

Response to Reviewer 1 – Sarah Barber

Gaussian mixture models for the optimal sparse sampling of offshore wind resource –
by Marcille et al.

Note: Your comments and questions are reported in this document, and we use bold text for our responses. The line numbers in the responses correspond to the corrected pdf file with highlighted differences.

General comments

- Scientific relevance: very relevant for offshore development, especially in countries with ambitious plans but not much of experience. It would be interesting to apply this to Brazil, for example.
- Scientific quality: generally high quality, explaining and highlighting differences or anomalies well and explaining alternative methods.
- Presentation quality: generally well-written and well-prepared. Please check your usage of apostrophes (I marked them at the beginning but then gave up).

Thank you for your positive comments on the scientific relevance. We indeed think that this can be a useful tool for regional planification for offshore wind and made the choice to present it application-driven. Thank you for the grammatical corrections, and apologies on the English. We did our best to correct typos and English idiomatic expressions.

Specific comments and technical corrections

- Please see the annotations in the attached file

Grammamy related comments throughout the manuscript:

All grammar related comments were implemented. A special care was given to apostrophes.

L.2: Insert one sentence here about what state of the art is and why your solution is needed

We added a sentence in this abstract to account for state of the art.

L. 2 – 4 “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”.

L.6: I'm not sure what you mean here. Do you mean that the result of applying the method is an optimal location for sensors that minimises the wind field reconstruction error? (in which case, you need to write "it yields an optimal....")

Yes it means that the result of applying the methodology is a collection of locations that together minimize the wind field reconstruction error. The sentence was reformulated.

L. 9: For someone reading this for the first time who is not experienced with sparse sampling methods, it's not really clear what this means. Can you describe it in a way that is more understandable to non-experts please?

The sentence was reformulated:

L. 9 – 14 “The described method applied to the study areas outputs sensors array 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.”

L12 – L 19: This part isn't necessary in a wind energy journal. I would replace this whole section with "Offshore wind is key to reaching ambitious net zero targets such as xxx (e.g. the European Energy Strategy 2050)"

The introductory paragraph was reduced to

L.16 – 23 “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)”. The focus is made on French waters afterwards for energetic roadmap definition and potential installed capacity purposes.

L33: for a given planned project? (or is this process done independent of individual project planning?)

It depends on the study case. In this paper we are focusing on a regional scale, with potentially several future projects (For planification).

L. 38 – 39 “Their number and siting thus need to be optimized in order to compose an optimal sensors network in an offshore wind development area.”

L.35 – Combined*

The point here is that the selection of k positions within N grid points is an intractable combinatorial* problem in the algebraic sense of the world, that can not be solved with convex optimization.

L.36: QR: explain

QR is the decomposition of a matrix into an orthogonal matrix Q and an upper triangular matrix R. The explanation is given in the next sentence

L.45: EOF's extrema – sensors' placement

All “EOFs extrema” were replaced by EOF extrema, removing the apostrophe (Empirical Orthogonal Functions' extrema → EOF extrema). It was detailed for first use

Seemingly “EOFs” was replaced by “EOF” to avoid confusion

“sensors placement” was replaced by sensors' placement throughout the document

L55: This sounds like you are referring to the improvement of the previous paper (Mohren et al.) but you can't be because this paper is older. Please explain this better.

Thanks for the comment. The two references (Chepuri and Leus 2014 and Mohren et al (2018) were switched to avoid confusion. The paragraph gathers 3 papers that bring innovations to “augment” the problem: Chepuri allow for continuous sensor placement (as opposed to grid point selection),

Mohren brings the temporal component in the problem, and Fukami (the most recent) uses a CNN super-resolution model from measurement sensors, selected with tessellation. L. 78 - 88

L.58 :Showing ?

In the reference from Fukami (2021), the sparse sampling problem is applied to a global sea surface temperature reconstruction problem. Being a very highly dimensional problem, it needs to go through tessellation to reduce the dimensionality. The obtained image is put on a grid for super-resolution using convolutional neural networks.

A sentence is added to specify : L. 87 – 88 “This reconstruction technique is then tested on sea surface temperature reconstruction globally, showing the possibility to use sparse sampling on very high dimension problems.”

L61 – 63: Please describe these more, as they are most relevant to this paper (including a summary of the conclusions, i.e. how well the methods worked)

The descriptions of those 2 relevant papers for wind energy application have been detailed and enriched with the reference of the algorithm used in Ali et. Al 2021 (Brunton et al. 2016 – SSPOC).

L. 89 – 100

“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”

L64: Is this because standards only require one measurement? Please explain in more detail.

Such methods are applied at the regional scale, for the deployment of several wind farms. It assumes the governmental deployment of several wind sensors for the wind resource assessment. At the wind farm level, this method is a bit overkill (at least using Numerical Weather Prediction data with low spatial resolution). In France it would be a great driver for the design of tendering areas, or for the financing of an observatory that would be beneficial to all players. Few precisions were added though the political reasons were quickly mentioned. But you are right, the way it is usually done, it’s 1 measurement point per development area, an area which is selected on several factors (including WRA). This study shows that a limited number of deployed sensors could help reconstruct the whole wind field at the regional scale.

L. 101 – 104 “However, to the best of our knowledge, such methods were never applied at the regional scale for wind energy resource assessment, to determine optimal sensors placement. 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, assuming spatial representativity."

L67-68: What does "sufficient" mean in this context? Please discuss and explain this. Does "optimal" not refer to the best compromise between costs and accuracy, in which case how can this be quantified?

Indeed, the number of sensors is "optimal" when representing a trade-off between wind field reconstruction and overall cost. This quantification is not the core of this paper and is mentioned in the discussion. Here the optimal number of sensors is obtained with user-defined error threshold as described latter-on. The "(i.e. sufficient)" was removed.

A precision is added: L. 116 – 123

"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."

The notion of "optimal" number of sensors is then discussed in section 4.1.

L69 – 77: How was this chosen and why? - Summarising the applied method at this point in the paper raises questions about why you chose it. You need to connect the literature review to the chosen method by summarising where the gaps in the current methods and how your method fills them. This requires a slight restructuring of this section.

The summary was changed to:

L. 101 - 131

"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 are not necessarily based on wind resource assessment only, 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"

L78: This section title is too general. I suggest "Database used for this study" of "Data sets applied" or similar.

Title changed to "Study data set" L. 132

L108: I don't understand this fully. Why does that justify 10 m measurement data, especially as we know that the wind turbines will be installed much higher? Please explain more clearly!

You are right, it does not justify it, but it does impose it. This point is mentioned in the discussion, and a precision is added with that regard:

L. 163 – 166

"The open-source MeteoNet data set only contains surface parameters of temperature, humidity, pressure and precipitation, and 10-meters wind speed (u_{10} , v_{10}), 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."

L114: Why? And how did you establish this?

L. 172 – 173 Simply missing data, it was mentioned "A total of 65 days (~ 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." It is a feature of the open-source data set basically

L116: Again, a slightly more specific section title would be good

This part was split in two: a "Background" section with problem statement and "Reduced Order Model" formalism, and a "Sparse sampling methods used in this study" describing the compared methods.

L117: You need to explain here that you first decide on three baseline methods for comparison purposes and then you apply a new method, Gaussian Mixture Model clustering. In the introduction we need to learn why this method is promising compared to existing methods (see my comments at the end of section 1).

The explanation on baseline methods is then given at the beginning of the "Sparse sampling methods used in this study" section. A more thorough explanation of the chosen technique is given in introduction.

Are you tackling this general problem, in which case you should write "...to reconstruct wind fields" or are you tackling a specific problem of just these wind fields, in which case you should write ".....to reconstruct the wind fields in the areas of xxxx in France" (or similar)?

Thanks for the comment. It refers to the general problem, that is why three different areas are tested.

L. 175 - 176

"The problem tackled in this paper is the finding of an optimal network of sensors to reconstruct offshore wind fields"

L119: Please explain this briefly

A precision is given:

L. 177 – 178 "The finding of D optimal input points from K grid points consists of a combinatorial optimization problem, the exhaustive search of which is computationally intractable."

L. 147: Not sure exactly what this means. I guess this is the "ground truth"? Please explain better.

Reformulated : L. 207 "assuming that the actual coefficients of the reduced basis are perfectly known. This is considered as the ground truth."

L166-168: This belongs in the introduction - see my comments at the end of section 1.

Thanks for the comment on the introduction's structure. It was reformulated in the introduction in a bigger paragraph about EOF

L. 41 – 57

"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))."

L184 – 187: Belongs in the introduction

This sentence was indeed redundant with the introduction and removed.

L.187: Please name these three methods here and say you are going to describe them in more detail in the next sections. Also, please explain why you chose these three. i.e. as a comparison for the new Gaussian Mixture Model clustering developed in this work

The introduction of the section was reformulated as follows:

L. 245 – L. 254

“4. Sparse sampling methods used in this study

In this section, the methods applied for the sensors' locations selection are described in detail. The novel data-driven method based on Gaussian Mixture Model is presented alongside 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).

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 Gaussian Mixture Model in this specific application.”

L.213: Reference?

Added reference to a paper retracing the Gram-Schmidt process:

Gram–Schmidt orthogonalization: 100 years and more – Leon et al. 2013

L. 280

L.228: Please explain exactly what is meant by a "cluster"

Precisions are given about the definition of cluster (group of point):

L. 295 – 297 “The model is a mixture of multivariate normal distributions, each distribution representing a cluster 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.”

L.263: Please refer to and explain this figure in the text.

The figure was described more thoroughly in the text

L.307 – 311

“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 (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”

L.266: Please give a brief introduction on what you will be talking about in this section

L. 338 – 340 “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.”

L268: What does this stand for?

BIC score is mentioned above, line 313

“The optimum number of clusters can be defined through the calculation of the Bayesian information criterion (BIC) score (Schwarz, 1978)”

L.270: Why?

It is a heuristic criterion with no theoretical guarantee of optimality. It hints a trade-off between accuracy and complexity, to avoid the over-fitting of the dataset by a high-dimensional model (1 cluster per grid point for example would yield to perfect reconstruction)

L.342 – 347

“The number of sensors to place on the grid is an input of the GMM. The BIC score described in section 4.2 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 criterion to hint the trade-off between accuracy and complexity, to avoid over-fitting.”

L.274: Is there a reference for this? And can it be quantified?

It can hardly be quantified. It is illustrated in figure 4, where the “elbow” of the curve is existing but not giving definitive results. We basically want to identify the moment when the gradient becomes almost constant (the elbow of the curve) i.e. the moment when adding the next cluster does not significantly change the BIC score.

The associated reference is from Thorndike in 1953 but we do not think it is clearly relevant for the paper <https://doi.org/10.1007/BF02289263>

L.278: How are they different? Random?

Yes random initialisations. “different” was changed to “random”

L. 288: It's not entirely clear to me if the layout of the sensors as well as the optimal number is included in this optimisation. I think it is choosing the optimal number out of the existing sensors, and not the optimal number in general, right? In which case, how do you know which ones have been chosen? Please clarify this. OK this comes in Section 4.2. Please refer to that here.

(if not, why is it helpful to know how many sensors are needed if the location isn't known or specified?)

The number of sensors is an input to the model. So we try to find out which number of sensors is best in terms of clustering. However, criterions such as BIC criterion are not sufficient to characterize completely an optimum, so the number of clusters is fixed as the minimum number of sensors to reach a certain error threshold (user-defined)

The clustering method is applied to a range of number of sensors, and the results are used to finally chose the number of sensors. It has the advantage that the obtained sensors' location yield to similar error levels between the three areas, which was not well captured by the clustering criterion.

We are not sure we understand your question and the reference to part 4.2 which is really about the comparison between the methods. The justification in our sense is given in part 4.1 :

L. 364 - 378

“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 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 RMSE 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.”

L. 292: Define RMSE

Root Mean Squared Error

Response to Reviewer 2

Gaussian Mixture Models for the Optimal Sparse Sampling of Offshore Wind Resource by Marcille et al

Note: Your comments and questions are reported in this document, and we use bold text for our responses. The line numbers in the responses correspond to the corrected pdf file with highlighted differences.

(1) According to the current situation of vigorously developing renewable energy, this article proposes a novel, effective and practical method for wind farm reconstruction and optimization of the optimal sensor network, both for the evaluation of offshore wind resources and the development of offshore wind energy. It has very important theoretical value and practical significance.

(2) The three locations selected by the author in the article are very representative. At the same time, as the focus of wind power development in France at present and in the future, the research results made by the author will have good reference significance.

(3) The author proposes four methods for selecting the position and number of sensors in the article, and at the same time gives a detailed introduction to these four methods, with concise language and clear organization.

Thank you for your positive comments.

revise opinion

(1) An explanation should be given at the end of the Introduction as to why the three regions of Normandy, South Brittany and the Bay of Lions in the Mediterranean were chosen.

The three areas are major areas for the future development of offshore wind in France and could require the development of a wind observational network for planning and execution of the projects. Precisions were added in the timeline at the beginning of the “Study Data set” section:

L. 136 – 137 “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).”

(2) The GMM method is good at reconstructing the weather situation while discarding points of high variability that may be associated with extreme events. How did you come to this conclusion?

This conclusion comes from the scores displayed in table 2. It shows that the GMM method is not systematically better than EOF extrema and QR pivoting in reconstructing the maximum wind speed of the map (Max wind speed RMSE in Table 2.) while it is clearly better for the mean wind speed and RMSE.

The interpretation is that the selected points in the GMM method are very different from the “salient” points of both the QR pivots and EOF extrema. Those point carry the most variability in the dataset but are not spatially representative of what’s happening at the regional scale. Coastal

points can be associated with extreme events due to the influence of the coast (Extreme in the sense of extreme variability compared to other grid points). While the centroids selected by the GMM methods are the most spatially representative points, discarding the extreme points that are not relevant for the regional reconstruction.

A precision was added:

L. 408 – 410 “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.”

(3) The author deleted a part of the data in the dataset, why did they delete them, and what are the criteria for deletion?

Basically 65 days in the open source dataset are corrupted files. The rest of the data was complete and coherent.

L. 172 – 173 “A total of 65 days (~ 6%) of the 3-year data set are unusable due to largely missing data. The days identified as erroneous are similar for each area and were removed from the analysis.”

(4) "Although the clustering itself may find the best of 5 clusters for the Mediterranean, this may result in a higher reconstruction error than the other regions." Why does this result?

This result is illustrated by figure 4 and 5, showing that the BIC score indicated number of sensors yield to a much higher error in the Med Sea than for other areas. It comes from the fact that the BIC score is not directly linked to the reconstruction error, but only from the likelihood of the obtained clustering. Precisions were added :

L. 359 – 363 “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”

(5) When testing the sensitivity of the method, an area 20 km from the coast was excluded, why choose 20 km instead of other ranges.

This range will be approximately the distance to the coast for future offshore projects in France. Furthermore, it is wide enough to exclude any impact from the coastal orography. In the paper from Barthelemie et. al in 2007 (Offshore Coastal Wind Speed Gradients: Issues for the Design and Development of Large Offshore Windfarms), the coastal zone is between 20 and 70 km in Europe (Results suggest that the distance from the coastline over which wind speed vertical profiles are not at equilibrium with the sea surface (which defines the coastal zone) extends to 20 km and possibly 70 km from the coast). The 20km buffer is considered, because of the proximity of next tender processes in France.

Added precision: L. 472 – 473 “It roughly corresponds to the minimum distance to the coast for future offshore wind parks and ensures that the impact of the orography is limited.”

