A k-nearest neighbor space-time simulator with applications to large-scale wind and solar power modeling

Yash Amonkar,1,2,5,* David J. Farnham,3,4 and Upmanu Lall1,2
1Columbia Water Center, Columbia University, New York, NY 10027, USA
2Department of Earth and Environmental Engineering, Columbia University, New York, NY 10027, USA
3Department of Global Ecology, Carnegie Institution for Science, Stanford, CA, USA
4Present address: ClimateAI, San Francisco, CA, USA
5Lead contact
*Correspondence: yva2000@columbia.edu
https://doi.org/10.1016/j.patter.2022.100454

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.2–5 At the same time, the costs of wind and solar electricity generation technologies have declined substantially over the last decade.6 These two factors are spurring increasing deployment of wind- and solar-based electricity generation.

A target system reliability requirement of 99.97%7 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.8,9 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.10,11

Long-duration storage (LDS), defined as storage needed to meet deficits for duration greater than 10 h,12,13 is one option to economically meet grid reliability targets while relying primarily on wind and solar generation.10 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.14 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 profiles. The estimation of these probabilities to assure high system reliability requires long data records, potentially over many decades. Collins et al.14 show the pitfalls of modeling energy reliability requires long data records, potentially over many decades. Collins et al.14 show the pitfalls of modeling energy generation facilities. We take the Electric Reliability Council of Texas (ERCOT)-Texas Interconnection region24 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 spatiotemporal data, including climate or environmental fields. An implicit assumption in the choice of timescale for the energy application is that chemical batteries help smooth out the sub-daily timescale shortages.12 The daily wind and solar capacity factors are computed at the hourly timescale and then averaged over the entire day (see the supplemental experimental procedures 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.

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 Figures 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) are more homogeneous and a function of the season, with low mean radiation and high variability in winter (December-January-February) and high mean radiation and low variability in the summer (June-July-August) (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. Yakowitz and Karlsson 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 datasets.

Lall and Sharma 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 introduced a k-nearest neighbor simulator for multivariate time series data following on earlier work, 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. developed a disaggregation method that 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 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 densities of the variables along with the 90th and 10th percentile divided by the mean for each grid.

The 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 used for univariate or low-dimensional multivariate simulations of non-Gaussian and nonlinear dependence that has been used extensively for other climate variables.

We apply our new KSTS algorithm to assess the severity, duration, and frequency of LDS 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 analyzes the issue of LDS sizing and requirement from the supply-side perspective of renewable energy producers to illustrate the utility of the proposed spatiotemporal 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 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 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-year gridded daily wind and solar data from the ERA-5 reanalysis dataset for 216 model 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 7,100-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 assets across all grid points is considered, with wind and solar having mean capacity factors of 0.28 and 0.19, respectively. For both the KSTS and KNN 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 correctly representing the space-time dependence 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 supplemental information.

Severity, duration, and frequency of energy droughts

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. 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 when 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 2A shows the annual exceedance probabilities for energy droughts of duration 20, 25, and 30 days with severity of 125% and 150% when the target threshold is the 25% 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) with the number of exceedances regressed against the duration and severity using a Poisson link function (see experimental procedures).

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 ∼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%–6% with the 5th and 95th percentiles roughly 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 2B. 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. The simulations show that these extreme events could occur far more frequently than would be estimated from relatively 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 assets, 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 threshold, 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 percentiles 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 (Figures S8 and S9), underlying probability distribution (Figure S10), autocorrelation 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.

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

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 (Figures S15 and S16).

DISCUSSION

The primary contribution of this paper is the presentation of 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 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. The KSTS 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.11,12,44 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 non-parametric 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 spatiotemporal bootstrap procedure, where a spatiotemporal 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
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. 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. 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. 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 $(n − 0.5)/(n + 1)$ 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 that is capable of capturing the heterogeneity in dynamics across even larger regions. This becomes important when the spatial scale of the simulation is expanded.

**Applicability to other problems**

The KSTS algorithm could be used for any spatiotemporal 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. 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. 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 timescales (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 modeling multiple streamflow or weather stations across a watershed while preserving the internal dependence structure. Such streamflow data exhibit spatiotemporal correlation patterns due to their position in the river network, altitude, and a host of other hydrological variables, 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. 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. 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 $(n − 0.5)/(n + 1)$ 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 that 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 with just Texas.

**EXPERIMENTAL PROCEDURES**

**Resource availability**

**Lead contact**

Further information and requests for resources and materials should be directed to and will be fulfilled by the lead contact, 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 years (1950–2020) across the Texas Interconnection and are derived from the ERA-5 reanalysis dataset, which can be accessed publicly. All code used in this study is made publicly available on Zenodo at https://doi.org/10.5281/zenodo.5912033.

**Wind and solar data**

The ERA-5 reanalysis variables used are wind speeds at 100-m altitude and downward surface solar radiation. The spatial grid size of the data is set at 0.5° latitude × 0.5° longitude and contains 216 grid points across the Texas Interconnection domain (Figure S17). The wind speed and solar radiation at each hour are averaged to hourly wind and solar power, respectively, and 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 years from January 1, 1950, to December 31, 2020.

The solar variable is the downward surface solar radiation (W/m²) and is converted to capacity factors at the hourly level by accounting for the dependence of photovoltaic performance on temperature (Figure S19). We then compute a capacity factor for each day by taking the mean of 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 DOY for each site. The deviation could be positive if that day’s value is greater than the selected threshold percentile value for that DOY or negative if it is lower. The daily energy deficit across the field is computed by aggregating the daily site deviation and is given by,

\[
y_d = \sum_{i=t}^{n}(X_d - \bar{X}_d),
\]

where \(y_d\) is the aggregated daily energy deviation at day \(t\); \(X_d\) is the normalized wind or solar value at site \(i\) and day \(t\); \(\bar{X}_d\) 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 \(y_d\) can take a positive (surplus) or negative (deficit) value on any day, while the cumulative deficit, the variable of interest is computed as

\[
z_t = \max(0, \ -y_t),
\]

where \(z_t\) and \(y_t\) are the cumulative deficit and aggregated 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

\[
C(e_i) = \sum_{j=t}^{\infty}(d_j > d_i) \cap \sum_{j=t}^{\infty}(s_j > s_i),
\]

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 loctf package.59

After the model fitting process, the count of exceedances \(C(e_i)\) is estimated using the fitted model for the required duration \(d_i\) and severity \(s_i\) for a desired drought event \(i\). The number of years of the record \(yr\) is then used to scale the number of exceedances to get the annual exceedance percentage \((p_i)\) using the formula:

\[
p_i = \frac{C(e_i) \times 100}{yr}
\]

where \(p_i\) is the annual exceedance percentage for a drought event \(i\) with severity \(s_i\) and duration \(d_i\).

**Fitting other models**

We considered and tested other strategies for spatiotemporal simulation with the Texas Interconnection data before 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.60 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 modeling 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).65 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 skilful temporal evolution characteristics, making a direct comparison infeasible.

Finally, different types of hidden Markov models (HMMs)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 spatiotemporal correlation structure.

**Simulator hyper-parameters**

For both KSTS and KNN, the seasonality in the data is accounted for by restricting the search of nearest neighbors to a ±30-day moving window across the years around the DOY. The number of nearest neighbors \((k)\) selected is approximately \(\sqrt{n}\). With 71 years and 61 days per year, \(\sqrt{n}\) is \(\sim 65\), where \(n\) is the number of possible candidate neighbors after accounting for the moving window.52 A lag-1 dependence structure for the state space is assumed. Forty-eight independent simulation realizations, each of the same
length as the reanalysis data (71 years = 25,933 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.\(^{32}\)

**KSTS**

The general structure and a cartoon example application of the KSTS algorithm is illustrated in Figures 5 and S20. 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 \(D_i\).

Define a state space \(D_i\) of dimension \(m\), which is the number of embedding delay lags. The state space can be a single lag, multiple

---

**Figure 5. Cartoon example application of the KSTS algorithm to a spatial dataset consisting of three grids/sites and data record (time) length of 7**
lags and/or disjoint lags allowing for custom time dependencies. The embedding selected for the simulator application could be,

Case 1 \( D_{ij} = (x_{t+1}, x_{t+2}); m = 2 \)

Case 2 \( D_{ij} = (x_{t+1}, x_{t+2}, x_{t+3}, x_{t+4}); m = 4, r = 1 \)

Case 3 \( D_{ij} = (x_{t+1}, x_{t+2}, x_{t+3}, x_{t+4}, x_{t+5}); m = 5 \)

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 \( D_{ij} \) is defined for each time series at site \( i \) and time \( t \), whereas \( D_{iT} \) are all the historic vectors that correspond to the selected embedding structure for site \( i \).

Step 2: compute the k-nearest neighbors for all sites at time \( t \).

At a time step for each site \( i \) using the current state space vector \( D_{ij} \), identify the k-nearest neighbors using the weighted Euclidean distance measure

\[
r_{ij} = \left( \sum_{j=1}^{m} w_j \left( D_{ij} - D_{2j} \right)^2 \right)^{1/2}
\]

where \( D_{ij} \) and \( D_{2j} \) are the \( j \)th components of \( D_{ij} \) and \( D_{ij} \) 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 \( r_{ij} \)) of site \( i \) at time \( t \) are stored in \( T_{ij} \).

Step 3: compute resampling probabilities for k-nearest neighbor indices using a discrete kernel \( p_i \),

\[
p_i = \frac{1}{\sum_{j=1}^{s} \bar{p}_j}
\]

where \( p_i \) is the resampling probability for the \( j \)th element (time instance of the \( j \)th nearest neighbor of \( D_{ij} \)) in \( T_{ij} \). The resampling kernel stays the same across all time \( t \) and across all sites and is pre-computed and stored before simulation. It is a function of the number of neighbors \( k \) and not the distances.

Step 4: define \( T_{ij} \) and similarity vector \( S_t \) for time \( t \).

Define \( T_{ij} \) as a matrix where the rows and columns correspond to the sites and unique time indices from the historical data, respectively. A row records the resampling probabilities associated with the time indices for the k-nearest neighbors in \( T_{ij} \) for each site \( j \) 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 \( T_{ij} \).

\[
S_t = \sum_{j=1}^{s} T_{ij}
\]

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 with 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.

\[
\bar{S}_{ij} = \frac{S_{ij}}{\sum_{j=1}^{s} S_{ij}}
\]

Step 6: resample the full spatial field for time \( t + 1 \).

Using the discrete probability mass function \( S_t \) sample a single value and resample 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 to supplemental experimental procedures for further details on the algorithm and hyper-parameter selection.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2022.100454.

ACKNOWLEDGMENTS

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, United States.U.L. acknowledges support from NSF Convergence Accelerator project number 2040613.

AUTHOR CONTRIBUTIONS

Y.A. developed the code and performed the computations, Y.A. and D.J.F. designed the analysis, conceived experiments, and conducted 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.

Received: July 19, 2021
Revised: November 15, 2021
Accepted: February 2, 2022
Published: February 28, 2022

REFERENCES

1. Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al. (2020). The ERAS global reanalysis. Q. J. R. Meteorol. Soc. 146, 1999–2049. https://doi.org/10.1002/qj.3803.
2. California Legislative Information (2018). Senate bill -100 California renewables portfolio standard program: emissions of greenhouse gases. https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=201720180SB100.
3. Deyette, J. (2019). States March toward 100% clean energy–who’s next? Union of concerned scientists. https://blog.ucsusa.org/jeff-deyette/states-march-toward-100-clean-energy-whos-next/.
4. European Commission (2020). Proposal for a 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). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020PCO080.
5. New York State Legislature (2019). The climate leadership and community protection act (CLCPA). https://climate.ny.gov/.
6. International Renewable Energy Agency (2020). Renewable power generation costs in 2019. https://www.irena.org/publications/2020/Jan/ReNewable-Power-Costs-in-2019.
7. North American Electric Reliability Corporation (2012). 2012 state of reliability. https://www.nerc.com/files/2012_sor.pdf.
8. Beaudin, M., Zareipour, H., Schellenberglaibe, A., and Rosehart, W. (2010). Energy storage for mitigating the variability of renewable electricity sources: an updated review. Energy Sustain. Dev. 14, 302–314. https://doi.org/10.1016/j.esd.2010.09.007.
9. Després, J., Mina, S., Kitous, A., Criqui, P., Hadjsaid, N., and Noirot, I. (2017). Storage as a flexibility option in power systems with high shares of variable renewable energy sources: a POLES-based analysis. Energy Econ. 64, 638–650. https://doi.org/10.1016/j.eneco.2016.03.006.
10. Jenkins, J.D., Luke, M., and Themstrom, S. (2018). Getting to zero carbon emissions in the electric power sector. Joule 2, 2498–2510. https://doi.org/10.1016/j.joule.2018.11.013.
48. Bett, P.E., and Thornton, H.E. (2016). The climatological relationships between wind and solar energy supply in Britain. Renew. Energy 87, 96–110. https://doi.org/10.1016/j.renene.2015.10.006.

49. Hyndman, R.J., and Athanasopoulos, G. (2018). Forecasting: Principles and Practice (OTexts).

50. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. Commun. ACM 63, 139–144. https://doi.org/10.1145/3422622.

51. Chen, Y., Wang, Y., Kirschen, D., and Zhang, B. (2018). Model-free renewable scenario generation using generative adversarial networks. IEEE Trans. Power Syst. 33, 3265–3275. https://doi.org/10.1109/TPWRS.2018.2794541.

52. Chen, Y., Wang, X., and Zhang, B. (2018). An unsupervised deep learning approach for scenario forecasts. In 2018 Power Systems Computation Conference (PSCC), pp. 1–7. https://doi.org/10.23919/PSCC.2018.8442500.

53. Rabiner, L., and Juang, B. (1986). An introduction to hidden Markov models. IEEE ASSP Mag. 3, 4–16. https://doi.org/10.1109/MASSP.1986.1165342.