A k-Nearest Neighbor Space-Time Simulator with applications to large-scale wind and solar power modelling

Yash Amonkar[1,2,5,*], David J Farnham[3,4], 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]Now at ClimateAi, San Francisco, USA
[5]Lead Author
[*]Correspondence: yva2000@columbia.edu

Abstract

Summary

We develop and present a k-Nearest Neighbor Space-Time Simulator that accounts for the spatiotemporal dependence in high dimensional hydroclimatic fields (e.g. wind and solar) and can simulate synthetic realizations of arbitrary length. We illustrate how this statistical simulation tool can be used in the context of regional power system planning under a scenario of high reliance on wind and solar generation and when long historical records of wind and solar power generation potential are not available. We show how our simulation model can be used 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 this estimation of supply side shortages for the Texas Interconnection.

Introduction

Many countries and individual states within the United 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 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 cost-effectively 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 one option 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 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 along with estimates of the demand profile. 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 interannual 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 United States[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. 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[23] 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 stochastic properties of the duration

and severity of wind and solar 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 capable of assessing the probability of the severity and duration of aggregate supply side energy shortages across a 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 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 [24] as a target example to explore the historical record and to demonstrate the performance of our algorithm. While LDS considerations motivate the use of daily data on potential wind and solar resource, the model can be used to simulate any spatio-temporal data, including climate or environmental fields. An implicit assumption in the choice of time scale for the energy application is that chemical batteries help smooth out the sub-daily time-scale shortages [11]. The daily wind and solar capacity factors are computed at the hourly timescale and then averaged over the entire day. (See the Experimental Procedure section for additional methodological details).

Over a large region (e.g., the Texas Interconnection), the wind and solar generation assets are likely to be spatially distributed throughout the region [25]. 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 wind and solar capacity factors from reanalysis data for our example application to Texas are illustrated in Figures 1 and S1.

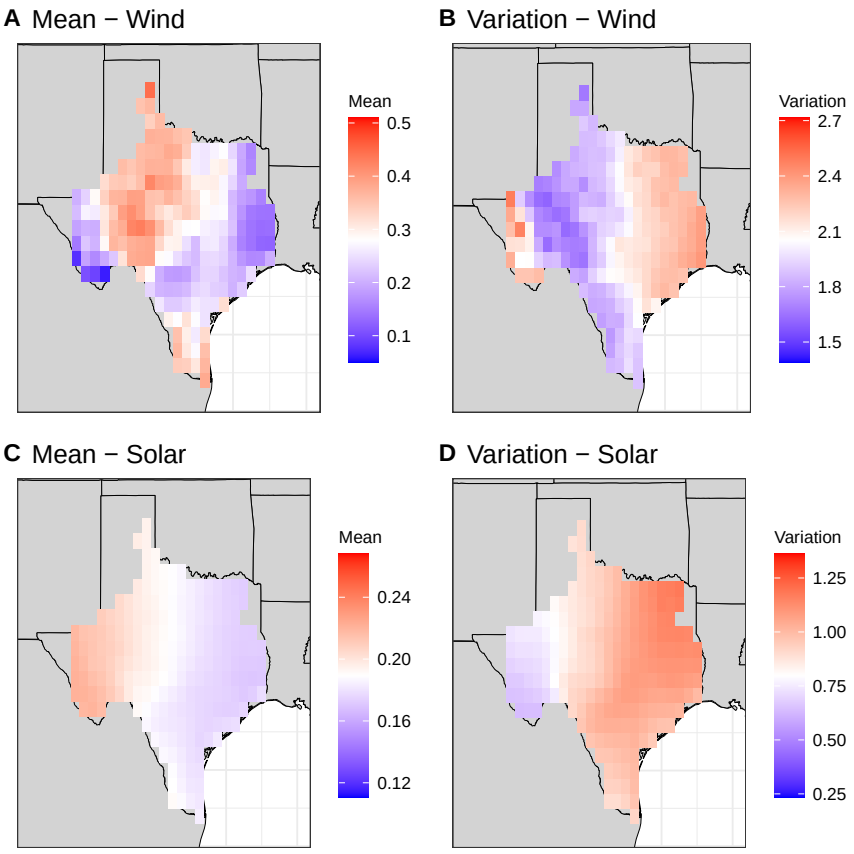

Figure 1: Mean and variation in daily wind and solar capacity factors across the Texas Interconnection. (A) Mean daily wind capacity factors. (B) Variation in daily wind capacity factors. (C) Mean daily solar capacity factors. (D) Variation in daily solar capacity factors. The variation is computed as the difference between the 90^{th} and 10^{th} percentile divided by the mean for each grid.

Daily wind and solar fields often exhibit variability that changes by location and time of year and needs to be accounted for in an analysis of potential renewable energy droughts or LDS system sizing[26]. As seen in Figure 1 and S1, wind and solar along the Gulf of Mexico and the land-area adjoining Louisiana are regions with relatively low generation potential but with relatively 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. Daily wind capacity factors are generally highest during spring while variation is highest during fall and lowest in summer and spring (Figure S1). The mean daily solar capacity factors and 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) (Figure S1).

The seasonal cross-field spatial correlation between wind and solar 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).

k-Nearest Neighbor Algorithm

We now discuss the historical development and associated literature of the k-nearest neighbor algorithm. The k-nearest neighbor algorithm, a non-parametric method, has been used in traditional problems of classification and regression across fields[27]. The algorithm serves as a simple first choice in most cases where the underlying data distribution characteristics are not known a priori. The algorithm has its origins in discriminant analysis[28]. Yakowitz[29] and Karlsson[30,31] first developed and utilized a nearest neighbor regression methodology in a time series context for use in rainfall-runoff forecasting. They showed that the method, when used in a time series context has attractive convergence properties, being asymptotically optimal for finite data sets.

Lall and Sharma[32] developed a nearest neighbor algorithm based simulator/resampling scheme for time series data, with applications for hydrological time series. The resampling scheme, referred to as nearest neighbor bootstrap in their work, preserves the dependence in a probabilistic sense, without making any assumptions about the distributional form and marginal densities of the underlying process. They also introduced a new resampling kernel to weigh the k successors rather than having uniform weights. They make the assumption that in the space of the nearest neighbors, the local density of the future resampled value can be approximated as a Poisson process. The kernel has the attractive properties of bandwidth and shape adapting to local sampling density changes along with the dimension of the feature vector; and decreases monotonically with distance of the neighbors.

Another study[33] introduced a k-nearest neighbors simulator for multivariate time series data following on earlier work[32], which was a univariate simulator. The multivariate knn simulation model, a non-parametric approximation of a multivariate lag-1 Markov process, was shown to simulate daily sequences of solar radiation, wind speed, maximum and minimum temperature and precipitation at a single site. The model simulations preserve the marginal densities of the variables along with the cross-correlations and spell lengths, crucial indices for climatological variables. Nowak et al[34] developed a disaggregation method which generates multi-site daily flows from a simulated annual value via the knn resampling scheme. While the above described algorithms were all non-parametric, Filho and Lall[35] developed a multivariate semi-parametric approach for multi-site streamflow forecasting conditional on external climate predictors using the knn resampling scheme. The key innovation in their work

included an adaptive strategy to compute scaling weights for the knn resampling approach, which are the regression coefficients of the external predictors from a parametric regression model. These scaling weights ensure that the relative importance of the predictor vectors is accounted for in the resampling scheme.

The structure of the new k-Nearest Neighbors Space Time Simulator (KSTS) algorithm that is presented here is 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 successors of the k-nearest neighbor 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 and solar capacity factors, have non-Gaussian skewed distributions and are bounded. The probabilistic sampling using k-nearest neighbors provides an effective approach to sampling from such a non-parametric 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 method generalizes to a higher dimensional space, the k-Nearest neighbor algorithm[32] used for univariate or low dimensional multivariate simulations of non-Gaussian and nonlinear dependence that has been used extensively for other climate variables[33,34,35,36,37].

We apply our new KSTS algorithm to assess the severity, duration, and frequency of long duration storage needs associated with 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 potential including the spatial correlation within and across the two fields and the temporal autocorrelation at each site. This study uses the issue of LDS sizing and requirement from the supply-side perspective of renewable energy producers to illustrate the utility of the proposed spatio-temporal field simulation algorithm. We recognize that both supply and demand (load) are needed to assess energy storage needs on a grid level, with the net load (demand – renewables) being of particular interest. As such, our application of the simulation algorithm should be viewed as illustrative but should not be seen as an estimation of the actual LDS needs on the Texas Interconnection. To properly contextualize our application, we consider a target firm energy contract from renewables across the domain, and

compute the drought statistics with reference to that contract. 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 underestimation of the regional LDS probabilities. The relative utility and performance of other statistical models relative to the KSTS model and relevant literature review is discussed in the Experimental Procedures section.

For the application presented, we use the 71-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 71-year simulations (or equivalently a 7100 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, which is defined as a shortage in aggregate energy produced across the grid relative to a target threshold. The probabilities of drought severity and duration can then be assessed from this derived set of events. If multiple simulations of 71 years are generated, then one can also get an estimate of the uncertainty associated with the probability of severity-duration given 71 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. While we make inferences on LDS statistics from a purely supply-side perspective, the primary purpose of the example is showing the application of the novel KSTS algorithm to a high dimensional problem of interest.

Results

We present an evaluation of the severity, duration, and 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 installed capacity allocation of wind and solar generation across all grid points is considered, with wind and solar having mean capacity factors of 0.28 and 0.19 respectively. For both types of simulations, we generated 48 realizations of 71 years of daily wind and solar data at each of the 216 sites. In an actual use case, a stochastic optimization model would use wind and solar capacities reflective of the Texas Interconnection along with the demand to allocate resources and estimate the 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.

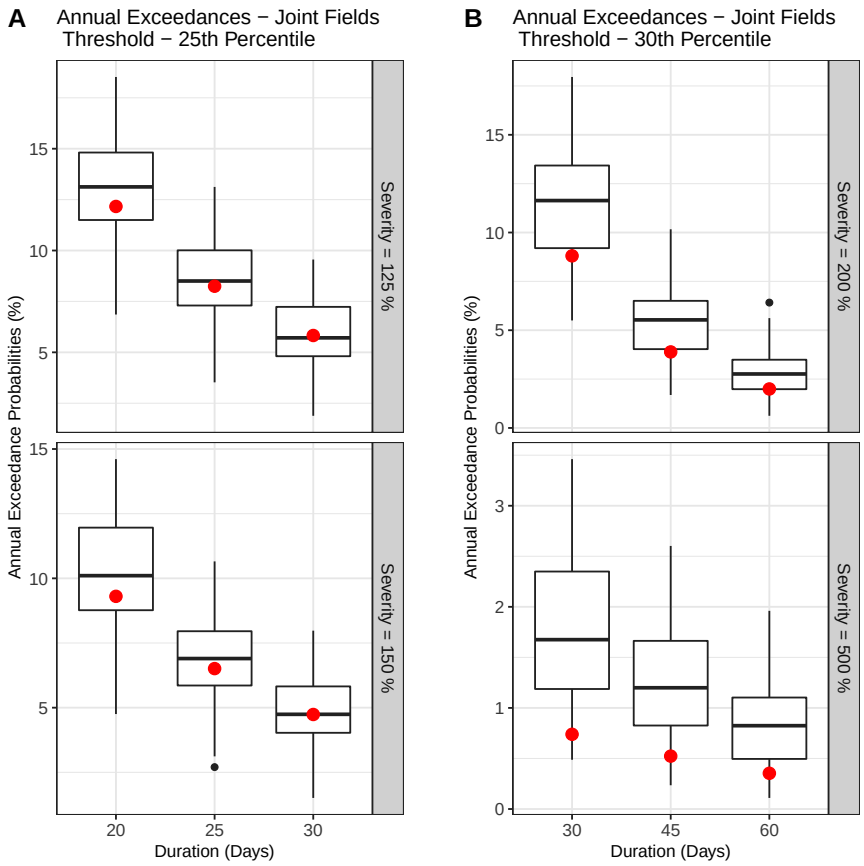

Figure 2: Probability of annual exceedances for energy droughts given a duration and severity with threshold values of (A) 25th percentile and (B) 30th percentile. The red dot denotes the exceedance probability calculated from the reanalysis data. The boxplots denote the uncertainty in the 48 generated 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 boxplot, 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 whiskers extend to the furthest point that is within 1.5 times the IQR below the 25th percentile and above the 75th percentile respectively.

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. Examples of such contracts would be where renewable power producers bid in the day ahead market but also buy options from natural gas producers (reliable sources) to hedge their risks in case of lower than anticipated production [38,39]. The forward contract example in our study is essentially a pre-bid power delivery promise (corresponding to the threshold) and the energy droughts are the periods where the producer will not be able to meet their obligations.

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) [40] 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 reanalysis 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 6% with the 5th and 95th percentiles extending from 2% to 8%, demonstrating the limitations of using solely the original 71 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 71 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 [41,42]. The simulations show that these extreme events could occur far more frequently than would be estimated from just the short historical records. In these illustrations, we consider specific thresholds for supply guarantees, 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 and wind-solar individual fields are provided in Figures S4 and S6. 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, and 40th percentile are shown in Figure S5. We see that KSTS is effective for representing the range of energy droughts. Similar boxplot estimates for the KNN algorithm generated simulations are not shown since the simulations show no occurrences of energy droughts at these thresholds.

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 S7). Both simulators are able to reproduce the quantiles (Figure S8, Figure S9), underlying probability distribution (Figure S10), auto-correlation structure (Figure S11), and site-level seasonality (Figure S12). The distribution of the aggregate generation over the full domain, however, is properly reproduced by the KSTS simulator but not by the KNN simulator.

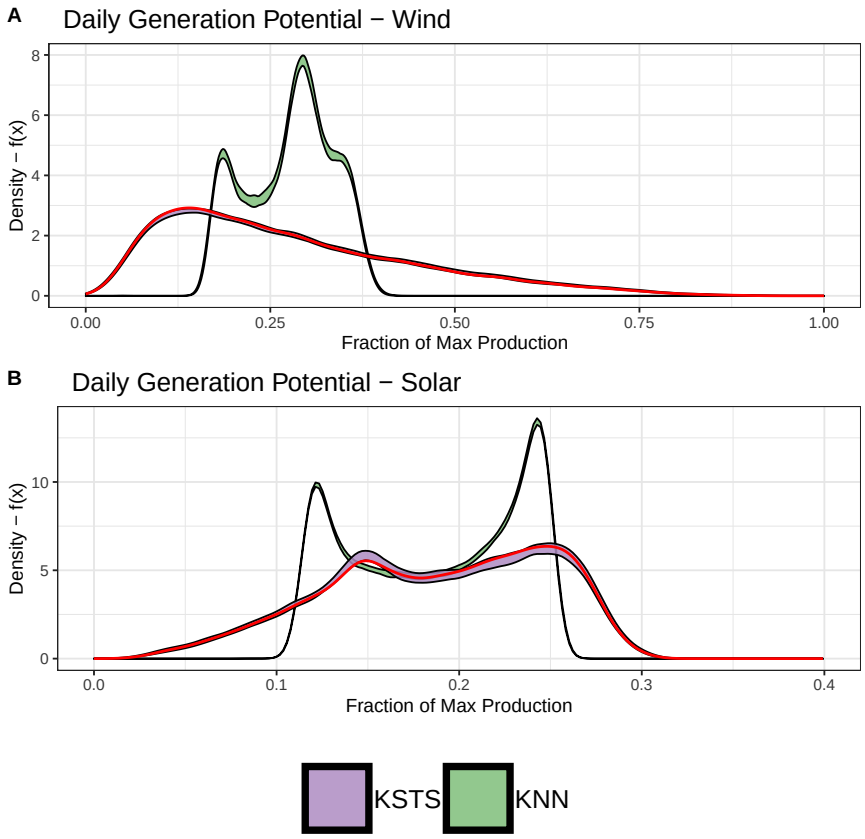

Figure 3: Kernel density estimate / Probability density function (PDF) of the daily aggregated energy production across the Texas Interconnection simulated using KSTS (purple) and KNN (green). The red line denotes the observed data pdf. The purple and green regions show the mid 90th (5th-95th) percentile interval regions from the individual pdfs computed from 48 simulations from each simulator. (A) Wind. (B) Solar.

The kernel density estimate of aggregated daily energy generation potential across the Texas Interconnection is shown in Figure 3 for the historical reanalysis record (red), 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

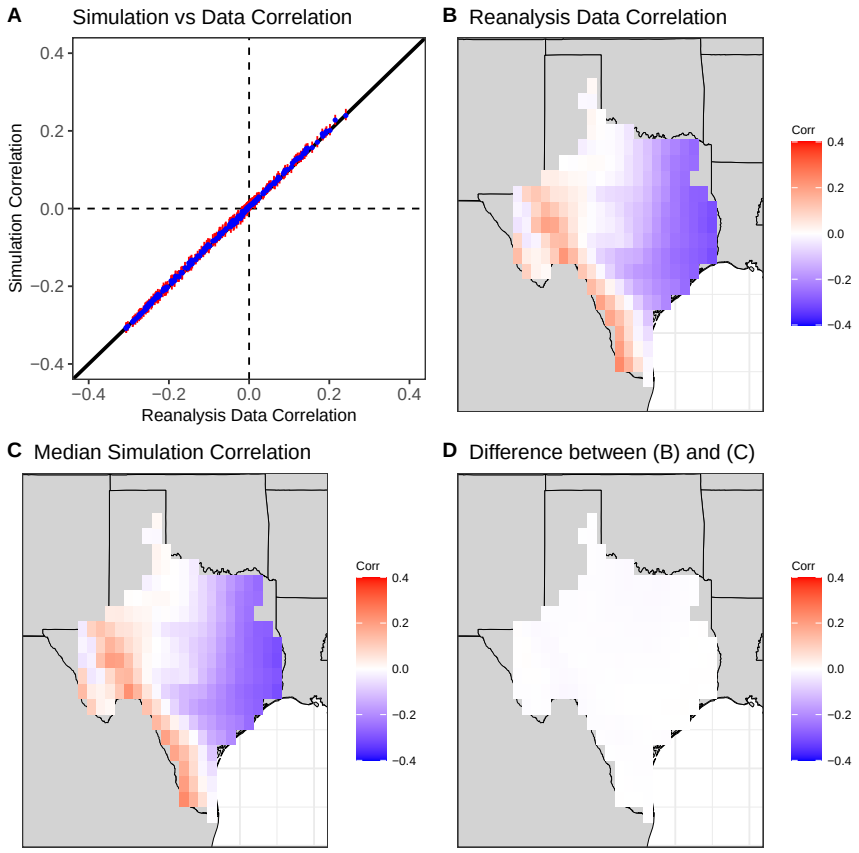

Figure 4: Pearson correlation between wind and solar at each grid point based on simultaneous simulations of wind and solar using KSTS. (A) Simulation correlation vs reanalysis data correlation between wind and solar where the red lines denote the mid 90th (5th-95th) percentile range and the blue dots 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 wind and solar across ERCOT is well reproduced by the KSTS simulations, which 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 S13). 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 S14). The seasonal variation in the correlation between wind and solar is also well modeled by the KSTS algorithm (Figure S15 and Figure S16).

Discussion

The primary contribution of this paper is the presentation of a k-Nearest Neighbor Space Time Simulator (KSTS) and its application to the joint wind-solar fields across the Texas Interconnection. We demonstrate 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. These resource droughts are analyzed from a purely supply side perspective in this study with demand (load) and installed solar and wind capacity data needed for further detailed analysis. This algorithm seeks to estimate the probability (and associated uncertainty) of the duration and severity of resource droughts integrated over the spatial domain, through simulation. 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[43]. As the scale is increased, there is evidence that LDS is an effective and economic component of the design of these regional systems[10,11,44]. However, most of the models developed and applied at these scales are deterministic and use relatively short records with a potential to lead to biased results[45]. 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 critical it is to model spatial dependence when assessing the characteristics of energy shortages on a grid.

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. These simulations show large uncertainties in the annual exceedance probabilities for the severity, duration, and threshold combinations considered, as well as potentially higher exceedance probabilities than computed from the 71 year data record for the more extreme severity, duration, and threshold 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, and frequency using an extended sample.

Applicability to Other Problems

The KSTS algorithm could be used for any spatio-temporal simulation problem where the preservation of spatial dependence is of interest, and the temporal dynamics are modeled through a Markovian process or through a time domain embedding, as illustrated in the methodology section. Typical examples would be any weather or climate fields where maintaining the space and time consistency across multiple variables is of interest. An example that is similar to the current context is a copula based model that was developed to model risk of national livestock losses in Mongolia using spatially distributed livestock loss data over time[46]. Many of the existing space-time simulators were developed in a Markovian framework with random variables considered to be drawn from the Exponential family of distributions.

Extension of the KSTS simulator to other time scales (e.g., hourly) is feasible. An hourly simulator would need to consider the diurnal cycle, in addition to 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 can also be applied to simultaneously modelling multiple streamflow or weather stations across a watershed while preserving the internal dependence structure. Such streamflow data exhibit spatio-temporal correlation patterns due to their position in the river network, altitude and a host of other hydrological variables[34], making application of KSTS attractive.

Limitations and Next Steps

Since KSTS is a hybrid resampling (bootstrap) method, it cannot simulate values not seen in the historical record. This is not a major issue for wind and solar capacity factors, since the lower and upper extremities of the distribution for both wind and solar are recorded in the reanalysis (historical) data enabling KSTS to generate daily simulations that span the entire distribution of both fields.

In the general case of other hydroclimatic variables, extrapolation to values not seen in the historical record is also possible[46]. If a parametric or non-parametric marginal probability distribution is fit to the time series of a variable, with parameters that may vary by season, one could draw observations from that

distribution that are consistent with the k-nearest neighbor value selected for simulation. If the rank (small to big) of the k-nearest neighbor value in the historical data is j, then an estimate of its corresponding cumulative distribution function F(x) is j/(n+1), where n is the sample size[47]. Accounting for uncertainty, one can consider that F(x) lies between (j-0.5)/(n+1) and (j+0.5)/(n+1). For the largest/smallest value on record the intervals would be ((n-0.5)/(n+1), 1) and (0,1.5/(n+1)), respectively. Consequently, if sampling values not seen in the historical record is of interest, one can first sample uniformly from this interval and then sample the corresponding value from the marginal distribution of x. This does not change the basic structure of the KSTS algorithm, but allows values to be simulated from an appropriate probability distribution for each variable considered.

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 in some sub-regions is driven by disparate atmospheric 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 71-yrs (1950-2020) 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 ERA-5 reanalysis variables used are wind speeds at 100 meter altitude and downward surface solar radiation[19]. The spatial grid size of the data is set at 0.5° lat × 0.5° lon and contains 216 grid points across the Texas Interconnection domain (Figure S17). The wind speed and solar radiation at each hour are converted to hourly wind and solar power respectively, and the daily capacity factors are computed as the average across 24 hours.

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 Figure S18). The data are converted to the daily time step by taking the mean of the hourly capacity factors for each day and the dataset spans the 71 years from January 1st, 1950 to December 31st, 2020.

The solar variable is the downward surface solar radiation (W/m^2) and is converted to capacity factor at the hourly level by accounting for the dependence of photovoltaic performance on temperature[48] (Figure S19). We then compute a capacity factor for each day by taking the mean of the 24 hourly values for that day.

Energy Deficits and Drought

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 that day's value 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[40].

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.

Fitting Other Models

We considered and tested other strategies for spatio-temporal simulation with the Texas Interconnection data prior to developing and testing the KSTS algorithm. A brief review of those efforts is presented below. The Autoregressive Integrated Moving Average (ARIMA) model was first fit to sites individually, using the Akaike Information Criterion to select model order[49]. The results, not displayed here, failed to capture the underlying data generating process with significant departures even from the base moments. ARIMA and other similar parametric approaches assume normality of the underlying distributions making them a bad fit to this joint modelling problem where both wind and solar capacity factors are non-normal, bounded and multi-modal distributions (Figure S10). The serial dependence structure of the wind and solar data is also nonlinear and the use of these linear models contributes to biases in the

simulations. Generalizations such as Vector Autoregressive processes and Space-Time autoregressive models suffer from the same problems, in addition to the challenge of fitting a high dimensional covariance matrix.

Another potential class of non-parametric machine learning based models applicable to the current problem are Generative Adversarial Networks (GANs) [50]. GANs have been used for renewable simulations (scenario generation) and do not assume normality of the underlying data [51,52]. However, while GANs can model complex spatial dependencies, they require a large amount of temporal data to fit the model. Our initial efforts at fitting GANs to the current data did not lead to a model that had skillful temporal evolution characteristics, making a direct comparison infeasible.

Finally, different types of Hidden Markov Models (HMM) [53] were also explored to simulate the wind-solar fields. The application of a non-homogeneous HMM with spatial covariance modeled using variograms led to unsatisfactory results due to inadequate representation of the spatio-temporal correlation structure.

Simulator Hyperparameters

For both KSTS and KNN, the seasonality in the data is accounted for by restricting the search of nearest neighbors to a ± 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 71-yrs and 61 days per year, $\sqrt{n}$ is ~ 65, where n is the number of possible candidate neighbors after accounting for the moving window [32]. A lag-1 dependence structure for the state space is assumed. 48 independent simulation realizations, each of the same length as the reanalysis data (71 yrs = 25933 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 [32].

k-Nearest Neighbor Space-Time Simulator (KSTS)

The general structure and a cartoon example application of the KSTS algorithm is illustrated in Figures S20 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 $\quad D_{i,t}$:- (x_{t-1}, x_{t-2}); m = 2

Case 2 $D_{i,t}$:- ($x_{t-\tau}$, $x_{t-2\tau}$, $x_{t-\phi}$, $x_{t-2\phi}$); m = 4, $\tau = 1$, $\phi = 12$
Case 3 $D_{i,t}$:- (x_{t-1}, x_{t-4} x_{t-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 interannual 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 at time t 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 vector 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 vector 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 vector S_t has the same length as the number of unique time indices in the data.

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

The similarity vector 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 next time step of 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.

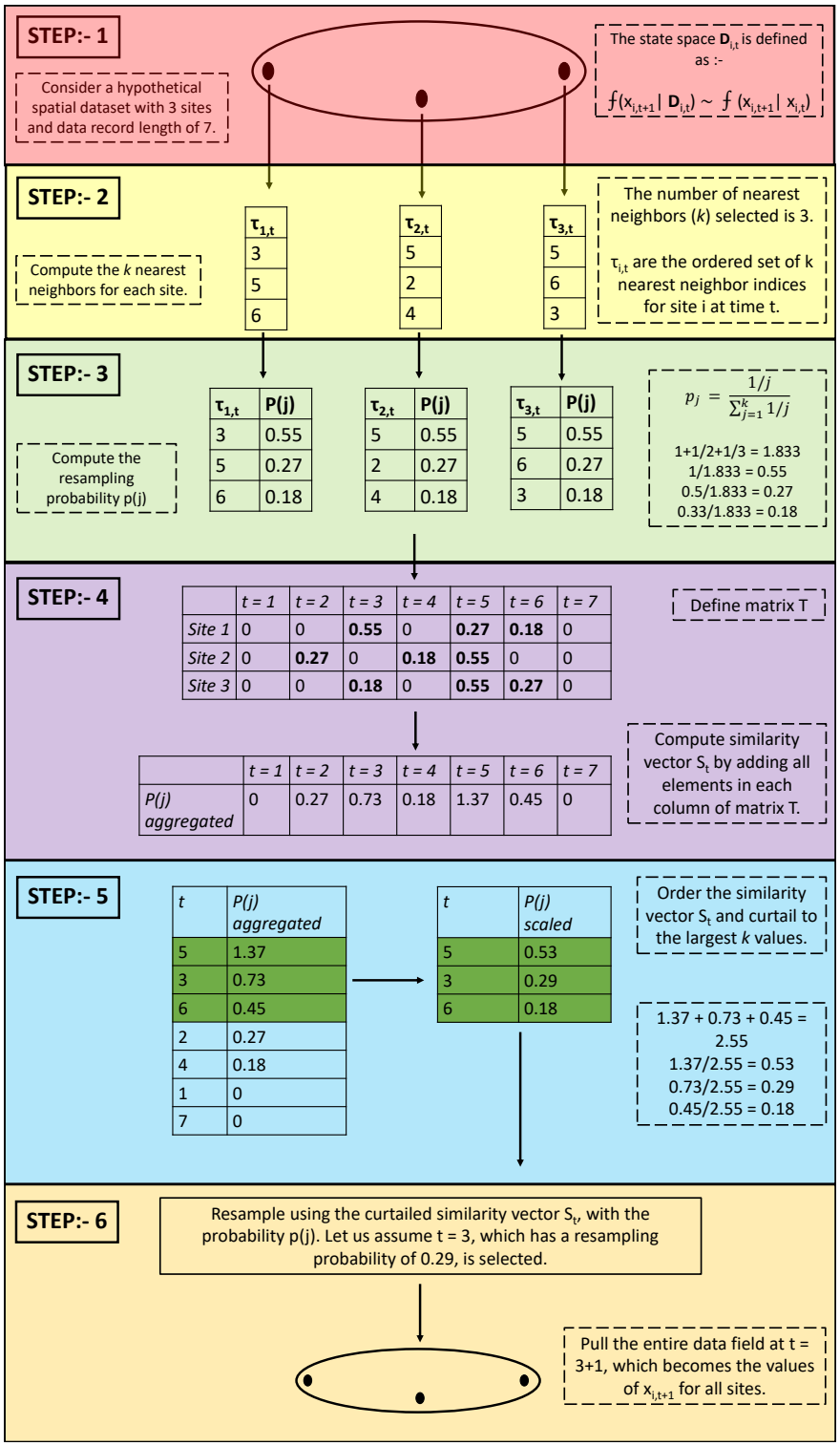

Figure 5: Cartoon example application of the KSTS algorithm to a spatial dataset consisting of 3 grids/sites and data record (time) length of 7.

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. who introduced the algorithm. 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=1588581905912&uri=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, Nouredine 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] 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.

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

[25] 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_One_Pager_June_2020.pdf.

[26] 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 values 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 Publishing.

[27] T. Cover and P. Hart. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1):21–27, January 1967. ISSN 1557-9654. doi: 10.1109/TIT.1967.1053964. Conference Name: IEEE Transactions on Information Theory.

[28] B. W. Silverman and M. C. Jones. E. Fix and J.L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951). International Statistical Review / Revue Internationale de Statistique, 57(3):233–238, 1989. ISSN 0306-7734. doi: 10.2307/1403796. URL https://www.jstor.org/stable/1403796. Publisher: [Wiley, International Statistical Institute (ISI)].

[29] S. Yakowitz. Nearest-Neighbour Methods for Time Series Analysis. Journal of Time Series Analysis, 8(2):235–247, 1987. ISSN 1467-9892. doi: 10.1111/j.1467-9892.1987.tb00435.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9892.1987.tb00435.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9892.1987.tb00435.x.

[30] M. Karlsson and S. Yakowitz. Nearest-neighbor methods for nonparametric rainfall-runoff forecasting. Water Resources Research, 23(7):1300–1308, 1987. ISSN 1944-7973. doi: 10.1029/WR023i007p01300. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/WR023i007p01300. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/WR023i007p01300.

[31] M. Karlsson and S. Yakowitz. Rainfall-runoff forecasting methods, old and new. Stochastic Hydrology and Hydraulics, 1(4):303–318, December 1987. ISSN 1435-151X. doi: 10.1007/BF01543102. URL https://doi.org/10.1007/BF01543102.

[32] Upmanu Lall and Ashish Sharma. A Nearest Neighbor Bootstrap For Resampling Hydrologic Time Series. Water 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.

[33] Balaji Rajagopalan and Upmanu Lall. A k-nearest-neighbor simulator for daily precipitation and other weather variables. Water 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.

[34] Kenneth Nowak, James Prairie, Balaji Rajagopalan, and Upmanu Lall. A nonparametric stochastic approach for multisite disaggregation of annual to daily streamflow. Water Resources Research, 46(8), 2010. ISSN 1944-7973. doi: 10.1029/2009WR008530. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2009WR008530. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2009WR008530.

[35] Francisco Assis Souza Filho and Upmanu Lall. Seasonal to interannual ensemble streamflow forecasts for Ceara, Brazil: Applications of a multivariate, 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.

[36] James Prairie, Balaji Rajagopalan, Upmanu Lall, and Terrance Fulp. A stochastic nonparametric technique for space-time disaggregation 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.

[37] James Prairie, Kenneth Nowak, Balaji Rajagopalan, Upmanu Lall, and Terrance Fulp. A stochastic nonparametric approach for streamflow generation combining observational and paleoreconstructed 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.

[38] Nayara Aguiar, Vijay Gupta, and Pramod P. Khargonekar. A Real Options Market-Based Approach to Increase Penetration of Renewables. IEEE Transactions on Smart Grid, 11(2):1691–1701, March 2020. ISSN 1949-3061. doi: 10.1109/TSG.2019.2942258. Conference Name: IEEE Transactions on Smart Grid.

[39] David D'Achiardi, Nayara Aguiar, Stefanos Baros, Vijay Gupta, and Anuradha M. Annaswamy. Reliability Contracts Between Renewable and Natural Gas Power Producers. IEEE Transactions on Control of Network Systems, 6(3):1075–1085, September 2019. ISSN 2325-5870. doi: 10.1109/TCNS.2019.2919857. Conference Name: IEEE Transactions on Control of Network Systems.

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

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] Upmanu Lall, Naresh Devineni, and Yasir Kaheil. An Empirical, Nonparametric Simulator for Multivariate Random Variables with Differing Marginal Densities and Nonlinear Dependence with Hydroclimatic Applications. Risk Analysis, 36(1):57–73, 2016. ISSN 1539-6924. doi: 10.1111/risa.12432. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/risa.12432. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/risa.12432.

[47] Lasse Makkonen. Bringing Closure to the Plotting Position Controversy. Communications in Statistics - Theory and Methods, 37(3):460–467, January 2008. ISSN 0361-0926. doi: 10.1080/03610920701653094. URL https://doi.org/10.1080/03610920701653094. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/03610920701653094.

[48] Philip E. Bett and Hazel E. Thornton. The climatological relationships between wind and solar energy supply in Britain. Renewable Energy, 87:96–110, March 2016. ISSN 0960-1481. doi: 10.1016/j.renene.2015.10.006. URL https://www.sciencedirect.com/science/article/pii/S0960148115303591.

[49] Rob J. Hyndman and George Athanasopoulos. Forecasting: principles and practice. OTexts, May 2018. ISBN 978-0-9875071-1-2. Google-Books-ID: _bBhDwAAQBAJ.

[50] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. Communications of the ACM, 63(11):139–144, October 2020. ISSN 0001-0782. doi: 10.1145/3422622. URL https://doi.org/10.1145/3422622.

[51] Yize Chen, Yishen Wang, Daniel Kirschen, and Baosen Zhang. Model-Free Renewable Scenario Generation Using Generative Adversarial Networks. IEEE Transactions on Power Systems, 33(3):3265–3275, May 2018. ISSN 1558-0679. doi: 10.1109/TPWRS.2018.2794541. Conference Name: IEEE Transactions on Power Systems.

[52] Yize Chen, Xiyu Wang, and Baosen Zhang. An Unsupervised Deep Learning Approach for Scenario Forecasts. In 2018 Power Systems Computation Conference (PSCC), pages 1–7, June 2018. doi: 10.23919/PSCC.2018.8442500.

[53] L. Rabiner and B. Juang. An introduction to hidden Markov models. IEEE ASSP Magazine, 3(1):4–16, January 1986. ISSN 1558-1284. doi: 10.1109/MASSP.1986.1165342. Conference Name: IEEE ASSP Magazine.

Supplementary Materials

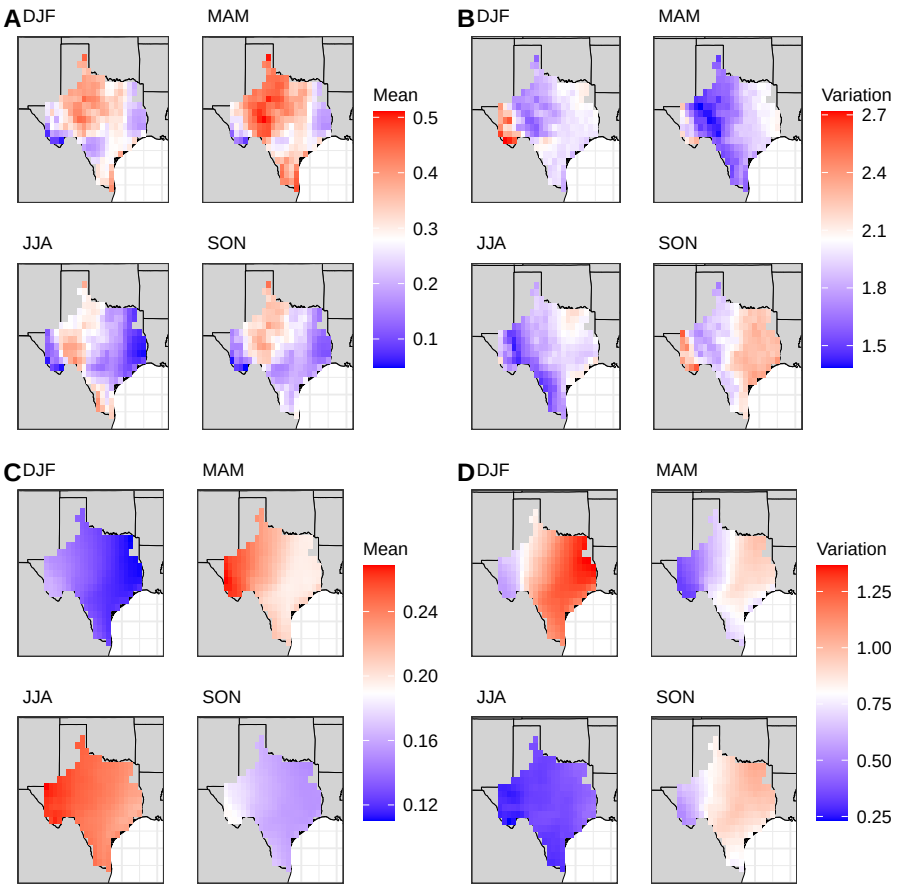

Figure S1: Seasonal mean and variation in daily wind and solar capacity factors across the Texas Interconnection. (A) Mean daily wind capacity factors by season. (B) Variation in daily wind capacity factors by season. (C) Mean daily solar capacity factors by season. (D) Variation in daily solar capacity factors 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).

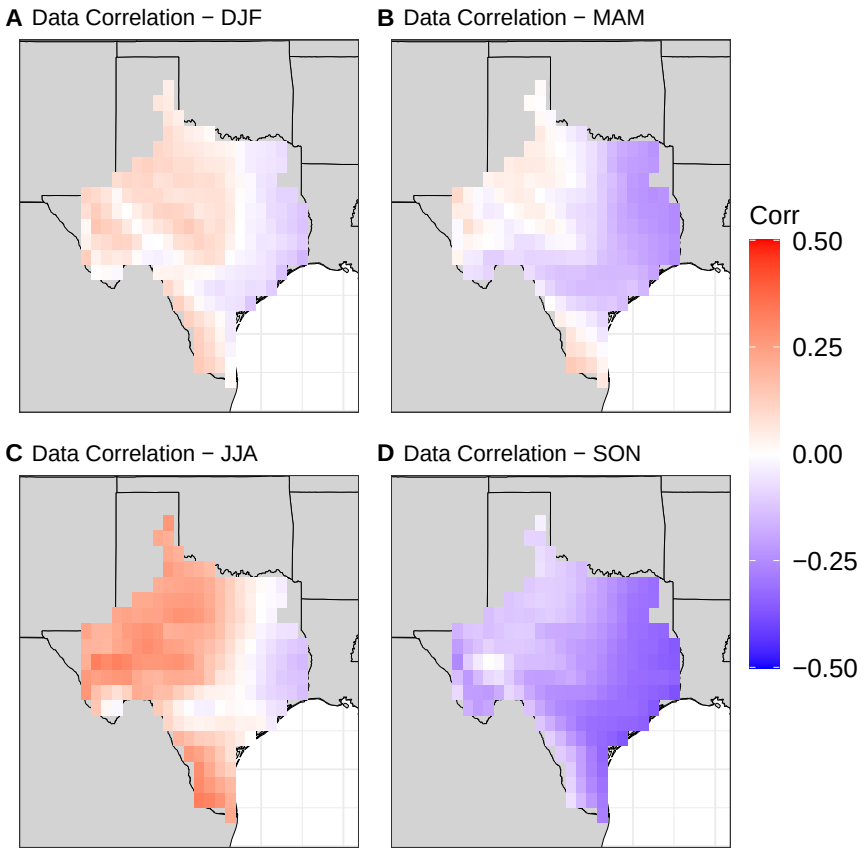

Figure S2: Seasonal correlation between daily wind and solar capacity factors in the ERA-5 reanalysis dataset at each grid point. (A) Dec-Jan-Feb (DJF). (B) Mar-Apr-May (MAM). (C) Jun-Jul-Aug (JJA). (D) Sep-Oct-Nov (SON). The correlations are computed using Pearson's method.

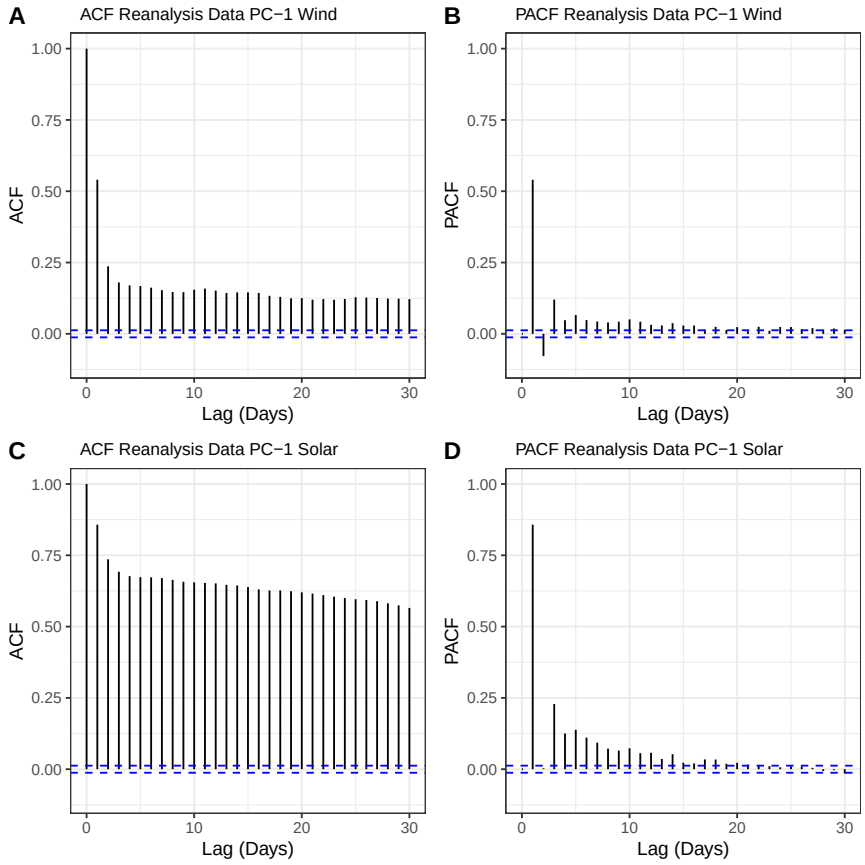

Figure S3: The auto-correlation (ACF) and partial auto-correlation (PACF) of the first Principal Component(PC) of the individual reanalysis data fields. (A) Wind PC-1 ACF. (B) Wind PC-1 PACF. (C) Solar PC-1 ACF. (D) Solar PC-1 PACF The fractional variance explained by PC-1 for solar and wind fields is 79 % and 63 % respectively. The blue dashed line denotes the significance level for the record length of 25933 days (71 yrs).

Annual Exceedance for Threshold-Duration-Severity

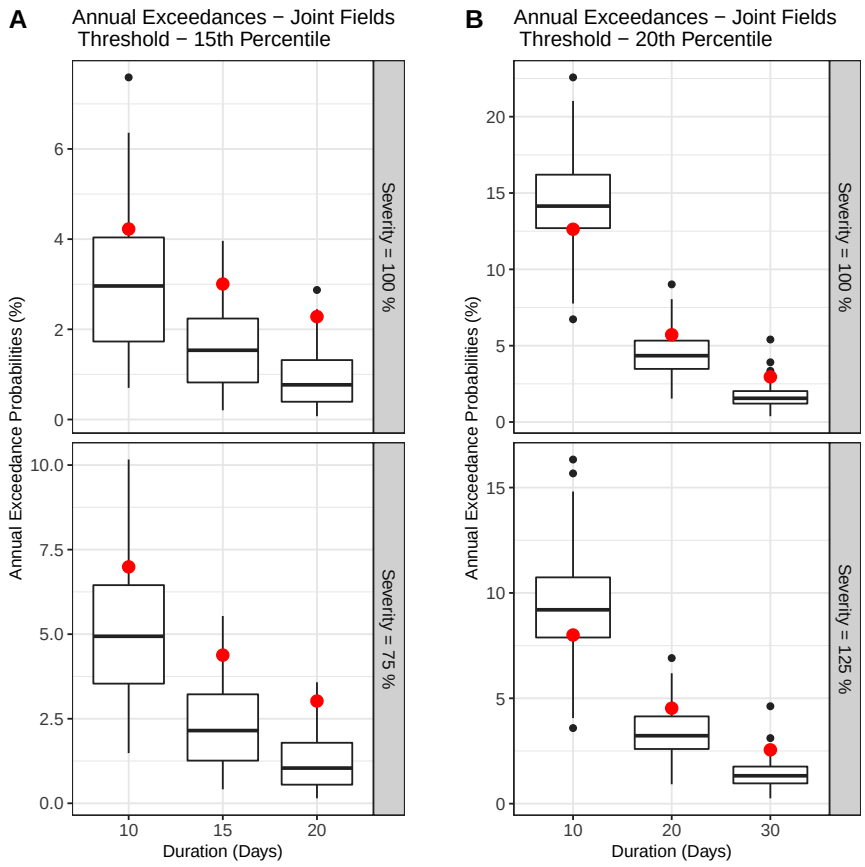

Figure S4: Probability of annual exceedances for energy droughts given a duration and severity with threshold values of (A) 15th percentile, and (B) 20th percentile. The red dot denotes the exceedance probability calculated from the reanalysis data. The boxplots denote the uncertainty in the 48 generated 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 boxplot, 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 whiskers extend to the furthest point that is within 1.5 times the IQR below the 25th percentile and above the 75th percentile respectively.

Severity vs Duration plots for different thresholds

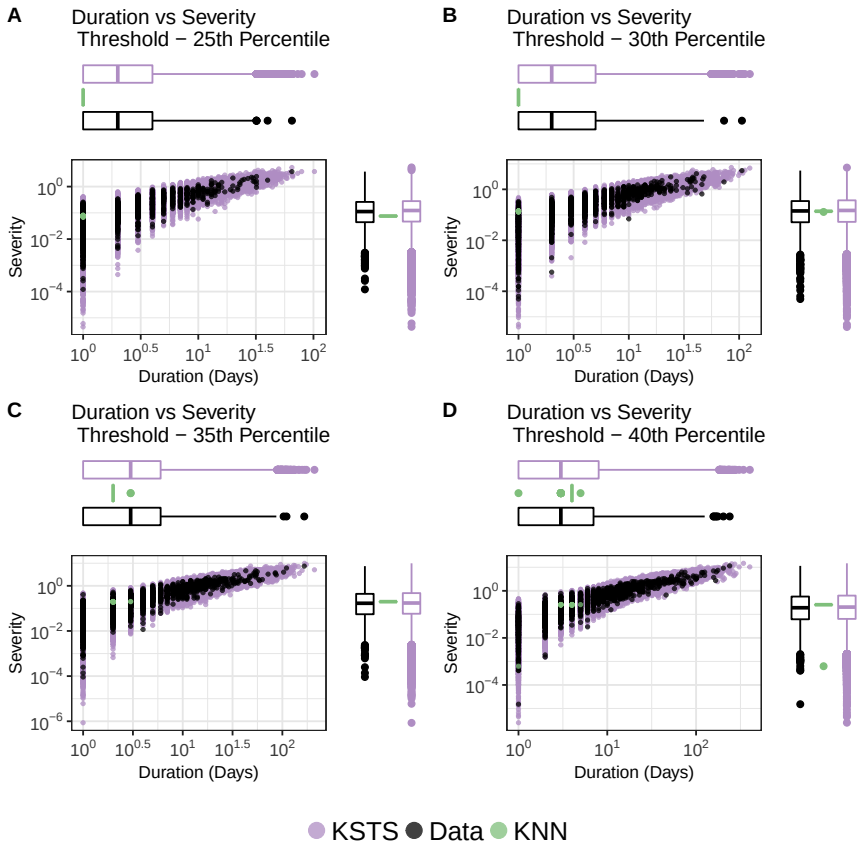

Figure S5: Duration versus Severity for all (wind and solar aggregated) energy droughts with marginal distributions (in boxplots for both variables) for the data (black), KSTS (purple) and KNN (green) simulations using threshold values of (A) 25th percentile, (B) 30th percentile, (C) 35th percentile, and (D) 40th percentile.

Energy Droughts for Wind and Solar Fields Individually

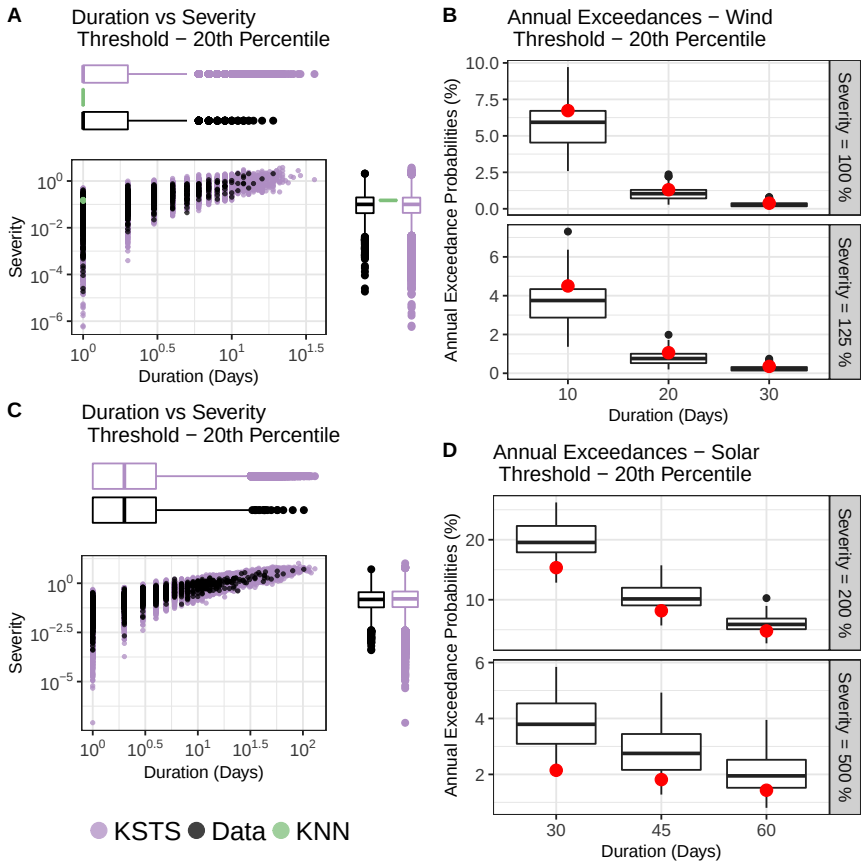

Figure S6: Duration versus Severity plots for all (A) wind and (C) solar energy droughts with marginal distributions (in boxplots for both variables) for the data (black), KSTS (purple) and KNN (green) simulations using 20th percentile as threshold. Probability of annual exceedances for (B) wind and (D) solar energy droughts given a duration and severity with threshold values of 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 boxplot, 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 whiskers extend to the furthest point that is within 1.5 times the IQR below the 25th percentile and above the 75th percentile respectively.

KSTS and KNN Simulations - Individual Site Characteristics

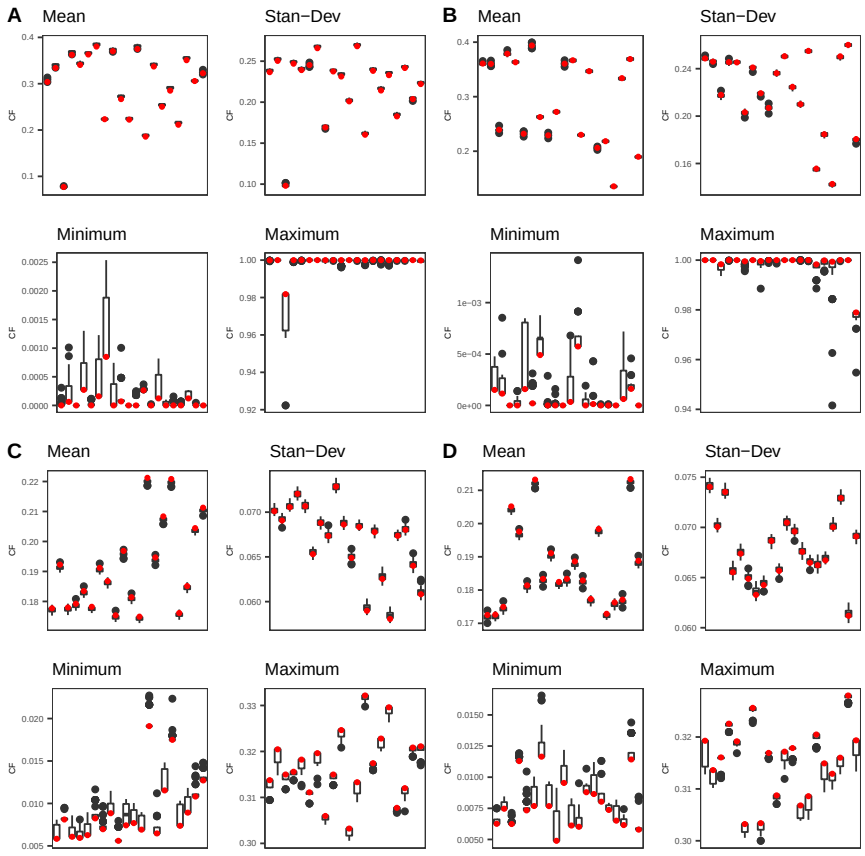

Figure S7: Simulation skill assessments for individual sites in the wind and solar fields for both KSTS and KNN simulations. (A) KSTS wind. (B) KNN wind. (C) KSTS solar. (D) KNN solar. For each sub-plot, we show the mean (top-left), the standard deviation (top-right), the minimum (bottom-left) and the 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.

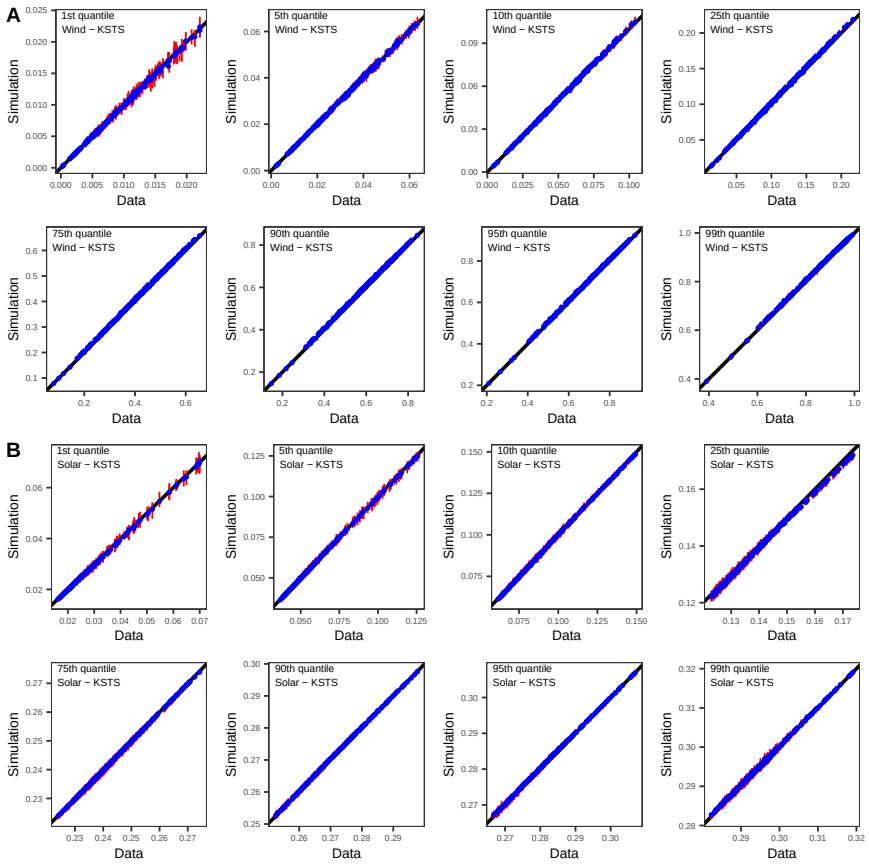

Figure S8: Simulation vs reanalysis data quantile plots for the 1st, 5th, 10th, 25th, 75th, 90th, 95th and 99th percentiles for KSTS simulations. (A) Wind KSTS - Top two rows. (B) Solar KSTS - Bottom two rows. The plots denote the quantiles for all 216 grid points in the wind and solar fields. The red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

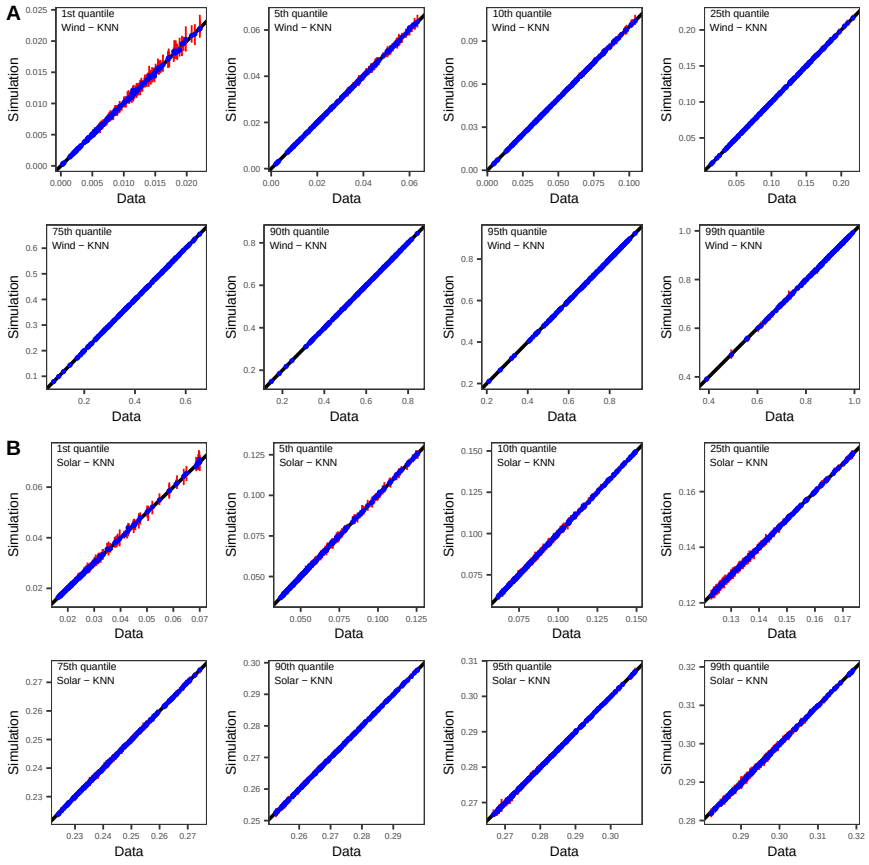

Figure S9: Simulation vs reanalysis data quantile plots for the 1st, 5th, 10th, 25th, 75th, 90th, 95th and 99th percentiles for KNN simulations. (A) Wind KSTS - Top two rows. (B) Solar KSTS - Bottom two rows. The plots denote the quantiles for all 216 grid points in the wind and solar fields. The red lines denote the mid 90th (5th-95th) percentile range and the blue dots denote the median value in the simulation spread.

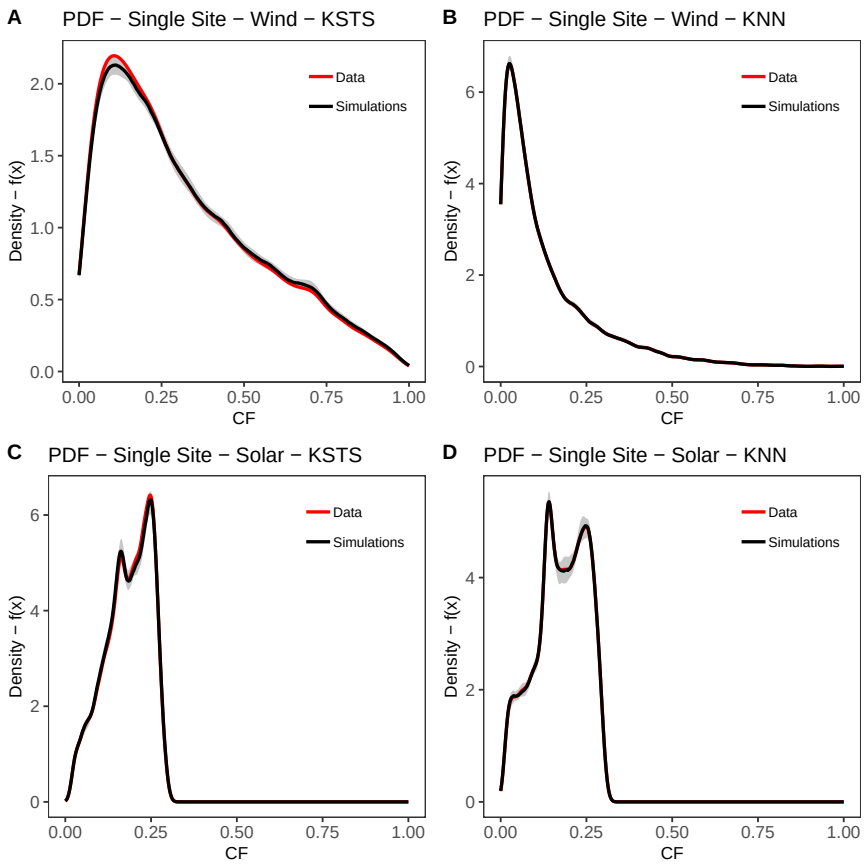

Figure S10: Kernel density estimate / Probability density function (PDF) plots for a single randomly selected grid for wind and solar. The red line denotes the reanalysis data probability density function for the selected site and the black line denotes the median simulation density. The grey region is the mid 90th (5th-95th) percentile range of the simulation spread. The grid point 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.

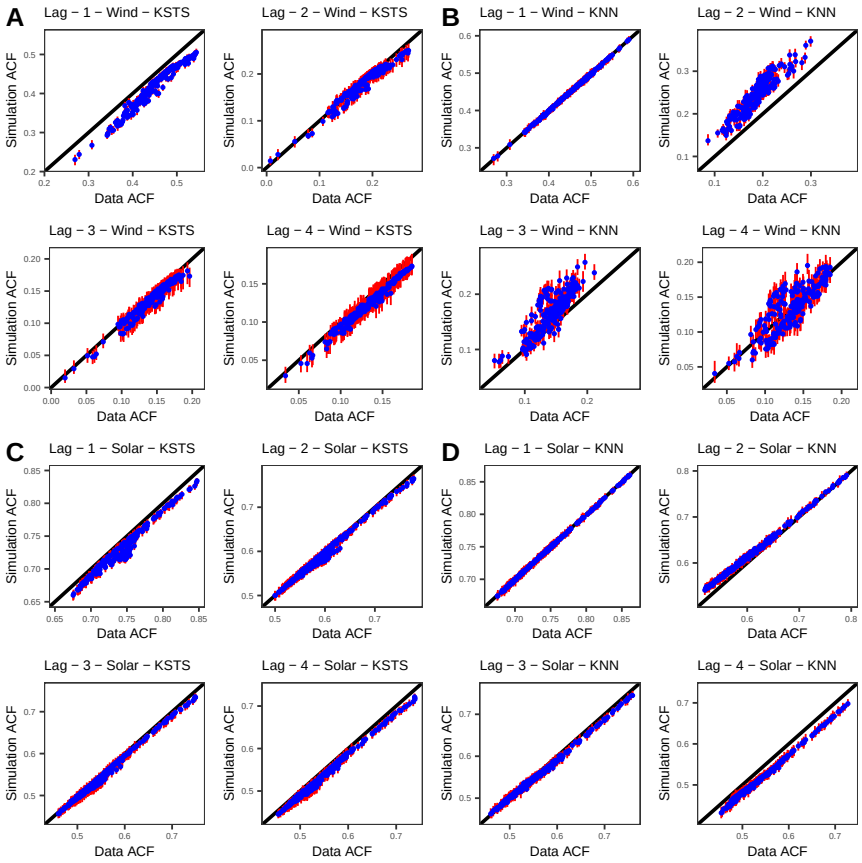

Figure S11: Simulation vs reanalysis data auto-correlation plots for lag 1,2,3 and 4 for all grid points. (A) Wind KSTS simulations. (B) Wind KNN simulations. (C) Solar KSTS simulations. (D) Solar KNN simulations. The red lines denote the mid 90th (5th-95th) percentile range and the blue dots denote the median value in the simulation spread.

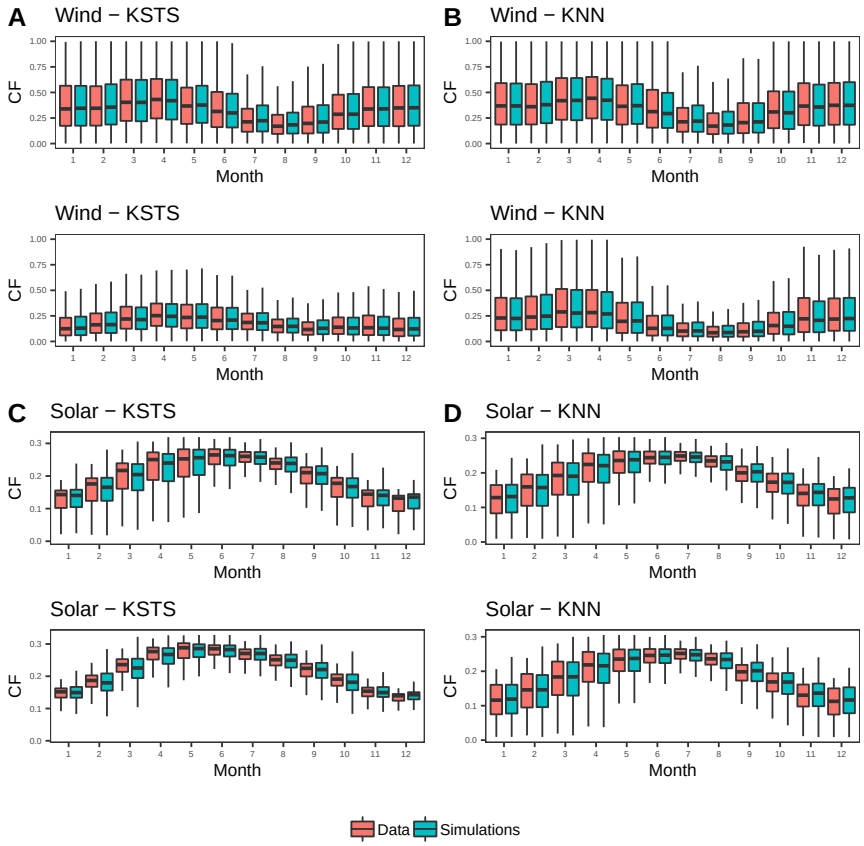

Figure S12: Seasonality / Monthly distribution of the reanalysis data and simulations. The red and green boxplots denote the reanalysis data and simulations respectively. (A) Wind KSTS simulations. (B) Wind KNN simulations. (C) Solar KSTS simulations. (D) Solar KNN simulations. Two grid points are randomly 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

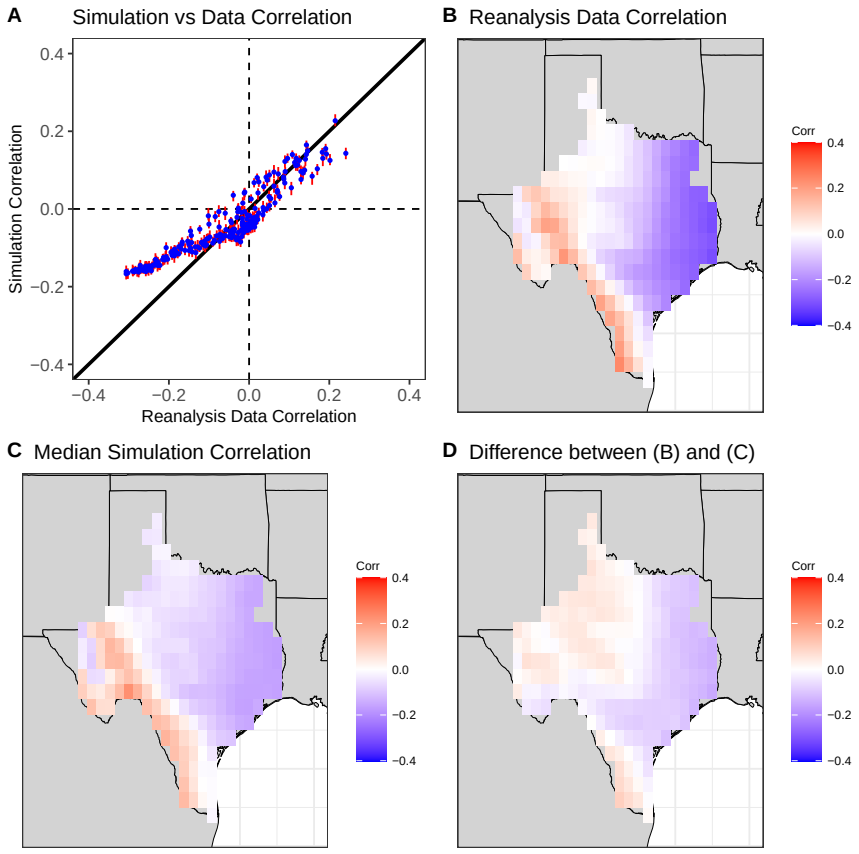

Figure S13: Pearson correlation between wind and solar at each grid point based on simultaneous simulations of wind and solar using KNN. (A) Simulation correlation vs reanalysis data correlation between wind and solar where the red lines denote the mid 90th (5th-95th) percentile range and the blue dots 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

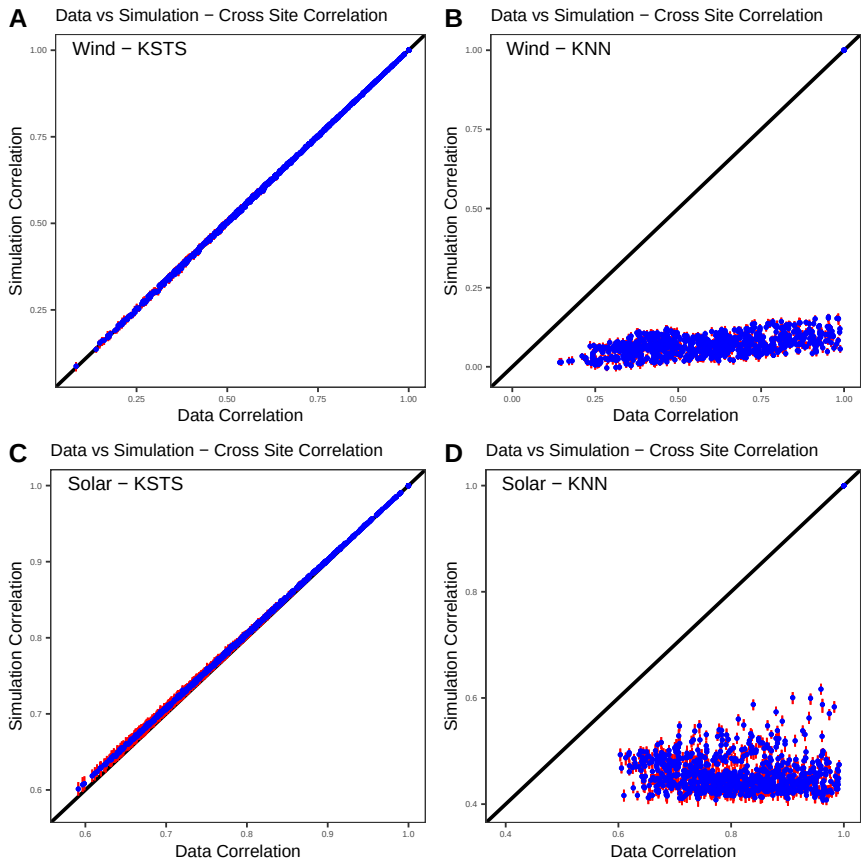

Figure S14: Simulation vs reanalysis data cross site correlation plots for individual fields. (a) Wind KSTS. (b) Wind KNN. (c) Solar KSTS (d) Solar KNN. 40 grids out of 216 are randomly selected and the 40x40 cross correlation values are computed and plotted instead of the entire 216 x 216 correlation values. The correlations are computed using Pearson's method. The red lines denote the mid 90th (5th-95th) percentile range and the blue dots denote the median value in the simulation spread.

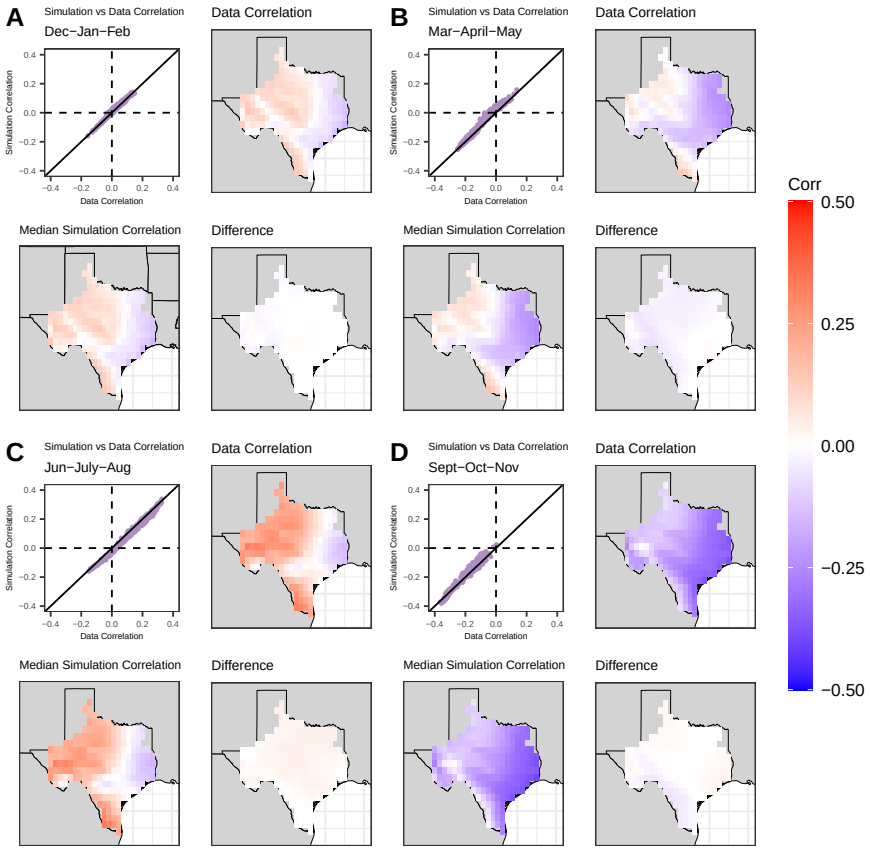

Figure S15: Seasonal correlation between wind and solar at each grid point for KSTS simulations. (A) Dec-Jan-Feb (DJF). (B) Mar-Apr-May (MAM). (C) Jun-Jul-Aug (JJA). (D) Sep-Oct-Nov (SON). For each subplot: (top-left) - median simulation vs reanalysis data correlation between wind and solar. (top-right)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

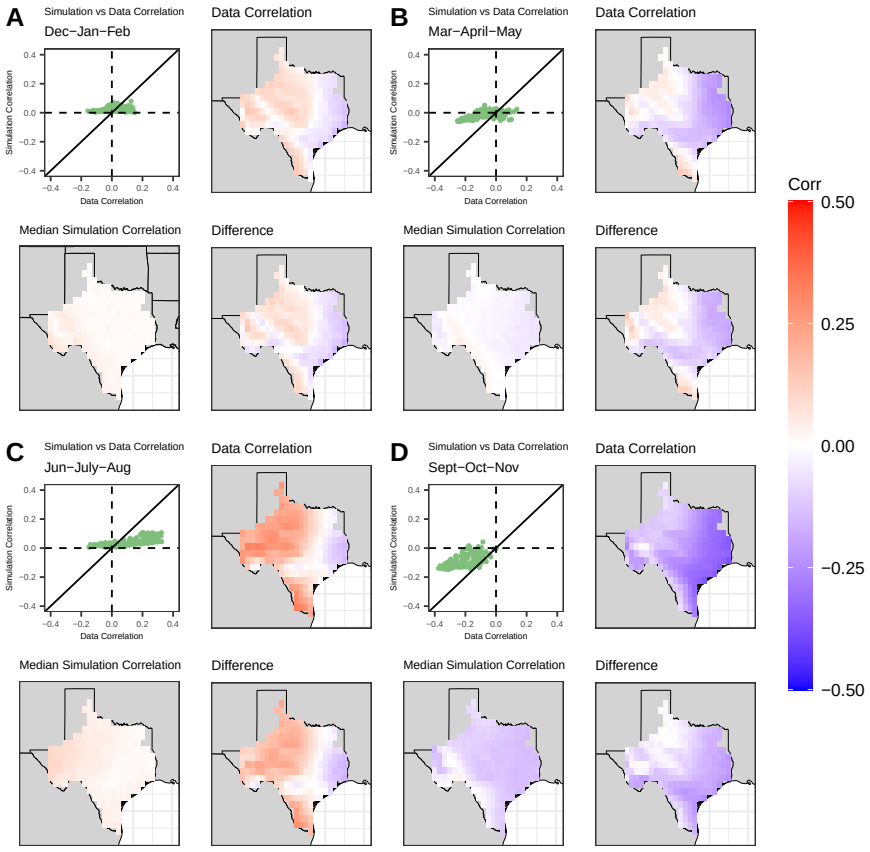

Figure S16: Seasonal correlation between wind and solar at each grid point for KNN simulations. (A) Dec-Jan-Feb (DJF). (B) Mar-Apr-May (MAM). (C) Jun-Jul-Aug (JJA). (D) Sep-Oct-Nov (SON). For each subplot: (top-left) - median simulation vs reanalysis data correlation between wind and solar. (top-right) - Plot of the reanalysis data correlations. (bottom-left) - Plot of the median simulation 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 - Figure S17), 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] .

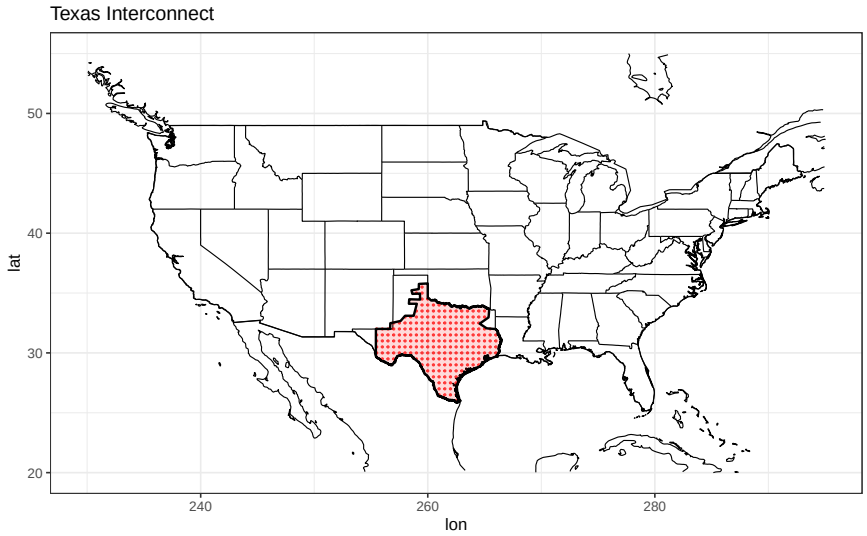

Figure S17: Texas Interconnection / ERCOT domain plot - The red shaded region denotes the area administered by ERCOT. The red dots (216) are the locations of the grid points (0.5° lat × 0.5° lon) from the ERA-5 reanalysis dataset.

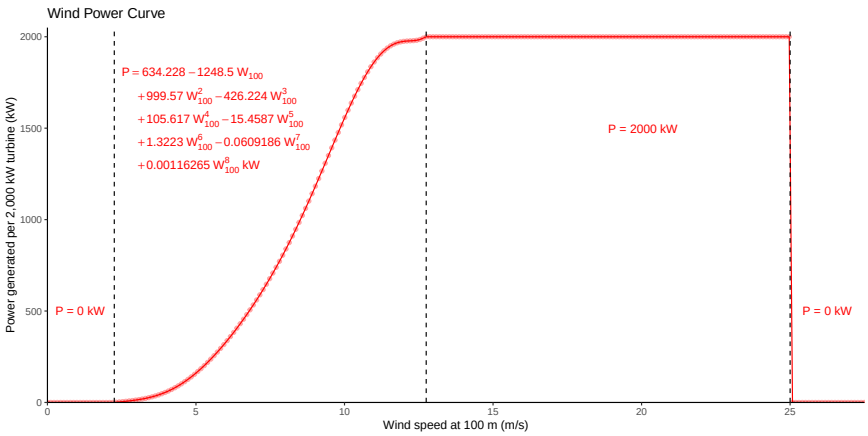

Figure S18: Wind Power Curve for a V90-2.0MW Vestas turbine.

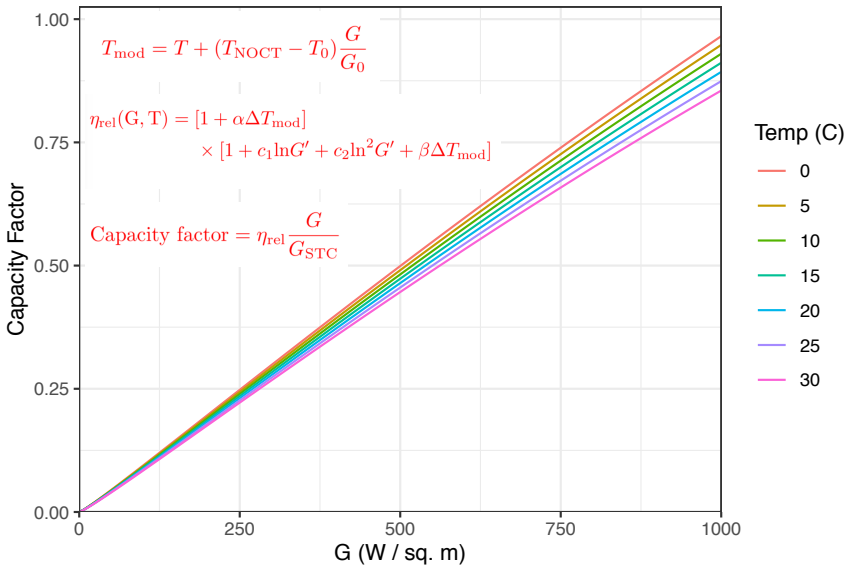

Figure S19: Relationship between solar capacity factor, hourly radiation data (G), and temperature as per Bett and Thornton [2]. Variables are defined in the following way:

$$G^{'} = \frac{G}{G_{STC}}$$

$$\Delta T_{mod} = T_{mod} - T_{STC}$$

. The constant values are the following $\alpha = 1.2{\times}10^{-3}K^{-1}$, $\beta = -4.6{\times}10^{-3}K^{-1}$, $c_1 = 0.033$, $c_2 = -0.0092$, $T_{NOCT} = 48^oC$, $T_o = 20^oC$, $G_o = 800\frac{W}{sq-m}$, $G_{STC} = 1000\frac{W}{sq-m}$, $T_{STC} = 25^oC$.

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 [3], 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 [3]. 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 of the nearest neighbors for a site with an embedding $D_{i,t}$, defined as (x_t, x_{t-1}) taking value of (10,9.8) at time t=100 days would be 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 in $\tau_{\mathrm{i,t}}$. The kernel $p_{i,j}$ 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 elsewhere. 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. The similarity vector S_t is the sum of the resampling probabilities associated with each unique time index across all sites. The curtailment of the similarity vector 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 vector S_t. Other measures of similarity of the spatial neighbors of the temporal process could also be considered.

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[3]. 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 Lall and Sharma[3] 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[3]. 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 [4]. 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 [3].

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 [5]. 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 [6].

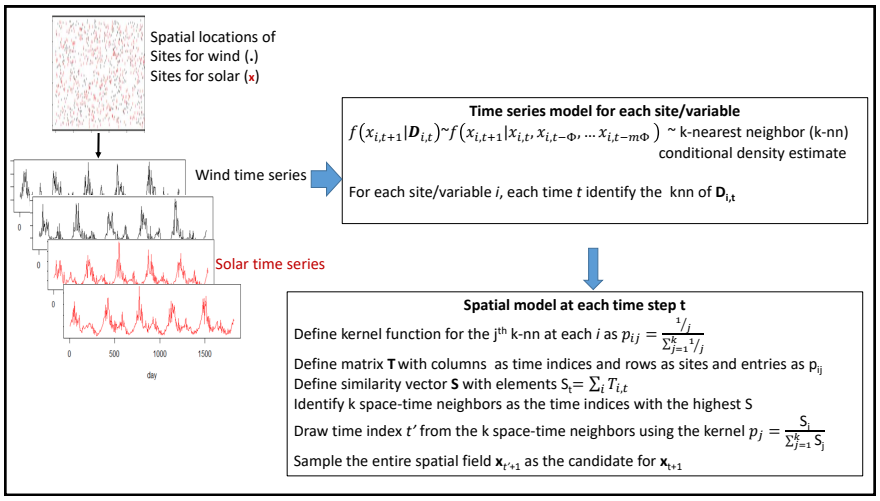

Figure S20: Schematic representation of the k-Nearest Neighbor Space Time Simulator (KSTS) with application to wind and solar fields.

Supplementary Materials - References

1. "About ERCOT." Electric Reliability Council of Texas. `http://www.ercot.com/about`. Accessed 10 July 2021.

2. Bett, Philip E., and Hazel E. Thornton. "The climatological relationships between wind and solar energy supply in Britain." Renewable Energy 87 (2016): 96-110.

3. Lall, Upmanu, and Ashish Sharma. "A nearest neighbor bootstrap for resampling hydrologic time series." Water resources research 32.3 (1996): 679-693.

4. 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.

5. 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.

6. 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).

7. Lall, Upmanu, Naresh Devineni, and Yasir Kaheil. "An empirical, nonparametric simulator for multivariate random variables with differing marginal

densities and nonlinear dependence with hydroclimatic applications." Risk Analysis 36.1 (2016): 57-73.