Predicting Spatiotemporal Impacts of Weather on Power Systems Using Big Data Science

Mladen Kezunovic, Zoran Obradovic, Tatjana Dokic, Bei Zhang, Jelena Stojanovic, Payman Dehghanian and Po-Chen Chen

Abstract Due to the increase in extreme weather conditions and aging infrastructure deterioration, the number and frequency of electricity network outages is dramatically escalating, mainly due to the high level of exposure of the network components to weather elements. Combined, 75% of power outages are either directly caused by weather-inflicted faults (e.g., lightning, wind impact), or indirectly by equipment failures due to wear and tear combined with weather exposure (e.g. prolonged overheating). In addition, penetration of renewables in electric power systems is on the rise. The country's solar capacity is estimated to double by the end of 2016. Renewables significant dependence on the weather conditions has resulted in their highly variable and intermittent nature. In order to develop automated approaches for evaluating weather impacts on electric power system, a comprehensive analysis of large amount of data needs to be performed. The problem addressed in this chapter is how such Big Data can be integrated, spatio-temporally correlated, and analyzed in real-time, in order to improve capabilities of modern electricity network in dealing with weather caused emergencies.

M. Kezunovic (✉) · T. Dokic · B. Zhang · P. Dehghanian · P.-C. Chen
Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA
e-mail: kezunov@ece.tamu.edu

T. Dokic
e-mail: tatjana.djokic@tamu.edu

B. Zhang
e-mail: adele.zhang@tamu.edu

P. Dehghanian
e-mail: payman.dehghanian@tamu.edu

P.-C. Chen
e-mail: pchen01@tamu.edu

Z. Obradovic · J. Stojanovic
Computer and Information Department, Temple University, Philadelphia, PA, USA
e-mail: zoran.obradovic@temple.edu

J. Stojanovic
e-mail: jelena.stojanovic@temple.edu

W. Pedrycz and S.-M. Chen (eds.), *Data Science and Big Data: An Environment of Computational Intelligence*, Studies in Big Data 24,
DOI 10.1007/978-3-319-53474-9_12

Keywords Aging infrastructure • Asset management • Big data • Data analytics • Data mining • Insulation coordination • Outage management • Power system • Solar generation forecast • Weather impact

1 Introduction

The Big Data (BD) in the power industry comes from multiple sources: variety of measurements from the grid, weather data from a variety of sources, financial data from electricity and other energy markets, environmental data, etc. The measurements from the physical network and weather data exhibit various aspects of BD: large volume, high velocity, increasing variety and varying veracity. For efficient condition-based asset and outage management, and preventive real-time operation, fast processing of large volumes of data is an imperative. The volume and velocity with which the data is generated can be overwhelming for both on-request and real-time applications. The heterogeneity of data sources and accuracy are additional challenges. The effective use of BD in power grid applications requires exploiting spatial and temporal correlations between the data and the physical power system network.

In this chapter, we will address unique fundamental solutions that will allow us to effectively fuse weather and electricity network data in time and space for the benefit of predicting the risk associated with weather impact on utility operation, generation, and asset and outage management. Computational intelligence approach plays a central role in such a task. Unstructured models like neural networks (NN), fuzzy systems and hybrid intelligent systems represent powerful tools for learning non-linear mappings. However, majority of those models assume independent and identically distributed random variables mainly focusing on the prediction of a single output and could not exploit structural relationships that exist between multiple outputs in space and time.

We will illustrate how more accurate predictions are possible by structured learning from merged heterogeneous BD compared to unstructured models, which is achieved by developing and characterizing several innovative decision-making tools. We will describe how the BD is automatically correlated in time and space. Then, the probabilistic graphical model called the Gaussian Conditional Random Fields (GCRF) will be introduced.

The proposed BD analytics will be examined for the following applications: (1) Assets management—predicts weather impacts on deterioration and outage rates of utility assets such as insulators, providing knowledge for optimal maintenance schedules and replacement strategies. The GCRF model is used to predict the probability of a flashover leading to probability of total insulator failure. (2) Solar Generation—the GCRF model is introduced to forecast the solar photovoltaic generation. Such data-driven forecasting techniques are capable of modeling both the spatial and temporal correlations of various solar generation stations, and

therefore they are performing well even under the scenarios with unavailable or missing data.

Benefits of the proposed method are assessed through a new risk-based frame-work using realistic utility data. The results show that spatio-temporal correlation of BD and GCRF model can significantly improve ability to manage the power grid by predicting weather related emergencies in electric power system.

2 Background

2.1 Power System Operation, Generation, Outage and Asset Management

The current context of the modern electric industry, characterized by competitive electricity markets, privatization, and regulatory or technical requirements man-dates power utilities to optimize their operation, outage and asset management practices and develop the requisite decision plans techno-economically. With the rapid deployment of renewable generation based on wind, solar, geothermal energy resources, etc., as well as the increasing demand to deliver higher quality electricity to customers, many electric utilities have been undergoing a paradigm shift regarding how the grid planning, operation, and protection should be reframed to enhance the resilience of the power delivery infrastructure in face of many threating risks. While the wide deployment of distributed renewable generation in the grid has brought about a huge potential for performance improvements in various domains, yet there are critical challenges introduced by renewables to overcome. An accurate forecast of the unpredictable and variable sources of renewable generation, as well as their strategic coordination in every-day normal and even emergency operation scenarios of the grid is of particular interest [1].

Asset management is said to be the process of cost minimization and profit maximization through the optimized operation of the physical assets within their life cycles. Asset management in electric power systems can be broadly classified into four main categories based on the possible time scales, i.e., real-time, short-term, mid-term, and long-term [2]. Real-time asset management mainly covers the key power system resiliency principles and deals with the unexpected outages of power system equipment and grid disruptions. System monitoring and operating condition tracking infrastructures, e.g. supervisory control and data acquisition systems (SCADA) and geographic information systems (GIS), play a vital role for a techno-economically optimized real-time asset management. By enhancing the situational awareness, they enable power system operators to effectively monitor and control the system. Contingency analysis as well as online outage management scheme is utilized to coordinate the necessary resource managements through automated control systems. Wide area measurement systems (WAMS), which have been recently realized by broad deployment of phasor

measurement units (PMUs), are among the new technologies in this context [3]. Intelligent Electronic Devices (IED) located in substations can record and store a huge amount of data either periodically with high sampling rates or on certain critical events such as faults. When properly interpreted, data gathered from such devices can be used to predict, operate, monitor and post-mortem analyze power system events. Yet, without knowledge on the meaning of such data, efficient utilization cannot be pursued. Thus, data mining techniques such as neural networks, support vector machines, decision trees, and spatial statistics should be considered for acquiring such knowledge.

Short-term asset management strives to maximize the rate of return associated with asset investments. The process, called asset valuation, is performed to incorporate company's investments in its portfolio. The value mainly depends on the uncertain market prices through various market realizations. Market risk assessment is a key consideration and the revenue/profit distributions are gained through a profitability analysis.

Optimized maintenance scheduling falls within the realm of mid-term asset management. It guides the maintenance plans and the associated decision making techniques toward satisfactorily meeting the system-wide desired targets. In other words, efforts are focused to optimize the allocation of limited financial resources where and when needed for an optimal outage management without sacrificing the system reliability. Smart grid concept has facilitated an extensive deployment of smart sensors and monitoring technologies to be used for health and reliability assessment of system equipment over time and to optimize the maintenance plans accordingly [4]. Two factors that are typically used for condition monitoring are the prognostic parameters and trends of equipment deterioration. Prognostic parameters that provide an indication of equipment condition such as ageing and deterioration are useful indicators of impending faults and potential problems. The trend of the deterioration of critical components can be identified through a trend analysis of the equipment condition data. In addition to the aforementioned factors, event triggered deterioration of equipment can be also considered. Gathering and interpreting historical data can provide useful information for evaluation of impacts that events had on given equipment over its lifetime. Data analytics can use such data as a training set and establish connection between causes of faults/disturbances and its impact on equipment condition. Knowledge gathered during training process can then be used for classification of new events generated in real time and prediction of equipment deterioration that is caused by these effects. With addition of historical data about outages that affected the network, a more efficient predictive based maintenance management can be developed [4].

Long-term investment in power system expansion planning as well as wide deployment of distributed generations fall within the scope of long-term asset management where the self-interested players, investors, and competitors are invited to participate in future economic plans.

Knowledge from outage and assets management can be combined by considering both IED recordings and equipment parameters and reliability characteristics for a more reliable decision making in practice. The goal is to explore limits of

available equipment and identify possible causes of equipment malfunction by gathering both data obtained through the utility field measurement infrastructure and additional data coming from equipment maintenance and operation records, as well as weather and geographic data sources. In this approach, decision about the state of the equipment is made based on the historical data of the events that affected the network, as well as real-time data collected during the ongoing events.

2.2 Weather Data Parameters and Sources

The measurement and collection infrastructure of weather data has been well developed over the past decades. In this subsection, as an example, the data from National Oceanic and Atmospheric Administration (NOAA) [5] will be discussed. Based on the type of measurements, the weather data may be categorized into:

- Surface observations using data collected from land-based weather stations that contain measurement devices that track several weather indices such as temperature, precipitation, wind speed and direction, humidity, atmospheric pressure, etc.
- Radar (Radio Detection and Ranging) provides accurate storm data using radio waves to determine the speed, direction of movement, range and altitude of objects. Based on radar measurements different reflectivity levels are presented with different colors on a map.
- Satellites generate raw radiance data. Unlike local-based stations, they provide global environmental observations. Data is used to monitor and predict wide-area meteorological events such as flash floods, tropical systems, tornadoes, forest fires, etc.

Table 1 [6–13] demonstrates the list of weather data sources from NOAA which may be useful for power system operations. A list of national support centers for more real-time and forecasting data may be found in [14]. A list of commercial

Table 1 List of weather data sources

Data source	Available access
National Weather Service (NWS)	GIS Portal [6]
	National Digital Forecast Database [7]
	Doppler Radar Images [8]
National Centers for Environmental Information (NCEI)	Data Access [9]
	Web Service [10]
	GIS Map Portal [11]
Office of Satellite and Product Operations (OSPO)	Satellite Imagery Products [12]
Global Hydrology Resource Center (GHRC)	Lightning and Atmospheric Electricity Research [13]

weather vendors for more specialized meteorological products and services may be found in [15]. More land and atmosphere data and images may be found in National Aeronautics and Space Administration (NASA) websites.

Specific power system operations are related to certain type of weather events. In such case, the most relevant weather data input is required for data analytics. An example below demonstrates the use of satellite and radar data.

The satellite meteorological detection is passive remote sensing in general, whereas the radar meteorological detection is active remote sensing. Radars can emit radio or microwave radiation and receive the back-scattering signals from a convective system. For a practical example regarding tropical cyclones, satellites can observe a tropical cyclone once it forms in the ocean, and radar can detect its inner structure as it moves near the continent and lands in.

Lightning data is gathered by the sensors that are typically located sparsely over the area of interest. There are three common types of lightning sensors: (1) Ground-based systems that use multiple antennas to determine distance to the lightning by performing triangulation, (2) Mobile systems that use direction and a sensing antenna to calculate distance to the lightning by analyzing surge signal frequency and attenuation, and (3) Space-based systems installed on artificial satellites that use direct observation to locate the faults.

Typical detection efficiency for a ground-based system is 70–90%, with a accuracy of location within 0.7–1 km, while space-based systems have resolution of 5–10 km, [16].

For example, The National Lightning Detection Network (NLDN) [17] uses ground-based system to detect lightning strikes across the United States. After detection data received from sensors in raw form is transmitted via satellite-based communication to the Network Control Center operated by Vaisala Inc. [18].

When it comes to the way data is received by the utility we can distinguish two cases: (i) the lightning sensors are property of the utility, and (ii) lightning data is received from external source. In the first case raw data are received from the sensors, while in second case external sources provide information in the format that is specific to the organization involved. Lightning data typically includes the following information: a GPS time stamp, latitude and longitude of the strike, peak current, lightning strike polarity, and type of lightning strike (cloud-to-cloud or cloud-to-ground).

2.3 Spatio-Temporal Correlation of Data

Any kind of data with a spatial component can be integrated into GIS as another layer of information. As new information is gathered by the system, these layers can be automatically updated. Two distinct categories of GIS data, spatial and attribute data can be identified. Data which describes the absolute and relative context of geographic features is spatial data. For transmission towers, as an example, the exact spatial coordinates are usually kept by the operator. In order to provide

additional characteristics of spatial features, the attribute data is included. Attribute data includes characteristics that can be either quantitative or qualitative. For instance, a table including the physical characteristics of a transmission tower can be described along with the attribute data.

In terms of spatial data representation, raster and vector data can be used. In case of vector data, polygons, lines and points are used to form shapes on the map. Raster presents data as a grid where every cell is associated with one data classification. Typically, different data sources will provide different data formats and types. Although modern GIS tools such as ArcGIS [19] are capable of opening and overlaying data in different formats, analysis of mixed database is a challenge [20].

GIS is often understood as a visualization tool of mapping geographical information. However, it also enables interpretation of spatio-temporal data for better decision making. An enterprise level GIS requires a GIS platform and a geospatial database management system [21]. The GIS platform will allow the execution of designed applications and enable access from various devices, and it is the key to interface with current utility decision-making tools such as outage management system. The geospatial database system will keep the most recent asset information and update the database from asset management.

In addition to spatial reference, data must also be time referenced in a unique fashion. Following factors are important for time correlation of data:

- Time scales: data can be collected with different time resolution: yearly, monthly, daily, hourly, once every few minutes or even seconds. Time scales for different applications and events of interest for this research are presented in Fig. 1 [22].
- Atomic Time: Standards and their characteristics are listed in Table 2. All standards use the same definition for 1 s.

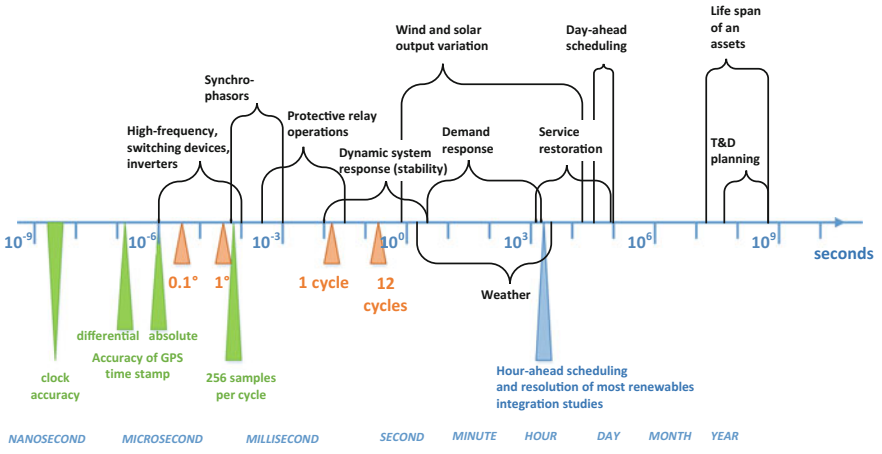

Fig. 1 Time scales of the Big Data and related applications of interest to the power sector

Table 2 Atomic Time Standards

Standard name	Leap seconds	Representation	Zero datum
UTC—Coordinated Universal Time	Included	Date and time: [ddmmyyyy, hhmmss]	N/A
GPS Satellite Time	Not included	Seconds, or week# + Seonds_of_Week	Midnight roll-over to 6 Jan 1980 UTC
TAI—International Atomic Time	Not included	Seconds	Midnight roll-over to 1 Jan 1970 UTC

Table 3 Time Synchronization Protocols

Protocol name	Media	Sync accuracy	Time standard	Description
Network Time Protocol (NTP), [23]	Ethernet	50–100 ms	UTC	Simplified version—SNTP assumes symmetrical network delay
Serial time code IRIG B-122, [24]	Coaxial	1–10 μs	TAI	Two versions used in substation: B12x and B00x "lagging" code
Precision Time Protocol (PTP), [25]	Ethernet	20–100 ns	TAI	Master to slave Hardware time-stamping of PTP packets is required

- Synchronization protocols: Accuracy of a time stamp is highly dependent on the type of the signal that is used for time synchronization. Different measuring devices that use GPS synchronization can use different synchronization signals. Table 3 lists properties of time synchronization protocols.

3 Weather Impact on Power System

3.1 Weather Impact on Outages

Main causes of weather related power outages are:

- Lightning activity: the faults are usually caused by cloud-to-ground lighting hitting the poles.
- Combination of rain, high winds and trees movement: in order to completely understand this event several data sources need to be integrated including precipitation data, wind speed and direction, and vegetation data.
- Severe conditions such as hurricanes, tornados, ice storms: in case of severe conditions multiple weather factors are recorded and used for analysis.
- In case of extremely high and low temperatures the demand increases due to cooling and heating needs respectively leading to the network overload.

Weather impact on outages in power systems can be classified into direct and indirect [26]. Direct impact to utility assets: This type of impact includes all the situations where severe weather conditions directly caused the component to fail. Examples are: lightning strikes to the utility assets, wind impact making trees or tree branches to come in contact with lines, etc. These types of outages are marked as weather caused outages. Indirect impact to utility assets: This type of impact accrues when weather creates the situation in the network that indirectly causes the component to fail. The examples are: hot weather conditions increasing the demand thus causing the overload of the lines resulting in the line sags increasing the risk of faults due to tree contact, exposure of assets to long term weather impacts causing component deterioration, etc. These types of outages are marked as equipment failure.

Weather hazard relates to the surrounding severe weather conditions that have a potential to cause an outage in the electric network. Key source of information for determining the weather hazard is weather forecast. Depending on the type of hazard that is under examination different weather parameters are observed. In case of a lightning caused outage, forecast for lightning probability, precipitation, temperature, and humidity needs to be considered, while in case of outages caused by wind impact, parameters of interest are wind speed, direction and gust, temperature, precipitation, humidity, probability of damaging thunderstorm wind.

National Digital Forecast Database (NDFD), [7] provides short-term (next 3–7 days) and long-term (next year and a half) weather prediction for variety of weather parameters such as temperature, precipitation, wind speed and direction, hail, storm probabilities etc. NDFD uses Numerical Weather Prediction (NWP), which is taking current weather observations and processing it using different numerical models for prediction of future state of weather. Overview of hazard analysis based on NDFD data is presented in Fig. 2 [27].

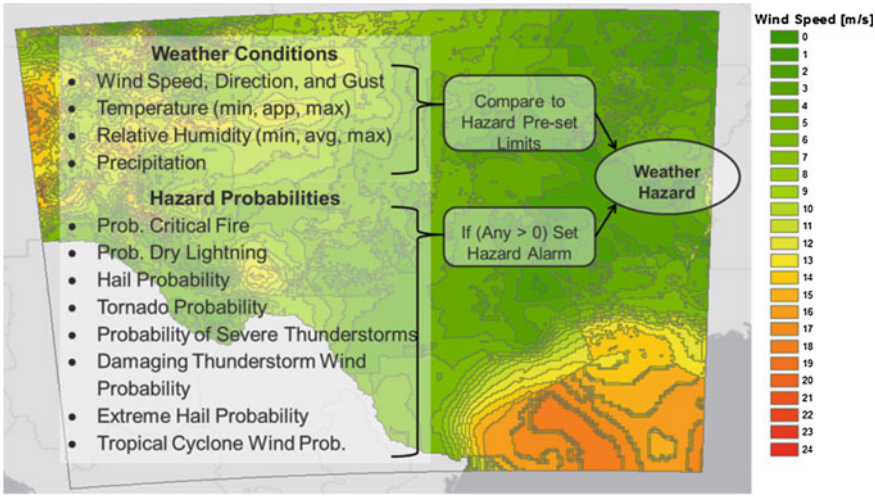

Fig. 2 Weather hazard using NDFD data, [27]

3.2 *Renewable Generation*

Due to its' environmentally friendly and sustainable nature, renewable generation, such as wind generation, solar generation, etc., has been advocated and developed rapidly during the past decades, as illustrated in Fig. 3. Many state governments are adopting and increasing their Renewable Energy Portfolio (RPS) standards [28]. Among the 29 states with the RPS, California government is holding the most aggressive one, aiming at reaching 33% renewables by 2020 [29].

Weather's impact on the renewable generation is quite evident. For example, solar generation largely depends on the solar irradiance, which could be affected by the shading effect due to the variability of the clouds. Besides, other weather condition such as the high temperature or snow can decrease the production efficiency of the solar panel. As another commonly utilized renewable energy, wind generation is sensitive to the availability of the wind resources, since small differences in wind speed lead to large differences in power. Besides, when the wind blows extremely hard, the wind turbine may switch off out of self-protection.

Renewable generation is quite sensitive to weather factors, and some of them are highly variable and unpredictable: the wind could blow so hard at this moment and suddenly stop at the next time step; the solar irradiance received by the solar panel could suddenly goes down because of a moving cloud, etc.

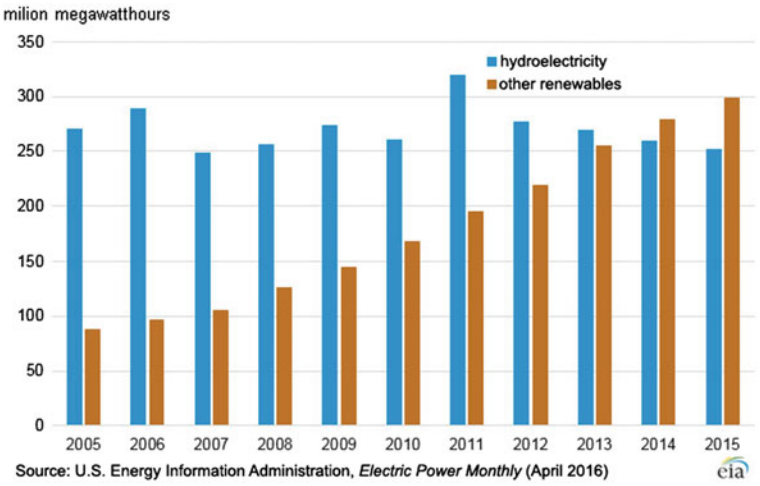

Fig. 3 Illustration on the development of renewable generation [30]

4 Predictive Data Analytics

4.1 Regression

4.1.1 Unstructured Regression

In the standard regression setting we are given a data set with N training examples, $D = \{(\mathbf{x}_i, y_i), i = 1, \ldots, N\}$, where $\mathbf{x}_i \in \mathbf{X} \subset R^M$ is an M dimensional vector of inputs and $y_i \in R$ is a real-valued output variable.

For example, in one of the applications described later in this chapter (Sect. 5.1.2.3), in the data set D, $\boldsymbol{x}$ are multivariate observations on a substation or a tower, like weather conditions or peak current and lightning strike polarity, while the output of interest y is the BIL (Basic Lightning Impulse Insulation Level) after occurrence of lightning strike.

The objective of regression is to learn a linear or non-linear mapping f from training data D that predicts the output variable y as accurately as possible given an input vector $\boldsymbol{x}$. Typically, the assumption about data-generating model is $y = f(\mathbf{x}) + \varepsilon, \varepsilon \sim \mathrm{N}(0, \sigma^2)$, where ε is Gaussian additive noise with constant variance σ^2. This setup is appropriate when data are independently and identically distributed (IID). The IID assumption is often violated in applications where data reveal temporal, spatial, or spatio-temporal dependencies. In such cases, the traditional supervised learning approaches, as linear regression or neural networks, could result in a model with degraded performances. Structured regression models are, therefore, used for predicting output variables that have some internal structure. Thus, in the following sections we introduce such methods.

4.1.2 Structured Regression (Probabilistic Graphical Models)

Traditional regression models, like neural networks (NN), are powerful tools for learning non-linear mappings. Such models mainly focus on the prediction of a single output and could not exploit relationships that exist between multiple outputs. In structured learning, the model learns a mapping f: $\mathbf{X}^N \rightarrow R^N$ to simultaneously predict all outputs given all input vectors. For example, let us assume that the value of y_i is dependent on that of y_{i-1} and y_{i+1}, as is the case in temporal data, or that the value of y_i is dependent on the values of neighboring y_h and y_j, as it is the case in spatially correlated data. Let us also assume that input x_i is noisy. A traditional model that uses only information contained in x_i to predict y_i might predict the value for y_i to be quite different from those of y_{i-1} and y_{i+1} or y_h and y_j because it treats them individually. A structured predictor uses dependencies among outputs to take into account that y_i is more likely to have value close to y_{i-1} and y_{i+1} or y_h and y_j, thus improving final predictions.

In structured learning we usually have some prior knowledge about relationships among the outputs y. Mostly, those relationships are application-specific where the

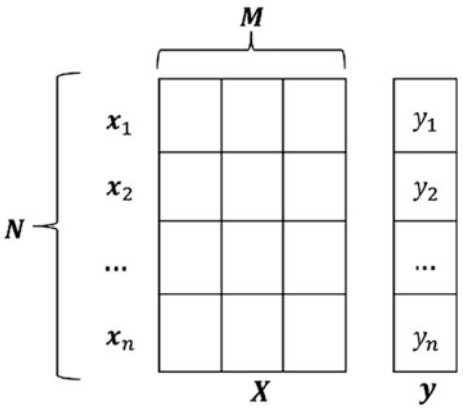

Fig. 4 Data set D in the standard regression setting with X as an input variable matrix and y as a real-valued output variable vector

dependencies are defined in advance, either by domain knowledge or by assumptions, and represented by statistical models.

Probabilistic Graphical Models

Relationships among outputs can be represented by graphical models. The advantage of the graphical models is that the sparseness in the interactions between outputs can be used for development of efficient learning and inference algorithms. In learning from spatial-temporal data, the Markov Random Fields [31] and the Conditional Random Fields (CRF) [32] are among the most popular graphical models (Fig. 4).

Conditional Random Fields

Originally, CRF were designed for classification of sequential data [32] and have found many applications in areas such as computer vision [33], natural language processing [34], and computational biology [35]. CRFs are a type of discriminative undirected probabilistic graphical model defined over graph G as

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \sum_{a=1}^{A} \psi_a(y_a, x_a), \quad \psi_a(y_a, x_a) = \exp\left(\sum_{k=1}^{K(A)} \theta_{ak} f_{ak}(y_a, x_a) \right), \tag{4.1}$$

where $Z(x)$ is a normalization constant and $\{\Psi_a\}$ is a set of factors in G. The feature functions f_{ak} and the weights θ_{ak} are indexed by the factor index a (each factor has its own set of weights) and k (there are K feature functions).

Construction of appropriate feature functions in CRF is a manual process that depends on prior beliefs of a practitioner about what features could be useful. The choice of features is often constrained to simple constructs to reduce the complexity

of learning and inference from CRF. In general, to evaluate $P(\mathbf{y}|\mathbf{x})$ during learning and inference, one would need to use time consuming sampling methods. This problem can be accomplished in a computationally efficient manner for real-valued CRFs.

CRF for regression is a less explored topic. The Conditional State Space Model (CSSM) [36], an extension of the CRF to a domain with the continuous multivariate outputs, was proposed for regression of sequential data. Continuous CRF (CCRF) [37] is a ranking model that takes into account relations among ranks of objects in document retrieval. In [33], a conditional distribution of pixels given a noisy input image is modeled using the weighted quadratic factors obtained by convolving the image with a set of filters. Feature functions in [33] were specifically designed for image de-noising problems and are not readily applicable to regression.

Most CRF models represent linear relationships between attributes and outputs. On the other hand, in many real-world applications this relationship is highly complex and nonlinear and cannot be accurately modeled by a linear function. CRF that models nonlinear relationship between observations and outputs has been applied to the problem of image de-noising [33]. Integration of CRF and Neural Networks (CNF) [38–40] has been proposed for classification problems to address these limitations by adding a middle layer between attributes and outputs. This layer consists of a number of gate functions each acting as a hidden neuron, that captures the nonlinear relationships. As a result, such models can be much more computationally expressive than regular CRF.

4.2 Gaussian Conditional Random Fields (GCRF)

4.2.1 Continuous Conditional Random Fields Model

Conditional Random Fields (CRF) provide a probabilistic framework for exploiting complex dependence structure among outputs by directly modeling the conditional distribution $P(\mathbf{y}|\mathbf{x})$. In regression problems, the output y_i is associated with input vectors $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ by a real-valued function called *association potential* $A(\boldsymbol{\alpha}, y_i, \mathbf{x})$, where $\boldsymbol{\alpha}$ is K-dimensional set of parameters. The larger the value of A is the more y_i is related to $\mathbf{x}$. In general, A is a combination of functions and it takes as input all input data $\mathbf{x}$ to predict a single output y_i meaning that it does not impose any independency relations among inputs $\mathbf{x_i}$ (Figs. 5 and 6).

To model interactions among outputs, a real valued function called *interaction potential* $I(\boldsymbol{\beta}, y_i, y_j, \mathbf{x})$ is used, where $\boldsymbol{\beta}$ is an L dimensional set of parameters. Interaction potential represents the relationship between two outputs and in general can depend on an input $\mathbf{x}$. For instance, interaction potential can be modeled as a correlation between neighboring (in time and space) outputs. The larger the value of the interaction potential, the more related outputs are.

For the defined association and interaction potentials, *continuous CRF* models a conditional distribution $P(\mathbf{y}|\mathbf{x})$:

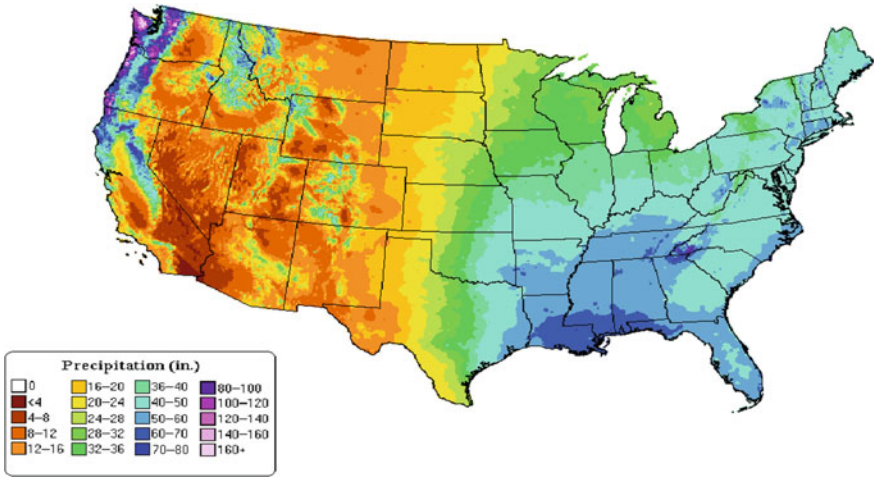

Fig. 5 U.S. average annual precipitation (1971–2000) [41]: spatial relationship between outputs of precipitation measurement stations over U.S.

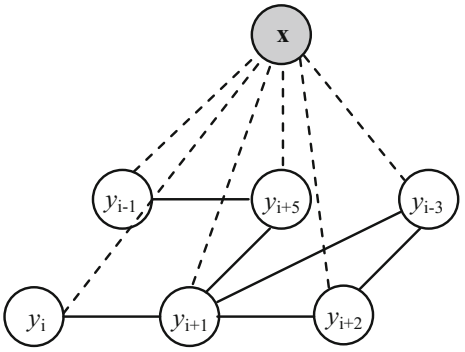

Fig. 6 Continuous CRF graphical structure. x-inputs (observations); y-outputs; *dashed lines*-associations between inputs and outputs; *solid lines*-interactions between outputs, [57]

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x},\alpha,\beta)} \exp\left(\sum_{i=1}^{N} A(\alpha, y_i, \mathbf{x}) + \sum_{j\sim i} I(\beta, y_i, y_j, \mathbf{x})\right) \tag{4.2}$$

where $\mathbf{y} = (y_1, \ldots, y_N)$, $j \sim i$ denotes the connected outputs y_i and y_j (connected with solid line at Fig. 6) and $Z(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is normalization function defined as

$$Z(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \int_{y} \exp\left(\sum_{i=1}^{N} A(\boldsymbol{\alpha}, y_i, \mathbf{x}) + \sum_{j\sim i} I(\boldsymbol{\beta}, y_i, y_j, \mathbf{x})\right) dy \tag{4.3}$$

The learning task is to choose values of parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to maximize the conditional log-likelihood of the set of training examples

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \sum \log P(\mathbf{y}|\mathbf{x}) \\ (\hat{\alpha}, \beta) &= \arg\max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} (L(\boldsymbol{\alpha}, \boldsymbol{\beta})). \end{aligned} \tag{4.4}$$

This can be achieved by applying standard optimization algorithms such as gradient descent. To avoid overfitting, L($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$) is regularized by adding $\boldsymbol{\alpha}^2/2$ and $\boldsymbol{\beta}^2/2$ terms to log-likelihood in formula (4.4) that prevents the parameters from becoming too large.

The inference task is to find the outputs $\mathbf{y}$ for a given set of observations $\mathbf{x}$ and estimated parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ such that the conditional probability $P(\mathbf{y}|\mathbf{x})$ is maximized

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} (P(\mathbf{y}|\mathbf{x})). \tag{4.5}$$

In CRF applications, A and I could be defined as linear combinations of a set of fixed features in terms of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ [42]

$$\begin{aligned} A(\alpha, y_i, \mathbf{x}) &= \sum_{k=1}^{K} \alpha_k f_k(y_i, \mathbf{x}) \\ I(\beta, y_i, y_j, \mathbf{x}) &= \sum_{l=1}^{L} \beta_l g_l(y_i, y_j, \mathbf{x}) \end{aligned} \tag{4.6}$$

This way, any potentially relevant feature could be included to the model because parameter estimation automatically determines their actual relevance by feature weighting.

4.2.2 Association and Interaction Potentials in the GCRF Model

If A and I are defined as quadratic functions of $\mathbf{y}$, $P(\mathbf{y}|\mathbf{x})$ becomes multivariate Gaussian distribution and learning and inference can be accomplished in a computationally efficient manner. In the following, the proposed feature functions that lead to Gaussian CRF are described.

Let us assume we are given K unstructured predictors (e.g. neural network, linear regression or any domain-defined model), $R_k(\mathbf{x})$, $k = 1, \ldots, K$, that predict single output y_i taking into account $\mathbf{x}$ (as a special case, only $\mathbf{x}_i$ can be used as $\mathbf{x}$). To model the dependency between the prediction and output, introduced are quadratic feature functions

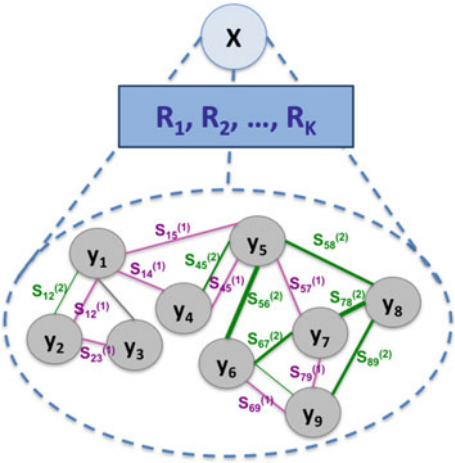

Fig. 7 Gaussian CRF graphical structure: association and interaction potentials are modeled as quadratic feature functions dependent on unstructured predictors R_k and similarity functions S(l), respectively

$$f_k(y_i, \mathbf{x}) = -(y_i - R_k(\mathbf{x}))^2, \quad k = 1, \ldots, K. \tag{4.7}$$

These feature functions follow the basic principle for association potentials—their values are large when predictions and outputs are similar. To model the correlation among outputs, introduced are the quadratic feature function

$$\begin{aligned} g_l(y_i, y_j, \mathbf{x}) = -e_{ij}^{(l)} S_{ij}^{(l)}(\mathbf{x})(y_i - y_j)^2, \quad & e_{ij}^{(l)} = 1 \quad \text{if } (i,j) \in \mathrm{G}_l, \\ & e_{ij}^{(l)} = 0, \quad \text{otherwise} \end{aligned} \tag{4.8}$$

that imposes that outputs y_i and y_j have similar values if they have an edge in the graph. Note that multiple graphs can be used to model different aspects of correlation between outputs (for example, spatial and temporal). $S_{ij}^{(l)}(\mathbf{x})$ function represents similarity between outputs y_i and y_j, that depends on inputs $\mathbf{x}$. The larger $S_{ij}^{(l)}(\mathbf{x})$ is, the more similar the outputs y_i and y_j are (Fig. 7).

4.2.3 Gaussian Canonical Form

$P(\mathbf{y}|\mathbf{x})$ for CRF model (4.2), which uses quadratic feature functions, can be represented as a multivariate Gaussian distribution. The resulting CRF model can be written as

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z}\exp\left(-\sum_{i=1}^{N}\sum_{k=1}^{K} \alpha_k (y_i - R_k(\mathbf{x}))^2 - \sum_{i,j}\sum_{l=1}^{L} \beta_l e_{ij}^{(l)} S_{ij}^{(l)}(\mathbf{x})(y_i - y_j)^2 \right) \tag{4.9}$$

The exponent in (4.9), which will be denoted as E, is a quadratic function in terms of $\mathbf{y}$. Therefore, $P(\mathbf{y}|\mathbf{x})$ can be transformed to a Gaussian form by representing E as

$$E = -\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{T}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}) = -\frac{1}{2}\mathbf{y}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{y} + \mathbf{y}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + const \tag{4.10}$$

To transform $P(\mathbf{y}|\mathbf{x})$ to the Gaussian form, $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ are determined by matching (4.9) and (4.10). The quadratic terms of $\mathbf{y}$ in the association and interaction potentials as represented as $-\mathbf{y}^{\mathrm{T}}\mathbf{Q}_1\mathbf{y}$ and $-\mathbf{y}^{\mathrm{T}}\mathbf{Q}_2\mathbf{y}$, respectively, and combined to get

$$\boldsymbol{\Sigma}^{-1} = 2(\mathbf{Q}_1 + \mathbf{Q}_2). \tag{4.11}$$

By combining the quadratic terms of $\mathbf{y}$ from the association potential, it follows that $\mathbf{Q}_1$ is diagonal matrix with elements

$$Q_{1ij} = \begin{cases} \sum_{k=1}^{K} \alpha_k, & i = j \\ 0, & i \neq j. \end{cases} \tag{4.12}$$

By repeating this for the interaction potential, it follows that $\mathbf{Q}_2$ is symmetric with elements

$$Q_{2ij} = \begin{cases} \sum_{k} \sum_{l=1}^{L} \beta_l e_{ik}^{(l)} S_{ik}^{(l)}(\mathbf{x}), & i = j \\ -\sum_{l=1}^{L} \beta_l e_{ij}^{(l)} S_{ij}^{(l)}(\mathbf{x}), & i \neq j. \end{cases} \tag{4.13}$$

To get $\boldsymbol{\mu}$, linear terms in E are matched with linear terms in the exponent of (4.9)

$$\boldsymbol{\mu} = \Sigma\mathbf{b}, \tag{4.14}$$

where $\mathbf{b}$ is a vector with elements

$$b_i = 2\left(\sum_{k=1}^{K} \alpha_k R_k(\mathbf{x})\right). \tag{4.15}$$

By calculating Z using the transformed exponent, it follows

$$Z(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{x}) = (2\pi)^{N/2}|\boldsymbol{\Sigma}|^{1/2}\exp(const). \tag{4.16}$$

Since exp(*const*) terms from Z and $P(\mathbf{y}|\mathbf{x})$ cancel out

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{(2\pi)^{N/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}(\mathbf{y}-\boldsymbol{\mu}) \right), \tag{4.17}$$

where $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ are defined in (4.11) and (4.14). Therefore, the resulting conditional distribution is Gaussian with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Observe that $\boldsymbol{\Sigma}$ is a function of parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, interaction potential graphs G_l, and similarity functions S, while $\boldsymbol{\mu}$ is also a function of inputs $\mathbf{x}$. The resulting CRF is called *the Gaussian CRF* (GCRF).

4.2.4 Learning and Inference

The **learning** task is to choose $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to maximize the conditional log-likelihood,

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) = \arg\max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} (L(\boldsymbol{\alpha}, \boldsymbol{\beta})), \text{ where } L(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum \log P(\mathbf{y}|\mathbf{x}). \tag{4.18}$$

In order for the model to be feasible, the precision matrix $\boldsymbol{\Sigma}^{-1}$ has to be positive semi-definite. $\boldsymbol{\Sigma}^{-1}$ is defined as a double sum of $\mathbf{Q}_1$ and $\mathbf{Q}_2$. $\mathbf{Q}_2$ is a symmetric matrix with the property that the absolute value of a diagonal element is equal to the sum of absolute values of non-diagonal elements from the same row.

By Gershgorin's circle theorem [43], a symmetric matrix is positive semi-definite if all diagonal elements are non-negative and if matrix is diagonally dominant. Therefore, one way to ensure that the GCRF model is feasible is to impose the constraint that all elements of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are greater than 0. In this setting, learning is a constrained optimization problem. To convert it to the unconstrained optimization, a technique used in [37] is adopted that applies the exponential transformation on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ parameters to guarantee that they are positive

$$\begin{aligned} \alpha_k &= e^{u_k}, \text{ for } k = 1, \ldots, K \\ \beta_l &= e^{v_l}, \text{ for } l = 1, \ldots, L, \end{aligned} \tag{4.19}$$

where u and v are real valued parameters. As a result, the new optimization problem becomes unconstrained.

All parameters are learned by the gradient-based optimization. Conditional log-likelihood form is

$$\log P = -\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{T}\boldsymbol{\Sigma}^{-\mathbf{1}}(\mathbf{y}-\boldsymbol{\mu}) - \frac{1}{2}\log|\boldsymbol{\Sigma}|. \tag{4.20}$$

Derivative of log-likelihood form is given by

$$d\ \log P = -\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T d\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}) + (d\mathbf{b}^T - \boldsymbol{\mu}^T d\boldsymbol{\Sigma}^{-1})(\mathbf{y}-\boldsymbol{\mu}) + \frac{1}{2}Tr(d\boldsymbol{\Sigma}^{-1}\cdot\boldsymbol{\Sigma}) \tag{4.21}$$

From (4.21) derivatives $\partial \log P/\partial\alpha_k$ and $\partial \log P/\partial\beta_l$ can be calculated. The expression for $\partial \log P/\partial\alpha_k$ is

$$\frac{\partial\ \log P}{\partial\alpha_k} = -\frac{1}{2}(\mathbf{y}-\mu)^T\frac{\partial\,\Sigma^{-1}}{\partial\alpha_k}(\mathbf{y}-\mu) + \left(\frac{\partial\mathbf{b}^T}{\partial\alpha_k} - \mu^T\frac{\partial\,\Sigma^{-1}}{\partial\alpha_k}\right)(\mathbf{y}-\mu) + \frac{1}{2}Tr\left(\Sigma\cdot\frac{\partial\,\Sigma^{-1}}{\partial\alpha_k}\right). \tag{4.22}$$

To calculate $\partial \log P/\partial\beta_l$, use $\partial\mathbf{b}/\partial\beta_l = 0$ to obtain

$$\frac{\partial\ \log P}{\partial\beta_l} = -\frac{1}{2}(\mathbf{y}+\boldsymbol{\mu})^T\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\beta_l}(\mathbf{y}-\boldsymbol{\mu}) + \frac{1}{2}Tr\left(\boldsymbol{\Sigma}\cdot\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\beta_l}\right). \tag{4.23}$$

Gradient ascent algorithm cannot be directly applied to a constrained optimization problem [37]. Here, previously defined exponential transformation on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is used and then gradient ascent is applied. Specifically, the maximization of log-likelihood with respect to $u_k = \log\alpha_k$ and $v_l = \log\beta$ instead to α_k and β_l is performed. As a result, the new optimization problem becomes unconstrained. Derivatives of log-likelihood function and updates of α's and β in gradient ascent can be computed as

$$\begin{aligned} &u_k = \log\alpha_k, \quad v_l = \log\beta_l \\ &u_k^{new} = u_k^{old} + \eta\frac{\partial L}{\partial u_k}, \quad \frac{\partial L}{\partial u_k} = \frac{\partial L}{\partial \log\alpha_k} = \alpha_k\frac{\partial L}{\partial\alpha_k} \\ &v_k^{new} = v_k^{old} + \eta\frac{\partial L}{\partial v_l}, \quad \frac{\partial L}{\partial v_l} = \frac{\partial L}{\partial \log\beta_l} = \beta_l\frac{\partial L}{\partial\beta_l}, \end{aligned} \tag{4.24}$$

where η is the learning rate.

The negative log-likelihood is a convex function of parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and its optimization leads to globally optimal solution.

In inference, since the model is Gaussian, the prediction will be expected value, which is equal to the mean $\boldsymbol{\mu}$ of the distribution,

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}) = \Sigma\mathbf{b}. \tag{4.25}$$

Vector $\boldsymbol{\mu}$ is a point estimate that maximizes $P(\mathbf{y}|\mathbf{x})$, while $\boldsymbol{\Sigma}$ is a measure of uncertainty. The simplicity of inference that can be achieved using matrix computations is in contrast to a general CRF model defined in (4.2) that usually requires advanced inference approaches such as Markov Chain Monte Carlo or belief

propagation. Moreover, by exploiting the sparsity of precision matrix Q, which is inherent to spatio-temporal data, the inference can be performed without the need to calculate $\mathbf{\Sigma}$ explicitly which reduces computational time to even linear with the dimensionality of $\mathbf{y}$ (depends on the level of sparsity).

4.2.5 GCRF Extensions

The previously defined GCRF model was first published in [37]. If size of the training set is N and gradient ascent lasts T iterations, training the GCRF model introduced at [37] requires $O(T \cdot N^3)$ time. The main cost of computation is matrix inversion, since during the gradient-based optimization we need to find $\mathbf{\Sigma}$ as an inverse of $\mathbf{\Sigma^{-1}}$. However, this is the worst-case performance. Since matrix $\mathbf{\Sigma^{-1}}$ is typically very sparse, the training time for sparse networks can be decreased to $O(T \cdot N^2)$. The total computation time of this model depends on the neighborhood structure of the interaction potential in GCRF. For example, GCRF computational complexity is $O(T \cdot N^{3/2})$ if the neighborhood is spatial and $O(T \cdot N^2)$ if it is spatio-temporal [44].

Several GCRF extensions were developed to speed-up the learning and inference, and increase the representational power of the model. A continuous CRF for *efficient* regression in large fully connected graphs was introduced via Variational Bayes *approximation* of the conditional distribution [45]. A distributed GCRF *approximation* approach was proposed based on *partitioning* large evolving graphs [46]. More recently, the *exact fast* solution has been obtained by extending the modeling capacity of the GCRF while learning a diagonalized precision matrix faster, which also enabled inclusion of *negative interactions* in a network [47].

Representational power of GCRF was further extended by *distance-based* modeling of interactions in structured regression to allow non-zero mean parameters in the interaction potential to further increase the prediction accuracy [48]. Structured GCRF regression was also recently enabled on *multilayer* networks by extending the model to accumulate information received from each of the layers instead of averaging [49]. A deep Neural GCRF extension was introduced to *learn* an *unstructured* predictor while learning the GCRF's objective function [50]. This was further extended by *joint learning* of *representation and structure* for sparse regression on graphs [51]. This is achieved by introducing hidden variables that are nonlinear functions of explanatory variables and are linearly related with the response variables. A semi-supervised learning algorithm was also developed for structured regression on partially observed attributed graphs by marginalizing out missing label instances from the joint distribution of labeled and unlabeled nodes [52]. Finally, a GCRF extension for modeling the *uncertainty propagation* in *long-term* structured *regression* was developed by modeling *noisy inputs,* and applied for regression on evolving networks [53].

5 Applications and Results

5.1 Insulation Coordination

5.1.1 Introduction

The integrity of an overhead transmission line is directly governed by the electrical and mechanical performance of insulators. More than 70% of the line outages and up to 50% of line maintenance costs are being caused by the insulator-induced outages [54].

In this section the Big Data application for insulation coordination is presented. Proposed method in [55] changes the insulator strength level during the insulator lifetime reflecting how weather disturbances are reducing the insulator strength. Historical data is analyzed to observe cumulative changes in the power network vulnerability. Regression is used to determine insulator breakdown probability. The data we used included Lightning Detection Network Data, weather data, utility fault locators' data, outage data, insulator specification data, GIS data, electricity market data, assets data, and customer impact data. The model includes economic impacts for insulator breakdown.

5.1.2 Modeling

Risk Based Insulation Coordination

The risk assessment framework used for this research is defined as follows [55]:

$$R = P[T] \cdot P[C|T] \cdot u(C) \tag{5.1}$$

where R is the State of Risk for the system (or component), T is the Threat intensity (i.e. lightning peak current), Hazard $P[T]$ is a probability of a lightning strike with intensity T, $P[C|T]$ is the Vulnerability or probability of an insulation total failure if lightning strike with intensity T occurred, and the Worth of Loss, $u(C)$, is an estimate of financial losses in case of insulation total failure.

Lightning Hazard

Probability of a lightning strike is estimated based on historical lightning data in the radius around the affected components. For each node, the lightning frequency is calculated as the ratio between the number of lightning strikes in the area with radius of 100 m around the node and the number of lightning strikes in the total area of the network, Fig. 8. As it can be seen in the Fig. 8, Hazard probability is calculated based on two factors: (1) probability that the lightning will affect the

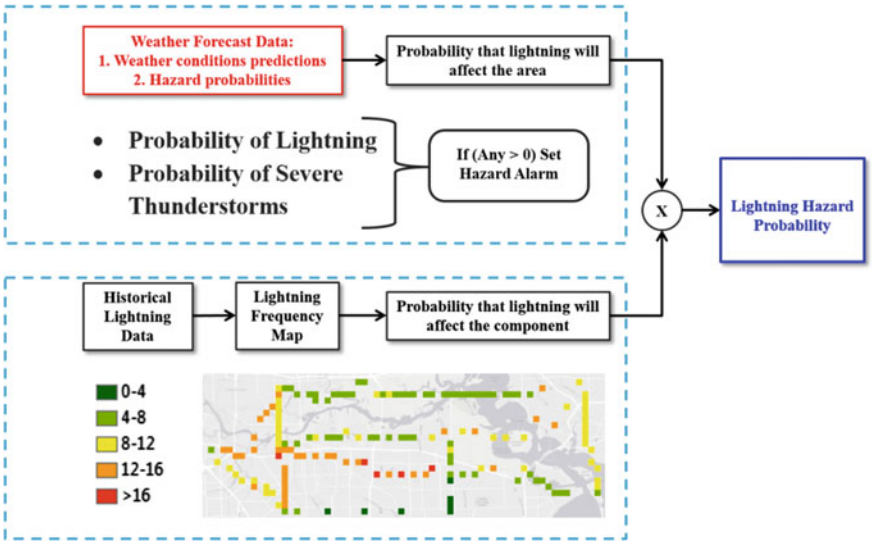

Fig. 8 Lightning Hazard probability

network area determined based on weather forecast, and (2) probability that the lightning will affect a specific component inside the network area determined based on historical lightning frequency data.

Prediction of Vulnerability

The goal of vulnerability part of risk analysis is to determine what impact the severe weather conditions will have on the network. In the case of lightning caused outages, the main goal is to determine what the probability of insulator failure is. Insulation coordination is the study used to select insulation strength to withstand the expected stress. Insulation strength can be described using the concept of Basic Lightning Impulse Insulation Level (BIL), (Fig. 9a) [56]. Statistical BIL represents a voltage level for which insulation has a 90% probability of withstand. Standard BIL is expressed for a specific wave shape of lightning impulse (Fig. 9b), and standard atmospheric conditions.

In order to estimate a new BIL as time progresses (BIL_{new}), data is represented in form of a graph where each node represents a substation or a tower and links between nodes are calculated using impedance matrix as illustrated in Fig. 10.

BILnew in our experiments is predicted using Gaussian Conditional Random Fields (GCRF) based on structured regression [52]. The model captures both the network structure of variables of interest (y) and attribute values of the nodes (x).

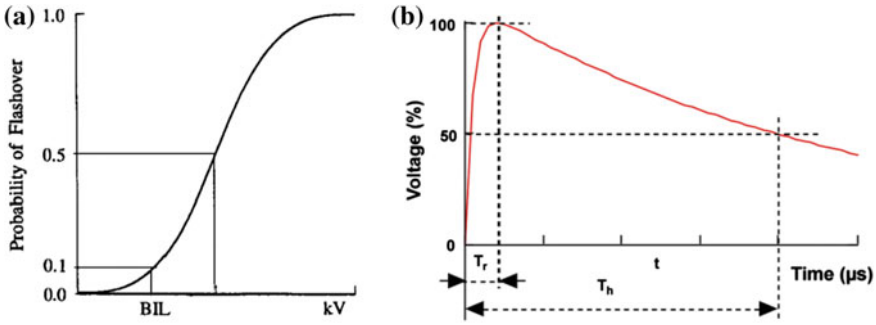

Fig. 9 **a** Basic lightning impulse insulation level (BIL), **b** Standard lightning impulse

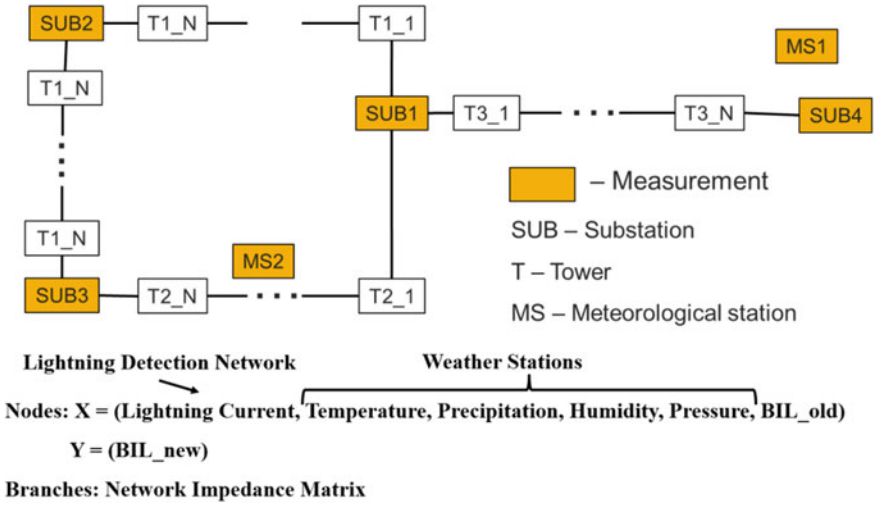

Fig. 10 Network graph representation [55]

It models the structured regression problem as estimation of a joint continuous distribution over all nodes:

$$P(y|x) = \frac{1}{Z}\exp\left(-\sum_{i=1}^{N}\sum_{k=1}^{K}\alpha_k(y_i - R_k(x))^2 - \sum_{i,j}\sum_{l=1}^{L}\beta_l e_{ij}^{(l)} S_{ij}^{(l)}(x)(y_i - y_j)^2\right) \tag{5.2}$$

where the dependence of target variables (y) on input measurements (x) based on k unstructured predictors $R_1, \ldots, R_k$ is modeled by the "association potential" (the first double sum of the exponent in the previous equation).

Economic Impact

Having a model that takes into account economic impact is the key for developing optimal mitigation techniques that minimize the economic losses due to insulation breakdown. Such analysis provides evidence of the Big Data value in improving such applications.

If electric equipment fails directly in face of severe weather changes or indirectly due to the wear-and-tear mechanism, an outage cost would be imposed to the system. For instance, the weather-driven flash-over might happen on an insulator k in the system leading to the corresponding transmission line i outage as well. In order to quantify the outage cost of an electric equipment, say failure of insulator k and accordingly outage of transmission line i at time t, $\Phi^t_{k,i}$, which is the total imposed costs, are quantified in (5.3) comprising of three monetary indices.

$$\Phi^t_{k,i} = C^t_{\mathrm{CM},k,i} + \sum_{\substack{d=1 \\ d \in LP}}^{D} \left(C^t_{\mathrm{LR},k,i} + C^t_{\mathrm{CIC},k,i} \right) \tag{5.3}$$

The first monetary term in (5.3) is a fixed index highlighting the corrective maintenance activities that needs to be triggered to fix the damaged insulator. Such corrective maintenance actions in case of an equipment failure can be the replacement of the affected insulator with a new one and the associated costs may involve the cost of required labor, tools, and maintenance materials. The variable costs [the second term in (5.3)] include the lost revenue cost imposed to the utility ($C^t_{\mathrm{LR},k,i}$) as well as the interruption costs imposed to the affected customers experiencing an electricity outage ($C^t_{\mathrm{CIC},k,i}$). In other words, the cost function $C^t_{\mathrm{LR},k,i}$ is associated with the cost imposed due to the utility's inability to sell power for a period of time and hence the lost revenue when the insulator (and the associated transmission line) is out of service during the maintenance or replacement interval. This monetary term can be calculated using (5.4) [58, 59].

$$C^t_{\mathrm{LR},k,i} = \sum_{\substack{d=1 \\ d \in \mathrm{LP}}}^{D} \left(\lambda^t_d . \mathrm{EENS}^t_{d,k,i} \right) \tag{5.4}$$

where, λ^t_d is the electricity price ($/MWh.) at load point d and $\mathrm{EENS}^t_{d,k,i}$ is the expected energy not supplied (MWh) at load point d due to the failure of insulator k and outage of line i accordingly at time t.

The last variable term of the cost function in (5.3) reflects the customer interruption costs due to the failure of insulator k and corresponding outage of transmission line i at time t which can be calculated in (5.5). $C^t_{\mathrm{CIC},k,i}$ is a function of the EENS index and the value of lost load (VOLL_d) at load point d which is governed by various load types being affected at a load point. The value of lost load

($/MWh.) is commonly far higher than the electricity price and is commonly obtained through customer surveys [59, 60].

$$C^{t}_{\mathrm{CIC},k,i} = \sum_{\substack{d=1 \\ d \in LP}}^{D} \left(\mathrm{VOLL}_{d}.EENS^{t}_{d,k,i}\right) \tag{5.5}$$

The cost function in (5.3), which is actually the failure consequence of an electric equipment (insulator in this case) can be calculated for each equipment failure in the network making it possible to differentiate the impact of different outages (and hazards) on the system overall economic and reliability performance.

5.1.3 Test Setup and Results

The network segment contains 170 locations of interest (10 substations and 160 towers). Historical data is prepared for the period of 10 years, starting from January 1st 2005, and ending with December 31st 2014. Before the first lightning strike, all components are assumed to have a BIL provided from the manufacturer. For each network component, risk value was calculated, assigned, and presented on a map as shown in Fig. 11. In part (a) of Fig. 11, the risk map on January 1st 2009 is presented, while in part (b), the risk map after the last recorded event is presented. With the use of weather forecast, the prediction of future Risk values can be accomplished. In Fig. 11c, the prediction for the next time step is demonstrated. For the time step of interest, the lightning location is predicted to be close to the line 11 (marked with red box in Fig. 11c. Thus, risk values assigned to the line 11 will have the highest change compared to that of the previous step. The highest risk change on line 11 happens for node 72 with changed from 22.8% to 43.5%. The Mean

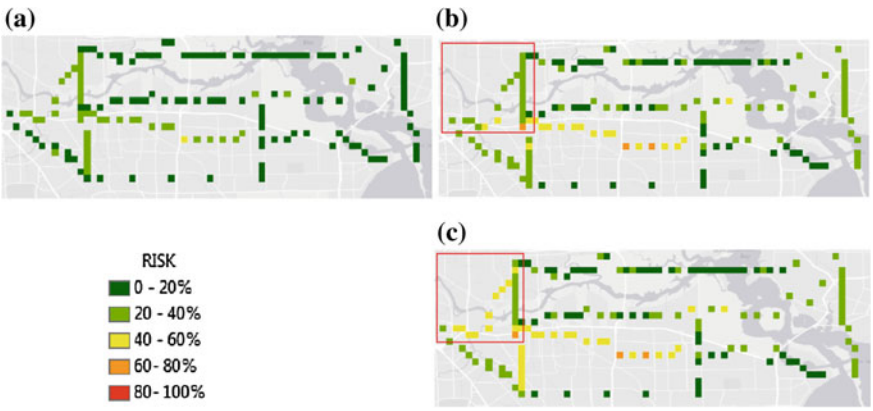

Fig. 11 Results—insulator failure risk maps [55]

Squared Error (MSE) of prediction of GCRF algorithm on all 170 test nodes is 0.0637 + 0.0301 kV when predicting the new value of BIL (BILnew).

5.2 Solar Generation Forecast

5.2.1 Introduction

The solar industry has been growing quite rapidly so far, and consequently, accurate solar generation prediction is playing a more and more important role, aiming at alleviating the potential stress that the solar generation may exert to the grid due to its variability and intermittency nature.

This section presents an example of how to conduct the solar prediction through the introduced GCRF model, while considering both the spatial and temporal correlations among different solar stations. As a data-driven method, the GCRF model needs to be trained with historical data to obtain the necessary parameters, and then the prediction can be accurately conducted through the trained model. The detailed modeling of the association and interaction potentials in this case is introduced, and simulations are conducted to compare the performance of GCRF model with two other forecast models under different scenarios.

5.2.2 Modeling

GCRF is a graphical model, in which multiple layers of graphs can be generated to model the different correlations among inputs and outputs. We are trying to model both the special and temporal correlations among the inputs and the outputs here, as shown in Fig. 12.

In Fig. 12, the red spots labeled in numbers locate different solar stations, in which historical measurements of solar irradiance are available as the inputs. Our goal is to predict the solar irradiance at the next time step as the outputs at different solar stations. Then the prediction of the solar generation can be obtained, since solar generation is closely related to the solar irradiance.

In next subsection, the relationship between the solar generation and solar irradiance is first introduced. Then, the modeling of both the temporal and spatial correlations by GCRF model is presented.

Solar Generation Versus Solar Irradiance

The relationship between the solar generation and the solar irradiance can be approximated in a linear form [61], as calculated in (5.6).

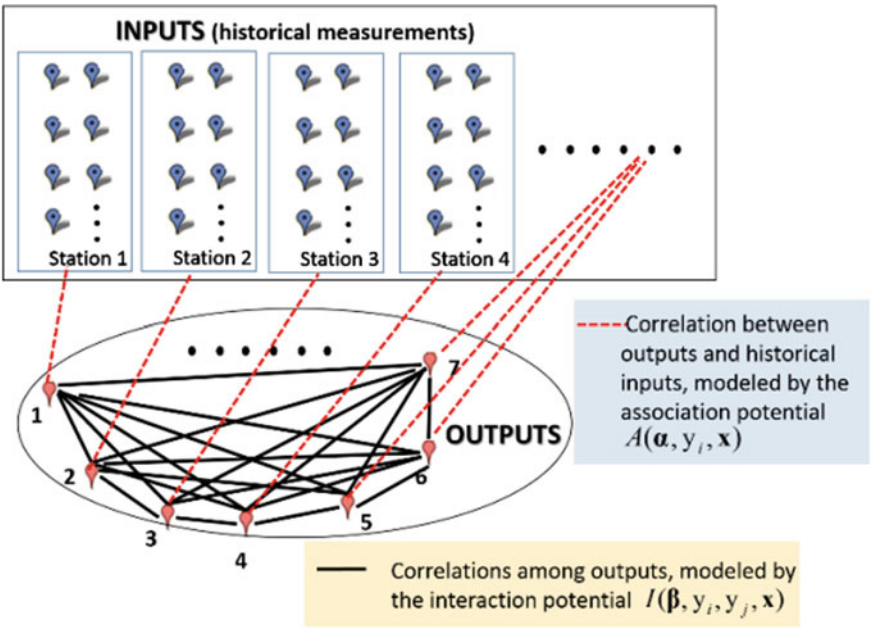

Fig. 12 Spatial and temporal correlations [61]

$$P_{solar} = I_{solar} \times S \times \eta \tag{5.6}$$

where P_{solar} is the power of the solar generation; I_{solar} is the solar irradiance (kWh/m^2); S is the area of the solar panel (m^2); and η is the generation efficiency of a given material.

Temporal Correlation Modeling

The temporal correlation lies in the relationship between the predicted solar irradiance of one solar station and its own historical solar irradiance measurements, as illustrated in red dot lines in Fig. 12. The autoregressive (AR) model [62] is adopted here to model the predictors $R_i(x)$ in the association potential, as denoted in (5.7).

$$R_i(x) = c + \sum_{m=1}^{p_i} \varphi_m y_i^{t-m} \tag{5.7}$$

where p_i is selected to be 10 to consider the previous 10 historical measurements; and φ_m is the coefficient of the AR model.

Spatial Correlation Modeling

The spatial correlation lies in the relationships among the predicted solar irradiance at different solar stations, as illustrated in the black lines in Fig. 12. It can be reasonably assumed that the closer the stations are, the more similar their solar irradiance will be. Therefore, the distance can be adopted to model the similarity between different solar stations, and the S_{ij} in the interaction potential can be calculated in (5.8).

$$S_{ij} = \frac{1}{D_{ij}^2} \tag{5.8}$$

where D_{ij} is the distance between station No. i and No. j.

5.2.3 Test Setup and Results

Hourly solar irradiance data in year 2010 from 8 solar stations have been collected from the California Irrigation Management Information System (CIMIS) [63]. The geographical information is provided in Fig. 13, where the station No. 1 is regarded as the targeted station, and two artificial stations (No. 9 and No. 10) are added very close to station No. 1.

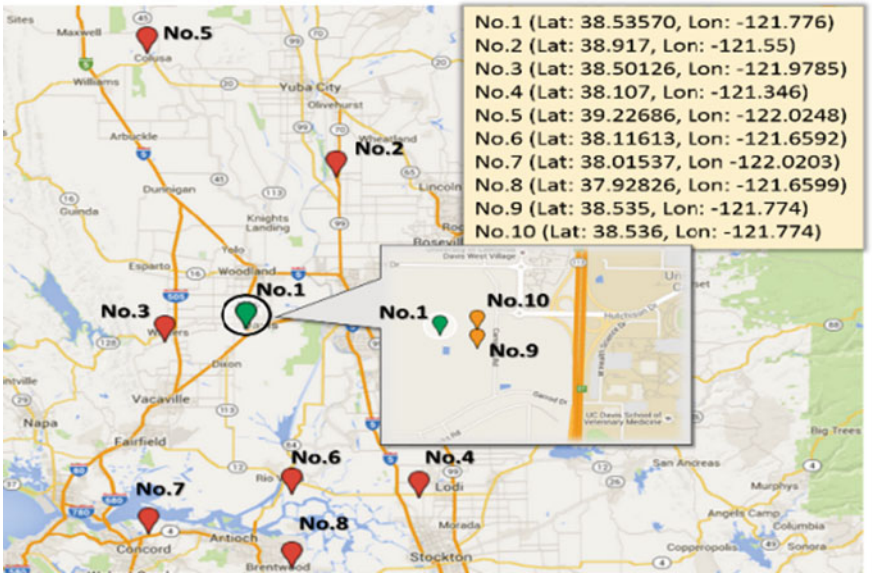

Fig. 13 Geographical information of the test system [61]

Table 4 Training and validation periods [61]

Case	1	2	3	4
Training period	January, March	May	July, September	November
Validation period	February, April	April, June	August, October	October, December

Table 5 Performance indices of various forecasting models: Scenario 1 [61]

Index	Cases	Forecast model		
		PSS	ARX	GCRF
MAE	Case 1	90.3676	56.5334	55.1527
	Case 2	98.1372	51.8562	40.4062
	Case 3	96.6623	35.5478	25.5906
	Case 4	92.8664	51.6816	29.6195
RMSE	Case 1	111.9337	76.7467	74.4007
	Case 2	116.5823	81.9164	60.6969
	Case 3	111.6060	55.8073	40.6566
	Case 4	108.1498	67.8648	43.7008

The scenarios are selected as follows: *Scenario 1*: no missing data; *Scenario 2*: missing data do exist (*Scenario 2-1*: one hourly data set is missing in station No. 1; *Scenario 2-2*: two successive hourly data sets are missing in station No. 1; *Scenario 2-3*: one hourly data set is missing in several stations; *Scenario 2-4*: no data is available in one of those stations.).

Besides, the data obtained have been divided into 4 cases during the training and validation periods, as listed in Table 4. And the performance of the GCRF model will be compared with that of two other models: Persistent (PSS) and Autoregressive with Exogenous input (ARX) models, through the index of the mean absolute errors (MAE) and the root mean square error (RMSE) defined in (5.9) and (5.10). The detailed information regarding to the PSS and ARX models can be found in [64].

$$\mathrm{MAE} = \frac{1}{Z}\sum_{t=1}^{Z} |\widehat{y}_t - y_t| \tag{5.9}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{Z}\sum_{t=1}^{Z} (\widehat{y}_t - y_t)^2}, \tag{5.10}$$

The performances of the three models regarding to Scenario 1 are listed in Table 5, and the detailed performances are illustrated in Fig. 14, in which the green line denotes the ideal prediction result, and the performance is better if it is closer to that line. We can observe clearly that GCRF model outperforms the other two models in Scenario 1.

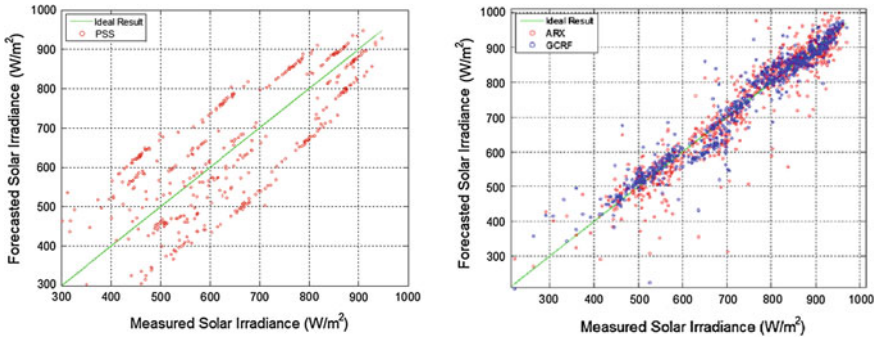

Fig. 14 Prediction performance of GCRF, ARX and PSS models: Scenario1 (Case3) [61]

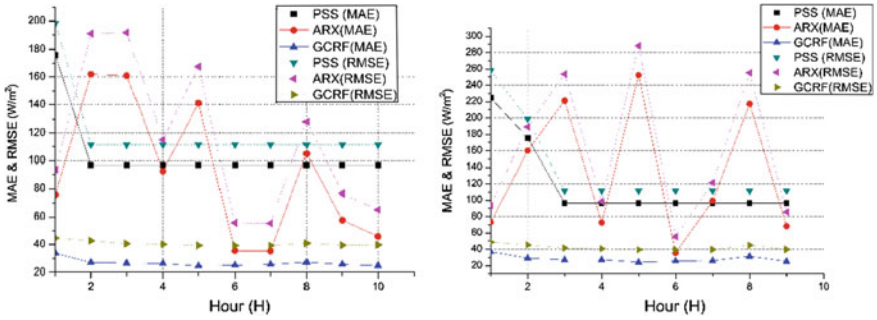

Fig. 15 Prediction performance of GCRF, ARX and PSS models: LEFT—Scenario 2-1 (Case3); RIGHT—Scenarios 2-2 (Case 3) [61]

Figures 15 and 16, Table 6 present the performance results of the three models in Scenario 2 (In Scenario 2-4, the data from Station No. 5 are totally not available). The simulation results under Scenario 2 shows that: (1) GCRF model still has the best performance when missing data exist; (2) the data from Station No. 9 and 10 play an important role. GCRF model works very well when there is no missing data in those two stations, while its performance may also compromise a bit when missing data occur in those two stations, though it still performs the best most of the time. The reason behind is the spatial correlations among those three stations (No. 1, 9 and 10) are strong, since they are physically close to each other. These spatial correlations are modeled and considered in the GCRF model, and therefore, the deviated prediction results, caused by the missing data, can be further adjusted by the strong spatial correlation.

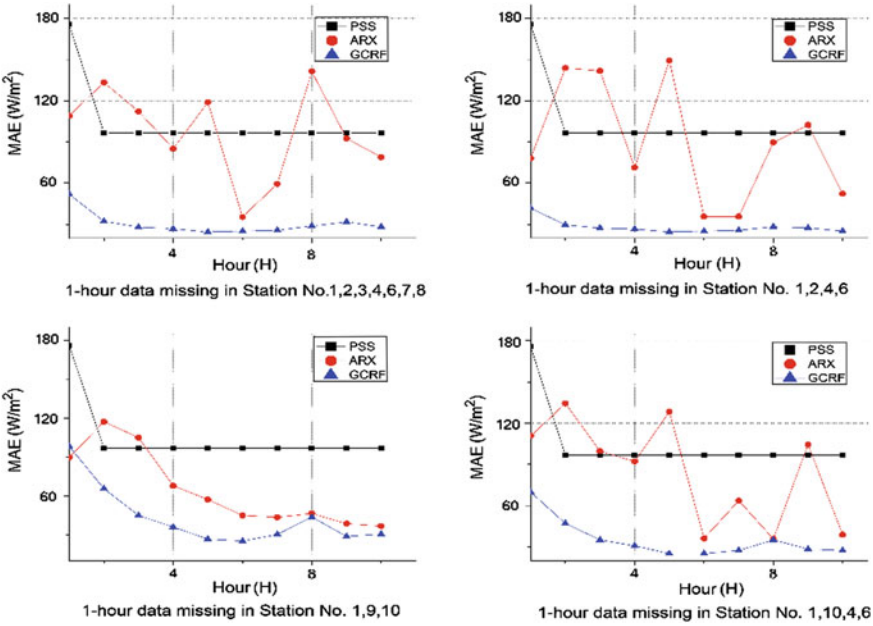

Fig. 16 Prediction performance of GCRF, ARX and PSS models: LEFT—Scenario 2-3 (Case3) [61]

Table 6 Performance indices of various forecasting models: Scenario 2-4 [61]

Index	Cases	Forecast model		
		PSS	ARX	GCRF
MAE	Case 1	96.6623	58.2602	55.9011
	Case 2	96.6623	47.7145	43.0316
	Case 3	96.6623	50.2712	26.8112
	Case 4	96.6623	56.3702	30.7201
RMSE	Case 1	111.6060	79.2030	75.6125
	Case 2	111.6060	76.5143	63.7870
	Case 3	111.6060	68.0714	41.8258
	Case 4	111.6060	72.0233	44.8699

6 Conclusions

In this chapter the application of the Big Data for analyzing weather impacts on power system operation, outage, and asset management has been introduced. Developed methodology exploits spatio-temporal correlation between diverse datasets in order to provide better decision-making strategies for the future smart grid. The probabilistic framework called Gaussian Conditional Random Fields (GCRF) has been introduced and applied to two power system applications:

(1) Risk assessment for transmission insulation coordination, and (2) Spatio-temporal solar generation forecast.

When applied to the insulation coordination problem, proposed model leads to the following contributions:

- Our data analytics changes the insulator strength level during the insulator lifetime reflecting how weather disturbances are reducing the insulator strength. We analyzed historical data to observe cumulative changes in the power network vulnerability to lightning. This allows for a better accuracy when predicting future insulator failures, since the impact of past disturbances is not neglected.
- We used GCRF to determine insulator breakdown probability based on spatiotemporally referenced historical data and real-time weather forecasts. The BD we used included Lightning Detection Network Data, historical weather and weather forecast data, data from utility fault locators, historical outage data, insulator specification data, Geographical Information System (GIS) data, electricity market data, assets replacement and repair cost data, and customer impact data. This was the first time that we are aware such an extensive Big Data set was used to estimate insulator failure probability.
- Our model included economic impacts for insulator breakdown. Having a model that takes into account economic impact is the key for developing optimal mitigation techniques that minimize the economic losses due to insulation breakdown. Such analysis provides evidence of the Big Data value in improving such applications.

When applied to the solar generation prediction, the proposed model leads to the following contributions:

- Not only the temporal but also the spatial correlations among different locations can be modeled, which leads to an improvement in the accuracy of the prediction performance.
- The adoption of the GCRF model can further ensure a good prediction performance even under the scenario with missing data, especially when the spatial correlations are strong. The reason behind is that the deviated prediction results, caused by the missing data, can be adjusted by the strong spatial correlation.

References

1. L. M. Beard et al., "Key Technical Challenges for the Electric Power Industry and Climate Change," IEEE Transactions on Energy Conversion, vol. 25, no. 2, pp. 465–473, June 2010.
2. M. Shahidehpour and R. Ferrero, "Time management for assets: chronological strategies for power system asset management," IEEE Power and Energy Magazine, vol. 3, no. 3, pp. 32–38, May-June 2005.
3. F. Aminifar et al., "Synchrophasor Measurement Technology in Power Systems: Panorama and State-of-the-Art," IEEE Access, vol. 2, pp. 1607–1628, 2014.

4. J. Endrenyi et al., "The present status of maintenance strategies and the impact of maintenance on reliability," IEEE Transactions on Power Systems, vol. 16, no. 4, pp. 638–646, Nov 2001.
5. National Oceanic and Atmospheric Administration, [Online] Available: http://www.noaa.gov/ Accessed 12 Feb 2017
6. National Weather Service GIS Data Portal. [Online] Available: http://www.nws.noaa.gov/gis/ Accessed 12 Feb 2017
7. National Digital Forecast Database. [Online] Available: http://www.nws.noaa.gov/ndfd/ Accessed 12 Feb 2017
8. National Weather Service Doppler Radar Images. [Online] Available. http://radar.weather.gov/ Accessed 12 Feb 2017
9. Data Access, National Centers for Environmental Information. [Online] Available: http://www.ncdc.noaa.gov/data-access Accessed 12 Feb 2017
10. Climate Data Online: Web Services Documentation. [Online] Available: https://www.ncdc.noaa.gov/cdo-web/webservices/v2 Accessed 12 Feb 2017
11. National Centers for Environmental Information GIS Map Portal. [Online] Available: http://gis.ncdc.noaa.gov/maps/ Accessed 12 Feb 2017
12. Satellite Imagery Products. [Online] Available: http://www.ospo.noaa.gov/Products/imagery/ Accessed 12 Feb 2017
13. Lightning & Atmospheric Electricity Research. [Online] Available: http://lightning.nsstc.nasa.gov/data/index.html Accessed 12 Feb 2017
14. National Weather Service Organization. [Online] Available: http://www.weather.gov/organization_prv Accessed 12 Feb 2017
15. Commercial Weather Vendor Web Sites Serving The U.S. [Online] Available: http://www.nws.noaa.gov/im/more.htm Accessed 12 Feb 2017
16. U. Finke, et al., "Lightning Detection and Location from Geostationary Satellite Observations," Institut fur Meteorologie und Klimatologie, University Hannover. [Online] Available: http://www.eumetsat.int/website/wcm/idc/idcplg?IdcService=GET_FILE&dDocName=pdf_mtg_em_rep26&RevisionSelectionMethod=LatestReleased&Rendition=Web Accessed 12 Feb 2017
17. K. L. Cummins, et al., "The US National Lightning Detection NetworkTM and applications of cloud-to-ground lightning data by electric power utilities," IEEE Trans. Electromagn. Compat., vol. 40, no. 4, pp. 465–480, Nov. 1998.
18. Vaisala Inc., "Thunderstorm and Lightning Detection Systems," [Online] Available: http://www.vaisala.com/en/products/thunderstormandlightningdetectionsystems/Pages/default.aspx Accessed 12 Feb 2017
19. Esri, "ArcGIS Platform," [Online] Available: http://www.esri.com/software/arcgis Accessed 12 Feb 2017
20. P.-C. Chen, T. Dokic, N. Stoke, D. W. Goldberg, and M. Kezunovic, "Predicting Weather-Associated Impacts in Outage Management Utilizing the GIS Framework," in Proceeding IEEE/PES Innovative Smart Grid Technologies Conference Latin America (ISGT LATAM), Montevideo, Uruguay, 2015, pp. 417–422.
21. B. Meehan, Modeling Electric Distribution with GIS, Esri Press, 2013.
22. A. von Meier, A. McEachern, "Micro-synchrophasors: a promising new measurement technology for the AC grid," i4Energy Seminar October 19, 2012.
23. Network Time Foundation, "NTP: The Network Time Protocol," [Online] Available: http://www.ntp.org/ Accessed 12 Feb 2017
24. IRIG Standard, "IRIG Serial Time Code Formats," September 2004.
25. IEEE Standards, IEEE 1588-2002, IEEE, 8 November 2002.
26. Q. Yan, T. Dokic, M. Kezunovic, "Predicting Impact of Weather Caused Blackouts on Electricity Customers Based on Risk Assessment," IEEE Power and Energy Society General Meeting, Boston, MA, July 2016.

27. T. Dokic, P.-C. Chen, M. Kezunovic, "Risk Analysis for Assessment of Vegetation Impact on Outages in Electric Power Systems," CIGRE US National Committe 2016 Grid of the Future Symposium, Philadelphia, PA, October–November 2016.
28. National Conference of State Legislatures (NCSL), [Online]. http://www.ncsl.org/research/energy/renewable-portfolio-standards.aspx Accessed 12 Feb 2017
29. International Electrotechnical Commission (IEC), "Grid integration of large-capacity Renewable Energy sources and use of large-capacity Electrical Energy Storage", Oct.1, 2012, [Online]. http://www.iec.ch/whitepaper/pdf/iecWP-gridintegrationlargecapacity-LR-en.pdf Accessed 12 Feb 2017
30. Johan Enslin, "Grid Impacts and Solutions of Renewables at High Penetration Levels", Oct. 26, 2009, [Online]. http://www.eia.gov/energy_in_brief/images/charts/hydro_&_other_generation-2005-2015-large.jpg Accessed 12 Feb 2017
31. A. H. S. Solberg, T. Taxt, and A. K. Jain, "A Markov random field model for classification of multisource satellite imagery," IEEE Transactions on Geoscience and Remote Sensing, vol. 34, no. 1, pp. 100–113, 1996.
32. J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in Proceedings of the 18th International Conference on Machine Learning, 2001, vol. 18, pp. 282–289.
33. M. F. Tappen, C. Liu, E. H. Adelson, and W. T. Freeman, "Learning Gaussian Conditional Random Fields for Low-Level Vision," 2007 IEEE Conference on Computer Vision and Pattern Recognition, vol. C, no. 14, pp. 1–8, 2007.
34. Sutton, Charles, and Andrew McCallum. "An introduction to conditional random fields for relational learning." Introduction to statistical relational learning (2006): 93–128.
35. Y. Liu, J. Carbonell, J. Klein-Seetharaman, and V. Gopalakrishnan, "Comparison of probabilistic combination methods for protein secondary structure prediction," Bioinformatics, vol. 20, no. 17, pp. 3099–3107, 2004.
36. M. Kim and V. Pavlovic, "Discriminative learning for dynamic state prediction," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 10, pp. 1847–1861, 2009.
37. T. Qin, T.-Y. Liu, X.-D. Zhang, D.-S. Wang, and H. Li, "Global Ranking Using Continuous Conditional Random Fields," in Proceedings of NIPS'08, 2008, vol. 21, pp. 1281–1288.
38. T.-minh-tri Do and T. Artieres, "Neural conditional random fields," in Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS), 2010, vol. 9, pp. 177–184.
39. F. Zhao, J. Peng, and J. Xu, "Fragment-free approach to protein folding using conditional neural fields," Bioinformatics, vol. 26, no. 12, p. i310-i317, 2010.
40. J. Peng, L. Bo, and J. Xu, "Conditional Neural Fields," in Advances in Neural Information Processing Systems NIPS'09, 2009, vol. 9, pp. 1–9.
41. http://www.prism.oregonstate.edu/inc/images/gallery_imagemap.png, Accessed 12 Feb 2017
42. S. Kumar and M. Hebert, "Discriminative Random Fields," International Journal of Computer Vision, vol. 68, no. 2, pp. 179–201, 2006.
43. G. H. Golub and C. F. Van Loan, Matrix Computations, vol. 10, no. 8. The Johns Hopkins University Press, 1996, p. 48.
44. H. Rue and L. Held, Gaussian Markov Random Fields: Theory and Applications, vol. 48, no. 1. Chapman & Hall/CRC, 2005, p. 263 p.
45. Ristovski, K., Radosavljevic, V., Vucetic, S., Obradovic, Z., "Continuous Conditional Random Fields for Efficient Regression in Large Fully Connected Graphs," Proc. The Twenty-Seventh AAAI Conference on Artificial Intelligence (AAAI-13), Bellevue, Washington, July 2013.
46. Slivka, J., Nikolic, M., Ristovski, K., Radosavljevic, V., Obradovic, Z. "Distributed Gaussian Conditional Random Fields Based Regression for Large Evolving Graphs," Proc. 14th SIAM Int'l Conf. Data Mining Workshop on Mining Networks and Graphs, Philadelphia, April 2014.

47. Glass, J., Ghalwash, M., Vukicevic, M., Obradovic, Z. "Extending the Modeling Capacity of Gaussian Conditional Random Fields while Learning Faster," Proc. Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16), Phoenix, AZ, February 2016.
48. Stojkovic, I., Jelisavcic, V., Milutinovic, V., Obradovic, Z. "Distance Based Modeling of Interactions in Structured Regression," Proc. 25th International Joint Conference on Artificial Intelligence (IJCAI), New York, NY, July 2016.
49. Polychronopoulou, A, Obradovic, Z. "Structured Regression on Multilayer Networks," Proc. 16th SIAM Int'l Conf. Data Mining (SDM), Miami, FL, May 2016.
50. Radosavljevic, V., Vucetic, S., Obradovic, Z. "Neural Gaussian Conditional Random Fields," Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Nancy, France, September, 2014.
51. Han, C, Zhang, S., Ghalwash, M., Vucetic, S, Obradovic, Z. "Joint Learning of Representation and Structure for Sparse Regression on Graphs," Proc. 16th SIAM Int'l Conf. Data Mining (SDM), Miami, FL, May 2016.
52. Stojanovic, J., Jovanovic, M., Gligorijevic, Dj., Obradovic, Z. "Semi-supervised learning for structured regression on partially observed attributed graphs" Proceedings of the 2015 SIAM International Conference on Data Mining (SDM 2015) Vancouver, Canada, April 30–May 02, 2015.
53. Gligorijevic, Dj, Stojanovic, J., Obradovic, Z."Uncertainty Propagation in Long-term Structured Regression on Evolving Networks," Proc. Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16), Phoenix, AZ, February 2016.
54. R. S. Gorur, et al., "Utilities Share Their Insulator Field Experience," T&D World Magazine, Apr. 2005, [Online] Available: http://tdworld.com/overhead-transmission/utilities-share-their-insulator-field-experience Accessed 12 Feb 2017
55. T. Dokic, P. Dehghanian, P.-C. Chen, M. Kezunovic, Z. Medina-Cetina, J. Stojanovic, Z. Obradovic "Risk Assessment of a Transmission Line Insulation Breakdown due to Lightning and Severe Weather," HICCS – Hawaii International Conference on System Science, Kauai, Hawaii, January 2016.
56. A. R. Hileman, "Insulation Coordination for Power Systems," CRC Taylor and Francis Group, LLC, 1999.
57. Radosavljevic, V., Obradovic, Z., Vucetic, S. (2010) "Continuous Conditional Random Fields for Regression in Remote Sensing," Proc. 19th European Conf. on Artificial Intelligence, August, Lisbon, Portugal.
58. P. Dehghanian, et al., "A Comprehensive Scheme for Reliability Centered Maintenance Implementation in Power Distribution Systems- Part I: Methodology", IEEE Trans. on Power Del., vol.28, no.2, pp. 761–770, April 2013.
59. W. Li, Risk assessment of power systems: models, methods, and applications, John Wiley, New York, 2005.
60. R. Billinton and R. N. Allan, Reliability Evaluation of Engineering Systems: Concepts and Techniques, 2nd ed. New York: Plenum, 1992.
61. B. Zhang, P. Dehghanian, M. Kezunovic, "Spatial-Temporal Solar Power Forecast through Use of Gaussian Conditional Random Fields," IEEE Power and Energy Society General Meeting, Boston, MA, July 2016.
62. C. Yang, and L. Xie, "A novel ARX-based multi-scale spatio-temporal solar power forecast model," in 2012 North American Power Symposium, Urbana-Champaign, IL, USA, Sep. 9–11, 2012.
63. California Irrigation Management Information System (CIMIS), [Online]. Available: http://www.cimis.water.ca.gov/ Accessed 12 Feb 2017
64. C. Yang, A. Thatte, and L. Xie, "Multitime-scale data-driven spatio-temporal forecast of photovoltaic generation," IEEE Trans. Sustainable Energy, vol. 6, no. 1, pp. 104–112, Jan. 2015.