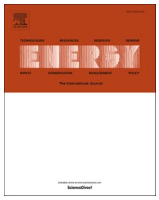
Energy 270 (2023) 126946

Contents lists available at ScienceDirect

Energy

journal homepage: www.elsevier.com/locate/energy



Scenario generation and risk-averse stochastic portfolio optimization applied to offshore renewable energy technologies

Victor A.D. Faria [a,*], Anderson Rodrigo de Queiroz [b], Joseph F. DeCarolis [c]

[a] Department of Operations Research at the NC State University, Raleigh, NC, USA
[b] Decision Sciences at NC Central University and the Department of Civil Construction and Environmental Engineering at NC State University, Durham, NC, USA
[c] Department of Civil Construction and Environmental Engineering at NC State University, Raleigh, NC, USA

ARTICLE INFO

Handling Editor: Neven Duic

Keywords:
Generative adversarial neural networks
Portfolio optimization
Risk assessment
Optimal site selection
Renewable energy

ABSTRACT

This work proposes an analytical decision-making framework considering scenario generation using artificial neural networks and risk-averse stochastic programming to define renewable offshore portfolios of wind, wave, and ocean current technologies. For the scenario generation, a generative adversarial neural network is developed to generate synthetic energy scenarios considering resources distributed over large geographic regions. These scenarios are then fed to a stochastic model, which objective to determine the optimal location and number of turbines for each technology. In the stochastic model formulation, a representation of the limits in the portfolio Levelized Cost of Energy and the maximization of the five percent lower energy generation conditions, also known as Conditional Value at Risk, is presented. The framework proposed here is tested considering data from a portion of the U.S. East coast, where the generative model was successful in creating energy scenarios statistically consistent with the historical data for wind, wave, and ocean current resources at more than 500 sites. Furthermore, the Conditional Value at Risk portfolio optimization model was used to construct efficient frontiers for a combination of different technologies, showing the significance of resource diversification as a tool to improve system security.

1. Introduction

With the increasing political and financial support directed to reducing the world's dependency on traditional energy resources such as coal, oil, and natural gas, the total participation of renewable energy generation more than double in the last decade [1], and according to the U.S. Energy Information Administration [2] by 2050 renewables will be the main source of primary energy consumption in the world. In this trend, renewables lead the investments in the power energy sector with approximately 350 Billion USD invested in 2020 [3].

Despite its potential, incorporating high levels of renewable energy into the power grid presents a number of challenges, including maintaining system stability and meeting demand cost-effectively. Various solutions have been proposed to address these issues. Reference [4] discusses the significance of energy storage in achieving a decarbonized energy matrix and analyzes different methods for optimal battery sizing. Reference [5] discusses the importance of improved hydropower forecasting for the operation of hydro-dominant systems and shows that ANNs can perform superior to more traditional methods in streamflow forecasting. Finally, reference [6] suggests exploring the complementarity of solar and wind energy to reduce energy variability and costs in the Mediterranean region.

Despite being in early stages of development and presenting costs that are still high, offshore energy technologies such as wind, wave, and ocean current may serve an important role in the future energy matrix by helping to reduce energy variability and increase system security, offering synergies within themselves and with other technologies. For instance, in Ref. [7], the complementarity between offshore wind and hydropower is identified for the different regions in Brazil. A significant reduction in energy variability is also identified in Ref. [6] when combining offshore wind and solar resources in the Mediterranean region. Reference [8] observed synergies between wind and wave energy in California, and [9] shows benefits in integrating offshore wind wave and ocean current in North Carolina using a mean-variance portfolio model.

From the perspective of assessing the resource quality of an energy generation portfolio, the Levelized Cost of Energy (LCOE), average capacity factor (CF), and variance are three of the most common measures found in the literature. While using these metrics, many works have

* Corresponding author.
E-mail addresses: vaduraes@ncsu.edu (V.A.D. Faria), adequeiroz@nccu.edu (A. Rodrigo de Queiroz), jfdecaro@ncsu.edu (J.F. DeCarolis).

https://doi.org/10.1016/j.energy.2023.126946
Received 22 June 2022; Received in revised form 6 February 2023; Accepted 11 February 2023
Available online 13 February 2023

Nomenclature

A. Abbreviations

ANN	Artificial Neural Network
BN	Batch Normalization
CF	Capacity Factor
CNN	Convolutional Neural Network
CVaR	Conditional Value at Risk
FID	Frechet Inception Distance
GANs	Generative Adversarial Neural Networks
IS	Inception Score
LCOE	Levelized Cost of Energy
MILP	Mixed-Integer Linear Programming
MMD	Maximum Mean Discrepancy
MMD_RN	MMD considering the real data and the data sampled from a multivariate gaussian distribution. Each MMD is computed over 365 (1-year) samples
MMD_RS	MMD considering the real data and the synthetic data created by the GAN model. Each MMD is computed over 365 samples
MMD_RR	MMD of the real data with itself, considering 365 samples
MMD_SS	MMD of the synthetic data created by the GAN with itself, considering 365 samples
PDF	Probability Density Function
PV	Photovoltaic
RBF	Radial Basis Function
VaR	Value at Risk

B. Indices and Sets

$e \in E$	Set of energy technologies (wind, wave, and ocean current)
$i \in I_e$	Set of feasible site locations associated with the technology e
$k \in D^R_{e,i}$	Set of site locations of the energy technology $e \in E$ that are less than R km from the site (e,i), $D^R_{e,i} \subset I_e$.
$x^h_i \in X_h$	Set of historical data samples for the MMD computation, $i = \{1,\ldots,N_h\}$
$x^s_i \in X_s$	Set of synthetic data samples for the MMD computation, $i = \{1,\ldots,N_s\}$
$\zeta_s \in \zeta$	Set of synthetic scenarios for the portfolio optimization model

C. Parameters

$C_{e,i}$	Annualized cost of deploying one turbine of the technology e at the site location $i \in I_e$ [$/Year-Turbine]
EG_{e,i,ζ_s}	Expected daily energy generation of one turbine of the technology e at the site location $i \in I_e$ and scenario ζ_s [MWh]
$\overline{LCOE}$	Upper bound in the Levelized Cost of Energy [$/MWh]
N_h	Number of days in the historical samples
N_s	Number of days in the synthetic samples
$\overline{Nt_{e,i}}$	Maximum number of turbines of the technology e that can be deployed at the site location $i \in I_e$
TNt_e	Total number of turbines deployed of the technology e
μ	Mean vector for the real/synthetic data (μ_r/μ_s)
Σ	Variance covariance matrix for the real/synthetic data (Σ_r/Σ_s)
σ	RBF kernel bandwidth (a hyperparameter)
$\alpha\%$	A pre-defined percentage value considered in the *CVaR* computation (e.g., 5%)

D. Decision Variables

c	Value at Risk (*VaR*). An auxiliary variable in the *CVaR* computation, $c \geq 0$
$x_{e,i} \in X$	Number of turbines of the technology $e \in E$ deployed at the site $i \in I_e$, $x_{(e,i)} \in Z^+$
$v_{e,i} \in V$	Binary variable responsible for controlling the center of the energy collection system for each energy technology, $v_{e,i} \in \{0,1\}$
$z_s \in Z$	An auxiliary variable in the *CVaR* computation, $z_s \geq 0$

E. Functions

$CVaR_\alpha(X,\zeta)$	Expected energy generation of the $\alpha\%$ lower generation scenarios (ζ) given a decision X
$MMD^2(X_h,X_s)$	Squared MMD of the historical data (X_h) and synthetic data (X_s)
$k(\cdot,\cdot)$	Kernel Function

followed a mean-variance portfolio approach for determining the optimal combination of site locations and technologies in a portfolio. For example [9], used the technique to minimize the variance in generation at given LCOE levels in the context of offshore energy deployments [10], used it in the context of onshore wind energy repowering, and [11] in the context of solar PV studies. In general, in a mean-variace model, one attempts to minimize the portfolio risk (variance of return) given a range of different expected return targets. However, one limitation of the mean-variance approach is that, since variance is a symmetric function, it penalizes both high and low generation conditions.

As an alternative to model asymmetric risks in portfolio optimization problems, Rockafellar and Uryasev [12] proposed a measure called Conditional Value at Risk (CVaR). The α-CVaR is defined as the conditional expectation of the "losses" of an investment at the $\alpha\%$ worse scenarios (e.g $\alpha = 5\%$). In the context of energy systems, if one wants to maximize the expected energy generation of a portfolio, the α-CVaR could be equationed to represent the expected energy generation at the $\alpha\%$ worse conditions. In this formulation, a mean-CVaR model [13] would attempt to find a portfolio that maximizes the total expected generation and a weighted version of α-CVaR (low generating scenarios).

Despite less frequent than the mean-variance formulation, CVaR models are gradually being incorporated into the literature to assist in the optimization and analysis of energy systems. In Ref. [14], the economic risks associated with PV generation are modeled in terms of a CVaR-LCOE metric, where yearly solar irradiation is characterized using a Weibull distribution for different regions of the Minas Gerais state in Brazil, and a Monte Carlo sampling procedure is used to compute the expected LCOE of solar PV investments given the 0.1% worse energy generation conditions. In Ref. [15], the authors investigated the portfolio optimization of solar and wind energy in Germany using a genetic algorithm to compute the efficient frontier of different portfolios considering the expected monthly return and risk determined by CVaR at the 10% worse scenarios. In Ref. [15], different from Ref. [14], no synthetic data generation is considered, and only historical data from 2015 to 2017 is used.

Another important component of portfolio optimization analysis is data availability. Renewable energy projects have an expected operating life of 20–40 years [16,17]. Ideally, the data used to optimize these portfolios would need to be long enough to capture different scenarios that can happen during the project lifetime. However, this can be challenging when only a few years of historical data are available. The problem becomes more significant if multiple technologies are evaluated simultaneously since all data used needs to be at the same time interval for consistency.

Different works have investigated alternatives to generate synthetic data for renewable energy resources. Hill et al. [18] explored the use of vector autoregressive models (VAR) in the synthetic generation of time

series for 14 wind energy sites in the United Kingdom. Dias et al. [19] proposed an autoregressive moving average model (ARMA) with state-space representation to generate synthetic wind energy data showing that this approach was able to partially reproduce the correlation between six wind turbine locations.

Chen et al. [20] and Qiao et al. [21] used Generative Adversarial Neural Networks (GANs) for the generation of wind energy time series at hourly and sub-hourly time scales considering the correlation of different wind energy sites. In these both studies, a maximum time horizon of 24 h was considered together with up to 24 site locations. It is interesting to mention that [20,21] also evaluated the capability of GANs in generating synthetic data for solar energy generation; however, in this case, the correlation of multiple site locations was not considered.

None of the previous references related to GANs [20,21] considered the generation of synthetic data for multiple site locations and energy resources simultaneously. This has an immediate application to portfolio optimization studies where an algorithm needs to find the optimal site selection for each energy conversion device across a large number of viable locations to minimize energy intermittency and increase the portfolio robustness. In this case, the generative model needs to ensure consistency not only between the distributions of different sites but also between the distributions of different renewable resources, which can be challenging as the number of sites increases.

This work proposes a CVaR stochastic programming formulation to optimize the site selection of offshore wind, wave, and ocean current devices to compose a renewable generation portfolio. This formulation aims to minimize the risk of low energy generation using CVaR, given different site and technology constraints, such that efficient frontiers could be computed in terms of the portfolio mean LCOE and CVaR. Other works, such as [22,23], have also investigated the use of CVaR in risk assessment and portfolio optimization; however, to our knowledge, the application of CVaR in the site selection of multiple energy resources over a large number of site candidates has not been explored before, and it represents an important contribution of this work.

To represent scenarios in the risk-averse stochastic program, a GAN model is developed to generate daily synthetic energy generation profiles capable of representing the interactions between all site locations (+500) and renewable resources analyzed (wind, wave, and ocean current). This GAN formulation takes advantage of the geometric distributions of the energy resources by using Convolutional Neural Networks (CNNs) [24]. Through this approach, the correlation of site locations that are closer to each other can be more easily characterized, allowing the GAN model to generate high-quality synthetic data at more than 500 locations. Finally, the techniques proposed in this work were evaluated using offshore resource data from a portion of the U.S. East coast.

The main contributions of this paper can be summarized as follows.

1) *CVaR Portfolio Modeling:* A modeling framework is proposed to integrate LCOE and CVaR in the construction of efficient frontiers for energy portfolios, considering optimal site selection and site/technology feasibility constraints.
2) *GAN Modeling Approach:* A new modeling representation of the input/output data of GANs is proposed, focused on applying the technique to large geographic regions and multiple energy resources using CNNs.
3) *Scenario Generation Case-Study:* The use of GANs in the generation of synthetic data is investigated considering more than 500 site locations and up to three energy resources (wave, wind, and ocean current), a condition frequently found in practice during site optimization/portfolio studies.
4) *Portfolio Optimization Case-Study:* Portfolio optimization studies are performed, evaluating the importance of sampling size (number of years of data) in the construction of efficient frontiers. The CVaR-optimized portfolios are compared with the more conventional variance-minimization Markowitz modeling [25], and different combinations of offshore wind, wave, and ocean current are investigated under our risk-return framework.

The remainder of this paper is divided as follows: Section 2 describes the models and methods used in this work, Section 3 details the simulation results, and Section 4 concludes the paper.

2. Models

This section details the generative model developed in this work, the performance measures used in the synthetic data evaluation, and the model designed for the portfolio optimization analysis.

2.1. Generative adversarial neural network model

Generative Adversarial Neural Networks (GANs) are a special type of neural networks capable of implicitly modeling high-dimensional distributions of data. This architecture is composed of two distinct models connected with each other, called Generator and Discriminator (Fig. 1). In GANs, the Generator is responsible for producing samples that are statistically similar to the real data, and the Discriminator is responsible for distinguishing between real and synthetic samples [24].

The training of GANs is usually divided into two stages that are iteratively repeated until the model improves its performance. In the first stage, the Generator receives an input noise vector, and after a series of mathematical operations, generates the first batch of synthetic samples. The samples produced by the Generator are labeled as “not real” and used to train the Discriminator together with a set of real samples. In this stage, the Discriminator learns to classify real and synthetic data.

In the second stage, the internal parameters of the Discriminator are frozen, and the synthetic data created by the Generator is deliberately labeled as “real data”. In this stage, the Generator is trained to improve the quality of its data output such that the Discriminator has more difficulty distinguishing between real and synthetic data. In this training scheme, the Generator learns by exploring information embedded in the Discriminator, and the Discriminator learns by comparing the samples created by the Generator with the real data.

A diagram of the GAN model implemented in this work to generate synthetic data of offshore wind, wave, and ocean current is shown in Fig. 1. Initially, historical resource data is processed to have the same grid resolution and to cover the same latitude and longitude (lat/long) regions in a matrix-like format. For the case shown, the area investigated is divided into 25×25 grid cells (each cell corresponding to a lat/long location). The 25×25 matrices of each renewable resource (three) are then concatenated so that a $25 \times 25 \times 3$ data sample can be generated for each day of the historical data. These $25 \times 25 \times 3$ matrices can be interpreted as images as they maintain the relative location of each site and resource (wind, wave & ocean current).

By organizing the data this way, CNNs can help GANs to recover the correlations between different site locations more efficiently. Sites that are closer to each other have a higher chance of being strongly correlated, and these geometric type of structures are known to be well-interpreted by CNNs. See, for example, the work of [26,27], which propose an understanding of CNNs computing process, showing the model’s capacity to isolate localized features, identify lines, edges, and complex shapes.

The GAN model shown in Fig. 1 with its different number of convolutions and deconvolutions is manually tuned by testing different parameter configurations. It was found that a latent space of 100 elements is sufficient to model the synthetic data through the Generator and that the use of Batch Normalization (BN) [28] improved the model training speed significantly (see Supplementary Note 1). BN normalizes the layer’s inputs during training allowing the gradient descent to take longer steps toward the objective function minimum [29], not only accelerating training but also reducing its dependency on the model

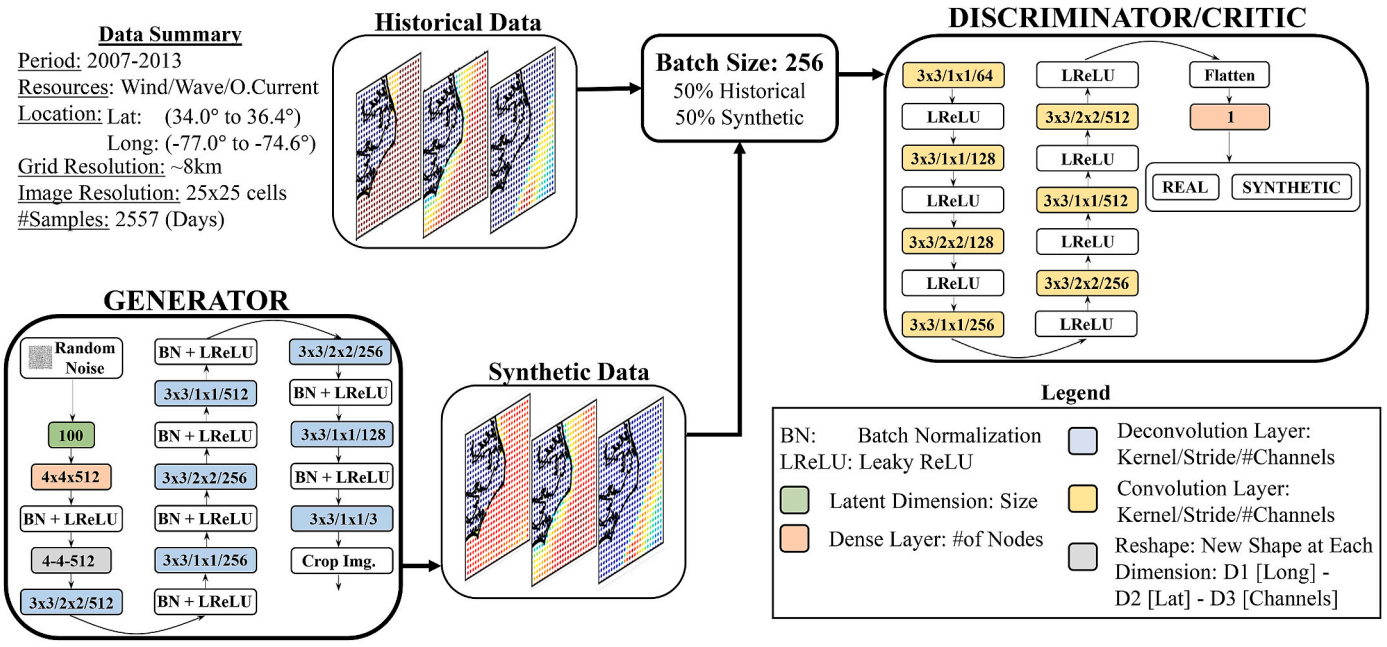


Fig. 1. Generative adversarial neural network model diagram.

initialization.

A Leaky ReLU activation function is used with the slope of the leak equals to 0.2, and the Wasserstein loss function [30] is used together with minibatch gradient descent, and Adam optimization algorithm [31] with a learning rate of 2e-3, momentum (β_1) of 0.5, and minibatch size of 256. The Wasserstein loss is used due to its better performance compared to more traditional error metrics [30].

Different from the binary cross-entropy loss function, the Wasserstein does not require a careful balance in the training and architecture of the Discriminator and Generator. On the other hand, the Wasserstein loss is highly intractable when the Discriminator is not a Lipshitz continuous function. In this work, the Lipshitz continuity of the Discriminator is enforced using a penalty term in the gradient norm as in Ref. [32].

It is important to mention that while using a Wasserstein loss function, the Discriminator does not output a zero (real) or one (synthetic) classification for the images. This function instead allows the Discriminator to assume any real value in order to compute the distance between the distribution of the real and synthetic data.

Still, concerning Fig. 1, a detailed description of the convolution and deconvolutional (inverse convolution) operations can be found in Refs. [33,34]. In this work, the convolution/deconvolutional layer parametrization (yellow/blue blocks) are described in terms of their kernel size, followed by stride and number of channels, as detailed in the figure legend.

The Generator model starts with a latent dimension of 100, which is used as input for a dense layer that creates the foundation for the synthetic image, a 4×4 matrix with 512 channels. This image is gradually extended to 8×8, 16×16, and 32×32 using deconvolutions. Lastly, the image is converted to 32×32 pixels with three channels and cropped to assume the format of the historical data $25 \times 25 \times 3$. The Discriminator follows the opposite path of the Generator; the $25 \times 25 \times 3$ image is gradually reduced to 13×13, 7×7, and 4×4, but at each of these steps the number of channels increases. In the end, the model uses a dense layer to output the evaluation of the Discriminator as a single number.

The framework described above is based on the ideas of well-known GANs architectures such as [35,36]. Finally, although the model shown in Fig. 1 is adjusted for a $25 \times 25 \times 3$ grid mapping (25 latitude pixels, 25 longitude pixels, and three energy resources), it can be modified to accommodate different grid resolutions and/or number of technologies by increasing/decreasing the number of convolution/deconvolutions.

The network presented in this section was implemented in Python using TensorFlow 2.6 [37], and its code is publicly available on [38], where more details regarding the model structure and parametrization can be obtained.

2.2. Performance measure of Generative Adversarial Neural Networks

The problem of evaluating generative models is an open research topic [39,40] that has gained considerable attention due to the impressive results shown by GANs. Comparing two sampling distributions in higher dimensional spaces is a difficult task, and classic approaches such as estimating the model log-likelihood are frequently impractical [41].

In the literature of GANs, the Maximum Mean Discrepancy (MMD) [39,42], Inception Score (IS) [40], and Frechet Inception Distance (FID) [43] are three of the most used metrics to compare the performance of different generative models. As previously discussed by many works, each of those three metrics has its pros and cons. References [42,44] show that the FID and IS are strongly biased with the sample size, only performing properly with very large samples. This bias problem is not evident in the MMD metric [42], which performs accordingly even for small data sets. The FID is considered to correlate very well with human perception and takes into consideration the statistics of the real and synthetic samples; the same is not true for the IS [43]. Finally, the FID can only capture up to two moments of distributions (from the feature space) due to Gaussian approximations made during the metric calculation, and while the FID and IS require a pre-ANN training (capable of identifying different classes of the real data-see Refs. [40,43]), the MMD requires the definition (and tunning) of a proper kernel function.

Considering the problem of training an additional ANN classification model just for computing the GAN performance (reducing the effective training set of the GAN), together with the limitations observed in Refs. [42,44] regarding bias in the IS and FID metrics, the authors decided to only use the MMD metric in this work. However, other statistical verification procedures, such as comparing the synthetic sampling distribution with the historical data at individual sites, are also provided in the supplementary file (Note 2) to improve the assessment of the GAN performance.

Finally, the squared MMD can be computed as (1) [42], where $k(\cdot)$, represents the kernel considered. The idea behind this metric is to compute the difference between two sample distributions by using a special class of functions capable of projecting the data in a higher-dimensional space. Taking the average distance between samples in this new space makes it possible to distinguish non-trivial differences in the distributions, such as those associated with the correlations between different sets of variables [45]. While computing the MMD, the Radial Basis Function (RBF) (2), rational quadratic, and polynomial kernels are frequently explored in the literature [42,45]. This work follows the approach of [45] and uses the RBF kernel (2) in the MMD computations.

$$MMD^2(X_h, X_s) = \frac{1}{N_h(N_h-1)}\sum_{i\neq j}^{N_h} k\left(x_i^h, x_j^h\right) + \frac{1}{N_s(N_s-1)}\sum_{i\neq j}^{N_s} k\left(x_i^s, x_j^s\right) - \frac{2}{N_h N_s}\sum_{i=1}^{N_h}\sum_{j=1}^{N_s} k\left(x_i^h, x_j^s\right) \quad (1)$$

$$k(x,y) = \exp\left(-\frac{1}{2\sigma^2}\|x-y\|^2\right) \quad (2)$$

2.3. Portfolio optimization model

The risk-averse stochastic optimization model developed in this work to perform the portfolio selection of offshore wind wave and ocean current technologies is presented in (3-10). The model objective is to find the optimal number of turbines per site location such that the expected energy generation of the portfolio in the worse α% scenarios is maximized.

Each energy scenario is equally likely in this work since the GAN model randomly generates them one by one. In this context, if a total of 10,000 scenarios are evaluated under an α equal to 5%, $CVaR_\alpha$ would be the expected energy generation during the 500 worse conditions. Fig. 2 illustrates the idea of *CVaR* showing in red the 500 worse conditions in a 10,000 sampling simulation.

In the formulation (3−10), constraint (4) enforces an upper bound for the LCOE of the portfolio, constraint (5) defines the total number of turbines per energy technology, constraint (6) limits the number of turbines per site location, and constraints (7–8) enforces a maximum radius for the energy collection system; (7–8) guarantee that turbines of the same technology will not be deployed very far from each other, which could lead to prohibitive configurations for the energy collection system, with high energy losses and cost.

Lastly, constraint (9−10) states that the number of turbines at each site location for each technology ($x_{e,i}$) is a non-negative integer, and the variable responsible for defining the energy collection system location ($v_{e,i}$) is binary.

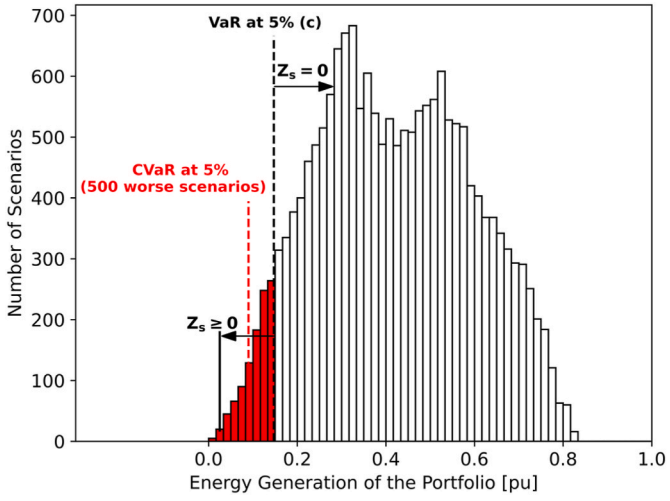


Fig. 2. Example of The Calculus of CVaR: For a Total of 10,000 Energy Scenarios, CVaR is the Expected Generation of The Portfolio Conditioned on the 5% Worse Scenarios (in red). Refer to Equations 11–13 for the interpretation of "c" (VaR) and "z_s".

$$\max_{(x,v)} \; CVaR_\alpha(X,\zeta) \quad (3)$$

$$\text{s.t.} \; \frac{\sum_{e\in E}\sum_{i\in I_e} C_{e,i}\, x_{e,i}}{\sum_{e\in E}\sum_{i\in I_e}\mathbb{E}\left[EG_{e,i,\zeta}\right]x_{e,i}} \leq \overline{LCOE} \quad (4)$$

$$\sum_{i\in I_e} x_{e,i} = TNt_e, \quad \forall \; e \in E \quad (5)$$

$$x_{e,i} \leq \overline{Nt_{e,i}}, \quad \forall \; (e,i) \in (E, I_e) \quad (6)$$

$$\sum_{k\in D_{e,i}^R} x_{e,k} \geq v_{e,i}\, TNt_e, \quad \forall \; (e,i) \in (E, I_e) \quad (7)$$

$$\sum_{i\in I_e} v_{e,i} = 1, \quad \forall \; e \in E \quad (8)$$

$$x_{e,i} \in \mathrm{Z}^+, \quad \forall \; (e,i) \in (E, I_e) \quad (9)$$

$$v_{e,i} \in \{0,1\}, \quad \forall \; (e,i) \in (E, I_e) \quad (10)$$

To solve the model (3−10) using and optimization solver, the *CVaR* metric needs to be represented arithmetically. Rockafellar [12] showed that by incorporating two auxiliary variables (z_s, and c) the objective function (3) can be written in terms of (11-13). In this formulation, c is known as Value at Risk (*VaR*), and it represents the energy generation at the 5% quantile for the example shown in Fig. 2 ($\alpha = 5\%$).

In terms of z_s, if the energy generated by the portfolio ($x_{e,i}, v_{e,i}$) in a given scenario (ζ_s) is larger than the *VaR* (c), z_s is zero; otherwise, it is positive and equal to the difference between the *VaR* (c) and the energy generated in the scenario. The combined influence of the objective function (11) and the constraints (12–13) leads to the $CVaR_\alpha$ value of the portfolio.

Formulated as (11–13 and 4–10), the stochastic portfolio optimization model presented here can be solved as a large-scale Mixed-Integer Linear Programming Problem (MILP), where the influence of each energy scenario ($\zeta_s \in \zeta$) is individually incorporated in the formulation through the constraint (12).

$$\max_{(x,v,c,z)} \; c - \frac{1}{\alpha\, N_s}\sum_{s=1}^{N_s} z_s \quad (11)$$

$$z_s \geq c - \sum_{e\in E}\sum_{i\in I_e} EG_{e,i,\zeta_s} x_{e,i}, \quad \forall \; s \in \{1,\ldots,N_s\} \quad (12)$$

$$c \geq 0 \; and \; z_s \geq 0 \quad (13)$$

By running the model (4−13) for different LCOE limits ($\overline{LCOE}$) it is possible to obtain an efficient frontier for the problem investigated. This curve characterizes the set of optimal portfolios that have the lowest average cost per MWh (LCOE) at a given risk level, here defined by $CVaR_\alpha(X,\zeta)$. Any portfolio outside the efficient frontier can be considered sub-optimal.

To compare the *CVaR* model (3−10) with the more traditional variance minimization framework [9,46], the objective function (3) can be substituted by (14), with no changes in constraints (4−10).

$$\min_{(x,v)} \; X^T \Sigma_s X \quad (14)$$

By changing (3) to (14), the model now aims to minimize the variance in the energy generation of the portfolio X, where X is represented

as a column vector of the decisions $x_{e,i}$, and Σ_s is the variance-covariance matrix for the synthetic data.

The models described in this section are implemented in Python using the Pyomo modeling language [47] and Gurobi [48] as optimization solver. The code is publicly available on [38].

3. Simulations

The simulations performed in this section are based on data from a portion of the U.S. East coast in North Carolina. North Carolina has recently received significant incentives for the deployment of renewable energy resources by establishing a target of 70% reduction in CO_2 emissions by 2030 and carbon neutrality by 2050 [49]. Furthermore, the state committed to a plan of deploying 2.8 GW of offshore wind energy by 2030 and 8 GW by 2040 [50]. These aspects mentioned above and the availability of three major offshore energy resources (wind, wave, and ocean current) made North Carolina the ideal region for investigating the models proposed in this work.

Fig. 3 shows a flow diagram summarizing the simulations and analysis performed in this section. First, historical data of wind speed, wave height/period, and ocean current speed [51–53] was converted to electrical energy considering a set of optimally selected turbines [9]. This data is then used in the training of a GAN model capable of implicitly estimating the probability distribution of the historical data. Next, the GAN is used to generate synthetic samples of energy generation, and this data is compared to the historical data for statistical consistency. Finally, the synthetic data is used in a portfolio optimization model to estimate the efficient frontiers of different renewable portfolios.

For the simulations performed in this work, a 16-core 5 GHz CPU with 64 GB of RAM and an RTX3080-10 GB GPU was used.

3.1. Data

In this work, offshore wind speed data comes from the NREL Wind Integration National Dataset (WIND) Toolkit [51], wave significant height and period comes from the WAVEWATCH III model [52], and ocean current speed comes from the HYCOM/NCODA model [53].

The conversion of raw energy resources to electrical energy generation is done using the energy conversion devices described in Ref. [9], which were carefully selected for optimal deployment on the North Carolina coast. The wind turbine has a rated capacity of 6 MW, the wave 1.5 MW, and the ocean current 4 MW. Furthermore, the annualized cost estimates, minimum/maximum deployment depths and other design/cost-related characteristics were also obtained from Ref. [9].

When considering the three energy resources simultaneously, a maximum seven-year overlapping between the different data sources is obtained for the period of 2007–2013 in daily time discretization. Consequently, the historical data used for the GAN training follow this seven-year overlapping. Finally, the daily energy generation data is upscaled for the resolution of $0.1^\circ \times 0.1^\circ$ and limited to a latitude of 34.0°–36.4° and longitude of -77.0° to -74.6°.

Fig. 4 shows the average capacity factor from 2007 to 2013 for the wind, wave, and ocean current technology. Any site location not represented in this figure is set equal to zero in the GAN training. After the scenario generation by the GAN model, only the site locations that satisfied the minimum/maximum deployment depths [9] are integrated into the portfolio optimization.

3.2. Generative neural network model

This section presents the results of the GAN model. Fig. 5 shows a box plot of the difference between the average energy generation of the synthetic and real data at each site location for the wind, wave, and ocean current technology. From this figure, it is possible to notice that the generative model produced scenarios that differ on average no more than 0.02 [p.u] from the real data.

Fig. 6 shows the variance-covariance matrix of the real (Fig. 6a) and synthetic data (Fig. 6b). In this figure, the x-axis and the y-axis are divided into three regions representing the wind, wave, and ocean current resources, such that it is possible to identify the interactions between any pair of technologies. Here, the results show a strong similarity between the variances and covariances of both datasets.

Finally, Fig. 7 shows a boxplot of 1000 samples of the MMDs computed using one year worth of data (365 data points randomly sampled). In this figure, the MMD_RR group compares the difference between two sampled distributions of the real data. The MMD_SS group compares two sampled distributions of the synthetic GAN data. The MMD_RS group compares the real data with the synthetic GAN data, and the MMD_RN group compares the real data distribution with samples

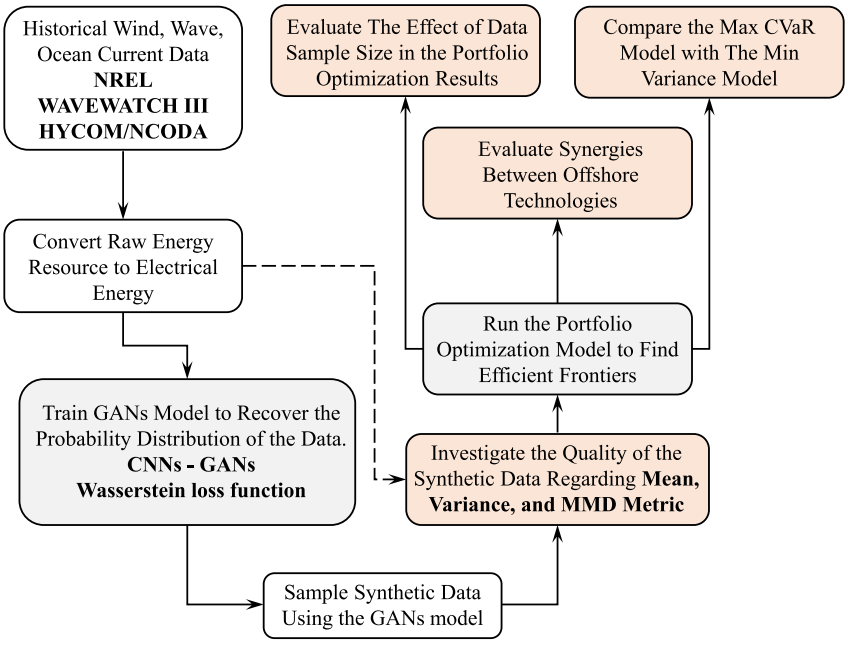


Fig. 3. Project flow diagram.

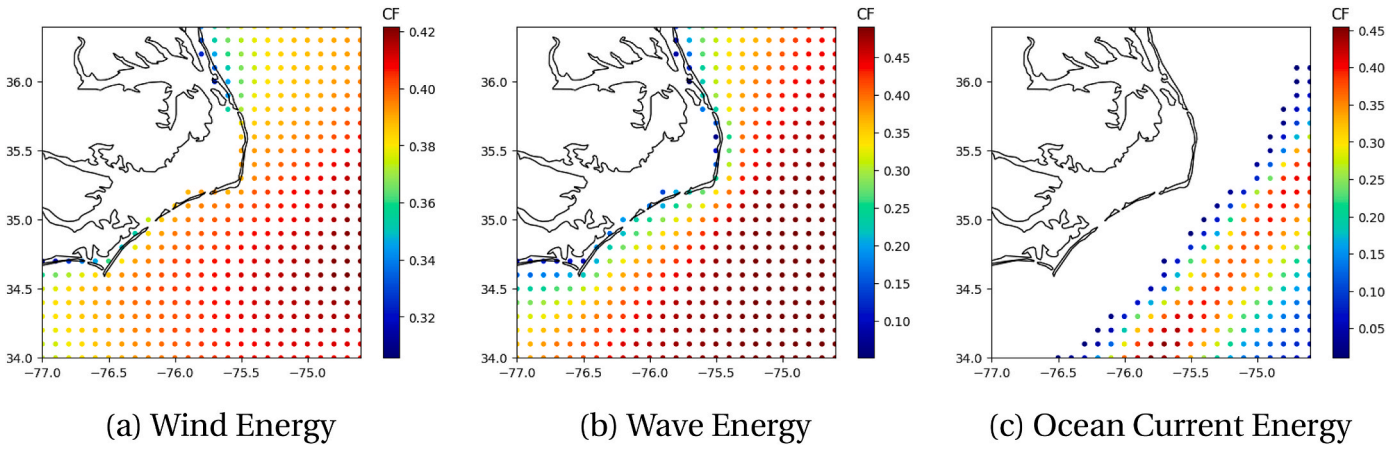


Fig. 4. Capacity factor for the wind, wave, and ocean current technology.

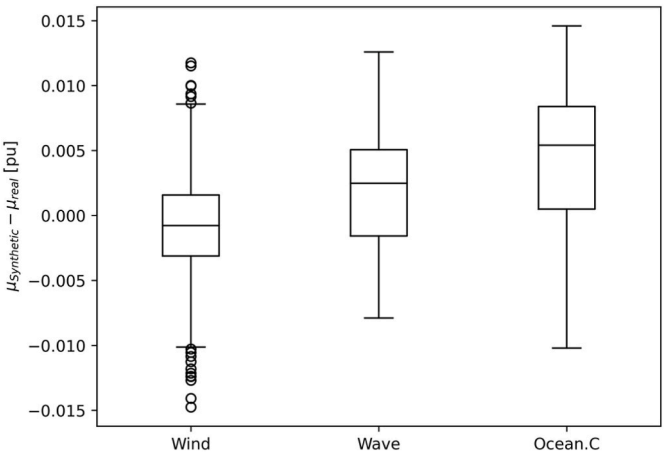


Fig. 5. Difference in the average energy generation of the synthetic and real data at each site location ($\mu_s - \mu_r$).

from a multivariate Gaussian model truncated from 0 to 1pu and fit with the real data itself.

The MMD metric has a non-intuitive meaning when used by itself (a single MMD Eq. (1)). On the other hand, this measure is very useful to compare the performance of two models. By looking at the MMD_RR group, it is possible to notice that random samples of the real data can differ themselves by up to 0.04 units of distance, but they differ on average only about 0.015 units of distance.

From Fig. 7, it is desired that both synthetic and real data behave similarly, which would translate to the MMD_SS and MMD_RS groups assuming values in the same range as the MMD_RR group. However, small differences not far from the scale seen in the real data (MMD_RR) are expected as the GAN model is capable of increasing the diversity of scenarios by sampling from its probability approximation of the real data.

From this figure, it is possible to see a great agreement between the data generated by the GAN model and the real data. However, this is not true for the samples generated using the multivariate normal distribution, as the MMD_RN group differs significantly from the MMD_RR group.

Since this work investigates the generation of synthetic data for multiple sites and technologies simultaneously, traditional statistical assessment tools such as comparing directly the probability distribution of both datasets becomes complex as the number of site/technology combinations increases fast with the number of sites, which is already large. However, in Supplementary Note 2.1, a small number of site locations is sampled, and the probability density function (PDF) estimated with the real data is plotted together with the PDF estimated with the GAN synthetic data, showing an excellent agreement between both PDFs.

Discussions about how the probability distribution of the real and synthetic data compare in terms of pairs of sites and/or technologies can be found in Supplementary Note 2.2. For example, how the energy generation of one site location changes given the generation in another site, and how this change differs with the real and synthetic data. Overall, for all site and technology pairs investigated (28 combinations), the GAN model continued to agree with the real data, well capturing non-linear interactions. PDFs for a series of portfolios (combinations of technologies and site locations) considering the real and synthetic datasets are shown in Supplementary Note 2.3. In these plots, it is possible to notice that the PDFs of the synthetic data are consistent with what is expected from a large sample of a probability distribution having smoother transitions in the PDFs when compared with the PDFs of the real data, which were estimates using a smaller number of samples (years 2007–2013).

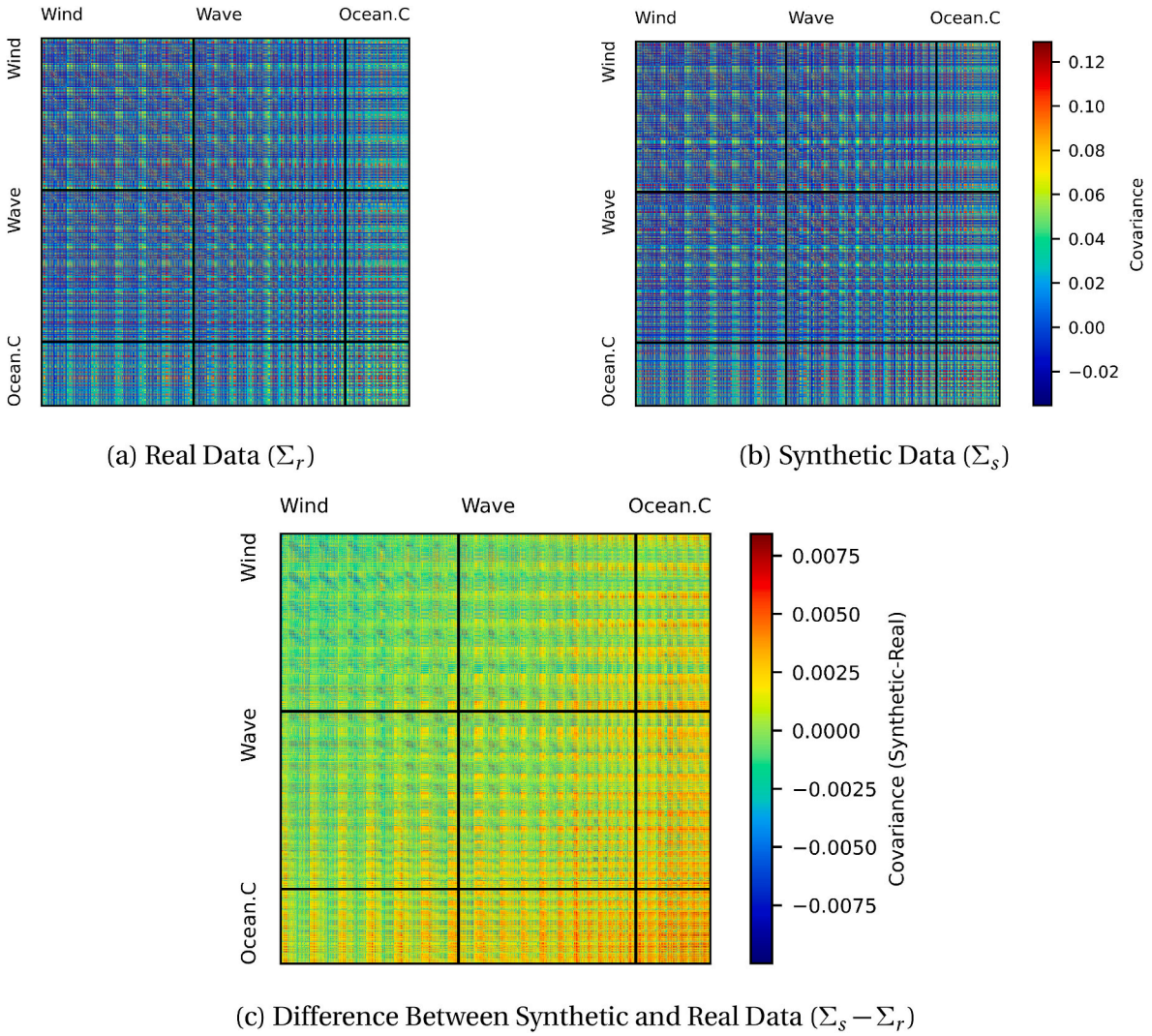
Overall, the results presented in this section and Supplementary Notes show that GANs are a powerful tool capable of adequately inferring the probability distribution of complex energy generation profiles over many technologies and large geographic regions, maintaining the statistical properties of its training data and providing diversity in its generated scenarios.

3.3. Portfolio optimization

In this section, the portfolio optimization models described in section 2.3 is used together with the GAN model to create efficient frontiers for different offshore renewable energy systems. These simulations consider the assumptions made in Ref. [9] regarding technology efficiency, rated capacity, cost, site feasibility, and maximum radius of the energy collection system ($D^R_{e,i}$). Furthermore, a maximum packing density of 50 turbines per site location per technology is enforced.

To understand the influence of the number of scenarios (years of data) in determining the efficient frontiers of the offshore energy portfolios, simulations with 10 and 30 years of energy generation are made considering the site location optimization of 200 wind, 200 wave and 200 ocean current turbines, equivalent to 2.3 GW of installed capacity.

Fig. 8a shows the simulations made using 10 years of data, and Fig. 8b shows the simulations with 30 years of data. In these figures, the x-axis represents the average energy generation in the 5% worse scenarios ($CVaR_{5\%}$), and the y-axis represents the average cost per MWh of the portfolio (LCOE). The curve in red represents the efficient frontier of the portfolio computed using the model (3−10), and each point on this curve represents a different solution, with turbines located at different



(a) Real Data (Σ_r)

(b) Synthetic Data (Σ_s)

(c) Difference Between Synthetic and Real Data ($\Sigma_s - \Sigma_r$)

Fig. 6. Covariance matrix of the synthetic and real data.

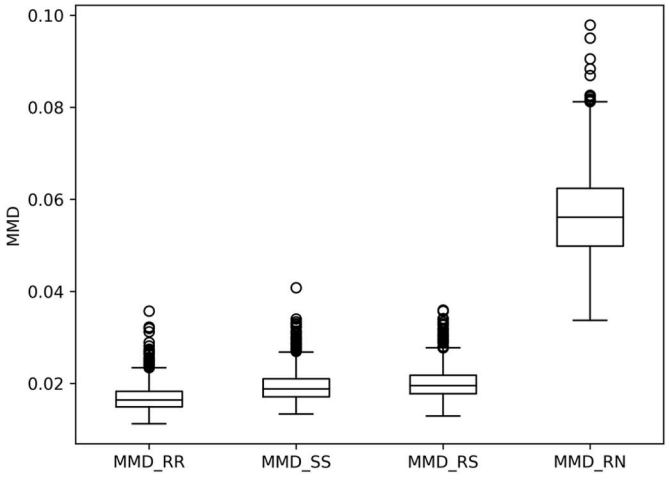


Fig. 7. MMD Values for one Year Worth of Data, When Comparing the Real Data With Itself (MMD_RR) the Synthetic GAN Data With Itself (MMD_SS), The Real With the Synthetic GAN Data (MMD_RS), and The Real With Multivariate Normal Sampled Data (MMD_RN).

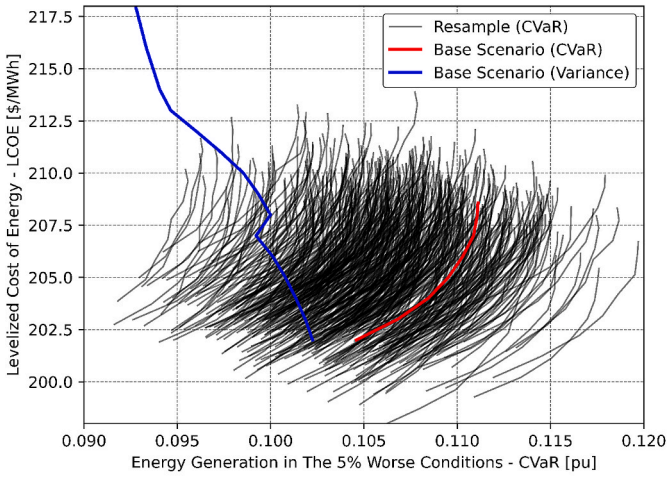
locations.

A sensitivity analysis is performed in the efficient frontier (red curve) by computing the CVaR and LCOE of each solution in the curve using a new sample generation data of the same size (number of years) as used in the optimization model (10 or 30 years). This resampling is done 500 times, and the results are shown as black lines in Fig. 8.

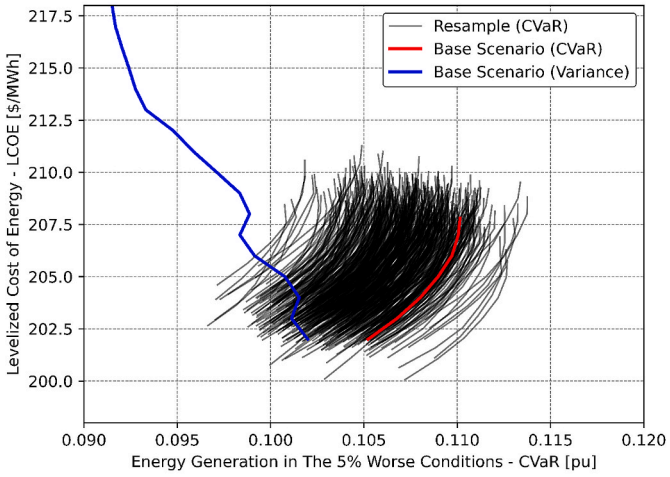
As can be seen in Fig. 8, the increase in the sample size from 10 to 30 years significantly reduced the uncertainty in the estimation of the efficient frontier. As most offshore energy projects have a lifetime of 20–30 years, the result of Fig. 8 clearly shows the importance of considering larger sample sizes to properly quantify the risk and return of these projects.

Fig. 8 also shows in blue the portfolio optimization results using the variance minimization model (14). It is interesting to notice that the efficient frontier constructed by the minimization of variance significantly underperforms in terms of energy generation in the 5% worse conditions. The model trades variance minimization for more severe low-energy generation scenarios. This result shows the importance of CVaR as a risk measure capable of modeling asymmetric risks in portfolio optimization problems.

Finally, Fig. 9 shows the efficient frontiers for different combinations of wind, wave, and ocean current considering seven years of historical (2007–2013) and synthetic data using the min-CVaR model. In this figure, an arrow indicates the installed capacity of each portfolio in terms of GW (Wind/Wave/O.Current). The simulations made with



(a) 10 Year Sample



(b) 30 Year Sample

Fig. 8. Efficient frontier given the deployment of 200 wind (1.2 GW), 200 wave (0.3 GW), and 200 ocean current turbines (0.8 GW), considering uncertainty quantification and the use of different number of samples in the portfolio optimization (10 and 30 Years).

historical data are represented as green curves, and the simulations made with synthetic data are represented as red curves. A sensitivity analysis was also performed on the red curves by computing the CVaR and LCOE of each solution point using a new seven-year sample from the GAN model. These results are shown as black lines and indicate the uncertainty in the synthetic data estimate.

From this figure, it is possible to notice that the worse portfolios in terms of CVaR are those with only wind and only ocean current. Despite having the lowest LCOEs the portfolios with only wind energy have CVaRs smaller than 0.02 [pu]. It is also possible to notice that the combination of different resources significantly improves the CVaR of the equivalent portfolio, as the deployments with wind, wave, and ocean current had the smallest CVaRs. This shows the importance of integrating different resources in order to improve system security and availability. Despite the high LCOEs of wave and ocean current compared to offshore wind energy, the continuous development of marine energy technologies is likely to lead to substantial cost reductions, as has been seen in other renewable energies such as solar and wind. In this condition, the complementarity between these different resources may become a cost-effective alternative to minimize the impact of energy variability and intermittency of renewables.

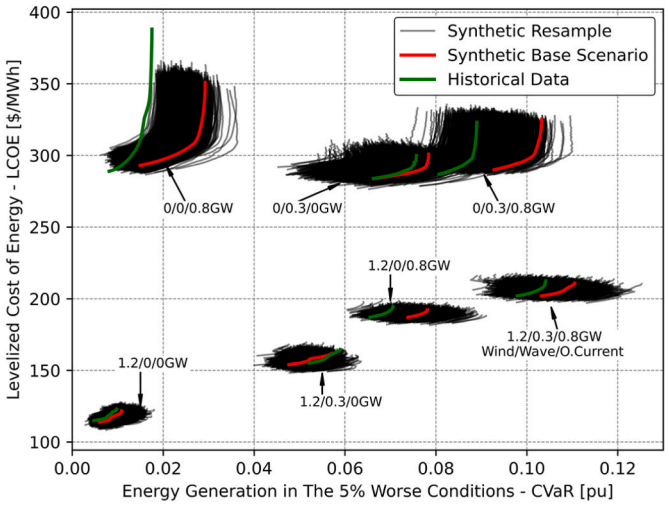


Fig. 9. Efficient Frontier for Different Combinations of Wind, Wave and Ocean Current Considering Seven Years of Historical and Synthetic Data. Only The min-CVaR Model (3−10) Was Used in These Simulations.

Fig. 9 also shows that portfolio optimizations made using the synthetic have a good agreement with the simulations made using historical data.

4. Conclusion

This work proposes the use of GAN models to generate synthetic data for multiple energy resources over large geographic regions, maintaining the statistical properties of each resource and site location such that the technique can be used in portfolio optimization studies. Furthermore, this work proposes a stochastic formulation for the site selection of renewable energy technologies considering the maximization of the average α%-worse energy generation conditions (CVaR) at pre-defined targets for the portfolio LCOE. The combination of both GAN and CVaR minimization models is proposed as a tool for the risk assessment of renewable energy portfolios.

The GAN model was tested considering the generation of synthetic data for wind, wave, and ocean current resources in a 25×25, 0.1° grid off the coast of North Carolina. Showing great agreement with the historical data, maintaining the statistical properties of each site and resource, as well as adequately capturing complex interactions between different locations/resources. The CVaR minimization model was used to generate the efficient frontier for different combinations of offshore portfolios considering the scenarios created by the GAN model. A batch of simulations was made using a different number of sampled years (10 and 30), showing the importance of sample size while performing portfolio optimization studies, as results may vary significantly depending on the size of the dataset. Our results also show that the traditional variance minimization modeling may lead to more severe low energy generation conditions, trading variance improvements for deterioration in the CVaR metric. Lastly, our results show that significant benefits in energy generation can be achieved by exploring the complementarity between different resources, as the portfolios with wind, wave and ocean current outperformed the less diversified portfolios.

Future works should attempt to apply the GAN technique for scenario generation of other renewable energy technologies and other locations. Another potential area of research is to improve the capacity of the GANs to deal with high-resolution information (e.g., 2×2km cells and larger regions), and to incorporate time dependency directly on the scenario generation of the GANs, such that more complex site selection and portfolio optimization studies could be performed and benefit from using the generated synthetic data.

In terms of assessing the performance of GAN models, it is important that metrics like FID, IS, and MMD continue to be investigated; for example, testing different kernel functions for the MMD metric can provide further insights into the generated samples. It is also recommended the comparison of the GAN synthetic data with more traditional models, even if this would require the use of datasets with fewer variables.

Finally, future research should also focus on improving the stochastic model formulation and techniques to speed up optimization, particularly when dealing with high-resolution data which increases model complexity.

Credit author statement

Victor. A.D. de Faria: Data curation, Methodology, Writing – original draft preparation, Software. **Anderson R. de Queiroz**: Methodology, Writing- Reviewing and Editing. **Joseph F. DeCarolis**: Methodology, Writing- Reviewing and Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors would like to thank the North Carolina Renewable Ocean Energy Program and the OR department at N.C. State University for their financial support of this work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.energy.2023.126946.

References

[1] IRENA - International Renewable Energy Agency. Renewable energy statistics 2020. IRENA; 2020.
[2] EIA - Energy Information Administration. International energy outlook 2019. EIA; 2019.
[3] IEA - International Energy Agency. World energy investment 2021. IEA; 2021.
[4] Cole W, Frazier AW. Cost projections for utility-scale battery storage. NREL; 2019.
[5] Faria VAD, Queiroz AR, Lima LM, Lima JWM, Silva BC. An assessment of multilayer perceptron networks for streamflow forecasting in large-scale interconnected hydros systems. Int J Environ Sci Technol 2022;19:5819–38.
[6] Soukissian TH, Karathanasi FE, Zaragkas DK. Exploiting offshore wind and solar resources in the Mediterranean using ERA5 reanalysis data. Energy Convers Manag 2021;237.
[7] Silva AR, Pimenta FM, Assireu AT, Spyrides MHC. Complementarity of Brazil's hydro and offshore wind power. Renew Sustain Energy Rev 2016;56:413–27.
[8] Stoutenburg ED, Jenkins N, Jacobson MZ. Power output variations of co-located offshore wind turbines and wave energy converters in California. Renew Energy 2010;35:2781–91.
[9] Faria VAD, Queiroz AR, DeCarolis J. Optimizing offshore renewable portfolios under resource variability. Appl Energy 2022;326.
[10] Santos-Alamillos F, Thomaidis N, Usaola-García J, Ruiz-Arias J, Pozo-Vázquez D. Exploring the mean-variance portfolio optimization approach for planning wind repowering actions in Spain. Renew Energy 2017;106:335–42.
[11] Shakouri M, Lee HW, Kim Y-W. A probabilistic portfolio-based model for financial valuation of community solar. Appl Energy 2017;191:709–26.
[12] Rockafellar RT, Uryasev S. Optimization of conditional value-at-risk. J. Risk; 2000.
[13] Salahi M, Mehrdoust F, Piri F. CVaR robust mean-CVaR portfolio optimization. ISRN Applied Mathematics; 2013.
[14] Aquila G, Coelho EdOP, Bonatto BD, Pamplona EdO, Nakamura WT. Perspective of uncertainty and risk from the CVaR-LCOE approach: an analysis of the case of PV microgeneration in Minas Gerais, Brazil. Energy 2021;226.
[15] Sinsel SR, Sinsel SR, Stephan A. Building resilient renewable power generation portfolios: the impact of diversification on investors' risk and return. Appl Energy 2019;254.
[16] Neary VS, Previsic M, Jepsen RA, Lawson MJ, Yu Y-H, Copping AE, Fontaine AA, Hallett KC, Murray DK. Methodology for design and economic analysis of marine energy conversion (MEC) technologies. Sandia National Laboratories; 2014.
[17] NREL-National Renewable Energy Laboratory. Energy analysis- useful life. NREL, [Online]. Available: https://www.nrel.gov/analysis/tech-footprint.html. [Accessed 1 November 2022]. Accessed.
[18] Hill DC, McMillan D, Bell KRW, Infield D. Application of auto-regressive models to U.K. Wind speed data for power system impact studies. IEEE Trans Sustain Energy 2012;3(1):134–41.
[19] Díaz G, Gómez-Aleixandre J, Coto J. Wind power scenario generation through state-space specifications for uncertainty analysis of wind power plants. Appl Energy 2016;162:21–30.
[20] Chen Y, Wang Y, Kirschen D, Zhang B. Model-free renewable scenario generation using generative adversarial networks. IEEE Trans Power Syst 2018;33(3):3265–75.
[21] Qiao J, Pu T, Wang X. Renewable scenario generation using controllable generative adversarial networks with transparent latent space. CSEE JOURNAL OF POWER AND ENERGY SYSTEMS 2021;7(1):66–77.
[22] Camargo LAS, Leonel LD, Rosa PS, Ramos DS. Optimal portfolio selection of wind power plants using a stochastic risk-averse optimization model, considering the wind complementarity of the sites and a budget constraint. Energy Power Eng 2020;12.
[23] Camal S, Teng F, Michiorri A, Kariniotakis G, Badesa L. Scenario generation of aggregated Wind, Photovoltaics and small Hydro production for power systems applications. Appl Energy 2019;242.
[24] Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge: MIT Press; 2016.
[25] Markowitz H. Portfolio selection. J Finance 1952;7.
[26] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs and H. Lipson, "Understanding neural networks through deep visualization".
[27] Qin Z, Yu F, Liu C, Chen X. How convolutional neural networks see the world — A survey of convolutional neural network visualization methods, vol. 1; 2018. https://arxiv.org/pdf/1804.11191.pdf.
[28] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. 2015. *arXiv:1502.03167*.
[29] Santurkar S, Tsipras D, Ilyas A, Madry A. How does batch normalization help optimization?. 2019. *arXiv:1805.11604*.
[30] Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. 2017. *arXiv:1701.07875v3*.
[31] Kingma DP, Ba JL. ADAM: a method for stochastic optimization. ICLR; 2015. 2015.
[32] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A. Improved training of Wasserstein GANs. 2017. *arXiv:1704.00028v3*.
[33] Goodfellow I, Bengio Y, Courville A. Convolutional networks. In: Deep learning. MIT Press; 2016. p. 322–61.
[34] Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation. 2015. *arXiv:1505.04366*.
[35] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. 2016. *arXiv:1511.06434*.
[36] Karras T, Aila T, Laine S, Lehtinen J. Progressive growing of gans for improved quality, stability, and variation. ICLR; 2018. 2018.
[37] TensorFlow. TensorFlow [Online]. Available: https://www.tensorflow.org/. Accessed November 2022.
[38] Faria VAD. GANs and CVaR portfolio implementation [Online]. Available: https://github.com/vaduraes/GANs_CVaR_Portfolio_Optimization. Accessed November 2022.
[39] Borji A. Pros and cons of GAN evaluation measures. 2018. *arXiv:1802.03446*.
[40] Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training GANs. In: 30th conference on neural information processing systems; 2016.
[41] Lucic M, Kurach K, Michalski M, Bousquet O, Gelly S. Are GANs created equal? A large-scale study. NeurIPS; 2018. 2018.
[42] Binkowski M, Sutherland DJ, Arbel M, Gretton A. DEMYSTIFYING MMD GANs. ICLR; 2018.
[43] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two time-scale update rule converge to a local nash equilibrium. 2018. *arXiv:1706.08500*.
[44] Chong MJ, Forsyth D. Effectively unbiased fid and inception score and where to find them. In: IEEE/CVF conference on computer vision and pattern recognition; 2020.
[45] Gretton A, Borgwardt KM, Rasch MJ, Scholkopf B, Smola A. A kernel two-sample test. J Mach Learn Res 2012;13.
[46] Li B, Queiroz AR, DeCarolis JF, He JB, Keeler AG, Neary VS. The economics of electricity generation from Gulf Stream currents. Energy 2017;134:649–58.
[47] Hart WE, Watson J-P, Woodruff DL. Pyomo: modeling and solving mathematical programs in Python. Mathematical Programming Computation 2011;3.
[48] Gurobi. Gurobi optimization [Online]. Available: https://www.gurobi.com/. Accessed November 2022.
[49] GENERAL ASSEMBLY OF NORTH CAROLINA. House bill 951. 2021.
[50] State of North Carolina. Executive order No. 218. 2021.

[51] NREL-National Renewable Energy Laboratory. Wind integration national dataset Toolkit [Online]. Available: https://www.nrel.gov/grid/wind-toolkit.html. [Accessed 1 November 2022]. Accessed.
[52] NOAA. NWW3 product viewer [Online]. Available: https://polar.ncep.noaa.gov/waves/viewer.shtml?-(none)-. [Accessed 1 November 2022]. Accessed.
[53] HYCOM. Gofs 3.1: 41-layer HYCOM + NCODA global 1/12° reanalysis [Online]. Available: https://www.hycom.org/dataserver/gofs-3pt1/reanalysis. [Accessed 1 November 2022]. Accessed.