Sensor Placement and Measurement of Wind for Water Quality Studies in Urban Reservoirs

WAN DU, ZIKUN XING, MO LI, BINGSHENG HE, and LLOYD HOCK CHYE CHUA,
Nanyang Technological University, Singapore
HAIYAN MIAO, Institute of High Performance Computing, Singapore

We study the water quality in an urban district, where the surface wind distribution is an essential input but undergoes high spatial and temporal variations due to the impact of surrounding buildings. In this work, we develop an optimal sensor placement scheme to measure the wind distribution over a large urban reservoir using a limited number of wind sensors. Unlike existing solutions that assume Gaussian process of target phenomena, this study measures the wind that inherently exhibits strong non-Gaussian yearly distribution. By leveraging the local monsoon characteristics of wind, we segment a year into different monsoon seasons that follow a unique distribution respectively. We also use computational fluid dynamics to learn the spatial correlation of wind. The output of sensor placement is a set of the most informative locations to deploy the wind sensors, based on the readings of which we can accurately predict the wind over the entire reservoir in real time. Ten wind sensors are deployed. The in-field measurement results of more than 3 months suggest that the proposed sensor placement and spatial prediction scheme provides accurate wind measurement that outperforms the state-of-the-art Gaussian model based on interpolation-based approaches.

Categories and Subject Descriptors: J.2 [**Physical Sciences and Engineering**]: Earth and Atmospheric Sciences; G.3 [**Probability and Statistics**]: Experimental Design; C.4 [**Performance of Systems**]: Measurement Techniques

General Terms: Experimentation, Measurement

Additional Key Words and Phrases: Sensor placement, spatial prediction, wind measurements, water quality, urban reservoir

ACM Reference Format:
Wan Du, Zikun Xing, Mo Li, Bingsheng He, Lloyd Hock Chye Chua, and Haiyan Miao. 2015. Sensor placement and measurement of wind for water quality studies in urban reservoirs. ACM Trans. Sensor Netw. 11, 3, Article 41 (February 2015), 27 pages.
DOI: http://dx.doi.org/10.1145/2700265

This work is supported by the Singapore National Research Foundation under its Environment and Water Technologies Strategic Research Programme and administered by the Environment and Water Industry Programme Office (EWI) of the PUB on project 1002-IRIS-09. This work is also supported in part by Singapore MOE AcRF Tier 2 MOE2012-T2-1-070 and NTU Nanyang Assistant Professorship (NAP) grant M4080738.020.
Part of this work was published in ACM/IEEE IPSN 2014 [Du et al. 2014c] and IEEE SECON 2014 [Du et al. 2014b].
Authors' addresses: W. Du, M. Li, and B. He, School of Computer Engineering, Nanyang Technological University, 50 Nanyang avenue, 639798, Singapore; emails: {duwan, limo, BSHE}@ntu.edu.sg; Z. Xing, School of Civil and Environmental Engineering, Nanyang Technological University, 50 Nanyang avenue, 639798, Singapore; email: ZKXing@ntu.edu.sg; L. H. C. Chua (Current address), School of Engineering, Deakin University, 221 Burwood Highway Burwood Victoria 3125, Australia; email: lloyd.chua@deakin.edu.au; H. Miao, Institute of High Performance Computing (IHPC), 1 Fusionopolis Way, #16-16 Connexis North, 138632, Singapore; email: miaohy@ihpc.a-star.edu.sg.

DOI: http://dx.doi.org/10.1145/2700265

1. INTRODUCTION

A healthy aquatic ecosystem and water-quality monitoring is essential for good understanding of the water resources and the security of social systems, especially for countries with limited water resources like Singapore. Recent limnological studies [Alexander and Imberger 2009; Xing et al. 2014a] reveal that the distribution of wind stress on the surface of a lake can significantly impact water hydrodynamics and affects water quality. In a previous study on sensitivity analysis [Xing et al. 2014b], based on uniform wind distributions, we also find that in the Marina Reservoir, wind-force variability has a significant impact on vertical and spatial variability of phytoplankton distribution, which can substantially change water quality. Accurate wind distribution measurement is thus critical for studying and predicting the water quality in the Marina Reservoir. To investigate the impact on water quality evolution numerically, the wind distribution above the water surface as well as other environmental parameters (e.g., air temperature and precipitation) are used as input to a three-dimensional hydrodynamics-ecological model, Estuary Lake and Coastal Ocean Model—Computational Aquatic Ecosystem Dynamics Model (ELCOM-CAEDYM) [Laval et al. 2003]. Based on the calculation in the model, we can obtain the distribution of water quality in the whole reservoir. We can also study the effect of different environmental parameters on water-quality evolution, and predict water quality of the reservoir in the future with a time step of 30 seconds.

Most existing limnological studies are conducted in rural lakes and based on simple assumptions of surface wind including uniform [Alexander and Imberger 2009] or interpolated surface wind distribution [Laval et al. 2003]. In this work, we collaborate with environmental scientists to understand the effect of wind on the water quality of the Marina Reservoir in Singapore, a typical urban water field. It is located in downtown Singapore with a water surface of 2.2km^2, as depicted in Figure 1. Due to seasonal effects and the urban landform created by a variety of high-rise buildings surrounding two basins of the reservoir, the wind field has high temporal and spatial variations. In Figure 2, from the wind roses drawn by the historical data in 2007 of the meteorological stations, we see that the wind patterns in one year at these 3 main basins of the Marina Reservoir are totally different. Furthermore, we will show in Section 3 that the wind patterns at locations close to each other are different inside each basin because of the impact from the surrounding buildings.

In order to obtain an accurate wind distribution over the water surface of the Marina Reservoir, we deploy many wind sensors to measure the wind direction and speed. Based on the sensor readings at some discrete locations, we derive the wind distribution over the entire reservoir. However, constrained by the budget and government restrictions, we cannot deploy plenty of wind sensors all over the Marina Reservoir. To maximize the accuracy of field measurements, we need to find the most informative locations to deploy a limited number of wind sensors, based on the observations from which we can accurately predict the wind at other unobserved locations. Optimal sensor placement together with spatial prediction is therefore the key problem this article will address.

The problem of optimal sensor placement has been studied in many applications that monitor spatial phenomena, such as temperature sensing [Krause et al. 2006] and field soil moisture estimation [Wu et al. 2012]. Techniques such as spatial statistics [Cressie 1993] and subset selection [Das and Kempe 2008] have been proposed in previous works. As commonly assumed in those studies, the underlying phenomenon at one location can be modeled by a Gaussian distribution and the phenomena over the target area is thus a Gaussian Process (GP), where the marginal and conditional distributions of a multi-variant Gaussian distribution are still Gaussian. The optimal sensor

Fig. 1. Water surface and surrounding topography of the Marina Reservoir in Singapore.

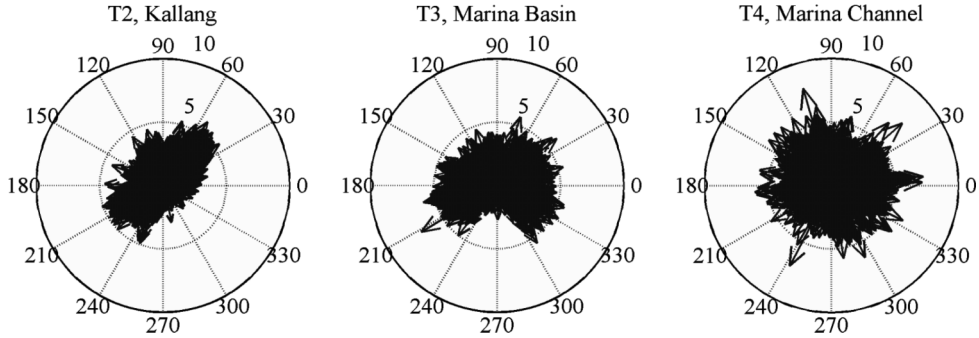


Fig. 2. Wind roses drawn by the historical data in 2007 from the meteorological stations at three different locations [Xing et al. 2014b].

placement is then calculated as the most informative locations by information theory criteria such as entropy [Ko et al. 1995] or mutual information [Guestrin et al. 2005; Krause et al. 2011]. Based on the sensor readings, spatial prediction is performed by estimating the posterior values of unobserved locations through Gaussian regression. In this article, we also refer to wind distribution as the wind field over the target area at a given point in time.

Existing GP-based approaches cannot be applied to wind measurement in this study mainly due to the following three challenges. First, as we will detail in Section 2, the wind directions in the field do not follow Gaussian process over time. Blindly applying GP-based approaches assuming Gaussian distribution of wind directions leads to suboptimal sensor placement and incurs large errors in spatial prediction. Second, existing approaches typically require sufficient prior knowledge on data distribution (usually collected from a denser predeployment) to train their GP model so as to capture pairwise correlations among different locations. Such prior knowledge is not available in our study. We do not possess historical wind distribution data of the field and it is cost prohibitive for us to predeploy numerous wind sensors to gain such knowledge. Third, in our study water quality in the reservoir has varied sensitivity to wind input at different locations due to diverse morphometrics and flow patterns, which calls for nonuniform measurement accuracy over the field. We need to optimize sensor

placement so that the sensors are deployed at locations with higher sensitivity to wind variations.

In this article, we propose a novel sensor placement and measurement approach to address these challenges. We propose a mixture model of wind as the sum of several Gaussian and uniform distributions. Inspired by the local monsoon characteristics of wind in Singapore, we do time series segmentation and divide one year into periods of different monsoon or intermonsoon seasons, during which the wind can be described or transformed to different Gaussian distributions and different prediction models can be trained. To derive the prediction model in each season, we obtain wind correlations among different locations in the field through Computational Fluid Dynamics (CFD) simulation instead of learning from predeployments. The optimal sensor placement is determined based on the information utility for all seasons and adjusted according to the sensitivity of water quality to wind in the field. When the sensor readings are collected in real time, we use an online clustering algorithm to flexibly determine the boundaries of these seasons with instant wind measurement in different years, performing proper spatial prediction accordingly. Finally, to further consider water quality prediction before deploying any wind sensors, we conduct a series of ELCOM-CAEDYM simulations for sensitivity analysis of water quality to the wind input and adjust accordingly the sensor placement scheme to factor the nonuniform accuracy requirement in wind measurement.

Ten wind sensors are finally deployed around or on the water surface of the Marina Reservoir according to the sensor placement scheme obtained from our analytical results. More than 3 months of in-field measurement results suggest that the proposed approach provides accurate spatial prediction of wind in both time and space. Compared with previous GP or interpolation-based approaches, our approach reduces average root-mean-squared error of measurement in wind direction by 81% and 33%, respectively.

The rest of this article is organized as follows. Section 2 gives the problem statement and presents the overview of the proposed approach. Section 3 presents the detailed design and analysis of the approach. Section 4 describes the in-field deployment experience and presents the experimental evaluation results. Section 5 summarizes the lessons we learned from this work and Section 6 discusses the applications and limitations of the proposed approach. Section 7 introduces related works. Section 8 presents our conclusions.

2. PROBLEM STATEMENT AND OVERVIEW

In this section, we formally formulate the sensor placement and spatial prediction problem. We present the unique challenges from our application and an overview of our approach.

2.1. Problem Statement

In this wind measurement application, we divide the Marina Reservoir into small grids of 20m*20m. We assume that each grid is a location with uniform wind field. It is also the smallest grid size required by the water quality study in the ELCOM-CAEDYM model. Even smaller grid size will not help the water quality study much. Totally, we need to cover more than 5k locations. The set of all locations over the Marina Reservoir is denoted as $\mathcal{V}$, where $|\mathcal{V}| = N$. The observations at each location $v_i \in \mathcal{V}$ can be modeled as a random variable X_i. All variables jointly form a random process. The objective of optimal sensor placement is to select a subset $\mathcal{A}$, $\mathcal{A} \subset \mathcal{V}$ and $|\mathcal{A}| = K << N$, from which we can predict the observations of the other locations, presented as $\mathcal{V} \backslash \mathcal{A}$, with minimal estimation errors.

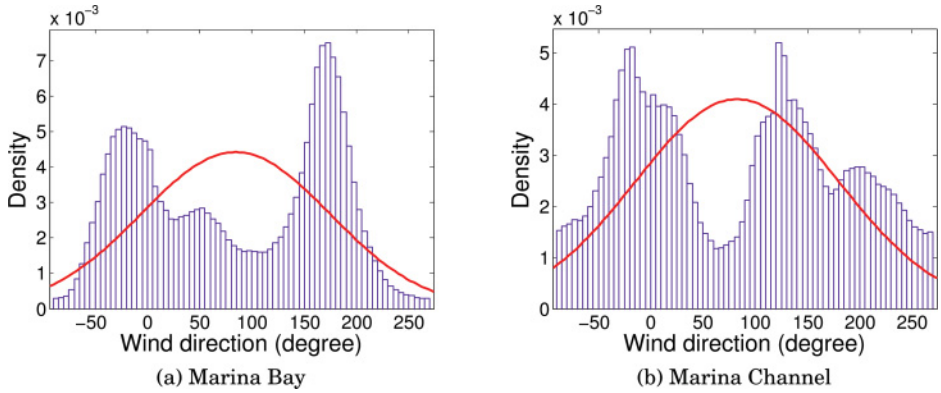


Fig. 3. Wind direction histograms for the year 2007. The red line is the Gaussian fitting curve of the wind direction density. Wind from the north is represented as 0 degrees.

Common approaches that have been applied to similar spatial prediction problems assume that the random variable X_i at each location follows a Gaussian distribution and the joint distribution of the variables over all locations can be modeled as a Gaussian process [Osborne et al. 2008; Krause et al. 2006]. With such GP assumptions, existing approaches benefit from the feature that the marginal and conditional distributions of a multi-variant Gaussian distribution are still Gaussian. Therefore, the most important sensor locations can be selected by some informative criteria such as entropy [Ko et al. 1995] or mutual information [Guestrin et al. 2005]. The observations on the other unobserved locations can then be predicted as the mean of conditional distribution $X_{\mathcal{V}\setminus\mathcal{A}}|X_{\mathcal{A}}$ with an uncertainty $\sigma^2_{v|\mathcal{A}}$:

$$\mu_{v|\mathcal{A}} = \mu_v + \sum_{v\mathcal{A}} \sum_{\mathcal{A}\mathcal{A}}^{-1} (x_{\mathcal{A}} - \mu_{\mathcal{A}}) \tag{1}$$

$$\sigma^2_{v|\mathcal{A}} = \sum\nolimits_{v,v} - \sum\nolimits_{v\mathcal{A}} \sum_{\mathcal{A}\mathcal{A}}^{-1} \sum\nolimits_{v\mathcal{A}}^{T}, \tag{2}$$

where $\sum_{v\mathcal{A}}$ is a vector of covariance between v and each element in $\mathcal{A}$, and $\sum_{\mathcal{A}\mathcal{A}}$ is the covariance matrix of $\mathcal{A}$.

This GP-based approach has been successfully used in many applications, such as temperature monitoring [Krause et al. 2006] and data collection tour planning [Meliou et al. 2007]. However, it cannot be directly applied in wind field measurements since the unique application features cannot support some of its prerequisites and assumptions. As depicted in Figure 3, the actual distribution of wind directions over one year is far from Gaussian. The data are collected by the meteorological stations in Marina Bay and Marina Channel over the year 2007. The inaccurate Gaussian fitting leads to large errors in understanding correlations within the wind field. As a result, it jeopardizes the results of sensor placement and spatial prediction. As will be shown in Section 4, the average prediction error of wind direction will reach as high as 89° if we blindly apply such a biased and mismodeled fitting.

In addition, to train the GP model, existing approaches require full prior knowledge of data distribution over the entire field such that the pairwise correlations of all locations in the field can be captured. For instance, in Guestrin et al. [2005], the training data for the GP model are collected with 54 temperature sensors predeployed for 5 days with a sampling interval of 0.5min. Another example is community sensing by Krause et al.

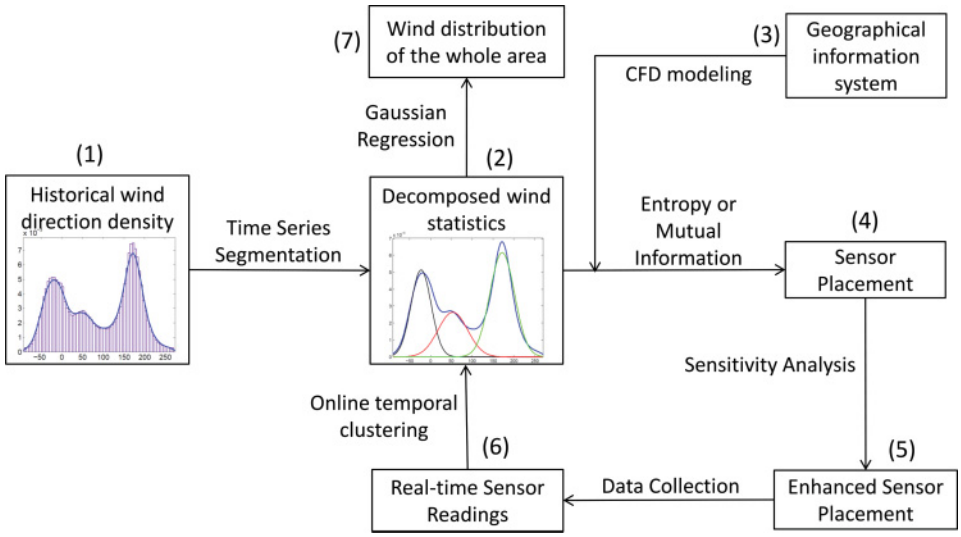


Fig. 4. Framework of the proposed approach. The design procedure is indicated by the sequence numbers in parentheses.

[2008], who provide the best route prediction based on a GP model trained by 2110 route planning requests obtained from volunteers during 2006 and 2007. Such full prior knowledge about the wind over the Marina Reservoir, however, is not available. Intrusively gaining such knowledge through predeploying sensors is also not possible. First, it is cost prohibitive to deploy an adequate number of sensors (more than 5k) to precisely cover the underlying field. In our study, the cost of a land sensor is about 6,000 USD and a floating sensor on the water surface costs about 8,000 USD. We can only plan the most informative sensor placement beforehand and then deploy a limited number of sensors (10 in this study). Second, due to topography and regulatory requirements in such an iconic center of the city, we are not able to deploy sensors at all desired places for full data survey. It took us several months to get permits from Singapore government agencies for deploying the wind sensors in the allowed areas, as shown in Figure 12, Section 3).

Finally, the wind measurements are often not the final objective but used to infer some consequential phenomena such as energy distribution [Burton et al. 2011] and water circulation in a lake [Laval et al. 2003]. We need to consider the water quality modeling while designing the optimal sensor placement scheme since the winds of different locations impose variant impact on water quality of a whole lake due to diverse morphometries and flow patterns. Although much effort is made to reduce the spatial prediction errors to the smallest range possible, the wind distributions obtained by estimating the observations through the readings of limited deployed sensors will inevitably contain some errors. Therefore, we intend to provide direct measurements by deploying sensors at locations with high impact on the final water quality studies and eliminate the prediction inaccuracy for the other unobserved locations maximally.

2.2. Approach Overview

We propose a novel approach to address the sensor placement and spatial prediction problem by considering the unique features of wind measurement applications. Figure 4 illustrates the main framework in steps. First, we possess the historical wind data from the two meteorological stations at Marina Bay (2007–2008) and Marina Channel

(2007–2008 and 2011–2013). We find a clear difference between the dominant wind directions in different time periods, which is consistent with the monsoon climate in Singapore.[1]

We develop a time series segmentation method and divide the sensor data of the whole year into two monsoon seasons and two intermonsoon seasons. In each segment, the wind at one location can thus be modeled or transformed to a Gaussian distribution. The optimal sensor locations are selected according to certain information criteria, for example, entropy in this work. The results of all seasons are combined to calculate the optimal sensor placement scheme in a whole year.

We incorporate CFD modeling to simulate the wind distributions above the Marina Reservoir based on 3D geographical information. CFD modeling can capture the detailed impact of surrounding high-rise buildings to the wind distribution by numerically solving the classic formulas of fluid mechanics [Anderson et al. 1995]. In this study, we perform offline CFD simulations to generate coarse wind distributions at different conditions and learn the correlations in the field rather than obtaining the final wind distribution in real time since CFD modeling is computationally complex and time-consuming.

To further consider the water quality sensitivity before deploying any wind sensors, we conduct a series of ELCOM-CAEDYM simulations to quantify the sensitivity of water quality to the wind input at different locations over the Marina Reservoir. We then adjust accordingly the sensor placement scheme to factor the nonuniform accuracy requirement in wind measurement.

Once we have obtained the optimal sensor locations, we deploy a certain number of wind sensors at the most critical locations. Based on a wireless data collection system [Du et al. 2014a], the technical details of which are beyond the scope of this article, we retrieve the real-time sensor readings from our server. Finally, we use an online clustering algorithm to dynamically identify the transitional point between different monsoon and intermonsoon seasons with instant wind measurements. Different spatial prediction parameters are applied in the identified seasons with real sensor readings.

3. WIND MEASUREMENT APPROACH

In this section, we present the detailed development procedure of our approach for wind measurements, including monsoon-based time series segmentation, dataset generation based on CFD modeling, optimized sensor placement, and spatial prediction.

3.1. Monsoon-Based Time Series Segmentation

Figure 5(e) presents the histograms of wind direction and speed for the entire year of 2007 drawn by the data of the meteorological station at Marina Channel. From the historical data, we see that two obvious peaks in the density of wind directions correspond to the two monsoon seasons in one year in Singapore, which are caused by the seasonal changes in global atmospheric circulation upon asymmetric heating of land and sea [Trenberth et al. 2000]. In each monsoon season, the wind is mainly from a dominant direction. It has been found from historical wind data of multiple years [Chia et al. 1991] that the wind directions of the two monsoon seasons are strongly Gaussian and the wind during the intermonsoon seasons is weak and more evenly distributed over all directions. The distribution of the whole year is the sum of all segments, exhibiting a mixture model. In this section, we introduce our monsoon-based

[1]Singapore has two monsoon seasons every year, Northeast (NE, roughly December–March) and Southwest (SW, roughly June-September). The names indicate the seasons' dominant wind direction. The monsoon seasons are separated by two intermonsoon periods, PreSW and PreNE, in which the wind is more evenly distributed.

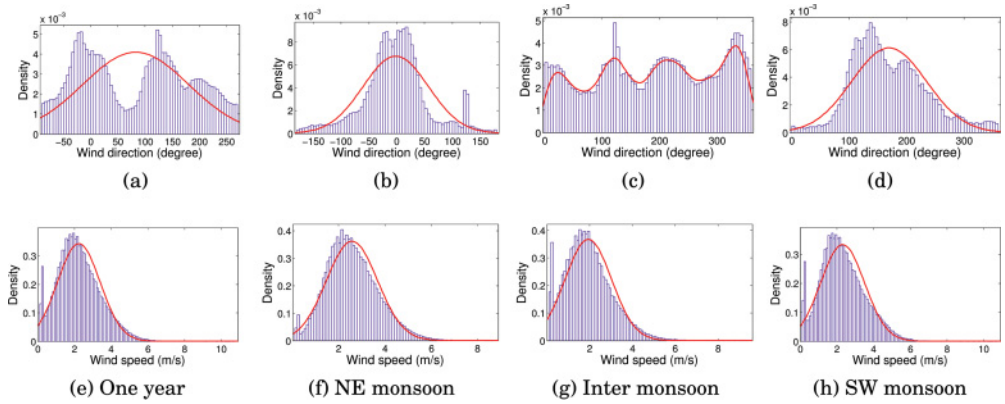


Fig. 5. Density of wind direction and speed over the year 2007 and its decomposed monsoon seasons. The two rows are the distributions of wind direction and wind speed, respectively. In each row, the first subfigure is the wind distribution of one year, and the following three subfigures present the wind distributions of the NE monsoon season, intermonsoon season, and SW monsoon season. In each subfigure, the red line is normal or the uniform fitting curve for relative distributions.

time series segmentation such that a whole year is segmented into different monsoon or intermonsoon seasons that follow different GP models.

3.1.1. Time Series Segmentation Algorithm. The traditional monsoon division scheme based on experience only provides month-level granularity. The start and end of a monsoon season may largely vary in different years. We thus need an accurate segmentation scheme to find the critical changing time points for monsoon season transitions.

The objective is to find four critical change points to make the wind directions in the monsoon seasons follow a Gaussian distribution as closely as possible and the wind directions in the intermonsoon seasons follow a uniform distribution as closely as possible. We use the Maximum Likelihood (ML) method to find the optimal time points that separate the one-year data from a meteorological station into four segments including M, N, K, and J samples, respectively, which maximize the likelihood function of the mixture model (two Gaussian and two uniform).

$$\begin{aligned}\mathcal{L}(\mu_1, \sigma_1, \theta_1, \mu_2, \sigma_2, \theta_2 | x_1, x_2, \ldots, x_{M+N+K+J}) \\ = \prod_{i=1}^{M} \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{1}{2\sigma_1^2}(x_i - \mu_1)^2\right] * \left[\frac{1}{\theta_1}\right]^N \\ * \prod_{i=1+M+N}^{K+M+N} \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left[-\frac{1}{2\sigma_2^2}(x_i - \mu_2)^2\right] * \left[\frac{1}{\theta_2}\right]^J,\end{aligned} \tag{3}$$

where $\mu_1 = (1/M)\sum_{i=1}^{M} x_i$ and $\sigma_1 = \big(1/(M-1)\big)\sum_{i=1}^{M}(x_i - \mu_1)^2$ are the unbiased estimation of parameters in the first Gaussian distribution including M samples, and $(1 + 1/N)max(x_{M<i\leq N+M})$ and $(1 + 1/J)max(x_{M+N+K<i\leq N+M+K+J})$ are the unbiased estimation of parameters (θ_1 and θ_2) in the two uniform distributions.

The computation complexity to solve Equation (3) is $\mathcal{O}(n^3)$, where n is the search space for each time point. Since we know the approximate start and end of each monsoon season, we can restrict the search space. Algorithm 1 presents the ML-based time series segmentation algorithm searching the optimal time points heuristically. We can obtain the same results with the method searching in the whole dataset exhaustively, but with much less computation. If we search in a 2-month span centered at the

ALGORITHM 1: Heuristic ML-Based Time Series Segmentation

Input: One year wind data.
Output: Time points, t_1, t_2, t_3 and t_4.
Initialization: $t_{1,old} = Mar.15$; $t_{2,old} = t_2 = Jun.1$; $t_{3,old} = Oct.1$; $t_{4,old} = t_4 = Dec.1$;
Step 0: concatenate the start and end of data, so that the last NE part is merged to the first NE part;
Step 1: In (t_4, t_2), search t_1 in a Gaussian/Uniform mixture model by an equation similar to Equation (3);
Step 2: In (t_2, t_4), search t_3 as **Step 1**;
Step 3: Based on the updated t_1 and t_3, search t_2 in (t_1, t_3) and search t_4 in (t_3, t_1);
if $(t_1 \neq t_{1,old} || t_2 \neq t_{2,old} || t_3 \neq t_{3,old} || t_4 \neq t_{4,old})$ **then**
    $t_{1,old} = t_1; t_{2,old} = t_2; t_{3,old} = t_3; t_{4,old} = t_4$;
    go to **Step 1**;
else
    return t_1, t_2, t_3 and t_4;
end if

experience-based time point that starts with the first day of the relative transitional month (e.g., April 1st for the transition from NE monsoon season to PreSW monsoon season), it takes less than 1 hour to converge.

Figure 5 presents the decomposed monsoon seasons for the year 2007. Two inter-monsoon seasons are combined since they present the same pattern. We see that the wind direction in each individual season is well fitted by a Gaussian or uniform model. Figure 5 also shows that the wind speed of each season can be perfectly modeled as a Gaussian distribution, because the wind speed of the whole year is also Gaussian distributed. Therefore, we mainly focus on the segmentation of wind direction. The wind speed will automatically follow a Gaussian distribution processed according to the results of wind direction.

3.1.2. Segmentation Result Analysis. Likelihood ratio, $D = 2(\ln \mathcal{L}_{new} - \ln \mathcal{L}_{null})$, is normally used to compare the fitness of two models. It expresses how many times more likely the data are under one model than the other. The likelihood ratio D between the mixture model derived by the proposed segmentation algorithm ($\mathcal{L}_{new}$) and the uni-Gaussian model ($\mathcal{L}_{null}$) used by traditional sensor placement approach is 16.8k. The winds divided by the proposed segmentation algorithm are modeled obviously better than the uni-Gaussian model. We also divide the winds by the fixed division scheme based on experience. The likelihood ratio between this experience-based mixture model and the uni-Gaussian model is 13.8k, which also shows that the division scheme derived by our segmentation algorithm can better fit the winds into proper statistical models. We will show in Section 4 that the mixture model derived from the proposed time series segmentation algorithm provides much more accurate sensor placement and spatial prediction than the uni-Gaussian model.

3.1.3. Application to Wind Measurements. To apply the segmentation results generated with the historical data of one meteorological station over one year, we need to answer two questions. Do the other locations in the target area hold the same segmentation scheme? Can the segmentation scheme generated by one year's worth of historical data be applied to other years or even to the current year?

All locations in the Marina Reservoir area share the same monsoon division scheme. The segmentation derived from the historical data of the Marina Channel meteorological station can be applied to other locations, since it is based on a general environmental phenomenon that is consistent across the region. The monsoon climates are caused by the seasonal changes in global atmospheric circulation due to the asymmetric heating

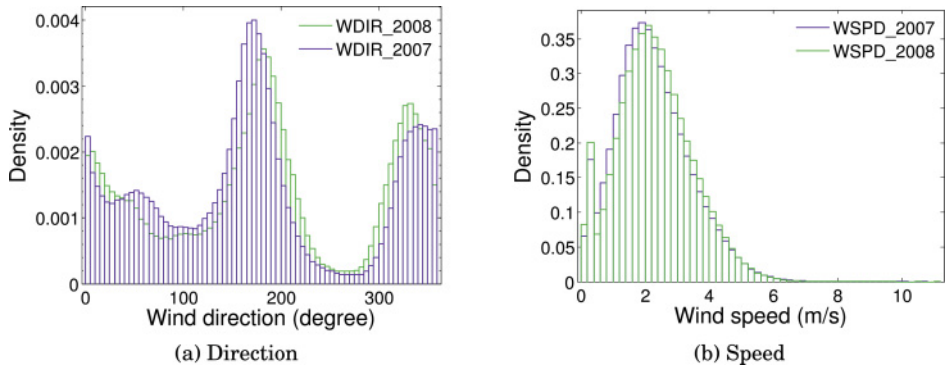


Fig. 6. Wind density for two consecutive years (2007 and 2008) collected at the Marina Bay meteorological station.

of land and sea [Trenberth et al. 2000]. Compared with the large-scale atmospheric circulation, the Marina Reservoir is small in size; therefore, all locations are dominated by the same monsoon pattern. For example, we study the historical data (2007) from both observatory sites at Marina Bay and Marina Channel and find exactly the same segmentation results (December 1 to March 15 for NE monsoon season and June 9 to September 15 for SW).

It has also been proven based on historical wind data [Chia et al. 1991] that the main directions of winds in monsoon seasons are stable for different years and the winds in intermonsoon seasons are evenly distributed. Figure 6 presents the wind statistics from the Marina Bay meteorological station for two consecutive years (2007–2008) which suggest highly similar distributions. As we find later, the derived parameters of GP models based on the data from different years are very close to each other for the same seasons. The most likely variation from year to year is the variance of Gaussian distribution for monsoon seasons. Such small fluctuation is easily flattened by looking at multi-year wind data. For the wind data of multiple years (e.g., 2007–2008 data from the Marina Channel meteorological station), the division is performed for each year respectively and the relative parts of different years are combined to derive the best Gaussian fitting. Beside wind direction, according to Figure 6, the wind speeds are stable for different years and always follow Gaussian.

3.2. Correlation Learning

Once the wind direction in every segment divided by the proposed segmentation algorithm follows a Gaussian distribution for each individual location, the multivariate Gaussian distribution formed by all locations can be referred to as a Gaussian process. To learn the pairwise correlation between any two locations in the target area, we need to obtain the parameters of the GP model. We apply CFD modeling to obtain simulated surface wind distributions above the Marina Reservoir for different wind directions above the atmospheric boundary layer. We build a dataset composed of many wind distributions and calculate the parameters of the GP model.

3.2.1. CFD Modeling. CFD studies the physical aspects of fluid flows by algebraically solving the fundamental governing equations like continuity and momentum conservation. Numerical results are finally obtained at discrete points in time and space. The CFD modeling of wind distribution needs two inputs: atmospheric flow and topography information of the land surface. On one hand, since the Marina Reservoir is relatively small in size compared with the large-scale atmospheric circulation, the atmospheric motion above this area can be treated as uniform. The atmospheric flow is therefore

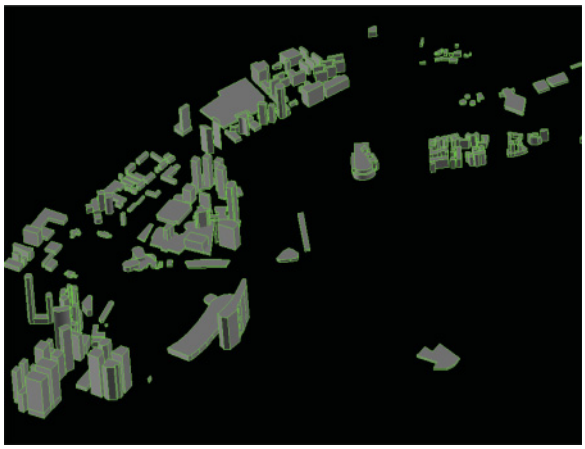


Fig. 7. 3D CFD model of geographical information around the Marina Reservoir.

a vector comprised of a dominant wind direction and speed. On the other hand, a three-dimensional CFD model of the Marina Reservoir area is developed based on the geographical information that contains building locations, building shapes as well as heights. Figure 7 depicts the built 3D CFD model of topography around the Marina Reservoir, which models all buildings located within 3 blocks of the reservoir offshore. The computational domain is 3.5km long, 2.5km wide and 0.8km high. The number of computational cells used for each simulation is approximately 40 million.

The commercial CFD software FLUENT 13.0 is used to calculate the surface wind distribution over the Marina Reservoir area. To capture the turbulent nature of the flow around buildings, the popular k-ϵ turbulence model is chosen because of its high computational efficiency [Anderson et al. 1995]. Standard and second-order discretization schemes are adapted for pressure interpolation. It takes almost 2 days for one simulation case to be converged using a workstation of 12 cores (running 8 parallel-Fluent licenses) and 32GB memory.

3.2.2. Dataset Generation. The CFD modeling results cannot provide accurate instant wind distribution due to the following two limitations. First, CFD requires real-time and accurate atmospheric circulation data as input to derive instant wind distribution, which is difficult to obtain. Second, CFD simulation is computationally complex and time-consuming, which makes the instant CFD computation impossible.

To capture the main characteristics of all possible wind distributions over the water surface, we run many simulations with different atmospheric flow inputs. A 16-point compass rose is used to categorize the incoming atmospheric flows into 16 directions evenly spanning 0° to 360°. For each direction, we run 10 gradually increasing speeds to explore all possible atmospheric motion velocities (0$\sim$9m/s) in Singapore. By doing this, we obtain a dataset of 160 independent surface wind distributions for the underlying area. Two examples are given in Figure 8, with an incoming atmospheric flow from north and south, respectively. We can see that the surface wind distributions have distinctive patterns for different incoming flows due to the influence of surrounding architectures.

For all wind distribution results of CFD simulations, the wind direction and speed at the location of the Marina Channel meteorological station is one-to-one mapped to the incoming atmospheric flow, because Marina Channel is in a relatively free space. In the dataset, we extract a wind vector at the location of the Marina Channel meteorological

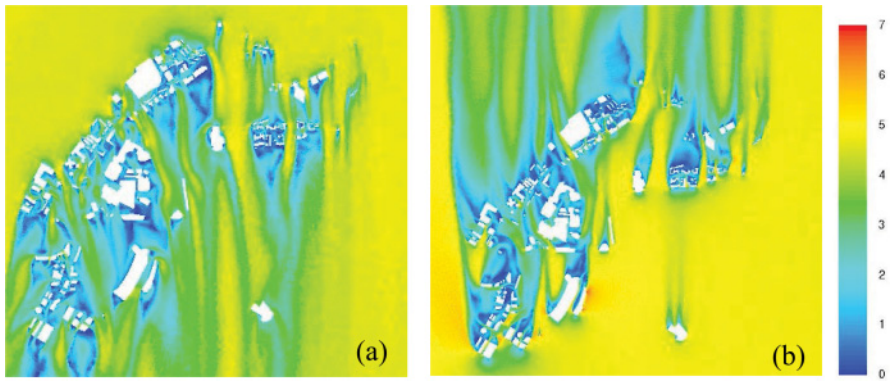


Fig. 8. CFD modeling results of wind velocity (m/s) distribution at a height of 1.5 m with an incoming atmospheric flow from north (a) and south (b).

station from each of the 160 CFD wind distributions. At the same time, these vectors divide the historical wind data of the Marina Channel meteorological station for the year 2007 into 160 segments. The occurrence frequency of each wind distribution in the whole year or in each monsoon or intermonsoon season can thus be computed. Based on this information and all 160 CFD wind distributions, we can generate a time series of wind distributions above the entire Marina Reservoir over many years as long as we possess the relative historical data of the meteorological station at Marina Channel.

3.2.3. Gaussian Process Model. Once we divided one year into different monsoon seasons and built the dataset as prior knowledge on the wind distributions in each segment, we could learn the spatial correlation between any two locations over the target area. The key parameters of the GP models, that is, mean vector and covariance matrix $\sum_{\mathcal{VV}}$, used in the consequential sensor placement and spatial prediction are thus obtained. We will show in Section 4 that the derived GP models are fine enough to provide high prediction accuracy.

3.3. Sensor Placement

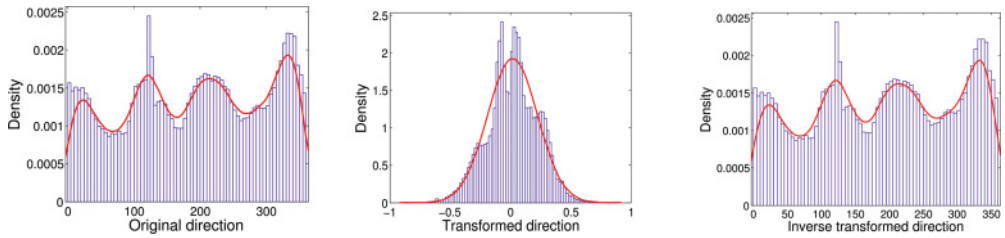
Through the GP model learned by the CFD-based dataset, we can find the optimal sensor locations in each monsoon season and intermonsoon season. For intermonsoon seasons, we need to first transform the uniform wind direction distribution to a Gaussian distribution. Finally, a permanent sensor placement scheme can be obtained by combining the results of all segments and considering the water quality sensitivity.

3.3.1. Sensor Placement for Single Monsoon Season. With the dataset of CFD modeling, we obtain a GP of wind for each season. It is NP-hard to select optimal sensor locations for predicting the mean, maximum, or minimum of other locations [Das and Kempe 2008]. Two widely used criteria to guide the sensor placement are entropy and mutual information. For entropy, the optimal sensor locations form a set that can provide the largest joint entropy.

$$\arg\max_{\mathcal{A}:|\mathcal{A}|=K} H(\mathcal{A}) \tag{4}$$

$$H(\mathcal{A}) = H(X_{a_k|a_{k-1},\ldots,a_{a_1}}) + \cdots + H(X_{a_2|a_1}) + H(X_{a_1})$$

Heuristic algorithms can be used to find the locations with largest entropy or conditional entropy iteratively. The selected locations provide the best prediction of observations at unobserved locations. For each location v, we treat the wind direction and



(a) Uniform (b) Transformed Gaussian (c) Inverse Uniform

Fig. 9. Transformation of uniform distribution in the intermonsoon season of the year 2007 at the Marina Channel meteorological station to Gaussian distribution, and inverse transformation of the transformed Gaussian distribution to a uniform distribution.

speed as a random variable vector. Its entropy is calculated as $H(X_v) = \frac{1}{2}log|2\pi e\sum_{v,v}|$, where $\sum_{v,v}$ is the covariance matrix of direction and speed at v.

The entropy criterion finds the most informative locations that are located far away from each other. An alternative [Guestrin et al. 2005], searches for locations that most significantly reduce the uncertainty of rest space through maximizing the mutual information between the selected locations and the rest, presented as $MI(\mathcal{V}\backslash\mathcal{A}, \mathcal{A}) = H(\mathcal{V}\backslash\mathcal{A}) - H(\mathcal{V}\backslash\mathcal{A}|\mathcal{A})$.

3.3.2. Transformation of Uniform Distribution. The uniform distribution of winds in intermonsoon seasons can be transformed to a Gaussian distribution using Inverse Transform Sampling (ITS). If X is uniformly distributed on $[0, 1]$ and $F(y_i) = x_i$, the random variable Y is drawn from a normal distribution described by its cumulative function $F = 1/2 + 1/2erf(y/\sqrt{2})$. Therefore, when we have the time series of wind direction in the intermonsoon season x_i, which follows a uniform distribution, we can transform it to a time series of Gaussian distribution, y_i, where $y_i = \sqrt{2}erf^{-1}(2x_i/360 - 1)$. Figure 9 shows that the transformed data by ITS can be fitted by a Gaussian distribution.

The advantage of ITS is that it supports bidirectional transformation. We can transform the wind data to Gaussian distribution to study the sensor placement and spatial prediction, and convert the estimated values of unobserved locations back to normal readings after processing. When we receive the real-time sensor readings, we first convert them to the transformed Gaussian domain and estimate the observations at the unobserved locations based on the transformed data. After that, we can convert the estimated data back to normal wind readings z_i through $z_i = 1/2 + 1/2erf(y_i/\sqrt{2})$. Figure 9 also shows that the inversely transformed data follows the exact distribution of the original data.

3.3.3. Sensor Placement for the Whole Year. The sets of sensor locations found for different monsoon or intermonsoon seasons are not exactly the same. Ideally, we deploy sensors in one season according to its optimal placement scheme and move the sensors according to another optimal placement scheme when the next season starts. However, in reality, we cannot do that due to high reinstallation cost in terms of both finance and time. We resort to providing a suboptimal solution to find a best balance among different seasons.

We consider this problem while calculating the entropy of each location. Assuming that the entropy of location v in jth time segment is $H(X_{v,j})$, the entropy of that location for all segments can be computed as:

$$H(X_v) = \sum_{j=1}^{3} w_j * H(X_{v,j}), \tag{5}$$

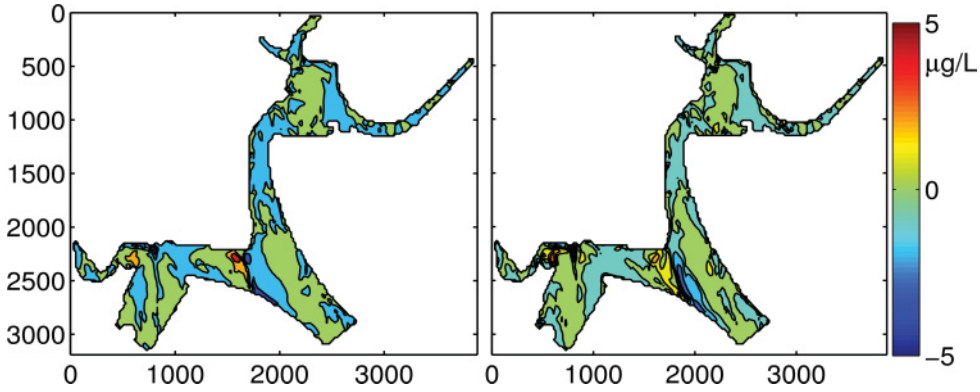


Fig. 10. The chlorophyll distribution differences between uniform wind and speed doubled at locations (1750, 2250) and (1850, 2250).

where w_j is the weight of jth time segment in the entire time series including two monsoon seasons and one combined intermonsoon season. From the viewpoint of information theory, by doing this, the information utility of each location is the sum of its entropies in each time segment weighted by the relative proportion. Once the location of the highest entropy is found, we search for the second and consequential locations by calculating the weighted conditional entropy until reaching the maximum number of sensors we can deploy. The optimal sensor placement scheme for one year is also the best solution for multiple years, as the wind pattern simply repeats with negligible changes for different years.

3.3.4. Sensitivity of Water Quality. To consider water quality during the design of the optimal wind sensor placement scheme, a sensitivity analysis is conducted to find the relative influence of wind at each location on the water quality in the Marina Reservoir. We first run the water-quality simulation with uniform wind distribution of the whole area and repeat that simulation by doubling the wind speed at one location. We record the differences of all water-quality parameters at each location between the two simulations. Figure 10 depicts the obvious differences in chlorophyll distributions for two scenarios. The chlorophyll sensitivity to the wind at location v is calculated as:

$$S_v = \sum_{j=1}^{N} \left| \frac{CHL_j^v - CHL_j}{CHL_j} \right|, \tag{6}$$

where N is the number of possible sensor locations and CHL_j^v is the chlorophyll value of jth location when the wind speed at location v is doubled. The sensitivity of water quality is the average of all water-quality parameters including chlorophyll, temperature, and dissolved oxygen. We find the water quality sensitivity to the wind at each location by repeating the experiments with doubled wind speed at that location. Figure 11 shows that the sensitivities of water quality at different locations are significantly distinct.

We factor the sensitivity analysis in calculating sensor placement by adjusting the information utility of each location with its normalized sensitivity.

$$H'(X_v) = S_v * H(X_v) \tag{7}$$

We normalize the raw sensitivity of each location by the highest sensitivity over the Marina Reservoir, which is at location (1750, 2250). From the viewpoint of information

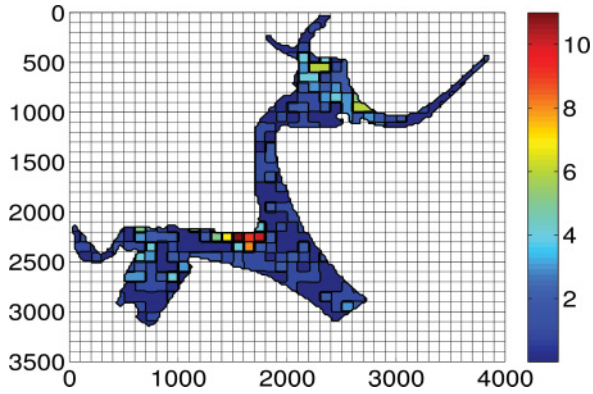


Fig. 11. Water-quality sensitivity to wind at different locations in the Marina Reservoir.

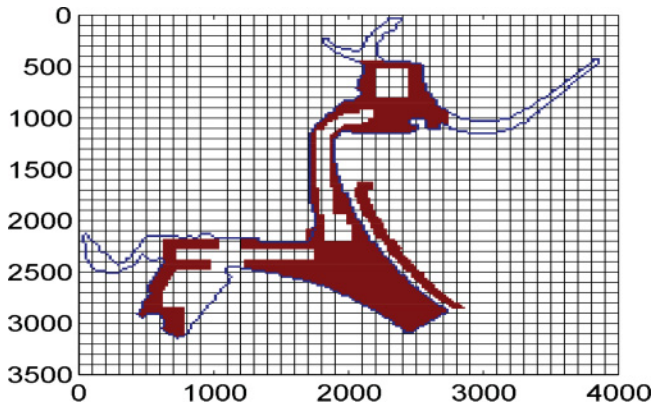


Fig. 12. The red region shows the area for which permission may be granted to install wind sensors.

theory, the information utility is the quantity one location can offer to eliminate the uncertainty of wind distribution of the whole reservoir. We reduce the information utility at one location if it has small impact on water quality. By doing this, the locations with high sensitivity will have more chance to be selected and wind sensors are deployed at the selected locations, which provide direct measurements with minimal error. The final studies of water quality will benefit from these wind fields with intended error distribution.

After the final adjustment to the sensor placement scheme according to sensitivity analysis, we calculate the entropy for all locations and obtain the final sensor placement with locations of the highest entropy or conditional entropy. Due to topography and regulatory constraints, we cannot install wind sensors at all desired locations. The area for which we may finally get permission to deploy sensors is depicted in Figure 12. We therefore choose the first location only if it has the highest entropy and is available to deploy sensors. If the location of the highest entropy is not permitted for sensor deployment, we turn to the next location with the highest entropy. We repeat this procedure until enough sensor locations are found. The number of sensors to deploy is constrained by the project budget and the prediction accuracy. In this study, we finally deployed 10 wind sensors which can provide acceptable prediction accuracy. The deployment layout of wind sensors is given in Section 4.

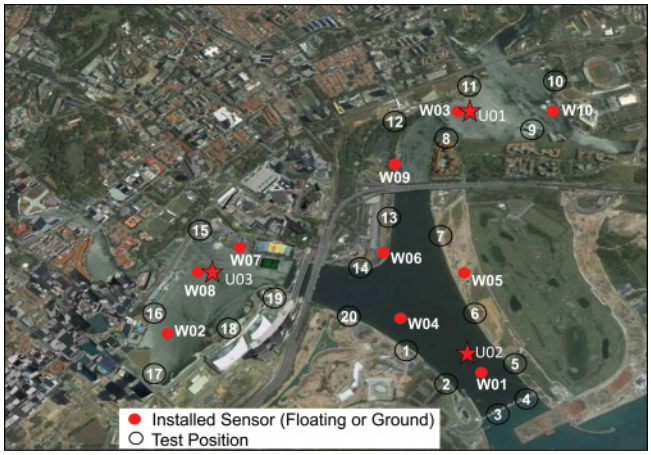


Fig. 13. Locations of deployed wind sensors and the test positions of the mobile sensor.

3.4. Spatial Prediction

We predict the observations at the unobserved locations as the mean of conditional distribution $X_{\mathcal{V}\setminus\mathcal{A}}|X_{\mathcal{A}}$ in Equation (1), with the decomposed Gaussian models and the input of real-time sensor readings. One problem is to determine which GP model should be used to perform the prediction. Because the start and end of monsoon seasons are variable for different years, we cannot cluster the sensor readings according to a fixed division scheme derived using the time series segmentation algorithm in Section 3.1.1.

An online temporal clustering algorithm is developed to dynamically search for the critical change point of monsoon seasons. When a new sensor reading is received, the likelihood of last N samples is calculated using the statistical model (Gaussian or uniform distribution) of current monsoon season. When the likelihood decreases to a user-defined threshold, τ, we infer that the transition of monsoon seasons occurs.

When a set of sensor readings measured by all deployed wind sensors at a given time point is categorized to a certain monsoon season, the relative GP model can then be applied to estimate the wind field on other unobserved locations using Equation (1).

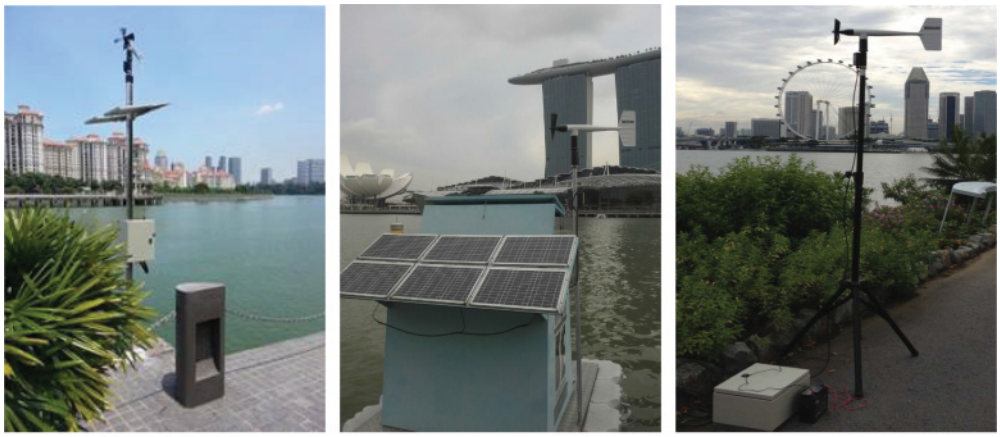
4. DEPLOYMENT AND EVALUATION

In this section, we introduce the in-field deployment of a wireless wind sensor network in the Marina Reservoir area and evaluate the performance of the proposed sensor placement and spatial prediction approaches with real measurement results.

4.1. Deployment of Wireless Sensor Network

The potential deployment area covers a water surface space of 2.2km^2 plus the terrain space within 100m from the water's edge since some locations on land may provide more information than those on the water surface to infer the wind observations on other locations. We divide the underlying area of the Marina Reservoir into small grids of 20m $*$ 20m, which provides the finest resolution. More than 5k locations need to be considered.

Ten wind sensors are finally deployed, as marked by the red dots in Figure 13, including 5 land sensors installed on the ground around the water and 5 floating sensors on the water surface. The locations are selected according to the proposed approach based on the historical data of the meteorological station on Marina Channel over two years (from 2007 to 2008). Due to the high computational complexity of calculating



(a) Land (W10) (b) Floating (W02) (c) Mobile (6)

Fig. 14. Three types of wind sensors.

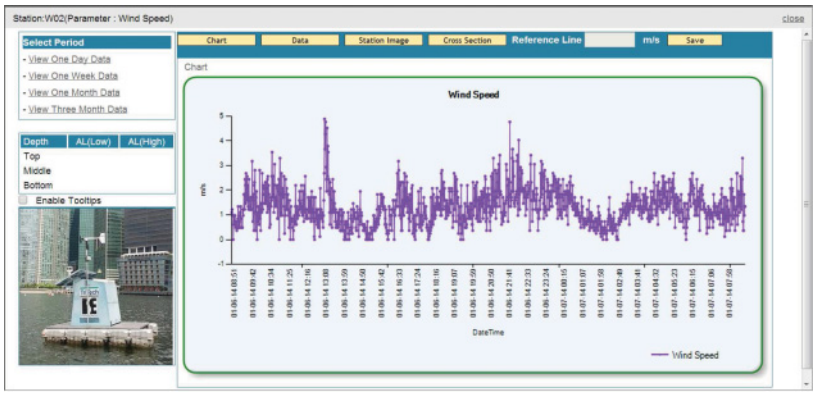


Fig. 15. The collected wind speed data drawn on the data access user interface.

mutual information over the large set of potential locations, entropy is used as the criterion for selecting the optimal sensor locations. Figure 13 also depicts the 3 sets of underwater sensors we deploy to measure some key parameters of water quality, such as dissolved oxygen, conductivity, chlorophyll, pH value, and temperature. The observations of underwater sensors will be used to study the water quality of the entire reservoir and validate our results on water quality.

Figure 14 depicts the wind sensors that we construct in this study. A land-based sensor, depicted in Figure 14(a), is fixed on the ground with an absolute reference direction. A floating sensor (Figure 14(b)) is anchored to the bottom of the water but floats on the water surface. It has limited rotational freedom. We add a compass of high accuracy for each floating sensor to determine the instant reference direction that will be used to calculate the absolute wind direction by offsetting the raw measurement. We also build a mobile wind sensor, depicted in Figure 14(c). It can be easily moved and set up temporarily at an arbitrary location. We use the mobile sensor to collect wind data for performance evaluation. Instead of a solar panel, a portable battery is used to provide energy. All the other components are the same as other permanent wind sensors.

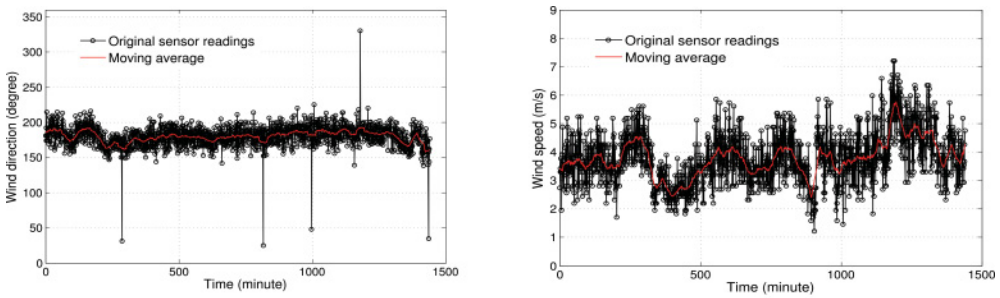


Fig. 16. Time series of wind direction and speed from 00:00 to 23:29 on September 1, 2013 and their smoothed values with a 30-minute moving average window.

For each sensor, the wind monitor model 05305L of R. M. Young is used, which provides an accuracy of 0.2m/s for speed and 3° for direction. In an early version of the data collection system, all wind sensors are equipped with a RTCU DX4 data logger. An accurate clock is provided in the data logger and all the sensor readings are instantly synchronized in-field. A solar panel is equipped to provide continuous power to the wind anemometer and data logger. For the mobile wind sensor, a portable battery is used to offer energy instead of a solar panel; all the other components are same as the land and floating wind sensors.

The minutely measured data are first logged and then transmitted back to our backend server directly through a cellular network. The real-time data are hosted in the server and can be accessed online through our data collection user interface. With this user interface, we can also monitor the status of each wind sensor. For example, Figure 16(b) presents the wind speed collected on January 6, 2014, shown by our data analysis tool on the Web site.

4.2. Experiment Setup

We evaluate the performance of the proposed approach by real measurements. With the deployed wireless wind sensor network, we have collected the wind data since July 2013. We study the accuracy of spatial prediction with reference to UniGau and linear interpolation. The latter method is widely used by current environmental analysis. The distance-weighted linear interpolation algorithm is adopted in our study, as shown in Equation (8):

$$x_{i\in\mathcal{V}|\mathcal{A}} = \frac{\sum_{j\in\mathcal{A}} x_j * \frac{1}{d_j}}{\sum_{j\in\mathcal{A}} \frac{1}{d_j}}. \tag{8}$$

The performance gain of our proposed approach comes from two aspects: optimal sensor placement and accurate spatial prediction. Spatial prediction is based on sensor placement, and they share the same system model. Since the advantages of Gaussian-based sensor placement over random deployment have been completely proven in previous works [Krause et al. 2006; Wu et al. 2012] and it is costly in terms of budget and time (more than 3 months) to reinstall the deployed wind sensors, we focus on evaluating the potential improvement of spatial prediction accuracy by comparing our approach (MIX) with UniGau and Interpolation based on the real wind measurements on the optimal sensor placement. The prediction error is measured by the average Root-Mean-Squared Error (RMSE) between the estimated values of unobserved

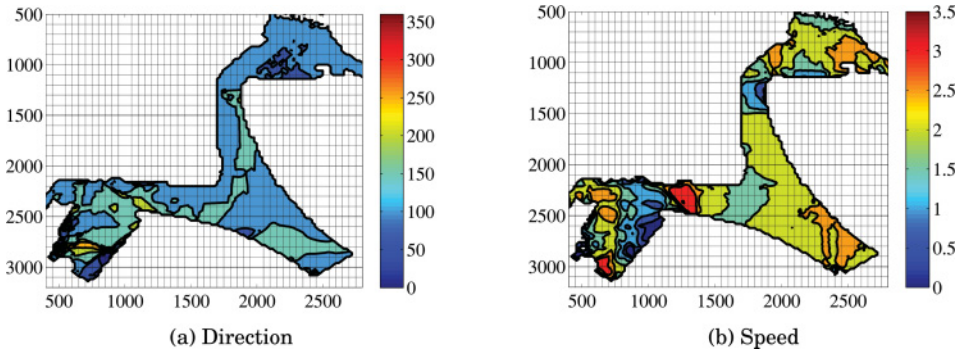


Fig. 17. Spatial prediction results derived by the proposed approach.

locations $\hat{X}_{\mathcal{V}\setminus\mathcal{A}}$ and their actual values $X_{\mathcal{V}\setminus\mathcal{A}}$:

$$RMSE(X_{\mathcal{V}\setminus\mathcal{A}}|X_{\mathcal{A}}) = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{\sum_{i\in\mathcal{V}\setminus\mathcal{A}}^{N}(\hat{X}_i^t - X_i^t)^2}{N}}. \quad (9)$$

Assume we have T sets of samples to conduct the evaluation and N locations are included in $\mathcal{V}\setminus\mathcal{A}$. The result is the average error in both temporal and spatial aspects.

4.3. Results

Figure 16 presents a snapshot of the measured wind distribution and speed for 24 hours at wind sensor W01. The minutely measured data is plotted as the black line and the smoothed data with a 30-minute moving average window is depicted as the red line. The latter is required as a stable input to feed the ELCOM-CAEDYM model. We will evaluate the performance of the proposed approach and the benchmark methods using the moving average data.

Figure 17 shows one example distribution of wind direction and speed derived by our proposed approach. From this example, we see that the spatial variation is large and our approach can provide a distribution with a fine-grained resolution. Since we cannot obtain the ground truth of the wind distribution over the whole reservoir at that time point, we will evaluate the accuracy of our approach based on the real measurements of many locations for long time periods.

Overall performance in space. To evaluate the spatial prediction accuracy at different locations, we measure wind direction and speed at 20 randomly selected locations along the water's edge of the Marina Reservoir using the mobile wind sensor. The test positions are depicted in Figure 13. At each location, per-minute wind data is collected for 1∼2 hours. Figure 18 presents the average RMSE of predicted direction and speed for each location.

Compared with UniGau and interpolation, the proposed approach reduces average RMSE of wind direction prediction by 81% and 33%, respectively. By the monsoon-based time series segmentation, MIX can accurately model the wind and provide high prediction accuracy. Because the wind direction distribution for the entire year is not Gaussian, UniGau produces large errors. The average RMSE of interpolation is relatively large because it does not consider the effect of surrounding buildings to wind field, thus cannot accurately capture the spatial variation of wind distribution. For wind speed prediction, the performance of MIX and UniGau is comparable since the wind speed of whole year is still Gaussian. They reduce the average RMSE of interpolation by 25%. From Figure 18, we can also see that the average RMSE of locations

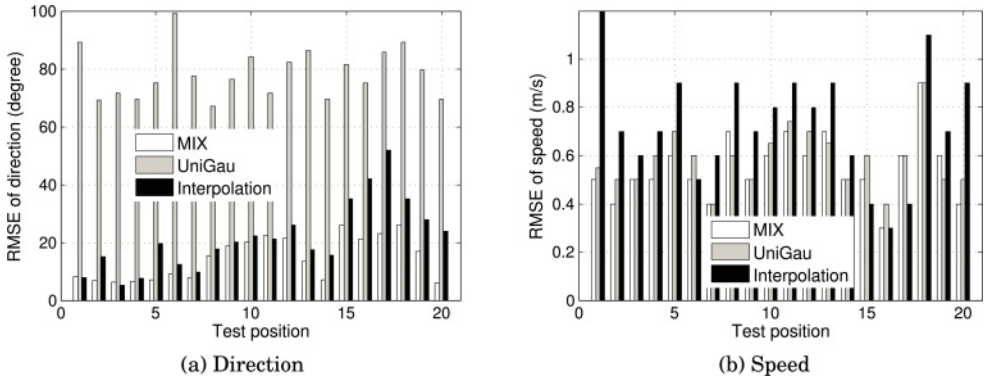


Fig. 18. Average prediction RMSE of wind direction and speed for 20 test positions.

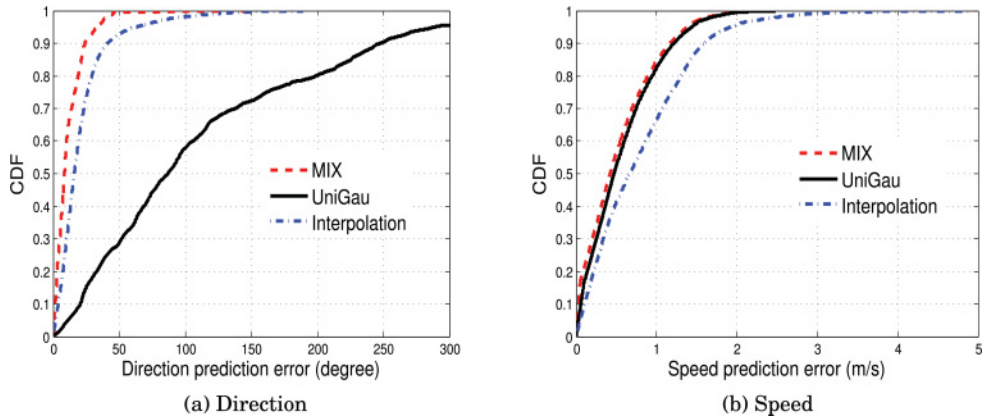


Fig. 19. Prediction error of wind direction and speed validated by the installed wind sensors.

near installed sensors or in open space is relatively small, because the wind patterns can be better captured by the statistic models and CFD modeling.

Overall Performance in Time. To further investigate the performance of the proposed approach for long-term wind measurement, we use the measurement data of all 10 sensors for 3 months. At one time, we choose one sensor and use the measurement data from the remaining 9 sensors to predict its wind direction and speed. We use its own measurement as the reference to calculate RMSE. We perform this evaluation in 10 rounds for all 10 sensors. We do not have data from W09 since it was missing a short time after installation. We are redeploying it to the opposite edge of the Kallang River, which is more secure and provides the same level of information for the spatial prediction over the target area.

Figure 19 presents the cumulative distribution of the absolute difference between predicted observation and the measured value for each sample. In this case, T and N in Equation (9) are both equal to 1. Similar to the results of the mobile sensor testing, MIX improves the prediction accuracy of wind direction by 87% for UniGau and 27% for interpolation, and the performances of MIX and UniGau for speed prediction are comparable and 21% higher than that of interpolation. MIX offers an average spatial prediction accuracy of 24°. The error is larger than the mobile test since the data of one installed wind sensor is used as the reference for evaluation but not included in the calculation of spatial prediction.

Table I. Computational Efficiency of Spatial Prediction

Performance metrics	Results
Total time for wind direction prediction(s)	7.09
Number of grids	30846
Time per grid(s)	2.3e-04

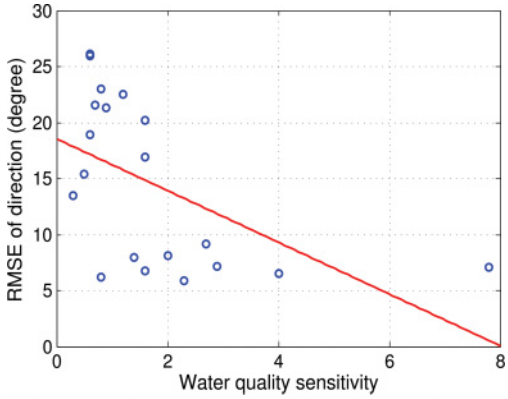


Fig. 20. The RMSE of direction prediction corresponding to the water-quality sensitivity for each mobile test location. The red line is the linear fitting curve of the scatter points.

Computational Efficiency. Table I presents the online computational efficiency of our spatial prediction approach. Since the sensor placement calculation is done offline before the deployment of sensors, we only measure the computational time of spatial prediction, which should be done as fast as possible once the real-time sensor readings are collected on the server. In Table I, we see that when a set of sensor readings measured at one time point from 10 locations are received, it takes 7.09s to calculate the distribution of wind direction over the entire Marina Reservoir. The spatial prediction of wind speed shares the same process with wind direction. Therefore, we need 14.18s to generate the final results for the following water quality studies. The test is conducted on a 64-bit server composed of a 3.2GHz Intel(R) Xeon(R) CPU and 16GB RAM. From the results, we can also see that the computation time for an even smaller granularity, for example, a grid size of 10m*10m, is 56.72s, which is still acceptable.

Sensitivity of water quality. We take into account water quality while solving the sensor placement problem so as to obtain an intended error distribution in space. Figure 20 presents the average RMSE at each test position in the experiment with the mobile sensor corresponding to the sensitivity of water quality at that location. The results show that the average RMSE is relatively low at locations with high water quality sensitivity. The linear regression between the water-quality sensitivities and the relative RMSEs of predicted direction for different locations reveals such an inverse trend.

Online clustering algorithm. We use historical data to evaluate the efficiency of our online clustering algorithm. The experiments are done using the historical wind data of the Marina Channel meteorological station over the year 2008. The likelihood calculated by the online clustering algorithm is the average likelihood of all N samples in the sliding window. Only the last N samples before the time point under consideration are used in the calculation. We set the sliding window size N to 1,440 samples corresponding to one day and the likelihood threshold τ to 5.45. The likelihood calculated offline is obtained using the monsoon-based time series segmentation algorithm introduced in Section 3.1 with the whole-year data. It is the sum likelihood of all samples. Figure 21 shows that the likelihoods calculated online and offline peak almost at the same time

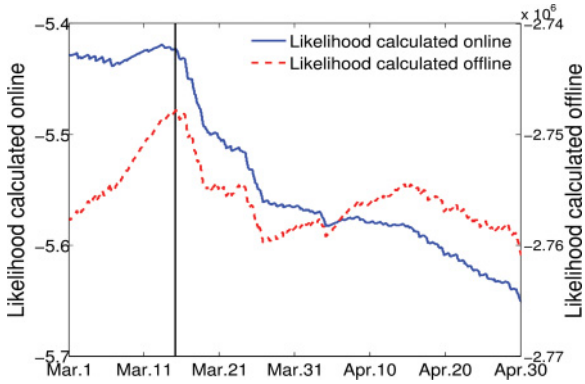


Fig. 21. Likelihood calculated online and offline.

point for the transition from NE monsoon season to PreSW intermonsoon season. The online temporal clustering algorithm can find the critical change point of time series with a small error of 3.6 days in this case.

5. LESSONS LEARNED

The wind we measure in this study exhibits a distinct non-Gaussian process from previously studied phenomena. Besides, we had no full prior knowledge to model the spatial correlation in the field. To work with these challenges, we developed a unique sensor placement approach including monsoon-based time series segmentation and the dataset generated using CFD modeling.

The proposed approach, on the other hand, demonstrates a complete procedure that extends the state-of-the-art sensor placement and spatial prediction methodology to solve a wider range of applications. There are many other phenomena that exhibit a similar non-Gaussian process captured by a mixture model over time, for example, city traffic flow [Sun et al. 2006], soil pollutant [Lin et al. 2010], and so forth. Generalizing the time series segmentation approach in such applications will significantly improve the accuracy of sensor placement and spatial prediction over clustered time periods.

Incorporating computational simulation, for example, CFD in this work, to build a dataset and study the field data correlation provides us a new method to gain prior knowledge when an intrusive way of learning such knowledge is not preferred. We do not need to run the time-consuming computational simulations online, but conduct just enough simulations to cover most cases and capture the statistical features of the target phenomena. In many applications in which it is impractical to predeploy enough sensors due to various constraints, the computational simulation procedure can provide coarse yet sufficient knowledge to statistically capture the spatial correlations that we need.

Wind measurements are often conducted in order to study some other environmental phenomena. We consider the water-quality modeling during sensor placement since the winds of different locations impose variant impact on water quality of the whole lake due to diverse morphometries and patterns. By considering the sensitivity of water quality during the calculation of entropy for each location, we deploy sensors to provide direct measurements at locations that have high impact on water quality.

6. DISCUSSION

In this section, we discuss the applications and limitations of the proposed solution.

Is the proposed solution able to scale to a much larger lake? Yes. According to the experiment results in Section 4.3, the computation complexity of the proposed solution

is low. The computation time of spatial prediction for both wind direction and wind speed over a reservoir composed of 30,846 grids is 14.18s by the computation in Matlab. If the area of the reservoir is increased 4 times, the number of grids is increased 4 times as well and the computation time is 56.72s, which is still acceptable.

How will the proposed solution work in other parts of the world? We believe the proposed solution best exploits the underlying nature of wind measurement in our study and achieves optimal accuracy with a limited number of sensors. Such a procedure can be generalized to measure wind in other regions or handle other applications in observing phenomena of a similar nature. The experiences we gained from time-series segmentation and water-quality sensitivity can be used in other studies.

Will the proposed solution be able to cope with frequent extreme weather (such as places with frequent hurricanes)? No. Our solution is mainly designed for wind measurement using normal wind sensors in the regions where the wind follows Gaussian distribution or can be decomposed to multiple Gaussian distributions. Because it is impossible to measure the wind speed and direction in hurricanes using current normal wind sensors, it is out of the scope of the proposed solution to measure hurricanes. Moreover, wind modeling is challenging for frequent extreme weather, such as hurricanes. It is very hard to obtain prior knowledge of the wind distribution over a large area during extreme weather.

7. RELATED WORKS

WSN for environmental monitoring. In the last decade, a large number of WSN deployments [Liu et al. 2013; Li and Liu 2009; Hu et al. 2009; Talzi et al. 2007; Barrenetxea et al. 2008; Ingelrest et al. 2010; Le Dinh et al. 2007] have been reported for environmental monitoring. They have also been used for pipeline monitoring [Lai et al. 2012] and mapping [Lai et al. 2013]. Some diagnostic approaches have also been designed for these applications [Liu et al. 2008; Dereszynski and Dietterich 2011]. Many large-scale systems with hundreds of nodes [Dutta et al. 2006; He et al. 2006; Liang et al. 2009; Liu et al. 2013] have been developed as well, such as military surveillance, temperature measurement in data centers, and forest monitoring. In this article, we focus on the optimal sensor placement problem for wind distribution measurements in large areas.

Sensor placement. The optimal sensor placement problem has been addressed in many previous works [Dhillon and Chakrabarty 2003; Lin and Chiu 2005; Joshi and Boyd 2009]. Among them, GP-based approaches [Das and Kempe 2008; Guestrin et al. 2005; Osborne et al. 2008] have been used in many applications monitoring spatial phenomena such as temperature [Krause et al. 2006] and soil moisture [Wu et al. 2012]. Communication efficiency is also be taken into account during the design of sensor placement [Krause et al. 2011]. The best location for base station is studied by Shi and Hou [2009] to maximize the network lifetime within an error bound. The nonuniform prediction accuracy problem is considered by Krause et al. [2008]. The reduction in the predicted variance over the unobserved locations is weighted according to their demand of accuracy. However, they cannot be directly applied to wind distribution measurement due to the temporal and spatial variations of wind. They also require the historical data of wind distribution over the whole area. In our case, they need to deploy more than 5,000 wind sensors above the Marina Reservoir with a grid of 20m*20m for more than 1 year. Even if they approximate the required Gaussian process model by some existing kernel functions, they also need the historical data from many wind sensors to validate their hypothesis. Due to the highly dynamic variation of wind in space over such an urban reservoir, the required number of wind sensors must be very large to achieve a high accuracy. It is impossible to deploy a large number of wind sensors in such an area beforehand.

Coverage. The coverage problem of sensor networks has been extensively studied [Kumar et al. 2004; Ganesan et al. 2006; Yan et al. 2008; Chen et al. 2013]. The

sensor placement problem for moving targets is considered by Wettergren and Costa [2012, 2009]. Data fusion is considered by Xing et al. [2004, 2009]. The coverage and connectivity in a duty-cycled sensor network are analyzed by Wang et al. [2003], Gu et al. [2013], and Yan et al. [2003]. Nevertheless, most of the existing theoretical works are based on the deterministic disc model or do not consider the unique features of target physical phenomena.

CFD modeling. It is a widely used tool to capture the fluid patterns in many applications, such as environmental engineering and aircraft design [Anderson et al. 1995]. Lim et al. [2012] have successfully applied CFD for geospatial risk assessment of wind channels in urban area with high accuracy. CFD modeling has also been used in sensor placement problems [Wang et al. 2011] and temperature forecasting [Chen et al. 2012] in data center environments. CFD models are built to capture extra hot-spot scenarios. A thermal forecasting model is proposed by Li et al. [2011] to model and predict temperatures around servers in data centers based on principles from thermodynamics and fluid mechanics.

Time series segmentation. Time series segmentation algorithms [Bernaola-Galván et al. 1996; Guralnik and Srivastava 1999] search for critical change points by iteratively dividing data into small segments with the same statistical model (e.g., Gaussian distribution). However, they cannot be applied for our mixture model of different statistic models, that is, Gaussian and uniform. Expectation maximum algorithms [Dempster et al. 1977] are utilized widely to divide a Gaussian mixture into individual Gaussian distributions. However, they cannot be used in the application of wind measurement either. First, the spatial correlation cannot be calculated since the samples of all locations at a given time point are not clustered in the same cluster. Second, the samples in the same cluster are not continuous in time. As a consequence, it is difficult to assign the online sensor readings to a proper cluster and apply the relative spatial prediction.

8. CONCLUSIONS

In this article, we propose a novel sensor placement and spatial prediction approach for wind distribution measurements. It leverages the monsoon characteristics of wind to study its statistic properties. A dataset is built using CFD modeling that captures the impact of surrounding buildings on wind distribution. Optimal sensor locations are selected through segmented wind statistical models and adjusted according to the sensitivity of water quality to wind at different locations. We deployed 10 wind sensors around or on the water surface of an urban reservoir. The observations of unobserved locations are predicted by the readings of deployed sensors clustered through an online algorithm. The in-field measurement results show that the proposed approach can significantly improve the accuracy of wind measurements.

REFERENCES

R. Alexander and J. Imberger. 2009. Spatial distribution of motile phytoplankton in a stratified reservoir: the physical controls on patch formation. *Journal of Plankton Research,* 31 (1), 101–118.

John D. Anderson. 1995. *Computational Fluid Dynamics: The Basics with Applications*. McGraw-Hill, New York, NY.

Guillermo Barrenetxea, François Ingelrest, Gunnar Schaefer, and Martin Vetterli. 2008. The hitchhiker's guide to successful wireless sensor network deployments. In *ACM SenSys*. 43–56.

Pedro Bernaola-Galván, Ramón Román-Roldán, and José L. Oliver. 1996. Compositional segmentation and long-range fractal correlations in DNA sequences. *Physical Review E,* 5181–5189.

Tony Burton, Nick Jenkins, David Sharpe, and Ervin Bossanyi. 2011. *Wind Energy Handbook*. John Wiley & Sons, Hoboken, NJ.

Jiming Chen, Junkun Li, Shibo He, Tian He, Yu Gu, and Youxian Sun. 2013. On energy-efficient trap coverage in wireless sensor networks. *ACM Trans. Sen. Netw.* 10, 1, Article 2, 29 pages.

Jinzhu Chen, Rui Tan, Yu Wang, Guoliang Xing, Xiaorui Wang, Xiaodong Wang, Bill Punch, and Dirk Colbry. 2012. A high-fidelity temperature distribution forecasting system for data centers. In *IEEE RTSS*. 215–224.

Lin Sien Chia, Ausafur Rahman, and Dorothy Bee Hian Tay. 1991. *The Biophysical Environment of Singapore*. Singapore University Press, Singapore.

Noel Cressie. 1993. *Statistics for Spatial Data*. John Wiley and Sons, Hoboken, NJ.

Abhimanyu Das and David Kempe. 2008. Sensor selection for minimizing worst-case prediction error. In *ACM/IEEE IPSN*. 97–108.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39, 1, 1–38.

Ethan W. Dereszynski and Thomas G. Dietterich. 2011. Spatiotemporal models for data-anomaly detection in dynamic environmental monitoring campaigns. *ACM Trans. Sen. Netw.* 8, 1, Article 3, 36 pages.

Santpal Singh Dhillon and Krishnendu Chakrabarty. 2003. Sensor placement for effective coverage and surveillance in distributed sensor networks. In *IEEE WCNC*. 1609–1614.

Wan Du, Zhenjiang Li, Jansen Christian Liando, and Mo Li. 2014a. From rateless to distanceless: Enabling sparse sensor network deployment in large areas. In *ACM SenSys*.

Wan Du, Mo Li, Zikun Xing, Bingsheng He, Lloyd Hock Chye Chua, Zhenjiang Li, Yuanqiang Zheng, and Pengfei Zhou. 2014b. Demo abstract: Wind measurements for water quality studies in urban reservoirs. In *IEEE SECON*.

Wan Du, Zikun Xing, Mo Li, Bingsheng He, Lloyd Hock Chye Chua, and Haiyan Miao. 2014c. Optimal sensor placement and measurement of wind for water quality studies in urban reservoirs. In *ACM/IEEE IPSN*. IEEE , 167–178.

Prabal Dutta, Jonathan Hui, Jaein Jeong, Sukun Kim, Cory Sharp, Jay Taneja, Gilman Tolle, Kamin Whitehouse, and David Culler. 2006. Trio: Enabling sustainable and scalable outdoor wireless sensor network deployments. In *ACM/IEEE IPSN*. 407–415.

Deepak Ganesan, Răzvan Cristescu, and Baltasar Beferull-Lozano. 2006. Power-efficient sensor placement and transmission structure for data gathering under distortion constraints. *ACM Trans. Sens. Netw.* 2, 2, 155–181.

Yu Gu, Long Cheng, Jianwei Niu, Tian He, and David H. C. Du. 2013. Achieving asymmetric sensing coverage for duty cycled wireless sensor networks. *IEEE Trans. Parallel Distrib. Syst.*

Carlos Guestrin, Andreas Krause, and Ajit Paul Singh. 2005. Near-optimal sensor placements in gaussian processes. In *ICML*.

Valery Guralnik and Jaideep Srivastava. 1999. Event detection from time series data. In *ACM KDD*. 33–42.

Tian He, P Vicaire, Ting Yan, Qing Cao, Gang Zhou, Lin Gu, Liqian Luo, R. Stoleru, J. A. Stankovic, and T. F. Abdelzaher. 2006. Achieving long-term surveillance in VigilNet. In *IEEE INFOCOM*. 1–12.

Wen Hu, Nirupama Bulusu, Chun Tung Chou, Sanjay Jha, Andrew Taylor, and Van Nghia Tran. 2009. Design and evaluation of a hybrid sensor network for cane toad monitoring. *ACM Trans. Sen. Netw.* 5, 1, Article 4, 28 pages.

François Ingelrest, Guillermo Barrenetxea, Gunnar Schaefer, Martin Vetterli, Olivier Couach, and Marc Parlange. 2010. SensorScope: Application-specific sensor network for environmental monitoring. *ACM Trans. Sen. Netw.* 6, 2, Article 17, 32 pages.

Siddharth Joshi and Stephen Boyd. 2009. Sensor selection via convex optimization. *IEEE Trans. Sig. Proc.* 57, 2, 451–462.

Chun-Wa Ko, Jon Lee, and Maurice Queyranne. 1995. An exact algorithm for maximum entropy sampling. *Operations Research* 43, 4, 684–691.

Andreas Krause, Carlos Guestrin, Anupam Gupta, and Jon Kleinberg. 2006. Near-optimal sensor placements: Maximizing information while minimizing communication cost. In *ACM/IEEE IPSN*. 2–10.

Andreas Krause, Carlos Guestrin, Anupam Gupta, and Jon Kleinberg. 2011. Robust sensor placements at informative and communication-efficient locations. *ACM Trans. Sen. Netw.* 7, 4, Article 31, 33 pages.

Andreas Krause, Eric Horvitz, Aman Kansal, and Feng Zhao. 2008. Toward community sensing. In *ACM/IEEE IPSN*. 481–492.

Santosh Kumar, Ten H. Lai, and József Balogh. 2004. On k-coverage in a mostly sleeping sensor network. In *ACM MobiCom*. 144–158.

Ted Tsung-Te Lai, Wei-Ju Chen, Yu-Han Tiffany Chen, Polly Huang, and Hao-Hau Chu. 2013. Mapping hidden water pipelines using a mobile sensor droplet. *ACM Trans. Sen. Netw.* 9, 2, Article 20, 33 pages.

Ted Tsung-Te Lai, Wei-Ju Chen, Kuei-Han Li, Polly Huang, and Hao-Hua Chu. 2012. TriopusNet: Automating wireless sensor network deployment and replacement in pipeline monitoring. In *ACM/IEEE IPSN*. ACM Press, New York, NY, 61–72.

Bernard Laval, Jörg Imberger, Ben R. Hodges, and Roman Stocker. 2003. Modeling circulation in lakes: Spatial and temporal variations. *Limnology and Oceanography* 48 (3), 983–984.

Tuan Le Dinh, Wen Hu, Pavan Sikka, Peter Corke, Leslie Overs, and Stephen Brosnan. 2007. Design and deployment of a remote robust sensor network: Experiences from an outdoor water quality monitoring network. In *IEEE LCN*. 799–806.

Lei Li, Chieh-Jan Mike Liang, Jie Liu, Suman Nath, Andreas Terzis, and Christos Faloutsos. 2011. ThermoCast: A cyber-physical forecasting model for datacenters. In *ACM KDD*. 1370–1378.

Mo Li and Yunhao Liu. 2009. Underground coal mine monitoring with wireless sensor networks. *ACM Trans. Sen. Netw.* 5, 2, Article 10, 29 pages.

Chieh-Jan Mike Liang, Jie Liu, Liqian Luo, Andreas Terzis, and Feng Zhao. 2009. RACNet: A high-fidelity data center sensing network. In *ACM SenSys*. 15–28.

Tian Kuay Lim, Haiyan Miao, Choonseng Chew, Kee Khoon Lee, and Durairaju Kumaran Raju. 2012. Environmental modeling for geospatial risk assessment of wind channels in Singapore. In *GSDI*.

Frank Y. S. Lin and Pei-Ling Chiu. 2005. A near-optimal sensor placement algorithm to achieve complete coverage-discrimination in sensor networks. *IEEE Commun. Lett.* 43–45.

Yu-Pin Lin, Bai-You Cheng, Guey-Shin Shyu, and Tsun-Kuo Chang. 2010. Combining a finite mixture distribution model with indicator kriging to delineate and map the spatial patterns of soil heavy metal pollution in Chunghua County, central Taiwan. *Elsevier Environ. Pollution.* 235–244.

Guojin Liu, Rui Tan, Ruogu Zhou, Guoliang Xing, Wen-Zhan Song, and Jonathan M. Lees. 2013. Volcanic earthquake timing using wireless sensor networks. In *ACM/IEEE IPSN*. 91–102.

Kebin Liu, Mo Li, Yunhao Liu, Minglu Li, Zhongwen Guo, and Feng Hong. 2008. Passive diagnosis for wireless sensor networks. In *ACM SenSys*. 113–126.

Yunhao Liu, Yuan He, Mo Li, Jiliang Wang, Kebin Liu, and Xiangyang Li. 2013. Does wireless sensor network scale? A measurement study on GreenOrbs. *IEEE Trans. Parallel Distrib. Syst.* 24, 10, 1983–1993.

Alexandra Meliou, Andreas Krause, Carlos Guestrin, and Joseph M. Hellerstein. 2007. Nonmyopic informative path planning in spatio-temporal models. In *AAAI*. 602–607.

Michael A. Osborne, Stephen J. Roberts, Alex Rogers, Sarvapali D. Ramchurn, and Nicholas R. Jennings. 2008. Towards real-time information processing of sensor network data using computationally efficient multi-output Gaussian processes. In *ACM/IEEE IPSN*. 109–120.

Yi Shi and Y. Thomas Hou. 2009. Optimal base station placement in wireless sensor networks. *ACM Trans. Sen. Netw.* 5, 4, Article 32, 24 pages.

Shiliang Sun, Changshui Zhang, and Guoqiang Yu. 2006. A Bayesian network approach to traffic flow forecasting. *IEEE Trans. Intell. Transport. Syst.* 7, 1, 124–132.

Igor Talzi, Andreas Hasler, Stephan Gruber, and Christian Tschudin. 2007. PermaSense: Investigating permafrost with a WSN in the Swiss Alps. In *ACM EmNets*. 8–12.

Kevin E. Trenberth, David P. Stepaniak, and Julie M. Caron. 2000. The global monsoon as seen through the divergent atmospheric circulation. *Journal of Climate* 13, 22, 3969– 3993.

Xiaodong Wang, Xiaorui Wang, Guoliang Xing, Jinzhu Chen, Cheng-Xian Lin, and Yixin Chen. 2011. Towards optimal sensor placement for hot server detection in data centers. In *IEEE ICDCS*. 899–908.

Xiaorui Wang, Guoliang Xing, Yuanfang Zhang, Chenyang Lu, Robert Pless, and Christopher Gill. 2003. Integrated coverage and connectivity configuration in wireless sensor networks. In *ACM SenSys*. 28–39.

Thomas A. Wettergren and Russell Costa. 2009. Optimal placement of distributed sensors against moving targets. *ACM Trans. Sen. Netw.* 5, 3, Article 26, 25 pages.

Thomas A. Wettergren and Russell Costa. 2012. Optimal multiobjective placement of distributed sensors against moving targets. *ACM Trans. Sen. Netw.* 8, 3, Article 21, 23 pages.

Xiaopei Wu, Mingyan Liu, and Yue Wu. 2012. In-situ soil moisture sensing: Optimal sensor placement and field estimation. *ACM Trans. Sen. Netw.* 8, 4, Article 33, 30 pages.

Guoliang Xing, Chenyang Lu, Robert Pless, and Joseph A. O'Sullivan. 2004. Co-grid: An efficient coverage maintenance protocol for distributed sensor networks. In *ACM/IEEE IPSN*. 414–423.

Guoliang Xing, Rui Tan, Benyuan Liu, Jianping Wang, Xiaohua Jia, and Chih-Wei Yi. 2009. Data fusion improves the coverage of wireless sensor networks. In *ACM MobiCom*. 157–168.

Zikun Xing, Derek A. Fong, Edmond Yat-Man Lo, and Stephen G. Monismith. 2014a. Thermal structure and variability of a shallow tropical reservoir. *Limnology and Oceanography* 59, 1, 115–128.

Zikun Xing, Cheng Liu, Lloyd H. C. Chua, Bingsheng He, and Jörg Imberger. 2014b. Impacts of variable wind forcing in urban reservoirs. In *7th International Symposium on Environmental Hydraulics*.

Ting Yan, Yu Gu, Tian He, and John A. Stankovic. 2008. Design and optimization of distributed sensing coverage in wireless sensor networks. *ACM Trans. Embed. Comput. Syst.* 7, 3, Article 33, 40 pages.

Ting Yan, Tian He, and John A. Stankovic. 2003. Differentiated surveillance for sensor networks. In *ACM SenSys*. 51– 62.

Received May 2014; revised August 2014; accepted September 2014