Meteorology and Atmospheric Physics (2022) 134:18
https://doi.org/10.1007/s00703-021-00857-3

ORIGINAL PAPER

Data-driven model of the local wind field over two small lakes in Jyväskylä, Finland

Takayuki Shuku[1] · Janne Ropponen[2] · Janne Juntunen[2] · Hiroshi Suito[3]

Received: 25 January 2021 / Accepted: 20 December 2021 / Published online: 18 January 2022

Abstract

This study presents a data-driven model of the local wind field over two small lakes in Jyväskylä, Finland. Five temporary monitoring stations installed during the summers of 2015 and 2016 observed wind speed/direction around the two lakes. In addition, an official meteorological station located 15 km north of the lakes is permanently available. Our goal was to develop a model that could evaluate wind speed and direction over the two lakes using only data from the permanent station. Statistical analysis for the spatio-temporal wind data revealed that (1) local wind speed is correlated with the elevation and its cyclic pattern is identical to that of the official-station data, and (2) the local wind direction field is spatially homogeneous and is strongly correlated with the official-station data. Based on these results, we built two regression models for estimating spatial distribution of local wind speed and directions based on the digital elevation model (DEM) and official-station data. We compared the predicted wind speeds/directions by the proposed model with the corresponding observation data and a numerical result for model validation. We found that the proposed model could effectively simulate heterogeneous local wind fields and considers uncertainty of estimates.

1 Introduction

Lake ecosystems are strongly influenced by circulation, diffusion, and mixing processes which are dominantly induced by wind shear stress. These wind-induced currents, also referred to as wind-driven currents, can be a dominant factor in aquatic ecosystems dynamic, particularly in shallow lakes, and the accurate assessment of the spatial distribution of wind shear stress over lakes is needed to better interpret limnological data and to conduct comprehensive studies of aquatic ecosystems (e.g., Bengtsson and Hellstrom 1992; Podsetchine and Schernewski 1999; Bachmann et al. 2000; Chao et al. 2017; Juntunen et al. 2019).

In practice, lake circulation models often assume wind fields to be spatially homogeneous, despite the recognized importance of spatial wind distributions (e.g., Podsetchine and Schernewski 1999). This is because high-resolution spatial data are usually unavailable for use in lake simulations, and there are no validated methods to evaluate local wind fields using only limited monitoring data. The application of computational fluid dynamics (CFD) models might be a promising approach to estimate the spatial heterogeneity of wind shear stress (Ratto et al. 1994; Kitada et al. 1998; Laird and Walsh 2003; Ferragut et al. 2011). Martinez-Garcia et al. (2021) extensively overviewed recent developments of local wind fields. However, issues with the practical application of CFD models include high computation costs, difficulties in determining reliable input parameters, and initial/boundary conditions that strongly impact simulation results. Data-driven modeling is another approach that has received much attention due to the development of machine learning methods and high-performance computers that can handle large amounts of data. Several studies have explored data-driven wind-field modeling. Robert et al. (2013) presented a general regression neural network for interpolating monthly wind speeds in complex Alpine orography. Bessac et al. (2015) proposed a multiscale stochastic generator for wind speed and demonstrated the model. Torma and Kramer

Responsible Editor: Clemens Simmer.

✉ Takayuki Shuku
shuku@cc.okayama-u.ac.jp

[1] Graduate School of Environmental and Life Science, Okayama University, 3-1-1 Tsushima naka, Kita-ku, Okayama, Okayama 700-8530, Japan

[2] Finnish Environment Institute SYKE, Jyväskylä Office, Survontie 9 A, 40500 Jyväskylä, Finland

[3] Advanced Institute for Materials Research (AIMR), Tohoku University, 2-1-1 Katahira, Aoba-ku, Sendai, Miyagi, Japan

(2017) estimated spatial wind fields based on an inverse distance-weighted method and validated the model through comparison with corresponding observation data.

Although data-driven modeling is promising for the estimation of local wind fields, a large dataset is usually required for accurate estimation. For example, Robert et al. (2013) used wind-speed data observed at more than 100 meteorological stations. In practice, however, the amount of available data is usually very limited, and monitoring stations are spatially distant. An approach that can reasonably evaluate local wind fields using small datasets is necessary.

This study presents a data-driven model for estimating local wind fields over two small lakes, Palokkajärvi and Tuomiojärvi, located in the northern part of Jyväskylä, Finland (Fig. 1) using a limited amount of data. Around the two lakes, five temporary monitoring stations installed during the summers of 2015 and 2016 observed wind speed/direction. In addition, an official meteorological station (Tikkakoski airport) located 15 km north of the two lakes was permanently available. This study aimed to develop a statistical model that could evaluate the local wind field, including wind speed and direction, over the two lakes using only data from the permanent station data and to demonstrate the model's applicability through a comparison of the model's predicted results and the corresponding observation data. In our previous research (Juntunen et al. 2019), a method for estimating wind field for the same area was proposed. Our previous method, however, is an interpolation method and requires local-station data. It cannot be applied to "prediction" that is the main focus in this study. The target problem in this study is unique, that is neither simple prediction nor simple interpolation, and existing data-driven methods, which are basically developed for interpolation, are difficult to directly apply to this problem, i.e., the target problem is not simple interpolation, it includes extrapolation. Thus, this study newly proposes a novel data-driven model for this unique problem. There are several machine learning methods that are applicable to data poor scenarios. For example, least absolute shrinkage selection operator (lasso, Tibshirani 1996) has received considerable attention in many research areas to solve the problems with "sparse" data, and Bayesian approach (e.g., Bishop 2006) is also promising to deal with data poor problems. This study, however, attempts to make the model as simple as possible for practical applications, and application of those advanced machine learning methods remains a topic for future study.

This paper is structured as follows. Within the "Materials and methods," Sect. 2.1 presents basic information on the target lakes and observation data, and Sect. 2.2 outlines the data-driven model for the target lakes in detail, including parameters of the wind data and the derivation of statistical regression models to estimate wind speed

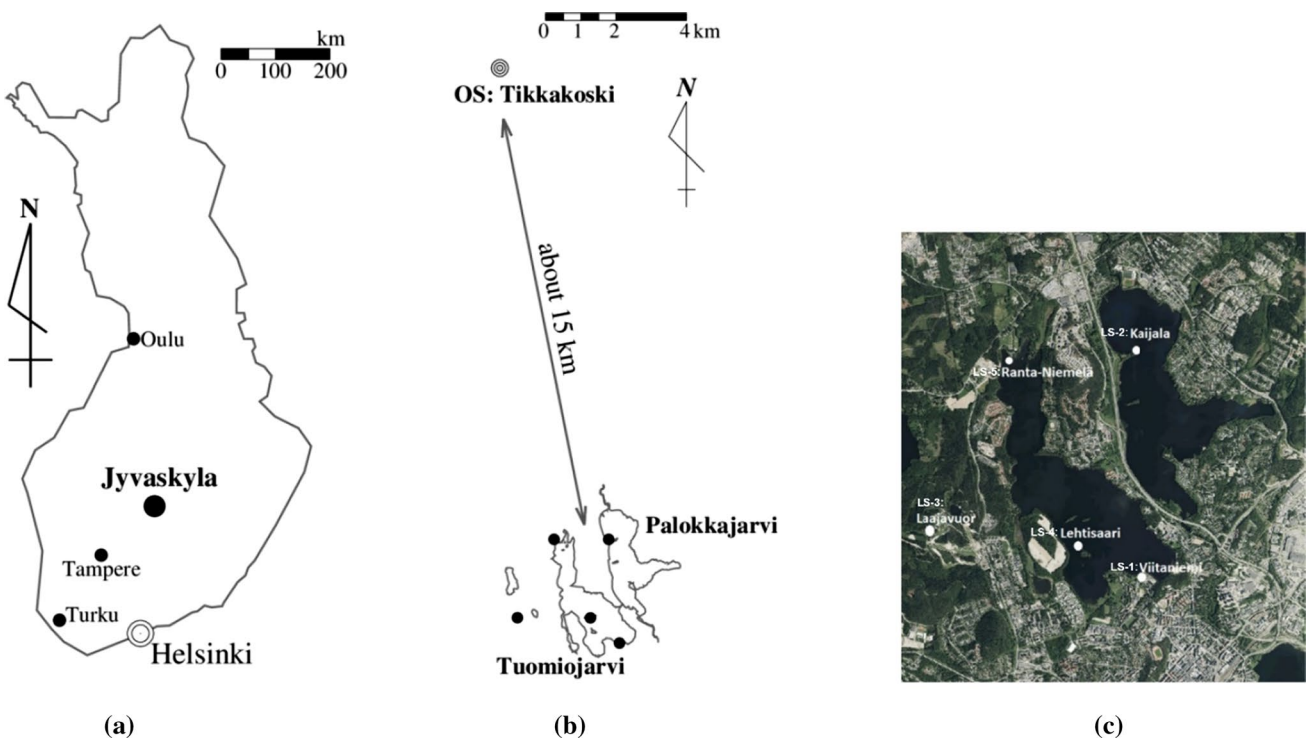


Fig. 1 Target lakes and local observation stations. (**a**) Location in Finland; (**b**) location of the official station; (**c**) target lakes and local observation station

and direction. In Sect. 3, the proposed model's simulations of local wind fields over the two lakes are presented and compared with the corresponding observation data for model validation, and some conclusions are presented in Sect. 4.

2 Materials and methods

2.1 Study area and observation data

The coupled Lakes Tuomiojärvi and Palokkajärvi are located in central Finland in the city of Jyväskylä. Both lakes are classified as small humic lakes. The surface area (A) of Lake Tuomiojärvi is 298 ha and its mean depth (H) is 3.5 m, maximum depth (MD) is 13.1 m, and volume (V) is 10.3×10^6 m^3. The corresponding characteristics for Lake Palokkajärvi are $A=258$ ha, $H=2.8$ m, MD = 10.1 m, and $V=7.2\times10^6$ m^3.

The Finish Environment Institute (Suomen Ympäristö KEskus, SYKE) constructed five temporary local stations around the two lakes and measured wind speed/direction during the summers (from June to October) of 2015 and 2016 to investigate the characteristics of local wind fields. These stations measured meteorological data including wind speed, wind direction, air temperature, relative humidity, and rainfall. Figure 2 shows the locations of the stations on a digital elevation model (DEM). In addition to this local data, meteorological data from Tikkakoski Airport, located approximately 15 km north of the study lakes, were obtained through the Finnish Meteorological Institute (FMI) open data services. The data consisted of wind speed and direction, air temperature, cloudiness, humidity, and rain intensity. Figure 3 shows the time series of wind speed and direction observed at the temporary and official stations in 2015. Local meteorological field stations were installed at and in the vicinity of Lake Tuomiojarvi and Lake Palokkajarvi by the authors during summer 2015 (Fig. 2). The stations were set up to measure wind speed and wind direction at 5 min intervals, and air temperature, relative humidity and rainfall at 15 min intervals. All observed values are 5 min averages. We used data from these stations for our modeling, and data from 2016 were used for validation. More details on the observed data and stations are reported by Juntunen et al. (2019).

2.2 Data-driven wind field model

Our data-driven model was based on wind data observed in 2015, and as shown in Fig. 3, these data are noisy and fluctuate widely. In this study, we developed a statistical model to deal with the noise inherent in the data. All simulation codes on the data-driven model are coded by the authors in Fortran 90.

2.2.1 Wind speed/direction model

This study aimed to develop a model for simulating a local wind field, including both wind speed and direction, over two lakes using official-station data. We assumed that local wind speeds/directions could be estimated by:

$$\left\{ \begin{array}{c} \mathbf{v}_t^{\mathrm{LS}-i} \\ \boldsymbol{\theta}_t^{\mathrm{LS}-i} \end{array} \right\} = \left\{ \begin{array}{c} f_{\mathbf{v}}(v_t^{\mathrm{OS}}, \theta_t^{\mathrm{OS}}) \\ f_{\boldsymbol{\theta}}(v_t^{\mathrm{OS}}, \theta_t^{\mathrm{OS}}) \end{array} \right\}, \quad (1)$$

where $\boldsymbol{v}$ and θ are the wind speed and direction vectors, respectively, that are expressed as $v_t^{\mathrm{LS}-i}$ and $\theta_t^{\mathrm{LS}-i}$(i corresponds to each local station number). The superscripts

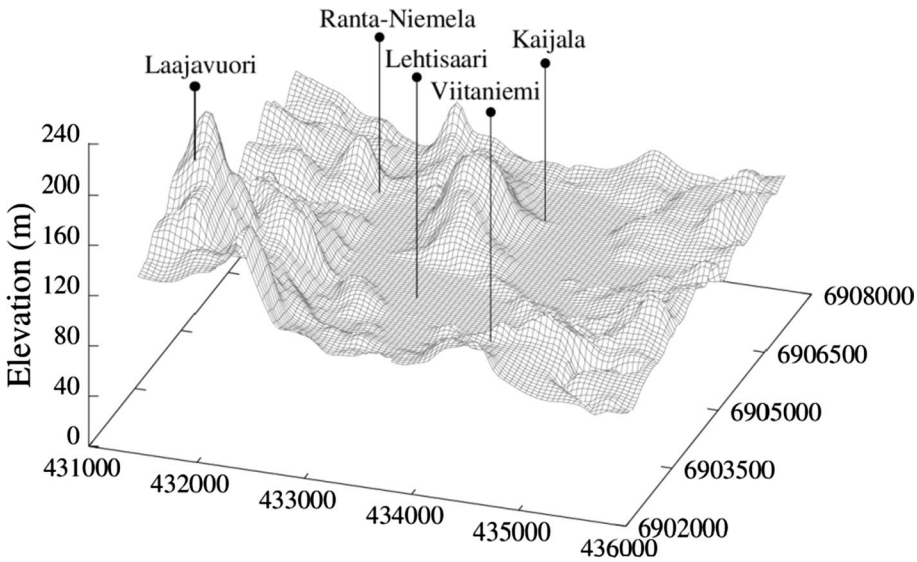


Fig. 2 Digital elevation model of the target area and location of local stations

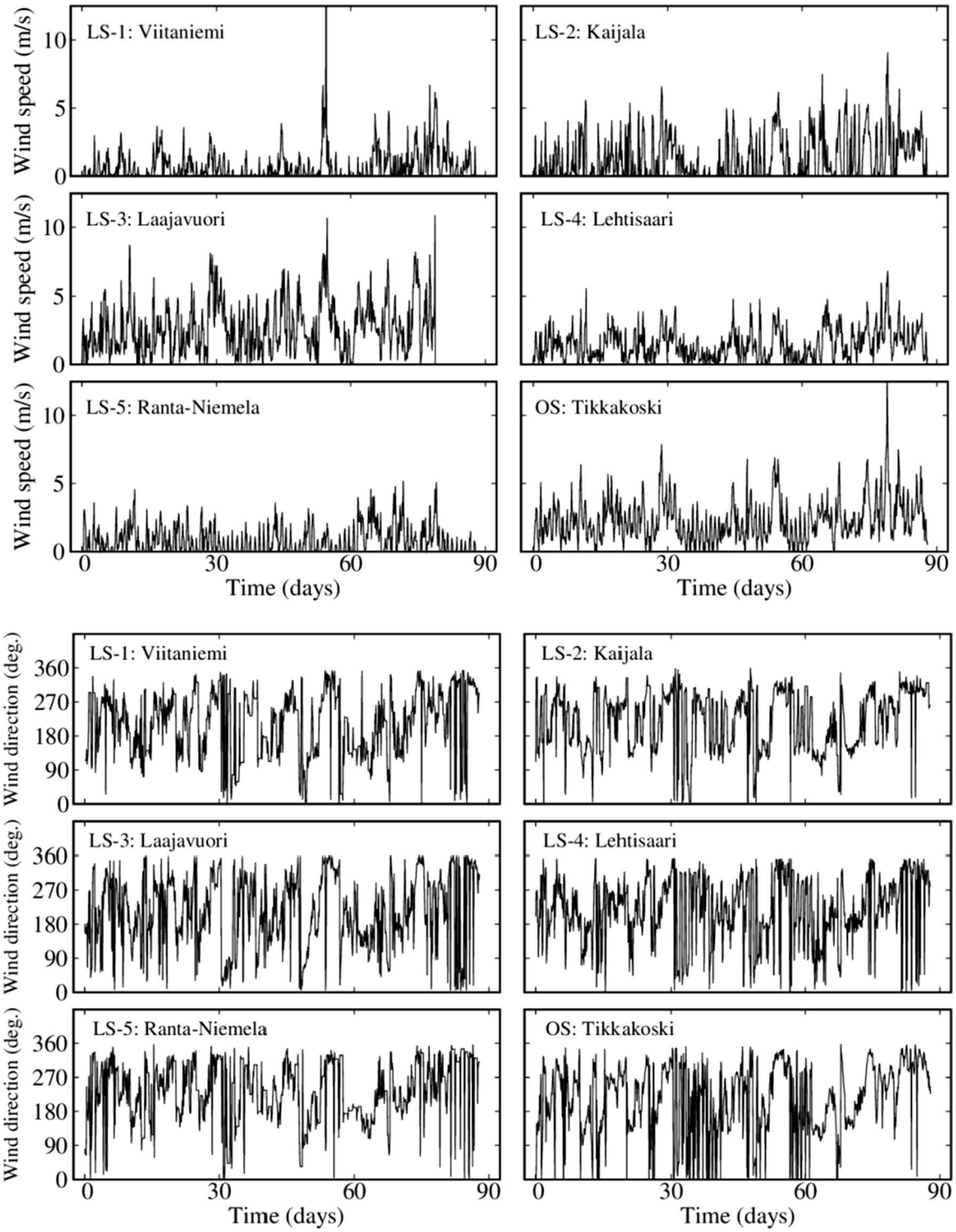


Fig. 3 Wind speed and direction

"LS" and "OS" mean "local station" and "official station," respectively, and $f_{v,\theta}$ is a linear/non-linear operator to estimate local-station data based on official-station data. The specific modeling procedure of $f_{v,\theta}$ is described in the following sections.

2.2.2 Correlation between wind speed and direction

We first investigated the correlation between wind speed and direction to judge whether the correlation structure should be modeled in Eq. (1). The correlation coefficient

Table 1 Correlation coefficient between wind speed and direction by station

Stations	$r_{AL}{}^2$
Viitaniemi (LS-1)	0.08972
Kaijala (LS-2)	0.02412
Laajavuori (LS-3)	0.03350
Lehtisaari (LS-4)	0.01188
Ranta-Niemelä (LS-5)	0.10060
Tikkakoski (OS)	0.03989

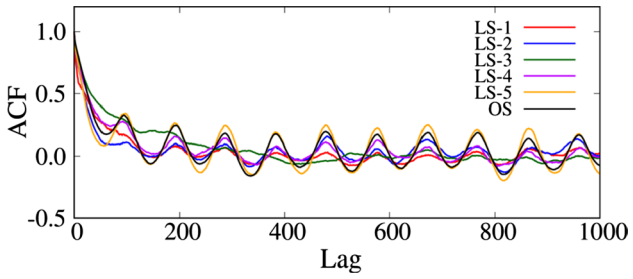


Fig. 4 Autocorrelations of wind speeds (Lag = 15 min)

$r_{AL}{}^2$ proposed by Johnson and Wherly (1977) was used for this purpose. This coefficient was designed to determine the correlation between circular and linear variables. Where $Y' = (Y_1, Y_2, Y_3) = (\cos\theta, \sin\theta, v)$, and s_{ij} is the sample covariance between Y_i and Y_j, the required canonical correlation coefficient between ($\cos\theta$, $\sin\theta$) and v is the positive square root of:

$$r_{AL}^2 = \frac{s_{11}s_{23}^2 + s_{22}s_{13}^2 - 2s_{12}s_{23}s_{31}}{s_{33}(s_{11}s_{22} - s_{12}^2)}. \quad (2)$$

The correlation coefficients for all stations are summarized in Table 1. Since the coefficients are less than 0.1 (except for Ranta-Niemelä station), we assumed that wind speed and direction are independent of each other; thus, the correlation terms could be treated as zero. Although, a correlation between wind speed and direction is not considered in modeling in some research (e.g., El-Fouly et al. 2008), its importance has been discussed by some researchers (e.g., Zhang et al. 2013). Introducing the correlation structure in data-drive modeling remains for future studies.

2.2.3 Wind speed modeling

Figure 4 shows the sample autocorrelation function (ACF) of wind speeds observed at all stations. Since the time resolution of the data is 15 min, 1 Lag corresponds to 15 min. The correlations decrease as Lag increases and become zero around Lag = 200. Most of the curves (except LS-3) have cyclic components with cycles lasting about one day (96 Lag), and these patterns are largely identical. In other words, the cyclic characteristics of locally measured wind speeds are identical to those of the official station's measurements. The data of LS-3, however, show a different pattern from the data of other local stations because LS-3 station is located at the top of a ski jump tower in a mountain (Fig. 2). Although LS-3 shows a slightly different pattern, this study used LS-3 data because the number of available data for the modeling was very limited. The effect of LS-3 data on the modeling will be discussed later. Based on this observation, we assume that local wind speeds can be simply expressed as:

$$v_t^{LS-i} = \gamma_t^{LS-i} v_t^O, \quad (3)$$

where γ_t^{LS-i} is the adjusting parameter for local-station data i (= 1,…, 5) and is calculated as

$$\gamma_t^{LS-i} = v_t^{LS-i} / v_t^O. \quad (4)$$

The histograms of γ_t^{LS-i} for all the local-station data are presented in Fig. 5 and seem to follow log-normal distributions. Figure 6 shows the QQ plot of γ^{LS-i} for all wind-speed data, and it is reasonable to assume that the γ^{LS-i} parameters follow log-normal distributions for wind-speed modeling. The results of Kolmogorov–Smirnov test also proves that the data follows normal distributions (except LS-2 data), and the p values of LS-1, 2, 3, 4, and 5 are 0.748, 2.170, 1.192, 1.641, and 0.919, respectively.

In general, wind speed increases with the height above ground, as friction against the ground is reduced as the height above ground increases. Thus, we developed a simple regression model for γ_t^{LS-i} that is expressed as:

$$\ln \gamma(z) = w_0 + w_1 z + \varepsilon_v, \quad (5)$$

where z is the elevation of the meteorological stations in meters, w_i is the model coefficient, and ε_v is model error according to the Gaussian distribution $N(0, \sigma_\gamma{}^2)$. Although other approaches to regression modeling exist, such as higher order polynomial functions, we attempted to make the model as simple as possible for practical applications and thus employed a linear model. Finally, the wind speed at local station i is given by:

$$v_t^{LS-i} = f_{11}(v_t^{OS}) = \exp(\ln \gamma(z^{LS-i}) + \sigma_\gamma^2 / 2) \times v_t^{OS}, \quad (6)$$

where z^{LS-i} is the elevation.

Since the model coefficients in Eq. (5) were calculated according to a linear-Gaussian model (e.g., Bishop 2006), a closed form solution is available and can be calculated using the ordinary least square method (LSM, e.g., Bishop 2006). Finally, $w_0 = -1.166$, $w_1 = 0.004$ and $\sigma_\gamma = 0.884$ are

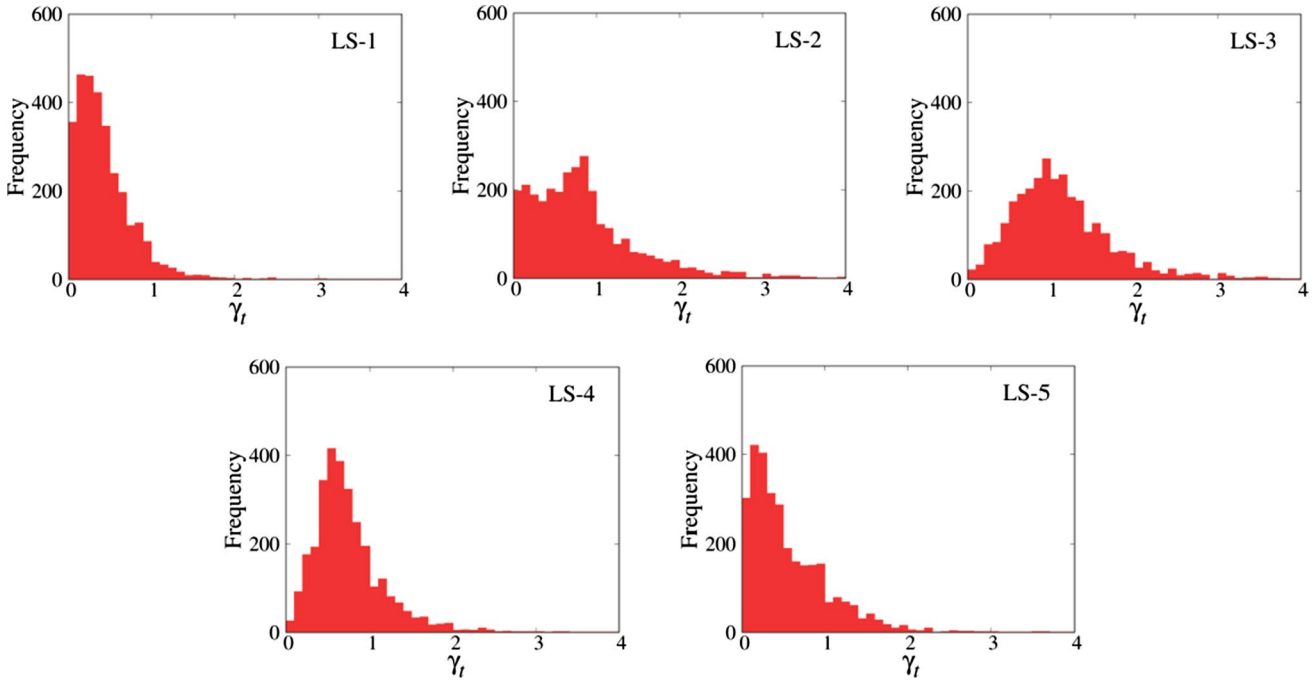


Fig. 5 Histograms of adjusting parameter $\gamma t^{LS\text{-}i}$ for each local station (LS-1–5)

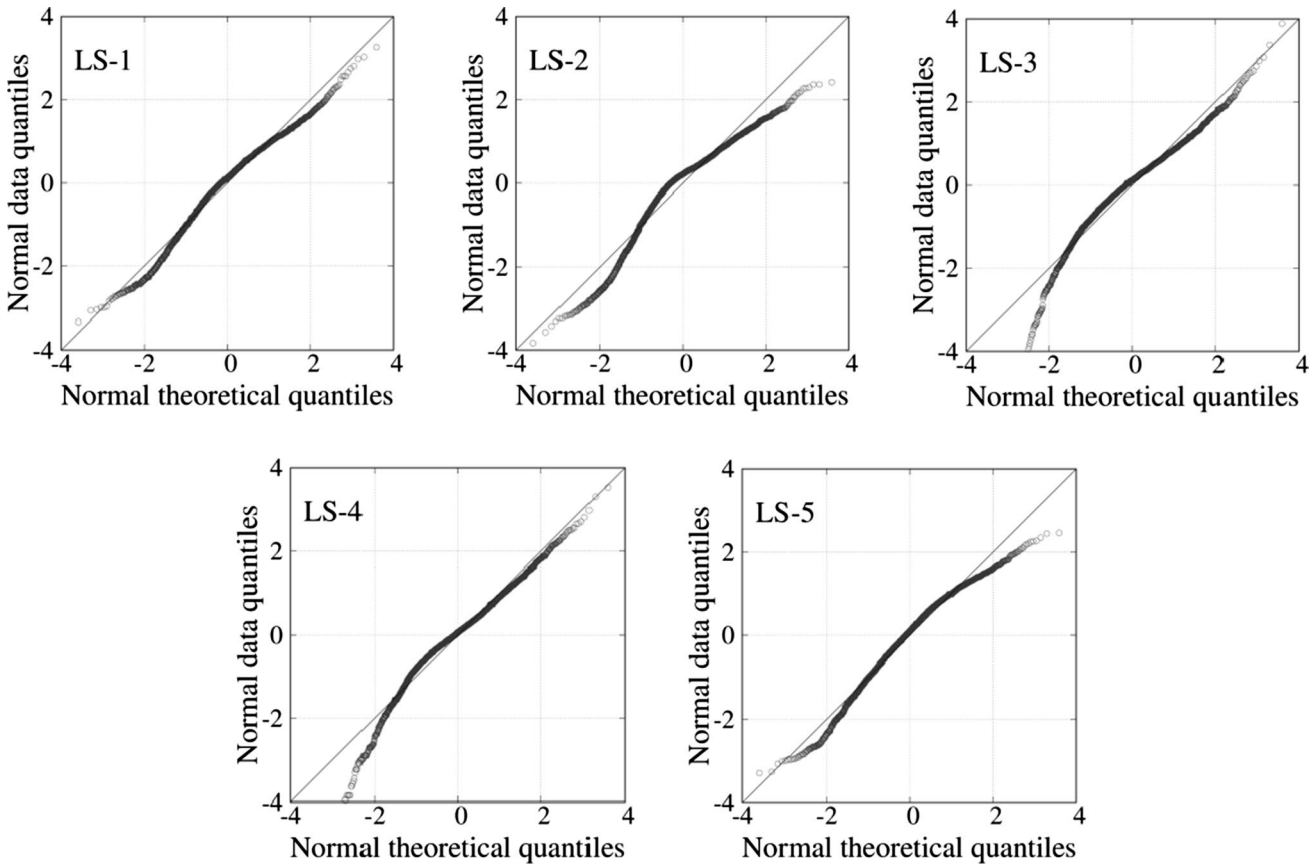


Fig. 6 QQ plot of adjusting parameter $\gamma^{tLS\text{-}i}$ for each local station (LS-1–5)

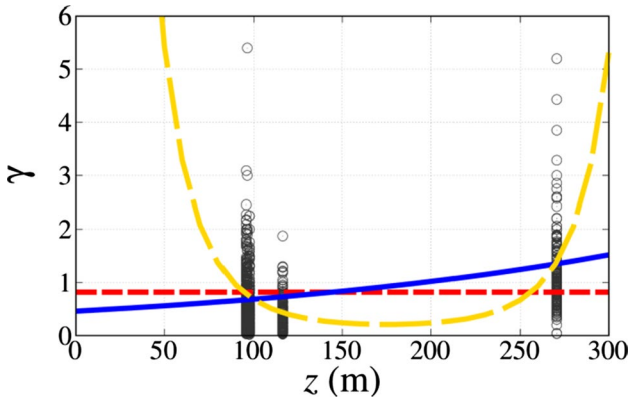


Fig. 7 Regression models for adjusting parameter γ.z = elevation (m)

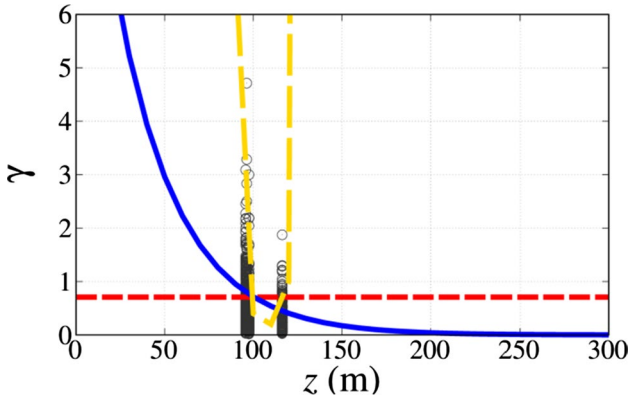


Fig. 8 Effect of LS-3 data on regression models

estimated, and the proposed regression models are shown in Fig. 7 as blue continuous line. A zero-degree polynomial (constant) model (red-dashed line) and a second-degree polynomial model (yellow broken line) are also shown in Fig. 7 for comparison. The constant model naturally cannot express the nature of wind velocity profile, wind speed increases with the height above ground, because the model is too simple. Whereas the second-degree model is too complex and shows an unreasonable relationship between γ and z, γ increases with decreasing height less than $z = 150$ m. Based on this comparison, we can say that the first-degree polynomial function (Eq. (5)) is the simplest and the most reasonable model for wind speed modeling. Used in conjunction with a DEM, the model depends only on elevation (m); therefore, the spatial distribution of wind speeds can be estimated via Eq. (3) using only the official-station data.

The effect of LS-3 data on the regression model was also studied herein. The zero-, first, and second-degree polynomial models for γ were created without LS-3 data and they are shown in Fig. 8. In the first-degree model, the parameter γ decreases with increasing z, and unreasonable relationship is obtained. Although LS-3 data show a different cyclic pattern from the data of other local stations (Fig. 4), it should be used for reasonable wind speed modeling.

2.2.4 Wind direction modeling

Figure 9 presents scatter plots based on the wind directions observed at the official station and five local stations. The wind direction is expressed in radians. Clearly, the wind directions observed at the local stations strongly correlate with the official-station data, and a simple regression model is thus applicable. This study employed the circular–circular regression model proposed by Downs and Mardia (2002):

$$\theta_t^{\mathrm{LS}} = f_{22}(\theta_t^{\mathrm{OS}}) = \beta + 2\mathrm{atan}\left\{\omega \tan(\theta_t^{\mathrm{OS}} - \alpha)/2\right\} + \varepsilon_\theta, \quad (7)$$

where θ^{LS} and θ^{OS} are wind directions at the local and official stations, α and β are angular location parameters, ω is a slope parameter, and ε_θ is the error term. We assumed that ε_θ follows the von Mises distribution with a mean θ^{OS} and a nonnegative concentration parameter κ. The von Mises distribution is known as the simplest probability density function in circular statistics and is defined as

$$p(\theta|\mu,\kappa) = \frac{1}{2\pi I_0(\kappa)} \exp\{\kappa \cos(\theta - \mu)\}, \quad (8)$$

where μ is the mean probability, and κ is the nonnegative concentration parameter (which is analogous to the inverse variance (precision) in the Gaussian distribution). The function I_0 is a zeroth-order Bessel modified function of the first order and is expressed as:

$$I_0(\kappa) = \frac{1}{2\pi} \int_0^{2\pi} \exp\{\kappa \cos\theta\}\, \mathrm{d}\kappa. \quad (9)$$

The circular–circular regression model has four unknown parameters, α, β, ω, and κ. We identified the parameters using the maximum likelihood (ML) method, and the likelihood function is given by:

$$l(\alpha,\beta,\omega,\kappa;\theta_1^{\mathrm{LS}},\ldots,\theta_n^{\mathrm{LS}}) = -n\log I_0(\kappa) + \kappa \sum_{j=1}^{n} \cos\left[\theta_j^{\mathrm{LS}} - \beta - 2\mathrm{atan}\left\{\omega \tan(\theta_j^{\mathrm{OS}} - \alpha)/2\right\}\right]. \quad (10)$$

Maximizing Eq. (10) is a typical global optimization problem with many local maxima. We maximized the likelihood function using the particle swarm optimization (PSO) method (Kennedy and Eberhart 1995), which has been widely used to solve global optimization problems. Other global optimization methods such as genetic algorithms (e.g., Banzhaf et al. 1998) and simulated annealing (e.g., van Laarhoven and Aarts 1987) are also applicable. We

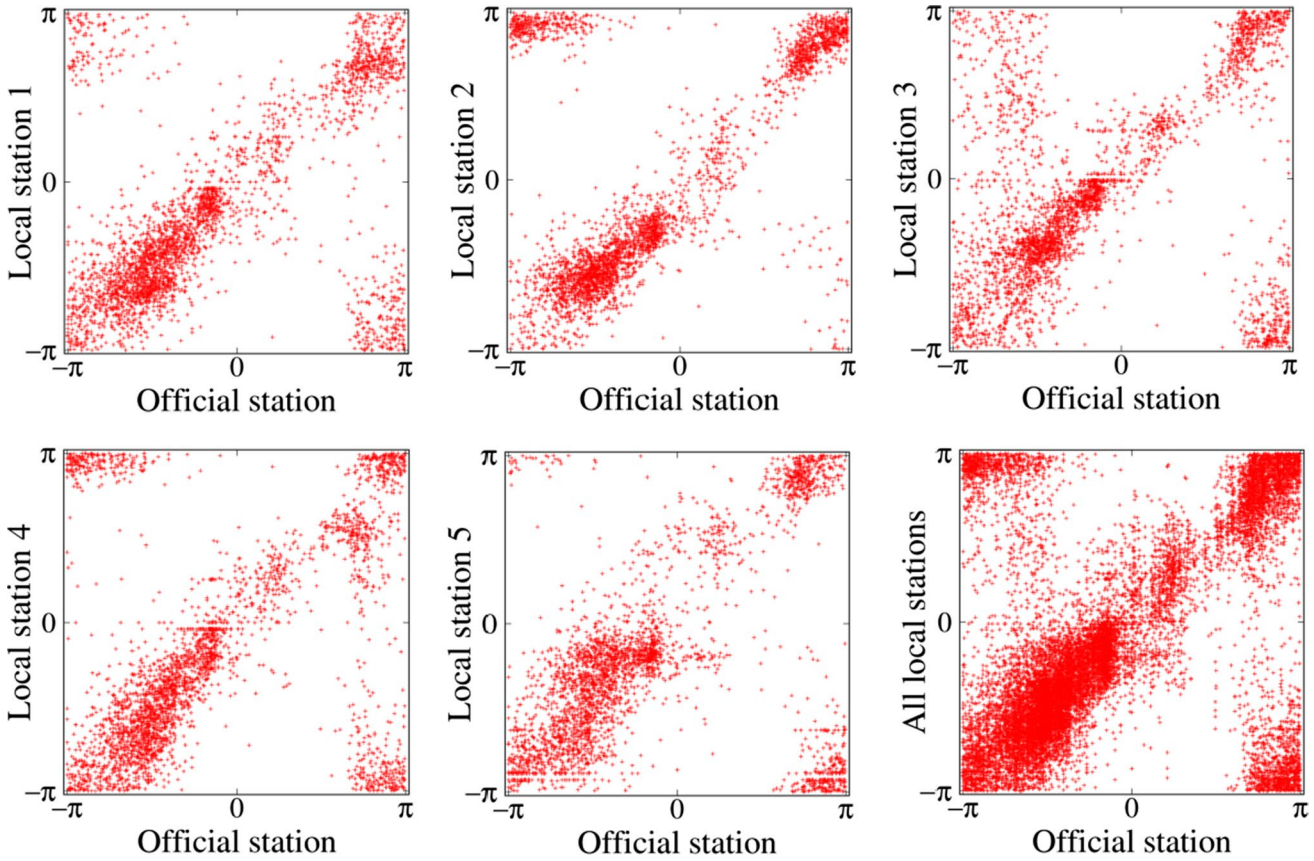


Fig. 9 Scatter plots of wind direction at local and official stations

Table 2 Particle swarm optimization parameters

Parameters	Value
Swarm size, n_s	100
Maximum number of iterations	1000
Cognitive parameter, c_1	2.0
Social parameter, c_2	2.0
Maximum velocity, k_{max}	0.9
Minimum velocity, k_{min}	0.4

performed the PSO with 10 different initial particles using the parameters listed in Table 2.

Figure 10 shows the convergence curve of the four identified parameters. The parameter curves fluctuate until around 200 iterations. After 200 iterations, however, all the parameters converge into constant values. These results imply that this problem has many local maxima, and global optimizers must be used. This result shows that 1000 iterations are sufficient to obtain reliable parameters in this problem. The identified parameters, α, β, ω, κ, are 1.336, 1.285, 1.240, and 2.63, respectively, and the proposed regression model with the identified parameters is shown in Fig. 11.

Several circular regression models and probability density functions have been proposed for wind direction modeling (e.g., Kato and Jones 2010; Shimizu and Wang 2013). When wind direction data are very complex, and the error term does not seem to follow a von Mises distribution, e.g., multimodal distributions or skewed distributions, more advanced models and probability distributions might be necessary.

This section built a wind direction model based on visualization of wind direction data and circular-circular regression analysis. The proposed wind direction model (Fig. 11) seems to reasonably capture the relationship between θ^{LS} and θ^{OS}. The performance of the model will be discussed in the subsequent section.

3 Results and discussion

The proposed model was developed based only on data observed in 2015, and we compared the model predictions of wind speed and direction with the corresponding observed data in 2016 for validation.

Figure 12 compares the 95% and 68% confidence intervals of model predictions with observation data for three

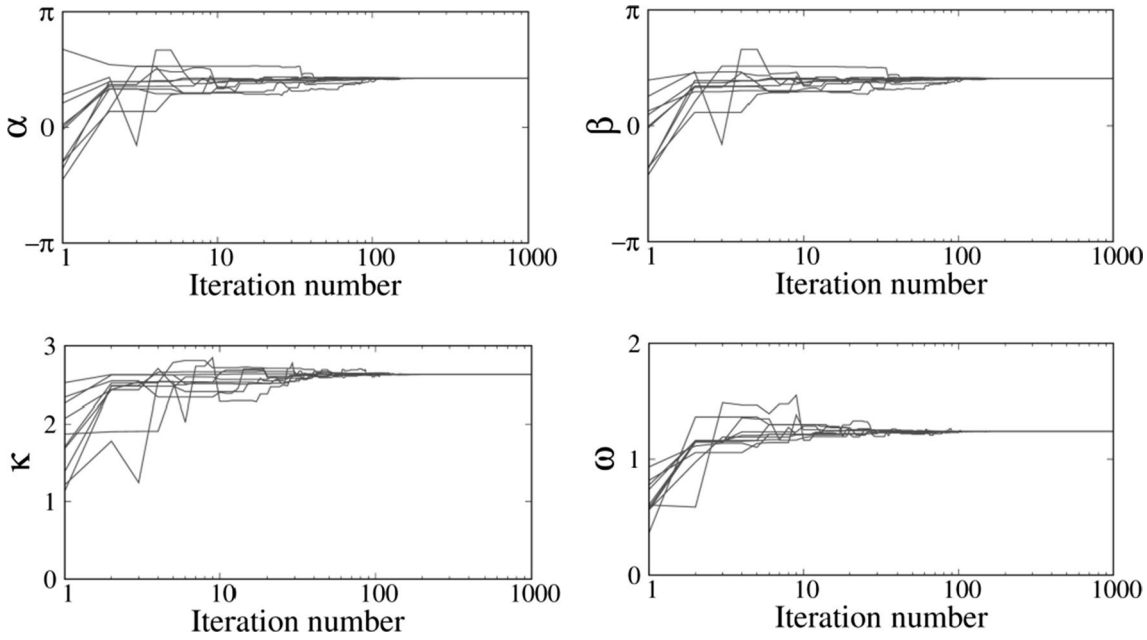


Fig. 10 Likelihood convergence curves for four unknown parameters (α, β, ω, and κ)

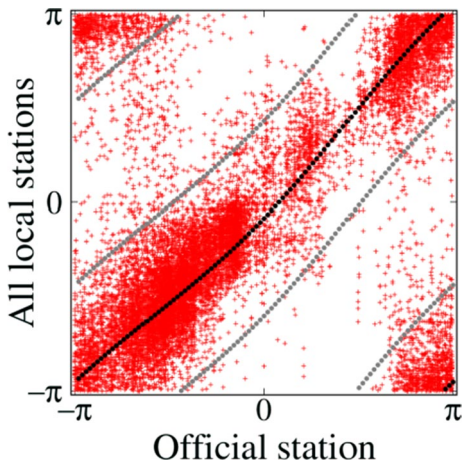


Fig. 11 Circular-regression model with parameter κ

observation stations (2016 data were only available for three observation stations (LS-2, 3, 4)). For both wind speed and direction, the observation data are within the 95% confidence intervals, and the proposed stochastic model can accurately capture the characteristics of time-series fluctuations in both of wind speed and directions. Figure 13 shows the comparison between the confidence intervals and observation data in a shorter period of time (0–9 days). Most of the data are within the 68% confidence intervals, and the predictions by the proposed model are reasonable. In particular, wind direction model well captures the corresponding observation data, and the data are around mode/mean of the estimations.

Since the existing data-driven methods are difficult to apply to this unique problem, and comparison between the proposed method and existing data-driven methods can be unreasonable and misleading. To shed light on the advantage of the proposed model, we compared the proposed model's prediction with the numerical prediction that was performed by SYKE using the Weather Research and Forecasting (WRF) Model. We used the WRF model of Version 3 (Skamarock et al. 2008). The horizontal and vertical mesh resolutions are about 100–200 m and 5–50 m, and time resolution is 5 min. Figure 14 is the comparison between the proposed model, the WRF model and observation data. The simulation results by the WRF model are strongly influenced by the simulation setup, such as mesh resolutions, input physical/empirical parameters, boundary conditions etc. This paper, however, does not detail the setup or how to set it and just focuses on the accuracy and how difference between two models. In the figure, "pro." and "wrf." mean the prediction by the proposed model and the WRF model, respectively. Both of two predictions agree well with the observation data, and the proposed model is as accurate as the numerical simulation. Figure 15 shows prediction–observation plots of wind speed and direction that corresponds to the results shown in Fig. 14. In the figure, white box

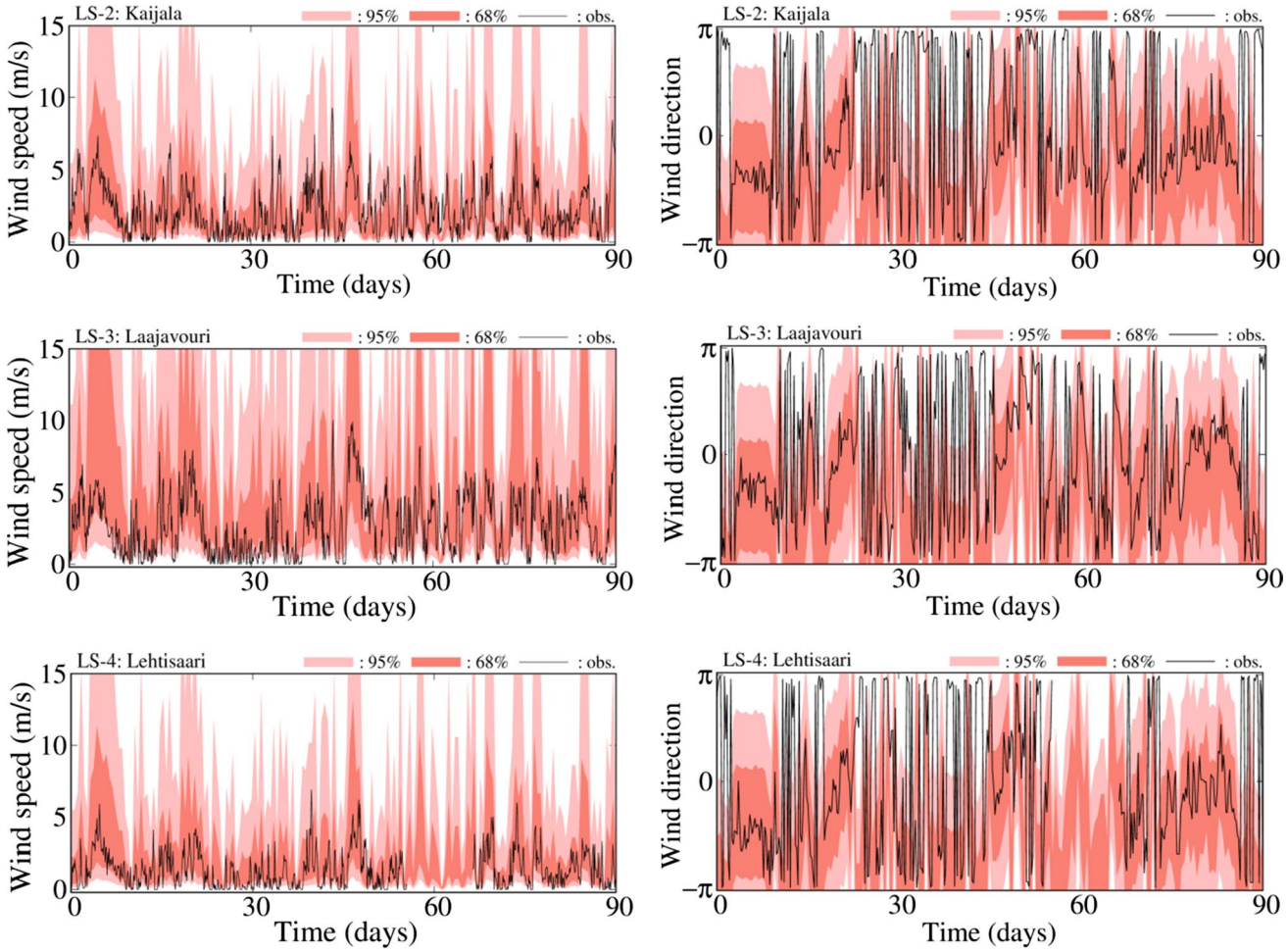


Fig. 12 Comparison of model predictions (68% and 95% confidence intervals) with observation data

indicates the result by the proposed model, red cross-mark indicates the results by the WRF model, and black dashed line indicates 45-degree line, i.e., prediction = observation line. In terms of practical applications, the proposed model can be more advantageous than existing methods because of its simplicity and less computation time. Through the comparisons, we can conclude that the proposed model can provide reasonable simulations of local wind fields around the target lakes using only the official-station data. If CFD is contrasted with data-driven modeling, data-driven modeling is better than CFD in terms of at least computation cost. Once a data-driven model for a target site is made, spatial–temporal behavior of wind speed/direction can be estimated without high computational cost. If wind simulation results are used as an input of lake simulations, computation cost would be a dominant factor in practice.

Though the proposed model performs well, this may be due to a relatively simple wind field around the target area; thus, modeling is relatively easy because the fluctuation patterns in the local- and official-station data are identical. The wind direction model is not a function of elevation like the wind speed model, and making the error term irrelevant in this model. Although the proposed model was designed for the target site, i.e., wind field around two lakes, it can be applied to other site. Wind fields in some area, however, are strongly influenced by geographical (terrain) effect and surface roughness. In that case, parameters describing terrain effects and surface roughness should be considered in modeling. Both of two results are scattered around the 45-degree line, and the WRF results tend to overestimate wind speed. For wind direction, the proposed and WRF models show similar scatter pattern. The Root-Mean-Square Errors

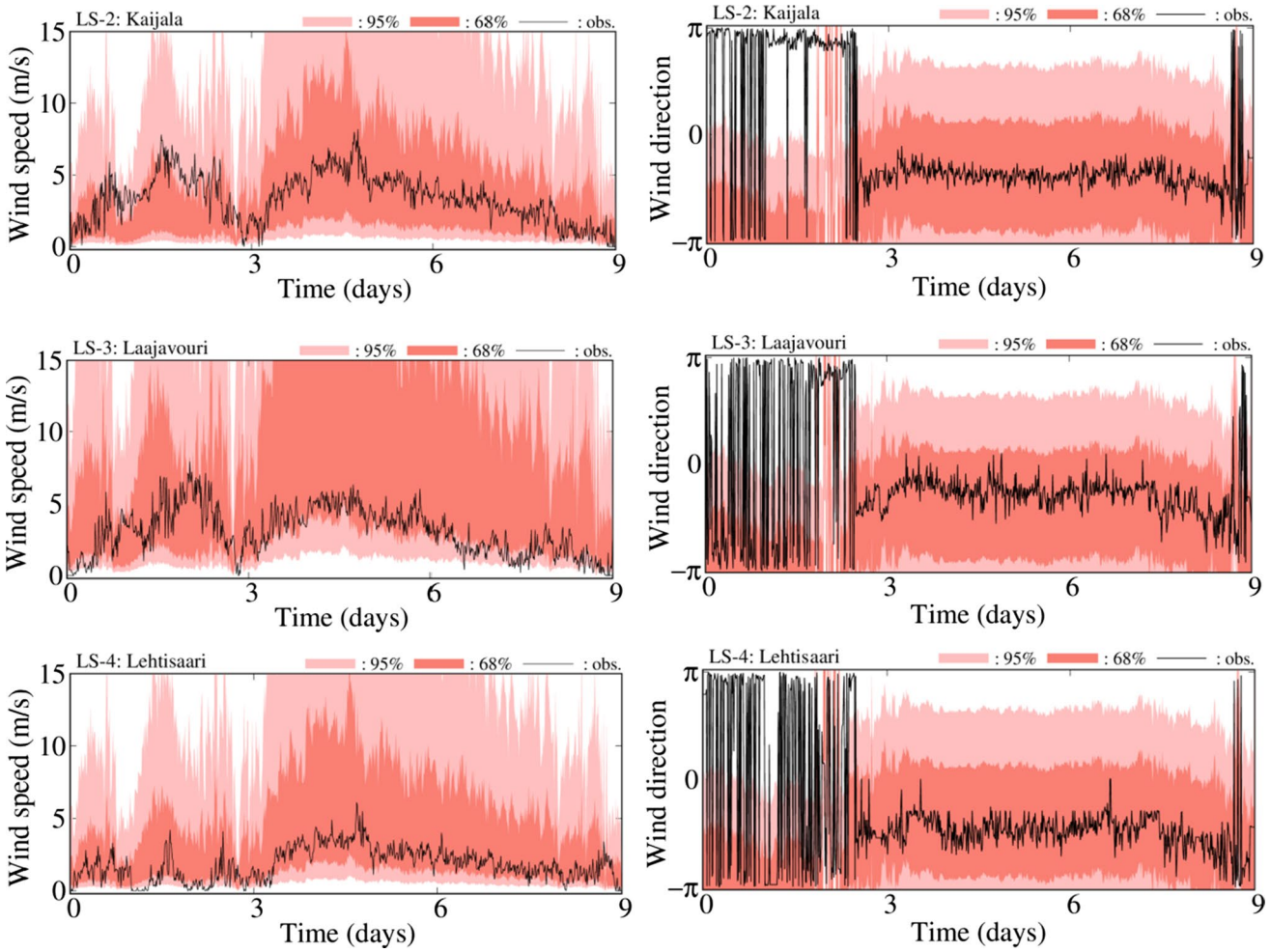


Fig. 13 Comparison of model predictions (68% and 95% confidence intervals) with observation data in a shorter period of time

(RMSEs) are summarized in Table 3. We calculated RMSEs of cos(θ) for wind direction data. Although two models show similar RMSEs for wind direction, the proposed model shows smaller RMSEs than the WRF model for wind speed. The prediction by the proposed model in Figs. 14 and 15 is based on the mean value. If we use other representative values such as most probable values or median, different results from the Figs. 14 and 15 can be obtained.

Figure 16 shows five random realizations of wind speed and direction at LS-4. In the figure, 68% and 95% confidence intervals are also shown for comparison. Most of the random realizations are within 95% confidence intervals. Although the proposed model can generate various fluctuation patterns, it sometimes outputs extremely large wind speeds due to the relatively large model error (standard deviation σ_γ). More complex modeling, e.g., modeling for specific directions (north, south, east, and west), might reduce the error. The advantage of the proposed model is its ability to generate various patterns in the local wind field, and the use of this model for lake circulation simulation could contribute to a better understanding of the lake dynamics and ecosystem.

Figure 17 shows a distribution map of wind vectors over the two lakes; Fig. 14(a) shows the statistical mean results, and (b) shows the results of a random sampling. The proposed model needs only an elevation and the wind speed & direction observed at the official station for prediction, and when a DEM data of the target area and official-station data are available, a distribution map like Fig. 15 can be created. The length of a wind vector indicates the intensity of the wind speed. Clearly, the proposed statistical model can generate heterogeneous wind fields. As noted, the proposed

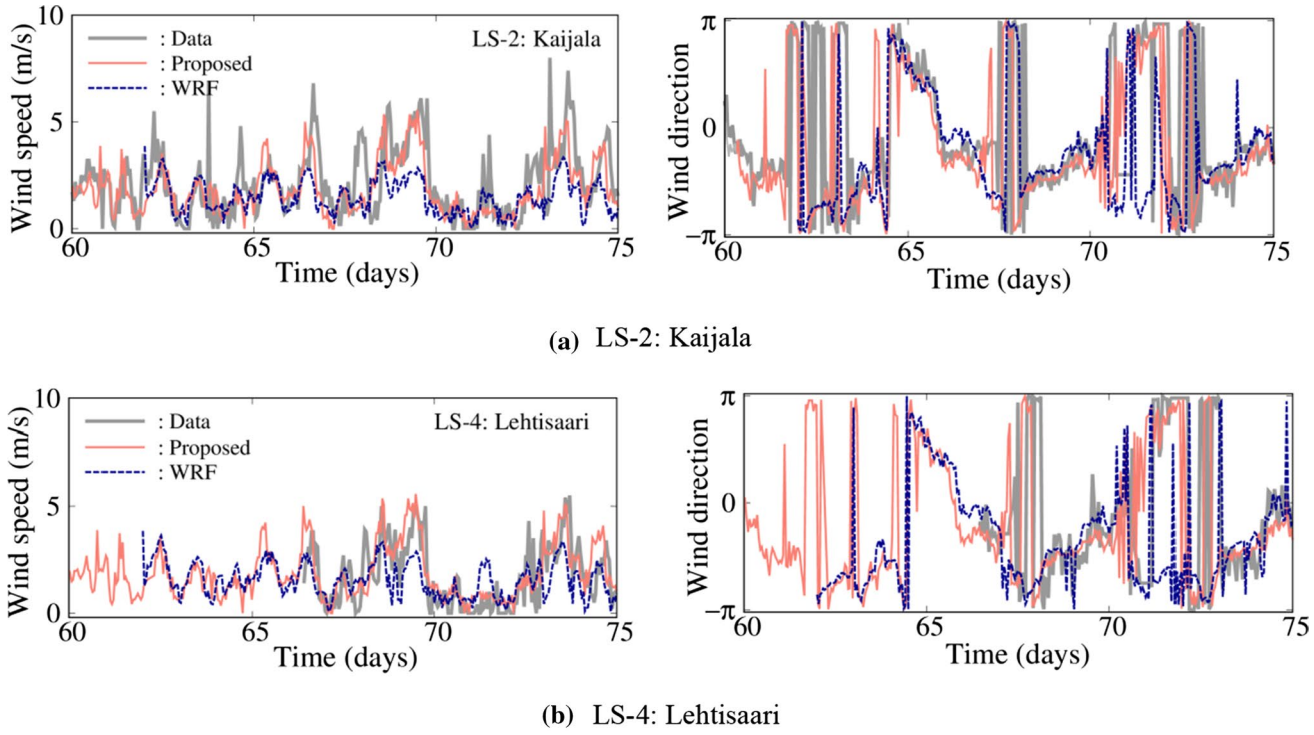


Fig. 14 Comparison of the model predictions and numerical simulation with 2016 observation data

model occasionally outputs extreme wind speeds due to the relatively large model uncertainty (σ_γ). Such extreme values might cause numerical instabilities when this model is used for lake circulation simulations. Further, the influence of high heterogeneity on the modeled lake dynamics remains a topic for future study.

Although characteristics of spatial and temporal variability of wind fields are not explicitly considered in the modeling, these characteristics have been well studied by some researchers (e.g., Kirchner-Bossi 2014; de Paula Gomez-Delgado et al. 2018; Garrido-Perez et al. 2018) and can be used to improve the performance of the proposed data-driven model. In addition, Kirchmeier et al (2014) proposed a method for downscaling of statistical characteristics of wind speed by analyzing large-scale and local-scale data. A downscaling method can be useful when the proposed method is applied to other sites that have smaller or larger area.

4 Conclusions

This study proposed a data-driven model for the local wind field over two small lakes in Jyväskylä, Finland. The findings of this study are summarized as follows:

(1) In the target area, the correlation coefficient between wind speed and direction is low, and we assumed that wind speed and direction are independent in the modeling. This characteristic, however, is site specific, and whether such assumption is reasonable or not must be judged based on the correlation analysis for each target site.
(2) We found that wind speed and direction around two lakes can be simply modeled using the adjusting parameter γ and official-station data through basic statistical analysis and data visualizations.

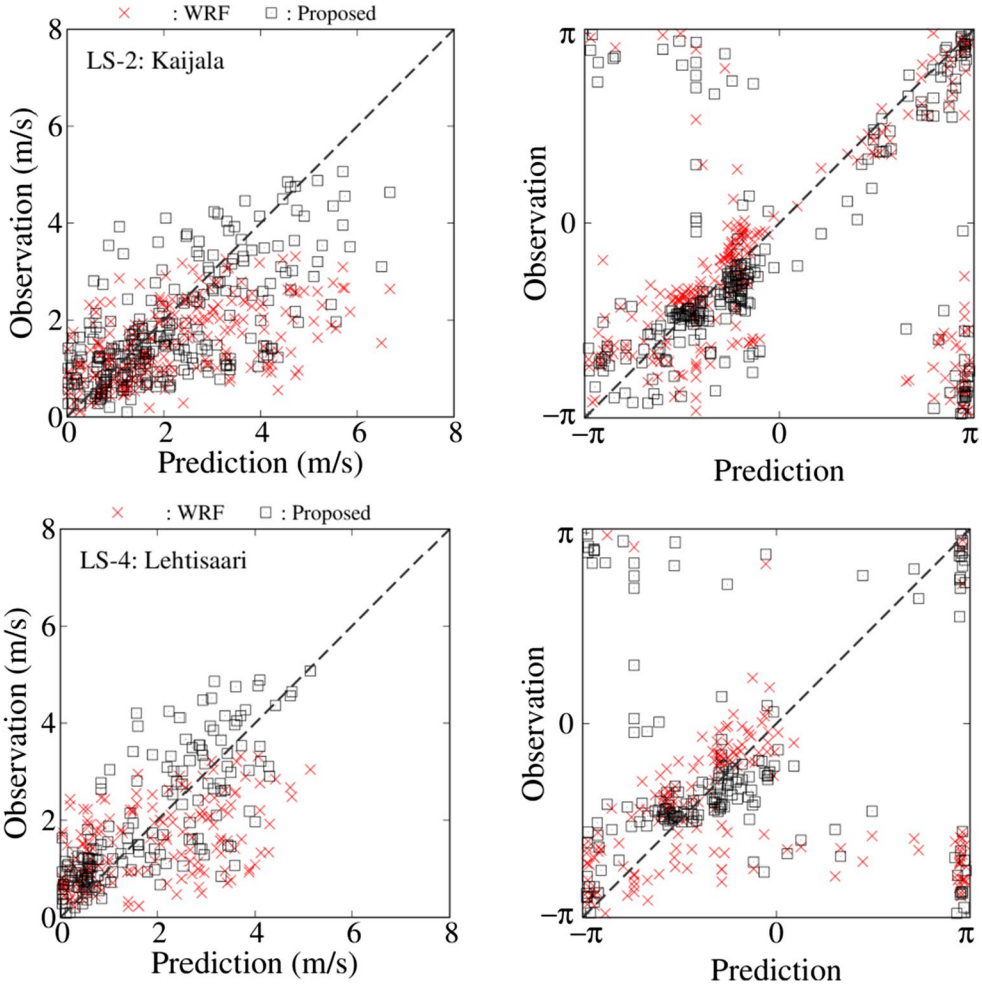


Fig. 15 Comparison of the model predictions and numerical simulation with 2016 observation data

Table 3 RMSEs of the proposed model and WRF model for wind speed and direction

	Proposed model		WRF	
	Speed	Direction	Speed	Direction
LS-2: Kaijala	1.174	0.482	1.429	0.491
LS-4: Lehtisaari	0.908	0.538	1.218	0.540

(3) We designed the wind-speed model using a linear regression model and the wind-direction using a circular–circular regression model. The model predictions were compared with corresponding observed data and a numerical simulation result (WRF) for validation. The predicted results by the proposed data-driven model agreed well with the observations and reasonably captured the fluctuation patterns in the data. In addition, the proposed model outperformed WRF simulation in terms of accuracy (RMSE) and computation cost.

(4) The proposed model is a stochastic model and can easily generate heterogeneous local wind fields.

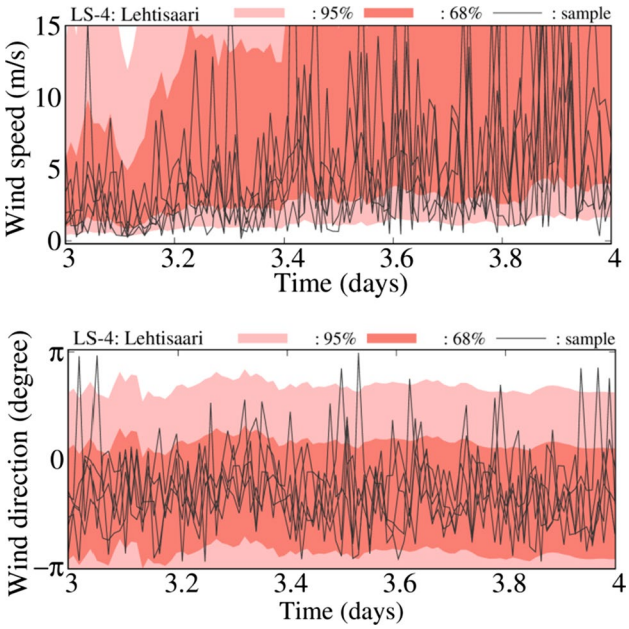
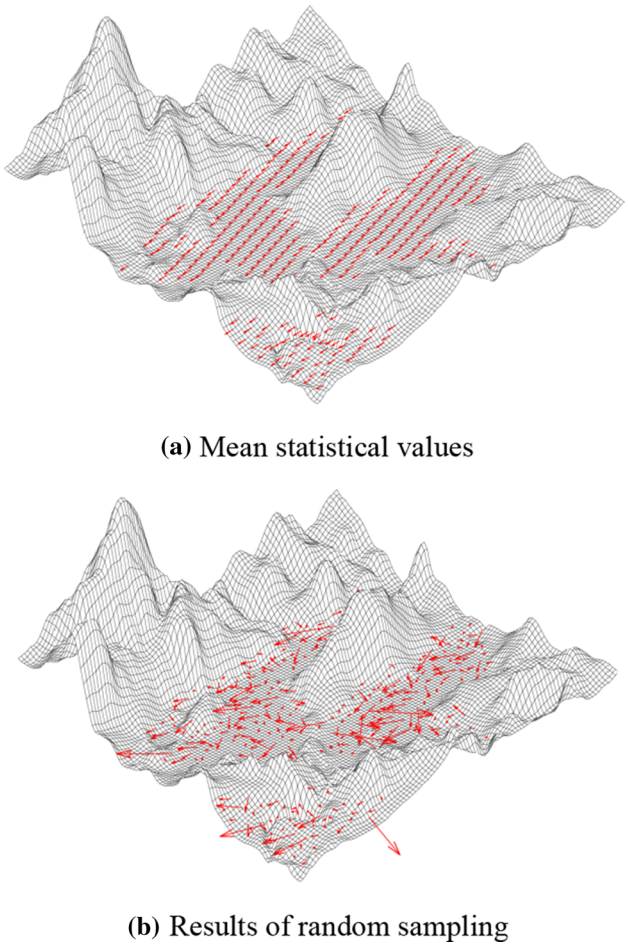


Fig. 16 Random realizations sampled from the proposed model



(a) Mean statistical values

(b) Results of random sampling

Fig. 17 Spatial distributions of wind vectors around the two lakes

Appendix 1

Surface roughness and topography of the target area

Parameters on surface roughness and topography (terrain effects) are not considered in the proposed data-driven modeling because (1) we have attempted to make the model as simple as possible for practical application and (2) spatially dense data are necessary in order to consider terrain effects in the modeling, but they were not available in this study. This Appendix provides some basic information on the surface roughness and topography around the local observation stations to interpret the model performance (why the proposed simple model worked well?) and discuss the applicability of the model in other sites.

The Corin land cover (CLC, Bossard et al. 2000) classes around the target area is shown in Fig. 18, and the CLC classes and their corresponding roughness lengths (Silva et al. 2007) are summarized in Table 4. The target area consists of several types of CLC classes, and the dominant CLC classes are 111: continuous urban fabric and 512: water bodies. Figure 19 shows the Laplacian filtered image (eight neighbors) of the DEM of the target area (Fig. 2). A Laplacian filter is often used to detect edges of the digital image, and we computed second derivatives of the elevation z in DEM (Fig. 2) using the following equation:

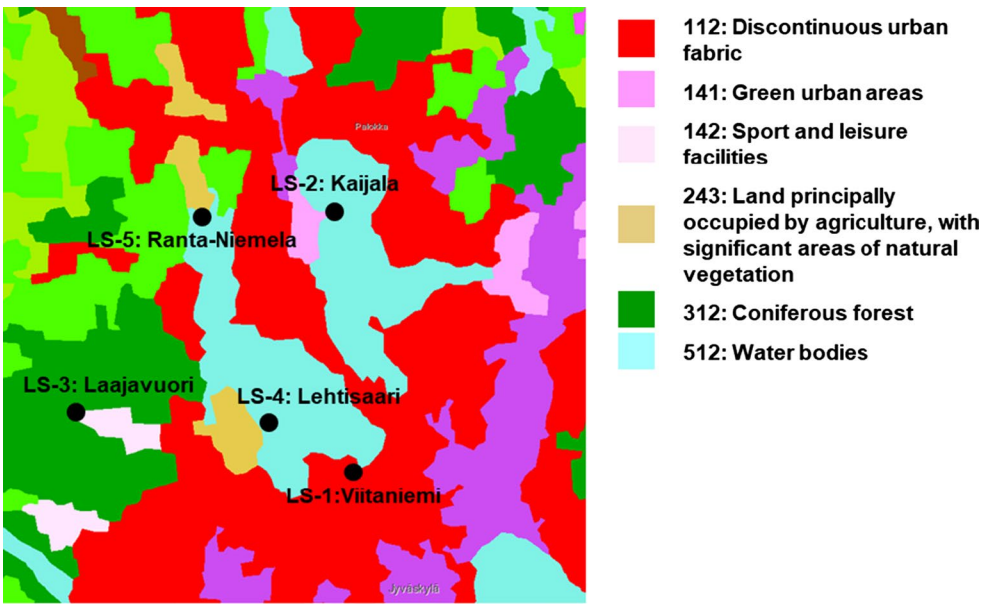


Fig. 18 CLC classes around local stations

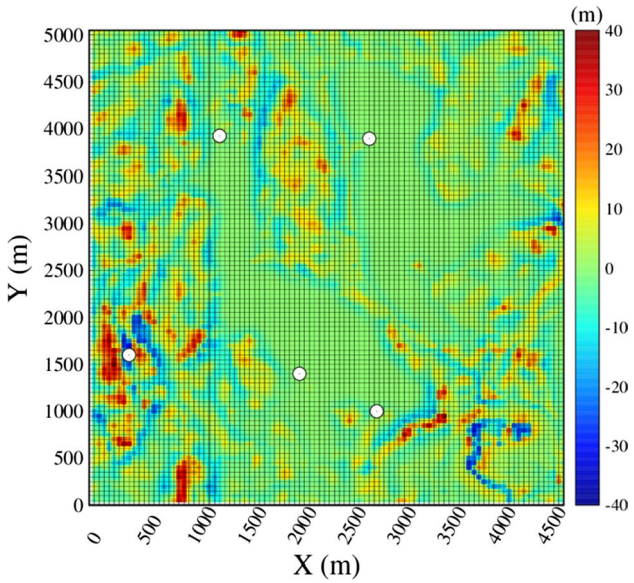


Fig. 19 Laplacian filtered image (eight neighbors) of the DEM of the target area

$$\begin{aligned}\nabla z(x_i, y_j) = 8 \times z(x_i, y_j) - z(x_{i-1}, y_{j-1}) - z(x_{i-1}, y_j) \\ - z(x_{i-1}, y_{j+1}) - z(x_i, y_{j-1}) - z(x_i, y_{j+1}) \\ - z(x_{i+1}, y_{j-1}) - z(x_{i+1}, y_j) - z(x_{i+1}, y_{j+1}).\end{aligned} \quad (11)$$

The effects of surface roughness (or land cover) and topography on wind behavior have been studied, and the importance of these two factors have been reported by many researchers (e.g., Ruel et al. 1998; Tian et al. 2015; Fu et al. 2020). Even though surface roughness and topography parameters were not considered in the modeling, the proposed model can reasonably predict the wind speed and direction over two lakes reasonably. The possible causes for this includes (1) the dominant CLC classes around the observation stations are 111 and 512 (Fig. 18) except LS-3, and there is no notable differences of wind behavior between stations, (2) the topography of the target area is not so complex (dominant color in the Laplacian filtered image is green), and terrain effect is very limited. Therefore, the proposed model may not work well in the area that have very complex land cover distribution and topography (e.g., deep valleys with large elevation differences).

Table 4 CLC classes and roughness lengths around local stations

Station	CLC class	Roughness length (m)
Viitaniemi (LS-1)	112 Discontinuous urban fabric	0.5
Kaijala (LS-2)	141 Green urban areas	0.6
Laajavuori (LS-3)	142, 312 Sports and leisure facilities, Coniferous forest	0.5, 0.6
Lehtisaari (LS-4)	243 Land principally occupied by agriculture, with significant areas of natural vegetation	0.3
Ranta-Niemelä (LS-5)	243 Land principally occupied by agriculture, with significant areas of natural vegetation	0.3

512: water bodies, roughness length = 0

Acknowledgements We would like to thank Dr. Timo Huttula for his valuable perspectives and advice on this study. We would like to also thank Editage (www.editage.com) for English language editing.

Author contributions TS: conceptualization, methodology, software, writing—original draft. JR: writing—review and editing, investigation. JJ: writing—review and editing, investigation. HS: writing—review and editing, supervision.

Funding This work was supported by a Japan Society for the Promotion of Science KAKENHI grant [grant number: 18K03408].

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Code availability All the codes used in this study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest There are no conflicts of interest to declare.

References

Bachmann RW, Hoyer MV, Canfield DE (2000) The potential for wave disturbance in shallow Florida lakes. Lake Reserv Manage 16(4):281–291

Banzhaf W, Nordin P, Keller RE, Francone FD (1998) Genetic programming: an introduction on the automatic evolution of computer programs and its application. Morgan Kaufmann Publishers, Inc., Burlington, p 398

Bengtsson L, Hellstrom T (1992) Wind-induced resuspension in a small shallow lake. Hydrobiologia 241:163–172

Bessac J, Ailliot P, Monbet V (2015) Gaussian linear state-space model for wind fields in the north-east Atlantic. Environmetrics 26:29–38

Bishiop CM (2006) Pattern recognition and machine learning. Springer, Berlin, p 738

Bossard M, Feranec J, Otahel J (2000) CORINE land cover technical guide—Addendum 2000. EEA Technical report No 40: 105p.

Chao JY, Zhang YM, Kong M, Zhuang W, Wang LM, Shao KQ (2017) Long-term moderate wind induced sediment resuspension meeting phosphorus demand of phytoplankton in the large shallow eutrophic Lake Taihu. PLoS ONE 12(3):e0173477

de Paula G-D, Gallego D, Pena-Ortiz C, Vega I, Ribera P, Garcia-Herrera R (2018) Long term variability of the northerly winds over the Eastern Mediterranean as seen from historical wind observations. Glob Planet Change 172:355–364

Downs TD, Mardia KV (2002) Circular regression. Biometrika 89(3):683–697

El-Fouly THM, El-Saadany E, Salama MMA (2008) One day ahead prediction of wind speed and direction. IEEE Trans Energy Conv 23(1):191–201

Ferragut L, Ascensio MI, Simon J (2011) High definition local adjustment model of 3D wind fields performing only 2D computations. Int J Num Meth Biomed Eng 27:510–523

Fu D, Liu Y, Li H, Liu S, Li B, Thapa S et al (2020) Evaluating the impacts of land cover and soil texture changes on simulated surface wind and temperature. Earth Space Sci 7:e2020EA01173

Garrido-Perez JM, Ordonez C, Gracia-Herrera R, Barriopedro D (2018) Air stagnation in Europe: spatiotemporal variability and impact on air quality. Sci Total Environ 645:1238–1252

Johnson RA, Wehrly TE (1977) Measures and models for angular correlation and angular-linear correlation. J Royal Stat Soc 39:222–229

Juntunen J, Ropponen J, Shuku T, Krogerus K, Huttula T (2019) The effect of local wind field on water circulation and dispersion of imaginary tracers in two small connected lakes. J Hydrol 579:124137

Kato S, Jones MC (2010) A family of distributions on the circle with links to, and applications arising from Möbius transformation. J Amer Stat Assoc 105:249–262

Kennedy J, Eberhart R (1995) Particle swarm optimization. Proc IEEE Int Conf Neural Networks, 1942–1948.

Kinrchner-Bossi N (2014) Centennial simulation of wind power output through soft computing algorithm, Doctoral Thesis, Universidad Complutense De Madrid, Madrid, 144p.

Kirchmeier MC, Lorenz DJ, Vimont DJ (2014) Statistical downscaling of daily wind speed variations. J Appl Meteorol Climatol 53:660–675

Kitada T, Okamura K, Tanaka S (1998) Effects of topography and urbanization on local winds and thermal environment in the Nohbi Plain, coastal region of central Japan: a numerical analysis by mesoscale meteorological model with a k–ε turbulence model. J Appl Meteorology 37:1026–1046

Laird N, Walsh JE (2003) Model simulations examining the relationship of lake-effect morphology to lake shape, wind direction, and wind speed. Mon Weather Rev 131:2102–2111

Martinez-Garcia FP, Contreras-de-Villar A, Munoz-Perez JJ (2021) Review ow wind model at a local scale: advantages and disadvantages. J Mar Sci Eng. https://doi.org/10.3390/jmse9030318

Podsetchine V, Schernewski G (1999) The influence of spatial wind inhomogeneity on flow patterns in a small lake. Wat Res 33(15):3348–3356

Ratto CF, Festa R, Romeo C, Frumento OA, Galluzi M (1994) Mass-consistent models for wind fields over complex terrain: the state of the art. Env Soft 9:247–268

Robert S, Foresti L, Kanevski M (2013) Spatial prediction of monthly wind speeds in complex terrain with adaptive general regression neural networks. Int J Climatol 33:1793–1804

Ruel JC, Pin D, Cooper K (1998) Effect of topography on wind behavior in a complex terrain. Forestry 71(3):261265

Shimizu K, Wang M (2013) Use of directional statistics in environmental science. Proc Inst Stat Math 61(2):289–305

Silva J, Ribeiro C, Guedes R (2007) Roughness length classification of Corine Land Cover classes. Proc Eur Wind Energy Conf 710:1–10

Skamarock WC, Klemp JB, Dudhia J, Gill DO, Barker DM, Duda MG, Huang XY, Wang W, Powers JG (2008) A description of the advanced research WRF Version 3. NCAR Tech. Note NCAR/TN-475+STR, 113. https://doi.org/10.5065/D68S4MVH

Tian W, Ozbay A, Hu H (2015) Terrain effects on characteristics of surface wind and wind turbin. Proc Eng 126:542–548

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Royal Statist Soc B 58(1):267–288

Torma P, Kramer T (2017) Wind shear stress interpolation over lake surface from routine weather data considering the IBL development. Period Polytech Civil Eng 61(1):14–26

van Laarhoven PJM, Aarts EHL (1987) Simulated annealing: theory and applications. Kluwer Academic Publisher, Dordrecht, p 198

Zhang J, Chowdhury S, Messac A, Castillo L (2013) A multivariate and multimodal wind distribution model. Renew Ener 51:436–447