Fast and Accurate Aggregated Modelling for Wind Farms Using Gaussian Mixture Model

Muhammad Helal Uddin, Junaid Khalid
School of Engineering
Cardiff University
Cardiff, UK
{uddinmh, khalidj1}@cardiff.ac.uk

Michael Smailes
Research & Technical Capabilities
Offshore Renewable Energy Catapult
Blyth, UK
michael.smailes@ore.catapult.org.uk

Jun Liang, Sheng Wang
School of Engineering
Cardiff University
Cardiff, UK
{liangj1, wangs9}@cardiff.ac.uk

***Abstract*—A detailed wind farm model that includes every wind turbine (WT) can significantly extend electromagnetic transient (EMT) simulation times and require substantial computational resources. Therefore, developing an appropriate dynamic aggregated equivalent model is crucial to replicate the overall dynamic behavior of wind farms. While WT aggregation in wind farms reduces the modelling and computational load, it may also compromise accuracy. Additionally, it can be challenging to precisely capture the dynamic behavior of WTs during power system disturbances. This paper combines gaussian mixture model clustering with information theoretic averaging method to address the aforementioned issues. The approach is validated by comparing the EMT simulation results of the detailed system with those of the proposed aggregated model. The results demonstrate that the proposed method accurately replicates both the steady-state and dominant transient responses of the detailed system while significantly improving computational efficiency.**

***Index Terms*—EMT simulation, aggregation modelling, gaussian mixture model, accuracy, fast computation.**

I. Introduction

As wind energy utilisation continues to rise, ensuring the dynamic performance of large-scale wind farms becomes critical for system design and stability assessment. A ‘detailed model’ represents a large scale wind farm by simulating each wind turbine’s control systems with power converter switching operations and its related cable interconnections. As a result, employing a detailed model in power system dynamic studies is difficult and time-consuming. At the system-level, integrating each individual wind turbine (WT) into the simulation model is impractical and unneeded. In fact, a large number of WTs should be modelled by aggregation [1]. The complexity in modelling and aggregation of wind farms (WFs) due to the stochastic nature of the wind, variability of wind speed and interactions between wind turbines is further complicated by different WF layouts and potential changes in the operating mode of WTs (i.e., normal, LVRT, disconnected) during faults in transmission lines.

Existing aggregation techniques can be categorised into single machine and multi-machine aggregation methods. In the single machine method, the whole wind farm is represented by one equivalent WT whose output power is scaled up to emulate the total output power of all WTs. It is computationally efficient and widely used in practical engineering. That said, a single aggregated generator and collector system impedance cannot adequately represent the variations in collector system voltage; the dynamic behavior of different types of WTs within a WF, or the differences in turbine output caused by wake effects. A multi-machine model approach where more than one aggregated machine is used to represent a WF is a more appropriate solution [1]. In contrast, in the multi-machine approach, WTs are clustered into groups based on the similarity of specific operational parameters. Compared with the single machine method, it is more accurate but computationally complex. Other aggregation approaches also include capacity weighted methods, probabilistic clustering and admittance model order reduction, which still have several restrictions, including reduced accuracy, and limited generality [2]–[4]. In the traditional WT clustering process, the clustering outcomes are based on the operating characteristics of the WT at a specific point in time. However, since these characteristics change dynamically over time, the clustering results derived from a single time point may become outdated as the operation progresses. Consequently, time-specific clustering results may lose their validity as the WT’s operational behavior evolves. Several researches on multi-machine based aggregation modelling have been conducted. A multi-machine dynamic equivalent method based on the fuzzy clustering algorithm is proposed for active power characteristic analysis of a DFIG based wind farm in [4]. Peng *et al.* [5] developed a geometric template matching based aggregation method but it needs more computational overhead, slower convergence rate. Considering to the limitation of existing researches, a Gaussian mixture model (GMM) based clustering technique with Information Theoretic Averaging (ITA) approach for PMSG based wind farm modelling is developed to achieve better accuracy and fast computational performance.

II. AGGREGATED MODELLING FOR WF

The complexity and accuracy of aggregation modelling (AM) both depend on the appropriate definition of the aggregation algorithm and the selection of indicators and parameters that appear in the final model. WT clustering is a powerful method to enhance the accuracy of WF aggregated models significantly [4]. To achieve effective WT clustering, two critical steps are involved: selecting appropriate clustering indicators and formulating the clustering algorithm. In this research, clustering indicators are derived from time series variables. A GMM clustering approach is proposed to analyse the time series data and derive the wind turbine clustering options.

A. Clustering Indicator

1) Wind Speed: WT performance is heavily influenced by input wind speed, which serves as a critical operational determinant. This parameter inherently captures important factors including the turbine's geographical placement and inter-turbine dynamics such as wake effects. Consequently, time-series data of input wind speed has been identified as one of the primary indicators for the clustering process.

2) Active Power: WT parameters can change over time with increased operation. To accurately capture the dynamic response of WTs, clustering indicators must reflect variations in WT parameters. Active power is generally considered the most effective in representing the output characteristics of WTs [4]. Therefore, the output active power of WTs is selected as another key clustering indicator.

3) Terminal residual voltage: A well-designed equivalent collector network is essential for the AM, as the dynamic response of each WT is influenced by line parameters and terminal residual voltages [6]. These voltages are measured on the low-voltage side of the step-up transformer, as depicted in Fig. 1. The line impedance Z_m is typically calculated under steady-state operation as $Z_m = [Z_{WT1}, Z_{WT2}, ..., Z_{WTm}]$. Furthermore, the capacity of the equivalent transformer scale up to the number of wind turbines. However, during a fault, the uniform terminal residual voltages for each WT affect the value of Z_m due to the parallel connection of the WTs, potentially leading to errors in the equivalent line impedance. Hence, terminal residual voltages chosen as another clustering indicator.

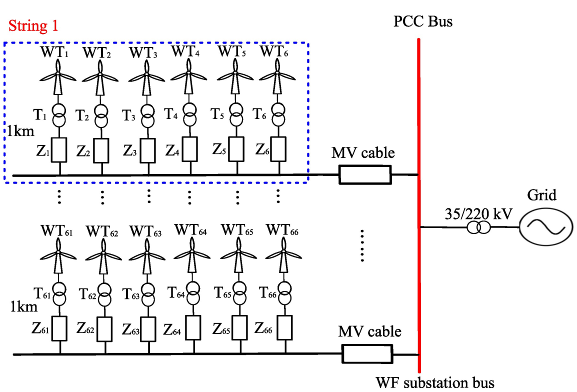

Fig. 1. WF layout considered in this study.

B. Clustering Algorithm

This section introduces a clustering technique based on GMM. This new algorithm uses time series data with clustering indicators, in contrast to conventional clustering methods that employ time spot data.

The core steps of the GMM algorithm is outlined below [7]:

Step 1: Create a dataset with the three clustering indicators: wind speed, active power, and terminal residual voltage. Represent the dataset as $X = (x_1, y_1, z_1), (x_2, y_2, z_2), \ldots, (x_n, y_n, z_n)$,where each triplet (x_i, y_i, z_i) corresponds to a data point with wind speed, active power, and terminal residual voltage.

Step 2: Initialise the parameters for each gaussian component (cluster) k, including:

- Mean vector $\mu_k = [\mu_{x_k}, \mu_{y_k}, \mu_{z_k}]$, corresponding to the three indicators.
- Covariance matrix Σ_k , describing the spread and correlation between wind speed, active power, and terminal residual voltage for each cluster.
- Mixing coefficient π_k , representing the weight (probability) of cluster k.

Step 3: Expectation-Maximization (EM) Algorithm: GMM uses the EM algorithm to iteratively refine the parameters to fit the data. The process involves two main steps:

a) **E-step (Expectation Step)**: For every data point, compute the posterior probability (responsibility) of membership in each cluster k, using the current parameter indicators.

$$\gamma_{ik} = \frac{\pi_k \mathcal{N}(X_i \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(X_i \mid \mu_j, \Sigma_j)} \tag{1}$$

where γi_k is the responsibility cluster k takes for data point i, and $\mathcal{N}(X_i \mid \mu_k, \Sigma_k)$ is the multivariate gaussian probability density function.

b) **M-step (Maximization Step)** : The EM algorithm updates each component's parameters (means μ_k, covariances Σ_k, and mixing weights π_k) using the responsibility values computed during the expectation step. These updates are derived from the responsibilities calculations performed in the E-step.

$$\begin{aligned} \mu_k &= \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{ik} X_i \\ \Sigma_k &= \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{ik} (X_i - \mu_k)(X_i - \mu_k)^T \\ \pi_k &= \frac{N_k}{n} \end{aligned} \tag{2}$$

where,$N_k = \sum_{i=1}^{n} \gamma_{ik}$ is the effective number of points assigned to cluster k.

Step 4: Repeat *Step 3* until the convergence condition is satisfied.

Step 5: After the EM algorithm converges, calculate the log-likelihood (LLL) of the model, which quantifies how well the GMM fits the data:

$$\ln(L) = \sum_{i=1}^{n} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(X_i \mid \mu_k, \Sigma_k) \right) \tag{3}$$

Step 6: The performance of the GMM classifier is significantly influenced by the selection of k. Model complexity must be balanced by choosing a suitable k value; too many components can cause overfitting, while too few can cause underfitting. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are the most widely recognized and frequently utilized approaches. These criteria provide quantitative methods to evaluate model quality and identify the most suitable number of clusters [7].

In order to balance the trade-off between model fit and model simplicity, both AIC and BIC use a penalty term to calculate the ideal value of k. Their definition is as follows:

$$\begin{aligned} \mathrm{AIC} &= 2\mathrm{p} - 2\ln(\mathrm{L}) \\ \mathrm{BIC} &= \mathrm{p}\ln(\mathrm{n}) - 2\ln(\mathrm{L}) \end{aligned} \tag{4}$$

Where L represents the fitted model's log-likelihood, p is the parameter count varies based on mixing weights, mean and covariance, and n equals the number of samples in dataset. When applied to the AM for WTs, selecting a model with lower AIC would typically result in a more accurate AM, but this could come at the cost of more clusters and potentially higher computational demand. On the other hand, a lower BIC suggests a model with fewer clusters (computationally faster) but penalises a more complex model (less accuracy).

Algorithm 1 summarises the algorithm based on AIC or BIC to determine the ideal number of components k_{opt}. Given a maximum value K_{max}, we executed the EM algorithm K_{max} times with $k = 1$ to K_{max}. For each k, we calculated the AIC or BIC, and the optimal value k_{opt} was determined by picking the k that corresponded to the lowest value of AIC or BIC.

Algorithm 1 Selection of the optimal number of gaussian components K_{opt} using AIC or BIC criterion. (Given K_{max} reprsents the maximum number of gaussian components)

Require: A set of data points X and K_{max}
Ensure: k_{opt}
for $k = 1$ to K_{max} **do**
Calculate the log-likelihood with k
Determine the AIC or BIC using the parameters that were acquired by employing GMM with k
end for
$k_{opt} = \arg\min_k(\mathrm{AIC}(k)$ or $\mathrm{BIC}(k))$

AIC and BIC provide the optimum cluster which could be different. To combine the advantages of both AIC and BIC, the system can be optimised using ITA approach [8]. It is based on the principle that instead of relying on a single model, we can average over a set of candidate models to account for uncertainty in model selection, leading to more robust estimates.

In ITA, Each model is assigned a weight, which is determined by how well it balances model fit and simplicity. Calculate AIC or BIC for each model:

$$\begin{aligned} \Delta AIC_K &= AIC - \min(AIC) \\ \Delta BIC_K &= BIC - \min(BIC) \end{aligned} \tag{5}$$

Compute models based on either AIC or BIC:

$$w_K = \frac{\exp(-0.5 \cdot \Delta_K)}{\sum_K \exp(-0.5 \cdot \Delta_K)} \tag{6}$$

where Δ_K can be either ΔAIC_K or ΔBIC_K, depending on which criterion is used. Now, cluster assignments from all models will be combined using the computed weights.

$$\gamma_{ik} = \sum_K (w_K \cdot \gamma_{ik}) \tag{7}$$

Step 7 : After selecting the optimal number of clusters k_{opt}, assign each data point to the cluster for which it has the highest responsibility (posterior probability):

$$\text{Cluster for } X_i = \arg\max_k (\gamma_{ik}) \tag{8}$$

The comprehensive steps of GMM algorithm showed in Fig. 2.

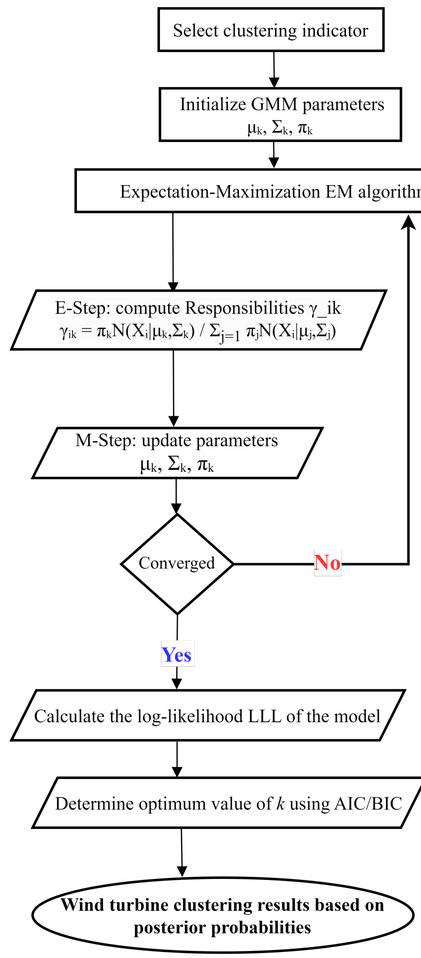

Fig. 2. Flowchart of Gaussian mixture model

III. SIMULATION AND DISCUSSION

A. Simulation Model for Case Study

The model of a wind farm consisting of 66 × 1.5 MW PMSGs ($WT_1 - WT_{66}$) is considered for GMM implementation as shown in Fig. 1. To compare results with the detailed model, one string of six wind generators are considered for EMT modelling due to practical slower simulation time of running whole WF (for instance 0.5 sec simulation time with whole WF will take 1500 s). Each WT generator operates with a terminal voltage of 0.69 kV and connected through a 35/0.69-kV transformer ($T_1 - T_6$) with 1 km apart. All power from each string is collected through the 30 km MV collection cable to the point of common coupling (PCC), and then output is delivered to the grid through 220/35-kV step up transformer.

TABLE I
WT AND COLLECTION CABLE PARAMETERS FOR CASE STUDY [4], [9]

Parameters	Values	Parameters	Values
Rated Power PMSG (MW)	1.5	Rated frequency	50 Hz
Stator resistance (p.u.)	0.0272	35/0.69 kV tran. rated power (MVA)	2
Stator reactance (p.u.)	0.5131	35/0.69 kV tran. impedadance (p.u.)	25.5
220/35 kV tran. rated Power (MVA)	50	220/35 kV tran.impedadance (p.u.)	500
Flux linkage (p.u.)	1.1884	Leakage reactance of T1 (p.u.)	0.0415
Inertia constant (s)	1.4393	DC Bus Voltage (V)	1150
Filter resistance (p.u.)	0.03	Filter reactance (p.u.)	0.3
DC capacitor (F)	0.01	Length of MV cable (km)	30
Collecting line resistance (p.u)	0.01371	Collecting line reactance (p.u)	0.00807
MV cable resistance (Ω/km)	0.46	MV cable reactance(Ω/km)	0.23

The detailed control of the rotor side, grid side, and pitch controller has been implemented but the focus is not to show the performance of these control systems and hence are not discussed in this paper. The parameters of PMSG and the collection system are shown in Table I.

B. WT Clustering

The WTs are clustered based on the time series of input wind speed, active power output, terminal residual voltage during the wind speed and fault disturbance. Based on Fig. 3, the optimal number of clusters is determined to be $k = 3$ and $k = 4$ using the AIC and BIC criterion respectively. These values provide the lowest criterion value which is a trade-off between accuracy and simplicity (fewer clusters). To optimise performance, $k = 4$ is selected as the optimal value using ITA, based on the results from both the AIC and BIC criteria. This means the whole WF can be represented by 4 aggregated WTs.

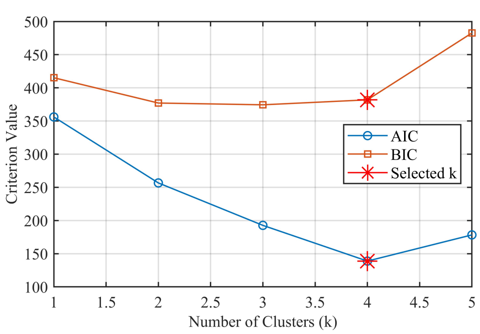

Fig. 3. Optimum number of cluster

The clustering results from the GMM algorithm are listed in Table II. Table II also provides a reference for the time spot-based classic clustering method, using 4s as the example time spot in wind speed disturbance. A unique clustering results is obtained using the proposed clustering method. However, the traditional time spot-based clustering result is not the same because it only captures turbine behavior at particular moments (4 s in this case) and consider turbine's operational characteristics constant at other times. To further validate results, a three phase to ground fault starting at 8 s and ending at 8.2 s is applied at PCC as shown in Fig. 1. The clustering options are presented in Table III.

TABLE II
WTS CLUSTERING OPTIONS WITH WIND SPEED DISTURBANCE

Clustering Criteria	Cluster No.	WT No.
Wind speed disturbance	Cluster 1	13,15,23,28∼30,34,37,38, 44,52,54,57,59,60,66
	Cluster 2	6,7,9,12,20,22,25,26,31,32, 35, 41,42,45,47,56,61
	Cluster 3	1∼5,14,18,24,27,33,43, 46,51,55,62∼65
	Cluster 4	8,10,11,16,17,19,21 36,39,40,48∼50,53,58
4s in wind speed disturbance	Cluster 1	13,15,23,28∼30,34,37,38, 44,52,54,57,59,60,66
	Cluster 2	6,7,9,12,20,22,25,26,31,32, 35, 41,42,45,47,56,61
	Cluster 3	1∼3,5,14,18,24,27,33,43, 46,49,51,55,64,65
	Cluster 4	4,8,10,11,16,17,19,21 36,39,40,48,50,53,58,62,63

TABLE III
WTS CLUSTERING OPTIONS WITH FAULT DISTURBANCE

Clustering Criteria	Cluster No.	WT No.
Fault disturbance	Cluster 1	13,15,23,28∼30,34,37,38, 44,52,54,57,59,60,66
	Cluster 2	6,7,9,12,20,22,25,26,31,32, 35, 41,42,45,47,56,61
	Cluster 3	1∼5,14,18,24,27,33,43, 46,51,55,62∼65
	Cluster 4	8,10,11,16,17,19,21 36,39,40,48∼50,53,58

As it can be seen, the clustering options given for a fault disturbance and a wind speed disturbance are identical. This is because wind turbines constantly experience random wind fluctuations, which have a similar effect on their dynamic behavior as faults do. As a result, a wind turbine can show different dynamic responses depending on the operating conditions (wind speed, power output), making wind disturbances resemble system faults in their effect on the system. So, an accurate clustering options for WTs can be achieved by the proposed GMM based clustering method under dynamic conditions.

C. Performance Evaluation of Proposed Model

To assess the robustness of the proposed WF aggregated model, we test its ability to represent dynamic responses under various disturbance conditions. The specific characteristics of these disturbances are outlined below:

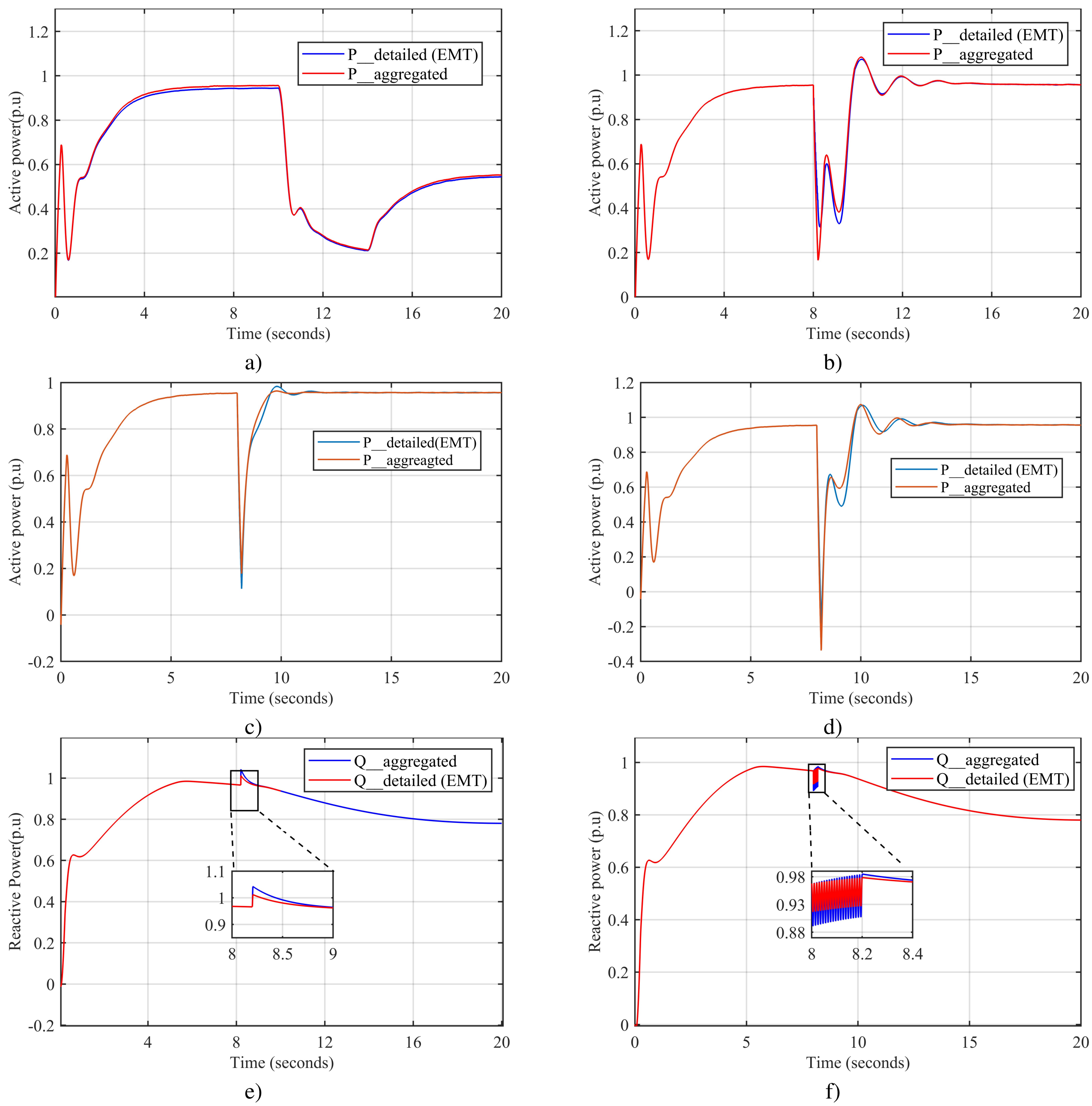

Fig. 4. Active power dynamics under (a) wind speed disturbance (b) three phase fault (c) single line to ground fault (d) double line to ground fault; Reactive power dynamics under (e) Three phase fault (f) single line to ground fault.

1) For the wind speed disturbance, the initial wind speed is 12 m/s and reduced to 3 m/s at 10 s then increased to 8 m/s at 14 s.

2) For the fault disturbance, a three-phase-to-ground fault and asymmetric fault occurs at the PCC at 8 s and recovers at 8.2 s.

These cases are shown in Fig. 4 which compares between EMT model and aggregated model.

To evaluate performance of aggregated models, Root mean squared error (RMSE) is commonly used as an effective method for error quantification [10]. Table IV gives the RMSE results which is calculated using (9).

$$
\begin{aligned}
RMSE_P &= \left[\frac{1}{t_2 - t_1}\int_{t_1}^{t_2}\left(\frac{p_{\text{aggregated}} - p_{\text{detailed}}}{S}\right)^2 dt\right]^{1/2} \\
RMSE_Q &= \left[\frac{1}{t_2 - t_1}\int_{t_1}^{t_2}\left(\frac{Q_{\text{aggregated}} - Q_{\text{detailed}}}{S}\right)^2 dt\right]^{1/2}
\end{aligned} \tag{9}
$$

Here, t_1 denotes the transient's start time and and t_2 its end time. $S = \sqrt{P^2_{\text{steady\_state}} + Q^2_{\text{steady\_state}}}$

Comparative analysis shows, the detailed and aggregated models produce nearly identical transient responses with

TABLE IV
COMPARISON OF EVALUATION INDEX WITH DETAILED MODEL

Case study	P_RMSE	Q_RMSE
Wind speed disturbance	1.9%	1.3%
Fault disturbance	1.7 %	1.4%

matching oscillation frequencies and damping properties, ultimately converging to the same steady state. There are minor differences giving very small RMSE of P and Q, compared with EMT model under wind speed and fault disturbance. This demonstrates the GMM ability to effectively model the complex interactions and nonlinearities in WT behavior, that closely mimic the detailed system's dynamic behavior.

TABLE V
COMPARISON OF COMPUTATIONAL TIME FOR 1 S SOLVING WITH DIFFERENT MODELS

Aggregation types	Computational time
Proposed model	0.14 s
Detailed model	354 s
Reduced order model [11]	0.306 s
Admittance based reduction model [3]	20 s

D. Computational Time Enhancement

The computational efficiency of the proposed aggregated model versus the detailed model was evaluated by measuring their respective simulation times. The computations are performed on a Dell latitude 5440 with 16 GB of RAM, 13th Gen Intel(R) Core(TM) i7-1355U running at 1.7 GHz, and a Windows 11, 64 bit operating system. The proposed model is more than 2500 times faster compared to detailed model as shown in Table V. The proposed model is the most computationally efficient among the models compared, making it an attractive option for scenarios requiring rapid simulations.

IV. CONCLUSION

This paper proposes an aggregation method based on GMM clustering, offering a distinct WT clustering options with optimum number of clusters (only 4 compared to the 11 in EMT model). The proposed model demonstrates high accuracy (less than 2% error) under dynamic conditions with computational burden factor reduction of 2500 times in compare with detailed EMT model enabling the analysis of larger-scale wind farms. To enhance the performance of WT aggregation modelling further, a multi-objective key parameter identification approach will be further studied, which can effectively reduce the number of parameters involved in the aggregation process.

ACKNOWLEDGMENT

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/X031144/1].

REFERENCES

[1] V. Jalili-Marandi, L.-F. Pak, and V. Dinavahi, "Real-time simulation of grid-connected wind farms using physical aggregation," *IEEE Transactions on industrial electronics*, vol. 57, no. 9, pp. 3010–3021, 2009.

[2] Y. Zhou, L. Zhao, I. B. Matsuo, and W.-J. Lee, "A dynamic weighted aggregation equivalent modeling approach for the dfig wind farm considering the weibull distribution for fault analysis," *IEEE Transactions on Industry Applications*, vol. 55, no. 6, pp. 5514–5523, Dec. 2019.

[3] J. Martínez-Turégano, S. Añó-Villalba, S. Bernal-Perez, and R. Blasco-Gimenez, "Aggregation of type-4 large wind farms based on admittance model order reduction," *Energies*, vol. 12, no. 9, p. 1730, 2019.

[4] J. Zou, C. Peng, H. Xu, and Y. Yan, "A fuzzy clustering algorithm-based dynamic equivalent modeling method for wind farm with dfig," *IEEE transactions on energy conversion*, vol. 30, no. 4, pp. 1329–1337, Dec. 2015.

[5] P. Wang, Z. Zhang, Q. Huang, N. Wang, X. Zhang, and W.-J. Lee, "Improved wind farm aggregated modeling method for large-scale power system stability studies," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 6332–6342, Nov. 2018.

[6] M. H. Uddin, J. Khalid, M. Fawad, M. Smailes, C. Jia, R. Simmonds, S. Wang, and J. Liang, "Reliability and cost assessment of ac/dc networks for offshore wind farm integration," in *2024 59th International Universities Power Engineering Conference (UPEC)*. IEEE, 2024, pp. 1–6.

[7] H. Wan, H. Wang, B. Scotney, and J. Liu, "A novel gaussian mixture model for classification," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 2019, pp. 3298–3303.

[8] K. P. Burnham, D. R. Anderson, K. P. Burnham, and D. R. Anderson, *Practical use of the information-theoretic approach*. Springer, 1998.

[9] J. Khalid, M. H. Uddin, M. Fawad, M. Smailes, C. Jia, R. Simmonds, S. Wang, and J. Liang, "Self-tunning converter control of pmsg-based wind turbine using multi-agent deep reinforcement learning," in *2024 59th International Universities Power Engineering Conference (UPEC)*. IEEE, 2024, pp. 1–6.

[10] W. Li, I. Kaffashan, A. M. Gole, and Y. Zhang, "Structure preserving aggregation method for doubly-fed induction generators in wind power conversion," *IEEE Transactions on Energy Conversion*, vol. 37, no. 2, pp. 935–946, Jun. 2022.

[11] M. G. Taul, X. Wang, P. Davari, and F. Blaabjerg, "Reduced-order and aggregated modeling of large-signal synchronization stability for multiconverter systems," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 9, no. 3, pp. 3150–3165, Jun. 2021.