Renewable Energy 202 (2023) 1215–1225

Contents lists available at ScienceDirect

Renewable Energy

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

Offshore wind resource assessment based on scarce spatio-temporal measurements using matrix factorization

Basem Elshafei [a,*], Alfredo Peña [c], Atanas Popov [a], Donald Giddings [a], Jie Ren [b], Dong Xu [b], Xuerui Mao [b]

[a] Faculty of Engineering, University of Nottingham, NG7 2RD, Nottingham, UK
[b] Beijing Institute of Technology — BIT, Beijing, PR China
[c] DTU Wind and Energy Systems, Technical University of Denmark, Roskilde, Denmark

ARTICLE INFO

Keywords:
Matrix factorization
Gaussian process regression
Spatiotemporal data fusion
Wind resource assessment

ABSTRACT

In the pre-construction of wind farms, wind resource assessment is of paramount importance. Measurements by lidars are a source of high-fidelity data. However, they are expensive and sparse in space and time. Contrarily, Weather Research and Forecasting models generate continuous data with relatively low fidelity. We propose a hybrid approach combining measurements and output from numerical simulations for the assessment of offshore wind. Firstly, the datasets were fed onto a matrix, with columns representing the spatial lidar and WRF points, and the rows representing the time steps. Entries of the matrix reflect the wind speed, empty entries represent unobserved data. Then, matrix factorization using Gaussian process was employed for filling the missing entries with statistically calculated estimates. The model was optimized with stochastic gradient descent to apply GP without approximation methods. To evaluate the method, wind speed data along the coast of Denmark were used. The proposed technique, evaluated using two experiments, resulted in 58% more accurate results than the industrial standard method with trivial increase of computational cost. The RMSE of the proposed method ranges between 0.35 and 0.52 m/s.

1. Introduction

Over the past decade, the energy industry has seen great changes due to a worldwide demand for sustainable energy. Clean energy is recognised as the pathway for a sustainable future, which leads to a dramatic expansion and an increase in renewable energy capacity, with a rise in global investments. In 2021, the global wind power market added 60 GW to its arsenal, the second largest wind power annual increase, reaching a total of 743 GW for both onshore and offshore sites [1]. In 2021, the UK wind energy accounted for an estimated 24.8% of electricity generation. As of 2020, the UK has a total set onshore record of 10.2 TWh and offshore of 9.2 TWh [2]. In addition, Europe intends to increase the demand for wind energy and its capacity by 35% within the next decade [3].

Evaluating the wind speed condition of a potential location is a critical early step before the construction of any wind farm. As minimal changes in speed can drastically have large deviations in the power output [4], and as the wind varies both geographically and temporally over a wide range of scales, an accurate wind resource assessment is essential and is considered of a paramount significance for a successful wind energy project [5]. Moreover, the assessment provides aid to the selection of wind turbines, their layouts, and for planning a wind project, which wind power developers use to estimate the future energy production of a wind farm to meet their demand [6].

Instruments that measure wind can yield accurate observations of the wind speed but can be expensive, and the data are generally sparse. This equipment includes lidars, which measure the line-of-sight (LOS) velocity by computing the Doppler shift of the signal of an infrared laser based on the movement of aerosols. However the lidar output is usually intermittent with large unavailability at offshore locations. Contrarily, numerical weather prediction (NWP) models offer output that covers large geographical areas and long-time horizons simultaneously and continuously, but the data are of a significantly lower fidelity [7].

Measurement instruments and numerical simulations complement each other, suggesting that hybrid data fusion techniques can be used to combine their merits. It is desirable to extend the information from coastal vertical lidars (wind profilers) for the reconstruction of offshore time series, as they are easier to maintain [8]. Information can be numerically extended from coastal measurements to offshore time series

* Corresponding author.
E-mail address: basem.elshafei2@nottingham.ac.uk (B. Elshafei).

https://doi.org/10.1016/j.renene.2022.12.006
Received 24 September 2022; Received in revised form 1 December 2022; Accepted 3 December 2022
Available online 8 December 2022

Nomenclature

R	$N \times M$ Data matrix
$Ri,$	*ith* row of **R**
$R_{:,j}$	*j*th column of **R**
N	Number of row in **R**
M	Number of columns in **R**
D	Dimension of the latent factors
$U\ N \times D$	Latent matrix for rows of **R**
$V\ M \times D$	Latent matrix for columns of **R**
$u_{i,}$	Latent factors for $R_{i,:}$
$v_{:,j}$	Latent factors for $R_{:,j}$

at lower cost and higher accuracy, compared to having complete data dependent on lidars at all the positions. This technique has been widely used in predictions of future developments based on various inputs [9]. Wind resource assessment is commonly required to cover large areas, over a long-time-interval (e.g. a year or a number of years) and spatio-temporal fusion of physically measured and numerically estimated wind [10].

Missing value estimation is a significant problem in many research areas, including recommender systems [11], geostatistics [12], and image restoration [13]. In most cases, the cost of acquiring high fidelity data or repeating an experiment, due to low availability, is high. Therefore, filling out missing data is the method of preference [14]. Most missing value estimation approaches include, but are not limited to, clustering algorithms and probabilistic matrix factorization (PMF) [15,16]. Matrix factorization techniques proved to be superior to clustering methods because the former allows incorporation of additional information [17]. In its basic form, PMF factorizes a matrix to find two lower rank matrices such that their dot product is the original matrix. After factorizing the partially observed matrix, each row and each column are assigned a latent vector, and the estimation of the missing cell becomes the inner product of the latent vectors for the corresponding row and the corresponding column. PMF characterizes both time steps and spatial points by vectors of factors inferred from point time series patterns. High correspondence between factors leads to an estimation. These methods are becoming increasingly popular by combining good scalability with predictive accuracy [18].

Based on available data, expectation maximization (EM) and maximum likelihood (ML) are two very common methods of estimating parameters of missing data. ML approaches the missing value estimation by finding the underlying probability distribution of the available data. Due to the sparseness of the dataset, it is imperative that both methods follow an iterative approach in estimating the missing values. The first step is to estimate the parameters of interest from the available data and the probable value of the missing data. Parameters are then recalculated using the available data along with estimates from the first round, and new parameters are applied to re-estimate the missing values, and so forth. This process is repeated until the estimated data have a high correlation with that of the previous cycle [15]. In the case of stochastic intermittent data, a mixture of Gaussian methods is a more suitable probability distribution method.

In machine learning, factorization-based methods are a well established and powerful technique for analysing data for matrix completion. A probabilistic framework for matrix factorization, was presented in Ref. [19], which was integrated to a fully Bayesian model later [20]. The model scales linearly with the number of observations in the original matrix and performed well on the Netflix dataset, where the rows represented users, the columns represented the items and the data is review-based. The model included adaptive prior on the model parameters and showed how the model capacity can be controlled automatically. MF was generalised to a full Bayesian model in Refs. [21,20], which incorporated multiple sources of side information and combined multiple priori estimates for the missing data using real-world drug--target interaction datasets. Additionally, Agathokleous and Tsapatsoulis [15] inspected the Voting Advice Application (VAA) data from the Cypriot presidential elections to estimate missing data for the party and candidate recommender system using several collaborative filtering methods. In this paper, we apply the idea of missing data completion in a matrix, in the wind industry sector.

The novelty of this work relies on the utilisation and deployment of a nonlinear probabilistic matrix factorization model with Gaussian process algorithm for the accurate assessment of offshore wind resources with reduced cost and high accuracy, testing multiple points both spatially and temporally. It combines the generally continuous but low-fidelity numerical data and high-fidelity but limited physical measurements. Efforts are also devoted to pre-processing the time series, taking into account additional information not considered in existing methods to lift the accuracy of the fusion. This algorithm enables the projection of limited nearshore measurements to offshore locations in light of numerical simulations and limited lidar measurements with significantly higher accuracy than the industry standard approach. The application uses tests for the prediction of wind speed in two space dimensions across time using multiple lidar, WRF, and the pre-processed data inputs, to deliver a model capable of performing spatial and temporal predictions in a single step. The model feeds 48 lidar points and 15 WRF spatial points along 12,960 time steps. The algorithm considers all neighboring data in the space and time domains for the prediction of every missing entry.

2. Methodology

2.1. Pre-processing: empirical wavelet transform

Empirical wavelet transform (EWT) [22] is an algorithm developed to process non-stationary time series. The algorithm can extract meaningful information from a given series by designing an appropriate wavelet filter bank, which decomposes the time series into a signal and additional residual components, resulting in a filter of the non-stationary signal. In the present work, the process starts by identifying and extracting the different intrinsic modes of the wind time series, by relying on robust preprocessing for peak detection. Then, spectrum segmentation is performed based on the detected maxima, hence constructing a corresponding wavelet filter bank. In this study, the EWT algorithm was used to preprocess 15 grid points, from output of a WRF-based numerical simulation. The process of extracting meaningful information from the signal was the first stage in building a forecasting model as shown in the matrix factorization flowchart in Fig. 2.

2.2. Multivariate Gaussian process regression

A Gaussian process $GP(m(t),k(t,s))$ is determined by a mean function $m(t)$ and a covariance function $k(t,s)$. Contrarily, a multivariate Gaussian is determined by a mean vector and a covariance matrix, $GP(m(t),k((t,s),(t',s')))$, where the algorithm considers the high-fidelity data, $f_h(t)$, in the multivariate set as a function of two variables (t,s), and s is the low-fidelity dataset, $f_l(t)$, $f_h(t) = g(t,f_l(t))$.

The Gaussian process works well for temporal data fusion and one dimensional time series predictions. However, in geostatistics and specifically in wind resource assessments, it is necessary to predict multiple points at different locations, and hence spatial extrapolation is required along temporal extrapolation. In previous studies [23], neural networks were trained to connect the time series of two spatial points to supply data that use measured time steps interpolated with WRF simulations and onshore lidar measurements to predict wind at locations offshore without lidar measurement. However, for this study, we experimented with 16 spatial points distributed across 4 km in three months at three different heights.

2.3. Non-linear probabilistic matrix factorization with Gaussian process

In this study, a non-linear probabilistic matrix factorization model was employed to map both space and time to a joint latent factor space of dimensionality, *F*, such that space-time interactions are modelled as inner products in that space. For the large and sparse matrices in this work, stochastic gradient descent was used to optimize the Gaussian process, which successfully handles large-scale and sparse machine learning problems, and the parameters are learned using maximum likelihood [16].

The results provided direct predictions of wind speed across multiple spatial points located at three different planes, representing three different heights, at any given time (See Fig. 1 for a graphical illustration). Previous forecasting models focused mainly on predicting the wind speed either spatially, by including multiple spatial points for a single timestep, which is known as point prediction, or temporally by interval predictions where the target is to forecast the wind speed for a specific spatial point for a specified time interval, dealing with a single measured wind speed time series.

2.3.1. Probabilistic matrix factorization (PMF)

For a dataset with *N* spatial points and *M* time steps, the matrix is considered as $\mathbf{R} \in N \times M$. The objective is to obtain a lower rank factorized form of $\mathbf{R}$,$\mathbf{R} = \mathbf{U}^{\top}\mathbf{V}$, where $\mathbf{U} \in \mathbf{R}^{D\times N}$ and $\mathbf{V} \in \mathbf{R}^{D\times M}$. The process of collecting the data of the rows and columns is the second stage in the flowchart of the matrix factorization process, this is shown as the data structuring process of the flow chart in Fig. 2. In this study, for the latent factor, *D*, two values were experimented, two and five. Predictions can then be performed on missing entries by estimating **(U,V)** from the training data and computing the resulting approximation to **R**. This is graphically demonstrated in Fig. 1 and is the third stage in the matrix factorization flowchart, where the missing entries are identified for predictions in Fifure 2. PMF favours a probabilistic perspective to solve the problem from the matrix factorization aspect. Let $R_{i,j}$ represent the wind speed of time-step *i* (time) for lidar point *j* (space), and $u_{i,:}$ and $v_{:,j}$, denoting the time-specific and space-specific latent feature vectors respectively:

1. For each row *i* in **R**, $[i]_1^N$, generate $u_{i,:} \sim \mathrm{N}(0,\sigma_U^2\,\mathbf{I})$, where **I** denotes the identity matrix.
2. For each column *j* in **R**, $[j]_1^M$, generate $v_{j,:} \sim \mathrm{N}(0,\sigma_V^2\,\mathbf{I})$. 3. For the non-missing cells (*i,j*), generate $R_{i,j} \sim N(u_{i,:}v_{j,:}^T, \sigma_2)$ where $\mathrm{N}(\mu,\sigma^2)$ represents the probability density function of the Gaussian distribution with mean μ, variance σ^2. A conditional distribution is defined over the observed wind speeds as:

$$p(\boldsymbol{R}|\boldsymbol{U},\boldsymbol{V},\sigma 2) = \prod_{ij}^{NM}\left[N\left(R_{ij}|U\top i,:V:,j,\sigma 2\right)\right]I_{ij} \tag{1}$$

where $^{\mathrm{Q}}$ is the product operator of a sequence, and I_{ij} is the indicator function that can either be equal to 1 when the matrix cell has a wind speed measurement or 0 otherwise. In PMF, matrix **R** is modelled as a low rank matrix with noise corruption, where the matrix factorization, $\mathbf{U}^{\top}\mathbf{V}$, is the mean of the distribution and the noise is Gaussian with variance σ^2.Thus, zero mean spherical Gaussian priors are placed on time and space feature vectors:

$$p(\boldsymbol{U}|\sigma U2\,) = Q_{Ni=1}N(ui,:|0,\sigma U2\,\boldsymbol{I}), p(\boldsymbol{V}|\sigma V2) = \prod_{j=1}^{N} YN\ (V:,|0,\sigma V2\,\boldsymbol{I}) \tag{2}$$

Then assume that both the space and time latent feature vector and the product latent feature vector obey the Gaussian prior distribution with zero mean. The log-posterior distribution over the latent matrices **U** and **V** is given by:

$$\begin{aligned} log\, p\left(U,V\middle|R,\sigma^2,\sigma_U^2,\sigma_V^2\right) = &-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{j=1}^{M}I_{ij}\left(R_{ij}-u_{i,:}v_{:,j}^{T}\right)^2 \\ &-\frac{1}{2\sigma_U^2}\sum_{i=1}^{N}u_{i,}^{T}u_{i,}-\frac{1}{2\sigma_V^2}\sum^{M}v_{:,j}^{T}v_{:,j} \\ &-\frac{1}{2}\left(A\ log\ \sigma^2+ND\ log\ \sigma_U^2+MD\ log\ \sigma_V^2\right)+C., \end{aligned} \tag{3}$$

where constant *C* does not depend on the latent parameters and is generated after expanding the function to collect all constants for the log of scaling factor by multiplying the coefficient variables.

2.3.2. Optimizer: stochastic gradient descent

Ideally, the marginal likelihood of the model would be calculated, but in practice this is not tractable. Instead, maximum a posteriori, MAP, inference maximizes the logarithmic likelihood with respect to **U** and **V**, which is equivalent to minimizing the sum of squared error function with quadratic regularization terms:

$$\begin{aligned} E=&\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M}I_{ij}\left(R_{ij}-u_{i,:}^{T}v_{:,j}\right)^2 \\ &+\frac{\lambda_U}{2}\sum_{i=1}^{N}\left|\left|u_{i,:}\right|\right|_F^2+\frac{\lambda_V}{2}\sum_{j=1}^{M}\left|\left|v_{:,j}\right|\right|_F^2 \end{aligned} \tag{4}$$

where $\lambda_U = \sigma^2/\sigma_U^2$, $\lambda_V = \sigma^2/\sigma_V^2$, and $\|.\|_F^2$ denotes the Frobenius norm. Performing gradient descent in **U** and **V** will give a local minimum of the objective function.

For each column of the latent matrices, the prior distribution is a zero-mean Gaussian process, which is a generalization of the multivariate Gaussian distribution.

The selection of likelihood is based entirely on the one with fewer parameters. In this case there were less columns since they represented the space domain, rather than the rows that represented the time steps. As using EM for a large matrix would be highly computationally expensive, it made sense to consider SGD, which converges much faster.

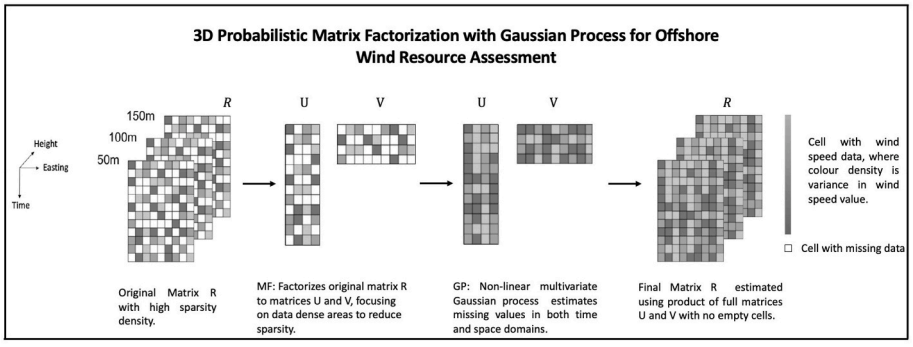

Fig. 1. Flow chart for 3D spatio-temporal probabilistic matrix factorization for wind resource assessment, where the amount of missing data is reduced and WRF data is fused with lidar measurements to estimate missing data.

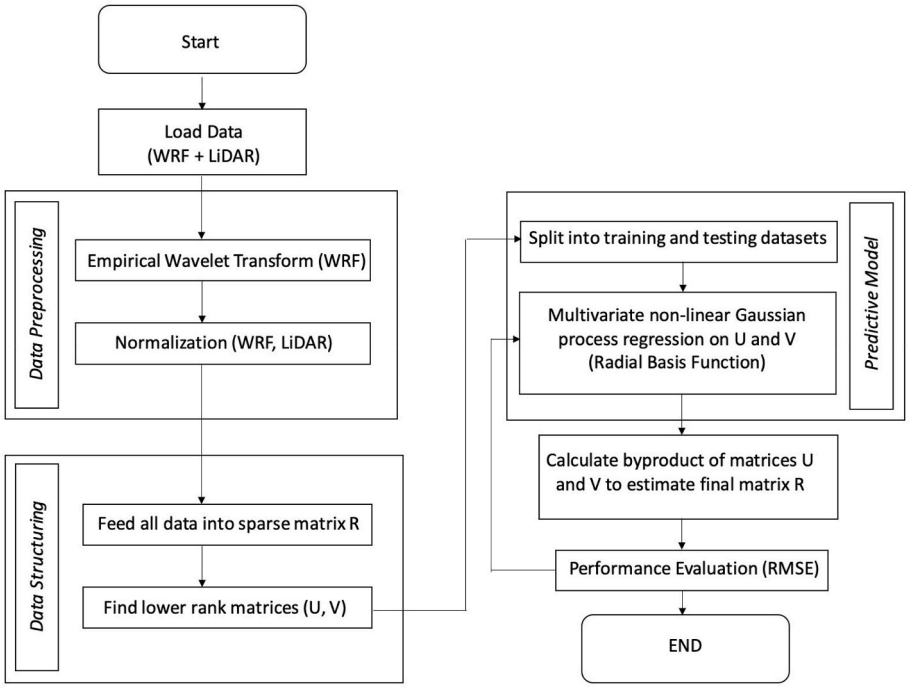

Fig. 2. Proposed matrix factorization with EWT preprocessing and Gaussian process algorithms flowchart for 3D wind speed predictions.

2.3.3. The model: non-linear PMF via GP-LVMs

A probabilistic matrix factorization with parameters marginalized belongs to a category of models called Gaussian process latent variable models (GP-LVM). The Gaussian process is a linear model that can be transferred nonlinear by replacing the inner product matrix by a Mercer kernel. Consequently, maximizing over the logarithmic likelihood can no longer be attained through an eigenvalue problem; but SGD in the manner described above is straightforward.

The regression model followed in Eq. (1) can be written in the form of a product of univariate Gaussian distributions, *g*,

$$p(\boldsymbol{R}|\boldsymbol{U},\boldsymbol{V},\sigma 2) = \prod_{i=1}^{ND}\prod_{j=1}^{YY\,N}\left(R_{ij}|ui,\top:v:j\right),(\sigma 2)I_{ij} \tag{5}$$

A Gaussian process with latent variable models can be recovered by recognising the placed prior distribution directly over the function through a Gaussian process. In a Gaussian process, the mean and covariance function specify the model, where the joint distribution for any given set of the function, *f*, is Gaussian. For a zero mean function, the distribution is N(**f**|0,**K**), where **K** represents the covariance function made up of elements, $k(\mathbf{x}_{i,:},\mathbf{x}_{j,})$, which represent the correlation between the two samples, $f_l(t)$ (low-fidelity), and $f_h(t)$ (high-fidelity) from **f** as a function of inputs associated with the samples $\mathbf{x}_{i,:}$ and $\mathbf{x}_{j,:}$.

A commonly used covariance function that gives a prior over nonlinear functions is known as the radial basis function (RBF) covariance,

$$k\left(X_{i,:},X_{j,:}\right)=\sigma^2\ exp\left(-\frac{1}{2}\left\|X_{i,:}-X_{j,:}\right\|^2\right)., \tag{6}$$

which can be substituted directly in the likelihood to give a probabilistic model, where parameters of the covariance function are presented in hyperparameters.

Following learning based on the SGD provides an estimate of the latent matrices **U** and **V**, where for a missing cell, R_{ij}, the maximum likelihood becomes the inner product of the corresponding latent vectors. This corresponds to the last stage in the matrix factorization flowchart in Fig. 2, where the byproduct of the inner matrices is calculated to estimate the final matrix.

3. Test case description

Two numerical experiments were run on the data acquired from the dualDoppler scans of the RUNE experiment [24](see Fig. 3). The scans were acquired between the period from December 2015 until March 2016. The dual-Doppler scans (36 in total per height measured) were performed with two scanning lidars (positions 1 and 3), which were configured to match their scanning patterns at three heights 50, 100, and 150 m above mean sea level (all heights are referred to this level unless otherwise stated). One 'virtual line', i.e. a line perpendicular to the coast was scanned in about 45 s. In particular, given a time-space matrix with missing entries, the goal is to predict the missing wind speeds at the unobserved time-steps by using both high-fidelity sparse lidar measurements and continuous simulation output from the WRF model (see Fig. 4 for a time series plot of lidar measurements and WRF simulations, and Figs. 5 and 6 for the distribution of the observations across the experiment period). Fig. 6 panel (a), shows the number of observations for every week of the experiment, while panel (b) shows the histogram for the number of observations. For all experiments, the data were partitioned as follows: 20% of the dataset for the validation, 20% for testing and 60% for training.

In addition, Fig. 7 demonstrates the wind speed data correlation between all lidar and WRF time series, reflecting that the scale of the weather patterns was similar to the scale of the measured volume. The correlations ranged from 0.80 to 0.99, indicating high correlation between all the sets, which reflects that even far points can be used to influence the prediction of any point in the matrix, and the strong relation between all the data points. High correlation between the independent features and dependent variable is a good indication that accurate estimations could be yielded.

The dual-Doppler points used were at the far most offshore 4-km range. In the first experiment, an evaluation on the effect of increasing the observations in the matrix and varying the lidar points by using three different matrices with different number of lidar points was tested. For each height, there were four dualDoppler points per km, here referred to as a lidar point. There were three matrices available: matrix A had the first 2 lidar points (from the west) per km, which led to a total of 8 new lidar points; matrix B had the first 3 lidar points per km leading to a total

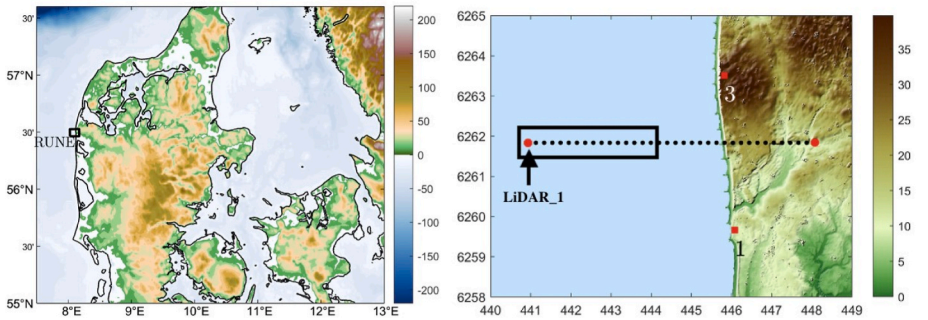

Fig. 3. (Left) RUNE experimental area (in the rectangle) in western Denmark. (Right) RUNE coastal experimental area on a digital model of the surface (UTM32 WGS84, Zone 32V). The positions of the lidars (1 and 3) are shown in squares and of the dual-Doppler scans (36) in black markers. The scans are taken at three heights, 50, 100 and 150 m. The 16 dual-Doppler scans used in this study are present in the black box. Colorbars indicated the height in meters above mean sea level [23,24].

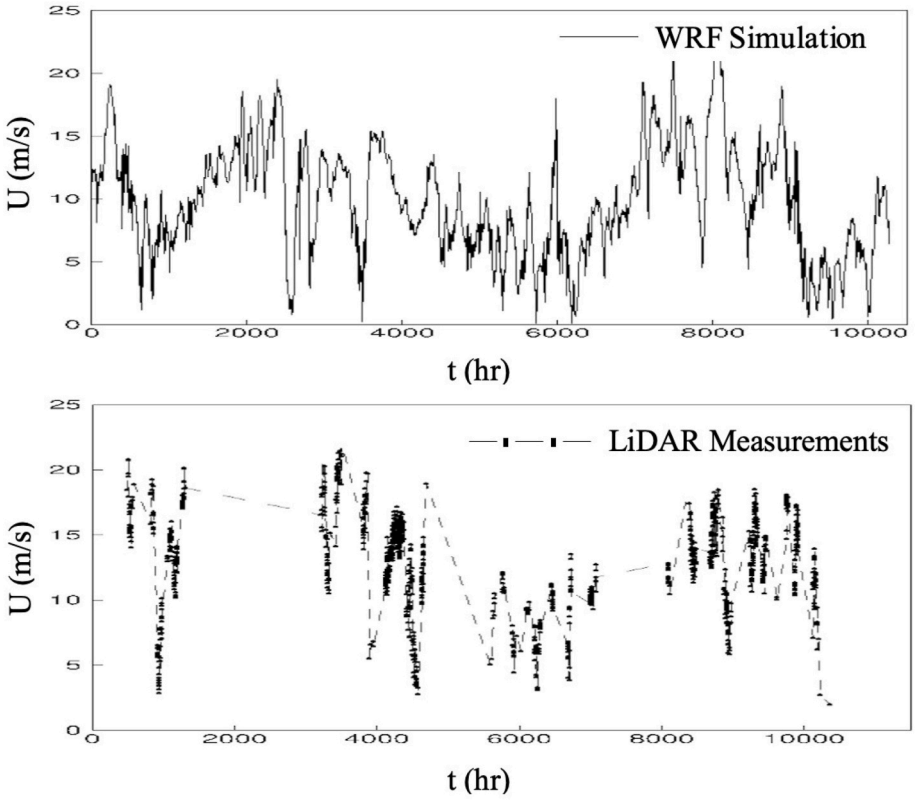

Fig. 4. (a) Low-fidelity data from numerical simulation output from the WRF model at the most offshore point. (b) High-fidelity data from the dual-Doppler lidar setup at the most offshore point. There are 5 continuous WRF time series and 16 intermittent lidar time series for each height, where data are missing at the same time intervals across all the points in the high-fidelity time series.

of 12 new lidar points; and matrix C had 4 lidar points per km leading to a total of 16 lidar points. The first experiment used data from the 50-m height level dataset. The statistics of the datasets for experiment 1 are given in Table 1, which also shows the division of data for all three matrices A, B and C with percentage of sparsity in every matrix.

For experiment 2, matrix C with 16 lidar points was used, at the three different heights (50, 100, and 150 m). Results from the experiment were compared to those using an academic prediction model for wind forecast predictions (EWT + GPR) and a leading industrial software, Windgrapher, a leading software for importing, visualising and analysing wind resource data [25]. Windgrapher follows the measure-correlate-predict (MCP) algorithms including linear least squares; the method is on correlating target and reference speed data, based in the linear least squares procedure. However, these techniques are one dimensional interval predictions techniques, hence can only be applied to one spatial point at a time and require 16 iterations for comparison with one iteration from the probabilistic matrix factorization with Gaussian process model.

4. Results and discussion

Results from both experiments 1 and 2 are discussed and compared in this section. First, the results from experiment 1, where three matrices had different number of dual-Doppler points to study the effect of adding lidar observations and increasing the sparseness of the matrix. Then, in experiment 2, results are compared by means of the root mean square error (RMSE) as metric for 16 dualDoppler points at 3 different heights using the proposed non-linear probabilistic matrix factorization with Gaussian process with two D-factors (2 and 5), the industrial software, the academic model, and the WRF simulation.

4.1. Experiment 1: different sparsity with additional lidar measurements

Three different datasets of the 16 dual-Doppler offshore points were

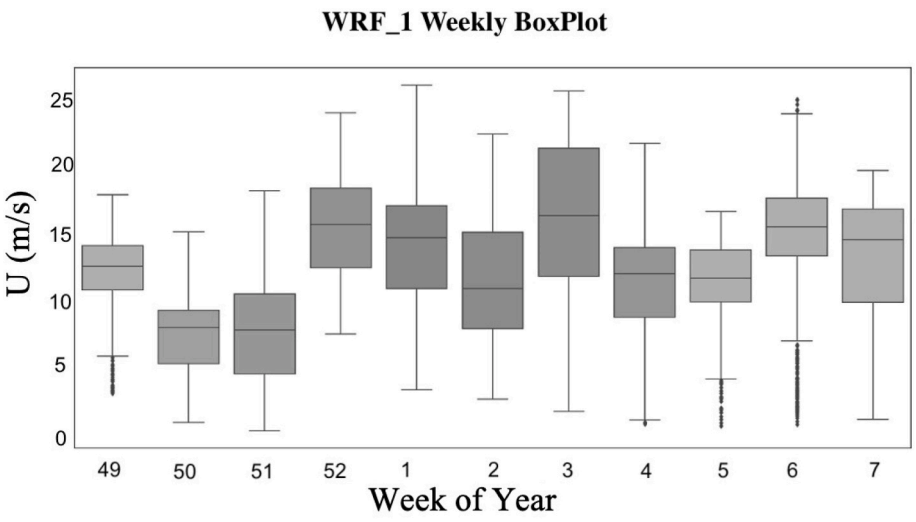

Fig. 5. (a) Weekly box plot for the WRF-based wind speed time series.

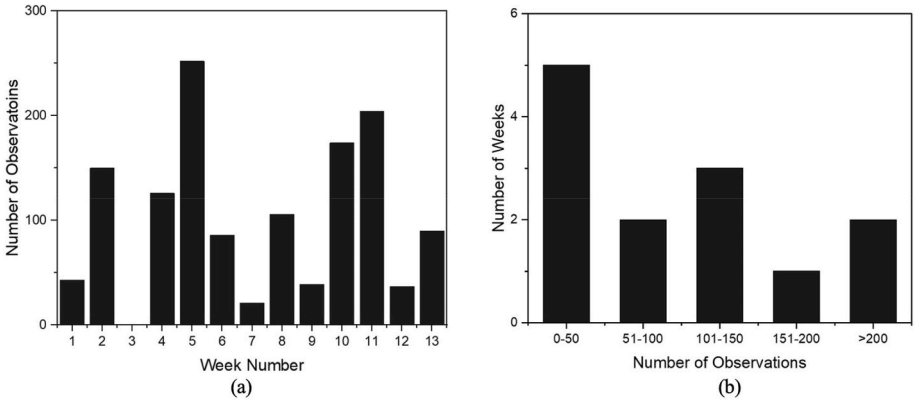

Fig. 6. Histograms of (a) wind speed measurement frequencies; and, (b) the division of observations by weeks.

used to evaluate three iterations of probabilistic matrix factorization. The performance of the three models was optimized, through the hyperparameters, by applying different learning rates and D-factors (2 and 5). Firstly, exploring the matrix with 5 WRF points and 2 lidar points per km, with a total of 8 dual-Doppler points of the total 16, the time steps represent 3 months worth of data at 10 min intervals. The dimensions of the matrix were 13 columns and 12,960 rows, represented as model (A). Then, for model (B), the number of lidar points was increased to 3 per km, and finally model (C) had 4 lidar points per km, including all 16 dual-Doppler points. The slight increase in the number of lidar points varied the sparsity of the models from 91% to 94%. This experiment was employed to test how the algorithm reacted to different sparsities (density of missing data) and how increasing the number of measurements affected the predictions.

Table 1 shows the statistics of the datasets used in Experiment 1. All models had a constant number of WRF data (5 points), while the lidar data and hence the total number of cells in the matrix varied, resulting in three different sparsity percentages for models A, B, and C; 91%, 93% and 94%, respectively.

Fig. 8 panel (a) shows the predicted time series of the first dual-Doppler point and the lidar measurements. Additionally, panel (b) shows the observed lidar measurements against their counterparts from the predicted time series for the 20% test dataset. Results from the matrix factorization process showed that the generated time series with predictions was able to follow the complex trends from the lidar data and follow a similar signal. Despite the fluctuations in the dataset, the generated time series in Fig. 8 panel (a) intersects with the lidar points at most of the lidar points, which is reflected in panel (b) of the same figure, where the variance between predicted-observed points is less than 0.5 m/s. Fig. 9 top panels (a), (b), and (c) shows the time series of the three original matrices with 8, 12, and 16 dual-Doppler points, respectively. Panels (d), (e), and (f) show the 3 result matrices with the predictions. The main purpose of Fig. 9 was to visually demonstrate the power of matrix factorization in wind speed prediction and observe how varying the percentage sparsity of the matrix affected the accuracy of predictions. The model digested a matrix with low data density (high sparsity) and was able to generate predictions for 16 different points at different locations in a single iteration. The RMSE for each point in the matrix with respect to the 20% hidden test data within the time series was measured and compared to the WRF output and lidar data. The results from all matrices outperform that of the WRF simulation, where the RMSEs of the time series generated by matrix factorization at the three heights ranged between 0.38 m/s and 0.46 m/s compared to 1.4 m/s and 0.8 m/s from the WRF simulation. The experiment showed that increasing the number of dual-Doppler points, hence the sparsity of the matrix and amount of high accurate data, did not affect the accuracy of the prediction as the RMSE was unchanged amongst all points. This was due to the multicollinearity in the dataset, since all lidar and WRF points were highly correlated. Note that using any of the neighboring points would be beneficial, despite the distances from the targeted point, and since the points were to an extent linearly correlated, the amount of missing data was not a problem.

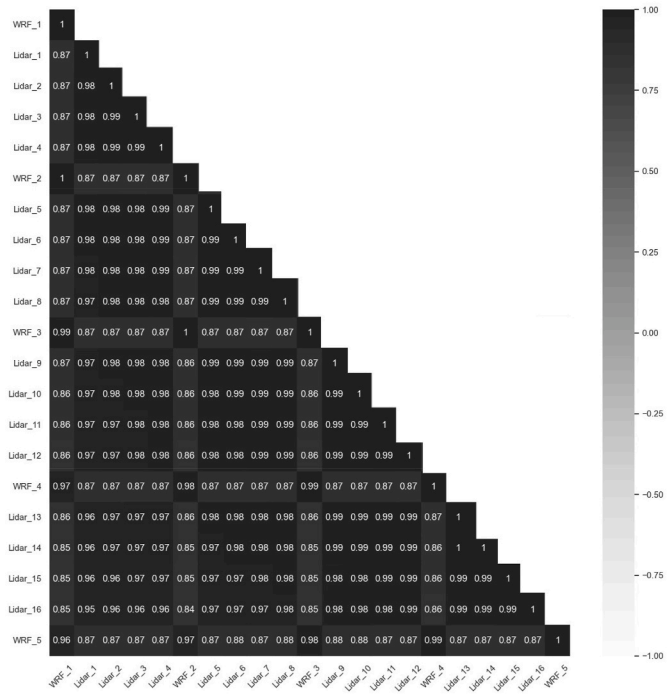

Fig. 7. Correlation between all lidar and WRF points.

Table 1
Statistics of the datasets used in Experiment 1.

	A	B	C
lidar Points	8	12	16
Measurements (lidar)	10624	15936	21248
Simulation data (WRF)	56880	56880	56880
Total No. of Cells	739440	966960	1194480
Missing data (sparsity)	91%	93%	94%

4.2. Experiment 2: fixed sparsity at different heights

The second numerical experiment aimed to test how the models performed at different heights compared to other academic and industrial algorithms. Three models were trained with 5 WRF points, the number of dual-Doppler points was constant at 4 points per km leading to a total of 16 in each matrix. The three different heights of the experiment are 50, 100, and 150 m. The sparsity of all three models was unchanged at 11.8%. Statistics of the dataset for this experiment can be found in Table 2.

Subsequently, the performance of the Matrix factorization algorithm was evaluated against results from Windgrapher. Another algorithm in the comparison was the empirical wavelet transform (EWT) and multi-fidelity Gaussian process regression (MF-GPR). The EWT was used to pre-process the WRF time series reducing the spikes and high frequency fluctuations, which results in a more accurate Gaussian process. Finally, for the PMF testing, different D numbers, 2 and 5 were used.

The RMSEs for each of the 16 dual-Doppler points were measured at one iteration from the Matrix using PMF with D = 2 and D = 5, then compared to their counterparts using EWT + MF-GPR, and Windgrapher, but after 16 different iterations (as they can predict for a single point at a time only). Fig. 10 shows the RMSE for each of the 16 dual-Doppler points using all algorithms tested, where panels (a), (b), and (c) represent the heights 50, 100, and 150 m, respectively. On average, the PMF algorithm using D = 2 and D = 5 was able to outperform the other algorithms at all measured points for all three heights. The RMSE of the PMF algorithm was reduced by at least 65% compared to that using Windgrapher and 40% compared to that of the academic algorithm for the height of 50 m. At the heights 100 and 150 m, the RMSEs were very similar; however, increasing the height reduced the RMSE and increased the percentage drop in RMSE, which was mainly because the accuracy of the WRF simulation increased at higher heights. Additionally, as matrix factorization predictions for the 16 points were performed at the same iteration and used the data from all lidar and WRF points, the 16 RMSEs for the NPMF results (C and D) in Fig. 10 had less variance in their values compared to that of the results from using other models, as other models performed predictions separately for each point with different hyperparameters resulting in different performances within the same model.

Fig. 11 shows the RMSE of the most (panel a) and least (panel b) offshore point using all the algorithms for all three heights; that is the 3 algorithms discussed above and the 2 PMF models. The results in the figure demonstrate the order of algorithms showing the least to most accurate, with PMF leading for both D numbers. Similarly, the results indicate that by increasing the height of the measurements more accurate predictions are achieved, which is due to higher ability of the WRF output to predict winds.

5. Conclusions

In this work, data fusion was performed between lidar measurements sparse in space and time with high-fidelity with output from the WRF model, which was continuous in space and time but of low-fidelity. The aim of the data fusion was to obtain spatio-temporal predictions at unobserved space and time points, suitable for offshore wind resource assessment. During the RUNE experiment dualDoppler scans were performed, which resulted in 36 lidar points across 10 km, both offshore and onshore, with 1331 measurements. Subsequently, the numerical

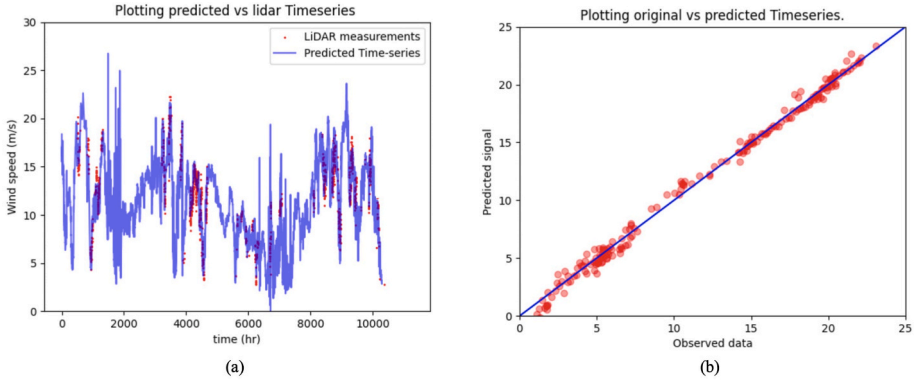

Fig. 8. (a) Predicted Time-series and lidar observations for the first dual-Doppler point marked red in Fig. 3(b). (b) Prediction points against counter observed measurements.

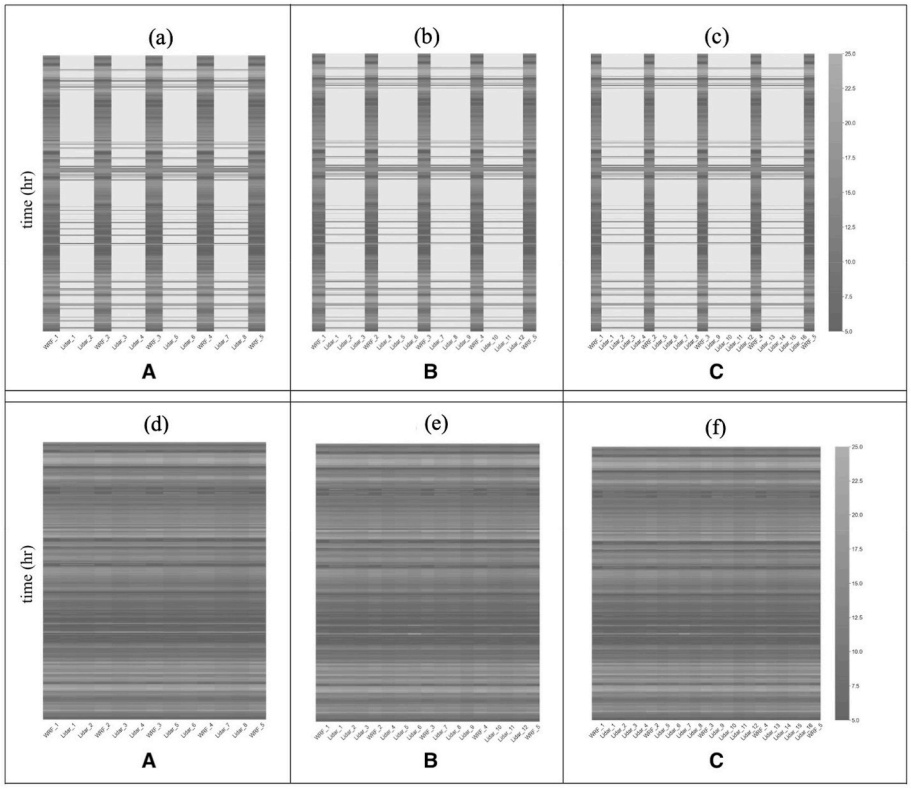

Fig. 9. Original matrices of all three setups with 2, 3, and 4 lidar points, A, B, and C, respectively (Top). Final processed matrices for all setups, A, B, and C, respectively (bottom).

Table 2
Statistics of the datasets used in Experiment 2.

	All heights
Lidar points	16
Time-steps	11376
Measurements observed	21248
Measurements per Point	1328
Measurements density (sparsity)	11.8%

simulations performed using the WRF model generated an instantaneous constant output every 10 min for the same period, resulting in a total of 11,376 data points.

In this study, the performance of the algorithm was tested on the offshore data at 16 points at 3 different heights, namely 50, 100, and 150 m.

In a first experiment, the number of lidar points per km was varied to test the accuracy of the model with less valuable lidar data of sparsity percentages, which varied between 91%, 93%, and 94%. As aforementioned, the high-fidelity lidar data was presented at unobserved regions and periods by exploiting the available data and the low-fidelity WRF output. The addition of more lidar points caused an increase in the sparsity of the matrix, despite giving more valuable lidar data. The RMSE of predictions was not affected as it ranged from 0.45 m/s and 0.52 m/s across all three matrices; in this experiment, only the data at 50 m was used.

Contrarily, the second experiment aimed to test the accuracy of prediction of the PMF model when the D-factor is varied at 2 and 5, when compared to the results of an academic model with pre-processing (EWT + multi-fidelity GPR) and those of an industrial leading software (Windgrapher). First, 16 lidar points and 5 WRF points with 11.8% sparsity were used. The results showed that using a lower D-factor was more accurate and resulted in improved predictions; this was observed significantly at lower heights. Additionally, the results for all models showed that by using data at higher heights more accurate predictions were achieved, as the WRF outputs were more accurate at higher levels at this particular site. Hence, the results for the heights of 50 and 150 m had the highest and lowest RMSEs, respectively. The difference RMSE

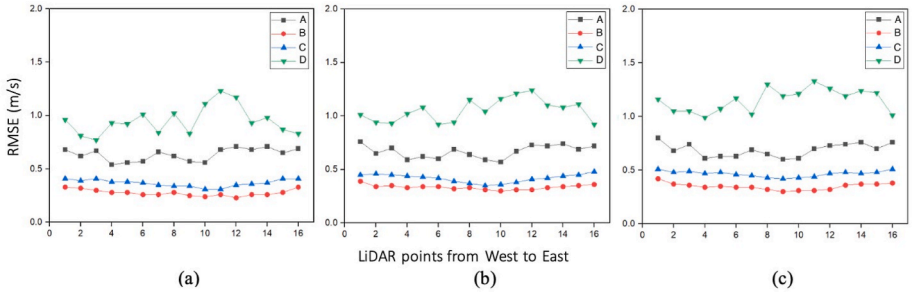

Fig. 10. Comparison between RMSE results from Windgrapher (A), EWT + MGPR (B), our tested method NPMF with Gaussian process (D = 5 and D = 2), (C) and (D), respectively, across all 16 grid points for heights 50 m (a), 100 m (b), and 150 m (c).

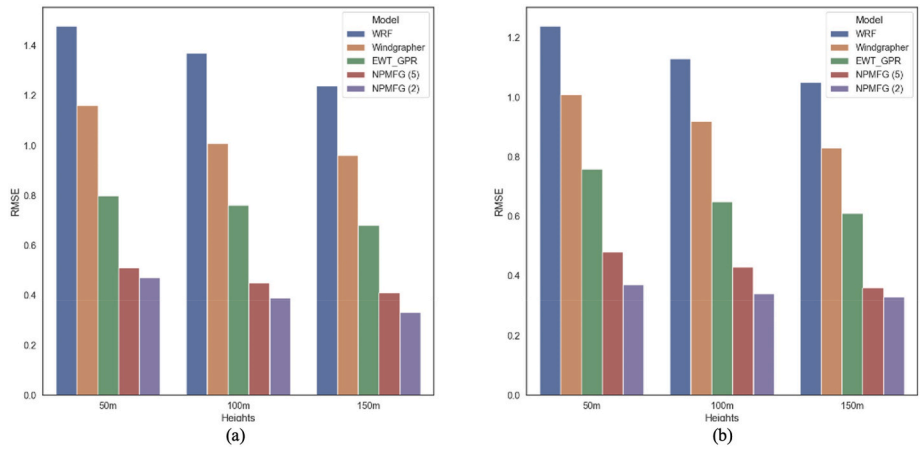

Fig. 11. (a) Comparison between RMSE results from the WRF output, Windgrapher, EWT + MGPR, our current method NPMF with Gaussian process (D = 5 and D = 2) for the far east lidar point for each height. (b) Comparison between RMSE results from the WRF output, Windgrapher, EWT + MGPR, NPMF with Gaussian process (D = 5 and D = 2) for the far west lidar point for each height.

between both heights was around 0.18 m/s, equivalent to a 15% drop in RMSE. Subsequently, both PMF models with 2 and 5 D-factors were able to outperform the results of both the industrial and the academic models by at least 58% and 40%, respectively.

There are three major limitations in this study, which could be addressed in future work. First, only data obtained from the RUNE experiment was addressed, which is a very sparse dataset, including only 1331 measurements for each lidar point equivalent to 220 h of measured data, reflecting several weeks with no data. Second, due to the resolution of the WRF model output (2 km), the generated WRF points are all to the nearest 100 m, which is a considerable distance, as points this far will have significant differences in the wind speed. Hence, interpolation was necessary to obtain WRF data at the corresponding lidar points. Third, the starting time for both datasets generated using the WRF model and from the lidar was not the same. The WRF output was available a few days earlier. This problem is called a 'cold' start and is a common issue in matrix completion problems.

Future work may also concern further development of the matrix input data. Additional datasets such as derivatives and other type of WRF output (including output from a higher resolution run) could be a great source of information. This would reduce the sparsity of the matrix and improve the Gaussian process, hence the accuracy of predictions. Furthermore, the preprocessing methods of the WRF data with EWT before processing in the matrix should be tested.

CRediT authorship contribution statement

Basem Elshafei: Writing – original draft, Performed the matrix factorization work and drafted the manuscript. **Alfredo Peña:** Writing – original draft, Processed the lidar and WRF data and drafted the case study part of the manuscript. **Atanas Popov:** Supervision, Methodology, Writing – original draft, Supervised the work and co-wrote the matrix factorization method. **Donald Giddings:** Writing – original draft, Methodology, Supervision, Supervised the work and co-wrote the Gaussian process method. **Jie Ren:** Writing – original draft, Designed the data structure to be fused and co-wrote the EWT model. **Dong Xu:** Methodology, Ran the Windgrapher tests for comparison with the proposed method. **Xuerui Mao:** Coordinated the work and summarized the findings.

Declaration of competing interest

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

Acknowledgement

This work has received funding from the European Union Horizon 2020 research and innovation programme under the Marie Skl odowska-Curie grant agreement No 777717 and the future and emerging technologies programme with agreement No.828799.

Appendix .0.1. EWT

The process consists of five main steps: extending the signal, Fourier transforms, extracting boundaries, building a filter bank, and extracting the sub band. The five level decomposition attained by the preprocessing algorithm, EWT, was able to describe the signal in a meaningful way with much less fluctuations, by extracting five uncorrelated filter modes from the wind speed signal and a residual from the extraction. The reconstructed signal will be used as additional input for the forecasting models.

Appendix .0.2. PMF is Bayesian PCA

With little changes to the notations, PMF is probabilistically equivalent to Bayesian PCA. The Bayesian treatment provides fully automatic complexity control as model parameters and hyperparameters are integrated. Considering a matrix of latent variables, $X \equiv U^{\top} \in R^{N \times D}$, and a mapping matrix that goes from the latent space to the space of observed data, $W \equiv V^{\top} \in R^{M \times D}$. Following the new notation, the probabilistic model can be written in the form:

$$p(R|W,X,\sigma^2) = \prod_{I=1}^{N} N(r_i|W_{Xi},:,\sigma^2 I), \tag{A1}$$

where X_i: is the ith column of U, and $r_{i,:}$ represents the column vector from the ith row of R containing wind speeds of the ith time-step for a point in space. The previous equation is a multi-output linear regression from a D dimensional feature matrix V to matrix targets R. Placing a prior over X gives the following, which can be marginalized later to give the marginal likelihood used in missing values imputation.

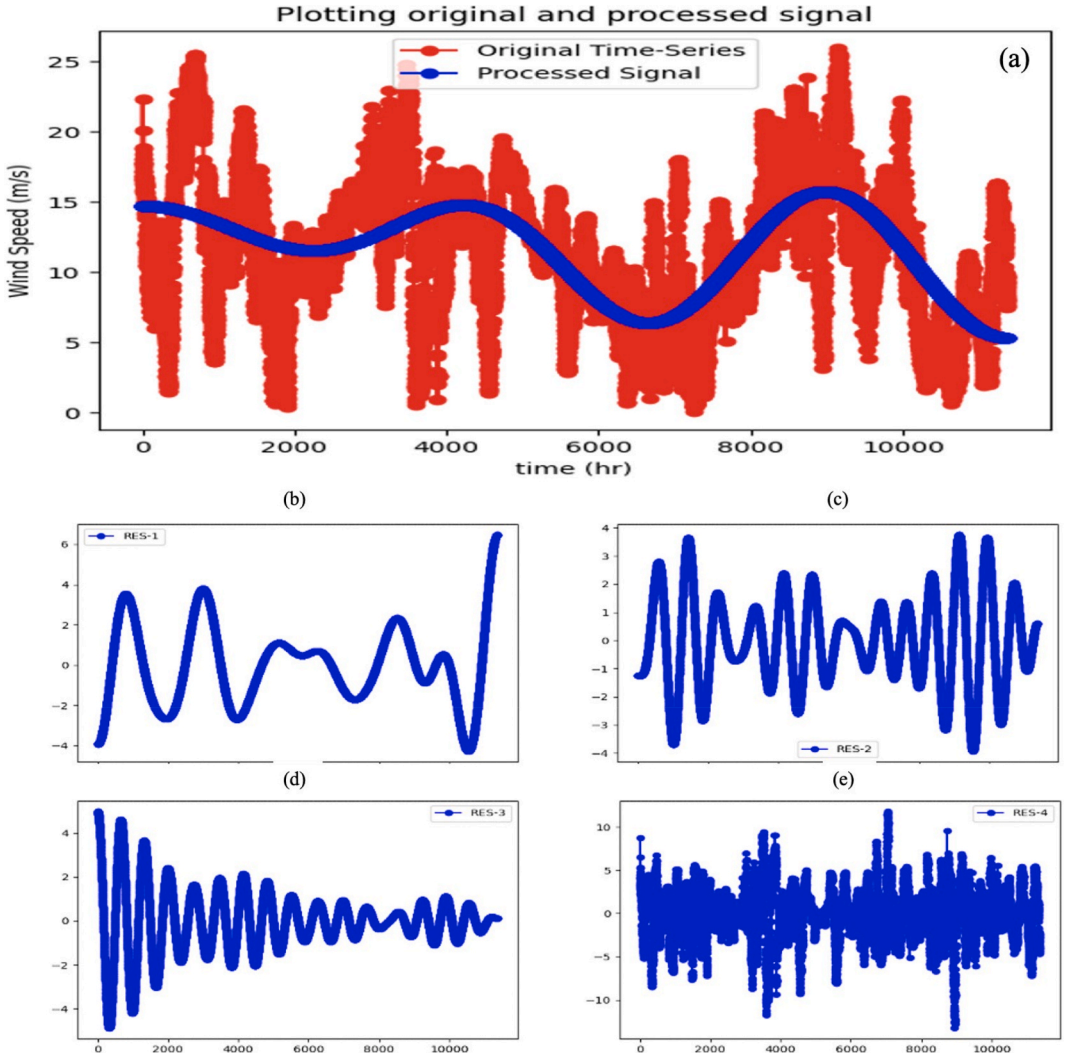

Fig. S12. a) Original and processed signals. (b–e) residuals of the preprocessing model.

$$p(X) = \prod_{i=1}^{ND} \prod_{j=1} N(x_{i,j}|0,\sigma_x{}^{-1}) \tag{A2}$$

Appendix .0.3. Missing values imputation

The method discussed is a Gaussian matrix factorization model with a particular covariance structure, which means marginalizing is straightforward in finding the missing values. A Gaussian distribution is considered over the following parameters: a vector y with mean μ and covariance σ, in the form y $N(\mu,\sigma)$. An observed subset of y is represented by y_i, where i is an index for the observed values. When marginalizing the missing values, getting the Gaussian form y_i $N(\mu_i,\sigma_{i,i})$ where μ_i, and $\sigma_{i,i}$, represent the mean vector with the rows for the sum of σ columns associated with the unobserved elements of the removed y. Hence, for a sparse data matrix, the likelihood is given by:

$$p(Y|W,\sigma^2,\alpha_x) = \prod_{i=1}^{N} N\left(y_{i,j_i}\middle|0,\alpha_x^{-1}W_{ji},W_{ji,:}^T+\sigma^2 I\right) \tag{A3}$$

Optimizing with respect to the parameters leads to α_x being part of W, which leaves the likelihood function associated with PPCA, and hence becomes intractable when marginalizing W. Instead the prior is taken over W,

$$p(W) = \prod_{I=1}^{MD} \prod_{j=1} N\left(W_{i,j}|0, \alpha W - 1\right) \tag{A4}$$

and the marginal likelihood is then in the form.

$$p(Y|X,\sigma,\alpha_w) = \prod_{M}^{D} N(y_{ij,j}|0,\alpha_w X_{ij,:} X_i^T + \sigma^2 I). \tag{.5}$$

which is the marginal likelihood of a Bayesian linear regression model with multiple outputs. These equivalences imply that with marginalisation of either W or X, will eventually lead to optimizing the resulting marginal likelihood for the remaining matrix and model hyperparameters.

References

[1] H.E. Murdock, D. Gibb, T. Andre, J.L. Sawin, A. Brown, F. Appavou, et al., International Nuclear Information System Renewables 2020-Global Status Report, 2020.

[2] Fs-unep-centre.org [online] Available at: https://www.fs-unepcentre.org/wp-content/uploads/2020/06/GTR2020.pdf, 2022.

[3] Chen Wang, Shenghui Zhang, Peng Liao, Tonglin Fu, Wind speed forecasting based on hybrid model with model selection and wind energy conversion, Renew. Energy 196 (2022) 763–781. ISSN 0960-1481.

[4] H. Victoria, S. Tomlin Alison, Cockerill Timothy, Improved near surface wind speed predictions using Gaussian process regression combined with numerical weather predictions and observed meteorological data, Renew. Energy 126 (2018).

[5] S.M. Weekes, A.S. Tomlin, S.B. Vosper, A.K. Skea, M.L. Gallani, J.J. Standen, Long-term wind resource assessment for small and medium-scale turbines using operational forecast data and measure–correlate–predict, Renew. Energy 81 (2015) 760–769. ISSN 0960-1481.

[6] A.M. Sempreviva, R.J. Barthelmie, S.C. Pryor, Review of methodologies for offshore wind resource assessment in European seas, Surv. Geophys. 29 (2008).

[7] S. Shikha, T.S. Bhatti, D.P. Kothari, A review of wind-resource-assessment technology, J. Energy Eng. 132 (2006).

[8] C. Zhang, W. Haikun, Z. Xin, L. Tianhong, Z. Kanjian, A Gaussian process regression based hybrid approach for short-term wind speed prediction, Energy Convers. Manag. 126 (2016).

[9] D.R. Drew, J.F. Barlow, T.T. Cockerill, M.M. Vahdati, The importance of accurate wind resource assessment for evaluating the economic viability of small wind turbines, Renew. Energy 77 (2015) 493–500, https://doi.org/10.1016/j.renene.2014.12.032. ISSN 09601481.

[10] J. Tastu, P. Pinson, E. Kotwa, M. Henrik, Aa Nielsen Henrik, Spatiotemporal analysis and modeling of short-term wind power forecast errors, Wind Energy 14 (2011).

[11] B.M. Marlin, R.S. Zemel, S.T. Roweis, M. Slaney, Recommender systems: missing data and statistical model estimation, IJCAI 12 (2011).

[12] B. Munoz, V.M. Lesser, R.A. Smith, B.M. Munoz Virginia Lesser Ruben A Smith, Applying multiple imputation with geostatistical models to account for item nonresponse in environmental data, J. Mod. Appl. Stat. Methods 9 (2010).

[13] Q. Zhang, Q. Yuan, C. Zeng, X. Li, Y. Wei, Missing data reconstruction in remote sensing image with a unified spatial-temporal-spectral deep convolutional neural network, IEEE Trans. Geosci. Rem. Sens. 56 (8) (2018) 4274–4288.

[14] X. Wang, A. Li, Z. Jiang, H. Feng, Missing value estimation for DNA microarray gene expression data by Support Vector Regression imputation and orthogonal coding scheme, BMC Bioinf. 7 (2006), https://doi.org/10.1186/1471-21057-32.

[15] M. Agathokleous, N. Tsapatsoulis, Voting Advice Applications: missing value estimation using matrix factorization and collaborative filtering, IFIP Adv. Inf. Commun. Technol. 412 (2013) 20–29, https://doi.org/10.1007/978-3-642-41142-7 3. Springer New York LLC.

[16] Ungar LH, Foster DP. Clustering methods for collaborative filtering. AAAI Workshop on Recommendation Systems, pp. 1-16 1998.

[17] Yehuda K, Robert B, Chris V. "Matrix factorization techniques for recommender systems," in Computer, vol. 42, no. 8, pp. 30-37, doi: 10.1109/MC.2009.263.2009.

[18] T. Zhou, H. Shan, A. Banerjee, G. Sapiro, Kernelized probabilistic matrix factorization: exploiting graphs and side information, in: SIAM International Confer- Ence on Data Mining, 2012, pp. 403–414.

[19] R. Salakhutdinov, A. Mnih, Probabilistic Matrix Factorization, 2007.

[20] B.B. Hu, A.B. Hu, EXTENDED BAYESIAN MATRIX FACTORIZATION WITH NON-random MISSING DATA bayesian matrix factorization with NonRandom missing data using informative Gaussian process priors and soft evidences bence Bolg'ar. Proceedings of the eighth international conference on probabilistic graphical models, in: Proceedings of Machine Learning Research vol. 52, 2016, pp. 25–36.

[21] H. Shan, A. Banerjee, Generalized probabilistic matrix factorizations for collaborative filtering, ICDM 3 (2010).

[22] J. Gilles, Empirical wavelet transform, IEEE Trans. Signal Process. 61 (16) (Aug.15, 2013) 3999–4010, https://doi.org/10.1109/TSP.2013.2265222.

[23] R. Sarkar, S. Julai, S. Hossain, W.T. Chong, M. Rahman, A comparative study of activation functions of NAR and NARX neural network for long-term wind speed forecasting in Malaysia, Math. Probl Eng. 2019 (2019).

[24] R. Floors, A. Peña, G. Lea, N. Vasiljević, E. Simon, M. Courtney, The RUNE experiment-a database of remote-sensing observations of near-shore winds, Rem. Sens. 8 (2016).

[25] B. Elshafei, A. Peña, D. Xu, J. Ren, J. Badger, F.M. Pimenta, et al., A hybrid solution for offshore wind resource assessment from limited onshore measurements, Appl. Energy 298 (2021), https://doi.org/10.1016/j.apenergy.2021.117245.