Energy Proceedings
Vol 54, 2025
ISSN 2004-2965

The Aggregation of Wind Time-Series in Capacity Expansion Models[#]

Dana Reulein*[1], Carmen Li[2], Dimitri Pinel[3], Christian Andresen[3], Hossein Farahmand[1], Masood Parvania[4]

1 Department of Electric Energy, Norwegian University of Science and Technology, Norway

2 Oxford Institute for Energy Studies, United Kingdom

3 Department of Energy Systems, SINTEF Energy Research, Norway

4 Department of Electrical and Computer Engineering, University of Utah, United States
(Corresponding Author: dana.reulein@ntnu.no)

ABSTRACT

Capacity Expansion Models (CEMs) are widely used in the academic literature to understand the needs and dynamics of highly renewable energy systems. Due to computational constraints, it is common to aggregate time-series data such as hourly power output from Variable Renewable Energy Sources (VRES) using clustering algorithms. However, there is evidence that the presence of wind data leads to increased clustering errors and biased investment decisions. With the above motivation, we combine two approaches from the literature and compare them against the state-of-the-art approach. For a small number of clusters, the proposed approach recovers 95% of the original variance and correlation. This leads to more robust investment decisions. However, we stress the increased computational burden involved.

Keywords: time-series aggregation, wind energy, power system modeling, optimization, clustering

NOMENCLATURE

Abbreviations	
CEM	Capacity Expansion Model
ED	Euclidean Distance
ENS	Energy Not Served
EU	European Union
DTW	Dynamic Time Warping
OCGT	Open-Cycle Gas Turbine
VRES	Variable Renewable Energy
Symbols	
$CorrE$	Correlation Error
i	Index of Hours
$\mathcal{K}$	Total Number of Clusters
k	Index of Clusters
$\mathcal{M}$	Dissimilarity Matrix
$\mathbb{M}_{k,r,t}$	Mean Vector for k,t,r
$\tilde{\mathbb{M}}_{k,r,t}$	Duration Curve for $\mathbb{M}_{k,r,t}$
$N(\mathcal{R})$	Number of Regions
$N(k)$	Number of Days belonging to k
$\mu_{t,r}$	Mean for t,r
$\hat{P}_{t,r,i}$	Normalized Value for $P_{t,r,i}$
r	Index of Regions
$\mathcal{R}$	Set of all Regions
$\sigma_{t,r}$	Standard Deviation for t,r
t	Index of Time-Series Types
$\mathcal{T}$	Set of all Time-Series Types
π	Set of all Paths
$VarC_t$	Covered Variance for t
$var^*_{t,r}$	Approximated Variance for t,r
$var_{t,r}$	Original Variance for t,r
w	Warping Window
w_k	Weight of k
$\tilde{\mathbb{X}}_{k,r,t}$	Duration Curve for k,t,r
$\tilde{\mathbb{Y}}_{k,r,t}$	Average Duration Curve for k,t,r

1. INTRODUCTION

The transition from fossil fuels to renewable energy sources has become an important part of the EU's long-term strategy to reach net zero emissions by 2050 [1]. In particular, electricity generation from onshore and offshore wind farms is expected to become more dominant in 2050's energy mix [2]. However, the expansion of wind energy requires substantial upfront capital investments in generation, storage, and transmission. To identify cost-efficient investment decisions, models based on mathematical optimization (e.g. CEMs) are widely used in the literature. However, the temporal simplifications required to make them computationally feasible [3] often lead to an inadequate representation of the uncertain and variable nature of VRES.

In particular, the aggregation of time-series data (e.g. VRES capacity factors, electricity demand) is common practice in the international energy research community, as the size of the input data directly influences the size of

This is a paper for the 16th International Conference on Applied Energy (ICAE2024), Sep. 1-5, 2024, Niigata, Japan.

the related optimization problem, and thus the computational cost [4]. Different methods to aggregate time-series data have been reported in the literature, each of which has its own set of advantages and disadvantages. Among these, clustering has become a well-established method for grouping and identifying similar periods (e.g. hours, days, weeks). In the context of CEMs, the Euclidean distance (ED) is often used to measure the similarity or dissimilarity between different periods. It refers to the hour-to-hour distance between two time-series in the Euclidean space. However, a premise for grouping similar periods is that the data must have hourly, daily, or seasonal patterns. Unlike PV generation or electricity demand, electricity generation from wind energy lacks clear patterns and has complex correlations over space and time (see Fig. 2) [5]. This issue has been recently raised in [6], where the authors show that wind data leads to clustering errors and thus to biased investment decisions. To surmount these issues, we propose two methodological improvements to the state-of-the-art approach.

First, we employ Dynamic Time Warping (DTW) to measure the dissimilarity or similarity between different periods. As opposed to a rigid hour-to-hour comparison with Euclidean matching, DTW performs an optimal matching that accounts for temporal shifts [7] and adds thus more flexibility with respect to random fluctuations of wind energy. The concept of DTW was first proposed by Sakoe and Chiba [8] to identify speech patterns and has been widely applied in various other fields. However, it is only sparely used for the aggregation of time-series data in CEMs, likely due to higher computational demands and the requirement for accurate parametrization. For example, the authors in [9] proposed two hierarchical clustering methodologies using DTW to aggregate load, solar, and wind time-series into a set of representative days. Both yielded better investment decisions compared to the k-means clustering algorithm with the ED. In [10], the authors employed a centroid-based partitional clustering algorithm [11] that uses DTW as its distance measure. However, a case study with representative days of electricity price data revealed that it did not perform as well as the k-shape method [12].

Secondly, we use a novel method for selecting the representative period as proposed in [13]. The prevailing approach within the energy research community is to use the medoid, the data point closest to the cluster's centroid [14]. Instead of relying on data from the original time-series, we calculate an artificial data point that better describes the distribution of the original time-series.

In view of the above, our main contribution lies in implementing and combining two methodologies for the aggregation of time-series data. In particular, the question is how to capture the statistical features of wind generation in CEMs using representative days. To tackle this question, the paper is organized as follows: Chapter 2 describes the generalized comparison framework for the aggregation of time-series data in CEMs. We test this approach in a case study, which is described in Chapter 3. In Section 4, data- and model-based indices are calculated and discussed. Finally, a conclusion is drawn.

2. METHODOLOGY

The proposed framework to identify a set of representative periods that captures the statistical features most relevant to the optimization problem is inspired by [10] and [14]. Our five-step approach is implemented in Julia 1.9.2[1] and can be applied to a discrete number of n-dimensional time-series vectors. In our case, we have $\mathcal{R} \times \mathcal{T}$ time-series vectors, each with 8,760 hourly values, where r denotes a specific region within all regions $\mathcal{R}$, and t represents a specific type of time-series data within all types $\mathcal{T}$ (e.g. solar PV, wind, electricity demand). In the following, the process for the aggregation of representative days is illustrated (see Fig. 1) and further described.

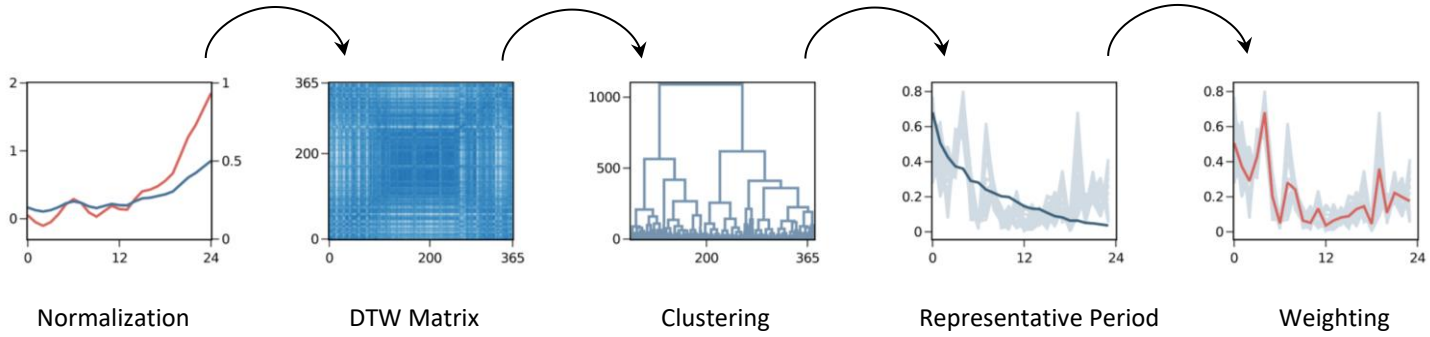

Fig. 1 Proposed Methodology

[1] https://github.com/danareu/ts_clustering

Step 1: Normalization

Normalization is a crucial step to identify patterns in time-series with different units (e.g. VRES capacity factors and electricity demand). Here, we apply z-normalization, which is typically used in the context of DTW [9], [10], to obtain a mean of zero and a standard deviation of one:

$$\hat{P}_{t,r,i} = \frac{1}{\sigma_{t,r}}(P_{t,r,i} - \mu_{t,r}) \quad (1)$$

where $\hat{P}_{t,r,i}$ is the normalized value in hour i, $P_{t,r,i}$ is the original value, $\sigma_{t,r}$ the standard deviation and $\mu_{t,r}$ the mean [6].

Step 2: DTW Similarity

Afterward, we transform the data in such a way that each day becomes a column vector that contains 24 hourly normalized values for each feature combination r and t. After each feature combination, we insert a 24-element zero vector. This limits the flexibility of the warping function to only allow time shifts within a day and between the same features, assuming a user-defined offset (warping window) $w \leq 24$.

Given two daily vectors d_1 and d_2, each of size $n \times 1$, the warping function defines a path through the $n \times n$ matrix $D_{1,2}$. The latter represents the hour-to-hour distances between d_1 and d_2 using any arbitrary distance metric. The optimal path has the least cumulative distances across $D_{1,2}$. Here, we use the Squared Euclidean distance such that the dissimilarity between d_1 and d_2 can be described as:

$$DTW(d_1, d_2) = \min \sum_{(i,j)\in\pi} (d_{1,i} - d_{2,j})^2 \quad (2)$$

where π is the set of all paths between all time-series elements. However, DTW as implemented in (2) suffers from a combinatory problem which has a $O(N^2)$ time and space complexity [15]. We thus use the Julia package "DynamicAxisWarping.jl"[2] which provides a well-optimized implementation of the dynamic programming problem. It further supports limiting the warping window and with it the requirement for computational resources. We retrieve the optimal cumulative distances for all possible vector combinations and store the result in the 365×365 dissimilarity matrix $\mathcal{M}$. Notice that the warping path for $w = 0$ coincides with the diagonal and thus yields the same result as Euclidean matching.

Step 3: Hierarchical Clustering

We apply agglomerative hierarchical clustering, the most common type of hierarchical clustering, to partition our distance matrix $\mathcal{M}$ [16]. Hierarchical clustering algorithms are deterministic and thus yield reproducible clusters [17], which is a desirable aspect for our comparison framework. We further employ Wards' linkage criterion [18] to determine the merging of clusters, which is often used in the context of time-series aggregation for CEMs [4], [10], [11]. Accordingly, clusters are step-wise merged into subgroups such that the increase in the sum of squared errors is minimized at each step [18]. Here, we built a dendrogram of nested clusters[3], from which we retrieve the desired number of clusters $\mathcal{K}$ and their cluster members.

Step 4: Representative Period

After grouping the data, we employ a novel approach from [13] to better capture the statistical features of the original time-series. It consists of five steps:

1. Determine the duration curve $\widetilde{\mathbb{X}}_{k,r,t}$ using the values of all cluster elements of cluster k and feature combination t and r.
2. Average $\widetilde{\mathbb{X}}_{k,r,t}$ accordingly to yield a 24-dimensional vector $\widetilde{\mathbb{Y}}_{k,r,t}$.
3. Calculate the mean $\mathbb{M}_{k,r,t}$ for each cluster k and feature combination t and r.
4. Determine the duration curve $\widetilde{\mathbb{M}}_{k,r,t}$ for $\mathbb{M}_{k,r,t}$ and store its index order.
5. Assign the index order from $\widetilde{\mathbb{M}}_{k,r,t}$ to the sorted values of $\widetilde{\mathbb{Y}}_{k,r,t}$ and sort its indexes in ascending order.

It is important to note that this method involves sorting, which adds on average a $O(N\ log(N))$ complexity component.

Step 5: Weighting

In CEMs, each representative day is weighted according to the relative size of their respective cluster [14]. The specific weight w_k for cluster k is the factor between the number of cluster elements of cluster k and the total number of clusters $\mathcal{K}$.

[2] https://github.com/baggepinnen/DynamicAxisWarping.jl
[3] https://github.com/JuliaStats/Clustering.jl

$$w_k = \frac{N(k)}{\mathcal{K}} \tag{3}$$

3. CASE STUDY

To measure the accuracy and computational cost of our proposed methodology, we evaluate both data- and model-based indices. Due to their central role in the future supply of offshore wind energy, we optimize investments in eight neighboring countries bordering the North Sea. First, we aggregate the data according to our proposed methodology. Next, we use the reduced data set as model input and analyze the solution quality.

3.1 Data

For our case study, we collect 8,760-dimensional time-series vectors from $\mathcal{R} \times \mathcal{T}$ feature combinations, where $\mathcal{T}$ includes PV, offshore wind, onshore wind, and electricity demand, and $\mathcal{R}$ includes Austria (AT), Belgium (BE), Switzerland (CH), Czech Republic (CZ), Germany (DE), Denmark (DK), Estonia (EE), Spain (ES), Finland (FI), France (FR), Ireland (IE), Lithuania (LT), Luxembourg (LU), Latvia (LV), Netherlands (NL), Norway (NO), Poland (PL), Sweden (SE), and United Kingdom (UK).

Furthermore, hourly power output for wind and solar PV farms are extracted from renewables.ninja [5], [19], [20] based on the weather year 2018. Renewables.ninja is a widely recognized open-access platform that uses NASA's MERRA-2 reanalysis data to simulate hourly power output from wind and solar PV farms. The profiles are based on the present wind and solar fleet and grouped to a country level.

We obtained electricity demand data using "entsoe-py"[4], a Python client for the ENTSO-E API. We resampled the time-series data to reduce the granularity from 15-minute intervals to hourly intervals.

To better understand the data, we computed the Pearson correlation matrix (see Fig. 2). For illustration purposes, we only considered eight regions. Accordingly, solar production is highly spatially correlated, as the effects of the Earth's rotation and orbit on diurnal and seasonal patterns outweigh local weather variations [21]. Wind energy, on the other hand, exhibits less pronounced but more complex spatial correlations. Interestingly, both wind and solar generation are slightly correlated with electricity demand, but not correlated with each other.

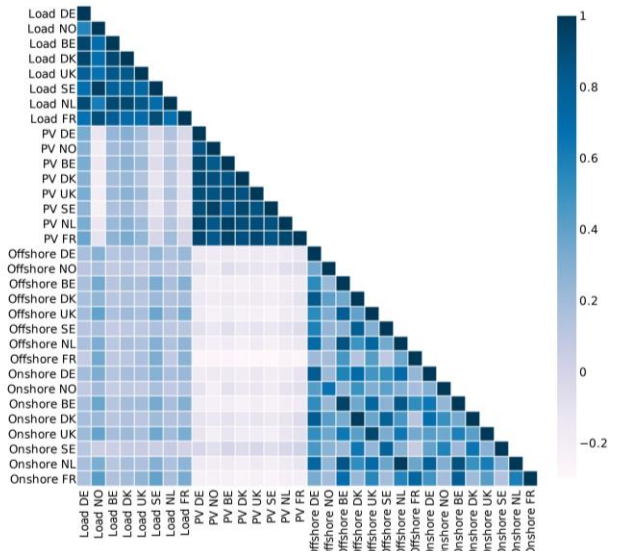

Fig. 2 Heatmap showing the correlations across all time-series

3.2 Capacity Expansion Model

The open-source energy system model GENeSYS-MOD is a branched version of OSeMOSYS, originally developed in 2011 by [22]. It is a linear model that optimizes dispatch and investment decisions in generation, storage, and transmission from 2018 to 2050 in 10-year intervals. The model has perfect foresight of future events and minimizes costs from a social planner's perspective with respect to increasing CO_2 prices. For more details, the interested reader is referred to [23].

GENeSYS-MOD models power, residential heating, industry, and transportation. However, we only consider the power sector, to limit the number of feature combinations, and thus the complexity of the problem. We also use its Julia implementation and solve the problem using the parallel barrier algorithm from CPLEX v1.0.3 to yield deterministic results. Both the dataset and the model are publicly available on GitHub[5].

3.3 Computational Complexity

If we were to run each year of operation in full hourly resolution, our model would require 229 million variables and 253 million constraints (including variable bounds and integrality constraints). However, it is not only the size of the model but also the hardware that determines the time it takes to build and solve the model [13]. We run our model on the following node of a high-performance cluster:

[4] https://github.com/EnergieID/entsoe-py

[5] https://github.com/GENeSYS-MOD

Tab. 1 Computational Resources

CPU Model	Intel Xeon Gold 6242
Number of Cores per Node	28
Threads per Core	4
Shared Memory [GB]	192
CPU Max Frequency [GHz]	3.9

4. RESULTS

In the following two sections, we evaluate data- and model-based indices for our proposed methodology and the state-of-the-art approach. For the latter, we use the same methodology as outlined in Section 2, except that we apply Euclidean matching ($w = 0$) and select the medoid as cluster representative.

4.1 Data-based Indices

First, we are interested in how well our representative periods describe the distribution of the original time-series. More specifically, we look at the spread of the distribution and use the variance as a metric for comparison. We take the representative periods and their weights to approximate a $8{,}760 \times 1$ vector for each t and r combination. This approach allows us to evaluate both the cluster representative and its weight. The mean variance $var^*_{t,r}$ across all $r \in R$ is then divided through the original mean variance $var_{t,r}$. Thus, in this analysis, we ignore spatial differences. We yield the covered variability $VarC_t$ for each $t \in T$ as described in [14]. A value of one indicates a perfect representation of the original variance or variability.

$$VarC_t = \frac{1}{N(\mathcal{R})} \sum_{r \in \mathcal{R}} \frac{var^*_{t,r}}{var_{t,r}} \tag{4}$$

Fig. 3 shows this value for different numbers of clusters. In particular, the first row displays our proposed methodology using the Euclidean distance (ED). The shaded region around the ED includes all possible values for $w \in [1{,}5{,}10{,}15{,}23]$. The second row shows the covered variability for the state-of-the-art (ED + medoid). Note the different scales between both approaches.

For all wind, PV, and load time-series, our proposed methodology underestimates the original variance, albeit only to a relatively small degree. Due to the scaling to the duration curve, only a small number of clusters is required to describe 99% of the original variability. The results imply that the use of DTW leads to a faster convergence with the original value for wind and load. Interestingly, this is not the case for load or PV, where typical diurnal patterns are best captured with an hourly comparison using ED.

Compared to our proposed methodology, the state-of-the art approach requires a higher number of clusters to reduce the error between the approximated and the original variance. More specifically, around 20 clusters are needed to limit this deviation to below 5%. For a smaller number of clusters, the approximated variance lies either 40% below or above the original value. Interestingly, the variance of wind is underestimated, whereas it is overestimated for PV and load.

Next, we calculate the Pearson correlation coefficient according to (5) as correlations and anticorrelations are important to determine the optimal mix of substitute and complementary energy sources [21].

$$CorrE = \frac{1}{N(\mathcal{T} \times \mathcal{R})^2} corr^* - corr \tag{5}$$

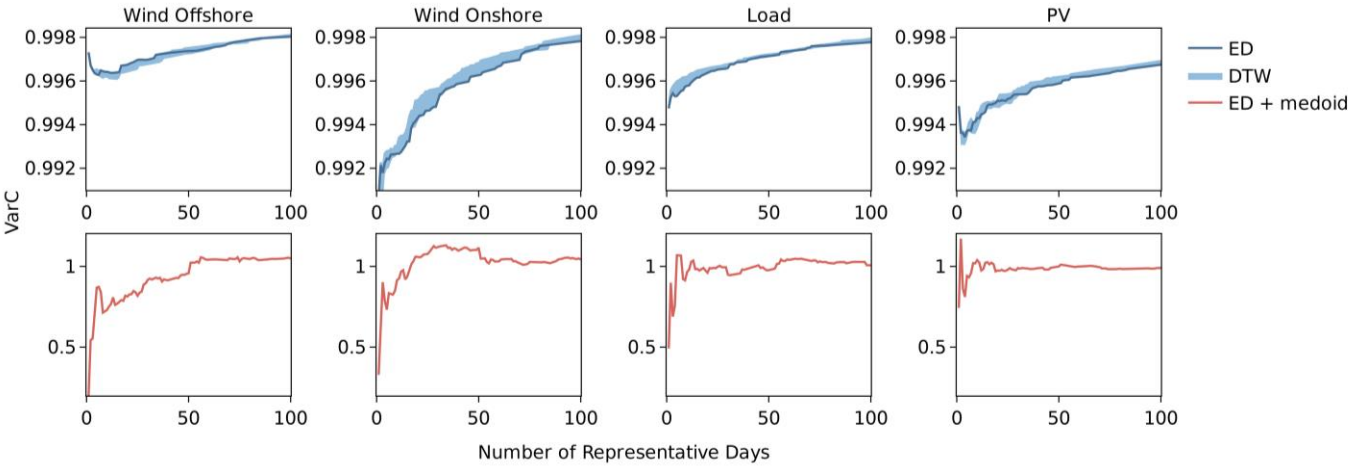

Fig. 3 Covered variance for Wind Offshore, Wind Onshore, Load and PV, respectively. All values are calculated as mean across all $r \in \mathcal{R}$.

We proceed exactly as before and approximate a $8{,}760 \times 1$ vector for each t and r combination. With $\mathcal{T} = 4$ and $\mathcal{R} = 19$, we compute a 76×76 correlation matrix for a specific number of clusters. Finally, we subtract the original correlation matrix from this matrix and calculate the mean value.

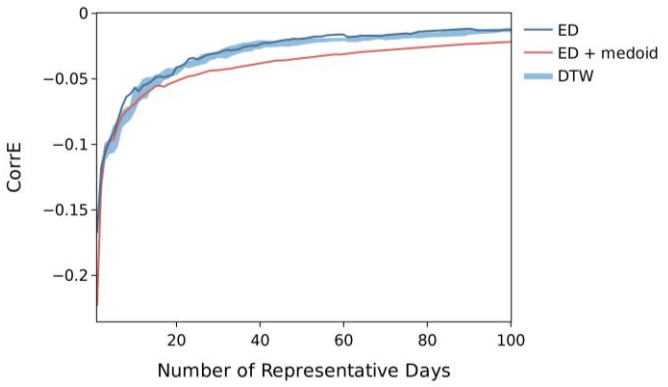

Fig. 4 Correlation Error

According to Fig. 4, all methods yield similar results for a small number of clusters. For $k > 10$, there is a difference between the state-of-the-art and ED. We observe that the new representative selection method converges faster and thus yields smaller errors. Also, the shaded regions indicate that the warping window can be beneficial in some cases.

We deduce that our proposed approach recovers the distribution and correlation of the original time-series to a greater degree and thus provides better results as opposed to the state-of-the-art. This applies to wind, solar PV, and load time-series, respectively.

4.2 Model-based Indices

Data-based metrics tell us how well each method describes the input data. Yet, in energy system modeling, it comes down to how accurately we can approximate the solution when we run the model at full hourly resolution. For this purpose, we set up both the model with different warping windows and number of clusters and use the objective function value as an indicator of the solution quality. In our case, it corresponds to the total discounted system costs.

However, we have no reference for the objective function value, as running the model at full hourly resolution would exceed our available computational resources. Therefore, we temporarily store the optimization result for transmission, storage, and production capacities for 2050. Then, we fix those investments in 2050 and run the model for 2050 with full hourly resolution. We further add slack variables to determine the amount of Energy Not Served (ENS). We use the level of ENS as an indicator of the robustness of investment decisions to meet demand in 2050. We argue that a solution with low ENS is close to the solution with full hourly resolution.

In Fig. 5, the results are structured column-wise for each model configuration (investment and dispatch mode/dispatch mode). We analyze the model results for $k \leq 20$. The graph in the first column shows the objective function value (left y-axis) and the total installed capacity of wind energy in 2050 (right y-axis). Except for the capacities, the shaded region around the ED includes all possible values for $w \in [1{,}5{,}10{,}15{,}23]$. Here, we calculated the mean value.

First, we look at the objective function value: It increases as more representative days are considered, which indicates that all methods analyzed in this paper are underestimating the system cost. Interestingly, the

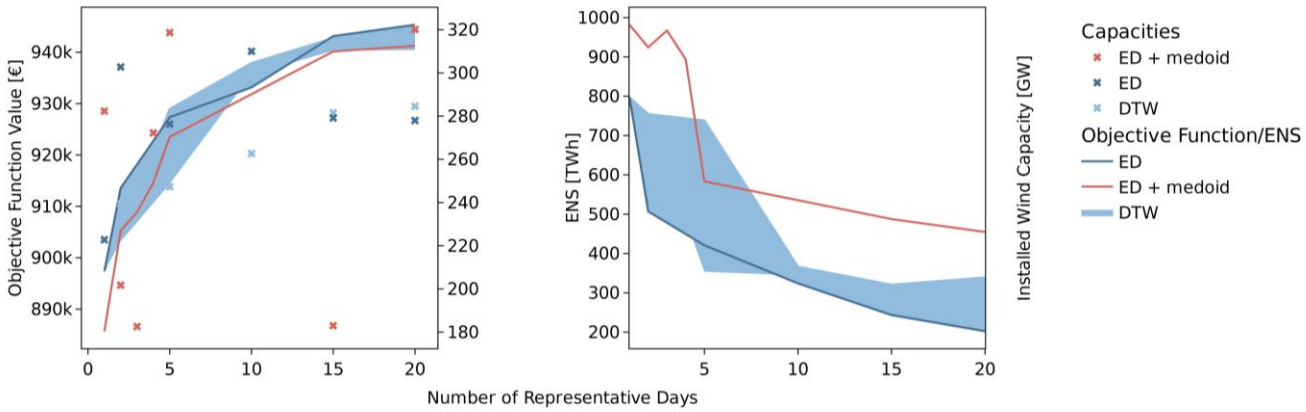

Fig. 5 Optimization Results with different configurations. The first column shows the objective function value and installed wind Capacity. The second column shows the amount of ENS in 2050.

state-of-the-art approach has lower costs compared to ED and DTW. This deviation is on average 0.5%. We can also see that a warping window $w > 0$ can sometimes lead to higher or lower costs compared to the ED. When looking at the wind capacity investments, we observe that the variability of wind investments is greater for the state-of-the-art approach. This can be attributed to the high variability among different clusters, especially for $k \leq 20$.

Similarly, as the number of clusters increases, the amount of ENS decreases. In general, we see that the state-of-the-art approach is above our proposed method and in most cases also above DTW. We deduce that our proposed approach leads to more robust investment decisions for a small number of clusters. However, we want to stress the increased computational burden involved.

5. Conclusion

In this paper, we proposed a methodology to capture the variable nature of wind energy in CEMs. In particular, we applied two approaches from the academic literature and used both data- and model-based indices to compare it against the state-of-the-art.

The results show that the proposed approach better recovers the variance and correlation of the original data. Furthermore, we show that our methodology outperforms the state-of-the-art and thus provides more robust investment decisions. In addition, we stress the additional $O(N\,log(N))$ and $O(N^2)$ complexity components.

Due to the scope of this paper, some research streams have been left open and remain unexplored. For example, future work should include more data-based indices. In addition, it would be insightful to investigate a range of potential investment options and to consider the feasibility of incorporating an approximation of long-term storage. Also, it was not investigated which parametrization is beneficial for the warping window. Lastly, we want to refer to FastDWT [15], an approximation of the DTW problem which has linear time and space complexity.

Acknowledgment

The work presented in this paper was conducted as part of the Ocean Grid project, a Green Platform project financed in part by The Research Council of Norway with project number 328750.

References

[1] European Commission, '2050 long-term strategy'. Accessed: Jul. 08, 2024. [Online]. Available: https://climate.ec.europa.eu/eu-action/climate-strategies-targets/2050-long-term-strategy_en

[2] D. Fraile *et al.*, 'Getting fit for 55 and set for 2050', 2021. [Online]. Available: https://etipwind.eu/files/reports/Flagship/fit-for-55/ETIPWind-Flagship-report-Fit-for-55-set-for-2050.pdf

[3] S. Collins *et al.*, 'Integrating short term variations of the power system into integrated energy system models: A methodological review', *Renewable and Sustainable Energy Reviews*, vol. 76, pp. 839–856, Sep. 2017, doi: 10.1016/j.rser.2017.03.090.

[4] L. Kotzur, P. Markewitz, M. Robinius, and D. Stolten, 'Time series aggregation for energy system design: Modeling seasonal storage', *Applied Energy*, vol. 213, pp. 123–135, Mar. 2018, doi: 10.1016/j.apenergy.2018.01.023.

[5] I. Staffell and S. Pfenninger, 'Using bias-corrected reanalysis to simulate current and future wind power output', *Energy*, vol. 114, pp. 1224–1239, Nov. 2016, doi: 10.1016/j.energy.2016.08.068.

[6] L. E. Kuepper, H. Teichgraeber, N. Baumgärtner, A. Bardow, and A. R. Brandt, 'Wind data introduce error in time-series reduction for capacity expansion modelling', *Energy*, vol. 256, p. 124467, Oct. 2022, doi: 10.1016/j.energy.2022.124467.

[7] S. Bothwell, A. Kaizer, R. Peterson, D. Ostendorf, V. Catenacci, and J. Wrobel, 'Pattern-based clustering of daily weigh-in trajectories using dynamic time warping', *Biometrics*, vol. 79, no. 3, pp. 2719–2731, 2023, doi: 10.1111/biom.13773.

[8] H. Sakoe and S. Chiba, 'Dynamic programming algorithm optimization for spoken word recognition', *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 1, pp. 43–49, Feb. 1978, doi: 10.1109/TASSP.1978.1163055.

[9] Y. Liu, R. Sioshansi, and A. J. Conejo, 'Hierarchical Clustering to Find Representative Operating Periods for Capacity-Expansion Modeling', *IEEE Trans. Power Syst.*, vol. 33, no. 3, pp. 3029–3039, May 2018, doi: 10.1109/TPWRS.2017.2746379.

[10] H. Teichgraeber and A. R. Brandt, 'Clustering methods to find representative periods for the optimization of energy systems: An initial framework and comparison', *Applied Energy*, vol. 239, pp. 1283–1293, Apr. 2019, doi: 10.1016/j.apenergy.2019.02.012.

[11] F. Petitjean, A. Ketterlin, and P. Gançarski, 'A global averaging method for dynamic time warping, with applications to clustering', *Pattern Recognition*, vol. 44, no. 3, pp. 678–693, Mar. 2011, doi: 10.1016/j.patcog.2010.09.013.

[12] J. Paparrizos and L. Gravano, 'k-Shape: Efficient and Accurate Clustering of Time Series', in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, Melbourne Victoria Australia: ACM, May 2015, pp. 1855–1870. doi: 10.1145/2723372.2737793.

[13] M. Hoffmann, L. Kotzur, and D. Stolten, 'The Pareto-optimal temporal aggregation of energy system models', *Applied Energy*, vol. 315, p. 119029, Jun. 2022, doi: 10.1016/j.apenergy.2022.119029.

[14] P. Nahmmacher, E. Schmid, L. Hirth, and B. Knopf, 'Carpe diem: A novel approach to select representative days for long-term power system modeling', *Energy*, vol. 112, pp. 430–442, Oct. 2016, doi: 10.1016/j.energy.2016.06.081.
[15] S. Salvador and P. Chan, 'FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space'.
[16] C. Ma and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*, vol. 20. 2007. doi: 10.1137/1.9780898718348.
[17] H. Teichgraeber and A. R. Brandt, 'Time-series aggregation for the optimization of energy systems: Goals, challenges, approaches, and opportunities', *Renewable and Sustainable Energy Reviews*, vol. 157, p. 111984, Apr. 2022, doi: 10.1016/j.rser.2021.111984.
[18] J. H. Ward, 'Hierarchical Grouping to Optimize an Objective Function', *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, Mar. 1963, doi: 10.1080/01621459.1963.10500845.
[19] 'Renewables.ninja'. Accessed: Jul. 11, 2024. [Online]. Available: https://renewables.ninja/
[20] S. Pfenninger and I. Staffell, 'Long-term patterns of European PV output using 30 years of validated hourly reanalysis and satellite data', *Energy*, vol. 114, pp. 1251–1265, Nov. 2016, doi: 10.1016/j.energy.2016.08.060.
[21] C. Li, C. K. Chyong, D. M. Reiner, and F. Roques, 'Taking a Portfolio approach to wind and solar deployment: The case of the National Electricity Market in Australia', *Applied Energy*, vol. 369, p. 123427, Sep. 2024, doi: 10.1016/j.apenergy.2024.123427.
[22] M. Howells *et al.*, 'OSeMOSYS: The Open Source Energy Modeling System', *Energy Policy*, vol. 39, no. 10, pp. 5850–5870, Oct. 2011, doi: 10.1016/j.enpol.2011.06.033.
[23] K. Löffler, K. Hainsch, T. Burandt, P.-Y. Oei, C. Kemfert, and C. Von Hirschhausen, 'Designing a Model for the Global Energy System—GENeSYS-MOD: An Application of the Open-Source Energy Modeling System (OSeMOSYS)', *Energies*, vol. 10, no. 10, Art. no. 10, Oct. 2017, doi: 10.3390/en10101468.