Journal of Great Lakes Research 46 (2020) 144–158

Contents lists available at ScienceDirect

Journal of Great Lakes Research

journal homepage: www.elsevier.com/locate/ijglr

# Improved thermal structure simulation and optimized sampling strategy for Lake Erie using a data assimilative model

Xinyu Ye [a], Philip Y. Chu [c], Eric J. Anderson [c], Chenfu Huang [a], Gregory A. Lang [c], Pengfei Xue [a,b,*]

[a] Department of Civil and Environmental Engineering, Michigan Technological University, Houghton, MI, United States
[b] Great Lakes Research Center, Michigan Technological University, Houghton, MI, United States
[c] NOAA – Great Lakes Environmental Research Laboratory, Ann Arbor, MI, United States

ARTICLE INFO

*Article history:*
Received 27 September 2018
Accepted 16 October 2019
Available online 11 January 2020

Communicated by Leon Boegman

*Keywords:*
Data assimilation
Thermal structure
Hydrodynamic modeling
Lake Erie
Great lakes

ABSTRACT

Lake Erie has experienced substantial environmental issues (e.g., hypoxia, harmful algal blooms) for decades, which are closely related to the lake's thermal characteristics. While three-dimensional (3D) hydrodynamic models have been widely applied to Lake Erie, challenges remain due to model representation of physical processes, errors and uncertainty in boundary conditions and forcing terms. The Great Lakes region has a relatively dense and long-term observational record, and these observational data have been used for model initialization and verification, but have not been incorporated into 3D model simulations through data assimilation (DA) to create reanalysis products or improve short-term forecasts. In this work, we developed and evaluated DA to improve thermal structure simulation of Lake Erie. Moored instrument data and satellite data are incorporated into a data-assimilative hydrodynamic model for analysis and evaluation. Results show that DA can effectively improve the model performance to create reanalysis fields when the DA formulation is appropriately developed in recognition of the dynamic complexities and anisotropic error covariances of Lake Erie. The data assimilative model also improves forecasting accuracy and restrains forecasting uncertainty to an acceptable level on a timescale of 1–7 days after being unleashed from DA. Lastly, data sampling strategies based on an error correlation map are examined. Results show the method can effectively reduce the sampling effort while still achieving similar model skills with potential for optimal design of an observation network or field sampling strategy.


## Introduction

Lake Erie, the shallowest and southernmost of the Great Lakes, has experienced substantial environmental issues for decades including large-scale low-oxygen (hypoxic) conditions (Burns et al., 2005; Daloglu et al., 2012) and algal blooms (Conroy et al., 2005; Michalak et al., 2013; Chaffin et al., 2013). The occurrences of hypolimnetic hypoxia, algal blooms and substantial eutrophication are closely related to Lake Erie's limnological and thermal characteristics. In the central basin, where the average water depth is 18.3 m, the position of the thermocline varies significantly from year to year in response to changes in meteorological conditions. This is a key factor affecting the development of hypoxia (Schertzer et al., 1987; Zhou et al., 2013). The western basin is the shallowest basin with an average depth of 7.3 m and reduced stratification (Schertzer et al., 1987). Water temperature and excessive nutrient loading significantly influence the timing and magnitude of phytoplankton blooms in this portion of the lake (Nicholls and Hopkins, 1993; Arnott and Vanni, 1996; Makarewicz et al., 1999; Vanderploeg et al., 2001; Smith et al., 2005; Millie et al., 2009; Chaffin et al., 2013).

To better understand Lake Erie's biophysical characteristics, accurate estimates of lake surface temperature (LST) and thermal structure are required. Significant advancements of 3D hydrodynamic models for the Great Lakes have been made since the nineties (Schwab and Bedford 1994; Beletsky et al., 2006; Wang et al., 2010; Huang et al., 2010; Chu et al., 2011; Fujisaki et al., 2013; Xue et al., 2015; Anderson et al., 2015; Xue et al., 2017, 2018; Ye et al., 2019; Kelley et al., 2018; Huang et al., 2019). However, challenges remain due to errors in model representation of physical processes, boundary conditions, or forcing terms. Ecologically, even relatively small changes in the thermal characteristic of lakes can cause significant shifts in phytoplankton, bacterioplankton, zooplankton populations, and associated metabolic processes in aquatic ecosys-

* Corresponding author at: Department of Civil and Environmental Engineering, Michigan Technological University, Houghton, MI, United States.
*E-mail address:* pexue@mtu.edu (P. Xue).

https://doi.org/10.1016/j.jglr.2019.10.018
0380-1330/

tems. (Tulonen et al., 1994; Drinkwater et al., 2003; Adrian et al., 2009; Arvola et al., 2009). Therefore, further improvement in simulating the lake's physical conditions is of great importance.

With regards to data, the Great Lakes region has a fairly dense and long-term observational record of meteorological and physical variables, as compared to other coastal seas and deep oceans. In-situ measurements and remotely sensed data have been widely used for model initialization and verification, but have not been incorporated into 3D model simulations to improve short-term forecasts or create reanalysis (Hawley et al., 2006; Zhang et al., 2007). Data assimilation (DA) is the most effective approach for statistically combining observational data and model dynamics to provide the best estimate of system state (Robinson and Lermusiaux, 2000; Li et al., 2008a,b; Chao et al., 2009; Houtekamer and Zhang 2016; Bannister et al., 2017). Furthermore, despite significant development of various observing approaches, time and space coverage of observational datasets, particularly subsurface measurements, remains very limited due to the high costs of building and maintaining observing networks. Optimized data sampling design is, therefore, one of the key research topics to the success of an integrated observing and forecasting system for the Great Lakes. DA can also be used to design appropriate monitoring, field sampling and management programs (Bishop et al., 2001; Zhang et al., 2007; Xue et al., 2011, 2012; Hoffman and Atlas 2016).

The purposes of this study are to develop and evaluate combined modeling-DA approaches for Lake Erie to improve lake thermal structure simulations, create hydrodynamic reanalysis products, improve the accuracy of short-term forecasts, and to provide guidance for optimizing sampling strategies. Various DA schemes have been developed and applied to ocean and atmosphere model systems, such as nudging, optimal interpolation (OI), three- or four-dimensional variational analysis (3D- or 4D-VAR), Kalman Filter (KF) and their variants. They differ in computational cost and optimality, and in their suitability for real-time data assimilation. From the perspective of estimation theory, nudging and OI can be regarded as simplified schemes of KF with empirically assigned gain matrix (Robinson and Lermusiaux, 2000), while in KF, the analysis gain is computed internally and updated continually. As the first step toward incorporating data assimilative capability into the National Oceanic and Atmospheric Administration (NOAA) Great Lakes Operational Forecasting System (GLOFS), we focus on nudging and OI data assimilation because both methods are fairly straightforward, powerful and computational efficient. They are often the preferred choices in the early stages of data-assimilative system development for a balance between efficiency, effectiveness, and flexibility in implementation.

## Observational data and hydrodynamic model

We use both moored instrument data and satellite data for analysis, assimilation, and verification. Specifically, the in-situ data are both assimilated and used for verification (withheld data), while the satellite-based LST is used for verification and for construction of the background error covariance matrix. Moored instrument data include continuous vertical temperatures profiles during summertime from a series of 13 thermistors deployed in Lake Erie during 2005 (Fig. 1a–c), as part of the International Field Years on Lake Erie (IFYLE) (Brandt and Lansing 2006). IFYLE was initiated and managed by NOAA's Great Lakes Environmental Research Laboratory (GLERL), in collaboration with researchers from the U.S., Canada, and Europe (www.glerl.noaa.gov/res/projects/ifyle/).

Satellite data used in this study are from the NOAA CoastWatch GLSEA (Great Lakes Surface Environmental Analysis) product. GLSEA uses cloud-free portions of Advanced Very High-Resolution Radiometer (AVHRR) SST imagery to create a daily digital map at 1024 × 1024 pixels. It utilizes linear geo-correction and a cell-based interpolation and extrapolation procedure, and has seven daytime and five nighttime cloud masks. The GLSEA data are updated daily with information from the cloud-free portions of the satellite imagery. A smoothing algorithm is applied to the map for days when no imagery is available. Detailed validation is documented by (Schwab et al., 1992, 1999). A comparison between GLSEA data and thermistor in-situ observations near the surface shows the GLSEA has accurate representation of LST (Fig. 2).

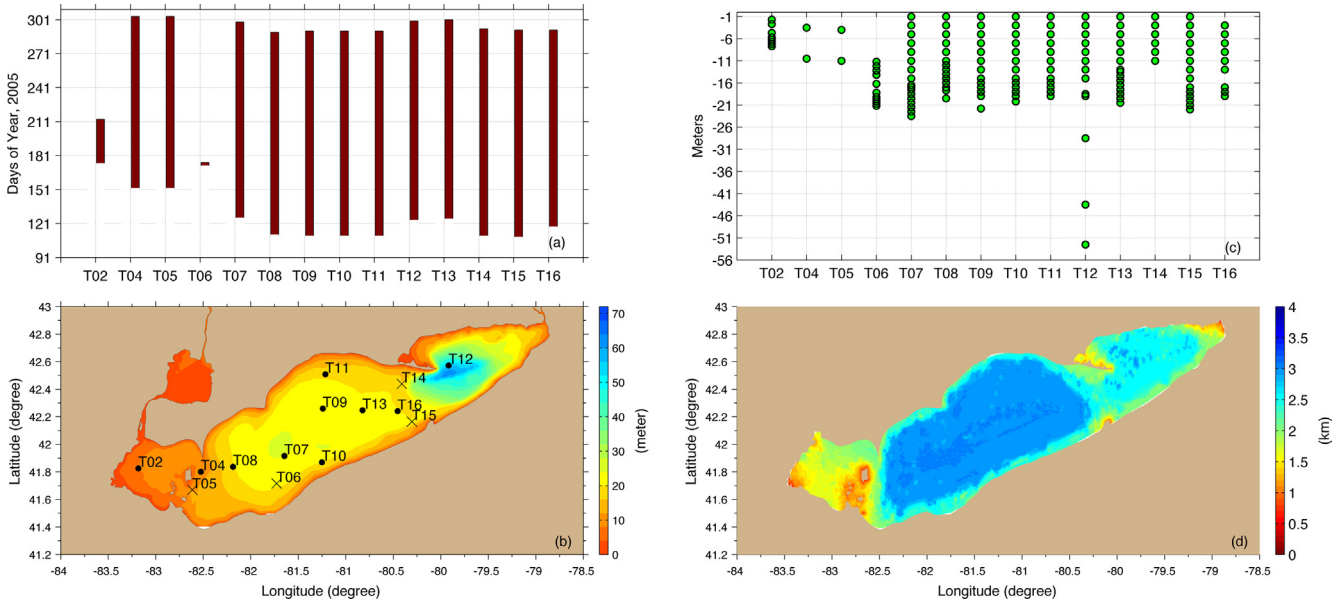


**Fig. 1.** Panel (a): Lake Erie thermistor data coverage for 2005; Panel (b): bathymetry with thermistor observations, data that are used for assimilation in various experiments are marked as black dots and data from stations (T05, T14, T15 noted with cross marks) are not assimilated into the model but only used to verify the regional impact of DA. Summer observational data is not available at T06; Panel (c): Vertical resolution of thermistor data; Panel (d): FVCOM model grid resolution.

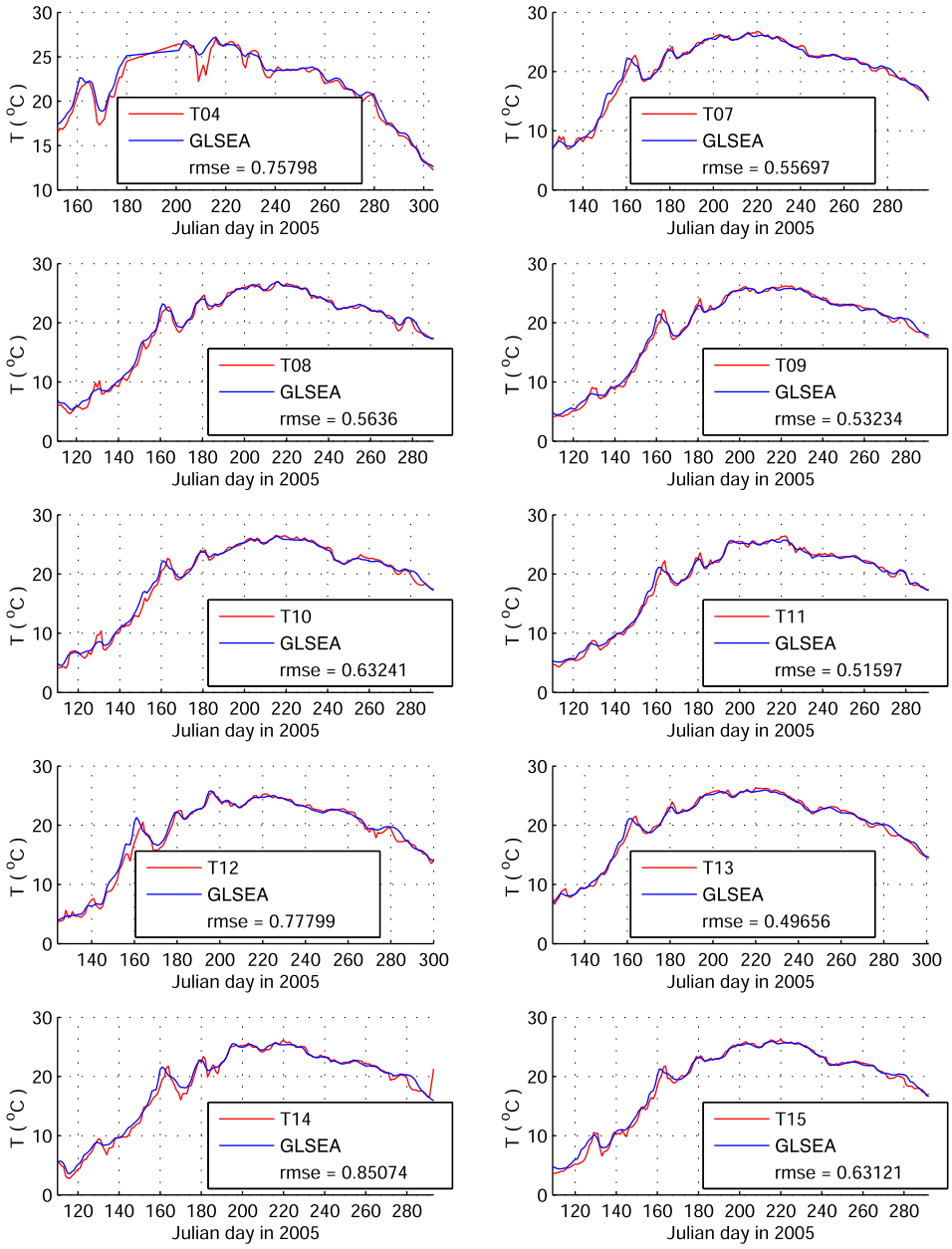


**Fig. 2.** Comparison between GLSEA data and thermistor in-situ observations when the thermistor data are available at 1-m from the surface. Notice T04 is measured at 3.5 m from the surface.

The hydrodynamic model is based on the NOAA Lake Erie Operational Forecast System (LEOFS; Kelley et al., 2018), a real-time nowcast and forecast system based on the Finite Volume Community Ocean Model (FVCOM; Chen, 2006). FVCOM is an oceanographic hydrodynamic model that solves the three-dimensional integral form of the governing equations on an unstructured, sigma-coordinate mesh. FVCOM has been applied in many coastal systems, including successful adaptation and implementation into the Great Lakes (Anderson and Schwab, 2013; Anderson et al., 2015; Xue et al., 2015; Anderson and Schwab, 2017; Xue et al., 2017; Ye et al., 2019; Huang et al., 2019)). In the upgraded NOAA operational model for Lake Erie (Kelley et al., 2018), the FVCOM model is developed with horizontal resolution ranging from 200 to 2500 m (Fig. 1d), and 21 vertical sigma (terrain-following) layers. The inflow and outflow to the lake are established via open-boundaries at connecting channels, Detroit and Niagara Rivers. At the Detroit River, the water level is prescribed along the open boundary from the NOAA National Ocean Service (NOS) water level gauge at Gibraltar, MI (9044020) using 6-minute observed data. The outflow is established by prescribing 6-minute water level at the head of the Niagara River, using a dynamic offset from observed water levels from the NOS gauge at Buffalo (9063020)

For this study, the model was run in hindcast mode using hourly surface forcing meteorology, interpolated from temporally- and spatially-varying coastal weather stations and in-lake buoys. This procedure has been adopted in many studies of Great Lakes hydrodynamics (Schwab and Bedford, 1994). Model simulation was carried out for the year 2005 using surface conditions for 10-meter wind, 2-meter air temperature, 2-meter dew-point temperature, and total cloud cover supplied to FVCOM. The LEOFS implementation of FVCOM uses the SOLAR heat flux subroutine (Liu and Schwab, 1987), developed specifically for the Great Lakes (Beletsky et al., 2003; Anderson and Schwab, 2013; Rowe et al., 2015). The model was run with both external and internal

mode time steps of 10 s. FVCOM uses the Mellor and Yamada level 2.5 (MY-2.5) and Smagorinsky turbulent closure schemes as default configurations for vertical and horizontal mixing, respectively (Mellor and Yamada, 1982; Smagorinsky, 1963).

## DA algorithms and design of numerical experiments

### *Assimilation schemes*

A DA system consists of three components: a set of observations, a dynamical model, and an assimilation scheme (Robinson and Lermusiaux, 2000). With the notation conventions suggested by Ide et al. (1997), vectors are represented by boldface lowercase letters, and matrices by boldface uppercase letters. We use the unbolded lowercase letters for scalars. The superscript $''f'', ''a'', ''t''$ stand for the forecast, the analysis, and the truth, respectively. The superscript "T" and "-1" denote matrix transpose and inverse. Consider a system with a model state vector $\boldsymbol{x}$ (in our case, it represents temperature values at n model grids in 3-D space),

$$\boldsymbol{x} = (x_1, x_2, \cdots, x_n)^T$$

Similarly, we denote the observation vector $\mathbf{y}$ in 3-D space,

$$\boldsymbol{y} = (y_1, y_2, \cdots, y_m)^T$$

where $m$ is the number of observational data points in 3-D space. An observation operator, $\boldsymbol{H}$, is required to map the model state to the observation locations to evaluate the model-data misfit, denoted as a vector $\boldsymbol{d} = (d_1, d_2, \cdots, d_m)^T$,where $\boldsymbol{d} = \boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}^f$. The operator $\boldsymbol{H}$ can be a linear or nonlinear function of the model variables, although it could often be a simple interpolation operator.

Sequential data assimilation provides a state estimate on an ongoing basis, iteratively alternating between a forecast (simulation) step and a state estimation (assimilation) step; the latter step is often called the "analysis" (Hunt et al., 2007). The analysis step at a given time $t_i$ combines results from a priori forecast (simulation) ($\boldsymbol{x}^f(t_i)$) and information from current observations $\boldsymbol{y}(t_i)$ to produce a current state estimate ($\boldsymbol{x}^a(t_i)$). This estimate $\boldsymbol{x}^a(t_i)$ is then used as initial condition for the next simulation cycle $\boldsymbol{x}^f(t_{i+1})$), which is subsequently used for estimating the next analysis $\boldsymbol{x}^a(t_{i+1})$, when new observations $\boldsymbol{y}(t_{i+1})$ are available.

To estimate the state of the system and reduce model simulation error, an essential component of *a priori* hypothesis is to define the statistics of model and observation errors with respect to the true state using error covariance matrices (Bouttier and Courtier, 2002). For an assimilation cycle (e.g. at a given time $t_i$), the optimal linear least-squares estimation theory states:

$$\boldsymbol{x}^a = \boldsymbol{x}^f + \boldsymbol{K}\boldsymbol{d} \tag{1}$$

$$\boldsymbol{K} = \boldsymbol{B}\boldsymbol{H}^T\left(\boldsymbol{H}\boldsymbol{B}\boldsymbol{H}^T + \boldsymbol{R}\right)^{-1} \tag{2}$$

where $\boldsymbol{x}^f$ and $\boldsymbol{x}^a$ are the forecast (simulation) state and analysis state as defined above. The linear operator $\boldsymbol{K}$, is the gain (or weight) matrix of the analysis. The weight matrix is determined by the model error covariance matrix $\boldsymbol{B}$ (referred to in Bouttier and Courtier, 2002 as the background error covariance) that is projected onto the observations by $\boldsymbol{H}$ and $H^T$, the observation error covariance matrix $\boldsymbol{R}$, and the misfit between the model simulation and observation $\boldsymbol{d}$. Under different assumptions and approximations, this leads to various sequential assimilation schemes such as nudging, OI and Kalman Filter (For detailed mathematical formulation, please see Lorenc, 1986; Ghil and Malanotte-Rizzoli, 1991; Ide et al., 1997; Robinson and Lermusiaux, 2000). In this study, we approximate and simplify it to two variants of nudging and OI with localization.

It is commonly assumed there is an influence area for each of the observations (i.e. localization), and if observations are sporadic and only have local influence, the background correlation does not necessarily need to be specified globally (Bouttier and Courtier, 2002). That is, for a given observation $y_j$, [j=1, 2 ⋯,m], only model results on the grids within the influence area of observation $y_j$ should be corrected based on the model-data misfit, $d_j$, [j=1, 2 ⋯, m].

### *Correction of 3D temperature fields using nudging assimilation*

In the nudging method, the analysis equation is simplified as,

$$\boldsymbol{x}^a_{m_i} = \boldsymbol{x}^f_{m_i} + K_{m_i} d_j \tag{3}$$

where $i = 1, 2, \cdots L$, and $m_1, m_2, ..m_L$ denote total L model grids within the influence area of a given observation station $y_j$, and $K_{m_i}$ is a product of weight functions expressed as:

$$K_{m_i} = \mathrm{w\_xy}_{m_i} \cdot \mathrm{w\_z}_{m_i} \cdot \mathrm{w\_t}_{m_i} \cdot \mathrm{G}_\alpha \tag{4}$$

where $\mathrm{w\_xy}_{m_i}$, $\mathrm{w\_z}_{m_i}$, $\mathrm{w\_t}_{m_i}$ are functions describing horizontal, vertical, and temporal temperature correlation, respectively, between the location of model grid $m_i[i = 1, 2, \cdots L]$ and the location of observation $y_j$. A nudging coefficient ($\mathrm{G}_\alpha = 7.6e^{-3}$) is used for each time step. Empirical correlation functions are constructed based on the following principles (Bouttier and Courtier, 2002): 1) the correlation functions must be smooth and dynamically constrained in physical space; 2) the correlation functions should be zero for long-distance separations if observations have only a local effect on the DA analysis; and 3) three-dimensional correlation can be built by combining separability considering different correlation physical scales on the horizontal and vertical.

Horizontally, for each observation, the localized correlation function w_xy is parameterized in two steps. In the first step, an empirical maximum characteristic correlation distance of 20 km (note that we also tested other empirical values of 15 km and 25 km; however, the results were similar or less desirable) is prescribed using the Cressman formula, a distance-dependent function that decays to zero beyond the specified correlation distance, so it shows decaying correlation strength from the center to 20 km away (Houtekamer and Mitchell, 2001; Hunt et al., 2007). Therefore, an observation only affects the model state within such a distance in DA analysis. A vector $\boldsymbol{c}_r$stores the Cressman-based correlation strength for all model grids ($m_1, m_2, ..m_L$) located within the 20 km correlation distance. This only represents an isotropic horizontal correlation pattern without considering the local physical process. In the second step, a local vector $\boldsymbol{c}_l$ for the L model grid points ($m_1, m_2, ..m_L$) within the 20 km correlation distance is used to store the local correlation pattern between each of these model grids and the observation location. The estimation of $\boldsymbol{c}_l$ is based on the Pearson correlation coefficient using the time series of model simulation of temperature over the summer 2005 for each model grid. For example, the $i$th element of vector $\boldsymbol{c}_l$ stores the correlation coefficient between the model grid point $m_i$ and the observation location, which is calculated based on the modeled temperature time series from the summer 2005 at the model grid point $m_i$ and the observation location. The final correlation w_xy is formed as $\boldsymbol{c}_r$ o $\boldsymbol{c}_l$, where the operator "o" is the Hadamard Element-wise multiplication.

Two examples of the formulated empirical horizontal correlation w_xy at two observations in the central basin (T09) and the coastal region (T10) are shown in Fig. 3a. The correlation scale is more isotropic in the central basin, while anisotropic correlation is formed in the nearshore region, with more significant correlation in the longshore direction along isobaths and weaker correlation across the local isobath.

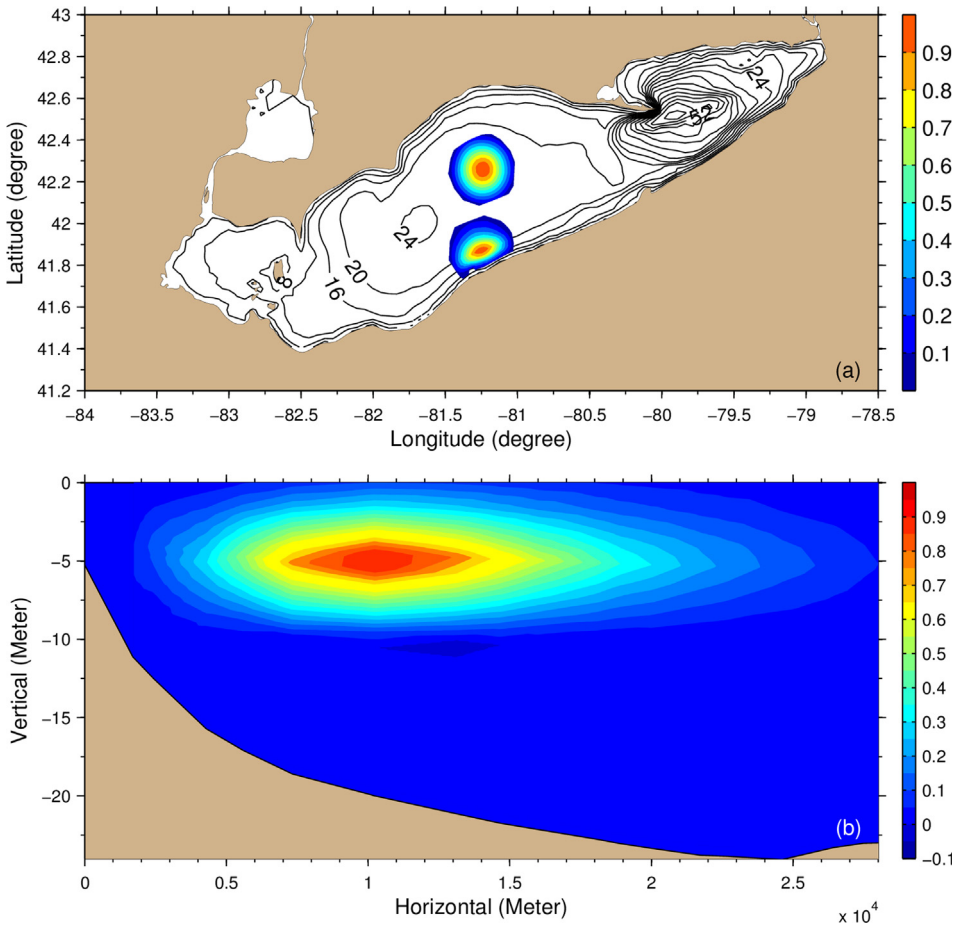


**Fig. 3.** Examples of the horizontal error correlation pattern at two observations at the central basin (T09) and the coastal region (T10) (a), and vertical correlation pattern at a depth of 5 m (b).

The vertical correlation w_z is specified explicitly using the Cressman formula with an impact radius of 5 m. The impact radius of 5 m is selected empirically based on the vertical correlation analysis at observations T04, T05 and T12 (Fig. 1c) where the vertical resolution of observations is relatively coarse. Sensitivity analysis also suggests the selection of a vertical impact radius is not very sensitive in our experiments because the vertical resolution of thermistor data is high (1–2 m) at most observation locations (Fig. 1c). It should be noted that the selection of a vertical impact radius needs to be further tested, particularly for cases when the vertical resolution of observational data is low. An underestimated impact radius could limit the ability of the observation to correct model bias beyond that length and lead to an incorrect shut down of convective mixing (Scott et al., 2018) while an overestimated impact radius could spread observational information to an extent that may be non-physical. The vertical and cross-shore correlation pattern for an observation in the coastal region is demonstrated in Fig. 3b.

Lastly, the temporal weighting function is defined such that the observation is given full weight over the first half of the assimilation window, with the weight decreasing linearly to zero over the second half of the assimilation window. The temporal weighting functions is defined as

$$w\_t = \begin{cases} 1, & |t - t_0| < T_w/2 \\ \frac{T_w - |t - t_0|}{\frac{T_w}{2}}, & T_w/2 \le |t - t_0| \le T_w \\ 0, & |t - t_0| > T_w \end{cases} \quad (5)$$

where $T_w$ is the assimilation time window (=96 h) and $t_0$ is the observation acquisition time.

*Further correction of the LST using OI assimilation*

LST is one of the most important physical variables for physical, biological, and climate studies for the Great Lakes. Limited in-situ observations can be available real-time, but are often too sparse to make direct spatiotemporal mapping of LST over the entire lake. One of our interests is how to use the information from the GLSEA to improve spreading of information from sparse, real-time, in-situ observations (e.g. thermistor data) over the lake surface. To that end, we develop an OI assimilating in-situ observations for further correction of the LST. The thermistor data are assimilated vertically using nudging, and then the corrections from the nudging at observation locations are further spread in the horizontal at the lake surface using OI.

For the OI, we first estimate a stationary model error covariance matrix $\boldsymbol{B}$ for the LST (notice $\boldsymbol{B}$ only needs to contain the 2-D horizontal information as it is for the correction of LST) using the difference between modeled LST results and GLSEA in the hindcast simulation of 2002–2004 (prior to the assimilation year 2005), and apply Eqs. (1) and (2) to conduct data assimilation for 2005, which allows us to make a correction of the LST by assimilating the in-situ thermistor data sampled in 2005. It is noted that the GLSEA data are never directly assimilated into our model. The GLSEA data in 2002–2004 is used to estimate the model error

covariance matrix **B**; the GLSEA data in 2005 is used for model-data comparison to evaluate the performance of data assimilative model. The model error covariance matrix $\boldsymbol{B}$ is defined as $\boldsymbol{B} = \overline{(\boldsymbol{x}^f - \boldsymbol{x}^t)(\boldsymbol{x}^f - \boldsymbol{x}^t)^T}$, where $\boldsymbol{x}^f$ and $\boldsymbol{x}^t$ denotes the model state and true state. Here we have used the GLSEA data as proximity to the true state $\boldsymbol{x}^t$ of LST for the summers of 2002–2004 in this experiment. Observation error covariance matrix $\boldsymbol{R}$ is assumed diagonal with the main diagonal value set as the square of the RMSE between GLSEA and thermistor data near the surface as shown in Fig. 2.

For the localization, we still follow the assumption that each observation has its own influence area, and we construct more natural, irregular influence areas that vary for each of the observations based on the strength of error covariance. In this case, the analysis equation becomes

$$\boldsymbol{x}^a_{m_i} = \boldsymbol{x}^f_{m_i} + \mathrm{corr}\left(m_i, m_{y_j}\right) \cdot \frac{\sigma_{m_i} \cdot \sigma_{m_{y_j}}}{\sigma^2_{m_{y_j}} + \sigma^2_{y_j}} \cdot d_j \tag{6}$$

where $i = 1, 2, \cdots L$, and $m_i$denote the total $L$ model grids within the influence area of a given observation station $y_j$; $\mathrm{corr}(m_i, m_{y_i})$ is the correlation between modeling error at model grid $m_i$ and modeling error at the location of observation $y_j$; $\sigma_{m_i}$ and $\sigma^2_{m_{y_j}}$ are variances of modeling error at grid $m_i$ and at the location of observation $y_j$; $\sigma^2_{y_j}$ is the observation error variance (the main diagonal value of $\boldsymbol{R}$) at observation $y_j$. The detailed derivation of Eq. (6) is provided in Appendix A. Eq. (6) shows that the DA can be directly associated with the error correlation map, which is detailed in Section 4.

## Design of numerical experiments

To analyze the DA effectiveness and efficiency in supporting model hindcasts, forecasts, and to assist in data sampling design, five numerical experiments are designed and carried out. All of the experiments use the same atmospheric forcing and hydrodynamic model configuration.

The first case is a control run (CR). In this case, the model is set up as a standard simulation with the default configuration of LEOFS. The model simulation starts from January 1st, 2002 and continues until September 30th, 2005. LST simulation results in the summers (July-September) of 2002–2004 are evaluated against GLSEA data to estimate the model error covariance matrix **B**; simulation results in year 2005 between July and September serve as a baseline for DA evaluation.

In case DA#1 (Data Assimilation case #1), the data assimilative model is configured using the nudging method. The model is restarted from July 1st, 2005 and run until the end of September using the CR restart file to make the two cases comparable for the analysis period of summer 2005. Assimilation is carried out at an approximately hourly interval, when the observational data are available. DA#1 is designed to evaluate the impact of DA on improving the model hindcast accuracy.

In case DA#2 (Data Assimilation case #2), the data assimilative model is configured using the OI approach for LST correction. This case is also restarted from July 1st, 2005 and run until the end of September using the CR restart file to make the two cases comparable for DA evaluation. DA#2 is designed to utilize historical remote-sensed GLSEA data to improve assimilation of scattered, real-time in-situ observations (e.g. thermistor data) and test optimal data sampling strategy for LST assimilation.

In cases of FC#1 (Forecast#1) and FC#2 (Forecast#2), the model is configured the same as in the case of DA1 and DA2, respectively. The assimilation is turned off on August 31, 2005. Afterward, the model is run for one month without DA (representing a forecast scenario), from September 1st to 30th. September 1st is selected because it is when the modeling simulation error from the control run starts to amplify quickly and results in large model bias. These two cases are designed to evaluate the impact of DA (when data are available for assimilation at the hindcast and nowcast stages, i.e., before September 1st) on improving short-term forecasting accuracy (no observational data available for forecasting, i.e., after September 1st). In all cases, model outputs have been interpolated to the same grid as the observations so direct comparisons can be made.

A summary of the configuration of these experiments is presented in Table 1.

## Results

In Lake Erie, the LST shows large seasonal and interannual variability (Fig. 4). The lake typically has the lowest water temperature during February and March, and the water temperature reaches its peak in August (occasionally July or September) with the LST of ~21–26 °C during the summer time. The interannual variability of the lake-mean summer LST can vary ± 5 °C relative to its climatological mean (Fig. 4-a). The most interannual variability is observed in the central basin with a standard deviation of >0.85 °C, while the shallow western basin shows a slightly lower interannual variability with a standard deviation of 0.75 °C (Fig. 4c). The spatial pattern of LST climatology is mainly latitude- and depth-dependent (Fig. 4b). The shallow western basin and the southwest coastal region are characterized with the highest LST of ~22.5–23 °C while less warm water (21.5–22.5 °C) exists in the majority of central basin and eastern basin. A band of relatively cold water (<21 °C) occupies the northern coast, approximately constrained within the local isobaths of the 0–20 m. The year of 2005 is one of the warmest years with relatively high summer LST.

The model results from the control run (CR) serves as a baseline to evaluate the model skill. The model configuration is described above. In the CR experiment, the simulated lake-mean LST shows good agreement with the GLSEA data (Fig. 5a,b,d). The model simulations reproduce the LST seasonal cycles in both magnitude and phase, and also capture cooling and warming events on synoptic time scales. The summer mean LST is 23.88 °C from GLSEA and 24.22 °C from the CR experiment with a root-mean-square-error (RMSE) of 0.85 °C over the summer. On a closer comparison, the major errors from the CR simulation during the summer arise when LST starts to decrease in September, with a noticeable warm bias up to 1.5 °C (RMSE of 1.21 °C). Spatially, a warm bias up to 2 °C is shown in the central basin and western basin, while a band of cold "bias" up to 2.5 °C is shown in the northern coast. (we note

**Table 1**
Summary of the design of numerical experiments.

| Experiment name | Assimilation scheme | Data assimilated | Experiment purpose |
|---|---|---|---|
| CR | Free run | | Baseline |
| DA#1 | Nudging | 10 out of 13 thermistor data (T02,T04,T07,T08,T09,T10, T11,T12,T13,T16) | Hindcast improvement |
| DA#2 | OI | 5 out of 13 thermistor data (T04, T07, T08, T11, T12) | Optimization of data sampling design |
| FC#1 | Nudging and free run | 10 out of 13 thermistor data (T02,T04,T07,T08,T09,T10, T11,T12,T13,T16) | Forecast improvement |
| FC#2 | OI and free run | 5 out of 13 thermistor data (T04, T07, T08, T11, T12) | Forecast improvement |

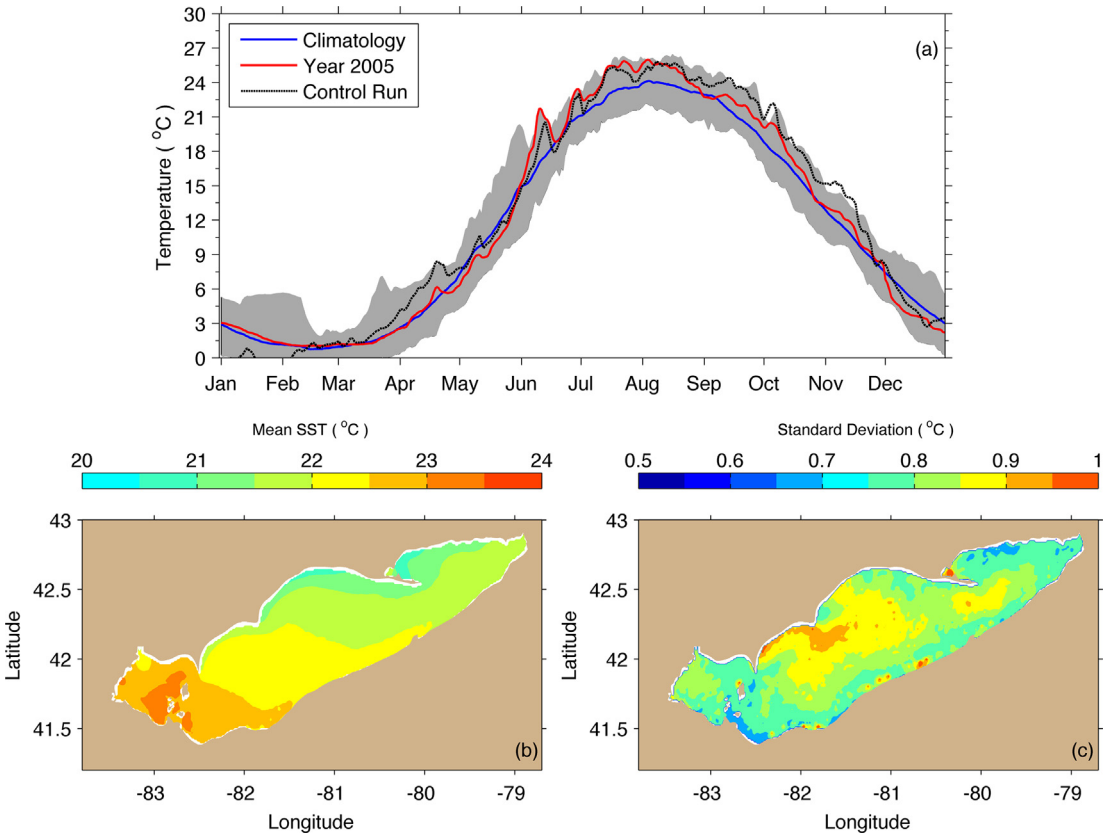


**Fig. 4.** Twenty-year climatology (1995–2017) envelopes of observed lake mean surface temperature from GLSEA (gray ribbon) with the climatological mean (blue line), LST of 2005 (red line), and model simulation results from CR (black line) in upper panel (a), climatological spatial pattern of LST of 1995–2017 (b) and corresponding standard deviation(c). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

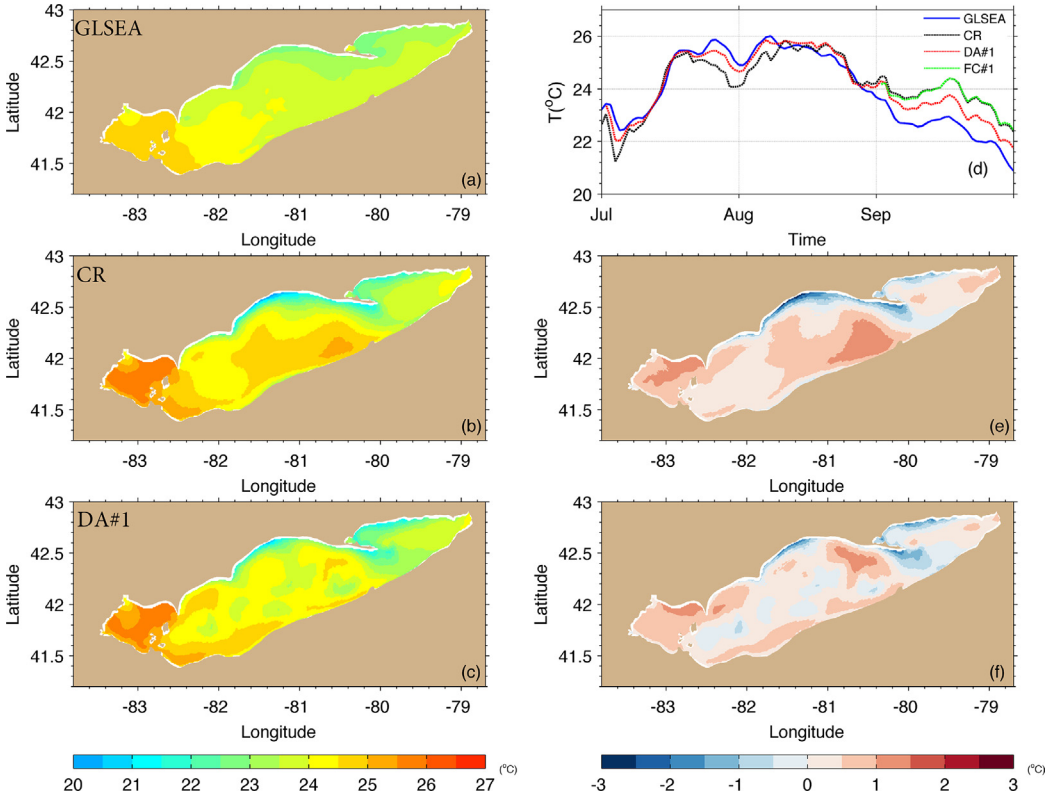


**Fig. 5.** Spatial pattern of LST during 2005 from GLSEA data(a), CR experiment (b), DA#1 experiment (c), lake-mean LST for four cases (d), model bias (model-GLSEA) for CR (e), for DA#1 (f).

**Table 2**
Skill Scores (SS) of modeled temperature profiles at the locations of the thermistor moorings in different numerical experiments. No observation available from T02 for forecasting comparison.

| Stations | SS in DA#1 | SS in DA#2 | SS in FC#1 | SS in FC#2 |
|---|---|---|---|---|
| T02 | 0.8631 | 0.0009 | N/A | N/A |
| T04 | 0.2402 | 0.5805 | 0.0851 | 0.5797 |
| T05 | 0.4314 | 0.2076 | 0.3883 | 0.5101 |
| T07 | 0.717 | 0.9652 | −0.0932 | −0.024 |
| T08 | 0.804 | 0.9856 | 0.7888 | 0.5652 |
| T09 | 0.8965 | 0.4899 | 0.2578 | 0.463 |
| T10 | 0.6859 | 0.1942 | 0.23 | 0.136 |
| T11 | 0.6189 | 0.967 | 0.3691 | 0.4005 |
| T12 | 0.783 | 0.9831 | 0.7919 | 0.6263 |
| T13 | 0.9373 | 0.2753 | 0.7441 | −0.154 |
| T14 | 0.5533 | 0.0108 | −0.2231 | 0.4023 |
| T15 | −0.1534 | 0.3354 | 0.3864 | 0.4113 |
| T16 | 0.8062 | −0.2996 | 0.6615 | −0.0588 |

there have been discussions that the nearshore cold band may be due to realistic upwellings but need to be further validated with in-situ observations).

The DA#1 experiment answers the question whether or not the DA can improve model simulation through the assimilation of thermistor data, in the sense of enhancing model accuracy not only at the observation locations but also on regional scales. Results shows the nudging method can effectively improve model simulations in LST (Fig. 5c,d,f) and thermal profile near observations when local decorrelation scale is properly specified. In DA#1, the data assimilative model provides an improved lake-mean LST over the entire summer (RMSE is reduced to 0.46 °C). In particular, the large warm bias during September is reduced by ~50% with considerably improved model accuracy on the lake-wide scale (Fig. 5d) with a RMSE reduced to 0.68 °C. Spatially, the aforementioned warm bias in the central basin is effectively removed (Fig. 5e, f) with a remnant bias below 0.5 °C in the majority of the central basin. In addition, the warm bias in the western basin is also reduced by 15%.

Modeled vertical thermal structures with and without data assimilation are also evaluated in comparison to observed thermistor data. Skill Score (SS) is used to quantitatively assess the improvement of modeled temperature profiles with DA runs over the CR (Table 2). The skill score (Murphy, 1988) is defined as:

$$SS = 1 - \frac{\overline{(DA - O)^2}}{\overline{(CR - O)^2}} \tag{7}$$

where DA represents results from the assimilation run, CR represents model results from the control run, and O stands for observations. A positive SS indicates the DA improvement over the control run, SS = 0 indicates no improvement, and a negative SS indicates the assimilation result became worse than control run. Table 2 shows that the assimilation in DA#1 improved 12 out of 13 stations, 10 stations received a skill score >0.5, i.e. the mean-square errors in these locations have been reduced at least 50%. It is noted that improvement is seen at stations (T05, T14) where data are not assimilated into the model.

Figs. 6–8 illustrate DA improvement at three different stations where model results from the CR experiment shows different error patterns. For example, in the eastern basin, the observation data (T12) shows strong near-surface (0–20 m) warming between July and September with a sharp thermocline, below which the water temperature drops abruptly below 6 °C (Fig. 6a). In the CR

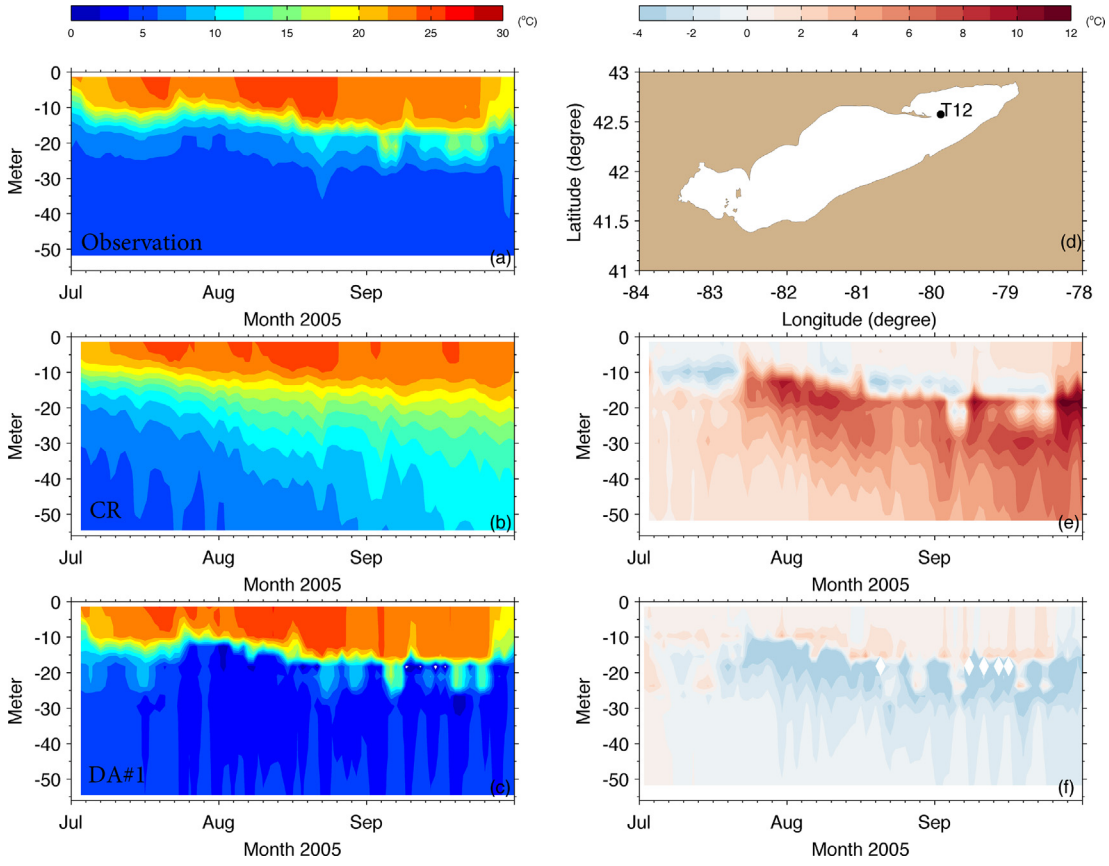


**Fig. 6.** Vertical temperature profiles from the thermistor observation (T12) (panel a; data location in panel d), from the CR experiment (b), and from the DA#1 (c), and model bias (model-observation) from CR e) and from the DA#1 (f).

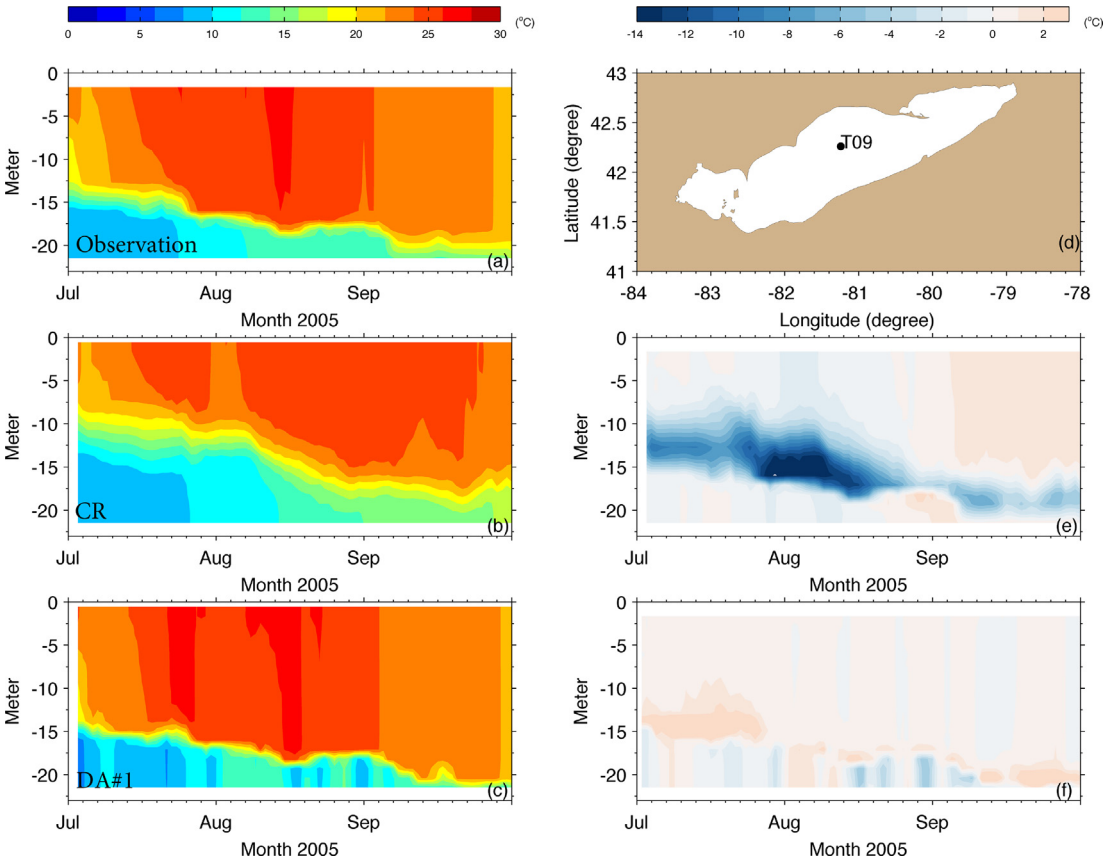


**Fig. 7.** The same as Fig. 5 but for the comparisons at the thermistor sampling station T09.

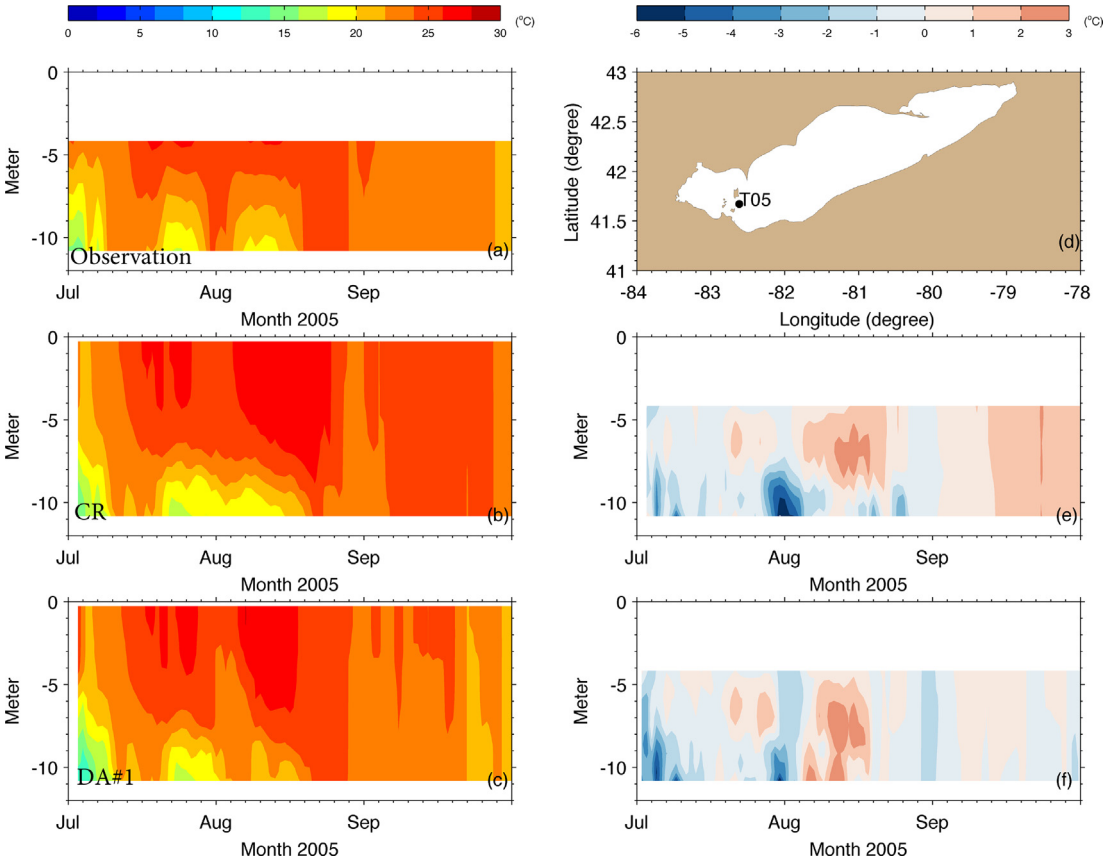


**Fig. 8.** The same as Fig. 5 but for the comparisons at the thermistor sampling station T05.

simulation, such a thermal gradient is not well resolved, and excessive heat is dissipated to the deep water, which causes a warm bias of 3–10 °C in the water column below the observed thermocline (Fig. 6b, e). In the central basin, the model CR simulation performs quite well in term of capturing the observed sharp thermocline but slightly underpredicts the mixed-layer depth. This is demonstrated by the model-data comparison at the observation (T09) in the central basin (Fig. 7). While the observed thermocline is located around 15 m in July and deepens to 18 m in August, the CR experiment shows the model underestimates the mixed-layer depth and predicts the thermocline at ~10 m in July and ~15 m in August (Fig. 7a, b). As the water temperature changes drastically around the thermocline, such a slight mismatch results in a significant cold bias of up to 10–14 °C at around 15 m in July and early August (Fig. 7e). In the western basin, where the water depth is much shallower, and the model-observation discrepancy is generally smaller (±1 °C) in the absence of a sharp thermocline, but a relatively large synoptic bias (e.g., on August 1st) is observed. More importantly, the general warm bias of ~1 °C persists (T05) in September (Fig. 8a, b, e). Note that the T05 data is not assimilated; it also demonstrated that model improvements are not limited to the locations where observational data have been assimilated.

In the DA#1 experiment, the data assimilative model successfully corrects all types of the model biases described above. The diffused thermal structure in the eastern basin is adjusted, and thermocline is correctly presented around 15 m, eliminating >90% of the warm bias in the CR experiment (Fig. 6c, f). The underestimated mixed-layer depth in the central basin in the CR experiment is also corrected, resulting in a similar pattern to the observed evolution of the thermocline change. Hence the substantial cold bias near the thermocline in the central basin is eliminated completely (Fig. 7c, f). In the western basin, the DA also successfully corrects both the long-term warm bias in September and the cold bias on the weather scale (Fig. 8c, f).

The results above show the nudging method can effectively improve model simulations in LST and thermal profile by assimilating thermistor data. Although the assimilation with prescribed anisotropic correlation pattern improves the LST hindcast across all lake basins, the local adjustment causes some disturbances in the coherence of the spatial pattern of LST: several "patches" are seen in the center basin, and the south-north temperature gradient in Fig. 5c is not as organized as in Fig. 5b. We hypothesize that the LST may be coherent on a large spatial scale and mainly controlled by surface heat fluxes (Xue et al., 2015). Therefore, it could be easier to estimate the error correlation pattern of LST than that of sub-surface layers, and consequently the improvement of LST on a basin scale may be achieved while retaining its spatial coherence. If this hypothesis is true, fewer observations may be needed for the LST assimilation if the sampling locations are strategically selected. Our next question is how to optimize sampling locations to minimize the sampling efforts while achieving similar assimilation effectiveness which would eventually lead to optimal sampling design. The DA#1 experiment is not able to address the issue as the construction of error covariance is based on the prescribed correlation length-scale. The above hypothesis and sampling design experiment are tested in DA#2 analysis, in which the error correlation map is derived from the difference of model simulation and GLSEA data in the simulation of previous summers (2002–2004).

The LST error correlation (i.e. $corr(i,j)$ in equation (6)) and corresponding correction weight (i.e. $corr(i,j) \cdot \frac{\sigma_i}{\sigma_j}$ in Eq. (6)) are estimated at 13 thermistor data sampling locations using model results and GLSEA data. Five locations (T04, T07, T08, T11, and T12) are selected for DA#2. The selection is based on choosing minimal number of sampling locations to give sufficient coverage of

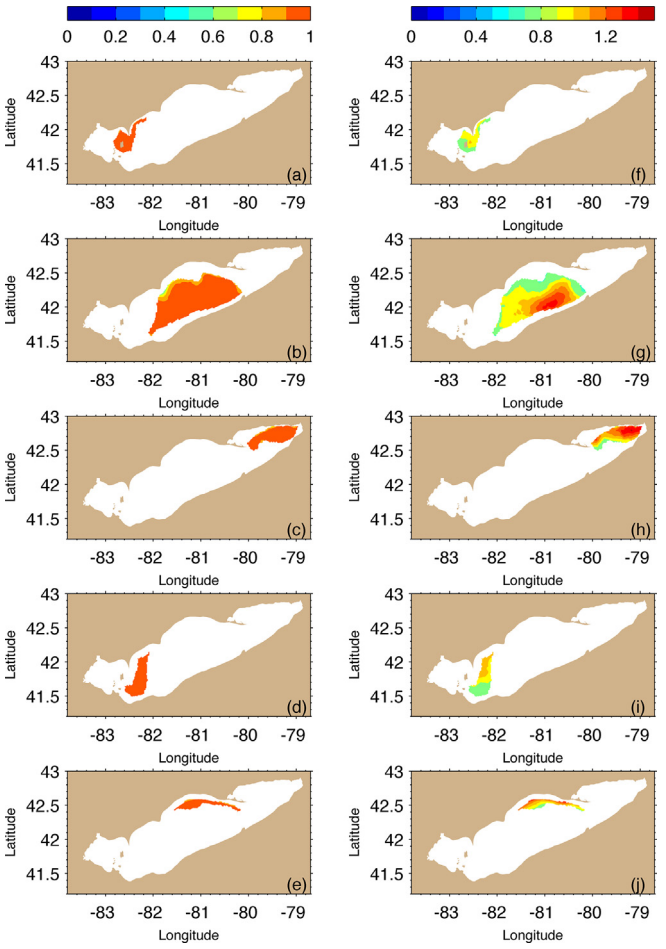


**Fig. 9.** Maps of error correlation coefficient (left column) and corresponding correction weight (right column) selected areas based on correlation coefficient >0.75 for five thermistor observations (T04, T07, T12, T08 and T11).

the lake using a combination of high influence areas of the selected observations, where high influence area is defined as the region with an error correlation coefficient >0.75 for that observation (Fig. 9). For example, the upper panels (Fig. 9) show the model error near the western basin is well correlated with model error at the sampling location T04 (Fig. 9a); therefore, the spatially varying corrections of model error near the western basin can be estimated based on the departure of model results from the observation at T04. Note that T02 would be a better choice than T04 for the western basin, but unfortunately the observational data at T02 are not available for the entire summer. In the majority of the central basin, the model error of LST is closely correlated with model error at the sampling location T07 (Fig. 9b). Fig. 9g provides the spatial correction weights within the high influence area of T07 based on Eq. (6). Similarly, the other three observations are selected to cover the eastern basin, western portion of the central basin, and the northern coastal zone (Fig. 9c–e, h, i, j).

Under such a design, the model error is efficiently reduced through DA. The modeled lake mean LST is in very good agreement with GLSEA data over the entire summer (RMSE of 0.45 °C) (Fig. 10d). In comparison, the DA#1 experiment only reduces errors by 50% in September using 12 thermistor data, while the DA#2 experiment eliminates >90% of the warm bias in September using only five selected thermistors with a RMSE down to 0.36 °C. The improvement is not only seen from the lake average LST but also from the spatial pattern of the summer LST. The warm biases in the central basin and western basin are removed to a large extent

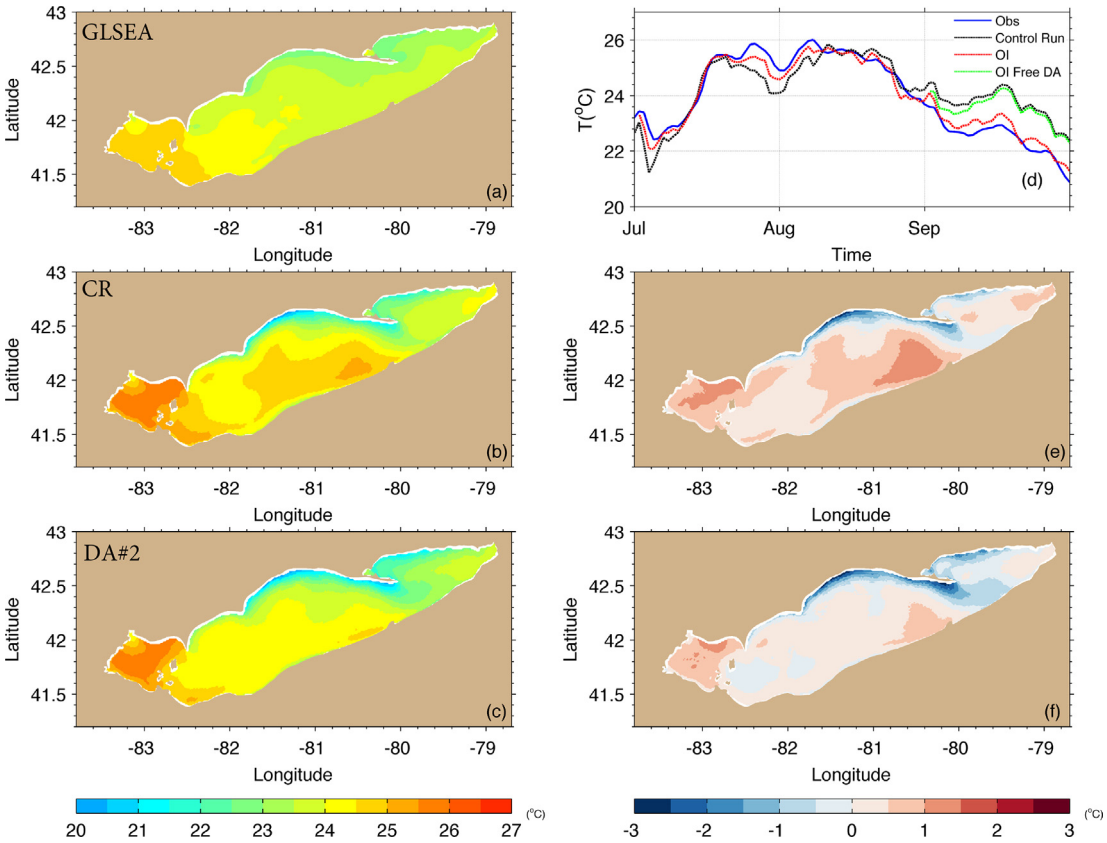


**Fig. 10.** Spatial pattern of LST during 2005 from GLSEA data (a), CR experiment (b), DA#2 experiment (c), lake-mean LST for four cases (d), model bias (model-GLSEA) for CR (e), for DA#2 (f).

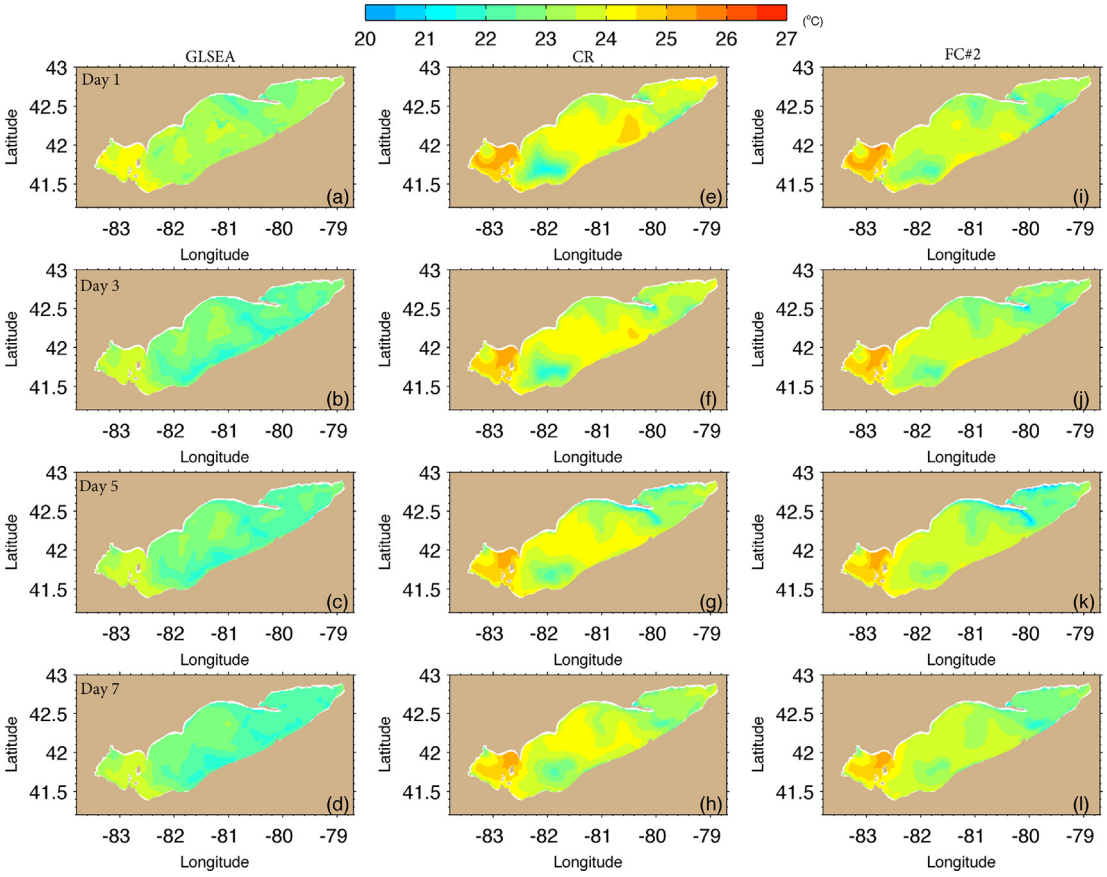


**Fig. 11.** Spatial pattern of LST on September 1st, 3rd, 5th, 7th, 2005 from GLSEA observation (a-d), CR (e-h), and FC#2 (i-l).

(Fig. 10f,g). More importantly, the coherent pattern of LST on basin scale is well preserved (Fig. 10c vs. Fig. 5c).

The DA#1 and DA#2 experiments show the effectiveness of DA in improving model hindcasts and reanalysis. The third question is to what extent the DA can help with the GLOFS short-term forecasting, which is evaluated in the experiments FC#1 and FC#2. To represent a forecast scenario, we start with the DA-enhanced nowcast conditions on Sept 1st, 2005. DA updates the model's initial condition for the subsequent forecast during which no data assimilation is used. September is selected for these experiments because the modeling simulation error was shown to be the most significant in the control run. The lake-wide average of LST in both FC#1 and FC#2 experiments improves in comparison to CR experiment. Results show that after the DA improves the forecast model initial condition, the forecast results within the first few days remain improved compared to the simulation (CR) (Figs. 5d and 10d). This can be clearly observed in Fig. 10d, which shows the improvement of forecast results compared to the control run in the DA#2 over the first 7 days of forecast. The error, of course, amplifies over time and the LST returns back to the CR simulation

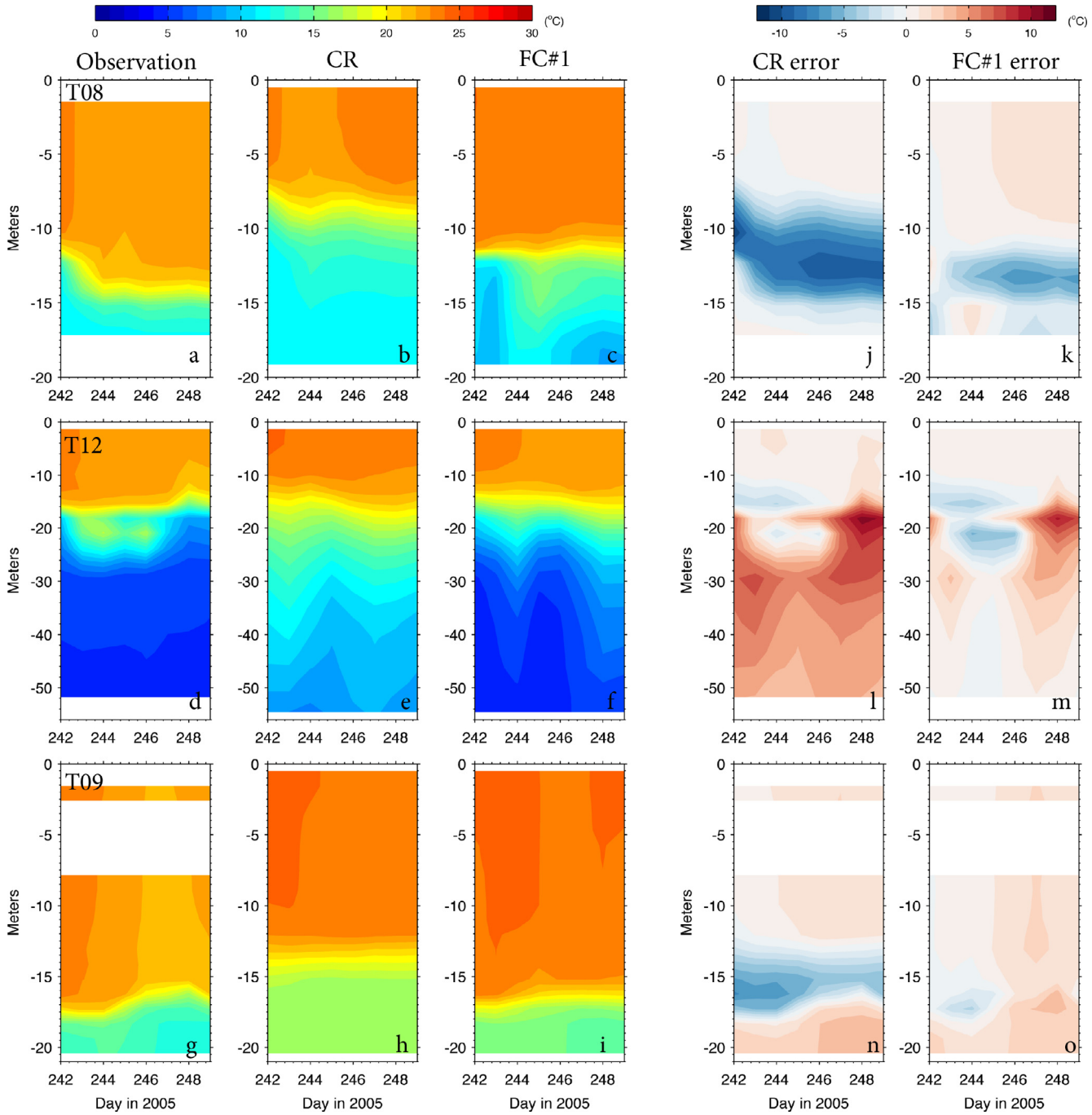


**Fig. 12.** The time evolution of the temperature profile at T08, T12 and T09 from observation (a,d,g), CR simulation(b,e,h), and FC#1 simulation (c,f,i) in the 7-day forecasting, respectively. Model errors in the CR and in the FC#1 at T08, T12 and T09 are shown in panels (j,l,n) and (k,m,o), respectively.

eventually after ~ 15 days, indicating the memory of impact of initial correction of temperature field is limited in the first 1–2 weeks. (Fig. 10d).

Spatially, the GLSEA data show that the LST in the majority of the lake stays below 24 °C with a constant cooling in the 7-day forecast period. The CR experiment shows a warm bias of 1.5–2.0 °C in the central and western basins over the same simulation period. When the warm bias is corrected on September 1st with DA, the LST error remains < 1 °C in the following 7 days without DA (Fig. 11). The importance of DA in assisting forecasts is also evidenced by the improved results in simulating the lake thermal structure. Fig. 12 demonstrates such improvement at various locations such as shallow water (T08), deep water in the eastern basin with abruptly changing bathymetry (T12) and open water in the central basin (T09). Although the model error is still amplified over time after the model is unleashed from the DA after September 1st, the magnitude of model forecast error is significantly reduced compared to the CR simulation. This is quantitatively demonstrated in the skill score for forecast experiments (Table 2). In the FC#1 and FC#2 cases, 11 and 10 out of 13 stations show forecast improvement, respectively.

## Discussion and conclusions

Physical dynamics in coastal regions are often highly nonlinear and vary significantly in time and space, where the results found in one system may not be directly applied to other coastal regions (Chen et al., 2009). The factors affecting success of DA include hydrodynamic characteristics of the system, level of model accuracy, representation of error covariance in assimilation experiments, effectiveness and efficiency of assimilating scheme, computational resources, and availability of data sites (Xue et al., 2011, 2012). This is why a feasibility study is crucial to the development of a new data assimilative forecasting system that is subject to a long-term incremental improvement (Chassignet et al., 2007; Martin et al., 2007; Chao et al., 2009; Farrara et al., 2013).

In this study, we developed and designed a series of DA experiments with the aim to improve Lake Erie thermal structure simulations. This is the first time, to our knowledge, multiple data assimilation techniques have been evaluated and applied to a 3D hydrodynamic model in the Great Lakes, and it serves as the foundation for developing future real-time GLOFS-DA forecasting systems and optimal sampling strategy. Results suggest that DA can effectively improve model performance with limited observational data when the DA formulation is developed appropriately. The formulation must take into account the dynamic characteristics of Lake Erie and its anisotropic error correlation pattern. Prediction skill of the data assimilative model is evaluated by examining model performance after the model is unleashed from DA. The correction of initial condition by DA positively influences the prediction results on a time scale of 1–7 days, and effectively constrains the amplification of the model error.

The feasibility of optimizing data sampling design to improve forecast models is also explored. The experiment targets the LST as GLSEA data is available for lake surface error correlation analysis. When the spatial error correlation is estimated appropriately, using only five strategically selected thermistors of the thirteen available, the DA can provide an accurate correction of the LST. Although this method, at its current stage of development, cannot be directly extended to 3D thermal structure assimilation because of the lack of sufficient data to estimate error covariance in the sub-surface layers, the experiment serves well as a proof-of-concept to demonstrate the potential to improve model accuracy with optimal data sampling efforts. It can also be extended to observing system simulation experiments (OSSEs) to evaluate the potential impact of new observing systems and alternative deployments of existing systems as a rigorous, cost-effective approach.

Dealing with error covariance is at the heart of DA (Derber and Bouttier, 1999; Weaver and Courtier, 2001; Fisher, 2003; Pereira and Berre, 2006; Fowler and Jan Van Leeuwen, 2013). In this study, we adopted stationary error covariance for the summer thermal structure as a case study. This is an efficient way to examine the feasibility of the DA implementation in a complex water system, toward improving a sophisticated operational forecasting system with assimilative capability. The construction of time-varying error covariance and the consideration of cross-correlation with dynamic constraints among multiple state variables need to be further investigated at the next stage (Kalnay, 2003; Li et al., 2008a, b; Brousseau et al., 2012; Waller et al., 2014). This also requires more observational data and large size ensemble simulations to determine the error spreading pattern. This flow-dependent covariance is particularly important when assimilating variables such as water currents or water level that are characterized by short decorrelation time scale and spatial scale (Kuragano et al., 2000; Xue et al., 2011; Li et al., 2015). Ultimately, developing a GLOFS-DA framework with a time-varying error covariance using rank-reduction (e.g. Fukumori and Malanotte-Rizzoli, 1995; Verlaan and Heemink, 1997; Pham et al., 1998; Fukumori, 2002; Cao et al., 2007; Cosme and et al., 2010; Xue et al., 2011) and ensemble representation (Evensen 2009; Anderson 2001; Houtekamer and Mitchell, 2001; Whitaker et al., 2002; Houtekamer et al., 2005; Torn and et al., 2009; Xue et al., 2011, 2012; Houtekamer and Zhang, 2016) is the long-term goal.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This is the contribution 68 of the Great Lakes Research Center at Michigan Technological University. This is NOAA GLERL contribution number 1932. The Michigan Tech high performance computing cluster, *Superior*, was used in obtaining the modeling results presented in this publication.

## Funding

Funding for this project was provided by the Great Lakes Restoration Initiative, through the University of Michigan Cooperative Institute for Great Lakes Research (CIGLR) cooperative agreement with the National Oceanic and Atmospheric Administration (NA17OAR4320152).

## Appendix A. Localized OI assimilation scheme

Under the condition that an observation $y_j$ has its own influence area and model correction are made only within the influence area of the observation based on the model-data misfit at the observation location, it becomes a specific case to apply Eqs. (1) and (2) for a single observation. Observation error covariance matrix $\boldsymbol{R}$ is reduced to a scalar as observation error variance $\sigma_{y_j}^2 = \overline{(y_j - \boldsymbol{H}\boldsymbol{x}^t)(y_j - \boldsymbol{H}\boldsymbol{x}^t)^T}$,$\boldsymbol{HBH^T}$ is also reduced to a scalar, which is the model error variance $\sigma_{m_{y_j}}^2$ at the observation location, defined as $\overline{(\boldsymbol{H}\boldsymbol{x}^f - \boldsymbol{H}\boldsymbol{x}^t)(\boldsymbol{H}\boldsymbol{x}^f - \boldsymbol{H}\boldsymbol{x}^t)^T}$. $\boldsymbol{BH^T}$ is reduced to a vector whose element is the model error covariance $\mathrm{cov}\left(m_i, m_{y_j}\right)$ of model error

at each model grid $m_i, i = 1, 2.., L$ within the influence area and model error at the observation location. In this case, the analysis Eqs. (1) and (2) is combined as

$$\boldsymbol{x}^a_{m_i} = \boldsymbol{x}^f_{m_i} + \mathrm{cov}\left(m_i, m_{y_j}\right) \cdot \frac{1}{\sigma^2_{m_{y_j}} + \sigma^2_{y_j}} \cdot d_j \tag{A1}$$

Since

$$Cov\left(m_i, m_{y_j}\right) = \mathrm{corr}\left(m_i, m_{y_j}\right) \cdot \sigma_{m_i} \cdot \sigma_{m_{y_j}} \tag{A2}$$

It follows that

$$\boldsymbol{x}^a_{m_i} = \boldsymbol{x}^f_{m_i} + corr\left(m_i, m_{y_j}\right) \cdot \frac{\sigma_{m_i} \cdot \sigma_{m_{y_j}}}{\sigma^2_{m_{y_j}} + \sigma^2_{y_j}} \cdot d_j \tag{A3}$$

where $i = 1, 2, \cdots L$, and $m_i$ denotes the total $L$ model grids within the influence area of a given observation station $y_j$, $corr(m_i, m_{y_i})$ is the correlation between modeling error at model grid $m_i$ and modeling error at the location of observation $y_j$, $\sigma_{m_i}$ and $\sigma^2_{m_{y_j}}$ are variances of modeling error at the grid $m_i$ and at the location of the observation $y_j$. $\sigma^2_{y_j}$ is the observation error variance at the observation $y_j$.

## References


Adrian, R. et al., 2009. Lakes as sentinels of climate change. Limnol. Oceanogr. 54, 2283–2297.

Anderson, E.J., Bechle, A.J., Wu, C.H., Schwab, D.J., Mann, G.E., Lombardy, K.A., 2015. Reconstruction of a meteotsunami in Lake Erie on 27 May 2012: roles of atmospheric conditions on hydrodynamic response in enclosed basins. J. Geophys. Res. Oceans 120, 8020–8038.

Anderson, E.J., Schwab, D.J., 2017. Meteorological influence on summertime baroclinic exchange in the Straits of Mackinac. J. Geophys. Res. Oceans 122, 2171–2182.

Anderson, E.J., Schwab, D.J., 2013. Predicting the oscillating bi-directional exchange flow in the Straits of Mackinac. J. Great Lakes Res. 39, 663–671.

Anderson, J.L., 2001. An ensemble adjustment Kalman filter for data assimilation. Mon. Weather Rev. 129 (12), 2884–2903.

Arnott, D.L., Vanni, M.J., 1996. Nitrogen and phosphorus recycling by the zebra mussel (Dreissena polymorpha) in the western basin of Lake Erie. Can. J. Fish. Aquat. Sci. 53, 646–659.

Arvola, L., George, G., Livingstone, D.M., Järvinen, M., Blenckner, T., Dokulil, M.T., Jennings, E., Aonghusa, C.N., Nõges, P., Nõges, T., 2009. The impact of the changing climate on the thermal characteristics of lakes. In: The Impact of Climate Change on European Lakes Anonymous. Springer, pp. 85–101.

Bannister, R., 2017. A review of operational methods of variational and ensemble-variational data assimilation. Q. J. R. Meteorol. Soc. 143, 607–633.

Beletsky, Dmitry, Schwab, David J., Roebber, Paul J., McCormick, Michael J., Miller, Gerald S., Saylor, James H., 2003. Modeling wind-driven circulation during the March 1998 sediment resuspension event in Lake Michigan: MODELING CIRCULATION IN LAKE MICHIGAN. J. Geophys. Res. 108 (C2), n/a–n/a. https://doi.org/10.1029/2001JC001159.

Beletsky, D., Schwab, D., McCormick, M., 2006. Modeling the 1998–2003 summer circulation and thermal structure in Lake Michigan. J. Geophys. Res. Oceans 111, C10010.

Bishop, C.H., Etherton, B.J., Majumdar, S.J., 2001. Adaptive sampling with the ensemble transform kalman filter. Part I: theoretical aspects. Mon. Weather Rev. 129, 420–436.

Bouttier, F., & Courtier, P., 2002, Data assimilation concepts and methods March 1999. Meteorological training course lecture series. ECMWF, 59.

Brousseau, P., Berre, L., Bouttier, F., Desroziers, G., 2012. Flow-dependent background-error covariances for a convective-scale data assimilation system. Q. J. R. Meteorol. Soc. 138, 310–322.

Burns, N.M., Rockwell, D.C., Bertram, P.E., Dolan, D.M., Ciborowski, J.J., 2005. Trends in temperature, Secchi depth, and dissolved oxygen depletion rates in the central basin of Lake Erie, 1983–2002. J. Great Lakes Res. 31, 35–49.

Brandt, S. and Lansing, M. The International Field Years on Lake Erie (IFYLE) (2006) https://www.glerl.noaa.gov/pubs/fulltext/2006/20060048.pdf.

Cao, Y., Zhu, J., Navon, I.M., Luo, Z., 2007. A reduced-order approach to four-dimensional variational data assimilation using proper orthogonal decomposition. Int. J. Numer. Meth. Fluids 53 (10), 1571–1583.

Chaffin, J.D., Bridgeman, T.B., Bade, D.L., 2013. Nitrogen constrains the growth of late summer cyanobacterial blooms in Lake Erie. Adv. Microbiol. 3 (06), 16.

Chao, Y., Li, Z., Farrara, J., McWilliams, J.C., Bellingham, J., Capet, X., Chavez, F., Choi, J., Davis, R., Doyle, J., 2009. Development, implementation and evaluation of a data-assimilative ocean forecasting system off the central California coast. Deep Sea Res. Part II 56, 100–126.

Chassignet, E.P., Hurlburt, H.E., Smedstad, O.M., Halliwell, G.R., Hogan, P.J., Wallcraft, A.J., Baraille, R., Bleck, R., 2007. The HYCOM (hybrid coordinate ocean model) data assimilative system. J. Mar. Syst. 65, 60–83.

Chen, C., Beardsley, R., Cowles, G., 2006. An Unstructured Grid, Finite-Volume Coastal Ocean Model. FVCOM User Manual.

Chen, C., Malanotte-Rizzoli, P., Wei, J., Beardsley, R.C., Lai, Z., Xue, P., Cowles, G.W., 2009. Application and comparison of Kalman filters for coastal ocean problems: an experiment with FVCOM. J. Geophys. Res. Oceans 114 (C5).

Chu, P., Kelley, J., Mott, G., Zhang, A.J., Lang, G., 2011. Development, implementation and skill assessment of the NOAA/NOS great lakes forecast System. J. Ocean Dynamics 61 (9), 1305–1316.

Conroy, J.D., Kane, D.D., Dolan, D.M., Edwards, W.J., Charlton, M.N., Culver, D.A., 2005. Temporal trends in Lake Erie plankton biomass: roles of external phosphorus loading and dreissenid mussels. J. Great Lakes Res. 31, 89–110.

Cosme, E., Brankart, J.M., Verron, J., Brasseur, P., Krysta, M., 2010. Implementation of a reduced rank square-root smoother for high resolution ocean data assimilation. Ocean Model. 33 (1–2), 87–100.

Daloğlu, I., Cho, K.H., Scavia, D., 2012. Evaluating causes of trends in long-term dissolved reactive phosphorus loads to Lake Erie. Environ. Sci. Technol. 46, 10660–10666.

Derber, J., Bouttier, F., 1999. A reformulation of the background error covariance in the ECMWF global data assimilation system. Tellus A Dyn. Meteorol. Oceanogr. 51, 195–221.

Drinkwater, K.F., Belgrano, A., Borja, A., Conversi, A., Edwards, M., Greene, C.H., Ottersen, G., Pershing, A.J., Walker, H., 2003. The Response of Marine Ecosystems to Climate Variability Associated with the North Atlantic Oscillation. Wiley Online Library.

Evensen, G., 2009. *Data assimilation: the ensemble Kalman filter*. Springer Science & Business Media. Prince William sound and an evaluation of its performance during sound Predictions 2009. Cont. Shelf Res. 63, S193–S208.

Fisher, M., 2003. Background Error Covariance Modelling, Seminar on Recent Development in Data Assimilation for. Atmosphere and Ocean.

Farrara, J. D., Y. Chao, Z. Li, X. Wang, X. Jin, H. Zhang, P. Li, Q. Vu, P. Q. Olsson, and G. C. Schoch (2013), A data-assimilative ocean forecasting system for the

Fowler, A., Jan Van Leeuwen, P., 2013. Observation impact in data assimilation: the effect of non-Gaussian observation error. Tellus A Dyn. Meteorol. Oceanogr. 65, 20035.

Fukumori, I., Malanotte-Rizzoli, P., 1995. An approximate Kaiman filter for ocean data assimilation: an example with an idealized Gulf Stream model. J. Geophys. Res. Oceans 100 (C4), 6777–6793.

Fukumori, I., 2002. A partitioned Kalman filter and smoother. Mon. Weather Rev. 130 (5), 1370–1383.

Fujisaki, A., Wang, J., Bai, X., Leshkevich, G., Lofgren, B., 2013. Model-simulated interannual variability of Lake Erie ice cover, circulation, and thermal structure in response to atmospheric forcing, 2003–2012. J. Geophys. Res. Oceans 118 (9), 4286–4304.

Ghil, M., Malanotte-Rizzoli, P., 1991. Data assimilation in meteorology and oceanography. In: Advances in geophysics, Vol. 33. Elsevier, pp. 141–266.

Hawley, N., Johengen, T.H., Rao, Y.R., Ruberg, S.A., Beletsky, D., Ludsin, S.A., Eadie, B. J., Schwab, D.J., Croley, T.E., Brandt, S.B., 2006. Lake Erie hypoxia prompts Canada-US study. Eos, Transactions American Geophysical Union 87, 313–319.

Hoffman, R.N., Atlas, R., 2016. Future observing system simulation experiments. Bull. Am. Meteorol. Soc. 97 (9), 1601–1616.

Houtekamer, P.L., Mitchell, H.L., Pellerin, G., Buehner, M., Charron, M., Spacek, L., Hansen, B., 2005. Atmospheric data assimilation with an ensemble Kalman filter: results with real observations. Mon. Weather Rev. 133 (3), 604–620.

Houtekamer, P.L., Mitchell, H.L., 2001. A sequential ensemble Kalman filter for atmospheric data assimilation. Mon. Weather Rev. 129, 123–137.

Houtekamer, P., Zhang, F., 2016. Review of the ensemble Kalman filter for atmospheric data assimilation. Mon. Weather Rev. 144, 4489–4532.

Huang, A., Rao, Y.R., Lu, Y., 2010. Evaluation of a 3-D hydrodynamic model and atmospheric forecast forcing using observations in Lake Ontario. J. Geophys. Res. Oceans 115, C02004.

Huang, C., Kuzincsky, A., Auer, M., O'Donnell, D., Xue, P., 2019. Management transition to the great lakes nearshore: insights from hydrodynamic modeling. J. Mar. Sci. Eng. 7 (5), 129. https://doi.org/10.3390/jmse7050129.

Hunt, B.R., Kostelich, E.J., Szunyogh, I., 2007. Efficient data assimilation for spatiotemporal chaos: a local ensemble transform Kalman filter. Phys. D: Nonlinear Phenomena 230, 112–126.

Ide, K., Courtier, P., Ghil, M., Lorenc, A.C., 1997. Unified Notation for Data Assimilation: Operational, Sequential and Variational. J. Meteorol. Soc. Jpn. Ser. II 75 (1B), 181–189.

Kalnay, E., 2003. Atmospheric Modeling, Data Assimilation and Predictability. Cambridge University Press.

Kelley, J.G.W., Chen, Y., Anderson, E.J., Lang, G.A., Xu, J., 2018. Upgrade of NOS lake erie operational forecast system (LEOFS) to FVCOM: model development and hindcast skill assessment. NOAA Tech. Memo. NOS CS 2018 (40), 92.

Kuragano, T., Kamachi, M., 2000. Global statistical space-time scales of oceanic variability estimated from the TOPEX/POSEIDON altimeter data. J. Geophys. Res. Oceans 105 (C1), 955–974.

Li, Z., Chao, Y., McWilliams, J.C., Ide, K., 2008a. A three-dimensional variational data assimilation scheme for the regional ocean modeling system. J. Atmos. Ocean. Technol. 25, 2074–2090.

Li, Z., Chao, Y., McWilliams, J.C., Ide, K., 2008b. A three-dimensional variational data assimilation scheme for the Regional Ocean Modeling System: Implementation and basic experiments. J. Geophys. Res. Oceans 113 (C5).

Li, Z., McWilliams, J.C., Ide, K., Farrara, J.D., 2015. A multiscale variational data assimilation scheme: formulation and illustration. Mon. Weather Rev. 143 (9), 3804–3822.
Liu, P.C., Schwab, D.J., 1987. A comparison of methods for estimating u* from given uz and air-sea temperature differences. J. Geophys. Res. Oceans 92, 6488–6494.
Lorenc, A.C., 1986. Analysis methods for numerical weather prediction. Q. J. Roy. Meteor. Soc. 112 (474), 1177–1194.
Makarewicz, J.C., Lewis, T.W., Bertram, P., 1999. Phytoplankton composition and biomass in the offshore waters of Lake Erie: pre-and post-Dreissena introduction (1983–1993). J. Great Lakes Res. 25, 135–148.
Martin, M., Hines, A., Bell, M., 2007. Data assimilation in the FOAM operational short-range ocean forecasting system: a description of the scheme and its impact. Q. J. R. Meteorol. Soc. 133, 981–995.
Michalak, A.M. et al., 2013. Record-setting algal bloom in Lake Erie caused by agricultural and meteorological trends consistent with expected future conditions. Proc. Natl. Acad. Sci. U.S.A. 110, 6448–6452.
Mellor, G.L., Yamada, T., 1982. Development of a turbulence closure model for geophysical fluid problems. Rev. Geophys. 20 (4), 851–875.
Millie, D.F., Fahnenstiel, G.L., Bressie, J.D., Pigg, R.J., Rediske, R.R., Klarer, D.M., Tester, P.A., Litaker, R.W., 2009. Late-summer phytoplankton in western Lake Erie (Laurentian Great Lakes): bloom distributions, toxicity, and environmental influences. Aquat. Ecol. 43, 915–934.
Nicholls, K.H., Hopkins, G.J., 1993. Recent changes in Lake Erie (north shore) phytoplankton: cumulative impacts of phosphorus loading reductions and the zebra mussel introduction. J. Great Lakes Res. 19, 637–647.
Pereira, M.B., Berre, L., 2006. The use of an ensemble approach to study the background error covariances in a global NWP model. Mon. Weather Rev. 134, 2466–2489.
Robinson, A. R. and P. F. J. Lermusiaux (2000), Overview of Data Assimilation, Harvard University Reports in Physical/Interdisciplinary Ocean Science, 62.
Rowe, M.D., Anderson, E.J., Wang, J., Vanderploeg, H.A., 2015. Modeling the effect of invasive quagga mussels on the spring phytoplankton bloom in Lake Michigan. J. Great Lakes Res. 41 (Supplement 3), 49–65.
Pham, D.T., Verron, J., Roubaud, M.C., 1998. A singular evolutive extended Kalman filter for data assimilation in oceanography. J. Mar. Syst. 16 (3–4), 323–340.
Smagorinsky, J., 1963. General circulation experiments with the primitive equations: I The basic experiment. Monthly Weather Rev. 91 (3), 99–164.
Schertzer, W., Saylor, J., Boyce, F., Robertson, D., Rosa, F., 1987. Seasonal thermal cycle of Lake Erie. J. Great Lakes Res. 13, 468–486.
Schwab, D.J., Bedford, K.W., 1994. Initial Implementation of the great lakes forecasting system: a real-time system for predicting lake circulation and thermal structure. Water Poll. Res. J. Canada 29, 203–220.
Schwab, D.J., Leshkevich, G.A., Muhr, G.C., 1992. Satellite measurements of surface water temperature in the great lakes: great lakes coastwatch. J. Great Lakes Res. 18, 247–258.
Schwab, D.J., Leshkevich, G.A., Muhr, G.C., 1999. Automated mapping of surface water temperature in the great lakes. J. Great Lakes Res. 25, 468–481.
Scott, K.A., Chen, C., Myers, P.G., 2018. Assimilation of argo temperature and salinity profiles using a bias-aware EnOI scheme for the labrador sea. J. Atmos. Oceanic Technol. 35, 1819–1834. https://doi.org/10.1175/JTECH-D-17-0222.1.
Smith, R.E., Hiriart-Baer, V.P., Higgins, S.N., Guildford, S.J., Charlton, M.N., 2005. Planktonic primary production in the offshore waters of dreissenid-infested Lake Erie in 1997. J. Great Lakes Res. 31, 50–62.
Tulonen, T., Kankaala, P., Ojala, A., Arvola, L., 1994. Factors controlling production of phytoplankton and bacteria under ice in a humic, boreal lake. J. Plankton Res. 16, 1411–1432.
Torn, R.D., Hakim, G.J., 2009. Ensemble data assimilation applied to RAINEX observations of Hurricane Katrina (2005). Mon. Weather Rev. 137 (9), 2817–2829.
Verlaan, M., Heemink, A.W., 1997. Tidal flow forecasting using reduced rank square root filters. Stochastic Hydrol. Hydraul. 11 (5), 349–368.
Vanderploeg, H.A., Liebig, J.R., Carmichael, W.W., Agy, M.A., Johengen, T.H., Fahnenstiel, G.L., Nalepa, T.F., 2001. Zebra mussel (Dreissena polymorpha) selective filtration promoted toxic Microcystis blooms in Saginaw Bay (Lake Huron) and Lake Erie. Can. J. Fish. Aquat. Sci. 58, 1208–1221.
Waller, J.A., Dance, S.L., Lawless, A.S., Nichols, N.K., 2014. Estimating correlated observation error statistics using an ensemble transform Kalman filter. Tellus A: Dyn. Meteorol. Oceanogr. 66, 23294.
Wang, J., Shen, Y., 2010. Modeling oil spills transportation in seas based on unstructured grid, finite-volume, wave-ocean model. Ocean Model. 35, 332–344.
Weaver, A., Courtier, P., 2001. Correlation modelling on the sphere using a generalized diffusion equation. Q. J. R. Meteorol. Soc. 127, 1815–1846.
Whitaker, J.S., Hamill, T.M., 2002. Ensemble data assimilation without perturbed observations. Mon. Weather Rev. 130 (7), 1913–1924.
Xue, P., Chen, C., Beardsley, R.C., 2012. Observing system simulation experiments of dissolved oxygen monitoring in Massachusetts Bay. J. Geophys. Res. Oceans 117, C05014.
Xue, P., Chen, C., Beardsley, R.C., Limeburner, R., 2011. Observing system simulation experiments with ensemble Kalman filters in Nantucket Sound, Massachusetts. J. Geophys. Res. Oceans 116, C01011.
Xue, P., Pal, J.S., Ye, X., Lenters, J.D., Huang, C., Chu, P.Y., 2017. Improving the simulation of large lakes in regional climate modeling: two-way lake-atmosphere coupling with a 3D hydrodynamic model of the great lakes. J. Climate 30, 1605–1627.
Xue, P., Schwab, D.J., Hu, S., 2015. An investigation of the thermal response to meteorological forcing in a hydrodynamic model of Lake Superior. J. Geophys. Res. Oceans 120, 5233–5253.
Xue, P., Schwab, D.J., Zhou, X., Huang, C., Kibler, R., Ye, X., 2018. A hybrid lagrangian-eulerian particle model for ecosystem simulation. J. Mar. Sci. Eng. 6 (4), 109. https://doi.org/10.3390/jmse6040109.
Ye, Xinyu, Anderson, Eric J., Chu, Philip Y., Huang, Chenfu, Xue, Pengfei, 2019. Impact of water mixing and ice formation on the warming of lake superior: a model-guided mechanism study. Limnol. Oceanogr. 64 (2), 558–574. https://doi.org/10.1002/lno.v64.210.1002/lno.11059.
Zhang, S., Harrison, M., Rosati, A., Wittenberg, A., 2007a. System design and evaluation of coupled ensemble data assimilation for global oceanic climate studies. Mon. Weather Rev. 135, 3541–3564.
Zhang, Z., Beletsky, D., Schwab, D.J., Stein, M.L., 2007b. Assimilation of current measurements into a circulation model of Lake Michigan. Water Resour. Res. 43 (11).
Zhou, Y., Obenour, D.R., Scavia, D., Johengen, T.H., Michalak, A.M., 2013. Spatial and temporal trends in Lake Erie hypoxia, 1987–2007. Environ. Sci. Technol. 47, 899–905.