Gaussian Mixture Models for the Optimal Sparse Sampling of Offshore Wind Resource

Robin Marcille [1, 2], Maxime Thiébaut [1], Pierre Tandeo [2], and Jean-François Filipot [1]

[1]France Énergies Marines, Technopôle Brest-Iroise, 525 Avenue Alexis de Rochon, 29280 Plouzané, France
[2]IMT Atlantique, Lab-STICC, UMR CNRS 6285, 29238 Plouzané, France

Correspondence: Robin Marcille (robin.marcille@france-energies-marines.org)

Abstract. Wind resource assessment is a crucial step for the development of offshore wind energy. It relies on the installation of measurement devices, whose placement is an open challenge for developers. Indeed, the optimal sensors' placement for field reconstruction is an open challenge in the field of sparse sampling. As for the application to offshore wind field reconstruction, no similar study was found, and standard strategies are based on semi-empirical choices. In this paper, a sparse sampling method using a Gaussian Mixture Model on Numerical Weather Prediction data is developed for offshore wind reconstruction. It is applied on France's main offshore wind energy development areas: Normandy, Southern Brittany, and the Mediterranean Sea. The study is based on 3 years of Meteo France AROME's data, available through the MeteoNet data set. Using a Gaussian Mixture Model for data clustering, it yields an optimal sensors' locations with regards to wind field reconstruction error. The proposed workflow is described and compared to state-of-the-art methods for sparse sampling. It constitutes a robust yet simple method for the definition of optimal sensor siting for offshore wind field reconstruction. The described method applied to the study areas outputs sensors arrays of respectively 7, 4, and 4 sensors for Normandy, Southern Brittany and the Mediterranean Sea. Those sensors arrays perform approximately 20% better than the median Monte Carlo case, and more than 30% better than state-of-the-art methods, with regards to wind field reconstruction error.

1 Introduction

Offshore wind energy is key in the decarbonation of the global energy production and the reaching of net-zero targets as developed in Shukla et al. (2022). With 11 million km^2 of territorial waters under French jurisdiction and 20,000 km of coastline, France has an extensive and windy seafront. It benefits from the second largest offshore wind potential in Europe, after the United Kingdom with up to 80 GW of foundation-based offshore wind and 140 GW of floating offshore wind that could be exploited according to IEA (2019). Offshore wind can then be a leading sector for the development of renewable energies in France. The french road-map currently plans 1 GW of tender per year from 2024 onwards for fixed and floating wind farms. This was confirmed and reinforced in early 2022, with 40 GW of installed capacity envisioned by 2050.

During the development phase of a wind project, the wind resource assessment is a key step to determine its financial feasibility. It can be carried out with numerical weather prediction (NWP) hindcast data such as WRF (Weather Research and

Forecasting model) data. However, field observations are necessary to estimate the uncertainties of the models and to assess higher resolution wind dynamics (Murthy and Rahi, 2017).

LiDARs, standing for Light Detection And Ranging, are remote sensing devices that measure wind speed using lasers. Floating LiDARs are certified devices for offshore wind resource assessment, they are LiDAR units integrated onto a standalone floating structure. These wind sensors offer the potential for reduced costs compared to meteorological masts (Gottschall et al., 2017), however, they can be expensive to install and require regular maintenance. Their number and siting thus need to be optimized in order to compose an optimal network of sensors in an offshore wind development area. Such networks are expected to capture most of the dominant wind dynamics from a minimum number of sensors.

Numerous efforts have been undertaken in different scientific fields to optimize sparse sensor siting, a combinatorial problem not solvable by standard approaches such as convex optimization. Sparse sampling is about selecting salient points in a highly dimensional system. It then requires a dimension reduction of the data, such as the use of Empirical Orthogonal functions (EOF). EOF analysis projects the original data onto an orthogonal basis derived by computing the eigenvectors of a spatially weighted anomaly covariance matrix. Therefore, EOF of a space-time physical process can represent mutually orthogonal space patterns where the data variance is concentrated, with the first pattern being responsible for the largest part of the variance, the second for the largest part of the remaining variance, and so on. EOF are then very useful for the data reduction of any complex data set such as climate data. By projecting the original data onto a limited subset of relevant orthogonal vectors, it reduces the dimensionality of the system and helps explain the variance of the data. In the past few decades, EOF analyses were used to study spatio-temporal patterns of climate variability, such as the North Atlantic oscillation, the Antarctic Oscillation or the variability of the Atlantic thermohaline circulation (e.g., Davis (1976); Thompson and Wallace (2000); Hawkins and Sutton (2007); Moore et al. (2013)).

EOF are often at the origin of methods employed to determine the optimal sensors' locations for signal reconstruction.In the field of geoscience, Yildirim et al. (2009) employed simulation results from different regional ocean models to define an efficient sensor placement. The authors used the EOF technique to determine the spatial modes of different simulated ocean dynamics systems. The extrema of the EOF spatial modes were found to be good locations for sensors' placement and accurate field reconstruction. Zhang and Bellingham (2008) added to the Empirical Orthogonal Functions' extrema (EOF extrema) method a constraint on the cross products of EOF to select the sensors' locations, and applied it to Pacific sea surface temperature reconstruction. Using the same kind of constrained EOF analysis, Castillo and Messina (2019) proposed a data-driven framework based on a Proper Orthogonal Decomposition (POD) to determine the optimal locations for power system oscillation monitoring and state reconstruction. In this study they selected iteratively the locations with highest POD amplitude and lowest cross coupling between the modes. In Yang et al. (2010), the EOF extrema are used for ocean dynamics reconstruction, introducing an exclusion volume to avoid redundancy, account for gappy data and for uncertainty.

Manohar et al. (2018) proposed a data-driven method based on a QR pivoting greedy algorithm on a reduced basis to determine optimal sensors' placements for face recognition, global sea surface temperature and flow reconstruction around a cylinder. The QR pivoting method decomposes a matrix into an orthogonal matrix and an upper triangular matrix using columns pivot. By iteratively selecting the column with the highest two-norm as pivot, this algorithm for QR factorization is suited for

the selection of salient points. In Clark et al. (2020), the QR-pivot decomposition is modified to include cost constraints and is applied on the three same data sets. The QR greedy algorithm described in these studies is a often used in recent studies, showing good capabilities and being very easily implementable.

Recent studies proposed innovative methods to improve the capabilities of sparse sampling. To improve the performance of the reconstruction, Chepuri and Leus (2014) proposed a method for grid augmentation to allow for continuous sensor placement, off-grid sensor selection and convex optimization problem formulation. In Mohren et al. (2018), the authors took inspiration from insects' neural activation during flights to derive a sparse sampling method in complex flows to create an encoder for flight mode classification including both the spatial and temporal dependencies of the data. In Fukami et al. (2021), the use of Voronoi tessellation, a method to optimally partition the space into n cells given n input points using a distance measure d, helps creating a viable input for super-resolution from sparse sensors using a Convolutional Neural Network (CNN). This reconstruction technique is then tested on sea surface temperature reconstruction globally, showing the possibility to use sparse sampling on very high dimension problems.

The problematic of optimal sensors' placement has also been investigated for wind energy measurements applications. Annoni et al. (2018) uses the QR greedy algorithm described in Manohar et al. (2018) to determine the optimal locations of sensors to improve the overall estimation precision of the flow field within a wind farm. In this study, the number of sensors is directly computed using a user-defined threshold with regards to reconstruction error. A similar strategy is implemented in this article as presented lower. The obtained results show good performance compared to randomly selected grid points, with an improvement of 8% in flow field reconstruction, and shows the interest in applying sparse sampling methods to the wind energy sector. At even finer scales, Ali et al. (2021) uses low-dimensional classifiers applied to the Proper Orthogonal Decomposition of a LES wake simulation to obtain sensors' locations for the reconstruction of wind turbine wakes. Using the method of sparse sensor placement optimization for classification described in Brunton et al. (2016), it shows the interest of sparse sampling for the control of wind turbines, using a Deep Learning algorithm to predict the wake fluctuations from sensors' measurements. Results show that most sensors are placed in the transition region, and the reconstruction yields to more than 92% correlation between predicted and real values.

However, to the best of our knowledge, such methods were never applied at the regional scale for wind energy resource assessment. In our opinion this is due to site selection procedures at the political level that do not necessarily rely on wind resource assessment at the regional level, and to smaller required spatial scales at the wind farm developer level, where only one or two sensors are deployed at the extremities of the area, assuming spatial representativity. The application of sparse sampling methodologies to offshore wind reconstruction is an addition of this work. Using NWP spatial wind data as input, the study proposes an unsupervised clustering framework for the identification of salient points in the spatial grid, similar to what can be obtain through EOF extrema analysis in Yildirim et al. (2009) or QR pivoting in Manohar et al. (2018). In the application-driven experimental set-up of this study, the two state-of-the-art methods fail to capture wind dynamics at the regional level. Unsupervised clustering automatically discriminates points that are too similar, making it a good candidate for sparse sampling in this case, while keeping the whole method simple and easily implementable.

The objective of the present study is twofold, and the associated problematic is formulated as the following - for conducting offshore wind resource assessment of any targeted area:

1. What is the optimal number of offshore wind sensors to be deployed to best characterize the wind resource?
2. What is the optimal location of each wind sensor?

The optimal number of sensors refers to a trade-off between wind field reconstruction accuracy, and overall cost and computational cost. The optimal locations given a certain number of sensors is the configuration giving the lowest reconstruction error. The two aspects are presented in this work, though realistic cost considerations are not covered.

To do so, this paper presents a data-driven method based on NWP data unsupervised clustering to estimate optimal sensors' locations for offshore wind field reconstruction using a Gaussian Mixture Model. It is compared to state-of-the-art methods used in the above literature (EOF extrema, QR pivoting, randomly selected sensors). The method is then implemented on three areas identified for offshore wind energy development in France. An optimal wind sensors network is proposed for each area, to help for the development of offshore wind energy in France.

2 Study data set

2.1 Study areas

The three areas investigated in this study are located off the coast Normandy, off the coast of Southern Brittany and in the Mediterranean sea, three major development areas for offshore wind in France with numerous planned offshore projects, listed in Table 1, with future tender processes for respectively 1.5GW of fixed offshore wind, 250MW of floating offshore wind and 2 x 250MW of floating (expected date of commissioning in 2030).

With water depth not exceeding 50 m (Fig. 1), the area located off the coast of Normandy area is particularly suitable for the deployment of fixed offshore wind farms. Current projects will be installed off the coast of Fécamp, Courseulles-sur-Mer and Dieppe - Le Tréport (Fig. 1). The total capacity of each wind farms will be 450-500 MW with a starting date of commissioning expected in 2023-24 (Table 1). In addition, the French Government has recently announced a new project of a wind farm located 32 km off the coast of Normandy (Fig. 1). This future wind farm will generate 1 GW. The starting date of commissioning is expected by 2028.

The area off the coast of Brittany is endowed with water depth up to 100 m which make it a very favourable area for the development of floating wind farms. ~~A pilot floating wind farm is planned off Groix-Belle Ile, to be commissioned in 2023. Furthermore, the~~ The French Government aims at developing ~~another~~ 250MW of floating wind energy in the area (Fig. 1).

Because of its very favorable and regular wind regimes and deep bathymetry, the Mediterranean Sea has significant wind potential for floating wind energy. This led to the development of three pilot floating wind farm projects (Leucate, Gruissan and Provence Grand Large) in the gulf of Lion. These projects will rely on 3 full scale 8-10 MW floating turbines, whose generated power will be injected in the French power grid by 2022-2023 (Table 1). In addition, two commercial wind farms with power capacity over 250 MW each will be in operation by 2029.

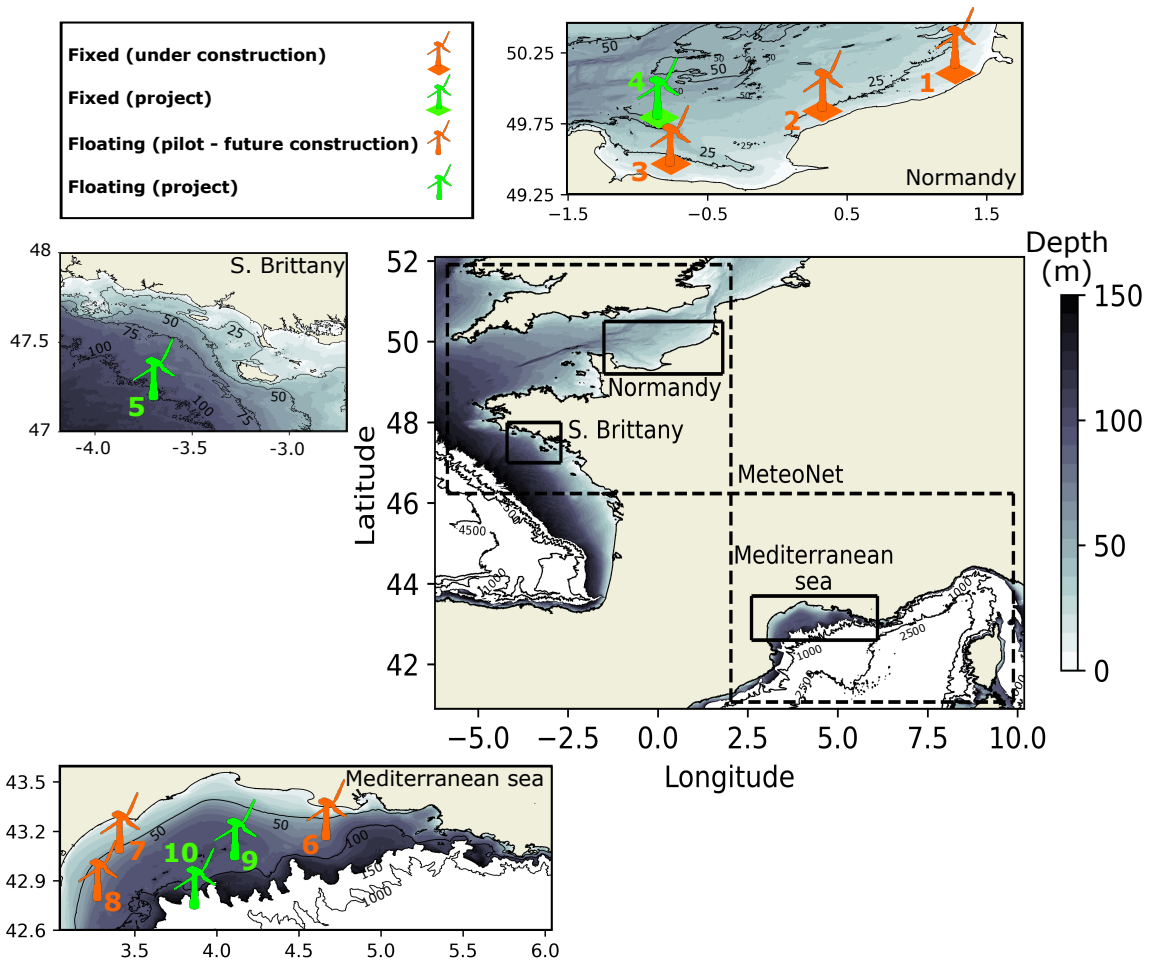

Figure 1. Overview of the French coasts with the bathymetry. The color shading shows the water depth until 150 m. Areas with water depth exceeding 150 m are shown in white. Black contours are used to identified depth of 1000m, 2500 m and 4500 m. The three study areas are shown by black rectangles. Each area is presented on different panels where the locations of the future foundation-based and floating wind farms are shown with their different stages of development. On the main panel, the dashed black lines delimits the areas covered by the MeteoNet data set.

2.2 The MeteoNet data set

MeteoNet is a meteorological data set developed and made available by Meteo-France (Larvor et al., 2020), the French national meteorological service. The data set contains full time series of satellite and radar images, NWP models and ground observations. The data covers two geographic areas of 550km x 550km on the Mediterranean and Brittany coasts (Fig. 1), and spans from 2016 to 2018. Hourly 10-meter wind output of the high resolution NWP model AROME are available. AROME is operational at Meteo-France since December 2008 (Seity et al., 2011). It was designed to improve short range forecasts of

Table 1. Characteristics of the foundation-based (Normandy) and floating (Southern Brittany, Mediterranean sea) future wind farms planned for the next decade on the study areas (CEREMA, 2022).

Areas	Wind Farm	Index in Fig. 1	Capacity (MW)	Number of wind turbines	Expected date of commissioning
Normandy	Dieppe- Le Tréport	1	496	62	~~2024~~2026
	Fécamp	2	497	71	2023
	Courseulles-sur-mer	3	448	64	~~2024~~2025
	AO4 call for tender	4	1 000	N/A	~~2028-29~~2030
Southern Brittany	~~Groix-Belle-Ile 5 28.5 3~~ ~~2022-23~~AO5 call for tender	~~6~~ 5	250	N/A	~~2028-2029~~ 2025-2030
Med. sea	Faraman - Port-Saint-Louis-du-Rhône	~~7~~ 6	24	3	~~2022~~2023
	Gruissan	~~8~~ 7	30	3	~~2022~~ 2024
	Leucate - Le Barcarès	~~9~~ 8	30	3	~~2022~~ 2024
	AO6 call for tender	~~10 and 11~~ 9 and 10	2 x 250	N/A	2028-2029

severe events such as intense Mediterranean precipitations, severe storms, fog or urban heat during heat waves. The physical parametrisations of the model come mostly from the Méso-NH research model whereas the dynamic core comes from the Non-Hydrostratic model ALADIN (Termonia et al., 2018). The resolution of the AROME grid is 1.3 km. The model is initialized from data assimilation derived from the ARPEGE-IFS variational assimilation system (Courtier et al., 1994) and adapted to the AROME finer resolution.

For each area of interest, the 10-meter zonal (u) and meridional (v) wind speed ~~were~~ are extracted from AROME. The open-source MeteoNet data set only contains surface parameters of temperature, humidity, pressure and precipitation, and 10-meters wind speed (u10, v10), which are considered in this study. The assumption is then made that relevant measurement points at 10 meters are equally relevant for hub height estimation, though this assumption should be tested with a suitable data set. Since the focus is on offshore wind, grid points at land were excluded from the analysis. The characteristics of each area are then the following:

- Normandy: 4272 grid points ($\sim$ 7 000 km^2).
- Southern Brittany: 1837 grid points ($\sim$ 3 000 km^2).
- Mediterranean sea: 3571 grid points ($\sim$ 5 800 km^2).

A total of 65 days ($\sim 6\%$) of the 3-year data set are unusable due to largely missing data. The missing data days are similar for each area and were removed from the analysis.

3 ~~BackgroundThe problem tackled in this paper is the finding of an optimal network of sensors to reconstruct offshore wind fields. It uses NWP gridded data of zonal and meridional wind above the study areas. The finding of~~ Preliminaries

3.1 Problem statement

The problematic of the presented work is to find D ~~optimal input points from~~ measurement points out of K ~~grid points consists of a combinatorial optimization problem, the exhaustive search of which is computationally intractable~~NWP grid points to minimize the reconstruction error of the offshore wind field. A formalism for this sparse sampling problem is proposed in this section.

In all that follows, the ~~NWP wind field over the study area $\mathbf{X}(t) \in \mathbb{R}^{2K}$ is the concatenation of~~ full state matrix $\mathbf{X}$ refers to the concatenation of zonal and meridional wind speeds on the ~~bi-dimensional wind speed at the~~ K grid points of the model ~~. The target is then expressed as the concatenation of reduced basis of r components for zonal and meridional wind speed. This target is to be reconstructed from a limited number of measurements $\mathbf{Y}(t) \in \mathbb{R}^D$. The~~ for all time steps. The list of sensors' locations γ, which is the output of the methods described in this ~~section try to define a set of input grid points locations $[\gamma_1, ..\gamma_D] \in [1, K]^D$ so that the wind field reconstructed from $\mathbf{Y}(t)$ best fits the target state of the system $\mathbf{X}(t)$.~~

3.2 ~~Problem statement~~

~~Let us consider a system described by its time-varying state $\mathbf{X}(t)$ that evolves according to unknown non-linear dynamics. It can be described on an orthogonal basis $\{\Phi_p\}$ as:~~

$$\mathbf{X}(t) = \sum_{p=1}^{P} \mathbf{a}_p(t)\phi_p$$

~~To reduce the complexity of the model, the state of the system can be approximated using the first r modes. The number of modes to use can be chosen using a threshold on the total variance of the data set such as:~~

$$\mathbf{X}(t) \approx \sum_{p=1}^{r} \mathbf{a}_p(t)\phi_p = \mathbf{X}^r$$

~~Leading to :~~

$$\mathbf{X}^r = \mathbf{\Phi}^r \mathbf{a}$$

~~Where $\mathbf{\Phi}^r$ is the reduced basis, r is the number of modes kept for approximation, and $\mathbf{a}_p \in \mathbb{R}^r$ are the time varying coefficients of the system's state on the reduced basis. The time index is omitted for simplicity.~~

The given system is then sampled according to a sampling matrix $\mathbf{C} \in \mathbb{R}^{D \times K}$ that extracts paper, contains the locations of the D input points out of K points from the grid. The sampling matrix is composed of lines of zero with ones at the sensors' locations. Given a set of locations $\gamma = [\gamma_1, ..\gamma_D] \in [1, K]^D$, $\gamma_i \neq \gamma_j$, and the associated canonical basis vectors $\mathbf{e}_{\gamma_i} = (\delta_{\gamma_i,k}) \in \mathbb{R}^K$, the sampling matrix is: sensors to sample the offshore wind field. The associated sparse measurement matrix $\mathbf{Y}_\gamma$ corresponds to the measured zonal and meridional wind speed at the γ locations for every time step.

$$\mathbf{C}_\gamma = \begin{bmatrix} \mathbf{e}_{\gamma_1} & \mathbf{e}_{\gamma_2} & \cdots & \mathbf{e}_{\gamma_D} \end{bmatrix}^T$$

The measured matrix containing the wind speeds measured at the locations defined in the $\mathbf{C}_\gamma$ matrix is then obtained by multiplying the full state by the sampling matrix. This measurements can then be approximated using the first r modes of the system:

$$\mathbf{Y} = \mathbf{C}_\gamma \mathbf{X} \approx \mathbf{C}_\gamma \mathbf{\Phi}^r \mathbf{a}$$

From the measurement matrix, the full state is reconstructed using the reduced basis coefficients $\mathbf{a}$ that can be obtained with the Moore-Penrose pseudo inverse of The formalism developed in this section is applied to the $\mathbf{C}_\gamma \mathbf{\Phi}^r$ matrix, noted $(\mathbf{C}_\gamma \mathbf{\Phi}^r)^\dagger$.

$$\mathbf{a}_{recons} = \mathbf{Y}(\mathbf{C}_\gamma \mathbf{\Phi}^r)^\dagger$$

This reconstruction $\mathbf{\Phi}^r \mathbf{a}_{recons}$ is then compared to the reconstruction with perfect knowledge on the reduced basis coefficients $\mathbf{\Phi}^r \mathbf{a}$, assuming that the actual coefficients of the reduced basis are perfectly known. This is considered as the ground truth. MeteoNet data set presented in section 2.2. The data set is split into a training and testing period. The training is performed on two thirds of the data set, composed of years 2016 and 2017, while the methods are scored on year 2018. By taking an integer number of year, the seasonality bias of weather data is limited.

Given a number of sensors D, the optimisation problem can then be stated as the minimization of the reconstruction error over all position combinations γ listed in the $\mathbf{C}_\gamma$ sampling matrix:

$$\underset{\gamma, D}{\arg\min} < \|\mathbf{\Phi}^r (\mathbf{C}_\gamma \mathbf{\Phi}^r)^\dagger \mathbf{Y} - \mathbf{\Phi}^r \mathbf{a}\|_2 >$$

With $< . >$ being the temporal mean. The reconstruction error is computed with regards to the reduced basis itself. The search of the optimal sampling matrix $\mathbf{C}_\gamma$ is a $\binom{K}{D}$ problem, intractable for the dimensions of this problem.

For what follows, the $\mathbf{Y}$ matrix is the concatenation of the zonal and meridional components of the wind speed on the K grid points as function of time.

3.2 Reduced order model

The reduced order model used to decrease the dimension of the input data is the Empirical Orthogonal Function analysis (EOF). Also known as Principal Component Analysis (PCA), it decomposes the data set onto an orthogonal basis. Practically, it is linked to the singular value decomposition of a matrix $\mathbf{X}$ such that:

$$\mathbf{X} = \mathbf{U}\underline{\mathbf{\Sigma}}\mathbf{\Sigma}\mathbf{V}^T \tag{1}$$

~~With $\mathbf{\Sigma}$~~ With $\mathbf{\Sigma}$ a diagonal matrix of positive σ_k singular values, $\mathbf{U}$ a matrix whose columns are ~~orthogonal, known as singular vectors,~~ the vectors of the orthogonal basis, and $\mathbf{V}$ the ~~loading~~ weights of the associated vectors.

The singular vectors are ~~chosen iteratively as the~~ orthogonal vectors on which the variance of the projected data is maximized. The diagonal elements of ~~$\mathbf{\Sigma}$~~ $\mathbf{\Sigma}$ are sorted per value, and are equal to the percentage of variance of the data set explained by each principal components. The variance explained by the first r EOF is then:

$$\underline{\frac{\sum_{i=1}^{r}\sigma_i^2}{\sum_j \sigma_j^2}}\ \frac{\sum_{i=1}^{r}\sigma_i^2}{\sum_i^K \sigma_i^2} \tag{2}$$

The number of EOF to use in the reduced order model can be set so that the variance explained by the reduced basis is above a certain threshold. ~~In this work, the data~~

For the study dataset, EOF of zonal and meridional wind speed are computed. In the NWP model, the grid points are strongly correlated spatially, hence, only a low number of EOF is needed to describe the vast majority of the data set variance. The number of EOF was set to 10, both for the zonal and meridional components of wind so that the reduced basis explains more than 95% of the total variance for the 3 areas.

~~For what follows the~~ The $\mathbf{\Phi}^r$ reduced basis is then the concatenation of the $\mathbf{\Phi}_u^{r/2}$ and $\mathbf{\Phi}_v^{r/2}$ ~~reduced basis~~ EOF for zonal and meridional wind speed, with $r = 20$.

3.3 Sparse sampling

3.3.1 State description

Let us consider a system described by its time-varying state $\mathbf{X}(t)$ that evolves according to unknown non-linear dynamics. It can be described on an orthogonal basis $\{\phi_i\}$ (e.g. the EOF) as:

$$\mathbf{X}(t) = \sum_{i=1}^{K} a_i(t)\phi_i \tag{3}$$

To reduce the complexity of the model, the state of the system can be approximated using the first r modes:

$$\mathbf{X}(t) \approx \sum_{i=1}^{r} a_i(t)\phi_i = \mathbf{X}^r(t) = \mathbf{\Phi}^r\mathbf{a}(t) \tag{4}$$

Where $\mathbf{\Phi}^r$ is the reduced basis matrix containing the first r modes, and $a_i(t)$ are the time varying coefficients of the system's state on the reduced basis.

The given system is then sampled according to a set of index $\boldsymbol{\gamma} = [\gamma_1, .. \gamma_D] \in [1, K]^D, \gamma_i \neq \gamma_j$ which represents the sensors' locations. From this, a sampling matrix is constructed $\mathbf{C}_{\boldsymbol{\gamma}} \in \mathbb{R}^{D\times K}$ that extracts the D measured locations out of the K grid

points of the full state. The sampling matrix is composed of lines of zero with ones at the sensors' locations. With the canonical basis vectors $\mathbf{e}_{\gamma_j} = (\delta_{\gamma_j,k}) \in \mathbb{R}^K$, the sampling matrix is:

$$\mathbf{C}_{\boldsymbol{\gamma}} = \begin{bmatrix} \mathbf{e}_{\gamma_1} & \mathbf{e}_{\gamma_2} & \cdots & \mathbf{e}_{\gamma_D} \end{bmatrix}^T \tag{5}$$

The sparse measurement matrix $\mathbf{Y}_{\boldsymbol{\gamma}}$ is then obtained by multiplying the full state $\mathbf{X}$ by the sampling matrix $\mathbf{C}_{\boldsymbol{\gamma}}$:

$$\mathbf{Y}_{\boldsymbol{\gamma}}(t) = \mathbf{C}_{\boldsymbol{\gamma}}\mathbf{X}(t) \tag{6}$$

3.3.2 Full state reconstruction from sparse measurements

From the sparse measurement matrix, the full state is reconstructed using the coefficients of the reduced basis. A linear model is constructed to link the matrix of EOF coefficients, $\mathbf{a}$, to the sparse measurement matrix $\mathbf{Y}_{\boldsymbol{\gamma}}$:

$$\mathbf{a} = \boldsymbol{\beta}\mathbf{Y}_{\boldsymbol{\gamma}} + \epsilon \tag{7}$$

With ϵ an additive Gaussian error.

The model fitting is performed on the training split of the data set. Let $\mathbf{Y}_{\text{train}}$ be the sparse measurement matrix on the training split, and $\mathbf{a}_{\text{train}}$ the true coefficients of the full state on the reduced basis for the training split. Using the Ordinary Least Squares formula, the $\boldsymbol{\beta}$ matrix can be estimated as:

$$\hat{\boldsymbol{\beta}} = (\mathbf{Y}_{\boldsymbol{\gamma},\text{train}}^T \mathbf{Y}_{\boldsymbol{\gamma},\text{train}})^{-1} \mathbf{Y}_{\boldsymbol{\gamma},\text{train}}^T \mathbf{a}_{\text{train}} \tag{8}$$

On the test data set, only the wind speed measurements at the $\boldsymbol{\gamma}$ locations are available. The coefficients of the reduced basis are computed using the least squares matrix estimated on the training data set:

$$\hat{\mathbf{a}}_{\boldsymbol{\gamma},\text{recons}} = \hat{\boldsymbol{\beta}}\mathbf{Y}_{\boldsymbol{\gamma},\text{test}} \tag{9}$$

And the full state is reconstructed using the reduced basis:

$$\hat{\mathbf{X}}_{\boldsymbol{\gamma},\text{recons}} = \boldsymbol{\Phi}^r \hat{\mathbf{a}}_{\boldsymbol{\gamma},\text{recons}} \tag{10}$$

3.3.3 Reconstruction error

The reconstruction $\hat{\mathbf{X}}_{\boldsymbol{\gamma},\text{recons}}$ is then compared to the reconstruction with perfect knowledge on the reduced basis coefficients $\mathbf{X}_{\text{real}} = \boldsymbol{\Phi}^r \mathbf{a}_{\text{real}}$, assuming that the actual coefficients of the reduced basis are perfectly known.

The reconstruction error associated with the sensors' locations $\boldsymbol{\gamma}$ is then the root mean squared error of the reconstructed state:

$$E_{\boldsymbol{\gamma},\text{recons}} = \frac{1}{T}\sum_{t=1}^{T} \sqrt{\frac{1}{K}\sum_{k=1}^{K} \left(\hat{\mathbf{X}}^k_{\boldsymbol{\gamma},\text{recons}}(t) - \mathbf{X}^k_{\text{real}}(t)\right)^2} \tag{11}$$

The optimisation problem that needs to be solved can then be stated as the minimization of the reconstruction error over all locations' combinations γ and number of sensors D:

$$\arg\min_{\gamma,D} E_{\gamma,\mathrm{recons}} \tag{12}$$

4 Sparse sampling methods used in this study

In this section, the methods applied for the ~~sensors' locations selection~~ sparse sampling are described in detail. The novel data-driven method based on Gaussian Mixture Model is presented ~~alongside~~ together with the three baselines emerging from the literature review. These are the random selection of locations (Monte Carlo), the dominant spatial modes' extrema (EOF extrema), and the QR greedy algorithm (QR pivots). All the methods described in this section should output a list of sensors' locations γ given a number of sensors D.

4.1 Baseline methods

The selected baseline methods are emerging from the literature as simple yet efficient methods for sparse sampling in ~~numerous~~ different situations. They are implemented to ~~measure the addition of the~~ compare their performances with the Gaussian Mixture Model ~~in~~ for this specific application.

4.1.1 Monte Carlo simulations

The first baseline consists in picking random sensors' locations. For each area, and for a number D of sensors ranging from 1 to 10, a hundred random combinations of locations ~~$\gamma \in P_D([1,K])$~~ $\gamma \in P_D([1,K])$ are considered. For each ~~γ~~ γ combination of sensors' locations, the reconstruction error is computed. From this ensemble of simulations, statistics on the reconstruction error are computed.

The median Monte-Carlo scenario for each area and number of sensors is then considered a benchmark for the study. It also gives information about the spread in reconstruction error resulting from all possible combinations.

4.1.2 Dominant spatial modes' extrema

In Yildirim et al. (2009), the extrema of the spatial dominant modes are found to be relevant locations if not optimal for the reconstruction of the flow field. Those points can be seen as salient points, that best characterize the spatial modes. It is then intuitive to select those to reconstruct the full state from the reduced basis. How many extrema are chosen from each variable and mode is studied specifically in Cohen et al. (2003), it is empirical and thus case specific.

In the case study of Cohen et al. (2003), the EOF decomposition gives modes that are highly spatially correlated. Moreover, in this study, points nearby the coast are influenced by the orography and show strong variability. Hence, sorting the points per coefficient and selecting the N first ones will lead to the selection of neighboring points, and/or irrelevant coastal points for our performance metric.

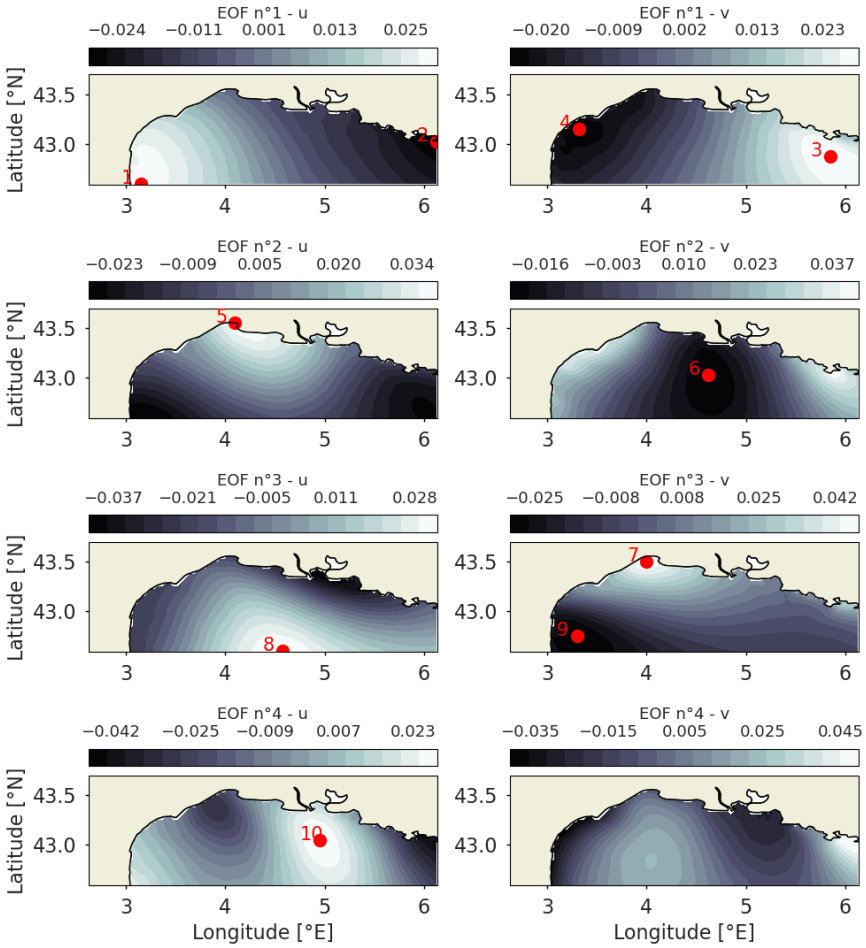

Figure 2. Selected sensors for the EOF extrema baseline in the Mediterranean Sea. The EOF coefficients are displayed as background, and the selected salient points, and their associated ranks are displayed as red dots. The two columns are zonal (u) and meridional (v) wind speed, and the rows correspond to the EOF rank.

The extrema are then chosen manually, as performed in Yildirim et al. (2009), from the visualisation of the first EOF for both zonal and meridional wind speed. For each parameter and EOF rank, the extrema are selected, and discarded if redundant (manual process). Then, they are sorted per absolute value and per EOF number for the two parameters.

285 The selected input points from EOF extrema for the Mediterranean Sea are shown in Fig. 2. From the first to the fourth EOF for zonal and meridional wind, the extrema are selected if they are not too redundant or close to the coast / border. This unsatisfactory workflow is a way to ensure minimum relevance for the obtained sensors array.

~~Selected sensors for the EOF extrema baseline in the Mediterranean Sea. The EOF coefficients are displayed as background, and the selected salient points, and their associated ranks are displayed as red dots. The two columns are zonal (u) and meridional (v) wind speed, and the rows correspond to the EOF rank.~~

4.1.3 QR pivots

The QR decomposition of the reduced basis matrix $\mathbf{\Phi}^r$ is the finding of two matrix $\mathbf{Q}$ orthogonal and $\mathbf{R}$ triangular superior such as:

$$\mathbf{\Phi}^r = \mathbf{QR} \tag{13}$$

The $\mathbf{Q}$ and $\mathbf{R}$ matrix are obtained using Gram-Schmidt process (Leon et al., 2013), which consists in iteratively removing each column's orthogonal projection onto a pivot column. The QR algorithm can be performed using column pivoting, i.e., at each iteration, the matrix $\mathbf{\Phi}^r$ is multiplied by a permutation matrix $\mathbf{P}$ such that the column taken for pivoting has the maximum two-norm. The decomposition is then:

$$\mathbf{\Phi}^r \mathbf{P}^T = \mathbf{QR} \tag{14}$$

The permutation matrix $\mathbf{P}$ is constructed so that the diagonal elements of $\mathbf{R}$ are decreasing. It is applied to the matrix of the reduced basis to identify pivot locations. It then contains the ranked index of sensors' locations to build the ~~sampling matrix.~~ sensors' locations list:

$$\gamma_j = \mathbf{P}_{jj} \quad \forall j \in [1, D] \tag{15}$$

The QR pivots method is described in Manohar et al. (2018) as a simple yet efficient method for sparse sensors' placement. It is used to determine model data driven sensors networks to reconstruct flow fields for the flow past a cylinder and the sea surface temperature retrieval, two situations that are analogous to the case study. It is even tuned to include costs constraints for the search of Pareto optimal sensors' placement in Clark et al. (2020). For wind fields estimation, it was applied to Computational Fluid Dynamics data in Annoni et al. (2018) to best reconstruct the flow in a wind farm. All in all, it represents a simple yet competitive baseline method for spare sensor placement.

4.2 Gaussian Mixture Model clustering

The proposed method in this study uses unsupervised clustering of the data to define sensors' locations. Gaussian Mixture Models use machine learning to fit multivariate normal distributions on the data.

4.2.1 Gaussian mixture

A Gaussian mixture model (GMM) is a probabilistic model for representing normally distributed sub-populations within an overall population (Reynolds, 2009). ~~The model is a mixture of multivariate normal distributions, each distribution representing~~

Each Gaussian distribution represents a group of points. Each point is then assigned to the distribution with highest likelihood, hence splitting the data points into clusters. GMM can be used in an unsupervised framework, allowing the model to select clusters automatically.

A GMM p is a weighted sum of M multivariate Gaussian distributions h:

, i.e., cluster. The model is a mixture, i.e., superposition, of multivariate Gaussian components which define a probability distribution $p(x)$ on the data:

$$p(\mathbf{x}) = \sum_{i=1}^{M} w_i h \sum_{j=1}^{D} \pi_j \mathcal{N}(\mathbf{x}|\mu_i \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_j) \tag{16}$$

where the multivariate Gaussian distribution in a D-dimensional space is given by: π_j being the mass of the Gaussian component j, with $0 \leq \pi_j \leq 1$ for all $j = 1, ..., D$ and $\sum_{j=1}^{D} \pi_j = 1$. $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ being the Gaussian density distribution such that:

$$h\mathcal{N}(\mathbf{x}|\mu\boldsymbol{\mu}, \boldsymbol{\Sigma}\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^{\mathrm{D}}|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)} \frac{1}{\sqrt{(2\pi)^{\mathrm{r}} det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{17}$$

with $\mathbf{x}$ being the ~~D-dimensional~~ r-dimensional input vector, ~~μ the D-dimensional~~ $\boldsymbol{\mu}$ the r-dimensional mean vector, and ~~$\boldsymbol{\Sigma}$ $(D \times D)$~~ $\boldsymbol{\Sigma}$ $(r \times r)$ the covariance matrix.

4.2.2 Expectation — Maximization algorithm

The core of GMM lies within Expectation Maximization (EM) algorithm, developped by **?**. It iteratively modifies the model's parameters to maximize the log-likelihood of the data.

The ~~number M of Gaussian components is an input of the model. Given this number of components,~~ log-likelihood, $\log(\mathcal{L})$, of the observations is given by:

$$\log(\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})) = \sum_{k=0}^{K} \log\left(\sum_{j=1}^{D} \pi_k \mathcal{N}(\mathbf{x}_k|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\right) \tag{18}$$

Then the empirical means, $\boldsymbol{\mu}_j$, covariances, $\boldsymbol{\Sigma}_j$ and weights, π_j of the different clusters are computed. The weights (mixing coefficients) represent the mass of the different clusters. The mass of the cluster is the proportion of data points assigned to this cluster. For the first iteration, the mean and covariance matrices are initialized ~~. Then,~~ randomly, and the weights matrix is equal for each cluster.

The second step of the algorithm is the ~~Expectation – Maximization algorithm is used to tune the components' parameters to maximize the likelihood. During the~~ expectation step, ~~the probability for each point to belong to each distribution is computed . Then the distributions' parameters are updated. These steps are repeated until the likelihood is constant.~~ E-step. The model parameters are updated to increase the log likelihood of the data. For each data point, x_k, the probability that this point belongs

to the cluster, c, is computed such that:

$$r_{kc} = \frac{\pi_c \mathcal{N}(\mathbf{x}_k|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)}{\sum_{j=1}^{D} \pi_j \mathcal{N}(\mathbf{x}_k|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{19}$$

~~Fig. 3 shows the workflow in this study. A two-dimensional dataset composed of $K = 3571$ grid points with 20 EOF features is used to feed the GMM.~~ The ~~20 features are composed of the 10 first EOF of zonal and meridional velocities. The clustering is then optimized spatially, so the entries (~~E-step computes those probabilities using the current estimates of the model's parameters. In this step, "responsibilities" of the Gaussian distributions are computed. They are represented by the variables r_{kc}. The responsibility measures how much the c-th Gaussian distribution is responsible for generating the k-th data point using conditional probability.

The third step is the maximization step, M-step. In this step, the ~~grid points) are assigned to clusters, based on their features (their coefficients on the first 20 EOF). The output~~ algorithm uses the responsibilities of the Gaussian distributions (computed in the E-step) to update the estimates of the model~~is then a list of labels for each grid points~~'s parameters. π_c, ~~creating spatial clusters in the study areas~~$\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ are updated using the following equations:

$$\pi_c = \frac{\sum_{k=1}^{K} r_{kc}}{K} \tag{20}$$

$$\boldsymbol{\mu}_c = \frac{\sum_{k=1}^{K} r_{kc} \mathbf{x}_k}{\sum_{k=1}^{K} r_{kc}} \tag{21}$$

$$\boldsymbol{\Sigma}_c = \frac{\sum_{k=1}^{K} r_{kc} (\mathbf{x}_k - \boldsymbol{\mu}_c)^2}{\sum_{k=1}^{K} r_{kc}} \tag{22}$$

These updated estimates are used in the next E-step to compute new responsibilities for the data points. This algorithm is applied iteratively until algorithm convergence, when the log likelihood of the data is maximized. The strict monotony of the likelihood in the E-M algorithm is demonstrated in **?**.

~~GMM require~~

4.2.3 Optimal number of clusters

GMM requires to impose as input the number of ~~components, i.e., the number of clusters ,~~ clusters in the model. The optimum number of clusters can be defined through the calculation of the Bayesian ~~information criterion~~ Information Criterion (BIC) score (Schwarz, 1978):

$$\mathrm{BIC} = -2\ln(\mathcal{L}) + G\ln(NK) \tag{23}$$

with $\mathcal{L}$, the maximized value of the likelihood function of the model, G the number of parameters ~~estimated by the model, i.e. the mean~~ in the mean vectors and covariance matrices of the Gaussian components, and ~~N~~ K, the number of data points~~,~~

~~i.e. the number of observations~~. The BIC score penalizes too complex models to avoid the over-fitting of the data set. In this way, it limits the number of components of the GMM with the ~~$G\ln(N)$ termwhile the likelihood could be maximized with one Gaussian component per grid point.~~ $G\ln(K)$ term.

The lower is the BIC, the better is the model. However, the curve of the BIC score can be monotone, and the identification of a minimum, i.e., the optimal number of clusters can be difficult. An alternative is the calculation of the gradient of the BIC score. The identification of the optimal number of clusters is hence done by the identification of an elbow in the curve of the gradient of the BIC score. The elbow is often not directly associated with one single specific number of clusters but rather encompassed two or three possible solutions. Thus, one can say that the gradient of the BIC score gives an indication on the range of optimal number of clusters. An extra step is required to determine the optimal number of clusters. In this study, this step is done through the determination of an error reconstruction threshold of the wind field. The number of clusters associated with the minimum error is considered as optimal for the GMM.

4.2.4 Implementation for the study case

Fig. 3 shows the workflow in this study. A two-dimensional data set composed of $K = 3571$ grid points with 20 EOF features is used to feed the GMM. The 20 features are composed of the 10 first EOF of zonal and meridional velocities. The clustering is then optimized spatially, so the entries (the grid points) are assigned to clusters, based on their features (their coefficients on the first 20 EOF). The output of the model is then a list of labels for each grid points, creating spatial clusters in the study areas.

The GMM procedure will find clusters of grid points that are correlated in the reduced basis. The centroids of the clusters, i.e., the point of maximum likelihood for a given cluster, are then chosen to be sensors' locations, as they are the most representative points of the clusters~~.~~

4.3 ~~Scoring~~

~~The data set is split into a training and testing period. The training is performed on two thirds of the data set, composed of years 2016 and 2017, while the methods are scored on year 2018. By taking an integer number of year, the seasonality bias of weather data is limited.~~

~~Given a set of locations γ, the reconstruction error is equal to the root mean squared error averaged over time between the reconstructed state of the system and the true state as approximated in the reduced basis. At each time step, the measured wind speed is $\mathbf{Y}(t)$ and the estimated coefficients on the reduced order basis are $(\mathbf{C}_\gamma\mathbf{\Phi}^r)^\dagger\mathbf{Y}(t)$. Going back to the $latitude \times longitude$ space, the zonal and meridional wind speed at every point of the grid at time t are: :~~

$$\mathbf{X}^{\text{recons}}(t)\gamma_j = \mathbf{\Phi}^r\mathbf{C}_\gamma\mathbf{\Phi}^{r\dagger}\mathbf{Y}\arg\max_{\mathbf{x}}\mathcal{N}(t\mathbf{x}|\boldsymbol{\mu}_j,\mathbf{\Sigma}_j) \quad \forall j \in [1, D]. \tag{24}$$

~~The real wind speed from the reduced basis coefficients at time t is:~~

$$\mathbf{X}^{\text{real}}(t) = \mathbf{\Phi}^r\mathbf{a}(t)$$

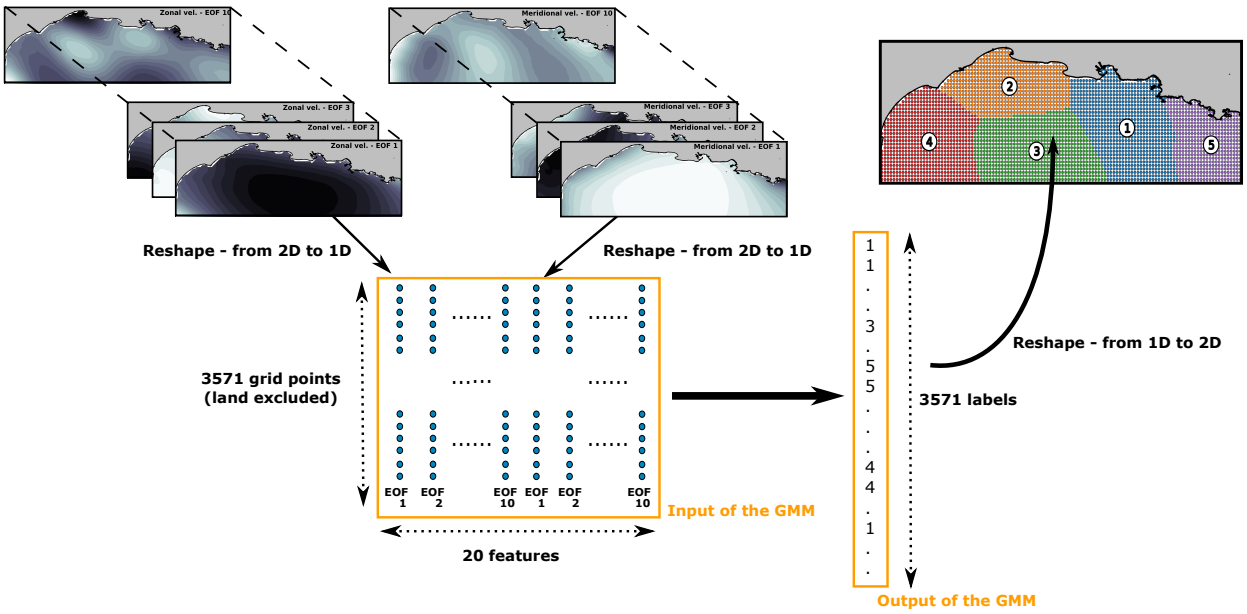

Figure 3. Schematic of the clustering procedure of wind data using Gaussian Mixture Models. It is illustrated for the unsupervised clustering of the Mediterranean Sea wind field.

~~The reconstruction error associated with the sampling matrix $\mathbf{C}_\gamma$ is then the temporal mean of the root mean squared error of the zonal and meridional wind speed over the full grid:~~

$$E_\gamma^{\text{recons}} = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{2K}\sum_{i=1}^{2K}\left(X_{\gamma,i}^{\text{recons}}(t) - X_i^{\text{real}}(t)\right)^2}$$

5 Results

In this section, the methods presented in section 4 are implemented on the three identified areas (Mediterranean Sea, Normandy and Southern Brittany) and compared with respect to the wind field reconstruction error. A method for the selection of an optimal number of sensors is described, and the suggested sensors' locations for the three areas are given.

5.1 Optimal number of sensors

The number of sensors to place on the grid is an input of the GMM. The BIC score described in section ~~4.2.1~~ 4.2.3 computes a trade-off between the likelihood of the obtained distribution, and the complexity of the model. Being sensible to the likelihood of the model and to its complexity, it is usually used to determine the number of clusters for the GMM by finding its minimum. However, there is no guarantee that there will be a minimum BIC score corresponding to an optimal number of clusters, and there is no guarantee that this number of clusters is actually optimal for the considered metric. Indeed, this metric is a heuristic

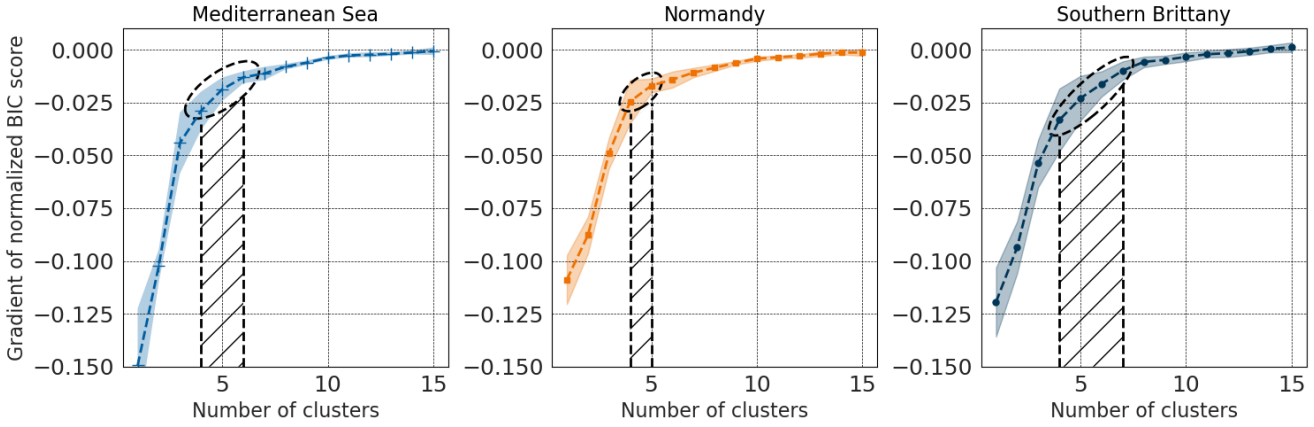

Figure 4. The gradient of normalized BIC score is shown for the three areas, for a number of clusters ranging from 1 to 15. The curves' envelopes are the 95% confidence interval obtained from 25 different initializations of the GMM training. The determination of an optimal number of sensors from these curves is uncertain, ranging from 4 to 6 for the Mediterranean Sea (shaded area), 4 to 5 for Normandy and 4 to 7 for Southern Brittany.

criterion to hint the trade-off between accuracy and complexity, to avoid over-fitting. If there is no minimum to the BIC score, one can look for an elbow in the BIC score's gradient, showing a number of clusters after which the marginal gain of BIC score is no longer significant. In this study, the BIC score showed no minimum up to 50 clusters, so its gradient was studied. However this technique is not very accurate, and the results should be interpreted carefully. For example, the knee identification is very dependant on the cut-in and cut-out of the curve for the definitions of the asymptotes. Furthermore, the GMM results are dependant on the initialization of the algorithm. As shown in Fig. 4, the obtained optimal number of sensors can range between 4 and 7, though it shows clear convergence for a number of clusters above 10. The gradient of BIC score was computed for 20 random GMM initializations for the three areas, and the mean gradient plotted as dashed line, with its 95% confidence interval as envelope. The BIC score was normalized to compare the three areas together. Similar trends can be observed, with stronger gradients in the Mediterranean Sea for the first clusters showing a bigger underlying complexity. For Southern Brittany, the associated uncertainty is bigger, showing weaker global minimum for the Expectation Maximization.

While the BIC score gives an indication on the range of optimal number of clusters, it does not necessarily translates into equivalent reconstruction for the wind fields. Although the clustering itself might find an optimum of 5 clusters for the Mediterranean Sea, this can lead to much higher reconstruction error than for the other areas as illustrated in Fig. 5(a). In particular for the Mediterranean Sea, the considered region is wider with several different wind regimes, which implies a higher variability. It then seems natural that more sensors than other areas would be needed to reach the same error level. Furthermore, the uncertainty on the optimal number of sensors shows an underlying property of this spatio-temporal data which has strong correlations between points, and for which clusters are not well separated. All in all, there is a need to cross-validate the computation of the optimal number of sensors. It is then proposed to validate the number of sensors from the

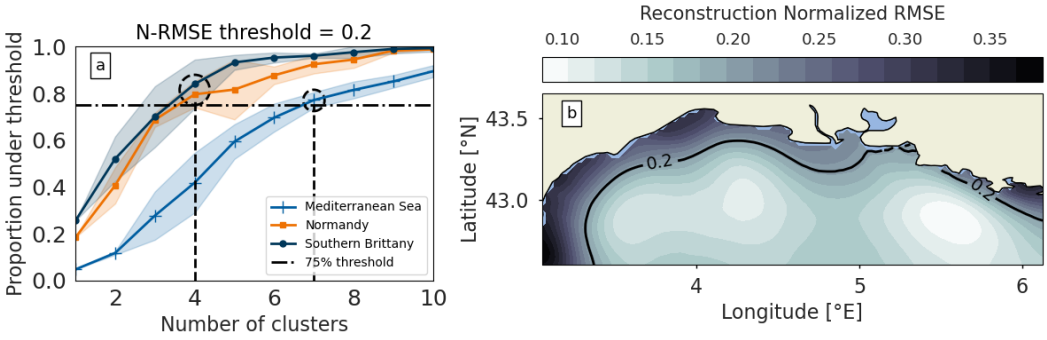

Figure 5. The computation of the proportion of the map under a certain error threshold for the three areas and 20 different GMM initializations allows for optimal number of sensors selection as the minimum number of sensors required to reach 75% of the map under threshold (a). The obtained reconstruction error map with the threshold contour shown illustrates the selection on the Mediterranean Sea (b).

computation of the reconstruction error. Exploring the range of number of clusters obtained through the BIC score gradient, the final number of sensors is chosen using a reconstruction error threshold.

To compare the three areas which have different wind regimes, the error threshold is defined as the reconstruction error of the normalized wind (Normalized Root Mean Squared Error or N-RMSE). The optimal number of clusters is then computed as the minimal number of clusters required to reconstruct 75% of the map with an error lower than the threshold.

It is then up to the final user to define an empirical error threshold to derive the optimal scenario. As shown in Fig. 5 (a), while the BIC score gradient curves are similar for the three areas, the normalized reconstruction error is significantly higher for the Mediterranean Sea for the same number of input points, thus necessitating a higher number of clusters to reach 75% of the map under threshold. The threshold of 0.2 normalized reconstruction error is shown in Fig. 5 (b). It yields to coherent results with regards to the BIC score analysis. The final numbers of clusters are then 4 for Normandy and Southern Brittany and 7 for the Mediterranean Sea. This workflow for the definition of the optimal number of sensors ensures similar performance between the three areas.

5.2 Clustering-derived sensors performance

The sensors' locations for the base case scenario with optimal number of sensors of 4 for Normandy and Southern Brittany and 7 for the Mediterranean Sea are then computed on the three areas for the four methods: Monte Carlo, QR pivoting, EOF extrema, and GMM.

The obtained sensors' locations are displayed in Fig. 6 as red dots. It can be visually noted that the sensors array derived from the GMM method (second row) is more evenly distributed than the benchmark sensors arrays. QR pivots locations (third row) are concentrated near the coast or at the maps' limits, and so are EOF extrema (fourth row). It shows how the GMM method allows for homogeneous sampling of the area, while benchmark methods tend to give too much weight to coastal and bordering points. This can be either artificial, due to spatial discontinuity at the limits of the maps, or because of the orographic

impact of the coast. Indeed, the wind near the coast shows more variability, and while those points are contained in wider spatial structures in the GMM, they can be considered as salient points in the QR pivoting method or in the EOF extrema.

The resulting reconstruction at different time steps is illustrated as background for the three areas and four methods in Fig. 6. The first row is the reference case, reconstructed with perfect knowledge on the 20 EOF coefficients (EOF reference). For the Mediterranean Sea, this specific time step shows a combined Mistral and Tramontane winds blowing in the Mediterranean Sea. It is a complex and standard situation with different wind regimes, strong offshore blowing winds in the North and West of the Gulf, and South-Eastern winds on the Eastern extremity. It can be noted that the GMM method correctly reproduces the intensity of those three phenomenons, while other techniques tend to overestimate or underestimate their effects.

For Normandy, the benchmark sensors array are largely off target on this specific case, predicting little to no wind offshore due to their exclusive coastal sampling, while the GMM method better captures both the coastal low winds and offshore wind cell.

For Southern Brittany, the effects of the sensors array is less clear, possibly explained by the smaller area, or by a simpler wind regime. However, the GMM method still performs largely better in terms of reconstruction error and wind patterns than benchmark methods.

Three different metrics are computed for the optimal scenarios on the three areas, and displayed in Table 2. Along with the reconstruction error described in Sect. **??**, the error in the reconstructed mean and maximum wind speed are displayed. For the three areas, the GMM method clearly leads to good reconstruction error and mean wind speed estimation. However, the EOF extrema method yields to better estimation of the maximum wind speed for Normandy and Southern Brittany. It illustrates the fact that the GMM method is good at reconstructing the synoptic situation, while discarding high variability points that can be relevant for extreme events. Indeed, coastal points that can have a high variability due to the coastal orographic effects, are selected as salient points by the EOF extrema and QR pivot, and discarded by the GMM that assign them to a wider cluster. This is efficient to reconstruct the mean situation in the whole map but can lead to higher errors on high variability areas.

The proposed GMM method is scored against the three baselines methods on the Mediterranean sea area, for a number of sensors ranging from 1 to 10. The results are displayed in Fig. 7, showing the great interest of the clustering derived method compared to benchmark methods for the offshore wind reconstruction from sparse sampling. QR pivoting sensors and EOF extrema sensors fail to surpass the Monte Carlo simulation for low number of sensors. The GMM method yield to reconstruction errors systematically below the minimum of the boxplots (i.e., first quartile minus 1.5 times the inter-quartile range which is equal to 99.65 % of the data in the Gaussian case.), showing the near-optimal reconstruction. The benchmark methods' errors eventually decrease for a high number of sensors and surpass the Monte Carlo median scenario for 10 sensors. However it is expected that the different methods should converge for high number of sensors, as the system is more and more constrained. It is illustrated by the decreasing spread within the Monte Carlo simulation.

For the GMM curve and the Monte Carlo boxplots, the reconstruction error seems to inflect for a number of sensors around 7, cross-validating the obtained optimal number of sensors for the base scenario. It can be noted that the reconstruction error for the EOF extrema method drastically decreases with the addition of the sixth sensor. As shown in Fig. 2, the sixth sensor is

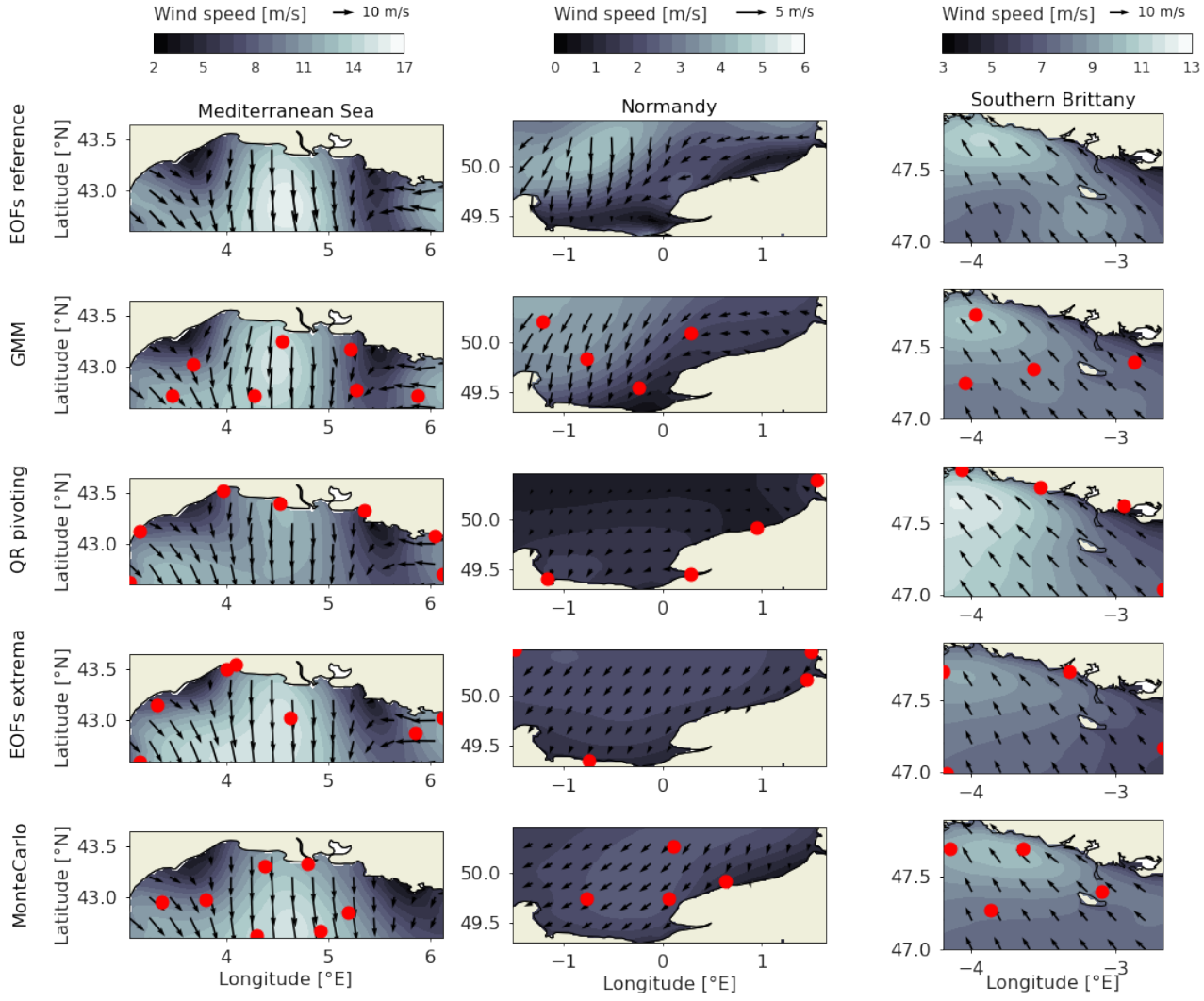

Figure 6. Reconstruction example for the optimal scenario on the three areas, from the reduced basis (EOF reference), from GMM clustering (GMM), and from the 3 baselines, QR pivoting, EOF extrema and Monte Carlo. The color grading shows the wind speed, the black arrows the wind direction, with length proportional to the wind speed, and the red dots are the locations of the sensors on the optimal scenario, with 7 sensors on the Mediterranean Sea, and 4 for Normandy and Southern Brittany.

a central offshore point, while the 5 first locations are near the coast. For number of sensors above 6, the EOF extrema method then compares to the Monte Carlo median scenario.

For low and optimal number of sensors, compared to state-of-the-art techniques, the GMM method allows for the efficient sensors' placement for offshore wind reconstruction. The obtained reconstruction errors are displayed in Table 3, along with the RMSE gain relative to the Monte Carlo median score. In the three areas, the GMM method improves significantly the

Table 2. Reconstruction errors computed for the 3 areas and the 4 sampling methods, including the random scenario displayed in Fig. 6. The best performing method is displayed in bold for each area. The reconstruction error (RMSE) and the errors on the max and mean wind speed at each time step are computed.

Area	Method	Max wind speed RMSE [m/s]	Mean wind speed RMSE [m/s]	RMSE [m/s]
Meditteranean Sea	GMM	**0.94**	**0.17**	**0.9**
	QR pivoting	1.42	0.42	1.77
	EOF extrema	1.28	0.2	1.37
	Monte Carlo	1.28	0.35	1.41
Normandy	GMM	2.0	**0.1**	**0.85**
	QR pivoting	1.84	0.56	1.83
	EOF extrema	**0.89**	0.42	1.44
	Monte Carlo	1.4	0.23	1.08
Southern Brittany	GMM	1.32	**0.09**	**0.7**
	QR pivoting	2.04	0.29	1.33
	EOF extrema	**0.89**	0.16	0.96
	Monte Carlo	1.87	0.17	0.93

Table 3. RMSE and RMSE percentage gain versus Monte Carlo median value for the base case scenario on the three areas, for number of clusters of 7 for the Mediterranean Sea and 4 for Normandy and Southern Brittany. The bold numbers show the best performances.

Score	Mediterranean Sea RMSE [m/s] | % gain	Normandy RMSE [m/s] | % gain	Southern Brittany RMSE [m/s] | % gain
GMM	**0.99 | -22%**	**0.90 | -24%**	**0.82 | -13%**
QR pivoting	1.60 | +27%	1.83 | + 55%	1.33 | + 39%
EOF extrema	1.37 | + 9%	1.44 | +22%	0.956 | +0.3%
Monte Carlo (Median)	1.26 | -	1.18 | -	0.952 | -

reconstruction error on the base case scenario by 13% for Southern Brittany, and more than 20% for Normandy and the Mediterranean Sea. The QR pivoting method proves irrelevant for this application with a 50% increase in reconstruction error in Normandy, and around 30% in Southern Brittany and the Mediterranean Sea. The extrema method is closer to the Monte Carlo median case, though above, probably thanks to the manual removal of irrelevant extrema.

To visualize the effect of the sensors' locations, and the origin of the reconstruction error, the reconstruction error is computed as the root mean square error for each grid point, and displayed for the main scenario on the three areas in Fig. 8.

The coastal sensors arrays from the QR pivoting method displayed in the second row do not allow for offshore wind reconstruction, as those points are strongly influenced by coastal effects. Strong reconstruction errors of more than 2 m/s are

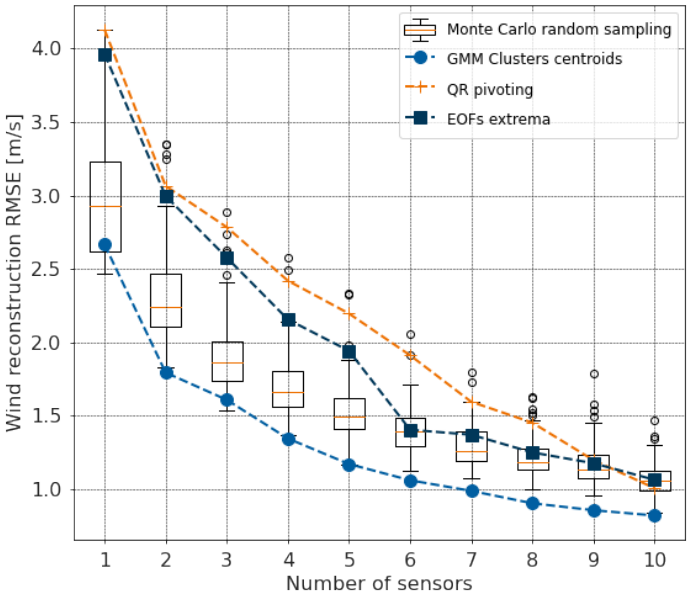

Figure 7. QR pivoting method (orange plus), EOF extrema method (deep blue squares) and GMM method (blue dots) are compared to Monte Carlo simulations, displayed as boxplots in terms of wind reconstruction error.

then obtained far offshore for Normandy and the Mediterranean Sea. For the EOF extrema method displayed in the third row, where some offshore sensors' locations are present in addition to coastal ones, the synoptic wind regime seems better captured with more homogeneous reconstruction error. The reconstruction error patterns show the strong spatial correlation of the input wind data with lower reconstruction errors around the sensors' locations. However, the radius of lowered reconstruction error depends on the location. It can be noted that coastal points in QR pivoting for the Mediterranean Sea have a small radius of influence, as opposed to some offshore points in the EOF extrema method. Coastal areas, where the wind is influenced by the coastal orography and thermodynamic effects, have lower spatial correlations or higher variability. As such, they are considered by the QR pivoting method and the extrema method as salient points. But in the end, they barely help to reconstruct the whole area's dynamics. It illustrates the importance of smart sparse sampling for the reconstruction, and the non-adequacy of QR pivots and EOF extrema as locations for the sparse sampling in this case.

For the three areas, the GMM obtained sensors' locations are homogeneously spatially distributed, allowing for good reconstruction on the whole map, though somewhat neglecting coastal locations. The locations, as centroids of clusters, are the most representative points of maximum likelihood clusters for a given number of sensors. As such, every point of the map belonging to a certain cluster, it allows for satisfactory reconstruction on most of the map. On the other hand, the QR pivoting method and EOF extrema method select salient points that do not necessarily correlate well with neighbouring points, hence lowering the performance for reconstruction.

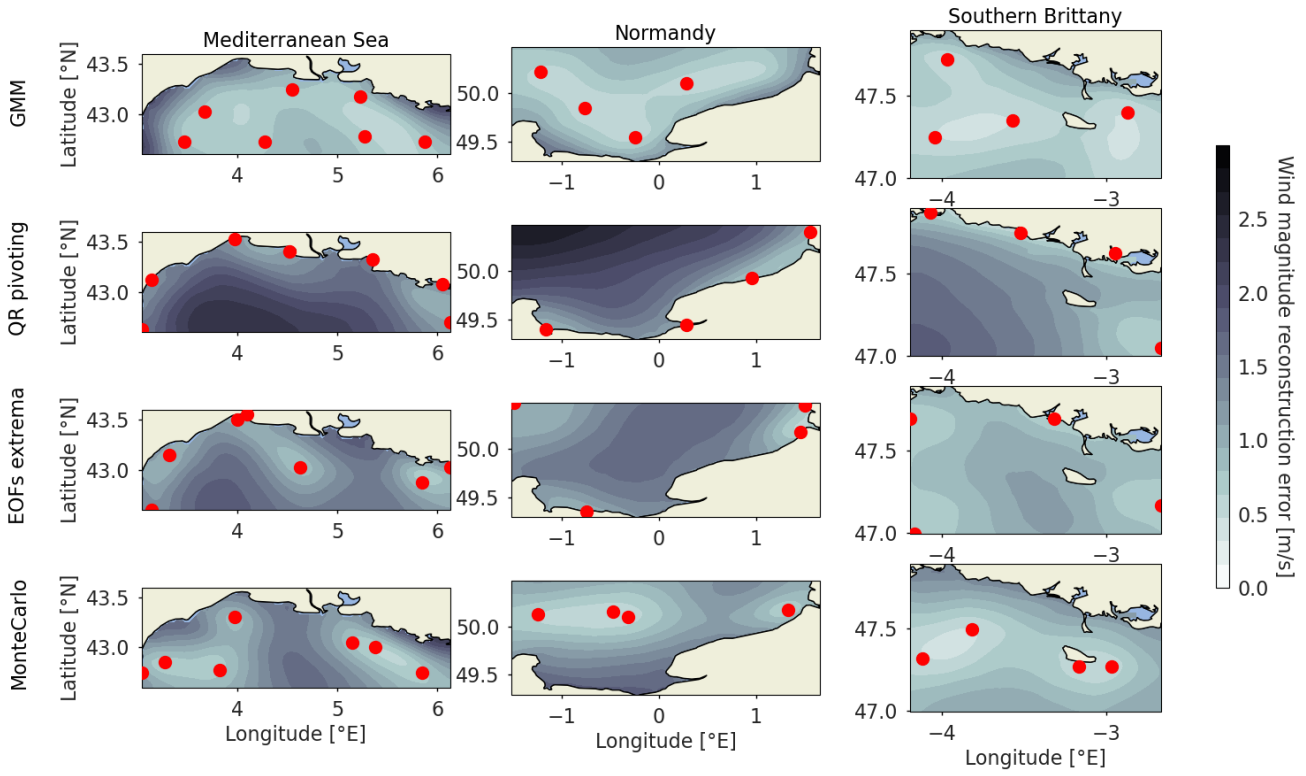

Figure 8. Wind magnitude reconstruction error temporally averaged per grid point on the Mediterranean Sea (left), Normandy (Center), Southern Brittany (right). The reconstruction error is computed for the optimal number of sensors determined in Sect. 5.1 using the 3 baselines and the proposed clustering method. The red dots are the grid points used as input for the least squares reconstruction.

The final suggested sensors' locations for the three offshore wind development areas are given in Table 4. These locations should be considered preferred locations for the deployment of floating LIDARs for wind resource assessment in the French offshore wind development areas.

6 Discussion

In this study, an optimal sparse sampling is proposed using a Gaussian Mixture Model on high-resolution NWP data from Meteo France's MeteoNet data set (AROME model). The method used is simple yet efficient for the optimal sparse sampling of offshore wind field. Applied to offshore wind resource assessment, it can be a useful tool for the design of observation networks. It is compared to state-of-the-art solutions that fail to efficiently sample this specific problem, and a method is

Table 4. Final locations selected for the deployment of floating LIDARs in French offshore wind development areas

sensor #	Mediterranean Sea (Latitude [°N], Longitude [°E])	Normandy (Latitude [°N], Longitude [°E])	Southern Brittany (Latitude [°N], Longitude [°E])
1	(42.775, 5.275)	(49.546, -0.242)	(47.721, -3.967)
2	(43.25, 4.55)	(50.096, 0.283)	(47.396, -2.867)
3	(42.725, 4.275)	(50.221, -1.217)	(47.346, -3.567)
4	(42.725, 3.475)	(49.846, -0.767)	(47.246, -4.042)
5	(43.175, 5.225)		
6	(43.025, 3.675)		
7	(42.725, 5.875)		

proposed to estimate the optimal number of sensors to deploy. The authors nonetheless raise attention on the following points to interpret and discuss the obtained results.

The metric that is used to measure the performance of the sparse sampling in this paper advantages the GMM method, because its homogeneous sampling allows for a correct reconstruction of the synoptic situation. Coastal points are not well reconstructed using the GMM method, but this does not reflect on the scoring. Since the metric averages the reconstructed wind field's error over the grid points, a method that performs fairly good over the entire area is preferred.

As a consequence, both the obtained sensors' location and its performance are depending on the selected area. In this study, the selected sites are simple rectangles over the future development areas. But it could be interesting to reconstruct the wind field and score the performance on specific sites defined by operational limits for example bathymetry on the Mediterranean Sea, or coastal exclusion area for the three cases. This could lead to different results, and the sensitivity of the proposed method would then be studied.

Given the high variability of the wind near the coast and the possible impact on the obtained results, the 20 first kilometers from the coast were excluded to test the sensitivity of the methods. It roughly corresponds to the distance to the coast for future offshore wind parks, and ensures that the impact of the coastal orography is limited. It turns out that it does not make the state-of-the-art methods more relevant for this application as they still tend to select bordering points as input points.

Now, in the context of the development of marine energies in French waters, not only the wind should be considered but other variables such as wave variables, physicochemical parameters (turbidity, sea surface temperature, salinity) which are important for environmental impact monitoring. The installation of a network of sensors would therefore gain traction if optimized with regards to multiple variables. A follow-up of this work would be to include model data for each of the variables of interest, and perform the clustering on the stacked 10 first EOF of each variable, for the design of a multi-parameters observation network.

Getting even closer to the industrial reality of the sensors' network, it would be of great interest to include a cost function dependant on the location (depth, distance from shore, other constraints). This method could be declined to find the Pareto optimal sensors network. The optimal number of sensors would therefore become the number at which the sensors marginal cost exceeds the reconstruction gain.

The data in this study is derived from the NWP model AROME data, with a 1.3km grid size and hourly definition. The parametrisation of the model offshore is not perfect, in particular for the sea/atmosphere coupling that can lead to discrepancies in surface parameters, as shown during the Mediterranean HyMex campaign (Rainaud et al., 2016). The learnt dynamics might then be a coarse description of the reality, and the derived sensors' locations be limited for real time wind reconstruction, since only trained on low spatio-temporal resolution patterns. The obtained locations are in this case optimal only for reconstructing the dynamics of the NWP model, and the representativeness of the used data compared to the reality needs to be questioned. It could then be of interest to run such study on higher resolution data, either from Synthetic Aperture Radar measurements or Large Eddy Simulations, though on shorter periods, or comparing the reconstruction on a set of measurements offshore.

Publicly available through the MeteoNet data set, only 10 meters wind speed are used in this paper. For offshore wind application, hub height wind speed are to be considered (heavy maintenance, loading, energy production). The described method is agnostic to the input data, though it would be of interest to validate the obtained sensors network with 100m wind speed data. It is not direct that the obtained sensor network will be the same with above 100m wind speed data, since the extrapolation will be depending on the grid point and the wind speed and direction (changing the sea surface roughness). The non-linear transformation of data can then change the weight given by the clustering model at each timestep and grid point.

Seemingly, performing the clustering with the power-curve transformed data can potentially lead to different results. The study focuses on wind speed, as this can apply to wind energy production but also maintenance operations planning or wind turbine loading. A specific study could focus on wind power, applying vertical extrapolation and wind power curve. The proposed method can be easily implemented to different data inputs.

The used data set compiles 3 years of data. The model is trained on 2 years and tested and scored on the third year. It could be of interest to carry the same study on a longer data set from global reanalysis models such as ERA5. The high-resolution regional AROME model from Météo France with its 3 years of open source data from the MeteoNet data set offers a higher number of grid point, making it more relevant for the sensor siting on small areas as the ones in this study. The features that need to be captured by the reconstruction are smaller scale than global models' grid size. Comparing the results obtained from the two sources could be of interest.

The used benchmark methods from the sparse-sampling literature i.e., the QR pivoting method and the EOF extrema method do not prove efficient for the stated problem. For generalization purposes, this method would need to be compared to state-of-the-art method on benchmark data set such as the simulated flow past a cylinder used in Manohar et al. (2018). This paper does not aim at generalizing a method, but develops an efficient solution to an identified problem, for which state-of-the-art methods seem to fail.

Eventually, the use of Gaussian Mixture Model seems appropriate for the sparse sampling of offshore wind resource. It is an easy method to implement with relatively low computational cost. It is flexible and can in principle be applied to higher dimensional systems. This could be of interest for offshore wind energy, allowing the inclusion of environmental parameters in the siting optimization. The method also shows good consistency on the three development areas tested with very different wind regimes. It is however important to stress the difficulty associated with the optimal number of sensors. As proposed in

this paper, the number of sensors is derived indirectly from an error threshold. In this context it seems difficult to include cost or environment constraints as such in the sensors siting.

7 Conclusions

A method for the finding of an optimal sensors network for offshore wind reconstruction is presented in this paper, and applied to three of the main offshore wind energy development area in France. The sparse sensors' placement problem is stated on a reduced basis of the 3 years AROME prediction of wind from the MeteoNet data set. State-of-the-art techniques of sparse sensors' placement for reconstruction (QR pivoting and extrema methods) are compared to the proposed method, based on the Gaussian Mixture Model clustering of Empirical Orthogonal Functions of zonal and meridional wind offshore. By selecting the clusters' centroids as proposed locations for sensors, the GMM method homogeneously partitions the domain into spatially correlated clusters. In this way, the reconstruction error on the whole domain is minimized, leading to a 20% decrease in wind reconstruction error compared to the median Monte Carlo case. On the other hand, state-of-the-art methods fail to reconstruct the whole wind field because they are attracted by salient points with high variability (bordering points). However, these points are not very spatially correlated to neighboring points, yielding to a reconstruction error higher than the median Monte Carlo case. The GMM clustering method gives indications on the optimal number of sensors to deploy, though this estimation should be refined either by the integration of cost or environmental constraints, or by the definition of a reconstruction error threshold.

GMM clustering method seems to be a simple yet efficient solution for sparse sensors' placement. Applied to offshore wind reconstruction, it allows for the optimal placement of sensors, and paves the way for smart marine monitoring in the era of offshore wind energy development. Further work should focus on the technique's generalization to benchmark problems, and question the representativeness of the used data set. For wind energy applications, the multivariate case should be studied for multi-instrumental sensors' placement, and the economic constraints should be implemented for the definition of the Pareto optimal number of sensors.

In the light of this study, the authors suggest the deployment of 7 sensors in the Mediterranean Sea, 4 sensors in Normandy and 4 sensors in Southern Brittany at optimal locations to reconstruct the offshore wind field and to help with the wind resource assessment on these areas.

Code and data availability. Meteorological data used in this study are available online through the MeteoNet data set. The code developed for offshore wind resource sparse sampling using Gaussian Mixture Models can be accessed through https://github.com/rmarcille/gmm_sparse_sampling.git

Author contributions. J-F. Filipot and M. Thiébaut identified the problematic. M. Thiébaut and P. Tandeo designed the experiment. R. Marcille carried out the experiment and performed the simulations under the supervision of M. Thiébaut and P. Tandeo. M. Thiébaut and R. Marcille developped the model code. R. Marcille prepared the manuscript with contribution from all co-authors

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work was supported by France Energies Marines and the French government, managed by the Agence Nationale de la Recherche under the Investissements d'Avenir program, with the reference ANR-10-IEED-0006-34. This work was carried out in the framework of the FOWRCE_SEA and POWSEIDOM projects.

References

Ali, N., Calaf, M., and Cal, R. B.: Clustering sparse sensor placement identification and deep learning based forecasting for wind turbine wakes, Journal of Renewable and Sustainable Energy, 13, 023 307, 2021.

Annoni, J., Taylor, T., Bay, C., Johnson, K., Pao, L., Fleming, P., and Dykes, K.: Sparse-sensor placement for wind farm control, in: Journal of Physics: Conference Series, vol. 1037, p. 032019, IOP Publishing, 2018.

Brunton, B. W., Brunton, S. L., Proctor, J. L., and Kutz, J. N.: Sparse Sensor Placement Optimization for Classification, SIAM Journal on Applied Mathematics, 76, 2099–2122, https://doi.org/10.1137/15M1036713, 2016.

Castillo, A. and Messina, A. R.: Data-driven sensor placement for state reconstruction via POD analysis, IET Generation, Transmission & Distribution, 14, 656–664, 2019.

CEREMA: Eoliennes en mer en France, https://www.eoliennesenmer.fr/, 2022.

Chepuri, S. P. and Leus, G.: Continuous sensor placement, IEEE Signal Processing Letters, 22, 544–548, 2014.

Clark, E., Kutz, J. N., and Brunton, S. L.: Sensor Selection With Cost Constraints for Dynamically Relevant Bases, IEEE Sensors Journal, 20, 11 674–11 687, https://doi.org/10.1109/JSEN.2020.2997298, 2020.

Cohen, K., Siegel, S., and McLaughlin, T.: Sensor Placement Based on Proper Orthogonal Decomposition Modeling of a Cylinder Wake, in: 33rd AIAA Fluid Dynamics Conference and Exhibit, vol. 4259, AIAA, https://doi.org/10.2514/6.2003-4259, 2003.

Courtier, P., Thépaut, J.-N., and Hollingsworth, A.: A strategy for operational implementation of 4D-Var, using an incremental approach, Quarterly Journal of the Royal Meteorological Society, 120, 1367–1387, 1994.

Davis, R. E.: Predictability of sea surface temperature and sea level pressure anomalies over the North Pacific Ocean, Journal of Physical Oceanography, 6, 249–266, 1976.

Fukami, K., Maulik, R., Ramachandra, N., Fukagata, K., and Taira, K.: Global field reconstruction from sparse sensors with Voronoi tessellation-assisted deep learning, Nature Machine Intelligence, 3, 945–951, 2021.

Gottschall, J., Gribben, B., Stein, D., and Würth, I.: Floating lidar as an advanced offshore wind speed measurement technique: current technology status and gap analysis in regard to full maturity, WIREs Energy and Environment, 6, e250, https://doi.org/10.1002/wene.250, 2017.

Hawkins, E. and Sutton, R.: Variability of the Atlantic thermohaline circulation described by three-dimensional empirical orthogonal functions, Climate Dynamics, 29, 745–762, 2007.

IEA: Offshore Wind Outlook 2019, Tech. rep., IEA, https://www.iea.org/reports/offshore-wind-outlook-2019, 2019.

Larvor, G., Berthomier, L., Chabot, V., Le Pape, B., Pradel, B., and Perez, L.: MeteoNet, An Open Reference Weather Dataset by Meteo-France. 2020, 2020.

Leon, S. J., Björck, , and Gander, W.: Gram-Schmidt orthogonalization: 100 years and more, Numerical Linear Algebra with Applications, 20, 492–532, https://doi.org/https://doi.org/10.1002/nla.1839, 2013.

Manohar, K., Brunton, B. W., Kutz, J. N., and Brunton, S. L.: Data-Driven Sparse Sensor Placement for Reconstruction: Demonstrating the Benefits of Exploiting Known Patterns, IEEE Control Systems Magazine, 38, 63–86, https://doi.org/10.1109/MCS.2018.2810460, 2018.

Mohren, T. L., Daniel, T. L., Brunton, S. L., and Brunton, B. W.: Neural-inspired sensors enable sparse, efficient classification of spatiotemporal data, Proceedings of the National Academy of Sciences, 115, 10 564–10 569, 2018.

Moore, G. W. K., Renfrew, I. A., and Pickart, R. S.: Multidecadal mobility of the North Atlantic oscillation, Journal of Climate, 26, 2453–2466, 2013.

Murthy, K. S. R. and Rahi, O. P.: A comprehensive review of wind resource assessment, Renewable and Sustainable Energy Reviews, 72, 1320–1342, 2017.

Rainaud, R., Lebeaupin Brossier, C., Ducrocq, V., Giordani, H., Nuret, M., Fourrié, N., Bouin, M.-N., Taupier-Letage, I., and Legain, D.: Characterization of air–sea exchanges over the Western Mediterranean Sea during HyMeX SOP1 using the AROME–WMED model, Quarterly Journal of the Royal Meteorological Society, 142, 173–187, 2016.

Reynolds, D. A.: Gaussian mixture models., Encyclopedia of biometrics, 741, 659–663, 2009.

Schwarz, G.: Estimating the dimension of a model, The annals of statistics, pp. 461–464, 1978.

Seity, Y., Brousseau, P., Malardel, S., Hello, G., Bénard, P., Bouttier, F., Lac, C., and Masson, V.: The AROME-France convective-scale operational model, Monthly Weather Review, 139, 976–991, 2011.

Shukla, P., Skea, J., Slade, R., Al Khourdajie, A., van Diemen, R., McCollum, D., Pathak, M., Some, S., Vyas, P., Fradera, R., Belkacemi, M., Hasijaand, A. Lisboa, G., Luz, S., Malley, J., and (eds.): Climate Change 2022: Mitigation of Climate Change. Working Group III Contribution to the IPCC Sixth Assessment Report, Tech. rep., IPCC, Cambridge, UK and New York, NY, USA, https://doi.org/10.1017/9781009157926, 2022.

Termonia, P., Fischer, C., Bazile, E., Bouyssel, F., Brozkova, R., Bénard, P., Bochenek, B., Degrauwe, D., Derková, M., Khatib, R., Hamdi, R., Mašek, J., Pottier, P., Pristov, N., Seity, Y., Smolikova, P., Španiel, O., Tudor, M., Wang, Y., and Joly, A.: The ALADIN System and its canonical model configurations AROME CY41T1 and ALARO CY40T1, Geoscientific Model Development, 11, 257–281, https://doi.org/10.5194/gmd-11-257-2018, 2018.

Thompson, D. W. and Wallace, J. M.: Annular modes in the extratropical circulation. Part I: Month-to-month variability, Journal of climate, 13, 1000–1016, 2000.

Yang, X., Venturi, D., Chen, C., Chryssostomidis, C., and Karniadakis, G. E.: EOF-based constrained sensor placement and field reconstruction from noisy ocean measurements: Application to Nantucket Sound, Journal of Geophysical Research: Oceans, 115, 2010.

Yildirim, B., Chryssostomidis, C., and Karniadakis, G. E.: Efficient sensor placement for ocean measurements using low-dimensional concepts, Ocean Modelling, 27, 160–173, 2009.

Zhang, Y. and Bellingham, J. G.: An efficient method of selecting ocean observing locations for capturing the leading modes and reconstructing the full field, Journal of Geophysical Research: Oceans, 113, 2008.