Increase Renewable Data Temporal Resolution Using Conditional GANs

Final Report - Zhiyuan Fan - zf2198

December 15, 2022

1 Introduction

As the world is moving towards the net-zero target, renewable electricity is playing an increasingly important role in the energy system, likely a dominating role in primary energy supply in the long-term future [IPCC, 2022]. However, a deeper penetration of renewable electricity is challenging for power supply from many different perspective: system/market scheduling, operation, and planning [Colon et al., 2019]. The fundamental challenge is the accurate modeling of renewable energy supplies, which have strong stochastic nature [Younesi et al, 2022]. The most common approach of addressing the stochastic challenge is through learning of large amount and wide ranges of scenarios: wind/solar/hydro renewable data from different geographies of different periods, i.e., the historical data. The data type of interest is time-series data, which is 1 dimensional with time period index T.

Despite the continuous effort of data collection, scenario generation still faces the great challenge of lacking data:

- Historical data is only available at certain locations for certain years, which is typically sparse spatially and uneven temporally [Aryanpur et al., 2021].
- Historical data is limited in exploring the future scenarios, especially under the climate change background that renewable patterns is expected to change [Amonkar et al., 2022].
- Probabilistic distribution model-based methods for data generation cannot preserve the temporal dynamics of time-series data, physical hard constraints, and spatial correlation [Chen et al., 2018].

All these demand a better way of understanding, replicating, and eventually manipulating the renewable scenarios according to the research purpose. General circulation models (GCMs) provides another pathway of scenario generation. GCMs are fundamentally based on fluid dynamics, or more specifically partial differential equations derived from the Navier-Stokes equation sets, which is widely used for forecasting climate change scenarios [Balaji et al., 2022]. Apart from the global perspective, GCMs contain rich data for local geographies for a long time horizons, such as wind/solar/hydro time series data for every location of planet surface for decades. Unfortunately, the computational expense for GCMs is substantial, meaning it's not capable of high resolution data generation (both spatially and temporally) while maintaining the global perspective accuracy [Balaji et al., 2022]. From temporal analysis, GCMs temporal solution typically does not go beyond daily [ISIMIP, 2022]. Similarly, some approaches were proven to work perfectly in long time horizon, high-dimensional data generation while preserving the spatial correlations and temporal dynamics, but also failed at higher resolution than daily [Amonkar et al., 2022]

Some recent progress of machine learning has shown great potential in generating synthetic data based on historical data using Generative Adversarial Networks (GANs) [Chen et al., 2018][Zhao et al., 2022]. From the [Chen et al., 2018], GANs was shown to have some desired features that may help with scenario generation:

- GANs are capable of generating new and distinct scenarios not perfectly match the historical data (training data)
- GANs can be conditional that allows further manipulation for scenario generation with certain prerequisite requirements, such as spatial correlation or temporal period
- GANs can generate high temporal resolution data, e.g., 5-mins temporal resolution data which is suitable for system planning

The contribution of this projects aims to bridge the challenge of data shortage and temporal resolution through GANs. The project investigated the feasibility and performance of using high resolution time-series as training data to generate and export data that has the same resolution as the training data (high temporal resolution), given only its daily average (low temporal resolution). Both 5-mins and hourly resolution were tested. However, with the same set-up of GANs, the 5-mins resolution results is dramatically different from the hourly resolution one.

2 Methodology and Data Description

GANs were initially developed to train and generative images [Creswell et al., 2018]. The fundamental intuition of GANs is to set up the minimax two player game between two deep neural networks (DNNs) that are trained simultaneously: the generator and the discriminator (figure 1). The relationship of two DNNs is in competition, which is why the its addressed as "adversarial": the generator updates its weights for each neuron to generate synthetic "fake" data, whose target is to mimic the temporal dynamics of training data and eventually "fool" the discriminator; the discriminator updates itself to capture the differences between synthetic data and training data, eventually maximize the chance of differentiate the two. The ideal outcome of GANs will be an converged discriminator has a 50-50 chance of differentiate the synthetic data from the training data, meaning that the synthetic data is effectively not indistinguishable. This lead to a generator capable of producing synthetic data that has as much as similarity of the training data, which is assumed realistic. As the generator takes random noise (from the same distribution) as input which is different each time.

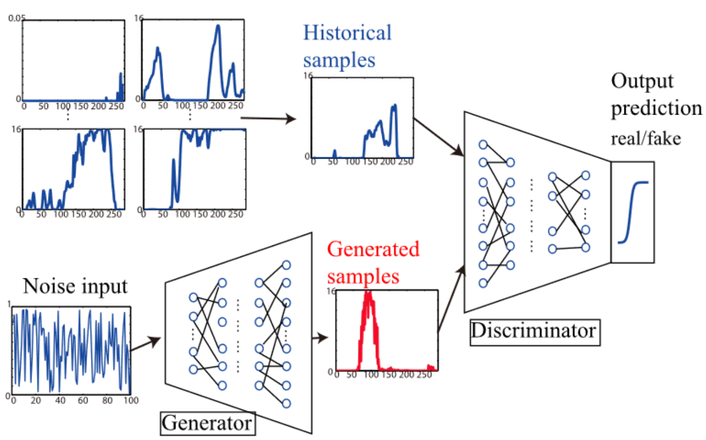


Figure 1: Architecture for GANs. Source: [Chen et al., 2018]

The key success of a GANs model is to find the suitable loss function for particular purpose, which is used to update the neural networks' weight. Loss function L_G for generator G and loss function L_D for discriminator D. Therefore, for each training data set X and random variable Z

sample from a fixed distribution (e.g., Gaussian) P_Z, the generator G is trying to minimize its expectation the discriminator can differentiate itself:

$$L_G = -E_Z[D(G(Z))]$$

Conversely, the discriminator is trying to minimizing the expectation of training data X and maximizing the expectation of generated data:

$$L_D = -E_X[D(X)] + E_Z[D(G(Z))]$$

And eventually, the overall set up of adversarial game is by constructing the value function *V(G,D)* of the minimax game between generator G (min) and discriminator D (max):

$$min_G max_D V(G, D) = E_X[D(X)] - E_Z[D(G(Z))]$$

For this study, the minimax objective is kept the same with the literature: Wasserstein distance (Earth-Mover distance), defined as the distance function, or cost of transporting one distribution to another:

$$W(X, Y) = inf \int |x - y| f_{XY}(x, y)\, dxdy$$

for all x,y belong to the same joint distribution.

So the updated version of value function *V(G,D)* using Wasserstein distance is expressed as:

$$W(D(X), D(G(Z))) = sup_D(E_X[D(X)] - E_Z[D(G(Z)])$$

Which is the overall loss function and objective that the model try to minimize:

$$objective = min(W(D(X), D(G(Z))))$$

The expectation is simply the empirical means for each batch, i.e., how many times the discriminator correctly identified the synthetic data from generator G apart from training data set X. With this mathematical set up and the GANs architecture, the rest of the work focuses on data reconfiguration.

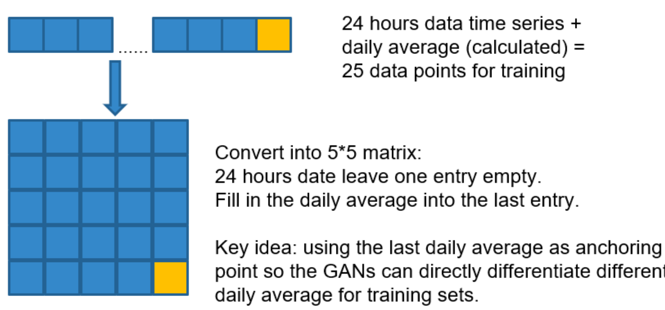


Figure 2: hourly resolution data reshape configuration

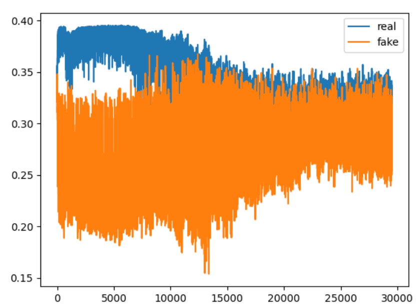
For this project, both 5-mins training data and hourly training data are applied. The 5-min data is extracted from NREL wind integration dataset []. Solar data extracted as well and tested without conditional label for concept proving, but not analyzed. The 5-mins wind data used for this study contains 1 year data for 52 different locations. The hourly data is downloaded from ERA 5 hourly data on single levels from CDS [CDS, 2022] as the same from the same paper with [Amonkar et al., 2022], which contains 256 locations hourly wind/solar data for over 40 years. For concept proving reason, only a 16 location subset of hourly resolution data was used, but kept the whole time series for the selected 16 locations. For this particular project, another input set is required, which is the daily average input used as label for conditional GANs. All the used label input data are generated from down-sampling by simply taking average of each day's input (both 5-mins or hourly).

The neural network set up uses convolution method with 3 layers of neural network. For capturing the daily temporal dynamics, the 5-min resolution data is reshaped to 24*24 size of square "image", containing 2-day of data (= 24*24/(12*24)). The hourly resolution data is reshaped to 5*5 size of square "image". Specifically, a 5*5 image require 24 + 1 input data. This study tried to input the daily average as the last required input entry. The convolution helps capture the temporal dynamics within each data section and gradually reduce/expand the dimension of the data "image" in the discriminator/generator. The data will be truncated if the last image cannot be filled.

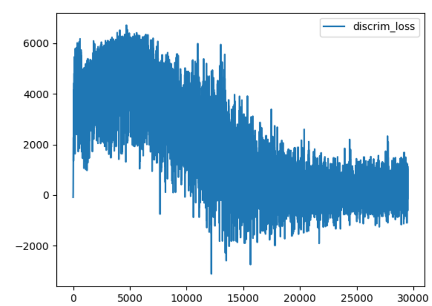
3 Results and Discussions

5-mins resolution results

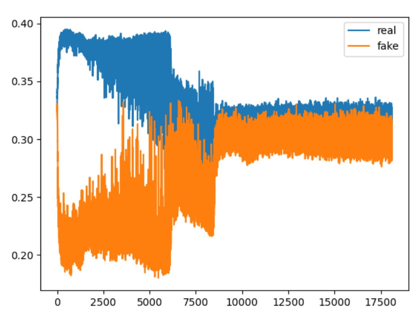
This study started with 5-mins resolution modeling and analysis for verification of concept. For both wind and solar unconditional case using the same hyperparameter inputs, the model converges to desired level, as reported by literature. While solar converges much faster than wind, also to a more stable state.



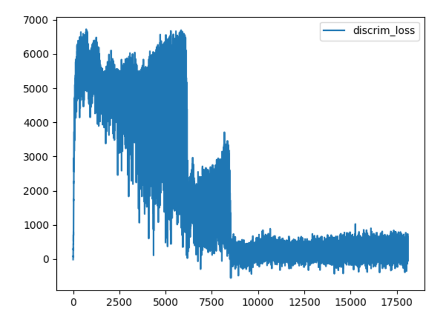
(a) Wind discriminator loss



(b) Wind loss function using Wasserstein distance



(c) Solar discriminator loss

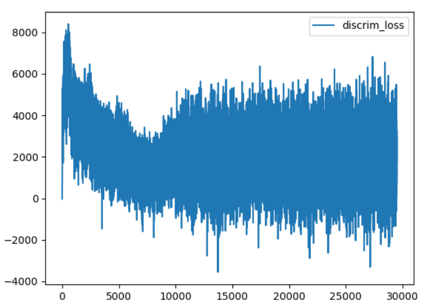


(d) Solar loss function using Wasserstein distance

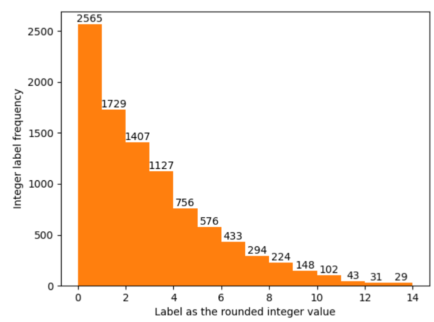
Figure 3: Wind/solar unconditional GANs convergence result with number of iterations with 5-mins resolution

As shown in figure 3, solar case converges within 10,000 iterations while wind case need more than 20,000 iterations. With the same step size (0.0001) of the gradient decent method, wind manage to converge to total loss within level of 2,000 while solar can control the total loss within 1,000. This suggest that wind has more complicated temporal dynamics than solar, which is intuitive correct and also proved by literature. Wind pattern is very stochastic but not truly random which requires more study [Qiu et al., 2017]. while solar pattern is well understood for the fixed solar angle and only affected by the local weather situation such as cloud. This eventually mean wind case needs longer time/iterations to train, smaller gradient descent step size to converge to the same level. Training both wind/solar in the same neural network with the same loss function or hyperparameter set up is not an ideal approach, however, a more common approach. For the sake of this study which tests the performance limit of GANs, we continue with wind case and investigate its potential.

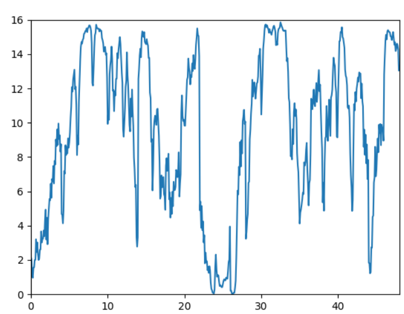
After the success for proof of concept and reproducing the GANs results for synthetic data generation, the next step of the project tests the conditional case using daily average as the condition label that differentiate the different renewable generation profiles, in this case, 5-mins resolution wind. The most intuitive try is by simply calculating the daily average (for 24*24 image size, 2 days average) and directly apply the numarical mean as the label, but the model will simply fail due to the fact that every numerical label is it's own category. In this imprudent trial of labeling case, the model is requiring GANs to treat every "image" with its unique distribution, then each distribution will have only 1 training data and definitely never converge.



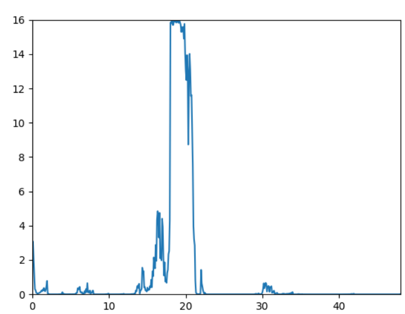
(a) Wind Wasserstein distance loss: integer conditional label



(b) Integer conditional label frequency histogram



(c) Sample output: Label 15, actual mean 9.57



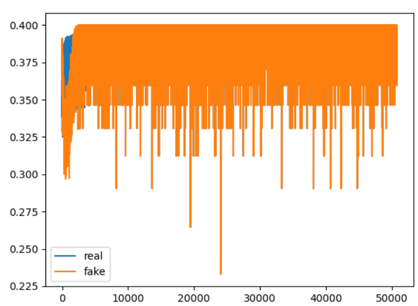
(d) Sample result: Label 1, actual mean 1.10

Figure 4: Wind 5-mins resolution conditional GANs: integer label results. (c)(d): sample data from generator plotting wind power vs. time horizon for 48 hours (2-day data per image)

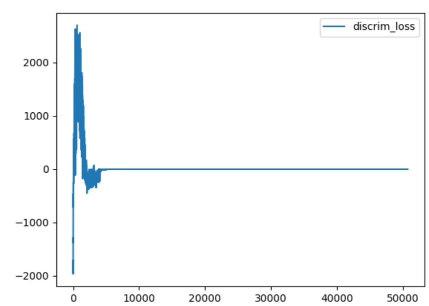
The next case tested was rounded integer label case (figure 4), which lead to a relatively good results, while the convergence of the model is still problematic. The label assign to each "image" in this case is the rounded integer number of the average value, which guarantees that the category of label will be limited the maximum mean, in this case 14 (figure 4b). Such number of labels is among the range that's proved by literature [Chen et al., 2018]. But as the study is using rounded integer of the mean as the label, there is very little flexibility of rearranging the distribution of cases among different labels. As shown in figure 4b, the distribution of each label varies dramatically, with >2,000 training samples for low-wind case but only dozens for high-wind case. This is a direct result of wind power distribution required by the problem set up. Eventually, this labeling technique will demonstrate convergence behavior in the early stage of the training but diverge again in the later stage (figure 4a). A different choice of hyperparameters such as step size won't converge the model, which indicate that the unevenly distributed label is the key problem, that the model is data hungry for the high-wind cases.

Still, the selected output of GANs' generator according to the label is generated and analysis (figure 4c,d). The temporal dynamics of the produced results fits the general performance of wind data. Both assigned label and actual mean of the generated samples are shown for two cases: label 1 and label 15. The label 1 actual mean is pretty close to the assigned label, meaning the GANs is capable of generating results with small label (small mean). However, the actual mean for the case of label 15 is far from the assigned label, the actual mean is only 9.57 while the desired output should have close to 15. The two sample results showing that it the data hungry part with high value label that's creating the model divergence: the model is not capable of generating data that matches the high label case and therefore picked up be the discriminator. This also explain how the tend to converge in the first stage of iterations as it did captures the wind temporal dynamics and behaviors of the low label value cases.

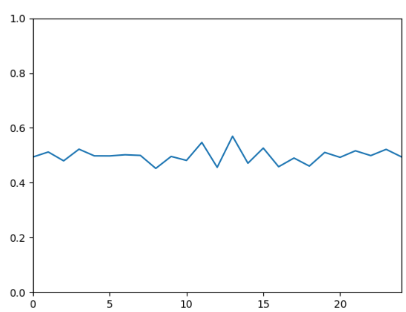
Hourly resolution results



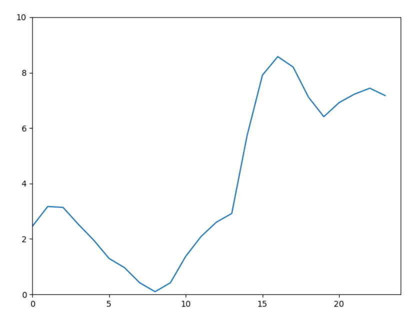
(a) Wind discriminator loss: hourly



(b) Wind Wasserstein distance loss: hourly



(c) Sample output from generator



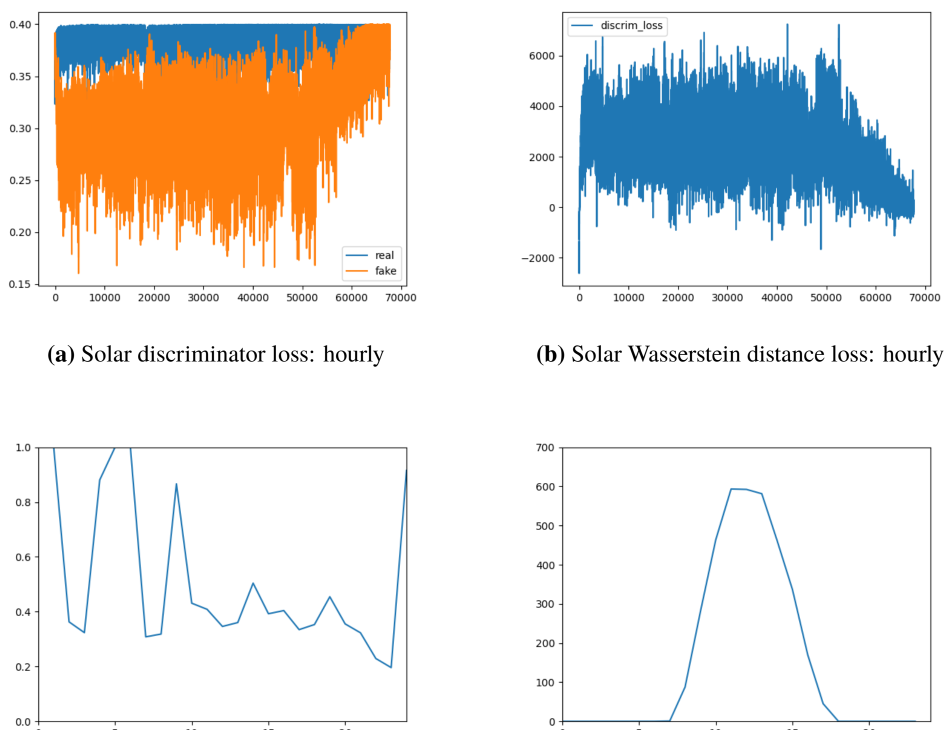
(d) Sample input training data

Figure 5: Wind hourly resolution unconditional GANs. (c)(d): sample output/input for hourly data training plotting wind power vs. 24 hours time horizon

The same methodology and GANs set up is then reconfigured to test the hourly data input with 5*5 image size using convolution. The only modification made is to the convolution function applied to the "image" to ensure the dimension matches for each step of down-sampling (discriminator) or up-sampling (generator).

Figure 5 shows the results of unconditional case. Unfortunately, the abnormality of quick convergence was observed that the discriminator's is totally not functional. The generator finds a way of generating one single output (figure 5c) that could totally fool the discriminator and kill the game forever. In which case, the generator generates a very flat data, whose biggest condition is that: the last (25th) value of the whole "image" data equals to the mean of the first 24 inputs, exactly how the training data is constructed in figure 2. Both generator and discriminator discovered this pattern and the game simply ends. The data reshape method used in figure 2 is found to be overwhelmingly powerful that both generator and discriminator cannot train itself to discover any other temporal patterns in the wind data.

Another way of looking at this is that: wind data is highly stochastic that the pattern is not obvious and strong. While a overwhelmingly strong pattern appears, all the other wind temporal dynamics are simply treated as random errors. This concludes that the idea of taking the mean and directly input to the training data is not applicable.



(a) Solar discriminator loss: hourly

(b) Solar Wasserstein distance loss: hourly

(c) Sample output from generator

(d) Sample input training data

Figure 6: Solar hourly resolution unconditional GANs. (c)(d): sample output/input for hourly data training plotting solar radiation vs. 24 hours time horizon

For exploration, several other testing was also performed: e.g., instead of input the mean value of the 24 hours data to the last entry of the 5*5 "image", dummy input such as 0 or 1 is filled instead. The model performs similarly that it will converge within 1 epoch and the generated data has the characteristic of last value equals to 0 or 1. Not just the mean value trick will be detected, any input with certain pattern will cover the trace of other temporal dynamics. This further exploration suggest that the hourly training data must not have other input, fitting exactly 24 hours data into the "image", e.g. 4*6.

Additional effort was given to the conditional GANs case for wind, which follows the same labeling technique with 5-mins resolution case. The conditional model for hourly wind using mean value as last input data will never converge.

As the wind temporal pattern is very stochastic that even filling the dummy value like 0 or 1 will fail the model, the case of solar may show very different results. Solar time series have very strong temporal patterns (figure 6d) that it naturally have 0 value for the hours in the night time, whose performance should not be affected by the dummy input to the last entry. As shown in figure 6, the model converges without conditional labels. However, the convergence of the model does not produce a reasonable output from generator (figure 6c and 6d): the generator output is dramatically different from input in both scale and shape. This might due to the selection of loss function (Figure 6a). Although the loss function for discriminator is gradually converging, the discriminator did not learn any real information from the training set, as the loss value for real data remains the same at 0.4 level and never changed. Other trial of conditional solar pattern will not converge.

4 Conclusion and Future Work

The findings of this project is summarized as below:

- The concept of Conditional GANs for increasing data resolution is proven to be possible. The limitation of the GANs' performance is the availability of data or loss function selection.
- Wind and solar data have very different behavior patterns that should not be trained under exactly the same GANs hyperparameters settings. Other methodologies such as Bayesian Optimization can help improve the performance of GANs hyperparameter tuning for the optimal results.
- Model convergence is problematic for 5-min resolution mainly due to the unevenly distributed labels. This problem can be potentially solve by: 1) reducing the number of labels to ensure a sufficient amount of training data for each label, which is a trade off between model accuracy and label resolution; 2) greatly increase the size of training data set to ensure a sufficient data, which takes longer time and data resources.
- The same GANs architecture based on convolution can be used for both hourly and 5-mins resolution data, as the model can show characteristics of identifying the patterns embedded in the data.
- The loss function described in the methodology section works for 5-mins resolution but not working for hourly resolution. Other choice of loss function shall be explore for hourly data training.
- In hourly resolution case, although generator cannot produce reasonable output for both cases, it still shows that solar data has stronger temporal pattern than wind. It's anticipated that an hourly resolution wind data training will be much harder than solar.

GANs is one of the most well established generative machine learning model developed for image processing and training. This project explores its capability in training and generating reconfigured time series data and uses convolutional neural networks to investigate the temporal dynamics in the time series. In the future, the task of increasing data temporal resolution can be further explored by trying different types of neural networks (e.g., Concurrent Neural Network) that is more suitable for time series data, or different loss functions which performs better for hourly resolution purpose.

5 Acknowledgement

Many thanks to Prof. Alp Kucukelbir who helped identifying the research scope and GANs as methodology, and Prof. Bolun Xu for the discussion and guidance during the project.

6 Bibliography

IPCC. "IPCC, 2022: Climate Change 2022: Impacts, Adaptation and Vulnerability," 2022. https://doi.org/10.1017/9781009325844.

Conlon, Terence, Michael Waite, and Vijay Modi. "Assessing New Transmission and Energy Storage in Achieving Increasing Renewable Generation Targets in a Regional Grid." Applied Energy 250 (September 2019): 1085–98. https://doi.org/10.1016/j.apenergy.2019.05.066.

Younesi, Abdollah, Hossein Shayeghi, Zongjie Wang, Pierluigi Siano, Ali Mehrizi-Sani, and Amin Safari. "Trends in Modern Power Systems Resilience: State-of-the-Art Review." Renewable and Sustainable Energy Reviews 162 (July 2022): 112397. https://doi.org/10.1016/j.rser.2022.112397.

Aryanpur, Vahid, Brian O'Gallachoir, Hancheng Dai, Wenying Chen, and James Glynn. "A Review of Spatial Resolution and Regionalisation in National-Scale Energy Systems Optimisation Models." Energy Strategy Reviews 37 (September 2021): 100702. https://doi.org/10.1016/j.esr.2021.100702.

Balaji, V., Fleur Couvreux, Julie Deshayes, Jacques Gautrais, Frédéric Hourdin, and Catherine Rio. "Are General Circulation Models Obsolete?" Proceedings of the National Academy of Sciences 119, no. 47 (November 22, 2022): e2202075119. https://doi.org/10.1073/pnas.2202075119.

ISIMIP, 2022. https://data.isimip.org/download/

Amonkar, Yash, David J. Farnham, and Upmanu Lall. "A K-Nearest Neighbor Space-Time Simulator with Applications to Large-Scale Wind and Solar Power Modeling." Patterns 3, no. 3 (March 2022): 100454. https://doi.org/10.1016/j.patter.2022.100454.

CDS, 2022. https://cds.climate.copernicus.eu/cdsapp!/dataset/reanalysis-era5-single-levels?tab=form

Zhao, Jin, Fangxing Li, Haoyuan Sun, Qiwei Zhang, and Hang Shuai. "Self-Attention Generative Adversarial Network Enhanced Learning Method for Resilient Defense of Networked Microgrids Against Sequential Events," October 25, 2022, 11.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A. A. (2018). Generative Adversarial Networks: An Overview. IEEE Signal Processing Magazine, 35(1), 53–65. doi:10.1109/msp.2017.2765202

Qiu, Ting, Bolun Xu, Yishen Wang, Yury Dvorkin, and Daniel S. Kirschen. "Stochastic Multistage Coplanning of Transmission Expansion and Energy Storage." IEEE Transactions on Power Systems 32, no. 1 (January 2017): 643–51. https://doi.org/10.1109/TPWRS.2016.2553678.