Systematic Integration of Large Data Sets for Improved Decision-Making

Final Project Report

Power Systems Engineering Research Center

Empowering Minds to Engineer
the Future Electric Energy System

Systematic Integration of Large Data Sets for Improved Decision-Making

Final Project Report

Project Team

Faculty:

Mladen Kezunovic, Project Leader
Le Xie
Texas A&M University

Santiago Grijalva, Polo Chau
Georgia Institute of Technology

Students:

Tatjana Dokic, Yingzhong Gu
Texas A&M University

Arjun Anand, Shang-Tse Cheng
Georgia Institute of Technology

PSERC Publication 15-05

September 2015

For information about this project, contact

Dr. Mladen Kezunovic
Department of Electrical and Computer Engineering
Texas A&M University
Phone: 979-845-7509
kezunov@ece.tamu.edu

Power Systems Engineering Research Center

The Power Systems Engineering Research Center (PSERC) is a multi-university Center conducting research on challenges facing the electric power industry and educating the next generation of power engineers. More information about PSERC can be found at the Center's website: http://www.pserc.org.

For additional information, contact:

Power Systems Engineering Research Center
Arizona State University
527 Engineering Research Center
Tempe, Arizona 85287-5706
Phone: 480-965-1643
Fax: 480-965-0745

Notice Concerning Copyright Material

Acknowledgements

This is the final report for the Power Systems Engineering Research Center (PSERC) research project titled "Systematic Integration of Large Data Sets for Improved Decision-Making" (project T-51). We express our appreciation for the support provided by PSERC's industry members and by the National Science Foundation under the Industry/University Cooperative Research Center program. In particular, we wish to thank industry advisors Feng Gao (ABB), Jay Giri (ALSTOM Grid), William Kouam Kamwa (AEP), David Martinez (Southern California Edison), Eugene Litvinov (ISO New England), Mahendra Patel (EPRI), and Bill Timmons (Western Area Power Authority).

Executive Summary

This project is focused on the Big Data uses in power systems. To be able to research this topic that spans across many applications and yet to stay within the scope of the project, three issues were addressed by the team of researchers: Outage and Asset Management (Mladen Kezunovic and his students), Short-term Spatio-temporal wind forecast (Le Xie and his students), and Distributed Database for Future Grid (Santiago Grijalva and Polo Chau and their students). Before specific results of the study are addressed, the report has offered a definition of the Big Data problem in the utility industry in Section 2. In Section 3, the improved fault location using lightning data for transmission network is presented. In Section 4, a methodology for evaluating risk of insulation breakdown is discussed. Section 5 describes big data application to wind power forecast and look-ahead power system dispatch. Section 6 describes implementation of distributed database for future grid using smart meter data as an example. Conclusions are given in Section 7.

As an introduction to the Big Data problem, it was pointed out that spatial and temporal aspects are most critical part of the use of Big Data. This is particularly challenging when the data from various sources are integrated. This is illustrated by an example where the data from various power system field measurement infrastructures is merged with data that comes from the sources outside the utility owned database. Examples of the utility sources of field measurement data of interest were supervisory control and data acquisition system, synchrophasor measurement system, automated smart metering system, etc. Examples of the non-utility owned databases are the lightning detection network, weather data coming from the various government services, landscape and vegetation data, etc. It has been pointed out that the uses of integrated data require innovative data analytics to deal with Big Data properties such as large volume, high velocity and increasing variety. The study has also pointed out that handling Big Data requires very close attention to the framework for spatial representation, namely the Geographic Information System (GIS) implementation, and the framework for temporal correlation, namely the Global Position System (GPS) time-synchronization.

In the area of outage and asset management, two important applications are addressed: fault location with improved accuracy and asset management with improved probability of failure assessment. For the fault location example, a traveling wave fault locator performance was enhanced by correlating the results with the location estimation from the lightning detection network data. It was demonstrated that using the data from the two sources enhances the accuracy of the fault location estimation. In the second example, monitoring of the failure probability of transmission line insulators was considered. It was shown that by correlating the insulator modeling data with the environmental data from the weather stations produces better failure rate estimates than what is possible to achieve with the traditional failure rate estimates. In both applications new data analytics were introduced, and it was shown how the Big Data had to be organized, managed and processed to obtain better results than what the traditional approaches offer.

The next part of the report is dealing with the use of enhanced data sets to improve the wind forecasts. Two classes of spatio-temporal wind forecast models, namely the

trigonometric direction diurnal (TDD) and TDD with geostrophic wind information (TDDGW) models are used and critically evaluated. It is shown that by incorporating spatial correlations of neighboring wind farms, the forecast quality in the near-term (hours-ahead) could be improved. The TDD and TDDGW models are incorporated into a robust look-ahead economic dispatch and a day-ahead reliability unit commitment. Compared with conventional temporal-only statistical wind forecast models, such as the PSS models, the spatio-temporal models consider both the local and geographical wind correlations. By leveraging both temporal and spatial wind historical data, more accurate wind forecasts can be obtained.

In the last section the issue how to organize and utilize Big Data in the distribution area, illustrated by the data coming from the automated meter reading systems is studied. The problem is framed in the context of the next generation Distributed Grid Control and Management Architecture. Particular focus was on the future database management techniques that can facilitate the use of Big Data. It was shown that by using open-source distributed databases such as HBase, we can get reasonably good performance in real-time data storage. The capability and speed of the database increases almost linearly as we increase the number of machines. Using HBase also comes with several other benefits, such as fault tolerance and the integration with other data processing tools.

The study has provided different aspect of the uses of Big Data ranging from the various applications in the transmission and distribution systems and as an outcome several conclusions may be reached:

- The use of Big Data in the utility industry requires new spatial and temporal integration approaches, and corresponding new data analytics
- The data sets that constitute Big Data are quite diverse in their properties, so special care needs to be placed on advanced database management techniques
- The use of Big data offers distinct performance benefits but also carries additional cost, so a judgment about most suitable applications needs to be exercised

Project Publications

- P. Chen, T. Dokic, N. Stokes, D. W. Goldberg, M. Kezunovic. “Predicting Weather-Associated Impacts in Outage Management Utilizing the GIS Framework,” IEEE/PES Innovative Smart Grid Technologies Latin America (ISGT-LA), Montevideo, Uruguay, October 2015.
- P. Chen, T. Dokic, M. Kezunovic. “The Use of Big Data for Outage Management in Distribution Systems”, CIRED Workshop on Challenges of Implementing Active Distribution System Management, Rome, Italy, June 2014.
- T. Dokic, P. Dehghanian, P. 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. (Accepted)
- M. Kezunovic, T. Dokic, P. Chen, V. Malbasa. "Improved Transmission Line Fault Location Using Automated Correlation of Big Data from Lightning Strikes and Fault-induced Traveling Waves," Hawaii International Conference on System Science (HICCS), Kauai, Hawaii, January 2015.
- M. Kezunovic; L. Xie, S. Grijalva. “The Role of Big Data in Improving Power System Operation and Protection,” 2013 IREP Symposium-Bulk Power System Dynamics and Control (IREP), Rethymnon, Greece, August 2013.
- X. Zhu, Genton, M. G., Gu, Y., and Xie, L. “Space-Time Wind Speed Forecasting for Improved Power System Dispatch,” *Test 23*, no. 1, pp. 1-25, Feb. 2014.

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 Project Objectives

The utility industry is encountering new challenges in dealing with extremely large data sets, often called big data. In [1], the authors have denoted that two groups of data can be distinguished: the first group stems from the utility measurement infrastructure, and the second group coming from other source of data not necessarily being part of utility infrastructure.

Examples of the first group of data experienced by the utility industry include the following: (1) synchrophasor measurement system (SMS) data that is used in addition to traditional supervisory control and data acquisition (SCADA) data for situational awareness, (2) data sets collected by poling automated revenue metering (ARM) systems for billing purposes, (3) transient recorder data used for fault location, (4) asset management data that may consist of condition-based measurements collected from intelligent electronic devices (IED) as well as nameplate and maintenance data entered off-line.

The second group of data sets, not obtained through the utility field measurement infrastructure but widely used in decision making, include weather data, Geographic Information System (GIS) data [2]-[4], data from the National Lightning Detection Network (NLDN) [5], landscape and vegetation data [ref], and electricity market data [6]. From [7], correlation of data in time and space and providing unified and generalized modeling, is essential in implementing data analytics for various electric utility applications.

The objective of this research is to illustrate the correlation of various data sets in order to satisfy specific requirements for data analytics in the electric power industry.

1.1.1 The Challenge of the Big Data Integration

The big data exhibits following characteristics, [7], [8]:

- Large volume: describes the quantity of collected data that can reach several gigabytes or even terabytes,
- High velocity: describes the rate at which data is collected and processed typically expressed in terms of number of samples per second,
- Increasing variety: describes the heterogeneity of analyzed data including many different data sources that follow different standards for data representation.

Table 1 lists estimation for big data characteristics of several data sources that are of interest for electric utilities. One may see that the amount of data and velocity with which the data is generated can be overwhelming for both on-request and real-time applications.

Table 1: The big data 3 Vs

		VOLUME	VELOCITY
V A R I E T Y	Smart meters	0.5 million devices can generate more than 120 GB of data per day	Collecting meter data at the interval of 5-15 min
	Synchronized phasor measurement	100 devices can collect more than 2.5 GB of phasor data per day	Up to 240 samples per second
	GIS and GPS data	Additional GIS layer for every type of data	GPS time tag accuracy of 100ns
	Weather data	Only one radar can generate over 50 GB of data per day	New data every 1 to 10 min
	Seismic data	Up to 200 GB of raw imagery per day	Sample rate from 0.01 to 100 samples per second

1.2 Targets of Research

1.2.1 Outage and Asset Management

The integration of weather data, condition based maintenance data and real-time substation IED data correlated within a geographic context can lead to improved outage and asset management decision-making. Useful data for the integration are:

- Geographic information about utility network asset placement;
- Condition based maintenance data for assets;
- Substation IED data concurrent to the lightning strikes that affected the network.

Additional data that is used to correlate the events are:

- Geographic Information System (GIS) data;
- Data from weather stations;
- Data from the National Lightning Detection Network (NLDN).

The goal of this effort is an integration of relevant data within a single unified model and an investigation of using such modeling to improve outage and asset management decision-making. Following are the outcomes of the research:

- Ability to evaluate performance of traveling wave fault location technique in the presence of lightning strikes and severe storms. This investigation is conducted to explore how additional data gathered from unconventional sources can improve the performance of this technique. Additional data can improve accuracy of the traveling wave fault location.
- Methodology for predicting impact on transmission line insulators, power and instrument transformer bushings, surge arresters, etc. Probability of failure after repeated lightning strikes is calculated predicting damage caused by the fault and overvoltage transient flashovers. Maintenance scheduling optimization approach based on risk based calculation. The risk based approach allows the crew scheduling prioritization for minimum cost and maximum improvement of asset reliability

1.2.2 Short-term Spatio-Temporal Wind Forecast

This segment of the project seeks a novel statistical wind power forecast framework, which leverages the spatio-temporal correlation in wind speed and direction data among geographically dispersed wind farms. Critical assessment of the performance of spatio-temporal wind power forecast was performed using realistic wind farm data from West Texas. It was used to justify whether spatio-temporal wind forecast models are numerically efficient approaches to improving forecast quality. The integrated forecast and economic dispatch framework was tested in a modified IEEE RTS 24-bus system. This research included the following detailed targets:

- Develop power system scenarios (day-ahead market operation scenario, intra-day operation scenario, and real-time market operation scenario) for the scheduling models. The advantages and disadvantages for various wind generation forecast models (PSS model, AR model, TDD model and TDDGW model) need to be evaluated.
- The wind data as well as the geographical and pattern information from West Texas needs to be processed to prepare for large scale simulation and model training.
- Based on the established benchmark system and power system scheduling, we have to conduct comprehensive numerical experiments for auto-correlation and cross-correlation analysis.
- We have to design a scheduling system to simulate the established benchmark system. With the scheduling system, we can compare the performance of different models in power system operation by incorporating spatio-temporal wind data. The benchmark system is a revised IEEE RTS system. We reconfigure the system parameters to mimic the electric power system in West Texas region. The system structure, resources mixture, load profiles as well as operating procedure of Texas are required to be well prepared for the scheduling system.

1.2.3 Distributed Database for Future Grid

There are a lot of potential applications for the data obtained from smart meters. Past experience particularly from blackouts has shown that there is a need for an enhanced decision making process, especially for the following:

- Tracking faults and defects in the system
- Utilizing intermittent power from renewable resources
- Outage management and restoration
- Monitoring voltage stability or oscillations

Data obtained from the various smart meter installations makes this a viable problem to solve. The first step involved however is to build a robust data management system that is able to process and store this data and make it available for analysis (either real-time or batch) in an easy to access format. We studied the feasibility of utilizing open source distributed database systems like HBase and Cassandra for storing the data.

1.3 Report Organization

The report is organized as follows: Section 2 provides background for data integration. In Section 3 the improved fault location using lightning data for transmission network is presented. In Section 4, a methodology for evaluating risk of insulation breakdown is presented. Section 5 describes big data application to wind power forecast and look-ahead power system dispatch. Section 6 describes implementation of distributed database for future grid. Contributions of the report are listed in the Section 7.

2. Technical Background

2.1 Data Sources

2.1.1 Utility Measurements

The SCADA system provides a set of measurements at the substation, including analog measurements such as bus voltages, flows (amps, MW, MVAR), frequency, transformer tap position, status (breaker switching state) signals, etc. The SCADA Data is sent to the energy management system (EMS) every two to ten seconds. New types of data coming from Phasor Measurement Unit (PMU) and IEDs, such as Digital protective relay (DPR), Digital fault recorder (DFR), Sequence of event recorder (SER), and other IEDs, have a much higher sampling rate and are able to record a larger amount of data. In addition, there are different configuration data that are stored together with these measurements.
As described in [9] the unified generalized representation of data and model may be performed in accordance with the standard formats for data exchange. The data exchange standard provides a description of the data syntax. Some examples are demonstrated below:

- SCL (IEC 61850-6) provides description for substation equipment and their configuration as well as data format for IEDs, [10],
- IEEE COMTRADE (IEEE C37.111) describes the data format for exchange of transient data captured by IEDs, [11],
- IEC CIM (IEC 61970) describes format for power system model and SCADA data, [12] and [13],
- IEEE COMFEDE (IEEE C37.239) describes format for event data exchange, [14],
- Measurement requirements and common data format for PMUs is described in IEEE PC37.118.1 [15], while format for communication of phasor measurements is described in IEEE PC37.118.2, [16].

2.1.2 GIS and GPS

Data involved in the power systems decision-making process have two main components, the temporal and spatial. GIS together with GPS enables spatial-temporal correlation as a base for all other data that may be associated or analyzed. As it is stated in [7] the spatial and temporal correlation of data plays an essential role in the process of integrating big data in the electric power industry applications. Spatial correlation of data is done by integrating different data sets as layers of GIS, while GPS is used for time synchronization between events.

Using GIS it is possible to systematically incorporate spatial data from different sources together [17]. Different data sources can be represented as individual layers, making data visualization and management easier. Several important aspects of GIS operation need to be taken into consideration [18]:

- Landbase Map: This is the data defining the physical features and spatial attributes of system components, and therefore it forms the framework of the overall GIS implementation.
- Data Conversion: For data to be useful it needs to be pre-processed since geographical attributes which are identified through imagery or other sources then need to go through the vectorization process.
- Layers: These are a GIS specific data structures allowing for heterogeneous data sources to be kept distinctly separate and stored in a single file.

Two distinct categories of GIS data, spatial and attribute data can be identified [19]. 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 example a table with the physical characteristics of a transmission tower can be described with the attribute 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. Some examples are:

- History of outages for specified components of transmission system,
- Lightning detection network data,
- Signals received from IEDs.

GPS is a system of 24 satellites installed by the US Department of Defense. It provides location and time information for GPS receivers located on the Earth. In order to use this service devices such as traveling wave recorders and lightning sensors are equipped with GPS receivers that supply information about longitude, latitude, and altitude, as well as a precise time tag. The latest equipment has a GPS time accuracy of 100 ns with a resolution of 10 ns [20].

2.1.3 Weather Data

With the developed modern science and technology, the measurements and the data collections infrastructure have been particularly designed and gradually improved for increasingly better operational weather forecasts. Not only localized observations (e.g. radar detection of tornados) but also large-scale weather patterns are necessary for predicting the weather over a specific small area.

The main weather factors affecting the power system are presented in Figure 1. Weather conditions have high impact on all parts of electric power system. Main causes of outages are:

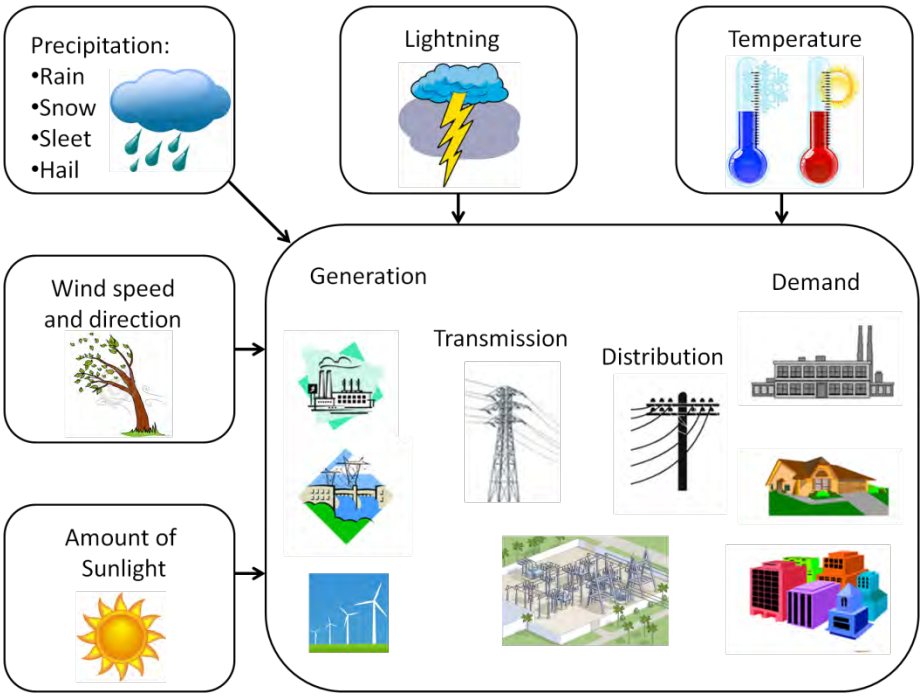

Figure 1: Weather factors affecting power system

- 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.
- Lightning activity: the faults are usually caused by cloud-to-ground lighting hitting the poles.
- 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 extremely low temperatures the demand increases which can lead to the network overload.

In the case of a large blackout, there is typically a combination of factors leading to the outage. For example, 2003 Northeast Blackout started with tree touching the line but then equipment failure led to the cascading distress on the electric grid. After the blackout, lack of good time synchronization system led to a lot of difficulties with investigation. Thus, in order to prevent large blackouts in the future, it is of great importance to connect all factors affecting the grid and correlate them in time and space. Weather impact together with equipment failure causes more than 50% of outages in electric network.

In order to efficiently exploit renewable energy resources such as wind and solar, weather information can be used as a direct indication of potentials for exploitation of these resources present in a targeted area. Energy suppliers are interested in coordinating the traditional power plants and weather-dependent energy sources. Thus, they need accurate predictions of weather conditions in locations where renewable energy plants are located.

For example, wind speed is crucial for energy production based on wind, while the sun's angle and cloud coverage are important information for solar based production.

Following factors are measured in a weather station:

- Air and soil temperature (min, max, average)
- Precipitation
- Wind speed and direction
- Lightning characteristics
- Air relative humidity (min, max, average)
- Solar radiation
- Pan evaporation
- Snow-water equivalent

Metrological stations (i.e. observation sites) worldwide routinely measure the basic meteorological parameters. The discussions below will demonstrate the data property primarily using ASOS, which the major source of automated surface observational data in the U.S. Most ASOS sites are located in the airports and currently there are more than 900 of them in the U.S [21]. The properties of data collection and interpolation are specifically discussed below:

- Data collection infrastructure:
    - Sensors: The group of sensors of ASOS is listed in [21] (e.g. Precipitation Identification Sensor). The specifications for sensor measuring range, accuracy, and resolution in ASOS are discussed in [21]. The requirements regarding sensor exposure and placement are listed in [22]. More information on station instrumentations may be found in [23].
    - Data exchange: The standard formats for weather data exchange among automated weather information systems are described in [24], and the configurations of ASOS network data flow are described in [21].
- Interpolation: Typically weather stations are sparsely located over the area. In order to generate maps with description of weather conditions for the whole area different interpolation techniques need to be applied [25].

Satellites detect different meteorological phenomena and the atmosphere in different temporal and spatial scales. The satellite meteorological detection is passive remote sensing in general, whereas the radar meteorological detection is active remote sensing. Satellites can measure shortwave (solar) radiation and longwave (terrestrial) radiation reflected and emitted by the earth-atmosphere system, respectively. Meteorological variables such as temperature and humidity profiles, cloud fraction, cloud top temperature can be derived from satellite observations. This process is called "retrieval" in remote sensing.

As a ground-based instrument, radar can detect the weather as far as hundreds of kilometers away. Radar detection has been widely applied for short-time weather forecasts. Radars

can emit radio or microwave radiation and receive the back-scattering signals from a convective system. The strength of radar echo is dependent on the type of hydrometeors and suggestive of the intensity of the convective system, within which precipitation and radial velocity can be retrieved from radar observations.

2.1.4 Lightning Data

The faults are usually caused by cloud-to-ground lighting hitting the poles. Depending on the area, the lightning may be very important in influencing electric power network faults. For instance, the research in UK [26] shows that lightning strikes are the second most common factor in weather-related distribution system faults. Due to predicted changes in operating conditions caused by weather and the change of power system infrastructure the percentage of faults induced by lightning is estimated to increase by 40% by 2080s [27]. In this case to minimize the effects of lightning proper protection of network structure (i.e. ground wires) and equipment (i.e. surge protectors) must be implemented by utilities [28].

The lightning detection network data can be used to correlate information about lightning characteristics with other event data gathered from the substation measurements. This provides better situational awareness during the critical events affecting the system and has the potential to improve automated fault location techniques.
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:

- Ground-based systems that use multiple antennas to determine distance to the lightning by performing triangulation.
- Mobile systems that use direction and a sensing antenna to calculate distance to the lightning by analyzing surge signal frequency and attenuation.
- 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 an accuracy of location within 0.7-1 km, while space-based systems have resolution of 5 to 10 km, [29].
For example, The National Lightning Detection Network (NLDN) [30] 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. [31].

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. No matter which source is used the 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).

2.2 The Big Data

The big data processing methodology consists of following steps [32]:

- Search
  - Implementing advance search process for fast access to data of interest.
  - Example: Gather all the data related to specific event.
- Machine Learning
  - Identifying important information using machine learning techniques such as classification and clustering.
  - Learning from the historical data.
  - Example: Learning from historical data about equipment and measurements at a given location.
- Knowledge Extraction
  - Extracting knowledge from information using different data mining techniques.
  - Extracting knowledge from the individual data sources without combining them. Each data set has its individual conclusions.
  - Example: Extracting knowledge separately from the lightning detection network, and separately from traveling wave recorders.
- Correlation
  - Combine individual knowledge gathered from different data sources to form a final conclusions and results based on reasoning.
  - Example: Correlating data from traveling wave recorders with data about lightning.
- Prediction
  - Identifying rules and trends in analyzed data that can be used to predict future behavior:
  - Linear Prediction
  - Neural Networks
  - Example: Predict what may be the response of equipment in the case of lightning strike in its vicinity.
- Visualization
  - Provide systematic way of presenting bots data and results.
  - Geographical and temporal representation.
  - Example: Showing all outage locations on the map.

3. Use of Big Data for Improved Fault Location

3.1 Introduction

Traveling wave fault location has been explored in literature [33-40] and claimed to be extremely accurate, which requires data sampling in the kilohertz and even megahertz range. Many utilities have either deployed such fault locators or are in the process of evaluating them. The GPS synchronization between two traveling wave recorders on two sides of the transmission line was discussed in [38-40]. In [41], the real time monitoring of transmission line transients under lightning strikes was presented. Real time electromagnetic transients were measured and correlated with lightning data recorded at the outage location to evaluate the impact on insulation coordination. Such measurements are very intensive exhibiting sampling rates of several megahertz.

Factors affecting accuracy of the traveling wave fault location methods are:

- Estimation of line length is a major cause of error. As it is presented in [33] not knowing exact line length and line topology can lead to the error close to 500 foot (150 m).
- The traveling waveform is assumed to travel at the speed of light, [42]. When it comes to the overhead transmission lines, velocity of the propagated wave is close to that of the light but not quite the same.
- Time stamping must be very precise to make the system work. As it is stated earlier the latest traveling wave fault locators have GPS time tag accuracy of 100ns, [42].
- Wave-detection error due to interpretation of the transient is a major source of error. This error results from misinterpretation of multiple transients and/or reflected transients. This is a significant concern in the case of lightning strikes. Lightning storms with multiple rapid strikes can cause confusion in terms of which transient was associated with which fault, [43]. In [39], the issue of multiple lightning strikes was investigated and it was reported that travelling wave recorders can produce incorrect results in such cases.
- Current transformers (CT) and capacitive voltage transformers (CVT) can affect the accuracy as well. In [44, 45] modeling techniques for transient response of CTs and CVTs are discussed. It has been pointed out that the differences in the length of the cabling from protection CT to the relay room at each end of the transmission line can affect accuracy [42]. Traveling wave fault location method used in this project extracts the traveling wave from the current signals collected on the secondary of CTs. The CTs have enough bandwidth to pass the transients, however they do affect accuracy of the method.
- Accuracy of the method is greatly affected in case of the faults with small inception angles (<5°). For the cases of fault inception at zero crossing, theoretically, no traveling wave from the fault location is generated [46].

This research demonstrates how the use of traveling wave and lightning surge measurements, correlated with data from Geographic Information System (GIS) and Global Positioning System (GPS) may bring major improvements in the outage

management. The utilities use such additional data types today, but data is processed manually and poorly correlated leading to delayed decisions and inaccuracies. The automated method has to address the big data problem due to heterogeneity of the data sets, as well as the high volume and velocity of data.

Because information coming from the lightning detection network is not a part of the conventional traveling wave fault location system it is not affected by all of the described errors. Thus, lightning detection network data may complement the fault location method and improves the accuracy of a complete system.

3.2 Data

Complete list of data is presented in Table 2, while Table 3 lists the big data properties of the presented application. The problem falls in the group of big data problems for the following reasons:

- Variety: The database includes sampled waveform data combined with reports from traveling wave fault locator units, lightning detection network, and geographical data. The data files come in different formats that are not compatible and information needs to be extracted so that they match the application. For example, lightning detection network provides the location of lightning strikes in terms of coordinates (longitude and latitude), while traveling wave recorder provides the information in terms of distance to the fault from line terminals.
- Volume: The implementation requires analysis of the extensive set of historical data in order to determine tradeoff between accuracy of traveling wave method and enhancement using the lightning data to determine the confidence of the data gathered during the event analyzed in real time. The lightning data is required for the period of time that covers all events from historical data and each lightning report will generate new map. This is just one level at which the volume of data can be overwhelming. In addition, during the fault, extensive set of data is received and

Table 2: List of data for improved fault location

<table>
<tr><th colspan="2">Dynamic</th><th colspan="2">Static</th></tr>
<tr><td>Lightning Detection Network</td><td>Traveling Wave Fault Locators</td><td>Geography</td><td>Simulation</td></tr>
<tr><td>Date and time of lightning strike, T_{light}</td><td rowspan="2">Date and time when event was recorded, T_A and T_B for two devices</td><td rowspan="3">Location of substations</td><td rowspan="2">Transmission line parameters</td></tr>
<tr><td>Location of a strike (latitude and longitude), L_{light}</td></tr>
<tr><td>Peak current and, I_{light}</td><td rowspan="2">Distance to the fault from the line terminal A, θ_A</td><td rowspan="2">Physical characteristic of a transmission line and towers</td></tr>
<tr><td>Lightning strike polarity, P_{light}</td><td rowspan="2">Geographical representation of the line</td></tr>
<tr><td>Type of lightning strike (cloud to cloud or cloud to ground), $Type_{light}$</td><td>Transient signals recorded at the line terminals</td><td>Line length, l</td></tr>
</table>

Table 3: The big data properties

		VOLUME	VELOCITY
V A R I E T Y	Traveling wave data	4 GB for storage of 2100 records from 8 line modules per substation device	Baud rate of 115200 bits per second
	Lightning data	40 MB of data per day	Sensor baud rate 4800 bits per second, event timing precision of 1μs
	GIS	Additional GIS layer for every type of data, each layer is few MB large	Up to 1000 maps per day can be generated for lightning data

not all of it is used for automated fault location. First, the important information needs to be extracted in an automated way. Typically, this process is done manually by utilities today. With methods used in big data analysis such as indexing for faster search and machine learning for extracting knowledge from data this process can be automated.

- Velocity: The velocity refers to the speed at which data is arriving to the central computing facility. During the fault, multiple sources will send a large amount of data that needs to be stored and ready for analysis. The examples are samples of traveling wave waveforms and coordinates of lightning strikes.

3.3 Methodology

3.3.1 Traveling Wave Fault Location

The GPS synchronized traveling wave method is used as one source of information for fault location. Traveling wave fault locator calculates distance to fault automatically based on recorded samples of traveling waves at one or both sides of the line. Mostly used method in modern devices is double ended Type D method with GPS synchronization. The locator calculates arrival time of the fault-induced waves using GPS as a reference. Then, these time tags are sent to the central station where fault location algorithm is used to determine distance to the fault from line terminals. In addition, samples of the recorded signal are transmitted. The accuracy of traveling wave method is highly dependent on the sampling rate. Modern devices use sampling frequency of 0.1 to 20 MHz. In case of Type D traveling wave method, GPS is primarily used for synchronization between signals received at two ends of the line. Conveniently, this information can be used for time correlation with lightning detection data that also uses GPS.

In order to implement traveling wave fault location the following steps are taken:

- Modeling of the power system: It is done according to the method given in ref, [47]. Transmission line modeling is done using J. Marti model, [48]. This is a frequency

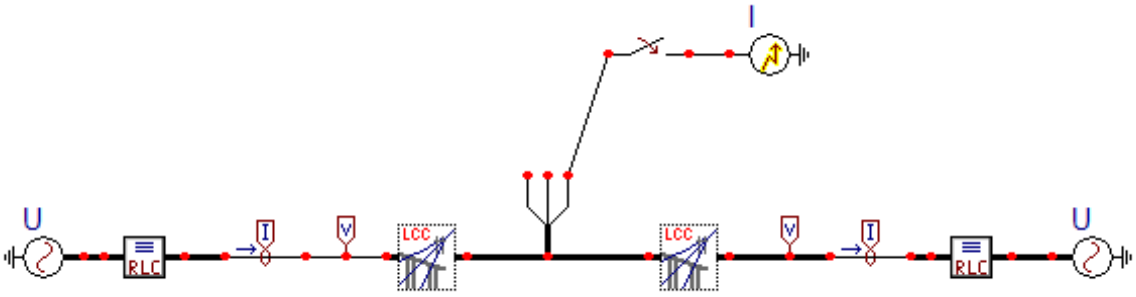

Figure 2: ATPDraw model of the tested line

dependent line model that uses analog filtering technique for identification of line parameters and can be simulated with ATP EMTP software, [49].

- Simulating the fault transients: Faults are introduced in various locations over the selected transmission line.
- Determining modal transformation for a three-phase system: Signals are transformed into modal components using Clark's transformation, [50]. After modal transformation a three-phase system is represented by an earth and two aerial modes. The aerial mode 1 is used for fault distance estimation.
- Computing the traveling wave velocity: Method that uses maximum of the first two consecutive peaks of the power delay profile (MPD method, [36]) is used.
- Calculating the arrival time: Wavelet transformation is used to determine the arrival time of the transient peak. The "mother" wavelet that is used is Daubechies wavelet, [37]. Wavelet Toolbox in MATLAB is used [51].
- Calculating fault location: The arrival times of the transient peaks at two TWRs that are located on two line terminals (T_A, T_B), line length between two TWRs (l) and calculated velocity of wave propagation (v) are used to calculate the distance θ to fault as

$$\theta = \frac{l + (T_A - T_B)v}{2} \tag{3.1}$$

- Performing time synchronization: Arrival times of two wave fronts are synchronized using GPS, [38-40].

The model of a 400 kV transmission line presented in Figure 2 is used for simulation in the experimental section. The sampling frequency was 1 MHz. The line length was 120 miles (~193 km). The faults were generated in the range from 10 to 110 miles from the terminal A.

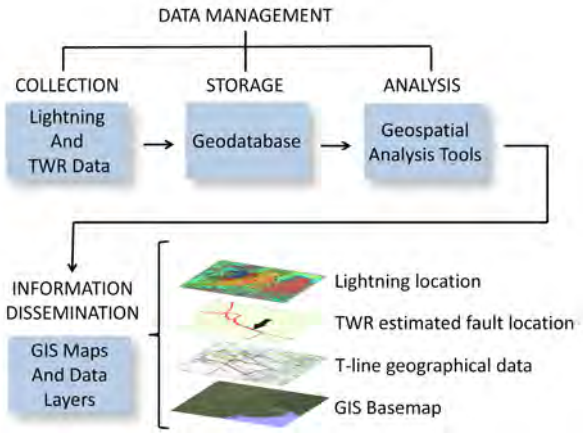

Figure 3: GIS data framework

Figure 4: Dataflow during the fault

3.3.2 Spatio-Temporal Correlation

Framework for GIS project is presented in Figure 3, [52]. Data collected by lightning detection network and traveling wave recorders is mapped and stored in Geodatabase. Geospatial Analysis Tools are used for manipulation of maps. Framework contains one layer for each type of data. Layers are classes or categories of data that can be organized in separate and distinct data structures, but integrated into a single file. These layers can be updated as the new information arrives to the system.

By comparing time stamps of events detected by traveling wave recorders and those obtained from querying the lightning detection system it may then be determined whether the disturbance is likely to be caused by lightning activity, as indicated by their closeness in time and space. The flow of information is illustrated in Figure 4. If it is determined that the disturbance is likely generated by lightning then the complete set of data about the event is gathered at the Central Station where correlation of data is leveraged together with analytics to improve fault location. In the Central Station the transient simulation of event is run and analysis of data is performed as described next.

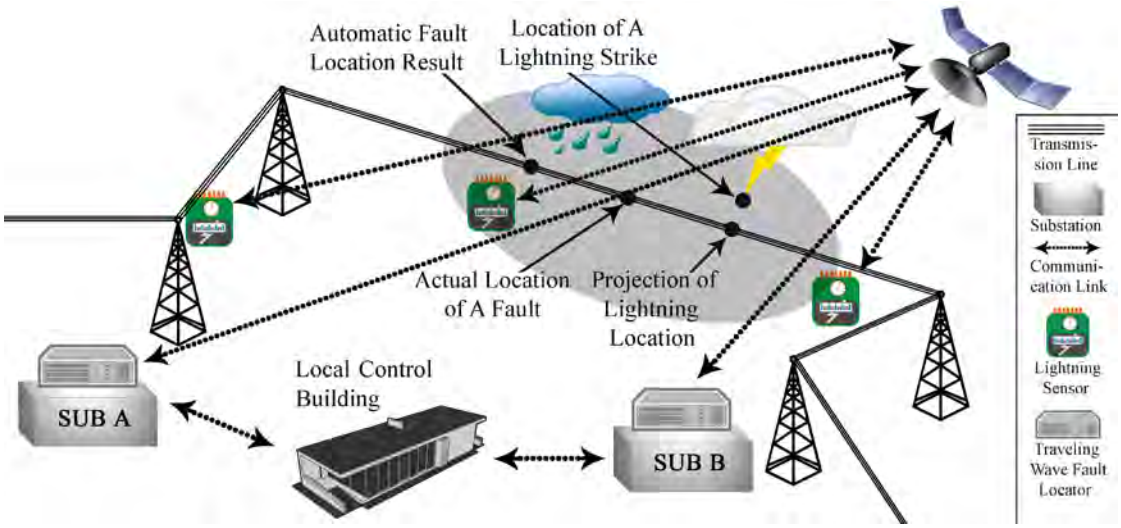

Figure 5: Spatio-temporal correlation of traveling wave fault locator data with lightning, GIS and GPS data

The data management for the correlation process is shown in Figure 5. Traveling wave fault recorders are located at both ends of the transmission line. On the other hand lightning sensors are typically not a part of the utility infrastructure and are located sparsely across a wide area. The traveling wave fault location system provides the estimate termed *Automatic Fault Location Result* in Figure 5. This result is implicitly allocated to the transmission line. Lightning sensors provide an estimated *Location of a Lightning Strike*. This result is presented in terms of longitude and latitude and it is not necessarily located on the transmission line, but rather somewhere in the vicinity of the line.

In Figure 6 processing steps for traveling wave fault location and lightning data integration are described. Before the beginning of spatio-temporal correlation process lightning data set is reduced to set containing only cloud-to-ground surges, where all instances of cloud-to-cloud surges are removed from data set. Then, the temporal correlation is done. After fault is detected, the time window that contains 2 seconds around the time stamp for fault beginning received from the traveling wave recorder (*FaultStart*) is created. The data received from lightning detection network is searched and only lightning strikes that satisfy following rule are collected inside the *Database A*:

$$|FaultStart - LightningTimeStamp| < 1s \qquad (3.2)$$

After that the spatial correlation is done. Based on location of line terminals and geographical representation of the line the buffer around the line is created that covers area going 300m on both sides of the line, Figure 7. This area has a shape of a polygon. The data in *Database A* is searched and only those lightning instances that are inside the buffer area are collected in *Database B*.

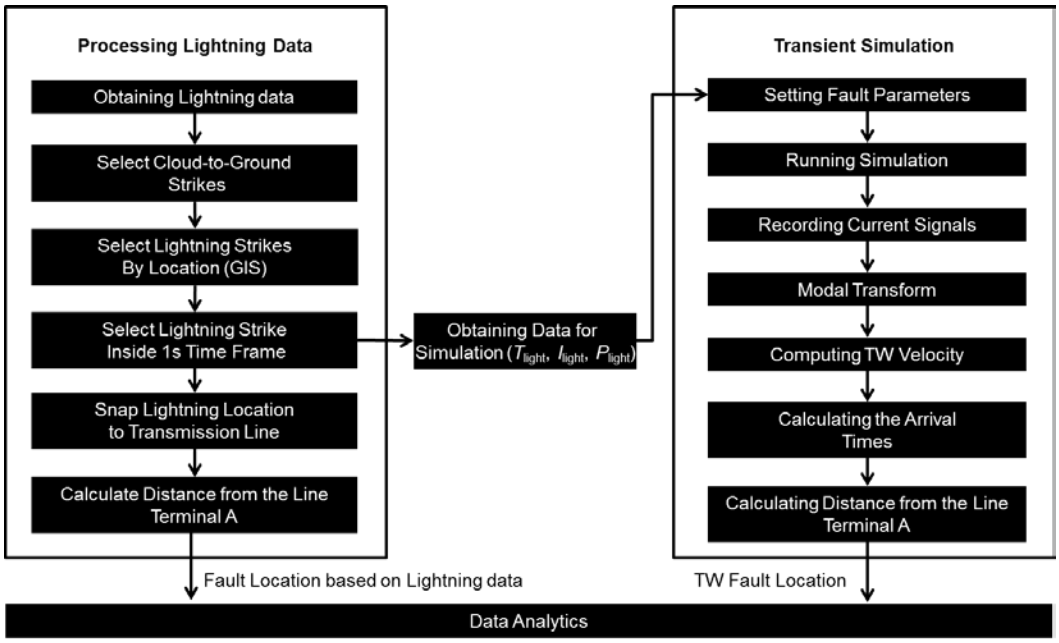

Figure 6: Processing steps

In the next step lightning instances from *Database B* are searched and the closest one to the traveling wave recorder result is chosen to be correlated as a *LightningDetectionResult,* Figure 8. The location of the lightning strike is projected to the closest point on the transmission line using a "snap" feature, Figure 9. The snap editing in GIS will move the point within the specific distance (tolerance) of the line to the closest point on the line. This snapped point is considered as the lightning detection network estimate of fault location so that the fault location can be described in terms of distance from the line terminals. For the tolerance 1 km distance from the line is chosen. Only lightning strikes that are within 1 km from the line will be used in the correlation process. Then, two fault location results one coming from the traveling wave fault locator and the other coming from the lightning

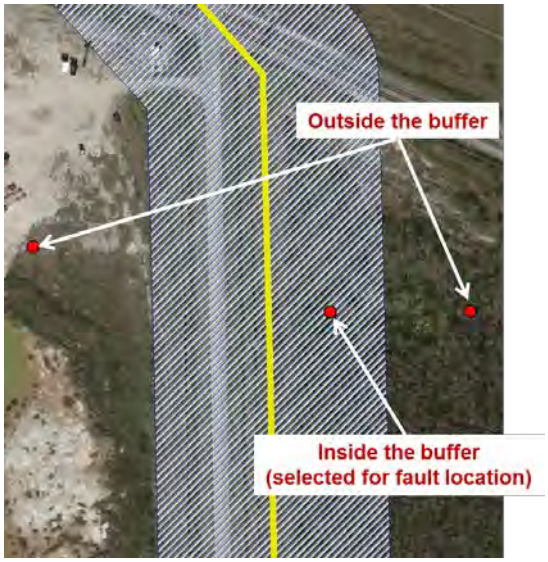

Figure 7: Buffer around the line

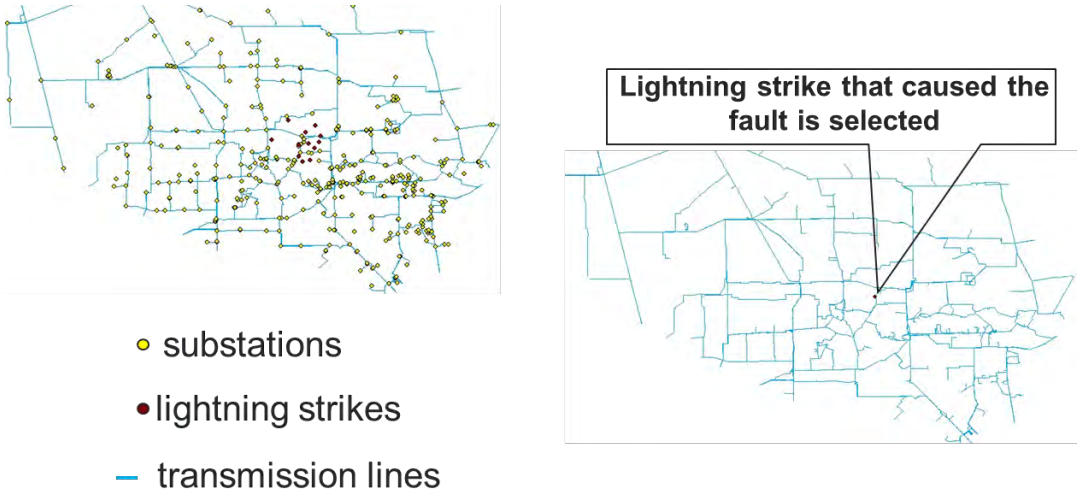

Figure 8: Selection of lightning strike

detection network are combined using a Bayesian framework in order to improve the accuracy of the prediction.

3.3.3 Data Analytics

We consider the traveling wave fault location to be the main source of information about the fault event. It processes the recorded data x, and makes the maximum likelihood estimate of the fault location based on this data. The precise value from (3.1) may be described as the following,

Figure 9: Projecting the lightning

$$FaultLocationResult \approx \arg\max_{\theta} p(x|\theta) \tag{3.3}$$

It is possible to discern the variance of θ either from historical records or through other means, but these methods may be unreliable and are beyond consideration in this study.

The lightning detection data is considered the prior probability, in this case coming from indirect, side-information and independent of the measurements x,

$$LightningDetectionResult \approx \arg\max_{\theta} p(\theta) \tag{3.4}$$

The posterior probability of the fault location can then be expressed using Bayes Theorem as,

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \tag{3.5}$$

In order to compute the necessary maximum a posteriori estimate of fault location,

$$ImprovedPrediction = \arg\max_{\theta} p(\theta|x) \tag{3.6}$$

it is not necessary to compute the normalization constant $p(x)$ because the same fault recorder data x is considered under all fault location positions θ.
By considering the posterior instead of only the likelihood better predictions are made because cross-domain data is integrated.

Taking the logarithm of the (3.5) and disregarding the normalization constant,

$$\log p(\theta|x) \sim \log p(x|\theta) + \log p(\theta) \tag{3.7}$$

Under the normal assumption for both distributions prior and likelihood in (3.6), the explicit computation of variance is not necessary. Instead it is computationally favorable to compute the optimal trade-off parameter *nu* from the interval [0, 1]. This parameter then controls the trade-off between a bigger or smaller variance in $p(x|\theta)$, and $p(\theta)$, but only in direct proportion to each other and irrespective of $p(x)$. At *nu* = 1 we completely trust the lightning detection network data, and then as *nu* is transitioned towards 0 more and more certainty is placed in the traveling wave fault location.

This approach is computationally favorable to fully Bayesian approaches such as Markov Chain Monte Carlo sampling, which would make the application infeasible for power systems.

Considering the monotonicity of the logarithmic function we may express the improved fault location as the linear combination of

$$\arg\max_{\theta} \log p(\theta|x) = \arg\max_{nu}\left[\arg\max_{\theta}\left[nu \cdot p(x|\theta)\right] + \arg\max_{\theta}\left[(1-nu)\cdot p(\theta)\right]\right] \tag{3.8}$$

The task becomes that of obtaining the precise *nu* to use. In order to compute *nu* a binary search along a line can be used to find optimal values since the problem is one dimensional. This process requires only $O(\log n)$ time to find the optimal nu among n given values. A simple linear combination like this has the advantage of high bias and low variance in machine learning terms, meaning that its predictions are not likely to be very imprecise in addition to having good generalization power across unseen examples. Because of the low computational complexity this kind of algorithm is directly applicable to big data scenarios.

3.4 Results

In order to assess the performance of the proposed fault location method it was necessary to evaluate its performance on a number of different fault scenarios. Using the model in Figure 1, 1000 fault scenarios were simulated. First, all fault scenarios were solved using only the traveling wave method for fault location. After simulation the error of this method was calculated as the relative mean absolute error,

$$e(\%) = \frac{|CalcDist - ActualDist|}{LineLength} \times 100 \tag{3.9}$$

Second, the results from the lightning detection network were calculated as explained in section 3.3 and (3.9) was used to quantify the error.

Algorithm flowchart for testing the proposed method is presented in Figure 10. First the dataset is split into training and testing set. Training data is used to compute the optimal tradeoff parameter nu. Error is then calculated using testing data.

After correlation of the two methods, as it was explained in Section 3.3.2, error of the improved result was calculated using (3.9). When dealing with a linear combination of predictors it is necessary to assess the generalization performance. Good generalization is indicated by the ability of a fault location method to locate faults accurately even for unforeseen fault locations. In order to quantify the generalization performance of the proposed fault location method it was necessary to compute the generalization error.

In order to estimate the generalization error of the improved fault location method it is necessary to split the data from many different scenarios into a training set and a testing set of data. Determining the optimal *nu* on the training set gives point estimates of the generalization error on the testing set when comparing the improved fault location to the true fault location, and therefore the procedure needs to be repeated for precise estimates, a process often calls for 2-fold cross-validation. The results in Figure 11 and 12 are average results, computed per scenario, from 100 replications of cross-validation.

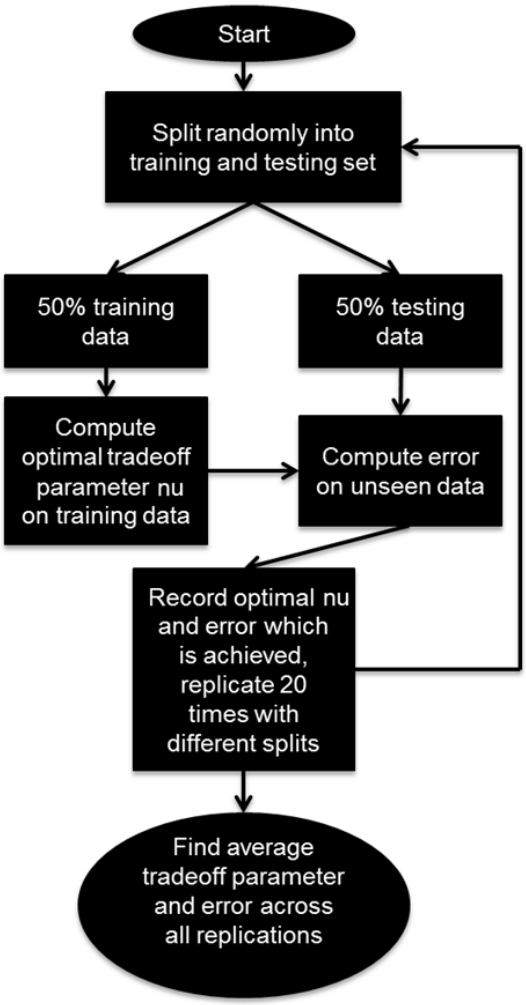

Figure 10: Algorithm flowchart for determining optimal trade-off parameter

A histogram of all three results is presented in Figure 11, where the *x*-axis represents the error and the *y*-axis represents the frequency of that error, where errors from different scenarios are binned according to a regular grid. From Figure 11, the proposed approach outperforms the traveling wave fault location, having the largest number of test cases with error that is closer to zero. For every test case our approach shoved better accuracy than the individual methods. Mean Square Error of distance to fault for the lightning data was 0.0076±3.1e-04 miles, for traveling wave it was 0.0012±4.3e-05 miles and the improved method showed 0.0011±4e-05 miles, both the variance and the mean of the error were smaller using the improved method on unseen fault scenarios, when compared to the traveling wave method.

3.5 Discussion

It is significant that the traveling wave method result has much higher accuracy than the one obtained from lightning data. The lightning detection data may only be useful in

enhancing the traveling wave fault detection method. Lightning data has very high variance as well compared to other two methods.

Additionally, the proposed method shows no bias in the predictions in Figure 10, indicating that the fault prediction location neither systematically over- or under-estimated. Because traveling waves are recorded on both sides of the transmission line, the error does not depend on the distance from the fault.

As it can be seen in Figure 12 the tradeoff parameter *nu* between accuracy of traveling wave method and lightning data is estimated to be optimal at 0.871±0.0133 on unseen examples. This can be interpreted as placing 87.1% trust in the result of the traveling wave method, 12.9% in the estimate from lightning. The low variance of *nu* is indicative of the low variance predictor used for improved fault location.

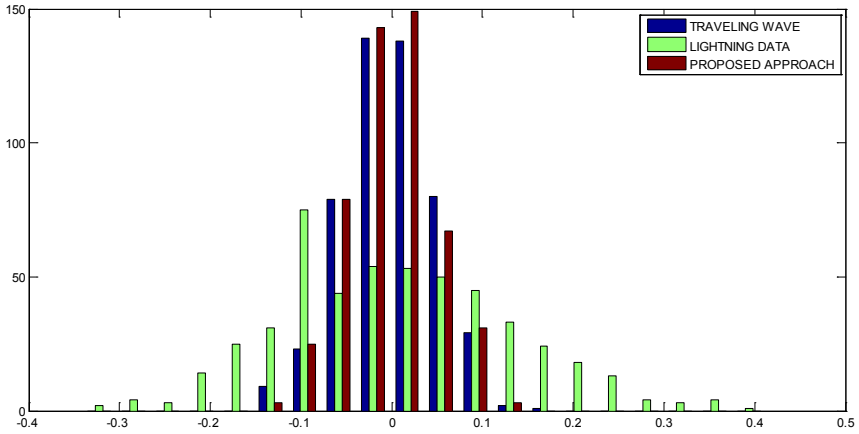

Figure 11: Histogram of an error distribution for individual traveling wave and lightning data; and our approach that combines two methods

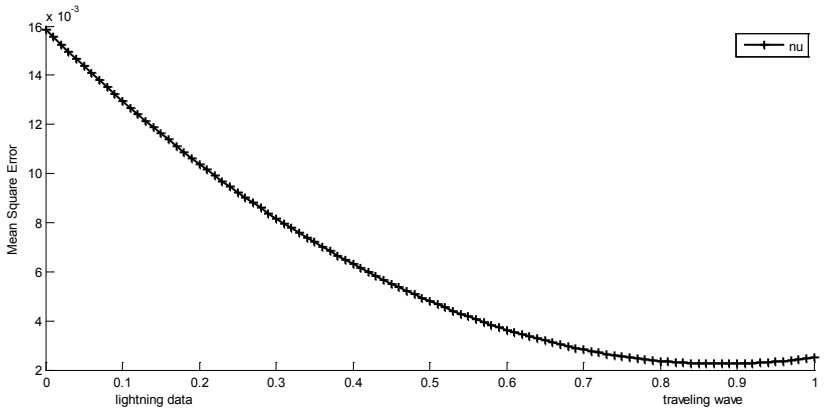

Figure 12: Comparison between traveling wave and lightning data using *nu*

3.6 Summary

The research demonstrates that correlating automatically multiple sources of data may help enhance fault location calculation. A method using correlation of cross-domain data for identifying which faults are likely to be caused by lightning strikes is presented. A method for locating faults using lightning detection data is presented and its precision quantified. A method of integrating lightning detection sensor data with traveling wave fault location measurements is presented and its precision quantified. The results indicate that integrating lightning detection sensor data with traveling wave fault detection data improves fault location accuracy. Proposed method that correlates traveling wave fault locator data and lightning data exhibits better performances than any of the methods alone.

4. Evaluating Impact of Weather Events to Insulation Coordination

4.1 Introduction

Lightning studies and experiences with insulation coordination have been reported in [53-57]. For the purpose of estimating probability of a lightning strike, historical lightning data has been used in [53, 54]. In [41], real time monitoring of transmission line transients under lightning strikes was presented, which allows for spatio-temporal correlation of lightning data and transient measurements to evaluate the impact on insulation coordination.

Correction factors for utilization of weather station data for insulation coordination have been described in [58]. In [59], the lightning-related risk analysis using artificial neural network in order to estimate insulator flashover rates has been performed. An optimization procedure to determine locations of line arresters that would minimize the risk has been implemented. In the mentioned work, the weather conditions have been taken into account; however, the statistical study has been performed based on a randomly generated data.

This paper builds on the study presented in previous chapter by adding the risk assessment to improve transmission system asset and outage management. The approach utilizes the weather data used for improved fault location to determine how the lightning protection assets deteriorate due to high frequency transients caused by lightning. By evaluating how past lightning strikes affect the condition of the equipment over time, quantitative means for prediction of potential insulator failure are derived allowing the measures that can be taken in order to improve lightning protection performances of the system to be specified.

The overarching goal of this research is to develop an early warning system (EWS) for prediction of insulation breakdown. The aim is to develop a risk-based EWS that can integrate real time data about lightning threats to the system, and determine system vulnerability of and economic impacts on a given area of interest. In today's practice the vulnerability of the system's components is determined in advance by the manufacturer after performing certain number of standardized tests under controlled conditions. The stress limit determined through the tests is assumed to be constant during the component's lifetime. In our research, accumulated impacts of past disturbances to the components are taken into account to assess unfolding component's vulnerability after exposure to continued weather threats such as lightning. In order to evaluate the current state of equipment as well as prediction of risk in case of future weather impact, a case with real data gathered from lightning detection network and weather stations is conducted.

4.2 Data

A complete list of data used in this study is presented in Table 4. All towers and substations were geographically referenced in order to spatially correlate their locations with those of lightning strikes and weather stations. The 100 m buffer is created around transmission lines. Only lightning strikes inside the buffer were selected and each of them was simulated as one fault scenario inside ATP. Location of lightning strike in the ATP model is determined based on lightning location in relation to the transmission network map.

Data obtained from three weather stations was used: Station NCHT2 - 8770777 - Manchester, TX, Station LYBT2 - 8770733 - Lynchburg Landing, TX, Station MGPT2 - 8770613 - Morgans Point, TX, [60]. The locations of stations are presented in Figure 13. Each weather station reports measured values at the location with certain time resolution.

4.3 Methodology

4.3.1 Integrating Weather Data

After values from multiple weather stations are collected, interpolation algorithms are used to estimate the values of parameters in an area of interest (i.e. transmission lines). The results are then presented as a vector or raster maps than can be overlaid with the network georeferenced map. Time instances of interest are times of lightning strikes obtained from lightning detection network. In order to temporally correlate lightning data with data from weather stations, linear interpolation is used. For each tower and substation, weather parameters at the locations were calculated based on distance to the weather stations as:

$$P = \frac{\frac{P_1}{d_1} + \frac{P_2}{d_2} + \frac{P_3}{d_3}}{\frac{1}{d_1} + \frac{1}{d_2} + \frac{1}{d_3}} \tag{4.1}$$

Table 4: List of data

Lightning Detection Network	Weather	Insulation Studies	Geography	Traveling Wave Fault Locators
Date and time of lightning strike	Temperature	Surge impedances of towers	Location of substations	Date and time when event was recorded
Location of a strike (latitude and longitude)	Atmospheric pressure	Surge impedances of ground wires	Geographical representation of the line	Distance to the fault from the line terminals
Peak current and lightning strike polarity	Relative humidity	Footing resistance	Location of towers	Transient signals recorded at the line terminals
Type of lightning strike (cloud to cloud or cloud to ground)	Precipitation	Components BIL (Basic Lightning Impulse Insulation Level)	Location of surge arresters	

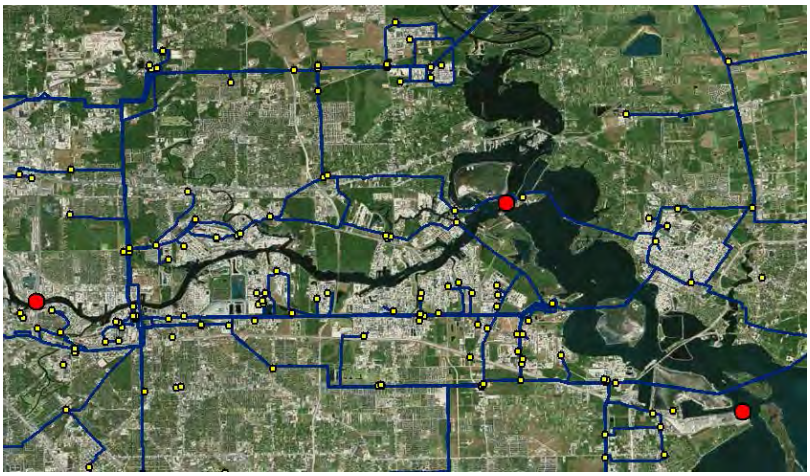

Figure 13: Location of three weather stations

where P is an estimated parameter value at the component location, P_i is a parameter value measured at weather station i; and d_i is a distance from the weather station i to the component. Weather data is used to calculate BIL under nonstandard atmospheric conditions [61], BIL_A as:

$$BIL_A = \delta \; H_C BIL_S \tag{4.2}$$

where BIL_A is the BIL under nonstandard conditions, BIL_S is the standard BIL, δ is the relative air density, and H_C is the humidity correction factor. Relative air density can be calculated using:

$$\delta = \frac{PT_S}{P_S T} \tag{4.3}$$

where T_S and P_S are standard temperature and pressure respectively; T and P are measured temperature and pressure respectively. Humidity correction factor is equal to 1 for rainy conditions and for dry conditions can be calculated using:

$$H_C = 1 + 0.0096 \cdot \left[\frac{H}{\delta} - 11 \right] \tag{4.4}$$

4.3.2 Network Modeling

The network is modeled using the ATP version of EMTP [49]. J. Marti's frequency dependent model [48] was used for modeling of transmission line segments between towers. For representing towers, multistory transmission tower model for lightning surge analysis proposed in [62] is used. Tower model parameters are calculated using (4.5.a-4.5.f):

$$Z_t = 60\left(\ln\frac{H}{r} - 1\right) \tag{4.5.a}$$

$$r = \left(\frac{r_1h_1 + r_2h_2 + r_3h_3}{2H}\right) \tag{4.5.b}$$

$$H = \sum_{i=1}^{3} h_i \tag{4.5.c}$$

$$R_i = \frac{-2Z_t\ln\sqrt{\gamma}}{h_1+h_2+h_3}h_i, \quad i = 1..3 \tag{4.5.d}$$

$$R_4 = -2Z_t\ln\sqrt{\gamma} \tag{4.5.e}$$

$$L_i = \frac{\alpha R_i 2H}{V_t}, \quad i = 1,4 \tag{4.5.f}$$

where:

Z_t – tower impedance,
H – tower height,
R_i – resistances of sections,
L_i – inductances of sections,
γ – attenuation coefficient,
α – damping coefficient,
V_t – propagation velocity,
h_1, h_2, h_3 – distances between stories,
h_4 – distance between lowest story and ground,
r_i – distances between tower center and tower edge at the level of a story i.
R_f – tower footing resistance.

Modeling of lightning impulse is done using current source based on CIGRE concave lightning model presented in Figure 14 [63]. Lightning peak current is obtained from the lightning detection network. A time characteristic is synchronized with the time of a lightning strike.

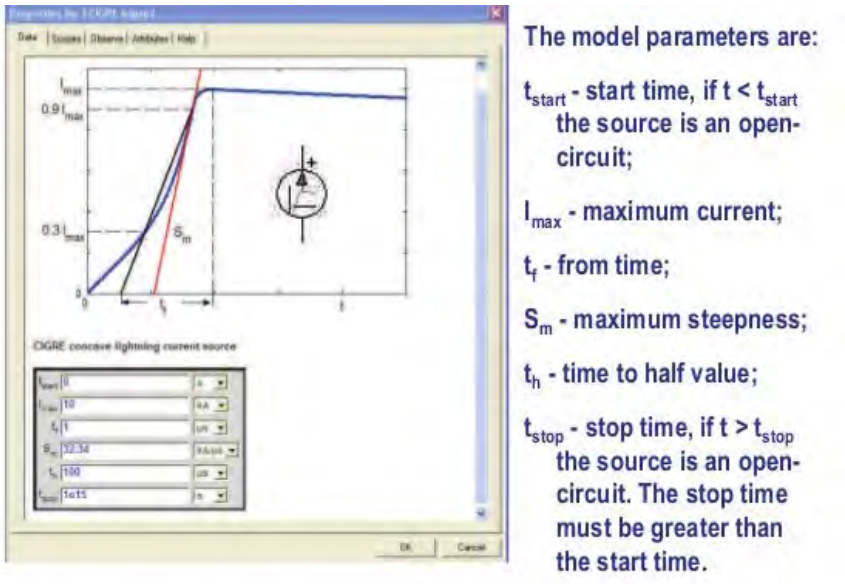

Figure 14: CIGRE concave lightning model, [63]

The simulation process is presented in Figure 15. First the lightning strike is selected and data from lightning detection network are sent to the ATP model in order to generate the fault. Then the simulation is run. After simulation for each node (tower) of interest, the maximum value of voltage is recorded. In parallel, the nonstandard *BIL* for the component is calculated using weather data. In the end, measured maximum voltage is compared to the component's nonstandard *BIL* and data is sent to the prediction model where it will be used as historical data for training.

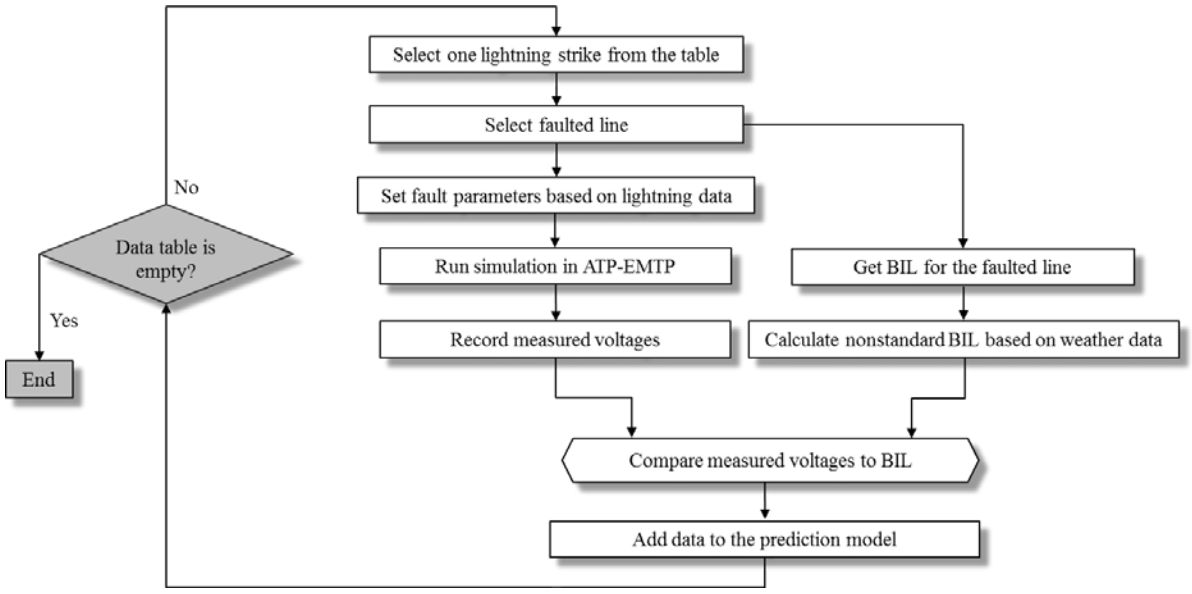

Figure 15: The simulation process

4.3.3 Risk Framework

Risk assessment framework is defined as follows:

$$R = P[T] \cdot P[C|T] \cdot u(C) \tag{4.6}$$

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 breakdown if lightning strike with intensity T occurred; and the Worth of Loss, $u(C)$, is an estimate of financial losses in case of insulation breakdown.

The proposed risk measure can be defined as a stochastic process and referenced in time and space as follows [64]:

$$\mathrm{R(X,t)} = \mathrm{P[T(X,t)]} \cdot \mathrm{P[C(X,t)|T(X,t)]} \cdot \mathrm{u(C(X,t))} \tag{4.7}$$

where X represents the spatial parameter (longitude and latitude) and t represents the time parameter obtained using GPS. As an example, the impact of lightning is associated with certain time and location. The impact that lightning will have on a component depends on the component's distance from the lightning strike. With the stochastic risk maps, the early warning system can be spatially and temporally mapped accordingly.
The spatio-temporal risk measures fit well with Bayesian paradigm [64]. The definition of Bayesian model is:

$$\pi(\mathrm{hypotesis|evidence}) = \\ = \frac{\mathrm{f(d_{obs}|\theta)\pi(\theta)}}{\int \mathrm{f(d_{obs}|\theta)\pi(\theta)d\theta}} \propto \beta \mathrm{f(d_{obs}|\theta)\pi(\theta)} \tag{4.8}$$

where $\pi(\theta)$ is the prior state of information or probability of the evidence associated with the set of parameters θ; $f(d_{obs}|\theta)$ is the likelihood of evidence or parameters θ will reproduce the observations d_{obs}; $\pi(\theta|d_{obs})$ is the posterior representing joint probability function between prior and likelihood.

Figure 16 shows the relationship model for risk assessment. Lightning data are indicating the probability of a lightning strike that is impacting probability of a backflashover. Probability of a backflashover is also under impact of weather conditions (temperature, pressure, humidity and precipitation). If there was a backflashover, the probability of a component failure is increased. Due to component failure, some losses are expected to be imposed. The chain of events will impact the final risk as discussed in [64] when using Bayesian Networks. The three components of risk analysis (hazard, vulnerability and worth of loss) can be identified in Figure 16 and are explained in the following sections.

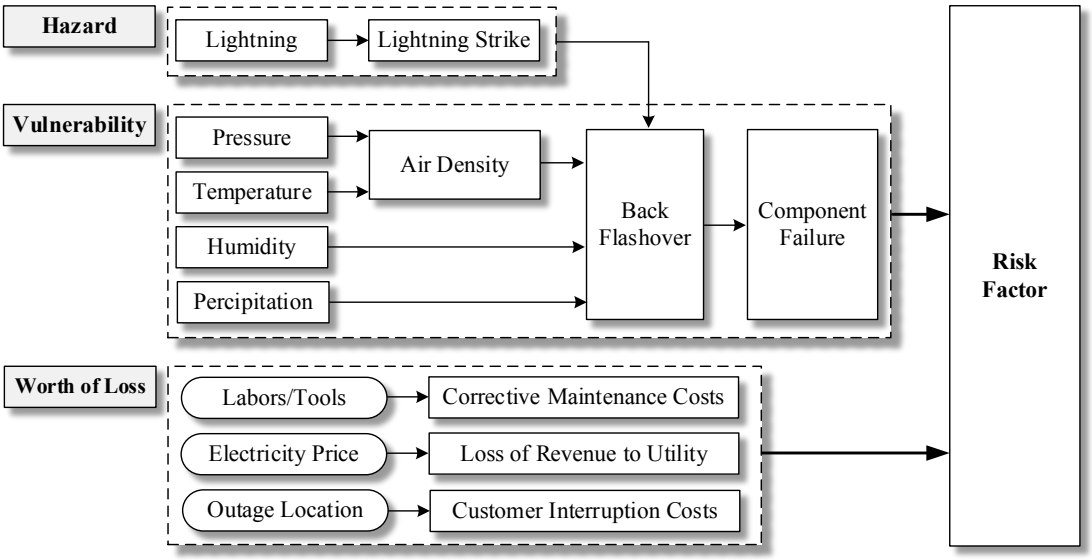

Figure 16: Risk analysis model

3.3.3.1. Hazard

Probability of a lightning strike is estimated based on historical lightning data in the radius around the affected components. Historical data for a period of 10 years were used. For each node, the lightning density is calculated as:

$$LD_i = \frac{L_A}{L_T} \tag{4.9}$$

where L_A is the number of lightning strikes in the area with radius of 100 m around the node and L_T is the number of lightning strikes in the total area of the network. Figure 17 shows the final lightning density map for the transmission network under study from which the value of Hazard is selected for each component of transmission network.

Figure 17: Lightning density map

3.3.3.2. Vulnerability

In order to estimate a new *BIL* as time progresses (*BIL_new*), data described above are represented here in form of a power system network where each node represents a substation or a tower and links between nodes are calculated using impedance matrix as illustrated in Figure 18. For each node in the graph, there are several input attribute values (x): temperature, atmospheric pressure, relative humidity, precipitation from the weather stations; Peak current and lightning strike polarity and the values of components BIL (Basic Lightning Impulse Insulation Level) prior to lightning strike. The output of interest (y) is the BIL after occurrence of lightning strike (*BIL_new*).

BIL_new in our experiments is predicted using Gaussian Conditional Random Fields (GCRF) based on structured regression [65, 66]. The model captures both the network structure of variables of interest (y) and attribute values of the nodes (x). It is a model based on a general graph structure and can represent the structure as a function of time, space, or any other user-defined structure. It models the structured regression problem as estimation of a joint continuous distribution over all nodes:

$$P(y|x) = \frac{1}{Z}\exp\left(-\sum_{i=1}^{N}\sum_{k=1}^{K}\alpha_k\left(y_i - R_k(x)\right)^2 - \right.$$
$$\left. -\sum_{i,j}\sum_{l=1}^{L}\beta_l e_{ij}^{(l)} S_{ij}^{(l)}(x)\left(y_i - y_j\right)^2\right) \tag{4.10}$$

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). The structure between outputs based on multiple layers of node inter-dependence is modeled by the "interaction potential" (the second double sum of the exponent in the equation). With such feature functions, the distribution can be expressed in Gaussian form that makes inference and learning of the

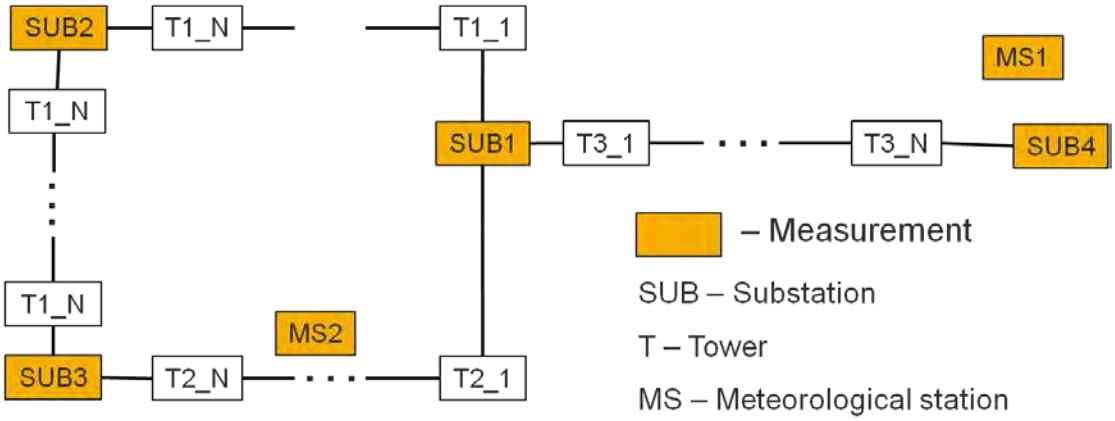

Figure 18: Illustration of a network data X = (lightning current, temperature, pressure, humidity, precipitation, BIL_old); Y = (BIL_new); links: impedance matrix.

model more feasible. The inference problem is then formulated as the mean of the Gaussian distribution that maximizes $P(y|x)$. Learning parameters ($\alpha_1, \ldots, \alpha_k; \beta_1, \ldots, \beta_l$) is done by convex optimization of the log likelihood. Outputs in terms of predictions of *BIL_new* variable are then used to calculate the probability of a flashover in case of a lightning strike $P[F|T]$.

The next step is calculating the probability of an insulation breakdown in case there was a flashover with a cumulative function:

$$\mathrm{P[C|F]} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\mathrm{V}} \mathrm{e}^{-\frac{(\mathrm{V}-\mathrm{V}_{50\%})^2}{2\sigma^2}} \mathrm{dV} \tag{4.11}$$

Vulnerability is calculated as a combination of probability of a flashover in case of a lightning strike $P[F|T]$ and a probability of an insulation breakdown after a flashover has occurred, $P[C|F]$ as:

$$\textit{Vulnerability} = P[C/T] = P[C/F] \cdot P[F|T] \tag{4.12}$$

As a result of Vulnerability analysis, the probability of insulation breakdown is expressed in terms of lightning peak current.

3.3.3.3. Worth of Loss Assessment

In case where the failure of an insulator ends up in a transmission line outage, the imposed outage cost can be quantified. The total imposed costs corresponding to the failure of insulator k and accordingly outage of transmission line i at time t, $\Phi^{\mathrm{t}}_{\mathrm{k,i}}$, are quantified in (4.13) comprising of three monetary indices.

$$\Phi^{t}_{k,i} = C^{t}_{\mathrm{CM},k,i} + \sum_{\substack{d=1 \\ d \in LP}}^{D} \left(C^{t}_{\mathrm{LR},k,i} + C^{t}_{\mathrm{CIC},k,i} \right) \tag{4.13}$$

The first monetary term in (4.13) is fixed and highlights the corrective maintenance activities to fix the damaged insulator. This cost index, which in some cases can be regarded as the replacement cost of the insulator, also includes the cost of required labor, regular tools, and maintenance materials. The variable costs (second term) include the lost revenue cost imposed to the utility ($\mathrm{C}^{\mathrm{t}}_{\mathrm{LR,k,i}}$) as well as the interruption costs imposed to the affected customers ($\mathrm{C}^{\mathrm{t}}_{\mathrm{CIC,k,i}}$). The cost function $\mathrm{C}^{\mathrm{t}}_{\mathrm{LR,k,i}}$ is associated with the cost imposed due to the utility's inability to sell power 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 (4.14) [67].

$$C^{t}_{\mathrm{LR},k,i} = \sum_{\substack{d=1 \\ d \in \mathrm{LP}}}^{D} \left(\lambda^{t}_{d} . \mathrm{EENS}^{t}_{d,k,i} \right) \tag{4.14}$$

where, λ_d^t is the electricity price ($/MWh.) at load point d and $EENS_{d,k,i}^t$ 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. Here, the EENS index of reliability is calculated by solving the following optimization problem [68]:

$$\min_{\theta,V,P,Q} \quad \sum_{g\in \mathrm{G}} C_g(P_g{}^t) + \sum_{g\in \mathrm{G_R}} C_g^R(r_g{}^t) \tag{4.15}$$

s.t.

$$\mathbf{g_P}(\boldsymbol{\theta},\mathbf{V},\mathbf{P}) = 0 \tag{4.16.a}$$

$$\mathbf{g_Q}(\boldsymbol{\theta},\mathbf{V},\mathbf{Q}) = 0 \tag{4.16.b}$$

$$\mathbf{h_F}(\boldsymbol{\theta},\mathbf{V}) \le 0 \tag{4.16.c}$$

$$\mathbf{h_T}(\boldsymbol{\theta},\mathbf{V}) \le 0 \tag{4.16.d}$$

$$\delta_n^{\min} \le \delta_n \le \delta_n^{\max}, \quad \forall n \in \mathrm{N} \tag{4.16.e}$$

$$V_n^{\min} \le V_n \le V_n^{\max}, \quad \forall n \in \mathrm{N} \tag{4.16.f}$$

$$P_g^{\min} \le P_g^t \le P_g^{\max}, \quad \forall g \in \mathrm{G} \tag{4.16.g}$$

$$Q_g^{\min} \le Q_g^t \le Q_g^{\max}, \quad \forall g \in \mathrm{G} \tag{4.16.h}$$

$$0 \le r_g^t \le \min(r_g^{\max}, \Delta_g), \quad \forall g \in \mathrm{G_R} \tag{4.16.i}$$

$$P_g^t + r_g^t \le P_g^{\max}, \quad \forall g \in \mathrm{G_R} \tag{4.16.j}$$

$$\sum_{g\in G, Z_m} r_g^t \ge R_{Z_m}^t, \quad \forall m \tag{4.16.k}$$

$$P_d^{\min} \le P_d \le P_d^{\max} \qquad \forall d \in \mathrm{LP} \tag{4.16.l}$$

$$\mathrm{IL}_d^t = P_{d_j} - P_{d_j}^{\text{supplied}} \quad \forall d \in \mathrm{LP}, \forall j \in \mathrm{N} \tag{4.16.m}$$

$$\mathrm{EENS}_{d,k,i}^t = \sum_{i\in \mathrm{N}} \sum_{h\in \mathrm{H}} P_h^t . \mathrm{IL}_{d,i,k}^t . RT_{d,i,k}^t \tag{4.16.n}$$

The optimization problem in (4.15) and (4.16) dispatches the energy and reserve to optimize the social welfare by minimizing the total cost of energy and reserves while satisfying AC power flow equations, ancillary service requirements, transmission and operating constraints. Constraints (4.16.a)-(4.16.b) represent the non-linear nodal active and reactive power balance equations. Network constraints (4.16.c)-(4.16.d) represent the branch flow limits for the "to" and "from" ends of each branch, respectively. Constraints (4.16.e)-(4.16.f) present the equality upper and lower limits on all bus voltage phase angles and magnitudes. Supply constraints are presented in (4.16.g)-(4.16.h) and (4.16.i)-(4.16.k) are capacity reserve constraints. Constraint (4.16.i) reflects the reserve for each generating unit that must be positive and limited above by a reserve offer quantity as well as the physical ramp rate (Δg) of the unit. Constraint (4.16.j) enforces that the total amount of

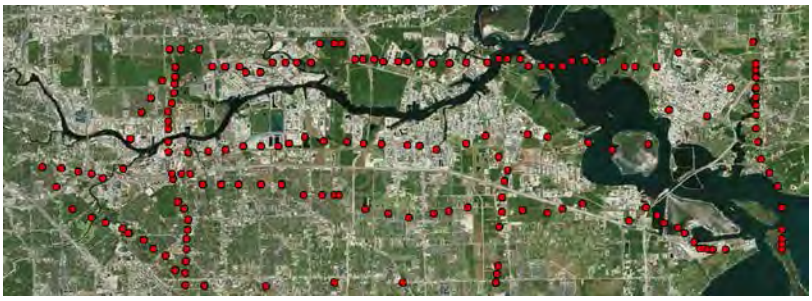

Figure 19: Location of transmission network components

energy plus reserve of the generating unit does not exceed its capacity. Constraint (4.16.k) is enforced to ensure that the right amount of capacity is procured according to the reserve requirements in each region. Constraint (4.16.l) restricts the demand at each load point between the lower and upper bounds. Constraint (4.16.m) calculates the interrupted load at each load point and constraint (4.16.n) evaluates the EENS corresponding to the outage condition.

The last variable term of the cost function in (4.13) 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 through (4.17). As it can be seen, $C^{t}_{CIC,k,i}$ is a function of the EENS index and the value of lost load ($VOLL_d$) which is governed by various load types being affected at each load point. The value of lost load ($/MWh.) is commonly far higher than the electricity price as obtained through customer surveys [69].

$$C^{t}_{\mathrm{CIC},k,i} = \sum_{\substack{d=1 \\ d \in LP}}^{D} \left(\mathrm{VOLL}_d . EENS^{t}_{d,k,i} \right) \tag{4.17}$$

The cost function in (4.13), which is actually the failure consequence of an insulator, can be calculated for each insulator failure in the network making it possible to differentiate the impact of different outages on the system overall economic performance.

4.4 Results

Part of the transmission network was geographically referenced as presented in Figure 19. The network segment contains 170 locations of interest (10 substations and 160 towers). Based on the geographical representation and network connectivity, the prediction graph is constructed as described in Figure 18. All towers were modeled using model described in section 4.3.2.

Historical data is prepared for the period of 10 years, starting from January 1st 2005, and ending with December 31st 2014. 1000 lightning strikes were assumed in the area of interest for the period of 10 years and used for Hazard calculation in the proposed risk framework. Out of 1000 strikes 100 strikes caused a flashover and were considered as an input data for prediction of vulnerability. For each instance of lightning strike, the weather parameters were obtained as described in section 4.3.1. The separate weather parameters

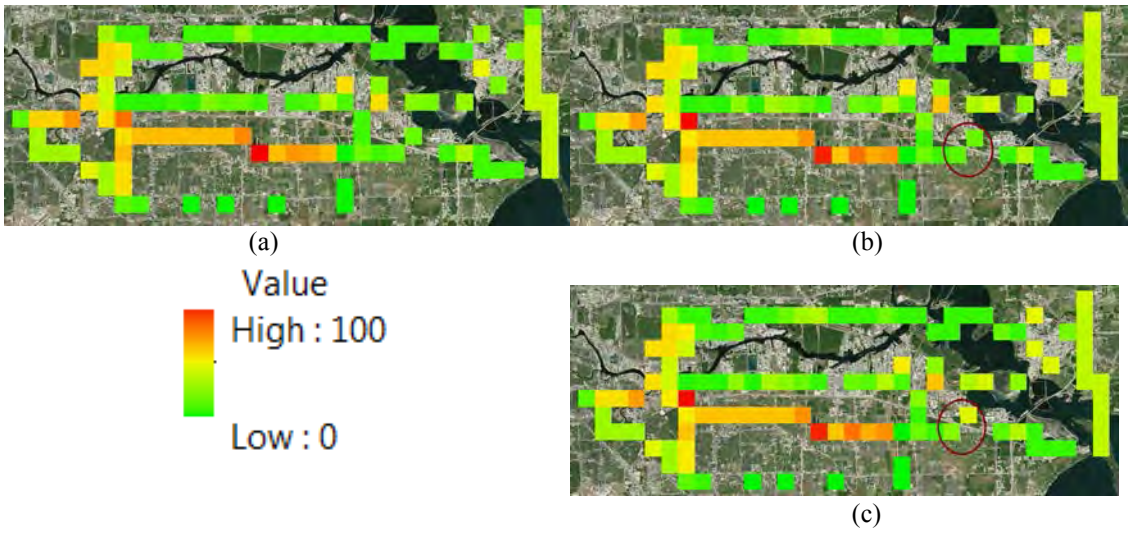

Figure 20: Total risk calculated on (a) January 1st 2009; (b) December 31st 2014; (c) January 5th 2015 (prediction after the next lightning strike)

were calculated for each component location. An example of weather data table is presented in Table 5.

Before the first lightning strike, all components are assumed to have a BIL provided from the manufacturer. In each time step, new BIL is calculated based on the old BIL and collected weather data for each lightning strike.

For each network component, risk value was calculated, assigned, and presented on a map as shown in Figure 20. The value of risk is presented as a percentage, where 100% is assigned to the component with highest risk for the future lightning-caused failures. In part (a) of Figure 20, the risk map at the January 1st 2009 is presented, while in part (b), the risk map after the last recorded event is presented. After several years of lightning impact, some of the zones that have experienced high rates of lightning activities have an increased value of risk. It is of utmost importance to observe the probability of future lightning strikes for such vulnerable zones.

With the use of weather forecast, the prediction of future Risk values can be accomplished. In Figure 20 (c), 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 node 141 (marked with red box in Figure 20 (c). Thus, risk value assigned to the node 141 will have the highest change compared to that of the previous step. The risk for node 141 changed from 22.8% to 35.2%. For the nearest node 140, the risk has changed from 21.6% to 25.6%. The Mean Squared Error (MSE) of prediction of GCRF algorithm on all 170 test nodes is 0.0637+0.0301.

4.5 Summary

The research introduces a new framework for predictive insulation breakdown risk assessment and maintenance plan. Methodology for calculation of equipment vulnerability to wide range of weather conditions surrounding lightning strikes is proposed. Insulation breakdown risk is assessed by analyzing time/space correlation between historical lightning and weather data, and network measurements. The algorithm is capable of predicting risk in case of future lightning strikes using Gaussian Conditional Random Fields (GCRF) structured regression model. With this model components geographical configuration is taken into account for a prediction. An automated early warning system (EWS) for prediction of insulation breakdown is developed and integrated with Geographical Information System (GIS).

5. Short-term Spatio-temporal Wind Power Forecast in Look-ahead Power System Dispatch

5.1 Introduction

Uncertainties and variabilities in renewable generation, such as wind energy, pose significant operational challenges to power system operators [70-74]. While conventional wisdom suggests that more spatially dispersed wind farms could be aggregated and "smooth out" total wind generation at any given time, the reality is that wind generation tends to be strongly correlated in many geographical regions [75,76]. As many regions/states are moving toward renewable portfolio standards (RPS) in the coming decade, the role of *accurate wind prediction* is becoming increasingly important for many regional transmission organizations (RTOs) [77].

The major uncertainty in conventional power grid operation comes from the demand side [78-80]. Nowadays, in power systems with high presence of intermittent generation, the main source of uncertainty comes from both demand and supply sides [70]. State-of-the-art load forecasts could achieve high accuracy in the day-ahead stage [81]. Compared with load forecasting, accurate forecast of wind generation still remains an open challenge. There exists a large body of literature on wind power forecasting, and state-of-the-art day-ahead wind forecast based on numerical weather prediction (NWP) models has enabled relatively accurate wind forecast with approximately 15\%-20\% of wind speed forecast mean absolute error (MAE) [82-85]. As the operating time moves closer to the near term (e.g., hour-ahead or 15 minute-ahead), at a high spatial resolution, the computation complexity (in terms of simulation time and memory requirements) often renders NWP models intractable [85].

In sharp contrast, data-driven statistical model is thought to be the most competitive method for near-term wind forecasting problems being able to capture the rapidly changing dynamics of the atmosphere and with nice model interpretation [86]. Statistical forecasting models could potentially provide *accurate and efficient* wind forecasts with MAE reduced to the range of around 5% or less [82]. A good set of references can be found in [87]. Our proposed spatio-temporal wind forecast model is directly targeted at computationally efficient near-term wind forecasts.

Starting from our preliminary work [88,89], the main objective of this project is to exploit a novel short-term spatio-temporal wind forecast model and quantify the dispatch benefits from improved short-term wind forecast. Wind generation is driven by wind patterns, which tend to follow certain geographical spatial correlations. For large-region wind farms, the wind generation forecast of the wind could significantly benefit from upstream wind power generation. Enabled by technological advances in sensing, communication, and computation, spatially correlated wind data could be leveraged for accurate system-wide short-term wind forecasts. This is potentially applicable to large-scale wind farms. The performance of such wind forecast model is critically assessed.

In order to fully exploit the advantage of spatio-temporal wind forecast, advanced power system scheduling is needed. In recent years, there are many valuable pieces of work along this direction. Currently, two major schools of methodologies exist: 1) based on stochastic optimization and 2) robust optimization. A security-constrained unit commitment algorithm is formulated by J. Wang et al., which considers the intermittency and volatility of wind power generation [90]. A two-stage stochastic programming model for reserves commitment in power systems with high penetration of wind generation is proposed by A. Papavasiliou et. al [91]. A stochastic optimization model is developed by P. Meibom et al. to study the operational impacts of high wind generation in Europe [92,93]. An adaptive robust optimization is proposed by D. Bertsimas et al. to solve security constrained unit commitment problems [94]. A robust unit commitment model is presented by Y. Guan et al. to schedule wind power and pumped hydro storage [95].

The advantage of the stochastic programming approach is to fully utilize the distribution of the uncertainty set to achieve optimal expected benefits. Compared with the stochastic approach, a robust optimization, focusing on optimal benefits under worst scenarios, has advantages in computation efficiency and low requirement for knowledge of full distribution [96,97]. The spatio-temporal forecast presented in this project is aiming at short-term power system application such as near-term (hour-ahead) or real-time economic dispatch which have high requirement of computation efficiency. Therefore, we propose and formulate a robust optimization based look-ahead economic dispatch model to quantify the economic benefits of improved forecast under uncertainties.

In this project, we propose a novel statistical wind power forecast framework, which leverages the spatio-temporal correlation in wind speed and direction data among geographically dispersed wind farms. Critical assessment of the performance of spatio-temporal wind power forecast is performed using realistic wind farm data from West Texas. It is shown that spatio-temporal wind forecast models are numerically efficient approaches to improving forecast quality. By reducing uncertainties in near-term wind power forecasts, the overall cost benefits on system dispatch can be quantified. We integrate the improved forecast with an advanced robust look-ahead dispatch framework. This integrated forecast and economic dispatch framework is tested in a modified IEEE RTS 24-bus system. Numerical simulation suggests that the overall generation cost can be reduced by up to 6% using a robust look-ahead dispatch coupled with spatio-temporal wind forecast as compared with persistent wind forecast models.

The suggested contributions of this project are:

- We propose to use two spatio-temporal correlated forecast models for short-term wind generation in power system operations, the TDD (trigonometric direction diurnal) and the TDDGW (TDD with geostrophic wind information) models. Both forecasting models take into account local and nearby wind farms' historical wind information. Additionally, based on atmospheric dynamic principles, the latter incorporates geostrophic wind information and has better forecasts than the former one. Both methods are tested with realistic wind data obtained in Texas, and they demonstrate improved forecast accuracy.

- We incorporate our spatio-temporal wind forecast into a robust look-ahead economic dispatch framework. Numerical study in a revised IEEE RTS 24-bus test system shows improved benefits compared with conventional static dispatch with time-persistent wind forecast models.

The rest of this chapter is organized as follows. In Section 5.2 we provide an overview of statistical wind forecast models, which is followed by the introduction of the proposed spatio-temporal wind forecast models. In Section 5.3 we compare the performance of spatio-temporal wind forecasts using realistic wind farm data obtained from West Texas. Section 5.4 presents the day-ahead reliability unit commitment model as well as a robust look-ahead economic dispatch formulation by incorporating available wind forecast. Numerical illustrations of the economic benefits of incorporating spatio-temporal wind forecast with robust look-ahead dispatch are presented in Section 5.5. Conclusions and future work are presented in Section 5.6.

5.2 Statistical Wind Forecasting

In this section, we provide an overview and critical assessment of several statistical approaches to short-term wind forecasting. Whereas NWP models play the key role in day-ahead to several hour-ahead wind forecasting, the computational burden and forecasting accuracy of NWP are still challenging in near-term forecasts (minutes-ahead to hour-ahead). As an alternative, data-driven statistical wind forecasting has gained increasing attention for near-term forecasts. Extensive research has been devoted to wind power forecasting problems [87, 98-100]. In short-term wind speed forecasting, statistical models that incorporate spatial information are the most competitive methods [86,87]. A regime-switching space-time model [101] improved forecasts by 29\% and 13\% compared with persistence forecasts and autoregressive in terms of root mean squared error (RMSE). It was generalized by the TDD model [102] by treating wind direction as a circular variable and including it in the model. Regime-switching models based on wind direction and conditional parametric models with regime-switching substantially reduced variance in the forecast errors [103]. Adaptive Markov-switching autoregressive models [104] were developed for offshore wind power forecasting problems in which the regime sequence is not directly observable but follows a first-order Markov chain.

Table 6: Site information

ID	Location	Area	Latitude	Longitude	Elevation
ROAR	3N Roaring Springs	Roaring S./Motley County		*W*100°50'43.38"	2615 ft.
SPUR	1W Spur	Spur/Dickens County	*N*33°28'51.05"		2287 ft.
JAYT	1SSE Jayton	Jayton (Kent Co. Airport)		*W*100°34'03.99"	2010 ft.
PITC	10WSW Guthrie	Guthrie/King County		*W*100°28'50.20"	1998 ft.

For wind speed forecasting problems, more realistic metrics that have penalization on underestimates and forecasts for small true values are desired for model evaluation [87] instead of RMSE and mean absolute errors (MAE). Power curve error [102] was proposed as a loss function, which links prediction of wind speed to wind power by a power curve and evaluates the loss based on the wind power with penalty on underestimates. The pros

and cons of the mean absolute percentage error and the mean symmetric absolute percentage error as loss functions to penalize both underestimates and forecasts for small true values were also discussed [87].

5.2.1 Wind Data Source in West Texas

The wind data we use here are the 5-minute averages of 3-second measurements of wind speed and direction collected by monitors placed at 10 meters above the ground from four sites in West Texas labeled ROAR, SPUR, PICT, and JAYT. Their locations are indicated by the red crosses in Figure 21, and more specific geographic information is listed in Table 6. The period of the wind data covers three years from January 1, 2008 to December 31, 2010. (The data sets are available at http://www.mesonet.ttu.edu/wind.html).

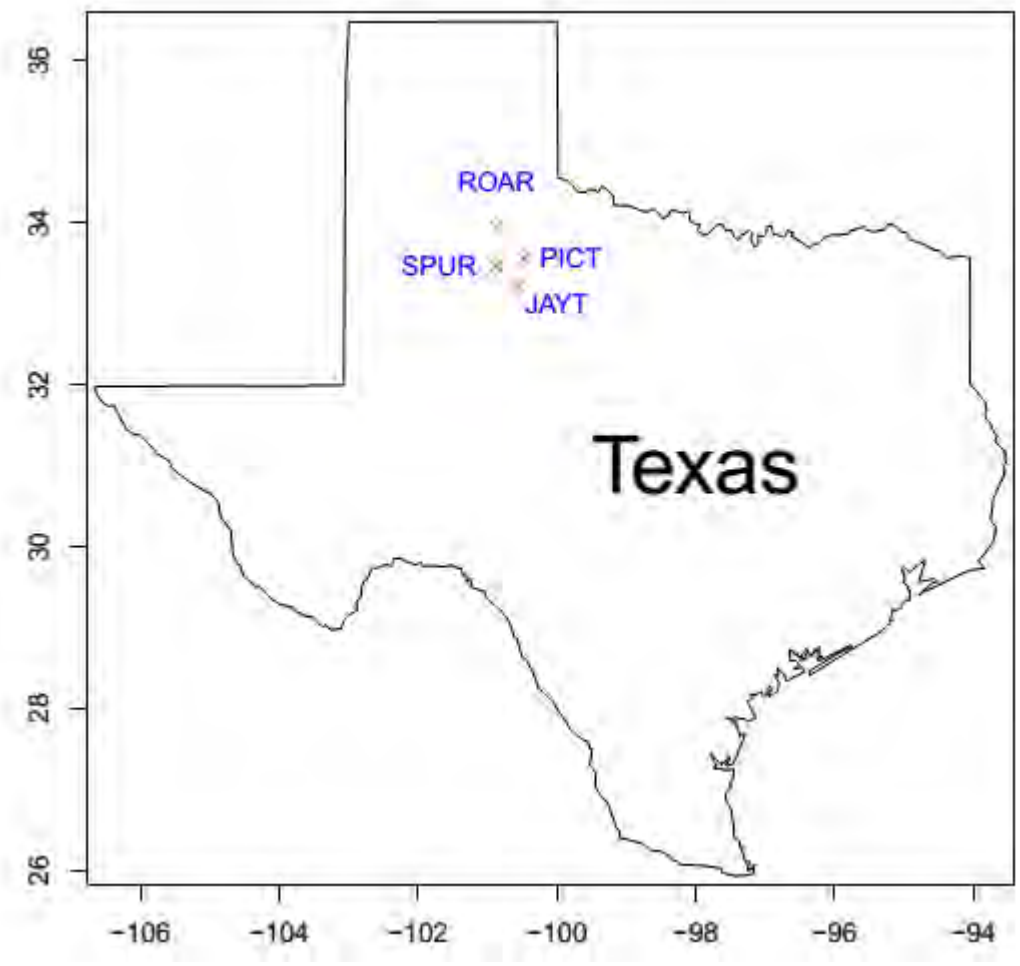

Figure 21: Map of the four locations in West Texas

Winds in this area are mainly from the south or north as shown by the wind roses in Figure 22, where the petals are the frequencies of wind blowing from a particular direction, and the colored bands are the ranges of wind speed. Given the flatness in this area, the spatial correlation in wind can be captured when a southerly wind is blowing: wind at ROAR will

mostly be just a shift from wind at SPUR. This means that to forecast the future wind speed at ROAR, it is definitely helpful to use the current and just past wind information at SPUR. Similarly, when the wind is blowing from the south or southeast, wind information at JAYT and PICT helps in predicting the wind speed at ROAR. A good forecasting model should take into account both spatial and temporal correlations in wind.

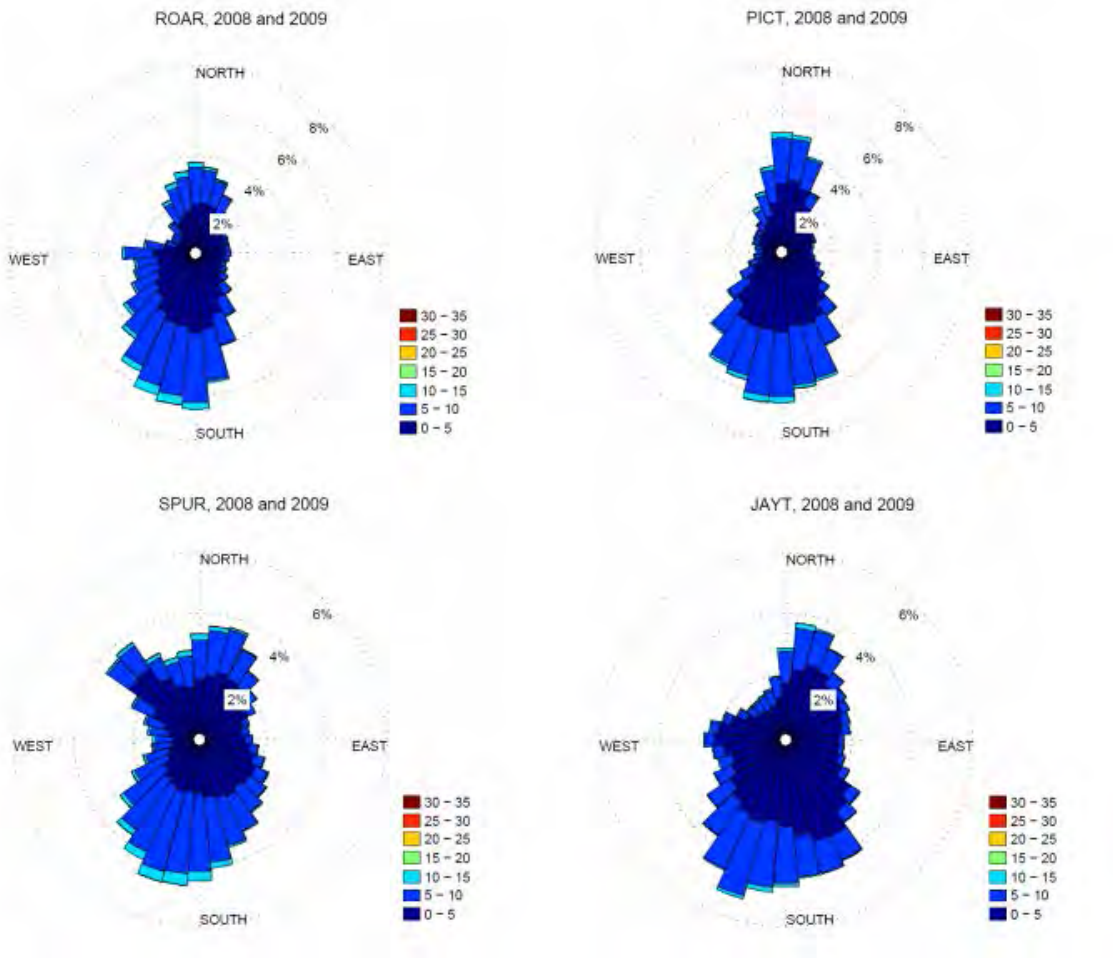

Figure 22: Wind roses of the four locations in West Texas

5.2.2 Space-time Statistical Forecasting Models

Four statistical models were used, PSS, AR, TDD and TDDGW, to forecast short-term wind speed at each of the four sites. In the first two models, only the temporal correlation in wind is considered, while the TDD and TDDGW models utilize wind information from the other three locations so that both spatial and temporal correlations in wind are taken into account. Moreover, the TDDGW model incorporates geostrophic information into the TDD model.

To make it simple, we describe the four models in the setting of forecasting wind speed at ROAR. Let $y_{R,t}$, $y_{S,t}$, $y_{J,t}$, and $y_{P,t}$ denote the wind speed at time t at ROAR, SPUR, JAYT, and PICT, respectively, and $\theta_{R,t}$, $\theta_{S,t}$, $\theta_{J,t}$, and $\theta_{P,t}$ denote the wind direction at time t. The goal is to estimate $y_{R,t+k}$, or the k-step-ahead wind speed at ROAR, denoted as $\hat{y}_{R,t+k}$, where each step is 5 minutes.

5.2.3 Persistent Forecasting

In the PSS model, it is assumed that the future wind speed is the same as the current one. For example, if $y_{R,t}$ is the wind speed at time *t* at ROAR, then the *k*-step future wind speed is predicted as $y_{R,t}$, or $\hat{y}_{R,t+k}=y_{R,t}$. PSS works very well for very short-term forecasting, such as 10-minute-ahead. The PSS model is usually treated as a reference and an advanced forecasting model is thought to be good if it outperforms PSS.

5.2.4 Autoregressive Models

AR models predict the future wind speed as a linear combination of past wind speeds. In our case, we apply AR to model the center parameter, $\mu_{R,t+k}$, in equation (5.2) (defined in the next part) as follows:

$$\mu^{r}_{R,t+k}=\alpha_0+\sum_{i=0}^{p}\alpha_{i+1}\mu^{r}_{R,t-i}. \tag{5.1}$$

The AR model assumes that future wind speed is related to historical wind information only at the same location, without considering the spatial correlation. Bayesian Information Criteria is used to selectee the order *p*.

5.2.5 Spatio-Temporal Trigonometric Direction Diurnal Model

The TDD model is an advanced space-time statistical forecasting model. It generalizes the Regime-Switching Space-Time model [101] by including wind direction in the model. As a probabilistic forecasting model, the TDD model estimates a predictive distribution for wind speed at time *t+k*, thus providing more information about the uncertainty in wind. More recently, the TDDGW model, which incorporates geostrophic wind information into the TDD model, was proposed [105] and more accurate forecasts are obtained than from the TDD model.

In the TDD model, it is assumed that $y_{R,t+k}$ follows a truncated normal distribution on the nonnegative real domain, that is, $y_{R,t+k}\sim N^{+}(\mu_{R,t+k},\sigma^{2}_{R,t+k})$ (this can be detected by the density plots in Figure 23), with center parameter $\mu_{R,t+k}$ and scale parameter $\sigma_{R,t+k}$. The key to achieve accurate forecasts lies in modeling these two parameters appropriately.

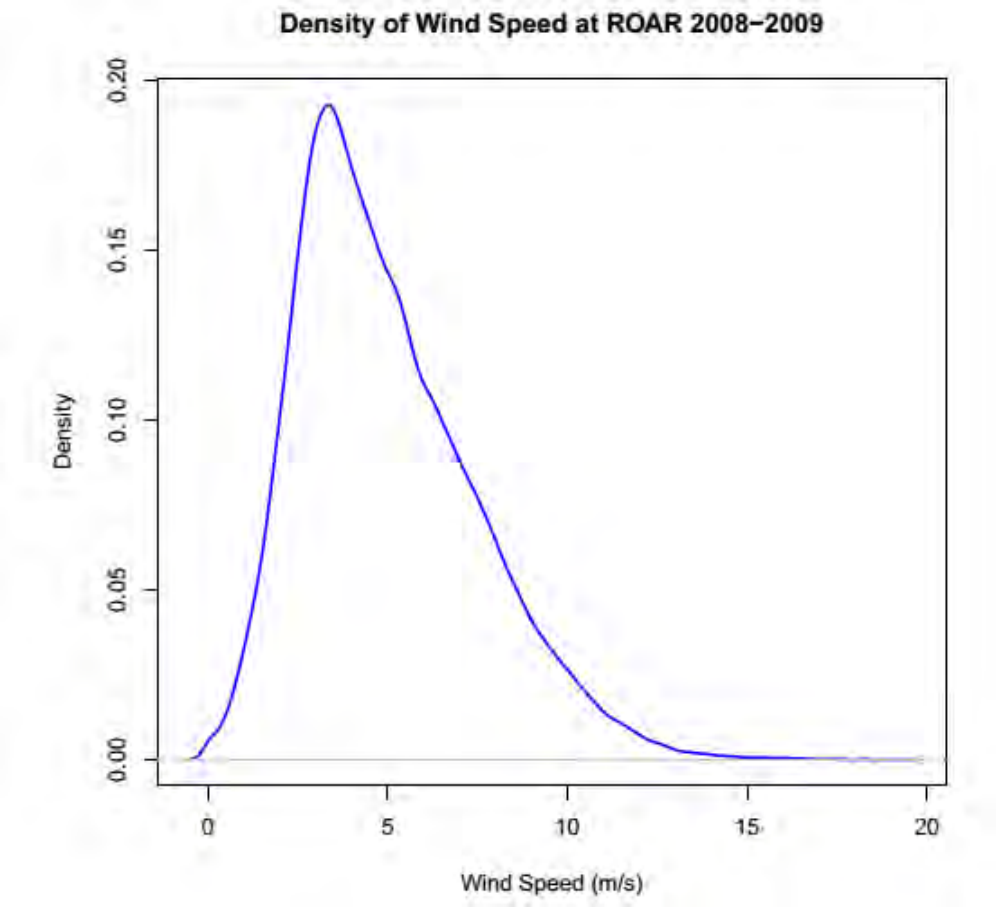

Figure 23: Wind speed density at ROAR 2008-2009

The center parameter, $\mu_{R,t+k}$, is modeled as $\mu_{R,t+k}=D_{R,h+k}+\mu^{r}_{R,t+k}$,
where $D_{R,h+k}$ is made of trigonometric functions to fit the diurnal pattern of the wind speed. Specifically,

$$D_{R,h} = d_0 + \sum_{j=1}^{2}\left\{ d_{2j-1}\sin\left(\frac{2\pi jh}{24}\right) + d_{2j}\cos\left(\frac{2\pi jh}{24}\right)\right\}$$

where h=1,2,…,24; see Figure 24. Figure 24 is the functional boxplot [106] of daily wind speed from 2008-2009 with the solid white line as the mean wind speed over 24 hours, the solid black line as the median, and the dashed green line as the fitted daily pattern.

The residual series after removing the diurnal pattern, $\mu^{r}_{R,t+k}$, is modeled as a linear function of current and past (up to time lag p) wind speed residuals and trigonometric functions of wind direction residuals at ROAR, as well as SPUR, JAYT, and PICT as follows:

$$\mu_{R,t+k}^{r} = \alpha_0 + \sum_{s\in\{R,S,J,P\}} \sum_{i=0}^{p} \alpha_{i+1} \mu_{s,t-i}^{r} + \sum_{s\in\{R,S,J,P\}} \sum_{j=0}^{p} \left[\beta_{j+1} \sin(\theta_{s,t-j}^{r}) + \gamma_{j+1} \cos(\theta_{s,t-j}^{r}) \right]. \quad (5.2)$$

The scale parameter, $\sigma_{R,t+k}$, is modeled as

$$\sigma_{R,t+k} = b_0 + b_1 v_t, \quad (5.3)$$

where b_0, $b_1 > 0$ and v_t is the volatility value:

$$v_t = \left[\frac{1}{8} \sum_{s\in\{R,S,J,P\}} \sum_{i=0}^{1} (\mu_{s,t-i}^{r} - \mu_{s,t-i-1}^{r})^2 \right]^{1/2}.$$

Error! Bookmark not defined. The coefficients in equation (5.2) along with b_0, b_1 in equation (5.3) are estimated by the continuous ranked probability score method (see [107] for more details). Predictors in (5.2) are selected with the Bayesian Information Criteria (see [102]).

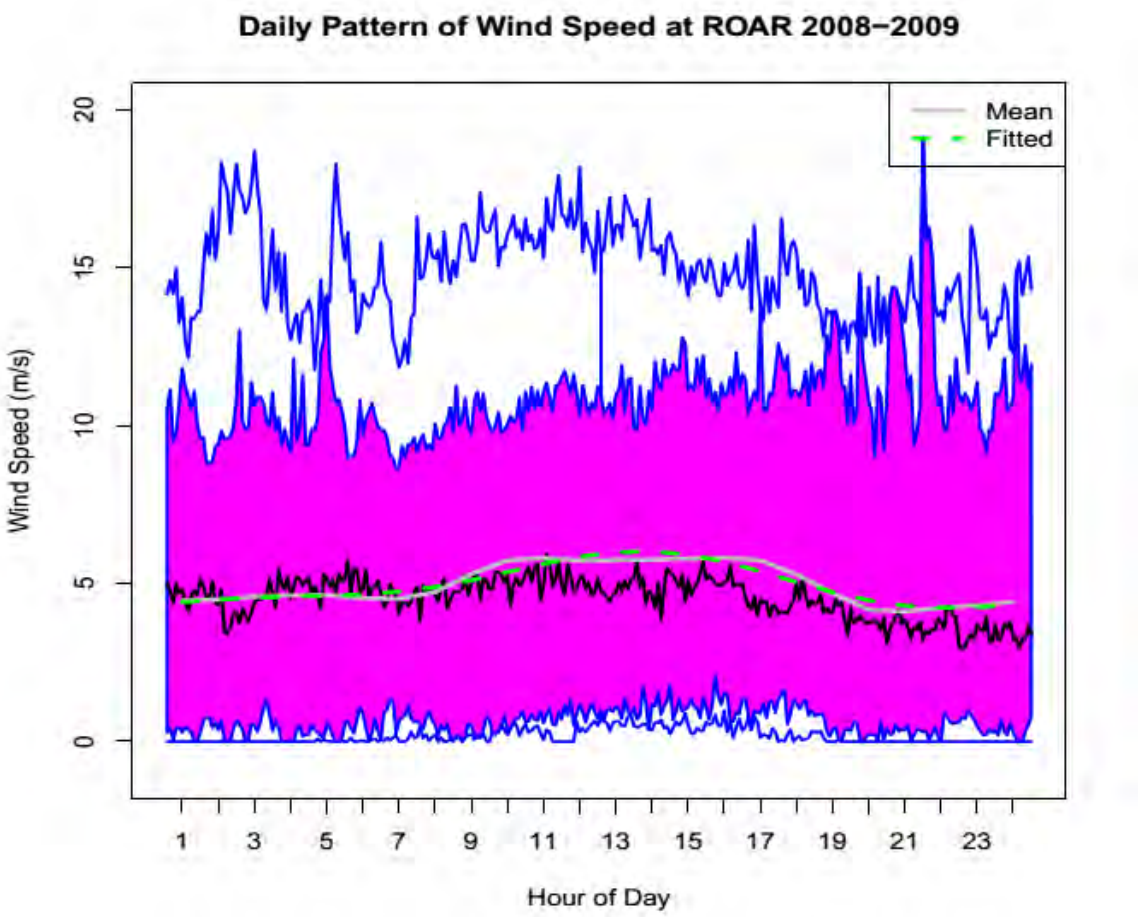

Figure 24: Functional boxplot [106] of daily wind speed at ROAR 2008-2009

As we know, pressure and temperature also have significant effects on wind speed. If this information could be taken into account in wind forecasting problems, more accurate forecasts would be expected. However, it was found that adding surface pressure and temperature directly into the center parameter model in (5.2) brings no improvement to the forecasting accuracy. This is the motivation of the TDDGW model. It takes geostrophic wind, which extracts information on pressure and temperature, into the TDD model as a predictor.

Geostrophic wind is the theoretical wind that results from an exact balance between the pressure gradient force (horizontal components) and the Coriolis force if there were no friction above the friction layer, and this balance is called the geostrophic balance. It is parallel to straight isobars. Figure 25 illustrates the difference between geostrophic wind (left) and real wind or surface wind (right).

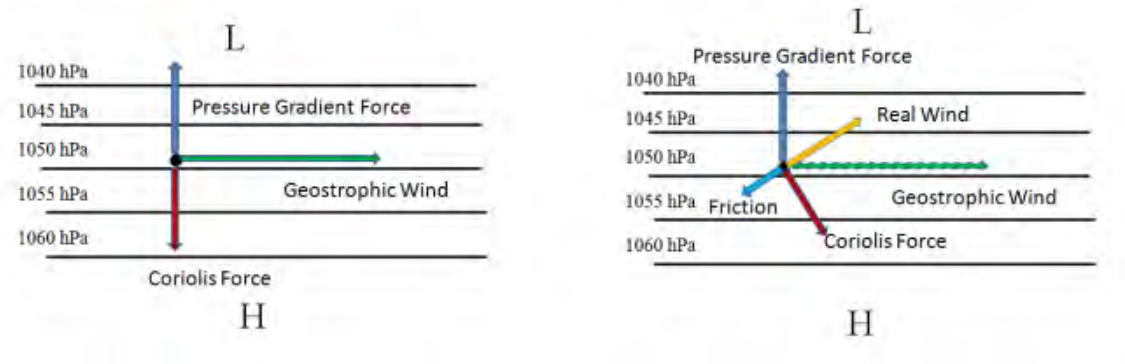

Figure 25: The pressure gradient, Coriolis, and friction forces influence the movement of air parcels. Geostrophic wind (left) and real wind (right)

The approximation of geostrophic wind is based on Newton's Second Law. It involves calculation of geopotential heights by referring to 850 hPa based on pressure, temperature and elevation, and fitting a plan of the geopotential height gradient in the region. Due to the space limitation, we refer readers to [105] for more detailed information.

The TDDGW model incorporates geostrophic wind into the TDD model, as shown in (5.4). This model not only includes important information on pressure and temperature, but it also has a clear and meaningful physical interpretation. Moreover, the TDDGW model keeps the advantage of the TDD model, namely to account for the spatio-temporal correlation in wind:

$$\begin{aligned}\mu^r_{R,t+k} &= \alpha_0 + \sum_{s\in\{R,S,J,P\}}\sum_{i=0}^{p}\alpha_{i+1}\mu^r_{s,t-i} + \sum_{k=0}^{p}c_{k+1}gw^r_{R,t-i} \\ &+ \sum_{s\in\{R,S,J,P\}}\sum_{j=0}^{p}\left[\beta_{j+1}\sin(\theta^r_{s,t-j}) + \gamma_{j+1}\cos(\theta^r_{s,t-j})\right],\end{aligned} \tag{5.4}$$

where $gw^r_{R,t-i}$ s are the residuals of the geostrophic wind after removing the diurnal pattern and the c_{k+1}s are the coefficients. Since geostrophic wind is above the friction layer, it

covers a large area. That means locations within the small area of our interests have very similar geostrophic values. We therefore use the geostrophic wind variable as a common predictor as shown in (5.5.4). The median of the truncated normal distribution is used as a point forecast:

$$z^{+}_{1/2}=\mu_{t+1}+\sigma_{t+1}\cdot\Phi^{-1}[1/2+(1/2)\Phi(-\mu_{t+1})/\sigma_{t+1}],$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution.

5.3 Forecasting Results and Comparison

In this section, the aforementioned four forecasting models are implemented to forecast 10-minute-ahead, 20-minute-ahead and up to 1-hour-ahead wind speed at the four locations in West Texas on one day each month except May 2010 (the days are chosen randomly).In the AR, TDD and TDDGW models, a 45-day sliding window of observations prior to the forecast is used to estimate coefficients in the models in which the variables are selected using the data from 2008 and 2009. For the diurnal pattern, the averages of 45 days' hourly wind speeds are used.

To evaluate the performance of the four forecasting models, mean absolute errors (MAE), defined below, are calculated from the forecasts on the 11 days and listed in Table 7:

$$MAE=\frac{1}{T}\sum_{t=1}^{T}|y_{R,t+k}-\hat{y}_{R,t+k}|,$$

where T=3168 for 11 days.

Table 7: MAE values of the 10-minute-ahead, 20-minute-ahead and up to 1-hour-ahead forecasts on 11 days' in 2010 from the PSS, AR, TDD and TDDGW models at the four locations (smallest in bold)

Location	Model	10 min	20 min	30 min	40 min	50 min	60 min
PICT	PSS	0.56	0.72	0.84	0.92	1.00	1.08
	AR	0.55	0.70	0.80	0.87	0.94	1.00
	TDD	**0.54**	**0.68**	**0.77**	0.84	0.90	0.95
	TDDGW	**0.54**	**0.68**	**0.77**	**0.83**	**0.89**	**0.94**
JAYT	PSS	0.50	0.63	0.71	0.78	0.83	0.89
	AR	0.48	0.60	0.68	0.75	0.8	0.86
	TDD	**0.47**	**0.57**	**0.64**	0.69	0.73	0.78
	TDDGW	**0.47**	**0.57**	**0.64**	**0.68**	**0.71**	**0.75**
SPUR	PSS	0.51	0.64	0.73	0.81	0.86	0.92
	AR	0.49	0.61	0.69	0.76	0.80	0.86
	TDD	**0.48**	**0.59**	**0.67**	0.72	0.76	0.81
	TDDGW	0.49	**0.59**	**0.67**	**0.71**	**0.75**	**0.79**
ROAR	PSS	0.55	0.71	0.82	0.92	0.98	1.02
	AR	**0.54**	0.68	0.78	0.86	0.92	0.96
	TDD	**0.54**	**0.67**	0.77	0.85	0.90	0.93
	TDDGW	**0.54**	**0.67**	**0.76**	**0.82**	**0.87**	**0.90**

From Table 7, we can see that MAE values increase by column, which means that the forecast accuracy reduces when the forecasting horizon, *k*, gets larger. Among the four models, the AR, TDD, and TDDGW models have smaller MAE values than the PSS and the space-time models, TDD and TDDGW, are more advanced than the PSS and AR models with smaller MAE values. As expected, by incorporating the geostrophic wind information, the TDDGW model increases its predictive accuracy. Its MAE values are reduced further compared with the TDD model, especially for 40-min-ahead or longer time lead forecasting. Relative to the MAE value of PSS, the TDDGW model obtains 15.7% reduction at JAYT for 1-hour-ahead forecasting, while it is 12.4% for the TDD model. This means that, by incorporating geostrophic wind information into the TDD model, we can further reduce the forecasting error up to 3.3%, based on the relative MAE value to PSS. The computational time for hour-ahead forecast using a laptop PC for one step of the TDDGW model is approximately 1.5 minutes, and the computational time for one step of TDD is approximately 1 minute. In contrast, recent literature suggests that it is currently impossible to compute the NWP models for hour-ahead scheduling purposes [85]. Therefore, data-driven statistical wind forecast models provide computationally feasible solutions for near-term operations for system operators. In the next two sections, the economic benefits of improved forecast are quantified in look-ahead dispatch models.

5.4 Power System Dispatch Model

With the spatio-temporal wind forecast models, we present in this section a critical assessment of the economic performance for power system operations. The power system scheduling framework formulated in this project is designed with two layers: 1) Day-ahead

reliability unit commitment (RUC) [108,109] and 2) robust look-ahead real-time (every 5 minutes) scheduling.

5.4.1 Day-ahead Reliability Unit Commitment

The structure of the two-layer dispatch model is described in Figure 26. The models of day-ahead reliability unit commitment (RUC) and real-time scheduling are presented below.

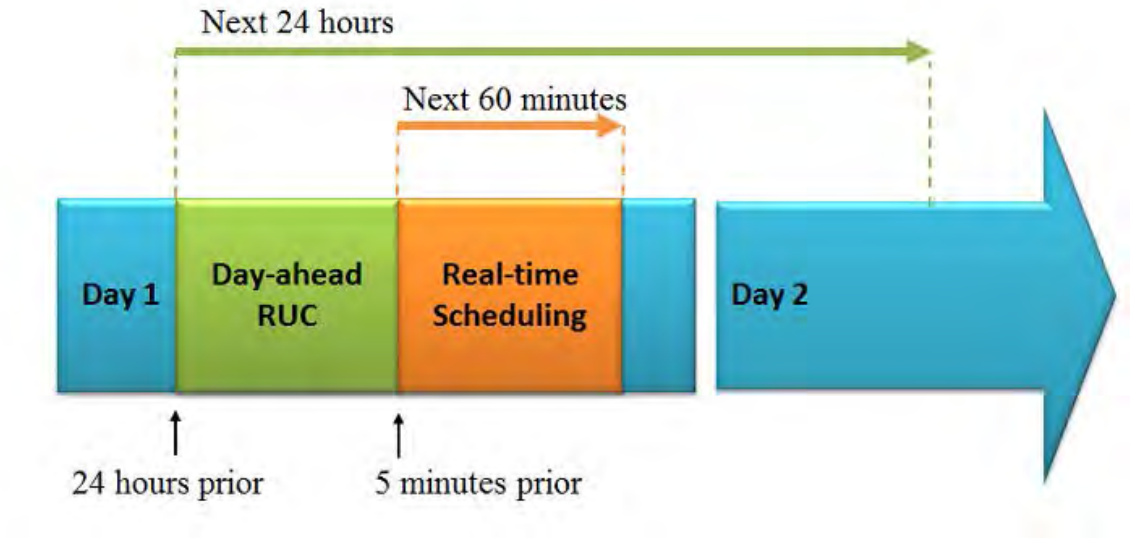

Figure 26: Two-layer dispatch model

The day-ahead reliability unit commitment ensures the reliability of the physical power system after clearing the day-ahead market. It takes place 24 hours prior to the real-time operation, as shown in Figure 26. Energy balancing and ancillary services (reserve services) are co-optimized with start-up/shut-down decisions. The model is formalized as follows:

$$\min_{P_{G_i}^k, P_{W_i}^k, P_{Rs_i}^k, x_i^k} : \sum_{k=k_0}^{T} [\sum_{i\in G} C_{G_i}(P_{G_i}^k) + \sum_{i\in G} C_{Rs_i}(P_{Rs_i}^k) + \sum_{i\in W} C_{W_i}(P_{W_i}^k) + \sum_{i\in F} C_{U_i}(x_{U_i}^k) + \sum_{i\in F} C_{D_i}(x_{D_i}^k)] \tag{5.5}$$

s.t.

$$\sum_{i\in G} P_{G_i}^k + \sum_{i\in W} P_{W_i}^k = \sum_{i\in D} P_{D_i}^k, k=k_0,\ldots,T \tag{5.6}$$

$$\sum_{i\in G} P_{Rs_i}^k \geq Rs^k, k=k_0,\ldots,T \tag{5.7}$$

$$|\mathbf{F}^k| \leqslant \mathbf{F}^{max}, k=k_0,\ldots,T \tag{5.8}$$

$$|P_{G_i}^{k}-P_{G_i}^{k-1}|\leqslant P_i^R\Delta T, i\in G, k=k_0,\ldots,T \tag{5.9}$$

$$x_i^k P_{G_i}^{\min}\leqslant P_{G_i}^{k}\leqslant x_i^k P_{G_i}^{\max}, i\in G, k=k_0,\ldots,T \tag{5.10}$$

$$P_{G_i}^{k}+P_{RS_i}^{k}\leqslant x_i^k P_{G_i}^{\max}, i\in G, k=k_0,\ldots,T \tag{5.11}$$

$$x_i^k-x_i^{k-1}\leqslant x_{U_i}^{k}, i\in G, k=k_0,\ldots,T \tag{5.12}$$

$$x_i^{k-1}-x_i^k\leqslant x_{D_i}^{k}, i\in G, k=k_0,\ldots,T \tag{5.13}$$

$$P_{W_i}^{min}\leqslant P_{W_i}^{k}\leqslant P_{W_i}^{max}, i\in W, k=k_0,\ldots,T \tag{5.14}$$

$$P_{W_i}^{k}\leqslant P_{W_i}^{k}=\tilde{f}(\mathbf{P}_W), i\in W, k=k_0,\ldots,T \tag{5.15}$$

$$x_i^k, x_{U_i}^k, x_{D_i}^k\in Binary, i\in G, k=k_0,\ldots,T. \tag{5.16}$$

In the proposed formulation, the objective function (5.5) is to minimize the power system operating costs including generation cost, reserve cost and start-up/shut-down cost of units. This scheduling problem is subject to various security constraints. (5.6) is the energy balancing equation. (5.7) is the system reserve requirement, which is often assessed according to system reliability requirement. (5.8) is the transmission capacity constraints. (5.9) are the ramping constraints of all generation units. (5.10) are the generators' capacity limits for generator units. (5.11) are the combined capacity constraints of generator units for providing energy and reserve services. (5.12) and (5.13) are start-up/shut-down indicator constraints. (5.14) is the capacity limit of wind farms. In this project, wind resources are assumed not to participate into ancillary services market providing reserve services. (5.15) is the wind forecast for each wind farm at time *k*, the details of which are explained in Section 5.3. Equation (5.16) gives the binary constraints to integer decision variables.

5.4.2 Robust Look-ahead Economic Dispatch

Following the day-ahead scheduling from the previous subsection, we assume that system operators conduct a real-time dispatch every 5 minutes. We formulate this dispatch model as a multi-stage robust look-ahead economic dispatch to utilize the information of advanced spatio-temporal forecast. The robust look-ahead dispatch minimizes system operation cost over a horizon of multiple steps (e.g., one hour) for worst cases under predefined uncertainty set. As other look-ahead economic dispatch, only the dispatch decisions of the first step are executed. The updated information, such as wind forecast, load forecast and system conditions will be fed into the dispatch model for future decision-making. The robust look-ahead economic dispatch is formulated as

$$\max_{\mathbf{u}\in \mathbf{U}} \min_{P_{G_i}^k, P_{W_i}^k, P_{SU_i}^k, P_{SD_i}^k} : \sum_{k=k_0}^{T} [\sum_{i\in G} C_{G_i}(P_{G_i}^k) + \sum_{i\in W} C_{W_i}(P_{W_i}^k)] \tag{5.17}$$

s.t.

$$\sum_{i\in G} P_{G_i}^k + \sum_{i\in W} P_{W_i}^k = \sum_{i\in D} \hat{P}_{D_i}^k, k=k_0,\ldots,T \tag{5.18}$$

$$|\mathbf{F}^k| \leqslant \mathbf{F}^{max}, k=k_0,\ldots,T \tag{5.19}$$

$$|P_{G_i}^k - P_{G_i}^{k-1}| \leqslant P_i^R \Delta T, i\in G\cup W, k=k_0,\ldots,T \tag{5.20}$$

$$\sum_{i\in G} P_{SU_i}^k \geq SU_D^k(\mathbf{u}), k=k_0,\ldots,T \tag{5.21}$$

$$\sum_{i\in G} P_{SD_i}^k \geq SD_D^k(\mathbf{u}), k=k_0,\ldots,T \tag{5.22}$$

$$P_{G_i}^k + P_{SU_i}^k \leqslant P_{G_i}^{\max}, i\in G, k=k_0,\ldots,T \tag{5.23}$$

$$P_{G_i}^k - P_{SD_i}^k \geq P_{G_i}^{\min}, i\in G, k=k_0,\ldots,T \tag{5.24}$$

$$P_{G_i}^{min} \leqslant P_{G_i}^k \leqslant P_{G_i}^{max}, k=k_0,\ldots,T \tag{5.25}$$

$$P_{W_i}^{min} \leqslant P_{W_i}^k \leqslant P_{W_i}^{max}, k=k_0,\ldots,T \tag{5.26}$$

$$P_{W_i}^k \leqslant \hat{P}_{W_i}^k, k=k_0,\ldots,T \tag{5.27}$$

$$0 \leqslant P_{SU_i}^k \leqslant P_{Ui}^R \Delta T, k=k_0,\ldots,T \tag{5.28}$$

$$0 \leqslant P_{SD_i}^k \leqslant P_{Di}^D \Delta T, k=k_0,\ldots,T. \tag{5.29}$$

The objective function (5.17) is to minimize the total operating cost for energy balancing. In real-time scheduling, various security constraints are considered. Energy balancing constraints are provided in (5.18). Transmission capacity constraints are given in (5.19). Ramping constraints of generators are presented in (5.20). We introduce short-term dispatchable (STDC) capacity to make sure the system has enough ramping capability to handle the uncertainty [110]. (5.21) and (5.22) are the upward/downward STDC balancing equations. The STDC are constrained by the ramping capability of each unit as presented in (5.28) and (5.29). Capacity constraints of conventional generators and wind farms are described in (5.25) and (5.26), respectively. (5.23) and (5.24) are combined capacity constraints between generation capacity and STDC. The dispatch points of wind generation should be no larger than the forecasted wind production potentials, as is shown in (5.27).

The uncertainty set U is given by (5.30).

$$
\begin{aligned}
&U(\hat{\mathbf{P}}_W^k, \overline{\mathbf{P}}_W^k, \hat{\mathbf{P}}_D^k, \overline{\mathbf{P}}_D^k, \overline{\Pi}_W^k, \underline{\Pi}_W^k, \overline{\Pi}_D^k, \underline{\Pi}_D^k, \overline{u}_W^k, u_W^k, \overline{u}_D^k, u_D^k) := \\
&\{\hat{\mathbf{P}}_W^k \in \mathbb{R}^{|\mathbf{W}|}, \hat{\mathbf{P}}_D^k \in \mathbb{R}^{|\mathbf{D}|} : \sum_{i \in \mathbf{W}} \frac{\hat{P}_{W_i}^k - \overline{P}_{W_i}^k}{\overline{u}_{W_i}^k - \overline{P}_{W_i}^k} \le \overline{\Pi}_W^k, \sum_{i \in \mathbf{W}} \frac{\overline{P}_{W_i}^k - \hat{P}_{W_i}^k}{\overline{P}_{W_i}^k - u_{W_i}^k} \le \underline{\Pi}_W^k, \\
&\sum_{i \in \mathbf{D}} \frac{\hat{P}_{D_i}^k - \overline{P}_{D_i}^k}{\overline{u}_{D_i}^k - \overline{P}_{D_i}^k} \le \overline{\Pi}_D^k, \sum_{i \in \mathbf{D}} \frac{\overline{P}_{D_i}^k - \hat{P}_{D_i}^k}{\overline{P}_{D_i}^k - u_{D_i}^k} \le \underline{\Pi}_D^k, \hat{P}_{W_i}^k \in [u_{W_i}^k, \overline{u}_{W_i}^k], \\
&\forall i \in \mathbf{W}, \hat{P}_{D_j}^k \in [u_{D_j}^k, \overline{u}_{D_j}^k], \forall j \in \mathbf{D}\}
\end{aligned}
\tag{5.30}
$$

Here $\hat{\mathbf{P}}_W^k$ is the vector of wind production potential forecasts fed into the dispatch model as presented in (5.27). $\overline{\mathbf{P}}_W^k$ is the vector of expectations of wind forecast for each location at each time step. $\bar{u}_W^k$ and $\underline{u}_W^k$ defines the upper bounds and lower bounds of wind forecast deviation from the expectation. $\overline{\Pi}_W^k$ is defined as the budget of uncertainty for wind forecast, which takes the value between 0 and |**W**|, where |**W**| is the number of wind sources modeled in the system. If the budget is set to be 0, the problem formulation turns out to be deterministic. As $\overline{\Pi}_W^k$ grows, the uncertainty set U enlarges, which indicates the system operation is toward more risk-averse, and the system is protected against higher degree of uncertain conditions.

Similarly, for the load forecast uncertainty, $\hat{\mathbf{P}}_D^k$ is the vector of load forecasts fed into the dispatch model. $\overline{\mathbf{P}}_D^k$ is the vector of expectations of load forecast for each bus at each time step. $\bar{u}_D^k$ and $\underline{u}_D^k$ defines the upper bounds and lower bounds of load forecast deviation from the expectation. $\overline{\Pi}_D^k$ is defined as the budget of uncertainty for load forecast, which takes the value between 0 and |**D**|.

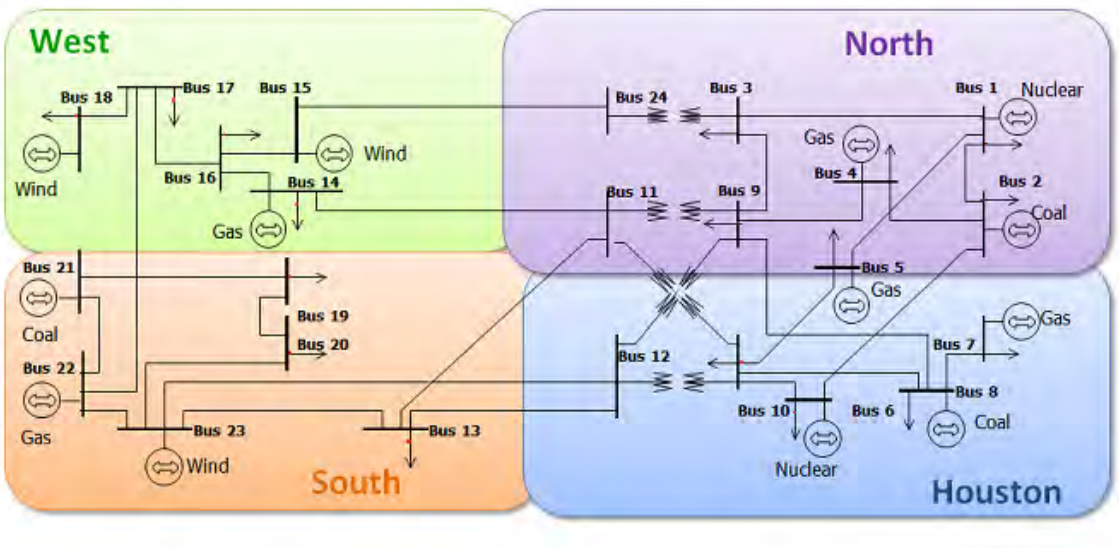

Figure 27: The IEEE RTS-24 system (modified)

5.5 Numerical Experiment

In this section, we conduct a numerical experiment on a 24-bus system to critically assess the *operational economic benefits* from improved short-term forecasts.

5.5.1 Simulation Platform Setup

The numerical example is modified from the IEEE Reliability Test System (RTS-24) [111]. The simulation duration is 24 hours. The operation interval in real-time scheduling is five minutes. The look-ahead horizon in real-time scheduling is 1 hour. Load profiles for 48 hours are collected from the ERCOT System [112]. Loads are scaled and factored out according to the portion of different buses. Wind forecasts are generated by various models discussed in Section 5.3 with forecast horizon which ranges from 10 minutes to 60 minutes. Then the wind power forecasts are transformed from the wind speed forecasts based on the Nordex 2.5 MW power curve.

The generator parameters are scaled according to [113]. The generator capacity portfolio (the installed capacity percentage of different technologies) is configured and scaled from the real ERCOT system [113]. The ramping rates and marginal costs are applied as shown in Table 8.

Table 8: Generator parameters

Bus	Type	Cap. (MW)	Cost ($/MWh)	Ramping (MW/min)
1	Nuclear	140	15	1.12
2	Coal	540	20	10.8
4	Natural Gas	300	40	15
5	Natural Gas	510	37	33.15
6	Nuclear	150	11	1.35
7	Natural Gas	490	39	34.3
8	Coal	165	23	3.135
14	Natural Gas	170	38	15.3
16	Wind(JYAT)	200	6	18
18	Wind(PICT)	240	4	24
21	Coal	300	21	5.4
22	Natural Gas	725	36	79.75
23	Wind(SPUR)	70	5	7.7

In the numerical studies, simulations of twelve sample days[1] are conducted. The twelve days are randomly selected as representative days for each month in 2010, as shown in Table 9.

Table 9: Sample days in simulation study

Sample	Date	Sample	Date	Sample	Date
Day 1	10-Jan	Day 5	9-May	Day 9	8-Sep
Day 2	27-Feb	Day 6	16-Jun	Day 10	19-Oct
Day 3	12-Mar	Day 7	1-Jul	Day 11	22-Nov
Day 4	21-Apr	Day 8	17-Aug	Day 12	6-Dec

[1] Day 5 for TDDGW model is not available due to the inaccessibility of measurement data. Therefore, for the averaged MAE comparison of wind speed forecasts, only 11 days are considered. For the independent studies of economic benefits in power system operation, Day 5 for models other than TDDGW are presented.

5.5.2 Results and Analysis

In this section, the simulation results of the numerical experiments are presented.

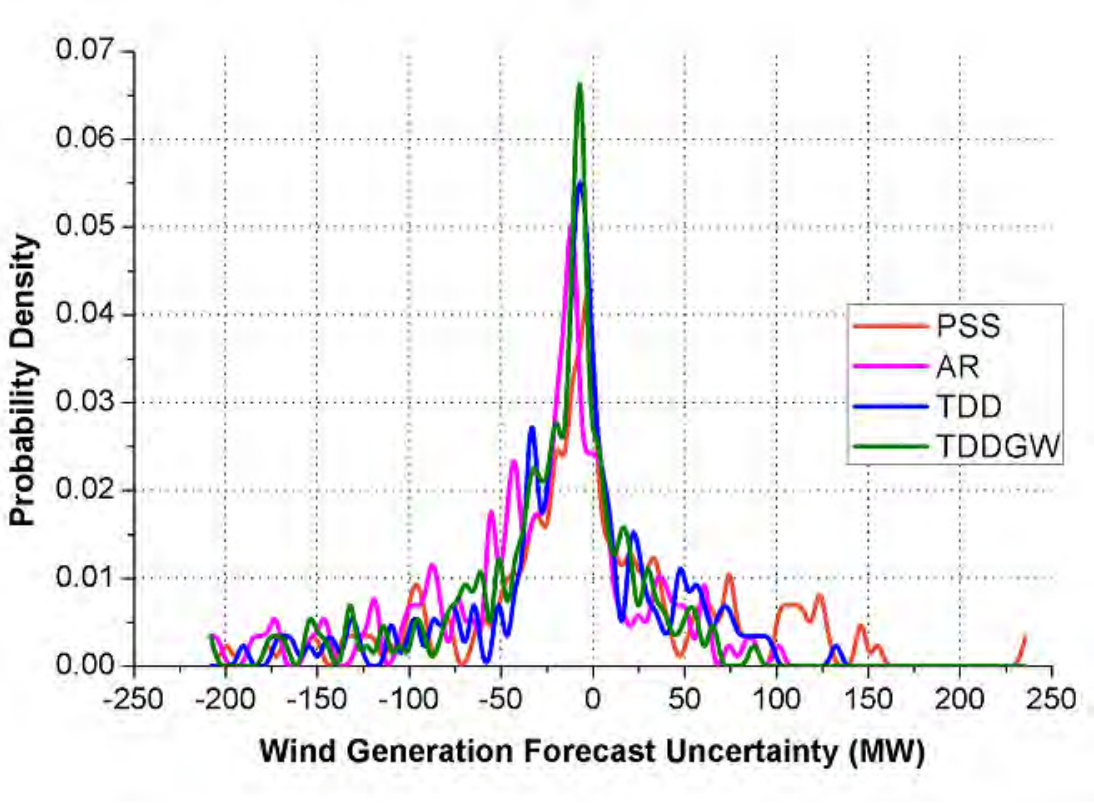

Figure 28: Distribution of forecast errors under different forecast models

The distribution of the forecast errors of the wind generation reveals the accuracy of the forecast approach. The distribution of its errors for the perfect forecast (PF) with 100% accuracy is a concentrated spike at the zero origin of the x-axis. The better the forecast accuracy the closer the distribution pattern is to the central spike. A forecast model with poor accuracy has its errors distributed widely. The probability density distributions of the wind generation forecast errors (for a 200 MW wind farm) using the PSS, AR, TDD and TDDGW models under various simulation days are presented in Figure 28. As we can observe, the distribution of the forecast errors of the PSS model is relatively spread out. The distribution of forecast errors of the TDD model is concentrated and has a higher central spike than do the AR and PSS models. The central spike of the TDDGW is higher than that of any other models. The shape of the forecast error distribution of the TDDGW model is closest to that of the perfect forecast. This is also verified by the wind speed forecast MAE presented in Table 7.

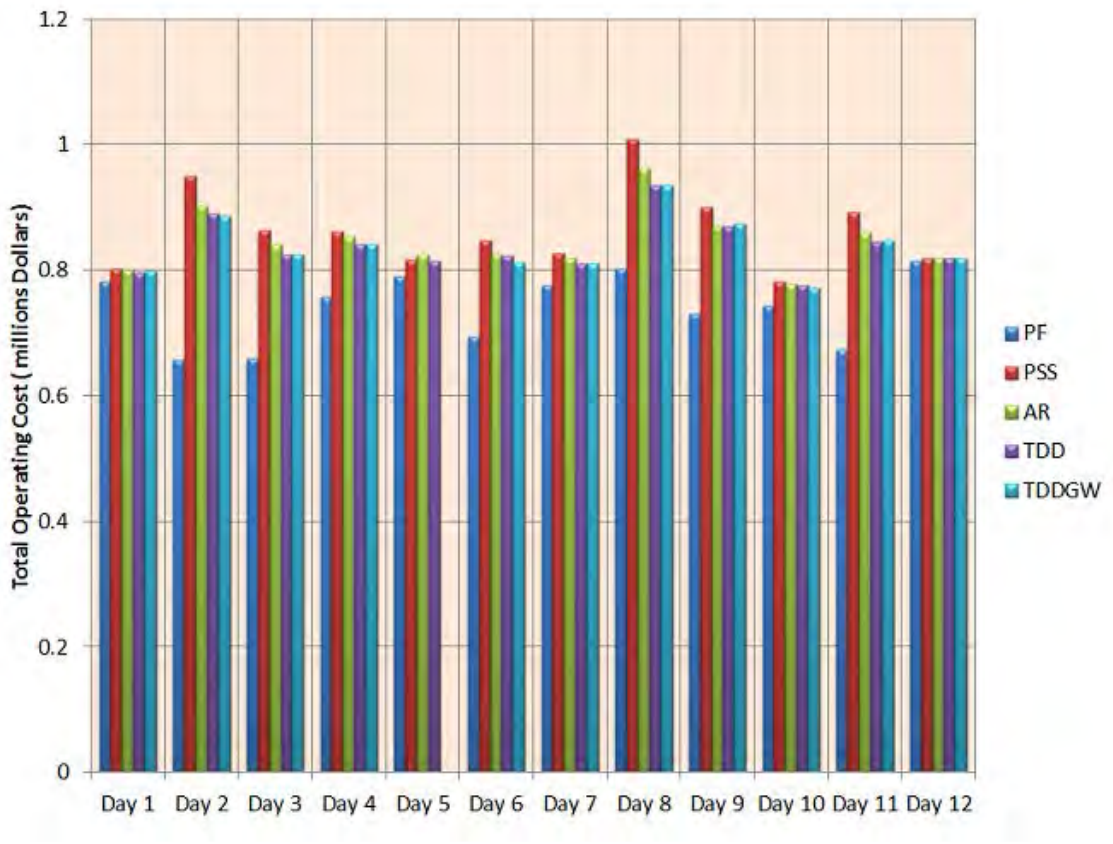

Figure 29: Total operating cost using different forecast models

By incorporating different forecast models into the power system economic dispatch, the economic performance differs. The economic performance results of Case A are presented in Figure 29, which includes the total operating cost of each simulation day. The costs of the perfect forecast, PSS, AR TDD and TDDGW models are represented by the blue bar, the red bar, the green bar, the purple bar and the cyan bar, respectively. As we can see, for most of the cases, the spatio-temporal forecasts (TDD and TDDGW) have lower operating costs than do the PSS and AR models.

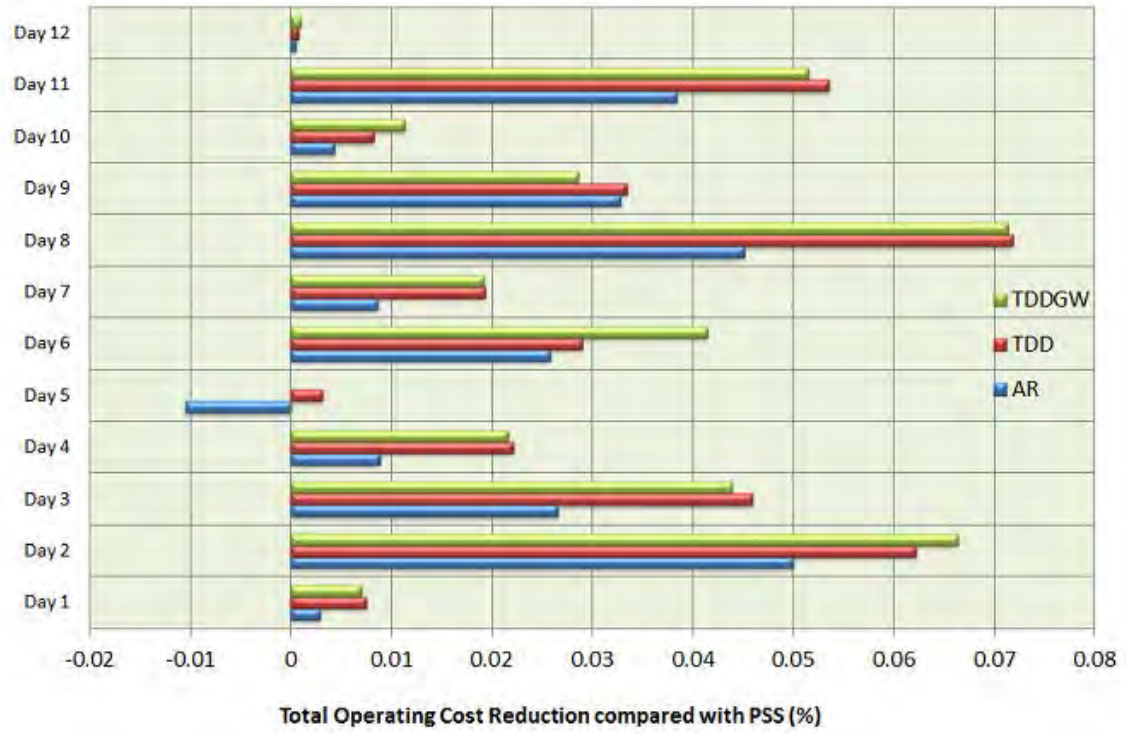

Figure 30: Operating cost reduction using different forecast models

Taking the PSS model as a benchmark, the reduction in operating cost by percentage using various forecast models is presented in Figure 30. As we can see, the TDD and TDDGW models, which consider spatio-temporal wind correlation, outperform the AR model and the PSS model in most of the cases. By incorporating the effect of geostrophic wind, the TDDGW model can have a lower system operating cost than the TDD model. For most of the days, the AR model performs better than the PSS model. However, it is observed that for some days (Day 5), the AR model does not produce as good a forecast as does the PSS model. That is the limitation of wind forecast based on purely historical data. In contrast, by incorporating spatial correlations, the TDD model can produce more accurate forecasts than can the PSS model and enable lower system operating costs.

5.6 Summary

Spatio-temporal wind forecast models (TDD and TDDGW models) are used and critically evaluated in this project. It is shown that by incorporating spatial correlations of neighboring wind farms, the forecast quality in the near-term (hours-ahead) could be improved. The TDD and TDDGW models are incorporated into a robust look-ahead economic dispatch and a day-ahead reliability unit commitment. Compared with conventional temporal-only statistical wind forecast models, such as the PSS models, the spatio-temporal models consider both the local and geographical wind correlations. By leveraging both temporal and spatial wind historical data, more accurate wind forecasts can be obtained. The potential economic benefits of advanced wind forecast are illustrated using a modified IEEE RTS 24 bus system. It is observed that the spatio-temporal model

can increase wind resources utilization, and reduce system costs against uncertainty. Such data-driven statistical methods for short-term wind forecast are also applicable in other similar regions with high wind penetration.

Future work will investigate the applicability of the proposed dispatch model to large-scale wind farms, such as offshore wind farms. Given the more consistent wind pattern over larger geographical areas, the potential benefits of the proposed method could be higher. Another important avenue for future research is to analyze the tradeoff between communication/computation burdens and the improved economic benefits by incorporating more spatially correlated wind data into power system dispatch models.

6. Distributed Database for Future Grid

6.1 Architectures

6.1.1 Distributed Grid Control and Management Architecture

The operation and control paradigm used today by the electricity industry is largely centralized, based on traditional Supervisory Control and Data Acquisition (SCADA) architecture originally proposed in the 60's, following the advent of the digital computer [114,115]. By using this centralized control paradigm the industry has been successful in achieving its objectives of providing reliable electricity at reasonable cost.

Requirements of secure integration of less predictable and variable renewable energy, deployment of smart grid sensing and communication infrastructure, and emerging consumer objectives result in substantially amplified communication, data management, and computation requirements and in highly complex decision-making problems [116,117,118].

The future grid will consist of billions of devices and millions of spatially distributed decision makers. These new (smart) devices are equipped with advanced electronics and embedded systems. The emerging decision makers, i.e., microgrids, buildings, homes, etc., are being instrumented with sensing and communication systems to enable automation, while electricity users have ever-growing access to ubiquitous information about energy use [119,120]. This is illustrated n Figure 31.

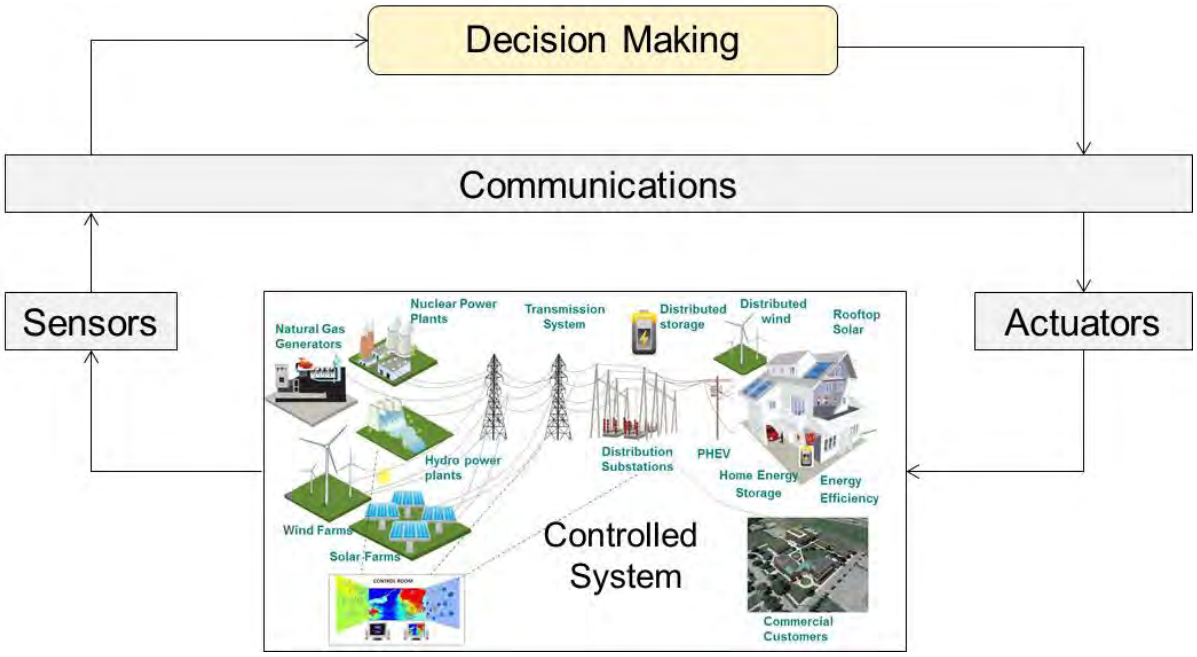

Figure 31: Illustration of the future grid.

Formidable benefits to the electricity system, the electricity industry at large, and consumers can be achieved if these actors and system devices can be coordinated in an intelligent manner. The current centralized architecture suffers from fundamental scalability limitations when the number of control points and decision-makers increases drastically. Thus, there is a need for an evolved model for managing the electricity infrastructure and the industry at large, one that reduces complexity, enables decentralized decision-making, allows for more flexible control, and supports desirable value propositions.

We describe a decentralized paradigm that will enable the electricity grid to operate with architectural characteristics similar to the internet: highly accessible, scalable to billions of actors, layered, and flexible. We adopt a broad and holistic approach, touching a wide domain of emerging concerns. Its architectural scope is therefore the operation and control of the entire electricity infrastructure.
Rather than proposing or designing architecture around a specific existing or developing technology, we start with the requirements both explicitly and implicitly stated in the academic literature and industrial community. We then discuss a framework that allows for many different technologies to contribute to grid operation and further propose some examples of what those might be. Thus, we seek to define an architecture that clearly reveals the challenges and needs for future grid design. Additionally, such a framework can be used to parse preexisting smart grid architecture proposals for the purpose of comparison, a task that can be difficult otherwise due to inconsistent vocabularies and different sets of requirements and objectives.

6.1.1.1. Limitations of the Centralized Architecture

The centralized grid control architecture, based on SCADA systems initially designed in the 60's, has grown and assimilated many new technologies without altering the underlying structure. However, this system will not continue to be scalable for the following reasons:

Expanding data requirements: The number of monitoring and control devices is increasing by several orders of magnitude over traditional data acquisition. In a centralized architecture, the control center faces a dilemma between incomplete information (e.g. coarse granularity) and an information tsunami [123], both of which prevent effective control action [123-126].

Communication bottlenecks: Centralized control will require moving massive amounts of data and hence expensive, mostly dedicated communications.
Intractable control and optimization problems: Traditional methods for real-time dispatch are based on instantaneous optimization without look-ahead capabilities and are deterministic; that is, they do not handle uncertainty and variability (as from renewable sources) [127]. Most current forms of stochastic optimization will be result in problem sets that are intractable in the required timeframe even with the most powerful supercomputers [128].

Risks of controlling large-scale renewable energy: It has been recognized that integration of large amounts of renewable energy poses operational challenges and can result in system events [129].

Growing complexity of system operations: Support for operator situational awareness is struggling to keep up [130]. The number and complexity of reliability and compliance procedures is growing rapidly as the industry integrates renewable energy and addresses concerns such as inter-area oscillations, the effects of demand response, and deployment of energy storage.

Growing complexity of market and regulatory framework: Current electricity markets exhibit fundamental market design limitations such as lack of direct interaction between consumers and producers, ad-hoc established market temporal scales, and conflict of interest between utility revenue and energy efficiency [131,132]. New propositions are needed that allow the markets to mature with direct participation of all the actors.

Cyber-security: Centralized control remains a cyber and physical security target [133,134]. It is based on the concept of bulk energy control centers, which require major infrastructure to be physically protected and usually redundant facilities, hardware and software infrastructure.

Data Privacy: A centralized framework results in the central organization controlling non-owned assets. This results in the need to send significant amounts of data from those non-owned control points. Data privacy concerns have been pointed out in smart grid pilots in the United States and have resulted in pushback from consumers [135].

6.1.1.2. Desired Architecture Attributes

The following are the desired attributes of the architecture as a cohesive framework to support efficient operation in the decades to come, as well as a smooth transition to that framework:

1. *Robustness*: The architecture must support the reliable operation of the grid under attack and loss of infrastructure modules including power, control, communications, computation, and markets components.
2. *Scalability*: the architecture must be scalable to billions of spatially distributed smart devices and millions of decision-makers.
3. *Generalizability*: the architecture must provide flexibility to support many types of power technologies and energy services.
4. *Technology independence*: the architecture must support interoperable technologies and deployment of innovative technology, even propositions not conceived today, without requiring architectural redesign.
5. *Backward-compatibility*: the architecture must be compatible with existing processes and must support continuity and integration of legacy systems.
6. *Incremental deployability:* the architecture must support incremental deployment of the various systems and technologies.

6.1.1.3. Need for Decentralized Framework

The deployment of massive sensing and communication infrastructure downstream from distribution systems to microgrids, feeders, and the customer facilitates a much more refined process of electricity use optimization. As a result, both the utility and the final electricity user will optimize objective functions associated with energy [121,122]. These objective functions of the user and the provider may not necessarily be aligned.

In the centralized architectural framework, there is one system level objective function, which corresponds to minimizing the total operating cost of a given geographic region for a given period of time. The various actors (producers, consumers, and distribution utilities) yield control of their assets to the Independent System Operator (ISO), who conducts centralized optimization and control. Smart grid pilots have pointed out the realization that users of electricity will no longer have the same objective function as the utility. In the long term, users may benefit from installation of local generation, such as rooftop solar panels or energy storage, to hedge against real-time pricing or pursue sustainability or environmental goals. Because of the diversity of objectives that the power grid actors have and the common goal of maintaining a reliable system, a control platform should optimize individual behaviors in concert to achieve system level objectives. Distributed operation and control design implementations fall across a spectrum between cooperative (as in frequency droop control) and competitive (as in price-based unit commitment or so-called transactive control [138]). Certainly, a combination of cooperative and competitive strategies will be necessary to obtain a reliable grid with equitable opportunities for users to maximize their profit.

Ultimately, the disparate classes of industrial, commercial, and residential consumers will evolve into similar economically motivated entities equipped with much more powerful information and control technologies. These actors will pursue their own long- and short-term energy objectives and will have all the incentives to invest, operate, and control more advanced technologies to meet their energy objectives [139]. The management architecture must provide the necessary elements to support such distributed decision making, and it must provide the mechanism so the various players can meet their objectives subject to system level constraints of reliability and sustainability.

Decision-making in the future grid will take place in a distributed manner, and it will be characterized by numerous actors pursuing their own energy objectives while adhering to protocols to address system level objectives and constraints.

6.1.2 Distributed Databases

A database management system (DBMS) is considered to be *distributed* if it provides access to data that resides at multiple (local) sites in a network. Some of these local data may not provide full support for schema integration, distributed query management and distributed transaction management. A distributed DBMS is considered to be *heterogeneous* if the local nodes have different types of computers, operating systems, database software, and different schemas.

Writing scalable, distributed applications that can take the advantage of distributed data and distributed computing requires formal and careful design. Databases can play a major role into the realization of distributed decision-making (DDM) and data-driven distributed control systems. This goal of distributed decision-making is supported by the natural evolution of modern enterprise systems. Today, enterprises are widely distributed, resulting in systems that are more reliable, more efficient, and highly responsive. A fundamental reason behind distributed processing is to be able to solve big and complicated problems by using a form of divide-and-conquer approach.

Today, the most advanced and complex businesses and industries are beginning to rely on distributed rather than centralized databases for the following reasons: a) cost, which continues to decrease, b) scalability, possible with off-the-shelf processors, c) high availability, achieved by mirroring and replicating data, d) integrability, of in-house systems with standard DBMS, e) support for legacy systems, f) support for new applications, g) flexible to support for new market forces, business reorganizations, and new business models.

Among the main challenges of distributed databases are:

a) Distributed database design
b) Distributed query processing
c) Distributed directory management
d) Distributed concurrency control and deadlock management
e) Reliability of distributed DBMS

We will discuss some of these issues as part of the design of a distributed database platform for smart grid.

6.1.3 Distributed Database Requirements

We assume that smart grid and future grid applications will involve a setting that includes heterogeneous distributed DBMSs that need to be integrated. Desirable features of these systems to achieve an integrated and seamless distributed DBMS are:

a) Schema integration
b) Distributed query processing
c) Distributed transaction management
d) Administrative functions
e) Security management

Furthermore, a critical requirement of these databases is soft-real-time capabilities necessary to support decision-making associated with controlling the actual electricity infrastructure. This needs to be facilitated by rapid production and consumption of data. A hierarchical structure of the database can be considered as a canonical design. A hierarchical system will aggregate data from numerous smaller units and can represent the

general state of a specific component of the system at various levels. The electricity grid control can be mapped into a hierarchical structure mapped to a hierarchical distributed DBMS. This structure would not only enable better observability and understanding of the overall system, but also provide better control of the specific components in question.

In order to better design a database for the smart grid, we would need to understand the extent of data that needs to be handled. We will consider a setting with a distribution utility serving residential or commercial consumers (prosumers). We assume an average of 50 appliances per prosumer, each recording a set of 20 values such as voltage, power, current, etc. If data was captured with a granularity of one minute, we would have approximately 44.6 million records in a month generated per prosumer. Given that this is the amount of data for a single prosumers, and considering that a large utility may have one million customers, it is reasonable to estimate that for a large utility the data generated will easily be in the trillions of records every month. We would therefore need a database system that is not only capable of handling such volumes of data, but also one that mimics the actual hierarchical and spatial structure of the grid control. Moreover, we would inevitably need an efficient archiving strategy for past data. This would entail ensuring that necessary data isn't lost but at the same time, data that is no longer directly relevant must be condensed/deleted as part of a retention policy.

6.1.4 Distributed Database Design

6.1.4.1. Main Use Cases

Different database systems have different strengths and weaknesses. Many databases are designed in a manner such that they are highly optimized for read operations, but very slow for write operations. It is therefore essential to do a detailed study of the various use cases in question and select a database package that can meet the requirements in an optimal manner.

In the case of the Smart Grid, we see that the data flow is dominant in one direction, i.e. from the clients to the database. The write requests outweigh the reads. Assuming a reporting frequency of 1 report every 10 seconds, we are looking at 360 reports being sent per hour from each client to the server. This is a pretty intense amount of data being sent to the server, all with the intention of writing to the database. Therefore, we need a database that performs very well under such a heavy load of write operations and this also gives us affordability in terms of slight delays with read operations (as compared to write operations).

6.1.4.2. Latency

Latency is defined as the total amount of time taken by a service to respond to a request from a client. Typically this consists of the time taken after the request has been dispatched from the client to the time it receives a response from the server. The overall latency of a system is determined by a variety of factors. The following are three primary factors:

1) Network time (connection bandwidth)

2) Response time from server
3) Load on the server

We now discuss the effect of the network on system latency in detail. Response time is the time taken by the server to react to and formulate a result from the time it receives a request. This includes the time spent in interacting with the database and any additional business logic (computation) that is performed on the data retrieved from the database.

The load on the server is directly related to the number of requests (and therefore the number of clients using the service). The higher the load on a server, the more the time it takes to respond. This also means that the server needs to have a queuing logic that determines what requests are handled first. Usually, a priority policy for requests is established. The most commonly used scheduling logic includes paradigms such as FIFO (First in First Out), LIFO (Last in First Out), random order, etc.

In the case of the Smart Grid, it is primarily the database response time that affects the server response time. This is because of the write nature of the operations performed on the smart grid. It is therefore essential for us to minimize the write latency. For this we would need to pick a database package that is write friendly and can handle numerous concurrent write requests. Write requests are also called *insert* operations.

Figure 32 compares the insert latencies of a few commercial database systems. The graph was built using data generated by an experiment that was performed on a local desktop machine. The experiment compares the insert latency of 2 main systems – Redis and Hbase, two very popular distributed data stores in the market today. The green and red bar refers to two different queries (one with the word "user" in it and one without it) that were used to break any possible caching mechanism that existed in the database. This helps us generate numbers from actual operations and not cached/optimized responses.

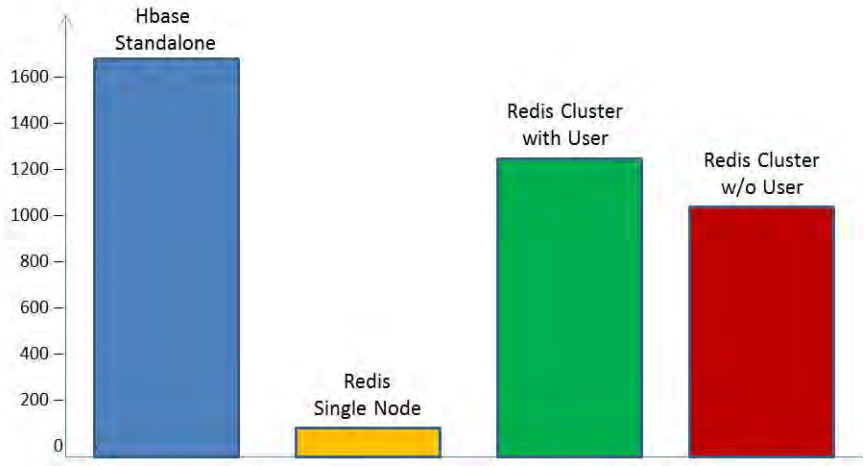

Figure 32: Insert latencies in commercial DBMSs

The average latency of a Google search result is 100 milliseconds. For most HTTP based internet services, the latency does not exceed 500 milliseconds. Latency is therefore indirectly, a measure of the quality of service being provided. Ensuring that a service has

minimal response time is the duty of the service provider. Network times are often not within the hands of the service provider, it is therefore essential to minimize the response time and optimize the positioning of servers.

6.1.4.3. Network Effects

The physical location of the server plays an extremely important role in the overall latency seen by the system. Network time is one of the most expensive and time consuming resources in the industry today. Therefore, to ensure minimum cost and time spent, we must ensure that the servers are located as close to the clients (geographically) as physically possible.

The graph in Figure 33 illustrates this point. The Milwaukee Digital Agency conducted an experiment in which they measured the average download speed (indirectly time spent) from various server locations to reach the client (located in Milwaukee). As expected, it was observed that proximity and latency were inversely correlated. This can also be used to illustrate a very important aspect about database systems. Server proximity in distributed database systems is of prime importance. If the time spent increases in the network for every instance of data exchange, not only does the client spend more time waiting for a response, but the server is also at risk of throttling. This is because the server is exposed to various requests from near and far clients. As a result, the requests tend to accumulate due to a largely random and unpredictable arrival time of the requests to the server. Therefore, the server could see sudden spikes in the request density. This could result in throttling of the database and in a more extreme case, Denial of Service (DoS) attacks.

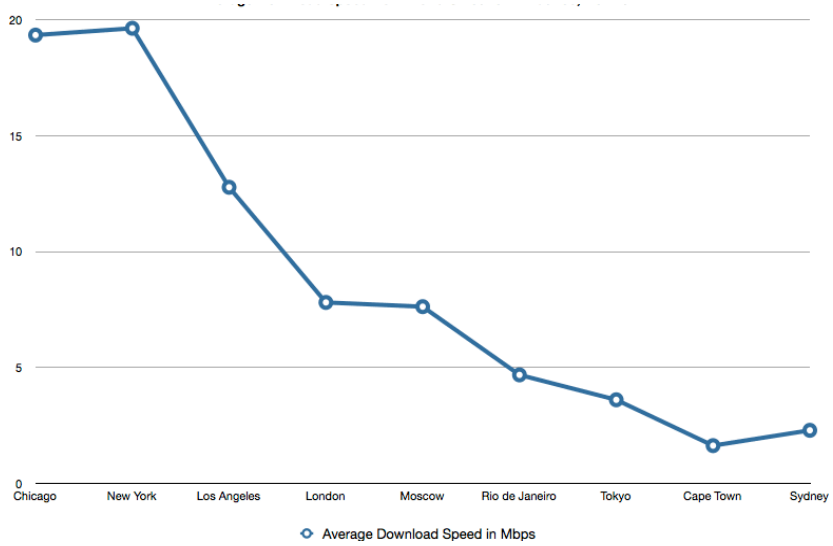

Figure 33: Average download speed from various cities to Milwakee
(image courtesy Milwaukee Digital Agency)

DoS attacks are a common technique used by malicious attackers to try and choke the server with several concurrent requests which the server does not possess enough capacity to handle, eventually leading to the server crashing. This could result in a significant loss of data and in addition, down time of the service.

In the case of the Smart Grid, (due to such a high frequency of reports being sent to the server) even a small amount of down time can result in a significant loss of data. The structure of DoS is illustrated in Figure 34.

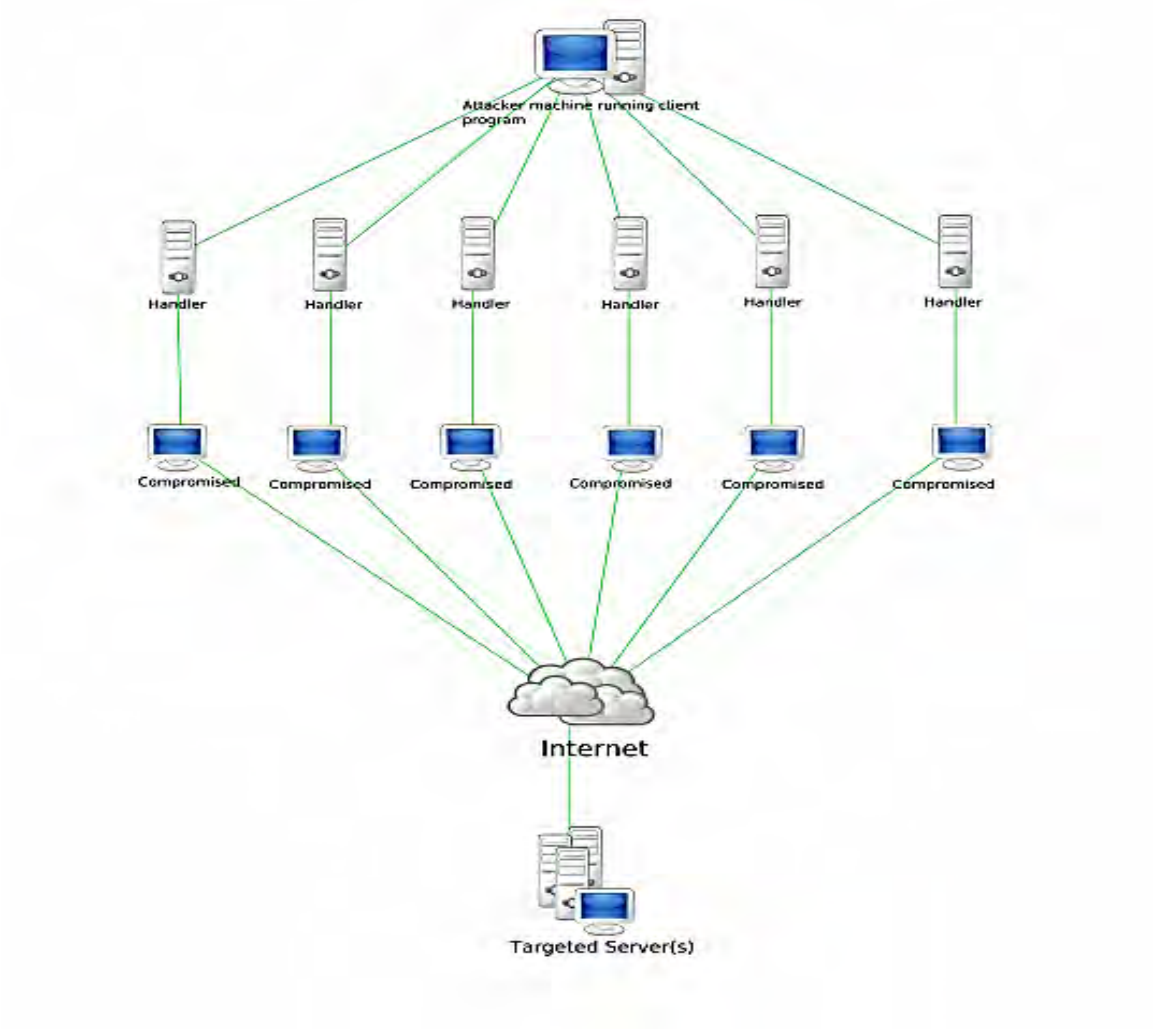

Figure 34: Structure of denial of services attack

6.1.4.4. Concurrency Control

Concurrency is one of the most common problems seen in high demand database systems. The main challenge with concurrency is ensuring that the integrity of the data is maintained regardless of the scheduling scheme used to handle the requests. This is an inevitable situation because with high frequency systems, there are bound to be numerous requests occurring at the same time (concurrently).

Figure 35 illustrates a common case of concurrency issues seen in a banking application. The square represents the bank balance of a joint account held by Jane and John. With an initial value of $1000, both John and Jane assume they can withdraw any amount lesser than $1000. As a result, both submit requests (simultaneously) to withdraw $700 and $500

respectively. Due to poor concurrency control measures, the application receives and processes both these requests simultaneously. As a result, a total sum of $1200 is withdrawn despite the balance being only $1000. As a result, the net balance reads -200.

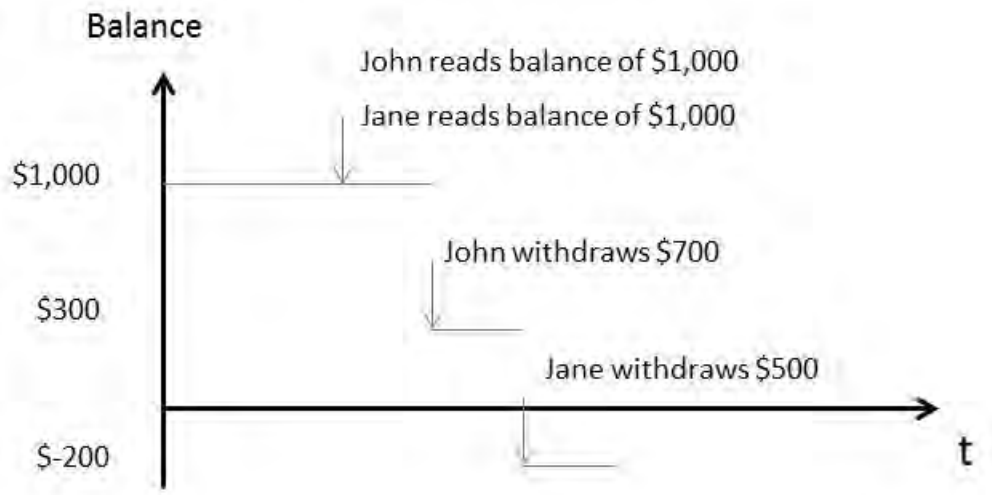

Figure 35: Example of concurrency management

To tackle the type of concurrency problems illustrated above, database systems follow a set of guidelines. These guidelines are also referred to as the ACID rule:

a) *Atomicity:* each operation is broken down into single step tasks and only after all the tasks in an operation are completed, the operation is marked as successful.
b) *Consistency:* every operation on the database must maintain a consistent state of the data after it is completed i.e. transition from one consistent state to another.
c) *Isolation:* each operation must perform in isolation to others. This ensures that values writes by one operation are not overwritten by another.
d) *Durability:* data written by operations must be persistent and survive system crashes and failures (typically done by using secondary storage mediums such as hard drives, disks, etc.)

Locking is a common technique used to handle concurrency in databases. Each transaction acquires a lock or a hold on a certain value that it intends to read from or write to. This serves the purpose of notifying other transactions of the intention. A transaction manager will usually establish rules or a priority order of locks on a value and ensure that only transactions that are eligible will be granted access to the values they are requesting for. There is also usually a time limit for which the transaction can hold a lock on a value. This ensures that multiple transactions don't wait for the same lock resulting in a deadlock situation.

Modern non-relational databases typically don't stress on the importance of data consistency. This is largely because ensuring consistency in the database requires a lot of checks and significantly slows down various data operations. As a result, a lot of the modern databases prioritize latency requirements over consistency. This can result in a few

instances of data misrepresentation to the users, but the cost is usually worth the benefit for those use cases.

This tradeoff is measured and documented. It is known as the CAP theorem. The CAP theorem states that "*It is impossible for any database system to provide Consistency, Availability and Partition Tolerance all at once*".

The image in Figure 36 conveys pictorially the CAP theorem, where the database systems of today are positioned with respect to the principles of the theorem. Most modern systems today prioritize high availability over the data consistency. This is expected, given the use case they deal with. Most internet companies such as Facebook, Google, Yahoo etc. aim for 100% availability. Looking at the nature of user data these firms receive, it is easily seen that consistency is not a significant issue. Therefore, it is possible to provide such low service latencies by compromising on strict rules about data integrity, which traditional models of databases hold very dear. Given the nature of data explosion happening in today's market, a majority of the systems are moving towards this model which ensures high availability and prioritizing minimal latency.

We see a similar situation in the case of smart grid applications. A system must be capable of storing one usage report (typically five values) every second from every single client connected to the grid. This is a significant amount of data and we must ensure that availability is as high as possible. Owing to the nature of the grid, every report sent out is critical and therefore we must ensure low latency as well. Therefore, as is the case of most large scale distributed systems today, it would be beneficial to use only a few consistency rules in certain specific layers of the database system.

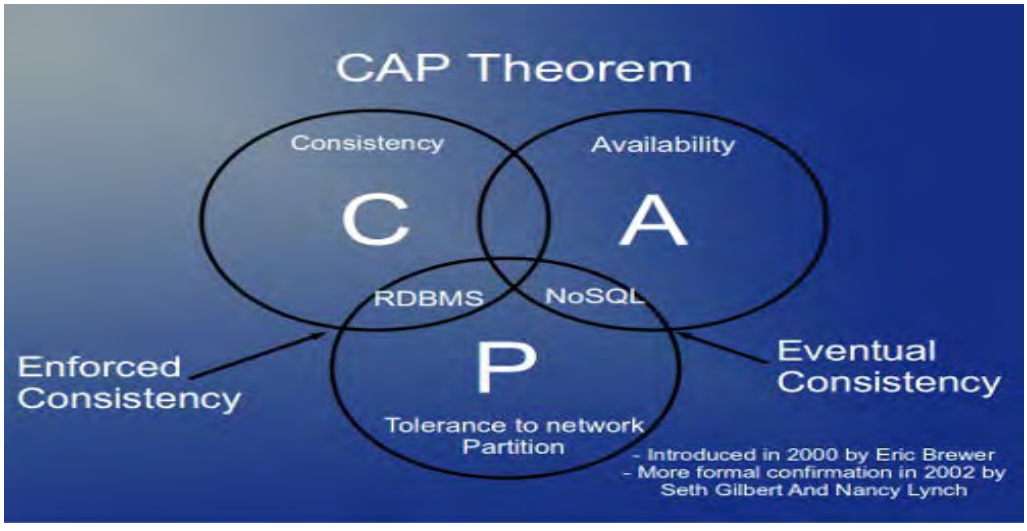

Figure 36: CAP theorem by Stephen Smith

6.1.4.5. Storage Mechanisms

Storing data in a distributed manner while maintaining logical consistency is one of the biggest challenges faced by the industry today. Numerous custom solutions have been developed and have also worked well, but most are designed with a specific use case in

mind. There is usually always some amount of calibration in the configuration required to optimize the data storage capabilities in distributed file systems. The most common and effective solutions are discussed below.

Distributed File Systems (DFS) are systems which typically operate over a network. A DFS appears as if it exists locally, but in reality is split up over multiple servers. Client server architecture is used for file sharing. Moreover, a fixed protocol is used for communication between the various servers, thus enabling multiple machines to coordinate regardless of platform or hardware. Google File System (GFS) and HDFS (Hadoop Distributed File System) are examples of DFS that are popular in the market. The image in Figure 37 represents a file being split over multiple servers. This enables access of the file to various locations spread across geographies.

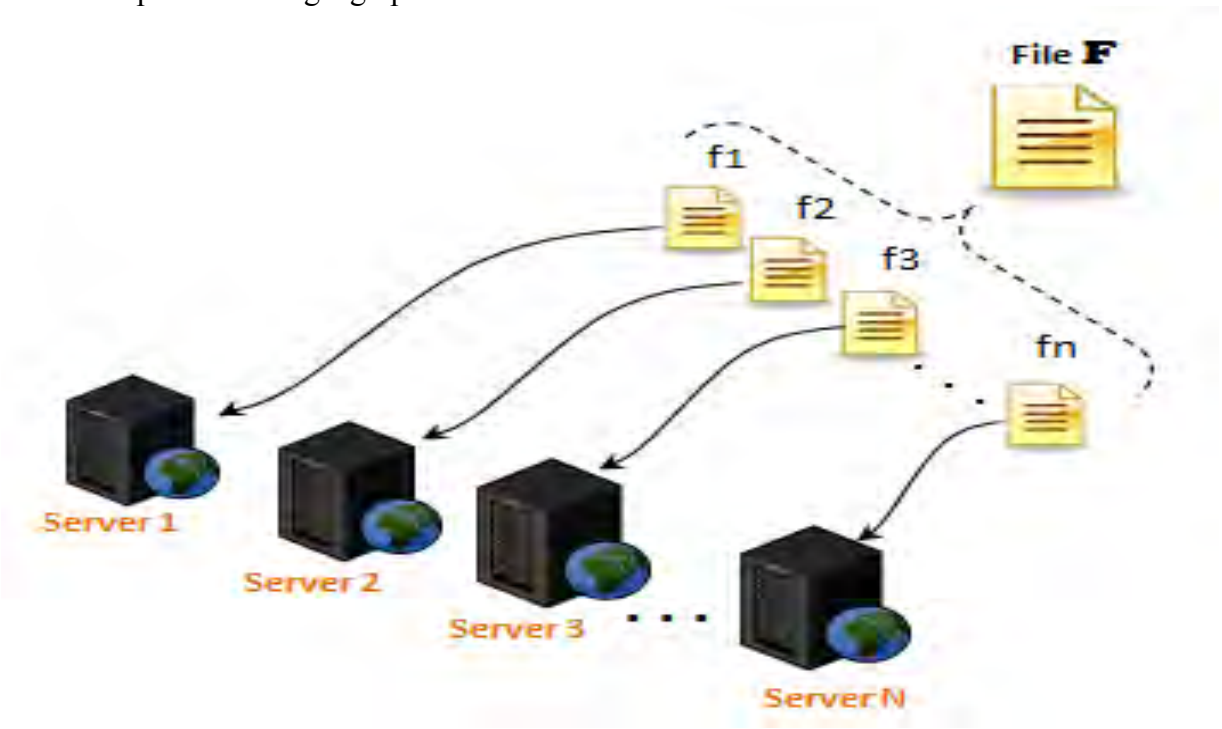

Figure 37: Split file in a distributed file system

Distributed Hash Tables are a good example of a distributed data structure. They are particularly useful when searching is a common operation on the data being stored. They provide an effective mechanism for distributing frequently access data across various locations. Latencies are usually low in distributed hash tables and that is why they are a preferred solution for many data retrieval problems.

The main logic behind a distributed hash table is the partitioning of keys into various ranges. This allows the placement of different keys (belonging to different ranges) in different locations. There is one collective node that has access to this metadata. This server receives the key being searched for and directs the query to the specific box that stores the key in question. This is a relatively simplistic model for storing data, but it is very effective. This also allows easy scaling by adding various levels of ranges of keys. Thus mimicking a tree like structure where each branch represents a range of values.

The image in Figure 38 illustrates what a typical hash hierarchy looks like.

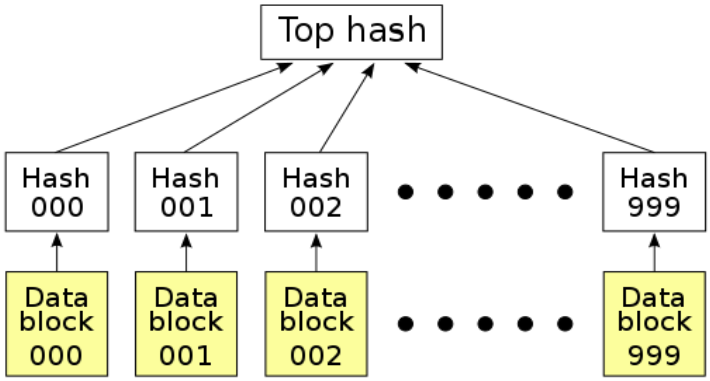

Figure 38: Hash list for distributed databases

Distributed Cache Stores are commonly used when there are certain values that are accessed repeatedly from servers located in multiple locations. They are typically smaller in capacity than DFS and distributed hash tables. They have a fixed lifetime, which means that they need to be refreshed if they are to persist in the cache store.

Data Sharding is a very useful technique that is used to split rows (of tables) of data horizontally, Figure 39. This means that a single table can be split up and stored in various locations and in different servers. This again allows the database to only access records that are relevant to that location or context. This greatly helps in minimizing latency and enabling faster lookup. A lot of the large scale databases rely heavily on sharding to split up the data into coherent portions. Shards can be created from the data based on a number of factors and not just frequency of access from a location. Factors like data integrity are also considered while sharding data.

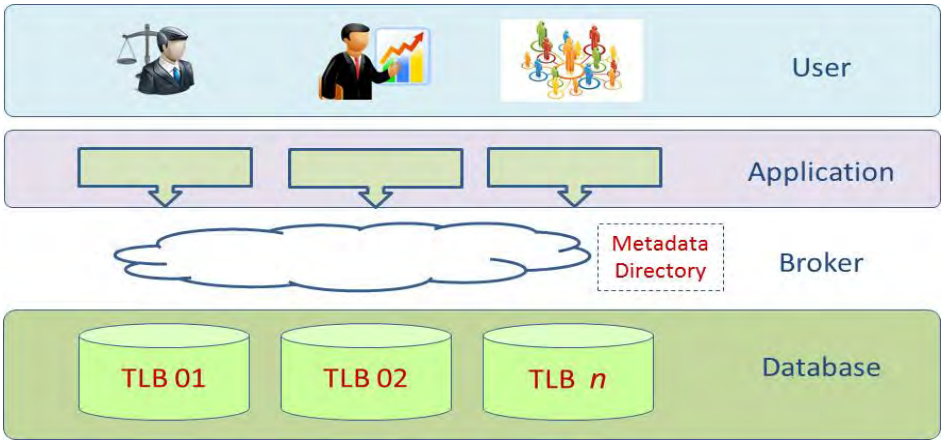

Figure 39: Database sharding

6.1.4.6. Security and Privacy of Data

With the ever increasing data demands of the industry today, it has become increasingly important to address security. Especially when user data is concerned, corporations take extra precaution to avoid breaches and privacy violations.

The most common form of **security** vulnerabilities are those that arise from unauthorized access. For this reason, authentication is a vital part of database security. Any kind of data that is inserted, edited or removed from the database must be highly regulated and monitored. Strong authentication mechanisms will also ensure that only programs with appropriate access are modifying specific tables and rows. It is also imperative to keep sufficient backup of the data. This will not only serve as a failsafe mechanism to retrieve lost data in case of a security breach, but will also be useful in cases where the data is corrupted or malformed due to programming errors.

Ensuring data **privacy** has become essential, especially where sensitive user data is concerned. The most common scope of privacy violations are loopholes in the communication platform. Therefore, most web based services rely on HTTPS now for data transmission. HTTPS is a more secure version of HTTP and it uses varying degrees of encryption to ensure that user data is protected from malicious intrusions and attacks during transmission. The main criticism for HTTPS is the slightly increased latency, but it is well worth the cost where user data security is concerned.

6.1.5 Distributed Database Simulation

A simulation approach was followed in order to obtain further insight into the performance of distributed DBMS for smart grid. The simulation was used to test the design of a distributed DBMS solution on a smaller scale. A web based simulation of the data generation process was developed to understand the quantity of data being produced and the techniques required to manage this data.

We have discussed that it is infeasible to send all the data directly to a utility database every minute due to factors such as network latency, write overheads, database overloads, etc. Instead, aggregation of data (a hierarchy) will enable the system to periodically communicate with the database and ensure that multiple records that form a logical partition can all be sent at once. This aggregation is facilitated by using multiple *Aggregator Nodes* (AN) that collect data from multiple *Devices* through their *Unit Meters* (UM) and pass the data on to a node higher in the hierarchy. Each aggregator node connects to its own database server (or a proxy server). The database server(s) are located relatively close to the nodes to minimize latency and maintain a sense of contextual segmentation of servers. *Computation Nodes* (CN) will run queries and computation against the data stored in these mid-level nodes. Results can be stored after aggregation at the highest level and past data can be archived with the desired level of granularity. This overall design for simulation is illustrated in Figure 40.

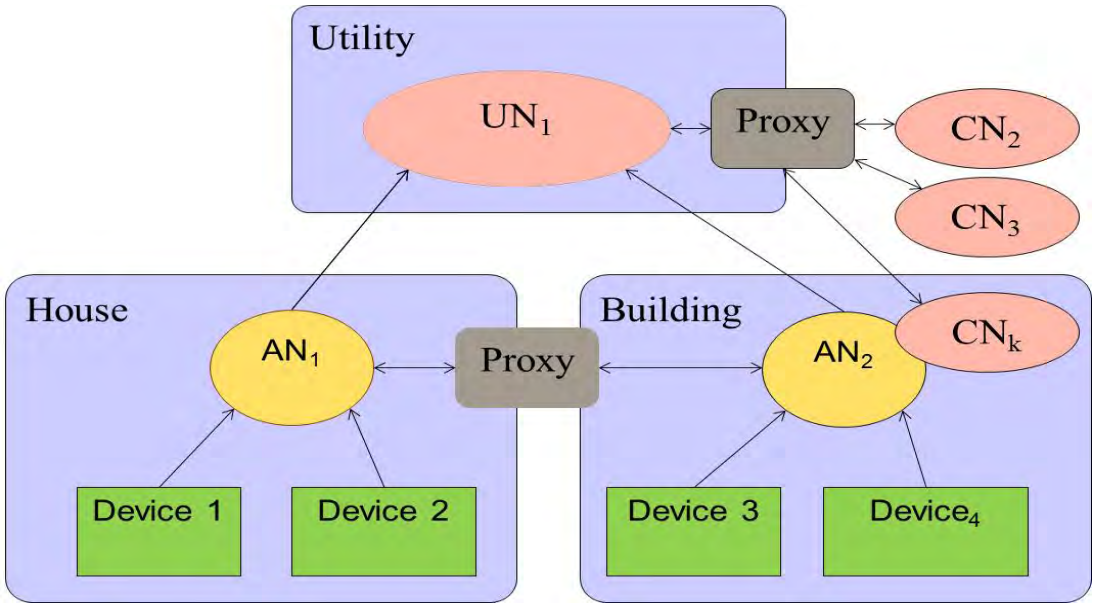

Figure 40: Structure of database simulation

Having a system with each collector node connecting to an isolated database could have potential issues. Fault tolerance would be low (one server going down could disrupt the whole system) and potential deadlocks and race conditions can be hard to recover from. Proxy server can route database requests to appropriate servers and enable a high efficiency system with very little scope for down time.

Consider a situation in which a specific server is suffering from overload or failure. This issue can be recognized, and the system will route all requests to that server to another server that is sitting idle. The data stored in the alternate server is pulled and restored to the appropriate server when functionality is back to normal. A monitoring agent that can detect potential issues before any damage/loss of data has occurred is required.

The system is designed such that the clients are agnostic to the internals of each database server. A communication protocol that allows data exchange to abide by a certain format that is fixed regardless of the storage mechanism is used. Each client makes read and write requests in a pre-specified format that the server can understand and respond to. Fluidity in the database design and makes it easy to implement improvements.

A system with these considerations was built that has the ability to simulate numerous clients (data usage units in the grid). Each window opened in the browser in assigned a new device identifier. This number uniquely identifies each device that sends reports to the server. There is also a location identifier that is assigned to a device based on the geographical location of the consuming unit. This helps the server aggregate data by region.

For instance, each client sends out two parameters: active power and voltage magnitude originated in a given appliance (identified by the device id). The simulation takes a maximum and minimum value for each of these parameters and generates a random number between these two values. There is another text box that directs the simulation on

the frequency at which these usage reports need to be sent. The client takes all of these input fields from the user and proceeds to send out these reports at the stipulated frequency. The number of requests sent till that instant of time are recorded and reported to the client. A "Trigger Computation" button exists on the page that sends a message to the server to perform an aggregation of data. This functionality will normally not exist with the client, but was included here for purposes of demonstration.

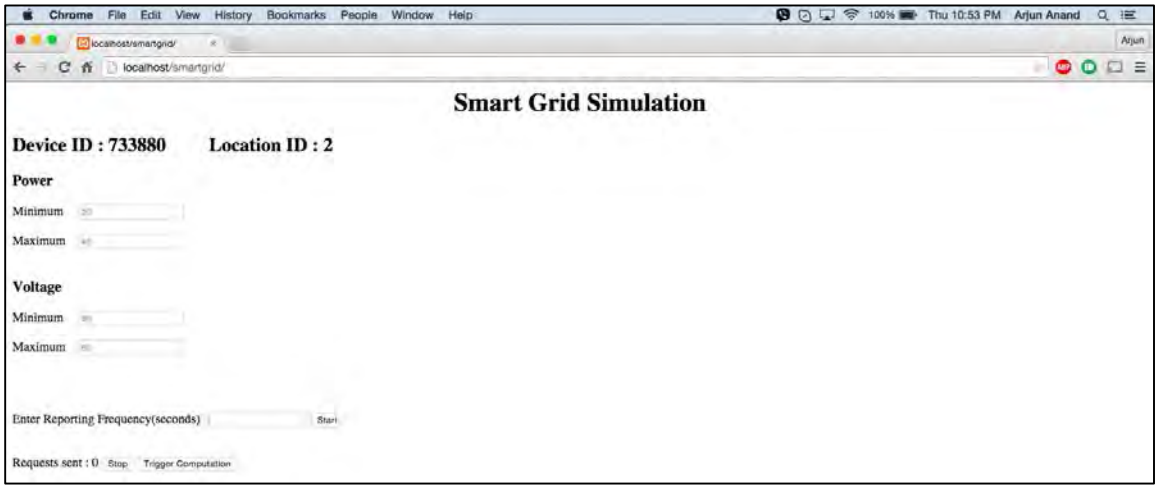

Figure 41: Simulation of user based clients in the Smart Grid

The data collected from the runs of the simulation is depicted in the figures below. The amount of data depends on the frequency of report generation. Varying this parameter really helps understand the fluctuation in the values and therefore helps in deciding what an appropriate frequency to receive reports is. This helps us ensure that there isn't a data redundancy in which we have more reports than are required.

The aggregation of data helps us to understand what kind of information we are able to extract from raw data generated by the clients. The grouping and averaging of values based on location helps us analyze and understand grid behavior. This can be really useful in predicting insecure conditions. This data depicted here can be grouped at one level higher so that the higher level node understands the situation at a macro scale. Therefore, we could perform a "zoom out" operation which would enable us to understand and analyze usage patterns in counties, cities, states, regions, etc.

DEVICE_ID	LOCATION_ID	REPORT_ID	P	V	RECEIVED_TIME
750011	1	12	52	126	2015-04-24 22:37:47
750011	1	13	56	99	2015-04-24 22:37:48
750011	1	14	57	100	2015-04-24 22:37:49
750011	1	15	36	109	2015-04-24 22:37:50
750011	1	16	23	125	2015-04-24 22:37:51
750011	1	17	53	79	2015-04-24 22:37:52
750011	1	18	52	104	2015-04-24 22:37:53
750011	1	19	42	127	2015-04-24 22:37:54
750011	1	20	55	66	2015-04-24 22:37:55
750011	1	21	37	85	2015-04-24 22:37:56
750011	1	22	58	123	2015-04-24 22:37:57
750011	1	23	58	102	2015-04-24 22:37:58
750011	1	24	36	101	2015-04-24 22:37:59
750011	1	25	34	139	2015-04-24 22:38:00
750011	1	26	41	67	2015-04-24 22:38:01
750011	1	27	41	119	2015-04-24 22:38:02
750011	1	28	43	89	2015-04-24 22:38:03
750011	1	29	40	66	2015-04-24 22:38:04
750011	1	30	38	137	2015-04-24 22:38:05
750011	1	31	29	79	2015-04-24 22:38:06
750011	1	32	38	100	2015-04-24 22:38:07
750011	1	33	27	129	2015-04-24 22:38:08
750011	1	34	44	71	2015-04-24 22:38:09
750011	1	35	34	73	2015-04-24 22:38:10
750011	1	36	57	114	2015-04-24 22:38:11
750011	1	37	24	135	2015-04-24 22:38:12
750011	1	38	34	127	2015-04-24 22:38:13
750011	1	39	35	65	2015-04-24 22:38:14
750011	1	40	44	104	2015-04-24 22:38:15
750011	1	41	30	119	2015-04-24 22:38:16
750011	1	42	25	136	2015-04-24 22:38:17
750011	1	43	31	65	2015-04-24 22:38:18
750011	1	44	26	87	2015-04-24 22:38:19
750011	1	45	27	72	2015-04-24 22:38:20
750011	1	46	36	135	2015-04-24 22:38:21
750011	1	47	54	87	2015-04-24 22:38:22
750011	1	48	33	113	2015-04-24 22:38:23
750011	1	49	30	75	2015-04-24 22:38:24
750011	1	50	34	105	2015-04-24 22:38:25
750011	1	51	21	127	2015-04-24 22:38:26

578 rows in table

Figure 42: Snapshot of reports generated by Prosumer Appliances

REPORT_ID	LOCATION_ID	P	V	TRIGGERING_LOCATION	TRIGGERING_DEVICE	COMPUTED_TIME
1	0	170.942	327.607	1	750011	2015-04-24 22:40:50
2	1	39.36	101.632	1	750011	2015-04-24 22:40:50
4	0	170.427	329.056	1	750011	2015-04-24 22:42:05
5	1	39.065	101.955	1	750011	2015-04-24 22:42:05
7	0	169.735	328.912	0	713766	2015-04-24 22:42:25
8	1	38.9398	101.463	0	713766	2015-04-24 22:42:25

Figure 43: Snapshot of aggregated data grouped by prosumers

The simulation developed was implemented using off-the-shelf hardware, MySQL database software and web services. The simulation illustrates that the proposed architecture would support the granularity, data update frequency, and size of the data give our assumptions. Fifty prosumers where simulated each with fifty appliances reporting at second granularity. The aggregation operations are done automatically using standard DBMS functions for presentation to the upper level. The experiments show that the proposed design is able to support this level of data intensity, velocity, and size, in a distributed and concurrent DMBS environment. Further simulations are required in order to identify breaking points and to determine the specific effects of various options regarding distributed database management.

6.2 Cloud-Based Performance of Smart Grid Data

Smart meter data represents a great opportunity for utility companies to monitor and respond to energy utilization spikes and failures almost instantaneously. However, deployment of these smart meters has resulted in a massive increase of the amount of data to be stored and processed. In this section, we take a deep look into how to effectively store the high frequency data transmitted by these smart meters. We focus on open-source distributed databases that can be easily accessed and deployed. We found that Apache HBase is a suitable and promising candidate.

6.2.1 Performance Evaluation Overview

The goal of the analytics infrastructure is to build a distributed system that can store, process, and analyze large amount of streaming data in real-time (e.g., smart meter data), as shown in Figure 44. To achieve this, the first step is to build a distributed database system with the following properties:

- Fast speed (both read and write)
- Scalability
- High availability
- Fault-tolerance
- Easy integration with other data processing frameworks

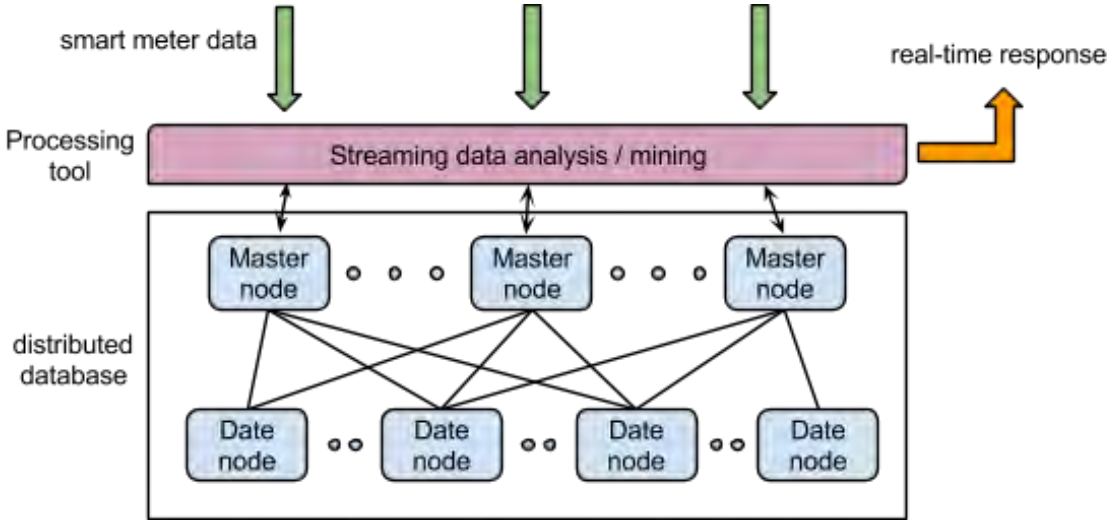

Figure 44: Overall analytics infrastructure

In order to store the data sent from the smart meters, several software were considered. Traditional relational database management system (RDBMS) like MySQL or SQLite because of the ACID properties they offer require data to be in a normalized form for optimal usage. This however comes at a cost of slower write time and also might cause insertions into multiple tables for each data point sent by the smart meter. This would hamper the speed at which the system is able to write especially when data from the smart

meters is sent several times per second. Also in order to effectively utilize the relational databases especially as a lot of data is to be written, sharding of the database is necessary which is complex.

Distributed storage systems like the Hadoop Distributed File System (HDFS) simplify this to a large extent. HDFS is a distributed, scalable, and portable file-system in the Apache Hadoop framework. There are several "NoSQL" distributed databases that built on top of HDFS. NoSQL is a general term meaning that the database isn't an RDBMS which supports SQL as its primary access language.

We have tested several open-source NoSQL distributed database systems to assess if they can meet the aforementioned properties. Among the many open-source distributed database systems available, HBase and Cassandra are the two that we have experimented that show promising results. We describe in detail the advantages and disadvantages in both systems in Section 6.2.3.

6.2.2 Dataset

The data used for our experiments is obtained from the smart meters installed in 172 different buildings on the Georgia Tech campus. Each line of records consists of the following:

- Building ID
- Timestamp
- P - The power reading
- Q - The reactive power reading
- V - The voltage reading

To evaluate the performance of the distributed databases for large-scale input, we replicate the data several times, each with a different time stamp.

6.2.3 Cloud-based Databases: HBase vs. Cassandra

HBase and Cassandra are the two most popular open-source distributed database systems. Both of them support high writing speed, but Cassandra has a slight advantage in this respect. By using a pseudo-distributed mode on a single machine, Hbase achieves a rate of 25000 writes per second (i.e., ingesting 25000 smart meter readings per second), and Cassandra with a slightly higher rate of 30000 writes per second.

However, HBase has several more important advantages. For example, HBase can be more flexibly configured, such as allowing the user to control when the data will be flushed from main memory to the hard disk. This greater flexibility may allow for easier construction of a sophisticated analytics pipeline in the later stage of this project. In terms of the integration with existing data processing frameworks, both of them can work with Hadoop, but HBase supports native HDFS file system while Cassandra uses its own CFS file system

(compatible with HDFS). Furthermore, HBase runs on top of the Apache Hadoop YARN ecosystem, so it is much easier to incorporate with other YARN applications, such as Apache Spark and Apache Strom. HBase is also by default supported by many cloud-computing platforms, such as Amazon AWS. For these reasons, our current experiments suggest that HBase may be a favorable choice for ingesting data in real time.

6.2.4 Evaluation Setup: Software & Hardware

We evaluate the performance on both a single machine (pseudo-distributed mode) and on a real distributed system (fully-distributed mode). For details about the setup and results on a single machine, please see the Appendix A.1.

The real distributed experiments were conducted on Amazon Elastic MapReduce (EMR) clusters of 16 and 32 machines. The specs of the machines used are listed below:

- Master node: c3.2xlarge
  - CPU: Intel Xeon E5-2680 v2 (Ivy Bridge), 8-cores
  - Memory: 15 GB
  - Storage: 2 * 80 GB SSD
- Slave nodes: m3.xlarge
  - CPU: Intel Xeon E5-2670 v2 (Ivy Bridge), 4-cores
  - Memory: 15 GB
  - Storage: 2 * 40 GB SSD

For the software, Amazon EMR provides a standard HBase version as well as a MapR M7 version. We choose the MapR M7 version since it is specifically optimized for high-performance storage and avoids a lot of overhead in the standard HBase.

6.2.5 Evaluation Results

Figure 45 shows the time needed to write various sizes of the data. All the results are averaged over 10 times. As one can see in Figure 45, with 16 machines, HBase can write up to 1 million records in less than 3 seconds. That is over 300000 writes per second. For 32 machines, the rate is slightly lower for smaller sized input, since the overhead outweighs the benefit of using more machines. However, when the input size exceeds a half million lines, it starts to outperform the cluster of 16 machines. We expect the difference to become larger for even large input. The performance is pretty stable and predictable even on the shared environment of Amazon EMR.

There are some configuration tricks used to achieve this result.

- Turn off auto flush
  - Otherwise HBase will flush every write operation immediately.

- Add prefix to the row keys
    - We use building ID + timestamp as row keys. Since it is well known to avoid incremental row keys in HBase, we add some prefixes (e.g. ‘A’, ‘B’, …, etc) to each row key. The number of different prefixes corresponds to the number of regions used to spread the data across.
- Pre-split regions
    - HBase by default use only 1 region initially, and splits the region once the size of the data exceeds some threshold. Therefore, in the beginning all the write operations are directed to a single machine. To solve this problem, we pre-split the regions with split points correspond to the prefixes added. This enforces records with different prefixes to be written to different regions, and hence nicely distributes the workload.

There are some other important configurations that may affect the performance, but are more application-dependent and hence we haven’t tried.

- HDFS replication factor.
    - HBase uses HDFS to store the data. HDFS by default uses a replication factor of 3, meaning all the data are stored in 3 different machines. One can reduce this factor for less important data.
- Write Ahead Log (WAL)
    - HBase uses WAL to ensure fault-tolerance. If one machine fails, HBase can use WAL to reconstruct the data before the failure. One can turn off this function to increase to writing speed, but may lose the un-flushed data if a machine fails.
- Data encryption
    - HBase provides the data encryption functionality. From the official HBase reference guide, one can expect a ~10% performance penalty for encrypted communication.

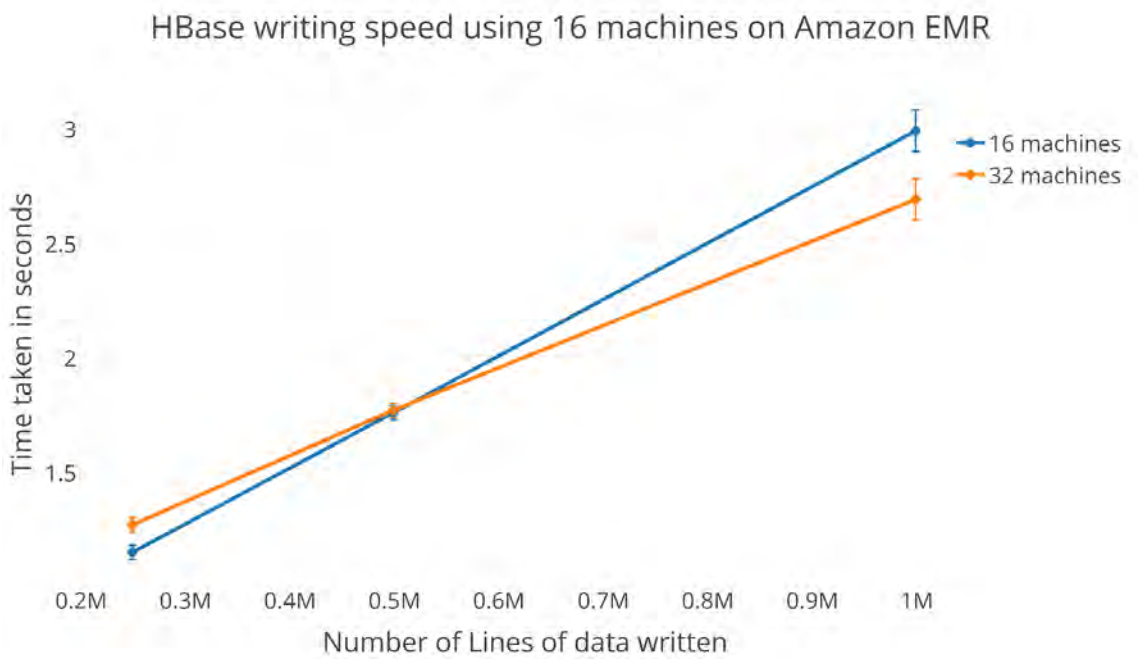

Figure 45: Performance of HBase on a distributed system

6.2.6 Summary

In this study, we show that by using open-source distributed databases such as HBase, we can get reasonably good performance in real-time data storage. The capability and speed of the database increases almost linearly as we increase the number of machines.

Using HBase also comes with several other benefits, such as fault tolerance and the integration with other data processing tools. Future work includes developing interactive data analysis tools that build on top of HBase and a real-time decision making systems based on streaming data analysis.

7. Conclusions

The report describes how large data sets can be used to improve power system operation and control. The spatio-temporal correlation of diverse data sets has been described together with data mining techniques for data analysis. The investigation showed that spatio-temporal correlation together with capabilities of Big Data technologies plays the most essential role for increasing value of collectable data. Following applications were used to demonstrate the usefulness of integration of Big Data in power system applications:

- Improved traveling wave fault location technique that combines conventional fault location techniques with data obtained by lightning detection network: It has been demonstrated that utilization of extended data set can improve accuracy of transmission line fault location.
- Risk assessment for evaluation of transmission insulation coordination and prediction of future outages due to insulation breakdown: With analysis of large historical data set and fast processing of incoming real-time data the early alarm system can be developed for evaluation of current state of lightning protection equipment as well as prediction of future lightning related failures in the system.
- Improved wind forecast for a robust look-ahead economic dispatch and a day-ahead reliability unit commitment: Spatio-temporal wind forecast models (TDD and TDDGW models) are used and critically evaluated. By leveraging both temporal and spatial wind historical data, more accurate wind forecasts can be obtained. The potential economic benefits of advanced wind forecast are illustrated using a modified IEEE RTS 24 bus system.
- Utilization of Big Data for distribution automated smart meter system: By using Apache HBase, an open-source distributed database, we can achieve real-time data storage with large-scale streaming data. In addition fault tolerance and the integration with other data processing tools are easily achievable.

8. References

[1] P.-C. Chen, T. Dokic, and M. Kezunovic, "The Use of Big Data for Outage Management in Distribution Systems," Int. Conf. on Electricity Distrib. (CIRED) Workshop, 2014, in press.

[2] J. Syllignakis, C. Adamakis, and T. M. Papazoglou, "A GIS Web - Application for Power System of Crete," 42nd Int. Universities Power Engineering Conf., pp. 414-418, Sep. 2007.

[3] X. Q. Li, Z. Y. Zeng, Y. C. Zhang, and X. J. Xu, "A Study of Distribution Load Transfer Operation Based on GIS," Int. Conference on Machine Learning and Cybernetics, pp. 1428-1433, Aug. 2007.

[4] A. N. Sekhar, K. S. Rajan, and A. Jain, "Spatial Informatics and Geographical Information Systems: Tools to Transform Electric Power and Energy Systems," 2008 IEEE Region 10 Conf. (TENCON), pp. 1-5, Nov. 2008.

[5] K. L. Cummins, E. P. Krider, and M. D. Malone, "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.

[6] T. Niimura, M. Dhaliwal, and K. Ozawa, "Fuzzy regression models to represent electricity market data in deregulated power industry," Joint 9th IFSA World Congr. and 20th NAFIPS Int. Conf., pp. 2556-2561, Jul. 2001.

[7] M. Kezunovic, L. Xie, and S. Grijalva, "The role of big data in improving power system operation and protection," 2013 IREP Symp. Bulk Power Syst. Dynamics and Control - IX Optimization, Security and Control of the Emerging Power Grid (IREP), pp. 1-9, Aug. 2013.

[8] IBM Software, "Managing big data for smart grids and smart meters," May 2012.

[9] M. Kezunovic, et al., "The Fundamental Concept of Unified Generalized Model and Data Representation for New Applications in the Future Grid", 45th HICCS Hawaii Int. Conf. on Syst. Sciences, Jan. 2012.

[10] IEC Standard 61850-6, "Communication networks and systems in Substations – Configu-ration description language for communication in electrical substations related to IEDs".

[11] IEEE Standards Office, IEEE Standard C37.111-1999, "IEEE Standard Common Format for Transient Data Exchange (COMTRADE) for Power Systems".

[12] IEC Standard 61970, "IEC 61970 Energy management system application program interface (EMS-API) - Part 301: Common Information Model (CIM) Base".

[13] IEC Standard 61968, "IEC 61968 Application integration at electric utilities - System in-terfaces for distribution management - Part 11: Common Information Model (CIM)".

[14] IEEE Standards Office, IEEE Std C37.239-2010, "IEEE Standard for Common Format for Event Data Exchange (COMFEDE) for Power Systems".

[15] IEEE Standards Office, IEEE Std C37.118.1-2011, "IEEE Standard for Synchrophasor Measurements for Power Systems".

[16] IEEE Standards Office, IEEE Std C37.118.2, "IEEE Standard for Synchrophasor Data Transfer for Power Systems".

[17] Esri, What is GIS?, 11/10/2013. [Online]. Available: http://www.esri.com/what-is-gis/overview#overview_panel
[18] United Nations, Department of Economic and Social Affairs, Geographic Information System for Power System Planning, United Nations, New York, 1997.
[19] David J. Buckey, Bgis Introduction to GIS, 11/10/2013. [Online]. Available: http://bgis.sanbi.org/gis-primer/
[20] D. Cole, et al., "Fault Location & System Restoration," Jul. 2013. [Online]. Available: http://www.qualitrolcorp.com/uploadedFiles/Siteroot/Products/Fault_Location_and_System_Restoration.pdf
[21] ASOS User Guide, NOAA, Mar. 1998.
[22] Federal Standard for Siting Meteorological Sensors at Airports, FCM-S4-1994, NOAA, Aug. 1994.
[23] Federal Meteorological Handbook No. 1 - Surface Weather Observations and Reports, FCM-H1-2005, NOAA, Sep. 2005.
[24] Standard Formats for Weather Data Exchange Among Automated Weather Information Systems, FCM-S2-1994, NOAA, Nov. 1994.
[25] E. Kalnay, Atmospheric modeling, data assimilation and predictability, Cambridge University Press, 2003.
[26] L. McColl, et al., "A Climate Change Risk Assessment for the UK Electricity Networks," presented at Int. Conf. Energy & Meteorology (ICEM), 2011.
[27] L. McColl, et al., "Assessing the potential impact of climate change on the UK's electricity network: electronic supplementary material 3." [Online] Available: http://link.springer.com/content/esm/art:10.1007/s10584-012-0469-6/file/MediaObjects/10584_2012_469_MOESM3_ESM.pdf
[28] Power Guide 2009/Book 07, "Protection against lightning effects – Legrand," Legrand 2009.
[29] 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
[30] 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.
[31] Vaisala Inc., "Thunderstorm and Lightning Detection Systems," [Online] Available: http://www.vaisala.com/en/products/thunderstormandlightningdetection systems/Pages/default.aspx
[32] G Shroff, "The Intelligent Web: Search, Smart Algorithms and Big Data," Oxford University Press, UK, 2013.
[33] E. O. Schweitzer, III, et al., "Locating Faults by the Traveling Waves They Launch", Texas A&M Conf. for Protective Relay Engineers, 2014, in press.
[34] P. F. Gale, et al., "Fault location based on travelling waves." 5th Int Conf. on Developments in Power Syst. Protection (DPSP), pp. 54–59, Apr. 1993.

[35] Y. J. Xia, et al., “A novel fault location scheme using voltage traveling-wave of CVTs,” 39th Int. Uni. Power Eng. Conf. (UPEC) (Vol. 2), vol. 1, pp. 768-772, 2004.
[36] A. Elhaffar and M. Lehtonen, “Multi-end Traveling Wave Fault Location Based on Current Traveling Waves,” 16th Power Syst. Computation Conf. (PSCC), Jul. 2008.
[37] F. H. Magnago and A. Abur, “Fault Location Using Wavelets,” IEEE Trans. Power Del., vol. 13, no. 4, Oct. 1998.
[38] A. Tabatabaei, et al., “Fault location techniques in power system based on traveling wave using wavelet analysis and GPS timing,” Electrical Review, vol. 88, no. 6, pp. 347-350, 2012.
[39] H. Lee and A. M. Mousa, “GPS travelling wave fault locator systems: investigation into the anomalous measurements related to lightning strikes,” IEEE Trans. Power Del., vol. 11, no. 3, pp. 1214-1223, 1996.
[40] W. Zhao, et al., “Improved GPS travelling wave fault locator for power cables by using wavelet analysis,” Int. J. of Electrical Power & Energy Syst., vol. 23, no. 5, pp. 403–411, Jun. 2001.
[41] T. Sadovic, et al., “Expert System for Transmission Line Lightning Performance Determination”, CIGRE Int. Colloq. on Power Quality and Lightning, Jun. 2012.
[42] D. Cole, et al., “Fault Location & System Restoration,” Jul. 2013. [Online]. Available: http://www.qualitrolcorp.com/uploadedFiles/Siteroot/Products/Fault_Location_and_System_Restoration.pdf
[43] M. M. Saha, et al., “Fault location – Basic concept and characteristic of methods,” Fault Location on Power Networks, Springer London, pp. 1-26, 2010.
[44] M. Kezunovic, et al., “Digital models of coupling capacitor voltage transformers for protective relay transient studies,” IEEE Trans. Power Del., vol. 7, no. 4, pp. 1927-1935, Oct. 1992.
[45] Y. Tang, et al., "Study on Effect of Current Transformer and Its Secondary Cable to Travelling Wave Propagation Characteristic of Electric Power Lines," 2nd Int. Conf. on Intell. Syst. Des. and Eng. Appl., pp.1495-1498, Jan. 2012.
[46] V. S. Kale, et al., “Faulted phase selection based on wavelet analysis of traveling waves’, Int. J. Computer and Elect. Eng., vol. 3, no. 3, pp. 421–425, Jun. 2011.
[47] T. Jinrui, et al., “Modeling Technology in Traveling-Wave Fault Location,“ TELEKOMNIKA, vol. 11, no. 6, pp. 3333-3340, Jun. 2013.
[48] J. R. Marti “Accurate modeling of frequency-dependent transmission lines in electromagnetic transient simulations,” IEEE Trans. Power Apparatus and Syst., vol. PAS-101, no. 1, pp. 147-157, Jan. 1982.
[49] Alternative Transients Program, ATP-EMTP, 2010. [Online]. Available: http://www.emtp.org
[50] E. Clarke, “Circuit Analysis of AC Power Systems, Symmetrical and Related Components,” Wiley, New York, 1943.
[51] Mathworks Inc. MATLAB R2012b User’s Guide. [Online]. Available: http://www.mathworks.com
[52] Program and Organizational Performance Division et al., “Best Practices in Geographic Information Systems-Based Transportation Asset Management,” Jan. 2012.

[53] I. M. Rawi, et al., "Lightning study and experience on the first 500kV transmission line arrester in Malaysia," 2014 International Conference on Lightning Protection (ICLP), Shanghai, China, 2014.
[54] W. Sones, S. M. Wong, “Overview on Transient Overvoltages and Insulation Design For a High Voltage Transmission System,” High Voltage Engineering and Application (ICHVE), 2010 International Conference on, New Orleans, LA, 2010.
[55] Z. G. Datsios, et al., "Estimation of the minimum shielding failure current causing flashover in overhead lines of the hellenic transmission system through ATP-EMTP simulations." 2012 International Conference on Lightning Protection (ICLP), Vienna, Austria, 2012.
[56] S. T. Mobarakei, T. Sami, B. Porkar, "Back flashover phenomenon analysis in power transmission substation for insulation coordination," 2012 11th International Conference on Environment and Electrical Engineering (EEEIC), Venice, May 2012.
[57] S. Bedoui, et al., "Analysis of lightning protection with transmission line arrester using ATP/EMTP: Case of an HV 220kV double circuit line." Universities Power Engineering Conference (UPEC), 2010 45th International. IEEE, 2010.
[58] Zhang, J., et al. "Application of hourly meteorological records to atmospheric correction factors in insulation coordination under switching impulse voltage." High Voltage Engineering and Application, 2008. ICHVE 2008. International Conference on. IEEE, 2008.
[59] R. Shariatinasab, et al., "Probabilistic evaluation of optimal location of surge arresters on EHV and UHV networks due to switching and lightning surges." Power Delivery, IEEE Transactions on, vol. 24, no. 4, pp. 1903-1911, 2009.
[60] National Oceanic and Atmospheric Administration’s national Data Buoy Center, “Station MGPT2 – Historical Data,” [Online] Available: http://www.ndbc. noaa.gov/station_history.php?station=mgpt2
[61] 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
[62] M. Ishii, et al. "Multistory transmission tower model for lightning surge analysis." Power Delivery, IEEE Transactions on vol. 6, no. 3, pp. 1327-1335, 1991.
[63] I. Uglesic, “Modeling of Transmission Line and Substation for Insulation Coordination Studies,” Simulation & Analysis of Power System Transients with EMTP-RV, Dubrovnik, April 2009.
[64] Z. Medina-Cetina and F. Nadim F, "Stochastic Design of an Early Warning System", Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, vol. 2, no. 4, pp. 223 – 236, 2008.
[65] J. Stojanovic, et al,. "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, (in press).
[66] K. Ristovski, et al., "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.
[67] P. Dehghanian, M. Fotuhi-Firuzabad, F. Aminifar, and R. Billinton, "A Comprehensive Scheme for Reliability Centered Maintenance Implementation in Power Distribution Systems- Part I: Methodology", IEEE Transactions on Power Delivery, vol.28, no.2, pp.761-770, April 2013.
[68] W. Li, Risk assessment of power systems: models, methods, and applications, John Wiley, New York, 2005.
[69] R. Billinton and R. N. Allan, Reliability Evaluation of Engineering Systems: Concepts and Techniques, 2nd ed. New York: Plenum, 1992.
[70] L. Xie, P. M. S. Carvalho, L. A. F. M. Ferreira, J. Liu, B. H. Krogh, N. Popli, and M. D. Ilic, "Wind integration in power systems: Operational challenges and possible solutions," Proceedings of the IEEE, vol. 99, no. 1, pp. 214–232, 2011.
[71] Y. Gu and L. Xie, "Fast sensitivity analysis approach to assessing congestion induced wind curtailment," IEEE Transactions on Power Systems, vol.29, pp. 101-110, 2014.
[72] Y. Chen, L. Xie, and P. R. Kumar, "Dimensionality reduction and early event detection using online synchrophasor data," in IEEE Power and Energy Society General Meeting, Vancouver, 2013.
[73] L. Xie, Y. Chen, and H. Liao, "Distributed online monitoring of quasi-static voltage collapse in multi-area power systems," IEEE Transactions on Power Systems, vol. 27, no. 4, pp. 2271–2279, 2012.
[74] Y. Zhang, Y. Chen, and L. Xie, "Multi-scale integration and aggregation of power system modules for dynamic security assessment," in IEEE Power and Energy Society General Meeting, Vancouver, 2013.
[75] M. Alexiadis, P. Dokopoulos, and H. Sahsamanoglou, "Wind speed and power forecasting based on spatial correlation models," IEEE Transactions on Energy Conversion, vol. 14, no. 3, pp. 836–842, 1999.
[76] S. Zhao, L. Xie, and C. Singh, "Cross-correlation study of onshore/offshore wind generation and load in texas," in North American Power Symposium, Manhattan, KS, 2013.
[77] "20% wind energy by 2030 increasing wind energy's contribution to U.S. electricity supply," U.S. Department of Energy, Tech. Rep. DOE/GO-102008-2567, July 2008.
[78] P. Dehghanian, M. Fotuhi-Firuzabad, S. Bagheri-Shouraki, and A. A. Razi Kazemi, "Critical component identification in reliability centered asset management of power distribution systems via fuzzy ahp," IEEE Systems Journal, vol. 6, no. 4, pp. 593–602, 2012.
[79] P. Dehghanian, M. Fotuhi-Firuzabad, F. Aminifar, and R. Billinton, "A comprehensive scheme for reliability centered maintenance in power distribution systems," IEEE Transactions on Power Delivery, vol. 28, no. 2, pp. 761–770, 2013.
[80] B. Zhang and M. Kezunovic, "Impact of available electric vehicle battery power capacity on power system reliability," in IEEE Power and Energy Society General Meeting, Vancouver, Canada, 2013.

[81] G. Gross and F. D. Galiana, “Short-term load forecasting,” Proceedings of the IEEE, vol. 75, pp. 1558 – 1573, 1987.
[82] J. Black, “Plans for wind integration in ISO-NE: Progress and challenges,” in 8th annual Carnegie Mellon Conference on the electricity industry, Pittsburgh, 2012.
[83] (2012) Deep Thunder-Precision Forecasting for Weather-Sensitive Business Operations. [Online]. Available: http://www.research.ibm.com/weather/DT.html =0pt
[84] W. P. Mahoney, K. Parks, G. Wiener, L. Yubao, W. L. Myers, S. Juanzhen, L. Delle Monache, T. Hopson, D. Johnson, and S. E. Haupt, “A wind power forecasting system to optimize grid integration,” IEEE Transactions on Sustainable Energy, vol. 3, no. 4, pp. 670–682, 2012.
[85] E. M. Constantinescu, V. M. Zavala, M. Rocklin, L. Sangmin, and M. Anitescu, “A computational framework for uncertainty quantification and stochastic optimization in unit commitment with wind power generation,” IEEE Transactions on Power Systems, vol. 26, no. 1, pp. 431–441, 2011.
[86] M. Genton and A. Hering, “Blowing in the wind,” Significance, vol. 4, pp. 11–14, 2007.
[87] X. Zhu and M. G. Genton, “Short-term wind speed forecasting for power system operations,” International Statistical Review, vol. 80, pp. 2–23, 2012.
[88] L. Xie, Y. Gu, X. Zhu, and M. G. Genton, “Power system economic dispatch with spatio-temporal wind forecasts,” in Energytech, 2011 IEEE, 2011, pp. 1–6.
[89] X. Zhu, M. G. Genton, Y. Gu, and L. Xie, “Space-time wind speed forecasting for improved power system dispatch (with discussion),” TEST, vol. 23, pp. 45-50, 2014.
[90] J. Wang, M. Shahidehpour, and Z. Li, “Security-constrained unit commitment with volatile wind power generation,” IEEE Transactions on Power Systems, vol. 23, no. 3, pp. 1319–1327, 2008.
[91] A. Papavasiliou, S. Oren, and R. O’Neill, “Reserve requirements for wind power integration: A scenario-based stochastic programming framework,” IEEE Transactions on Power Systems, vol. 26, no. 4, pp. 2197–2206, 2011.
[92] P. Meibom, R. Barth, B. Hasche, H. Brand, C. Weber, and M. O’Malley, “Stochastic optimization model to study the operational impacts of high wind penetrations in ireland,” IEEE Transactions on Power Systems, vol. 26, no. 3, pp. 1367–1379, 2011.
[93] A. Tuohy, P. Meibom, E. Denny, and M. O’Malley, “Unit commitment for systems with significant wind penetration,” IEEE Transactions on Power Systems, vol. 24, no. 2, pp. 592–601, 2009.
[94] D. Bertsimas, E. Litvinov, X. Sun, J. Zhao, and T. Zheng, “Adaptive robust optimization for the security constrained unit commitment problem,” IEEE Transactions on Power Systems, vol. 28, no. 1, pp. 52–63, 2013.
[95] R. Jiang, J. Wang, and Y. Guan, “Robust unit commitment with wind power and pumped storage hydro,” IEEE Transactions on Power Systems, vol. 27, no. 2, pp. 800–810, 2012.
[96] A. A. Thatte, L. Xie, D. E. Viassolo, and S. Singh, “Risk measure based robust bidding strategy for arbitrage using a wind farm and energy storage,” IEEE Transactions on Smart Grid, 2014.

[97] A. A. Thatte and L. Xie, “Towards a unified operational value index of energy storage in smart grid environment,” IEEE Transactions on Smart Grid, vol. 3, no. 3, pp. 1418–1426, 2012.
[98] G. Giebel, R. Brownsword, G. Kariniotakis, M. Denhard, and C. Draxl, “The state-of-the-art in short-term prediction of wind power : A literature overview, 2nd edition,” ANEMOS.plus, p. 109, 2011.
[99] G. Kariniotakis, P. Pinson, N. Siebert, G. Giebel, and R. Barthelmie, “The-state-of-the-art in short-term prediction of wind power - from an offshore perspective,” in Symposium ADEME, IFREMER, 2004.
[100] C. Monteiro, H. Keko, R. Bessa, V. Miranda, A. Botterud, J. Wang, and G. Conzelmann, “A quick guide to wind power forecasting: Stat-of-the-art 2009,” 2009.
[101] T. Gneiting, K. Larson, K. Westrick, M. G. Genton, and E. Aldrich, “Calibrated probabilistic forecasting at the stateline wind energy center,” Journal of the American Statistical Association, vol. 101, no. 475, pp. 968–979, 2006.
[102] A. S. Hering and M. G. Genton, “Powering up with space-time wind forecasting,” Journal of the American Statistical Association, vol. 105, no. 489, pp. 92–104, 2010.
[103] J. Tastu, P. Pinson, E. Kotwa, H. Nielsen, and H. Madsen, “Spatio-temporal analysis and modeling of wind power forecast errors,” International Statistical Review, vol. 14, pp. 43–60, 2011.
[104] P. Pinson and H. Madsen, “Adaptive modeling and forecasting of wind power fluctuations with markov-switching autoregressive models,” Journal of Forecasting, vol. 31, pp. 281–313, 2012.
[105] X. Zhu, K. Bowman, and M. G. Genton, “Incorporating geostrophic wind information for improved space-time wind speed forecasting,” Annals of Applied Statistics, 2013, invited revision.
[106] Y. Sun and M. G. Genton, “Functional boxplots,” Journal of Computational and Graphical Statistics, vol. 20, pp. 316–334, 2011.
[107] T. Gneiting and A. E. Raftery, “Strictly proper scoring rules, prediction, and estimation,” Journal of the American Statistical Association, vol. 102, no. 477, pp. 359–378, 2007. [Online]. Available: http://pubs.amstat.org/doi/abs/10.1198/016214506000001437 =0pt
[108] H. Zhong, Q. Xia, Y. Wang, and C. Kang, “Dynamic economic dispatch considering transmission losses using quadratically constrained quadratic program method,” IEEE Transactions on Power Systems, vol. 28, no. 3, pp. 2232–2241, 2013.
[109] H. Zhong, L. Xie, and Q. Xia, “Coupon incentive-based demand response: Theory and case study,” IEEE Transactions on Power Systems, vol. 28, no. 2, pp. 1266–1276, 2013.
[110] Y. Gu and L. Xie, “Early detection and optimal corrective measures of power system insecurity in enhanced look-ahead dispatch,” IEEE Transactions on Power Systems, vol. 28, no. 2, pp. 1297–1307, 2013.
[111] C. Grigg, P. Wong, P. Albrecht, R. Allan, M. Bhavaraju, R. Billinton, Q. Chen, C. Fong, S. Haddad, S. Kuruganty, W. Li, R. Mukerji, D. Patton, N. Rau, D. Reppen, A. Schneider, M. Shahidehpour, and C. Singh, “The IEEE reliability

test system-1996. a report prepared by the reliability test system task force of the application of probability methods subcommittee," IEEE Transactions on Power Systems, vol. 14, no. 3, pp. 1010–1020, 1999.
[112] M. Giberson. (2009) Power prices in ERCOT's west zone: a mix of wind power, natural gas prices, transmission constraints, and (inefficient) congestion management practices. [Online]. Available: http://knowledgeproblem.com/2009/07/22/ercot-west-power-prices-2009-jan-jun/ =0pt
[113] Electric Reliability Council of Texas. ERCOT 2009 Annual Report. [Online]. Available: http://www.ercot.com/news/presentations/2010/index =0pt
[114] F. C. Schweppe, R. D. Tabors, J. L. Kirtley, H. R. Outhred, F. H. Pickel, and A. J. Cox, "Homeostatic Utility Control," Power Apparatus and Systems, IEEE Transactions on, vol. PAS-99, no. 3, pp. 1151-1163, 1980.
[115] A. H. Rosenfeld, D. A. Bulleit, and R. A. Peddie, "Smart Meters and Spot Pricing: Experiments and Potential," Technology and Society Magazine, IEEE, vol. 5, no. 1, pp. 23-28, 1986.
[116] DoE, "DoE Smart Grid System Report," 2009.
[117] G. A. Council, "GridWise Interoperability Path Forward," 2005.
[118] NIST, "Framework and Roadmap for Smart Grid Interoperability Standards, Release 1.0," 2010.
[119] M. Finney. "PG&E Acknowledges Smart Meter Problems." April, 2010. Available: http://abclocal.go.com/kgo/story?section=news/7_on_your_side&id=7406652
[120] DoE, "Communications Requirements of Smart Grid Technologies," 2010.
[121] T. Hubert and S. Grijalva, "Realizing smart grid benefits requires energy optimization algorithms at residential level," in Innovative Smart Grid Technologies (ISGT), 2011 IEEE PES, 2011, pp. 1-8.
[122] T. Hubert and S. Grijalva, "Home energy manager: A consumer-oriented interactive tool to optimize energy use," in Consumer Electronics (ICCE), 2011 IEEE International Conference on, 2011, pp. 505-506.
[123] G. Lambert-Torres, "Application of rough sets in power system control center data mining," in Power Engineering Society Winter Meeting, 2002. IEEE, 2002, pp. 627-631 vol.1.
[124] P. M. Mahadev and R. D. Christie, "Envisioning power system data: vulnerability and severity representations for static security assessment," Power Systems, IEEE Transactions on, vol. 9, no. 4, pp. 1915-1920, 1994.
[125] T. J. Overbye and J. D. Weber, "Visualization of power system data," in System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on, 2000, p. 7 pp.
[126] A. Venkatesh, G. Cokkinides, and A. P. S. Meliopoulos, "3D-Visualization of Power System Data Using Triangulation and Subdivision Techniques," in System Sciences, 2009. HICSS '09. 42nd Hawaii International Conference on, 2009, pp. 1-8.
[127] F. Capitanescu, J. L. Martinez Ramos, P. Panciatici, D. Kirschen, A. Marano Marcolini, L. Platbrood, and L. Wehenkel, "State-of-the-art, challenges, and

future trends in security constrained optimal power flow," Electric Power Systems Research, vol. 81, no. 8, pp. 1731-1741, 2011.
[128] E. M. Constantinescu, V. M. Zavala, M. Rocklin, L. Sangmin, and M. Anitescu, "A Computational Framework for Uncertainty Quantification and Stochastic Optimization in Unit Commitment With Wind Power Generation," Power Systems, IEEE Transactions on, vol. 26, no. 1, pp. 431-441, 2011.
[129] J. K. Kaldellis, "The wind potential impact on the maximum wind energy penetration in autonomous electrical grids," Renewable Energy, vol. 33, no. 7, pp. 1665-1677, 2008.
[130] M. Panteli, P. A. Crossley, D. S. Kirschen, and D. J. Sobajic, "Assessing the Impact of Insufficient Situation Awareness on Power System Operation," Power Systems, IEEE Transactions on, vol. PP, no. 99, pp. 1-11, 2013.
[131] R. H. Boroumand and G. Zachmann, "Retailers' risk management and vertical arrangements in electricity markets," Energy Policy, vol. 40, no. 0, pp. 465-472, 2012.
[132] A. N. Kleit, A. V. Shcherbakova, and X. Chen, "Restructuring and the retail residential market for power in Pennsylvania," Energy Policy, vol. 46, no. 0, pp. 443-451, 2012.
[133] T. Chee-Wooi, G. Manimaran, and L. Chen-Ching, "Cybersecurity for Critical Infrastructures: Attack and Defense Modeling," Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, vol. 40, no. 4, pp. 853-865, 2010.
[134] G. N. Ericsson, "Cyber Security and Power System Communication - Essential Parts of a Smart Grid Infrastructure," Power Delivery, IEEE Transactions on, vol. 25, no. 3, pp. 1501-1507, 2010.
[135] C. Nunez. "Who's Watching? Privacy Concerns Persist as Smart Meters Roll Out." 2012. Available: http://news.nationalgeographic.com/news/energy/2012/12/121212-smart-meter-privacy/
[136] N. Capodieci, G. Cabri, G. A. Pagani, and M. Aiello, "An Agent-Based Application to Enable Deregulated Energy Markets," in Computer Software and Applications Conference (COMPSAC), 2012 IEEE 36th Annual, 2012, pp. 638-647.
[137] C. Tiansong, H. Goudarzi, S. Hatami, S. Nazarian, and M. Pedram, "Concurrent optimization of consumer's electrical energy bill and producer's power generation cost under a dynamic pricing model," in Innovative Smart Grid Technologies (ISGT), 2012 IEEE PES, 2012, pp. 1-6.
[138] H. Pu, J. Kalagnanam, R. Natarajan, M. Sharma, R. Ambrosio, D. Hammerstrom, and R. Melton, "Analytics and Transactive Control Design for the Pacific Northwest Smart Grid Demonstration Project," in Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on, 2010, pp. 449-454.
[139] S. Aman, Y. Simmhan, and V. K. Prasanna, "Energy management systems: state of the art and emerging trends," Communications Magazine, IEEE, vol. 51, no. 1, pp. 114-119, 2013.

[140] S. Grijalva, M. Costley, and N. Ainsworth, "Prosumer-Based Control Architecture for the Future Electricity Grid," in IEEE Multi-conference on Systems and Control, 2011.
[141] S. Grijalva and M. U. Tariq, "Prosumer-based smart grid architecture enables a flat, sustainable electricity industry," in Innovative Smart Grid Technologies (ISGT), 2011 IEEE PES, pp. 1-6

A.1. Appendix: Evaluation using Single-Machine Simulated Cloud

In this section, we describe the experiments of HBase and Cassandra on a single machine.

System Specification

All the experiments described in this section have been run on a MacBook Pro having the following configuration:

- MacBook Pro Retina, 13-inch, Late 2012
- Running OS X Version 10.9.2
- 2.9 GHz Intel Core i7 Processor
- 8 Gb 1600 Mhz DDR3 RAM
- 250 Gb SSD

Hadoop

- Hadoop version 2.2.0 (Yarn) was installed in Pseudo-Distributed Mode.

HBase

- HBase version 0.96.1.1 was installed on top of the HDFS installation and configured to use the existing HDFS for its operations.

Cassandra

- Cassandra version 2.0.6 and configured to run on a single system mode.

HBase Performance

The write mechanism in HBase follows what is known as the write path. As soon as the client makes a put call for writing to HBase, the request is routed to the corresponding region server (region servers are logical segmentations of the HBase table with each region server responsible for its own part of the table). The region server writes the data to a temporary volatile memory called the memstore; this is very fast as writing to the memstore is cheap. Once enough data has been accumulated in the memstore, the data is then written to HFiles in the HDFS file system. In order to optimize the writes it is advisable to batch writes on the client side and once enough writes are reached flush them into the HBase in one go. If the auto flush is not manually turned off, HBase tries to flush each and every line it receives through the put operation thereby decreasing the write times. In our experiments flush was called once a pre decided number of lines was cached on the client side. This threshold is generally equal to the number of lines loaded from the database. This allows us to measure the time taken to write these N lines to HBase.

The time taken varies with iterations to find the closest measure of time taken. The N lines were written to the database a 100 times and the provided estimate of time taken was an average of these 100 attempts to write data to HDFS. As shown in there is a significant improvement in the write performance on turning off auto flush (and thereby ensuring that

the put commands are batched on the client side) and with auto flush on (which caused each put operation to be flushed to disk before the next put operation was processed).

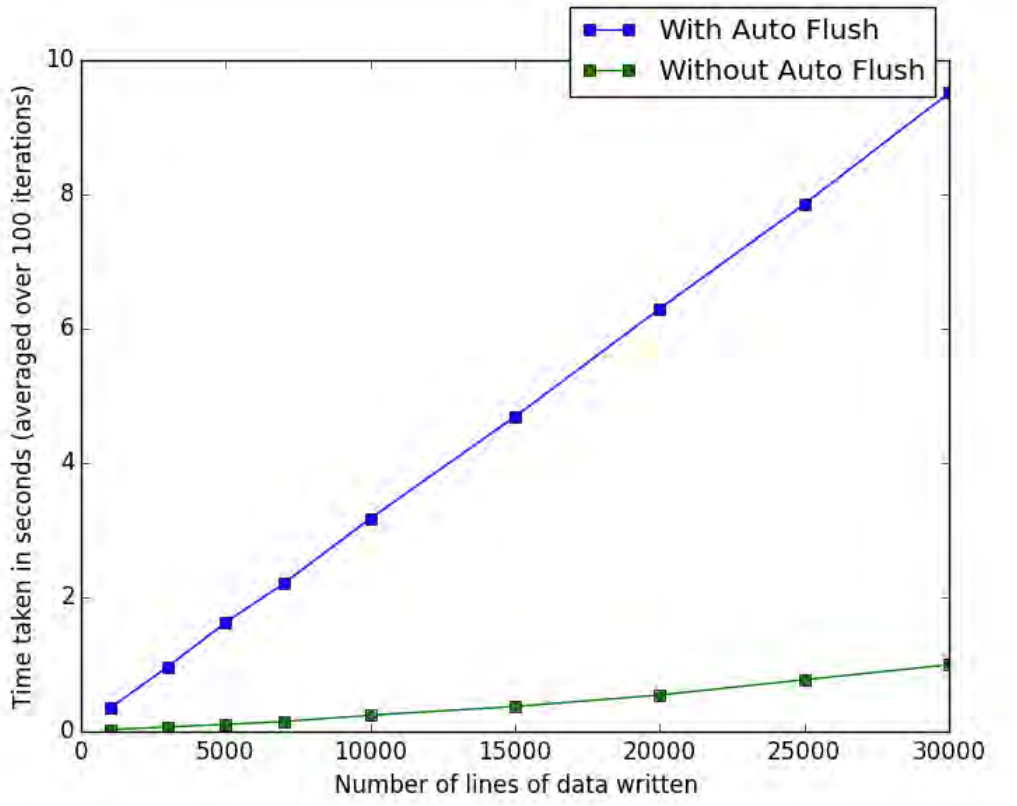

Figure 46: HBase: with and without auto-flushing

Table 10 shows the exact write times for HBase in the optimal scenario on a pseudo-distributed cluster. As we can see writing 30000 lines takes just below one second with auto flush turned off. This indicates that HBase is able to comfortably achieve our target write times even with just a single computer in pseudo-distributed mode.

Table 10: HBase write time with auto flush turned off

Lines of Data written (per iteration)	**Time taken (averaged over 100 iterations in ms)**
10000	242.82
15000	372.01
20000	543.17
25000	770.96
30000	994.01

Cassandra Performance

The way writes are handled in Cassandra is not too dissimilar to the way HBase handles them. Like in HBase data sent for writes in Cassandra is first written to an in memory table structure called the memtable and also a commit log which is very similar to the WAL in HBase. Writes are considered successful if they are written into both the memtable and the commit log. Data is batched in the memtable and is periodically written to disk into a persistent table structure called the SSTable. Both the memtable and the SSTable are maintained in the column family format, which means that any particular row of data would be stored across multiple SSTables. In order to determine if a particular SSTable has data from a given row Cassandra uses another data structure called a Bloom Filter. Each SSTable is associated with a bloom filter, which stores metadata about the data in the SSTable. In the background Cassandra regularly merges smaller SSTables into larger SSTables in a process called compaction.

HBase as unlike HBase, Cassandra does not expose any method to the client which causes the memtables to be flushed and data to be written to the disk. Therefore all writes happened only to the memtable. So in order to get accurate readings N lines of data was selected and written to the database for a period of one hour. At the end of this period using the values of the total time taken and the number of lines written we can calculate the per second throughput. Also, to ensure that as many lines as possible were written to the SSTable in this period the maximum buffer space available to the Cassandra memtables was restricted to 1Mb. These settings were configured in the cassandra.yaml file present in the config folder under the root Cassandra installation by modifying the memtable_total_space_in_mb property. The output obtained is outlined in . Cassandra seems to perform better than HBase when it comes to writing as we can see that it even manages to write 30000 lines within a second.

Table 11: Cassandra write times

Lines of Data written (per iteration)	**Time taken (ms) (estimated by writing for one hour)**
20000	543.17
25000	770.96
30000	994.01

One can see from Figure 47 that both Cassandra and HBase are comparable when it comes to write speeds, with Cassandra having a slight edge for higher values of number of lines (N). This is as expected as Cassandra is optimized for writing data and is known to perform better than HBase when it comes to writing data. However, HBase has several other important advantages as described in Section 6.2.3.

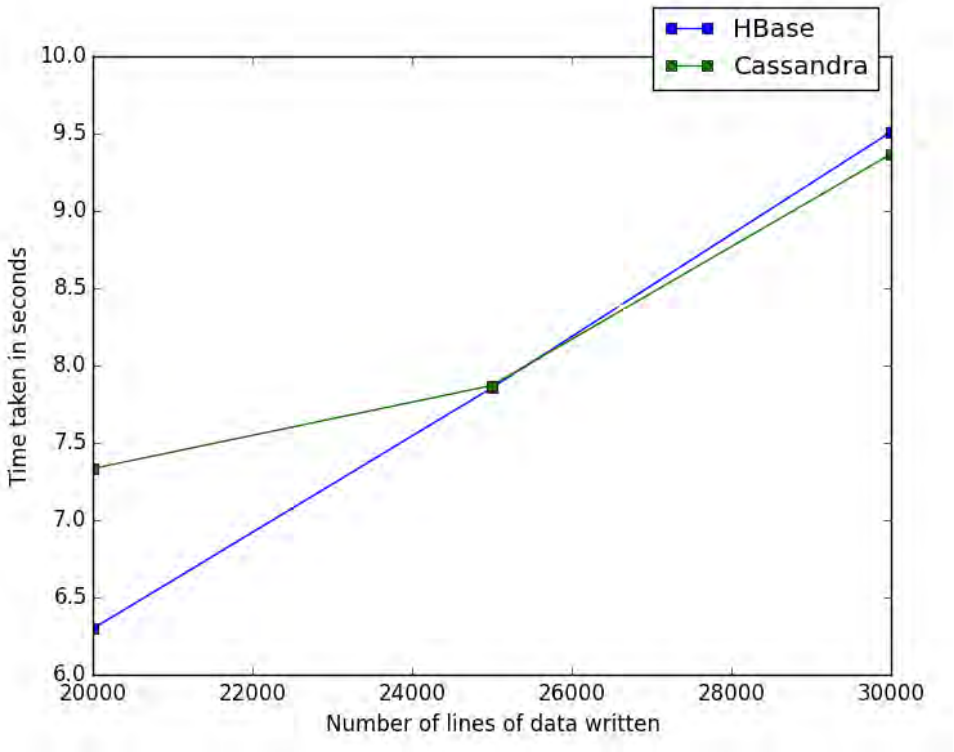

Figure 47: Comparing write performance of HBase and Cassandra