Response to comments: S. C. Pryor
Title: Gaussian Mixture Models for the Optimal Sparse Sampling of Offshore Wind Resource
Authors: R. Marcille, M. Thiébaut, P. Tandeo, J-F. Filipot

We thank the associate editor for the constructive comments. Major revisions have been suggested before potential publication. Please see below our responses and implemented changes.

While I accept some model tools are proprietary I would kindly ask the authors (1) consider releasing their code (as strongly encouraged by the journal but not a requirement):

The FOWRCE-SEA consortium accepted to share the code associated to the presented results. All functions used to generate the presented results are shared in the repository, and a jupyter notebook showing example of use on the Mediterranean Sea. Data for the Mediterranean Sea is available in the repository, while the data for the two other study areas can directly be downloaded from the Meteo Net Dataset.

Github repository:

https://github.com/rmarcille/gmm_sparse_sampling.git

Added paragraph:

L.531 – L.533

"*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"

I am not an expert in Gaussian mixture models but was not able to follow the methodology applied so I engaged a review from a notable statistician (who also works in the domain) and they also find the methodology to be poorly described. I would kindly ask the authors to substantially improve the traceability of the description of the method.

The whole methodology part was reformulated to improve readability.

The Background section was reformulated – Preliminaries, and subsections were added to improve traceability. The following sections are now proposed.

3. Preliminaries

3.1 Problem Statement

L.149 - 157

The problem addressed in the paper is stated in this section. The full state (X), sparse measurement matrix (Y), and locations matrix (γ) notations are introduced. The train/test split of the dataset is described.

This section introduced the problem's notation before presenting the formalism to increase the understanding.

3.2 Reduced Order Model

L.158 - 175

The data reduction method is presented there. The application case (10 first EOFs of both zonal and meridional wind speed) is detailed at the end.

3.3 Sparse sampling

L.176 - 214

In this section, the core of the proposed formalism is described, always linking to the proposed application.

3.3.1 State description

L.176 – L.191

How is the full state described, reduced and sampled. In this section, the EOF decomposition is applied to the full state of the offshore wind field.

Then a locations matrix (γ) associated with a sampling matrix **C** is introduced to formalize the sampling from the ensemble of NWP grid points. This locations matrix is the output of the methods described in section 4.

Eventually the sparse measurement matrix (**Y**) is derived from the full state using the sampling matrix

3.3.2 Full state reconstruction

L.192 – L.205

The methodology used to reconstruct the full state of the system from the sparse measurement matrix is described in this section. A linear model is fitted between the sparse measurement matrix and the matrix of EOF coefficients (**a**) on the training dataset using ordinary least squares. The obtained coefficients matrix $\hat{\beta}$ is used to reconstruct the full state from sparse measurements.

3.3.3 Reconstruction error

L.205 – L.214

From the obtained reconstruction, a reconstruction error is defined, eventually stating the minimization problem at the core of this article. The reconstruction error is to be minimized through all locations matrix (γ) and number of input *D*

In part ***4. Sparse sampling methods used in this study*** – the introduction was modified with the addition of a precision concerning the goal of those methods :

L.219

“All the methods described in this section should output a matrix of sensors’ locations γ given a number of sensors D.”

For the baseline method QR pivoting, the formalism is detailed for the sensors location definition. Equation is given to find the gamma matrix. L. 257

Then the description of the Gaussian Mixture Model (section 4.2) was largely reformulated. A step-by-step description of the methodology is given.

4.2.1 Gaussian mixture

L.268 - 275

Here the basic definition of a Gaussian mixture model is given with mathematical expressions of the probability distribution and the Gaussian distribution.

4.2.2 Expectation – Maximization algorithm

L.277 – L.301

The Expectation-Maximization (EM) algorithm is decomposed into 3 steps. Each step is described in this section. The first step is the random initialization of the means and covariances and iteratively determination of the weights. The second step is the Expectation step (E-step), where the probability that a given point belongs to a cluster is computed. The "responsibilities" of the Gaussian distributions are computed in the E-step. The third step is the Maximization step (M-step) where the algorithm uses the responsibilities of the Gaussian distribution to update the means, covariances and weights. These updated estimates are used in the next E-step to compute new responsibilities for the data points. So on and so forth, this process will repeat until algorithm convergence, typically achieved when the model parameters do not change significantly from one iteration to the next

4.2.3 Bayesian Information Criterion

L.302 – L.317

The Gaussian mixture require as input to set to number of clusters. It is calculated with the Bayesian Information Criterion (BIC) score. Definition of the BIC is given as well as its mathematical expression. It is explained that the BIC score is a trade-off between likelihood of the obtained distribution, and the complexity of the model.

4.2.4 Implementation for the study case

L.319 – L.327

In this new section, the method to apply GMM for the study case is detailed. Both the workflow for the clustering, and the extraction of future sensors locations.

I believe it would also be very useful to elaborate a little further regarding the relative pros and cons of this specific method.

A paragraph is added at the end of the discussion to put the stress on the pros and cons of the method.

L.502 – L.508

"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."