Manifold Methods for Assimilating Geophysical and Meteorological Data in Earth System Models and their Components

Ammar Safaie[1], Chinh Dang[2], Han Qiu[1], Hayder Radha[2], Mantha S. Phanikumar[1*]

Abstract

A novel manifold method of reconstructing dynamically evolving spatial fields is presented for assimilating data from sensor networks in integrated land surface – subsurface, oceanic / lake models. The method was developed based on the assumption that data can be mapped onto an underlying differential manifold. In this study, the proposed method was used to reconstruct meteorological forcing over Lake Michigan, the bathymetry of an inland lake (Gull Lake), and precipitation over the Grand River watershed in Michigan. In the first case study, hourly interpolated meteorological forcing data were used to run a three-dimensional hydrodynamic model of Lake Michigan to quantify the improvement that results from the use of the new interpolation method. In the second example, the bathymetry of Gull Lake was reconstructed from measured scatter point data using the manifold technique. A hydrodynamic model of Gull Lake was developed and further improved using improved bathymetry. In the last case study, daily participation data over a six-year period were interpolated over the Grand River watershed and used as input to an integrated, distributed hydrologic model. All three examples illustrate the superior performance of the manifold method over standard methods in terms of accuracy and computational efficiency. Our results also indicate that evaluating the relative performance of interpolation methods using the cross-validation method can lead to misleading conclusions about the relative performance of the methods.

1 Introduction

Water continuously circulates between the Earth's surface and the atmosphere and moves through watersheds to become stored underground as groundwater or in surface reservoirs such as lakes. This circulation is a key aspect of Earth system models or components thereof and meteorological forcing plays an important role in modeling coupled physical - biogeochemical processes. Process-based modeling approaches describe land-lake-atmosphere interactions by explicitly considering the spatiotemporal variability of meteorological forcing fields (Xue et al., 2015). There are a variety of numerical ocean models, such as POM (Blumberg, and Mellor, 1987), ROMS (Haidvogel and Beckmann, 1999), and FVCOM (Chen et al., 2006), and their performance is highly dependent on how realistic the distributions of surface forcing fields (wind stress, heat flux, precipitation, evaporation) are. These forcing fields can be obtained from observational data, the output of a weather forecast model, or a combination thereof (Xue et al., 2015). In all cases it is crucial to use a suitable interpolation method in order to assimilate observations into the models, and estimate variables at unsampled locations and/or times. For instance, currents in large lakes such as the Laurentian Great Lakes are mostly controlled by wind. Therefore, by improving the representation of wind fields in models of lake circulation, we expect to describe coupled biophysical processes in lakes more accurately. For example, Safaie et al. (2016) demonstrated that improved representation of meteorological fields based on natural neighbor interpolation of weather station data produced superior results for currents and bacterial concentrations relative to similar results based on a nearest neighbor interpolation of the same data.

In situ observations generally have sparse and inhomogeneous distribution in space and time, and it is often infeasible to accurately reconstruct the true field from the data. However, more information about the structure of the field and its evolution, allows for better approximations (Barth et al., 2008). Various deterministic [e.g., nearest neighbor, natural neighbor, inverse distance weighting (IDW), spline, polynomial] and geostatistical (e.g. kriging) interpolation methods have been developed to generate spatial fields. There have been numerous efforts to compare different spatial interpolation methods in order to identify the best method for a given model application. Many researches have used cross-validation for assessing the performance of the interpolation methods. In this method, a subset of the original dataset is withheld to be used later for validating the interpolated field constructed from the rest of the observational data. Mean error (ME), root mean square error (RMSE) and the coefficient of determination (R^2) are commonly used to evaluate the performance of each interpolation method (Suparta and Rahman, 2016). However, every problem has a unique method of interpolation that works best for a given distribution of observations and the intended use of the interpolated data. Density of a sensor network, spatial variability of the variable of interest and its distribution, and observational errors, all influence the accuracy of the interpolated field (MacEachren and Davidson, 1987). For example, Luo et al. (2008) compared seven spatial interpolation techniques to identify which method produced the best estimation of the wind speed data recorded across England and Wales. Their study showed that kriging is the best method, and that the thin plate spline method had higher ME and RMSE values. However, in (Suparta and Rahman, 2016) the performance of the thin plate spline interpolation based on the RMSE and R^2 values was found to be better than kriging for less dense data points over the selected interpolation surface. Therefore, comparing interpolation methods using the cross-validation method without considering the data structure

75 and the purpose of interpolation is not guaranteed to produce the best representation of the
76 underlying data.

77 Precipitation is another important component of the water cycle. Spatial distribution of
78 precipitation influences the hydrologic response of a watershed (Daly et al., 2002). Basin
79 responses of rainfall-runoff processes are closely related to the spatial variability of precipitation
80 (Anquetin et al., 2010; Bell and Moore, 2000; Beven and Hornberger, 1982; Obled et al., 1994;
81 Schuurmans and Bierkens, 2007; Syed et al., 2003; Tetzlaff and Uhlenbrook, 2005).
82 Hydrologists usually use rain gauge data to obtain rainfall patterns over a watershed. Bell and
83 Moore (2000) found that the responses of a distributed hydrologic model are sensitive to the
84 locations of the rain gauges within the catchment and hence to the spatial variability of rainfall.
85 Nicótina et al. (2008) performed numerical experiments to study the effects of different spatial
86 resolutions of rainfall on various catchments. They found that the catchment response is sensitive
87 to the spatial distribution of rainfall only when water residence time in the channels is
88 comparable to the hillslope travel time; thus rainfall spatial heterogeneity likely plays a more
89 important role in affecting the runoff response of large watersheds (typically larger than 103
90 km^2) than smaller watersheds.

91 Various efforts have been devoted to improving the representation of spatially-distributed
92 rainfall fields in hydrologic modeling. For example, Ly et al. (2011) compared seven
93 interpolation methods for daily rainfall and found that geostatistical methods such as kriging and
94 IDW algorithms significantly outperformed the Thiessen polygon method, which is also known
95 as the nearest neighbor method. Sun et al. (2000) tested different rainfall estimation methods on
96 a 1060 km^2 catchment in Australia to evaluate the flood modeling capabilities of a hydrologic
97 model and found that blending radar and rain gauge data in a co-kriging framework provided

98 better performance compared with an approach based on kriging of the rain gauges data alone.
99 Masih et al. (2011) used a semi-distributed model, Soil Water Assessment Tool (SWAT), to
100 simulate the Karkheh River basin in Iran with two different precipitation interpolation methods.
101 Their results, based on a comparison of simulated and observed discharges using the metrics of
102 coefficient of determination and the Nash-Sutcliffe efficiency (NASH), showed that precipitation
103 interpolation with inverse distance and elevation weighting technique produced better
104 performance relative to the SWAT default method of nearest neighbor method.

105 Accurate representation of geophysical features such as topography and bathymetry is also
106 important in earth system models and their components, and model performance depends on the
107 interpolation method used to assign the topographic information over a numerical mesh in
108 processed-based models. Yan et al. (2014) compared different interpolation methods, including
109 IDW, global polynomial interpolation, local polynomial interpolation, radial basis functions,
110 ordinary kriging (OK), simple kriging (SK), universal kriging (UK), and co-kriging (CK) to
111 determine the water/land boundary point elevation based on *in situ* water level data from 14
112 control stations in Dongting Lake. They used a cross-validation method to select the optimal
113 method, which was found to be the OK method. Merwade (2009) studied the effect of spatial
114 trend on interpolation of river bathymetry, and compared the performance of different
115 interpolation methods. The number of measurements and their spatial arrangement, as well as
116 channel morphology and geology were found to influence the accuracy of the interpolation
117 results (Merwade, 2009). Due to the effects of these and other factors on the performance of
118 various methods, comparisons of different spatial interpolation methods could not point out the
119 best universal interpolation method (Li and Heap, 2008; Šiljeg et al., 2015).

120 In this paper, we propose a novel manifold method to assimilate different types of
121 spatiotemporal data in integrated earth system models based on the hypothesis that an
122 environmental dataset (including independent variables such as longitude, latitude, and time, and
123 the measured variables of interest) can be mapped onto an underlying differential manifold. A
124 manifold (M) is an n-dimensional topological space such that each point of M and its
125 neighborhood can be approximated by a small flat piece in the Euclidean space, $\mathbb{R}^n$. We can
126 think of a manifold as a set of low-dimensional curves and surfaces within higher dimension
127 Euclidean spaces (Victor and Pollack, 2010). Some typical examples of manifolds are smooth
128 surfaces, such as a torus (Fig. 1a) or a sphere (Fig. 1b), where each point and its neighborhood
129 can be approximated by a small flat linear-subspace within the three-dimensional Euclidean
130 space. Another example of a manifold in a high dimensional space is a Calabi-Yau manifold
131 which has found important applications in theoretical physics (e.g. superstring theory). Fig. 1c
132 shows a two-dimensional cross-section of a six-dimensional Calabi-Yau manifold. Surfaces of
133 all these three manifolds are not a Euclidean space. The laws of the Euclidean geometry,
134 however, are valid locally.

135 Working directly in the high dimensional space generally involves dealing with complex
136 algorithms. Modeling the high dimensional data using manifolds with fewer degrees of freedom
137 has captured a great deal of attention recently (Zhang et al., 2016). The use of low-dimensional
138 manifolds not only reduces computational load for further processing, but also helps visualize the
139 entire dataset, which is an important initial step to make sense of the data before proceeding with
140 more goal-directed modeling and analyses (Ma et al., 2011).

141 The classic method of principal component analysis (PCA), including the more recent Robust
142 PCA (Candès et al., 2011), is arguably the most popular framework for approximating a set of

143 high-dimensional data points by a low dimensional linear space. PCA does not work well when
144 the underlying data structure is non-linear. Under such scenario, PCA methods require
145 approximating the input data points using higher dimensional linear spaces to reduce the
146 approximation error. Instead of assuming that the input data follow a linear structure, an
147 alternative solution is learning the true underlying low-dimensional structure of the data. The
148 problem of non-linear dimensionality reduction for a set of high dimensional data points is
149 known as manifold learning. Examples of early works for non-linear dimensionality reduction
150 include Isomap (Tenenbaum et al., 2000), local linear embedding (LLE) (Roweis, 2000), and
151 Eigenmaps (Belkin and Niyogi, 2003), which have been used to learn the manifold structure of
152 data. Since then, the manifold model has been exploited extensively in numerous applications
153 such as face recognition, action classification, segmentation, image denoising, image/video
154 super-resolution, and multi-scale image analysis (Carin et al., 2011; Dang et al., 2013, 2014;
155 Dang and Radha, 2015).

156 Most of the above manifold learning methods have been inspired by linear techniques, mainly
157 based on the assumption that non-linear manifolds can be approximated by locally linear parts
158 (Mordohai and Medioni, 2010). Two pioneering works in this area are the Isomap approach
159 (Tenenbaum et al., 2000) and the LLE algorithm (Roweis, 2000). The Isomap algorithm aims to
160 preserve the geodesic distance among points from the input dataset. On the other hand, the LLE
161 algorithm targets the local linear geometry of neighbors in a manifold. Numerous works on
162 manifold learning have been further developed. A comprehensive review of prior works can be
163 found in (van der Maaten et al., 2009).

164 Data assimilation methods seek to estimate both model parameters and model states and
165 significant progress has been made in the development of joint state-parameter estimation

166 methods such as the ensemble Kalman filter (EnKF) and its variants (Moradkhani et al., 2005;
167 Evensen, 2006; Pathiraja et al., 2016). The parameters estimated using these approaches are
168 usually time-dependent although most dynamical models used in earth sciences use time-
169 invariant parameters. Yang et al. (2007) describe an alternative optimization-assimilation
170 approach for soil moisture in which they first estimate parameters within a long optimization
171 window and then estimate model states within a short assimilation window. Due to the
172 computational nature of the three-dimensional process-based models considered in the present
173 study, we do not focus on the use of methods similar to the EnKF method. Instead we estimate
174 optimal (relative to the observations) model parameters and states over the simulation period.
175 Future papers will focus on the use assimilation methods such as the EnKF with the framework
176 of manifold methods.

177 In this paper, the effectiveness of the presented manifold algorithm is evaluated through
178 assimilation of geophysical and meteorological data in integrated land surface – subsurface and
179 lake models, although the methods described are general and can be used in many other areas of
180 computational geosciences. We first apply the proposed method to reconstruct wind fields (time-
181 varying vector fields) over Lake Michigan. Since currents in Lake Michigan are primarily driven
182 by wind, we expect to improve the simulation of hydrodynamic and biophysical variables of
183 interest by improving the wind fields. Instead of relying on the cross-validation of interpolated
184 wind data, however, we use a well-tested hydrodynamic model of Lake Michigan and compare
185 current measurements with simulated currents to test the interpolation methods. The manifold
186 methods are used to reconstruct meteorological data including wind fields, cloud cover, dew
187 point, pressure, shortwave and longwave solar radiation, relative humidity, and air temperatures
188 for improved simulation of circulation in Lake Michigan. Then the method is applied to

assimilate bathymetry data as a scalar field for use in a hydrodynamic model of Gull Lake - a relatively large (8 km^2 surface area) and deep (34 m maximum depth) clear water lake in the in Kalamazoo County in southwestern Michigan. In the third example, time-dependent fields of participation are interpolated over the Grand River watershed and used as input for an integrated, land surface – subsurface processes model (PAWS+CLM; Shen and Phanikumar, 2010; Shen et al., 2013). Grand River watershed is located in the middle of Michigan's Lower Peninsula and is the second largest watershed in Michigan. In this example, the manifold method is tested using stream discharge outputs of the PAWS+CLM model which has been tested in several catchments in the past (Niu et al., 2014; Niu and Phanikumar, 2015; Shen et al., 2014, 2013).

2 Materials and Methods

2.1 Manifold approach

Based on Einstein's theory of relativity, physical events are located on the continuum (manifold) of space-time. Therefore, station locations and times of observations form a space-time manifold viewed as a four-dimensional vector space. One way to handle spatiotemporal interpolation problems, inspired by this concept, is to integrate space and time simultaneously (Li and Revesz, 2004). An underlying assumption behind this approach is that time and space dimensions can be treated as equally important (Li et al., 2014). In order to add time as another dimension of space, time values are needed to be scaled for a spatiotemporal dataset by a scaling speed (Li et al., 2014; Schwab and Beletsky, 1998). For a point measurement, we can then define a four-vector $P^{\mu} = (ct, \vec{x})$ where c is a time scale, t is the time coordinate and $\vec{x}$ is a three-dimensional vector space. We assume that the set of high-dimensional data points P (and the estimated data points P_0) belongs to a differential manifold M , which may be curved and have

a complicated topology, but the neighborhood of each point is approximately similar to a small piece of Euclidean space (resembles $\mathbb{R}^D$). Since a traditional distance measure is built upon the geometry of Euclidean space, we adapt the calculation to a neighborhood or a small region of the assumed manifold.

An example of a one-dimensional curve in Fig. 2 illustrates the general idea of the manifold estimation approach. The set of points P in Fig. 2 includes sample data points where we have measured data as well as a point P_0 where data are missing. For example, in the context of the wind field data, one full measurement (or data point) includes five components: time, longitude, latitude, wind speed, and wind direction. The partially missing data point may contain known components (time, longitude, latitude) and unknown or missing components (wind speed and wind direction).

Consider a smooth *n*-dimensional manifold *M* embedded in a *D*-dimensional Euclidean space. Suppose that it is desired to estimate the wind field for a data point $P_0 \in \mathbb{R}^n$ from a set of training data points that belong to a manifold M. The space/time coordinates of the point (the independent variables) are known, however, the data (the dependent variable) are missing. We denote $P_0 = \begin{bmatrix} P_0^{\mu} \\ P_0^{\nu} \end{bmatrix} \in \mathbb{R}^n$ as the data point using the superscript μ to denote the independent variables and the superscript ν to denote the dependent variable which is the missing component of interest here. $P_0^{\mu} \in \mathbb{R}^{n_\mu}$ is the sub-vector of the known components, and $P_0^{\nu} \in \mathbb{R}^{n_\nu}$ $(n_\mu + n_\nu = n)$ is the corresponding sub-vector (e.g., wind velocity vector) for the missing component where $P_0^{\nu} = \vec{V} = (u, v)$ and u, v are the orthogonal components of the wind velocity

($\vec{V}$). The training data points, for example $P = \{P_1, P_2, ..., P_7\}$ in Fig. 2, also include the two components $P_i = \begin{bmatrix} P_i^{\mu} \\ P_i^{\nu} \end{bmatrix} \in \mathbb{R}^n$, but there is no missing component here since both dependent and independent variables are assumed to be known at the nearby stations. Given a point $P_0^{\mu} \in \mathbb{R}^{n_\mu}$, the algorithm locates a set of nearest points to P_0^{μ} based on the distances $d(P_i^{\mu}, P_0^{\mu})$ between pairs of points P_i^{μ} and P_0^{μ}. In order to determine local neighbors of P_0^{μ}, we can calculate the distances between P_0^{μ} and either all other points within a fixed radius ε, or all of its *k* nearest neighbors (Tenenbaum et al., 2000). Then, a tangent space (linear subspace) of the manifold M at the point P_0 is created from the set of nearest points (Fig. 3a), denoted by $T_{P_0}(M) = \begin{bmatrix} T^{\mu} \\ T^{\nu} \end{bmatrix}$ where T^{μ}, T^{ν} denote the tangent spaces for the independent and dependent variables in the data at P_0^{μ} and P_0^{ν}. Finally, the point $P_0 = \begin{bmatrix} P_0^{\mu} \\ P_0^{\nu} \end{bmatrix} \in \mathbb{R}^n$ will be located as the closest point that belongs to that tangent space.

To represent the closest distance between a point and a tangent space, we use the Euclidean distance of an orthogonal projection from that point to the tangent space. Since a tangent space is a linear space (or affine space in a more general case), one point can orthogonally project into that space. The question is how to define neighbors for each data point? The underlying idea is how to define similarity distance among the training data points, and then the overall similarity matrix. Several methods have been considered in the past, such as k-nearest neighbors (Press et al., 2007), ϵ-ball method (Allard et al., 2012) or the use of sparse representation theory (Dang et al., 2014, 2013; Dang and Radha, 2015). To approximate the wind field, we do not focus on

analyzing a predetermined set of tangent spaces as was done earlier (Dang et al., 2014, 2013), but instead create a tangent space for each input data point as creating a tangent space for a given input data leads to a better approximation of the manifold.

The estimation of P_0^{ν} is performed using the following steps:

1. Given a set of neighboring points, estimate the tangent space T^{μ} at the point of interest, P_0 :
   Details of the method for creating a tangent space from a set of data points are described in Appendix-A and in Dang et al. (2014). One simple method is to create a tangent space using singular value decomposition (SVD, Press et al., 2007). By way of an example in Fig. 2, a tangent space (the line *b*) is created for P_0^{μ} from a set of its neighboring points (P_5^{μ} and P_6^{μ}). This tangent space at $P_0^{\mu} \in M$ is denoted by T^{μ}.
2. Find the orthogonal projection of P_0^{μ} onto the tangent space:

The closest point $P' \in \mathrm{T}^{\mu}$ to the given point P_0^{μ} is located at the intersection of the line *b* and the line perpendicular to it which passes through the point P_0^{μ}. P' which is a projection of P_0^{μ} onto the subspace T^{μ} can be represented as an approximation of point P_0^{μ}. The orthogonal projection of vector point P_0 in a high-dimensional space onto a low-dimensional vector subspace is given by:

$$\prod\nolimits_{\mathrm{T}^{\mu}}(P_0^{\mu}) = A(A^T A)^{-1} A^T P_0^{\mu} = AA^{+} P_0^{\mu} \qquad (1)$$

where $A = T^{\mu} \in \mathbb{R}^{D \times n}$ is a full rank matrix with $n < D$ containing the set of points on the tangent space of P_0^{μ} and $\prod\nolimits_{\mathrm{T}^{\mu}}(P_0^{\mu})$ denotes the projection of P_0^{μ} onto the subspace T^{μ}. This

projection is derived from the solution of the normal equation $A^T Ax = A^T P_0^\mu$ which is equivalent to the associated least squares solution of $Ax = P_0^\mu$. Due to the difficulty associated with inverting a general matrix that may be singular or non-square depending on the number of neighboring points selected in the manifold method, the problem (1) can be posed as a minimization problem in which the Moore-Penrose pseudoinverse A^+ (Golub and Van Loan, 2013) of the original matrix A is used. The pseudoinverse A^+ generalizes the concept of matrix inverse and arises in the minimum norm (that is, approximate as opposed to exact) or best-fit (in a least squares sense) solution to a system of linear equations. The problem: $\underset{\mathbf{x}}{\text{minimize}} \left\| Ax - P_0^\mu \right\|_2$ has the solution $x = A^+ P_0^\mu$. The pseudoinverse can be computed using SVD as follows: if $A = U\Sigma V^T$, where U, V denote unitary matrices and Σ is a diagonal matrix containing the singular values of A, then $A^+ = V\Sigma^+ U^T$. We used the function `pinv` to compute the pseudoinverse in MATLAB.

3. Find a linear representation coefficient vector α of that projection onto the tangent space:

This coefficient is calculated by solving the following equation:

$$\prod_{T^\mu} (P_0^\mu) = A \cdot \alpha \qquad (2)$$

4. Estimate the missing components of the point P_0 (P_0^ν):

The last step is finding a point on the subspace T^ν that is closest (in norm) to the point P_0. In order to do that, T^ν is projected using the projection coefficient α computed in step 3:

$$P_0^\nu \cong T^\nu . \alpha \qquad (3)$$

The result of this projection is the closest point to P_0^ν that belongs to its subspace. In this algorithm, high-dimensional coordinates of selected neighborhoods on the manifold are

projected to a low-dimensional subspace. An alternative to this approach is to use kernel regression to assign a weight to each neighbor based on the distance from P_0^{μ} (Fig. 3b). A weight for each selected neighborhood can be computed using the following Gaussian kernel function:

$$W_i = e^{-\frac{\left(P_i^{\mu} - P_o^{\mu}\right)^2}{2\sigma^2}},\ \sigma = \sqrt{\mathrm{var}\left(d\left(P_i^{\mu}, P_o^{\mu}\right)\right)} \tag{4}$$

Examples of manifolds representing geophysical (bathymetry) and meteorological (wind) data are shown in Figures 4(a) and (b). These figures support the assumption that the manifold can be considered as being linear locally, but with complicated topology overall.

2.2 Test case: Analytical function

Before applying the manifold method to reconstruct complex geophysical and meteorological data, we first evaluate the effectiveness of the method in reproducing an analytical function, since errors can be computed relative to the known function values; therefore, the F7 function suggested by Lazzaro and Montefusco (2002) and Renka and Brown (1999) is used:

$$F7(x, y) = 2\cos(10x)\cdot\sin(10y) + \sin(10x\cdot y) \tag{5}$$

where the domain of F7 is restricted to $0 \le x \le 1$ and $0 \le y \le 1$ (Fig. 5a) . Three sets of sparse random points from a normal distribution were generated in the domain with numbers of sampling points of 30, 60, and 90. The F7 function was sampled randomly as shown in Fig. 5b and the manifold method was tested by withholding one point at a time and estimating its associated value from the remaining points using the manifold method, in addition to other

311 methods such as the natural neighbor, nearest neighbor, and IDW interpolations. Since known
312 components of the scatter points are located in the two-dimensional X-Y plane, at least two
313 neighboring points are needed to form a tangent space for the manifold method. Therefore, for
314 simplicity, only two nearest neighbors are used in both manifold and IDW interpolation methods.

315 ## 2.3 Assimilating Meteorological Data for Improved Lake Circulation Modeling:
316 ## Lake Michigan

317 The proposed method was first applied for the reconstruction of wind fields (time-varying
318 vector fields) over Lake Michigan. Hourly wind speed and direction data during April-
319 September 2008 were obtained from the National Data Buoy Center (NDBC) weather stations
320 surrounding the lake (Fig. 6). The wind measurements were adjusted to a 10 m anemometer
321 height using the profile methods described in Schwab (1987). Since the aerodynamic roughness
322 over the lake is much lower compared to its counterpart over the land, an empirical overland-
323 overlake adjustment was applied to the wind speeds recorded by overland stations (Schwab and
324 Beletsky, 1998). The datasets of wind speed and direction were converted to two coordinates in
325 the Cartesian coordinate system (x and y directions).

326 Instead of using the cross-validation method to evaluate the interpolated wind data, results from
327 the hydrodynamic model of the lake were compared with current measurements to test the
328 applied method. To this end, a well-tested three-dimensional hydrodynamic model of the lake
329 (Safaie et al., 2016) was used. The model was based on the unstructured grid Finite Volume
330 Community Ocean Model (FVCOM; Chen et al., 2006) which was successfully used in the past
331 in ocean (Li et al., 2014), lake (Nguyen et al., 2014) and river (Anderson and Phanikumar, 2011)

332 modeling. Details of the unstructured mesh used in the hydrodynamic model are presented in
333 Table 1.

334 Wind fields from April to September 2008 were reconstructed at the locations of nodes in the
335 numerical mesh. Other hourly meteorological observations related to heat flux fields, including
336 air temperature, cloud cover, dew point, long-wave solar radiation, short-wave solar radiation,
337 and relative humidity, obtained from the National Climatic Data Center (NCDC) and NDBC
338 stations, were interpolated over the computational grid using a smoothed natural neighbor
339 method with a smoothing radius of 30 km. Air pressure was assumed to be constant (10^5 Pa)
340 through the course of the study and a constant startup water temperature with a value of 2.5 ^{o}C
341 was used in the model. The overlake dew points were estimated from overland observations
342 using an empirical formula described in (Schwab and Beletsky, 1998). Air temperature and cloud
343 cover were used to estimate long-wave solar radiation (Parkinson and Washington, 1979) and
344 short-wave solar radiation was modeled using the clear-sky value and cloud cover (Nguyen et al.,
345 2014). Six arc-second bathymetric data obtained from the NOAA National Geophysical Data
346 Center (NGDC) combined with two-meter resolution LIDAR data along the Indiana coast from
347 the National Oceanic and Atmospheric Administration (NOAA) were interpolated to the
348 numerical mesh using the natural neighbor method (Safaie et al., 2015).

349 Three bottom-mounted, upward-looking Acoustic Doppler Current Profilers (ADCPs) were
350 deployed at stations M, BB and S (Fig. 6) in southern Lake Michigan from early June to late
351 August 2008 to measure nearshore currents for model testing (Thupaki et al., 2013; Safaie et al.,
352 2016). The hydrodynamic model was run from April to August 2008 to have a two-month spin-
353 up period. Evaluation of the manifold method was carried out by comparing the simulated
354 currents with data collected by the ADCPs. Comparisons between simulated and observed

currents can be improved by identifying an optimal set of parameters in the manifold method. These parameters include: an optimum number of the nearest neighbors to create a tangent space, the time scale c, and parameters of the Gaussian kernel function. In addition, the method used for creating a tangent space from a set of data points (Appendix-A) can be changed to improve the agreement between simulated and observed currents. The manifold method for the reconstruction of wind fields was directly applied to reconstruct the other six scalar observations to calculate the heat flux fields. This time, however, P^v is a scalar, rather than a vector.

2.4 Assimilating Geophysical Data for Improved Lake Circulation Modeling: Gull Lake

In the second example, the bathymetry of Gull Lake was reconstructed using a manifold method. The lake bathymetry data were collected using a SonTek RiverSurveyor M9 system. The M9 system has an Acoustic Doppler Profiler (ADP) with two sets of four profiling beams and one vertical acoustic beam (0.5-MHz echo-sounder) for river discharge measurements and bathymetric surveys. The system was equipped with differential GPS with sub-meter precision and mounted on a SonTeck hydroboard to avoid high pitch and roll angles. The vertical acoustic beam has a range of 0.2 m to 80 m with an accuracy of 1% and a resolution of 0.001 m. The bathymetry survey was performed in four days (June 9 – June 12, 2015) by collecting data along longitudinal and transverse transects of the lake with an approximate interval of 200 m between each transect pair and sampling interval of 0.2 m- 2 m along the transects depending on the boat speed (Fig. 7).

In order to assimilate the bathymetry of the lake, a three-dimensional hydrodynamic model based on FVCOM has been developed for the lake during the period of thermal stratification

377 (June-August of 2014). The hydrodynamic equations were solved by the numerical model on an
378 unstructured grid and details are given in Table 1.

379 The meteorological observations for calculation of wind and heat flux fields were obtained
380 from NCDC, Weather Underground (https://www.wunderground.com), and the Kellogg
381 Biological Station Long-Term Ecological Research (KBS LTER, http://lter.kbs.msu.edu)
382 stations, a total of 22 locations surrounding Gull Lake from May to August (Fig. 8). Instead of a
383 constant air pressure, hourly air pressure data recorded by the KBS LTER stations were used to
384 improve the performance of the model. This also helped in the calculation of water density in
385 FVCOM based on a polynomial expression (Jackett and Mcdougall, 1995) that takes pressure
386 into account. After applying overland-overlake adjustments, all observations were interpolated
387 over the numerical mesh using a smoothed natural neighbor method with a smoothing radius of
388 15 km. This radius provided the best simulated results between the ranges of 0 to 30 km. Air
389 temperatures were adjusted using the empirical formula of $T_a = 0.4T_{la} + 0.6T_w$ (Schwab and
390 Beletsky, 1998), where T_a is the adjusted air temperature over water, T_{la} is the air temperature
391 reported by overland stations, and T_w is the surface water temperature. The surface water
392 temperature was collected using an Onset HOBO Pro v2 sensor with an accuracy of 0.2^{o}C. A
393 linearly varying startup water temperature was used with a value of 12^{o}C at the water surface and
394 4^{o}C at the depth of 10 m. The hydrodynamic model was tested using observed current data
395 measured using a Teledyne - RDI Sentinel-V ADCP (1000 kHz frequency with a bin size of 0.3
396 m) deployed in the nearshore waters of the lake in approximately 10 m of water (Fig. 7b).
397 Finally, the bathymetry of the lake interpolated to grid nodes using the manifold method was
398 assimilated into the model.

2.5 Assimilating Precipitation Data for an Integrated, Distributed Hydrologic model

In the third example, we simulate the hydrology of a watershed in the Great Lakes region using the spatial distribution of six-year (2000-2005) daily participation data over the Grand River (GR) watershed in Michigan using an integrated, process-based hydrologic model, PAWS+CLM (Shen and Phanikumar, 2010; Shen et al., 2013; Niu et al., 2014). The model is able to simulate different hydrologic components and states including channel discharge, surface runoff, evapotranspiration, groundwater, soil moisture, soil temperature, and vegetation growth. PAWS+CLM uses a structured finite-volume grid to solve the governing partial differential equations for different hydrologic components. Governing equations and numerical details of PAWS have been described in (Shen and Phanikumar, 2010) and in Table 2 of (Niu et al., 2014).

The Grand River (GR) watershed (Fig. 9) was selected as our study domain. GR watershed is located in the middle of Michigan's Lower Peninsula and it is the second largest watershed in Michigan. The watershed has a drainage area of 14, 431 km^2 and drains portions of 15 counties in Michigan. The GR stretches 420 kilometers to the outlet at Grand Haven on Lake Michigan and it is the longest river in Michigan. Shen et al. (2014) described the details of data input and integration algorithms of PAWS+CLM; thus we simply introduce the basic data input and processing information for our model in this section. We used a grid resolution of 1 km x 1 km for horizontal discretization which produced a 170×195 mesh for the GR watershed and 20 vertical layers to simulate the vadose zone dynamics by solving the Richards equation and 2 layers for the groundwater domain (unconfined and confined aquifers) (Table 1).

For topographic calculations (e.g. surface slope and overland flow), the 30 m resolution National Elevation Dataset (NED, http:/ned.usgs.gov) from U.S. Geological Survey (USGS) was used as the Digital Elevation Model (DEM). For river network simulation, National Hydrography Dataset (NHD) from USGS was assimilated and reorganized as 'river segments' with a length of one kilometer. We used the 30 m resolution raster data provided by the Michigan Department of Natural Resources, i.e. the Integrated Forest Monitoring Assessment and Prescription (IFMAP) dataset as the land use and land cover (LULC) layout (MDNR, 2010). Soil information was obtained from Soil Survey Geographic (SSURGO) (Soil Survey Staff) database from U.S. Department of Agriculture. This information was processed by the pedotransfer functions provided in ROSETTA (Schaap et al., 2001) to provide soil properties of water retention and unsaturated conductivities. Climate driven data (e.g. precipitation, daily maximum temperature and minimum temperature, wind speed) are acquired as point input (Fig. 9) from National Climatic Data Center (NCDC, 2010) of the National Oceanic and Atmospheric Administration (NOAA).

In this study, 14 rain gauges in the GR were selected to obtain the spatial distribution of rainfall over the watershed and for assimilation into the model. Previous applications of the PAWS+CLM model used the nearest neighbor method as the default for processing precipitation data. In our work, the manifold method was tested by evaluating the stream discharge outputs of PAWS+CLM against USGS data. The parameters of PAWS+CLM are listed in Table 1 of (Shen et al., 2013), most of which are spatially distributed and have taken into account the spatial heterogeneity of parameters such as the soil parameters and groundwater hydraulic conductivities.

3 Results and Discussion

3.1 Analytical function

True values of the analytical function at each of the randomly selected sampling locations were compared with the estimated values obtained by the manifold method as well as other standard interpolation methods. The performance statistics for this example are provided in Table 2. For all methods, the approximation of the F7 function improved by increasing the number of sampling points. In this particular example, the results show that the manifold method produced better overall performance compared to the other three methods considered. However, the best method in this example might perform differently on another test function or for a different sampling point distribution; therefore, we examine the performance of the method for other datasets in the following sections.

3.2 Lake Michigan

Due to the sparse distribution of weather stations around Lake Michigan, it was not clear *a priori* how many neighboring stations would provide an adequate representation of the data. Since choosing a relatively few (e.g., three) neighboring stations in this situation would involve using information from stations that are far apart as neighbors, we used kernel regression to assign weights to each station depending on the distance from the point of interest. For each node of the numerical grid of Lake Michigan, k number of nearest neighbors were selected and their assigned weights were projected to a low-dimensional subspace. The free parameters in the method are c (time scale), σ (the parameter used in kernel regression), and k. The standard deviation of weather station distances from the point of interest was used for the parameter σ in kernel regression. Performance of the manifold method as measured by a comparison of

simulated and observed currents in Lake Michigan is summarized in Table 3 relative to the other standard methods considered. We note that the manifold method based on kernel weighting considering all stations produced the best overall performance as measured by the root mean squared error (RMSE) between the observed and simulated currents. The performance of the method without kernel regression and with only three neighboring stations was comparable to the other methods but slightly inferior to the natural and nearest neighbor methods. Fig. 10 shows the RMSE and R^2 values for different numbers of nearest neighbors at different ADCP locations. Having all stations to create the tangent space for the manifold method resulted in a better representation of wind fields, and improved the results of the hydrodynamic model (Fig. 11 shows the comparison at station M). We can see that the manifold method performs better than the IDW method at two of the offshore stations (M and BB) but not at the nearshore location S. We believe that the reason for this has to do with the fact that in the nearshore region there are a number of additional processes (waves, wave-current interactions etc) which are not simulated in our model. Therefore model performance in that region cannot be directly attributed to the wind field. At the other two offshore stations M and BB, where the flow is predominantly wind-driven, an improvement in the simulated hydrodynamic fields can be seen.

Finally, cross-validation was used to compare the performance of the manifold method with other standard methods for the same Lake Michigan datasets. The performance metrics are summarized in Table 4. In this cross-validation method, one weather station was withheld to be used later for validating the manifold method, and all other stations surrounding the lake were used for the manifold training set. This process was repeated so that each weather station was given a chance to be part of this validation process. Based on these results the proposed manifold method with three nearest neighbors gave better results compared to other standard methods.

However, the performance of the hydrodynamic model based on these methods was relatively inferior compared to the performance of the model when the manifold method used all neighboring points. All different versions of the manifold methods had reasonable computational efficiency. The computational time for the O-kriging was high due to the time needed for finding the best variogram at each time step.

3.3 Gull Lake

Wind and heat flux fields of the Gull Lake were interpolated over the numerical mesh of the lake using the natural neighbor method. Then, the bathymetry of the lake was interpolated over the mesh using the same natural neighbor method to develop the initial version of the lake hydrodynamic model. The raw bathymetry data, which has some regions of steep bathymetry change, created artificial currents in the model due to an error in the pressure gradient force introduced by the sigma-coordinate system of FVCOM (Mellor et al., 1998). Therefore, the interpolated bathymetry was smoothed with a radius of 100 m in order to reduce the errors. The results of the developed model using natural neighbor method and IDW with three nearest neighbors are presented in Fig. 12.

The model was used to assimilate the bathymetry of the lake based on the manifold method. The bathymetry data were reconstructed from the tangent space of the manifold with three nearest neighbors and smoothed with the same method described above. The hydrodynamic model was run with the reconstructed bathymetry. The final comparisons of the vertically-averaged velocity profiles at the ADCP location are presented in Fig. 12. The best value of σ used in kernel regression was equal to the standard deviation of distances of observational points where water depth values are available within a search radius of 50 m from the point of interest.

When the number of samples within this radius was smaller than 100, σ value was calculated based on locations of 100 nearest samples. This method is more accurate when enough samples are available around an estimated point, unless selecting 100 samples itself does a reasonable job. RMSE values (m/s) of eastward and northward velocities in Gull Lake for comparison of the manifold method with other standard methods used in limnology and oceanography are presented in Table 5. The statistics of cross-validation for all (=71) measured longitudinal and transverse transects are shown in Table 6. The cross-validation was performed by omitting one transect at each step and calculating the bathymetry for that transect from the rest of the observation data and repeating the process for all other transects.

3.4 Grand River watershed

The daily distributions of precipitation over the GR watershed for the period 2000 to 2005 were reconstructed from selected rain gauges using the manifold method. The precipitation fields over a period of six years were tested by comparing the simulated and observed stream discharges at selected USGS gaging stations within the watershed. In this example, kernel regression with a standard deviation of all rain gauge distances from the estimated points was used in the manifold method. The tangent space at each grid point of the numerical model was obtained from three-nearest neighbors around that point. The PAWS+CLM model was run with precipitation distributions built using the manifold, natural neighbor, nearest neighbor, and inverse distance methods. . The final comparisons of simulated and observed stream flows for USGS gauges are presented in Figs. 13 and 14. Model performance metrics (NASH, Absolute Percent Bias (APB), and RMSE) were computed to evaluate the performance of the manifold method with other standard methods (Table 7). The manifold method provided superior results as can be seen from the improved representation of baseflow over the simulation period. This can be seen clearly from semi-log

plots of the stream hydrograph comparison. The cross-validation results for precipitation are presented in Table 8. In this method, data from one rain gauge was withheld and data from the rest of the gauges were used for manifold training. This process was repeated for all 14 rain gauges. Since rain gauges have a sparse distribution, the cross-validation method could not identify the best method for this case. The results of the integrated, distributed hydrologic model, however, clearly demonstrate the efficacy of the manifold method in the reconstruction of precipitation fields. In particular, the better simulated baseflow shows the strength of using precipitation data based on the manifold method to improve the simulation of heterogeneous partitioning of surface runoff and infiltration processes.

4 Conclusions

We presented a novel manifold method of reconstructing spatio-temporal data for assimilating geophysical and meteorological data in integrated land surface subsurface, and oceanic/lake models. All three case studies illustrate the superior performance of the presented manifold algorithm over standard methods in terms of accuracy and computational efficiency. The hydrodynamic model of Lake Michigan based on the manifold method of reconstructing wind fields produced better performance relative to the other methods. The best results were obtained using kernel regression applied to all weather stations (neighbors). However, the cross-validation results show that the results of the three nearest neighbors were better than the other methods.

The Gull Lake model results indicated that the proposed method has the ability to reconstruct geophysical data at unsampled locations. The use of spatiotemporal precipitation fields constructed using the manifold method produced better stream discharge simulations compared to similar results from the nearest neighbor, natural neighbor and IDW interpolation methods in a

554 large watershed (> 10000 km^2). Finally, all three examples show that evaluating the performance
555 of interpolation methods using the cross-validation method without considering the data structure
556 and the purpose of interpolation can lead to misleading conclusions about the relative
557 performance of the methods considered. The comparisons presented here indicate that manifold
558 methods show promise for modeling Earth system processes based on data from sensor
559 networks. Our future work will combine manifold methods with approaches such as the EnKF
560 method to further improve process-based modeling of land surface, subsurface and lake/ocean
561 models.

562 Based on the results presented, we note that: (1) Details of the manifold method such as the
563 tangent space estimation, the distance metric that defines spatiotemporal proximity and other
564 details can be further improved to improve the performance of the manifold method; however,
565 these topics are beyond the scope of the present paper. (2) We do not claim that the manifold
566 method provides superior performance on all datasets and for all performance metrics but from
567 the examples considered here it appears that the manifold method may offer an attractive method
568 that is comparable or superior to other standard methods. More research is needed to understand
569 the relative strengths and weaknesses of different manifold-based approaches compared to
570 standard methods.

571 # Acknowledgments

572 This work was supported by a grant from the National Science Foundation, CyberSEES program
573 (Award # 1331852). We thank Elena Litchman, Pam Woodruff, Jim Allen, Mike Gallagher,
574 Andrew Fogiel, and Tuan D. Nguyen for their assistance with field data collection in Gull Lake.

575 We gratefully acknowledge the use of data from the Kellogg Biological Station LTER which is
576 supported by NSF (DEB 1027253) and by Michigan State University AgBioResearch.

577 Appendix-A. Tangent Space Estimation

578 To understand the local geometry of the surface $f(x)$ near a point $x \in \mathbb{R}^n$, we consider the first-
579 order Taylor series expansion of the surface:

580 $$f(\bar{x}) = f(x) + \frac{\partial f(x)}{\partial x}(\bar{x} - x) + O(\|\bar{x} - x\|^2) = f(x) + J_f(\bar{x} - x) + O(\|\bar{x} - x\|^2) \quad \text{(A1)}$$

581 where $J_f(x) \in \mathbb{R}^{D \times n}$ is the Jacobian matrix of f at the point x. If the components of $f(x)$ can
582 be written as: $f(x) = [f_1(x), f_2(x), f_3(x) \cdots f_D(x)]^T$ and $x = [x_1, x_2, x_3 \cdots x_n]^T$, then the Jacobian
583 can be written as:

584 $$J_f(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \vdots & \vdots \\ \frac{\partial f_D}{\partial x_1} & \cdots & \frac{\partial f_D}{\partial x_n} \end{bmatrix} \quad \text{(A2)}$$

585 To understand the local shape of the surface in equation (A1), we seek to determine the space
586 $(\bar{x} - x)$, such that as we move away from x, the value of the function doesn't change to within
587 first order. This is equivalent to finding the space T such that:

588 $$T = \left\{ (\bar{x} - x) \middle| J_f(x)(\bar{x} - x) = 0 \right\} \quad \text{(A3)}$$

This space is the tangent space to the surface at point x and is the right null space of the Jacobian matrix $J_f(x)$. The space orthogonal to the tangent space is the row space of the Jacobian and orthogonal representations of these spaces can be obtained from SVD. The right null space of J_f is the columns of V corresponding to zero singular values. Therefore, the tangent space of the manifold M at $y = f(x)$ is:

$$T(M) = \text{span}\left(J_f(x)\right) \tag{A4}$$

From a practical computation point of view, given a set of sample points $y = \{y_1, y_2, y_3, \cdots y_m\}$, a simple method of constructing the tangent space is to approximate it as the line/surface obtained by joining the local neighboring points. The tangent space can also be directly estimated using SVD. If C^m denotes the local covariance matrix:

$$C^m = \frac{1}{m}\sum_{i=1}^{m} y_i y_i^T = U\Sigma U^T \tag{A5}$$

where $U = [u_1, u_2, u_3, \cdots u_D]$ and $\Sigma = \text{diag}[\lambda_1, \lambda_2, \lambda_3 \cdots \lambda_D]$ denote the eigenvector and eigenvalue matrices respectively, then the optimal (in a least-squares sense) *n*-dimensional linear subspace is the span of the *n*-largest eigenvectors in *U*:

$$T(M) \cong \text{span}\{u_1, u_2, u_3, \cdots u_n\} \tag{A6}$$

5. References

Allard, W.K., Chen, G., Maggioni, M., 2012. Multi-scale geometric methods for data sets II: Geometric Multi-Resolution Analysis. Appl. Comput. Harmon. Anal. 32, 435–462. doi:10.1016/j.acha.2011.08.001

608 Anderson, E.J., Phanikumar, M.S., 2011. Surface storage dynamics in large rivers: Comparing
609 three-dimensional particle transport, one-dimensional fractional derivative, and multirate
610 transient storage models. Water Resources Research 47, 1–15. doi:10.1029/2010WR010228

611 Anquetin, S., Braud, I., Vannier, O., Viallet, P., Boudevillain, B., Creutin, J.-D., Manus, C.,
612 2010. Sensitivity of the hydrological response to the variability of rainfall fields and soils for the
613 Gard 2002 flash-flood event. J. Hydrol. 394, 134–147. doi:10.1016/j.jhydrol.2010.07.002

614 Barth, A., Azcárate, A. A., Joassin, P., Jean-Marie, B., Troupin, C., 2008. Introduction to
615 Optimal Interpolation and Variational Analysis. Presented at the SESAME Summer School,
616 SESAME Summer School, Varna, Bulgaria.

617 Belkin, M., Niyogi, P., 2003. Laplacian eigenmaps for dimensionality. Speech Commun. 1, 349–
618 367.

619 Bell, V.A., Moore, R.J., 2000. The sensitivity of catchment runoff models to rainfall data at
620 different spatial scales. Hydrol. Earth Syst. Sci. 4, 653–667. doi:10.5194/hess-4-653-2000

621 Beven, K.J., Hornberger, G.M., 1982. Assessing the Effect of Spatial Pattern of Precipitation in
622 Modeling Stream Flow Hydrographs. J. Am. Water Resour. Assoc. 18, 823–829.
623 doi:10.1111/j.1752-1688.1982.tb00078.x

624 Blumberg, A.F., Mellor, G.L., 1987. A description of a three-dimensional coastal ocean
625 circulation model. Am. Geophys. Union 1–16.

626 Candès, E.J., Li, X., Ma, Y., Wright, J., 2011. Robust principal component analysis? J. ACM 58,
627 1–37. doi:10.1145/1970392.1970395

628 Carin, L., Baraniuk, R.G., Cevher, V., Dunson, D., Jordan, M.I., Sapiro, G., Wakin, M.B., 2011.
629 Learning Low-Dimensional Signal Models. IEEE Signal Process. Mag. 28.
630 doi:10.1109/MSP.2010.939733

631 Chen, C., Beardsley, R., Cowles, G., 2006. An Unstructured Grid, Finite-Volume Coastal Ocean
632 Model (FVCOM) System. Oceanography 19, 78–89. doi:10.5670/oceanog.2006.92

633 Daly, C., Gibson, W.P., Taylor, G.H., Johnson, G.L., Pasteris, P., 2002. A knowledge-based
634 approach to the statistical mapping of climate. Clim. Res. 22, 99–113.

635 Dang, C., Aghagolzadeh, M., Radha, H., 2014. Image Super-Resolution via Local Self-Learning
636 Manifold Approximation. IEEE Signal Process. Lett. 21, 1245–1249.
637 doi:10.1109/LSP.2014.2332118

638 Dang, C., Radha, H., 2015. Fast Image Super Resolution via Selective Manifold Learning of
639 High Resolution Patches. Presented at the IEEE Proceedings of International Conference on
640 Image Processing (ICIP15), IEEE Proceedings of International Conference on Image Processing
641 (ICIP15), Québec City, Canada.

642 Dang, C.T., Aghagolzadeh, M., Moghadam, A.A., Radha, H., 2013. Single image super
643 resolution via manifold linear approximation using sparse subspace clustering, in: Global
644 Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE. IEEE, Austin, Texas,
645 U.S.A., pp. 949–952.

646 Evensen, G., 2006, Data Assimilation: The Ensemble Kalman Filter, Springer, N. Y.

647 Golub, G.., Van Loan, C.F., 2013. Matrix Computations, 4th ed. Baltimore: Johns Hopkins, pp.
648 756.

649 Haidvogel, D.B., Beckmann, A., 1999. Numerical Ocean Circulation Modeling. Imp. Coll. Press.

650 Zhang, H., Mendoza-Sanchez, I., Miller, E.L., Abriola, L.M., 2016. Manifold Regression
651 Framework for Characterizing Source Zone Architecture. IEEE Transactions on Geoscience and
652 Remote Sensing 54, 3–17. doi:10.1109/TGRS.2015.2448086

653 Jackett, D.R., Mcdougall, T.J., 1995. Minimal Adjustment of Hydrographic Profiles to Achieve
654 Static Stability. J. Atmospheric Ocean. Technol. 12, 381–389. doi:10.1175/1520-
655 0426(1995)012<0381:MAOHPT>2.0.CO;2

656 Lazzaro, D., Montefusco, L.B., 2002. Radial basis functions for the multivariate interpolation of
657 large scattered data sets. J. Comput. Appl. Math., Int. Congress on Computational and Applied
658 Mathematics 2000 140, 521–536. doi:10.1016/S0377-0427(01)00485-X

659 Li, J., Heap, A.D., 2008. A review of spatial interpolation methods for environmental scientists.
660 Geoscience Australia, Canberra.

661 Li, L., Losser, T., Yorke, C., Piltner, R., 2014. Fast Inverse Distance Weighting-Based
662 Spatiotemporal Interpolation: A Web-Based Application of Interpolating Daily Fine Particulate
663 Matter PM2:5 in the Contiguous U.S. Using Parallel Programming and k-d Tree. Int. J. Environ.
664 Res. Public. Health 11, 9101–9141. doi:10.3390/ijerph110909101

665 Li, L., Revesz, P., 2004. Interpolation methods for spatio-temporal geographic data. Comput.
666 Environ. Urban Syst. 28, 201–227. doi:10.1016/S0198-9715(03)00018-8

667 Li, R. et al., 2014. Observed wintertime tidal and subtidal currents over the continental shelf in
668 the northern South China Sea, J. Geophys. Res. Oceans, 119(8), pp. 5289-5310, doi:10.1002/
669 2014JC009931.

670 Luo, W., Taylor, M.C., Parker, S.R., 2008. A comparison of spatial interpolation methods to
671 estimate continuous wind speed surfaces using irregularly distributed data from England and
672 Wales. Int. J. Climatol. 28, 947–959. doi:10.1002/joc.1583

673 Ly, S., Charles, C., Degré, A., 2011. Geostatistical interpolation of daily rainfall at catchment
674 scale: the use of several variogram models in the Ourthe and Ambleve catchments, Belgium.
675 Hydrol. Earth Syst. Sci. 15, 2259–2274. doi:10.5194/hess-15-2259-2011

MacEachren, A.M., Davidson, J.V., 1987. Sampling and isometric mapping of continuous geographic surfaces. Am. Cartogr. 14, 299–320.

Masih, I., Maskey, S., Uhlenbrook, S., Smakhtin, V., 2011. Assessing the Impact of Areal Precipitation Input on Streamflow Simulations Using the SWAT Model1. JAWRA J. Am. Water Resour. Assoc. 47, 179–195. doi:10.1111/j.1752-1688.2010.00502.x

Ma, Y., Niyogi, P., Sapiro, G., Vidal, R., 2011. Dimensionality reduction via subspace and submanifold learning. IEEE Signal Process. Mag. 28, 14–126. doi:10.1109/MSP.2010.940005

Mellor, G.L., Oey, L.Y., Ezer, T., 1998. Sigma coordinate pressure gradient errors and the seamount problem. Journal of Atmospheric and Oceanic Technology 15, 1122–1131.

Merwade, V., 2009. Effect of spatial trends on interpolation of river bathymetry. J. Hydrol. 371, 169–181. doi:10.1016/j.jhydrol.2009.03.026

Mordohai, P., Medioni, G., 2010. Dimensionality estimation, manifold learning and function approximation using tensor voting. J. Mach. Learn. Res. 11, 411–450.

Moradkhani, H., Sorooshian, S., Gupta, H.V., Houser, P.R., 2005. Dual state–parameter estimation of hydrological models using ensemble Kalman filter, Adv. Water Resour. 28, 135–147.4

Nguyen, T.D., Thupaki, P., Anderson, E.J., Phanikumar, M.S., 2014. Summer circulation and exchange in the Saginaw Bay-Lake Huron system. J. Geophys. Res. Oceans 119, 2713–2734. doi:10.1002/2014JC009828

Nicótina, L., Alessi Celegon, E., Rinaldo, A., Marani, M., 2008. On the impact of rainfall patterns on the hydrologic response. Water Resour. Res. 44, 1–14. doi:10.1029/2007WR006654

Niu, J., Phanikumar, M.S., 2015. Modeling watershed-scale solute transport using an integrated, process-based hydrologic model with applications to bacterial fate and transport. J. Hydrol. 529, 35–48. doi:10.1016/j.jhydrol.2015.07.013

Niu, J., Shen, C., Li, S.-G., Phanikumar, M.S., 2014. Quantifying storage changes in regional Great Lakes watersheds using a coupled subsurface-land surface process model and GRACE, MODIS products. Water Resour. Res. 50, 7359–7377. doi:10.1002/2014WR015589

Obled, C., Wendling, J., Beven, K., 1994. The sensitivity of hydrological models to spatial rainfall patterns: an evaluation using observed data. J. Hydrol. 159, 305–333. doi:10.1016/0022-1694(94)90263-1

Oleson, K.W., Lawrence, D.M., Gordon, B., Flanner, M.G., Kluzek, E., Peter, J., Levis, S., Swenson, S.C., Thornton, E., Feddema, J., others, 2010. Technical description of version 4.0 of the Community Land Model (CLM) (No. NCAR/TN-478+STR), NCAR Technical Note. National Center for Atmospheric Research, Boulder, Colorado.

710 Parkinson, C.L., Washington, W.M., 1979. A large-scale numerical model of sea ice. J. Geophys.
711 Res. Oceans 84, 311–337. doi:10.1029/JC084iC01p00311

712 Pathiraja, S., Marshall, L., Sharma, A., Moradkhani, H., 2016, Hydrologic modeling in dynamic
713 catchments: A data assimilation approach, Water Resour. Res. 52, 1-23.
714 doi:10.1002/2015WR01719

715 Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T., 2007. *Numerical Recipes in*
716 *C++: The Art of Scientific Computing*, third ed. Cambridge University Press, New York.

717 Renka, R.J., Brown, R., 1999. Algorithm 792: Accuracy Test of ACM Algorithms for
718 Interpolation of Scattered Data in the Plane. ACM Trans Math Softw 25, 78–94.
719 doi:10.1145/305658.305745

720 Roweis, S.T., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science
721 290, 2323–2326. doi:10.1126/science.290.5500.2323

722 Safaie, A., Wendzel, A., Ge, Z., Nevers, M.B., Whitman, R.L., Corsi, S.R., Phanikumar, M.S.,
723 2016. Comparative Evaluation of Statistical and Mechanistic Models of Escherichia coli at
724 Beaches in Southern Lake Michigan. Environ. Sci. Technol., 50, 2442–2449.
725 doi:10.1021/acs.est.5b05378

726 Schaap, M.G., Leij, F.J., van Genuchten, M.T., 2001. rosetta: a computer program for estimating
727 soil hydraulic parameters with hierarchical pedotransfer functions. J. Hydrol. 251, 163–176.
728 doi:10.1016/S0022-1694(01)00466-8

729 Schuurmans, J.M., Bierkens, M.F.P., 2007. Effect of spatial distribution of daily rainfall on
730 interior catchment response of a distributed hydrological model. Hydrol. Earth Syst. Sci.
731 Discuss. 11, 677–693.

732 Schwab, D.J., 1987. Simulation and forecasting of Lake Erie storm surges. Mon. Weather Rev.
733 106, 1476–1487.

734 Schwab, D.J., Beletsky, D., 1998. Lake Michigan Mass Balance Study: Hydrodynamic modeling
735 project (No. NOAA Technical Memorandum ERL GLERL-108). Great Lakes Environmental
736 Research Laboratory, Ann Arbor, MI.

737 Shen, C., Niu, J., Fang, K., 2014. Quantifying the effects of data integration algorithms on the
738 outcomes of a subsurface–land surface processes model. Environ. Model. Softw. 59, 146–161.
739 doi:10.1016/j.envsoft.2014.05.006

740 Shen, C., Niu, J., Phanikumar, M.S., 2013. Evaluating controls on coupled hydrologic and
741 vegetation dynamics in a humid continental climate watershed using a subsurface-land surface
742 processes model. Water Resour. Res. 49, 2552–2572. doi:10.1002/wrcr.20189

743 Shen, C., Phanikumar, M.S., 2010. A process-based, distributed hydrologic model based on a
744 large-scale method for surface–subsurface coupling. Adv. Water Resour. 33, 1524–1541.
745 doi:10.1016/j.advwatres.2010.09.002

746 Šiljeg, A., Lozić, S., Šiljeg, S., 2015. A comparison of interpolation methods on the basis of data
747 obtained from a bathymetric survey of Lake Vrana, Croatia. Hydrol. Earth Syst. Sci. 19, 3653–
748 3666. doi:10.5194/hess-19-3653-2015

749 Sun, X., Mein, R.G., Keenan, T.D., Elliott, J.F., 2000. Flood estimation using radar and
750 raingauge data. J. Hydrol. 239, 4–18. doi:10.1016/S0022-1694(00)00350-4

751 Suparta, W., Rahman, R., 2016. Spatial interpolation of GPS PWV and meteorological variables
752 over the west coast of Peninsular Malaysia during 2013 Klang Valley Flash Flood. Atmospheric
753 Res. 168, 205–219. doi:10.1016/j.atmosres.2015.09.023

754 Syed, K.H., Goodrich, D.C., Myers, D.E., Sorooshian, S., 2003. Spatial characteristics of
755 thunderstorm rainfall fields and their relation to runoff. J. Hydrol. 271, 1–21. doi:10.1016/S0022-
756 1694(02)00311-6

757 Tenenbaum, J.B., De Silva, V., Langford, J.C., 2000. A global geometric framework for
758 nonlinear dimensionality reduction. Science 290, 2319–2323.

759 Tetzlaff, D., Uhlenbrook, S., 2005. Significance of spatial variability in precipitation for process-
760 oriented modelling: results from two nested catchments using radar and ground station data.
761 Hydrol. Earth Syst. Sci. 9, 29–41. doi:10.5194/hess-9-29-2005

762 Thupaki, P., M.S. Phanikumar and R.L. Whitman, 2013. Solute dispersion in the coastal
763 boundary layer of southern Lake Michigan, J. Geophys. Res. Oceans, vol. 118, No. 3, pp. 1606-
764 1617, doi: 10.1002 / jgrc.20136 (2013)

765 Thupaki, P., Phanikumar, M.S., Schwab, D.J., Nevers, M.B., Whitman, R.L., 2013. Evaluating
766 the role of sediment-bacteria interactions on *Escherichia coli* concentrations at beaches in
767 southern Lake Michigan. , J. Geophys. Res. Oceans118, 7049–7065.
768 doi:10.1002/2013JC008919

769 van der Maaten, L.J., Postma, E.O., van den Herik, H.J., 2009. Dimensionality reduction: A
770 comparative review. J. Mach. Learn. Res. 10, 66–71.

771 Victor, G., Pollack, A., 2010. Differential topology. American Mathematical Soc.

772 Wendzel, A., 2014. Constraining mechanistic models of indicator bacteria at recreational
773 beaches in Lake Michigan using easily-measurable environmental variables (M. Sc.
774 Dissertation). Michigan State University, East Lansing. MI.

775 Xue, P., Schwab, D.J., Hu, S., 2015. An investigation of the thermal response to meteorological
776 forcing in a hydrodynamic model of Lake Superior. J. Geophys. Res. Oceans 120, 5233–5253.
777 doi:10.1002/2015JC010740

778 Yan, Y., Xiao, F., Du, Y., 2014. Construction of lake bathymetry from MODIS satellite data and
779 GIS from 2003 to 2011. Chin. J. Oceanol. Limnol. 32, 720–731. doi:10.1007/s00343-014-3185-4

780 Yang, K., Watanabe, T., Koike, T., Li, X., Fujii, H., Tamagawa, K., Ma, Y., Ishikawa, H., 2007.
781 Auto-calibration system developed to assimilate AMSR-E data into a land surface model for
782 estimating soil moisture and the surface energy budget. J. Meteorol. Soc. Jpn. 85A, 229–242.

783

784

785

786

787

788

789

790 # List of Figures

Fig. 6. Locations of the ADCPs deployed during summer 2008 and weather stations surrounding Lake Michigan.

Fig. 7(a). Bathymetry of Gull Lake. (b) Boat tracks generated during the sampling survey in Gull Lake.

Fig. 8. Selected weather stations surrounding Gull Lake.

Fig. 9. Map of Grand River watershed showing the locations of the USGS gauges, rain gauges and National Hydrography Dataset (NHD) streams.

Fig. 10. Performance of the manifold method evaluated using observed and simulated currents at different stations in Lake Michigan. Different number of nearest neighbors were used to reconstruct the wind field using the manifold method with kernel regression.

Fig. 11. Comparison of simulated (black lines) and observed (red lines) vertically averaged currents at the location M in Lake Michigan. (a) Alongshore velocity (b) Cross-shore velocity

Fig. 12. Comparison of simulated (black lines) and observed (red lines) vertically averaged currents at the ADCP location in Gull Lake. (a) Eastward velocity and (b) Northward velocity

Fig. 13. Comparison of simulated and observed stream flows for USGS gauge #04116000 in (a) a linear scale, and (b) logarithmic scale.

Fig. 14. Comparison of simulated and observed stream flows for USGS gauge #04119000 in (a) a linear scale, and (b) logarithmic scale.

List of Tables

Table 1. Properties of the numerical grids used for the hydrodynamic and hydrologic

Model	Grid Classification	Element shape	Grid Resolution	# Nodes	# Elements	#Vertical layers
FVCOM (Lake Michigan)	Unstructured	Triangle	4m -5 km	12,684	23,602	20
FVCOM (Gull Lake)	Unstructured	Triangle	8-100m	5,132	9,361	20
PAWS (Grand River)	Structured	Quadrilateral	1 km	33,150	32,786	22

Table 2. Cross-validation results for the analytical function based on different sampling points selected randomly.

Sample size	Method	R^2	RMSE	Fn	PBIAS	NASH	APB (%)
30	Manifold	0.667	0.778	0.710	14.420	0.416	0.652
	Natural neighbor	0.582	0.876	0.799	-8.866	0.259	0.713
	Nearest neighbor	0.619	0.882	0.804	-14.689	0.250	0.669
	IDW	0.577	0.870	0.793	35.847	0.270	0.727
60	Manifold	0.846	0.579	0.531	-33.924	0.703	0.472
	Natural neighbor	0.816	0.615	0.564	16.912	0.664	0.466
	Nearest neighbor	0.779	0.720	0.660	-34.524	0.540	0.512
	IDW	0.832	0.603	0.553	-44.292	0.677	0.469
90	Manifold	0.891	0.502	0.432	-14.103	0.791	0.400
	Natural neighbor	0.874	0.539	0.464	-10.666	0.759	0.344
	Nearest neighbor	0.867	0.571	0.491	-28.777	0.730	0.446
	IDW	0.859	0.567	0.487	-5.974	0.735	0.416

Table 3. RMSE values (m/s) of alongshore and cross-shore velocities for comparison of the manifold method with other standard methods used in limnology and oceanography

	Loaction: M		Location: BB		Location: S	
Method	RMSE u	RMSE v	RMSE u	RMSE v	RMSE u	RMSE v
O-kriging	0.0385	0.0290	0.0590	0.0349	0.0540	0.0152
Nearest Neighbor	0.0363	0.0286	0.0580	0.0348	0.0545	0.0152
Natural Neighbor	0.0366	0.0275	0.0553	0.0334	0.0515	0.0158
Manifold (3 NBR)	0.0383	0.0276	0.0594	0.0346	0.0568	0.0158
Manifold+Kernel (3 NBR)	0.0371	0.0268	0.0576	0.0341	0.0559	0.0158
Manifold+Kernel (all NBR)	0.0304	0.0265	0.0531	0.0312	0.0568	0.0154
IDW (all NBR)	0.0328	0.0267	0.0535	0.0316	0.0498	0.0155

833

834 **Table 4.** Cross-validation results for wind field over Lake Michigan.

Method	R^2u	R^2v	RMSEu	RMSEv	Computational time (s)
O-kriging	0.441	0.572	3.497	3.853	92463.8
Nearest Neighbor	0.666	0.743	2.792	3.044	18.6
Natural Neighbor	0.693	0.794	2.558	2.750	183.6
Manifold (3 NBR)	0.690	0.801	2.433	2.595	28.1
Manifold+Kernel (3 NBR)	0.710	0.806	2.392	2.566	55.1
Manifold+Kernel (all NBR)	0.547	0.681	2.884	3.129	77.3
IDW (3 NBR)	0.724	0.822	2.278	2.458	69.3

835

836

837 **Table 5.** RMSE values (m/s) of eastward and northward velocities in Gull Lake for comparison of the
838 manifold method with other standard methods used in limnology and oceanography

Method	RMSE u	RMSE v
Natural Neighbor	0.0090	0.0205
Manifold+Kernel (3 NBR)	0.0098	0.0200
IDW (3 NBR)	0.0095	0.0204

839

840

841

842

843 **Table 6.** Cross-validation results for Gull Lake bathymetry.

Method	R^2	RMSE (m)	Fn	NASH	PBIAS
Manifold	0.890	2.011	0.222	0.678	-14.016
Natural Neighbor	0.925	1.288	0.170	0.468	-13.301
Nearest Neighbor	0.888	2.039	0.230	0.670	-17.132
IDW (3 NBR)	0.839	3.282	0.6065	0.540	-15.918

844

845

846 **Table 7.** Comparison of the manifold method with other standard methods for precipitation over the
847 Grand River watershed.

	USGS04116000			USGS04119000		
Method	NASH	RMSE	APB(%)	NASH	RMSE	APB(%)
Nearest Neighbor	0.56	37.13	36.47	0.58	58.15	30.54
Natural Neighbor	0.33	44.52	54.71	0.36	71.88	48.72
Inverse Distance	0.38	44.22	54.18	0.37	70.29	48.01
Manifold	0.59	35.56	33.76	0.63	54.57	27.87

848

849

850 **Table 8.** Cross-validation results for precipitation over the Grand River watershed.

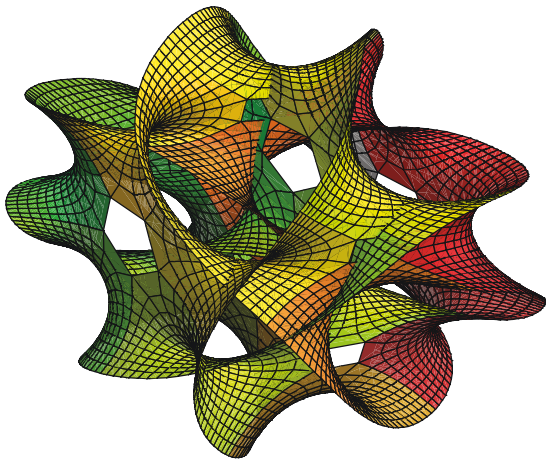
Method	R^2	RMSE (cms)	Fn	NASH	APB (%)
Manifold	0.543	7.933	0.910	0.624	72.4
Natural Neighbor	0.543	7.920	0.901	0.574	73.6
Nearest Neighbor	0.471	8.802	0.930	0.604	81.8
IDW (3 NBR)	0.567	7.542	0.907	0.619	70.0

851

Graphical Abstract (for review)

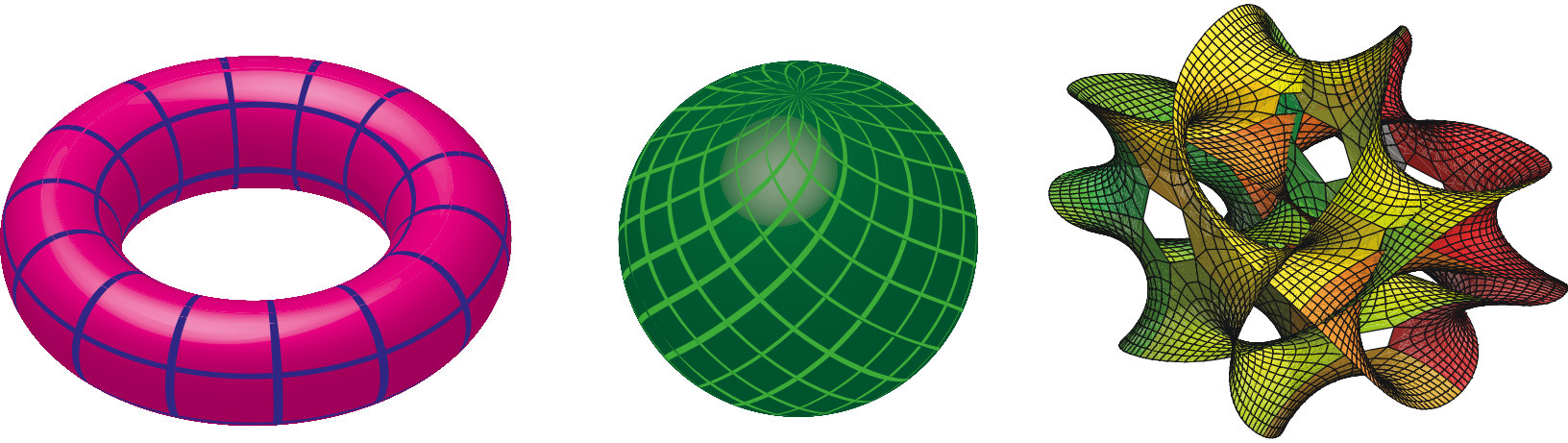
(a)

(b)



(c)

Figure 1



(a) (b) (c)

Figure 2

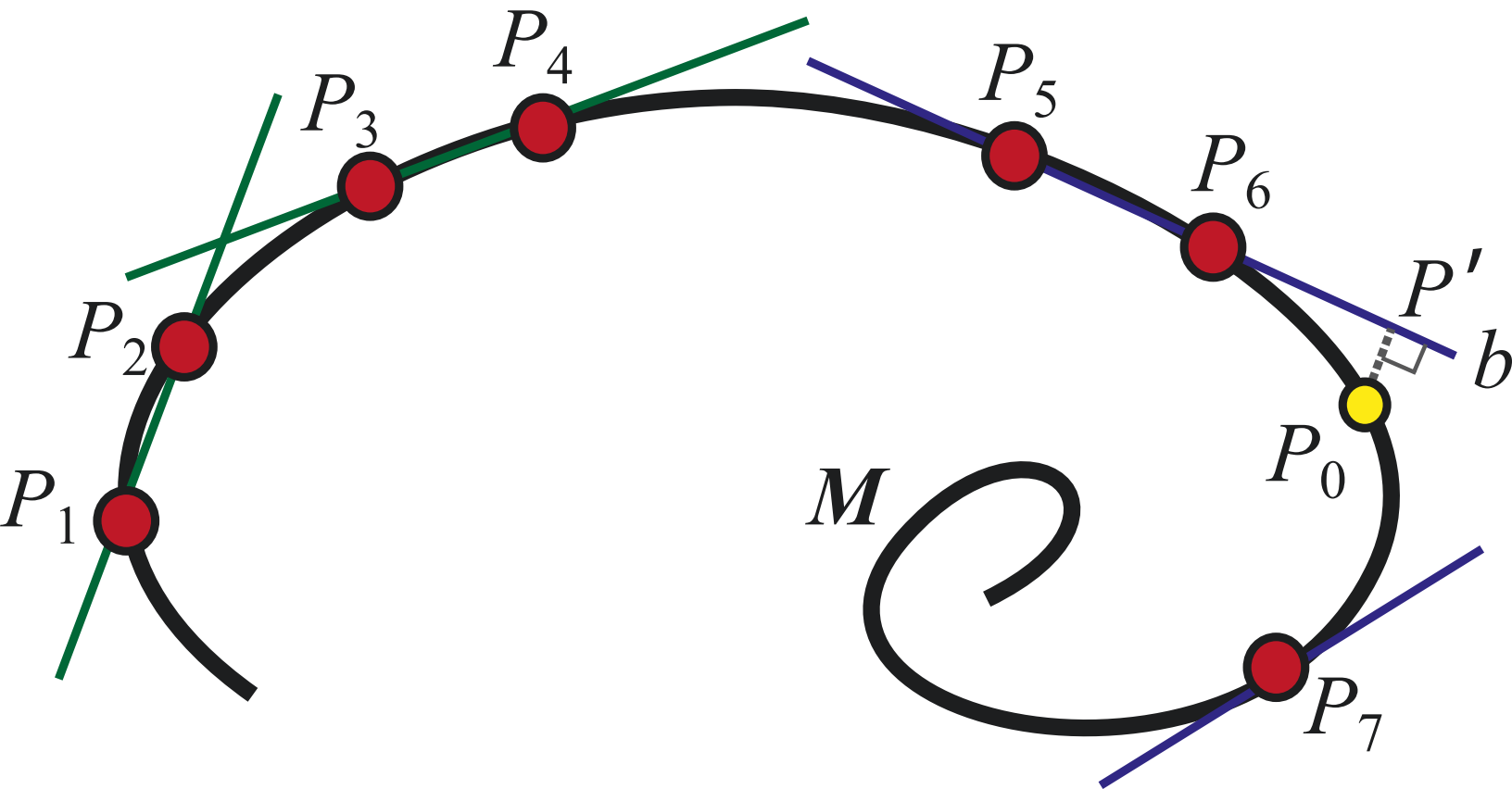


Figure 3

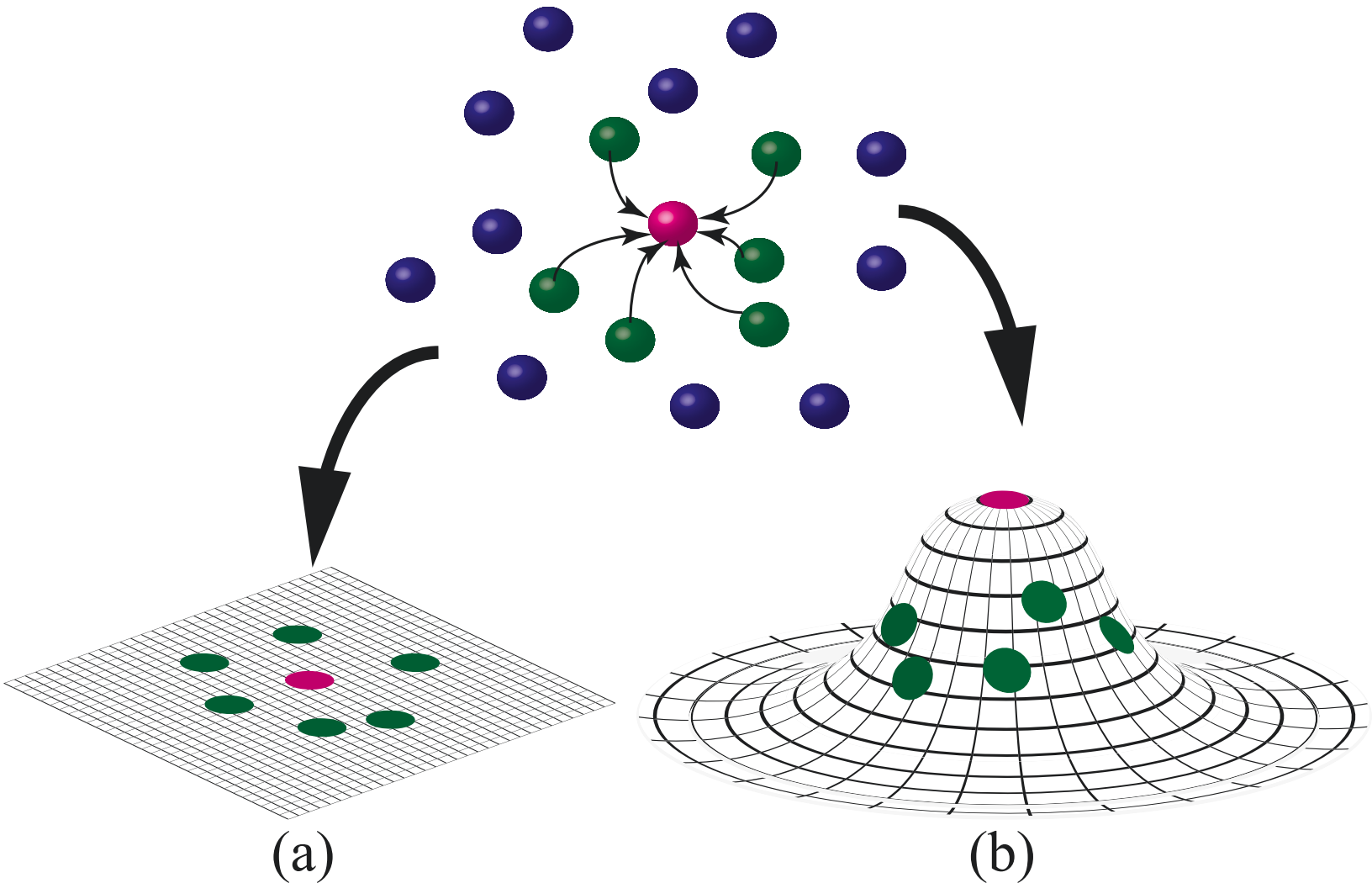


Figure 4

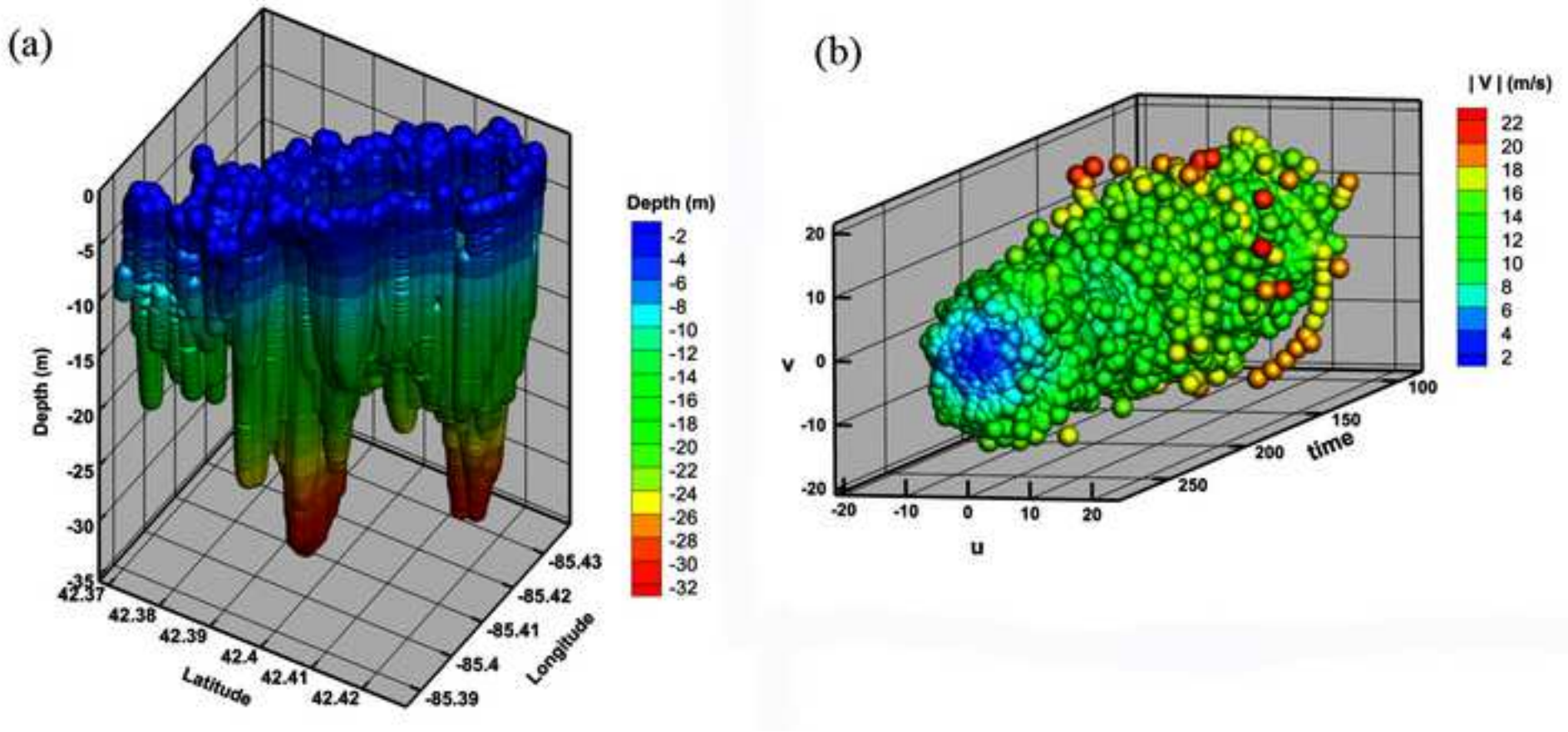

(a)
Depth (m)
0
-5
-10
-15
-20
-25
-30
-35
42.37
42.38
42.39
42.4
42.41
42.42
Latitude
-85.43
-85.42
-85.41
-85.4
-85.39
Longitude
Depth (m)
-2
-4
-6
-8
-10
-12
-14
-16
-18
-20
-22
-24
-26
-28
-30
-32
(b)
| V | (m/s)
22
20
18
16
14
12
10
8
6
4
2
v
20
10
0
-10
-20
-20
-10
0
10
20
u
100
150
200
250
time

Figure 5

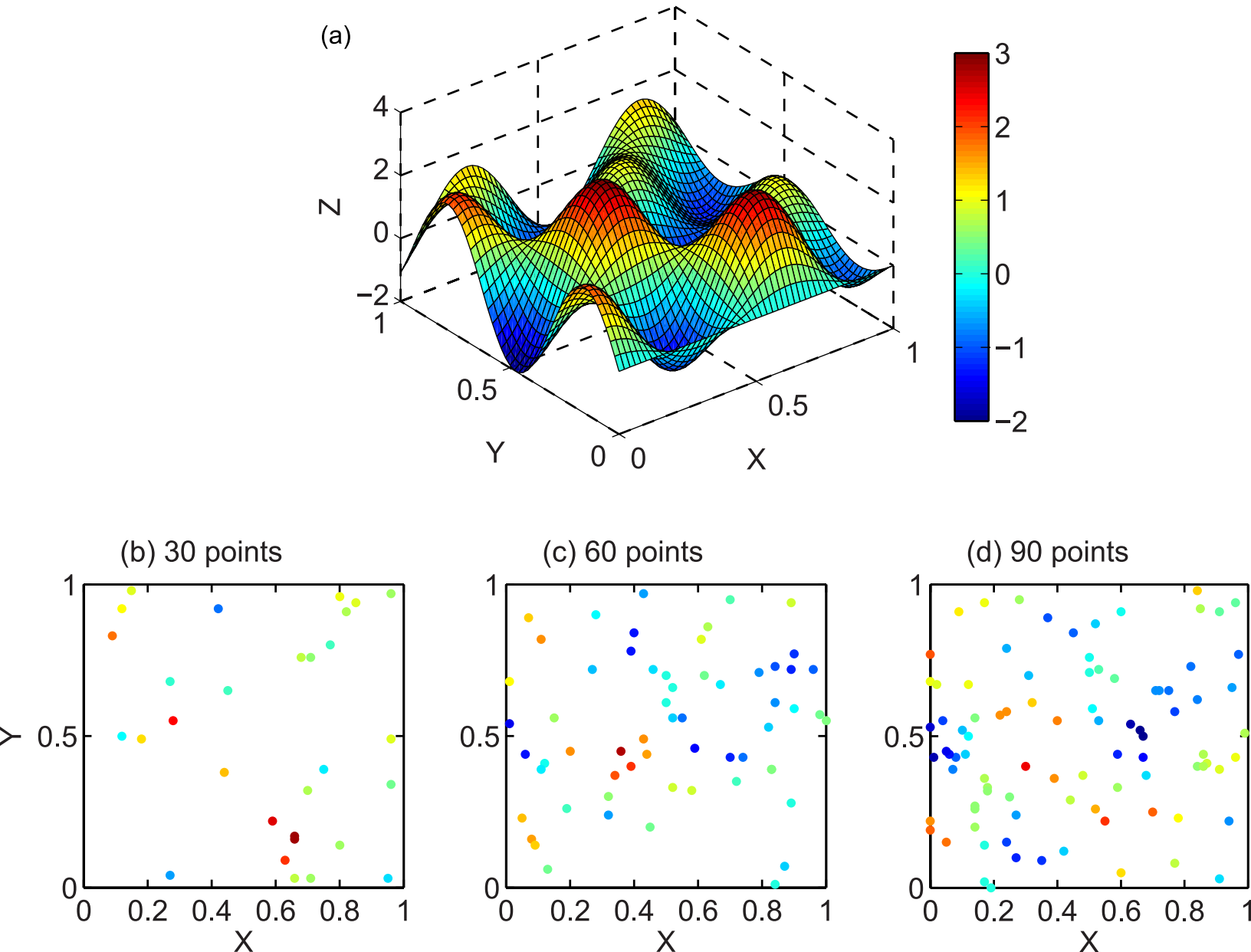

(a)
Z
4
2
0
−2
1
0.5
Y
0
0
0.5
1
X
3
2
1
0
−1
−2
(b) 30 points
(c) 60 points
(d) 90 points
Y
0
0.5
1
0
0.2
0.4
0.6
0.8
1
X

Figure 6

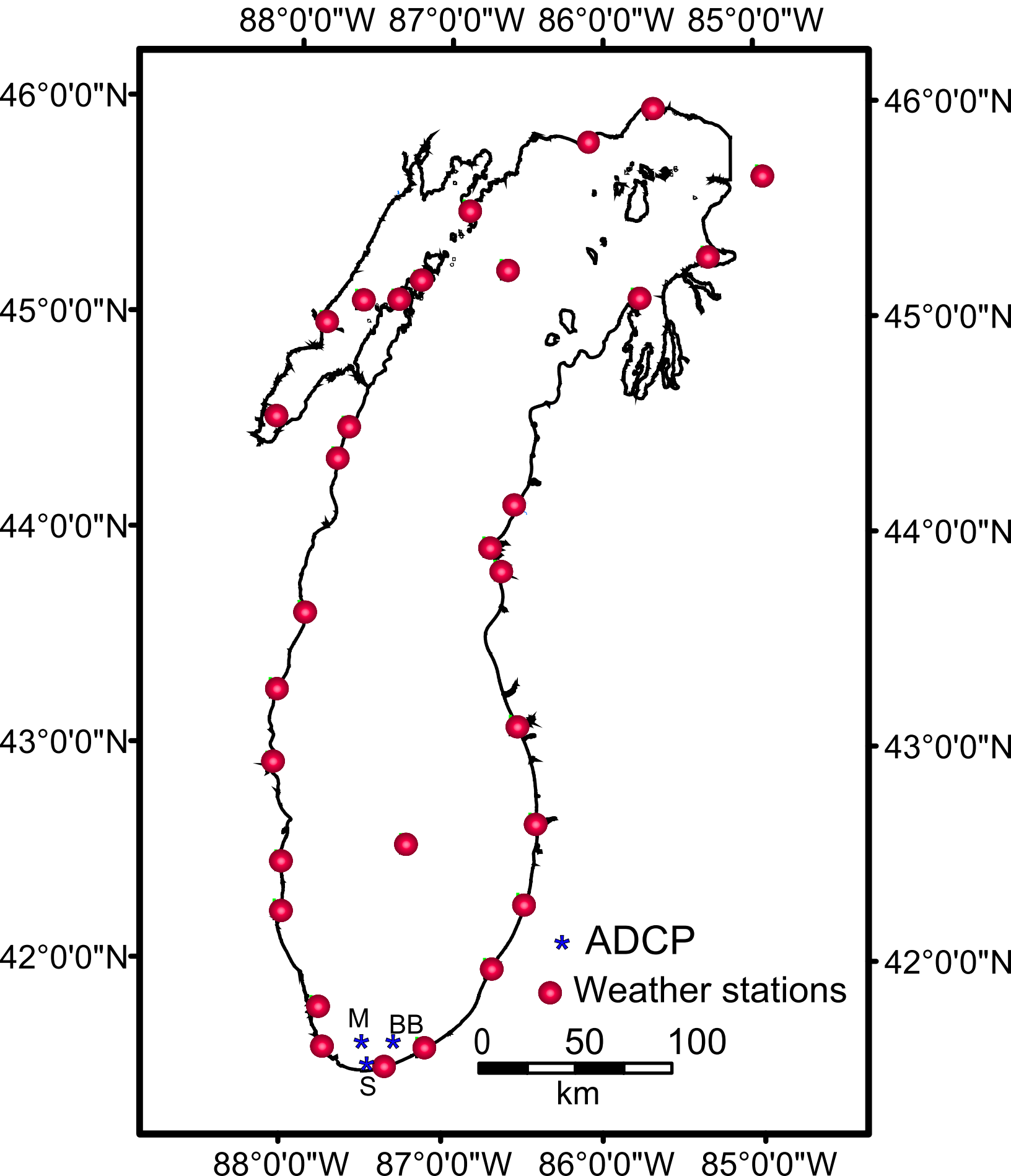
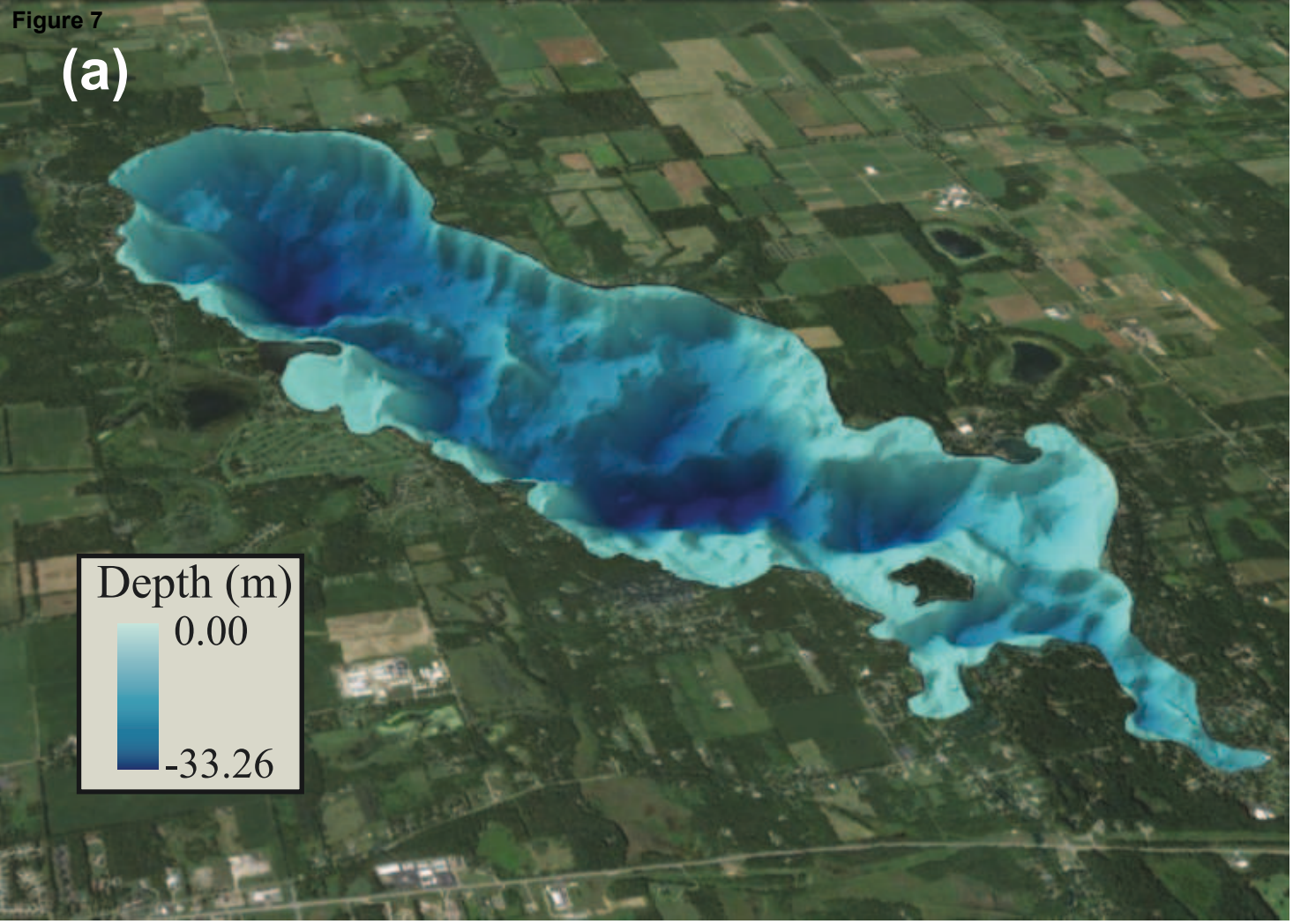

88°0'0"W
87°0'0"W
86°0'0"W
85°0'0"W
46°0'0"N
45°0'0"N
44°0'0"N
43°0'0"N
42°0'0"N
ADCP
Weather stations
M
BB
S
0
50
100
km

Figure 7


(a)
Depth (m)
0.00
-33.26

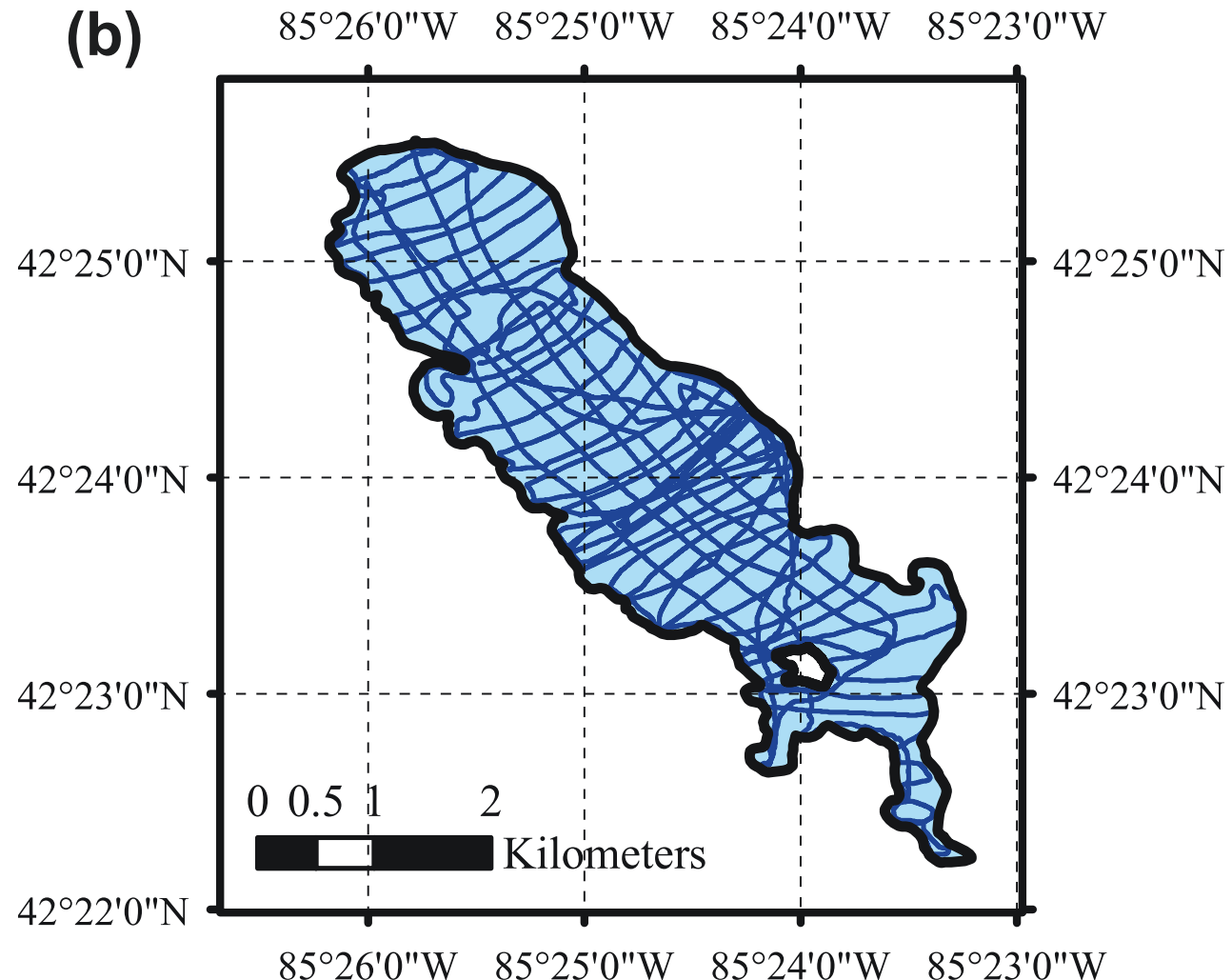

(b)
85°26'0"W
85°25'0"W
85°24'0"W
85°23'0"W
42°25'0"N
42°24'0"N
42°23'0"N
42°22'0"N
N
0 0.5 1 2
Kilometers

Figure 8

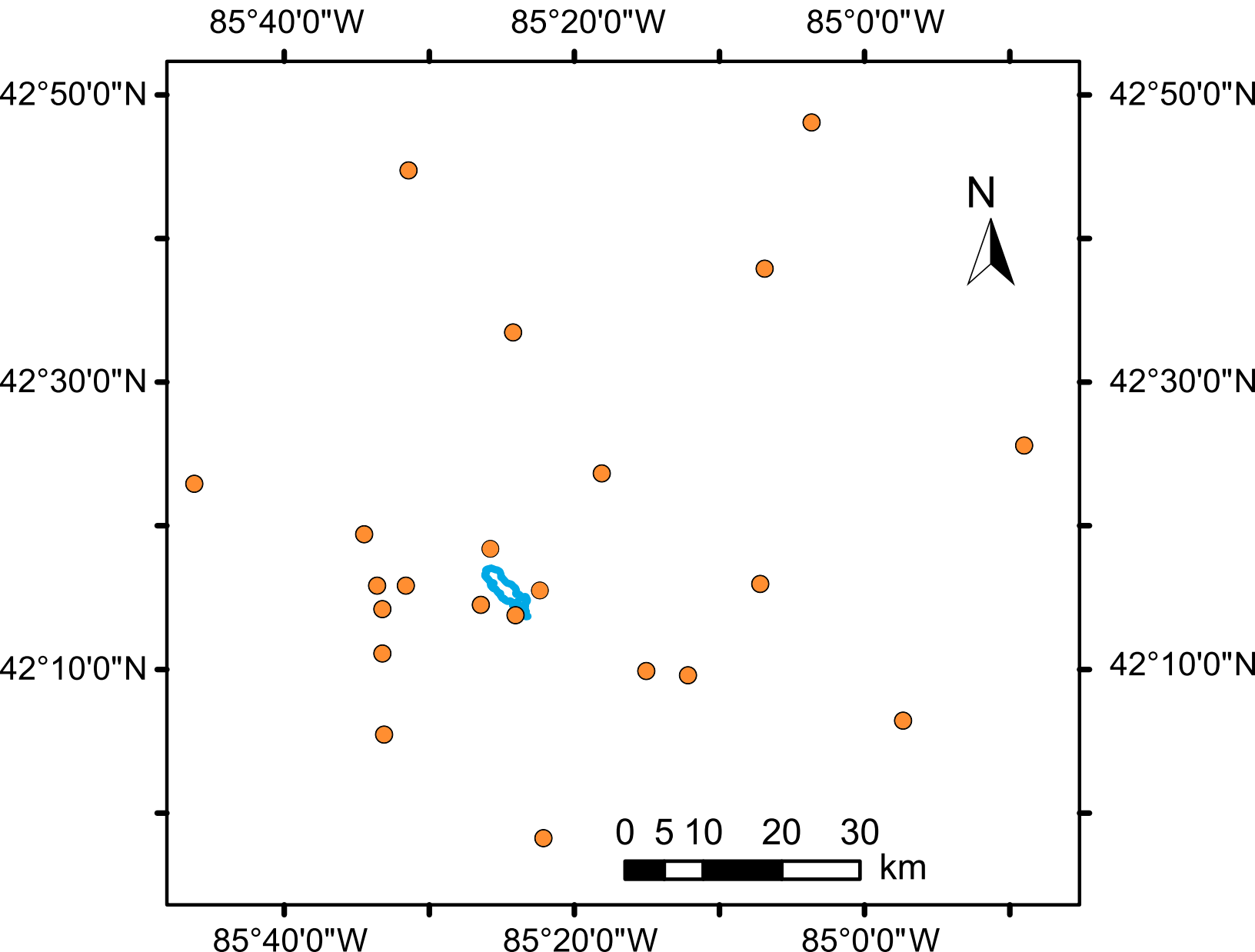


Figure 9

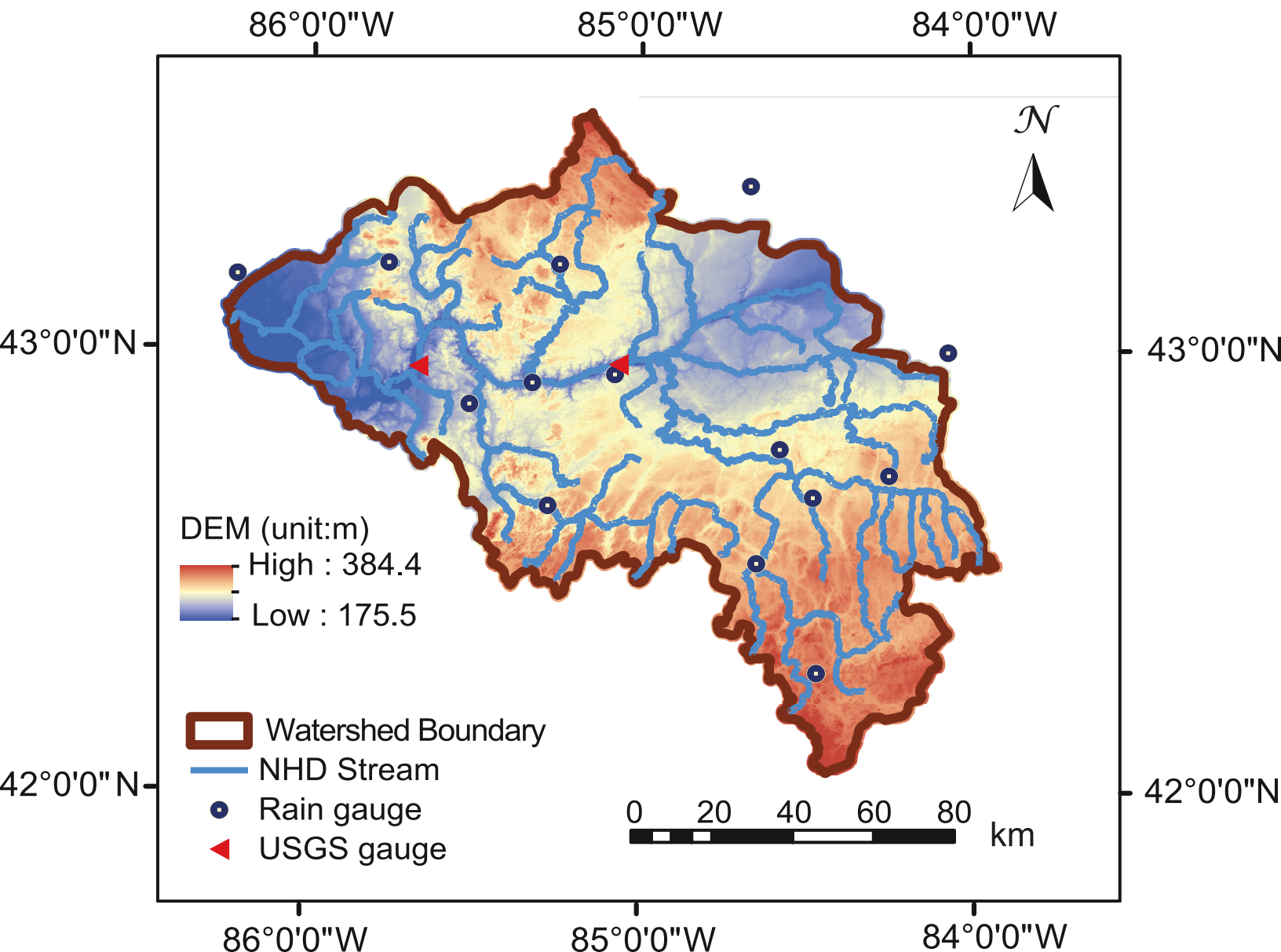

86°0'0"W
85°0'0"W
84°0'0"W
43°0'0"N
42°0'0"N
N
DEM (unit:m)
High : 384.4
Low : 175.5
Watershed Boundary
NHD Stream
Rain gauge
USGS gauge
0 20 40 60 80
km

Figure 10

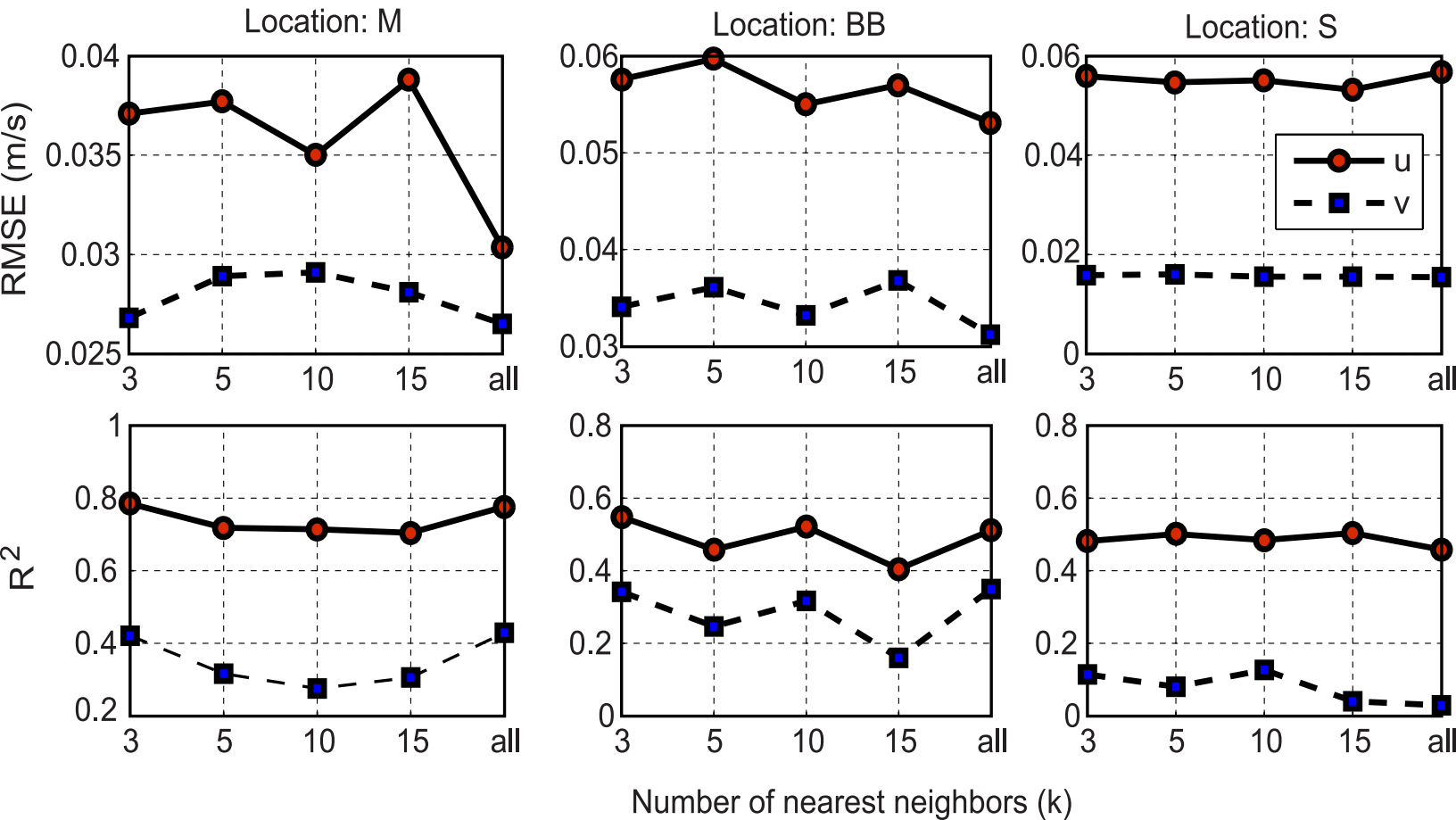

Location: M
Location: BB
Location: S
RMSE (m/s)
R2
u
v
Number of nearest neighbors (k)
3
5
10
15
all

Figure 11

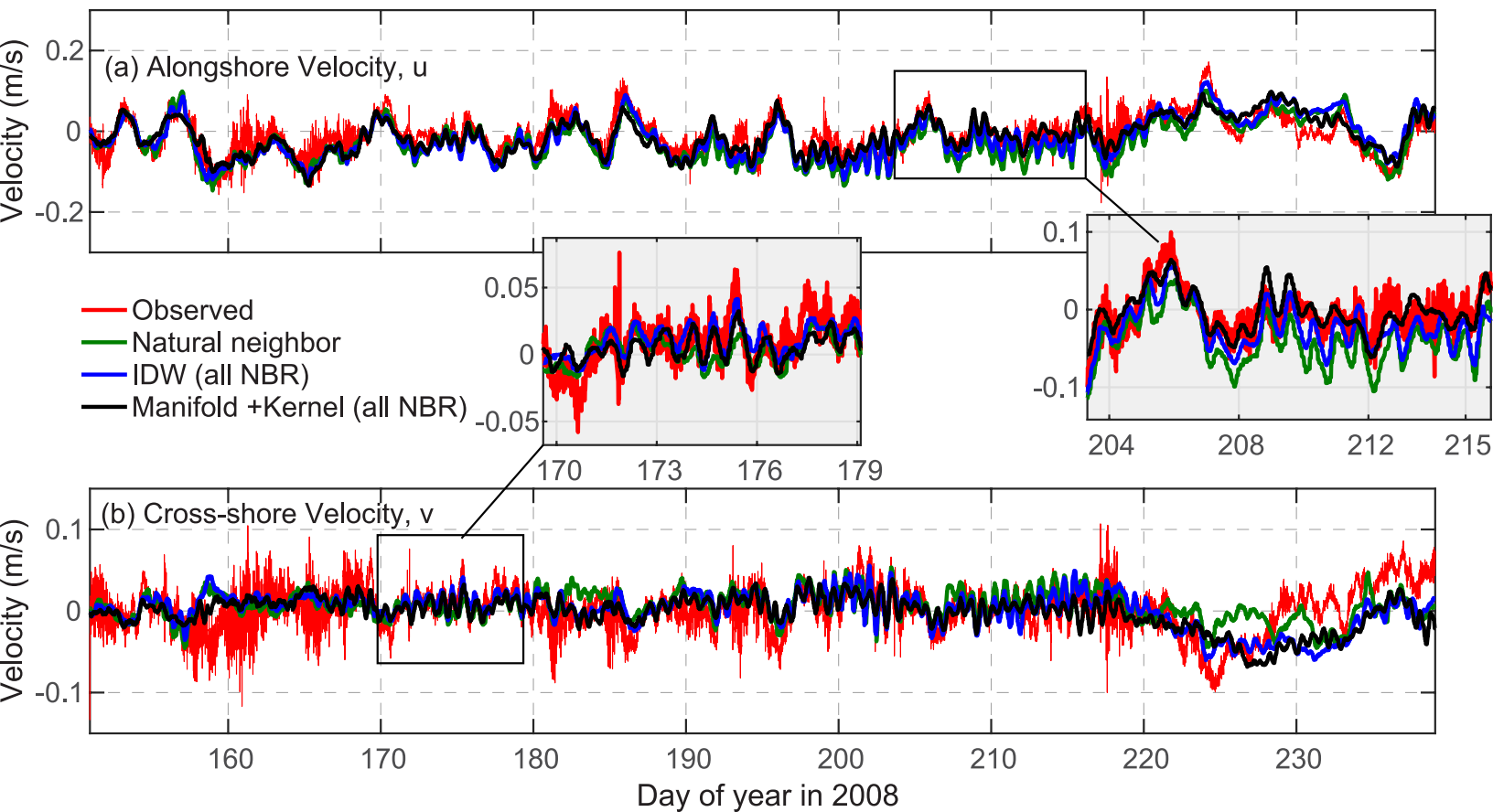

(a) Alongshore Velocity, u
(b) Cross-shore Velocity, v
Velocity (m/s)
Day of year in 2008
Observed
Natural neighbor
IDW (all NBR)
Manifold +Kernel (all NBR)

Figure 12

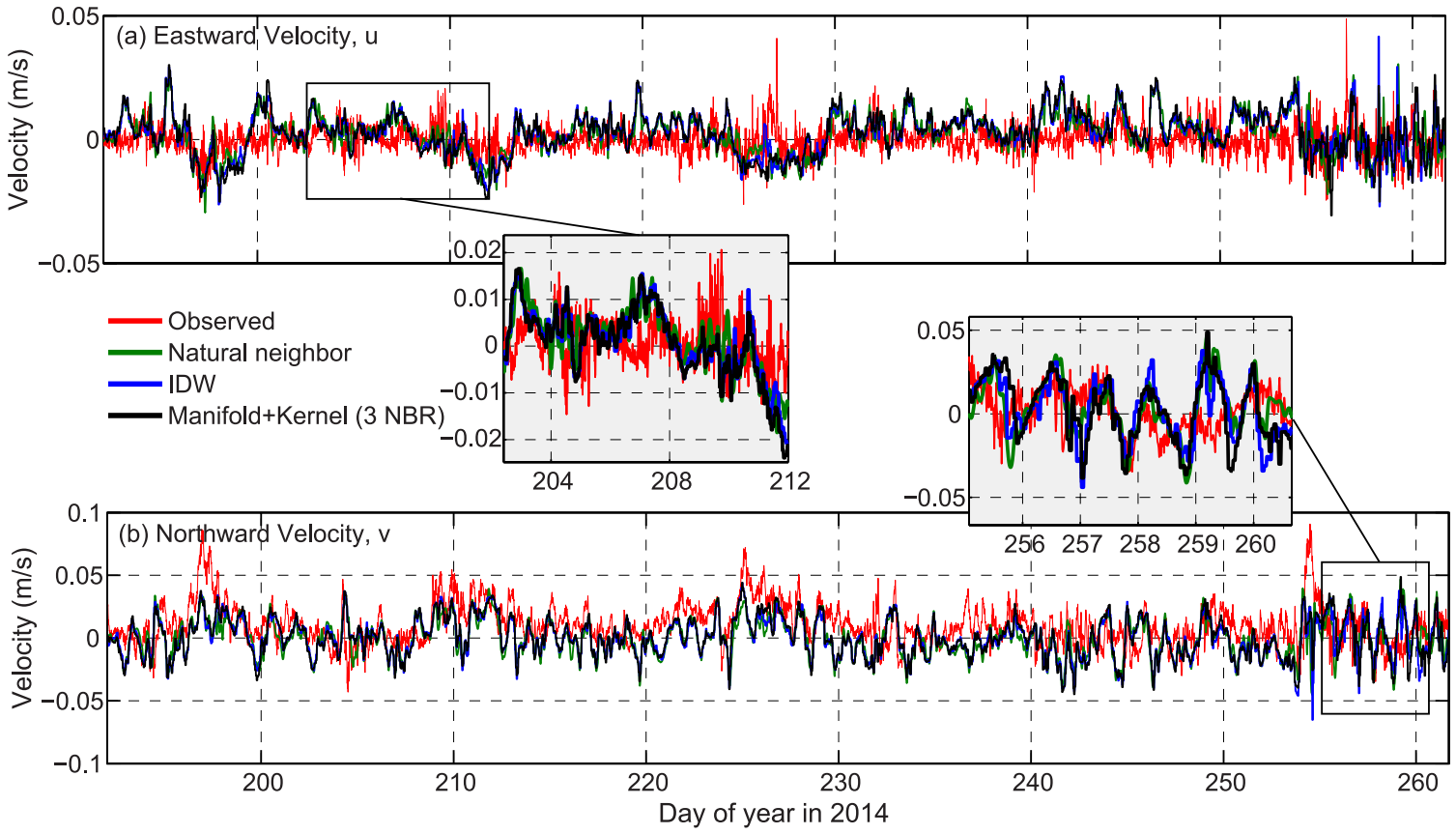

(a) Eastward Velocity, u
Velocity (m/s)
0.05
0
−0.05
Observed
Natural neighbor
IDW
Manifold+Kernel (3 NBR)
0.02
0.01
0
−0.01
−0.02
204
208
212
0.05
0
−0.05
256
257
258
259
260
(b) Northward Velocity, v
Velocity (m/s)
0.1
0.05
0
−0.05
−0.1
200
210
220
230
240
250
260
Day of year in 2014

Figure 13

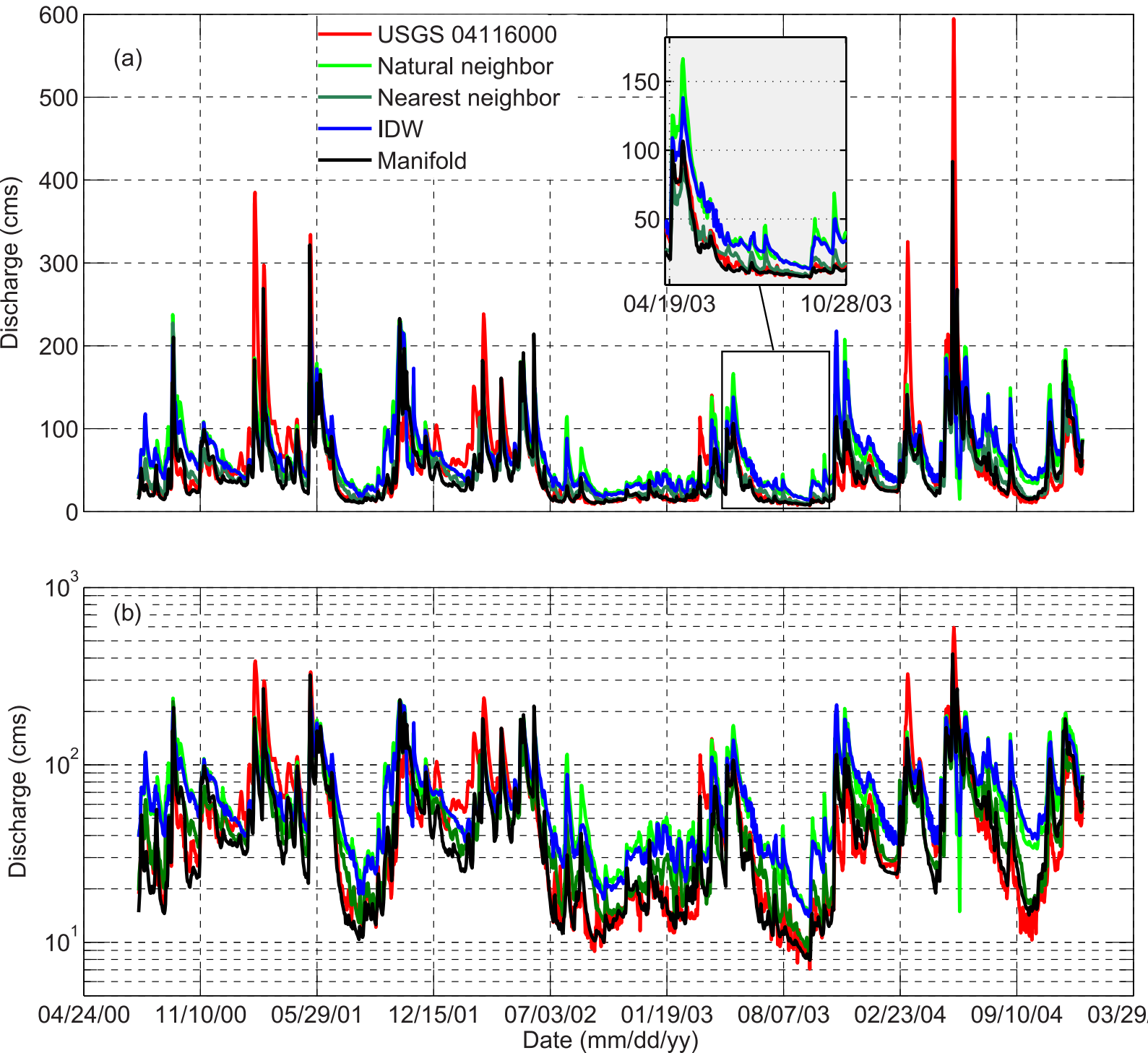
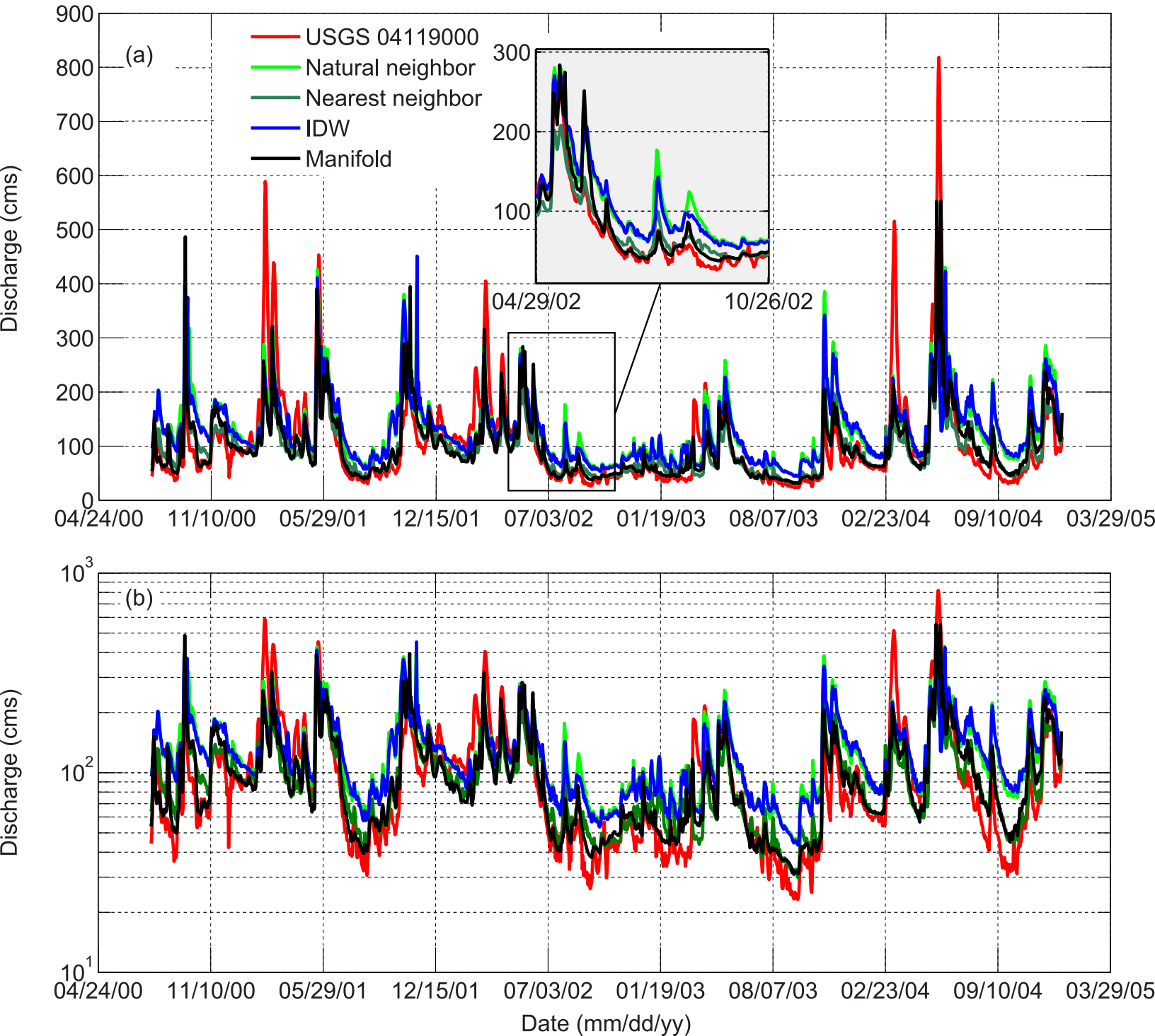

(a)
USGS 04116000
Natural neighbor
Nearest neighbor
IDW
Manifold
Discharge (cms)
600
500
400
300
200
100
0
150
100
50
04/19/03
10/28/03
(b)
10^3
10^2
10^1
Discharge (cms)
04/24/00
11/10/00
05/29/01
12/15/01
07/03/02
01/19/03
08/07/03
02/23/04
09/10/04
03/29/05
Date (mm/dd/yy)

Figure 14


(a)
USGS 04119000
Natural neighbor
Nearest neighbor
IDW
Manifold
Discharge (cms)
900
800
700
600
500
400
300
200
100
0
300
200
100
04/29/02
10/26/02
04/24/00
11/10/00
05/29/01
12/15/01
07/03/02
01/19/03
08/07/03
02/23/04
09/10/04
03/29/05
(b)
10^3
10^2
10^1
Discharge (cms)
Date (mm/dd/yy)