds.]. 2010. Data assimilation and information, p. 3–12. *In* , Data assimilation: Making sense of observations. Springer Science & Business Media.

Li, H., E. Kalnay, and T. Miyoshi. 2009. Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter. Q. J. R. Meteorol. Soc. **135**: 523–533. doi:10.1002/qj.371

Logue, J. B., C. A. Stedmon, A. M. Kellerman, N. J. Nielsen, A. F. Andersson, H. Laudon, E. S. Lindstrom, and E. S. Kritzberg. 2016. Experimental insights into the importance of aquatic bacterial community composition to the degradation of dissolved organic matter. ISME J. **10**: 533–545. doi:10.1038/ismej.2015.131

Masutani, M., and others. 2010. Observing system simulation experiments. *In*, Data assimilation: Making sense of observations, Eds. W.A. Lahoz, B. Khattatov and R. Ménard, Springer, 647–679.

Monteith, D. T., and others. 2007. Dissolved organic carbon trends resulting from changes in atmospheric deposition chemistry. Nature **450**: 537–541. doi:10.1038/nature06316

Mostovaya, A., B. Koehler, F. Guillemette, A. K. Brunberg, and L. J. Tranvik. 2016. Effects of compositional changes on reactivity continuum and decomposition kinetics of lake dissolved organic matter. J. Geophys. Res. Biogeosci. **121**: 1733–1746. doi:10.1002/2016JG003359

Niu, S., Y. Luo, M. C. Dietze, T. F. Keenan, Z. Shi, J. Li, and F. S. C. Iii. 2014. The role of data assimilation in predictive ecology. Ecosphere **5**: 1–16. doi:10.1890/ES13-00273.1

Obrador, B., P. A. Staehr, and J. P. C. Christensen. 2014. Vertical patterns of metabolism in three contrasting stratified lakes. Limnol. Oceanogr. **59**: 1228–1240. doi:10.4319/lo.2014.59.4.1228

Ojala, A., J. L. Bellido, T. Tulonen, P. Kankaala, and J. Huotari. 2011. Carbon gas fluxes from a brown-water and a clear-water lake in the boreal zone during a summer with extreme rain events. Limnol. Oceanogr. **56**: 61–67. doi: 10.4319/lo.2011.56.01.0061

Porter, J. H., P. C. Hanson, and C. C. Lin. 2012. Staying afloat in the sensor data deluge. Trends Ecol. Evol. **27**: 121–129. doi:10.1016/j.tree.2011.11.009

Quay, P. D., S. R. Emerson, B. M. Quay, and A. H. Devol. 1986. The carbon cycle for Lake Washington—a stable isotope study. Limnol. Oceanogr. **31**: 596–611. doi:10.4319/lo.1986.31.3.0596

Core Team, R. 2016, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Available from http://www.R-project.org/. Accessed date: 2018-10-11.

Randerson, J. T., F. S. Chapin, J. W. Harden, J. C. Neff, and M. E. Harmon. 2002. Net ecosystem production: A comprehensive measure of net carbon accumulation by ecosystems. Ecol. Appl. **12**: 937–947. doi:10.1890/1051-0761(2002)012[0937:NEPACM]2.0.CO;2

Read, E. K., and others. 2017. Water quality data for national-scale aquatic research: The water quality portal. Water Resour. Res. **53**: 1735–1745. doi:10.1002/2016WR019993

Read, J. S., and others. 2012. Lake-size dependency of wind shear and convection as controls on gas exchange. Geophys. Res. Lett. **39**: L09405. doi:10.1029/2012GL051886

Schindler, D. E., and R. Hilborn. 2015. Prediction, precaution, and policy under global change. Science **347**: 953–954. doi: 10.1126/science.1261824

Solomon, C. T., and others. 2013. Ecosystem respiration: Drivers of daily variability and background respiration in lakes around the globe. Limnol. Oceanogr. **58**: 849–866. doi:10.4319/lo.2013.58.3.0849

Søndergaard, M., and M. Middelboe. 1995. A cross-system analysis of labile dissolved organic carbon. Mar. Ecol. Prog. Ser. **118**: 283–294. doi:10.3354/meps118283

Soranno, P. A., and others. 2015. Building a multi-scaled geospatial temporal ecology database from disparate data sources: Fostering open science and data reuse. GigaScience **4**: **28**. doi:10.1186/s13742-015-0067-4

Striegl, R. G., P. Kortelainen, J. P. Chanton, K. P. Wickland, G. C. Bugna, and M. Rantakari. 2001. Carbon dioxide partial pressure and ^{13}C content of northe temperate and boreal lakes at spring ice melt. Limnol. Oceanogr. **46**: 941–945. doi:10.4319/lo.2001.46.4.0941

Tranvik, L. J., and others. 2009. Lakes and reservoirs as regulators of carbon cycling and climate. Limnol. Oceanogr. **54**: 2298–2314. doi:10.4319/lo.2009.54.6_part_2.2298

Travis, J., and others. 2014. Integrating the invisible fabric of nature into fisheries management. Proc. Natl. Acad. Sci. USA **111**: 581–584. doi:10.1073/pnas.1402460111

USEPA. 2009, National lakes assessment: A collaborative survey of the nation's lakes. United States Environmental Protection Agency. doi:10.1109/ISBI.2009.5193217

Vachon, D., and Y. T. Prairie. 2013. The ecosystem size and shape dependence of gas transfer velocity versus wind speed relationships in lakes. Can. J. Fish. Aquat. Sci. **70**: 1757–1764. doi:10.1139/cjfas-2013-0241

Vachon, D., and P. A. del Giorgio. 2014. Whole-lake CO_2 dynamics in response to storm events in two morphologically different lakes. Ecosystems **17**: 1338–1353. doi:10.1007/s10021-014-9799-8

Vachon, D., Y. T. Prairie, F. Guillemette, and P. A. del Giorgio. 2016. Modeling allochthonous dissolved organic carbon mineralziation under variable hydrologic regimes in boreal lakes. Ecosystems **20**: 781–795. doi:10.1007/s10021-016-0057-0

Vachon, D., C. T. Solomon, and P. A. del Giorgio. 2017. Reconstructing the seasonal dynamics and relative contribution of the major processes sustaining CO_2 emissions in northern lakes. Limnol. Oceanogr. **62**: 706–722. doi:10.1002/lno.10454

Vallino, J. J. 2010. Ecosystem biogeochemistry considered as a distributed metabolic network ordered by maximum entropy produciton. Philos. Trans. R. Soc. B **365**: 1417–1427. doi:10.1098/rstb.2009.0272

Van de Bogert, M. C., D. L. Bade, S. R. Carpenter, J. J. Cole, M. L. Pace, P. C. Hanson, and O. C. Langman. 2012. Spatial heterogeneity strongly affects estimates of ecosystem metabolism in two north temperate lakes. Limnol. Oceanogr. **57**: 1689–1700. doi:10.4319/lo.2012.57.6.1689

Zwart, J. A., N. Craig, P. T. Kelly, S. D. Sebestyen, C. T. Solomon, B. C. Weidel, and S. E. Jones. 2016. Metabolic and physiochemical responses to a whole-lake experimental increase in dissolved organic carbon in a north-temperate lake. Limnol. Oceanogr. **61**: 723–734. doi:10.1002/lno.10248

Zwart, J. A., S. D. Sebestyen, C. T. Solomon, and S. E. Jones. 2017. The influence of hydrologic residence time on lake carbon cycling dynamics following extreme precipitation events. Ecosystems **20**: 1000–1014. doi:10.1007/s10021-016-0088-6

Acknowledgments

We thank the University of Notre Dame, Environmental Research Center for hosting our study. The chemical analyses were conducted at the Center for Environmental Science and Technology at University of Notre Dame. Synthetic data assimilation analyses were run with the support of Notre Dame's Center for Research Computing. Technical assistance was provided by J. J. Coloso, B. Conner, S. McCarthy, E. Mather, S. Elser, C. J. Humes, J. Lerner, and M. F. Ebenezer. Discussions with P. del Giorgio, A. M. Raiho, participants in the Near-Term Ecological Forecasting 2018 summer course, members of the Jones Lab, and comments from two anonymous reviewers significantly improved the manuscript. Our work on this project was supported by a team grant from the Fonds de Recherche du Québec – Nature et Technologies (FRQNT 191167) to C.T.S. and Y.T.P. and P. del Giorgio, and also supported by the National Science Foundation Graduate Research Fellowship under NSF DGE-1313583 and NSF Earth Sciences Postdoctoral Fellowship under NSF EAR-PF-1725386 to J.A.Z. and NSF award DEB-1547866 to S.E.J. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Conflict of Interest

None declared

Submitted 24 April 2018
Revised 11 October 2018
Accepted 08 December 2018

Associate editor: Gordon Holtgrieve