Joule

Inferences on Long Duration Storage Requirements for Wind and Solar Energy from a Novel Space-Time Simulator

--Manuscript Draft--

Manuscript Number:	JOULE-D-21-00787
Full Title:	Inferences on Long Duration Storage Requirements for Wind and Solar Energy from a Novel Space-Time Simulator
Article Type:	Research Article
Keywords:	Long Duration Storage; Wind and Solar Generation; Texas Interconnection; Stochastic Simulators; Macro-Energy Systems
Corresponding Author:	Yash Vijay Amonkar Columbia University New York, NY UNITED STATES
First Author:	Yash Amonkar
Order of Authors:	Yash Amonkar
	David J Farnham
	Upmanu Lall
Abstract:	The space-time variability of potential electricity generation is a critical factor for the design of wind and solar dominated power systems. Limited historical data is available on incoming solar radiation and wind. Consequently, a stochastic model for simulating the co-variation of wind and solar fields is needed to assess the probability distribution of the severity and duration of energy "droughts" at the network scale that need to be managed by long duration storage or alternate energy sources. We present a novel k-Nearest Neighbor Space-Time Simulator that accounts for the space-time dependence in high dimensional wind and solar fields and can simulate synthetic wind and solar fields of arbitrary length. Probability distributions of the severity-duration-frequency of regional energy droughts relative to a target are produced. As expected, the severity-duration-frequency of shortfalls is much greater when space-time dependence is properly considered than when it is not.
Opposed Reviewers:	
Suggested Reviewers:	Ashish Sharma, Ph.D. Professor, University of New South Wales Faculty of Engineering ashish.sharma@unsw.edu.au The proposed algorithm builds up on the paper by Lall and Sharma (1996). Dr. Sharma will have a good understanding of the algorithm and be able to provide critiques too.
	Jesse Jenkins, Ph.D. Assistant Professor, Princeton University jessejenkins@princeton.edu Lead Author of "Getting to Zero Carbon Emissions in the Electric Power Sector" - Joule (2018)
	Stefan Pfenninger, Ph.D. Assistant Professor, TU Delft: Technische Universiteit Delft S.Pfenninger@tudelft.nl Author of the paper - "Impacts of Inter-annual Wind and Solar Variations on the European Power System." Joule (2018)
Additional Information:	
Question	**Response**
Standardized datasets A list of datatypes considered standardized under Cell Press policy is available here. Does this manuscript	No

report new standardized datasets?	
Original Code Does this manuscript report original code?	Yes
Reviewers must have anonymous access to these original code that is free-of-cost. Please provide code location and instructions for access here. Please consult this Author's guide for more information: "How standardized datasets and original code accompany Cell Press manuscripts from submission through publication" or email us at joule@cell.com. as follow-up to "**Original Code** Does this manuscript report original code?"	https://github.com/yashamonkar/LDS-Inferences

Department of Earth and Environmental Engineering
Columbia University
New York, NY 10027
June 2021

Dear Editor,

I am writing to submit the manuscript entitled, “Inferences on Long Duration Storage Requirements for Wind and Solar Energy from a Novel Space-Time Simulator” by Yash Amonkar, David J. Farnham, and Upmanu Lall for consideration for publication in *Joule*.

The need for long duration storage to help manage the variability of wind and solar power generation has been increasingly discussed, including in Dowling et al. (2020) in *Joule*. Estimating the necessary storage capacity for a given system requires relatively long data records to identify the statistics of wind and solar “droughts”. However, long records of wind and solar potential are often not available. We address this issue by proposing a novel simulator capable of producing synthetic, realistic, and long-record wind and solar spatio-temporal fields. These fields can then be used to estimate the severity and duration of events when wind and solar generation potential is low, and storage may be needed to bridge the gap in supply.

Our attached paper presents a novel k-nearest neighbors’ model capable of simulating joint wind and solar fields with hundreds of individual sites. The application of the simulator to ERCOT region with 40-yrs of gridded reanalysis data is demonstrated. The ability of the simulator to infer long-duration storage requirements and the corresponding sampling uncertainty is shown. Finally, the importance of the properly simulating the spatial dependence when modelling such large fields is demonstrated by comparing our novel simulator with another model that adequately fits individual sites but fails to capture the spatial dependence.

We have selected the journal *Joule* due to its focus on sustainable energy and interdisciplinary solutions for providing sustainable energy solutions. We think that the novel methodologies and ideas presented in our paper will be of interest and utility to many researchers, industry professionals and academics working at the intersection of sustainable energy, energy storage and its interaction with grid reliability.

This manuscript has not been submitted elsewhere. If you require additional information regarding this manuscript, please contact me at yva2000@columbia.edu.

Thank you for your time and consideration,
Yash Amonkar

Inferences on Long Duration Storage Requirements for Wind and Solar Energy from a Novel Space-Time Simulator

Yash Amonkar[1,2,4,*], David J Farnham[3], and Upmanu Lall[1,2]

[1]Columbia Water Center, Columbia University, New York, New York, USA - 10027
[2]Department of Earth and Environmental Engineering, Columbia University, New York, New York, USA - 10027
[3]Department of Global Ecology, Carnegie Institution for Science, Stanford, CA, USA
[4]Lead Author
[*]Correspondence: yva2000@columbia.edu

Abstract

Summary

The space-time variability of potential electricity generation is a critical factor for the design of wind and solar dominated power systems. Limited historical data is available on incoming solar radiation and wind. Consequently, a stochastic model for simulating the co-variation of wind and solar fields is needed to assess the probability distribution of the severity and duration of energy "droughts" at the network scale that need to be managed by long duration storage or alternate energy sources. We present a novel k-Nearest Neighbor Space-Time Simulator that accounts for the space-time dependence in high dimensional wind and solar fields and can simulate synthetic wind and solar fields of arbitrary length. Probability distributions of the severity-duration-frequency of regional energy droughts relative to a target are produced. As expected, the severity-duration-frequency of shortfalls is much greater when space-time dependence is properly considered than when it is not.

Context and Scale

Many regions plan to integrate more wind and solar generation into the energy grid. Greater reliance on wind and solar generation introduces power supply variability that can pose risks of undersupply. Designing a reliable grid thus requires estimating the frequency, duration, and severity of periods of low wind and/or solar generation potential. We develop and apply a novel statistical simulation model that can produce realistic, synthetic realizations of wind and solar potential across a region, which

allows us to estimate the probability of extreme events that are not necessarily represented in the relatively short observational records available. This tool allows us to estimate the probability of regional wind and solar energy "droughts", and hence allows us to estimate the long duration storage (LDS) needed to achieve desired grid reliability. An application to the Texas Interconnection demonstrates how better estimates of LDS requirements can be obtained through simulation.

Highlights

Space-time correlations in wind and solar fields can be spatially and seasonally non-homogeneous in a region like Texas.

The probability distribution of the aggregate long duration storage (LDS) needed to withstand renewable energy shortages in this setting needs to account for this variable dependence.

A machine learning approach that can capture the joint co-variation of wind and solar fields is presented.

An example application to the Texas Interconnection demonstrates the utility of the approach.

Key Words

Long Duration Storage, Wind and Solar Generation, Texas Interconnection, Stochastic Simulators, Macro-Energy Systems

Introduction

Many countries and US states are mandating reductions in carbon emissions to mitigate anthropogenic climate change, especially from the power sector[1 2 3 4]. At the same time, the costs of wind and solar electricity generation technologies have declined substantially over the last decade[5]. These two factors are spurring increasing deployment of wind and solar based electricity generation.

A target system reliability requirement of 99.97 %[6] necessitates the addition of energy storage, fossil or hydro power sources or significant overcapacity to buffer supply variations if there is a high penetration of variable solar and wind generation.[7 8]. Studies show future scenarios with wind-heavy and/or solar-heavy grid mixes would need long term and even seasonal storage to meet current reliability standards[9 10].

Long Duration Storage (LDS), defined as storage needed to meet deficits for duration greater than 10 hours[11 12], is critical to economically meet grid reliability targets while relying primarily on wind and solar generation[9]. Many recent macro scale electricity studies focusing on renewable electric grids and economy wide de-carbonization models commonly include LDS and expansion of long-distance transmission capacity across the continent to smooth the variation in renewable production[9]. Such an approach necessitates proper consideration of the temporal and spatial dependence structure of available wind and solar energy including their cross-dependence.

Given a candidate regional configuration of wind and solar generators, sizing LDS economically for a regional grid requires estimates of the probability of potential energy shortages for different durations. The estimation of these

probabilities to assure high system reliability requires long data records, potentially over many decades. Collins et al[13] show the pitfalls of modelling energy systems that rely on variable generation using short data records, and note the substantial impact on European power generation costs due to interannual climate variability. Dowling et al[11] analyzed LDS sizing and found that the estimated requirement increased as the record length was increased from 1 to 6 years, emphasizing that long data records are needed to properly estimate LDS requirements. This observation is unsurprising given the low frequency behavior of weather and climate, that is well known to have quasi-periodic modes at seasonal to inter-annual to decadal time scales[14 15 16].

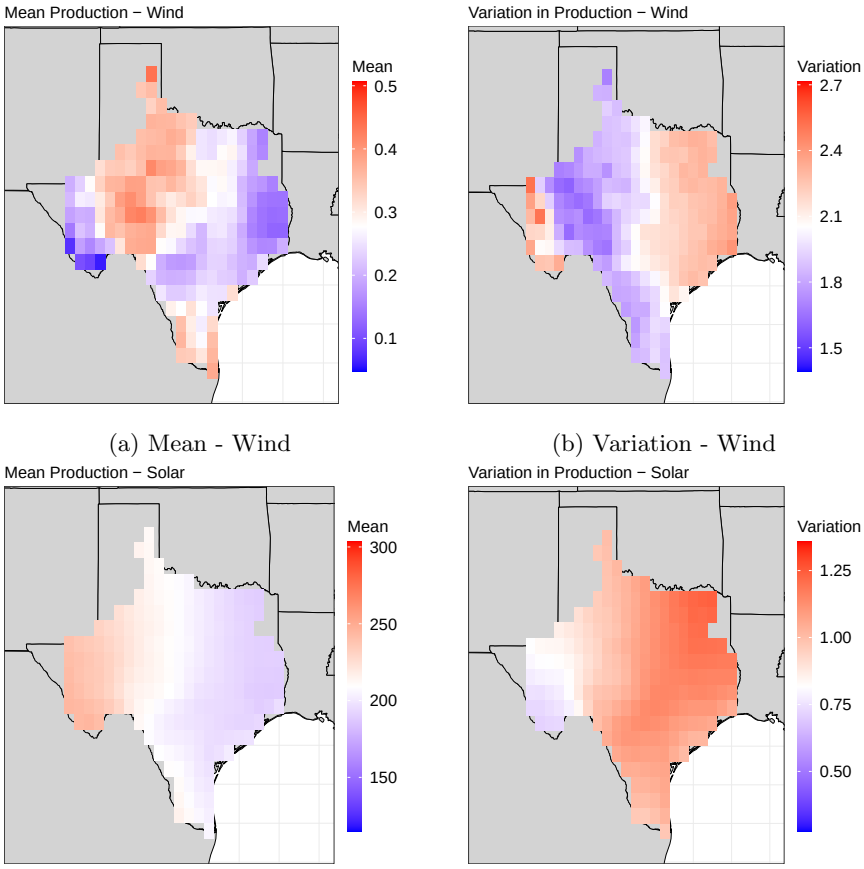
The potential for persistent and long duration solar and wind "droughts" and their potential teleconnection to climate modes was illustrated using several long record stations in the USA[17]. The availability of long-historical wind and solar data records, however, is restricted to a few sites, for example, airports in the United States[18]. Decades long reanalysis datasets[19 20 21 22] are consequently used to generate gridded wind and solar data records. The reliability of reanalysis products is time-dependent with the post 1970s satellite measurement era considered to be the most reliable, effectively setting the data record length to 40 years[23]. This data can used with deterministic optimization methods to compute reliability, capacity allocation, siting and least cost optimization solutions.

An analysis of 39 years of hourly historical (reanalysis) wind and solar data demonstrated the importance of LDS to reduce costs for a wind-solar based electricity system if high reliability is desired[11]. A subsequent paper[24] focused on the Western Interconnection and derived the frequency of solar and wind droughts of different durations using a 39 year historical record. They define a drought when the production from a source drops below a specified threshold. However, they do not explicitly consider the spatial or temporal correlation across wind and solar sources, or the stochastic properties of the duration and severity of aggregate energy droughts. Further, their analysis is limited to what can be extracted from the historical record. A primary goal of our paper is to provide a stochastic analysis capability to assess the probability of the severity and duration of the aggregate long duration energy shortage across a spatial region with both wind and solar generators. In other words, our goal is to develop a flexible methodology capable of estimating the exceedance probability (including its uncertainty) for any wind and/or solar generation shortage event of any given duration and severity and for any portfolio distribution of wind and solar collectors over a domain.

While the instrumental data themselves encode the space-time dependence structure which arises due to seasonality, geography and other climate variations, a finite record is basically a sample or realization from the underlying stochastic process. In this paper we address the challenge of developing a stochastic simulator that can synthetically extend these reanalysis data records while reproducing the space and time dependence structure of the wind and solar fields, so that more reliable estimates of the severity and duration of regional total wind and solar energy potential and their uncertainty can be estimated.

The wind and solar data from the ERA-5 renalysis product [19] are used for the development and testing of a stochastic spatio-temporal model that can provide insights as to the variation of the aggregate energy production from a set of spatially distributed wind and solar generation facilities. We take the Electric Reliability Council of Texas (ERCOT) - Texas Interconnection region [25] as a target example to explore the historical record and to demonstrate the performance of our algorithm. LDS considerations motivate the use of daily data on potential wind and solar resource. The implicit assumptions is that chemical batteries help smooth out the sub-daily time-scale shortages [11]. The potential wind power is derived from hourly wind speed data at 100m using a turbine rating curve and then averaged for the day. The downward surface solar radiation is taken as the mean value for the day (see the Experimental Procedure section).

Over a large region (e.g., the Texas Interconnection), the wind and solar generation assets are likely to be spatially distributed throughout the region [26]. Non-homogeneous and non-local space and time correlations in the potential energy production across the assets utilized by a grid operator are possible. The annual and seasonal variation of the daily wind and solar energy potential across 216 grid points using daily averages of daily wind and solar reanalysis data for our example application to Texas is illustrated in Figures 1 and S1.



(a) Mean - Wind

(b) Variation - Wind

(c) Mean - Solar

(d) Variation - Solar

Figure 1: Mean and variation in daily wind capacity factors and downward surface solar radiation (W/m-sq) across the Texas Interconnection. (a) Mean daily wind capacity factors. (b) Variation in daily wind capacity factors. (c) Mean daily downward surface solar radiation. (d) Variation in daily downward surface solar radiation. The variation is computed as the difference between the 90^{th} and 10^{th} percentile divided by the mean for each grid.

Daily wind and solar radiation often exhibit significant spatial correlation that varies by location and time of year and needs to be accounted for in an analysis of potential renewable energy droughts or LDS system sizing[27]. As seen in Figure 1 and S1, wind and solar along the Gulf of Mexico and the land-area adjoining Louisiana are regions with lower generation potential but with high variability. The mean wind capacity factor and its variability (Figure 1) is non-homogeneous. The highest capacity factors are in the north-western and southern-most portions of the interconnection. The highest variability is in the eastern portion of the interconnection (Figure S1). Daily wind capacity

factors are generally highest during spring and lowest during summer while wind variation is highest during fall and lowest in summer and spring. The mean daily radiation and its variability (Figure 1) is more homogeneous and a function of the season, with low mean radiation and high variability in winter (DJF) and high mean radiation and low variability in the summer (JJA).

The cross field spatial correlation along with seasonality in the wind and the solar fields is illustrated in Figure S2 where significant local and non-local spatial correlation structures are evident. The temporal dependence structure explored through the dominant principal component of each field also shows heterogeneity between fields (Figure S3).

The new k-Nearest Neighbors Space-Time Simulator (KSTS) algorithm is structured as follows. A model for temporal variability at each site and for each variable (wind and solar) is considered first. This entails defining a state space through an embedding of the time series. A time series simulation can then be achieved by sequentially drawing from the k nearest neighbor successors of the embedding at each time step, but this will not preserve spatial dependence. Spatial dependence is then introduced by identifying the most likely neighbors of the full spatial field by aggregating neighbor likelihoods for each site/variable. If the state space evolution at two sites is similar (i.e., identified by the same neighbors in time), then the evolution of those two sites would be fully synchronous. Thus the similarity in the selection of neighbors reflects the similarity in dynamics and provides a useful basis for space-time conditioning of a random field's dynamics. The k-nearest neighbors identified across all the sites as the most similar at a given time, are then used to randomly draw a full spatial field for the next time step, using a kernel function that accounts for their degree of similarity through a probability measure. The process is repeated sequentially to generate a time series simulation of the spatial field.

The target variables, wind power and downward surface solar radiation, have non-Gaussian and skewed distributions. They are also potentially bounded, since the maximum wind a generator can use is limited and the maximum incoming solar radiation on a calendar day is also limited. Consequently, the probabilistic sampling using k-nearest neighbors provides an effective approach to sampling from a nonparametric distribution applied to each target variable. The seasonality in the variables is accounted for by restricting search of k-nearest neighbors using a moving window around the Day of Year (DOY). This new algorithm generalizes a k-Nearest neighbor algorithm [28] used for univariate or low dimensional multivariate simulations of non-Gaussian and nonlinear dependence that has been used extensively for other climate variables [29 30 31 32].

We apply our new KSTS algorithm to assess the severity-duration-frequency of long duration storage needs associated with the aggregate regional energy production. We show that the simulator captures the regional aggregate as well as the site by site probabilities of wind and solar energy produced including the spatial correlation within and across the two fields and the temporal autocorrelation at each site. We also run a simulation (henceforth termed KNN) that preserves the time series structure but not the spatial structure or the wind-solar dependence. As one may expect, this demonstrates a significant underestima-

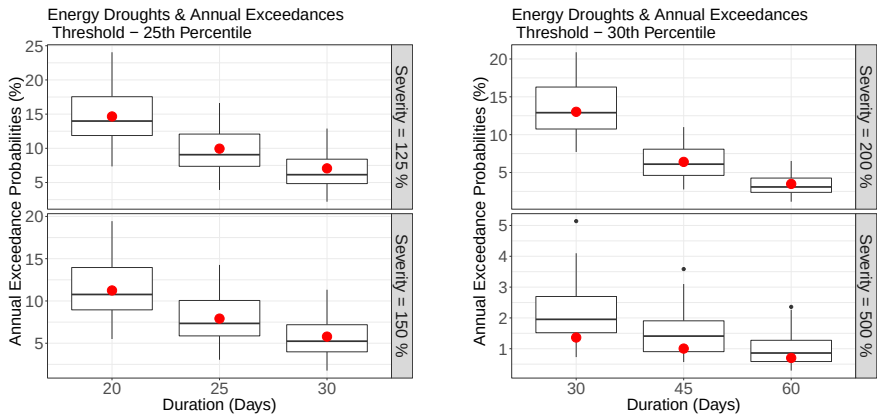
tion of the regional LDS probabilities.

For the application presented, we use the 40-yr gridded daily wind and solar data from the ERA-5 [19] reanalysis dataset for 216 sites (grids/nodes) in the Texas Interconnection. Using the KNN or KSTS algorithm one can generate a large number (e.g., 100) of synthetic 40-year simulations (or equivalently a 4000 year simulation) of the daily wind and solar fields, without and with spatial dependence preserved, respectively. From each simulation we extract the duration and severity of each drought event as a shortage in aggregate energy produced across the grid, relative to a target threshold. The probabilities of LDS severity and duration can then be assessed from this derived set of events. If multiple simulations of 40 years are generated, then one can get also get an estimate of the uncertainty associated with the probability of a LDS severity-duration given 40 years of data. If a single long simulation is generated, then we can estimate LDS severity-duration probabilities with reduced uncertainty using the longer synthetic record.

Results

We present an evaluation of the severity-duration-frequency of the aggregate energy droughts for the Texas Interconnection with (KSTS) and without (KNN) preserving the spatial structure and wind-solar dependence in simulations. For illustrative purposes a uniform land allocation (and consequently capacity within wind and solar separately) for generation is considered. For both types of simulations, we generated 48 realizations of 40 years of daily wind and solar data at each of the 216 sites. In an actual use case, an optimization model may be used to allocate wind and solar capacity across the sites and estimate the regional size of LDS capacity using the simulations developed. The results presented here illustrate the importance of getting the space-time dependence right in the simulations for a proper estimation of the regional LDS capacity given a candidate spatial configuration of wind and solar generation. Detailed performance statistics of the simulator are presented in the supplement.

Severity-Duration-Frequency of Energy Droughts



(a) 25th percentile threshold (b) 30th percentile threshold

Figure 2: Probability of joint annual exceedances for energy droughts given a duration and severity with threshold values of (a) 25th percentile and (b) 30th percentile. The red dot denotes the exceedance probability calculated from the reanalysis data. The boxplots denote the uncertainty in the 48 generated simulations using KSTS. The duration is in days and the severity is denoted in terms of percentage of the mean historical reanalysis value. For each box-plot, the thick black horizontal line across the box denotes the median of the annual exceedance probabilities from the simulations and the edges of the box denote the 25th and 75th percentiles, and the lower and upper extents of the vertical lines outside the box denote the 5th and 95th percentiles.

Energy droughts are defined as continuous periods when the daily production falls below a target threshold. The threshold value, changing every calendar day in a year, can be thought of as a forward contract's daily obligation to be supplied based on the seasonality of the historical reanalysis data. The severity of the drought is the accumulated deficit in production over the duration of the event, i.e., the level of default on a potential contract covering the period, while the duration of the event is the duration during which the deficit exists. Figure 2 (a) shows the annual exceedance probabilities for energy droughts of duration 20, 25 and 30 days with severity of 100% and 150% when the target threshold is the 25th percentile of the distribution of energy that could be produced over that period based on the historical data. The severity of energy droughts was scaled by the mean daily historical production, with a severity of 100% denoting a shortfall equal to the mean daily historical value. The annual exceedance probabilities were computed using local regression (Locfit)[33] with the number of exceedances regressed against the duration and severity using a Poisson link function. (see Experimental Procedures Section and Supplementary Materials)

The KSTS simulations bracket the exceedance probabilities seen in the re-

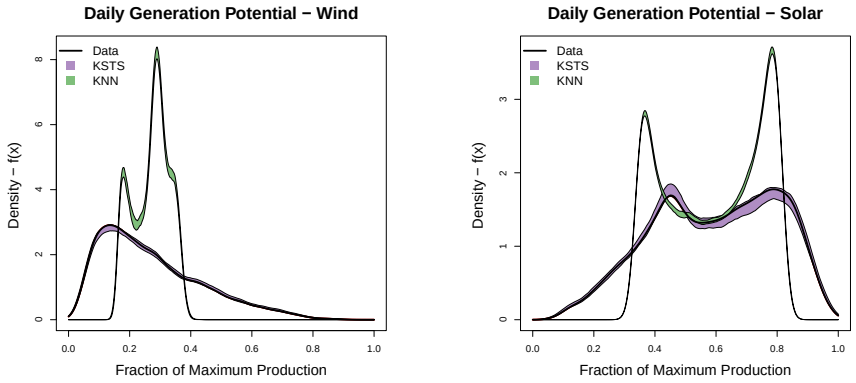
analysis data (Figure 2). For example, an energy drought with duration over 30 days with a severity of 150% relative to a threshold guaranteeing delivery set at the 25th percentile of daily regional generation, has an annual exceedance probability of $\sim 5\%$, based on the reanalysis data. This corresponds to an event that may be expected to be exceeded once every 20 years. The median exceedance probability from the simulations is quite close to this, but with considerable uncertainty around that value. The 25th to 75th percentiles from the simulations are around 4% to 7.5% with the 5th and 95th percentiles extending from 2% to 11%, demonstrating the limitations of using solely the original 40 year record for such evaluations.

Results from increasing the target threshold to the 30th percentile of daily regional energy production and looking at higher severity and longer duration droughts are shown in Figure 2 (b). The KSTS simulations bracket the exceedance probabilities seen in the reanalysis data for the severity of 200%. The simulations show higher exceedance probabilities than the data for the 500% case, which is not surprising considering these are rare events with mean annual exceedance probabilities of 0.5-1.5% and thus are difficult to identify given relatively short data records. The severity/duration probabilities from the historical record of 40 years have high uncertainty for events that are rarer than perhaps once every 10 years (annual exceedance probability of 0.1) given this record length[34,35]. The simulations show that these extreme events could occur far more frequently than if we rely on just the short historical records. In these illustrations, we considered specific thresholds for supply guarantee, specific drought durations and severity levels and present the range of probabilities of exceedance from the simulations. In a system design optimization model, for a candidate spatial configuration of generation, the simulator would provide the probability distribution for a candidate LDS capacity that is considered to meet the deficit over a specified duration (e.g., specified by a contract). Alternately, one could also compute the probability distribution of the shortage beyond the candidate LDS to assess potential penalties for non-delivery, if those were considered in the optimization model.

Annual exceedance probabilities for different combinations of duration and severity, and for multiple thresholds are provided in Figure S4. The entire joint distribution of duration and severity for all energy droughts in the data and the generated simulations relative to a threshold for thresholds at the 25th, 30th, 35th, 40th percentile are shown in Figure S5. We see that KSTS is effective for representing a range of energy droughts. Similar boxplot estimates (Figure 2) for the KNN algorithm generated simulations are not shown since the simulations show no occurence of energy droughts at these thresholds unlike the reanalysis data.

KSTS Reproduces the Aggregate Generation

The simulations from both KSTS and KNN reproduce temporal dynamics and data characteristics across both wind and solar fields at individual sites. The moments (mean and standard deviation), minimum and maximum for individual sites in KSTS and KNN simulations are representative of the underlying data (Figure S6). Both simulators are able to reproduce the quantiles (Figure S7, Figure S8), underlying probability distribution (Figure S9), auto-correlation structure (Figure S10), and site-level seasonality (Figure S11).

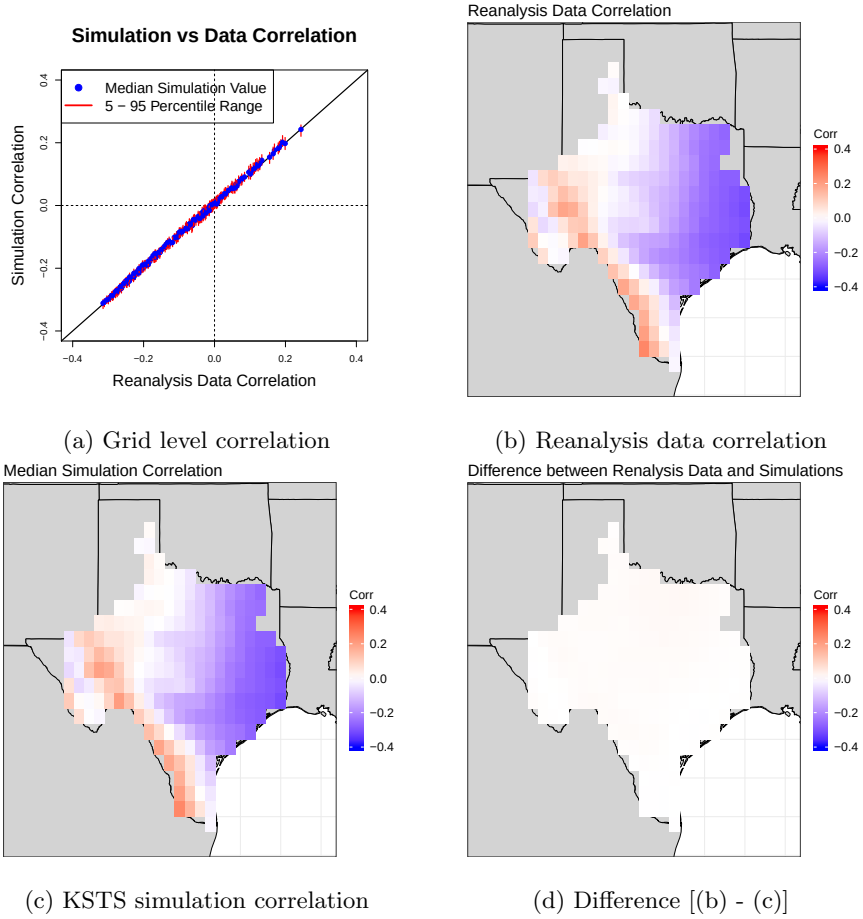


(a) Wind (b) Solar

Figure 3: Kernel density estimate / Probability density function (PDF) of the daily aggregated energy production across the Texas Interconnection simulated using KSTS (purple) and KNN (green). The black line denotes the observed data pdf. The purple and green regions show the mid 90th (5th-95th) percentile interval regions from the individual pdfs computed from 48 simulations from each simulator. (a) Wind, (b) Solar. Equal land and capacity allocation is assumed across all the nodes/grids.

The kernel density estimate of aggregated daily energy generation potential across the Texas Interconnection is shown in Figure 3 for the historical reanalysis record (black), and for the KSTS (purple) and KNN (green) simulations. The degree to which adequate consideration of the spatial dependence and the wind-solar correlation leads to a proper representation of the potential for energy production is illustrated through the fidelity of the KSTS simulations to the density function from the observations, and the marked departure of the KNN based simulations. It is clear that modeling spatial and cross field dependence is important to get the right frequency of the tail events (i.e., for LDS probabilities), even if the site-level production is adequately simulated without considering spatial dependence.

KSTS Reproduces Cross-Field Dependence



(a) Grid level correlation

(b) Reanalysis data correlation

(c) KSTS simulation correlation

(d) Difference [(b) - (c)]

Figure 4: Pearson correlation between wind and solar at each grid point based on simultaneous simulations of wind and solar using KSTS. (a) Simulation correlation vs reanalysis data correlation between wind and solar where the red lines denote the mid 90th (5th-95th) percentile range and the blue dots denote the median value in the simulation spread generated using the KSTS method. (b) Map of the grid-wise correlations in the reanalysis data record. (c) Map of the grid-wise median simulation correlations using KSTS. (d) Map of the difference between (b) and (c).

From Figure 4 we note that the grid-wise correlation between daily wind and solar across ERCOT is well reproduced by the KSTS simulations that are based on simultaneous modeling of the wind and solar fields. By comparison, the KNN simulations do not exhibit grid-wise wind-solar correlations consistent with the

reanalysis data (Figure S12). Furthermore, the spatial correlation structure across all grids within a field for both wind and solar is also well reproduced by the KSTS simulations unlike the KNN simulations (Figure S13). The seasonal variation in the correlation between wind and solar is also well modeled by the KSTS algorithm (Figure S14 and Figure S15).

Discussion

The primary contribution of this paper is the presentation of a novel k-Nearest Neighbor Space Time Simulator (KSTS) and its application to the joint wind-solar fields across the Texas Interconnection. We demonstrated the importance of using a stochastic simulator that can properly reproduce the marginal probability densities of wind and solar at each site, as well as the cross-field spatial dependence structure if good estimates of the severity-duration and frequency of long duration renewable energy droughts, are of interest. So far, much of the development of renewable electricity sources has focused on local microgrids, but there has been growing interest in national and regional grids [36]. As the scale is increased, there is evidence that LDS is an effective and economic component of the design of these regional systems [10 11 37]. However, most of the models developed and applied at these scales are deterministic and use relatively short records [38]. They do not consider the possible contracting structures for guaranteed delivery and the associated default penalties. The probabilities of the severity and duration of defaults as well as the penalties and LDS costs would ultimately determine economically optimal resource allocations. We anticipate and are planning to develop stochastic simulation-optimization models to address a range of questions associated with such designs and contracts. The KSTS simulator is motivated by this context, and it was important to understand how important modeling spatial dependence to assess rare probabilities of the duration and severity of shortages from a grid was for a real system.

From the application to the Texas Interconnection, we note that there is substantial seasonal variability in the spatial expression of potential wind and solar resource. This is not a surprise. The point by point wind-solar correlation varies substantially by location and by season, as does the spatial correlation structure for wind and solar and their cross-dependence. If these factors are ignored, then the resulting regional LDS probability distributions are compromised quite significantly. The simulations assuming a uniform land and capacity allocation scenario show large uncertainties in the annual exceedance probabilities for the severity-duration-threshold combinations considered, as well as potentially higher exceedance probabilities than computed from the 40 year data record for the more extreme threshold-severity-duration combinations.

The KSTS simulator is nonparametric and is appropriate for this setting where the target variables are bounded with non-Gaussian distributions with space and time dependence across variables changing by season. Since KSTS is based on sampling the observed data, it can be thought of as a spatio-temporal bootstrap procedure, where a spatio-temporal kernel is used at each time step

to sample a historical field with probabilities determined by the kernel and a distance metric applied to the temporal state space for each variable. The temporal sequences of potential energy produced at each site and across the region are different even though the individual daily values are resampled from the historical record. This allows the analysis of the range of drought severity-duration-frequency using an extended sample. Extensions of the simulator to hourly or other time scales are feasible. They would need to consider the diurnal cycle as well as the seasonal cycle, and we are exploring computationally efficient strategies for an algorithm that can address this while maintaining spatial and cross field dependence.

The KSTS simulator exploits the similarity in the temporal evolution across the fields and grid points. The potential next step would be developing an algorithm which is capable of capturing the heterogeneity in dynamics across even larger regions. This becomes important when the spatial scale of the simulation is expanded from the Texas Interconnection to either the Western or Eastern Interconnection or the entire North American continent. Such a large scale makes it more likely that the wind and solar availability is driven by different climate dynamics and consequently their temporal evolution structure would be heterogeneous when compared to just Texas.

Experimental Procedures

Resource Availability

Lead Contact

Further information and requests for resources and materials should be directed to Yash Amonkar yva2000@columbia.edu

Materials Availability

This study did not generate new unique materials.

Data and Code Availability

The KSTS and KNN generated simulations use wind and solar data spanning 40-yrs (1979-2018) across the Texas Interconnection and are taken from the ERA-5 reanalysis dataset[19], which can be accessed publicly. All code used in this study is made publicly available on Github at https://github.com/yashamonkar/LDS-Inferences.

Wind and Solar Data

The variables used are wind speed at 100 meter altitude and downward surface solar radiation from the ERA-5 reanalysis dataset[19]. The spatial grid size of the data is set at 0.5° lat $\times$ 0.5° lon and contains 216 grid points across the Texas Interconnection domain (Figure S16).

Wind power is estimated by converting the 100 m wind speed to wind power using the wind turbine power curve from a V90-2.0MW Vestas turbine (as shown in Fig S17). We then estimate wind power availability per unit area (W/m^2) by assuming a required land area per turbine equal to four times the rotor diameter in one direction and seven times the rotor diameter in the other direction [39]. This means that each turbine is assumed to take up 0.2268 km^2 given the rotor diameter of 90 m for the V90-2.0MW Vestas turbine. The data are converted to the daily time step by taking the mean of the hourly data for each day and the dataset spans the 40 years from January 1st, 1979 to December 31st, 2018.

The solar variable is the downward surface solar radiation (W/m^2) and is converted to the daily time step by taking the mean of the hourly data. The solar data are not converted to capacity factors. The conversion to capacity factors or solar power involves making assumptions about the PV technology, tilt and tracking capabilities of the array and are not made in this study.

Our example application assumes a uniform land allocation for both wind and solar generators.

Energy Deficits and Drought Spells

The daily energy deficit is defined as the daily deviation below a percentile threshold for that day of year (DOY) for each site. The deviation could be positive if it is greater than the selected threshold percentile value or negative if it is lower. The daily energy deviation across the field is computed by aggregating the daily site deviation and is given by,

$$y_t = \sum_{i=1}^{n} (x_{i,t} - \widetilde{x_{i,T}})$$

where, y_t is the aggregated daily energy deviation at day t; $x_{i,t}$ is the normalized wind or solar value at site i and day t; $\widetilde{x_{i,T}}$ is the normalized DOY percentile based on the selected threshold for site i and day DOY(t); n is the total number of grid points (216) times the fields (wind and solar). The aggregated deviation y_t can take a positive (surplus) or negative (deficit) value on any particular day, while the cumulative deficit, the variable of interest is computed as,

$$z_1 = \max(0, -y_1)$$
$$z_t = \max(0, z_{t-1} - y_t)$$

where, z_t and y_t are the cumulative deficit and daily deviation at day t respectively. While y_t can either be positive or negative, the cumulative deficit takes a lower value of 0 (surplus) and is restricted to positive values (periods of energy deficit). Energy Droughts for a selected threshold percentile are defined to occur during instances of consecutive days with positive values of cumulative deficit. Severity of a drought event is defined as the maximum cumulative deficit during the drought period, while the duration is the spell length in days.

Annual Exceedance Probability

The previous section is used to compute the duration and severity for all energy droughts in the data and the generated simulations. The number of exceedances (e_i) for each drought i include all drought events in the data record (or individual simulation realizations) having a greater severity and greater duration than event i, which are computed as,

$$C(e_i) = \sum_{j=1}^{n}(d_i > d_j) \bigcap \sum_{j=1}^{n}(s_i > s_j) \tag{1}$$

where, $C(e_i)$ is the count of exceedances for drought event i with duration d_i and severity s_i and n is the total number of drought events. The count of exceedances $C(e_i)$ is regressed against the severity s_i and duration d_i using Poisson regression. The methodology used is local regression using the locfit package[33].

After the model fitting process, the count of exceedances $C(e_t)$ is estimated using the fitted model for the required duration d_t and severity s_t for a desired drought event t. The number of years of the record (yr) is then used to scale the number of exceedances to get the annual exceedance probability (p) using the formula:-

$$p_t = \frac{C(e_t) \times 100}{yr} \tag{2}$$

where, p_t is the annual exceedance percentage for a drought event t with severity s_t and duration d_t.

Simulator Hyperparameters

For both KSTS and KNN methods, the seasonality in the data is accounted for by restricting the search of nearest neighbors to a $\pm$30 days moving window across the years around the day of the year (DOY). The number of nearest neighbors (k) selected is approximately $\sqrt{n}$. With 40 years and 61 days per year, $\sqrt{n}$ is ~ 50, where n is the number of possible candidate neighbors after accounting for the moving window[28]. A lag-1 dependence structure for the state space is assumed. 48 independent simulation realizations, each of the same length as the reanalysis data (40 yrs = 14610 days), are generated using the KSTS and KNN algorithms.

The KNN algorithm is fit to each grid individually using the hyper-parameters specified above with the algorithm which is outlined in Lall and Sharma[28].

k-Nearest Neighbor Space-Time Simulator (KSTS)

The general structure and a cartoon example application of the KSTS algorithm is illustrated in Figures S18 and 5 respectively. The algorithm leads to a space-time simulation process that is Markovian (or corresponds to a state space formed by the embedding) in time.

KSTS Algorithm

Step 1:- Define the composition of the state space $D_{i,t}$.

Define a state space $D_{i,t}$ of dimension m which is the number of embedding delay lags. The state space can be a single lag, multiple lags and/or disjoint lags allowing for custom time dependencies. The embedding selected for the simulator application could be,

Case 1 $D_{i,t}$:- ($x_{t\text{-}1}$, $x_{t\text{-}2}$); m = 2
Case 2 $D_{i,t}$:- ($x_{t\text{-}\tau}$, $x_{t\text{-}2\tau}$, $x_{t\text{-}\phi}$, $x_{t\text{-}2\phi}$); m = 4, $\tau = 1$, $\phi = 12$
Case 3 $D_{i,t}$:- ($x_{t\text{-}1}$, $x_{t\text{-}4}$ $x_{t\text{-}7}$); m = 3

Case 1 represents simple dependence on the two previous values. Case 2 represents dependence on the past two values and values 12 and 24 steps before the current value allowing for monthly and inter-annual dependence for monthly data. Case 3 represents incorporation of a temporal dependence structure unique to the data. The state space $D_{i,t}$ is defined for each time series at site i and time t, whereas $D_{i,T}$ are all the historic vectors which correspond to the selected embedding structure for site i.

Step 2:- Compute the k-nearest neighbors for all sites at time t.

At time step t and site i using the current state space vector $D_{i,t}$, identify the k-nearest neighbors using the weighted Euclidean distance measure,

$$r_{i,t} = \left(\sum_{j=1}^{m} w_j([D_{i,t}]_j - [D_{i,T}]_j)^2 \right)^{1/2}$$

where, $[D_{i,t}]_j$ and $[D_{i,T}]_j$ are the j^{th} components of $D_{i,t}$ and $D_{i,T}$ respectively and w_j are the weights assigned to each of the embedding lags j. This is repeated for all sites. The ordered set of time indices which correspond to the k nearest neighbors (as defined by the euclidean distances stored in $r_{i,t}$) of site i and time t computed above are stored in $\tau_{i,t}$.

Step 3:- Compute resampling probabilities for k nearest neighbor indices using a discrete kernel p_j.

$$p_j = \frac{1/j}{\sum_{j=1}^{k} 1/j}$$

where p_j is the resampling probability for the jth element (time instance of the jth nearest neighbor of $D_{i,t}$) in $\tau_{i,t}$. The resampling kernel stays the same across all time t and across all sites, and is pre-computed and stored prior to simulation. It is a function of the number of neighbors k and not the distances.

Step 4:- Define $T_{i,t}$ and similarity matrix S_t for time t.

Define $T_{i,t}$ as a matrix where the rows and columns correspond to the sites and unique time indices from the historical data respectively. The columns record the resampling probabilities associated with the time indices for the k-nearest neighbors in $\tau_{i,t}$ for each site i, with values being 0 for other time indices. The similarity matrix S_t is then defined as the sum of all elements in each column in $T_{i,t}$.

$$S_t = \sum_{i=1}^{s} T_{i,t}$$

where s is the total number of sites. The similarity matrix S_t has the same length as the number of unique time indices in the data.

Step 5:- Curtail and scale the similarity matrix S_t.

The similarity matrix S_t is ordered and curtailed to its highest k values. The time indices associated the k highest values of S_t are selected as the k-nearest neighbor candidates for the entire spatial field. The probabilities of the associated k neighbors are scaled to add up to 1.

$$[S_t]_j = \frac{[S_t]_j}{\sum_{j=1}^{k} [S_t]_j}$$

Step 6:- Re-sample the full spatial field for time $t + 1$.

Using the discrete probability mass function S_t, sample a single value and re-sample entire fields across all sites from the time index which corresponds to the selected value in S_t as data for the simulation at time $t + 1$. Return to Step 2 if further time-steps are needed for the simulation..

Refer supplementary materials for further details on the algorithm and hyper-parameter selection.

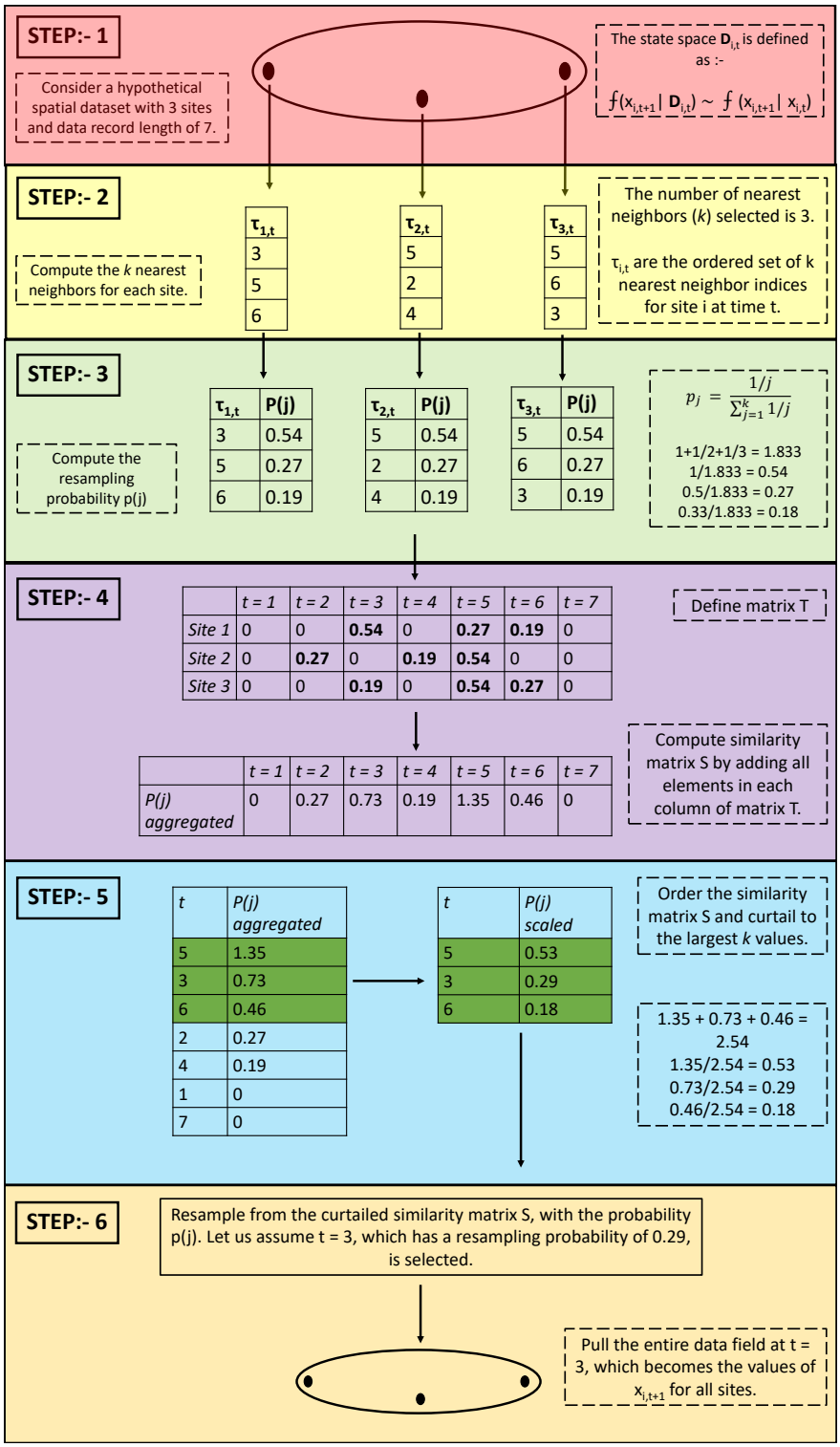


Figure 5: Cartoon of the KSTS algorithm for an example application with a spatial dataset consisting of 3 grids/sites and data record (time) length of 7.

Acknowledgment

Y.A acknowledges support from the Cheung-Kong Innovation Doctoral Fellowship. D.J.F. was supported by a gift from Gates Ventures LLC to the Carnegie Institution for Science.

Author Contributions

Y.A developed the code and performed the computations. Y.A and D.J.F designed the analysis, conceived experiments and simulation checks with supervision from U.L. D.J.F provided the data. Y.A took the lead in writing the manuscript with all authors discussing and contributing to the final manuscript.

Declaration of Interests

The authors declare no competing interests.

References

[1] California Legislative Information. Senate Bill -100 California Renewables Portfolio Standard Program: emissions of greenhouse gases., 2018. URL https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill$_id = 201720180SB100$.

[2] Jeff Deyette. States March toward 100% Clean Energy–Who's Next?, August 2019. URL https://blog.ucsusa.org/jeff-deyette/states-march-toward-100-clean-energy-whos-next. Section: Energy.

[3] European Comission. REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL establishing the framework for achieving climate neutrality and amending Regulation (EU) 2018/1999 (European Climate Law), 2020. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1588581905912uri=CELEX:52020PC0080.

[4] New York State Legislatur. New York's Climate Leadership and Community Protection Act (CLCPA), 2019. URL https://climate.ny.gov/.

[5] International Renewable Energy Agency. Renewable Power Generation Costs in 2019. Technical report, 2020. URL /publications/2020/Jun/Renewable-Power-Costs-in-2019.

[6] North American Electric Reliability Corporation. (NERC). 2012 State of Reliability. Technical report, 2012. URL https://www.nerc.com/files/2012$_sor.pdf$.

[7] Marc Beaudin, Hamidreza Zareipour, Anthony Schellenberglabe, and William Rosehart. Energy storage for mitigating the variability of renewable electricity sources: An updated review. *Energy for Sustainable Development*, 14(4):302–314, December 2010. ISSN 0973-0826. doi: 10.1016/j.esd.2010.09.007. URL https://www.sciencedirect.com/science/article/pii/S0973082610000566.

[8] Jacques Després, Silvana Mima, Alban Kitous, Patrick Criqui, Noureddine Hadjsaid, and Isabelle Noirot. Storage as a flexibility option in power systems with high shares of variable renewable energy sources: a POLES-based analysis. *Energy Economics*, 64:638–650, May 2017. ISSN 0140-9883. doi: 10.1016/j.eneco.2016.03.006. URL https://www.sciencedirect.com/science/article/pii/S0140988316300445.

[9] Jesse D. Jenkins, Max Luke, and Samuel Thernstrom. Getting to Zero Carbon Emissions in the Electric Power Sector. *Joule*, 2(12):2498–2510, December 2018. ISSN 2542-4351. doi: 10.1016/j.joule.2018.11.013. URL https://www.sciencedirect.com/science/article/pii/S2542435118305622.

[10] Matthew R. Shaner, Steven J. Davis, Nathan S. Lewis, and Ken Caldeira. Geophysical constraints on the reliability of solar and wind power in the United States. *Energy & Environmental Science*, 11(4): 914–925, April 2018. ISSN 1754-5706. doi: 10.1039/C7EE03029K. URL https://pubs.rsc.org/en/content/articlelanding/2018/ee/c7ee03029k. Publisher: The Royal Society of Chemistry.

[11] Jacqueline A. Dowling, Katherine Z. Rinaldi, Tyler H. Ruggles, Steven J. Davis, Mengyao Yuan, Fan Tong, Nathan S. Lewis, and Ken Caldeira. Role of Long-Duration Energy Storage in Variable Renewable Electricity Systems. *Joule*, 4(9):1907–1928, September 2020. ISSN 2542-4351. doi: 10.1016/j.joule.2020.07.007. URL https://www.sciencedirect.com/science/article/pii/S2542435120303251.

[12] ARPA-E. Duration Addition to electricitY Storage, 2018. URL https://arpa-e.energy.gov/technologies/programs/days.

[13] Seán Collins, Paul Deane, Brian Ó Gallachóir, Stefan Pfenninger, and Iain Staffell. Impacts of Inter-annual Wind and Solar Variations on the European Power System. *Joule*, 2(10):2076–2090, October 2018. ISSN 2542-4351. doi: 10.1016/j.joule.2018.06.020. URL https://www.sciencedirect.com/science/article/pii/S254243511830285X.

[14] Luc Bonnafous, Upmanu Lall, and Jason Siegel. A water risk index for portfolio exposure to climatic extremes: conceptualization and an application to the mining industry. *Hydrology and Earth System Sciences*, 21 (4):2075–2106, April 2017. ISSN 1027-5606. doi: 10.5194/hess-21-2075-2017. URL https://hess.copernicus.org/articles/21/2075/2017/. Publisher: Copernicus GmbH.

[15] Shaleen Jain and Upmanu Lall. Floods in a changing climate: Does the past represent the future? *Water Resources Research*, 37(12):3193–3205, 2001. ISSN 1944-7973. doi: https://doi.org/10.1029/2001WR000495. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2001WR000495. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2001WR000495.

[16] James Doss-Gollin, David J. Farnham, Scott Steinschneider, and Upmanu Lall. Robust Adaptation to Multiscale Climate Variability. *Earth's Future*, 7(7):734–747, 2019. ISSN 2328-4277. doi: https://doi.org/10.1029/2019EF001154. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019EF001154. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019EF001154.

[17] David J. Farnham. *Identifying and Modeling Spatio-temporal Structures in High Dimensional Climate and Weather Datasets with Applications to Water and Energy Resource Management*. PhD thesis, Columbia University, 2018. URL https://doi.org/10.7916/D8321CTB.

[18] Scott Chamberlain. 'NOAA' Weather Data from R [R package rnoaa version 1.3.2], February 2021. URL https://CRAN.R-project.org/package=rnoaa. Publisher: Comprehensive R Archive Network (CRAN).

[19] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. ISSN 1477-870X. doi: https://doi.org/10.1002/qj.3803. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803. _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3803.

[20] Patrick Laloyaux, Eric de Boisseson, Magdalena Balmaseda, Jean-Raymond Bidlot, Stefan Broennimann, Roberto Buizza, Per Dalhgren, Dick Dee, Leopold Haimberger, Hans Hersbach, Yuki Kosaka, Matthew Martin, Paul Poli, Nick Rayner, Elke Rustemeier, and Dinand Schepers. CERA-20C: A Coupled Reanalysis of the Twentieth Century. *Journal of Advances in Modeling Earth Systems*, 10(5):1172–1195, 2018. ISSN 1942-2466. doi: https://doi.org/10.1029/2018MS001273. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018MS001273. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2018MS001273.

[21] Ronald Gelaro, Will McCarty, Max J. Suárez, Ricardo Todling, Andrea Molod, Lawrence Takacs, Cynthia A. Randles, Anton Darmenov, Michael G. Bosilovich, Rolf Reichle, Krzysztof Wargan, Lawrence Coy, Richard Cullather, Clara Draper, Santha Akella, Virginie Buchard, Austin Conaty, Arlindo M. da Silva, Wei Gu, Gi-Kong Kim, Randal Koster, Robert Lucchesi, Dagmar Merkova, Jon Eric Nielsen, Gary Partyka, Steven Pawson, William Putman, Michele Rienecker, Siegfried D. Schubert, Meta Sienkiewicz, and Bin Zhao. The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *Journal of Climate*, 30(14):5419–5454, July 2017. ISSN 0894-8755, 1520-0442. doi: 10.1175/JCLI-D-16-0758.1. URL https://journals.ametsoc.org/view/journals/clim/30/14/jcli-d-16-0758.1.xml. Publisher: American Meteorological Society Section: Journal of Climate.

[22] D. P. Dee, S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. C. M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haimberger, S. B. Healy, H. Hersbach, E. V. Hólm, L. Isaksen, P. Kållberg, M. Köhler, M. Matricardi, A. P. McNally, B. M. Monge-Sanz, J.-J. Morcrette, B.-K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J.-N. Thépaut, and F. Vitart. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137(656):553–597, 2011. ISSN 1477-870X. doi: https://doi.org/10.1002/qj.828. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.828. _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.828.

[23] Andreas Sterl. On the (In)Homogeneity of Reanalysis Products. *Journal of Climate*, 17(19):3866–3873, October 2004. ISSN 0894-8755, 1520-0442. doi: 10.1175/1520-0442(2004)017¡3866:OTIORP¿2.0.CO;2. URL https://journals.ametsoc.org/view/journals/clim/17/19/1520-0442$_2$004$_0$17$_3$866$_o$*tiorp*$_2$.0.*co*$_2$.*xml* $American Meteorological Society Section : Journal of Climate$.

[24] Katherine Z. Rinaldi, Jacqueline A. Dowling, Tyler H. Ruggles, Ken Caldeira, and Nathan S. Lewis. Wind and Solar Resource Droughts in California Highlight the Benefits of Long-Term Storage and Integration with the Western Interconnect. *Environmental Science & Technology*, 55 (9):6214–6226, May 2021. ISSN 0013-936X. doi: 10.1021/acs.est.0c07848. URL https://doi.org/10.1021/acs.est.0c07848. Publisher: American Chemical Society.

[25] Electric Reliability Council of Texas. About ERCOT, 2021. URL http://www.ercot.com/about.

[26] Electric Reliability Council of Texas. Impact of increased wind resrouces in the ERCOT region. Technical report, June 2020. URL http://www.ercot.com/content/wcm/lists/200196/Wind$_O$*ne*$_P$*ager*$_J$*une*$_2$020.*pdf*.

[27] Andrew Kumler, Ignacio Losada Carreño, Michael T. Craig, Bri-
Mathias Hodge, Wesley Cole, and Carlo Brancucci. Inter-annual
variability of wind and solar electricity generation and capacity val-
ues in Texas. *Environmental Research Letters*, 14(4):044032, April
2019. ISSN 1748-9326. doi: 10.1088/1748-9326/aaf935. URL
https://doi.org/10.1088/1748-9326/aaf935. Publisher: IOP Publish-
ing.

[28] Upmanu Lall and Ashish Sharma. A Nearest Neighbor
Bootstrap For Resampling Hydrologic Time Series. *Wa-
ter Resources Research*, 32(3):679–693, 1996. ISSN 1944-
7973. doi: https://doi.org/10.1029/95WR02966. URL
https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/95WR02966.
_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/95WR02966.

[29] Balaji Rajagopalan and Upmanu Lall. A k-nearest-neighbor sim-
ulator for daily precipitation and other weather variables. *Wa-
ter Resources Research*, 35(10):3089–3101, 1999. ISSN 1944-
7973. doi: https://doi.org/10.1029/1999WR900028. URL
https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/1999WR900028.
_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/1999WR900028.

[30] James Prairie, Balaji Rajagopalan, Upmanu Lall, and Terrance
Fulp. A stochastic nonparametric technique for space-time disag-
gregation of streamflows. *Water Resources Research*, 43(3), 2007.
ISSN 1944-7973. doi: https://doi.org/10.1029/2005WR004721. URL
https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005WR004721.
_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2005WR004721.

[31] James Prairie, Kenneth Nowak, Balaji Rajagopalan, Upmanu
Lall, and Terrance Fulp. A stochastic nonparametric approach
for streamflow generation combining observational and paleore-
constructed data. *Water Resources Research*, 44(6), 2008. ISSN
1944-7973. doi: https://doi.org/10.1029/2007WR006684. URL
https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2007WR006684.
_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2007WR006684.

[32] Francisco Assis Souza Filho and Upmanu Lall. Seasonal to interannual
ensemble streamflow forecasts for Ceara, Brazil: Applications of a mul-
tivariate, semiparametric algorithm. *Water Resources Research*, 39(11),
2003. ISSN 1944-7973. doi: https://doi.org/10.1029/2002WR001373. URL
https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2002WR001373.
_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2002WR001373.

[33] Catherine Loader. Local Regression, Likelihood and Density Es-
timation [R package locfit version 1.5-9.4], March 2020. URL
https://CRAN.R-project.org/package=locfit. Publisher: Comprehen-
sive R Archive Network (CRAN).

[34] Gary D. Tasker. Effective record length for the T-year event. *Journal of Hydrology*, 64(1):39–47, July 1983. ISSN 0022-1694. doi: 10.1016/0022-1694(83)90059-8. URL https://www.sciencedirect.com/science/article/pii/0022169483900598.

[35] Robert Link, Thomas B. Wild, Abigail C. Snyder, Mohamad I. Hejazi, and Chris R. Vernon. 100 years of data is not enough to establish reliable drought thresholds. *Journal of Hydrology X*, 7:100052, April 2020. ISSN 2589-9155. doi: 10.1016/j.hydroa.2020.100052. URL https://www.sciencedirect.com/science/article/pii/S2589915520300031.

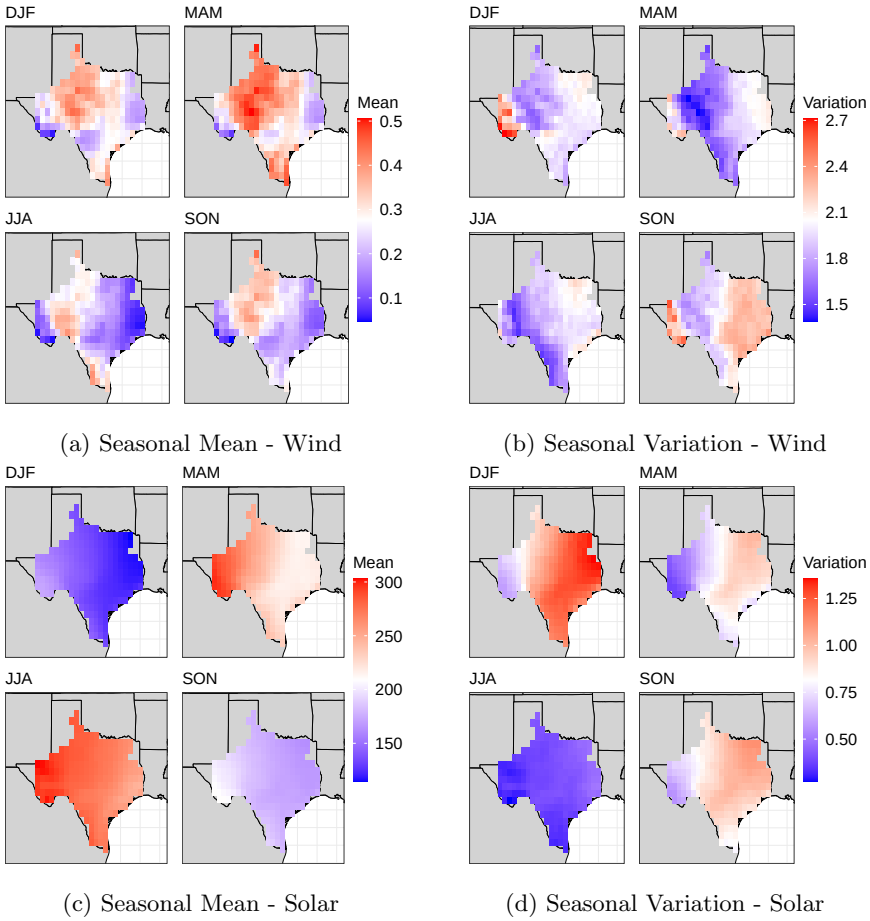
[36] Patricia J. Levi, Simon Davidsson Kurland, Michael Carbajales-Dale, John P. Weyant, Adam R. Brandt, and Sally M. Benson. Macro-Energy Systems: Toward a New Discipline. *Joule*, 3(10):2282–2286, October 2019. ISSN 2542-4351. doi: 10.1016/j.joule.2019.07.017. URL https://www.sciencedirect.com/science/article/pii/S2542435119303617.

[37] Micah S. Ziegler, Joshua M. Mueller, Gonçalo D. Pereira, Juhyun Song, Marco Ferrara, Yet-Ming Chiang, and Jessika E. Trancik. Storage Requirements and Costs of Shaping Renewable Energy Toward Grid Decarbonization. *Joule*, 3(9):2134–2153, September 2019. ISSN 2542-4351. doi: 10.1016/j.joule.2019.06.012. URL https://www.sciencedirect.com/science/article/pii/S2542435119303009.

[38] Leonard Göke and Mario Kendziorski. The adequacy of time-series reduction for renewable energy systems. *arXiv:2101.06221 [econ, q-fin]*, January 2021. URL http://arxiv.org/abs/2101.06221. arXiv: 2101.06221.

[39] Xi Lu and Michael B. McElroy. Chapter 4 - Global Potential for Wind-Generated Electricity. In Trevor M. Letcher, editor, *Wind Energy Engineering*, pages 51–73. Academic Press, January 2017. ISBN 978-0-12-809451-8. doi: 10.1016/B978-0-12-809451-8.00004-7. URL https://www.sciencedirect.com/science/article/pii/B9780128094518000047.

Supplementary Materials



(a) Seasonal Mean - Wind

(b) Seasonal Variation - Wind

(c) Seasonal Mean - Solar

(d) Seasonal Variation - Solar

Figure S1: Seasonal mean and variation in daily wind capacity factors and downward surface solar radiation (W/m-sq) across the Texas Interconnection. (a) Mean daily wind capacity factors by season. (b) Variation in daily wind capacity factors by season. (c) Mean downward surface solar radiation by season. (d) Variation in downward surface solar radiation by season. The seasonal variation is computed as the difference between the 90^{th} and 10^{th} percentile divided by the mean for each grid point for each season. The sub-plots are arranged as follows :- top left - Dec-Jan-Feb (DJF), top right - Mar-Apr-May (MAM), bottom left - Jun-July-Aug (JJA), bottom right - Sept-Oct-Nov (SON).

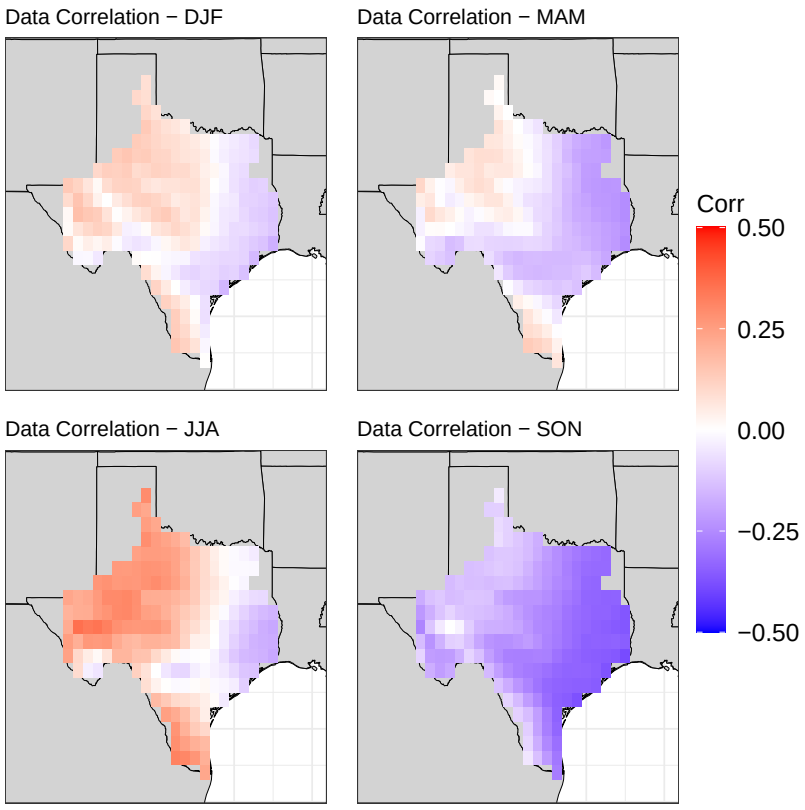
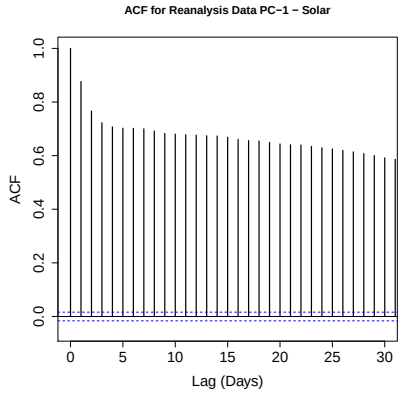
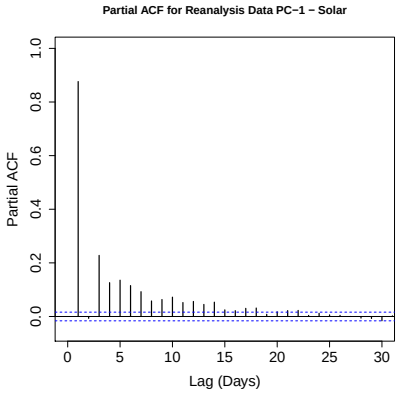


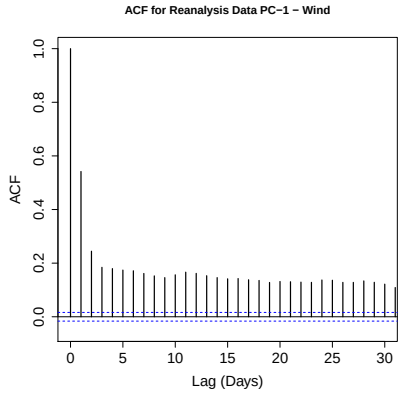
Figure S2: Seasonal correlation between daily wind capacity factors and downward surface solar radiation in the ERA-5 renealysis dataset at each grid point. (top-left) Dec-Jan-Feb (DJF). (top-right) Mar-Apr-May (MAM). (bottom-left) Jun-Jul-Aug (JJA). (bottom-right) Sep-Oct-Nov (SON). The correlations are computed using Pearson's method.



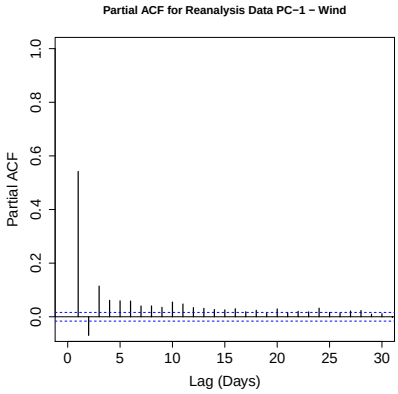
(a) Solar PC1 - ACF



(b) Solar PC1 - Partial ACF



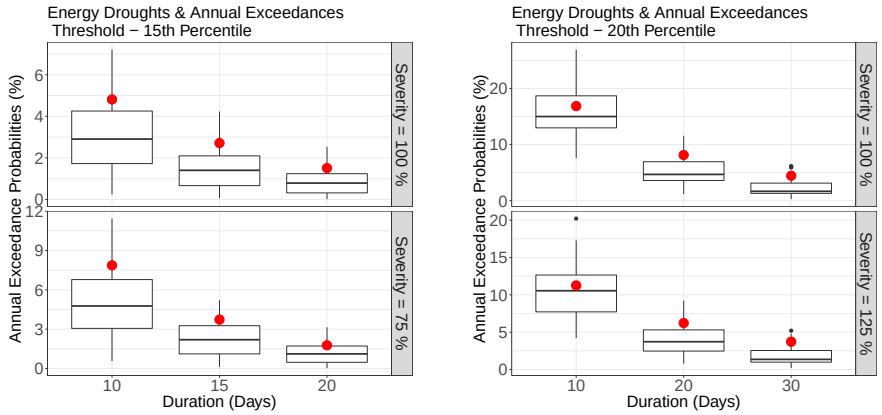
(c) Wind PC1 - ACF



(d) Wind PC1 - Partial ACF

Figure S3: The auto-correlation (ACF) and partial auto-correlation (PACF) for PC-1 of the reanalysis data. (a) Solar PC-1 ACF (b) Solar PC-1 partial ACF (c) Wind PC-1 ACF (d) Wind PC-1 partial ACF. The fractional variance explained by PC-1 for solar and wind fields is 80 % and 63 % respectively. The blue dashed line denotes the significance level for the record length of 14610 days (40 yrs).

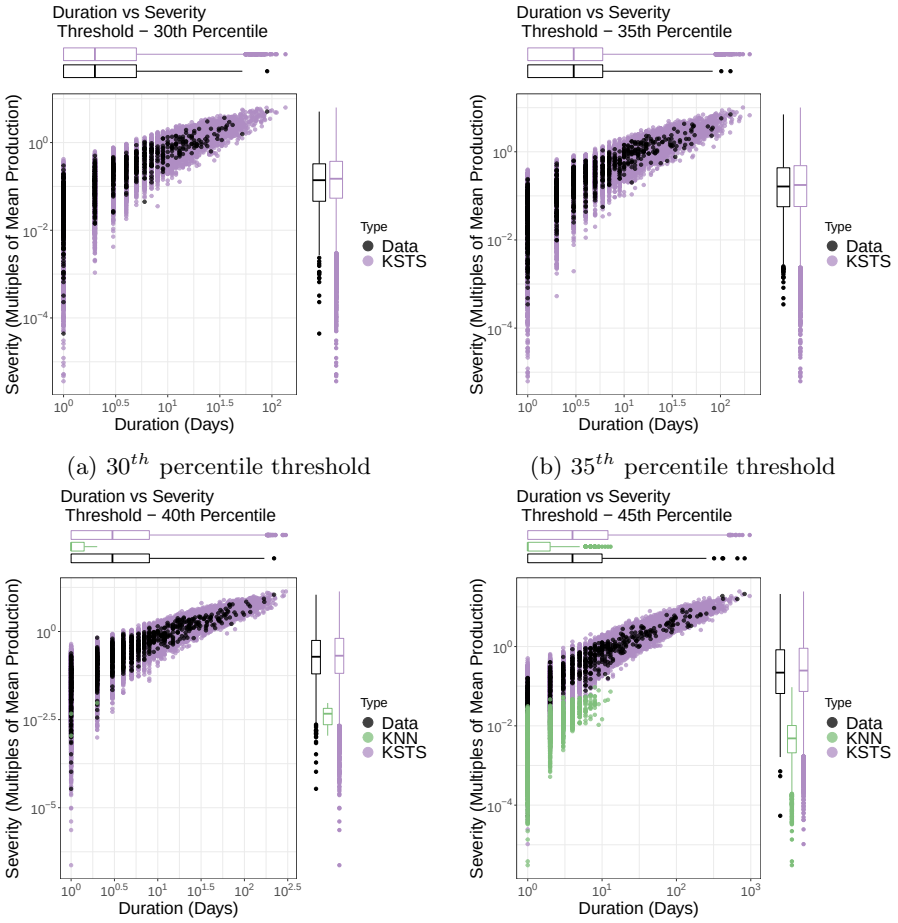
Annual Exceedance for Threshold-Duration-Severity



(a) 15th percentile threshold (b) 20th percentile threshold

Figure S4: Probability of joint annual exceedances for energy droughts given a duration and severity with threshold values of (a) 15th percentile, and (b) 20th percentile. The red dot denotes the exceedance probability calculated from the reanalysis data. The boxplots denote the uncertainty in the 48 generated simulations using KSTS. The duration is in days and the severity is denoted in terms of percentage of the mean historical reanalysis value. For each box-plot, the thick black horizontal line across the box denotes the median of the annual exceedance probabilities from the simulations and the edges of the box denote the 25th and 75th percentiles, and the lower and upper extents of vertical lines outside the box denote the 5th and 95th percentiles.

Severity vs Duration plots for different thresholds



(a) 30^{th} percentile threshold

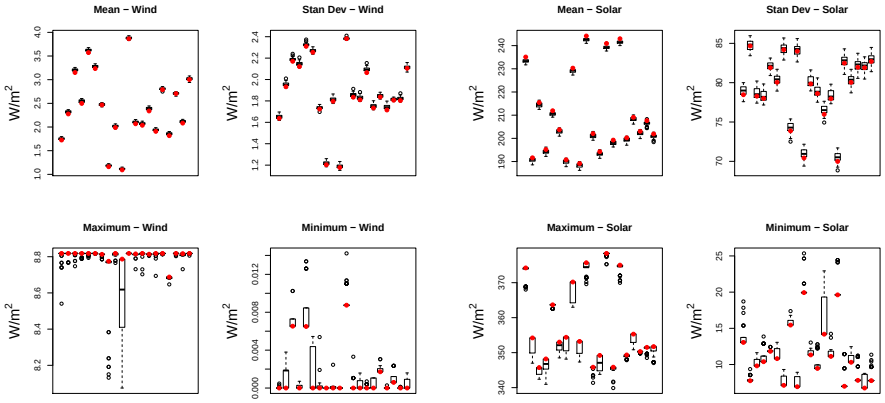
(b) 35^{th} percentile threshold

(c) 40^{th} percentile threshold

(d) 45^{th} percentile threshold

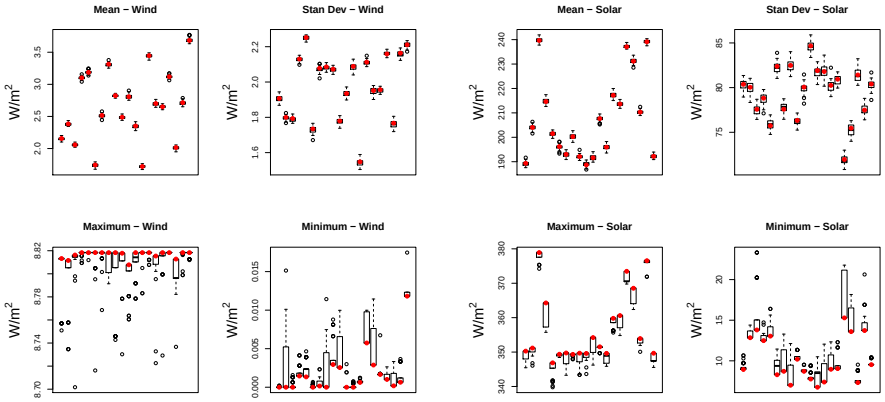
Figure S5: Severity versus duration for all energy droughts with marginal distributions (in boxplots for both variables) for the data (black), KSTS (purple) and KNN (green) simulations using threshold values of (a) 30th percentile, (b) 35th percentile, (c) 40th percentile, and (d) 45th percentile. For the lower thresholds, the KNN simulations do not have occurrence of energy droughts and hence do not appear in the plots.

KSTS and KNN Simulations - Individual Site Characteristics



(a) KSTS - Wind

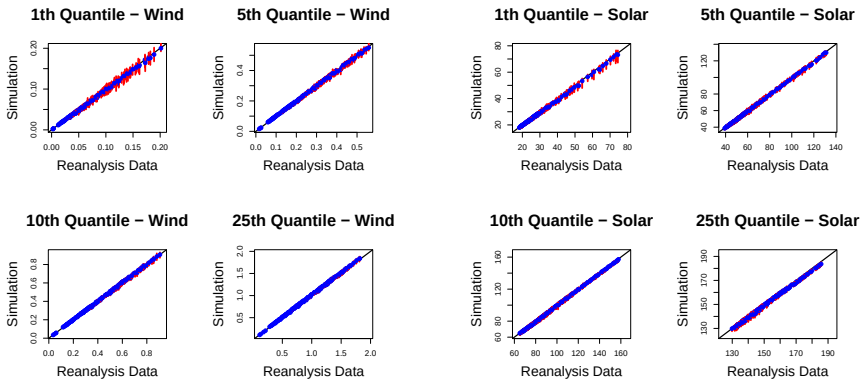
(b) KSTS - Solar



(c) KNN - Wind

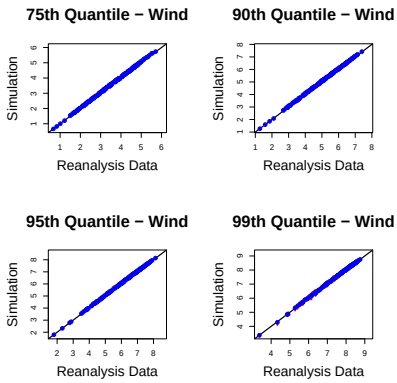
(d) KNN - Solar

Figure S6: Simulation skill assessments for individual sites in the wind and solar fields for both KSTS and KNN simulations. (a) KSTS wind. (b) KSTS solar. (c) KNN wind. (d) KNN solar. For each sub-plot, we show the mean (top-left), the standard deviation (top-right), the minimum (bottom-left) and the maximum (bottom-right). Red dots denote the reanalysis data value and box-plots denote spread among the 48 simulations. Each subplot includes results for 20 randomly selected grid-points out of the 216 total grids.

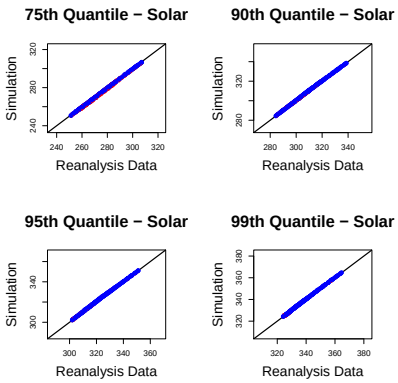


(a) KSTS - Wind Lower Quantiles

(b) KSTS - Solar Lower Quantiles

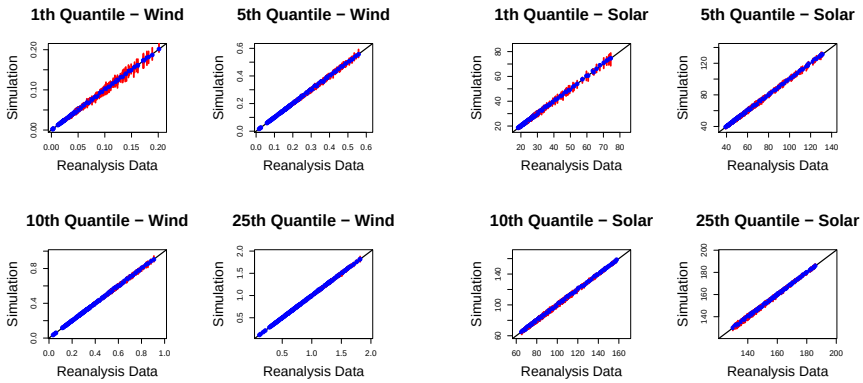


(c) KSTS - Wind Upper Quantiles



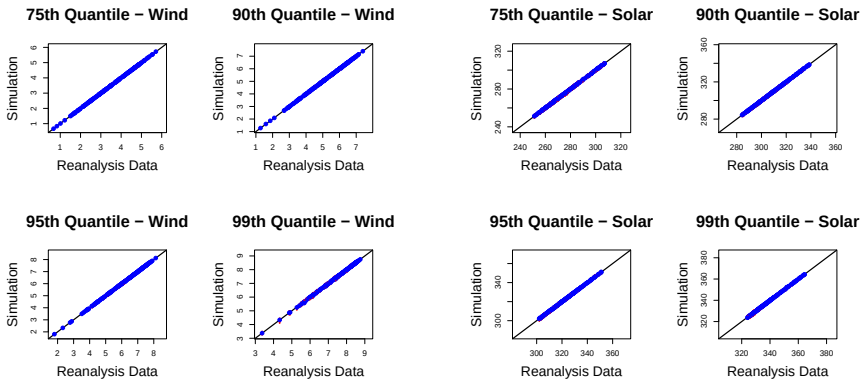
(d) KSTS - Solar Upper Quantiles

Figure S7: Simulation vs reanalysis data quantiles plots for the 1st, 5th, 10th, 25th, 75th, 90th, 95th and 99th percentiles for KSTS simulations. The plots denote the quantiles for all 216 grid points in the wind and solar fields. The red lines denote the mid 90th (5th-95th) percentile range and the blue dots denote the median value in the simulation spread.



(a) KNN - Wind Lower Quantiles

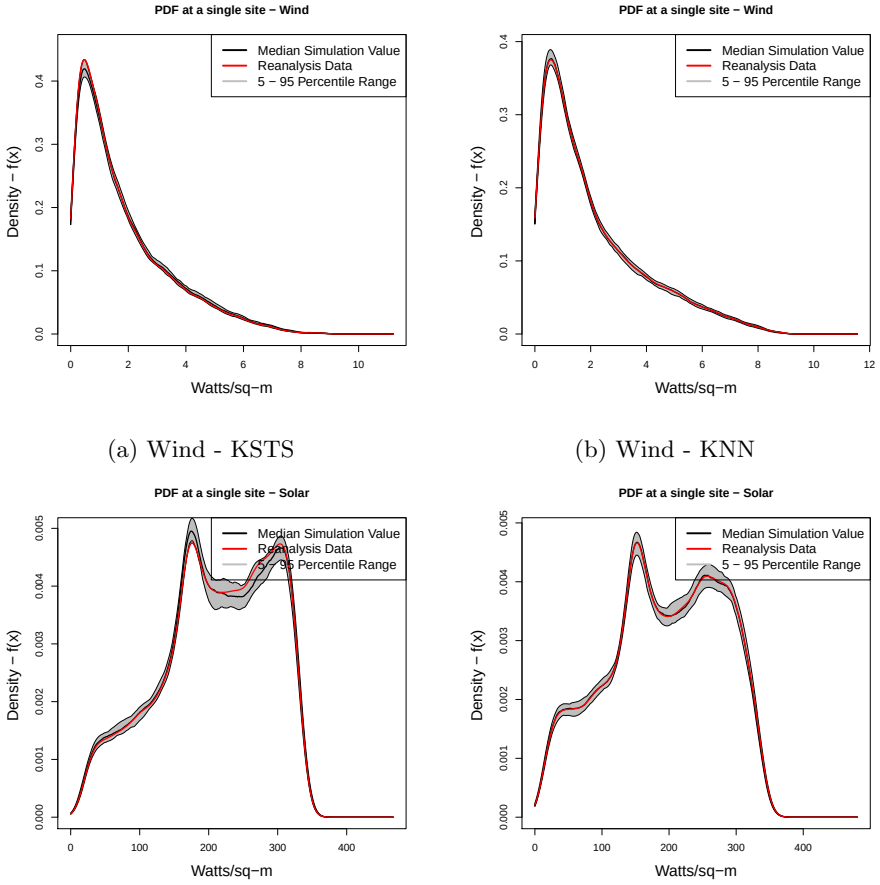
(b) KNN - Solar Lower Quantiles



(c) KNN - Wind Upper Quantiles

(d) KNN - Solar Upper Quantiles

Figure S8: Simulation vs reanalysis data quantiles plots for the 1st, 5th, 10th, 25th, 75th, 90th, 95th and 99th percentiles for KNN simulations. The plots denote the quantiles for all 216 grid points in the wind and solar fields. The red lines denote the mid 90th (5th-95th) percentile range and the blue dots denote the median value in the simulation spread.



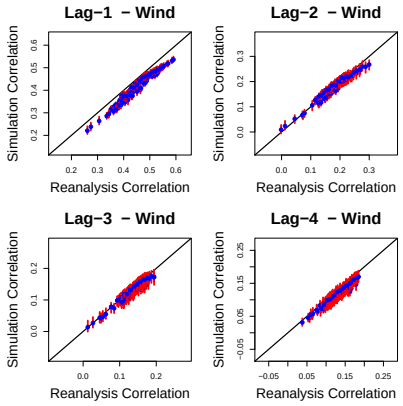
(a) Wind - KSTS

(b) Wind - KNN

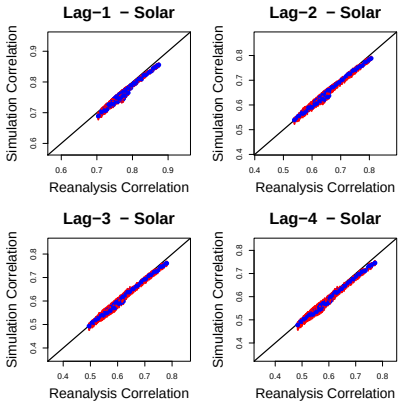
(c) Solar - KSTS

(d) Solar - KNN

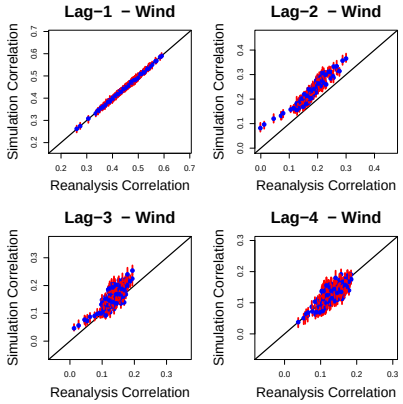
Figure S9: Kernel density estimate / Probability density function (PDF) plots for a single randomly selected grid for wind and solar. The red line denotes the reanalysis data probability density function for the selected site and the black line denotes the median simulation density. The grey region is the mid 90th (5th-95th) percentile range of the simulation spread. The grid that is shown in this figure is selected at random separately for KSTS and KNN. (a) Wind KSTS simulation. (b) Wind KNN simulation. (c) Solar KSTS simulation. (d) Solar KNN simulation.



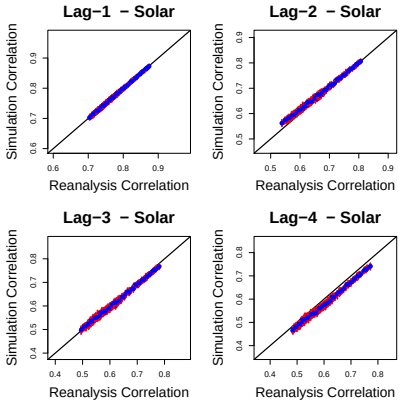
(a) KSTS - Wind



(b) KSTS - Solar

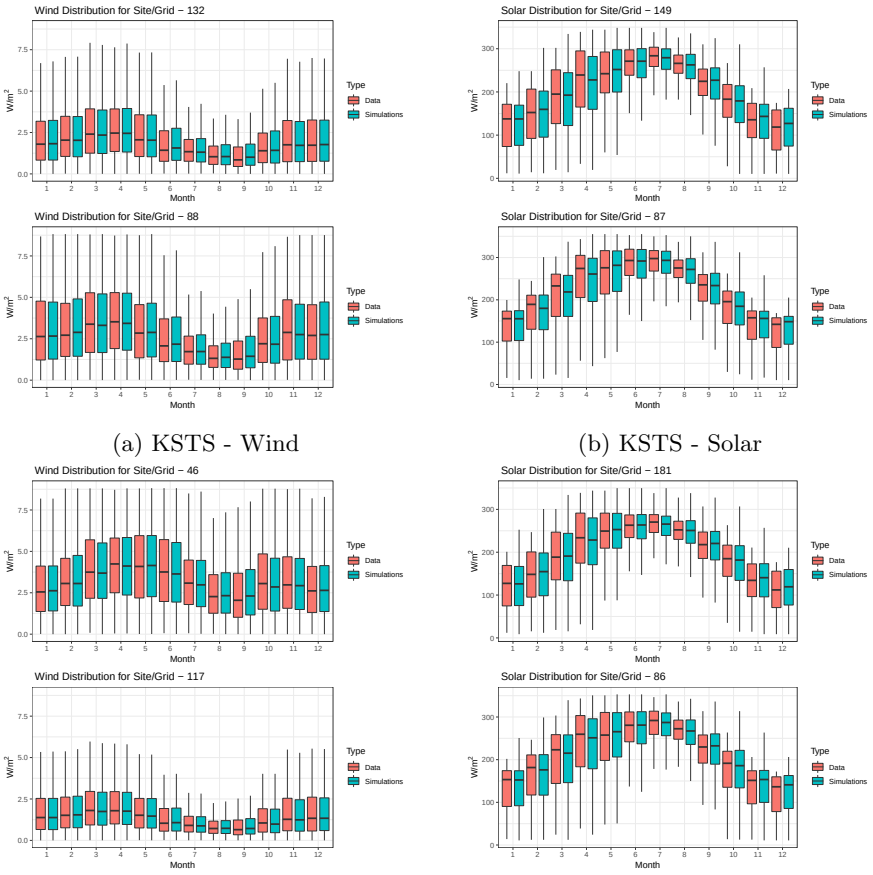


(c) KNN - Wind



(d) KNN - Solar

Figure S10: Simulation vs reanalysis data auto-correlation plots for lag 1,2,3 and 4 for all grid points. (a) Wind KSTS simulations. (b) Solar KSTS simulations. (c) Wind KNN simulations. (d) Solar KNN simulations. The red lines denote the mid 90th (5th-95th) percentile range and the blue dots denote the median value in the simulation spread.



(a) KSTS - Wind

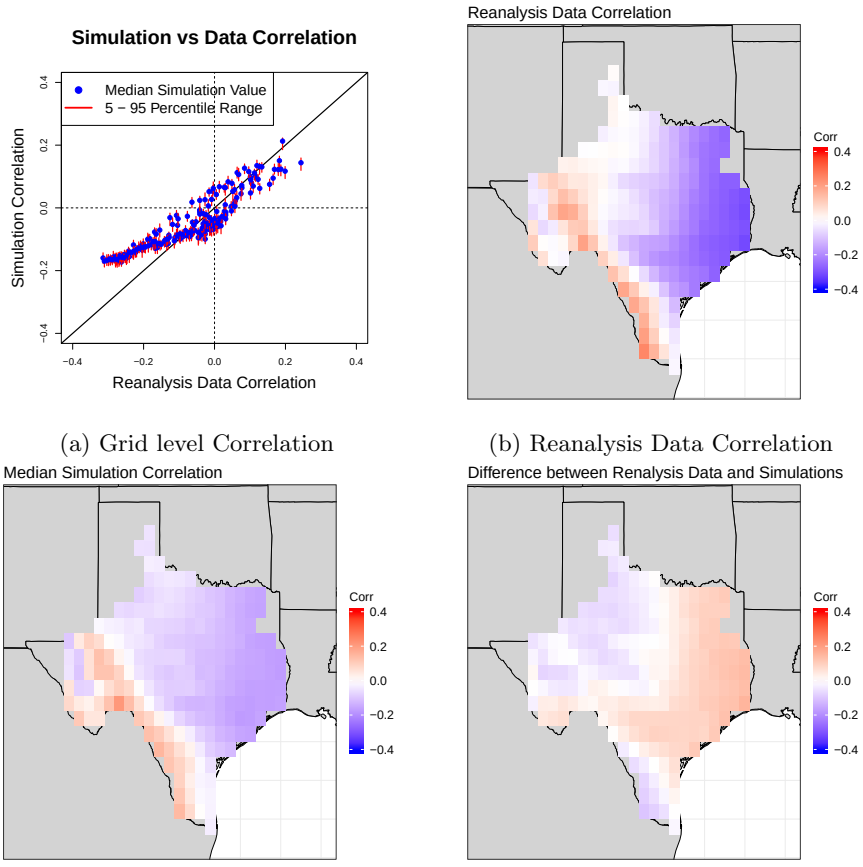
(b) KSTS - Solar

(c) KNN - Wind

(d) KNN - Solar

Figure S11: Seasonality / Monthly distribution of the reanalysis data and simulations. The red and green boxplots denote the reanalysis data and simulations respectively. (a) Wind KSTS simulations. (b) Solar KSTS simulations. (c) Wind KNN simulations. (d) Solar KNN simulations. Two grids are randomly selected for wind and solar. The grids are selected at random separately for KSTS and KNN. Months are numbered in accordance with the Gregorian calendar

Cross-Field Dependence



(a) Grid level Correlation

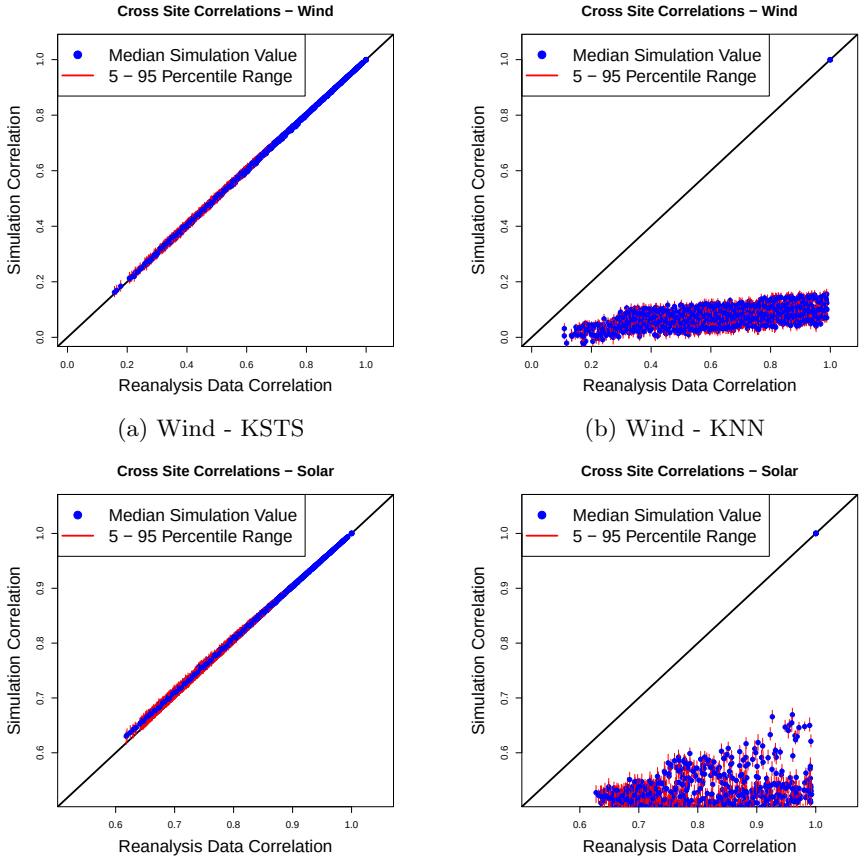
(b) Reanalysis Data Correlation

(c) KNN Simulations Correlation

(d) Difference [(b) - (c)]

Figure S12: Pearson correlation between wind and solar at each grid point based on simultaneous simulations of wind and solar using KNN. (a) Simulation correlation vs reanalysis data correlation between wind and solar where the red lines denote the mid 90th (5th-95th) percentile range and the blue dots denote the median value in the simulation spread generated using the KNN Method. (b) Map of the grid-wise correlations in the reanalysis data record. (c) Map of the grid-wise median simulation correlations using KNN. (d) Map of the difference between (b) and (c).

Individual Field (Wind and Solar) Spatial Correlations



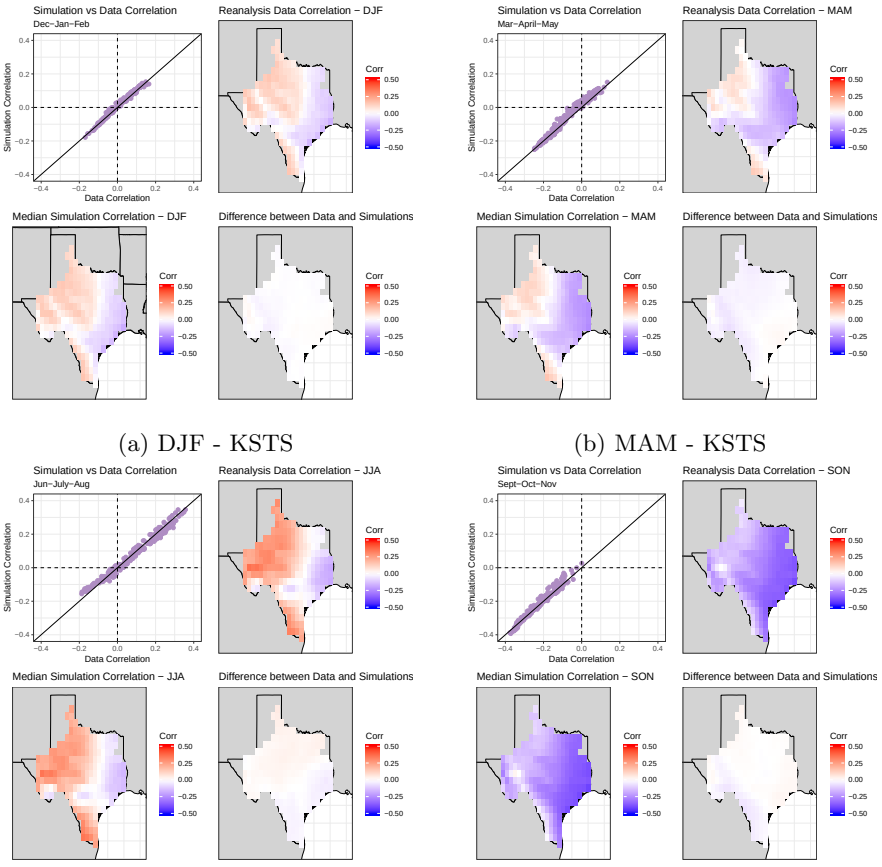
(a) Wind - KSTS

(b) Wind - KNN

(c) Solar - KSTS

(d) Solar - KNN

Figure S13: Simulation vs reanalysis data cross site correlation plots for individual fields. (a) Wind KSTS. (b) Wind KNN. (c) Solar KSTS (d) Solar KNN. 40 random grids out of 216 are selected and the 40x40 cross correlation values are computed and plotted instead of the entire 216 x 216 correlation values. The correlations are computed using Pearson's method.



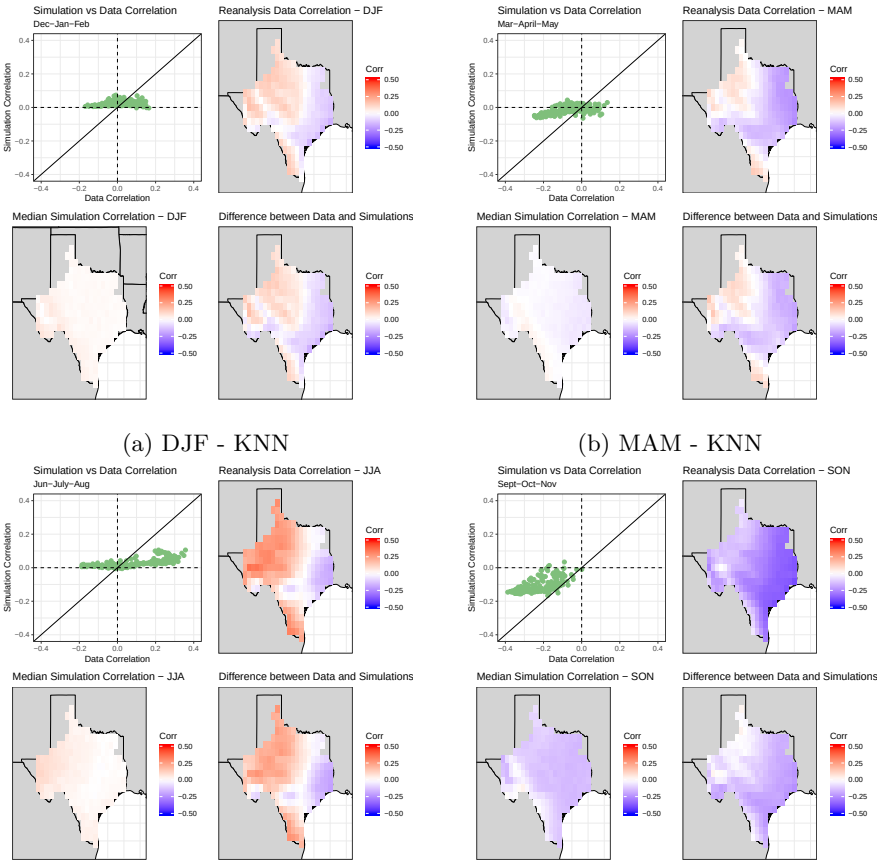
(a) DJF - KSTS

(b) MAM - KSTS

(c) JJA - KSTS

(d) SON - KSTS

Figure S14: Seasonal correlation between wind and solar at each grid point for KSTS simulations. (a) Dec-Jan-Feb (DJF). (b) Mar-Apr-May (MAM). (c) Jun-Jul-Aug (JJA). (d) Sep-Oct-Nov (SON). For each subplot: (top-left) - median simulation vs reanalysis data correlation between wind and solar. (bottom-left) - Plot of the median simulation correlations. (top-right) - Plot of the reanalysis data correlations. (bottom-right) - Plot of the difference between data and median simulations correlations. The correlations are computed using Pearson's method.



(a) DJF - KNN

(b) MAM - KNN

(c) JJA - KNN

(d) SON - KNN

Figure S15: Seasonal correlation between wind and solar at each grid point for KSTS simulations. (a) Dec-Jan-Feb (DJF). (b) Mar-Apr-May (MAM). (c) Jun-Jul-Aug (JJA). (d) Sep-Oct-Nov (SON). For each subplot: (top-left) - median simulation vs reanalysis data correlation between wind and solar. (bottom-left) - Plot of the median simulation correlations. (top-right) - Plot of the reanalysis data correlations. (bottom-right) - Plot of the difference between data and median simulations correlations. The correlations are computed using Pearson's method.

Data

The Electric Reliability Council of Texas (ERCOT - Fig. S16), functions as an Independent System Operator and the balancing authority for the Texas Interconnection and manages about 90% of state's electric load. ERCOT covers about 75% of the land area in Texas[1] .

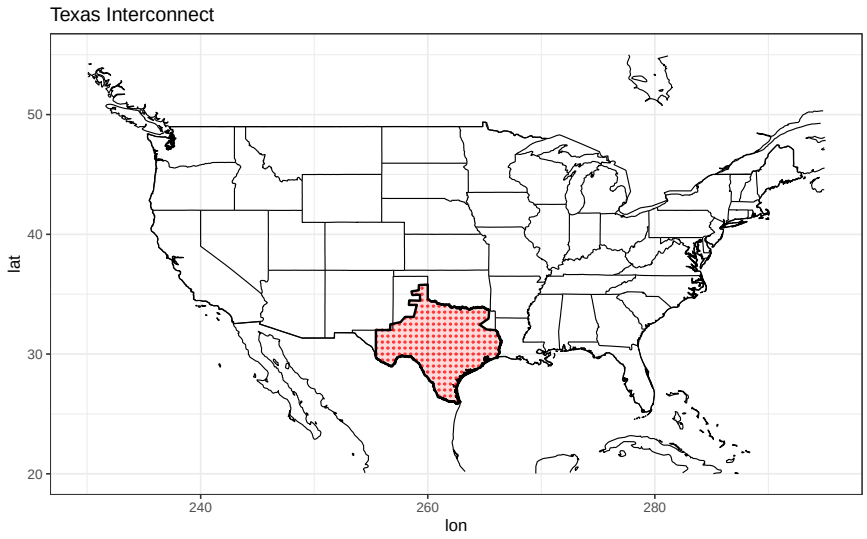


Figure S16: Texas Interconnection / ERCOT domain plot - The red shaded region denotes the area administered. The red dots (216) are the locations of the grid points (0.5° lat × 0.5° lon) from the ERA-5 reanalysis dataset.

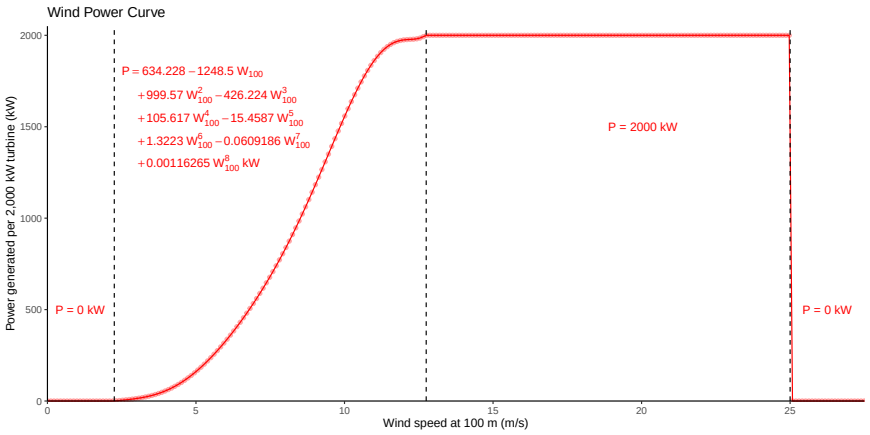


Figure S17: Wind power curve for a V90-2.0MW Vestas turbine.

KSTS Description

The algorithm first models the temporal variability at each site and for each variable field. A state space $D_{i,t}$ is defined through an embedding of the time series with a delay parameter ϕ and an embedding dimension m, where i is the index for the site/variable combination. Following Lall and Sharma [2], we consider that the conditional density $f(x_{i,t+1}|D_{i,t})$ is defined through the $x_{i,t'+1}$ corresponding to the time indices t' associated with the k-nearest neighbors (knn) of $D_{i,t}$ in the historical data set. The kernel function p_j associated with the jth nearest neighbor is proportional to the rank j of the neighbor [2]. The sequential drawing from the knn successors at each time step using the specified kernel leads to a simulation of $x_{i,t}$ in a time series dependence structure. Since the procedure leads to a resampling of the historical data, the algorithm can be considered to be a bootstrap which preserves the time dependence in serial data.

Given a state space $D_{i,t}$ at site i and time t, the k-nearest neighbor algorithm is used to identify a set of time indices $\tau_{i,t}$ that correspond to the time instances corresponding to the nearest neighbors of site i at time t. An example, for site, with an embedding $D_{i,t}$, defined as (x_t, x_{t-1}) which takes values (10,9.8) at time t=100 days, the nearest neighbors would be identified as the closest historical vectors to (10,9.8) using the data for this site.

The time instances at which these neighbors occur in the historical time series are then recorded in the order of nearest neighbors in $\tau_{\mathrm{i,t}}$. The kernel $p_{i,j}$ then associates a probability proportional to 1/j for the jth element (time instance of the jth nearest neighbor of $D_{i,t}$) in $\tau_{i,t}$, for the first k neighbors and 0 else. For space-time neighbors across all sites, i.e. to address spatial dependence, we now identify appropriate k-nearest neighbors by finding the time indices in the historical data that have the highest likelihood of being selected across all sites given their associated resampling probabilities.

Define $T_{i,t}$ as a matrix such that the rows and columns are pointers for sites and unique time indices from the historical data respectively. The columns record the resampling probabilities associated with the time indices of the k-nearest neighbors for all sites at time t. Other measures of similarity of the spatial neighbors of the temporal process could also be considered. The similarity matrix S_t is the sum of the resampling probabilities associated with each unique time index across all sites. The curtailment of the similarity matrix S_t is carried out by selection of the time indices which correspond to the k highest resampling probability values in S_t, now designated as the k-nearest neighbor candidates for the entire spatial field. The full spatial field of the simulation for the next time step is resampled after re-scaling probability values (such that they add to 1) of the curtailed similarity matrix S_t.

Hyper-parameters of the Algorithm

Resampling Kernel Weight Function

The resampling kernel p_j selected for the simulator is the one proposed

by Lall and Sharma[2]. This resampling kernel decreases monotonically with increase in distance, with the bandwidth and kernel shape varying with the local sampling density. Overall, the kernel is adaptive to the dimensionality of the state space, with implicit dependence through the distance calculations. Further, the resampling weights need to be computed only once and stored, which significantly reduces computation time.

$$p_j = \frac{1/j}{\sum_{j=1}^{k} 1/j}$$

Other options for the kernel include a uniform kernel ($p_j = 1/k$) or a power kernel based on the distances of the k neighbors. Refer to Lall and Sharma[2] for further details on the behavior of the kernel in the boundary region, for bounded data and comparison to a uniform kernel.

Number of Neighbors (k) and Model Order (m)

One method to choose model hyper-parameters involves criterion that minimize the mean squared error in forecast. The generalized cross validation (GCV) score was suggested to select k and m[2]. The selected number of nearest neighbors k and the order of the feature vector m are the ones which minimize the GCV score, which is given by

$$GCV = \frac{\sum_{i=1}^{n} e_i^2/n}{\left(1 - \frac{1}{\sum_{j=1}^{k} 1/j}\right)^2}$$

where, e_i is the forecast error at point i for the model fit to all the data without it and n is the total number of points. The selection of these parameters by GCV is most appropriate if the model errors e_i are normally distributed or if the variables are transformed such that model errors are normally distributed. Non-normality of the errors may lead to sub-optimal choice of k and m with respect to it's conditional mean and variance.

Another method to select the model lags in the feature vector is the false nearest neighbors algorithm which determines the embedding dimension for the process [3]. Finally, an ad-hoc choice of $k = n^{0.5}$ is suggested across the knn literature, with low sensitivity around this value. Further suggestions include trying various combinations of k and m followed by visual examinations of the simulation attributes and data [2].

Scaling Weights (w)

The simplest selection choice for the weights w, which weigh the euclidean distance of the selected lags m is to be specified *a priori* with uniform values. The weights can also be selected such that they minimize the forecast error in least squares sense when used in a knn regression setup [4]. An alternate adaptive strategy is to compute scaling weights (w) for the knn resampling approach such that they are the regression coefficients of the selected external predictors from a parametric regression model [5].

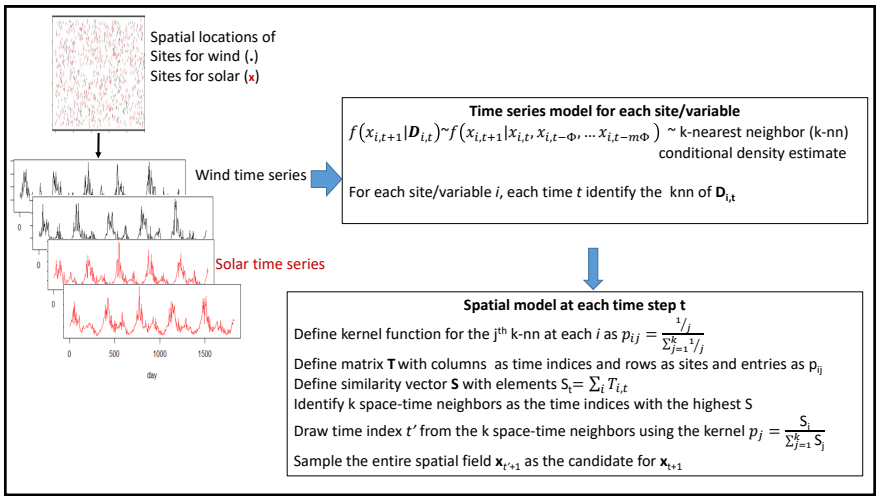


Figure S18: Schematic representation of the k-Nearest Neighbor Simulator (KSTS) with application to wind and solar fields.

Supplementary Materials - References

1. Electric Reliability Council of Texas. About ERCOT, 2021. URL http://www.ercot.com/about

2. Upmanu Lall and Ashish Sharma. A Nearest Neighbor Bootstrap for Resampling Hydrologic Time Series. Water Resources Research 32.3 (1996): 679-693.

3. Kennel, Matthew B., Reggie Brown, and Henry DI Abarbanel. "Determining embedding dimension for phase-space reconstruction using a geometrical construction." Physical review A 45.6 (1992): 3403.

4. Yakowitz, S., and M. Karlsson. "Nearest neighbor methods for time series, with application to rainfall/runoff prediction." Advances in the statistical sciences: Stochastic hydrology. Springer, Dordrecht, 1987. 149-160.

5. Souza Filho, Francisco Assis, and Upmanu Lall. "Seasonal to interannual ensemble streamflow forecasts for Ceara, Brazil: Applications of a multivariate, semiparametric algorithm." Water Resources Research 39.11 (2003).